A Flexible and Tractable State-Space Model with Application to Stochastic Volatility

Christian Gouriéroux (CREST and University of Toronto) and Yang Lu
(Aix-Marseille University)

November 6, 2016

Abstract

We introduce a general, state-space (or latent factor) model for time series and panel data. The state process has a flexible dynamics capable of approximating any Markov process arbitrarily well. It is of finite dimensional dependence, with a latent, endogenous switching regime interpretation. This latter leads to simple recursive formulas for prediction and filtering, which is generically faster than existing, simulation based methods such as the particle filter. Under some further constraints, the model can be estimated by maximum composite likelihood, 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 conditional skewness, volatility leverage effect, as well as time non-reversibility. The methodology is illustrated using Apple stock data, which confirms the improvement of our model with respect to existing SV models.

Key words: Endogenous regime switch, polynomial expansion, finite dimensional dependence, composite likelihood, time reversibility, leverage effect.

1 Introduction

There is a great concern of developing flexible state-space, or latent factor models in Economics, Finance, and Insurance. Examples of applications include both time series and panel data, such as:

- the return of asset prices, with the stochastic volatility as the state variable [see e.g. Ruiz (1994); Kim et al. (1998)].
- panel binary event data, such as the corporate defaults [see e.g. Duffie et al. (2009)].

- (panel or time series) count data, such as the number of (buy or sell) transactions of a specific market, the number of lapses (resp. redemptions) of a life insurer (resp. investment fund), the number of defaults in a specific sector [see e.g. Darolles et al. (2013)] during each interval, or the annual number of accidents of each policyholder. In these cases the state variable is the stochastic intensity.
- (panel or time series) duration data with stochastic intensity, see e.g. Ghysels et al. (2004); Bauwens and Hautsch (2006).
- panel data on individuals' dynamic discrete choices [see e.g. Abbring (2010); Hu and Shum (2012); Norets and Tang (2013)], in which the state variable arises as the unobservable *taste*, or *belief* variable.

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 applications.

Nevertheless, compared to observation-driven models, the estimation and forecasting of parameter-driven, state-space models are generically computationally intensive, and necessitates simulation based techniques such as particle filters or MCMC [see e.g. Chib and Winkelmann (2012)]. Our paper introduces a general family of models that are *i*) sufficiently flexible to fit a wide range of data, *ii*) associated with simple estimation and forecasting procedure. We propose to 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 Markov processes arbitrarily well. Moreover it is of finite dimensional dependence, and is associated with a latent, endogenous switching regime interpretation under some further assumptions. The resulting state-space model allows for simple recursive formulas for prediction, filtering and smoothing, which is generically 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 existing methods.

While the methodology of the paper can be applied to either panel or time series data, our paper will be focused on a time series application: the stochastic volatility. In particular, we contribute to the stochastic volatility (SV) literature by proposing a model allowing for both conditional skewness, leverage effect and time non-reversibility. While the first two properties have already been studied, most models consider only time reversible dynamics, and this assumption is generically violated by financial and economic time series data.

The rest of 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 discuss the time reversibility assumption. Examples are discussed in Section 3. Forecasting, filtering formulas are derived in Section 4. In particular, we note that the derivation of the term structure of predictions requires only the computation of a *finite* number of univariate integrals. The use of models to approximate any Markov dynamics of the state variable is explored in Section 6. 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 where 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, $\mathcal{X}$ could be the whole real space $\mathbb{R}$, the positive half-space $\mathbb{R}_{>0}$, or a bounded interval such as $]0,1[$, whereas $\mathcal{Y}$ can be an infinite set such as $\mathbb{R}$ (for instance, stochastic volatility data), $\mathbb{R}_{>0}$, or a discrete set, such as $\mathbb{N}$ (count data), or $\{0,1,....,N\}$ (binomial or categorical data). For expository purpose, we assume that there are no other observable covariates.[1] 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 $\underline{Y_{t-1}}$ as well as 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 and MCMC.

Our paper proposes a general family of 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

[1] Indeed, the forecasting and estimation of the model is similar when covariates are included.

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 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 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 finance and insurance applications since a density function with negative values spells arbitrage opportunities and could lead to dangerous financial consequences. 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, in the sense that when J goes to infinity, the proposed expansion-based density can approximate *any* bivariate density function arbitrarily well [see Section 5].

As a special case, 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 B can be written as $B = \beta\beta'$, where β is a $J + 1$ dimensional vector, then we have $f(x_t, x_{t+1}) = \Big[\phi(x_t)\sum_{i=0}^{J}\beta_i x_t^i\Big]\Big[\phi(x_{t+1})\sum_{i=0}^{J}\beta_i x_{t+1}^i\Big]$.

The choice of the benchmark density can be motivated by the prior knowledge about the process (X_t), or more frequently, by mathematical convenience. For instance, if (X_t) takes value 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 distribution, or log-normal distribution; if (X_t) takes value in the real space, 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.

The flexible distribution (2.3) is valid if the moments $\mu_j = \int x^j \phi(x) \mathrm{d}x, j = 1, 2, ..., 2J$ exist. 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 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 form:

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

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: $\int \phi(x) U(x) \mathrm{d}x = e$.

Expression (2.6) assumes implicitly that all the 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. Thus, without loss 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 and sufficient that it defines two identical margins.

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

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

In this case, the marginal distribution is:

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

Proof. See Appendix 1.1. □

[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:

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

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

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 B is 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 will derive 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. Thus the one-step-ahead conditional distribution has a simple matrix product form. This type of formula can be extended to longer horizon.

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 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. First, in order for it to be well defined, let us 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 left eigenvector of* Π *with unitary eigenvalue.*

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

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

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. □

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 is always a linear combination of densities which are 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 $\Pi^{h-1}e = e$. 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:

[4]For instance by using the command "integrate" in R. Note that 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.

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

The proof of this result is omitted, as it is based on the same arguments as for a finite state Markov chain [see e.g. Seneta (2006)], even if the entries of Π are not necessarily nonnegative.

Moreover, by equation (2.14), we can 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 quickly (slowly) 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 can be 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 interpretation. Let us denote by* S_t *the regime indicator at date* t *with* $2J+1$ *possible regimes. We can write:*

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

whereas the conditional density of X_{t+1} *given* $\underline{S_t}$ *and* $\underline{X_t}$ *is:*

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

Proof. The proof is immediate, since the components of vector $\frac{U'(X_t)D}{U'(X_t)De}$ sum up to unity. □

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

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

In particular, the latent process (S_t) is also a Markov chain with respect to its own history, called the embedded chain of (X_t). This Markov chain 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:

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

Note also that when the embedded chain (S_t) exists, by Corollary 1, 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 use 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):

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

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.

This 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.

Moreover, this tractability remains true in the general case, 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 is of finite dimensional dependence (FDD) [see Gourieroux and Jasiak (2001); Gourieroux et al. (2014)], that is, the conditional density is a linear combination

of a finite number of products of functions of X_t and X_{t+1}.

Let us finally interpret 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 following two special 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 time reversibility of the process, and is left for the next subsection.

2.4 Time reversibility

The symmetry of D implies the symmetry of the joint distribution 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 process (X_t) is time-reversible.

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 antisymmetric*[6]*.*

Proof. Let us denote $V(x) = (1, x, ..., x^J)$. Then 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 bivariate polynomials. Thus the previous identity is satisfied if and only if at least one of the two multiplicative terms are identically zero, that is to say, B is antisymmetric, or symmetric. □

[6] Also called skew-symmetric.

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

The time-reversibility [see e.g. Darolles et al. (2004)] means that the distribution of (X_t, X_{t+1}) is the same as that of (X_{t+1}, X_t). It is satisfied by a large number of time series models considered in the financial literature, including: (log-)normal ARMA processes, autoregressive gamma process[7], 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) for evidences of time irreversibility of macro and asset return data, respectively]. In order to allow for time non-reversibility in the dynamics of the process (X_t), let us now characterize the conditions on B, such that $(D - D')e = 0$ is still satisfied, without D being symmetric.

To this end, we introduce an equivalent parametrisation of B, which is based on the idea is that any square matrix (here B) can be uniquely decomposed into the sum of a symmetric matrix $B_1 = \frac{1}{2}(B + B')$, and an antisymmetric matrix $B_2 = \frac{1}{2}(B - B')$. In particular, all the diagonal entries of the latter are zero. Thus instead of parametrizing $B = B_1 + B_2$, we will 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:

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

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) are symmetric (resp. antisymmetric), that is,

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

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

$b_{1,j_1,k_1} = b_{1,k_1,j_1}$, 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 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 a kind of orthogonality conditions between the entries of matrices B_1 and B_2, which have $\frac{(J+1)(J+2)}{2} - 1$ and $\frac{J(J+1)}{2}$ parameters, respectively. At the same time, since the sum of all entries of D_2 is zero, the condition $D_2 e = 0$ involves at most $(2J+1)-1 = 2J$ orthogonality conditions. These conditions are fairly easy to handle, as they are linear in the parameters B_1 and B_2.

To summarize, we have obtained a simple parametrisation of our model in the time non reversible case. 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 fails to 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 the 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 A special case with $J = 1$

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

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

Thus $(D' - D)e = 0$ is equivalent to, after simplification:

$$\begin{aligned} (b_{01} - b_{10})(2b_{00}\mu_1 + \mu_2(b_{10} + b_{01})) &= 0, \\ (b_{01} - b_{10})\mu_2(b_{01} + b_{10} + 2b_{11}\mu_1) &= 0, \\ (b_{01} - b_{10})\mu_1(b_{00} - b_{11}\mu_2) &= 0. \end{aligned} \tag{3.1}$$

Under the assumption that $\mu_1\mu_2 \neq 0$, this system can be simplified to two cases. That is, either $i)$ $b_{01} = b_{10}$, which means that B is symmetric, or $ii)$, $b_{01} + b_{10} + 2b_{11}\mu_1 = b_{00} - b_{11}\mu_2 = 0$. This latter contains the case where B is anti-symmetric, that is $b_{01} + b_{10} = b_{00} = b_{11} = 0$.

Thus a stationary dynamics compatible with (3.1) and a symmetric B is, for instance:

$$f(x_t, x_{t+1}) = \frac{1}{M}\phi(x_t)\phi(x_{t+1})\Big(1 + b_{01}(x_t + x_{t+1})\Big)^2. \tag{3.2}$$

Let us now derive the form of the matrix Π of model (3.2). 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 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:

$$Y_t = \sigma(X_t)\epsilon_t, \tag{3.3}$$

where the ϵ_t's are $IIN(0,1)$ and independent of the nonnegative volatility process $\sigma_t = \sigma(X_t)$, which is a function of process (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 (2015)]. In model (3.3), 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. Nevertheless, in our paper we will consider an inverse, (rooted) Gamma type specification, that is:

$$\sigma(X_t) = \frac{1}{\sqrt{X_t}}.$$

Indeed, it will be shown in the application section that this model has the advantage that it leads to closed form expression of the marginal, or pairwise joint density.[9]

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 degree of freedom α:

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

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.

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)]. A (bivariate) copula is the CDF of a bivariate

[9] Moreover, 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.

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 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 density $f(x_{t+1}|x_t)$.

A family of copulas that solves this difficulty is called 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:

$$C(x_1,x_2)=x_1x_2(1+\theta(1-x_1)(1-x_2)),\qquad |\theta|<1/3. \tag{3.4}$$

Our model has a different philosophy. Its specification 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 generically not uniform. Indeed, if the marginal density $f_0(x_t)=\phi(x_t)\frac{U'(x_t)De}{e'De}$ is 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 :

$$(D'-D)e=0 \tag{3.5}$$

$$\int_0^1 \frac{e'De}{U'(x_t)De}U(x_t)\mathrm{d}x_t=e. \tag{3.6}$$

In general, equation (3.6) 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 predictive future values 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 X_T given $\underline{Y_T}$.

4.1 Forecasting

The unconditional density of Y_t given the past is:

Proposition 7.

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

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

$$P'(\underline{y_0}) = \frac{e'D}{e'De}, \tag{4.3}$$

$$P'(\underline{y_t}) = P'(\underline{y_{t-1}}), \Pi(\underline{y_t}), \tag{4.4}$$

with matrix $\Pi(\underline{y_t})$ *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. □

Let us provide some comments. First, the dependence of y_t on its whole past $\underline{y_{t-1}}$ is summarized by a finite-dimensional vector $P'(\underline{y_{t-1}})$, which is called *mimicking factor* [see e.g. Gourieroux and Jasiak (2001); Huberman et al. (1987)] in finance.

Second, when the state process has the embedded switching regime interpretation, We have the following causal chain:

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

Moreover, 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 observation of $\underline{y_{t-1}}$:

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

As a consequence, formula (4.4) is the analogy of the Kitagawa filter [see Kitagawa (1987)] for hidden Markov models (HMM), except that the discrete chain (S_t) is endogenous. 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. This matrix is *state-dependent*. Indeed, given S_{t-1}, the future value of the state variable S_t, and the observable variable Y_t are dependent since both

depend on X_t. Thus the transition of the latent state variable depends on the current value of Y_t.This is the analogy of *volatility feedback*, or asymmetric volatility in the stochastic volatility literature [see e.g. Bollerslev et al. (2006)]. Second, $\Pi(\underline{y_t})$ is generically *not* a stochastic matrix. Indeed, the sums of each of its rows are:

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

which is generically not equal to e. Nevertheless, by construction, $\Pi(\underline{y_t})$ is such that 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). Moreover, let us remark that in the special case when Y_t is independent from X_t, that is $l(y_t|x_t) = l(y_t)$,[10] we have

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

and the normalization constant

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

that is $\Pi(\underline{y_t}) = \Pi$.

Third, entries of the matrix in equation (4.4) can be computed numerically using a routine, adaptive quadrature method. In total, it suffices to compute a finite number (indeed $4J+2$) of univariate integrals, that are respectively: $\int \phi(x)x_t^n\frac{l(y_t|x_t,y_{t-1})}{U'(x_t)De}\mathrm{d}x_t, \forall n \in [|0,4J|]$. 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 analogy of the updating of the population in a particle filter [see e.g. Pitt and Shephard (1999); Koopman et al. (2015)]. Indeed, in the latter algorithm, 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$\int \phi(x)x_t^n\frac{l(y_t|x_t,y_{t-1})}{U'(x_t)De}\mathrm{d}x_t$ via Monte-Carlo simulation. We refer to Appendix 3 for a more detailed comparison between the two methods.

Fourth, when exogenous covariates (Z_t) are included, that is, the conditional distribution of

[10] For instance, when one observation y_t is missing, we can write, say, $l(y_t|x_t) = l(y_t) = 1$, that is, the observation y_t is non informative of the state variable.

the observable variable becomes $l(y_t|x_t, z_t, \underline{y_{t-1}})$, the previous proposition remains true, once the transition matrix $\Pi(\underline{y_t})$ is replaced by $\Pi(\underline{y_t}, z_t)$, whose expression is the same as (4.5), except that $l(y_t|x_t, \underline{y_{t-1}})$ should be replaced by $l(y_t|x_t, z_t, \underline{y_{t-1}})$.

Finally, our state-space model can be viewed as a flexible generalization of the method of Nieto-Barajas and Walker (2002); Creal (2015), 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, since the probability that S_t takes large values is numerically small. This allows the author to derive similar filtering and forecasting formulas. Our model generalizes the result of Creal 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 infinite transition matrix, and computationally speaking, a very large number of possible values for the intermediate variable S_t,[11] whereas in our case no approximation is involved, and the number of the "pseudo" states is fixed and much smaller.

4.2 Filtering and smoothing

The simple forecasting formula is associated with simple expression of the predictive density of X_t given the past observation. 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:

Corollary 4. *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:

[11] Fore instance, Creal (2015) propose to use a finite chain with 3000 states to approximate the dynamics of the process S_t.

Proposition 8. *The smoothing density is, for $t < T$:*

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

Thus this smoothing density can be computed recursively. Indeed, we introduce the vector Q'_t backwardly by:

$$Q_T = g(y_T|\underline{y_{T-1}}), \tag{4.9}$$

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

then we have:

$$l(x_t|\underline{y_T}) = \frac{1}{P'(\underline{y_{t-1}})g(y_t|\underline{y_{t-1}})} \frac{P'(\underline{y_{t-1}})\left[\phi(x_t)\frac{U(x_t)U'(x_t)D}{U'(x_t)De}l(y_t|x_t)\right]Q_{t+1}}{P'(\underline{y_{t-1}})\Pi(\underline{y_t})Q_{t+1}} \tag{4.11}$$

Proof. See Appendix 1.7. □

Thus equation (4.11) is the analogy of the forward-backward smoothing algorithm for hidden Markov models and particle filters. Again, it can be interpreted by the switching regime interpretation. To this end, let us make the simplifying assumption that $l(y_t|\underline{y_{t-1}}, x_t) = l(y_t|x_t)$. In this case, vector $g(\underline{y_T}|\underline{y_{T-1}})$, that is the conditional distribution of Y_T given $\underline{Y_{T-1}}$ and S_T, is also, up to a normalization constant, the vector of conditional probabilities of S_{T-1} given y_T only. Thus similarly, Q_t is, up to normalization, the vector of conditional probabilities of S_{t-1} given $\overline{y_t} := y_t, ..., y_T$ only, and equation (4.10) provides the backward updating formula of Q_t, in the same spirit as the forward updating formula of $P'(\underline{y_t})$ [see equation (4.4)].

Then we remark that we have the causal scheme:

$$\begin{array}{ccccccccccccccc}
\cdots & \to & X_{t-2} & \to & X_{t-1} & \to & S_{t-1} & \to & X_t & \to & S_{t+1} & \to & X_{t+1} & \to & X_{t+2} & \to & \cdots \\
\cdots & & \downarrow & & \downarrow & & & & \downarrow & & & & \downarrow & & \downarrow & & \cdots \\
\cdots & & Y_{t-2} & & Y_{t-1} & & & & Y_t & & & & Y_{t+1} & & Y_{t+2} & & \cdots
\end{array}$$

In the first line, we have a Markov chain, in the sense that each variable depends on its past only

via its nearest left neighbour. Then we have:

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

4.3 A special case

Let us now consider the case, 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, in this case, the conditional transition matrix $\Pi(\underline{y_t})$ depends, up to a multiplicative constant $\frac{1}{l(y_t|y_{t-1})}$, only on y_t. Thus when $\mathcal{Y}$, the domain of Y_t is finite, such as when Y_t is conditionally binomial or categorical, then 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. If $l(y_t|\underline{y_{t-1}}, x_t)$ depends also on $\underline{y_{t-1}}$, then generically no analytic formula exists for the predictive density $l(y_{T+h}|\underline{y_T})$. In this case future trajectories should be simulated in order to get an approximated value of $l(y_{T+h}|\underline{y_T})$. However, under the extra assumption (4.12), we get following formula formula:

Lemma 2. Under Assumption (4.12), the $h-$step-ahead conditional 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. □

This lemma indicates that 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 (2015) 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, let us explain how model (2.3) approximates the dynamics of *any* given Markov process. It is based on the orthonormal decomposition of a *square rooted* density. Then similar approaches in the literature, especially those based on the decomposition of the density, are briefly discussed.

5.1 Polynomial decomposition of a square rooted density

Let us assume that[12] 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 product space L^2 associated with $\phi(x_t)\phi(x_{t+1})\mathrm{d}x_t\mathrm{d}x_{t+1}$. 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 $\phi(x_t)\phi(x_{t+1})\mathrm{d}x_t\mathrm{d}x_{t+1}$:

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

Thus we get the following orthonormal decomposition:

$$\sqrt{\frac{f(x_t,x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}=\sum_{i,j=0}^{\infty} a_{i,j}P_i(x_t)P_j(x_{t+1}), \tag{5.1}$$

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 (5.1) 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 unitary norm. Thus a natural approximation of $f(x_t,x_{t+1})$ is obtained by truncating the RHS of equation (5.1), and taking the square:

$$f(x_t,x_{t+1})\approx \phi(x_t)\phi(x_{t+1})\Big[\sum_{i,j=0}^{J} a_{i,j}P_i(x_t)P_j(x_{t+1})\Big]^2. \tag{5.2}$$

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

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 $\{P_i(x_t)P_j(x_{t+1}), 0 \leq i,j \leq J\}$. This function is not yet a density, but should be normalized to obtain the following approximating density:

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

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

Proposition 9. *The sequence of densities f_J approximates f arbitrarily well in terms of the Hellinger distance, when J goes to infinity:*

$$\iint \Big|\sqrt{f_J(x_t, x_{t+1})} - \sqrt{f(x_t, x_{t+1})}\Big|^2 \mathrm{d}x_t \mathrm{d}x_{t+1} \longrightarrow 0. \tag{5.4}$$

Proof. See Appendix 1.6. □

This approximation result suggests joint choices of the benchmark density and the orthonormal polynomials. For instance when ϕ is a gamma density, the orthonormal polynomials are called 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 process, with the approximation of (univariate) densities by means of polynomial expansion *directly* [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 $X_{t+1}|X_t$ can be decomposed into the product of, say, the Gaussian density, and an infinite polynomial sum:

$$f(x_{t+1}|x_t) = \phi(x_{t+1})\sum_{i=0}^{\infty} a_i(x_t)P_i(x_{t+1}), \tag{5.5}$$

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 (5.5) 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, a major downside of this approach is that the obtained truncated version of (5.5) is *not* a proper density, since it is generically not nonnegative. This leads to several difficulties. First, it is not possible to evaluate the statistical quality of approximation 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 aspects. First, our method guarantees the positivity of the conditional density. Second, the decomposition (5.5) 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 conditions. Finally, instead of considering the conditional distribution $X_{t+1}|X_t$,[13] 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 ensures the stationarity of the dynamics, as well as easy-to-verify ergodicity conditions, whereas these properties are typically much harder, if possible, to derive for the approach by conditional distribution.

Our density expansion approach is also to be compared with the recently proposed Bernstein copula [see Sancetta and Satchell (2004)]:

$$C(x_1, x_2) = \sum_{i=0}^{J} \sum_{j=0}^{J} \alpha(i,j) P_{i,J}(x_1) P_{j,J}(x_2), \qquad \forall x_1, x_2 \in [0,1], \tag{5.6}$$

where $P_{i,J}(x)$ is the Bernstein polynomial: $P_{i,J}(x) = C_J^i x^i (1-x)^{J-i}$, and $\alpha(i,j)$ is a sequence of constants satisfying some constraints in order for (5.6) to define a genuine copula.

The Bernstein copula is based on the Bernstein-Weierstrass theorem, which approximates a given bivariate distribution function $C \in \mathcal{C}([0,1]^2)$ by the following function:

$$C(x_1, x_2) \approx \sum_{i=0}^{J} \sum_{j=0}^{J} C(\frac{i}{J}, \frac{j}{J}) P_{i,J}(x_1) P_{j,J}(x_2).$$

[13] 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.

Thus the corresponding joint copula density is

$$c(x_1, x_2) = \sum_{i=0}^{J} \sum_{j=0}^{J} \alpha(i,j)(P_{i,J}(x_1))'(P_{j,J}(x_2))', \tag{5.7}$$

which is also of *finite dimensional dependence*. Thus (5.7) can also be used to specify the dynamics of the state process, once the latter is normalized to have uniform margin. 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 $g(y_t|\underline{y_t}, x_t)$.

6 Application

6.1 Statistical inference

6.1.1 Maximum likelihood estimation

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. Thus the log-likelihood function can be computed recursively, with a computational cost that is lower than using simulation-based techniques such as particle filters.

6.1.2 Composite likelihood estimation

Let us now introduce an alternative, composite likelihood (CL) estimation method [see e.g. Varin and Vidoni (2008); Varin et al. (2011) for a review] that is particularly suited for our model. Although it is (slightly) less efficient[14] than the full maximum likelihood estimation, its computational cost is extremely low, and is similar to that of the Generalized Method of Moments. The composite likelihood is based on the following closed form expression of the joint distribution of (Y_t, Y_{t+h}):

Lemma 3. 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. □

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

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 interpretation of (6.2) is $f_Y(y_t, y_{t+h}) = \mathbb{E}[f_Y(y_t, y_{t+h}|S_t, S_{t+h})]$. That is, the joint density is first computed conditional on the regimes of S_t and S_{t+h}, before integrating them out.

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 composite likelihood estimator:

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

Varin and Vidoni (2008) show that under mild regularity and identification conditions, the estimator (6.3) is asymptotically consistent and normally distributed.

Let us now talk about the choice of integer m, and of the weights $w_k, k = 1, ..., m$. As far as we know, there is no general theory on the optimal choice of these parameters [see however Varin et al. (2011) for a discussion]. Nevertheless, if m is too small, then the amount of information contained in the composite likelihood function is reduced, and hence the efficiency of CL estimation would be low. At the same time, the computational effort required is proportional to m, thus in the application, we take $m = 5$ in the application. As for the weights w_k, we set them to be $w_k = 0.9^k, \forall k$. In other words, pairs with a smaller distance have a higher weight, although we let the weight w_k to decrease slowly, in order to reflect the high persistence of financial returns.

Finally, the pairwise composite likelihood is easy to compute, when function g allows for closed form expression [see the discussion below Lemma 2]. This will be 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[15]. Moreover, Varin and Vidoni (2008) show, via a simulation experiment, that the efficiency loss with respect to the MLE is rather small, compared to usual GMM [see e.g. Andersen and Sørensen (1996)], which requires a similar computational cost as the CL method.

[15] We can remark that its computation is parallelisable.

6.2 A stochastic volatility application

6.2.1 Time reversible models

i) 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 the same constant. Therefore, we set $b_{0,0} = 1$.[16]

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

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

since $\mu_j = \frac{\Gamma(\alpha+j)}{\Gamma(\alpha)}$. Thus each component function $g_j(\cdot)$ is the density of a rescaled Student's t distribution, which has been widely used in Finance to account for the heavy-tail of the 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 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 M1 assumes that the conditional return is normal. This distribution is symmetric, and thus the model does not allow 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

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

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. That is, $\nu_k = 0$ if k is odd, and $\nu_k = (k-1)(k-3)\cdots 1$ if k is even. Thus this density function $h(\epsilon)$ has the same spirit 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:

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

It can be shown that under this distributional assumption, the skewness

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

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

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

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

A consequence of the density specification (6.5) is that 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)].

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

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

In order to compute the variance $\mathbb{E}[Y_t^2] = [\int g'(y_t)y_t^2\mathrm{d}y_t]\frac{De}{e'De}$ under model M2 as well as the nested model M1, let us compute the components of vector $\int g'(y_t)y_t^2\mathrm{d}y_t$. When $I=1$ [see equation (6.6)], we have:

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

Note that the previous formula 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:

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

As a consequence of these formulas, we get the joint moment

$$
\mathbb{E}[y_t^2 y_{t+h}^2] = \Big[\int g(y_t)y_t^2\mathrm{d}y_t\Big]'\frac{D\Pi^{h-1}}{e'De}\Big[\int g(y_{t+h})y_{t+h}^2\mathrm{d}y_{t+h}\Big],
$$

which can be used to compute the autocorrelation function of the squared return, that is $\mathrm{corr}[Y_t^2, Y_{t+h}^2]$.

Similarly, the first, third and fourth moments of the processes are given by $\mathbb{E}[Y_t] = [\int g'(y_t)y_t\mathrm{d}y_t]\frac{De}{e'De}$,

$\mathbb{E}[Y_t^3] = [\int g'(y_t) y_t^3 \mathrm{d}y_t] \frac{De}{e'De}$, $\mathbb{E}[Y_t^4] = [\int g'(y_t) y_t^4 \mathrm{d}y_t] \frac{De}{e'De}$, with

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

We can remark that first, as expected, the existence of higher moment require stronger conditions. For $\mathbb{E}[y_t^2]$ to exist, it suffices that $\alpha > 1$. For $\mathbb{E}[y_t^3]$ to exist, α should be larger than $3/2$. Finally, for $\mathbb{E}[y_t^4]$ to exist, α should be larger than 2; second, the different regimes $S_t = 0, ..., 2J$ lead to different tail behavior of $Y_t|S_t$, since under these regimes, the conditional densities of $X_t|S_t$, that are, $\phi(x_t)\frac{x_t^j}{\mu_j}$, have different behavior at zero. In particular, the benchmark regime $J = 0$ corresponds to the heaviest tail of $Y_t|S_t$.

iii) As a natural comparison with our flexible model, we also estimate the following benchmark model M3:

$$y_t = \frac{1}{\sqrt{x_t}} \epsilon_t, \tag{6.8}$$

where (X_t) follows an ARG process [see Appendix 6 for the definition and the representation with embedded switching regimes of an ARG process]. While Creal (2015) has shown that such ARG based models can be estimated using (approximate) maximum likelihood, for the sake of comparison with other models, we employ composite likelihood. Indeed, It is shown in Appendix 5 that the joint p.d.f. of (X_t, X_{t+1}) in the ARG model can be expressed as an infinite mixture of gamma product densities [see equation (eq. a.5)].

6.2.2 A time non reversible model

Let us now investigate whether allowing for time non reversible dynamics improves the fit of the data. To this end, we use the alternative parametrisation of B proposed in Section 2.4, that is, $B = B_1 + B_2$, where B_1 is symmetric, whereas B_2 is antisymmetric. The previous time reversible model corresponds to the special, benchmark case where $B = B_1$, and $B_2 = 0$. As is shown in equation (2.20), the equal margin constraint $(D - D')e = 0$ implies a set of *linear* constraints on entries of B_2, once entries of B_1 are given. Thus compared to the model with symmetric B, introducing asymmetry leads to at least

$$n = \frac{J(J+1)}{2} - 2J \tag{6.9}$$

more degrees of freedom. 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.[17]

The number n in (6.9) is positive if and only if

$$\frac{J+1}{2} > 2, \qquad \Leftrightarrow J > 3.$$

In other words, when J is smaller (resp. larger) than 4, introducing symmetry reduces (increases) the number of free parameters compared to the model with symmetric B. Thus to analyse the potential improvement of allowing for asymmetric B, we estimate the model with $J = 4$. Then we can 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, as well as the rest $\frac{J(J+1)}{2} - 2J$ free parameters.

This model will be denoted Model M4 in the following, and J is set to be 4, that is

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

When $J = 4$, we have $\frac{J(J+1)}{2} - 2J = 2$ degrees of freedoms for the matrix B_2. 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 simplified the equal margin constraint $D_2 e = 0$, and hence avoided constrained optimisation of the composite likelihood function.

[17] 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 $k = J$.

6.2.3 Analysis of the volatility of Apple stock

All models are estimated by composite likelihood, on the data of Apple stock (AAPL), traded at the New York Stock Exchange (NYSE). The data is obtained freely from Yahoo Finance, with an observation window spanning from 2000/10/2 up to 2016/9/29. We take Y_t to be the daily adjusted return[18]. As in the literature, we express, without loss of generality, the returns in percentage. The following figure plots the evolution of the daily return.

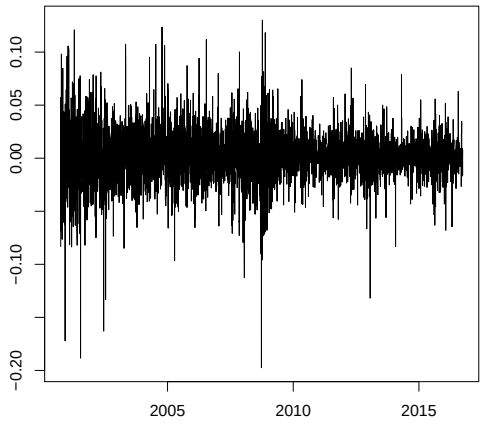

Figure 1: Daily return of the Apple stock between 2000/10/2 and 2016/9/29.

The following figure plots the histogram of the marginal distribution of the Y_t^2, which is a proxy of the volatility $1/X_t$, since $Y_t^2 = \frac{1}{X_t}\epsilon_t^2$.

[18]That is, the return adjusted for the dividend payment.

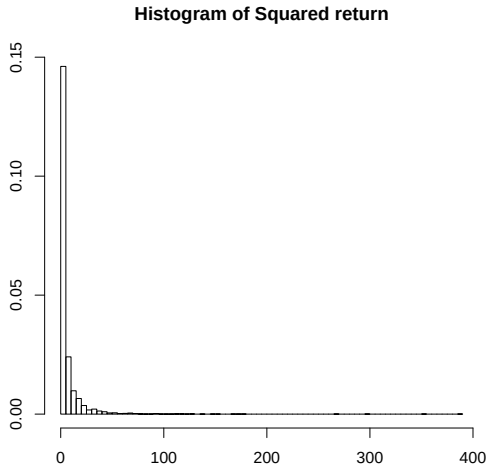

Figure 2: Histogram of Y_t^2. The range of the data has been divided into 100 subintervals of equal length.

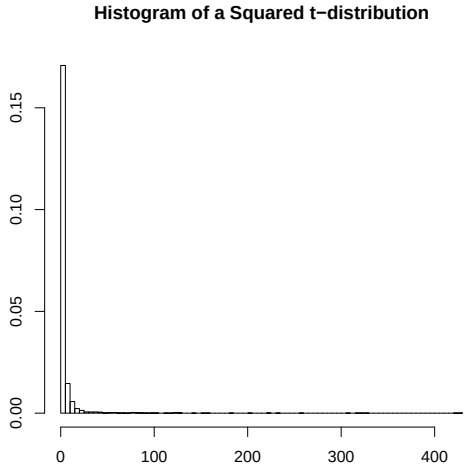

Figure 3: Histogram of a simulated sample of size 4000, following the standard symmetric $t-$distribution with 2.5 degrees of freedom. The range of the data has been divided into 100 subintervals of equal length.

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

Let us now report some summary statistics of Y_t, in particular its four first empirical 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 can further compute the empirical 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. We can also remark 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).

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 usually 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 analogy 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. Thus 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.

The following table reports the parameter estimate of the models.

Model	M1 $J=2$	M1 $J=3$	M2 $J=3$	M2 $J=4$	M4 $J=4$	M4 $J=4$
Residual	Gaussian	Gaussian	skewed $I=1$	skewed $I=1$	skewed $I=1$	skewed $I=2$
Symmetry of B	yes	yes	yes	yes	no	no
c	0.102	0.0803	0.0741	0.0751	0.0503	0.0549
α	2.517	2.84	3.09	2.92	3.14	2.80
b_{01}	-1.57	-3.63	-3.47	-3.72	-2.92	-1.59
b_{02}	6.30	-3.02	5.28	-2.77	1.76	1.67
b_{03}	$--$	-0.979	-0.96	-0.936	-19.1	-2.60
b_{04}	$--$	$--$	$--$	-0.260	11.8	-2.46
b_{10}	$= b_{01}$	$= b_{01}$	$= b_{01}$	$= b_{01}$	-3.03	-1.60
b_{11}	-0.892	3.79	4.56	3.72	-1.73	-1.08
b_{12}	-0.339	-0.376	-3.90	48.9	-9.63	-9.63
b_{13}	$--$	10.02	9.92	10.2	71.4	-19.06
b_{14}	$--$	$--$	$--$	-0.899	108	12.52
b_{20}	$= b_{02}$	$= b_{02}$	$= b_{02}$	$= b_{02}$	1.62	1.66
b_{21}	$= b_{12}$	$= b_{12}$	$= b_{12}$	$= b_{12}$	-30.1	13.8
b_{22}	7.74	6.62	12.9	7.11	15.2	4.17
b_{23}	$--$	-0.940	-0.97	-0.87	140	143
b_{24}	$--$	$--$	$--$	-0.870	160	825
b_{30}	$--$	$= b_{03}$	$= b_{03}$	$= b_{03}$	18.7	1.66
b_{31}	$--$	$= b_{13}$	$= b_{13}$	$= b_{13}$	-73.0	17.3
b_{32}	$--$	$= b_{23}$	$= b_{23}$	$= b_{23}$	-143	-144
b_{33}	$--$	-0.961	-0.955	-0.991	-1.56	-1.55
b_{34}	$--$	$--$	$--$	-0.269	59.0	4101
b_{40}	$--$	$--$	$--$	$= b_{04}$	-10.8	2.66
b_{41}	$--$	$--$	$--$	$= b_{14}$	-108	-12.4
b_{42}	$--$	$--$	$--$	$= b_{24}$	-160	-825
b_{43}	$--$	$--$	$--$	$= b_{34}$	-59.7	-4101
b_{44}	$--$	$--$	$--$	0.401	-0.206	-0.0917
β_1	$--$	$--$	0.017	0.0125	0.0128	-0.0273
β_2	$--$	$--$	$--$	$--$	$--$	-0.00295
ℓ_{CL}	-74356	-74264	-74105	-73861	-73480	-73407
AIC_{CL}	148726	148548	148232	147756	146998	146854

Table 1: Parameter estimate of the four models. 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).

Let us now analyse the Model M3 with $J = 4$ and $I = 2$. Below we report the 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.

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

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

We can remark that the antisymmetric matrix D_2 is non zero, which confirms the time non reversibility of the Apple return data. Nevertheless, the largest entry of $|D_1|$ is $|d_{1,88}| = 1.47$, whereas the largest (in absolute value) entry of $|D_2|$ is $|d_{2,53}| = 0.122$, which is significantly smaller than 1.47.

Next, let us report the theoretical marginal moments predicted by these models, and compare them to the corresponding empirical moments.

Model	M1 $J=2$	M1 $J=3$	M2 $J=3$	M2 $J=4$	M4 $J=4$	Real data
Residual	Gaussian	Gaussian	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	0.0934	0.119	0.106
$\mathbb{E}[y_t^2]$	5.02	5.34	5.64	5.24	6.12	6.02
$\mathbb{E}[y_t^3]$	0	0	0.95	0.50	-0.72	-1.40
$\mathbb{E}[y_t^4]$	205	216	298	318	391	304
$\text{corr}[y_t^2, y_{t+1}^2]$	0.043	0.045	0.054	0.075	0.091	0.121

Table 2: Comparison of some summary statistics with their theoretical values predicted by various models.

From the previous table, we can see that, as expected, increasing J, or adding a parameter capturing the skewness both lead to significant improvement of the fit, in terms of the composite likelihood, or 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. In this case, 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 compare the empirical marginal density with the density given by model M3. 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 unitary mass, then the marginal distribution of Y_t is given by:

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

where range$Y = [\min_t Y_t, \max_t Y_t]$ is the range of 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 table reports the comparison with the kernel density with the model implied density.

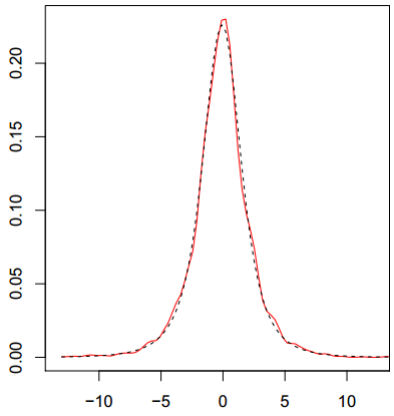

Figure 4: Comparison of the model M3 implied marginal density of $f(y_t)$ with the kernel density estimator. Full line: kernel density estimator; dashed line: model implied density.

We can see that the model M3 provides a very good fit of the marginal density.

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-densities 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.

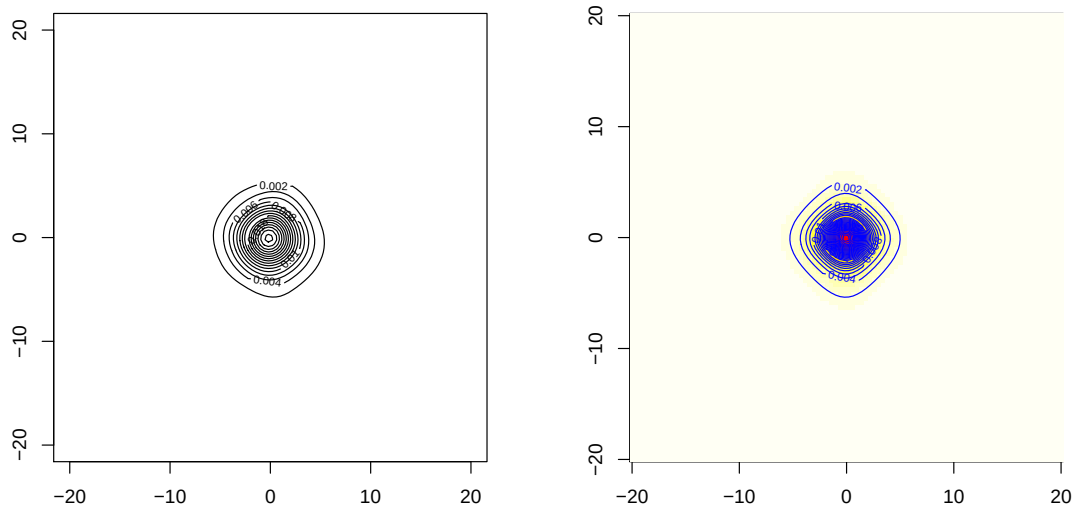

Figure 5: Left panel: Iso-densities of the kernel-based estimate of $f(y_t, y_{t+1})$. Right panel: Iso-densities of the model implied joint density.

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:

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

  thus we have:

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

- the smoothed mean $\mathbb{E}[\frac{1}{X_t}|\underline{y_T}]$. By Appendix 5, 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 6: Filtering, smoothing and forecasting of the conditional variance.

7 Conclusion

The aim of the paper was twofold. First, we have introduced a general class of state-space models. It 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, within the same framework, both time non reversibility and leverage effect.

Finally, a bi-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 Gourieroux 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, due to the positivity constraint of the density, which is automatically satisfied in our model. The flexibility of this new specification will be 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:

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

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 that $U'(x) D e = e' D U(x)$, 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 the identity for $h + 1$. Thus we have proven Property 2.

Appendix 1.3 Proof of Lemma 1

Let us first remark that the function $x \mapsto U'(x) D e$ is continuous and lower bounded by zero, and goes to infinity when $|x|$ goes to infinity[19]. 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 of the marginal distribution, $U'(x) D e$ is null if and only if the marginal

[19] 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.

density $f_0(x) = 0$ for some x, or equivalently, the polynomial in y

$$\sum_{j=0}^{J}\sum_{k=0}^{J} b_{j,k}x^j y^k = 0$$

for almost surely all values of y. This is equivalent to all the coefficients before the terms $1, y, y^2, ..., y^J$ being null, that is: or $B(1, x, x^2, ..., x^J)' = 0$ for a certain x, where matrix $B = (B_{j,k})$.

Appendix 1.4 Proof of Proposition 5

If all the entries of Π is positive, then by Perron-Frobenius theorem, Π has a unique, simple eigenvalue with a right eigenvector of only positive entries, which is e. Hence 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, under the Assumptions of the Proposition, nonnegative. 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 attained. 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 attained. 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. We 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, then process (X_t) is time reversible: $f(x_t, x_{t+1}) = f(x_{t+1}, x_t)$, and 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[20], 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 4. The operator $g \mapsto Ag$ defined by $:Ag(x) = \mathbb{E}[g(X_{t+1})|x_t = x]$ 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.

[20]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.

Conversely, if g is an eigenfunction with eigenvalue λ, that is:

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

Multiplying both sides by $\phi(x)U(x)$ and integrating with respect to x, we get: $\lambda\int\phi(s)g(s)U(s)\mathrm{d}s = \Pi\int\phi(s)g(s)U(s)\mathrm{d}s$. Thus each eigenvalue of process (X_t) is an eigenvalue of Π. □

Thus by this lemma, the eigenfunctions ϕ_j of the operator is 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 7

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

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

which is formula (4.3). 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 (4.4).

Appendix 1.6 Proof of Corollary 4

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

Appendix 1.7 Proof of Proposition 8

Let us first compute the joint distribution:

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

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

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

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:

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

Appendix 1.8 Proof of Lemma 2

We have:

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

Appendix 1.9 Proof of Lemma 3

We have:

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

Appendix 1.10 Proof of Proporition 9

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

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

Appendix 2 Comparison with gamma mixture model

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

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

can be attained from the joint p.d.f. of a gamma mixture, since all the coefficients of polynomial. Whereas it cannot be attained by a "pseudo-mixture". 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 the joint p.d.f. of a pseudo-mixture, but 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, it can be shown that 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 in-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, 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. One direct method is to compute numerically the integral, using, say, the adaptive quadrature approach. This method is 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} seconds	10 seconds
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 now discuss the simulation of trajectories of the state process. In the case where 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 $q_j(x) \propto \phi(x)x^j$. If ϕ is gamma, then q_j becomes the gamma density Gamma$(c, \alpha + S_t)$ and can be simulated directly; if the general case, draws from this distribution can be obtained from a simple acceptance-rejection method.

This approach is not directly applicable when some entries of D are zero or negative. However, in this general case, the process (X_t) can still be simulated using the acceptance-rejection method, inspired by the method of Gallant and Tauchen (1993). Let us first 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. Alternative, if

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

then 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 (2006); Creal (2015)] based state process. The ARG process is the exact time-discretization of 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 $\gamma(\delta + Z_t, c)$, 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 (2006) that the ergodicity condition of the ARG is $\beta c < 1$, with a gamma $\gamma(\delta, \frac{c}{1-\beta c})$ stationary distribution. Thus we have, under stationarity:

$$f_0(x_t) = \frac{x_t^{\delta-1} e^{-\frac{x(1-\beta c)}{c}} c^\delta}{\gamma(\delta)(1-\beta c)^\delta} = \phi(x_t) \frac{e^{\beta x_t}}{(1-\beta c)^\delta} \qquad \text{(eq. a.1)}$$

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

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

where

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

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

$$\begin{aligned} f(x_t, x_{t+1}) &\approx \phi(x_t)\phi(x_{t+1}) \sum_{j=0}^{J} \frac{(\beta c)^j \Gamma(\delta)}{\Gamma(j+1)\Gamma(\delta+j)(1-\beta c)^\delta} x_t^j x_{t+1}^j \\ &= \phi(x_t)\phi(x_{t+1}) U(x_t)' Q_{2J} U(x_{t+1}), \end{aligned} \qquad \text{(eq. a.4)}$$

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

Similarly, the joint distribution of (X_t, X_{t+h}) has the same form as (eq. a.4), except that the correlation coefficient is changed to $(\beta c)^h$ [see Gouriéroux and Jasiak (2006)]:

$$f(x_t, x_{t+1}) \approx \phi(x_t)\phi(x_{t+1}) \sum_{j=0}^{J} \frac{(\beta c)^{jh} \Gamma(\delta)}{\Gamma(j+1)\Gamma(\delta+j)(1-\beta^h c^h)^\delta} x_t^j x_{t+1}^j. \qquad \text{(eq. a.5)}$$

This equation can be used to conduct composite likelihood estimation.

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.

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.

Chen, X. and Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335.

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*.

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*, pages 93–115.

Creal, D. D. (2015). 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*.

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: Journal of the Econometric Society*, pages 363–390.

Gallant, A. R. and Tauchen, G. (1989). Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications. *Econometrica*, pages 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.

Gourieroux, 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. (2006). Autoregressive Gamma Processes. *Journal of Forecasting*, 25(2):129–152.

Gourieroux, C. and Jasiak, J. (2006). Multivariate jacobi process with application to smooth transitions. *Journal of Econometrics*, 131(1):475–505.

Gourieroux, C. and Monfort, A. (2015). Pricing with finite dimensional dependence. *Journal of Econometrics*, 187(2):408–417.

Gourieroux, C., Monfort, A., and Renne, J.-P. (2014). Pricing Default Events: Surprise, Exogeneity and Contagion. *Journal of Econometrics*, 182(2):397–411.

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.

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.

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*, pages 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*, pages 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.

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*, pages 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.

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.