Revealing the arcane: an introduction to the art of stochastic volatility models

Tsyplakov, Alexander
Novosibirsk State University, Economics Department

28. September 2010

Revealing the arcane: an introduction to the art of stochastic volatility models

Alexander Tsyplakov

Novosibirsk State University, Economics Department

September 28, 2010

Abstract

This essay is aimed to provide a straightforward and sufficiently accessible demonstration of some known procedures for stochastic volatility model. It reviews the important related concepts, gives informal derivations of the methods and can be useful as a cookbook for a novice. The exposition is confined to classical (non-Bayesian) framework and discrete-time formulations.

1 Stochastic volatility modeling preliminaries

1.1 Introduction

A well-known phenomenon for financial time series is volatility clustering. The phenomenon can be accounted for by GARCH which—with its various modifications—is the most popular model of volatility (for an early overview see Bollerslev et al. (1994)). However, a more natural and conceptually simple model of volatility is probably the model of autoregressive stochastic volatility (ARSV or simply SV). Unlike GARCH, log-volatility is modeled as a first-order autoregression (see below). Similarly to GARCH, stochastic volatility model can be applied to various financial time series like stock prices or exchange rates.

We illustrate our discussion of stochastic volatility modeling with examples. Here the two real-data examples are introduced.

Example 1 (daily RTS stock market index, 1996–2009). RTSI is a stock market index of RTS ("Russian Trading System") stock exchange. It is "the main benchmark for the Russian securities industry and is based on the Exchange's 50 most liquid and capitalized shares".[1] We apply stochastic volatility model to continuously compounded returns computed from the daily RTSI close data. The returns are defined as $y_t = (\ln RTSI_t - \ln RTSI_{t-1}) \times 100$. The length of the series is $T = 3494$ observations.

Example 2 (daily pound/dollar exchange rates from October 1981 to June 1985). Next dataset is a series of weekdays close exchange rates.[2] The data we use are $y_t = (\ln E_t - \ln E_{t-1}) \times 100$, where E_t is the exchange rate. The length of the series is $T = 946$. The dataset initially appeared in an empirical application in Harvey et al. (1994). Subsequently it was analyzed extensively in the literature on stochastic volatility and states-space models.[3]

[1] See `http://www.rts.ru/`.

[2] The data can be found at `http://www.estima.com/textbooks/durkoop.zip`, `http://www.ssfpack.com/dkbook/dkdata/sv.dat` or `http://www.nuffield.ox.ac.uk/users/shephard/EXCH.ZIP`. The series is also distributed with popular EViews econometric program as `svpdx.dat`.

[3] For example, Shephard & Pitt (1997), Kim et al. (1998), Durbin & Koopman (2000), Meyer & Yu (2000), Durbin & Koopman (2001), Meyer et al. (2003), Davis & Rodriguez-Yam (2005), Liesenfeld & Richard (2006)

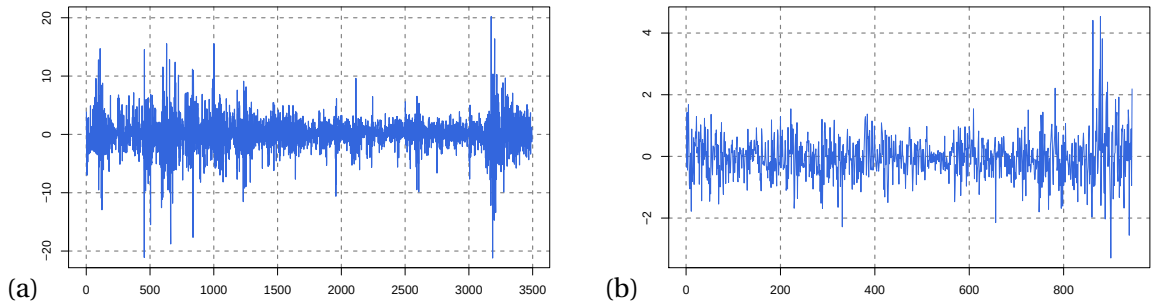

Figure 1: *(a) RTSI daily returns, 1996–2009, (b) £/$ daily rates of change, October 1981—June 1985.*

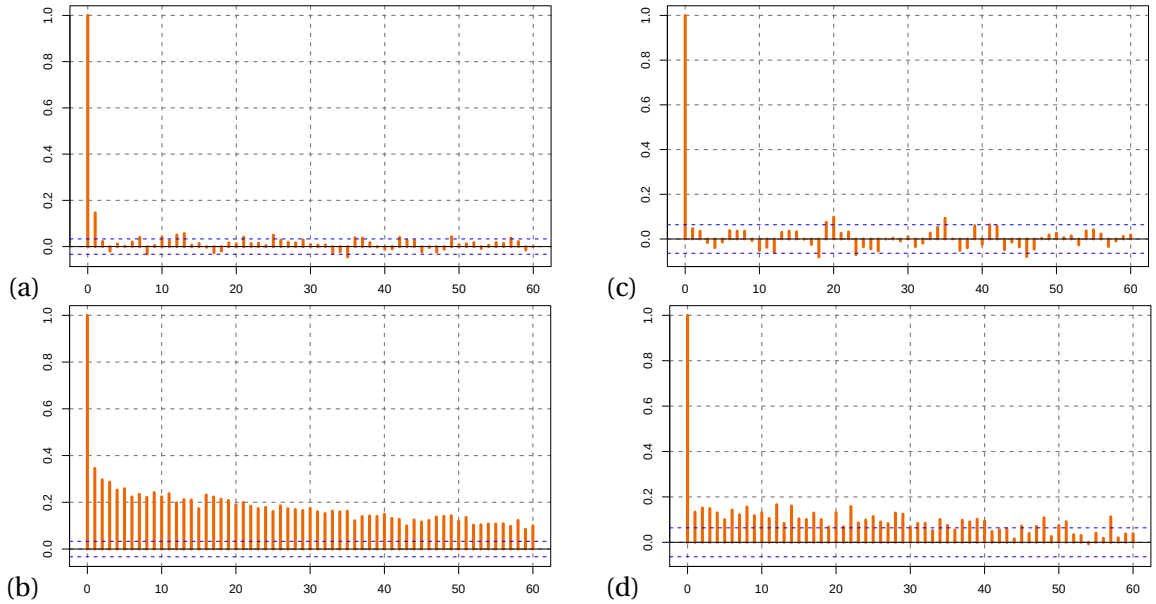

Figure 2: *(a) RTSI, correlogram, (b) RTSI, correlogram for absolute values, (c) £/$, correlogram, (d) £/$, correlogram for absolute values.*

Both example series (Figure 1(a), (b)) do not show strong autocorrelation. This can be seen from their correlograms (Figure 2(a), (c)). RTSI series has significant, but not very large first-order autocorrelation. However, variance of the two series changes over time leading to volatility clusters. For instance, for the exchange rates volatility is larger at the end of the period. This effect can be measured by the autocorrelation functions of $|y_t|$, y_t^2 or $\ln y_t^2$. Figure 2(b), (d) shows correlograms of absolute values $|y_t|$. Serial correlation is quite significant. This justifies the use of volatility modeling.

The origins of the model are not very clear. Possibly, the model was a very natural one and several researchers came to the idea independently. Discrete-time stochastic volatility models which we discuss[4] can be viewed as approximations to continuous-time models developed in mathematical finance literature.

Some of the early uses of the model can be found in Taylor (1982), Taylor (1986), Scott (1987), Hull & White (1987), Nelson (1988). Several pioneering papers on the subject are collected in Shephard (2005).

Stochastic volatility modeling is an active research area. Moreover, SV model is a popular showcase example in the flourishing literature on non-linear non-Gaussian state-space models, hidden Markov models and other related subjects. Therefore it is not possible to cover all the methods and ideas which are connected to SV model. Our task in this essay is somewhat limited. We are

[4]Continuous-time stochastic volatility models are reviewed in Ghysels et al. (1996) and Shephard & Andersen (2009).

trying to make SV modeling more accessible by collecting in one place several useful instruments for a practitioner to start with.

1.2 Basic SV model

SV model based on first-order autoregression (Markov chain) can be written as[5]

$$\begin{aligned} y_t &= \sigma_\xi \xi_t \exp(h_t/2), \\ h_t &= \delta h_{t-1} + \sigma_\eta \eta_t. \end{aligned} \tag{1}$$

Here h_t is scaled log-volatility (conditional variance[6] of y_t for this model is given by $\sigma_t^2 = \sigma_\xi^2 \exp(h_t)$ if $\mathsf{Var}\,\xi_t = 1$). It is assumed that scale parameters σ_ξ and σ_η are positive and that log-volatility autoregressive coefficient $|\delta| < 1$ (close to plus unity in applications). Disturbances in the basic SV model are assumed to be two independent series of Gaussian white noise

$$\xi_t \sim \mathcal{N}(0,1) \qquad \text{and} \qquad \eta_t \sim \mathcal{N}(0,1).$$

The model is often called *the* stochastic volatility model as it is the most intensively studied model of the SV class of models.

In what follows $\mathbf{y} = (y_1, ..., y_T)$ is a vector of observations, $\mathbf{h} = (h_1, ..., h_T)$ is a vector of unobserved volatility process and $\boldsymbol{\theta} = (\sigma_\xi, \delta, \sigma_\eta)$ is a parameters vector.

Example 3 (simulation example). We take $\delta = 0.98$, $\sigma_\eta = 0.2$, $\sigma_\xi = 1$ and $T = 500$ and simulate SV process. One realization (of both y_t and $\sigma_t^2 = \sigma_\xi^2 \exp(h_t)$) is shown in Figure 3. It can be seen that the regions of higher σ_t^2 correspond to more dispersed y_t while the regions of lower σ_t^2 correspond to less dispersed y_t. For these parameters the coefficient of variation of conditional variance, defined as[7]

$$CV = \frac{\sqrt{\mathsf{Var}\,\sigma_t^2}}{\mathsf{E}\,\sigma_t^2} = \frac{\sqrt{\mathsf{Var}[\exp(h_t)]}}{\mathsf{E}[\exp(h_t)]} = \sqrt{\exp\left(\sigma_\eta^2/(1-\delta^2)\right) - 1},$$

is 1.32 which is rather high, but is realistic for financial time series. The coefficient of variation measures how volatile is volatility. When CV is close to zero the volatility is almost constant.

It was suggested that SV-type models can provide a more adequate description of the behavior of many time series than GARCH-type models. The reason is that in a SV-type model volatility is not determined functionally by the lagged disturbances of the mean equation. Instead, it is modeled as a separate stochastic process driven by its own disturbances η_t. As a result for SV process (unlike GARCH-type process) next period volatility is not fully known (h_t cannot be forecasted exactly given information available at time $t-1$). Yet SV-type models are not as popular in empirical research as GARCH-type models, which is explained by the difficulties with statistical analysis of the former. In 1.3 we discuss the roots of this problem.

---

[5] Alternatively we could work with

$$\begin{aligned} y_t &= \xi_t \exp(h_t/2), \\ h_t - \omega &= \delta(h_{t-1} - \omega) + \sigma_\eta \eta_t. \end{aligned}$$

The equation for h_t can also be written as $h_t = \omega + \delta h_{t-1} + \sigma_\eta \eta_t$. These specifications are equivalent to (1).

[6] The term "conditional variance" is ambiguous for an SV model (unlike GARCH). By conditional variance here and below we mean the variance of y_t conditional on h_t and previous history $y_{t-1}, h_{t-1}, y_{t-2}, h_{t-2}, \ldots$ For the basic SV model (1) it is the same as the variance of y_t conditional on h_t. It is clearly not the same as the variance of y_t conditional on $y_{t-1}, y_{t-2}, \ldots$

[7] The expressions for the moments which are needed for deriving the coefficient of variation formula can be found in Appendix C. When CV is small it is approximately equal to the unconditional standard deviation of h_t which is $\sqrt{\mathsf{Var}\,h_t} = \sigma_\eta/\sqrt{1-\delta^2}$.

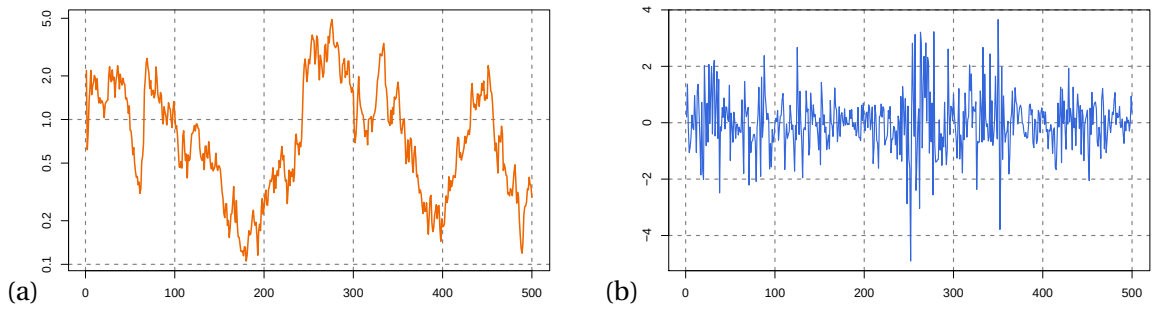

Figure 3: *A realization of stochastic volatility process with* $\delta = 0.98$, $\sigma_\eta = 0.2$, $\sigma_\xi = 1$ *and* $T = 500$*; (a) conditional variance* σ_t^2 *(logarithmic scale), (b)* y_t.

There are many different ways to estimate SV models. Below we focus on several methods of approximating the likelihood function. Given an estimate of the likelihood function one can use well-known optimization techniques[8] (like the quasi-Newton BFGS algorithm with numerical first derivatives[9] or the derivatives-free Nelder–Mead simplex-reflection algorithm) to maximize the obtained function with respect to parameters $\boldsymbol{\theta}$. The method of moments approach which can deliver feasible initial estimates of parameters is also discussed. Broto & Ruiz (2004) and Jungbacker & Koopman (2009) give a survey of estimation methods.

1.3 SV model as a model with unobserved components

Many applied statistical models are stated in terms of disturbances and parameters. If $\mathbf{u}$ is a $N \times 1$ vector of disturbances and $\boldsymbol{\theta}$ is a $m \times 1$ vector of parameters then it is assumed that the dependent variable $\mathbf{y}$ is a $n \times 1$ vector which is generated according to some known mapping $\mathscr{F}$: $\mathbf{y} = \mathscr{F}(\mathbf{u};\boldsymbol{\theta})$. Probabilistic assumptions are made in terms of $\mathbf{u}$, rather than in terms of $\mathbf{y}$. However, by definition $\mathbf{u}$ is not directly observed. Instead, $\mathbf{y}$ is observed. In some popular models $\mathscr{F}$ specifies a one-to-one mapping between $\mathbf{u}$ and $\mathbf{y}$ so that $\mathbf{u}$ can be obtained indirectly given some vector of parameters $\boldsymbol{\theta}$. For example, for the classical linear regression $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$.

In many models information about $\mathbf{u}$ is partially lost. For example, it can be that $N > n$, which means that a one-to-one mapping between $\mathbf{u}$ and $\mathbf{y}$ can not exist. For some models $\mathbf{u}$ can be partitioned as $\mathbf{u} = (\boldsymbol{\varepsilon}, \boldsymbol{\eta})$ where $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector such that $\mathbf{y} = \mathscr{F}(\boldsymbol{\varepsilon}, \boldsymbol{\eta}; \boldsymbol{\theta})$ specifies a one-to-one mapping between $\boldsymbol{\varepsilon}$ and $\mathbf{y}$ given $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$. Here $\boldsymbol{\eta}$ is a $(N-n) \times 1$ vector of unobserved components (or latent variables). To analyze this kind of models when $N - n$ is small it can be convenient to throw away information about the probabilistic properties of $\boldsymbol{\eta}$. Two common approaches are:

- assigning $\boldsymbol{\eta}$ some reasonable values (like expectations $\mathsf{E}\,\boldsymbol{\eta}$),
- treating $\boldsymbol{\eta}$ as parameters and estimating them together with regular parameters $\boldsymbol{\theta}$.

For example, in the MA(1) model $y_t = u_t + \mu u_{t-1}$ one can take $u_0 = 0$ and then calculate $u_1, \ldots, u_n$ recursively from $y_1, \ldots, y_n$: $u_t = y_t - \mu u_{t-1}$. In the GARCH model prehistoric values ε_t^2, σ_t^2 for $t < 1$ are commonly replaced by the unconditional variance.

However, if $N - n$ is not small such a loss of information can be inadmissible. Moreover, if $N - n$ is of the same order as n then throwing away information is of no help. This is the case with the SV model because one observable series y_t is determined by two disturbance series, ε_t and η_t, so that $N = 2n$. Hence the difficulties in estimation of SV model compared to oft-used GARCH.

[8] We do not discuss optimization algorithms here. See the literature on numerical optimization like Nocedal & Wright (2006).

[9] Some of the methods discussed can be used to get analytical derivatives of the approximate likelihood function. However, finding needed analytical derivatives can be an intricate problem so we will not explore the possibility in this essay.

In general, one has to deduce probabilistic properties of $\mathbf{y}$ from the assumptions about probabilistic properties of $\mathbf{u}$. For the generalized method of moments (GMM) one needs to obtain moment conditions on $\mathbf{y}$. For the method of maximum likelihood the probability density function $f(\mathbf{y}|\boldsymbol{\theta})$ of the observable data $\mathbf{y}$ is needed. Obtaining $f(\mathbf{y}|\boldsymbol{\theta})$ in general needs integration. For some models the integration can be done analytically to yield a closed-form expression. For other models like SV this is unfeasible.

One eminent model for which obtaining $f(\mathbf{y}|\boldsymbol{\theta})$ is straightforward is the Gaussian linear model. Assume that $\mathbf{u}$ has a multivariate normal distribution $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that the link between $\mathbf{y}$ and $\mathbf{u}$ is given by a linear (affine) function

$$\mathbf{y} = \mathbf{A}\mathbf{u} + \mathbf{b}. \tag{2}$$

Here $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{A}$, $\mathbf{b}$ can all depend non-linearly on $\boldsymbol{\theta}$. By the properties of multivariate normal distribution $\mathbf{y}$ is also multivariate normal

$$\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}).$$

Its log-density (log-likelihood function) is

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \ln f(\mathbf{y}|\boldsymbol{\theta}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}\right| - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T})^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}). \tag{3}$$

The conditional distribution $\mathbf{u}|\mathbf{y}$ summarizes information on $\mathbf{u}$ which can be inferred by observing $\mathbf{y}$. This conditional distribution is also multivariate normal:

$$\mathbf{u}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T})^{-1}(\mathbf{y} - \mathbf{b}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T})^{-1}\mathbf{A}\boldsymbol{\Sigma}).$$

Mean of the conditional distribution $\bar{\mathbf{u}}(\mathbf{y}) = \mathsf{E}(\mathbf{u}|\mathbf{y}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T})^{-1}(\mathbf{y} - \mathbf{b})$ is called the smoothed value of $\mathbf{u}$. It is the best mean-square predictor of $\mathbf{u}$ based on $\mathbf{y}$.

There is at least one weak point in this reasoning. The matrix $\boldsymbol{\Sigma}$ is $N \times N$, the matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}$ is $n \times n$. These can be quite huge in some financial applications. Time series of length $n = 10000$ leading to 10000×10000 matrices are not that uncommon nowadays.

The Linear Gaussian state-space models are special cases of the linear Gaussian models. They enable one to use low-dimensional recursions for evaluating likelihood functions. A well-known algorithm for doing this is Kalman filter[10].

Let us return to the SV model. In this model it is not possible to derive the distribution of $\mathbf{y}$ from the distributions of ξ_t and η_t in a closed form. MLE is a natural method for estimating the SV model, because the distributions of disturbances are known exactly (given parameters). However, the knowledge of the distributions of disturbances cannot immediately give the knowledge of the distribution of the observable data $\mathbf{y}$.

SV models belong to the class of non-linear non-Gaussian state-space models. The log-volatility component h_t is called the unobservable (latent, hidden) state of the system at time t. Below we treat h_t as unobservable components instead of corresponding disturbances η_t. This has some advantages in the case of state-space models.

The likelihood function is defined as $L(\boldsymbol{\theta}; \mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})$. For the SV model it cannot be expressed in a closed form. In the theory likelihood function can be found from $f(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})$ by integrating out $\mathbf{h}$. That is, it can be expressed by a multidimensional integral

$$f(\mathbf{y}|\boldsymbol{\theta}) = \int f(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})\, d\mathbf{h} = \int f(\mathbf{y}|\mathbf{h}, \boldsymbol{\theta}) f(\mathbf{h}|\boldsymbol{\theta})\, d\mathbf{h}.$$

The joint distribution of $\mathbf{y}$ and $\mathbf{h}$ described by density $f(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})$ is called distribution of complete data. "Complete data" means the data on both observable $\mathbf{y}$ and unobservable $\mathbf{h}$. Both $f(\mathbf{y}|\mathbf{h}, \boldsymbol{\theta})$

[10] See Commandeur & Koopman (2007), Durbin & Koopman (2001) and Harvey & Proietti (2005) on state-space models and Kalman filter.

and $f(\mathbf{h}|\boldsymbol{\theta})$ (and thus $f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$) are known for the basic SV model (see below). However, the integral cannot be calculated analytically[11]. Consequently, one needs to use numerical integration to obtain $L(\boldsymbol{\theta};\mathbf{y})$. Difficulties with devising and programming of efficient algorithms and substantial computational costs lead to low popularity of SV models in applied areas. However, as computers become faster and new methods are developed the use of SV modeling increases.

For future exposition we introduce the terminology which can often be found in the SV literature. For a given vector of parameters $\boldsymbol{\theta}$ one can consider various (marginal, joint, conditional) distributions of $\mathbf{y}$ and $\mathbf{h}$. For the SV model the marginal distribution of $\mathbf{h}$ is known (given $\boldsymbol{\theta}$). When observing the data $\mathbf{y}$ we obtain some additional information on the value of $\mathbf{h}$. This is summarized by the conditional distribution $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$. In Bayesian terms[12] $\mathbf{h}|\boldsymbol{\theta}$ is the prior distribution of unobserved $\mathbf{h}$ (beliefs on $\mathbf{h}$ held before the arrival of new information) and $\mathbf{h}|\mathbf{y},\boldsymbol{\theta}$ is the posterior distribution of $\mathbf{h}$ (beliefs on $\mathbf{h}$ held after obtaining the new information $\mathbf{y}$).

An important fact is that the posterior density is proportional to the density of complete data (both considered as functions of $\mathbf{h}$ for some given $\mathbf{y}$) where the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ provides the proportionality coefficient:

$$f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = \frac{f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta})}. \tag{4}$$

This proportionality is the key to some methods described below. First, it turns out that a good approximation for $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ can provide a good estimate of the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$. Second, the distribution of $\mathbf{h}|\mathbf{y},\boldsymbol{\theta}$ can itself be of interest for the various tasks of smoothing, filtering and forecasting.

1.4 Various densities for SV model

Here we write out densities for the basic SV model (1) which are useful for an (approximate) maximum likelihood estimation.

Consider the model (1). Let $\Omega_t = (y_1,\dots,y_t,h_1,\dots,h_t)$ be the history of SV process until time t. The distribution of the complete data $\mathbf{y},\mathbf{h}$ corresponding to parameters $\boldsymbol{\theta}$ is given by the density

$$f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{h},\boldsymbol{\theta})f(\mathbf{h}|\boldsymbol{\theta}) = \prod_{t=1}^{T} f(y_t|h_t,\Omega_{t-1},\boldsymbol{\theta}) \prod_{t=1}^{T} f(h_t|\Omega_{t-1},\boldsymbol{\theta}).$$

Here $f(y_t|h_t,\Omega_{t-1},\boldsymbol{\theta})$ is the density of $\mathcal{N}(0,\sigma_\xi^2 e^{h_t})$, $f(h_t|\Omega_{t-1},\boldsymbol{\theta})$ is the density of $\mathcal{N}(\delta h_{t-1},\sigma_\eta^2)$. The density $f(h_1|\Omega_0,\boldsymbol{\theta}) = f(h_1|\boldsymbol{\theta})$ is a special case. Stationarity of the AR(1) process describing h_t implies that $h_1|\boldsymbol{\theta} \sim \mathcal{N}(0,\sigma_\eta^2/(1-\delta^2))$. We see that for the basic SV model (1) the component densities simplify to $f(y_t|h_t,\Omega_{t-1},\boldsymbol{\theta}) = f(y_t|h_t,\boldsymbol{\theta})$ and $f(h_t|\Omega_{t-1},\boldsymbol{\theta}) = f(h_t|h_{t-1},\boldsymbol{\theta})$ so that

$$f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) = \prod_{t=1}^{T} f(y_t|h_t,\boldsymbol{\theta}) f(h_1|\boldsymbol{\theta}) \prod_{t=2}^{T} f(h_t|h_{t-1},\boldsymbol{\theta}). \tag{5}$$

The component log-densities are

$$\ln f(y_t|h_t,\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi\sigma_\xi^2) - \frac{h_t}{2} - \frac{y_t^2}{2\sigma_\xi^2 e^{h_t}},$$

$$\ln f(h_1|\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi\sigma_\eta^2) + \frac{1}{2}\ln(1-\delta^2) - \frac{1-\delta^2}{2\sigma_\eta^2} h_1^2$$

[11] Shephard (1994) proposed a SV-type model for which this integral can be calculated. His model contains a random walk in volatility equation and thus similar to model (1) with $\delta = 1$.

[12] Do not be misled by the similarity with the terminology used in a Bayesian inference on $\boldsymbol{\theta}$. For the Bayesian approach p.d.f. $f(\boldsymbol{\theta})$ describes the prior distribution of $\boldsymbol{\theta}$ and $f(\boldsymbol{\theta}|\mathbf{y})$ describes the posterior distribution of $\boldsymbol{\theta}$ given some data $\mathbf{y}$.

and

$$\ln f(h_t|h_{t-1},\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi\sigma_\eta^2) - \frac{1}{2\sigma_\eta^2}(h_t - \delta h_{t-1})^2.$$

Using these we write the log-density of complete data:

$$\begin{aligned}\ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) = &-\frac{T}{2}\ln(2\pi\sigma_\xi^2) - \frac{1}{2}\sum_{t=1}^{T}\left(h_t + \frac{y_t^2}{\sigma_\xi^2 e^{h_t}}\right)\\ &-\frac{T}{2}\ln(2\pi\sigma_\eta^2) + \frac{1}{2}\ln(1-\delta^2) - \frac{1}{2\sigma_\eta^2}\left[(1-\delta^2)h_1^2 + \sum_{t=2}^{T}(h_t - \delta h_{t-1})^2\right]\end{aligned} \tag{6}$$

2 Estimation using a Gaussian approximation for $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$

In this essay we consider only Gaussian approximations for $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$. Such approximations are the simplest and most widely used. Other approximations (for example, those employing the Student's t distribution) can be treated by analogy with Gaussian ones.

If $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ is a Gaussian approximating density then $\ln g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ is quadratic in $\mathbf{h}$ by the properties of the multivariate normal distribution. This allows to find $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ without knowing $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$. By writing $\ln f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = \ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) - \ln f(\mathbf{y}|\boldsymbol{\theta})$ one can see that only the log-density of the complete data $\ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$ is needed to find an approximation, because the log-likelihood $\ln f(\mathbf{y}|\boldsymbol{\theta})$ does not depend on $\mathbf{h}$.

Let $\ln f_a(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$ be some approximation to $\ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$ which is quadratic in $\mathbf{h}$. Such an approximation can be written as

$$\ln f_a(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) = u(\mathbf{y}) + \mathbf{h}^\top\mathbf{v}(\mathbf{y}) - \frac{1}{2}\mathbf{h}^\top\mathbf{W}(\mathbf{y})\mathbf{h},$$

where $u(\mathbf{y})$, 1×1, $\mathbf{v}(\mathbf{y})$, $T\times 1$, $\mathbf{W}(\mathbf{y})$, $T\times T$ are some functions of $\mathbf{y}$ only. We assume that $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ is multivariate normal with mean $\bar{\mathbf{h}}(\mathbf{y})$ and covariance matrix $\boldsymbol{\Sigma}(\mathbf{y})$. Then the log-density is given by

$$\ln g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}(\mathbf{y})\right| - \frac{1}{2}(\mathbf{h}-\bar{\mathbf{h}}(\mathbf{y}))^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y})(\mathbf{h}-\bar{\mathbf{h}}(\mathbf{y})).$$

Equating the coefficients for the second-order and first-order terms we obtain $\boldsymbol{\Sigma}(\mathbf{y}) = \mathbf{W}^{-1}(\mathbf{y})$, $\bar{\mathbf{h}}(\mathbf{y}) = \mathbf{W}^{-1}(\mathbf{y})\mathbf{v}(\mathbf{y})$. Thus,

$$\begin{aligned}\ln g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) &= -\frac{T}{2}\ln(2\pi) + \frac{1}{2}\ln\left|\mathbf{W}(\mathbf{y})\right| - \frac{1}{2}(\mathbf{h}-\mathbf{W}^{-1}(\mathbf{y})\mathbf{v}(\mathbf{y}))^\top\mathbf{W}(\mathbf{y})(\mathbf{h}-\mathbf{W}^{-1}(\mathbf{y})\mathbf{v}(\mathbf{y}))\\ &= -\frac{T}{2}\ln(2\pi) + \frac{1}{2}\ln\left|\mathbf{W}(\mathbf{y})\right| - \frac{1}{2}\mathbf{v}(\mathbf{y})^\top\mathbf{W}^{-1}(\mathbf{y})\mathbf{v}(\mathbf{y}) + \mathbf{h}^\top\mathbf{v}(\mathbf{y}) - \frac{1}{2}\mathbf{h}^\top\mathbf{W}(\mathbf{y})\mathbf{h}\end{aligned}$$

(Obviously, this approximation will work only if $\mathbf{W}(\mathbf{y})$ is symmetric and positive definite).

Then an approximation for $\ln f(\mathbf{y}|\boldsymbol{\theta})$ is given by

$$\ln f_a(\mathbf{y}|\boldsymbol{\theta}) = \ln f_a(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) - \ln g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$$

following the analogy with the

$$\ln f(\mathbf{y}|\boldsymbol{\theta}) = \ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) - \ln f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$$

identity. So the approximate log-likelihood function is

$$\ell_a(\boldsymbol{\theta};\mathbf{y}) = \ln f_a(\mathbf{y}|\boldsymbol{\theta}) = u(\mathbf{y}) + \frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{W}(\mathbf{y})\right| + \frac{1}{2}\mathbf{v}(\mathbf{y})^\top\mathbf{W}^{-1}(\mathbf{y})\mathbf{v}(\mathbf{y}).$$

The idea of a Gaussian approximation is very general and it has to be elaborated upon to make it applicable to the case of the SV model. For the basic SV model distribution of $\mathbf{h}|\boldsymbol{\theta}$ is already a multivariate normal one so that $\ln f(\mathbf{h}|\boldsymbol{\theta})$ is quadratic in $\mathbf{h}$. Consequently, we only need quadratic approximations of $\ln f(y_t|h_t,\boldsymbol{\theta})$ with respect to h_t.

Suppose that

$$\ln f(y_t|h_t,\boldsymbol{\theta}) = A_t + A_t^0 h_t + A_t^{00} h_t^2 + R_t(h_t; y_t, \boldsymbol{\theta}),$$

where A_t, A_t^0, A_t^{00} are coefficients.[13] We replace $\ln f(y_t|h_t,\boldsymbol{\theta})$ in (5) by

$$\ln f_a(y_t|h_t,\boldsymbol{\theta}) = A_t + A_t^0 h_t + A_t^{00} h_t^2 \tag{7}$$

to get a quadratic approximation for $\ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$. Combining powers of $h_1,\ldots,h_T$ we can write the approximation as[14]

$$\ln f_a(\mathbf{y},\mathbf{h}) = \sum_{t=1}^{T} (B_t^0 h_t + B_t^{00} h_t^2 + B_t^{01} h_t h_{t-1}) + B. \tag{8}$$

The formulas connecting coefficients B_t^0, B_t^{00} and B_t^{01} with A_t^0 and A_t^{00} are given in Appendix A.

Then quadratic approximation for log-density of $\mathbf{h}|\mathbf{y}$ has a form similar to (8):

$$\ln g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = \sum_{t=1}^{T} (B_t^0 h_t + B_t^{00} h_t^2 + B_t^{01} h_t h_{t-1}) + const.$$

It is possible to decompose a multivariate distribution $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ into a chain of conditional univariate distributions as follows:

$$g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = \prod_{t=1}^{T} g(h_t|h_1,\ldots,h_{t-1},\mathbf{y},\boldsymbol{\theta}).$$

Since only terms with $h_t h_{t-k}$ for $k=0$ and $k=1$ are present, the decomposition is simply

$$g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = \prod_{t=1}^{T} g(h_t|h_{t-1},\mathbf{y},\boldsymbol{\theta}),$$

where $h_t|h_{t-1},\mathbf{y} \sim \mathcal{N}(K_t + L_t h_{t-1}, M_t)$, $t = 1,\ldots,T$ for some coefficients K_t, L_t, M_t (with $L_1 = 0$). This is a time-inhomogeneous Markov chain or AR(1) process. The elementary univariate densities are given by

$$\ln g(h_t|h_{t-1},\mathbf{y},\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi M_t) - \frac{1}{2M_t}(h_t - K_t - L_t h_{t-1})^2. \tag{9}$$

Approximate Gaussian log-density is the sum of logarithms of these elementary densities:

$$\ln g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\ln(M_t) - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{M_t}(h_t - K_t - L_t h_{t-1})^2. \tag{10}$$

From this we obtain an approximate log-likelihood (see Appendix A):

$$\ell_a(\boldsymbol{\theta};\mathbf{y}) = B + \frac{T}{2}\ln(2\pi) + \frac{1}{2}\sum_{t=1}^{T}\ln(M_t) + \frac{1}{2}\sum_{t=1}^{T}\frac{K_t^2}{M_t} \tag{11}$$

[13] The notation for coefficients is a bit strange at first glance, but it is mnemonic and allows to economize on symbols.

[14] We accept a non-strict notation for the terms corresponding to $t=1$ (and $t=T$). Any term containing h_{t-1} for $t=1$ (or h_{t+1} for $t=T$) should be removed and the corresponding coefficient should be equated to zero. Also $f(h_t|h_{t-1})$ for $t=1$ is just $f(h_1)$.

or

$$\ell_a(\boldsymbol{\theta};\mathbf{y}) = \sum_{t=1}^{T} A_t - T\ln\sigma_\eta + \frac{1}{2}\ln(1-\delta^2) + \frac{1}{2}\sum_{t=1}^{T}\ln(M_t) + \frac{1}{2}\sum_{t=1}^{T}\frac{K_t^2}{M_t}. \tag{12}$$

One can be also interested in an estimate of $\mathbf{h}$ given the observable data $\mathbf{y}$. It is easy to compute the mean $\bar{\mathbf{h}} = \bar{\mathbf{h}}(\mathbf{y})$ of an approximating distribution $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ (which is also the median and the mode by the properties of multivariate normal distributions). This "smoothed" $\mathbf{h}$ can be obtained by the following recursion

$$\bar{h}_1 = K_1, \qquad \bar{h}_t = K_t + L_t\bar{h}_{t-1}, \quad t = 2,\ldots,T. \tag{13}$$

Similarly estimates of the variance of h_t are given by

$$s_{h,1}^2 = M_1, \qquad s_{h,t}^2 = M_t + L_t^2 s_{h,t-1}^2, \quad t = 2,\ldots,T. \tag{14}$$

Assuming the log-normal distribution $e^{h_t} \sim \mathscr{LN}(\bar{h}_t, s_{h,t}^2)$ we can also obtain approximate the smoothed conditional variance as[15]

$$\mathsf{E}[\sigma_\xi^2 e^{h_t}|\mathbf{y},\boldsymbol{\theta}] \approx \bar{\sigma}_t^2 = \sigma_\xi^2\exp(\bar{h}_t + s_{h,t}^2/2). \tag{15}$$

More generally, the chain of univariate Gaussian distributions $\mathscr{N}(K_t + L_t h_{t-1}, M_t)$ can be considered as a simple "smoother".[16]

3 Gaussian approximation for $\ln(\chi_1^2)$ and QML estimation

We can square y_t in (1) and take logarithms. Then

$$\ln(y_t^2) = \ln\sigma_\xi^2 + h_t + \ln(\xi_t^2).$$

Since ξ_t is standard normal it follows that $\ln(\xi_t^2) \sim \ln(\chi_1^2)$. The mean and variance of $\ln(\chi_1^2)$ distribution are[17] $\mathscr{C} \approx -1.27036$ and $\pi^2/2$. Thus, we can write this equation as

$$\ln(y_t^2) = \ln\sigma_\xi^2 + \mathscr{C} + h_t + \omega_t, \tag{16}$$

where $\omega_t = \ln(\xi_t^2) - \mathscr{C}$. This together with

$$h_t = \delta h_{t-1} + \sigma_\eta\eta_t$$

makes up a linear state-space model.[18] The only problem with it is that the error ω_t is not Gaussian. Consequently it is not possible to write out the exact likelihood function.

Harvey et al. (1994) suggest using the quasi maximum likelihood (QML) method to estimate the model (see also Scott (1987), Nelson (1988)). The QML method approximates the distribution of $\omega_t = \ln(\xi_t^2) - \mathscr{C}$ by $\mathscr{N}(0,\pi^2/2)$. Thereby the SV model is approximated by a linear Gaussian state-space model. The approximation is not very accurate, as $\ln(\chi_1^2)$ has a thick left tail and thin right tail (see Figure 4).

For another illustration of the approximation we turn to generated data.

Example 3 (continued). We take the realization of SV process from Figure 3. In Figure 5 both $h_t + \ln\sigma_\xi^2$ and $\ln(y_t^2) - \mathscr{C} = h_t + \ln\sigma_\xi^2 + \omega_t$ are plotted. The log-volatility $h_t + \ln\sigma_\xi^2$ is an AR(1) process while $\ln(y_t^2) - \mathscr{C}$ is an AR(1) plus noise process. The noise ω_t is not Gaussian which shows up in a disproportional number of "negative outliers" in the plot.

[15] Alternatively the smoothed value of conditional variance can be defined as $\bar{\sigma}_t^2 = \sigma_\xi^2\exp(\bar{h}_t)$ which corresponds to a geometric mean rather than an arithmetic mean.

[16] The method is equivalent to a more widely known Kalman smoother, but its computation omits an additional Kalman filtering step.

[17] More exactly, $\mathscr{C} = \psi(1/2) - \ln(1/2)$ where $\psi(\cdot)$ is the digamma function.

[18] It is also possible to rewrite this as the ARMA(1, 1) model for $\ln(y_t^2)$.

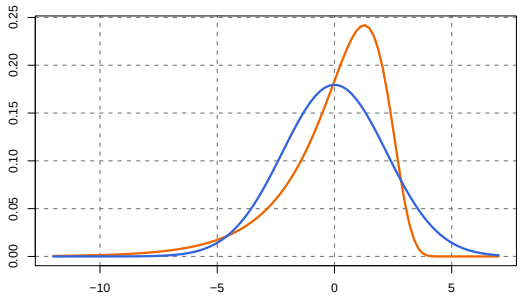

Figure 4: *Densities of* $\ln(\chi_1^2) - \mathscr{C}$ *and* $\mathcal{N}(0, \pi^2/2)$ *distributions compared.*

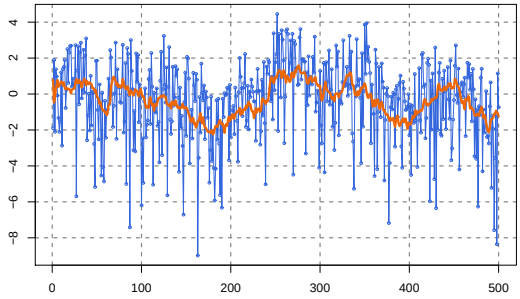

Figure 5: *Log-volatility* $h_t + \ln\sigma_\xi^2$ *and* $\ln(y_t^2) - \mathscr{C}$ *for Example 3, illustration of QML.*

The quasi log-likelihood for $(\ln(y_1^2), \ldots, \ln(y_T^2))$ can be defined similarly to (3) in the linear model (2):

$$\ell_Q(\boldsymbol{\theta}) = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\mathbf{z}(\boldsymbol{\theta})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\mathbf{z}(\boldsymbol{\theta}).$$

Here $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix of $(\ln(y_1^2), \ldots, \ln(y_T^2))$ and $\mathbf{z}(\boldsymbol{\theta})$ consists of $z_t = \ln(y_t^2) - \ln\sigma_\xi^2 - \mathscr{C}$.

Harvey et al. (1994) employ Kalman the filter technique to do the calculations. Here we show how to obtain the QML estimates by assuming that $\ln(\xi_t^2)$ is approximately normally distributed without writing out the full Kalman filter equations.

We do not need the error component corresponding to ξ_t to have a zero mean so we write simply

$$\ln(y_t^2) = \ln\sigma_\xi^2 + h_t + \varepsilon_t$$

where $\varepsilon_t = \ln(\xi_t^2)$. The exact distribution of $\varepsilon_t = \ln(\xi_t^2)$ is given by the density function

$$f(\varepsilon_t) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}\varepsilon_t - \frac{1}{2}e^{\varepsilon_t}\right)$$

and is approximated by $\mathcal{N}(\mu_\varepsilon, \sigma_\varepsilon^2)$ where $\mu_\varepsilon = \mathscr{C} \approx -1.27036$ and $\sigma_\varepsilon^2 = \pi^2/2$. Thus,

$$\ln f(\ln(y_t^2)\,|\,h_t, \boldsymbol{\theta}) = \ln f(\varepsilon_t) = -\frac{1}{2}\ln(2\pi) + \frac{1}{2}\varepsilon_t - \frac{1}{2}e^{\varepsilon_t} \approx -\frac{1}{2}\ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}(\varepsilon_t - \mu_\varepsilon)^2$$

or

$$\ln f(\ln(y_t^2)\,|\,h_t, \boldsymbol{\theta}) \approx -\frac{1}{2}\ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}(\ln(y_t^2) - h_t - 2\ln\sigma_\xi - \mu_\varepsilon)^2$$

The link between densities of y_t and $\ln(y_t^2)$ (conditional on h_t) is given by

$$f(\ln(y_t^2)\,|\,h_t, \boldsymbol{\theta}) = f(y_t\,|\,h_t, \boldsymbol{\theta}) \cdot \big|y_t\big|.$$

So we can write

$$\ln f(y_t|h_t,\boldsymbol{\theta}) = \ln f(\ln(y_t^2)|h_t,\boldsymbol{\theta}) - \frac{\ln(y_t^2)}{2} \approx -\frac{1}{2}\ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}(\ln(y_t^2) - h_t - 2\ln\sigma_\xi - \mu_\varepsilon)^2 - \frac{\ln(y_t^2)}{2}$$
$$= -\frac{1}{2}\ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}(\ln(y_t^2) - 2\ln\sigma_\xi - \mu_\varepsilon)^2 - \frac{\ln(y_t^2)}{2} + \frac{1}{\sigma_\varepsilon^2}(\ln(y_t^2) - 2\ln\sigma_\xi - \mu_\varepsilon)h_t - \frac{h_t^2}{2\sigma_\varepsilon^2}.$$

In terms of (7) we have

$$A_t = -\frac{1}{2}\ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}(\ln(y_t^2) - 2\ln\sigma_\xi - \mu_\varepsilon)^2 - \frac{\ln(y_t^2)}{2},$$
$$A_t^0 = \frac{1}{\sigma_\varepsilon^2}(\ln(y_t^2) - 2\ln\sigma_\xi - \mu_\varepsilon),$$
$$A_t^{00} = -\frac{1}{2\sigma_\varepsilon^2}.$$

The QML estimates are obtained by maximizing (12) with respect to parameters $\boldsymbol{\theta}$.[19]

A practical difficulty with the method is that y_t^2 for real data can have an excessive proportion of observations which are close to zero (or equal to zero if the values are rounded or if holidays are not accounted for). For such observations (so-called inliers) $\ln(y_t^2)$ would assume large negative values (or would be undefined). To cope with the difficulty, one can simply truncate small values of y_t^2 by replacing y_t^2 with $\max\{y_t^2, \alpha s_y^2\}$ were α is a small positive number and s_y^2 is the sample mean of y_t^2 (e.g. see Sandmann & Koopman (1998)). Breidt & Carriquiry (1996) propose to replace y_t^2 with

$$\ln(y_t^2 + \lambda s_y^2) - \lambda s_y^2/(y_t^2 + \lambda s_y^2)$$

for a small positive λ. Their choice for λ is 0.005. They show that the transformation reduces the excess kurtosis and improves the performance of the QML estimator.

From QML we can obtain a smoothed value of $\mathbf{h}$. Suppose that K_t, L_t and M_t correspond to the QML approximation. Then we can use $\mathsf{E}(\mathbf{h}|\mathbf{y}) \approx \bar{\mathbf{h}}(\mathbf{y})$, where $\bar{\mathbf{h}}(\mathbf{y})$ is given by (13). This estimator is the best mean-square linear predictor of $\mathbf{h}$ in terms of $\{\ln(y_t^2)\}$.

By means of Kalman filter one can obtain a decomposition of the quasi log-likelihood function:

$$\ell_Q(\boldsymbol{\theta}) = \sum_{t=1}^{T} \ell_{Qt}(\boldsymbol{\theta}). \tag{17}$$

It can be demonstrated that for each t

$$\mathsf{E}[\nabla_{\boldsymbol{\theta}} \ell_{Qt}(\boldsymbol{\theta})] = \mathbf{0}.$$

This representation shows that the QML estimator can be viewed as a particular case of the generalized method of moments estimator. This suggests consistency and asymptotic normality of the QML estimator.

The most intricate aspect of the QML approach to SV modeling is obtaining the covariance matrix and the standard errors of QML estimates $\hat{\boldsymbol{\theta}}_Q$. We cannot just use the minus inverse Hessian $(-\hat{\mathbf{H}}_Q^{-1})$ of the quasi log-likelihood $\ell_Q(\boldsymbol{\theta})$, where

$$\hat{\mathbf{H}}_Q = \nabla_{\boldsymbol{\theta}}^2 \ell_Q(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_Q},$$

[19] It was suggested to include σ_ε^2 in $\boldsymbol{\theta}$ and estimate it along with the other parameters rather than fixing it at the known value $\sigma_\varepsilon^2 = \pi^2/2$. The purpose is to improve the small-sample properties of the estimates. See Jungbacker & Koopman (2009).

as an estimator of the covariance matrix which is usual for the maximum likelihood estimation. It is inconsistent. The literature on extremum estimators (including the literature on QML estimators; see White (1984)) suggests that asymptotic distribution of $\hat{\boldsymbol{\theta}}_Q$ is given by

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_Q - \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, (\mathscr{H}_Q^\infty)^{-1} \mathscr{I}_Q^\infty (\mathscr{H}_Q^\infty)^{-1}), \tag{18}$$

where $\mathscr{H}_Q^\infty$ is the asymptotic expected Hessian

$$\mathscr{H}_Q^\infty = \mathscr{H}_Q^\infty(\boldsymbol{\theta}) = \lim_{T\to\infty} \frac{1}{T} \mathscr{H}_Q^T(\boldsymbol{\theta}), \qquad \mathscr{H}_Q^T(\boldsymbol{\theta}) = \mathsf{E}[\nabla_{\boldsymbol{\theta}}^2 \ell_Q(\boldsymbol{\theta})]$$

and $\mathscr{I}_Q^\infty$ is the asymptotic information matrix

$$\mathscr{I}_Q^\infty = \mathscr{I}_Q^\infty(\boldsymbol{\theta}) = \lim_{T\to\infty} \frac{1}{T} \mathscr{I}_Q^T(\boldsymbol{\theta}),$$

$$\mathscr{I}_Q^T(\boldsymbol{\theta}) = \mathsf{Var}_\theta[\nabla_{\boldsymbol{\theta}} \ell_Q(\boldsymbol{\theta})] = \mathsf{E}_\theta[\nabla_{\boldsymbol{\theta}} \ell_Q(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^\intercal \ell_Q(\boldsymbol{\theta})].$$

For the genuine maximum likelihood we have the information matrix identity $\mathscr{I}_Q^T(\boldsymbol{\theta}) = -\mathscr{H}_Q^T(\boldsymbol{\theta})$ and its asymptotic variant $\mathscr{I}_Q^\infty = -\mathscr{H}_Q^\infty$. For QML this is no more true and we get a sandwich covariance matrix which is typical for misspecified models.

Several estimators of $\mathscr{H}_Q^\infty$ and $\mathscr{I}_Q^\infty$ are available. A straightforward (but computationally intensive) method is to use $\mathscr{H}_Q^T(\hat{\boldsymbol{\theta}}_Q)$ and $\mathscr{I}_Q^T(\hat{\boldsymbol{\theta}}_Q)$ where the expectations should be approximated by Monte Carlo simulations.

Another way is to use the "spectral" approximations to $\mathscr{H}_Q^T(\hat{\boldsymbol{\theta}}_Q)$ and $\mathscr{I}_Q^T(\hat{\boldsymbol{\theta}}_Q)$ which can be obtained analytically, but require a rather tedious derivation. See Appendix B for the final expressions without intermediate calculations. (The derivation is available from the author upon request.)

By passing to the limit in the spectral approximations one can obtain the analytical expressions for $\mathscr{H}_Q^\infty(\boldsymbol{\theta})$ and $\mathscr{I}_Q^\infty(\boldsymbol{\theta})$. This allows to use $\mathscr{H}_Q^\infty(\hat{\boldsymbol{\theta}}_Q)$ and $\mathscr{I}_Q^\infty(\hat{\boldsymbol{\theta}}_Q)$ as estimates of $\mathscr{H}_Q^\infty$ and $\mathscr{I}_Q^\infty$. Formulas for $\mathscr{H}_Q^\infty(\boldsymbol{\theta})$ and $\mathscr{I}_Q^\infty(\boldsymbol{\theta})$ are given in Ruiz (1994), but she uses a slightly different parametrization of the SV model.

Another way to estimate $\mathscr{I}_Q^\infty$ is to use (17) to write

$$\mathscr{I}_Q^T(\boldsymbol{\theta}) = \sum_{s=1}^T \sum_{t=1}^T \mathsf{E}(\nabla_{\boldsymbol{\theta}} \ell_{Qs}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^\intercal \ell_{Qt}(\boldsymbol{\theta})).$$

Taking into account this representation we can write the following asymptotic estimate:

$$\mathscr{I}_Q^\infty \approx \frac{1}{T} \sum_{s=1}^T \sum_{t=1}^T k\Big(\frac{|t-s|}{L+1}\Big) \nabla_{\boldsymbol{\theta}} \ell_{Qs}(\hat{\boldsymbol{\theta}}_Q) \nabla_{\boldsymbol{\theta}}^\intercal \ell_{Qt}(\hat{\boldsymbol{\theta}}_Q),$$

where $k(z)$ is a kernel function which is usually chosen in such a way that $k(0) = 1$ and $k(z) = 0$ for $|z| > 1$ and L is lag truncation parameter. A popular kernel[20] is the Bartlett kernel defined as

$$k(z) = \begin{cases} 1 - |z|, & |z| \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

This way of estimating $\mathscr{I}_Q^\infty$ is naturally complemented by a simple Hessian estimator of $\mathscr{H}_Q^\infty$:

$$\mathscr{H}_Q^\infty \approx \frac{1}{T} \hat{\mathbf{H}}_Q.$$

[20] See Andrews (1991) for a discussion of the estimator and examples of other popular kernels.

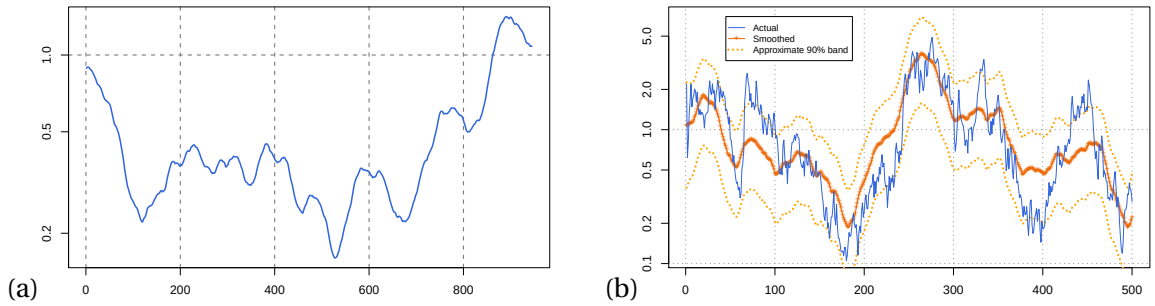

Figure 6: *(a) The smoothed value of conditional variance from QML estimates, Example 2, (b) the smoothed value, an approximate confidence band and the actual conditional variance, Example 3. (Logarithmic axes are used for the conditional variance).*

Table 1: *QML estimates*

	Example 2		Example 3			Example 3, simulation		
	estimates	std. err.	estimates	true values	std. err.	mean	RMSE	mean std.err.
δ	0.9889	0.0092	0.9732	0.9800	0.0209	0.9370	0.0844	0.0401
σ_η	0.0934	0.0345	0.1901	0.2000	0.0735	0.2709	0.1403	0.1037
σ_ξ	0.6654	0.0725	0.8036	1.000	0.1117	1.0349	0.2246	0.1354

The derivatives needed to obtain the estimates of covariance matrix can be evaluated numerically.

Example 2 (continued). We programmed[21] the QML method in the Ox programming language.[22] The approximate log-likelihood function was maximized using the BFGS algorithm implementation built-in in Ox. Figure 6(a) shows the smoothed value of the conditional variance $\bar{\sigma}_t^2$ based on the QML estimate for exchange rates (see (15) above). The left part of Table 1 shows the estimates and their standard errors (based on the "spectral" estimator of covariance matrix). Note that the proximity of the estimated δ to 1 where the quasi likelihood functions has a singularity can lead to serious distortions in the standard errors for short series.

Example 3 (continued). The central part of Table 1 shows the QML estimates for the realization of the SV process from Figure 3. The right part of the table reports the root mean squared errors (RMSE) for the QML estimator. The RMSEs were estimated from 1000 Monte Carlo simulations with the same true values of the parameters. Figure 6(b) compares the smoothed conditional variance based on the QML estimates $\bar{\sigma}_t^2$ with actual one. An approximate pointwise confidence band based on $\hat{\sigma}_\xi^2 \exp(\bar{h}_t \pm 1.64 s_{h,t})$ is also shown (see (13) and (14)) which would correspond to the 0.05 and 0.95 quantiles if the QML approximation for the posterior distribution were correct. Here $\hat{\sigma}_\xi^2$ is the QML estimate of σ_ξ^2 and $\bar{h}_t$, $s_{h,t}^2$, $\bar{\sigma}_t^2$ are given by (13), (14) and (15).

4 Quadratic expansion around the mode. Laplace's approximation

A natural method of finding a Gaussian approximation of $\ln f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ is to use the second-order Taylor expansion of $\lambda(\mathbf{h}) = \ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$ around some point $\mathbf{h}^*$:

$$\lambda(\mathbf{h}) \approx \nabla\lambda(\mathbf{h}^*)^\intercal(\mathbf{h}-\mathbf{h}^*) + \frac{1}{2}(\mathbf{h}-\mathbf{h}^*)^\intercal\nabla^2\lambda(\mathbf{h}^*)(\mathbf{h}-\mathbf{h}^*) + const,$$

where $\nabla\lambda(\mathbf{h})$ is the gradient and $\nabla^2\lambda(\mathbf{h})$ is the Hessian matrix of $\lambda(\mathbf{h})$.

[21] The source code of all the programs for this essay is available from the author.

[22] Doornik (2009). A free Ox Console version can be downloaded from `http://www.doornik.com/download.html`.

Recall that the log-density of $y_t|h_t,\boldsymbol{\theta}$ is

$$\ln f(y_t|h_t,\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi\sigma_\xi^2) - \frac{h_t}{2} - \frac{y_t^2}{2\sigma_\xi^2 e^{h_t}},$$

To get the quadratic approximation of $\ln f(y_t|h_t,\boldsymbol{\theta})$ as a function of h_t it is necessary to approximate e^{-h_t}. The second-order expansion of e^{-h_t} around h_t^* is given by

$$e^{-h_t} \approx e^{-h_t^*}\left(1 - h_t + h_t^* + \frac{1}{2}(h_t - h_t^*)^2\right).$$

Thus, we write

$$\begin{aligned}\ln f_a(y_t|h_t,\boldsymbol{\theta}) &= -\frac{1}{2}\ln(2\pi\sigma_\xi^2) - \frac{1}{2}\left(h_t + \tilde{y}_t^2\left(1 - h_t + h_t^* + \frac{1}{2}(h_t - h_t^*)^2\right)\right)\\ &= -\frac{1}{2}\ln(2\pi\sigma_\xi^2) - \frac{\tilde{y}_t^2}{2}\left(1 + h_t^* + \frac{1}{2}h_t^{*2}\right) + \left(\frac{\tilde{y}_t^2}{2}\left(1 + h_t^*\right) - \frac{1}{2}\right)h_t - \frac{\tilde{y}_t^2}{4}h_t^2,\end{aligned}$$

where

$$\tilde{y}_t^2 = \frac{y_t^2}{\sigma_\xi^2 e^{h_t^*}}.$$

In terms of (7) we have

$$\begin{aligned}A_t &= -\frac{1}{2}\ln(2\pi\sigma_\xi^2) - \frac{\tilde{y}_t^2}{2}\left(1 + h_t^* + \frac{1}{2}h_t^{*2}\right),\\ A_t^0 &= \frac{\tilde{y}_t^2}{2}\left(1 + h_t^*\right) - \frac{1}{2},\\ A_t^{00} &= -\frac{\tilde{y}_t^2}{4}.\end{aligned}$$

Davis & Rodriguez-Yam (2005), Shimada & Tsukuda (2005) suggest using the mode $\hat{\mathbf{h}}$ of the posterior distribution $\mathbf{h}|\mathbf{y},\boldsymbol{\theta}$ as $\mathbf{h}^*$. Although the p.d.f. $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ is not directly known, the proportionality (4) allows to acquire the mode by maximizing $f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$ with respect to $\mathbf{h}$:

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = \arg\max_{\mathbf{h}} f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}).$$

The idea of such an approximation can be found in Durbin & Koopman (1997). See also Meyer et al. (2003).

There is a simple iterative algorithm for finding $\hat{\mathbf{h}}$. Suppose that we have an approximate mode $\mathbf{h}^*$. We already considered a quadratic expansion of $\ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$ as a function of $\mathbf{h}$. The expansion of $\lambda(\mathbf{h}) = \ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$ around $\mathbf{h}^*$ is given by

$$\ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) \approx \nabla\lambda(\mathbf{h}^*)^\intercal(\mathbf{h} - \mathbf{h}^*) + \frac{1}{2}(\mathbf{h} - \mathbf{h}^*)^\intercal\nabla^2\lambda(\mathbf{h}^*)(\mathbf{h} - \mathbf{h}^*) + const.$$

Then the next approximation for the mode, $\mathbf{h}^{**}$, is the maximum of this quadratic function:

$$\mathbf{h}^{**} = \mathbf{h}^* - \left(\nabla^2\lambda(\mathbf{h}^*)\right)^{-1}\nabla\lambda(\mathbf{h}^*).$$

This is the classical Newton's method of nonlinear optimization (see Nocedal & Wright, 2006). If the current step does not give an improvement, that is, if $\lambda(\mathbf{y},\mathbf{h}^{**}) < \lambda(\mathbf{y},\mathbf{h}^*)$ then a new approximate value of the mode can be obtained by a line search over $\mathbf{h}^* + \alpha(\mathbf{h}^{**} - \mathbf{h}^*)$.

It is not necessary to invert the $T \times T$ Hessian matrix $\nabla^2 \lambda(\mathbf{h}^*)$ directly. Note that the Hessian is a band (tridiagonal) matrix. The step $\mathbf{h}^{**} - \mathbf{h}^*$ is found as the solution of a system of linear equations

$$\nabla^2 \lambda(\mathbf{h}^*)(\mathbf{h}^{**} - \mathbf{h}^*) = -\nabla \lambda(\mathbf{h}^*),$$

which is simple for a tridiagonal symmetric matrix $\nabla^2 \lambda(\mathbf{h}^*)$. Actually, we already have all necessary data to solve the system. From $\mathbf{h}^*$ we get K_t, L_t, M_t. Then the next approximation $\mathbf{h}^{**}$ can be constructed recursively from the modes of $\mathcal{N}(K_1, M_1)$ and $\mathcal{N}(K_t + L_t h_{t-1}^{**}, M_t)$, $t = 2, \ldots, T$. That is

$$h_1^{**} = K_1, \qquad h_t^{**} = K_t + L_t h_{t-1}^{**}, \qquad t = 2, \ldots, T. \tag{19}$$

(Here we skip the derivation of these formulas from that of the Newton's method. It is an ordinary, but a bit lengthy exercise.)

For the typical data several iterations of the Newton's algorithm are enough. In order to control the convergence we can inspect

$$\nabla \lambda(\mathbf{h}^*)^\intercal \left(\nabla^2 \lambda(\mathbf{h}^*)\right)^{-1} \nabla \lambda(\mathbf{h}^*) / T = -\nabla \lambda(\mathbf{h}^*)^\intercal \left(\mathbf{h}^{**} - \mathbf{h}^*\right) / T.$$

If it is close to zero (say, less than 10^{-12}) then the iterations can be stopped. The gradient $\nabla \lambda(\mathbf{h})$ can be found, for example, by differentiating (8) with respect to $\mathbf{h}$. An element of $\nabla \lambda(\mathbf{h})$ is given by $B_t^0 + 2B_t^{00} h_t + B_t^{01} h_{t-1} + B_{t+1}^{01} h_{t+1}$.

Davis & Rodriguez-Yam (2005), Shimada & Tsukuda (2005) do not prove statistical properties of their estimator. However, empirical examples show that the method can give estimates which are quite close to the exact maximum likelihood estimates, as reported in Davis & Rodriguez-Yam (2005), Shimada & Tsukuda (2005) and Skaug & Yu (2007).

The method is very similar to the Laplace's approximation (LA; it is also known as saddle-point approximation). The Laplace's method is used for an approximate evaluation of integrals of the form

$$\int e^{Mf(\mathbf{x})} d\mathbf{x}.$$

We assume that $f(\mathbf{x})$ is a vector-function with a unique global maximum at $\hat{\mathbf{x}}$ and $\mathbf{x}$ is a $n \times 1$ vector. Point $\hat{\mathbf{x}}$ is characterized by the first-order condition $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}$. The function $f(\mathbf{x})$ is approximated by the second-order expansion around $\hat{\mathbf{x}}$:

$$f(\mathbf{x}) \approx f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^\intercal \nabla^2 f(\hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}}) = f(\hat{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^\intercal \nabla^2 f(\hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}}).$$

Accordingly, the integral is approximated by

$$\int e^{Mf(\mathbf{x})} d\mathbf{x} \approx e^{Mf(\hat{\mathbf{x}})} \int \exp\left(\frac{M}{2}(\mathbf{x} - \hat{\mathbf{x}})^\intercal \nabla^2 f(\hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})\right) d\mathbf{x}.$$

The integral in the right-hand side is closely related to the probability density function of the multivariate normal distribution $\mathcal{N}\left(\hat{\mathbf{x}}, -\left(M\nabla^2 f(\hat{\mathbf{x}})\right)^{-1}\right)$. Knowing that the integral of the density is one, we can write

$$\int e^{Mf(\mathbf{x})} d\mathbf{x} \approx \left(\frac{2\pi}{M}\right)^{n/2} \left|-\nabla^2 f(\hat{\mathbf{x}})\right|^{-1/2} e^{Mf(\hat{\mathbf{x}})}.$$

The Laplace's approximation is valid asymptotically as $M \to \infty$.

It is clear that the above argument is not applicable to SV model. There is no multiplier similar to M which can be assumed to be "sufficiently large" to allow an asymptotic justification of the Laplace's approximation. It is wise therefore to be a bit cautious when using this estimator, as its bias would not vanish in large samples. Davis & Rodriguez-Yam (2005) propose to use a bootstrap to reduce the bias.

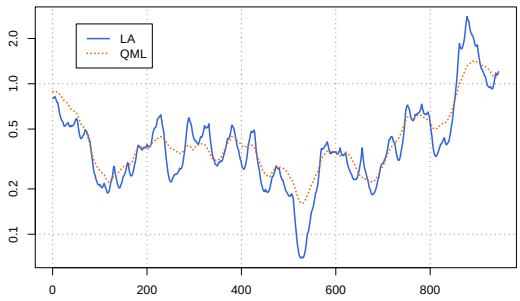

Figure 7: *The smoothed value of the conditional variance from the Laplace's approximation (solid) and the QML estimates (dotted), Example 2.*

Table 2: *Estimates for the Laplace's approximation method*

	Example 2		Example 3			Example 3, simulation		
	estimates	std. err.	estimates	true values	std. err.	mean	RMSE	mean std.err.
δ	0.9750	0.0122	0.9613	0.9800	0.0180	0.9653	0.0361	0.0186
σ_η	0.1632	0.0363	0.2397	0.2000	0.0486	0.2120	0.0538	0.0495
σ_ξ	0.6360	0.0685	0.8031	1.0000	0.1106	1.0133	0.2167	0.1731

The covariance matrix of the estimates based on the Laplace's method can be approximated by the minus inverse Hessian as is common in the maximum likelihood estimation. Of course, consistency of this estimator cannot be assured. Judging from the results of the theory of extremum estimators it can be conjectured that there should be asymptotic normality similar to (18):

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{LA} - \boldsymbol{\theta}^*_{LA}) \sim \mathcal{N}(\mathbf{0}, (\mathscr{H}^\infty_{LA})^{-1} \mathscr{I}^\infty_{LA} (\mathscr{H}^\infty_{LA})^{-1}),$$

where $\mathscr{I}^\infty_{LA}$ and $\mathscr{H}^\infty_{LA}$ are defined similarly to $\mathscr{I}^\infty_Q$ and $\mathscr{H}^\infty_Q$ and $\boldsymbol{\theta}^*_{LA}$ is the pseudo-true value of the parameters vector. The finite-sample analogues, $\mathscr{I}^T_{LA}$ and $\mathscr{H}^T_{LA}$, can be straightforwardly estimated by Monte Carlo. This would provide a consistent estimator of covariance matrix. Another possibility is to estimate the covariance matrix using a bootstrap (see Davis & Rodriguez-Yam (2005)).

Example 2 (continued). Figure 7 shows the smoothed value of the conditional variance based on the Laplace's approximation estimate for the exchange rates data. The left part of Table 2 shows the estimates and their standard errors (based on the minus inverse Hessian).

Example 3 (continued). The central part of Table 2 shows the Laplace's approximation estimates for the realization of the SV process from Figure 3. The right part of the table reports the root mean squared errors for the estimator based on the Laplace's approximation. The RMSEs were estimated from 1000 Monte Carlo simulations with the same true values of the parameters. The RMSEs are lower than the RMSEs for the QML estimator.

5 Simulation-based likelihood approximation

5.1 Introduction

The maximum likelihood method has clear advantages in the case of the SV model as the probability distribution of the data is fully specified by the assumptions of the model. The maximum likelihood is a classical and well-understood method for which a rich theory and a battery of useful procedures are available. However, it requires resorting to computer-intensive techniques. With steady increase of computer power computer-intensive techniques become more practicable, thus making the maximum likelihood a method of choice for stochastic volatility modeling.

To apply numerical optimization algorithms to the problem of finding the (arg)maximum of the likelihood one needs a method to evaluate the likelihood for a given value of parameters vector $\boldsymbol{\theta}$. To evaluate a multidimensional integral

$$L(\boldsymbol{\theta};\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}) = \int f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})\,d\mathbf{h}$$

one needs numerical integration algorithms. Ordinary deterministic algorithms of numerical integration are not very suitable for multidimensional integrals due to the "curse of dimensionality". Consequently, the most practical family of algorithms is based on Monte Carlo simulations. Estimators of $\boldsymbol{\theta}$ which are defined as solutions to

$$L_{MC}(\boldsymbol{\theta};\mathbf{y}) \to \max_{\boldsymbol{\theta}},$$

where $L_{MC}(\boldsymbol{\theta};\mathbf{y})$ is a Monte Carlo approximation to the likelihood function $L(\boldsymbol{\theta};\mathbf{y})$ are called simulated maximum likelihood (SML) or Monte Carlo maximum likelihood estimators.

Monte Carlo methods were introduced to the SV literature by Danielsson & Richard (1993), Danielsson (1994), Shephard (1993). Simulation-based likelihood approximations were first developed in Danielsson & Richard (1993), Danielsson (1994). Other important contributions to the classical (non-Bayesian) simulation-based maximum likelihood approach are Durbin & Koopman (1997), Shephard & Pitt (1997), Sandmann & Koopman (1998), Durbin & Koopman (2000), Liesenfeld & Richard (2003), Durham (2006). For the Bayesian approach to the SV model see Jacquier et al. (1994), Shephard & Pitt (1997), Kim et al. (1998), Durbin & Koopman (2000), Meyer & Yu (2000), Chib et al. (2002), Hautsch & Ou (2008).[23]

5.2 Monte Carlo integration and importance sampling explained

The basic idea of the Monte Carlo integration is that an integral

$$I = \int f(\mathbf{x})\,d\mathbf{x}$$

can be rewritten as

$$I = \int \frac{f(\mathbf{x})}{\mu(\mathbf{x})}\mu(\mathbf{x})\,d\mathbf{x} = \mathsf{E}_\mu \frac{f(\mathbf{x})}{\mu(\mathbf{x})} = \mathsf{E}_\mu\, v(\mathbf{x}),$$

where $\mu(\mathbf{x})$ is the p.d.f. of some suitable distribution (called a proposal distribution[24]), E_μ is the expectation taken under the assumption that $\mathbf{x} \sim \mu(\mathbf{x})$ and

$$v(\mathbf{x}) = f(\mathbf{x})/\mu(\mathbf{x}).$$

It is assumed that $\mu(\mathbf{x})$ is known in a closed form and there exist efficient methods of generation (pseudo-)random variables from μ. Given a sample of size S of random variables $\mathbf{x}^s \sim \mu(\mathbf{x})$ we can compute a Monte Carlo approximation to I as

$$I = \mathsf{E}_\mu\, v(\mathbf{x}) \approx \frac{1}{S}\sum_{s=1}^{S} v(\mathbf{x}^s)$$

or

$$I \approx \bar{v},$$

[23] In fact, almost all of the techniques described in this essay can be adapted to the Bayesian inference after suitable modifications. The widespread use of MCMC methods (see footnote 36) for Bayesian computations is explained mostly by historical reasons. The importance sampling is no less adequate for the task, but it is possibly more intuitive due to a simpler probability theory used.

[24] Other terms are instrumental distribution, importance distribution, importance sampler or just sampler.

where $\bar{v}$ is the arithmetic mean of S values $v^s = v(\mathbf{x}^s) = f(\mathbf{x}^s)/\mu(\mathbf{x}^s)$. (Below we write $\mathbf{x}^s \leftsquigarrow \mu(\mathbf{x})$ to show that independent random variables $\mathbf{x}^s$, $s = 1, \ldots, S$ are to be generated according to a distribution with the density $\mu(\mathbf{x})$).

This approximation is based on the Law of large numbers from which it follows that $\bar{v}$ converges almost surely to I. Of course, there is no guarantee that this approximation would be good for an arbitrary $\mu(\mathbf{x})$ unrelated to $f(\mathbf{x})$. The values of v^s in Monte Carlo samples can be too different, some very small and some very large, but rare, which makes the sample mean a poor estimate.

In probabilistic terms, there is no guarantee that $\bar{v}$ has a finite variance. Note that

$$\mathsf{Var}\,\bar{v} = \frac{1}{S^2}\sum_{s=1}^{S}\mathsf{Var}\,v^s = \frac{1}{S}\mathsf{Var}_\mu\, v(\mathbf{x}).$$

It is advisable to choose μ for which the variance $\mathsf{Var}_\mu\, v(\mathbf{x})$ (and hence $\mathsf{Var}_\mu\,\bar{v}$) is finite and low.

In practice a badly chosen proposal distribution would show up in the problems with the speed of convergence of $\bar{v}$ to the limit I. As S goes to infinity one would see from time to time extremely large values of $v^s = f(\mathbf{x}^s)/\mu(\mathbf{x}^s)$ which would lead to leaps in $\bar{v}$.

The minimal variance of $\bar{v}$ is attained when $f(\mathbf{x})$ and $\mu(\mathbf{x})$ are proportional so that $v(\mathbf{x})$ does not depend on $\mathbf{x}$. In this case $\mathsf{Var}_\mu\,\bar{v} = 0$ and $I = \bar{v}$ with probability one. This seems paradoxical. Explanation of this seeming paradox is that if we know exactly a density function $\mu(\mathbf{x})$ such that $\mu(\mathbf{x}) \propto f(\mathbf{x})$ then $f(\mathbf{x}) = \mu(\mathbf{x})I$ (because by definition $\int \mu(\mathbf{x})d\mathbf{x} = 1$) which would mean that I is known.

It follows that a good proposal density $\mu(\mathbf{x})$ should be approximately proportional to $f(\mathbf{x})$ (assuming that $f(\mathbf{x})$ is non-negative). A good approximation would lead to a small Monte Carlo variance and a fast root-S convergence of $\bar{v}$ to I. A bad approximation would lead to a large Monte Carlo variance even for large S and lack of convergence of $\bar{v}$ to I.

The importance sampling (IS) is a particular case of Monte Carlo integration which refers to the situation when the integral I to be evaluated is represented from the start in the form of the expectation of some function $\tau(\mathbf{x})$ with $\mathbf{x}$ distributed according to some p.d.f. $\pi(\mathbf{x})$, that is,

$$I = \int \tau(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = \mathsf{E}_\pi\,\tau(\mathbf{x}).$$

There is no guarantee that the direct approximation

$$I \approx \frac{1}{S}\sum_{s=1}^{S}\tau(\mathbf{x}^s) \quad \text{with } \mathbf{x}^s \leftsquigarrow \pi(\mathbf{x})$$

would be accurate enough. The reason is the same as was set forth for the general Monte Carlo integration. A suitable choice of the proposal distribution $\mu(\mathbf{x})$ can improve the accuracy. The integral is written as

$$I = \int \tau(\mathbf{x})W(\mathbf{x})\mu(\mathbf{x})d\mathbf{x} = \mathsf{E}_\mu\,\tau(\mathbf{x})W(\mathbf{x}),$$

where $W(\mathbf{x}) = \pi(\mathbf{x})/\mu(\mathbf{x})$. Then the new approximation is

$$I \approx \frac{1}{S}\sum_{s=1}^{S}\tau(\mathbf{x}^s)W(\mathbf{x}^s) \quad \text{with } \mathbf{x}^s \leftsquigarrow \mu(\mathbf{x}).$$

This is a weighted average with weights $W^s = W(\mathbf{x}^s)$ (called the importance weights). Note that in general the weights are unnormalized; they do not sum to one. It is also possible to use the normalized importance weights

$$w^s = \frac{W^s}{\sum_{k=1}^{S} W^k}$$

so that

$$I \approx \frac{\sum_{s=1}^{S} \tau(\mathbf{x}^s) W^s}{\sum_{s=1}^{S} W^s} = \sum_{s=1}^{S} \tau(\mathbf{x}^s)\, w^s \quad \text{with } \mathbf{x}^s \leftsquigarrow \mu(\mathbf{x}).$$

For the importance sampling to provide good accuracy the proposal distribution should be chosen in such a way that its p.d.f. $\mu(\mathbf{x})$ is approximately proportional to $\tau(\mathbf{x})\pi(\mathbf{x})$. This should work for positive[25] functions $\tau(\mathbf{x})$. When $\mu(\mathbf{x})$ is approximately proportional to $\tau(\mathbf{x})\pi(\mathbf{x})$ the function $\tau(\mathbf{x})W(\mathbf{x})$ is approximately constant and the variance of the Monte Carlo estimator is small.

Another use of the importance sampling applies to the case where the p.d.f. $\pi(\mathbf{x})$ is known only in an unnormalized form, that is, only $\Pi(\mathbf{x})$ is known where $\Pi(\mathbf{x}) = C\pi(\mathbf{x})$ and C is an unknown constant given by $C = \int \Pi(\mathbf{x})\, d\mathbf{x}$. The goal is to estimate $I = \mathsf{E}_\pi\, \tau(\mathbf{x})$. One can write I as

$$I = \frac{\int \tau(\mathbf{x})\Pi(\mathbf{x})\, d\mathbf{x}}{\int \Pi(\mathbf{x})\, d\mathbf{x}} = \frac{\int \tau(\mathbf{x}) W(\mathbf{x})\mu(\mathbf{x})\, d\mathbf{x}}{\int W(\mathbf{x})\mu(\mathbf{x})\, d\mathbf{x}},$$

where $W(\mathbf{x}) = \Pi(\mathbf{x})/\mu(\mathbf{x})$. The importance sampling approximation for I is the same as above:

$$I \approx \frac{\sum_{s=1}^{S} \tau(\mathbf{x}^s) W^s}{\sum_{s=1}^{S} W^s} = \sum_{s=1}^{S} \tau(\mathbf{x}^s)\, w^s \quad \text{with } \mathbf{x}^s \leftsquigarrow \mu(\mathbf{x}),$$

with[26]

$$w^s = \frac{W^s}{\sum_{k=1}^{S} W^k} = \frac{\Pi(\mathbf{x}^s)/\mu(\mathbf{x}^s)}{\sum_{k=1}^{S} \Pi(\mathbf{x}^k)/\mu(\mathbf{x}^k)}.$$

When $\tau(\mathbf{x})$ does not vary much, a good choice of the proposal distribution would ensure that all the weights w^s are approximately the same (about $1/S$) so that $\{\mathbf{x}^s\}$ represent approximately an equally weighted sample from $\pi(\mathbf{x})$.

If integrals should be estimated for a set of different functions $\tau(\mathbf{x})$ it would be time-consuming to adapt $\mu(\mathbf{x})$ to each new function. Suppose that the corresponding expectations do exist and the IS estimates have low enough variances when $\pi(\mathbf{x})$ is used directly as $\mu(\mathbf{x})$ (if $\pi(\mathbf{x})$ were known). Then it would be natural to fit $\mu(\mathbf{x})$ to $\Pi(\mathbf{x})$. A popular sample characteristic of the quality of such approximation (given a sample $\mathbf{x}^s \leftsquigarrow \mu(\mathbf{x})$ with weights w^s) is the effective sample size

$$\mathrm{ESS} = \frac{1}{\sum_{s=1}^{S} (w^s)^2}.$$

When all of the weights w^s are $1/S$ one has $\mathrm{ESS} = S$. If $\mathrm{ESS} \ll S$ then $\mu(\mathbf{x})$ is a poor approximation to $\pi(\mathbf{x})$.[27] (One can also use the coefficient of variation for w^s, the variance of $\ln w^s$, the entropy and other accuracy measures.)

Additional information about Monte Carlo integration and importance sampling can be found in Evans & Swartz (1995), Gentle (2003), Rubinstein & Kroese (2008).

[25] For functions which are sometimes negative and sometimes positive $\mu(\mathbf{x})$ should be chosen approximately proportional to $|\tau(\mathbf{x})|\,\pi(\mathbf{x})$. However, this would not make the variance of the Monte Carlo estimator close to zero.

[26] There is a minor technical point in computing the normalized importance weights. The weights can be quite huge in some situations. So it is better to obtain them in logarithmic form as $\ln W^s = \ln \Pi(\mathbf{x}^s) - \ln \mu(\mathbf{x}^s)$. Then one can find the largest weight W^L and use the following formula for the normalized weights:

$$w^s = \frac{\exp(\ln W^s - \ln W^L)}{\sum_{k=1}^{S} \exp(\ln W^k - \ln W^L)}$$

to avoid an arithmetic overflow (or underflow).

[27] If $\mathrm{ESS} \ll S$ then empirical ESS as given in the text is also a poor estimator of the theoretical effective sample size (which we don't define here). Thus, it is hard to decide which of two poor proposal distributions is better on the basis of empirical ESS values.

5.3 Monte Carlo integration for SV model

Monte Carlo integration for dynamic models with unobserved components like the SV model comprises simulation of several trajectories for the unobserved dynamic components. In the case of the SV model a typical Monte Carlo method uses a sample $\mathbf{h}^1,\dots,\mathbf{h}^S$ of trajectories generated according to some distribution which resembles the posterior distribution $\mathbf{h}|\mathbf{y}$.

For the SV model a crude ("brute force") approach to Monte Carlo evaluation of $f(\mathbf{y}|\boldsymbol{\theta})$ is to use

$$f(\mathbf{y}|\boldsymbol{\theta}) = \int f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})\,d\mathbf{h} = \int f(\mathbf{y}|\mathbf{h},\boldsymbol{\theta})\,f(\mathbf{h}|\boldsymbol{\theta})\,d\mathbf{h} = \mathsf{E}_{f(\mathbf{h}|\boldsymbol{\theta})}\, f(\mathbf{y}|\mathbf{h},\boldsymbol{\theta}).$$

This gives a crude approximation

$$f(\mathbf{y}|\boldsymbol{\theta}) \approx \frac{1}{S}\sum_{s=1}^{S} f(\mathbf{y}|\mathbf{h}^s,\boldsymbol{\theta})$$

with $\mathbf{h}^s \leftsquigarrow f(\mathbf{h}|\boldsymbol{\theta})$. However, this direct approach is not usable. Even for enormous number of simulations S the approximation would be inaccurate.

To get a better Monte Carlo approximation we can use some other proposal density $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$:

$$f(\mathbf{y}|\boldsymbol{\theta}) = \int \frac{f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})}{g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})}\,g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})\,d\mathbf{h} = \mathsf{E}_g\,\frac{f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})}{g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})}.$$

Denote

$$v(\mathbf{h};\mathbf{y},\boldsymbol{\theta}) = \frac{f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})}{g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})}.$$

Then

$$f(\mathbf{y}|\boldsymbol{\theta}) = \mathsf{E}_g\, v(\mathbf{h};\mathbf{y},\boldsymbol{\theta})$$

and the Monte Carlo approximation for the likelihood function $L(\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta})$ is given by the corresponding sample average[28]

$$L_{MC}(\boldsymbol{\theta}) = \bar{v}(\mathbf{y},\boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S} v(\mathbf{h}^s;\mathbf{y},\boldsymbol{\theta}) \quad \text{with } \mathbf{h}^s \leftsquigarrow g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}). \tag{20}$$

The Monte Carlo approximation for log-likelihood is then

$$\ell_{MC}(\boldsymbol{\theta}) = \ln \bar{v}(\mathbf{y},\boldsymbol{\theta}). \tag{21}$$

As was explained above, $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ should be chosen to be approximately proportional to $f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{h},\boldsymbol{\theta})f(\mathbf{h}|\boldsymbol{\theta})$. The ideal choice of $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ is $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ because then $v(\mathbf{h};\mathbf{y},\boldsymbol{\theta})$ is a constant equal to $f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})/f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta})$. However, $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ is no more known than $f(\mathbf{y}|\boldsymbol{\theta})$. Therefore, the key requirement for using Monte Carlo integration to evaluate the likelihood function is to find a good approximation to $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$.

Note that the problem of finding a good approximation to $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ should be solved anew for each value of the parameters vector $\boldsymbol{\theta}$. Also such an approximation should depends on the available data $\mathbf{y}$. We emphasized this in our notation by writing the proposal distribution as $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$, not simply as $g(\mathbf{h}|\mathbf{y})$ or $g(\mathbf{h})$.

We can compare the use of a general proposal distribution $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ with the crude approach based on prior distribution of $\mathbf{h}$. Denote $W(\mathbf{h};\mathbf{y},\boldsymbol{\theta}) = f(\mathbf{h}|\boldsymbol{\theta})/g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$. Then $f(\mathbf{y}|\boldsymbol{\theta})$ can be written as

$$f(\mathbf{y}|\boldsymbol{\theta}) = \mathsf{E}_g\left[f(\mathbf{y}|\mathbf{h},\boldsymbol{\theta})\,W(\mathbf{h};\mathbf{y},\boldsymbol{\theta})\right].$$

[28] Some $v^s = v(\mathbf{h}^s;\mathbf{y},\boldsymbol{\theta})$ can be quite large to be dealt directly. When implementing the method one would prefer to obtain the weights in the logarithmic form as $\ln v^s = \ln f(\mathbf{y},\mathbf{h}^s|\boldsymbol{\theta}) - \ln g(\mathbf{h}^s|\mathbf{y},\boldsymbol{\theta})$ and take precautions similar to those described in footnote 26.

This demonstrates that (20) is the importance sampling with respect to sampling from the prior distribution $f(\mathbf{h}|\boldsymbol{\theta})$. That is why in the SV literature the Monte Carlo methods of approximating $f(\mathbf{y}|\boldsymbol{\theta})$ by numerical integration are called importance sampling methods. However, there is no good reason to consider the crude proposal distribution $f(\mathbf{h}|\boldsymbol{\theta})$ as a natural one. It is not difficult to find much better approximations to $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$.

If the distribution $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ is T-dimensional normal $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ for some $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\mathbf{y},\boldsymbol{\theta})$ and $\boldsymbol{\mu} = \boldsymbol{\mu}(\mathbf{y},\boldsymbol{\theta})$ and $\boldsymbol{\zeta}^s \rightsquigarrow \mathcal{N}(\mathbf{0}_T,\mathbf{I}_T)$ for $s = 1,\ldots,S$ is a set of initial standard normal random numbers then a Monte Carlo set of trajectories $\mathbf{h}^s \rightsquigarrow g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ can be obtained by

$$\mathbf{h}^s = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}^s,$$

where $\boldsymbol{\Sigma}^{1/2}$ is some square root of $\boldsymbol{\Sigma}$. (The most natural square root of $\boldsymbol{\Sigma}$ can be obtained by the Cholesky decomposition). Obviously, the dimensionality of $\boldsymbol{\Sigma}$ can be too high which makes the direct method unsuitable for the actual computations. However, we have a decomposition

$$g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = g(h_1|\mathbf{y},\boldsymbol{\theta})\prod_{t=2}^{T} g(h_t|h_{t-1},\mathbf{y},\boldsymbol{\theta})$$

which allows to sample from $g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ recursively using $h_1^s \rightsquigarrow g(h_1|\mathbf{y},\boldsymbol{\theta})$, $h_t^s \rightsquigarrow g(h_t|h_{t-1}^s,\mathbf{y},\boldsymbol{\theta})$ ($t = 2,\ldots,T$) or

$$h_1^s \rightsquigarrow \mathcal{N}(K_1,M_1), \qquad h_t^s \rightsquigarrow \mathcal{N}(K_t + L_t h_{t-1}^s, M_t), \quad t = 2,\ldots,T.$$

Given an initial standard normal random vector $\boldsymbol{\zeta}^s$ we can obtain $\mathbf{h}^s$ as follows:

$$h_1^s = K_1 + \zeta_1^s\sqrt{M_1} \qquad \text{and} \qquad h_t^s = K_t + L_t h_{t-1}^s + \zeta_t^s\sqrt{M_t}, \quad t = 2,\ldots,T.$$

Note that $L_{MC}(\boldsymbol{\theta})$ is to be maximized with respect to $\boldsymbol{\theta}$ and that it most probably would be used to evaluate numerical derivatives. So it is important that $L_{MC}(\boldsymbol{\theta})$ is smooth with respect to $\boldsymbol{\theta}$. If for each evaluation of the Monte Carlo likelihood we used a newly generated set of $\boldsymbol{\zeta}^s$, it would make the maximization very troublesome due to random noise. In practice to avoid the Monte Carlo "chatter" the same sample of initial random numbers $\boldsymbol{\zeta}^1,\ldots,\boldsymbol{\zeta}^S$ is used for each likelihood evaluation. This is called the method of common random numbers.

The most popular proposal distribution in the SV literature is the one based on the Laplace's approximation. We will call the corresponding SML method SML-LA. It can utilize the Kalman filter for needed calculations as in Durbin & Koopman (1997), Shephard & Pitt (1997), Sandmann & Koopman (1998), Durbin & Koopman (2000). Alternatively, Durham (2006), Skaug & Yu (2007) develop a direct approach utilizing the well-known properties of band matrices. Our discussion above which utilizes a simple factorization of the multivariate Gaussian density is a convenient reformulation of this later approach.

The simulated maximum likelihood method provides estimates which asymptotically coincide with the maximum likelihood estimates if S grows to infinity together with T at a sufficiently fast rate.[29] Under this assumption an asymptotic approximation to the distribution of SML estimates $\hat{\boldsymbol{\theta}}_{MC}$ is given by

$$\hat{\boldsymbol{\theta}}_{MC} \sim \mathcal{N}(\boldsymbol{\theta}, -\hat{\mathbf{H}}_{MC}^{-1}), \tag{22}$$

where

$$\hat{\mathbf{H}}_{MC} = \nabla_{\boldsymbol{\theta}}^2 \ell_{MC}(\hat{\boldsymbol{\theta}}_{MC}) = \nabla_{\boldsymbol{\theta}}^2 \ell_{MC}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MC}}$$

is the Hessian matrix of the Monte Carlo log-likelihood. However, Monte Carlo method for finite S has an associated numerical error. In practice we have some finite S and T and would like to estimate the size of the Monte Carlo errors in the parameters estimates.

[29] More specifically, the requirement is $T \to \infty$, $S \to \infty$ and $\sqrt{T}/S \to 0$. See Gourieroux & Monfort (1997), Proposition 3.2.

Table 3: *SML-LA estimates (S = 1000)*

	Example 2			Example 3			
	estimates	std. err.	MC std. err.	estimates	true values	std. err.	MC std. err.
δ	0.9753	0.0121	0.00015	0.9613	0.9800	0.0180	0.00021
σ_η	0.1630	0.0360	0.00064	0.2417	0.2000	0.0491	0.00101
σ_ξ	0.6363	0.0690	0.00020	0.8027	1.000	0.1112	0.00005

Table 4: *SML-LA estimates, Example 3, simulation*

	mean	true values	RMSE	mean std.err.
δ	0.9628	0.9800	0.0324	0.0274
σ_η	0.2191	0.2000	0.0539	0.0504
σ_ξ	1.0192	1.000	0.2058	0.2049

A straightforward (but computationally demanding) way to evaluate the Monte Carlo errors is to use the Monte Carlo method. At first several SML estimates $\hat{\boldsymbol{\theta}}_{MC}$ for independent sets of initial random numbers are obtained. Then the standard errors due to Monte Carlo are computed as the standard deviations of these estimates. For example, see Liesenfeld & Jung (2000).

Durbin & Koopman (1997) propose the following approximation for the mean squared error matrix due to Monte Carlo (that is, the mean squared error matrix with respect to the unknown exact maximum likelihood estimate $\hat{\boldsymbol{\theta}}$):

$$\mathsf{E}[(\hat{\boldsymbol{\theta}}_{MC}-\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{MC}-\hat{\boldsymbol{\theta}})^{\intercal}] \approx \hat{\mathbf{H}}_{MC}^{-1}\Big[\frac{1}{S^2\bar{v}^2}\sum_{s=1}^{S}(\mathbf{q}^s-\bar{\mathbf{q}})(\mathbf{q}^s-\bar{\mathbf{q}})^{\intercal}\Big]\hat{\mathbf{H}}_{MC}^{-1}, \tag{23}$$

where $\bar{v}=\bar{v}(\mathbf{y},\hat{\boldsymbol{\theta}}_{MC})$ given by (20), $\mathbf{q}^s=\nabla_{\boldsymbol{\theta}}v(\mathbf{h}^s;\mathbf{y},\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MC}}$ and $\bar{\mathbf{q}}=\frac{1}{S}\sum_{s=1}^{S}\mathbf{q}^s$. The details can be found in Durbin & Koopman (2001), pp. 217-219.

Example 2 (continued). The left part of Table 3 shows the SML-LA estimates for the exchange rates data and their standard errors (based on (22)). The estimator uses $S = 1000$ simulations. The results are very similar to those for the Laplace's approximation method without simulation (see Table 2). ESS at the maximum is about 300 which shows that the proposal distribution is reliable. The Monte Carlo standard errors are the square roots of the diagonal elements of the Durbin–Koopman estimate (23). These standard errors due to Monte Carlo are rather low compared to the standard errors of the parameters estimates. Actually, for practical purposes we could take much smaller number of simulations, $S = 100$ or less. Sandmann & Koopman (1998) recommend to choose S as low as 5.

Example 3 (continued). The right part of Table 3 shows the SML-LA estimates based on $S = 1000$ simulations for the realization of SV process from Figure 3. In this example ESS is about 211, which confirms that the proposal distribution is reliable. Table 4 reports the root mean squared errors for the SML-LA estimator based on $S = 100$ simulations. The RMSEs were estimated from 300 Monte Carlo simulations with the same true values of the parameters. The RMSEs are very close to the RMSEs for the parent LA estimator (see Table 2).

Besides the Laplace's approximation we could obtain a proposal distribution from the QML approximation. However, its performance is hopelessly inferior. For example, for the exchange rates data with $S = 10000$ and the same parameters as in Table 3 its application resulted in an ESS value as low as 1.74.

An interesting development of the idea of simulation with the QML proposal distribution is the method proposed in Kim et al. (1998). The distribution of $\ln(\xi_t^2)$ can be approximated as a mixture of normals. If s_t is a variable corresponding to the index of a normal distribution in a mixture for

time t then conditionally on $s_1, \ldots, s_T$ one has a linear Gaussian state-space model, which is easy to handle. We will not explain this method further; see Kim et al. (1998).

5.4 Efficient importance sampling

Liesenfeld & Richard (2003), Liesenfeld & Richard (2006) propose to use the efficient importance sampling (EIS) technique due to Richard & Zhang (2007) to estimate stochastic volatility models. The idea is to select a proposal distribution used in Monte Carlo integration in such a way that it approximately minimizes the variance of the estimate. This approach to SV modeling can be traced back to Danielsson & Richard (1993) and Danielsson (1994) where a special case of it is developed under the name of "accelerated Gaussian importance sampling".

Suppose that there is a family of possible proposal distributions $\mu(\mathbf{x}, \boldsymbol{\psi})$ used for Monte Carlo integration which depends on a vector of parameters $\boldsymbol{\psi}$. The integral $I = \int \phi(\mathbf{x}) d\mathbf{x}$ is estimated as

$$\hat{I} = \frac{1}{S} \sum_{s=1}^{S} v(\mathbf{x}^s, \boldsymbol{\psi}) \quad \text{with } v(\mathbf{x}^s, \boldsymbol{\psi}) = \frac{\phi(\mathbf{x}^s)}{\mu(\mathbf{x}^s, \boldsymbol{\psi})}, \qquad \mathbf{x}^s \leftsquigarrow \mu(\mathbf{x}^s, \boldsymbol{\psi}).$$

As the realizations $\mathbf{x}^s$ are drawn independently it follows that

$$\mathsf{Var}\, \hat{I} = \frac{1}{S} \mathsf{Var}_{\boldsymbol{\psi}}\, v(\mathbf{x}, \boldsymbol{\psi}).$$

(We use $\mathsf{E}_{\boldsymbol{\psi}}$ ($\mathsf{Var}_{\boldsymbol{\psi}}$) to denote the expectation (variance) with respect to $\mu(\mathbf{x}, \boldsymbol{\psi})$). We want to find the value of $\boldsymbol{\psi}$ for which the variance is approximately minimal. It can be seen that the variance $\mathsf{Var}\, \hat{I}$ is proportional to

$$\mathsf{Var}_{\boldsymbol{\psi}}\, v(\mathbf{x}, \boldsymbol{\psi}) = \mathsf{E}_{\boldsymbol{\psi}} \left[(v(\mathbf{x}, \boldsymbol{\psi}) - I)^2 \right] = \int \left(v(\mathbf{x}, \boldsymbol{\psi}) - I \right)^2 \mu(\mathbf{x}, \boldsymbol{\psi}) d\mathbf{x}.$$

The integral would not be known in a closed form, but it can be approximated by the sample average of $(v(\mathbf{x}^s, \boldsymbol{\psi}) - I)^2$ with $\mathbf{x}^s \leftsquigarrow \mu(\mathbf{x}, \boldsymbol{\psi})$. However, using $\mu(\mathbf{x}, \boldsymbol{\psi})$ as a proposal distribution[30] creates difficulties for minimization of the estimated variance with respect to $\boldsymbol{\psi}$. To circumvent these difficulties, we can use a proposal distribution with some preliminary parameters vector $\boldsymbol{\psi}^*$. If $\boldsymbol{\psi}^*$ is the current vector of parameters then

$$\int \left(v(\mathbf{x}, \boldsymbol{\psi}) - I \right)^2 \frac{\mu(\mathbf{x}, \boldsymbol{\psi})}{\mu(\mathbf{x}, \boldsymbol{\psi}^*)} \mu(\mathbf{x}, \boldsymbol{\psi}^*) d\mathbf{x} = \mathsf{E}_{\boldsymbol{\psi}^*} \left[\left(v(\mathbf{x}, \boldsymbol{\psi}) - I \right)^2 \frac{\mu(\mathbf{x}, \boldsymbol{\psi})}{\mu(\mathbf{x}, \boldsymbol{\psi}^*)} \right],$$

where the expectation is taken with respect to $\mu(\mathbf{x}, \boldsymbol{\psi}_0)$. This can be approximated by

$$\frac{1}{S} \sum_{s=1}^{S} \left(v(\mathbf{x}^s, \boldsymbol{\psi}) - I \right)^2 \frac{\mu(\mathbf{x}^s, \boldsymbol{\psi})}{\mu(\mathbf{x}^s, \boldsymbol{\psi}^*)} = \frac{1}{S} \sum_{s=1}^{S} \left(\frac{\phi(\mathbf{x}^s)}{\mu(\mathbf{x}^s, \boldsymbol{\psi})} - I \right)^2 \frac{\mu(\mathbf{x}^s, \boldsymbol{\psi})}{\mu(\mathbf{x}^s, \boldsymbol{\psi}^*)} \tag{24}$$

with $\mathbf{x}^s \leftsquigarrow \mu(\mathbf{x}, \boldsymbol{\psi}^*)$. The function can be minimized with respect to $\boldsymbol{\psi}$ (and I) to get a better proposal distribution than $\mu(\mathbf{x}, \boldsymbol{\psi}^*)$. The procedure can be repeated until convergence by replacing $\boldsymbol{\psi}^*$ with the estimated $\boldsymbol{\psi}$.

The problem of minimizing (24) can be roughly approximated by a least squares problem for log-densities. The corresponding regression is

$$\ln \phi(\mathbf{x}) = \gamma + \ln \mu(\mathbf{x}, \boldsymbol{\psi}) + \text{residual}.$$

(See Richard & Zhang (2007). They also give a better approximation by a weighted least squares problem). So one can simply fit $\ln \mu(\mathbf{x}, \boldsymbol{\psi})$ to $\ln \phi(\mathbf{x})$ (with an additional constant term γ) at a set of points $\mathbf{x} = \mathbf{x}^s$, $s = 1, \ldots, S$, where $\mathbf{x}^s \leftsquigarrow \mu(\mathbf{x}, \boldsymbol{\psi}^*)$.

[30] In some cases it is possible. We need generated $\mathbf{x}^s$ to depend smoothly on $\boldsymbol{\psi}$.

Table 5: *EIS estimates ($S = 100$)*

	Example 2			Example 3			
	estimates	std. err.	MC std. err.	estimates	true values	std. err.	MC std. err.
δ	0.9751	0.0122	0.00017	0.9615	0.9800	0.0179	0.00020
σ_η	0.1640	0.0364	0.00068	0.2408	0.2000	0.0490	0.00084
σ_ξ	0.6360	0.0689	0.00023	0.8027	1.000	0.1114	0.00006

In the case of the stochastic volatility model this approach cannot be applied directly. Suppose that the proposal distribution for $\mathbf{h}|\mathbf{y}$ is multivariate normal. In general a T-dimensional multivariate normal distribution has $T(T+1)/2$ parameters. We can take into account the dynamic structure of the $\mathbf{h}|\mathbf{y}$ distribution for our basic SV model. There is an immediate link between h_t and h_{t-1}, but there is no direct link between h_t and h_{t-k} for $k > 1$. So we can assume a tridiagonal covariance matrix. This reduces the number of parameters to $3T-1$. However, this is still a fairly large number taking into account that in general we need no less simulations than there are parameters of the proposal distribution.

To resolve this problem it is reasonable to use a simpler piecemeal approach for the basic SV model. Note that

$$\ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) = \sum_{t=1}^{T} [\ln f(y_t|h_t,\boldsymbol{\theta}) + \ln f(h_t|h_{t-1},\boldsymbol{\theta})].$$

The terms $\ln f(h_t|h_{t-1},\boldsymbol{\theta})$ are already quadratic in h_{t-1} and h_t. We need only quadratic approximations for $\ln f(y_t|h_t,\boldsymbol{\theta})$ (as a function of in h_t) to obtain a quadratic approximation of $\ln f(\mathbf{y},\mathbf{h},\boldsymbol{\theta})$. We already discussed this approach. So we can simply run the following linear regression:

$$\ln f(y_t|h_t,\boldsymbol{\theta}) = A_t + A_t^0 h_t + A_t^{00} h_t^2 + \text{residual}$$

and calculate K_t, L_t and M_t as before (see section 2 and Appendix A). The regressions are run one by one independently of each other for $t = 1,\dots,T$. The observations for the regressions are obtained from simulated h_t^s, $s = 1,\dots,S$. A single h_t^s for a particular t is taken from $\mathbf{h}^s$, where $\mathbf{h}^s$, $s = 1,\dots,S$ are drawn from the current proposal distribution. Several iterations of the method are made. New K_t, L_t and M_t give a proposal distribution, from which new h_t^s are taken. New h_t^s are used as the data in the EIS regressions leading to new K_t, L_t and M_t and so on. Finally, the approximate log-likelihood for given $\boldsymbol{\theta}$ is obtained from (21). As we noted earlier, the problem of finding a good approximation to $f(\mathbf{h}|\mathbf{y},\boldsymbol{\theta})$ should be solved anew for each value of parameters vector $\boldsymbol{\theta}$.

Actually the method described is largely a heuristics. It is linked only indirectly to the problem of minimizing the variance. Nevertheless, below we call it "efficient importance sampling" following Liesenfeld & Richard (2003). The method based on the normal distribution was first proposed in Danielsson & Richard (1993) as the "accelerated Gaussian importance sampling".

Example 2 and Example 3 (continued). Table 5 is an analogue of Table 3 for EIS. In the EIS algorithm we used $S = 100$ simulations and 3 iterations starting from the Laplace's approximation estimates. The estimates are very similar. ESS is about 79 for the exchange rates and 75 for the data of Example 3, which shows that EIS provides better proposal distributions than the Laplace's approximation (ESS/S of 79% (75%) versus 30% (21%) for LA) and would need less simulations than SML-LA to attain the same accuracy.

Although EIS needs less simulations than SML-LA to attain the same accuracy, it includes an additional computation of regressions coefficients. Which algorithm is faster depends on a computer, programming implementations, data and other circumstances. A Monte Carlo comparison favoring SML-LA as a method of parameter estimation of SV model can be found in Lee & Koopman (2004). In any case, a better proposal distribution provided by EIS is an important advantage for tasks other than the SML estimation (see below).

Table 6: *The method of moments estimates for Example 3, simulation*

		MM		GMM	
	true values	mean	RMSE	mean	RMSE
δ	0.9800	0.6790	0.4021	0.9556	0.0441
σ_η	0.2000	0.5827	0.4731	0.2473	0.1024
σ_ξ	1.000	1.0175	0.2279	1.0268	0.2324

6 Method of moments estimation

It is not hard to derive analytical expressions for various moments of a process y_t described by the basic SV model (1) (see Appendix C). In particular, for $n > -1$

$$\mathsf{E}\left|y_t\right|^n = \frac{\sigma_\xi^n 2^{n/2}\Gamma((n+1)/2)}{\sqrt{\pi}} \exp\left(\frac{n^2\sigma_\eta^2}{8(1-\delta^2)}\right)$$

and for $m > -1$, $n > -1$ and lag $k > 0$

$$\mathsf{E}\left[\left|y_t\right|^m \left|y_{t-k}\right|^n\right] = \frac{1}{\pi}\sigma_\xi^{m+n} 2^{(m+n)/2}\Gamma((m+1)/2)\Gamma((n+1)/2) \exp\left(\frac{(m^2+n^2+2mn\delta^k)\sigma_\eta^2}{8(1-\delta^2)}\right).$$

Moments of $\ln(y_t^2)$ can also be employed:

$$\mathsf{E}\ln(y_t^2) = \ln\sigma_\xi^2 + \mathscr{C}, \qquad \mathsf{Var}[\ln(y_t^2)] = \sigma_\eta^2/(1-\delta^2) + \pi^2/2$$

and for $k > 1$

$$\mathsf{Cov}(\ln(y_t^2), \ln(y_{t-k}^2)) = \sigma_\eta^2\delta^k/(1-\delta^2).$$

To apply the method of moments one calculates theoretical moments of y_t from the SV model as functions of the parameters $\boldsymbol{\theta}$ and then equates these theoretical moments to their sample analogues. If the number of the moments is the same as the number of the unknown parameters this gives a system of nonlinear equations from which parameter estimates can be obtained. Examples of using this technique for estimating the SV model are Scott (1987), Dufour & Valéry (2006).

For example, if m is the sample mean of $\ln(y_t^2)$, s^2 is the sample variance and c_k is the k-th sample autocovariance then a method of moments estimator of the parameters of the basic SV model is given by

$$\hat{\delta} = c_2/c_1, \qquad \hat{\sigma}_\eta = \sqrt{(s^2 - \pi^2/2)(1-\hat{\delta}^2)}, \qquad \hat{\sigma}_\xi = \exp((m - \mathscr{C})/2). \tag{25}$$

The vanilla MM estimates behave poorly, but for long enough series they can be used as reasonable starting values for more complicated algorithms.

Example 3 (continued). We use (25) to estimate the basic SV model for 10000 realizations of the SV process with $\delta = 0.98$, $\sigma_\eta = 0.2$, $\sigma_\xi = 1$ and $T = 500$. Very often (in 51% of all realizations) valid estimates cannot be computed at all, because either $s^2 < \pi^2/2$ or $c_2^2 > c_1^2$. The RMSEs for the valid estimates are reported in Table 6. The simulations results show that the MM estimator given by (25) is almost useless for these settings.

There are infinitely many moments and one can propose infinitely many MM estimators most of which would have inferior statistical properties. The generalized method of moments (GMM)[31] is an extension of the ordinary method of moments which allows to use more moments than there are parameters. See Melino & Turnbull (1990), Andersen (1994), Jacquier et al. (1994), Hall (2005)

[31] See Hansen (1982), Hall (2005).

for applications of GMM to the SV model. Andersen & Sørensen (1996) is an extensive simulation study of the properties of GMM. We do not discuss the use of GMM in the case of the SV model. It is more or less straightforward application of the standard GMM toolkit. The weighting matrix of GMM can be selected optimally and obtained in a closed form for moments based on various powers of $|y_t|$ and $\ln y_t^2$; see Dhaene & Vergote (2003). Popular improvements of the basic GMM can be readily used (the continuously updating GMM, the iterated GMM, the empirical likelihood method).

Example 3 (continued). We employ a modification of the method proposed in Taylor (1986) to estimate the basic SV model for $\delta = 0.98$, $\sigma_\eta = 0.2$, $\sigma_\xi = 1$ and $T = 500$. The parameters δ and σ_η are estimated by minimizing

$$\sum_{k=1}^{K} \left(c_k - \sigma_\eta^2 \delta^k / (1-\delta^2) \right)^2 .$$

This is a simple nonlinear regression. Here K is some chosen number of autocovariances; it should be much smaller than T. As K is much smaller than T, nonlinear regression estimation is much faster than QML estimation. For σ_ξ the estimator is $\hat{\sigma}_\xi = \exp((m - \mathscr{C})/2)$ as above. We used 10000 realizations of the SV process and $K = 50$. The realizations with $|\hat{\delta}| \geq 1$ were rejected. This was observed only for 0.5% of all realizations. RMSEs for remaining estimates are reported in Table 6.

This simple GMM estimator can provide good starting values for other algorithms.

If the moments of a model cannot be obtained analytically one can estimate them using Monte Carlo integration provided the model allows direct simulation (which is the case with the SV model). This leads to the simulated method of moments of Duffie & Singleton (1993). It can be useful for some extended SV models.

It is well-known from the GMM literature that the best choice of the moments should be based on the score vector (the gradient of the log-likelihood function). Then GMM estimation is equivalent to ML estimation and is asymptotically efficient. The generalized method of moments is then called the efficient method of moments (EMM). Gallant & Tauchen (1996), Gallant et al. (1997) propose a Monte Carlo approximation to full EMM based on the score vector of an auxiliary model with the known likelihood function which fits the data sufficiently well (called a score generator). They use the SNP (semi-nonparametric) model as a score generator for the SV model. Andersen et al. (1999) consider several alternative score generators and conduct an extensive simulation study of their performance.

Monfardini (1998), Calzolari et al. (2004) use the indirect inference to estimate SV model. The idea of this method (see Gourieroux et al. (1993)) is to estimate a simple auxiliary model and then find by means of Monte Carlo simulation the parameters of the underlying model which provide the parameters of the auxiliary model as close as possible to those obtained from the original data.

It should be noted that the use of Monte Carlo simulations in a moment-based estimation makes these methods not very competitive compared to the simulated maximum likelihood methods considered in the other sections of this essay. To give a summary, the moment-based methods either provide estimates which are not very accurate or use Monte Carlo simulations which make them almost as computationally expensive as simulated maximum likelihood methods. However, for SV-type models, which are not fully parametrically specified, the moment-based estimation can be preferred as it requires less assumptions to be valid.

Moment-based methods have yet another limitation. They usually do not provide directly information which can be used for smoothing, filtering and forecasting.

For a review of various moment-related techniques for stochastic volatility models see Renault (2009).

7 Extending the basic model

7.1 An extended stochastic volatility model

In this section we will explore a more general SV model

$$\begin{aligned} y_t &= \mathbf{X}_t\boldsymbol{\beta} + \kappa r(h_t) + \sigma_\xi \xi_t \exp(h_t/2), \\ h_t &= \delta h_{t-1} + \alpha\xi_{t-1} + \sigma_\eta \eta_t. \end{aligned} \tag{26}$$

Compared to the basic SV model (7) this formulation includes several additional effects: exogenous variables in the mean, an in-mean effect, asymmetry and fat tails.

The term with $\kappa r(h_t)$ corresponds to an in-mean effect similar to that in the GARCH-M model (cf. Engle et al. (1987)). The idea of this extension is that returns on assets can be related to the degree of riskiness of the assets as risk-averse investors need a compensation for additional risk. The SVM model was proposed in Koopman & Uspensky (2002). Possible choices of the in-mean function $r(\cdot)$ are $r(h_t) = \exp(h_t/2)$, $r(h_t) = \exp(h_t)$ and $r(h_t) = h_t$.

We assume that $\eta_t \sim \mathcal{N}(0,1)$ and ξ_t are independent white noise processes. For ξ_t one can choose a more fat-tailed distribution than $\xi_t \sim \mathcal{N}(0,1)$. Popular choice of the distribution is $\xi_t \sim t_\nu$ (the Student's distribution with ν degrees of freedom). Conditional variance of SV series with Student's t errors is

$$\sigma_\xi^2 \exp(h_t)\,\mathsf{Var}\,\xi_t = \sigma_\xi^2 \exp(h_t)\frac{\nu}{\nu-2}.$$

The time-varying variance in the SV model allows to capture to some great extent the fat tails observed in financial time series. However, as shown by the extensive experience with GARCH-type models using a time-varying variance could be insufficient to fully capture the kurtosis of the observed financial time series. Bollerslev (1987) introduces GARCH-t model, which is the GARCH model with Student's t innovations. Assuming that $\xi_t \sim t_\nu$ in (1) produces a similar generalization for the basic SV model. SV models with fat tails are studied in Harvey et al. (1994), Sandmann & Koopman (1998), Liesenfeld & Jung (2000), Chib et al. (2002), Liesenfeld & Richard (2003), Jacquier et al. (2004) and Durham (2006) among others. An important fact is that, as discussed in Carnero et al. (2004), the SV model with Gaussian errors can be more adequate empirically than the GARCH model with Gaussian errors. Therefore, one would expect to find a relatively large degrees of freedom parameter ν in the SV model with Student's t errors.[32]

Model (26) with $\alpha = 0$, $\kappa = 0$ and $\boldsymbol{\beta} = 0$ is similar to (7) in many aspects and shares with it most of the methods described earlier. We will call it the basic SV-t model.

The $\alpha\xi_{t-1}$ item in the volatility equation of (26) captures an asymmetric effect of innovations on volatility. It is assumed that a negative shock to ξ_{t-1} can lead to a higher level of future volatility. One explanation is that if a stock price is lowered by some shock then the financial leverage (which can be measured by the debt-to-equity ratio) is increased, which tend to raise the volatility in the future. This phenomenon is called the leverage effect.[33] Various aspects of models with asymme-

[32] Another way of introducing fat tails into the SV model is by including an additional latent factor (see Durham (2006)). The second factor could be a white noise or weakly autocorrelated series. In particular, one can use

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + \kappa r(h_t) + \sigma_\xi \xi_t \sqrt{\lambda_t} \exp(h_t/2),$$

where λ_t is the second factor which is i.i.d. with $\nu/\lambda_t \sim \chi^2_\nu$ (see Jacquier et al. (1999), Jacquier et al. (2004)). This imitates (26) with the Student's distribution since $\xi_t\sqrt{\lambda_t} \sim t_\nu$.

[33] Without the fat-tailness of ξ_t we could model asymmetric effect by introducing a correlation between ξ_{t-1} and η_t to the basic SV model (1), that is, by assuming that

$$\begin{pmatrix} \xi_{t-1} \\ \eta_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

There is a question of timing of the asymmetric effect. In Jacquier et al. (2004) it is assumed that ξ_t and η_t are

try and leverage are studied in Jacquier et al. (1994), Harvey & Shephard (1996), Yu (2005), Asai & McAleer (2005), Durham (2006), Omori et al. (2007).

The term $\mathbf{X}_t\boldsymbol{\beta}$ allows y_t to depend on a set of explanatory variables $\mathbf{X}_t$. These can include an intercept term, seasonal dummies. Sandmann & Koopman (1998) mention option implied volatility, trade volume data. Inclusion of the lags of y_t can help to capture autocorrelation.

The presence of a mean component $\mathbf{X}_t\boldsymbol{\beta}$ in (26) does not lead to much difficulty. The coefficients $\boldsymbol{\beta}$ can be estimated consistently before the other parameters by the ordinary least squares when $\kappa = 0$. See Harvey & Shephard (1993) for a further discussion and application of GLS. Alternatively in the maximum likelihood context one can work with the residuals $y_t - \mathbf{X}_t\boldsymbol{\beta}$ instead of y_t and maximize the (approximate) likelihood function with respect to all the parameters jointly.

Below we suppress the dependence on $\boldsymbol{\theta}$ in our notation for the densities.

The distribution of y_t conditional on h_t is based on distribution of ξ_t with a scale $\sigma_\xi \exp(h_t/2)$ and a shift $\mathbf{X}_t\boldsymbol{\beta} + \kappa r(h_t)$. Thus, the log-density for $y_t|h_t$ is given by

$$\ln f(y_t|h_t) = \ln \rho(\xi_t(y_t, h_t)) - \ln \sigma_\xi - h_t/2,$$

where $\rho(\cdot)$ is the density function of ξ_t which can depend on the distribution parameters (like ν for the Student's distribution) and

$$\xi_t(y_t, h_t) = \frac{y_t - \mathbf{X}_t\boldsymbol{\beta} - \kappa r(h_t)}{\sigma_\xi \exp(h_t/2)}. \tag{27}$$

The mean equation disturbance ξ_t is fixed conditional on y_t and h_t and is given by $\xi_t = \xi_t(y_t, h_t)$. Consequently, the distribution of h_t conditional on y_{t-1} and h_{t-1} is normal with the mean $\delta h_{t-1} + \alpha\xi_{t-1}(y_{t-1}, h_{t-1})$ and the variance σ_η^2. The log-density for $h_t|y_{t-1}, h_{t-1}$ is given by

$$\ln f(h_t|y_{t-1}, h_{t-1}) = -\frac{1}{2}\ln(2\pi\sigma_\eta^2) - \frac{1}{2\sigma_\eta^2}(h_t - \delta h_{t-1} - \alpha\xi_{t-1}(y_{t-1}, h_{t-1}))^2.$$

About the distribution of h_1 one can assume that $h_1 \sim \mathcal{N}(0, \sigma_{\eta 1}^2)$ where $\sigma_{\eta 1}^2$ is a known variance, so that

$$\ln f(h_1) = -\frac{1}{2}\ln(2\pi\sigma_{\eta 1}^2) - \frac{1}{2\sigma_{\eta 1}^2}h_1^2.$$

Asymmetry in the volatility equation creates the most serious problems for the estimation of the extended model (26) compared to the basic SV model. The main reason for this is that $\ln f(h_t|y_{t-1}, h_{t-1})$ is no more quadratic in h_t, h_{t-1}.

7.2 QML estimation for the extended model

QML as described above is easily modified for the case of the basic SV-t model (see Ruiz (1994)). QML is based on the assumption that $\varepsilon_t = \ln(\xi_t^2)$ is approximately distributed as $\mathcal{N}(\mu_\varepsilon, \sigma_\varepsilon^2)$ with

correlated:

$$\begin{pmatrix} \xi_t \\ \eta_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

This alternative specification can be also written as

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + \kappa r(h_t) + \sigma_\xi(\xi_t + \alpha(h_t - \delta h_{t-1}))\exp(h_t/2),$$
$$h_t = \delta h_{t-1} + \sigma_\eta\eta_t.$$

See a discussion of timing issues and the corresponding empirical evidence in Yu (2005), Durham (2006). Overall, the difference between the two specifications is not very considerable.

$\mu_\varepsilon = \mathsf{E}\,\varepsilon_t$ and $\sigma_\varepsilon^2 = \mathsf{Var}\,\varepsilon_t$. For $\xi_t \sim t_\nu$ we can rewrite ε_t as $\varepsilon_t = \ln(\xi_t^2) = \ln x_1 - \ln(x_2/\nu)$ where x_1 and x_2 are independent, $x_1 \sim \chi_1^2$ and $x_2 \sim \chi_\nu^2$. This allows to calculate mean and variance of ε_t:

$$\mu_\varepsilon = \mathscr{C} - \psi(\nu/2) + \ln(\nu/2)$$

and

$$\sigma_\varepsilon^2 = \pi^2/2 + \psi'(\nu/2),$$

where $\mathscr{C} = \psi(1/2) - \ln(1/2) \approx -1.27036$, $\psi(\cdot)$ is the digamma function and $\psi'(\cdot)$ is the trigamma function.

Harvey & Shephard (1996) demonstrate how to take into account asymmetry when estimating the SV model by QML. Kirby (2006) propose a method which allows to account for asymmetric effects in several SV-type models. Using this logic model (26) with $\kappa = 0$ can be transformed into a linear state-space form as follows:

$$\begin{aligned} y_t - \mathbf{X}_t\boldsymbol{\beta} &= \sigma_\xi \xi_t \exp(h_t/2), \\ \ln((y_t - \mathbf{X}_t\boldsymbol{\beta})^2) &= 2\ln\sigma_\xi + h_t + \ln(\xi_t^2), \\ h_{t+1} &= \delta h_t + \alpha\xi_t + \sigma_\eta \eta_{t+1}. \end{aligned}$$

The regression coefficients $\boldsymbol{\beta}$ can be estimated before the other parameters. Then the Kalman filter can be used to calculate the quasi likelihood of the model. Note that the error term of the transition equation $\alpha\xi_t + \sigma_\eta\eta_{t+1}$ is correlated with the error terms of the two measurement equations (which are $\sigma_\xi\xi_t \exp(h_t/2)$ and $\ln(\xi_t^2) - \mathsf{E}\ln(\xi_t^2)$). This requires a variant of the Kalman filter with correlated errors.

7.3 Laplace's approximation

The log-density of the complete data for the extended model can be written as

$$\ln f(\mathbf{y}, \mathbf{h}) = \sum_{t=1}^{T} \ln\phi_t(y_t, y_{t-1}, h_t, h_{t-1}),$$

where

$$\ln\phi_t(y_t, y_{t-1}, h_t, h_{t-1}) = \ln f(y_t, h_t | y_{t-1}, h_{t-1}) = \ln f(y_t | h_t) + \ln f(h_t | y_{t-1}, h_{t-1}).$$

Each term $\ln\phi_t$ here depends only on h_t and h_{t-1}. This suggests that similar to the basic SV model the approximate log-density of complete data would be of the form (8). The corresponding multivariate Gaussian density $g(\mathbf{h}|\mathbf{y})$ can also be represented as a product of univariate conditional densities $g(h_t | h_{t-1}, \mathbf{y})$ each of them being univariate normal $\mathcal{N}(K_t + L_t h_{t-1}, M_t)$ for some K_t, L_t, M_t.

The idea is to approximate $\ln f(\mathbf{y}, \mathbf{h})$ by its quadratic expansion around some point $\mathbf{h}^*$:

$$\begin{aligned} \ln\phi_{at} = {} & F_t + F_t^0(h_t - h_t^*) + F_t^1(h_{t-1} - h_{t-1}^*) \\ & + \frac{1}{2}F_t^{00}(h_t - h_t^*)^2 + F_t^{01}(h_t - h_t^*)(h_{t-1} - h_{t-1}^*) + \frac{1}{2}F_t^{11}(h_{t-1} - h_{t-1}^*)^2, \end{aligned}$$

where we denote

$$F_t = \ln\phi_t\big|_{\mathbf{h}=\mathbf{h}^*}, \qquad F_t^i = \left.\frac{d\ln\phi_t}{dh_{t-i}}\right|_{\mathbf{h}=\mathbf{h}^*}, \qquad F_t^{ij} = \left.\frac{d^2\ln\phi_t}{dh_{t-i}h_{t-j}}\right|_{\mathbf{h}=\mathbf{h}^*}.$$

Table 7: *Laplace's approximation estimates of extended SV models, Example 1*

	Model I		Model II		Model III		Model IV	
	estimates	std. err.	estimates	std. err.	estimates	std. err.	estimates	std. err.
δ	0.9711	0.0061	0.9708	0.0061	0.9672	0.0065	0.9743	0.0058
σ_η	0.2516	0.0226	0.2554	0.0227	0.2559	0.0230	0.2190	0.0233
σ_ξ	2.1205	0.1575	2.1107	0.1574	2.1101	0.1428	2.0034	0.1528
α	—	—	—	—	−0.0376	0.0137	−0.0328	0.0120
κ	—	—	0.1763	0.0389	0.1619	0.0371	0.1718	0.0369
ν	—	—	—	—	—	—	16.901	6.0354
log-lik.	−7847.62	0.0012	−7836.20	0.0012	−7832.09	0.0072	−7828.81	0.0002

The analytical expressions for F_t, F_t^i and F_t^{ij} are given in Appendix D. Alternatively one can use numerical methods to evaluate the derivatives matrices F_t^i and F_t^{ij} if taking derivatives analytically turns out to be cumbersome.[34] From $\mathbf{h}^*$ we can get next a approximation $\mathbf{h}^{**}$ using (19). By iterating the procedure we get approximately the mode $\hat{\mathbf{h}}$ of $\ln f(\mathbf{y},\mathbf{h})$.

Appendix D provides formulas for obtaining coefficients B, B_t^0, B_t^{00} and B_t^{01} of approximation (8) from F_t, F_t^i and F_t^{ij}. Parameters K_t, L_t, M_t are obtained from B, B_t^0, B_t^{00} and B_t^{01} in the same way as for the basic SV model (see Appendix A).

Example 1 (continued). We estimated the basic model and several extended SV models for the RTSI series using the Laplace's approximation method. The in-mean effect is modeled as $r(h_t) = \exp(h_t/2)$. Table 7 shows the results. Both the in-mean and the leverage effects are significant at 1% level. There is also some evidence of fat-tailed innovations. (The log-likelihood estimates are discussed below). In the extended model with leverage effect the coefficient of correlation between $\alpha\xi_t + \sigma_\eta\eta_{t+1}$ and ξ_t is $\alpha/\sqrt{\alpha^2+\sigma_\eta^2}$. From the estimates of Model IV in Table 7 we get an estimate of −0.148 for this correlation coefficient.

7.4 Efficient importance sampling for the extended SV model

Richard & Zhang (2007) propose a piecemeal approach to fitting of a proposal distribution in high-dimensional models. Here we describe their approach in a somewhat more general form.

Suppose that we need to evaluate $I = \int \phi(\mathbf{x})d\mathbf{x}$ where $\mathbf{x}$ is T-dimensional. We assume that $\phi(\mathbf{x})$ can be factorized as $\phi(\mathbf{x}) = \prod_{t=1}^T \phi_t(\mathbf{x}_{\le t})$. (Here and below we use the following shortcut notation: $\mathbf{x}_{\le t} = (x_1,\dots,x_t)$ and $\mathbf{x}_{<t} = (x_1,\dots,x_{t-1})$). The functions $\phi_t(\mathbf{x}_{\le t})$ should be non-trivial as functions of x_t. (We use subscript t in ϕ_t to indicate that it is not assumed to be a legitimate probability density function). Conformably, the proposal distribution $\mu(\mathbf{x})$ can be factored as $\mu(\mathbf{x}) = \prod_{t=1}^T \mu(x_t|\mathbf{x}_{<t})$.

The piecemeal method runs backwards from T to 1, and for each observation t an elementary distribution $\mu(x_t|\mathbf{x}_{<t})$ is estimated. Suppose that we want to fit $\ln\mu(x_T|\mathbf{x}_{<T})$ to $\ln\phi_T(\mathbf{x}_{\le T})$. To do so it is important to add some function which would capture additional dependence on $\mathbf{x}_{<T}$. We will call this addition a stopgap function and denote $\ln\tilde{\mu}_T(\mathbf{x}_{<T})$. Because $\ln\tilde{\mu}_T(\mathbf{x}_{<T})$ is added to $\ln\mu(x_T|\mathbf{x}_{<T})$, it should be added to $\ln\phi_{T-1}(\mathbf{x}_{\le T-1})$. Therefore, for observation $T-1$ log-density $\ln\mu(x_{T-1}|\mathbf{x}_{<T-1})$ plus the stopgap function $\ln\tilde{\mu}_{T-1}(\mathbf{x}_{<T-1})$ should be fitted to $\ln\phi_{T-1}(\mathbf{x}_{\le T-1}) + \ln\tilde{\mu}_T(\mathbf{x}_{<T})$. In general a regression for $t = T,\dots,1$ is given by

$$\ln\phi_t(\mathbf{x}_{\le t}) + \ln\tilde{\mu}_{t+1}(\mathbf{x}_{\le t};\hat{\boldsymbol{\psi}}_{t+1}) = \ln\mu(x_t|\mathbf{x}_{<t};\boldsymbol{\psi}_t) + \ln\tilde{\mu}_t(\mathbf{x}_{<t};\boldsymbol{\psi}_t) + R_t, \tag{28}$$

[34] See Nocedal & Wright (2006) on methods of numerical differentiation. Durham (2006) use the Maple computer algebra system to analytically find derivatives for a more complicated SV-type model. Skaug & Yu (2007) propose to use automatic differentiation.

where $\hat{\boldsymbol{\psi}}_{t+1}$ are the estimates of the parameters which are already obtained for $t+1$. (At the start, for $t=T$, we set $\ln\tilde{\mu}_{T+1}(\mathbf{x}_{\leq T})=0$). The parameters estimates $\hat{\boldsymbol{\psi}}_t$ are found using this nonlinear regression.

In order to obtain an "efficient" Gaussian proposal distribution $g(\mathbf{h}|\mathbf{y})$ for the SV model we assume that $g(h_t|\mathbf{y},\mathbf{h}_{<t})$ for $t=1,\ldots,T$ are normal, that the mean depends linearly on h_{t-1} so that

$$h_t|\mathbf{y},\mathbf{h}_{<t}\sim\mathcal{N}(K_t+L_t h_{t-1},M_t)$$

and that the logarithm of stopgap, $\ln\tilde{\mu}_t$, is a quadratic function of h_{t-1}.

Then the regression (28) can be rewritten for $t=2,\ldots,T$ as

$$\ln\phi_t+\ln\tilde{\mu}_{t+1}=D_t+D_t^0h_t+D_t^1h_{t-1}+D_t^{00}h_t^2+D_t^{01}h_th_{t-1}+D_t^{11}h_{t-1}^2+R_t. \tag{29}$$

For $t=1$ the regression is simply

$$\ln\phi_1+\ln\tilde{\mu}_2=D_1+D_1^0h_1+D_1^{00}h_1^2+R_1. \tag{30}$$

The parameters K_t, L_t and M_t can be recovered from the coefficients of the EIS regressions by equating the coefficients of h_t^2, h_t and h_th_{t-1} to that in (9). It follows that

$$M_t=-\frac{1}{2D_t^{00}},\qquad K_t=D_t^0M_t,\qquad L_t=D_t^{01}M_t.$$

The value of stopgap function is obtained after estimation of period t regression as

$$\ln\tilde{\mu}_t=\ln\phi_t+\ln\tilde{\mu}_{t+1}-\ln g(h_t|\mathbf{y},\mathbf{h}_{<t})-R_t,$$

where R_t are the residuals from the regression.

The EIS method is started from some proposal distribution described by K_t, L_t and M_t. Generated trajectories $\mathbf{h}^s$ provide data points for EIS regressions. The regressions produce new K_t, L_t and M_t. Several iterations are made to achieve convergence.

Example 1 (continued). We apply the Monte Carlo method with $S=10000$ simulations to estimate the log-likelihood for the estimates in Table 7. The proposal distribution is obtained by the EIS method with $S=1000$ simulations and 10 iterations. The estimates with corresponding standard errors due to simulation are given in the last row of the table. These results confirm that in-mean and leverage effects are significant. The likelihood ratio statistics are

$$\begin{aligned}LR(\text{model I against model II})&=22.84\ [<10^{-5}],\\ LR(\text{model II against model III})&=8.22\ [0.0041],\\ LR(\text{model III against model IV})&=6.56\ [0.0104].\end{aligned}$$

P-values from χ_1^2 distribution are in square brackets. The last p-value is not reliable as ν for the normal distribution is $+\infty$, which is the right boundary of admissible values for SV model with the Student's t distribution. In any case the use of the Student's t distribution is helpful as it improves the quality of the proposal distribution. For model IV ESS is 2894.8 while for model III it is as low as 137.2.

8 Smoothing, filtering and forecasting

8.1 Introduction

An important task in SV modeling is inferring information on $\mathbf{h}$ from $\mathbf{y}$. In other words, one can be interested in the distribution of the latent state $\mathbf{h}$ conditional on the observable data $\mathbf{y}$. The

calculation of various characteristics of $\mathbf{h}|\mathbf{y}$ is generically called smoothing. We already discussed finding the mode of $\mathbf{h}|\mathbf{y}$. However, other characteristics like $\mathsf{E}(\mathbf{h}|\mathbf{y})$ or quantiles of $\mathbf{h}|\mathbf{y}$ can be also of interest. Monte Carlo simulations can be used for the task of smoothing the latent state of SV model.

Filtering refers to exploring characteristics of a sequence of conditional distributions $\mathbf{h}_{\leq t}|\mathbf{y}_{\leq t}$, where $t = 1,2,\ldots$ Filtering imitates inference in the situation of sequential flow of information. If we know the observable variable up to time t, $\mathbf{y}_{\leq t}$, we can explore $\mathbf{h}_{\leq t}|\mathbf{y}_{\leq t}$. With the arrival of the next observation y_{t+1} we can explore $\mathbf{h}_{\leq t+1}|\mathbf{y}_{\leq t+1}$, and so on.

Filtering can be useful for on-line inference in the SV model (for example, for monitoring of the current latent state). The results of on-line filtering can be used for on-line forecasting and hence for financial decision-making. (Of course, this rises the problems of updating parameters estimates and obtaining approximating functions $g(\mathbf{h}_{\leq t}|\mathbf{y}_{\leq t})$ in a sequential manner). Some applications could require imitation of on-line forecasting (for example, in order to estimate the behavior of the implied forecast uncertainty).

Forecasting in the SV model is closely related to smoothing and filtering and can be implemented by means of Monte Carlo simulation.

An important use of filtering is for obtaining residuals from one-step-ahead forecasts for the purpose of model diagnostic checking. This is by far the most popular approach to SV model diagnostics (and also to diagnostics of time series models in general). Multistep forecasts can also be used for diagnostics, but there is a problem of serial dependence.

We discuss the tasks of smoothing, filtering and forecasting under the assumption that the vector of parameters $\boldsymbol{\theta}$ is known. In practice one would substitute some suitable estimate (for example, an estimate obtained from simulated maximum likelihood method). Of course, the consequences of this substitution can be not very innocuous for short series. The methods of taking into account parameters uncertainty are yet to be developed.[35]

A discussion of smoothing, filtering and forecasting in the non-linear non-Gaussian state-space models by means of importance sampling can be found, for example, in Tanizaki (2003), Creal (2009).

8.2 Smoothing

The posterior distribution $\mathbf{h}|\mathbf{y}$ is not known in a closed form. We only know $f(\mathbf{y},\mathbf{h})$ which (as a function of $\mathbf{h}$) is proportional to $f(\mathbf{h}|\mathbf{y})$. The knowledge of $f(\mathbf{y},\mathbf{h})$ allows to apply the importance sampling to the task of smoothing.

If $\boldsymbol{\tau}(\mathbf{h})$ is some function of $\mathbf{h}$ then its expected value is

$$\mathsf{E}\big(\boldsymbol{\tau}(\mathbf{h})|\mathbf{y}\big) = \int \boldsymbol{\tau}(\mathbf{h}) f(\mathbf{h}|\mathbf{y}) d\mathbf{h} = \frac{1}{f(\mathbf{y})} \int \boldsymbol{\tau}(\mathbf{h}) f(\mathbf{y},\mathbf{h}) d\mathbf{h} = \frac{\int \boldsymbol{\tau}(\mathbf{h}) f(\mathbf{y},\mathbf{h}) d\mathbf{h}}{\int f(\mathbf{y},\mathbf{h}) d\mathbf{h}}. \tag{31}$$

After estimation of a SV model we have a density function $g(\mathbf{h}|\mathbf{y})$ which is an approximation to $f(\mathbf{h}|\mathbf{y})$. Rewrite the expectation in terms of $g(\mathbf{h}|\mathbf{y})$ as

$$\mathsf{E}\big(\boldsymbol{\tau}(\mathbf{h})|\mathbf{y}\big) = \frac{\int \boldsymbol{\tau}(\mathbf{h}) v(\mathbf{h};\mathbf{y}) g(\mathbf{h}|\mathbf{y}) d\mathbf{h}}{\int v(\mathbf{h};\mathbf{y}) g(\mathbf{h}|\mathbf{y}) d\mathbf{h}} = \frac{\mathsf{E}_g[\boldsymbol{\tau}(\mathbf{h}) v(\mathbf{h};\mathbf{y})]}{\mathsf{E}_g[v(\mathbf{h};\mathbf{y})]},$$

where

$$v(\mathbf{h};\mathbf{y}) = \frac{f(\mathbf{y},\mathbf{h})}{g(\mathbf{h}|\mathbf{y})}.$$

This expectation can be estimated by means of Monte Carlo as a weighted average

$$\mathsf{E}\big(\boldsymbol{\tau}(\mathbf{h})|\mathbf{y}\big) \approx \bar{\boldsymbol{\tau}} = \frac{\sum_{s=1}^{S} \boldsymbol{\tau}(\mathbf{h}^s) v(\mathbf{h}^s;\mathbf{y})}{\sum_{s=1}^{S} v(\mathbf{h}^s;\mathbf{y})}$$

[35] One possibility is to use Bayesian approach with "uninformative" prior.

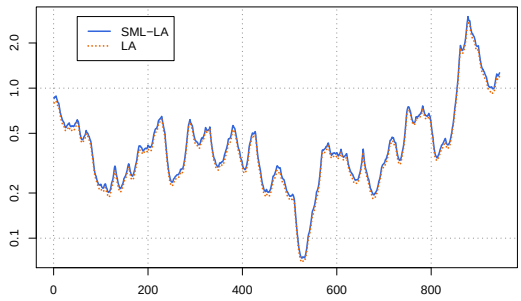

Figure 8: *Smoothed value of the conditional variance from Monte Carlo simulations (solid) and the Laplace's approximation (dotted), Example 2.*

with $\mathbf{h}^s \rightsquigarrow g(\mathbf{h}|\mathbf{y})$. In terms of normalized weights

$$w^s = w(\mathbf{h}^s;\mathbf{y}) = \frac{v(\mathbf{h}^s;\mathbf{y})}{\sum_{k=1}^{S} v(\mathbf{h}^k;\mathbf{y})}$$

the estimate can be rewritten as

$$\mathsf{E}\left(\boldsymbol{\tau}(\mathbf{h})|\mathbf{y}\right) \approx \bar{\boldsymbol{\tau}} = \sum_{s=1}^{S} \boldsymbol{\tau}(\mathbf{h}^s)\, w^s. \tag{32}$$

The method of importance sampling essentially approximates the posterior distribution $\mathbf{h}|\mathbf{y}$ by a discrete distribution which associates probability w^s with trajectory $\mathbf{h}^s$ from a finite set of trajectories $\{\mathbf{h}^1,\ldots,\mathbf{h}^S\}$.[36] Theoretical moments are estimated by weighted sample moments (which are theoretical moments for an approximating discrete distribution).

Example 2 (continued). We use the exchange rates example to estimate the expected conditional variance from the smoothing distribution, $\sigma_\xi^2 \mathsf{E}[\exp(h_t)|\mathbf{y}]$. We take the SML-LA estimates of the basic SV model from Table 3 and use the corresponding proposal distribution to make 10000 Monte Carlo simulations for smoothing purposes. Figure 8 plots the estimate and compares it with a similar estimate from the parent Laplace's approximation without Monte Carlo defined as (15). The two series are fairly close.

Quantiles of the posterior distribution $h_t|\mathbf{y}$ can be estimated from a sorted[37] Monte Carlo sample $h_t^{(1)} < h_t^{(2)} < \cdots < h_t^{(S)}$ with associated weights $w_t^{(s)}$. A possible estimate of p-quantile is $h_t^{(k)}$ for which

$$\sum_{s=1}^{k-1} w_t^{(k)} < p < \sum_{s=1}^{k} w_t^{(k)}.$$

Example 3 (continued). We take the EIS estimates of the basic SV model from Table 5 and the corresponding proposal distribution to find the 0.05 and 0.95 quantiles with $S = 10000$ simulations. The results are shown in Figure 9 together with the actual realization of the conditional variance from Figure 3(a). This is analogous to Figure 6(b) for QML.

[36] A Markov chain Monte Carlo (MCMC) algorithm can also be used to generate from the posterior distribution $\mathbf{h}|\mathbf{y}$. (See Tierney (1994), Chib & Greenberg (1996), Gentle (2003), Rubinstein & Kroese (2008) for a discussion of MCMC.) For some proposal p.d.f. $g(\mathbf{h}|\mathbf{y})$ approximating the unknown posterior p.d.f. $f(\mathbf{h}|\mathbf{y})$ one can use so called independence chain algorithm which is a simple variant of Metropolis–Hastings algorithm. MCMC can produce a set of trajectories $\mathbf{h}_1,\ldots,\mathbf{h}_S$ which are almost independent of each other and are distributed approximately according to $f(\mathbf{h}|\mathbf{y})$. Then one can approximate the posterior distribution by a discrete distribution which associates probability $1/S$ with trajectory $\mathbf{h}^s$. Similarly to the importance sampling theoretical moments are estimated by sample moments. See Liesenfeld & Richard (2006) for a discussion of parallels between the importance sampling and the Metropolis–Hastings algorithm. Liesenfeld & Richard (2006) following Tierney (1994) propose to enhance the independence chain by an accept/reject step.

[37] Sorting requires $O(S \ln S)$ operations which can be large for large S. There are faster methods of finding weighted sample quantiles, but we do not consider them in this essay.

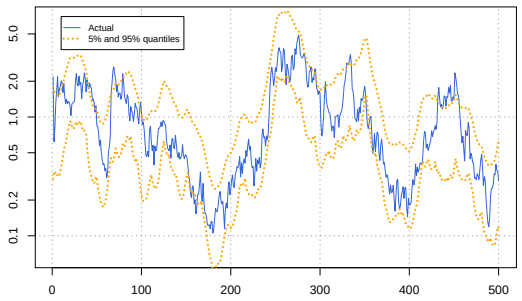

Figure 9: *Confidence band from Monte Carlo smoothing estimates based on EIS method and actual conditional variance, Example 3.*

To reduce the number of simulations S one needs to reduce the variance of $\bar{\boldsymbol{\tau}}$. Some improvement can be obtained by choosing $g(\mathbf{h}|\mathbf{y})$ to be an accurate approximation to $f(\mathbf{h}|\mathbf{y})$. For example, one can use EIS at this step even if it was not used during estimation of SV parameters. However, in general the variance of $\bar{\boldsymbol{\tau}}$ is not zero here even when $g(\mathbf{h}|\mathbf{y}) = f(\mathbf{h}|\mathbf{y})$ exactly (that is, when all w^s are equal to $1/S$).

One can use various other variance reduction techniques (like control variates) to reduce the number of simulations. However, such techniques are less fruitful than fitting $g(\mathbf{h}|\mathbf{y})$ to $f(\mathbf{h}|\mathbf{y})$.

8.3 Filtering

The basic formula for filtering is the same as for smoothing (see (31))

$$\mathsf{E}\left(\boldsymbol{\tau}_t(\mathbf{h}_{\le t})|\mathbf{y}_{\le t}\right) = \frac{\int \boldsymbol{\tau}_t(\mathbf{h}_{\le t}) f(\mathbf{y}_{\le t}, \mathbf{h}_{\le t}) d\mathbf{h}_{\le t}}{\int f(\mathbf{y}_{\le t}, \mathbf{h}_{\le t}) d\mathbf{h}_{\le t}}.$$

A expectation is approximated as a weighted average

$$\bar{\boldsymbol{\tau}}_{wt} = \sum_{s=1}^{S} \boldsymbol{\tau}_t(\mathbf{h}^s_{\le t})\, w^s_t \tag{33}$$

with $\mathbf{h}^s_{\le t} \rightsquigarrow g(\mathbf{h}_{\le t}|\mathbf{y}_{\le t})$ and weights given by

$$v^s_t = v(\mathbf{h}^s_{\le t}; \mathbf{y}_{\le t}) = f(\mathbf{y}_{\le t}, \mathbf{h}^s_{\le t}) / g(\mathbf{h}^s_{\le t}|\mathbf{y}_{\le t})$$

and

$$w^s_t = \frac{v^s_t}{\sum_{k=1}^{S} v^k_t}. \tag{34}$$

Note that for filtering we have to use a family of proposal distributions $g(\mathbf{h}_{\le t}|\mathbf{y}_{\le t})$ indexed by t. For (33) to be a good enough approximation for moments of the filtering distribution it is desirable to use a proposal density $g(\mathbf{h}_{\le t}|\mathbf{y}_{\le t})$ which is approximately proportional to the filtering density $f(\mathbf{h}_{\le t}|\mathbf{y}_{\le t})$ (in other words, to the density $f(\mathbf{y}_{\le t}, \mathbf{h}_{\le t})$ viewed as a function of $\mathbf{h}_{\le t}$). Thus, a full filtering procedure consists of choosing each $g(\mathbf{h}_{\le t}|\mathbf{y}_{\le t})$ to approximate $f(\mathbf{h}_{\le t}|\mathbf{y}_{\le t})$ and then using (33) for $t = 1, \ldots, T$. This amounts to applying smoothing as described above to a sequence of time series $(\mathbf{y}_{\le t})$, $t = 1, \ldots, T$. Of course, the full procedure can be very time-consuming. Since each smoothing step requires $O(tS)$ operations, the full filtering procedure requires $O(T^2 S)$ operations for a series of length T.

A less time-consuming procedure can be based on a single distribution $g(\mathbf{h}|\mathbf{y}) = g(\mathbf{h}_{\le T}|\mathbf{y}_{\le T})$. The distribution can be presented recursively:

$$g(\mathbf{h}_{\le t}|\mathbf{y}) = g(h_t|\mathbf{y}, \mathbf{h}_{<t}) g(\mathbf{h}_{<t}|\mathbf{y}).$$

We assume that it is possible to directly generate h_t from $g(h_t|\mathbf{y},\mathbf{h}_{<t})$. The proposal distribution for time t is just $g(\mathbf{h}_{\le t}|\mathbf{y}_{\le t}) = g(\mathbf{h}_{\le t}|\mathbf{y})$. (In what follows we simplify our notation by omitting the dependence of the proposal distribution on the full length of observed data $\mathbf{y} = \mathbf{y}_{\le T}$). If trajectories $\mathbf{h}^s_{<t}$, $s = 1,\dots,S$ are already generated from $g(\mathbf{h}_{<t})$ then it is possible to append these trajectories: $\mathbf{h}^s_{\le t} = (\mathbf{h}^s_{<t}, h^s_t)$, where $h^s_t \rightsquigarrow g(h_t|\mathbf{h}^s_{<t})$. Noting that $f(\mathbf{y}_{\le t},\mathbf{h}_{\le t})$ can be represented recursively as

$$f(\mathbf{y}_{\le t},\mathbf{h}_{\le t}) = f(y_t,h_t|\mathbf{y}_{<t},\mathbf{h}_{<t})f(\mathbf{y}_{<t},\mathbf{h}_{<t}),$$

we see that it is possible to evaluate the unnormalized weights of the trajectories recursively:

$$v^s_t = \frac{f(\mathbf{y}_{\le t},\mathbf{h}^s_{\le t})}{g(\mathbf{h}^s_{\le t})} = \frac{f(y_t,h^s_t|\mathbf{y}_{<t},\mathbf{h}^s_{<t})f(\mathbf{y}_{<t},\mathbf{h}^s_{<t})}{g(h^s_t|\mathbf{h}^s_{<t})g(\mathbf{h}^s_{<t})} = \frac{f(y_t,h^s_t|\mathbf{y}_{<t},\mathbf{h}^s_{<t})}{g(h^s_t|\mathbf{h}^s_{<t})}v^s_{t-1}$$

or simply

$$v^s_t = u^s_t v^s_{t-1},$$

where $u^s_t = f(y_t,h^s_t|\mathbf{y}_{<t},\mathbf{h}^s_{<t})/g(h^s_t|\mathbf{h}^s_{<t})$ are called the incremental weights. The recursion for the weights is started with $v^s_1 = u^s_1 = f(y_1,h^s_1)/g(h^s_1)$.

The approach can be described as follows: initially a set of trajectories $\mathbf{h}^s \rightsquigarrow g(\mathbf{h})$ is generated and then only the weights are updated recursively.

The problem with a single proposal distribution is that it would be adapted to the series of length T. For arbitrary t the quality of approximation could be inferior with a very non-uniform distribution of weights. This can be measured by the effective sample size

$$\mathrm{ESS}_t = \frac{1}{\sum_{s=1}^S (w^s_t)^2}.$$

A partial remedy for the problem of inadequacy of a single proposal distribution can be proposed. The proposal distribution can be adapted to current t by tuning the conditional distributions corresponding to several last observations, $t-K+1,\dots,t$, and using these modified proposal distributions to replace the K last observations in the simulated trajectories. We only consider $K = 1$ case. We take $g(\mathbf{h}_{\le t}|\mathbf{y}_{\le t}) = g(h_t|\mathbf{y}_{\le t},\mathbf{h}_{<t})g(\mathbf{h}_{<t}|\mathbf{y})$ where $g(h_t|\mathbf{y}_{\le t},\mathbf{h}_{<t})$ is tuned in such a way, that $g(\mathbf{h}_{\le t}|\mathbf{y}_{\le t})$ is a better proposal distribution for the task of filtering at time t. For the methods we considered earlier (of which EIS is the most useful for the task of obtaining a good proposal distribution) this does not lead to $O(tS)$ computation complexity for time t. Only $O(S)$ operations are required for one t (and $O(KS)$ if lag K is used). Then importance weights for time t are

$$\check{v}^s_t = \frac{f(\mathbf{y}_{\le t},\check{\mathbf{h}}^s_{\le t})}{g(\check{\mathbf{h}}^s_{\le t}|\mathbf{y}_{\le t})} = \frac{f(y_t,\check{h}^s_t|\mathbf{y}_{<t},\mathbf{h}^s_{<t})f(\mathbf{y}_{<t},\mathbf{h}^s_{<t})}{g(\check{h}^s_t|\mathbf{y}_{\le t},\mathbf{h}^s_{<t})g(\mathbf{h}^s_{<t})} = \frac{f(y_t,\check{h}^s_t|\mathbf{y}_{<t},\mathbf{h}^s_{<t})}{g(\check{h}^s_t|\mathbf{y}_{\le t},\mathbf{h}^s_{<t})}v^s_{t-1},$$

where $\check{h}^s_t \rightsquigarrow g(h_t|\mathbf{y}_{\le t},\mathbf{h}^s_{<t})$, $\mathbf{h}^s_{<t} \rightsquigarrow g(\mathbf{h}_{<t})$ and $\check{\mathbf{h}}^s_{\le t} = (\check{h}^s_t,\mathbf{h}^s_{<t})$ or

$$\check{v}^s_t = \check{u}^s_t v^s_{t-1},$$

where $\check{u}^s_t = f(y_t,\check{h}^s_t|\mathbf{y}_{<t},\mathbf{h}^s_{<t})/g(\check{h}^s_t|\mathbf{y}_{\le t},\mathbf{h}^s_{<t})$. This approach is fruitful, because the filtering proposal distributions $g(\mathbf{h}_{\le t}|\mathbf{y}_{\le t})$ usually differ appreciably from the smoothing proposal distribution $g(\mathbf{h}_{\le t}|\mathbf{y})$ only for a few last observations.[38]

[38] Liesenfeld & Richard (2003) note similar proximity of $g(\mathbf{h}_{<t}|\mathbf{y}_{<t})$ and $g(\mathbf{h}_{\le t}|\mathbf{y}_{\le t})$ proposal distributions obtained by EIS.

8.4 Forecasting

Assume that the distributions of $y_t|y_{t-1},h_T$ and $h_t|h_{t-1},y_{t-1}$ are determined by the model and there is an algorithm to generate random variables from these distributions. Then given $\mathbf{y}$ and h_T one can generate future values y^s_{T+1}, h^s_{T+1}, y^s_{T+2}, h^s_{T+2},... sequentially, where s is an index of a trajectory. This produces a Monte Carlo sample of forecasting trajectories $(\mathbf{y}^s_{>T},\mathbf{h}^s_{>T})$ generated according to $f(\mathbf{y}_{>T},\mathbf{h}_{>T}|\mathbf{y},h^s_{T-1},h^s_T)$.

Of course, one should start the recursion from some h^s_T. This can be the last element of vector $\mathbf{h}^s$ generated according to $g(\mathbf{h}|\mathbf{y})$. Because we draw $\mathbf{h}^s$ from an approximation $g(\mathbf{h}|\mathbf{y})$ instead of true $f(\mathbf{h}|\mathbf{y})$, the generated forecasting trajectories $(\mathbf{y}^s_{>T},\mathbf{h}^s_{>T})$ have associated unequal importance weights w^s. When estimating an expectation of some function of $(\mathbf{y}_{>T},\mathbf{h}_{>T})$ by sample mean (that is, when using the importance sampling), one should use the weighted sample mean with weights w^s.

For a sample of future trajectories $(\mathbf{y}^s_{>T},\mathbf{h}^s_{>T})$, $s=1,\dots,S$ with importance weights $\{w^s\}$ one can estimate various forecast statistics like point forecasts, interval forecasts and so on. For example, to get an interval forecast for $Y_H=\sum_{i=1}^H y_{t+i}$ one simulates a sample of Y^s_H and calculates the relevant sample quantiles.

If $\boldsymbol{\tau}(\mathbf{y}_{>T},\mathbf{h}_{>T})$ is some function of a future trajectory then its expected value can be written as

$$\mathsf{E}\left(\boldsymbol{\tau}(\mathbf{y}_{>T},\mathbf{h}_{>T})|\mathbf{y}\right)=\int \boldsymbol{\tau}(\mathbf{y}_{>T},\mathbf{h}_{>T})f(\mathbf{h}|\mathbf{y})f(\mathbf{y}_{>T},\mathbf{h}_{>T}|\mathbf{h})d(\mathbf{h},\mathbf{y}_{>T},\mathbf{h}_{>T}).$$

Similarly to smoothing and filtering this expectation can be estimated by means of Monte Carlo as a weighted average

$$\mathsf{E}\left(\boldsymbol{\tau}(\mathbf{y}_{>T},\mathbf{h}_{>T})|\mathbf{y}\right)\approx\bar{\boldsymbol{\tau}}=\sum_{s=1}^S\boldsymbol{\tau}(\mathbf{y}^s_{>T},\mathbf{h}^s_{>T})w^s,$$

where $\mathbf{h}^s\backsim g(\mathbf{h}|\mathbf{y})$, $(\mathbf{y}^s_{>T},\mathbf{h}^s_{>T})\backsim f(\mathbf{y}_{>T},\mathbf{h}_{>T}|\mathbf{h}^s)$ and $\{w^s\}$ are corresponding normalized importance weights.

One can also produce interval forecasts from weighted sample quantiles (see a description of possible algorithm above, in subsection 8.2 on smoothing).

Example 1 (continued). We illustrate dynamic forecasting in the context of the SV model using the RTSI data. We forecast dynamically for horizons $H=1,\dots,200$ at two different dates, January 30, 2007 and January 30, 2009. The estimates are obtained by the Laplace's approximation method for the shortened series. The proposal distribution is obtained by EIS. What we want to forecast is not the return y_{T+H}, but the level stock index itself. For a sample of initial Monte Carlo trajectories we can obtain Monte Carlo trajectories of RTSI as

$$RTSI^s_{T+H}=RTSI_T\exp\Big(\sum_{i=1}^H y^s_{T+i}/100\Big).$$

The interval forecasts are the 10% and 90% weighted sample quantiles of $RTSI^s_{T+H}$. Figure 10 shows the results.

8.5 SV model diagnostics

Denote the c.d.f. of forecast distribution $y_t|y_1,\dots,y_{t-H}$ by $F(y_t|y_1,\dots,y_{t-H})$. If the model is correct then $v_{t,H}=F(y_t|y_1,\dots,y_{t-H})$ is uniformly distributed $U[0,1]$. This is called the probability integral transform (PIT). For $H=1$ the series $v_t=v_{t,1}=F(y_t|y_1,\dots,y_{t-1})$ should be independent. For $H>1$ one can use $v_{t,H}$, but the series in general would be dependent. It can be useful to convert $v_{t,H}$ to the standard normal form $z_{t,H}=\Phi^{-1}(v_{t,H})$ were $\Phi(\cdot)$ is the standard normal c.d.f. as many diagnostic tests have more power under normality. Also useful is the "folded" PIT $v'_{t,H}=\left|2v_{t,H}-1\right|$ and corresponding $z'_{t,H}=\Phi^{-1}(v'_{t,H})$ which should be distributed as $U[0,1]$ and $\mathcal{N}(0,1)$ respectively.

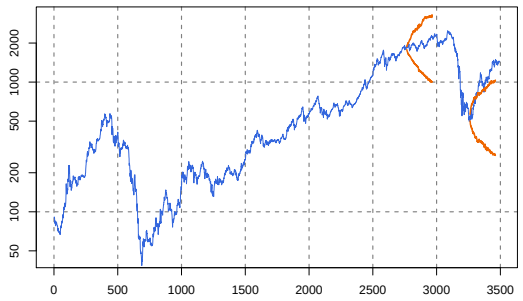

Figure 10: *Interval forecasts of RTSI, January 30, 2007 and January 30, 2009.*

See Diebold et al. (1998) for a general discussion. PIT-based tests are used in Kim et al. (1998), Liesenfeld & Richard (2003), Durham (2006) for the purpose of SV model diagnostics.

In Monte Carlo forecasting described above the forecast distribution is approximated by a discrete distribution produced from Monte Carlo sample (with associated weights). If a forecast of y_t is made at time $t-H$ then we denote an element of Monte Carlo forecast sample by $y^s_{t|t-H}$ and its normalized weight by w^s_{t-H}. A simple estimate of $v_{t,H}$ is given by

$$\sum_{s=1}^{S} w^s_{t-H} I(y^s_{t|t-H} < y_t),$$

where $I(A)$ is a 0/1 indicator of condition A. A better estimate can be obtained by averaging the theoretical probabilities $\Pr(y^s_{t|t-H} < y_t | h^s_{t|t-H})$ instead of 0/1 indicator. These probabilities are determined by the model (26):

$$\Pr(y^s_{t|t-H} < y_t | h^s_{t|t-H}) = \Pr(\xi_t < \xi_t(y_t, h^s_{t|t-H})) = F(\xi_t(y_t, h^s_{t|t-H})),$$

where function $\xi_t(y_t, h_t)$ is defined in (27) and $F(\xi_t)$ is cumulative distribution function of ξ_t (standard normal or Student's t). The estimate of $v_{t,H}$ is given by

$$\hat{v}_{t,H} = \sum_{s=1}^{S} w^s_{t-H} F(\xi_t(y_t, h^s_{t|t-H})).$$

For diagnostic purposes we need to obtain a series of $\hat{v}_{t,H}$ for $t = H+1, \ldots, T$. This is done by applying the filtering procedure discussed earlier.

One can use the PIT series for various diagnostic tests. The most important uses are detecting autocorrelation, autoregressive conditional heteroskedasticity and violation of distributional assumptions. Also PIT-based diagnostics can help to check "calibration" of density forecasts in general; see Gneiting et al. (2007). Folded PIT corresponds to even moments and can help to reveal fat tails, autoregressive conditional heteroskedasticity and lack of forecast calibration.

Example 1 (continued). We apply PIT-based diagnostics to the estimates obtained by the LA method for the basic SV model and the RTSI data. The proposal distribution is obtained by EIS with $S = 100$. Approximations to the forecast distributions are obtained from $S = 10000$ simulations. Figure 11 shows some graphical results. The histogram of v_t series shows inadequate calibration: the distribution is somewhat biased to the right. The correlogram of $z_{t,1}$ series reveals significant first-order autocorrelation. This agrees with Figure 2(a) as the basic SV model cannot capture autocorrelation. The correlogram of $z'_{t,1}$ series does not reveal autocorrelation. This correlogram can be confronted with the correlogram of $|y_t|$ in Figure 2(b) (which reveals volatility clustering). The comparison suggests that the basic SV model adequately captures volatility dynamics.

We also apply several more formal PIT-based diagnostic tests. The following notation is used: m_k is k-th central moment of $z_t = z_{t,1}$, $\bar{z}$ is sample mean of z_t and $\tilde{T}$ is the number of observations.

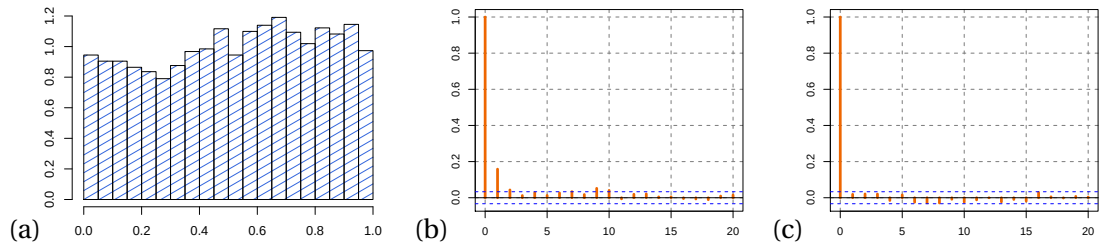

Figure 11: *PIT-based diagnostics for the basic SV model, Example 1: (a) histogram of PIT v_t series; (b) correlogram of PIT $z_{t,1}$ series; (c) correlogram of PIT $z'_{t,1}$ series.*

1. Statistic $\bar{z}/\sqrt{m_2}\cdot\sqrt{\tilde{T}}$ is approximately distributed as $\mathcal{N}(0,1)$ and can help to detect bias in one-step-ahead forecasts. For the current example it is 3.97 with p-value less than 0.1%. Hence, there is an upward bias in the forecast distribution of the model.

2. A similar statistic for $z'_t = z'_{t,1}$ can help to detect whether the forecast distribution is too sharp or too fuzzy. For the current example it is -1.13 which is not significant at 20% level. Hence, there are no signs of inadequacy in this aspect of forecasts calibration.

3. Statistic $m_3/m_2^{3/2}\cdot\sqrt{\tilde{T}/6}$ (based on the skewness coefficient $m_3/m_2^{3/2}$) is approximately distributed as $\mathcal{N}(0,1)$ and can help to detect unmodeled asymmetry in the distribution of model innovations. For the current example it is -4.35 with p-value less than 0.1%. The distribution is visibly asymmetric.

4. Statistic $(m_4/m_2^2-3)\cdot\sqrt{\tilde{T}/24}$ (based on kurtosis coefficient m_4/m_2^2) is approximately distributed as $\mathcal{N}(0,1)$ and can help to detect unmodeled kurtosis in the distribution of model innovations. For the current example it is 3.16 with p-value less than 1%. There are signs of fat-tailness.

5. Ljung–Box statistic $Q=\tilde{T}\left(\tilde{T}+2\right)\sum_{i=1}^{k} r_i^2/(\tilde{T}-i)$ based on a.c.f. r_i for z_t is approximately distributed as $\chi^2(k)$ and can help to detect unmodeled autocorrelation. For the current example Q for $k=10$ autocorrelation coefficients is 120.9 with p-value less than 0.1%. The autocorrelation is rather significant.

6. Ljung–Box statistic based on a.c.f. for z'_t can help to detect unmodeled autoregressive conditional heteroskedasticity. For the current example Q for $k=10$ autocorrelation coefficients is 16.2 which is not significant at 10% level.

A word of caution should be said about the use of PIT-based test statistics. Actually little is known about their asymptotic distribution. The distributions and p-values mentioned here are only rough approximations.

We can conclude that the basic SV model is not quite adequate for the RTSI data. We need to model the conditional mean, not only the conditional variance. Diagnostic tests suggest that the distribution for innovations should be skewed and with somewhat fatter tails.

9 Other extensions of SV model

One can find numerous extensions of the basic SV model in the literature. We would not attempt to provide a representative survey in this essay. We just mention some interesting directions.

An SV model with multiple factors instead of a single latent factor h_t in (1) can be used as an alternative to the SV-t model and as a way to approximate long-range dependence. For example, see Liesenfeld & Richard (2003), Durham (2006), Jungbacker & Koopman (2009). Usually two factors are used, one of which is highly persistent.

Continuous time models with jumps are popular in the mathematical finance literature. Discrete time stochastic volatility models with jumps can be obtained by discretization of these continuous time models; for example, see Chernov et al. (1999), Eraker et al. (2003). Chib et al. (2002) deal with a discrete time formulation from the start. Jumps can be added to the innovations of the mean equation and can capture fat tails. Jumps in the innovations of the volatility equation can also be important.

For some (long enough) financial series a slow decay in sample autocorrelation function of absolute returns is observed. This can be captured by a long memory process for h_t such as ARFIMA. See Breidt et al. (1998), Harvey (2007), Brockwell (2007), Hurvich & Soulier (2009) among others. These models are analogues of GARCH-type long memory models (for a discussion of such models see Davidson (2004)). Harvey et al. (1994), Ruiz (1994) consider a random walk specification for h_t which can be likened to IGARCH.

In this essay we discussed only univariate SV models. Yet in the context of financial time series joint analysis of several series can provide some benefits. This is documented by the huge literature on multivariate GARCH-type modeling. Behavior of financial time series can exhibit a large degree of mutual correlation. First, these correlations can be important for various financial applications like portfolio management. Second, joint modeling increases statistical efficiency. Third, one can explore whether the joint behavior of multiple series is driven by a much smaller number of underlying factors and try to uncover those factors. Multivariate SV models were studied and/or surveyed in Harvey et al. (1994), Danielsson (1998), Liesenfeld & Richard (2003), Asai et al. (2006), Yu & Meyer (2006), Chib et al. (2009) among others.

SV model is similar to other models which contain an unobservable factor described by the first-order autoregression. Some of the methods for such models are also similar. These include stochastic conditional duration (Bauwens & Veredas (2004)) and "parameter-driven" dynamic count data models (for example, see Zeger (1988) and Jung et al. (2006)).

References

Andersen, T. G. (1994). Stochastic Autoregressive Volatility: A Framework for Volatility Modeling. *Mathematical Finance* 4, 75–102.

Andersen, T. G., H.-J. Chung & B. E. Sørensen (1999). Efficient Method of Moments Estimation of a Stochastic Volatility Model: A Monte Carlo Study. *Journal of Econometrics* 91, 61–87.

Andersen, T. G. & B. E. Sørensen (1996). GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study. *Journal of Business and Economic Statistics* 14, 328–352.

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica* 59, 817–858.

Asai, M. & M. McAleer (2005). Dynamic Asymmetric Leverage in Stochastic Volatility Models. *Econometric Reviews* 24, 317–332.
URL http://www.informaworld.com/10.1080/07474930500243035

Asai, M., M. McAleer & J. Yu (2006). Multivariate Stochastic Volatility: A Review. *Econometric Reviews* 25, 145–175.
URL http://www.informaworld.com/10.1080/07474930600713564

Bauwens, L. & D. Veredas (2004). The Stochastic Conditional Duration Model: A Latent Variable Model for the Analysis of Financial Durations. *Journal of Econometrics* 119, 381–412.

Bollerslev, T. (1987). A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *Review of Economics and Statistics* 69, 542–547.

Bollerslev, T., R. F. Engle & D. B. Nelson (1994). ARCH Models. In R. F. Engle & D. McFadden, eds., Handbook of Econometrics, Elsevier, vol. IV.

Breidt, F. J. & A. L. Carriquiry (1996). Improved Quasi-Maximum Likelihood Estimation for Stochastic Volatility Models. In J. C. Lee, W. O. Johnson & A. Zellner, eds., Modelling and Prediction: Honoring Seymour Geisser. Springer, 228–247.

Breidt, F. J., N. Crato & P. de Lima (1998). The Detection and Estimation of Long Memory in Stochastic Volatility. *Journal of Econometrics* 83, 325–348.

Brockwell, A. E. (2007). Likelihood-based Analysis of a Class of Generalized Long-Memory Time Series Models. *Journal of Time Series Analysis* 28, 386–407.

Broto, C. & E. Ruiz (2004). Estimation Methods for Stochastic Volatility Models: A Survey. *Journal of Economic Surveys* 18, 613–649.

Calzolari, G., G. Fiorentini & E. Sentana (2004). Constrained Indirect Estimation. *Review of Economic Studies* 71, 945–973.

Carnero, M. A., D. Peña & E. Ruiz (2004). Persistence and Kurtosis in GARCH and Stochastic Volatility Models. *Journal of Financial Econometrics* 2, 319–342.

Chernov, M., A. Gallant, E. Ghysels & G. Tauchen (1999). A New Class of Stochastic Volatility Models with Jumps: Theory and Estimation. Working Paper 99s-48, CIRANO.

Chib, S. & E. Greenberg (1996). Markov Chain Monte Carlo Simulation Methods in Econometrics. *Econometric Theory* 12, 409–431.

Chib, S., F. Nardari & N. Shephard (2002). Markov Chain Monte Carlo Methods for Stochastic Volatility Models. *Journal of Econometrics* 108, 281–316.

Chib, S., Y. Omori & M. Asai (2009). Multivariate Stochastic Volatility. In T. G. Andersen, R. A. Davis, J. Kreiß& T. Mikosch, eds., Handbook of Financial Time Series, Springer, 365–400.

Commandeur, J. J. F. & S. J. Koopman (2007). An Introduction to State Space Time Series Analysis. Oxford University Press.

Creal, D. D. (2009). A Survey of Sequential Monte Carlo Methods for Economics and Finance. *Econometric Reviews*, forthcoming.

Danielsson, J. (1994). Stochastic Volatility in Asset Prices: Estimation with Simulated Maximum Likelihood. *Journal of Econometrics* 64, 375–400.

Danielsson, J. (1998). Multivariate Stochastic Volatility Models: Estimation and a Comparison with VGARCH Models. *Journal of Empirical Finance* 5, 155–173.

Danielsson, J. & J. F. Richard (1993). Accelerated Gaussian Importance Sampler with Application to Dynamic Latent Variable Models. *Journal of Applied Econometrics* 8, S153–S173. (Supplement: Special Issue on Econometric Inference Using Simulation Techniques).

Davidson, J. (2004). Moment and Memory Properties of Linear Conditional Heteroscedasticity Models, and a New Model. *Journal of Business & Economic Statistics* 22, 16–29.

Davis, R. A. & G. Rodriguez-Yam (2005). Estimation for State-Space Models Based on a Likelihood Approximation. *Statistica Sinica* 15, 381–406.
URL http://www3.stat.sinica.edu.tw/statistica/J15N2/J15N25/J15N25.html

Dhaene, G. & O. Vergote (2003). Asymptotic Results for GMM Estimators of Stochastic Volatility Models. Center for Economic Studies Discussions Paper Series 03.06, K. U. Leuven.
URL http://www.econ.kuleuven.ac.be/ces/discussionpapers/Dps03/Dps0306.pdf

Diebold, F. X., T. A. Gunther & A. S. Tay (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review* 39, 863–883.

Doornik, J. A. (2009). Ox 6 — An Object-Oriented Matrix Programming Language. Timberlake Consultants Ltd.

Duffie, J. & K. Singleton (1993). Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica* 61, 929–952.

Dufour, J.-M. & P. Valéry (2006). On a Simple Two-Stage Closed-Form Estimator for a Stochastic Volatility in a General Linear Regression. In D. Terrell & T. B. Fomby, eds., Econometric Analysis of Financial and Economic Time Series, Elsevier JAI, vol. 20, Part A of *Advances in Econometrics*, 259–288.

Durbin, J. & S. J. Koopman (1997). Monte Carlo Maximum Likelihood Estimation of Non-Gaussian State Space Models. *Biometrika* 84, 669–684.

Durbin, J. & S. J. Koopman (2000). Time Series Analysis of Non-Gaussian Observations Based on State Space Models from Both Classical and Bayesian Perspectives. *Journal Of The Royal Statistical Society Series B* 62, 3–56.

Durbin, J. & S. J. Koopman (2001). Time Series Analysis by State Space Methods. Oxford University Press.

Durham, G. B. (2006). Monte Carlo Methods for Estimating, Smoothing, and Filtering One- and Two-Factor Stochastic Volatility Models. *Journal of Econometrics* 133, 273–305.

Engle, R. F., D. M. Lilien & R. P. Robins (1987). Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model. *Econometrica* 55, 391–407.

Eraker, B., M. Johannes & N. Polson (2003). The Impact of Jumps in Volatility and Returns. *The Journal of Finance* 58, 1269–1300.

Evans, M. & T. Swartz (1995). Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems. *Statistical Science* 10, 254–272.

Gallant, A. & G. Tauchen (1996). Which Moments to Match? *Econometric Theory* 12, 657–681.

Gallant, A. R., D. Hsieh & G. Tauchen (1997). Estimation of Stochastic Volatility Models with Diagnostics. *Journal of Econometrics* 81, 159–192.

Gentle, J. H. (2003). Random Number Generation and Monte Carlo Methods. Springer, 2nd edn..

Ghysels, E., A. Harvey & E. Renault (1996). Stochastic Volatility. In G. Maddala & C. Rao, eds., Handbook of Statistics: Statistical Methods in Finance, Vol. 14, North-Holland.

Gneiting, T., F. Balabdaoui & A. E. Raftery (2007). Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society: Series B* 69, 243–268.

Gourieroux, C. & A. Monfort (1997). Simulation-Based Econometric Methods. Oxford University Press.

Gourieroux, C., A. Monfort & E. Renault (1993). Indirect Inference. *Journal of Applied Econometrics* 8, S85–S118. Supplement: Special Issue on Econometric Inference Using Simulation Techniques.

Hall, A. R. (2005). Generalized method of moments. Oxford University Press.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* 50, 1029–1054.

Harvey, A., E. Ruiz & N. Shephard (1994). Multivariate Stochastic Variance Models. *Review of Economic Studies* 61, 247–264.

Harvey, A. C. (2007). Long Memory in Stochastic Volatility. In J. Knight & S.Satchell, eds., Forecasting Volatility in the Financial Markets, Oxford: Butterworth-Heinemann, 351–363. 3rd edn..

Harvey, A. C. & T. Proietti, eds. (2005). Readings in Unobserved Components Models. Advanced Texts in Econometrics. Oxford University Press.

Harvey, A. C. & N. Shephard (1993). Estimation and Testing of Stochastic Variance Models. STICERD Econometrics Discussion Paper 93-268, London School of Economics.

Harvey, A. C. & N. Shephard (1996). Estimation of an Asymmetric Stochastic Volatility Model for Asset Returns. *Journal of Business & Economic Statistics* 14, 429–434.

Hautsch, N. & Y. Ou (2008). Stochastic Volatility Estimation Using Markov Chain Simulation. In W. K. Härdle, N. Hautsch & L. Overbeck, eds., Applied Quantitative Finance, Berlin, Heidelberg: Springer, chap. 12, 249–274. 2nd edn..

Hull, J. & A. White (1987). The Pricing of Options on Assets with Stochastic Volatilities. *Journal of Finance* 42, 281–300.

Hurvich, C. M. & P. Soulier (2009). Handbook of Financial Time Series. In T. G. Andersen, R. A. Davis, J. Kreiß& T. Mikosch, eds., Stochastic Volatility Models with Long Memory, Springer, 345–354.

Jacquier, E., N. G. Polson & P. E. Rossi (1994). Bayesian Analysis of Stochastic Volatility Models. *Journal of Business & Economic Statistics* 12, 69–87.

Jacquier, E., N. G. Polson & P. E. Rossi (1999). Stochastic Volatility: Univariate and Multivariate Extensions. Working Paper 99s-26, CIRANO.

Jacquier, E., N. G. Polson & P. E. Rossi (2004). Bayesian Analysis of Stochastic Volatility Models with Fat-Tails and Correlated Errors. *Journal of Econometrics* 122, 185–212.

Jung, R. C., M. Kukuk & R. Liesenfeld (2006). Time Series of Count Data: Modeling, Estimation and Diagnostics. *Computational Statistics & Data Analysis* 51, 2350–2364.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0167947306002581`

Jungbacker, B. & S. J. Koopman (2009). Parameter Estimation and Practical Aspects of Modeling Stochastic Volatility. In T. G. Andersen, R. A. Davis, J. Kreiß& T. Mikosch, eds., Handbook of Financial Time Series, Springer.

Kim, S., N. Shephard & S. Chib (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies* 65, 361–393.

Kirby, C. (2006). Linear Filtering for Asymmetric Stochastic Volatility Models. *Economics Letters* 92, 284–292.

Koopman, S. J. & E. H. Uspensky (2002). The Stochastic Volatility in Mean Model: Empirical Evidence from International Stock Markets. *Journal of Applied Econometrics* 17, 667–689.

Lee, K. M. & S. J. Koopman (2004). Estimating Stochastic Volatility Models: A Comparison of Two Importance Samplers. *Studies in Nonlinear Dynamics & Econometrics* 8, Article 5, 1–15.

Liesenfeld, R. & R. C. Jung (2000). Stochastic Volatility Models: Conditional Normality versus Heavy-Tailed Distributions. *Journal of Applied Econometrics* 15, 137–160.

Liesenfeld, R. & J.-F. Richard (2003). Univariate and Multivariate Stochastic Volatility Models: Estimation and Diagnostics. *Journal of Empirical Finance* 10, 505–531.

Liesenfeld, R. & J.-F. Richard (2006). Classical and Bayesian Analysis of Univariate and Multivariate Stochastic Volatility Models. *Econometric Reviews* 25, 335–360.

Melino, A. & S. M. Turnbull (1990). Pricing Foreign Currency Options with Stochastic Volatility. *Journal of Econometrics* 45, 239–265.

Meyer, R., D. A. Fournier & A. Berg (2003). Stochastic Volatility: Bayesian Computation Using Automatic Differentiation and the Extended Kalman filter. *Econometrics Journal* 6, 408–420.

Meyer, R. & J. Yu (2000). BUGS for a Bayesian Analysis of Stochastic Volatility Models. *Econometrics Journal* 3, 198–215.

Monfardini, C. (1998). Estimating Stochastic Volatility Models Through Indirect Inference. *Econometrics Journal* 1, 113–128.

Nelson, D. B. (1988). The Time Series Behavior of Stock Market Volatility and Returns. Ph. D. dissertation, Massachusetts Institute of Technology, Dept. of Economics, Cambridge, MA. URL `http://hdl.handle.net/1721.1/14363`

Nocedal, J. & S. J. Wright (2006). Numerical Optimization. Springer-Verlag, 2nd ed. edn..

Omori, Y., S. Chib, N. Shephard & J. Nakajima (2007). Stochastic Volatility with Leverage: Fast and Efficient Likelihood Inference. *Journal of Econometrics* 140, 425–449.

Renault, E. (2009). Moment-Based Estimation of Stochastic Volatility Models. In T. G. Andersen, R. A. Davis, J. Kreiß& T. Mikosch, eds., Handbook of Financial Time Series, Springer, 269–311.

Richard, J.-F. & W. Zhang (2007). Efficient High-Dimensional Importance Sampling. *Journal of Econometrics* 141, 1385–1411.

Rubinstein, R. Y. & D. P. Kroese (2008). Simulation and the Monte Carlo Method. Wiley-Interscience, 2nd ed. edn..

Ruiz, E. (1994). Quasi-Maximum Likelihood Estimation of Stochastic Volatility Models. *Journal of Econometrics* 63, 289–306.

Sandmann, G. & S. J. Koopman (1998). Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood. *Journal of Econometrics* 87, 271–301.

Scott, L. O. (1987). Option Pricing when the Variance Changes Randomly: Theory, Estimation, and an Application. *Journal of Financial and Quantitative Analysis* 22, 419–438.

Shephard, N. (1993). Fitting Non-Linear Time Series Models, with Applications to Stochastic Variance Models. *Journal of Applied Econometrics* 8, 135–152.

Shephard, N. (1994). Local Scale Models: State Space Alternative to Integrated GARCH Processes. *Journal of Econometrics* 60, 181–202.

Shephard, N., ed. (2005). Stochastic Volatility: Selected Readings. Advanced Texts in Econometrics. Oxford University Press.

Shephard, N. & T. G. Andersen (2009). Stochastic Volatility: Origins and Overview. In T. G. Andersen, R. A. Davis, J.-P. Kreiß& T. Mikosch, eds., Handbook of Financial Time Series, Springer, 233–254.

Shephard, N. & M. K. Pitt (1997). Likelihood Analysis of Non-Gaussian Measurement Time Series. *Biometrika* 84, 653–668.

Shimada, J. & Y. Tsukuda (2005). Estimation of Stochastic Volatility Models: An Approximation to the Nonlinear State Space Representation. *Communications in Statistics — Simulation and Computation* 34, 429–450.

Skaug, H. & J. Yu (2007). Automated Likelihood Based Inference for Stochastic Volatility Models. Working paper, Sim Kee Boon Institute for Financial Economics, Singapore Management University.

Tanizaki, H. (2003). Nonlinear and Non-Gaussian State-Space Modeling with Monte Carlo Techniques: A Survey and Comparative Study. In D. N. Shanbhag & C. R. Rao, eds., Handbook of Statistics, Elsevier, vol. 21, chap. 22, 871–929.

Taylor, S. J. (1982). Financial Returns Modelled by the Product of Two Stochastic Processes: A Study of Daily Sugar Prices, 1961–79. In O. D. Anderson, ed., Time Series Analysis: Theory and Practice 1, North-Holland.

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

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Annals of Statistics* 22, 1701–1728.

White, H. L. (1984). Maximum Likelihood Estimation of Misspecified Dynamic Models. In T. K. Dijkstra, ed., Misspecification Analysis, New York: Springer, 1–19.

Yu, J. (2005). On Leverage in a Stochastic Volatility Model. *Journal of Econometrics* 127, 165–178.

Yu, J. & R. Meyer (2006). Multivariate Stochastic Volatility Models: Bayesian Estimation and Model Comparison. *Econometric Reviews* 25, 361–384.

Zeger, S. L. (1988). A Regression Model for Time Series of Counts. *Biometrika* 75, 621–629.

A Some formulas for Gaussian approximation

By replacing $\ln f(y_t|h_t,\boldsymbol{\theta})$ in (5) with

$$\ln f_a(y_t|h_t,\boldsymbol{\theta}) = A_t + A_t^0 h_t + A_t^{00} h_t^2.$$

(see (7)) we get a quadratic approximation for $\ln f(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})$:

$$\begin{aligned}\ln f_a(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}) = &\sum_{t=1}^{T}\left(A_t + A_t^0 h_t + A_t^{00} h_t^2\right)\\ &-\frac{T}{2}\ln(2\pi\sigma_\eta^2) + \frac{1}{2}\ln(1-\delta^2) - \frac{1}{2\sigma_\eta^2}\left[(1-\delta^2)h_1^2 + \sum_{t=2}^{T}(h_t-\delta h_{t-1})^2\right].\end{aligned}$$

This can be rearranged as

$$\ln f_a(\mathbf{y},\mathbf{h}) = B + \sum_{t=1}^{T} B_t^0 h_t + \sum_{t=1}^{T} B_t^{00} h_t^2 + \sum_{t=2}^{T} B_t^{01} h_t h_{t-1}.$$

where

$$\begin{aligned}B &= \sum_{t=1}^{T} A_t - \frac{T}{2}\ln(2\pi\sigma_\eta^2) + \frac{1}{2}\ln(1-\delta^2),\\ B_t^0 &= A_t^0, \quad t=1,\ldots,T,\\ B_t^{00} &= A_t^{00} - \frac{1+\delta^2}{2\sigma_\eta^2}, \quad t=2,\ldots,T-1,\\ B_1^{00} &= A_1^{00} - \frac{1}{2\sigma_\eta^2}, \quad B_T^{00} = A_T^{00} - \frac{1}{2\sigma_\eta^2},\\ B_t^{01} &= \frac{\delta}{\sigma_\eta^2}, \quad t=2,\ldots,T.\end{aligned}$$

Now, according to (10) approximate Gaussian log-density is (ignoring the terms which do not depend on $h_1,\ldots,h_T$)

$$\begin{aligned}\ln g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) &= -\frac{1}{2}\sum_{t=2}^{T}\frac{1}{M_t}(h_t^2 + L_t^2 h_{t-1}^2 - 2L_t h_t h_{t-1} - 2K_t h_t + 2K_t L_t h_{t-1}) + const\\ &= \sum_{t=1}^{T-1}\left(\frac{K_t}{M_t} - \frac{K_{t+1}L_{t+1}}{M_{t+1}}\right)h_t + \frac{K_T}{M_T}h_T\\ &\quad - \sum_{t=1}^{T-1}\left(\frac{1}{2M_t} + \frac{L_{t+1}^2}{2M_{t+1}}\right)h_t^2 - \frac{1}{2M_T}h_T^2 + \sum_{t=2}^{T}\frac{L_t}{M_t}h_t h_{t-1} + const.\end{aligned}$$

This has the form

$$\ln g(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}) = \sum_{t=1}^{T}(B_t^0 h_t + B_t^{00} h_t^2 + B_t^{01} h_t h_{t-1}) + const.$$

By equating the coefficients we write a system of equations for parameters K_t, L_t, M_t:

$$\frac{K_t}{M_t} - \frac{K_{t+1}L_{t+1}}{M_{t+1}} = B_t^0, \quad t=1,\ldots,T-1, \qquad \frac{K_T}{M_T} = B_T^0,$$

$$-\frac{1}{2M_t} - \frac{L_{t+1}^2}{2M_{t+1}} = B_t^{00}, \quad t=1,\ldots,T-1, \qquad -\frac{1}{2M_T} = B_T^{00},$$

$$\frac{L_t}{M_t} = B_t^{01}, \quad t = 2, \dots, T.$$

This system can be readily solved for K_t, L_t, M_t by means of backward recursion:

$$M_t = -\frac{1}{2B_t^{00} + B_{t+1}^{01} L_{t+1}}, \qquad K_t = (B_t^0 + B_{t+1}^{01} K_{t+1}) M_t, \qquad L_t = B_t^{01} M_t, \qquad t = T, \dots, 1$$

assuming that $B_{T+1}^{01} = 0$ (and $B_1^{01} = 0$ to get $L_1 = 0$).

Now we have both $\ln f_a(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})$ and $\ln g(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})$. Approximate log-likelihood is the difference between them. The difference does not depend on $\mathbf{h}$, because all terms with $\mathbf{h}$ must cancel out by construction. So we simply use $\mathbf{h} = \mathbf{0}$ to get $\ln f_a(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})|_{\mathbf{h}=\mathbf{0}} = B$ and

$$\ln g(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})|_{\mathbf{h}=\mathbf{0}} = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T} \ln(M_t) - \frac{1}{2}\sum_{t=1}^{T} \frac{K_t^2}{M_t}$$

Finally, the approximate log-likelihood is

$$\begin{aligned}
\ell_a(\boldsymbol{\theta}; \mathbf{y}) &= \ln f_a(\mathbf{y}|\boldsymbol{\theta}) = \ln f_a(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})|_{\mathbf{h}=\mathbf{0}} - \ln g(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})|_{\mathbf{h}=\mathbf{0}} \\
&= B + \frac{T}{2}\ln(2\pi) + \frac{1}{2}\sum_{t=1}^{T} \ln(M_t) + \frac{1}{2}\sum_{t=1}^{T} \frac{K_t^2}{M_t} \\
&= \sum_{t=1}^{T} A_t - T\ln\sigma_\eta + \frac{1}{2}\ln(1-\delta^2) + \frac{1}{2}\sum_{t=1}^{T} \ln(M_t) + \frac{1}{2}\sum_{t=1}^{T} \frac{K_t^2}{M_t}.
\end{aligned}$$

B "Spectral" approximation for covariance matrix of QML estimates

The covariance matrix of QML estimates $\hat{\boldsymbol{\theta}}_Q$ is estimated as $\tilde{\mathscr{H}}^{-1}(\hat{\boldsymbol{\theta}}_Q)\tilde{\mathscr{I}}(\hat{\boldsymbol{\theta}}_Q)\tilde{\mathscr{H}}^{-1}(\hat{\boldsymbol{\theta}}_Q)$. We assume that the first element of $\boldsymbol{\theta}$ is σ_ξ.

Denote

$$\mu_i = -\ln\left(\sigma_\omega^2 + \sigma_\eta^2\left(1 + \delta^2 - 2\delta\cos\left(\frac{\pi i}{T+1}\right)\right)^{-1}\right), \qquad m_i^r = e^{\mu_i}\frac{\partial \mu_i}{\partial \theta_r},$$

$$\varphi = \frac{2}{T+1}\sum_{i=1}^{\lfloor (T+1)/2 \rfloor} \frac{e^{\mu_{2i-1}}}{\tan^2\left(\frac{\pi(2i-1)}{2(T+1)}\right)}, \qquad \gamma_3 = \mathsf{E}\,\omega_t^3/\sigma_\omega^3, \qquad \gamma_4 = \mathsf{E}\,\omega_t^4/\sigma_\omega^4.$$

Then (for $r \neq 1$, $s \neq 1$)

$$\tilde{\mathscr{I}}_{11} = -\tilde{\mathscr{H}}_{11} = \frac{4\varphi}{\sigma_\xi^2}, \qquad \tilde{\mathscr{H}}_{1r} = 0, \qquad \tilde{\mathscr{I}}_{1r} = \frac{\gamma_3\sigma_\omega^3\varphi}{\sigma_\xi} \cdot \sum_{i=1}^{T} m_i^r,$$

$$\tilde{\mathscr{H}}_{rs} = -\frac{1}{2}\sum_{i=1}^{T} \frac{\partial \mu_i}{\partial \theta_r}\frac{\partial \mu_i}{\partial \theta_s},$$

$$\tilde{\mathscr{I}}_{rs} = -\tilde{\mathscr{H}}_{rs} + \frac{\sigma_\omega^4(\gamma_4 - 3)}{4(T+1)}\left(\sum_{i=1}^{T} m_i^r \sum_{i=1}^{T} m_i^s + \frac{1}{2}\sum_{i=1}^{T} m_i^r m_i^s + \frac{1}{2}\sum_{i=1}^{T} m_i^r m_{T+1-i}^s\right).$$

Derivatives $\frac{\partial \mu_i}{\partial \theta_r}$ can be evaluated numerically.

For the basic SV model $\mathsf{E}\ln(\xi_t^2) = \mathscr{C} = \psi(1/2) - \ln(1/2)$, $\sigma_\omega^2 = \mathsf{E}\,\omega_t^2 = \pi^2/2$, $\mathsf{E}\,\omega_t^3 = -14\zeta(3) \approx -16.829$, where $\zeta(z)$ is the Riemann zeta function, $\mathsf{E}\,\omega_t^4 = \frac{7}{4}\pi^4$ (see Dhaene & Vergote (2003)). Thus, $\gamma_3 = -28\sqrt{2}\zeta(3)/\pi^3 \approx -1.5351$, $\gamma_4 = 7$.

The sums can be further approximated by integrals to obtain analytical expressions for the asymptotic matrices $\mathscr{I}_Q^\infty(\boldsymbol{\theta})$ and $\mathscr{H}_Q^\infty(\boldsymbol{\theta})$.

C Moments of the basic SV model

Assuming stationarity of the log-volatility process $h_t = \delta h_{t-1} + \sigma_\eta \eta_t$ we can write

$$h_t \sim \mathcal{N}\left(0, \frac{\sigma_\eta^2}{1-\delta^2}\right).$$

From $y_t = \sigma_\xi \xi_t \exp(h_t/2)$ and the assumption that ξ_t and h_t are independent it follows that

$$\mathsf{E}\left|y_t\right|^n = \sigma_\xi^n \mathsf{E}|\xi_t|^n \mathsf{E}\exp(nh_t/2).$$

Here $\exp(nh_t/2)$ is log-normal:

$$\exp(nh_t/2) \sim \mathcal{LN}\left(0, \frac{n^2\sigma_\eta^2}{4(1-\delta^2)}\right)$$

and thus

$$\mathsf{E}\exp(nh_t/2) = \exp\left(\frac{n^2\sigma_\eta^2}{8(1-\delta^2)}\right).$$

As mentioned in Harvey (2007) if $x \sim \chi_\nu^2$ then (for $\alpha > -\nu/2$)

$$\mathsf{E}\, x^\alpha = \frac{2^\alpha \Gamma(\nu/2+\alpha)}{\Gamma(\nu/2)}.$$

For the basic SV model $\xi_t^2 \sim \chi_1^2$. It follows that

$$\mathsf{E}|\xi_t|^n = \mathsf{E}[(\xi_t^2)^{n/2}] = \frac{2^{n/2}\Gamma((n+1)/2)}{\Gamma(1/2)} = \frac{2^{n/2}\Gamma((n+1)/2)}{\sqrt{\pi}}.$$

Combining these results we have (for $n > -1$)

$$\mathsf{E}\left|y_t\right|^n = \frac{\sigma_\xi^n 2^{n/2}\Gamma((n+1)/2)}{\sqrt{\pi}} \exp\left(\frac{n^2\sigma_\eta^2}{8(1-\delta^2)}\right).$$

Specifically, for $n = 1$ and $n = 2$ (using $\Gamma(1) = 1$ and $\Gamma(3/2) = \sqrt{\pi}/2$)

$$\mathsf{E}\left|y_t\right| = \sigma_\xi \sqrt{\frac{2}{\pi}} \exp\left(\frac{\sigma_\eta^2}{8(1-\delta^2)}\right), \qquad \mathsf{E}\, y_t^2 = \sigma_\xi^2 \exp\left(\frac{\sigma_\eta^2}{2(1-\delta^2)}\right).$$

It is also possible to derive autocovariances of $\left|y_t\right|$ and y_t^2. In general

$$\mathsf{E}\left[\left|y_t\right|^m \left|y_{t-k}\right|^n\right] = \sigma_\xi^{m+n} \mathsf{E}|\xi_t|^m \mathsf{E}|\xi_{t-k}|^n \mathsf{E}\exp((mh_t + nh_{t-k})/2) \qquad (k > 0).$$

Here $(mh_t + nh_{t-k})/2$ is normally distributed with zero mean and variance $\frac{(m^2+n^2+2mn\delta^k)\sigma_\eta^2}{4(1-\delta^2)}$. Its exponent is log-normal:

$$\exp((mh_t + nh_{t-k})/2) \sim \mathcal{LN}\left(0, \frac{(m^2+n^2+2mn\delta^k)\sigma_\eta^2}{4(1-\delta^2)}\right).$$

Hence

$$\mathsf{E}\exp((mh_t + nh_{t-k})/2) = \exp\left(\frac{(m^2+n^2+2mn\delta^k)\sigma_\eta^2}{8(1-\delta^2)}\right)$$

and

$$\mathsf{E}\left[\left|y_t\right|^m\left|y_{t-k}\right|^n\right]=\frac{1}{\pi}\sigma_\xi^{m+n}2^{(m+n)/2}\Gamma((m+1)/2)\Gamma((n+1)/2)\exp\left(\frac{(m^2+n^2+2mn\delta^k)\sigma_\eta^2}{8(1-\delta^2)}\right)\qquad(k>0).$$

In particular, for $m=1$ and $n=1$

$$\mathsf{E}\left[\left|y_t\right|\left|y_{t-k}\right|\right]=\frac{2\sigma_\xi^2}{\pi}\exp\left(\frac{(1+\delta^k)\sigma_\eta^2}{4(1-\delta^2)}\right)\qquad(k>0).$$

These are non-central autocovariances. Similar expressions can be derived for y_t^2.

We write $\ln(y_t^2)$ as

$$\ln(y_t^2)=\ln\sigma_\xi^2+\mathscr{C}+h_t+\omega_t,$$

where $\omega_t=\ln(\xi_t^2)-\mathscr{C}$, $\mathsf{E}\,\omega_t=0$, $\mathsf{Var}\,\omega_t=\pi^2/2$. From $\mathsf{E}\,h_t=0$ it follows that

$$\mathsf{E}\ln(y_t^2)=\ln\sigma_\xi^2+\mathscr{C}.$$

Next, h_t and ω_t are two independent stationary processes. The process h_t is AR(1) with the autoregression coefficient δ and the innovations variance σ_η^2, while ω_t is white noise. Consequently, second moments of $\ln(y_t^2)$ can be easily obtained:

$$\mathsf{Var}[\ln(y_t^2)]=\mathsf{Var}\,h_t+\mathsf{Var}\,\omega_t=\sigma_\eta^2/(1-\delta^2)+\pi^2/2,$$

$$\mathsf{Cov}(\ln(y_t^2),\ln(y_{t-k}^2))=\mathsf{Cov}(h_t,h_{t-k})=\sigma_\eta^2\delta^k/(1-\delta^2),\qquad k>0.$$

D Some formulas for extended SV model

For $t=2,\dots,T$

$$\ln\phi_t=\ln\rho(\xi_t)-\ln\sigma_\xi-h_t/2-\frac{1}{2}\ln(2\pi\sigma_\eta^2)-\frac{1}{2}\eta_t^2.$$

where $\xi_t=\xi_t(y_t,h_t)$ defined by (27) and

$$\eta_t=\eta_t(h_t,y_{t-1},h_{t-1})=\frac{h_t-\delta h_{t-1}-\alpha\xi_{t-1}(y_{t-1},h_{t-1})}{\sigma_\eta}.$$

The derivatives are given by

$$\frac{d\ln\phi_t}{dh_t}=(\ln\rho(\xi_t))'\xi_t'-\frac{1}{2}-\frac{1}{\sigma_\eta}\eta_t,\qquad\frac{d\ln\phi_t}{dh_{t-1}}=\frac{1}{\sigma_\eta}\eta_t(\delta+\alpha\xi_{t-1}'),$$

$$\frac{d^2\ln\phi_t}{dh_t^2}=(\ln\rho(\xi_t))''(\xi_t')^2+(\ln\rho(\xi_t))'\xi_t''-\frac{1}{\sigma_\eta^2},\qquad\frac{d^2\ln\phi_t}{dh_t\,dh_{t-1}}=\frac{1}{\sigma_\eta^2}(\delta+\alpha\xi_{t-1}'),$$

$$\frac{d^2\ln\phi_t}{dh_{t-1}^2}=\frac{1}{\sigma_\eta}\eta_t\alpha\xi_{t-1}''-\frac{1}{\sigma_\eta^2}(\delta+\alpha\xi_{t-1}')^2.$$

The derivatives of

$$\xi_t(y_t,h_t)=\frac{y_t-\mu-\kappa r(h_t)}{\sigma_\xi\exp(h_t/2)}.$$

with respect to h_t are given by

$$\xi_t'=-\frac{1}{2}\xi_t-\frac{\kappa}{\sigma_\xi\exp(h_t/2)}r'(h_t),$$

$$\xi_t''=\frac{1}{4}\xi_t+\frac{\kappa}{\sigma_\xi\exp(h_t/2)}\left(r'(h_t)-r''(h_t)\right).$$

The derivatives of in-mean function are

- for $r(h_t) = \exp(h_t/2)$: $r'(h_t) = r(h_t)/2$, $r''(h_t) = r(h_t)/4$,
- for $r(h_t) = \exp(h_t)$: $r'(h_t) = r''(h_t) = r(h_t)$,
- for $r(h_t) = h_t$: $r'(h_t) = 1$, $r''(h_t) = 0$.

For the standard normal distribution with the density function $\ln\rho(\xi) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\xi^2$ the derivatives obviously are

$$(\ln\rho)' = -\xi, \qquad (\ln\rho)'' = -1.$$

For the Student's t distribution with the density function

$$\ln\rho(\xi) = -\ln B(\nu/2, 1/2) - \frac{1}{2}\ln(\nu) - \frac{\nu+1}{2}\ln\left(1 + \frac{\xi^2}{\nu}\right).$$

where

$$B(\nu/2, 1/2) = \frac{\Gamma(\nu/2)\Gamma(1/2)}{\Gamma((\nu+1)/2)} = \frac{\Gamma(\nu/2)\sqrt{\pi}}{\Gamma((\nu+1)/2)},$$

the derivatives are

$$(\ln\rho)' = -\frac{\xi(\nu+1)}{\nu+\xi^2}, \qquad (\ln\rho)'' = -\frac{(\nu-\xi^2)(\nu+1)}{(\nu+\xi^2)^2}.$$

The elementary quadratic approximation for the complete data log-density is

$$\begin{aligned}\ln\phi_{at} = {} & F_t + F_t^0(h_t - h_t^*) + F_t^1(h_{t-1} - h_{t-1}^*) \\ & + \frac{1}{2}F_t^{00}(h_t - h_t^*)^2 + F_t^{01}(h_t - h_t^*)(h_{t-1} - h_{t-1}^*) + \frac{1}{2}F_t^{11}(h_{t-1} - h_{t-1}^*)^2.\end{aligned}$$

It can be written as

$$\ln\phi_{at} = C_t + C_t^0 h_t + C_t^1 h_{t-1} + C_t^{00} h_t^2 + C_t^{01} h_t h_{t-1} + C_t^{11} h_{t-1}^2,$$

where

$$\begin{gathered}C_t = F_t - F_t^0 h_t^* - F_t^1 h_{t-1}^* + \frac{1}{2}F_t^{00} h_t^{*2} + F_t^{01} h_t^* h_{t-1}^* + \frac{1}{2}F_t^{11} h_{t-1}^{*2}, \\ C_t^0 = F_t^0 - F_t^{00} h_t^* - F_t^{01} h_{t-1}^*, \qquad C_t^1 = F_t^1 - F_t^{01} h_t^* - F_t^{11} h_{t-1}^*, \\ C_t^{00} = \frac{1}{2}F_t^{00}, \qquad C_t^{01} = F_t^{01}, \qquad C_t^{11} = \frac{1}{2}F_t^{11}\end{gathered}$$

with obvious modifications for $t = 1$. Summing $\ln\phi_{at}$ up, we obtain

$$\ln f_a(\mathbf{y}, \mathbf{h}) = \sum_{t=1}^{T} \ln\phi_{at} = \sum_{t=1}^{T} \left[C_t + C_t^0 h_t + C_t^1 h_{t-1} + C_t^{00} h_t^2 + C_t^{01} h_t h_{t-1} + C_t^{11} h_{t-1}^2\right].$$

This sum can be rearranged to obtain (8). The coefficients of this representation are

$$B_t^0 = C_t^0 + C_{t+1}^1, \quad t = 1, \ldots, T,$$

$$B_t^{00} = C_t^{00} + C_{t+1}^{11}, \quad t = 1, \ldots, T, \qquad B_t^{01} = C_t^{01}, \quad t = 2, \ldots, T$$

with $C_{T+1}^1 = 0$ and $C_{T+1}^{11} = 0$) and

$$B = \sum_{t=1}^{T} C_t.$$