A Flexible State-Space Model with Application to Stochastic Volatility

Christian Gouriéroux*, and Yang Lu†

November 18, 2016

Abstract

We introduce a general state-space (or latent factor) model for time series and panel data. The state process has a polynomial expansion based dynamics that can approximate any Markov dynamics arbitrarily well, and has a latent, endogenous switching regime interpretation. The resulting state-space model is associated with simulation-free, recursive formulas for prediction and filtering, as well as the maximum composite likelihood estimation method, with an extremely low computational cost. When applied to the stochastic volatility (SV) of asset returns, the model can capture, in a unified framework, stylized facts such as heavy tailed return, volatility feedback, as well as time irreversibility. The methodology is illustrated using Apple stock return data, which confirms the improvement of our model with respect to a benchmark SV model.

Key words: Endogenous regime switching, polynomial expansion, composite likelihood, time irreversibility, volatility feedback, copula.

JEL code: C32, C14

Acknowledgement: This research has benefited from the support of the Chair ACPR: Regulation and Systemic Risks.

*CREST and University of Toronto. christian.gourieroux@ensae.fr

†Aix-Marseille University. Corresponding author: luyang000278@gmail.com

1 Introduction

There is a growing concern of developing flexible state-space (or latent factor, dynamic random effect) models in Economics, Finance, and Insurance. Potential applications concern both time series and panel data, such as:

1. the returns of financial assets [see e.g. Ruiz (1994); Kim et al. (1998)], or the incomes of individual workers [see e.g. Jensen and Shore (2011)], with the stochastic volatility as the state variable.

2. panel binary event data, such as corporate defaults [see e.g. Duffie et al. (2009)], with the stochastic default probability as the state variable.

3. (panel or time series) count data, such as the numbers of (buy or sell) transactions of a specific market, the numbers of lapses (resp. redemptions) of a life insurer (resp. investment fund), the numbers of defaults in a specific sector [see e.g. Darolles et al. (2013)] during each time interval, or the annual numbers of accidents of car insurance policyholders. In these cases the state variable is the stochastic intensity.

4. (panel or time series) duration data with stochastic intensity [see e.g. Ghysels et al. (2004); Bauwens and Hautsch (2006)].

5. panel data on individuals' dynamic discrete choices [see e.g. Kasahara and Shimotsu (2009), Abbring (2010), Hu and Shum (2012), Norets and Tang (2013)], in which the state variable arises as the unobservable taste, or belief variable.[1]

Such models are called parameter-driven in the time series literature [see e.g. Cox (1981)], as opposed to observation-driven models such as ARMA and GARCH processes, in which the conditional forecasting density is a simple, deterministic function of past values. Parameter-driven models have the advantage of being more intuitive, more flexible [see Koopman et al. (2016) for a discussion], and easily applicable to a wide range of data, with potentially irregular features such as missing data. In Finance, many parameter-driven, stochastic volatility models have the further advantages (over GARCH models) of providing closed form formulas for derivatives prices [see e.g. Heston (1993), Gouriéroux and Monfort (2015)] and accounting for volatility risk.

Nevertheless, compared to observation-driven models, the estimation and forecasting of parameter-driven state-space models are often computationally intensive, and necessitates simulation based techniques such as particle filters or MCMC [see e.g. Chib and Winkelmann (2012)]. Our paper

[1] This literature has emphasized on flexible models with few distributional constraints. Most papers cited above show that such models are non-parametrically identified. However, once the identification proved, few tractable models have been proposed for the estimation purpose.

introduces a family of models that i) is sufficiently flexible to fit a wide range of data, and ii) leads to simple, simulation-free procedures for estimation and forecasting. We specify the dynamics of the Markov state process via the joint density function, which has a polynomial expansion form. We show that this model is flexible and capable of approximating any univariate Markov dynamics arbitrarily well. Moreover it is of finite dimensional dependence, and has a latent endogenous switching regime interpretation. As a consequence, the resulting state-space model allows for simple recursive formulas for prediction, filtering and smoothing, which is faster than simulation based methods such as the particle filter. Under some further constraints, the model can be estimated by maximum composite likelihood, whose computational cost is extremely low compared to simulation-based methods.

While the methodology of the paper can be applied to both panel and time series data, we will be focused on a time series application: the stochastic volatility of asset returns. In particular, we contribute to the stochastic volatility (SV) literature by proposing a unified framework for heavy-tailed return, volatility feedback and time irreversibility. While the first two properties have already been separately studied in the literature, most models consider only time reversible dynamics, and this assumption is often violated by financial and economic time series.

The paper is organized as follows. The Markov state process is introduced in Section 2. We discuss the stationarity and ergodicity of the state process, provide interpretation of the dynamics in terms of underlying switching regimes, and characterize the time (ir-)reversibility condition. Illustrative examples are provided in Section 3. Recursive forecasting, filtering and smoothing formulas are derived in Section 4. The non-parametric background of the state process dynamics, in particular its capability to approximate any Markov dynamics, is explored in Section 5. Section 6 discusses the maximum composite likelihood estimation, and proposes a stochastic volatility application on Apple stock return. Section 7 concludes. Proofs and technical details are gathered in Appendices.

2 The model

2.1 The state-space representation

We consider a state-space model in which X_t is a one-dimensional state variable with domain $\mathcal{X}$, whereas Y_t is the observable variable with domain $\mathcal{Y}$. Depending on the application, $\mathcal{X}$ can be the whole real line $\mathbb{R}$, the positive half-line $\mathbb{R}_{>0}$, or a bounded interval such as $]0,1[$, whereas $\mathcal{Y}$ can be an infinite set such as $\mathbb{R}$ (such as for return data), $\mathbb{R}_{>0}$, or a discrete set, such as $\mathbb{N}$ (count data), or $\{0,1,....,N\}$ (binomial or categorical data). This dynamic system has the following

representation:

$$Y_t|\underline{Y_{t-1}}, \underline{X_t} \sim l(y_t|x_t, \underline{y_{t-1}}), \tag{2.1}$$

$$X_t|\underline{Y_{t-1}}, \underline{X_{t-1}} \sim l(x_t|\underline{x_{t-1}}), \tag{2.2}$$

where $\underline{y_{t-1}}$ (resp. $\underline{x_{t-1}}$) is the past trajectory of process (Y_t) (resp. (X_t)) up to time $t-1$.In other words, process (X_t) is exogenous, and the conditional distribution of Y_t given its own past $\underline{Y_{t-1}}$ and the whole trajectory of (X_t) depends only on $\underline{Y_{t-1}}$ and the current value of the state variable X_t.

Such a model is usually difficult to estimate, except in some special cases: *i*) if process (X_t, Y_t) is jointly Gaussian, then the estimation is conducted via the Kalman filter; *ii*) if the domain $\mathcal{X}$ of the state variable X_t is finite, for instance if process (X_t) is a discrete Markov chain, then we can use the Kitagawa filter [see e.g. Kitagawa (1987)]. Besides these two cases, the estimation of the model involves simulation based techniques such as particle filter or MCMC.

Our paper introduces a new family of dynamic models, in which the dynamics of the state variable (X_t) is Markov with a flexible joint density f, approaching any unknown density function by a benchmark density times a positive squared polynomial. More precisely, we assume:

Assumption 1. Process (X_t) is Markov, stationary, with the joint distribution of (X_t, X_{t+1}) given by:

$$f(x_t, x_{t+1}) = \frac{1}{M}\phi(x_t)\phi(x_{t+1})\Big[\sum_{j=0}^{J}\sum_{k=0}^{J} b_{j,k}x_t^j x_{t+1}^k\Big]^2, \tag{2.3}$$

where $\phi(\cdot)$ is a positive benchmark density, whose integer moments are all finite, J is an integer, the matrix of coefficients $b_{j,k}$ is real and the normalization constant is equal to:

$$M = \sum_{j_1,k_1,j_2,k_2=0}^{J} b_{j_1,k_1}b_{j_2,k_2}\mu_{j_1+j_2}\mu_{k_1+k_2} = e'De > 0, \tag{2.4}$$

where $\mu_k := \int x^k\phi(x)\mathrm{d}x$, $k = 0, 1, ..., 2J$, are the power moments of distribution ϕ, matrix $D = (d_{j,k})_{0\leq j,k\leq 2J}$ is defined by:

$$d_{j,k} = \mu_j\mu_k \sum_{\substack{j_1+j_2=j \\ 0\leq j_1,j_2\leq J}} \sum_{\substack{k_1+k_2=k \\ 0\leq k_1,k_2\leq J}} b_{j_1,k_1}b_{j_2,k_2}, \tag{2.5}$$

and vector $e = (1, 1, ..., 1)' \in \mathbb{R}^{2J+1}$ is the vector with unitary components.

The density function $f(x_t, x_{t+1})$ in (2.3) is the product of two terms. The first term $\phi(x_t)\phi(x_{t+1})$ is a product density. The second term is a squared polynomial. We take the square of the polynomial in order to avoid negative values for the density function. Indeed, ensuring positivity

is especially important in financial applications since a density function with negative values spells arbitrage opportunities and could lead to dangerous financial strategies. The background of this flexible distribution is the approximation of a rooted density function $\sqrt{f(x_t, x_{t+1})}$ by the product between a rooted benchmark product density $\sqrt{\phi(x_t)\phi(x_{t+1})}$ and a polynomial expansion in the two arguments. The model is semi-parametric, since the proposed expansion-based density can approximate any bivariate density function arbitrarily well when J goes to infinity (see Section 5).

If the squared polynomial is equal to one, then we get $f(x_t, x_{t+1}) = \phi(x_t)\phi(x_{t+1})$, which corresponds to an i.i.d. sequence with marginal density ϕ. If the rank of B is 1, say $B = \beta\beta'$, where β is a $J+1$ dimensional vector, then we have $f(x_t, x_{t+1}) = \left[\phi(x_t)\sum_{i=0}^{J}\beta_i x_t^i\right]\left[\phi(x_{t+1})\sum_{i=0}^{J}\beta_i x_{t+1}^i\right]$. We still get the independence with a modified marginal distribution.

The choice of the benchmark density can be motivated by the prior knowledge about the process (X_t), and/or mathematical convenience. For instance, if (X_t) takes values between 0 and 1, then we can use the uniform density $\phi(x) = \mathbb{1}_{0<x<1}$, or a beta density $\frac{x^{\alpha-1}(1-x)^{\beta}}{B(\alpha,\beta)}$; if (X_t) takes values in the space of positive real numbers $\mathbb{R}_{>0}$, then we can use the Gamma, Weibull, or log-normal distribution; if (X_t) takes value in the real line, we can also use the Gaussian density. It will be shown, in Section 5, that the polynomial expansion can be conducted with respect to any benchmark density.

In a state-space model with unobservable state variable, the state variable is defined up to an invertible (nonlinear) transform. We assume that such a transformation has been first applied to ensure the existence of all power moments[2].

Thanks to the introduction of D, the joint density of (X_t, X_{t+1}) in equation (2.3) can be more conveniently rewritten in the matrix product form:

$$f(x_t, x_{t+1}) = \phi(x_t)\phi(x_{t+1})\frac{U'(x_t)DU(x_{t+1})}{e'De}, \tag{2.6}$$

where the symbol $'$ denotes the transpose of a matrix, and vector function U is defined by:

$$U_j(x) = \frac{x^j}{\mu_j}, \qquad \forall j = 0, 1, ..., 2J,$$

that satisfies: $\int \phi(x)U(x)\mathrm{d}x = e$. This definition assumes implicitly that all moments $\int \phi(s)s^j\mathrm{d}s$ are non zero. This can be achieved by an appropriate choice of the state variable[3], for instance with values between 0 and $+\infty$, or 0 and 1, and of the benchmark density ϕ. Thus, without loss

[2]Note that the special form (2.3) is not invariant by nonlinear transform of the state variable.

[3]Nevertheless, if, for instance, ϕ is a normal density so that all the odd moments are zero, we can still write the density in a similar way as:

$$f(x_t, x_{t+1}) = \phi(x_t)\phi(x_{t+1})\frac{V'(x_t)CV(x_{t+1})}{\omega'C\omega},$$

of generality, we focus on the case where the density can be written into (2.6).

In order for the density in equation (2.3) to be the joint distribution of the neighbouring terms of a stationary Markov process, it is necessary that it defines two identical margins.

Proposition 1. *The joint density function* (2.3) *defines identical margins if and only if matrix* D *satisfies:*

$$(D - D')e = 0. \tag{2.7}$$

Then the marginal distribution is:

$$f_0(x_t) = \phi(x_t)\frac{U'(x_t)De}{e'De} = \phi(x_t)\frac{e'DU(x_t)}{e'De}. \tag{2.8}$$

Proof. See Appendix 1.1. □

This condition usually depends on the choice of the benchmark density ϕ. However, if D is symmetric, then condition $(D-D')e = 0$ is automatically satisfied. It will be shown in Section 2.4 that a necessary and sufficient condition for D to be symmetric, is that matrix B is symmetric, $(b_{i,j} = b_{j,i},\ \forall i, j)$, or anti-symmetric, $(b_{i,j} = -b_{j,i},\ \forall i, j)$. However, the symmetry of D is not a necessary condition to ensure $(D - D')e = 0$. In particular, we derive in Section 2.4 a simple characterization on entries of matrix B to ensure the equal margin condition $(D - D')e = 0$.

2.2 Ergodicity of the state process

From the joint and marginal distributions, we get the conditional distribution of X_{t+1} given X_t:

$$f_1(x_{t+1}|x_t) = \frac{f(x_t, x_{t+1})}{f_0(x_t)} = \phi(x_{t+1})\frac{U'(x_t)DU(x_{t+1})}{U'(x_t)De}. \tag{2.9}$$

Since $e'De = M > 0$ from (2.4), we have $U'(x_t)De > 0$, and $U(x_t)'DU(x_{t+1}) \geq 0$ almost surely, due to the expressions of the marginal and conditional densities. This one-step-ahead conditional density has a simple matrix product form. This type of formula can be extended to longer horizons.

Proposition 2 (Conditional distribution at horizon h)**.** *The conditional distribution of* X_{t+h} *given* X_t *is:*

$$f_h(x_{t+h} \mid x_t) = \phi(x_{t+h})U(x_t)'\frac{D\Pi^{h-1}U(x_{t+h})}{U'(x_t)De}, \tag{2.10}$$

where $V_j = \phi(x)x^j$, $\omega_j = \int \phi(s)s^j \mathrm{d}s$, and matrix C is such that $c_{i,j} = \sum_{\substack{j_1+j_2=j \\ 0\leq j_1,j_2\leq J}} \sum_{\substack{k_1+k_2=k \\ 0\leq k_1,k_2\leq J}} b_{j_1,k_1}b_{j_2,k_2}$. Then the derivation of the main properties of the model are largely identical.

where the $(2J+1) \times (2J+1)$ *matrix* Π *is defined by:*

$$\Pi = \int \frac{U(x)U(x)'D}{U'(x)De}\phi(x)\mathrm{d}x. \tag{2.11}$$

Proof. See Appendix 1.2. □

Let us discuss the form of matrix Π. In order for this matrix to be well defined, let us first assume that the denominator $U'(x)De$ is bounded away from zero. This is satisfied under rather mild conditions, for instance:

Lemma 1. If the null space of the $(J+1) \times (J+1)$ matrix B does not contain any vector of the form $(1, x, ..., x^{J-1}, x^J)'$, then $U'(x)De$ is lower bounded by a positive constant.

Proof. See Appendix 1.3. □

From now on let us assume that the assumption in Lemma 1 holds. The entries of matrix Π usually do not allow for closed form expression, but they only involve univariate integrals and thus can be computed very efficiently[4]. The following corollary provides the left and right eigenvectors of matrix Π associated with the unitary eigenvalue.

Corollary 1. *1. The rows of* Π *sum up to one:* $\Pi e = e$*. Therefore* e *is a right eigenvector of* Π *with unitary eigenvalue.*

2. The vector De *is a left eigenvector of* Π *associated with the unitary eigenvalue:* $(De)'\Pi = (De)'$.

Proof. We have:

$$\Pi e = \int \frac{U(x)U'(x)De}{U'(x)De}\phi(x)\mathrm{d}x = \int U(x)\phi(x)\mathrm{d}x = e. \tag{2.12}$$

Similarly, we have:

$$(De)'\Pi = \int \frac{e'D'U(x)}{U'(x)De}U(x)'D\phi(x)\mathrm{d}x = e'D = e'D', \tag{2.13}$$

since $(D' - D)e = 0$. Thus e and De are respectively the left and right eigenvectors of Π with the unitary eigenvalue. □

Roughly speaking, the first property says that Π is row stochastic, if its entries are all non-negative[5]. In other words, in this case Π is associated with a Markov chain. This Markov chain

[4]For instance by using the command "integrate" in R. These quantities can also be computed by Monte-Carlo simulation, but the latter approach is much slower in order to get a similar degree of numerical accuracy. See Appendix 3 for a comparison.

[5]Nevertheless, in our model, Π is allowed to have negative entries.

will be formally introduced in the next Section. The second property provides the stationary distribution of the chain.

Let us now compare the conditional distribution and the marginal distribution. We have:

Corollary 2.

$$f_h(x_{t+h}|x_t) - f_0(x_{t+h}) = \frac{\phi(x_{t+h})U'(x_t)D}{U(x_t)'De}\Big(\Pi^{h-1} - \frac{ee'D}{e'De}\Big)U(x_{t+h}). \tag{2.14}$$

As a consequence, the conditional density of X_{t+h} given X_t [see equation (2.10)] is always a linear combination of densities which are the components of $\phi(x_{t+h})U(x_{t+h})$, with coefficients $\frac{U(x_t)'D\Pi^{h-1}}{U'(x_t)De}$ that sum up to unity, since $\Pi^{h-1}e = e$ by induction. However, these coefficients are not necessarily positive.

As the weak ergodicity of the process is equivalent to the convergence of $f_h(x_{t+h}|x_t) - f_0(x_{t+h})$ to zero, we get the following sufficient condition for ergodicity:

Proposition 3. *The state process (X_t) is weakly ergodic if 1 is a simple eigenvalue of matrix Π, and all other eigenvalues are strictly smaller than 1 in modulus.*

The ergodicity is necessary in order for likelihood type estimators to be consistent and asymptotically normal. The proof of this proposition is omitted, as it is based on the same arguments as for a finite state Markov chain [see e.g. Seneta (2006), Chapter 1], even if the entries of Π are not necessarily nonnegative.

Moreover, by equation (2.14), we see that the second largest (in modulus) eigenvalue of Π has a large impact on the serial correlation of the process (X_t). If this eigenvalue is large (resp. small), then the conditional distribution $f(x_{t+h}|x_t)$ converges slowly (resp. quickly) to the stationary distribution $f_0(x_{t+h})$.

2.3 Interpretation in terms of latent regimes

The dynamics of the process defined in the previous subsection is easily interpreted in terms of a process with switching regimes, under the following assumption:

Assumption 2. Process (X_t) is positive, and the components of $U'(x_t)D$ are nonnegative for all $x_t \in \mathcal{X}$.

A sufficient condition for $U'(x_t)D$ to be nonnegative is that entries of D are nonnegative. This assumption also ensures that the entries of Π are nonnegative.

Proposition 4. *Under Assumptions 1 and 2, the dynamics of process (X_t) admits a switching regime S_t, which takes values in $0, ..., 2J$. We can write the conditional density of X_{t+1} given $\underline{S_t}$*

and $\underline{X_t}$ *as:*

$$l(x_{t+1}|s_t) \sim \frac{\phi(x_{t+1})x_{t+1}^{s_t}}{\mu_{s_t}} = \phi(x_{t+1})U_{s_t}(x_{t+1}),$$

where the conditional probabilities of S_t *given* $\underline{S_{t-1}}$ *and* $\underline{X_t}$ *are:*

$$\Big(\mathbb{P}[S_t = 0|\underline{S_{t-1}}, \underline{X_t}], ..., \mathbb{P}[S_t = 2J|\underline{S_{t-1}}, \underline{X_t}]\Big)' = \frac{U'(X_t)D}{U'(X_t)De}.$$

Proof. The proof is immediate, since the components of the row vector $\frac{U'(X_t)D}{U'(X_t)De}$ are nonnegative and sum up to one, and the components of column vector $\phi(x)U(x)$ are density functions. □

To summarize, under Assumptions 1 and 2, we have the following causal scheme:

$$\dots S_{t-1} \to X_t \to S_t \to X_{t+1} \to S_{t+1} \dots . \tag{2.15}$$

This combined process is Markov, in the sense that each variable depends on all the variables on its LHS only via its nearest left neighbour. Moreover, the latent process (S_t) is also a Markov chain with respect to its own history. This Markov chain, called the embedded chain of (X_t), is characterized by its transition matrix, which is exactly matrix Π since:

$$\Pi = \int_0^\infty \Big[U(x)\phi(x)\Big]\Big[\frac{U'(x)D}{U'(x)De}\Big]\mathrm{d}x.$$

Within the integral, the first term $\phi(x)U(x)$ is the column vector of densities of X_{t+1} given $S_t = i$, whereas the second term $\frac{U'(x)D}{U'(x)De}$ is the row vector of conditional probabilities $\mathbb{P}[S_{t+1} = j|X_{t+1}]$, for $i, j = 0, 1, ..., 2J$, respectively. To summarize we have:

$$\Pi_{i,j} := \mathbb{P}[S_{t+1} = j \mid S_t = i] = \int \mathbb{P}[S_{t+1} = j|S_t = i, x_{t+1}]l(x_{t+1}|S_t = i)\mathrm{d}x_{t+1}. \tag{2.16}$$

By Corollary 1, when the embedded chain (S_t) exists, the stationary distribution of this Markov chain is $\frac{De}{e'De}$. Moreover, by using the joint density formula $f(x_t, x_{t+1}) = \phi(x_t)\phi(x_{t+1})U'(x_t)\frac{D}{e'De}U(x_{t+1})$, we know that the re-normalized matrix $\frac{D}{e'De}$ is such that: $\Big(\frac{D}{e'De}\Big)_{i,j} = \mathbb{P}[S_t = i, S_{t+1} = j]$. This matrix has been termed the "joint probability matrix" by McCausland (2007), who uses it to study the time-reversibility of a Markov chain (see also the next subsection, as well as the application section on time-reversibility).

The switching regime representation also provides an interpretation of the form of the $h-$step-

ahead conditional density (2.10):

$$f_h(x_{t+h}|x_t) = \mathbb{E}\Big[f_h(x_{t+h}|x_t, S_t, S_{t+1}, ..., S_{t+h-1}) \mid x_t\Big]$$
$$= \underbrace{\frac{U'(x_t)D}{U'(x_t)De}}_{\text{conditional probabilities of } S_t \text{ given } X_t} \Pi^{h-1} \underbrace{\phi(x_{t+h})U(x_{t+h})}_{\text{conditional density of } X_{t+h} \text{ given } S_{t+h-1}}, \quad (2.17)$$

where Π^{h-1} is the transition matrix between S_t and S_{t+h-1}. Thus, instead of integrating out the continuously valued intermediate variables $X_{t+1}, ..., X_{t+h-1}$, which is a $(h-1)$ dimensional integral, we can integrate out the embedded discrete variables $S_t, ..., S_{t+h-1}$. This latter is much easier, and has a tractable matrix form because (S_t) is a Markov chain.

The embedded chain representation is the key property of the state process (X_t). Besides this conditional density formula, it is shown in Appendix 4 that the simulation of the trajectory of the process can be conducted via the intermediate variable, whereas the estimation, filtering, smoothing and forecasting formulas of the resulting state-space model are easy to derive without computing high-dimensional integrals.

This tractability remains when D has negative entries, that is, when the embedded switching regime no longer exists. In this case, the conditional distribution of X_{t+1} given X_t satisfies the finite dimensional dependence (FDD) property [see Gouriéroux and Jasiak (2001); Gouriéroux and Monfort (2015)], that is, the conditional density is a linear combination of a finite number of products of functions of X_t and X_{t+1}.

Note that the causality scheme in equation (2.15) is different from the usual Markov switching model, which has the chain:

$$\begin{array}{ccccccc} \dots & S_{t-1} & \longrightarrow & S_t & \longrightarrow & S_{t+1} & \dots \\ & \downarrow & & \downarrow & & \downarrow & \\ \dots & X_{t-1} & & X_t & & X_{t+1} & \dots \end{array}$$

Indeed, first, in our model the switching regime S_t may not exist, since D can have negative entries. Secondly, even when it exists, its transition probabilities are endogenous, since $l(S_{t+1}|S_t, Y_t)$ depends on the observable stochastic variable Y_t. Thus our paper contributes to the literature on endogenous switching regime models [see e.g. Kim et al. (2008); Chang et al. (2017)]. We refer to Appendix 2 for a more detailed comparison of the two models in terms of the goodness of fit of the joint distribution of (X_t, X_{t+1}), in the special case of a gamma benchmark density.

Let us finally discuss the ergodicity of the process. In general, the eigenvalues of Π should be computed numerically to check the conditions in Proposition 3. Nevertheless, the ergodicity

condition is automatically satisfied in the two following cases:

Proposition 5. *If the state process (X_t) is positive, then any of the two following conditions implies the condition of Proposition 3, and hence the ergodicity of (X_t):*

- *all entries of D are nonnegative;*
- *D is symmetric.*

Proof. See Appendix 1.4. □

The first condition is linked to the embedded Markov chain representation of the process. The second condition concerns the time reversibility of the process, and is deeply discussed in the next subsection.

2.4 Time reversibility

Roughly speaking, process (X_t) is time reversible if the reversely ordered process (X_{-t}) has the same dynamics as (X_t). Under the Markov assumption of (X_t), the time reversibility is equivalent to the symmetry of the joint distribution $f(x_t, x_{t+1})$ in the two arguments, that is

$$f(x_t, x_{t+1}) = \phi(x_t)\phi(x_{t+1})\frac{U'(x_t)DU(x_{t+1})}{e'De} = \phi(x_t)\phi(x_{t+1})\frac{U'(x_{t+1})DU(x_t)}{e'De} = f(x_{t+1}, x_t), \tag{2.18}$$

for all values of x_t, x_{t+1}, or equivalently matrix D is symmetric.

How can we characterize the symmetry of D, and hence that of the joint p.d.f. $f(x_t, x_{t+1})$, in terms of the entries of matrix B? We have the following proposition:

Proposition 6. *The joint p.d.f. $f(x_t, x_{t+1})$ is symmetric if and only if B is symmetric, or is antisymmetric*[6]*.*

Proof. Let us denote $V(x) = (1, x, ..., x^J)$. We have $f(x_t, x_{t+1}) = \phi(x_t)\phi(x_{t+1})(V(x_t)BV(x_{t+1}))^2$. Thus the symmetry of $f(x_t, x_{t+1})$ is equivalent to:

$$\Big[V(x_t)BV(x_{t+1}) + V(x_{t+1})BV(x_t)\Big]\Big[V(x_t)BV(x_{t+1}) - V(x_{t+1})BV(x_t)\Big] = 0, \qquad \forall x_t, x_{t+1} > 0.$$

The LHS is the product of two polynomials in (x_t, x_{t+1}). The previous identity is satisfied if and only if at least one of the two multiplicative terms is identically zero, that is to say, B is antisymmetric, or symmetric. □

Thus, when B is symmetric or antisymmetric, the condition $(D - D')e = 0$ is automatically satisfied.

[6] Also called skew-symmetric.

The time-reversibility is satisfied by a large number of time series models considered in the financial literature, including the (log-)normal ARMA processes, the autoregressive gamma process[7], the autoregressive Jacobi process[8] as well as Gaussian and Archimedean copula based time series models considered by Chen and Fan (2006). However, both the theoretical [see e.g. Maskin and Tirole (1988)] and empirical literature have rejected this assumption [see e.g. Ramsey and Rothman (1996), Chen et al. (2000), Darolles et al. (2004), Racine and Maasoumi (2007), Beare and Seo (2014) for evidences of time irreversibility of macro and financial data].

In our framework, the condition $(D-D')e = 0$ can also be satisfied for a non reversible process (X_t). To understand this point, let us introduce the (unique) decomposition of matrix B into the sum of a symmetric matrix $B_1 = \frac{1}{2}(B + B')$, and an antisymmetric matrix $B_2 = \frac{1}{2}(B - B')$. Thus instead of parametrizing $B = B_1 + B_2$, we can parametrize B_1 and B_2, which involves the same number of parameters. This decomposition of B provides us with the corresponding decomposition of D into its symmetric part D_1 and antisymmetric part D_2:

$$
\begin{aligned}
d_{j,k} &= \mu_j \mu_k \sum_{\substack{j_1+j_2=j \\ 0\le j_1,j_2\le J}} \sum_{\substack{k_1+k_2=k \\ 0\le k_1,k_2\le J}} b_{j_1,k_1} b_{j_2,k_2} \\
&= \mu_j \mu_k \sum_{j_1,j_2} \sum_{k_1,k_2} (b_{1,j_1,k_1} + b_{2,j_1,k_1})(b_{1,j_2,k_2} + b_{2,j_2,k_2}) \\
&= \underbrace{\mu_j \mu_k \sum_{j_1,j_2} \sum_{k_1,k_2} (b_{1,j_1,k_1} b_{1,j_2,k_2} + b_{2,j_1,k_1} b_{2,j_2,k_2})}_{:=D_1,\ \text{symmetric}} \qquad (2.19) \\
&\quad + \underbrace{2\mu_j \mu_k \sum_{j_1,j_2} \sum_{k_1,k_2} b_{1,j_1,k_1} b_{2,j_2,k_2}}_{:=D_2,\ \text{antisymmetric}}, \qquad (2.20)
\end{aligned}
$$

where for expository purpose, from the second line on we have omitted the constraints on the indices, that are $j_1 + j_2 = j, 0 \le j_1, j_2 \le J$ and $k_1 + k_2 = k, 0 \le k_1, k_2 \le J$.

Then let us remind that B_1 (resp. B_2) is symmetric (resp. antisymmetric), that is, $b_{1,j_1,k_1} = b_{1,k_1,j_1}$, and $b_{2,j_1,k_1} = -b_{2,k_1,j_1}$. Thus the term D_1 in (2.19) is symmetric in j, k, whereas the term D_2 in (2.20) is antisymmetric in j, k. In other words, we have obtained the symmetric/antisymmetric decomposition of matrix $D = D_1 + D_2$.

As a consequence, the equal margin constraint $(D - D')e = 0$ is equivalent to $D_2 e = 0$, that

[7]The autoregressive gamma process is the time discretized Cox-Ingersol-Ross process, and takes value in $\mathbb{R}_{\ge 0}$, see e.g. Gouriéroux and Jasiak (2006a) for details.

[8]The autoregressive Jacobi process takes values in $[0, 1]$, see e.g. Gouriéroux and Jasiak (2006b); Demni and Zani (2009).

is, the sum of each row of D_2 is zero, with

$$d_{2,j,k} = 2\mu_j\mu_k \sum_{\substack{j_1+j_2=j \\ 0\le j_1,j_2\le J}} \sum_{\substack{k_1+k_2=k \\ 0\le k_1,k_2\le J}} b_{1,j_1,k_1} b_{2,j_2,k_2}, \qquad j,k = 0,...,2J.$$

Thus the condition $D_2 e = 0$ is an orthogonality condition between B_1 and B_2.

Proposition 7. *The equal margin condition $(D-D')e = 0$ is satisfied if and only if the symmetric and antisymmetric components of matrix B are orthogonal.*

These orthogonality conditions would be easy to test in practice. Let us discuss how many restrictions have to be considered. Compared to the model with symmetric B, say, introducing asymmetry leads to at least:

$$n = \frac{J(J+1)}{2} - 2J \tag{2.21}$$

more degrees of freedom, which is positive if and only if $\frac{J+1}{2} > 2$, that is when $J > 3$.. Indeed, matrix B_2 has zero on the diagonal, and $\frac{J(J+1)}{2}$ entries above the diagonal. On the other hand, the system of linear constraints $(D - D')e = 0$ is composed of $2J + 1$ equations. But since $e'(D - D')e = 0$ for any matrix D, the sum of these $2J + 1$ equations are zero. Thus these constraints correspond to $2J$ linearly independent linear equations in entries of B_2.[9]

To summarize, we have obtained a simple parametrisation of our model in the general case including reversible as well as irreversible process. A similar approach has initially been suggested by McCausland (2007) in the context of finite-state Markov chains, in which D is the joint probability distribution of the chain: $d_{i,j} = \mathbb{P}[S_t = i, S_t = j]$. However, McCausland does not provide a parametrization for D_1 and D_2. This is due to the difficulty of satisfying both the linear constraint $D_2 e = 0$, and the non-linear constraint that entries of $D_1 + D_2$ should be nonnegative in his framework. In our model, although D_1, D_2 depend on parameters B_1, B_2 in a quadratic way, the positivity of the entries of $D_1 + D_2$ is not required, and the constraint $D_2 e = 0$ is linear in parameters B_2, and hence can be handled rather easily. This illustrates one of the advantages of our specification of (X_t) compared to, say, an exogenous Markov switching model.

3 Examples

3.1 Case $J = 1$

Let us study the case where $J = 1$.

[9] The number $2J$ is generically attained, except when B or D are of reduced rank. That is, for instance, when $b_{j,k} = 0$ so long as $j = J$ or $k = J$.

The equal margin condition. First, let us discuss the implications of the equal margin condition $(D - D')e = 0$ on the entries of matrix $B = (b_{i,j})_{0 \le i,j \le J}$. We have:

$$f(x_t, x_{t+1}) = \frac{1}{M}\phi(x_t)\phi(x_{t+1})\Big(b_{00} + b_{10}x_t + b_{01}x_{t+1} + b_{11}x_t x_{t+1}\Big)^2.$$

Then matrix D is given by:

$$D = \begin{bmatrix} b_{00}^2 & 2b_{00}b_{01}\mu_1 & b_{01}^2\mu_2 \\ 2b_{00}b_{10}\mu_1 & 2b_{00}b_{11}\mu_1^2 + 2b_{01}b_{10}\mu_1^2 & 2b_{01}b_{11}\mu_1\mu_2 \\ b_{10}^2\mu_2 & 2b_{10}b_{11}\mu_1\mu_2 & b_{11}^2\mu_2^2 \end{bmatrix}.$$

The equal margin condition $(D' - D)e = 0$ is equivalent to:

$$(b_{01} - b_{10})(2b_{00}\mu_1 + \mu_2(b_{10} + b_{01})) = 0, \tag{3.1}$$

$$(b_{01} - b_{10})\mu_2(b_{01} + b_{10} + 2b_{11}\mu_1) = 0, \tag{3.2}$$

$$(b_{01} - b_{10})\mu_1(b_{00} - b_{11}\mu_2) = 0. \tag{3.3}$$

Note that, equation (3.3) can be obtained by summing (3.1) and (3.2).[10] Under the assumption $\mu_1\mu_2 \neq 0$, we have two cases. Either *i*) $b_{01} = b_{10}$, which means that B is symmetric, or *ii*),

$$b_{01} + b_{10} + 2b_{11}\mu_1 = 0 \tag{3.4}$$

$$b_{00} - b_{11}\mu_2 = 0. \tag{3.5}$$

Thus in the case $J = 1$, either B is symmetric, in this case B_1 can take any symmetric matrix value. Or B is anti-symmetric, in this case $B_2 = \frac{b_{01}-b_{10}}{2}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ can take any antisymmetric matrix values, but entries of $B_1 = \begin{bmatrix} b_{00} & \frac{b_{01}+b_{10}}{2} \\ \frac{b_{01}+b_{10}}{2} & b_{11} \end{bmatrix}$ should satisfy the constraints (3.4) and (3.5). This is expected, since the number of extra degree of freedom $n = \frac{J(J+1)}{2} - 2J$ is negative when $J = 1$ [see (2.21)].

The form of Π. A stationary dynamics compatible with $(D - D')e = 0$ is, for instance:

$$f(x_t, x_{t+1}) = \frac{1}{M}\phi(x_t)\phi(x_{t+1})\Big(1 + b_{01}(x_t + x_{t+1})\Big)^2. \tag{3.6}$$

[10] Since $e'D_2e = 0$, see Section 2.4 for details.

Let us now derive the form of the matrix Π of model (3.6). We have:

$$\Pi = \int \frac{\phi(x)}{U'(x)De} \begin{bmatrix} b_{01}^2x^2 + 2b_{01}x + 1 & 2\mu_1 x b_{01}^2 + 2\mu_1 b_{01} & b_{01}^2\mu_2 \\ \frac{x}{\mu_1}(b_{01}^2x^2 + 2b_{01}x + 1) & \frac{x}{\mu_1}(2\mu_1 x b_{01}^2 + 2\mu_1 b_{01}) & \frac{b_{01}^2\mu_2 x}{\mu_1} \\ \frac{x^2}{\mu_2}(b_{01}^2x^2 + 2b_{01}x + 1) & \frac{x^2}{\mu_2}(2\mu_1 x b_{01}^2 + 2\mu_1 b_{01}) & b_{01}^2x^2 \end{bmatrix} \mathrm{d}x,$$

with

$$U'(x)De = b_{01}^2x^2 + 2b_{01}x + 1 + 2\mu_1 x b_{01}^2 + 2\mu_1 b_{01} + b_{01}^2\mu_2.$$

The computation of this matrix involves five numerical integrations of fractional functions, that are $\int \phi(x)\frac{x^j}{U'(x)De}\mathrm{d}x$, where $j = 0, ..., 4$, respectively. We refer to Appendix 3 for a discussion of this numerical integration.

3.2 A stochastic volatility model

A benchmark model with stochastic volatility defines the asset return y_t as:

$$Y_t = \sigma(X_t)\epsilon_t, \tag{3.7}$$

where the ϵ_t's are $IIN(0,1)$ and independent of the nonnegative volatility process $\sigma_t = \sigma(X_t)$. In the existing SV literature, the volatility process σ_t is usually assumed Markov with a conditional distribution of a gamma type [see e.g. Madan and Seneta (1990), Feunou and Tédongap (2012), Creal (2016)]. In model (3.7), the gamma type volatility model can be obtained when we take $\sigma(X_t) = X_t$, where the joint distribution of (X_t, X_{t+1}) follows the polynomial expansion form, with a gamma benchmark density.[11]

Let us assume that the joint pdf of (X_t, X_{t+1}) is of the form (2.3), where ϕ is the p.d.f. of a gamma distribution with shape parameter α and scale parameter c:

$$\phi(x) = \frac{x^{\alpha-1}\exp(-x/c)}{\Gamma(\alpha)c^\alpha} := g(x, \alpha, c), \qquad \text{say.}$$

Under the gamma density assumption, the function $\phi(x)x^j$ is still a gamma density (up to a multiplicative constant). Thus the conditional density $f(x_{t+1}|x_t)$ is a weighted average of gamma densities, if the latent switching regime exists. These gamma densities share the same scale parameter, and their degrees of freedom are $\alpha, \alpha+1, ..., \alpha+2J$, respectively. They correspond to different levels of risk: the larger S_t, the larger the conditional expectation of X_{t+1} given S_t.

[11] Nevertheless, in our application we will consider an inverse-gamma type specification, that is: $\sigma(X_t) = 1/\sqrt{X_t}$. This specification has the advantage of leading to closed form expressions of the marginal, and pairwise joint densities. The gamma-type model was initially introduced by Madan and Seneta (1990), who acknowledges that such model has the disadvantage of not being able to capture the heavy tail of asset returns. In our specification, the marginal return is heavy tailed when the benchmark density is gamma.

As an illustration, let us consider the joint density:

$$f(x_t, x_{t+1}) = \frac{1}{M}\phi(x_t)\phi(x_{t+1})\Big(1 + b_{11}x_t x_{t+1}\Big)^2,$$

where $b_{11} > 0$. In this case the conditional probabilities of S_t given X_t are the components of vector $\frac{U'(x_t)D}{U'(x_t)De}$, and are equal to:

$$\frac{1}{U'(x_t)De}\Big(1, 2b_{01}\mu_1 x_t, b_{01}^2\mu_2 x_t^2\Big)'.$$

3.3 Comparison with the copula literature

The specification of the joint distribution (X_t, X_{t+1}) of a Markov process is also used in copula-based time series models [see e.g. Beare (2010), Joe (2014)]. A (bivariate) copula is the cdf of a bivariate distribution on $[0,1] \times [0,1]$ with uniform margins. In this literature, the marginal distribution of the process is flexible, but up to now the copula is usually assumed to be of a rather simple form such as Archimedean [see e.g. Chen and Fan (2006)], which, unlike our model, does not lead to tractable formulas for the transition densities $f(x_{t+h}|x_t)$ at all horizons h.

A family of copulas that solves this difficulty is the family of polynomial copulas. That is, the copula function (and hence the corresponding copula density) is a polynomial in the two arguments. The simplest polynomial copula is the Farlie-Gumbel-Morgenstern (FGM) copula[12]

$$C(x_1, x_2) = x_1 x_2(1 + \theta(1 - x_1)(1 - x_2)), \qquad |\theta| < 1/3. \tag{3.8}$$

Since the corresponding copula density is of finite dimensional dependence, the transition density $f(x_{t+h}|x_t)$ has a tractable form at all horizons h.[13]This approach has not been chosen by us, since, in many applications, the uniform margin of the state process does not lead to a simple expression for the conditional density $g(y_t|\underline{y_t}, x_t)$

Our specification of (X_t) has a different philosophy. It is based on a (in practice, simple) benchmark density rather than a simple marginal density. As a consequence, our model usually does not lead to a tractable copula. That is, the resulting marginal density $\phi(x_t)\frac{U'(x_t)De}{e'De}$ is not

[12]See also Sancetta and Satchell (2004) for a large, flexible family of polynomial copulas, the Bernstein copula.

[13]Indeed, the copula density is $c(x_1, x_2) = 1 + \theta(1 - 2x_1)(1 - 2x_2)$. Thus if a Markov process (X_t) has uniform margin and joint density c for the pair (x_t, x_{t+1}), then the one-step-ahead conditional density is:

$$\begin{aligned} f(x_{t+2}|x_t) &= \int_0^1 \Big[1 + \theta(1 - 2x_t)(1 - 2x_{t+1})\Big]\Big[1 + \theta(1 - 2x_{t+1})(1 - 2x_{t+2})\Big]\mathrm{d}x_{t+1} \\ &= 1 + \theta^2(1 - 2x_{t+2})(1 - 2x_t)\int_0^1 (1 - 2x_{t+1})^2\mathrm{d}x_{t+1} = 1 + \frac{\theta^2}{3}(1 - 2x_{t+2})(1 - 2x_t). \end{aligned}$$

Similarly, the $h-$step-ahead conditional density is: $f(x_{t+h}|x_t) = 1 + \frac{\theta^h}{3^{h-1}}(1 - 2x_t)(1 - 2x_{t+h})$.

uniform in general. Indeed, if the marginal density $f_0(x_t) = \phi(x_t)\frac{U'(x_t)De}{e'De}$ were uniform on the domain $\mathcal{X} = [0,1]$, then the benchmark ϕ should have the form:

$$\phi(x_t) = \frac{e'De}{U'(x_t)De}, \qquad \forall x_t \in [0,1].$$

The RHS of this equation depends on $B = (b_{i,j})_{i,j}$ as well as the moment parameters $\mu_1, ..., \mu_{2J}$. They satisfy the constraints :

$$(D' - D)e = 0 \tag{3.9}$$

$$\int_0^1 \frac{e'De}{U'(x_t)De} U(x_t)\mathrm{d}x_t = e. \tag{3.10}$$

In general, equation (3.10) does not lead to a tractable relationship between μ_j's and B.

4 Forecasting and filtering

Let us now derive the predictive formulas of our model. This includes *i*) the forecasting, that is, the distribution of Y_{T+h} given $\underline{Y_T}$; *ii*) the smoothing, that is, the conditional distribution of the state variables X_t given $\underline{Y_T}$ for each $t = 1, ..., T-1$; *iii*) the filtering, that is, the conditional distribution of X_T given $\underline{Y_T}$.

4.1 Forecasting

Proposition 8. *The conditional density of Y_t given the past is:*

$$l(y_t|\underline{y_{t-1}}) = P'(\underline{y_{t-1}})g(y_t|\underline{y_{t-1}}), \tag{4.1}$$

where the column vector $g(y_t|\underline{y_{t-1}})$ is defined by:

$$g(y_t|\underline{y_{t-1}}) = \int l(y_t|x_t, \underline{y_{t-1}})\phi(x_t)U(x_t)\mathrm{d}x_t, \tag{4.2}$$

and the row vector $P'(\underline{y_{t-1}})$ is computed recursively by:

$$P'(\underline{y_t}) = P'(\underline{y_{t-1}})\Pi(\underline{y_t}), \tag{4.3}$$

$$\text{with initial condition } P'(\underline{y_0}) = \frac{e'D}{e'De}, \tag{4.4}$$

and matrix $\Pi(\underline{y_t})$ *is given by:*

$$\Pi(\underline{y_t}) := \frac{1}{l(y_t|\underline{y_{t-1}})} \int \phi(x_t) \frac{U(x_t)U'(x_t)D}{U'(x_t)De} l(y_t|x_t, \underline{y_{t-1}}) \mathrm{d}x_t. \tag{4.5}$$

Proof. See Appendix 1.5. □

The dependence of y_t on its whole past $\underline{y_{t-1}}$ is summarized by a finite-dimensional vector $P(\underline{y_{t-1}})$, often called *mimicking factor* in Finance [see e.g. Huberman et al. (1987), Gouriéroux and Jasiak (2001)].

When the state process has the embedded switching regime interpretation, we have the following causal chain:

$$\begin{array}{ccccccc} \dots & (X_{t-1} \to S_{t-1}) & \longrightarrow & (X_t \to S_t) & \longrightarrow & (X_{t+1} \to S_{t+1}) & \dots \\ \dots & \downarrow & & \downarrow & & \downarrow & \dots \\ \dots & Y_{t-1} & \longrightarrow & Y_t & \longrightarrow & Y_{t+1} & \dots \end{array}$$

The vector $P(\underline{y_{t-1}})$ is the vector of conditional probabilities of embedded chain S_{t-1} belonging to the $2J+1$ different regimes given $\underline{y_{t-1}}$:

$$P'(\underline{y_{t-1}}) = \Big(\mathbb{P}[S_{t-1} = 0 \mid \underline{y_{t-1}}], ..., \mathbb{P}[S_{t-1} = 2J \mid \underline{y_{t-1}}]\Big)'. \tag{4.6}$$

Thus recursive formula (4.3) is the analogue of the Kitagawa filter for hidden Markov models [see Kitagawa (1987)], except that the discrete chain (S_t) is endogenous in our framework. More precisely, conditional on the history of Y_t, S_t is a (time-inhomogeneous) Markov chain, with transition matrix $\Pi(\underline{y_t})$ at each date t:

$$\pi_{i,j}(\underline{y_t}) = \mathbb{P}[S_{t+1} = i | S_t = j, \underline{y_t}].$$

This transition matrix is state-dependent. Indeed, given S_{t-1}, the future state variable S_t, and the observable variable Y_t are dependent by means of X_t. Thus the transition of the latent state variable depends on the current value of Y_t. The matrix $\Pi(\underline{y_t})$ is not a stochastic matrix. Indeed, the sums of the entries of each of its row are:

$$\Pi(\underline{y_t})e = \frac{1}{l(y_t|\underline{y_{t-1}})} \int \phi(x_t) \frac{U(x_t)U'(x_t)De}{U'(x_t)De} l(y_t|x_t, \underline{y_{t-1}}) \mathrm{d}x_t = \frac{1}{l(y_t|\underline{y_{t-1}})} g(y_t|\underline{y_{t-1}}),$$

is not equal to e. Nevertheless, by construction, $\Pi(\underline{y_t})$ is such that the entries of vector $P(\underline{y_t})$

sum up to one:

$$P'(\underline{y_t})e = P'(\underline{y_{t-1}})\Pi(\underline{y_t})e = \frac{1}{l(y_t|\underline{y_{t-1}})}P'(\underline{y_{t-1}})g(y_t|\underline{y_{t-1}}) = 1,$$

by equation (4.1).

The entries of the matrix in recursive equation (4.3) can be computed numerically by using an adaptive quadrature method. It suffices to compute a finite number $4J + 2$ of univariate integrals, that are: $\int \phi(x)x_t^n \frac{l(y_t|x_t,\underline{y_{t-1}})}{U'(x_t)\overline{De}}\mathrm{d}x_t, \forall n \in [|0, 4J|]$. This method is much faster than the standard Monte-Carlo simulation. Nevertheless, the latter approach is interesting to discuss, as it is, roughly speaking, the analogue of the updating of the population in a particle filter [see e.g. Pitt and Shephard (1999); Koopman et al. (2015)]. Indeed, in the Monte Carlo approach, a large number of trajectories of (X_t) are generated in order to provide an approximation of the conditional distribution $l\Big((x_t)_{t=1,\dots,T}|(y_t)_{t=1,\dots,T}\Big)$. This has a computational cost that is similar to the computation of the integral $\int \phi(x)x_t^n \frac{l(y_t|x_t,\underline{y_{t-1}})}{U'(x_t)\overline{De}}\mathrm{d}x_t$ via Monte-Carlo simulation. We refer to Appendix 3 for a more detailed comparison between the two numerical methods.

Finally, our state-space model is a flexible generalization of the model of Nieto-Barajas and Walker (2002), Creal (2016), who specify (X_t) as an ARG process (see Appendix 6). The ARG process is also associated with an embedded latent regime variable (Z_t), which is discrete (but infinite) valued. Then they propose to approximate its dynamics by a Markov chain with a finite (but large) number of states. This allows the authors to derive similar filtering and forecasting formulas. Our model generalizes Creal's result in several aspects. First, the specification of the domain, as well as the dynamics of our state process is flexible. Second, Creal's method requires truncating an infinite transition matrix and keeping a very large number of possible values for the intermediate variable S_t,[14] whereas in our approach no approximation is involved, and the number of the "pseudo" states is fixed and much smaller.

Proposition 8 provides the one-step-ahead nonlinear forecasting formula. Due to the feedback effect, that is the dependence of $l(y_t|\underline{y_{t-1}}, x_t)$ in the lagged observations $\underline{y_{t-1}}$, the longer horizon forecast formula $l(y_{t+h}|\underline{y_{t-1}})$ has no simple expression[15]. Nevertheless, simulation of the trajectories of (X_t) (and hence also that of (Y_t)) can be conducted rather easily (see Appendix 4).

[14] Fore instance, Creal (2016) proposed to use a finite chain with 3000 states to approximate the dynamics of process (S_t).

[15] Except in the case without feedback effect $l(y_t|\underline{y_{t-1}}, x_t) = l(y_t|x_t)$, which will be analysed in subsection 4.3.

4.2 Filtering and smoothing

The simple forecasting formula in Proposition 8 is associated with a simple expression for the predictive density of X_t given the past observations. We have:

Corollary 3.

$$l(x_t|\underline{y_{t-1}}) = P'(\underline{y_{t-1}})\phi(x_t)U(x_t). \tag{4.7}$$

Proof. See Appendix 1.5. □

Similarly, the filtering density is:

Proposition 9. *The filtering density of X_t given the observables $\underline{Y_t}$ is:*

$$l(x_t|\underline{y_t}) = \phi(x_t)\frac{P'(\underline{y_{t-1}})U(x_t)l(y_t|x_t,\underline{y_{t-1}})}{l(y_t|\underline{y_{t-1}})}. \tag{4.8}$$

Proof. See Appendix 1.6. □

Let us now infer, for each $t < T$, the smoothing distribution $l(x_t|\underline{y_T})$. We have the following proposition:

Proposition 10. *The smoothing density is:*

$$l(x_t|\underline{y_T}) = \frac{1}{P'(\underline{y_{t-1}})g(y_t|\underline{y_{t-1}})}\frac{P'(\underline{y_{t-1}})\Big[\phi(x_t)\frac{U(x_t)U'(x_t)D}{U'(x_t)De}l(y_t|\underline{y_{t-1}},x_t)\Big]Q_{t+1}}{P'(\underline{y_{t-1}})\Pi(\underline{y_t})Q_{t+1}}, \qquad \textit{for } t < T \tag{4.9}$$

where vector Q_t is defined backward by:

$$Q_{t-1} = \Pi(\underline{y_{t-1}})Q_t, \qquad \forall t < T, \tag{4.10}$$

$$\textit{with terminal condition } Q_T = g(y_T|\underline{y_{T-1}}). \tag{4.11}$$

Proof. See Appendix 1.7. □

This smoothing equation is similar to the forward-backward smoothing algorithm for hidden Markov model [see e.g. Scott (2002)]. Again, it can be interpreted in terms of switching regime. Indeed, we have:

$$\begin{aligned} l(x_t|\underline{y_T}) &\propto l(x_t|\underline{y_t})l(\overline{y_{t+1}}|x_t,\underline{y_t}) \propto l(x_t|\underline{y_t})\sum_{j=0}^{2J}\mathbb{P}[S_t=j|x_t]l(\overline{y_{t+1}}|x_t,S_t=j) \\ &\propto \Big(P'(\underline{y_{t-1}})\phi(x_t)U(x_t)l(y_t|\underline{y_{t-1}},x_t)\Big)\Big(\frac{U'(x_t)D}{U'(x_t)De}Q_{t+1}\Big), \end{aligned}$$

where $\overline{y_{t+1}} = (y_{t+1}, ..., y_T)$, and $Q_{t+1} = \Pi(\underline{y_{t+1}}) \cdots \Pi(\underline{y_{T-1}}) g(y_T|\underline{y_{T-1}})$ is the vector of densities of $\overline{y_{t+1}}$ given $S_t = j$, for $j = 0, ..., 2J$.

4.3 Model without feedback

Let us now consider the case without feedback, when the conditional density of the observable variable depends only on x_t:

$$l(y_t|\underline{y_{t-1}}, \underline{x_t}) = l(y_t|x_t). \tag{4.12}$$

This assumption can lead to two potential simplifications. First, the conditional transition matrix $\Pi(\underline{y_t})$ depends on y_t only, up to a multiplicative constant $\frac{1}{l(y_t|y_{t-1})}$. Thus, when the domain $\mathcal{Y}$ of Y_t is finite, the recursive forecasting algorithm involves only a finite number of numerical integrals to be conducted only once.

The second simplification concerns the prediction of the observable variable y_{T+h} at any horizon h. We have an explicit formula:

Proposition 11. *In the model without feedback, the $h-$step-ahead predictive density is:*

$$l(y_{T+h}|\underline{y_T}) = P'(\underline{y_T})\Pi^{h-1} g(y_{T+h}), \qquad \forall h \in \mathbb{N}, \tag{4.13}$$

where the vector function $g(y_t) = \int U(x_t) l(y_t|x_t)\phi(x_t)\mathrm{d}x_t$ is the same as in (4.2)*.*

Proof. See Appendix 1.8. □

Then the computation of the predictive distribution at all horizons necessitates only the computation of Π, as well as the vector function $g(y_t)$. If the latter has a simple form, then the total computational cost is very low. This is for instance the case, when:

- ϕ is a gamma density, and the conditional distribution $l(y_t|x_t)$ belongs to the exponential family, such as Poisson $\mathcal{P}(x_t)$, normal $\mathcal{N}(0, x_t^2)$, or $\mathcal{N}(0, 1/x_t^2)$ [see e.g. Creal (2016) for a larger list and applications to state-space time series].
- ϕ is the uniform density, or a beta density on $[0, 1]$, and $l(y_t|x_t)$ is binomial $Bin(n, x_t)$ for an integer n, or multivariate Gaussian with stochastic correlation coefficient x_t, in a stochastic correlation model.

5 A non-parametric approach

Our specification of the state process is, in some sense, non-parametric. In the first subsection, we explain how model (2.3) approximates the dynamics of any given Markov process. The

approximation is based on the orthonormal decomposition of a square rooted density. Then similar dynamics in the literature, especially those based on the decomposition of the density itself, are briefly discussed.

5.1 Polynomial decomposition of a square rooted density

Let us assume that[16] there exists an orthonormal basis of polynomials $P_i(x)$, with $deg(P_i) = i$ for the L^2 space associated with measure $\phi(x)\mathrm{d}x$. Then $(P_i(x_t)P_j(x_{t+1}))$, where i, j varying, is an orthonormal polynomial basis for the L^2-space associated with the product measure $\phi(x_t)\phi(x_{t+1})\mathrm{d}x_t\mathrm{d}x_{t+1}$. Since $f(x_t, x_{t+1})$ integrates to unity, the ratio $\sqrt{\frac{f(x_t,x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}$ belongs to the L^2-space associated with the probability measure $\phi(x_t)\phi(x_{t+1})\mathrm{d}x_t\mathrm{d}x_{t+1}$:

$$\iint \Big[\sqrt{\frac{f(x_t, x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}\Big]^2 \phi(x_t)\phi(x_{t+1})\mathrm{d}x_t\mathrm{d}x_{t+1} = 1. \tag{5.1}$$

Thus we get the orthonormal decomposition:

$$\sqrt{\frac{f(x_t, x_{t+1})}{\phi(x_t)\phi(x_{t+1})}} = \sum_{i,j=0}^{\infty} a_{i,j} P_i(x_t)P_j(x_{t+1}), \tag{5.2}$$

where the coordinates $a_{i,j}$ are the inner products between $P_i(x_t)P_j(x_{t+1})$ and $\sqrt{\frac{f(x_t,x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}$:

$$a_{i,j} = \iint \phi(x_t)\phi(x_{t+1})\sqrt{\frac{f(x_t, x_{t+1})}{\phi(x_t)\phi(x_{t+1})}} P_i(x_t)P_j(x_{t+1})\mathrm{d}x_t\mathrm{d}x_{t+1}.$$

The infinite sum in (5.2) converges in the sense of L^2, and $\sum_{i,j=0}^{\infty} a_{i,j}^2 = 1$, since function $\sqrt{\frac{f(x_t,x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}$ is of unit norm [see equation (5.1)]. Thus a natural approximation of $f(x_t, x_{t+1})$ is obtained by truncating the RHS of equation (5.2), and taking the square:

$$f(x_t, x_{t+1}) \approx \phi(x_t)\phi(x_{t+1})\Big[\sum_{i,j=0}^{J} a_{i,j} P_i(x_t)P_j(x_{t+1})\Big]^2. \tag{5.3}$$

In other words, the RHS is the orthogonal projection of $\sqrt{\frac{f(x_t,x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}$ onto the linear space generated by of $\{P_i(x_t)P_j(x_{t+1}), 0 \le i, j \le J\}$. This function is not yet a density, but can be normalized to obtain the following approximating density:

$$f_J(x_t, x_{t+1}) = \frac{1}{M_J}\phi(x_t)\phi(x_{t+1})\Big[\sum_{i,j=0}^{J} a_{i,j} P_i(x_t)P_j(x_{t+1})\Big]^2, \tag{5.4}$$

[16] Such an orthonormal basis exists under rather mild conditions [see e.g. Filipović et al. (2013), Thm 1].

with $M_J = \sum_{i,j=0}^{J} a_{i,j}^2$, such that function $f_J(x_t, x_{t+1})$ integrates to one. This approximating density can be as precise as possible, when J is large. More precisely we have:

Proposition 12. *The sequence of densities f_J approximates f arbitrarily well in terms of the Hellinger distance*[17]*, when J goes to infinity:*

$$\iint \left| \sqrt{f_J(x_t, x_{t+1})} - \sqrt{f(x_t, x_{t+1})} \right|^2 \mathrm{d}x_t \mathrm{d}x_{t+1} \longrightarrow 0. \tag{5.5}$$

Proof. See Appendix 1.6. □

This approximation result suggests joint choices of the benchmark density and orthonormal polynomials. For instance when ϕ is a gamma density, the orthonormal polynomials can be the generalized Laguerre polynomials [see e.g. Szeg (1939)]. In practice, however, since the expression of the density involves the square of the polynomial $\sum_{i,j=0}^{J} a_{i,j} P_i(x_t) P_j(x_{t+1})$, it is much more convenient to re-parameterize the orthonormal polynomials using the canonical, power polynomials. Thus we get a density of the same form as (2.3).

5.2 Comparison with the polynomial decomposition of a density

It is interesting to compare our specification of the state dynamics with the direct approximation of (univariate) densities by means of polynomial expansion [see e.g. Jarrow and Rudd (1982), Corrado and Su (1996), Aït-Sahalia (2002), Filipović et al. (2013), Xiu (2014)]. This literature is based on the idea that, under integrability conditions, each univariate density function f, such as the conditional density of X_{t+1} given X_t can be decomposed into the product of a benchmark density, and an infinite polynomial sum:

$$f(x_{t+1}|x_t) = \phi(x_{t+1}) \sum_{i=0}^{\infty} a_i(x_t) P_i(x_{t+1}), \tag{5.6}$$

where ϕ is the benchmark density (which is usually Gaussian in the Finance literature), and P_i are the corresponding orthonormal polynomials. Then this literature proposes to truncate the decomposition (5.6) up to a finite order J to obtain an approximation of f. While this approach is simpler since it does not involve the square of the polynomials, its major drawback is that the truncated version of (5.6) is not a proper density, since it is not nonnegative. This leads to several difficulties. First, it is not possible to evaluate the accuracy of this expansion, since neither the Kullback distance, nor the Hellinger distance between the expanded density and the benchmark density can be defined. Second, negative probabilities typically lead to arbitrage opportunities when it comes to derivative pricing. Our modelling strategy differs from this literature in two

[17] See e.g. Beran (1977) for a discussion of the Hellinger distance.

respects. First, our method guarantees the positivity of the conditional density. Second, the decomposition (5.6) requires a rather strong integrability condition:

$$\int \frac{f^2(x_{t+1}|x_t)}{\phi(x_{t+1})}\mathrm{d}x_{t+1} < \infty,$$

which has been shown to be sometimes violated [see Aït-Sahalia (2002) for a discussion], whereas our decomposition of $\sqrt{\frac{f(x_t,x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}$ does not require such extra condition. Finally, instead of considering the conditional distribution of X_{t+1} given X_t,[18] we specify the joint distribution of (X_t, X_{t+1}). While the dynamics of a stationary Markov process can be characterized by either of these two distributions, the approach by joint distribution facilitates the derivation of stationarity conditions and of the stationary density.

6 Application

Let us now discuss the implementation of this type of state-space models. We first discuss the estimation approaches, and then consider the application to the stochastic volatility model.

6.1 Estimation methods

6.1.1 Maximum likelihood approach

The log-likelihood function has the form:

$$\log \ell(\theta) = \sum_{t=1}^{T} \log l(y_t|\underline{y_{t-1}}, \theta) = \sum_{t=1}^{T} \log \int_0^\infty l(y_t|\underline{y_{t-1}}, x_t, \theta) l(x_t|\underline{y_{t-1}}, \theta)\mathrm{d}x_t, \tag{6.1}$$

where the predictive density $l(x_t|\underline{y_{t-1}})$ is given by the forecasting formula of Section 4. The maximum likelihood estimator is consistent, asymptotically normal and asymptotically efficient. The log-likelihood function can be computed recursively, with a computational cost that is lower than the cost of simulation-based techniques such as particle filters.

6.1.2 Maximum composite likelihood approach

Let us now introduce the maximum composite likelihood estimation (MCLE) method [see e.g. Varin and Vidoni (2008), Varin et al. (2011), Gouriéroux and Monfort (2016), Gouriéroux et al. (2016) for reviews] that is particularly suited for our model. Although it is (slightly) less efficient[19]

[18] Gallant and Nychka (1987); Gallant and Tauchen (1989) also propose to specify the conditional distribution, although their model are based on the decomposition of its square root, and hence ensures nonnegativity.

[19] In most financial applications, where the sample size of the data is extremely large, the efficiency loss is largely compensated by the computational gain.

than the full maximum likelihood estimation, its computational cost is extremely low, and is similar to that of the Generalized Method of Moments (GMM). The composite likelihood is based on the following closed form expression of the joint distribution of (Y_t, Y_{t+h}):

Lemma 2. Under Assumption (4.12), we have, for each $h \geq 1$:

$$f_Y(y_t, y_{t+h}) = g'(y_t) \frac{D\Pi^{h-1}}{e'De} g(y_{t+h}), \tag{6.2}$$

Proof. See Appendix 1.9. □

This simple expression is to be compared with the joint distribution of (X_t, X_{t+h}), that is $f(x_t, x_{t+h}) = \phi(x_t)\phi(x_{t+1})U'(x_t)\frac{D\Pi^{h-1}}{e'De}U(x_{t+h})$.

The (order m) pairwise composite likelihood function is defined by:

$$\ell_{CL}(\theta) = \arg\min \sum_{t=1}^{T} \sum_{h=1}^{\min(m,T-t)} w_h \log f(y_t, y_{t+h}|\theta),$$

where θ denotes the set of parameters of the model, and w_h are nonnegative weights. This is a pseudo-log-likelihood function, which evaluates the joint densities of all pairs (y_t, y_{t+h}), so long as h is smaller than an integer m. Maximizing this function leads to the MCLE:

$$\hat{\theta} = \arg\min \ell_{CL}(\theta). \tag{6.3}$$

Varin and Vidoni (2008) show that under mild regularity and identification conditions, the estimator (6.3) is asymptotically consistent and normally distributed. It is however typically not efficient [see e.g. Gouriéroux and Monfort (2016) for a discussion on similar more efficient methods]. Nevertheless, Varin and Vidoni (2008) show, via a simulation experiment, that the efficiency loss with respect to, say, the MLE is rather small, compared to usual GMM [see e.g. Andersen and Sørensen (1996)], which requires a similar computational cost as the MCLE method.

Let us talk about the choice of m. Roughly speaking, if m is too small, the amount of information contained in the composite likelihood function is reduced, and hence the efficiency of the MCLE would be low. At the same time, the computational effort required is proportional to m; thus in the application, we set $m = 5$. As for the weights w_h, we set them to be $w_h = 0.9^h, \forall h$. In other words, pairs with a smaller distance have a higher weight, although we let the weight w_h decrease slowly, in order to reflect the high persistence of financial returns.

The pairwise composite likelihood is easy to compute, when function g allows for closed form expression (see the discussion below Lemma 2). This is made possible by carefully choosing the benchmark density ϕ and the conditional density $l(y_t|x_t)$. Indeed, according to the polynomial

decomposition approach of Section 5, we can choose any benchmark density subject to some minimal integrability constraints. When g is explicit, the computational gain of the method with respect to MLE is significant[20].

6.2 A stochastic volatility application

6.2.1 The models

i) **Model M1**

Let us first consider the model M1:

$$y_t = \frac{1}{\sqrt{x_t}}\epsilon_t, \tag{6.4}$$

where (ϵ_t) is i.i.d. standard normal, independent of (X_t), and the process (X_t) follows the Markov dynamics introduced in Section 2, with a gamma benchmark density $\phi(x_t) = \frac{1}{\Gamma(\alpha)c^\alpha} x_t^{\alpha-1} e^{-x_t/c}$, with $\alpha, c > 0$.

Since we can multiply X_t by a constant and accordingly divide ϵ_t by the same constant, for identification purpose we assume without loss of generality: $\mathbb{E}[\epsilon_t^2] = 1$. Secondly, in the joint p.d.f. of equation (2.3), we can multiple all the coefficients $b_{i,j}$ by a same constant. Therefore, we set $b_{0,0} = 1$.[21]

Thus $1/\sqrt{x_t}$ is the stochastic, latent volatility of y_t, that is, $1/x_t = \mathbb{V}[y_t|\underline{y_{t-1}}, \underline{x_{t-1}}]$. Under the specification (6.4), the marginal density of Y_t is $g(y_t)'\frac{De}{e'De}$, where the components of $g(y_t)$ are the conditional densities of y_t in each "regime" given by:

$$\begin{aligned} g_j(y_t) &= \int_0^\infty \phi(x_t)\frac{x_t^j}{\mu_j}\frac{\sqrt{x_t}}{\sqrt{2\pi}} e^{-\frac{y_t^2 x_t}{2}} \mathrm{d}x_t = \frac{\Gamma(\alpha+j+\frac{1}{2})}{\Gamma(\alpha)(\frac{y_t^2}{2}+1/c)^{\alpha+j+\frac{1}{2}}\sqrt{2\pi}\mu_j} \\ &= \frac{\Gamma(\alpha+j+\frac{1}{2})}{c^{\alpha+j}\Gamma(\alpha+j)(\frac{y_t^2}{2}+1/c)^{\alpha+j+\frac{1}{2}}\sqrt{2\pi}} = \frac{\sqrt{c}\Gamma(\alpha+j+\frac{1}{2})}{\Gamma(\alpha+j)(\frac{cy_t^2}{2}+1)^{\alpha+j+\frac{1}{2}}\sqrt{2\pi}}, \end{aligned}$$

since $\mu_j = \frac{\Gamma(\alpha+j)}{\Gamma(\alpha)}$. Each component function $g_j(\cdot)$ is the density of a re-scaled Student's $t-$distribution, which has been widely used in Finance to account for the heavy-tails of asset returns [see e.g. Harvey et al. (1994)]. More precisely, we have $g_j(y) = \sqrt{c(\alpha+j)}h_j(\sqrt{c(\alpha+j)}y)$, where h_j is the density of the standard $t-$distribution with $2\nu + 2j$ degrees of freedom.

Finally, in Model M1, we assume that matrix D is symmetric. By Proposition 6, this is equivalent to B being symmetric, or antisymmetric. Since the polynomial expansion approach

[20] Note that its computation is parallelisable.

[21] This normalization constraint implies that $b_{0,0} \neq 0$. This implicit assumption is motivated by the polynomial expansion formula developed in Section 5.

suggests that the coefficient $b_{0,0}$ is non zero, we only consider the case where B is symmetric. We estimate model M1 for $J = 2, 3, 4$ in order to illustrate the improvement of the fit when J increases.

ii) **Model M2.**

Model M1 assumes that the return is conditionally normal. This latter distribution is symmetric, and thus the model does not allow for conditional skewness. A simple, yet flexible generalization of Model M1 is Model M2, where we keep the same dynamics for the state variable (X_t), but the standard normal density of error ϵ_t is replaced by:

$$h(\epsilon) = \frac{1}{M_\epsilon}\psi(\epsilon)\big(\beta_0 + \sum_{i=1}^{I}\beta_i\epsilon^i\big)^2, \tag{6.5}$$

where $\beta_0 = 1$, ψ is the standard normal density, and the normalization constant M_ϵ is such that h integrates to unity. That is,

$$M_\epsilon = 1 + \sum_{k=2}^{2I} \sum_{i,l=0, i+l=k}^{I} \beta_i\beta_l\nu_k,$$

where $\nu_k = \int_{-\infty}^{\infty}\psi(\epsilon)\epsilon^k \mathrm{d}\epsilon$ is the $k-$th moment of the standard normal distribution; we have, $\nu_{2k+1} = 0$, and $\nu_{2k} = \frac{(2k)!}{2^k k!}$. Thus the density function $h(\epsilon)$ has the same form as that of (X_t, X_{t+1}). It is obtained by squaring and renormalizing the polynomial expansion of a square rooted given univariate density, with respect to the benchmark density ψ. It is therefore a new, flexible alternative to the parametric skewed distributions proposed in the literature [see e.g. Fernández and Steel (1998), Zhu and Galbraith (2010), Ferreira and Steel (2012)]. For instance, when $I = 1$, the density of the error is:

$$h(\epsilon) = \psi(\epsilon)\frac{1 + 2\beta_1\epsilon + \beta_1^2\epsilon^2}{1 + \beta_1^2}. \tag{6.6}$$

Under this distributional assumption, the skewness of β_1 is equal to:

$$\frac{\mathbb{E}[(\epsilon - \mathbb{E}[\epsilon])^3]}{(\mathbb{V}[\epsilon])^{3/2}} = \frac{4\beta_1^3(1 - 3\beta_1^2)}{\sqrt{(1 + \beta_1^2)(1 + 3\beta_1^2)}}.$$

Thus ϵ has negative skewness if $\beta_1 \in]-\frac{1}{\sqrt{3}}, 0[$.

Similarly, when $I = 2$, we get:

$$h(\epsilon) = \psi(\epsilon)\frac{1 + 2\beta_1\epsilon + (\beta_1^2 + 2\beta_2)\epsilon^2 + 2\beta_1\beta_2\epsilon^3 + \beta_2^2\epsilon^4}{1 + \beta_1^2 + 2\beta_2 + 3\beta_2^2}. \tag{6.7}$$

Under the density specification (6.5), the distribution of ϵ_t is no longer symmetric with respect to 0. Therefore, positive and negative past returns have different impacts on the forecast of the future volatility. This is the so-called leverage effect, volatility feedback, or asymmetric volatility [see e.g. Harvey and Shephard (1996), Bollerslev et al. (2006)]. As a comparison, the standard SV literature addresses the leverage effect using continuous time diffusion models [see e.g. Harvey and Shephard (1996)]. This literature has several drawbacks. First, numerical approximations, such as the Euler scheme have to be used to time-discretize the data. This induces approximation error, as well as time reversible discretized process[22]. Second, in these discretized diffusion models, the return cannot have heavy tail. Third, the model estimation is computationally cumbersome.

Under model (6.7), the components of the marginal distribution of y_t in each regime are:

$$\begin{aligned} g_j(y_t) &= \int_0^\infty \phi(x_t)\frac{x_t^j}{\mu_j}\frac{\sqrt{x_t}}{\sqrt{2\pi}(1+\beta_1^2+2\beta_2+3\beta_2^2)}e^{-\frac{y_t^2 x_t}{2}}\Big[1+2\beta_1\sqrt{x_t}y_t+(\beta_1^2+2\beta_2)x_t y_t^2+2\beta_1\beta_2 x_t^{\frac{3}{2}}y_t^3+\beta_2^2 y_t^4\Big]\mathrm{d}x_t \\ &= \frac{1}{c^{\alpha+j}\sqrt{2\pi}(1+\beta_1^2+2\beta_2+3\beta_2^2)\Gamma(\alpha+j)}\Big[\frac{\Gamma(\alpha+j+\frac{1}{2})}{(\frac{y_t^2}{2}+1/c)^{\alpha+j+\frac{1}{2}}}+2\beta_1 y_t\frac{\Gamma(\alpha+j+1)}{(\frac{y_t^2}{2}+1/c)^{\alpha+j+1}} \\ &\quad +(\beta_1^2+2\beta_2)y_t^2\frac{\Gamma(\alpha+j+\frac{3}{2})}{(\frac{y_t^2}{2}+1/c)^{\alpha+j+\frac{3}{2}}}+2\beta_1\beta_2\frac{\Gamma(\alpha+j+2)}{(\frac{y_t^2}{2}+1/c)^{\alpha+j+2}}+\beta_2^2\frac{\Gamma(\alpha+j+\frac{5}{2})}{(\frac{y_t^2}{2}+1/c)^{\alpha+j+\frac{5}{2}}}\Big] \end{aligned}$$

Let us now consider the moments of the process (Y_t). We have:

$$\mathbb{E}[Y_t^p]=\Big[\int g'(y_t)y_t^p\mathrm{d}y_t\Big]\frac{De}{e'De}=\Big(\int g_0(y_t)y_t^p\mathrm{d}y_t,...,\int g_{2J}(y_t)y_t^p\mathrm{d}y_t\Big)\frac{De}{e'De},$$

or for the joint moments:

$$\mathbb{E}[Y_t^p Y_{t+h}^p]=\Big[\int g(y_t)y_t^p\mathrm{d}y_t\Big]'\frac{D\Pi^{h-1}}{e'De}\Big[\int g(y_{t+h})y_{t+h}^p\mathrm{d}y_{t+h}\Big]. \tag{6.8}$$

They are immediately deduced from the corresponding moments in each regime, whose expressions are derived in Appendix 6. The existence of these moments depend on the shape parameter α of the benchmark distribution ϕ of X_t. The moments of order p exists if $\alpha > \frac{p}{2}$. What matters for the existence of moments is the behavior at zero of the conditional distribution $l(X_t|S_t)$, which is gamma with shape parameter $S_t+\alpha$. In particular, under regime $S_t=0$, the density of $l(X_t|S_t)$ has the heaviest tail at zero.

iii) **Model M3.**

[22] All discrete time discretization of univariate diffusion processes are time reversible [see McCausland (2007)].

As a benchmark, we also estimate the following model M3:

$$y_t = \frac{1}{\sqrt{x_t}}\epsilon_t, \tag{6.9}$$

where (X_t) follows an ARG process (see Appendix 5 for details), characterized by its stationary distribution $\gamma(\alpha, \frac{c}{1-\rho})$, and its autocorrelation coefficient ρ. As in model M2, we let (ϵ_t) to be non asymmetric, with $I = 1$ in the expansion (6.5).

While Creal (2016) has shown that such ARG based models can be estimated using (approximate) maximum likelihood, for the sake of comparison and for computational tractability (see Section 4.1 for a discussion), we employ the maximum composite likelihood. It is shown in Appendix 5 that the joint p.d.f. of (Y_t, Y_{t+h}) in the ARG model can be expressed as an infinite mixture of gamma product densities [see equation (eq. a.5)], hence leading to similar quasi-closed forms (subject to truncation) of the joint p.d.f. of (Y_t, Y_{t+h}).

iv) **Model M4**

The three previous models assume a time reversible dynamics for process (X_t). Let us now consider the model M4 that allows for time irreversibility. This model is a generalisation of M2, with non symmetric matrix B. We use the parametrisation of B as $B = B_1 + B_2$, where B_1 is symmetric, and B_2 antisymmetric (see Section 2.4). The time reversible model M1 corresponds to the special case where $B_2 = 0$. As shown in equation (2.20), the constraint $(D - D')e = 0$ for equal margins implies a set of linear constraints on B_2, once B_1 is given. Thus to analyse the potential improvement of allowing for partially asymmetric B, we estimate the model with $J = 4$. Then we use the orthogonal condition $D_2 e = 0$ to express $2J$ out of $\frac{J(J+1)}{2}$ different unknown entries of B_2 as a function of the entries of B_1, and of the additional $\frac{J(J+1)}{2} - 2J$ free parameters.

Let us consider model M4 with integer $J = 4$. We have $\frac{J(J+1)}{2} - 2J = 2$ degrees of freedoms for the matrix B_2. Let us denote B_1 and B_2 by:

$$B_1 = \begin{bmatrix} 1 & b_1 & b_2 & b_3 & b_{10} \\ b1 & b_4 & b_5 & b_6 & b_{11} \\ b_2 & b_5 & b_7 & b_8 & b_{12} \\ b_3 & b_6 & b_8 & b_9 & b_{13} \\ b_{10} & b_{11} & b_{12} & b_{13} & b_{14} \end{bmatrix}, \qquad B_2 = \begin{bmatrix} 0 & a_1 & a_2 & a_3 & a_4 \\ -a_1 & 0 & a_5 & a_6 & a_7 \\ -a_2 & -a_5 & 0 & a_8 & a_9 \\ -a_3 & -a_6 & -a_8 & 0 & a_{10} \\ -a_4 & -a_7 & -a_9 & -a_{10} & 0 \end{bmatrix}$$

Therefore, we solve the condition $D_2 e = 0$ in $a_3, ..., a_{10}$, where a_1, a_2 are the free parameters. This is a linear system with 8 unknowns and $2J = 8$ equations. Its solution is quite complicated and

omitted, but can be easily obtained using a computer program. Therefore, the set of parameters of the new model M4, with $J = 4$ and $I = 2$, say, is:

$$\theta = (c, \alpha, (b_j)_{j=1,...,14}, a_1, a_2, \beta_1, \beta_2),$$

and the other coefficients $a_3, ..., a_{10}$ are known functions of the components of θ. In other words, we have successfully solved the equal margin constraint $(D - D')e = 0$, to avoid constrained optimisation of the composite likelihood function.

6.2.2 Analysis of the volatility of Apple stock return

i) **The estimates for models M1-M4**

All models are estimated by maximum composite likelihood on daily return data of the Apple stock (AAPL), traded at the New York Stock Exchange (NYSE). The data are downloaded from Yahoo Finance, with an observation window spanning from 2000/10/2 up to 2016/9/29. Y_t is the (non annualized) daily adjusted return, that is the return adjusted for the dividend payment. As in the literature, we express, without loss of generality, the returns in percentage. Figure 1 provides the evolution of the daily return.

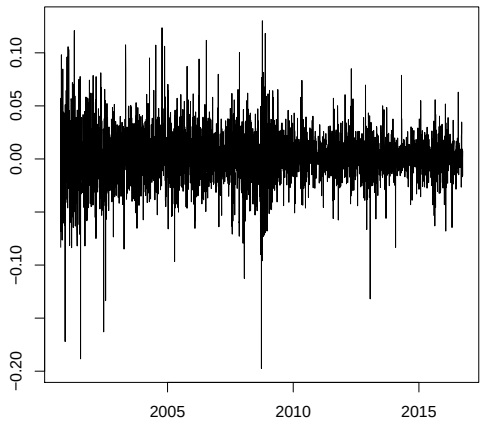

Figure 1: Daily return of the Apple stock between 2000/10/2 and 2016/9/29.

Let us now plot in Figure 2 the histogram of the marginal distribution of Y_t.

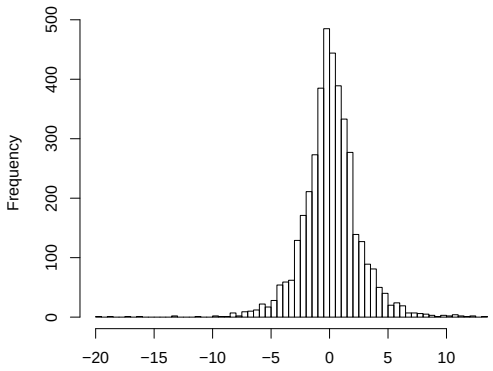

Figure 2: Histogram of Y_t.

The distribution is not symmetric with respect to the origin and has different left and right tails. This motivates the introduction of a flexible conditional distribution for the error term [see equation (6.5)].

In Figure 3 we provide the histogram of the historical distribution of the Y_t^2, as a proxy of the volatility $1/X_t$, since $\mathbb{E}[Y_t^2|X_t] = 1/X_t$ in Model M1 and M3, where error (ϵ_t) has unitary variance.

Figure 3: Histogram of Y_t^2.

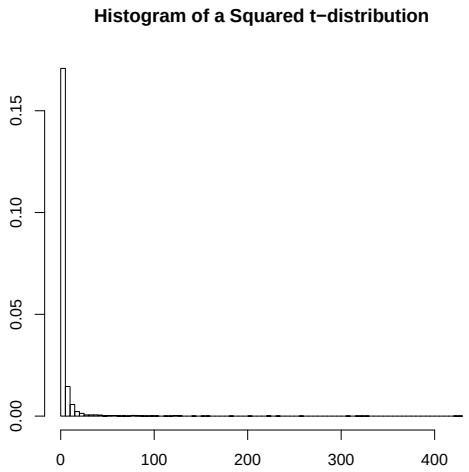

Figure 4: Histogram of a simulated sample of size 4000, following the standard symmetric $t-$distribution with 2.5 degrees of freedom.

Figure 3 shows that the distribution of Y_t^2 is heavy tailed. This feature is well replicated, in Figure 4, by a simulated sample from a $t-$distribution. This justifies our specification (6.4), under which the density of Y_t is a linear combination of $t-$ distribution densities.

Let us now report some summary statistics of Y_t, in particular its four first historical moments.

$$\begin{array}{ll} \frac{1}{T}\sum_{t=1}^{T} y_t = 0.10, & \frac{1}{T}\sum_{t=1}^{T} y_t^2 = 6.02 \\ \frac{1}{T}\sum_{t=1}^{T} y_t^3 = -1.40, & \frac{1}{T}\sum_{t=1}^{T} y_t^4 = 304 \end{array} \tag{6.10}$$

We deduce the historical kurtosis of Y_t, which is equal to 8.43 and significantly larger than 3, as well as the empirical skewness, which is equal to -0.22, indicating a heavier left tail. Note that the sign of the skewness is different from the sign of the empirical mean. This suggests that we need at least the two first terms in the expansion of the density (6.5).

Besides the tractability of the composite likelihood function, as well as the heavy tail property of the return, another motivation of our inverse-gamma type SV model (6.4) is summarized by the historical autocorrelation function of $1/Y_t^2$, as well as that of Y_t^2.

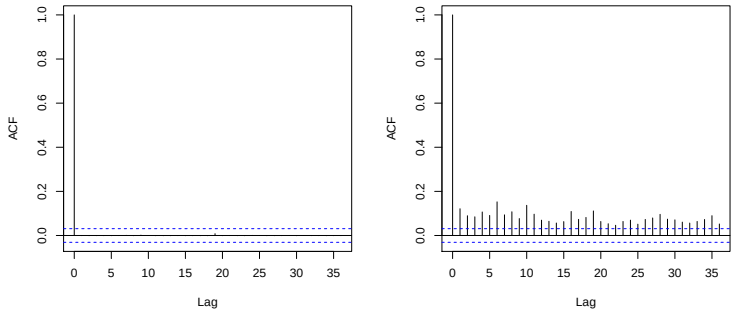

Figure 5: Historical autocorrelation function of $1/Y_t^2$ (left panel) and of Y_t^2, on the right panel

As expected, Y_t^2 has a significant autocorrelation. Nevertheless, the left panel shows that $1/Y_t^2$ has virtually no autocorrelation. This suggests that the serial dependence of the process (Y_t) is rather non-linear, and in particular, standard, non flexible SV models such as model M3 are unlikely to capture this pattern.

For the numerical optimization of the composite likelihood functions, we start with the simplest, nested model M1 with $J = 2$. Once an estimate has been found, we use it as a starting point for the more general model. In terms of computational time, one evaluation of the composite likelihood function takes less than 0.5 second on a PC with 2 GB RAM. In order to compare different models, we introduce the concept of composite Akaike Information Criterion (AIC_{CL}), that is the analogue of the standard likelihood-based AIC [see Varin and Vidoni (2005) for details]:

$$AIC_{CL} = -2\ell_{CL}(\hat{\theta}) + 2\dim(\theta),$$

where $\ell_{CL}(\hat{\theta})$ denotes the optimum of the composite likelihood and $\dim(\theta)$ the dimension of the parameter space. This information criterion favours models which fit well the set of pairwise densities $f(y_t, y_{t+h})$, for lags h ranging from 1 up to $m = 10$, and penalizes the number of parameters, in the same way as the standard AIC.

We report in Table 1 the parameter estimates for the different models.

Model	M1 $J=2$	M1 $J=3$	M2 $J=3$	M2 $J=4$	M3 $--$	M4 $J=4$	M4 $J=4$
Error Density	Gaussian $I=0$	Gaussian $I=0$	skewed $I=1$	skewed $I=1$	skewed $I=1$	skewed $I=1$	skewed $I=2$
Symmetry of (X_t)	yes	yes	yes	yes	yes	no	no
c	0.102(*)	0.0803(*)	0.0741 (*)	0.0751(*)	0.201(*)	0.0503(*)	0.0549(*)
α	2.51(*)	2.84(*)	3.09(*)	2.92(*)	1.73(*)	3.14(*)	2.80 (*)
ρ	$--$	$--$	$--$	$--$	0.085(*)	$--$	$--$
$b_{1,01}$	-1.57(*)	-3.63(*)	-3.47(*)	-3.72(*)	$--$	-2.97(*)	-1.59(*)
$b_{1,02}$	6.30(*)	-3.02(*)	5.28(*)	-2.77(*)	$--$	1.69(*)	1.67(*)
$b_{1,03}$	$--$	-0.979(*)	-0.96(*)	-0.936(*)	$--$	-0.2(*)	-0.47(*)
$b_{1,04}$	$--$	$--$	$--$	-0.260(*)	$--$	0.501(*)	0.102(*)
$b_{1,11}$	-0.892(*)	3.79 (*)	4.56(*)	3.72(*)	$--$	-1.73(*)	-1.08(*)
$b_{1,12}$	-0.339(*)	-0.376(*)	-3.90(*)	48.9(*)	$--$	-19.8(*)	2.08(*)
$b_{1,13}$	$--$	10.02(*)	9.92(*)	10.2(*)	$--$	-0.8(*)	-1.15(*)
$b_{1,14}$	$--$	$--$	$--$	-0.899(*)	$--$	0	0.0610(*)
$b_{1,22}$	7.74(*)	6.62(*)	12.9(*)	7.11(*)	$--$	15.2(*)	4.17(*)
$b_{1,23}$	$--$	-0.940 (*)	-0.97 (*)	-0.87(*)	$--$	-1.5(*)	0.502(*)
$b_{1,24}$	$--$	$--$	$--$	-0.870(*)	$--$	0	0
$b_{1,33}$	$--$	-0.961(*)	-0.955(*)	-0.991(*)	$--$	-1.56(*)	-1.55(*)
$b_{1,34}$	$--$	$--$	$--$	-0.269(*)	$--$	-0.35(*)	0
$b_{1,44}$	$--$	$--$	$--$	0.401(*)	$--$	-0.206(*)	-0.0917(*)
$b_{2,10}$	$--$	$--$	$--$	$--$	$--$	-0.055(*)	0.005(*)
$b_{2,20}$	$--$	$--$	$--$	$--$	$--$	0.07(*)	0.005(*)
$b_{2,30}$	$--$	$--$	$--$	$--$	$--$	-18.9(*)	-2.13(*)
$b_{2,40}$	$--$	$--$	$--$	$--$	$--$	11.3(*)	-2.56(*)
$b_{2,21}$	$--$	$--$	$--$	$--$	$--$	10.2(*)	-11.7(*)
$b_{2,31}$	$--$	$--$	$--$	$--$	$--$	72(*)	-18.4(*)
$b_{2,41}$	$--$	$--$	$--$	$--$	$--$	108(*)	12.5(*)
$b_{2,32}$	$--$	$--$	$--$	$--$	$--$	141.5(*)	143(*)
$b_{2,42}$	$--$	$--$	$--$	$--$	$--$	160(*)	825(*)
$b_{2,43}$	$--$	$--$	$--$	$--$	$--$	59.3(*)	4101(*)
β_1	$--$	$--$	0.017 (*)	0.0125(*)	0.0265(*)	0.0128(*)	0.0273(*)
β_2	$--$	$--$	$--$	$--$	$--$	$--$	0.0295(*)
ℓ_{CL}	-74356	-74264	-74105	-73861	-73660	-73480	-73407
AIC_{CL}	148726	148548	148232	147756	147328	146998	146854

Table 1: Parameter estimates. The first two columns report the estimates of the models with Gaussian innovations, whereas the third and fourth columns report the models with skewed innovations introduced in equation (6.6). The fifth column reports the model M3 with ARG based state process. The last two columns report the model M4 with time irreversible state process. The symbol $--$ indicates that a parameter is set to zero in a model. $*$ indicates that the parameter is significant at the 5 % level. In all models, matrix B is specified in terms of the symmetric matrix B_1 and the antisymmetric matrix B_2.

From the previous table, we see that increasing J, or adding a parameter capturing the skewness both lead to a significant improvement of the fit, in terms of the composite likelihood, and the composite AIC. Moreover, the model with antisymmetric B has a significant better fit than comparable models with symmetric B. This result shows the advantage of our model, by

allowing the coefficients of D to be negative. Although the switching regime can no longer be defined, the quality of approximation of the squared polynomial largely dominates the case with only nonnegative entries of D.

Let us now focus on the column for the ARG based model M3. The estimate of the autocorrelation coefficient of (X_t) is $\rho \approx 0.08$, which is rather weak. This can be explained by the left panel of Figure 5. Indeed, the serial dependence of the ARG model is characterized by one parameter ρ, and the lack of autocorrelation of $1/Y_t^2$ suggests a rather small $\hat{\rho}$. As a consequence, model M3 cannot well capture the substantial autocorrelation of Y_t^2.

ii)**Analysis of Model M4.**

Let us now analyse the time irreversible Model M4 with $J = 4$ and $I = 2$. Below we report the estimated values of symmetric part D_1 and antisymmetric part D_2 of matrix D. For expository and comparison purpose, we round off all the entries to three decimal places. We get:

$$D_1 = \begin{bmatrix} 1 & -0.492 & 0.19 & -0.053 & 0.012 & -0.002 & 0.002 & 0.002 & 0.001 \\ -0.492 & 0.07 & 0.012 & -0.014 & 0.004 & 0.007 & 0.009 & 0.001 & -0.001 \\ 0.19 & 0.012 & 0.002 & -0.003 & 0.01 & 0.012 & -0.008 & -0.032 & -0.015 \\ -0.053 & -0.014 & -0.003 & -0.021 & -0.012 & -0.022 & -0.088 & -0.078 & 0 \\ 0.012 & 0.004 & 0.01 & -0.012 & 0.044 & 0.078 & -0.067 & 0.062 & 0.239 \\ -0.002 & 0.007 & 0.012 & -0.022 & 0.078 & 0.216 & -0.047 & 0.228 & 0.773 \\ 0.002 & 0.009 & -0.008 & -0.088 & -0.067 & -0.047 & -0.699 & -0.615 & 0.823 \\ 0.002 & 0.001 & -0.032 & -0.078 & 0.062 & 0.228 & -0.615 & -1.479 & 0 \\ 0.001 & -0.001 & -0.015 & 0 & 0.239 & 0.773 & 0.823 & 0 & 0 \end{bmatrix},$$

$$D_2 = \begin{bmatrix} 0 & 0.002 & 0 & -0.036 & 0.005 & 0.001 & -0.003 & 0 & 0 \\ -0.002 & 0 & -0.116 & 0.01 & 0.024 & -0.015 & 0.003 & 0 & 0 \\ 0 & 0.116 & 0 & 0.11 & 0.097 & -0.074 & 0.037 & -0.005 & 0.001 \\ 0.036 & -0.01 & -0.11 & 0 & 0.106 & -0.122 & 0.063 & -0.009 & 0.001 \\ -0.005 & -0.024 & -0.097 & -0.106 & 0 & -0.046 & 0.032 & -0.005 & 0 \\ -0.001 & 0.015 & 0.074 & 0.122 & 0.046 & 0 & 0.017 & -0.002 & 0 \\ 0.003 & -0.003 & -0.037 & -0.063 & -0.032 & -0.017 & 0 & -0.001 & 0 \\ 0 & 0 & 0.005 & 0.009 & 0.005 & 0.002 & 0.001 & 0 & 0 \\ 0 & 0 & -0.001 & -0.001 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The estimated antisymmetric matrix D_2 is significantly non zero, which confirms the time irreversibility of the Apple return data. Nevertheless, the largest entry of $|D_1|$ is $|d_{1,88}| = 1.479$,

whereas the largest (in absolute value) entry of $|D_2|$ is $|d_{2,53}| = 0.122$, which is significantly smaller than 1.479.

iii) **Estimation of marginal moments.**

Let us now check how these estimated models are able to reconstitute the true marginal moments. For this purpose, we report the theoretical marginal moments predicted by these models, and compare them to the corresponding historical moments.

Model	M1 $J=2$	M1 $J=3$	M2 $J=3$	M3 $--$	M4 $J=4$	M4 $J=4$	Real data
Residual	Gaussian	Gaussian	skewed $I=1$	skewed $I=1$	skewed $I=1$	skewed $I=2$	$--$ $--$
Symmetry of B	yes	yes	yes	$--$	no	no	$--$
$\mathbb{E}[y_t]$	0	0	0.21	0.108	0.0934	0.119	0.106
$\mathbb{E}[y_t^2]$	5.02	5.34	5.64	6.08	5.24	6.12	6.02
$\mathbb{E}[y_t^3]$	0	0	- 0.95	-0.96	-0.11	-1.21	-1.40
$\mathbb{E}[y_t^4]$	205	216	298	303	318	314	304
$\text{corr}[y_t^2, y_{t+1}^2]$	0.043	0.045	0.054	0.0001	0.075	0.091	0.121

Table 2: Comparison of historical marginal moments with their theoretical counterparts predicted by various models.

Increasing I or J, or introducing non symmetric matrix B leads to moments that are closer to their historical values. On the other hand, the benchmark model M3 can satisfactorily fit the marginal moments (this is expected, since under model M3 with standard Gaussian error, the marginal distribution of Y_t is Student, and by Figures 2 and 3, we know that the Student distribution is a good proxy of the marginal distribution.), but fails to predict the autocorrelation coefficient $\text{corr}[y_t^2, y_{t+1}^2]$, due to the small estimate of the autocorrelation coefficient ρ of the state process: $\hat{\rho} \approx 0.08$.

iii)**Estimation of the marginal density**

Let us now compare the empirical marginal density with the density predicted by the different models. The empirical marginal density is obtained from a kernel-based non-parametric density estimator [see e.g. Rosenblatt (1975)]. More precisely, we take a positive kernel function K, which is defined on $\mathbb{R}$, and has unit mass, then the marginal distribution of Y_t is estimated by:

$$\hat{f}(y_1) = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{1}{h_T} K\Big(\frac{y_t - y_1}{h_T}\Big), \qquad \forall y_1, \tag{6.11}$$

where the bandwidth h_T depends on T. Under mild conditions [see e.g. Darolles et al. (2004)], in particular if h_T goes to zero at an appropriate rate in T, such an estimator is asymptotically consistent. In the application we use the Gaussian kernel, and set the number of equal-lengthed intervals to be 100. The following three figures compare the model implied marginal densities with the historical kernel density estimators for three models: M3, M4 ($J = 4, I = 1$), and M4 ($J = 4, I = 2$).

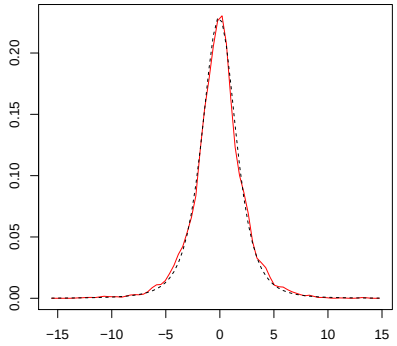

Figure 6: Comparison of the model M3 implied marginal density of $f(y_t)$ with the kernel density estimator. Full line: kernel density estimator; dashed line: model implied density.

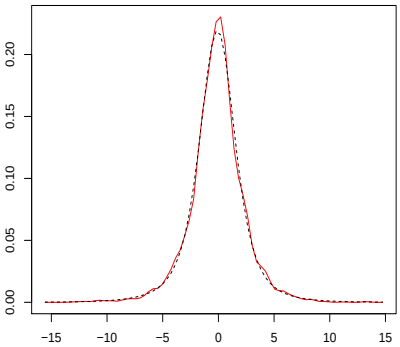

Figure 7: Comparison of the model M4 ($J = 4, I = 1$) implied marginal density of $f(y_t)$ with the kernel density estimator. Full line: kernel density estimator; dashed line: model implied density.

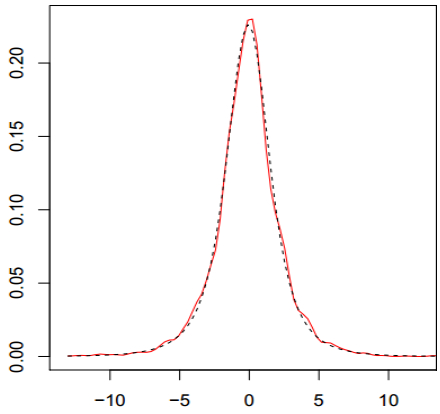

Figure 8: Comparison of the model M4 ($J = 4, I = 2$) implied marginal density of $f(y_t)$ with the kernel density estimator. Full line: kernel density estimator; dashed line: model implied density.

Model M4 provides a better fit of the marginal density than the ARG based model M3. Within Models M4, increasing I also leads to a slight improvement of the fit. Similarly, we use the kernel method to estimate the joint density of (Y_t, Y_{t+1}):

$$\hat{f}(y_1, y_2) = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{1}{h_T^2} K\Big(\frac{y_t - y_1}{h_T}\Big) K\Big(\frac{y_{t+1} - y_2}{h_T}\Big), \qquad \forall y_1, y_2 \in \text{range } Y. \tag{6.12}$$

The following figure plots the iso-density curves of the obtained empirical kernel density estimate, and compare it with the model implied joint density. It confirms that the model provides a good fit of the joint density function.

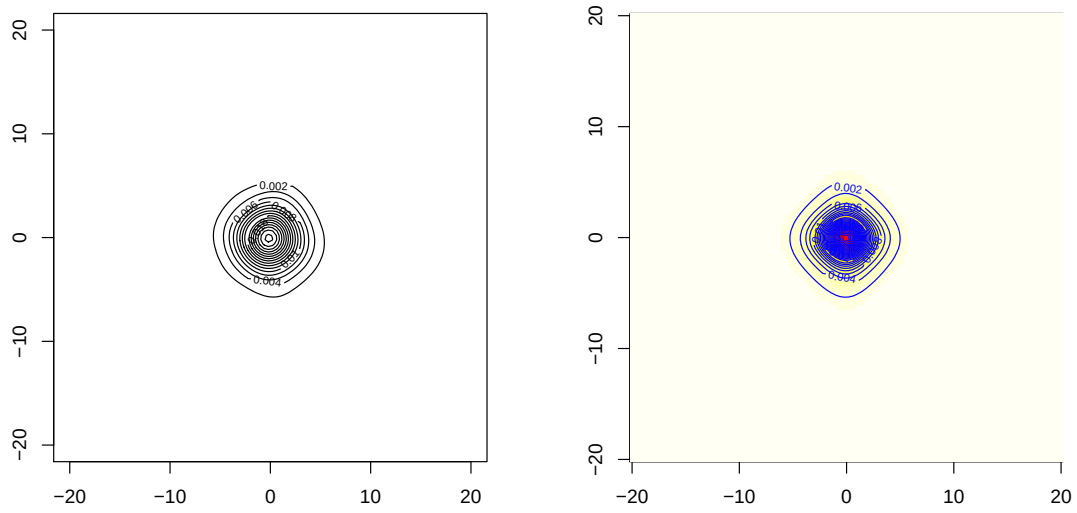

Figure 9: Left panel: Iso-density curves of the kernel-based estimate of $f(y_t, y_{t+1})$. Right panel: Iso-density curves of the model implied joint density.

iv) **Filtering** Let us now apply the recursive formula described in Section 4.1 to compute:

- the filtered mean of the past squared volatility, that is $\mathbb{E}[\frac{1}{X_t}|\underline{y_t}]$.

  By Corollary 4, the conditional p.d.f. of X_t given $\underline{y_t}$ is:

$$l(x_t|\underline{y_t}) = \frac{P'(\underline{y_{t-1}})}{P'(\underline{y_{t-1}})g(y_t)}\Big(\frac{e^{-(\frac{y_t^2}{2}+\frac{1}{c})x_t}x_t^{\alpha+\frac{1}{2}-1}}{\sqrt{2\pi}\Gamma(\alpha+0)c^{\alpha+0}},...,\frac{e^{-(\frac{y_t^2}{2}+\frac{1}{c})x_t}x_t^{\alpha+2J+\frac{1}{2}-1}}{\sqrt{2\pi}\Gamma(\alpha+2J)c^{\alpha+2J}}\Big).$$

  Thus we have:

$$\mathbb{E}[\frac{1}{X_t}|\underline{y_t}] = \frac{P'(\underline{y_{t-1}})}{P'(\underline{y_{t-1}})g(y_t)}\Big(\frac{\Gamma(\alpha+0-\frac{1}{2})(\frac{c}{1+\frac{cy_t^2}{2}})^{\alpha+0-\frac{1}{2}}}{\sqrt{2\pi}\Gamma(\alpha+0)c^{\alpha+0}},...,\frac{\Gamma(\alpha+2J-\frac{1}{2})(\frac{c}{1+\frac{cy_t^2}{2}})^{\alpha+2J-\frac{1}{2}}}{\sqrt{2\pi}\Gamma(\alpha+2J)c^{\alpha+2J}}\Big).$$

- the smoothed mean $\mathbb{E}[\frac{1}{X_t}|\underline{y_T}]$.

  By Proposition 10, this mean is equal to:

$$\mathbb{E}[\frac{1}{X_t}|\underline{y_T}] = \frac{P'(\underline{y_{t-1}})\Big[\int \frac{\phi(x_t)}{x_t}\frac{U(x_t)U'(x_t)D}{U'(x_t)De}\frac{\sqrt{x_t}}{\sqrt{2\pi}}e^{-\frac{y_t^2x_t}{2}}\mathrm{d}x_t\Big]\Pi(\underline{y_{t+1}})\cdots\Pi(\underline{y_{T-1}})g(y_T)}{P'(\underline{y_{t-1}})\Pi(\underline{y_t})\Pi(\underline{y_{t+1}})\cdots\Pi(\underline{y_{T-1}})g(y_T)}.$$

- the term structure of predictive mean of the future volatility $\mathbb{E}[\frac{1}{X_{T+h}}|\underline{y_T}]$, where $h \in \mathbb{N}$.

  By the proof of Lemma 2, the conditional distribution $x_{T+h}|\underline{y_T}$ has the density $l(x_{T+h}|\underline{y_T}) =$

$\phi(x_{T+1})P'(\underline{y_T})\Pi^{h-1}U(x_{T+1})$. Thus we have:

$$\mathbb{E}[\frac{1}{X_{T+h}}|\underline{y_T}] = P'(\underline{y_T})\Pi^{h-1}\Big(\frac{1}{c(\alpha+0-1)},...,\frac{1}{c(\alpha+2J-1)}\Big).$$

Figure 10 plots these filtered/smoothed/predicted volatility for model M4, with $J = 2$ and $I = 2$.

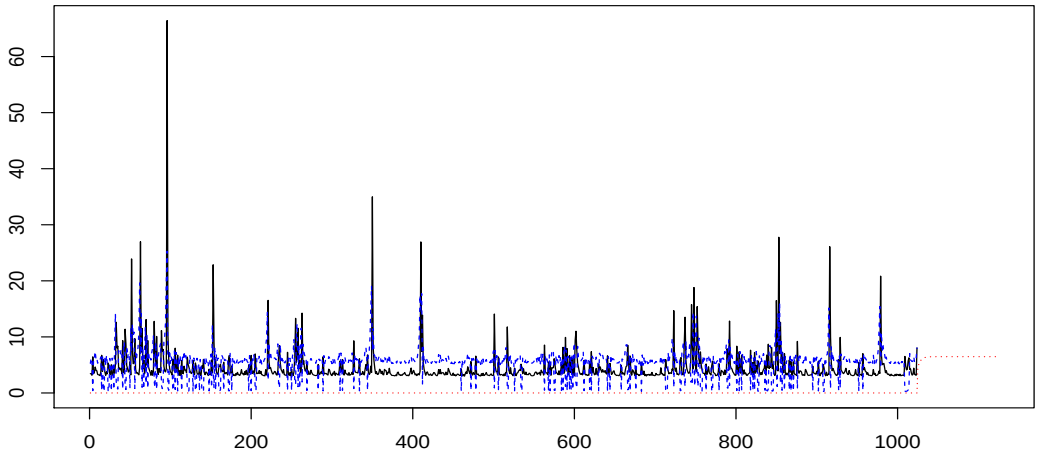

Figure 10: Filtering (dotted line), smoothing (dashed line) of the conditional variance for the latest 1000 dates, along with the term structure of volatility forecast for $h = 1$ to $h = 100$.

7 Conclusion

The aim of the paper was twofold. First, we have introduced a general class of state-space models. This class is flexible enough to capture any Markov dynamics of the state variable, and has an intuitive endogenous switching regime interpretation. Moreover, the model is associated with simple simulation-free methods for filtering, forecasting, smoothing and estimation. Second, we have investigated a new stochastic volatility model that is capable of capturing, in a unified framework, the heavy tail, the volatility feedback, as well as the time irreversibility.

A by-product of our model is the introduction of a flexible specification for univariate Markov processes. This model is of finite dimensional dependence, which leads to simple linear and non-linear conditional moments. Such a model can also be applied to observable time series, such as the (historical and risk-neutral) dynamics of short term interest rate. It has recently been shown by Gouriéroux and Monfort (2015) that FDD models have the potential of becoming a serious competitor of affine term structure models [such as the CIR/ARG model]. However, up

to now appropriate FDD models are rather sparse. The flexibility of this new specification is an essential advantage in order to fit the whole term structure of interest rates. This is left for future research.

Appendices

Appendix 1 Proofs of the propositions

Appendix 1.1 Proof of Proposition 1

From the joint distribution, we derive the marginal distribution of X_t:

$$f_0(x_t) = \phi(x_t) \int \phi(x_{t+1}) \frac{U'(x_t) D U(x_{t+1})}{e' D e} \mathrm{d}x_{t+1} = \phi(x_t) \frac{U'(x_t) D e}{e' D e}.$$

Similarly, the marginal distribution of X_{t+1} is $\tilde{f}_0(x_{t+1}) = \phi(x_{t+1}) \frac{e' D U(x_{t+1})}{e' D e}$. Thus the condition for $\tilde{f}_0 = f_0$ is $U'(x) D e = e' D U(x)$, for all $x \in \mathcal{X}$. This is equivalent to $(D - D')e = 0$, so long as the support $\mathcal{X}$ contains an infinity of points.

Appendix 1.2 Proof of Proposition 2

Let us proceed by induction. Assume that identity (2.10) is valid for a given $h \geq 1$, then we have:

$$\begin{aligned} f_{h+1}(x_{t+h+1} \mid x_t) &= \int f_1(x_{t+h+1} \mid x_{t+1}) f_1(x_{t+1} \mid x_t) \mathrm{d}x_{t+1} \\ &= \int \phi(x_{t+1}) \frac{U'(x_t) D}{U'(x_t) D e} U(x_{t+1}) \frac{U'(x_{t+1}) D \Pi^{h-1}}{U'(x_{t+1}) D e} U(x_{t+h+1}) \mathrm{d}x_{t+1} \\ &= \phi(x_{t+h}) \frac{U'(x_t) D \Pi^h}{U'(x_t) D e} U(x_{t+h+1}), \end{aligned}$$

that is identity (2.10) for $h + 1$. Thus we have proven Proposition 2.

Appendix 1.3 Proof of Lemma 1

Let us first remark that the function $x \mapsto U'(x) D e$ is continuous, lower bounded by zero, and goes to infinity when $|x|$ goes to infinity[23]. Thus in order to show that it is lower bounded by a positive constant, it suffices to show that $U'(x) D e$ cannot take value zero. By the expression

[23] Indeed, $U'(x) D e$ is a polynomial. Thus the nonnegativity implies that its dominant coefficient is positive; thus this polynomial goes to infinity when x goes to infinity.

of the marginal distribution, $U'(x)De$ is null if and only if $\sum_{i,j=0}^{J} b_{i,j}x^i y^j = 0$ almost surely in y. This is equivalent to all coefficients before the terms $1, y, y^2, ..., y^J$ being null, that is: $B(1, x, x^2, ..., x^J)' = 0$ for a certain x.

Appendix 1.4 Proof of Proposition 5

If all the entries of Π are positive, then by Perron-Frobenius theorem, Π has a unique, simple eigenvalue with a right eigenvector of only positive entries, which is e. We deduce the ergodicity of (S_t), as well as the convergence of Π^h towards the projector matrix $\frac{ee'D}{e'De}$ [see equation (2.14)]. Let us now show that, if the entries of Π are nonnegative, then the chain is still aperiodic and irreducible, once the potential isolated states, that are states that are almost surely never reached, are discarded. It suffices to show that, if there exist i, j belonging to $[|0, 2J|]$ such that the transition probability $\mathbb{P}[S_{t+1} = j|S_t = i]$ is zero, then j is necessarily an isolated state. That is, the probability of reaching j from any state k is zero. To prove this, let us remark that:

$$\mathbb{P}[S_{t+1} = j|S_t = i] = \int \phi(x) \frac{x^i}{\mu_i} \Big(\frac{U'(x)D}{U'(x)De} \Big)_j \mathrm{d}x,$$

where $\left(\frac{U'(x)D}{U'(x)De} \right)_j$ denotes the $j-$th component of the vector $\frac{U'(x)D}{U'(x)De}$, which is nonnegative since entries of D are nonnegative under the Assumptions of the Proposition. Thus $\mathbb{P}[S_{t+1} = j|S_t = i] = 0$ implies that $\left(\frac{U'(x)D}{U'(x)De} \right)_j$ is zero almost everywhere. This latter in turn implies that $\mathbb{P}[S_{t+1} = j|S_t = k] = 0$ for any state k. As a consequence, state j is never reached. Thus the space $[|0, 2J|]$ can be partitioned into the union of a regular class, in which the probability of moving from one state to another is always positive, and potentially several isolated states that are never reached. Thus the Markov chain (S_t) is ergodic. Hence the unitary eigenvalue is simple, and the largest in modulus of Π.

Let us now consider the case where D is symmetric and denote by H the Hilbert space of functions g such that $\mathbb{E}[g^2(X_t)]$ is finite. Upon this space, we can define the one-step-ahead conditional expectation operator $\mathcal{T}$ by, for all $g \in H$:

$$\mathcal{T}g(x) = \mathbb{E}[g(X_{t+1})|X_t = x] = \frac{U'(x)D}{U'(x)De} \int \phi(s)g(s)U(s)\mathrm{d}s.$$

Since D is symmetric, process (X_t) is time reversible: $f(x_t, x_{t+1}) = f(x_{t+1}, x_t)$, and the operator $\mathcal{T}$ is self-adjoint, that is, for all functions g_1, g_2 we have:

$$< g_1, \mathcal{T}g_2 >= \mathbb{E}[g_1(X_t)g_2(X_{t+1})] = \mathbb{E}[g_2(X_t)g_1(X_{t+1})] =< g_2, \mathcal{T}g_1 > .$$

Thus, under mild conditions[24], the operator $\mathcal{T}$ is diagonalizable and we have the following spectral decomposition [see Lancaster (1958); Hansen et al. (1998); Darolles et al. (2004)]:

$$\mathcal{T}g = \sum_{j=0}^{2J} \rho_j < \psi_j, g > \psi_j,$$

or in terms of conditional density:

$$f(x_{t+1}|x_t) = f_0(x_t) \sum_{j=0}^{2J} \rho_j \psi_j(x_t)\psi_j(x_{t+1}),$$

where (ψ_j) is an orthonormal family of real eigenfunctions, and (ρ_j) is a corresponding sequence of eigenvalues. For instance, the first one ρ_0 is equal to 1 and is associated with the constant function $g = 1$. The spectral decomposition usually involves an infinity of eigenvalues ρ_j, however since (X_t) has finite dimensional dependence, at most the first $2J+1$ terms are non zero. Finally, since $\mathcal{T}$ is self-adjoint, the eigenvalues are all real. Moreover, by the definition of the operator $\mathcal{T}$, the eigenvalues are no larger than 1 in modulus.

In the rest of the proof, let us show that the unitary eigenvalue is simple and -1 cannot be an eigenvalue. If any of the other ρ_i, say ρ_1, is equal to 1 or -1, then by definition we have $\text{corr}[\phi_1(X_t), \phi_1(X_{t+1})] = 1$ or -1. This means that $\phi_1(X_t) = \phi_1(X_{t+1})$, or $\phi_1(X_t) = -\phi_1(X_{t+1})$ almost surely. Let us now study the form of the eigenfunctions ϕ_j and show a contradiction. We have the following property:

Lemma 3. The operator $g \mapsto \mathcal{T}g$ defined by $:\mathcal{T}g(x) = \mathbb{E}[g(X_{t+1})|X_t = x]$ and the matrix Π have the same spectrum.

Proof. If $\Pi V = \lambda V$ for a non zero vector V then the function $g(x) := \frac{U'(x)DV}{U'(x)De}$ is such that

$$\mathcal{T}g(x) = \frac{U'(x)D}{U'(x)De} \int \phi(s)U(s) \frac{U'(s)DV}{U'(s)De} \mathrm{d}s = \lambda g(x).$$

In other words, g is an eigenfunction of process (X_t), associated with eigenvalue (of process (X_t)) λ_j.

Conversely, if g is an eigenfunction with eigenvalue λ, we have:

$$\mathcal{T}g(x) = \lambda g(x) = \frac{U'(x)D}{U'(x)De} \int \phi(s)g(s)U(s)\mathrm{d}s,$$

Multiplying both sides by $\phi(x)U(x)$ and integrating with respect to x, we get: $\lambda \int \phi(s)g(s)U(s)\mathrm{d}s =$

[24]This condition is that the joint density satisfies $\iint \frac{f^2(x_t,x_{t+1})}{f_0(x_t)f_0(x_{t+1})}\mathrm{d}x\mathrm{d}y < \infty$. It can be easily checked that the process (X_t) satisfies this condition.

$\Pi \int \phi(s)g(s)U(s)\mathrm{d}s$. Thus each eigenvalue of operator $\mathcal{T}$ is an eigenvalue of Π. □

By Lemma 3, the eigenfunctions ϕ_j of the operator $\mathcal{T}$ are necessarily of the form $\phi_j(x) = \frac{U'(x)DV_j}{U'(x)De}$, where V_j is a right eigenvector of Π. Thus $\phi_1(X_t) = \phi_1(X_{t+1})$, or $\phi_1(X_t) = -\phi_1(X_{t+1})$ is equivalent to $\frac{U'(x_t)DV_j}{U'(x_t)De} = \frac{U'(x_{t+1})DV_j}{U'(x_{t+1})De}$ or $\frac{U'(x_t)DV_j}{U'(x_t)De} = -\frac{U'(x_{t+1})DV_j}{U'(x_{t+1})De}$. This is a non degenerate curve on the plan (X_t, X_{t+1}), which implies a degenerate joint distribution of (X_t, X_{t+1}). This is a contradiction. Therefore, all other eigenvalues ρ_j, $j = 1, 2, ..., 2J$ are smaller than 1 in modulus. As a consequence, the symmetry of D implies the conditions of Proposition 3, and hence the ergodicity of the state process (X_t).

Appendix 1.5 Proof of Proposition 8

The predictive density $l(x_t|\underline{y_{t-1}})$ is linked to the posterior density $l(x_{t-1}|\underline{y_{t-1}})$ via:

$$\begin{aligned} l(x_t|\underline{y_{t-1}}) &= \int l(x_t|\underline{y_{t-1}}, x_{t-1}) l(x_{t-1}|\underline{y_{t-1}})\mathrm{d}x_{t-1} \\ &= \int l(x_t|x_{t-1}) l(x_{t-1}|\underline{y_{t-1}})\mathrm{d}x_{t-1} \\ &= \phi(x_t) \int \frac{U'(x_{t-1})DU(x_t)}{U'(x_{t-1})De} l(x_{t-1}|\underline{y_{t-1}})\mathrm{d}x_{t-1} \\ &= \phi(x_t) P'(\underline{y_{t-1}}) U(x_t), \end{aligned}$$

where $P'(\underline{y_{t-1}}) := \int \frac{U'(x_{t-1})D}{U'(x_{t-1})De} l(x_{t-1}|\underline{y_{t-1}})\mathrm{d}x_{t-1}$. It remains to derive the recursive formula for this latter. For the initial condition we can remark that $l(x_0|\underline{y_0}) = f_0(x_0) = \phi(x_0)\frac{U'(x_0)De}{e'De}$. Thus

$$P'(\underline{y_0}) = \int \frac{U'(x_0)D}{U'(x_0)De} l(x_0|\underline{y_0})\mathrm{d}x_0 = \int \phi(x_0)\frac{U'(x_0)D}{e'De}\mathrm{d}x_0 = \frac{e'D}{e'De}.$$

Let us now derive the updating formula. First, we remark that the posterior density is linked to the predictive density via:

$$l(x_t|\underline{y_t}) = l(x_t|y_t, \underline{y_{t-1}}) = \frac{l(x_t, y_t|\underline{y_{t-1}})}{l(y_t|\underline{y_{t-1}})} = \frac{l(x_t|\underline{y_{t-1}}) l(y_t|x_t, \underline{y_{t-1}})}{l(y_t|\underline{y_{t-1}})}. \qquad (*)$$

Thus we have:

$$\begin{aligned}
P'(\underline{y_t}) &= \int \frac{U'(x_t)D}{U(x_t)'De} l(x_t|\underline{y_t})\mathrm{d}x_t \\
&= \frac{\int \frac{U'(x_t)D}{U'(x_t)De} l(x_t|\underline{y_{t-1}}) l(y_t|x_t, \underline{y_{t-1}})\mathrm{d}x_t}{\int l(x_t|\underline{y_{t-1}}) l(y_t|x_t)\mathrm{d}x_t} \\
&= \frac{P'(\underline{y_{t-1}})\Big[\int \phi(x_t)\frac{U(x_t)U'(x_t)D}{U'(x_t)De} l(y_t|x_t, \underline{y_{t-1}})\mathrm{d}x_t\Big]}{P'(\underline{y_{t-1}})\Big[\int \phi(x_t)U(x_t) l(y_t|x_t, \underline{y_{t-1}})\mathrm{d}x_t\Big]},
\end{aligned}$$

which is formula (4.3)-(4.5).

Appendix 1.6 Proof of Corollary 9

This corollary is a direct consequence of formula (*).

Appendix 1.7 Proof of Proposition 10

Let us first compute the joint distribution:

$$\begin{aligned}
& l(y_T, x_T, y_{T-1}, x_{T-1}, ..., y_{t+1}, x_{t+1}, x_t|\underline{y_t}) \\
&= l(x_t|\underline{y_t}) l(x_{t+1}|x_t) l(y_{t+1}|\underline{y_t}, x_{t+1}) \cdots l(y_{T-1}|\underline{y_{T-2}}, x_{T-1}) l(x_T|x_{T-1}) l(y_T|\underline{y_{T-1}}, x_T) \\
&\propto P'(\underline{y_{t-1}})\phi(x_t)U(x_t) l(y_t|\underline{y_{t-1}}, x_t)\phi(x_{t+1})\frac{U'(x_t)DU(x_{t+1})}{U'(x_t)De} l(y_{t+1}|\underline{y_t}, x_{t+1}) \\
&\qquad \phi(x_{t+2})\frac{U'(x_{t+1})DU(x_{t+2})}{U'(x_{t+1})De} l(y_{t+2}|\underline{y_{t+1}}, x_{t+2}) \cdots \phi(x_T)\frac{U'(x_{T-1})DU(x_T)}{U'(x_{T-1})De} l(y_T|\underline{y_{T-1}}, x_T)\phi(x_T).
\end{aligned}$$

Then by integrating out $x_{t+1}, ..., x_T$, we obtain:

$$\begin{aligned}
& l(y_T, y_{T-1}, ..., y_{t+1}, x_t|\underline{y_t}) \\
&= \Big[\phi(x_t)\frac{U(x_t)U'(x_t)D}{U'(x_t)De} l(y_t|\underline{y_{t-1}}, x_t)\Big]\Big[\int \frac{U(x_{t+1})U'(x_{t+1})D}{U'(x_{t+1})De} l(y_{t+1}|\underline{y_t}, x_{t+1})\phi(x_{t+1})\mathrm{d}x_{t+1}\Big] \times \cdots \\
&\quad \Big[\int \frac{U(x_{T-1})U'(x_{T-1})D}{U'(x_{T-1})De} l(y_{T-1}|\underline{y_{T-2}}, x_{T-1})\phi(x_{T-1})\mathrm{d}x_{T-1}\Big]\Big[\int l(y_T|\underline{y_{T-1}}, x_T)\phi(x_T)U(x_T)\mathrm{d}x_T\Big] \\
&\propto \Big[\phi(x_t)\frac{U(x_t)U'(x_t)D}{U'(x_t)De} l(y_t|\underline{y_{t-1}}, x_t)\Big]\Pi(\underline{y_{t+1}}) l(y_{t+1}|\underline{y_t})\Pi(\underline{y_{t+2}}) l(y_{t+2}|\underline{y_{t+1}}) \cdots \Pi(\underline{y_{T-1}}) l(y_T|\underline{y_{T-1}}) g(y_T|\underline{y_{T-1}}).
\end{aligned}$$

Finally, the smoothing density is obtained by taking the ratio between the RHS of the last

equation and its integral with respect to x_t:

$$\begin{aligned}
l(x_t|\underline{y_T}) &= \frac{l(y_T, y_{T-1}, ..., y_{t+1}, x_t|\underline{y_t})}{l(y_T, y_{T-1}, ..., y_{t+1}|\underline{y_t})} \\
&= \frac{P'(\underline{y_{t-1}})\Big[\phi(x_t)\frac{U(x_t)U'(x_t)D}{U'(x_t)De}l(y_t|\underline{y_{t-1}}, x_t)\Big]\Pi(\underline{y_{t+1}})\Pi(\underline{y_{t+2}})\cdots\Pi(\underline{y_{T-1}})g(y_T|\underline{y_{T-1}})}{P'(\underline{y_{t-1}})\Pi(\underline{y_t})l(y_t|\underline{y_{t-1}})\Pi(\underline{y_{t+1}})\cdots\Pi(\underline{y_{T-1}})g(y_T|\underline{y_{T-1}})} \\
&= \frac{1}{P'(\underline{y_{t-1}})g(y_t|\underline{y_{t-1}})}\frac{P'(\underline{y_{t-1}})\Big[\phi(x_t)\frac{U(x_t)U'(x_t)D}{U'(x_t)De}l(y_t|x_t)\Big]\Pi(\underline{y_{t+1}})\Pi(\underline{y_{t+2}})\cdots\Pi(\underline{y_{T-1}})g(y_T|\underline{y_{T-1}})}{P'(\underline{y_{t-1}})\Pi(\underline{y_t})\Pi(\underline{y_{t+1}})\cdots\Pi(\underline{y_{T-1}})g(y_T|\underline{y_{T-1}})}
\end{aligned}$$

Then we can remark that this formula can be rewritten in the recursive form (4.9).

Appendix 1.8 Proof of Proposition 11

$$\begin{aligned}
& l(y_{T+h} \mid \underline{y_T}) \\
&= \int l(x_{T+1}|\underline{y_T})l(x_{T+h}|x_{T+1})l(y_{T+h}|x_{T+h})\mathrm{d}x_{T+1}\mathrm{d}x_{T+h} \\
&= \int P'(\underline{y_T})\phi(x_{T+1})U(x_{T+1})\phi(x_{T+h})\frac{U'(x_T)D\Pi^{h-2}U(x_{T+h})}{U'(x_T)De}l(y_{T+h}|x_{T+h})\mathrm{d}x_{T+1}\mathrm{d}x_{T+h} \\
&= P'(\underline{y_T})\underbrace{\Big[\int \phi(x_{T+1})\frac{U(x_{T+1})U'(x_{T+1})D}{U(x_{T+1})'De}\mathrm{d}x_{T+1}\Big]}_{=\Pi}\Pi^{h-1}\Big[\int l(y_{T+h}|x_{T+h})\phi(x_{T+h})U'(x_{T+h})\mathrm{d}x_{T+h}\Big] \\
&= P'(\underline{y_T})\Pi^{h-1}g(y_{T+h}).
\end{aligned}$$

Appendix 1.9 Proof of Lemma 2

$$\begin{aligned}
f_Y(y_t, y_{t+h}) &= \int l(y_t|x_t)l(y_{t+h}|x_{t+h})l(x_t, x_{t+h})\mathrm{d}x_t\mathrm{d}x_{t+h} \\
&= \int l(y_t|x_t)l(y_t|x_{t+h})\phi(x_t)\phi(x_{t+h})\frac{U'(x_t)D\Pi^{h-1}U(x_{t+h})}{e'De}\mathrm{d}x_t\mathrm{d}x_{t+h} \\
&= \Big[\int U'(x_t)l(y_t|x_t)\phi(x_t)\mathrm{d}x_t\Big]\frac{D\Pi^{h-1}}{e'De}\Big[\int l(y_{t+h}|x_{t+h})\phi(x_{t+h})U(x_{t+h})\mathrm{d}x_{t+h}\Big] \\
&= \frac{g'(y_t)D\Pi^{h-1}g(y_{t+h})}{e'De}.
\end{aligned}$$

Appendix 1.10 Proof of Proposition 12

When J goes to infinity, $M_J = \sum_{i,j=0}^{J} a_{i,j}^2$ converges to $\sum_{i,j=0}^{\infty} a_{i,j}^2 = 1$. Thus we have:

$$
\begin{aligned}
&\iint |\sqrt{f_J(x_t, x_{t+1})} - \sqrt{f(x_t, x_{t+1})}|^2 \mathrm{d}x_t \mathrm{d}x_{t+1} \\
&\leq \iint \Big|\sqrt{f_J(x_t, x_{t+1})} - \sqrt{\frac{f_J(x_t, x_{t+1})}{M}}\Big|^2 \mathrm{d}x_t \mathrm{d}x_{t+1} + \iint \Big|\sqrt{\frac{f_J(x_t, x_{t+1})}{M_J}} - \sqrt{f(x_t, x_{t+1})}\Big|^2 \mathrm{d}x_t \mathrm{d}x_{t+1} \\
&= (1 - \frac{1}{\sqrt{M}}) \iint f_J(x_t, x_{t+1}) \mathrm{d}x_t \mathrm{d}x_{t+1} \\
&\qquad + \iint \Big||\sum_{i,j=0}^{J} a_{i,j} P_i(x_t) P_j(x_{t+1})| - |\sum_{i,j=0}^{\infty} a_{i,j} P_i(x_t) P_j(x_{t+1})|\Big|^2 \phi(x_t)\phi(x_{t+1}) \mathrm{d}x_t \mathrm{d}x_{t+1} \\
&\leq 1 - \frac{1}{\sqrt{M}} + \iint \Big|\sum_{i>J, \text{ or } j>J} a_{i,j} P_i(x_t) P_j(x_{t+1})\Big|^2 \phi(x_t)\phi(x_{t+1}) \mathrm{d}x_t \mathrm{d}x_{t+1} \\
&= 1 - \frac{1}{\sqrt{M}} + \sum_{i>J, \text{ or } j>J} a_{i,j}^2 \longrightarrow 0, \qquad \text{when } J \text{ goes to infinity.}
\end{aligned}
$$

Appendix 2 Comparison with the gamma mixture model

Let us compare the univariate [see (6.5)] and bivariate [see (2.3)] densities based on polynomial expansions with a gamma benchmark with that of a mixture of gamma densities. Without loss of generality, let us only consider the univariate case.

First, let us remark that the two types of models are non-nested. For instance, the marginal density

$$\tilde{f}_0(x) \propto e^{-cx} x^{\alpha-1}(1 + x + x^2/5)$$

is a gamma mixture, but cannot be obtained from the polynomial expansion based density (6.5). Indeed, the marginal density of a pseudo-mixture is $e^{-cx} x^{\alpha-1} \frac{U'(x)De}{e'De}$, where the polynomial $\frac{U'(x)De}{e'De}$ takes nonnegative value for all positive and negative x, which is not the case for $1 + x + x^2/5$. On the other hand, it can be checked that the marginal density

$$\hat{f}_0(x) \propto e^{-cx} x^{\alpha-1}(1 - x + x^2)$$

can be attained from (6.5), but is not a gamma mixture.

The standard argument concerning the approximation of a positive, univariate distribution by Gamma mixtures is given by Tijms (1994). Let us denote by F the cumulative distribution function, then Tijms shows that the infinite mixture:

$$f_\theta(x) = \sum_{i=1}^{\infty} \Big[F(i\theta) - F((i-1)\theta)\Big] \frac{x^{i-1} e^{-x/\theta}}{\theta^i (i-1)!},$$

defines a distribution that converges weakly to the initial distribution F, when θ goes to 0. Indeed, its characteristic function converges to that of F, when θ goes to 0. Thus the finite gamma mixture, which is obtained by truncating the previous infinite sum and re-normalizing, can be utilized as an approximation of the initial distribution.

This approximation scheme, however, has the inconvenience that the limiting case $\theta = 0$ does not define a proper density function. Thus since a positive θ has to be chosen in a finite gamma mixture model, there remains an approximation error even if we leave the infinite sum un-truncated. This is not the case when we consider the pseudo-mixture, for which the only approximation error comes from the truncation. This explains why the gamma mixture is less efficient in terms of density approximation.

Appendix 3 Numerical integration Vs Monte-Carlo

Our experiment concerns the computation of the entries of matrix Π, or, equivalently, of the following integrals:

$$\int_0^\infty e^{-x/c} \frac{x^{i+j+\alpha-1}}{U'(x)De} \mathrm{d}x,$$

for $i, j \in [|0, 2J|]$. To this end we compare two approaches. The first approach is the adaptive quadrature approach, implemented in most statistical packages. The other approach is the standard Monte-Carlo simulation method. If we denote by $(Y_i, i = 1, 2, ...)$ i.i.d. samples following the gamma distribution with shape parameter α and scale parameter c, then we have, by the law of large numbers:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{Y_i^n}{U'(Y_i)De} \longrightarrow \frac{1}{\Gamma(\alpha)c^\alpha} \int_0^\infty e^{-x/c} \frac{x^{j+\alpha-1}}{U'(x)De} \mathrm{d}x, \qquad \forall j \in [|0, 4J|],$$

when the sample size N goes to infinity. The following table compares the computational time and relative accuracy of the two approaches, conducted with the statistical package R on a standard PC.

	Numerical Integration	Monte-Carlo (10^8 simulations)
Time used	10^{-4} s	10 s
Relative Accuracy	10^{-10}	10^{-4}

Table 3: Comparison of the performance of the two methods. The relative accuracy of the numerical integration is directly obtained from R, whereas that of the Monte-Carlo is obtained from the central limit theorem.

Appendix 4 Simulation of the trajectory of the process

Let us discuss the simulation of trajectories of the state process. When the entries of Π are nonnegative, the process can be simulated quite easily, using the chain structure (2.15). Indeed, given X_t, we can i) simulate S_t using elementary probabilities $\frac{U'(X_t)D}{U'(X_t)D(X_t)e}$; ii) given $S_t = j$ simulate X_{t+1} by drawing from the density $q_j(x) \propto \phi(x)x^j$. For instance, if ϕ is gamma, then q_j is also a gamma density $\gamma(c, \alpha + S_t)$; in the general case, draws from this distribution can be obtained from a simple acceptance-rejection method.

This approach is no longer applicable when some entries of D are negative. Let us propose the acceptance-rejection method, inspired by the work of Gallant and Tauchen (1993). First, we remark that:

$$f_1(x_{t+1} \mid x_t) = \frac{\phi(x_{t+h})U'(x_t)DU(x_{t+h})}{U'(x_t)De} \leq \phi(x_{t+h})\frac{U'(x_t)D_2U(x_t)}{U'(x_{t+h})De} := b(x_{t+h}|x_t),$$

where D_2 is the matrix obtained in a similar way as D, but by replacing all entries of B by the corresponding absolute value:

$$d_{2,j,k} = \frac{\Gamma(j+\alpha_1)\Gamma(k+\alpha_2)}{c_1^{j+\alpha_1}c_2^{k+\alpha_2}} \sum_{\substack{j_1+j_2=j \\ 0\leq j_1,j_2\leq J}} \sum_{\substack{k_1+k_2=k \\ 0\leq k_1,k_2\leq J}} |b_{j_1,k_1}b_{j_2,k_2}|, \qquad \forall j,k$$

Let us denote the new conditional density $g(x|x_t) := \phi(x)\frac{U'(x_t)D_2U(x)}{U'(x_t)D_2e}$, which is a mixture of, say, Gamma densities and can be simulated exactly. Then we generate an independent pair (v_1, v_2) such that v_1 follows the uniform distribution on $[0, 1]$ and v_2 follows $g(\cdot|x_t)$. If

$$v_1 > f_1(v_2 \mid x_t)/b(v_2|x_t),$$

then we reject the pair (v_1, v_2) and try again. Otherwise, if

$$u \leq f_1(v_2 \mid x_t)/b(v_2|x_t),$$

we accept v_2 as a sample from the conditional distribution $f_1(\cdot \mid x_t)$.

Appendix 5 Autoregressive gamma process

The literature has already considered state space models with an Autoregressive Gamma process (ARG) [see e.g. Pitt et al. (2002), Gouriéroux and Jasiak (2006a), Creal (2016)] based state process. The ARG process is the exact time-discretization of the Cox-Ingersoll-Ross process and

its dynamics is defined as follows:

- conditional on X_t, count variable Z_t follows a Poisson distribution with parameter βX_t.
- conditional on Z_t, variable X_{t+1} follows $\gamma(\alpha + Z_t, c)$, where c is the scale parameter.

Thus the ARG process has a causal scheme analogous to equation(2.15):

$$\dots Z_{t-1} \to X_t \to Z_t \to X_{t+1} \to Z_{t+1} \dots$$

It has been shown by Gouriéroux and Jasiak (2006a) that the ergodicity condition of the ARG is $\rho = \beta c < 1$, with a gamma $\gamma(\alpha, \frac{c}{1-\rho})$ stationary distribution. Thus under stationarity we have:

$$f_0(x_t) = \frac{x_t^{\alpha-1} e^{-\frac{x_t(1-\rho)}{c}} (1-\rho)^\alpha}{\Gamma(\alpha) c^\delta} = \phi(x_t) e^{\rho/c x_t} (1-\rho)^\delta \qquad \text{(eq. a.1)}$$

$$f(x_{t+1}|x_t) = \sum_{j=0}^{\infty} \frac{e^{-\rho/c x_t} (\beta x_t)^j}{\Gamma(j+1)} \frac{(1-\rho)^\alpha}{\Gamma(\alpha+j) c^{\alpha+j}} x_{t+1}^{\alpha+j-1} e^{-\frac{x_{t+1}}{c}} = \phi(x_{t+1}) \sum_{j=0}^{\infty} \frac{(\beta/c)^j \Gamma(\alpha)(1-\rho)^\alpha}{\Gamma(j+1)\Gamma(\alpha+j)} e^{-\beta x_t} x_t^j x_{t+1}^j \qquad \text{(eq. a.2)}$$

$$f(x_t, x_{t+1}) = \phi(x_t)\phi(x_{t+1}) \sum_{j=0}^{\infty} \frac{(\beta/c)^j \Gamma(\alpha)(1-\rho)^\alpha}{\Gamma(j+1)\Gamma(\alpha+j)} x_t^j x_{t+1}^j, \qquad \text{(eq. a.3)}$$

where

$$\phi(x) = \frac{1}{c^\alpha \Gamma(\alpha)} e^{\alpha-1} e^{-x/c}.$$

In practice the expression of the joint density $f(x_t, x_{t+1})$ is often truncated at a high order J:

$$\begin{aligned} f(x_t, x_{t+1}) &\approx \phi(x_t)\phi(x_{t+1}) \sum_{j=0}^{J} \frac{(\beta/c)^j \Gamma(\alpha)(1-\rho)^\alpha}{\Gamma(j+1)\Gamma(\alpha+j)} x_t^j x_{t+1}^j \\ &= \phi(x_t)\phi(x_{t+1}) U'(x_t) Q_J U(x_{t+1}), \end{aligned} \qquad \text{(eq. a.4)}$$

where matrix Q_J is the $(J+1) \times (J+1)$, with $j-th$ diagonal entry $\frac{(\beta/c)^j \Gamma(\alpha)(1-\rho)^\alpha}{\Gamma(j+1)\Gamma(\alpha+j)} \mu_j^2 = \frac{\rho^j \Gamma(\alpha+j)(1-\rho)^\alpha}{\Gamma(j+1)\Gamma(\alpha)}$. In this expression, the parameter βc is the correlation coefficient between X_t and X_{t+1}. It is easily shown [see e.g. Gouriéroux and Jasiak (2006a)] that the joint distribution of X_t and X_{t+h} has a similar expression as (eq. a.4), except that $\rho = \beta c$ should be replaced by $\rho_h = \rho^h$:

$$f(x_t, x_{t+h}) \approx \phi(x_t)\phi(x_{t+1}) U'(x_t) \text{Diag}\Big(\frac{\rho^0 \Gamma(\alpha+0)(1-\rho^h)^\alpha}{\Gamma(0+1)\Gamma(\alpha)}, \dots, \frac{\rho^{Jh} \Gamma(\alpha+J)(1-\rho^h)^\alpha}{\Gamma(2J+1)\Gamma(\alpha)} \Big) U(x_{t+h}) \qquad \text{(eq. a.5)}$$

As a consequence, when the ARG model is applied to the SV model (6.4), the joint p.d.f. of (Y_t, Y_{t+h}) is equal to:

$$f(y_t, y_{t+h}) \approx g'(y_t)\text{Diag}\Big(\frac{\rho^0\Gamma(\alpha+0)(1-\rho^h)^\alpha}{\Gamma(0+1)\Gamma(\alpha)}, ..., \frac{\rho^{Jh}\Gamma(\delta+J)(1-\rho^h)^\alpha}{\Gamma(J+1)\Gamma(\alpha)}\Big)g(y_{t+h}).$$

This equation can be used to conduct maximum composite likelihood estimation. As noted by Creal (2016), the quality of approximation of this latter equation depends on the value of J. Roughly speaking, the closer ρ to unity, the more slowly the infinite summation (eq. a.3) converges, hence the larger value of J we should take [see also the discussion at the end of Section 4.1]. In the application, we take $J = 100$.

Appendix 6 Moments of Y_t under different regimes of S_t in the stochastic volatility model

Let us compute the components of vector $\int g'(y_t)y_t^2 \mathrm{d}y_t$, where (Y_t) follows the model (6.4), with non symmetric error term ϵ_t [see equation (6.6)]. When $I = 1$, we have:

$$\begin{aligned}
\int g_j(y_t)y_t^2 \mathrm{d}y_t &= \frac{1}{1+\beta_1^2}\Big[\frac{1}{c(\alpha+j-1)} + c\beta_1^2(\alpha+j)\sqrt{c(\alpha+j+1)}\int y_t^4 h_{j+1}(\sqrt{c(\alpha+j+1)}y_t)\mathrm{d}y_t\Big] \\
&= \frac{1}{1+\beta_1^2}\Big[\frac{1}{c(\alpha+j-1)} + c\beta_1^2\frac{\alpha+j}{c^2(\alpha+j+1)^2}\int z_t^4 h_{j+1}(z)\mathrm{d}z\Big] \\
&= \frac{1}{1+\beta_1^2}\Big[\frac{1}{c(\alpha+j-1)} + \beta_1^2\frac{\alpha+j}{c(\alpha+j+1)^2}\frac{3(2\alpha+2j+2-2)}{2\alpha+2j-2}\big(\frac{2\alpha+2j+2}{2\alpha+2j}\big)^2\Big] \\
&= \frac{1+3\beta_1^2}{c(1+\beta_1^2)(\alpha+j-1)}.
\end{aligned}$$

The formula above is only valid, when $\alpha + j > 1$. Since j takes values in $0, ..., 2J$, when $\alpha \in]0, 1]$, variable Y_t has an infinite variance. Similarly, when $I = 2$, we have:

$$\int g_j(y_t)y_t^2 \mathrm{d}y_t = \frac{1+3(\beta_1^2+2\beta_2)+15\beta_2^2}{c(\alpha+j-1)(1+\beta_1^2+2\beta_2+3\beta_2^2)}.$$

Similarly, the first, third and fourth moments of the components of vector function g are:[25]

$$\int g_j(y_t) y_t \mathrm{d}y_t = \frac{1}{c^{1/2}} \frac{\Gamma(\alpha + j - \frac{1}{2})}{\Gamma(\alpha + j)} \frac{2\beta_1 + 6\beta_1\beta_2}{1 + \beta_1^2 + 2\beta_2 + 3\beta_2^2},$$
$$\int g_j(y_t) y_t^3 \mathrm{d}y_t = \frac{1}{c^{3/2}} \frac{\Gamma(\alpha + j - \frac{3}{2})}{\Gamma(\alpha + j)} \frac{6\beta_1 + 15\beta_1\beta_2}{1 + \beta_1^2 + 2\beta_2 + 3\beta_2^2},$$
$$\int g_j(y_t) y_t^4 \mathrm{d}y_t = \frac{\Gamma(\alpha + j - 2)}{c^2 \Gamma(\alpha + j)} \frac{3 + 15(\beta_1^2 + 2\beta_2) + 105\beta_2^2}{1 + \beta_1^2 + 2\beta_2 + 3\beta_2^2}.$$

References

Abbring, J. H. (2010). Identification of Dynamic Discrete Choice Models. *Annu. Rev. Econ.*, 2(1):367–394.

Aït-Sahalia, Y. (2002). Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-form Approximation Approach. *Econometrica*, 70(1):223–262.

Andersen, T. G. and Sørensen, B. E. (1996). GMM Estimation of a Stochastic Volatility Model: a Monte Carlo Study. *Journal of Business & Economic Statistics*, 14(3):328–352.

Bauwens, L. and Hautsch, N. (2006). Stochastic Conditional Intensity Processes. *Journal of Financial Econometrics*, 4(3):450–493.

Beare, B. K. (2010). Copulas and Temporal Dependence. *Econometrica*, 78(1):395–410.

Beare, B. K. and Seo, J. (2014). Time Irreversible Copula-Based Markov Models. *Econometric Theory*, 30(05):923–960.

Beran, R. (1977). Minimum Hellinger Distance Estimates for Parametric Models. *The Annals of Statistics*, pages 445–463.

Bollerslev, T., Litvinova, J., and Tauchen, G. (2006). Leverage and Volatility Feedback Effects in High-Frequency Data. *Journal of Financial Econometrics*, 4(3):353–384.

Chang, Y., Choi, Y., and Park, J. Y. (2017). A New Approach to Model Regime Switching. *Journal of Econometrics*, 196(1):127–143.

Chen, X. and Fan, Y. (2006). Estimation of Copula-Based Semiparametric Time Series Models. *Journal of Econometrics*, 130(2):307–335.

[25] These formulas can be easily checked using a symbolic computation package such as Mathematica, and thus their proof is omitted.

Chen, Y.-T., Chou, R. Y., and Kuan, C.-M. (2000). Testing Time Reversibility Without Moment Restrictions. *Journal of Econometrics*, 95(1):199–218.

Chib, S. and Winkelmann, R. (2012). Markov Chain Monte Carlo Analysis of Correlated Count Data. *Journal of Business & Economic Statistics*, 19(4):428–435.

Corrado, C. J. and Su, T. (1996). Skewness and Kurtosis in S&P 500 Index Returns Implied by Option Prices. *Journal of Financial Research*, 19(2):175–192.

Cox, D. R. (1981). Statistical Analysis of Time Series: Some Recent Developments. *Scandinavian Journal of Statistics*, 8:93–115.

Creal, D. D. (2016). A Class of Non-Gaussian State Space Models with Exact Likelihood Inference. *forthcoming Journal of Business & Economic Statistics*.

Darolles, S., Florens, J.-P., and Gouriéroux, C. (2004). Kernel-based Nonlinear Canonical Analysis and Time Reversibility. *Journal of Econometrics*, 119(2):323–353.

Darolles, S., Gouriéroux, C., and Gagliardini, P. (2013). Survival of Hedge Funds: Frailty vs Contagion. *CREST DP*, 1:54.

Demni, N. and Zani, M. (2009). Large Deviations for Statistics of the Jacobi Process. *Stochastic Processes and their Applications*, 119(2):518–533.

Duffie, D., Eckner, A., Horel, G., and Saita, L. (2009). Frailty Correlated Default. *Journal of Finance*, 64(5):2089–2123.

Fernández, C. and Steel, M. F. (1998). On Bayesian Modeling of Fat Tails and Skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Ferreira, J. T. S. and Steel, M. F. (2012). A Constructive Representation of Univariate Skewed Distributions. *Journal of the American Statistical Association*, 101(474):823–829.

Feunou, B. and Tédongap, R. (2012). A Stochastic Volatility Model with Conditional Skewness. *Journal of Business & Economic Statistics*, 30(4):576–591.

Filipović, D., Mayerhofer, E., and Schneider, P. (2013). Density Approximations for Multivariate Affine Jump-Diffusion Processes. *Journal of Econometrics*, 176(2):93–111.

Gallant, A. R. and Nychka, D. W. (1987). Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica*, 55(2):363–390.

Gallant, A. R. and Tauchen, G. (1989). Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications. *Econometrica*, 57(5):1091–1120.

Gallant, A. R. and Tauchen, G. (1993). *A Nonparametric Approach to Nonlinear Time Series Analysis: Estimation and Simulation*. Springer.

Ghysels, E., Gourieroux, C., and Jasiak, J. (2004). Stochastic Volatility Duration Models. *Journal of Econometrics*, 119(2):413–433.

Gouriéroux, C. and Jasiak, J. (2001). State-space Models with Finite Dimensional Dependence. *Journal of Time Series Analysis*, 22(6):665–678.

Gouriéroux, C. and Jasiak, J. (2006a). Autoregressive Gamma Processes. *Journal of Forecasting*, 25(2):129–152.

Gouriéroux, C. and Jasiak, J. (2006b). Multivariate Jacobi Process with Application to Smooth Transitions. *Journal of Econometrics*, 131(1):475–505.

Gouriéroux, C. and Monfort, A. (2015). Pricing with Finite Dimensional Dependence. *Journal of Econometrics*, 187(2):408–417.

Gouriéroux, C. and Monfort, A. (2016). Composite Indirect Inference with Application to Corporate Risks. *CREST DP*.

Gouriéroux, C., Monfort, A., and Renault, E. (2016). Consistent Pseudo-Maximum Likelihood Estimators. *forthcoming Annals of Economics and Statistics*.

Hansen, L. P., Scheinkman, J. A., and Touzi, N. (1998). Spectral Methods for Identifying Scalar Diffusions. *Journal of Econometrics*, 86(1):1–32.

Harvey, A., Ruiz, E., and Shephard, N. (1994). Multivariate Stochastic Variance Models. *The Review of Economic Studies*, 61(2):247–264.

Harvey, A. C. and Shephard, N. (1996). Estimation of an Asymmetric Stochastic Volatility Model for Asset Returns. *Journal of Business & Economic Statistics*, 14(4):429–434.

Heston, S. L. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *Review of Financial Studies*, 6(2):327–343.

Hu, Y. and Shum, M. (2012). Nonparametric Identification of Dynamic Models with Unobserved State Variables. *Journal of Econometrics*, 171(1):32–44.

Huberman, G., Kandel, S., and Stambaugh, R. F. (1987). Mimicking portfolios and exact arbitrage pricing. *Journal of Finance*, 42(1):1–9.

Jarrow, R. and Rudd, A. (1982). Approximate Option Valuation for Arbitrary Stochastic Processes. *Journal of Financial Economics*, 10(3):347–369.

Jensen, S. T. and Shore, S. H. (2011). Semiparametric Bayesian Modeling of Income Volatility Heterogeneity. *Journal of the American Statistical Association*, 106(496):1280–1290.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman and Hall.

Kasahara, H. and Shimotsu, K. (2009). Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices. *Econometrica*, 77(1):135–175.

Kim, C.-J., Piger, J., and Startz, R. (2008). Estimation of Markov Regime-Switching Regression Models with Endogenous Switching. *Journal of Econometrics*, 143(2):263–273.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies*, 65(3):361–393.

Kitagawa, G. (1987). Non-Gaussian State-Space Modeling of Nonstationary Time Series. *Journal of the American Statistical Association*, 82(400):1032–1041.

Koopman, S. J., Lucas, A., and Scharth, M. (2015). Numerically Accelerated Importance Sampling for Nonlinear Non-Gaussian State-Space Models. *Journal of Business & Economic Statistics*, 33(1):114–127.

Koopman, S. J., Lucas, A., and Scharth, M. (2016). Predicting Time-Varying Parameters with Parameter-Driven and Observation-Driven Models. *Review of Economics and Statistics*, 98(1).

Lancaster, H. (1958). The Structure of Bivariate Distributions. *The Annals of Mathematical Statistics*, 29(3):719–736.

Madan, D. B. and Seneta, E. (1990). The Variance Gamma (V.G.) Model for Share Market Returns. *The Journal of Business*, 63(4):511–24.

Maskin, E. and Tirole, J. (1988). A Theory of Dynamic Oligopoly, II: Price Competition, Kinked Demand Curves, and Edgeworth Cycles. *Econometrica*, 56(3):571–599.

McCausland, W. J. (2007). Time Reversibility of Stationary Regular Finite-State Markov Chains. *Journal of Econometrics*, 136(1):303–318.

Nieto-Barajas, L. E. and Walker, S. G. (2002). Markov Beta and Gamma Processes for Modelling Hazard Rates. *Scandinavian Journal of Statistics*, 29(3):413–424.

Norets, A. and Tang, X. (2013). Semiparametric Inference in Dynamic Binary Choice Models. *The Review of Economic Studies*, 81(3):1229–1262.

Pitt, M. K., Chatfield, C., and Walker, S. G. (2002). Constructing First Order Stationary Autoregressive Models via Latent Processes. *Scandinavian Journal of Statistics*, 29(4):657–663.

Pitt, M. K. and Shephard, N. (1999). Filtering via Simulation: Auxiliary Particle Filters. *Journal of the American Statistical Association*, 94(446):590–599.

Racine, J. S. and Maasoumi, E. (2007). A Versatile and Robust Metric Entropy Test of Time-Reversibility, and Other Hypotheses. *Journal of Econometrics*, 138(2):547–567.

Ramsey, J. B. and Rothman, P. (1996). Time Irreversibility and Business Cycle Asymmetry. *Journal of Money, Credit and Banking*, 28(1):1–21.

Rosenblatt, M. (1975). A Quadratic Measure of Deviation of Two-Dimensional Density Estimates and a Test of Independence. *Annals of Statistics*, 3(1):1–14.

Ruiz, E. (1994). Quasi-Maximum Likelihood Estimation of Stochastic Volatility Models. *Journal of Econometrics*, 63(1):289–306.

Sancetta, A. and Satchell, S. (2004). The Bernstein Copula and its Applications to Modeling and Approximations of Multivariate Distributions. *Econometric Theory*, 20(03):535–562.

Scott, S. L. (2002). Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century. *Journal of the American Statistical Association*, 97(457):337–351.

Seneta, E. (2006). *Non-Negative Matrices and Markov Chains*. Springer Science & Business Media.

Szeg, G. (1939). *Orthogonal Polynomials*, volume 23. American Mathematical Society.

Tijms, H. C. (1994). *Stochastic Models: An Algorithmic Approach*, volume 303. John Wiley & Sons Inc.

Varin, C., Reid, N., and Firth, D. (2011). An Overview of Composite Likelihood Methods. *Statistica Sinica*, 21(1):5–42.

Varin, C. and Vidoni, P. (2005). A Note on Composite Likelihood Inference and Model Selection. *Biometrika*, 92(3):519–528.

Varin, C. and Vidoni, P. (2008). Pairwise Likelihood Inference for General State Space Models. *Econometric Reviews*, 28(1-3):170–185.

Xiu, D. (2014). Hermite Polynomial Based Expansion of European Option Prices. *Journal of Econometrics*, 179(2):158–177.

Zhu, D. and Galbraith, J. W. (2010). A Generalized Asymmetric Student-t Distribution with Application to Financial Econometrics. *Journal of Econometrics*, 157(2):297–305.