A Flexible State-Space Model with Application to Stochastic Volatility

Christian Gouriéroux* and Yang Lu†

June 15, 2017

Abstract

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

Key words: Endogenous Regime Switching, Polynomial Expansion, Composite Likelihood, Time Irreversibility, Copula.

JEL code: C32, C14

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

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

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

1 Introduction

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

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

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

iii) (panel or time series) count data, such as the numbers of (buy or sell) transactions of a specific market, the numbers of lapses (resp. redemptions) of a life insurer (resp. investment fund), during each time interval, or the annual numbers of accidents of car insurance policyholders [see Lu (2016)]. In these cases the state variable is the stochastic intensity.

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

v) panel data on individuals' dynamic discrete choices [see e.g. 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 data, with potentially irregular features such as missing data. In Finance, many parameter-driven, stochastic volatility models have the further advantages (over GARCH models) of providing closed form formulas for derivatives prices [see e.g. Heston (1993), Gouriéroux and Monfort (2015)] and accounting for volatility risk.

Nevertheless, compared to observation-driven models, the estimation and forecasting of parameter-driven state-space models are often computationally intensive, and simulation based [see e.g. Chib and Winkelmann (2001)]. Our paper introduces a family of models that *i*) is sufficiently flexible to fit a wide range of data, and *ii*) leads to simple, simulation-free estimation and forecasting procedures. We specify the dynamics of the Markov state process via a polynomial expansion based joint density, which can approximate any univariate Markov dynamics arbitrarily well. Moreover the dynamics has a latent endogenous switching regime interpretation, and the resulting state-space model allows for simple recursive formulas for prediction and filtering, which is faster than simulation based methods such as the particle filter. Under some further constraints, the model can be estimated by maximum composite likelihood, whose computational cost is extremely low compared to simulation-based methods.

Our methodology can be applied to both time series and panel data. In particular, in some panel data applications, the cross-sectional dimension is very large. This is for instance the case for automobile insurance [see item *iv*) in the first paragraph], for which the number of policy-holders can attain 100,000 or even larger. This makes the standard simulation-based approach computationally cumbersome for estimation compared to the composite likelihood approach on which our model relies. However, in the application section, our illustration only concerns the stochastic volatility (SV) of asset returns, and we contribute to this literature by allowing for time irreversibility of the volatility process.

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

2 The model

2.1 The state-space representation

We consider a state-space model in which X_t is a one-dimensional state variable with domain $\mathcal{X}$, whereas Y_t is the observable variable with domain $\mathcal{Y}$. Depending on the application, $\mathcal{X}$ can be the whole real line $\mathbb{R}$, the positive half-line $\mathbb{R}_{>0}$, or a bounded interval such as $]0,1[$, whereas $\mathcal{Y}$ can be uni- or multivariate, continuous or discrete. 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 past $\underline{Y_{t-1}}$ and the whole trajectory of (X_t) depends only on $\underline{Y_{t-1}}$ and the current value of the state variable X_t.

Such a model is usually difficult to estimate, except *i*) if process (X_t, Y_t) is jointly Gaussian, where 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); Hamilton (1989)]. Besides these two cases, the estimation of the model involves simulation techniques. Let us now focus on the state process and introduce a model in which (X_t) is Markov with a flexible polynomial expansion based joint density. 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 benchmark density whose integer moments are all finite, J is an integer, the matrix of coefficients $b_{j,k}$ is real and the normalisation constant M is equal to:

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

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

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

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

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. In the applications, the benchmark density ϕ is specified parametrically, and some convenient parametric forms, depending on the domain $\mathcal{X}$ and the conditional density $l(y_t|\underline{y_{t-1}}, x_t)$, will be discussed in Section 4.3. The second term is a squared polynomial, which ensures the positivity of the density function. Similar univariate densities have been introduced in the micro literature by Gallant and Nychka (1987), who call it a semi-nonparametric (SNP) density. The background of (2.3) as the polynomial expansion in an L^2 space, as well as its link with the SNP approach, are discussed in Section 5. The approach via the joint density is also encountered in the copula-based time series models [see Chen and Fan (2006)], and the online appendix of this paper provides a comparison of these two approaches.

If the rank of B is 1, say $B = \beta\beta'$, where β is a $J+1$ dimensional vector, then we have $f(x_t, x_{t+1}) = \frac{1}{M}\left[\phi(x_t)\sum_{i=0}^{J}\beta_i x_t^i\right]\left[\phi(x_{t+1})\sum_{i=0}^{J}\beta_i x_{t+1}^i\right]$, which corresponds to an i.i.d. sequence with marginal density $f(x_t) = \frac{1}{\sqrt{M}}\phi(x_t)\left[\sum_{i=0}^{J}\beta_i x_t^i\right]^2$. Finally, when entries of D are all nonnegative and (x_t) is positive, the joint density is a mixture distribution. We refer to Section 2 of the online Appendix for a detailed comparison between our model and the standard mixture model in terms of the goodness of fit of the joint distribution of (X_t, X_{t+1}), in the special case of a gamma benchmark density.

Using D, the density in equation (2.3) can be more conveniently rewritten in the matrix product form:

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

where vector function U is defined by:

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

and satisfies: $\int \phi(x)U(x)\mathrm{d}x = e$. This definition assumes implicitly that all moments $\int \phi(s)s^j\mathrm{d}s$ are non zero. This can be achieved by an appropriate choice of the state variable[1], for instance with values between 0 and $+\infty$, or 0 and 1, and of the benchmark density ϕ.

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

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

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

Then the marginal distribution is:

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

and it is nonnegative by construction.

Proof. See Appendix 1.1. □

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

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

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 are largely identical.

This condition depends both on the benchmark density ϕ and B. A simpler, equivalent characterization of this condition will be derived in Section 2.4.

2.2 Ergodicity of the state process

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

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

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

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

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

where 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. □

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

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

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

Proof. See Appendix 1.3. □

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

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

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

Proof. See Appendix 1.4. □

Roughly speaking, the first property says that Π is row stochastic, if its entries are all nonnegative[2]. 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, when the latter exists.

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

Corollary 2.

$$f_h(x_{t+h}|x_t) - f_0(x_{t+h}) = \frac{\phi(x_{t+h})U'(x_t)D}{U(x_t)'De}\Big(\Pi^{h-1} - \frac{ee'D}{e'De}\Big)U(x_{t+h}). \tag{2.12}$$

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

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

The ergodicity is necessary in order for likelihood type estimators to be consistent and asymptotically normal. The proof of Proposition 3 is omitted, as it is based on the same arguments as for a finite state Markov chain [see e.g. Seneta (2006), Chapter 1], even if the entries of Π are not necessarily nonnegative.

Moreover, by equation (2.12), we see that the second largest (in modulus) eigenvalue of Π has a large impact on the serial correlation of the process (X_t). If this eigenvalue is large (resp. small), then the conditional distribution $f(x_{t+h}|x_t)$ converges slowly (resp. quickly) to the stationary distribution $f_0(x_{t+h})$.

In general, the eigenvalues of Π should be computed numerically to check the conditions in Proposition 3. Nevertheless, in Section 1 of the online Appendix, we discuss special cases where this condition is automatically satisfied.

2.3 Interpretation in terms of latent regimes

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

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

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 the entries of D are nonnegative. This latter also ensures that the entries of Π are nonnegative.

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

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

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

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

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

To summarize, under Assumptions 1 and 2, we have the following causal scheme[3]:

$$\ldots S_{t-1} \to X_t \to S_t \to X_{t+1} \to S_{t+1} \ldots . \tag{2.13}$$

In this chain, each variable depends on all the variables on its LHS only via its nearest left neighbour. Moreover, the latent process (S_t) is also a Markov chain with respect to its own history, and the transition matrix of this embedded Markov chain is $\Pi = \int_0^\infty \Big[U(x)\phi(x)\Big]\Big[\frac{U'(x)D}{U'(x)De}\Big]\mathrm{d}x$. Indeed, we have:

$$\Pi_{i,j} := \mathbb{P}[S_{t+1} = j \mid S_t = i] = \int \mathbb{P}[S_{t+1} = j|S_t = i, x_{t+1}]l(x_{t+1}|S_t = i)\mathrm{d}x_{t+1}. \tag{2.14}$$

By Corollary 1, when the embedded chain (S_t) exists, the stationary distribution of this Markov chain is $\frac{De}{e'De}$.

Let us now discuss the interpretation of matrix D. Since the joint density $f(x_t, x_{t+1}) = \phi(x_t)\phi(x_{t+1})U'(x_t)\frac{D}{e'De}U(x_{t+1})$ has a mixture interpretation, each entry (i, j) of the re-normalized matrix $\frac{D}{e'De}$ is, roughly speaking, the probability that "X_t and X_{t+1} belong to the i (resp. j-th) regime. Nevertheless, in general we do not have the identity:

$$\left(\frac{D}{e'De}\right)_{i,j} \neq \mathbb{P}[S_t = i, S_{t+1} = j]. \tag{2.15}$$

[3]Such a causal scheme is introduced by Pitt et al. (2002), who consider several parametric, reversible examples, such as the Autoregressive Gamma.

This latter matrix is given by $\mathrm{Diag}(\frac{De}{e'De})\Pi$, where $\mathrm{Diag}(\frac{De}{e'De})$ is the diagonal matrix with diagonal vector $\frac{De}{e'De}$, that is the stationary distribution of the Markov chain (S_t). It has been termed the "joint probability matrix" by McCausland (2007), who uses it to study the time-reversibility of a Markov chain. The inequality (2.15) can be explained by the fact that the Markov chain (S_t) in our model is endogenous, thus the joint stationary distribution of (S_t, S_{t+1}) has a complicated relationship with that of (X_t, X_{t+1}). Nevertheless, it will be shown in the next subsection that matrix $\frac{D}{e'De}$ plays a similar role in characterizing the time reversibility of our Markov process (X_t).

The switching regime representation also explains the tractability of the $h-$step-ahead conditional density (2.10):

$$f_h(x_{t+h}|x_t) = \underbrace{\frac{U'(x_t)D}{U'(x_t)De}}_{\text{conditional probabilities of } S_t \text{ given } X_t} \Pi^{h-1} \underbrace{\phi(x_{t+h})U(x_{t+h})}_{\text{conditional density of } X_{t+h} \text{ given } S_{t+h-1}}, \quad (2.16)$$

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 discrete ones $S_t, ..., S_{t+h-1}$.

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

The causality scheme in equation (2.13) is different from the usual Markov switching model, which has the causal chain:

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

Indeed, in our model the switching regime S_t may not exist, since D can have negative entries. Secondly, even if it exists, its transition probabilities are endogenous, since $l(S_{t+1}|S_t, Y_t)$ depends on the observable stochastic variable Y_t. Thus our paper contributes to the literature on endogenous switching regimes [see e.g. Kim et al. (2008); Chang et al. (2017)].

2.4 Time reversibility

Roughly speaking, process (X_t) is time reversible if the reversely ordered process (X_{-t}) has the same dynamics as (X_t). It is a simplifying assumption satisfied by many models, including normal ARMA processes, time discretized diffusion processes such as the autoregressive gamma process[4], as well as Gaussian and Archimedean copula based time series. However, this property is usually not predicted by theoretical models [see e.g. Maskin and Tirole (1988)]. Moreover, the empirical literature on financial and economic data [see e.g. Ramsey and Rothman (1996); Chen et al. (2000); Darolles et al. (2004); Beare and Seo (2014); Racine and Maasoumi (2007); McCausland (2007)] usually reject this assumption. However, once the reversibility rejected, very few alternatives have been proposed to account for the irreversibility. Let us now discuss how the irreversibility of the state process (X_t) can be addressed in our framework. First, let us look at the condition for reversibility.

Under the Markov assumption of (X_t), the time reversibility is equivalent to the symmetry of the joint distribution $f(x_t, x_{t+1})$ in the two arguments, that is,

$$f(x_t, x_{t+1}) = \phi(x_t)\phi(x_{t+1})\frac{U'(x_t)DU(x_{t+1})}{e'De} = \phi(x_t)\phi(x_{t+1})\frac{U'(x_{t+1})DU(x_t)}{e'De} = f(x_{t+1}, x_t), \tag{2.17}$$

for all values of x_t, x_{t+1}, or equivalently matrix D is symmetric. How can we characterize the symmetry of D in terms of B? We have the following proposition:

Proposition 5. *D is symmetric if and only if B is symmetric, or is antisymmetric*[5].

Proof. See Appendix 1.5. □

We can also remark that when D is symmetric, condition $(D - D')e = 0$ is automatically satisfied. This explains the simplification, when the reversibility is assumed. Let us now look for models of (X_t) with a irreversible dynamics. It suffices to show that the condition $(D-D')e = 0$ can also be satisfied for a non reversible process (X_t). To derive the equivalent condition of the latter condition, let us introduce the (unique) decomposition of matrix B into the sum of a symmetric matrix $B_1 = \frac{1}{2}(B + B')$, and an antisymmetric matrix $B_2 = \frac{1}{2}(B - B')$. This

[4]The autoregressive gamma process is the time discretized Cox-Ingersol-Ross process, and takes value in $\mathbb{R}_{\geq 0}$, see Section 5 of the online Appendix for details.

[5]Also called skew-symmetric.

decomposition of B leads to the corresponding decomposition of D:

$$\begin{aligned}
d_{j,k} &= \mu_j\mu_k \sum_{\substack{j_1+j_2=j\\ 0\le j_1,j_2\le J}} \sum_{\substack{k_1+k_2=k\\ 0\le k_1,k_2\le J}} b_{j_1,k_1}b_{j_2,k_2}\\
&= \mu_j\mu_k \sum_{j_1+j_2=j}\sum_{k_1+k_2=k} (b_{1,j_1,k_1}+b_{2,j_1,k_1})(b_{1,j_2,k_2}+b_{2,j_2,k_2})\\
&= \underbrace{\mu_j\mu_k \sum_{j_1+j_2=j}\sum_{k_1+k_2=k} (b_{1,j_1,k_1}b_{1,j_2,k_2}+b_{2,j_1,k_1}b_{2,j_2,k_2})}_{:=d_{1,j,k},\ \text{symmetric}} + \underbrace{2\mu_j\mu_k \sum_{j_1+j_2=j}\sum_{k_1+k_2=k} b_{1,j_1,k_1}b_{2,j_2,k_2}}_{:=d_{2,j,k},\ \text{antisymmetric}},
\end{aligned} \tag{2.18}$$

where the term $d_{1,j,k}$ (resp. $d_{2,j,k}$) in (2.18) is symmetric (resp. antisymmetric) in j, k, since B_1 (resp. B_2) is symmetric (resp. antisymmetric). As a consequence, the equal margin constraint $(D - D')e = 0$ is equivalent to $D_2e = 0$, that is, the sum of each row of D_2 is zero. This is an orthogonality condition between B_1 and B_2.

Proposition 6. *The condition $(D - D')e = 0$ is satisfied if and only if the symmetric and antisymmetric components of matrix B are orthogonal.*

These orthogonality conditions are easy to check in practice. Compared to the model with symmetric B, say, introducing asymmetry leads to:

$$n = \frac{J(J+1)}{2} - 2J, \tag{2.19}$$

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

3 Examples

3.1 Case $J = 1$

Let us study the case where $J = 1$, and

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

The equal margin condition. First, let us discuss the implications of the condition $(D - D')e = 0$ on matrix $B = (b_{i,j})_{0\leq i,j\leq J}$. Matrix D is given by: $D = \begin{bmatrix} b_{00}^2 & 2b_{00}b_{01}\mu_1 & b_{01}^2\mu_2 \\ 2b_{00}b_{10}\mu_1 & 2b_{00}b_{11}\mu_1^2 + 2b_{01}b_{10}\mu_1^2 & 2b_{01}b_{11}\mu_1\mu_2 \\ b_{10}^2\mu_2 & 2b_{10}b_{11}\mu_1\mu_2 & b_{11}^2\mu_2^2 \end{bmatrix}$.
The equal margin condition $(D' - D)e = 0$ is equivalent to:

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

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

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

Note that, equation (3.3) can be obtained by summing (3.1) and (3.2).[6] Under the assumption $\mu_1\mu_2 \neq 0$, we have two cases. Either *i*) $b_{01} = b_{10}$, which means that B is symmetric, or *ii*),

$$b_{01} + b_{10} + 2b_{11}\mu_1 = 0, \qquad \text{and } b_{00} - b_{11}\mu_2 = 0. \tag{3.4}$$

Thus, in the case $J = 1$, either B is symmetric, and B_1 can be any symmetric matrix, or B is anti-symmetric, and $B_2 = \frac{b_{01}-b_{10}}{2}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ can be any antisymmetric matrix, but the entries of $B_1 = \begin{bmatrix} b_{00} & \frac{b_{01}+b_{10}}{2} \\ \frac{b_{01}+b_{10}}{2} & b_{11} \end{bmatrix}$ should satisfy the constraints (3.4). This is expected, since the number of extra degree of freedom $n = \frac{J(J+1)}{2} - 2J$ is negative when $J = 1$ [see (2.19)].

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

$$f(x_t, x_{t+1}) = \frac{1}{M}\phi(x_t)\phi(x_{t+1})\Big(1 + b_{01}(x_t + x_{t+1})\Big)^2. \tag{3.5}$$

Let us now derive the form of the matrix Π in model (3.5). 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

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

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

The hidden regimes. Under model (3.5), the conditional probabilities of S_t given X_t are the components of vector:

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

3.2 A stochastic volatility model

A benchmark one-dimensional stochastic volatility model defines the asset return y_t as:

$$Y_t = \sigma(X_t)\epsilon_t, \tag{3.6}$$

where the ϵ_t's are $IIN(0,1)$ and independent of the nonnegative volatility process $\sigma_t = \sigma(X_t)$. In the existing SV literature developped for derivative pricing, 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 (2017)]. In model (3.6), the gamma type volatility model is obtained when we take $\sigma(X_t) = X_t$, where the joint distribution of (X_t, X_{t+1}) follows the polynomial expansion form [see (2.3)], with a gamma benchmark density $\phi(x) = \frac{x^{\alpha-1}\exp(-x/c)}{\Gamma(\alpha)c^{\alpha}}$.[7]

Under the joint pdf specification (2.3) with gamma benchmark density, the conditional density of X_{t+1} given $S_t = j$, that is $\frac{1}{\mu_j}\phi(x)x^j$, is still a gamma density with the same scale parameter c and different shape parameters $\alpha + j$, $j = 0 =, ..., 2J$, respectively. They correspond to different levels of risk: the larger S_t, the larger the conditional expectation $\mathbb{E}[X_{t+1}|S_t]$.

4 Forecasting and filtering

Let us now derive the predictive formulas. This includes: *i*) the forecasting, that is $l(y_{T+h}|\underline{Y_T})$; *ii*) the filtering, that is, $l(x_t|\underline{Y_t})$, and the smoothing, that is, $l(x_t|\underline{Y_T})$.

4.1 Forecasting

Proposition 7. *The conditional density of Y_t given the past is:*

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

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

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

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

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

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

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

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

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

Proof. See Appendix 1.6. □

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

When (X_t) has the embedded switching regime interpretation, we have the following causal chain:

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

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

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

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

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

Given S_{t-1}, the future state variable S_t, and the observable variable Y_t are linked through X_t. Thus the transition of the latent state variable depends on the current value of Y_t. The matrix

$\Pi(\underline{y_t})$ is not a stochastic matrix. Indeed, the sums of the entries of each of its row are:

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

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

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

Entries of matrix $\Pi(\underline{y_t})$ can be computed numerically by using an adaptive quadrature method. This method is much faster than the standard Monte-Carlo simulation (see Section 3 of the online Appendix). Nevertheless, the latter approach is interesting to discuss, as it is, roughly speaking, the analogue of the updating of the population in a particle filter [see e.g. Pitt and Shephard (1999)]. Indeed, in the latter approach, a large number of trajectories of (X_t) are generated in order to provide an approximation of the conditional distribution $l\Big((x_t)_{t=1,\dots,T}|(y_t)_{t=1,\dots,T}\Big)$. This has a computational cost that is similar to the numerical integral of the entries of $\Pi(\underline{y_t})$ via Monte-Carlo simulation.

Finally, our state-space model is a generalization of the model of Creal (2017), who specify (X_t) as an autoregressive gamma (ARG) process (see Section 5 of the online Appendix). The ARG process is also associated with an embedded latent regime variable (Z_t), which is discrete (but infinite) valued. Then Creal proposes to approximate its dynamics by a Markov chain with a large number of states. More precisely, Creal (2017) proposed to use a finite chain with 3000 states to approximate the dynamics of process (S_t). This allows the authors to derive similar forecasting formula. Our model generalizes Creal's result in several aspects. First, the dynamics of our state process is flexible and allows for time irreversibility. Second, our filtering algorithm involves a much smaller number of states for (S_t).

Proposition 7 provides the one-step-ahead nonlinear forecasting formula. Due to the feedback effect, that is the dependence of $l(y_t|\underline{y_{t-1}}, x_t)$ in the lagged observations $\underline{y_{t-1}}$, the longer horizon forecast formula $l(y_{t+h}|\underline{y_{t-1}})$ has no simple expression[8]. Nevertheless, simulation of the trajectories of (X_t) can be conducted rather easily using an acceptance-rejection method (see Section 4 of the online Appendix).

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

4.2 Filtering and smoothing

The simple forecasting formula in Proposition 7 is associated with a simple expression for the predictive density of X_t given the past observations. We have:

Corollary 3.

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

Proof. See Appendix 1.6. □

Similarly, the filtering density is:

Proposition 8. *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 1.7. □

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

Proposition 9. *The smoothing density is:*

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

where vector Q_t is defined backward by:

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

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

Proof. See Appendix 1.8. □

Due to the endogenous switching regime interpretation, the smoothing equation is similar to the forward-backward smoothing algorithm for hidden Markov models [see e.g. Scott (2002)]. Indeed, we have:

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

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

4.3 Model without feedback

Let us now consider the model without feedback, that is:

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

In particular, we have also $g(y_t|\underline{y_{t-1}}) = g(y_t) = \int U(x_t) l(y_t|x_t)\phi(x_t)\mathrm{d}x_t$. This assumption leads to two simplifications. First, the matrix $\Pi(\underline{y_t})$ depends on y_t only, up to a multiplicative constant $\frac{1}{l(y_t|\underline{y_{t-1}})}$. Thus, when the domain $\mathcal{Y}$ of Y_t is finite, the recursive forecasting algorithm involves only a finite number of numerical integrals. Secondly, the prediction of the observable variable y_{T+h} at any horizon h has an explicit formula:

Proposition 10. *In the model without feedback, the $h-$step-ahead predictive density is:*

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

Proof. See Appendix 1.9. □

The predictive distributions are simple so long as function $g(y_t)$ is simple. This is rendered possible by carefully choosing the form of ϕ and $l(y_t|x_t)$. For instance, g has a closed form when:

- ϕ is a gamma density, and $l(y_t|x_t)$ 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 (2017) for an exhaustive list].
- ϕ is the uniform density, or a beta density, and $l(y_t|x_t)$ is binomial $Bin(n, x_t)$.

5 A non-parametric approach

Our specification (2.3) of the state process is, in some sense, non-parametric. In this section, we explain how it approximates the dynamics of any given Markov process, and compare to some competing approaches of the literature.

5.1 Polynomial decomposition of a square rooted density

Let us assume that[9] 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 or-

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

thonormal polynomial basis for the L^2 space associated with the measure $\phi(x_t)\phi(x_{t+1})\mathrm{d}x_t\mathrm{d}x_{t+1}$. Since $f(x_t, x_{t+1})$ integrates to unity, the ratio function $\sqrt{\frac{f(x_t,x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}$ belongs to this L^2space, with unitary norm:

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

Thus we get the orthonormal decomposition:

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

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

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

The infinite sum in (5.2) converges in the sense of L^2, and $\sum_{i,j=0}^{\infty} a_{i,j}^2 = 1$. Then we approximate $f(x_t, x_{t+1})$ by truncating the RHS of equation (5.2), and taking the square:

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

In other words, the RHS is the orthogonal projection of $\sqrt{\frac{f(x_t,x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}$ onto the linear space generated by $\{P_i(x_t)P_j(x_{t+1}), 0 \leq i, j \leq J\}$. This function is not yet a density, thus we normalize it to get:

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

with $M_J = \sum_{i,j=0}^{J} a_{i,j}^2$. This approximating density can be as accurate as possible, when J is large. More precisely we have:

Proposition 11. *The sequence of densities f_J approximates f arbitrarily well in terms of the Hellinger distance*[10]*, when J goes to infinity:*

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

Proof. See Appendix 1.6. □

This convergence result holds for any choice of benchmark density. It also suggests joint

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

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

5.2 Comparison with the polynomial decomposition of a density

Let us compare our specification of (X_t) with the direct approximation of (univariate) densities by means of polynomial expansion [see e.g. Aït-Sahalia (2002), Filipović et al. (2013)]. This literature decomposes, under integrability conditions, the conditional density $f(x_{t+1}|x_t)$ into:

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

where ϕ is the benchmark density (which is usually Gaussian in the Finance literature), and P_i are the corresponding orthonormal polynomials. Then this literature truncates the RHS up to a finite order J to obtain an approximation f_J of f. The major drawback of this approach is that, unlike in our model, the truncated version of (5.6) is not a proper density, since it is not nonnegative. This leads to several difficulties. First, it is not possible to evaluate the accuracy of the approximation, since neither the Kullback distance, nor the Hellinger distance between f_J and f can be defined. Second, negative probabilities typically lead to arbitrage opportunities when it comes to derivative pricing. Our model differs from this literature in three aspects. First, our method guarantees the positivity of the densities. Second, the decomposition (5.6) requires a rather strong integrability condition[11] which has been shown to be sometimes violated [see Aït-Sahalia (2002)], whereas our decomposition of $\sqrt{\frac{f(x_t,x_{t+1})}{\phi(x_t)\phi(x_{t+1})}}$ does not require such extra condition. Finally, instead of specifying the conditional distribution of X_{t+1} given X_t,[12] we specify the joint distribution of (X_t, X_{t+1}). This approach facilitates the derivation of stationarity conditions and of the stationary density.

6 Illustration

Let us now discuss the implementation of this type of models. We first discuss the estimation approaches, and then consider a stochastic volatility model.

[11]This condition is $\int \frac{f^2(x_{t+1}|x_t)}{\phi(x_{t+1})} \mathrm{d}x_{t+1} < \infty$.

[12]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.

6.1 Estimation methods

6.1.1 Maximum likelihood approach

The log-likelihood function has the form:

$$\log \ell(\theta) = \sum_{t=1}^{T} \log l(y_t|\underline{y_{t-1}}, \theta) = \sum_{t=1}^{T} \log \int_{0}^{\infty} l(y_t|\underline{y_{t-1}}, x_t, \theta) l(x_t|\underline{y_{t-1}}, \theta) \mathrm{d}x_t, \tag{6.1}$$

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

6.1.2 Maximum composite likelihood approach

Let us now introduce the maximum composite likelihood estimation (MCLE) method [see e.g. Varin and Vidoni (2008), Gouriéroux and Monfort (2017), Gouriéroux et al. (2017) for reviews] that is particularly suited for our model. Although it is (slightly) less efficient[13] than MLE, its computational cost is extremely low, and is similar to that of the Generalized Method of Moments (GMM). The composite likelihood is based on the following closed form expression of the joint distribution of (Y_t, Y_{t+h}):

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

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

Proof. See Appendix 1.10. □

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

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

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

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

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

as h is smaller than an integer m. Maximizing this function leads to the MCLE:

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

Function $\ell_{CL}(\theta)$ is easy to compute, when function g allows for closed form expression (see the discussion below Proposition 10). Varin and Vidoni (2008) show that under mild conditions, the estimator (6.3) is asymptotically consistent and normally distributed. It is however typically not efficient [see e.g. Gouriéroux and Monfort (2017) for a discussion on similar more efficient methods]. Nevertheless, Varin and Vidoni (2008) show, via a simulation experiment, that the efficiency loss with respect to, say, the MLE is rather small, compared to usual GMM [see e.g. Andersen and Sørensen (1996)], which requires a similar computational cost as the MCLE method.

Finally, let us discuss the choice of m. Roughly speaking, if m is too small, the amount of information contained in the composite likelihood function is reduced, and hence the efficiency of the MCLE would be low. At the same time, the computational effort required is proportional to m; thus in the application, we set $m = 5$. As for the weights w_h, we set them to be $w_h = 0.9^h, \forall h$. in order to reflect the high persistence of financial returns.

6.2 A stochastic volatility application

6.2.1 The models

We consider the following competing models:

i) **Model M1**

Let us first consider the model M1:

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

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

Since we can multiply $\sqrt{x_t}$ by a constant and accordingly multiply ϵ_t by the same constant, for identification purpose we assume without loss of generality: $\mathbb{E}[\epsilon_t^2] = 1$. Secondly, in the joint p.d.f. of equation (2.3), we can multiple all the coefficients $b_{i,j}$ by a same constant. Therefore, we set $b_{0,0} = 1$.[14]

[14]This normalization constraint implies that $b_{0,0} \neq 0$. This implicit assumption is motivated by the polynomial

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

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

since $\mu_j = \frac{\Gamma(\alpha+j)}{\Gamma(\alpha)}$. Each component function $g_j(\cdot)$ is the density of a re-scaled Student's $t-$distribution, which has been widely used in Finance to account for the heavy-tails of asset returns [see e.g. Fernández and Steel (1998)]. 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 5, 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 M2.**

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

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

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

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

where $\nu_k = \int_{-\infty}^{\infty}\psi(\epsilon)\epsilon^k\mathrm{d}\epsilon$ is the $k-$th moment of the standard normal distribution; we have, $\nu_{2k+1} = 0$, and $\nu_{2k} = \frac{(2k)!}{2^k k!}$. Thus the density function $h(\epsilon)$ has a similar form as the density of

expansion formula developed in Section 5.

(X_t, X_{t+1}), obtained by squaring and renormalizing the polynomial expansion of a square rooted univariate density, with respect to the benchmark density ψ. It is a new, flexible alternative to the parametric skewed distributions [see e.g. Fernández and Steel (1998)]. For instance, when $I = 1$, the density of the error is:

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

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

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

Thus ϵ has negative skewness if $\beta_1 \in] - \frac{1}{\sqrt{3}}, 0[$. Similarly, when $I = 2$, we get:

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

Under the density specification (6.5), the distribution of ϵ_t is no longer symmetric with respect to 0. Therefore, positive and negative past returns have different impacts on the forecast of the future volatility.

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

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

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

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

or for the joint moments:

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

They are immediately deduced from the corresponding moments in each regime, whose expressions are derived in Section 6 of the online Appendix. The existence of these moments depend on the shape parameter α of the benchmark density ϕ. The moments of order p exists if $\alpha > \frac{p}{2}$.

What matters for the existence of moments is the behavior at zero of the conditional distribution $l(X_t|S_t)$, which is gamma with shape parameter $S_t + \alpha$. In particular, under regime $S_t = 0$, the density of $l(X_t|S_t)$ has the heaviest tail at zero.

iii) **Model M3.**

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

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

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

iv) **Model M4**

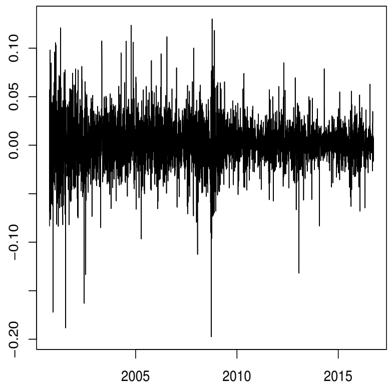
The three previous models assume a time reversible dynamics for process (X_t). The following model M4 allows for time irreversibility. This model is a generalisation of M2, with non symmetric matrix B. We use the parametrisation of B as $B = B_1 + B_2$, where B_1 is symmetric, and B_2 antisymmetric (see Section 2.4). The time reversible model M1 corresponds to the special case where $B_2 = 0$. As shown in equation (2.18), the constraint $(D - D')e = 0$ for equal margins implies a set of linear constraints on B_2, once B_1 is given. To analyse the potential improvement of allowing for partially asymmetric B, we estimate the model with $J = 4$. Then we use the orthogonal condition $D_2 e = 0$ to express $2J$ entries of B_2 above the diagonal as functions of the entries of B_1, and of the additional $\frac{J(J+1)}{2} - 2J = 2$ entries of B_2, which we choose to be $b_{2,1,0}$ and $b_{2,2,0}$. Thus we have a linear system with 8 unknowns and $2J = 8$ equations. Therefore, the set of parameters of the model M4, with $J = 4$ and $I = 2$, say, is:

$$\theta = (c, \alpha, (b_{1,j,k})_{0\leq j\leq k\leq 4}, b_{2,1,0}, b_{2,2,0}, \beta_1, \beta_2).$$

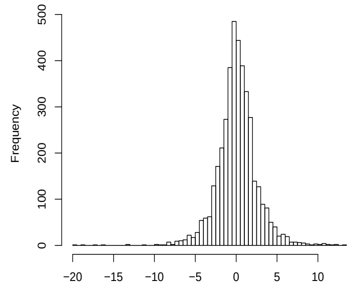
6.2.2 The volatility of Apple stock return

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

All models are estimated by maximum composite likelihood on daily return data of the Apple stock (AAPL). The data are downloaded from Yahoo Finance, observed from 2000/10/2 up to 2016/9/29. Y_t is the (non annualized) daily adjusted return, that is the return adjusted for the dividend payment. We express the returns in percentage. The left panel of Figure 1 provides the evolution of the daily return, whereas the right panel provides the histogram of its marginal distribution.



(a) Daily return of the Apple stock between 2000/10/2 and 2016/9/29.



(b) Histogram of Y_t.

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

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

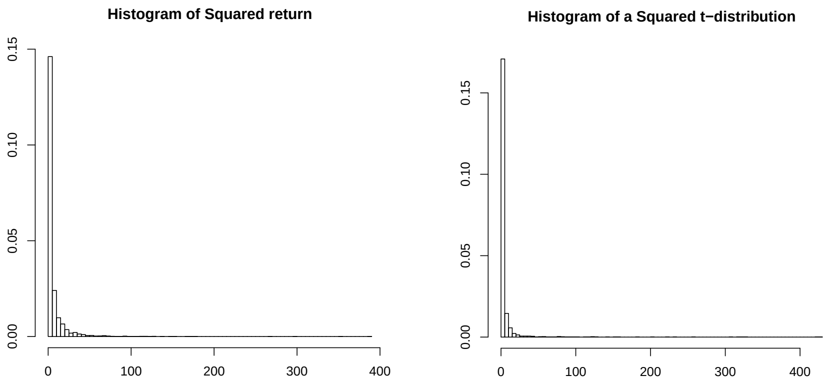


Figure 2: Left panel: Histogram of Y_t^2. Right panel: Histogram of a simulated sample of size 4000, following the standard symmetric $t-$distribution with 2.5 degrees of freedom

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

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

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

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

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

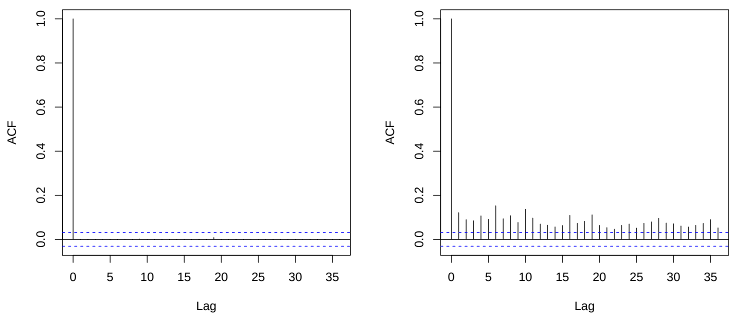


Figure 3: Historical autocorrelation function of $1/Y_t^2$ (left panel) and of Y_t^2, on the right panel

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

In order to compare different models, we introduce the concept of composite Akaike Information Criterion (AIC_{CL}), that is the analogue of the standard likelihood-based AIC [see Varin and Vidoni (2005) for details]:

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

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

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

Model	M1 $J=2$	M1 $J=3$	M2 $J=3$	M2 $J=4$	M3 $--$	M4 $J=4$	M4 $J=4$
Error Density	Gaussian $I=0$	Gaussian $I=0$	skewed $I=1$	skewed $I=1$	skewed $I=1$	skewed $I=1$	skewed $I=2$
Symmetry of (X_t)	yes	yes	yes	yes	yes	no	no
c	0.102(*)	0.0803(*)	0.0741 (*)	0.0751(*)	0.201(*)	0.0503(*)	0.0549(*)
α	2.51(*)	2.84(*)	3.09(*)	2.92(*)	1.73(*)	3.14(*)	2.80 (*)
ρ	$--$	$--$	$--$	$--$	0.085(*)	$--$	$--$
$b_{1,01}$	-1.57(*)	-3.63(*)	-3.47(*)	-3.72(*)	$--$	-2.97(*)	-1.59(*)
$b_{1,02}$	6.30(*)	-3.02(*)	5.28(*)	-2.77(*)	$--$	1.69(*)	1.67(*)
$b_{1,03}$	$--$	-0.979(*)	-0.96(*)	-0.93(*)	$--$	-0.2	-0.47(*)
$b_{1,04}$	$--$	$--$	$--$	-0.26	$--$	0.50(*)	0.102
$b_{1,11}$	-0.892(*)	3.79 (*)	4.56(*)	3.72(*)	$--$	-1.73(*)	-1.08(*)
$b_{1,12}$	-0.339(*)	-0.376(*)	-3.90(*)	48.9(*)	$--$	-19.8(*)	2.08
$b_{1,13}$	$--$	10.02(*)	9.92(*)	10.2(*)	$--$	-0.8(*)	-1.15(*)
$b_{1,14}$	$--$	$--$	$--$	-0.89(*)	$--$	0.01	0.061
$b_{1,22}$	7.74(*)	6.62(*)	12.9(*)	7.11(*)	$--$	15.2(*)	4.17(*)
$b_{1,23}$	$--$	-0.940 (*)	-0.97 (*)	-0.87(*)	$--$	-1.5(*)	0.50(*)
$b_{1,24}$	$--$	$--$	$--$	-0.87(*)	$--$	0.01	0.01
$b_{1,33}$	$--$	-0.961(*)	-0.95(*)	-0.99(*)	$--$	-1.56(*)	-1.55(*)
$b_{1,34}$	$--$	$--$	$--$	-0.26	$--$	-0.35(*)	0.01
$b_{1,44}$	$--$	$--$	$--$	0.40(*)	$--$	-0.206(*)	-0.091(*)
$b_{2,10}$	$--$	$--$	$--$	$--$	$--$	-0.055	0.005
$b_{2,20}$	$--$	$--$	$--$	$--$	$--$	0.07	0.005
$b_{2,30}$	$--$	$--$	$--$	$--$	$--$	-18.9(*)	-2.13(*)
$b_{2,40}$	$--$	$--$	$--$	$--$	$--$	11.3(*)	-2.56(*)
$b_{2,21}$	$--$	$--$	$--$	$--$	$--$	10.2(*)	-11.7(*)
$b_{2,31}$	$--$	$--$	$--$	$--$	$--$	72	-18.4
$b_{2,41}$	$--$	$--$	$--$	$--$	$--$	108(*)	12.5(*)
$b_{2,32}$	$--$	$--$	$--$	$--$	$--$	141(*)	143(*)
$b_{2,42}$	$--$	$--$	$--$	$--$	$--$	160(*)	825(*)
$b_{2,43}$	$--$	$--$	$--$	$--$	$--$	59.3(*)	4101(*)
β_1	$--$	$--$	0.017 (*)	0.0125(*)	0.0265(*)	0.0128(*)	0.0273(*)
β_2	$--$	$--$	$--$	$--$	$--$	$--$	-0.0295(*)
ℓ_{CL}	-74356	-74264	-74105	-73861	-73660	-73480	-73407
AIC_{CL}	148726	148548	148232	147756	147328	146998	146854

Table 1: Parameter estimates. The first two columns report the estimates of model M1 with Gaussian errors, the third and fourth columns report the models M2 with skewed errors. The fifth column reports the model M3 with ARG based state process and skewed error. The last two columns report the model M4 with time irreversible state process. The symbol $--$ indicates that a parameter is set to zero in a model. $*$ indicates that the parameter is significant at the 5 % level. In all models, matrix B is specified in terms of the symmetric matrix B_1 and the antisymmetric matrix B_2.

Increasing J, or adding a parameter capturing the skewness both lead to a significant improvement of the fit, in terms of the composite likelihood and composite AIC. Moreover, the model with antisymmetric B has a significant better fit than comparable models with symmetric

B. This shows that the volatilit process is not reversible. This result shows the advantage of our model, by allowing the coefficients of D to be negative. Although the switching regime can no longer be defined, the quality of approximation of the squared polynomial largely dominates the case with only nonnegative entries of D.

Let us now focus on the column for the ARG based model M3. The estimate of the autocorrelation coefficient of (X_t) is $\rho \approx 0.08$, which is rather weak. This can be explained by the left panel of Figure 3. Indeed, the serial dependence of the ARG model is characterized by one parameter ρ, and the lack of autocorrelation of $1/Y_t^2$ suggests a rather small $\hat{\rho}$. As a consequence, model M3 cannot well capture the substantial autocorrelation of Y_t^2.

ii) **Analysis of Model M4.**

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

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

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

The estimated antisymmetric matrix D_2 is significantly non zero, which confirms the time irreversibility of the Apple return data. Nevertheless, the largest entry of $|D_1|$ is $|d_{1,88}| = 1.479$, whereas the largest (in absolute value) entry of $|D_2|$ is $|d_{2,53}| = 0.122$, which is significantly smaller than 1.479.

iii) **Estimation of marginal moments.**

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

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

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

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

iii) **Estimation of the marginal density**

Let us now compare the kernel-based empirical marginal density with the density predicted by the different models. . We take a positive Gaussian kernel K, which is defined on $\mathbb{R}$, and has unit mass, then the marginal distribution of Y_t is estimated by:

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

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

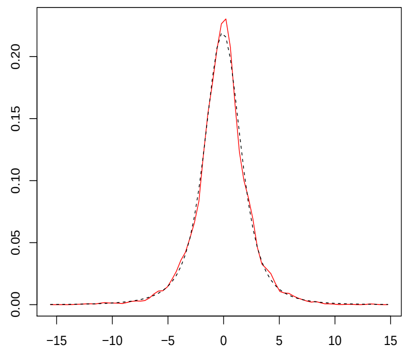


Figure 4: Comparison of the model M3 implied marginal density of $f(y_t)$ with the kernel density estimator. Full line: kernel density estimator; dashed line: model implied density.

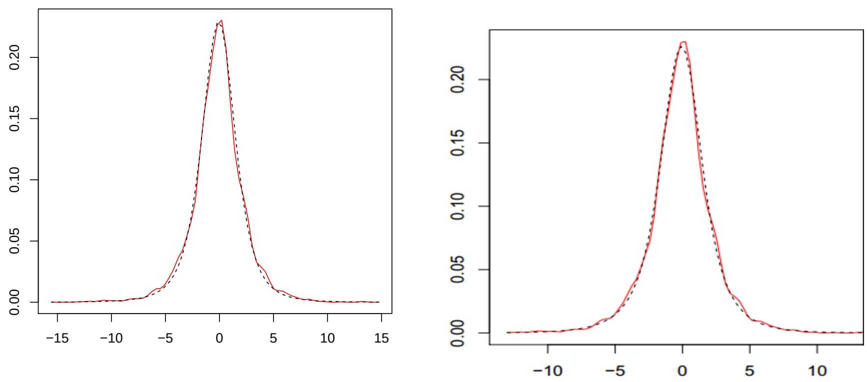


Figure 5: Comparison of the model M4 ($J = 4$) implied marginal density of $f(y_t)$ with the kernel density estimator. Full line: kernel density estimator; dashed line: model implied density. Left panel: $I = 1$; right panel: $I = 2$.

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

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

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

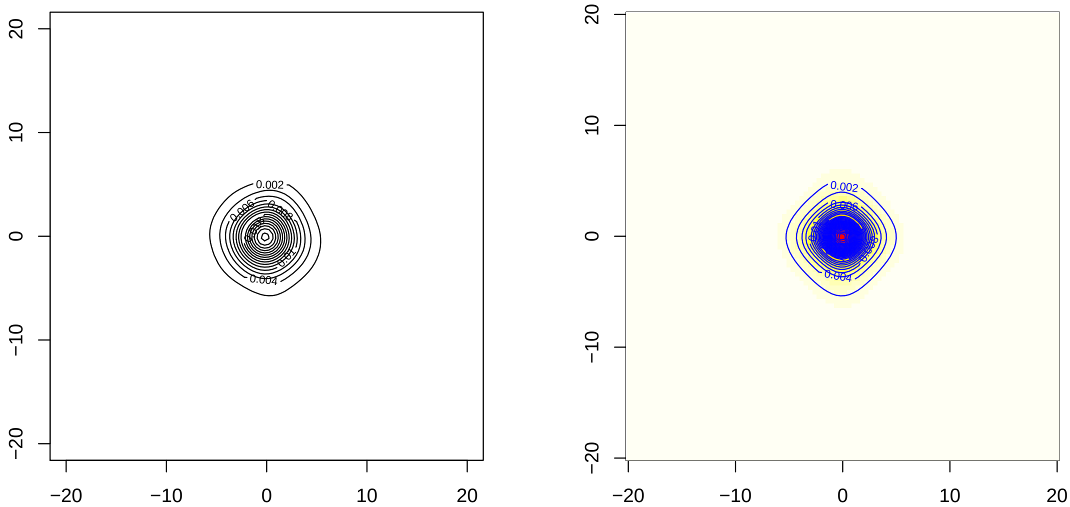


Figure 6: Left panel: Iso-density curves of the kernel-based estimate of $f(y_t, y_{t+1})$. Right panel: Iso-density curves of the model implied joint density.

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

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

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

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

  Thus we have:

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

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

  By Proposition 9, this mean is equal to:

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

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

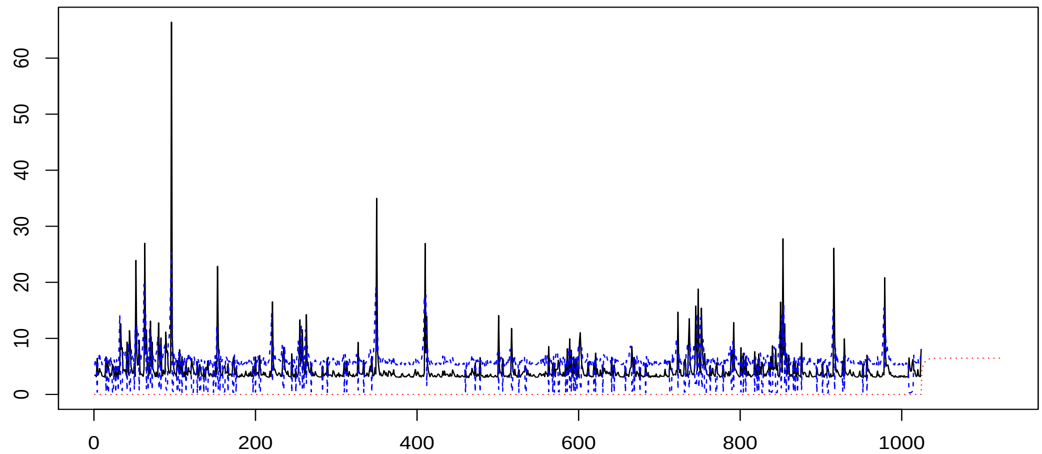


Figure 7: Filtering (dotted line), smoothing (dashed line) of the conditional variance for the latest 1000 dates, along with the term structure of volatility forecast for $h = 1$ to $h = 100$.

7 Conclusion

We have introduced a state-space model based on a flexible dynamics of the state process. This latter has an intuitive endogenous switching regime interpretation and allows for time irreversibility. The resulting state-space model is associated with simple composite likelihood based estimation, as well as simulation-free algorithms for filtering, forecasting, and smoothing.

Our model is illustrated on the stochastic volatility of asset return. From this financial point of view, we shall note that the same model can also be used for pricing derivatives. Indeed, it has recently been shown by Gouriéroux and Monfort (2015) that models with endogenous regime switching usually lead to simple pricing formulas. The flexibility of our new state process dynamics is an essential advantage for calibration, when the prices of a large number of derivatives are available. This is left for future research.

Appendix 1 Proofs of the propositions

Appendix 1.1 Proof of Proposition 1

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

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

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

Appendix 1.2 Proof of Proposition 2

Assume that identity (2.10) is valid for a given $h \geq 1$, then we have:

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

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

Appendix 1.3 Proof of Lemma 1

Function $x \mapsto U'(x) D e$ is continuous, lower bounded by zero, and goes to infinity when $|x|$ goes to infinity[15]. Thus in order for it to be 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, $U'(x) D e$

[15]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.

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

Appendix 1.4 Proof of Corollary 1

We have:

$$\Pi e = \int \frac{U(x)U'(x)De}{U'(x)De}\phi(x)\mathrm{d}x = \int U(x)\phi(x)\mathrm{d}x = e. \qquad \text{(eq. a.1)}$$

Similarly, we have:

$$(De)'\Pi = \int \frac{e'D'U(x)}{U'(x)De}U(x)'D\phi(x)\mathrm{d}x = e'D = e'D', \qquad \text{(eq. a.2)}$$

since $(D' - D)e = 0$.

Appendix 1.5 Proof of Proposition 5

Let us denote $V(x) = (1, x, ..., x^J)'$. We have $f(x_t, x_{t+1}) = \frac{1}{M}\phi(x_t)\phi(x_{t+1})(V'(x_t)BV(x_{t+1}))^2$. Thus the symmetry of $f(x_t, x_{t+1})$ is equivalent to:

$$\Big[V'(x_t)BV(x_{t+1}) + V'(x_{t+1})BV(x_t)\Big]\Big[V'(x_t)BV(x_{t+1}) - V'(x_{t+1})BV(x_t)\Big] = 0, \qquad \forall x_t, x_{t+1} > 0.$$

The LHS is the product of two polynomials in (x_t, x_{t+1}). The previous identity is satisfied if and only if at least one of the two multiplicative terms is identically zero[16], that is to say, B is antisymmetric, or symmetric.

Appendix 1.6 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:

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

[16]Indeed, at least one of the two terms are null for an infinity of x_t and an infinity of x_{t+1}. This latter term can be rearranged as a polynomial of x_t whose coefficients are polynomials of x_{t+1}. Thus this polynomial in x_t has an infinity of roots, thus all its coefficients are zero for an infinity of x_{t+1}). Thus this polynomial in x_t is identically zero.

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

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

As for the updating formula, we remark that the posterior density is linked to the predictive density via:

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

Thus we have:

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

which is formula (4.3)-(4.5).

Appendix 1.7 Proof of Corollary 8

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

Appendix 1.8 Proof of Proposition 9

Let us first compute the joint distribution:

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

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

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

Finally, the smoothing density is obtained by taking the ratio between the RHS of the last equation and its integral with respect to x_t:

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

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

Appendix 1.9 Proof of Proposition 10

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

Appendix 1.10 Proof of Lemma 2

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

Appendix 1.11 Proof of Proposition 11

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

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

References

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

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

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

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

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. (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data. *Journal of Business & Economic Statistics*, 19(4):428–435.

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

Creal, D. D. (2017). 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.

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.

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

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

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

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

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

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

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

Gouriéroux, C. and Monfort, A. (2017). Composite Indirect Inference with Application to Corporate Risks. *forthcoming Econometrics and Statistics*.

Gouriéroux, C., Monfort, A., and Renault, E. (2017). Consistent Pseudo-Maximum Likelihood Estimators. *forthcoming Annals of Economics and Statistics*.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, pages 357–384.

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

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

Huberman, G., Kandel, S., and Stambaugh, R. F. (1987). Mimicking Portfolios and Exact Arbitrage Pricing. *Journal of Finance*, 42(1):1–9.

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

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

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

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

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

Lu, Y. (2016). Dynamic frailty count process in insurance: Estimation, pricing and forecasting. *forthcoming Journal of Risk and Insurance*.

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

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

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

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

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

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

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

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

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

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

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

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

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.