Modeling time-varying parameters using artificial neural networks: A GARCH illustration

Morvan Nongni Donfack[,a] and Arnaud Dufays[*a,b,c,d]

[a] *Département d'économique, Université Laval, Canada.*
[b] *Département des sciences de gestion, Université Namur, Belgium.*
[c] *CeReFiM associate researcher.*
[d] *CRREP associate researcher.*

Version: September 1, 2020

Abstract We propose a new volatility process in which parameters vary over time according to an artificial neural network (ANN). We prove the process's stationarity as well as the global identification of the parameters. Since ANNs require economic series as input variables, we develop a shrinkage approach to select which explanatory variables are relevant to forecast volatility. Empirically, the proposed model favorably compares with other flexible processes in terms of in-sample fit on six financial returns. It also delivers accurate short-term volatility predictions in terms of root mean squared errors and the predictive likelihood criterion. For long-term forecasts, it can be competitive with the Markov-switching GARCH model if appropriate exogenous variables are used. Since our new type of time-varying parameter (TVP) process is based on a universal approximator, the approach can readily revisit and potentially improve many standard TVP applications.

JEL classification: C11, C15, C22, C45, C58

Keywords: GARCH, Neural network, Shrinkage priors, Time-varying parameters

Arnaud Dufays is a 'chargé de recherches' of the 'fonds de la Recherche Scientifique - FNRS'. He acknowledges financial support from the F.R.S-FNRS via the contract CR 144. He also acknowledges financial supports from the Fonds de recherche du Québec – Société et culture and from the SSHRC-CRSH via the project 430-2017-00215. The authors thank participants for their helpful comments at the conference entitled 'Recent Advances in Econometrics: International Conference in Honor of Luc Bauwens'.

[*] Corresponding author.

1 Introduction

Volatility is critically important for risk management and the pricing of derivatives (stocks, options, etc.). As such, a significant amount of research has been devoted to modeling the volatility of financial returns. This includes the generalized autoregressive conditional heteroskedastic (GARCH) model (Engle, 1982; Bollerslev, 1986), in which volatility evolves deterministically over time according to past volatility and past financial log-returns, the stochastic volatility (SV) model (Taylor, 2008), and many other alternatives. However, the majority of existing models typically rely on fixed parameters, ignoring the impact of changes in the financial system. This may lead to bad forecasts and decisions (e.g. Dieobold (1986), Lamoureux and Lastrapes (1990) and Hillebrand (2005)).

To account for economic changes, researchers have proposed models that allow parameters to vary over time. Two particular models are widely used: the Markov-switching (MS) process and the time-varying parameter (TVP) model. The former assumes parameters are fixed over a particular period, and detects abrupt changes to these parameters at specific times (e.g., Francq et al. (2001), Haas et al. (2004a) and Bauwens et al. (2010)). Conversely, in the TVP process, each parameter evolves in each time period. To alter parameters over time, a dynamic is postulated. A popular choice is an autoregressive (AR) process, since it balances parsimony and flexibility of the model (e.g. Kim et al. (1999) and Krolzig (2013)).[1] This paper proposes a class of TVP models in which parameters evolve according to artificial neural networks (ANNs) instead of the standard AR process.

Our TVP GARCH model, called the TVP_{ANN}-GARCH process, relies on an ANN function for the parameter dynamics of the GARCH volatility equation, instead of assuming an AR process for parameter dynamics as is done in standard TVP models. ANNs are appealing because they are universal approximators (for example, see Cybenko (1989), Bishop (2006)). Put simply, this means that if a true dynamic of the parameters exists and is deterministic, an ANN can arbitrarily closely approximate this function given the true explanatory variables.

[1] The AR process includes the random walk process.

Consequently, we believe that ANNs are natural choices for specifying parameter dynamics. Moreover, since any TVP application can potentially be revisited, the new approach could be widely applicable.

In addition to being universal approximators, ANNs have two other advantages over AR process: inference tractability and interpreting parameter dynamics. With respect to inference tractability, since ANNs are deterministic functions of the input variables, one can directly evaluate the likelihood function, which is the key ingredient for inferring model parameters. In contrast with TVP models, the AR process has to be integrated out. This integration can only be exact in particular frameworks (under linear models and Gaussian innovations). This limitation has generated numerous papers focusing on the TVP model estimation (see Broto and Ruiz (2004) for a review), and it remains one of the most active lines of TVP research (e.g. Chan and Eisenstat (2018)). Additionally, ANNs require a researcher to choose input variables (explanatory variables such as past volatility, book-to-market ratios, previous trading volumes or other financial factors) for modeling the financial log-return volatilities through the parameter dynamics. These variables are of particular interest to economists, since they create a network of crucial variables that Granger cause the output value of the ANNs. This direct interpretation is lacking in current TVP models.

With regard to the TVP literature, closely related processes are different kinds of smooth transition models. The latter process was originally proposed by Luukkonen et al. (1988), and consists of a TVP model in which parameters can change over time according to a logistic function. This smooth transition model exhibits the appealing feature of modeling both abrupt or smooth changes of the parameters, making a link between MS and standard TVP frameworks. In Medeiros and Veiga (2005), the transition function of the parameters is shown to be a one-layer ANN with a logistic function used in multiple neural units. In many subsequent works (for example, see Scharth and Medeiros (2009), McAleer and Medeiros (2008) and Medeiros and Veiga (2009)), the model is extended to different specifications such as long-memory processes, heterogeneous autoregressive models and GARCH processes.

The approach in this paper is distinct from smooth transition models. Smooth transition models use a one-layer ANN with logistic functions, despite the ANN literature finding that deeper ANNs generally produce better estimates. Moreover, these papers assume that all model parameters have to change with the same ANN, an assumption that limits the flexibility and the interpretation of the parameter dynamics (see change-point examples in Peluso et al. (2016) or Dufays and Rombouts (2018)). Smooth transition models have been limited to at most three ANN input variables, whereas we propose to include far more input variables. In fact, the model shrinks irrelevant parameters toward zero such that it can potentially include many explanatory variables. The shrinkage method is another deviation from the smooth transition literature, as most of the papers are developed in the frequentist paradigm (an exception is Deschamps, 2008). We propose a sound Bayesian estimation procedure that includes the computation of the marginal likelihood for model comparison purposes. An empirical exercise based on the price of US bank stocks highlights that the TVP_{ANN}-GARCH model is superior to commonly used processes in terms of marginal likelihood. In a forecasting exercise, the TVP_{ANN}-GARCH model stands out for predicting the volatility at short horizons (up to two days ahead) and it remains competitive at longer horizons with respect to the MS-GJR-GARCH model in terms of predictive likelihoods.

The paper is organized as follows. Section 2 presents the TVP_{ANN}-GARCH process, discusses its statistical properties (conditions for weak and strong stationarity) and reviews the related literature. Section 3 is devoted to the model identification conditions. Section 4 shows how to estimate the model. We illustrate the algorithm using simulated data in Section 5, and provide empirical examples in Section 6. Section 7 concludes the paper. The Appendix provides technical proofs for our theoretical results.

2 Model definition and theoretical results

We consider a GARCH(1,1) model specified as

$$\begin{aligned} y_t &= \sigma_t \eta_t \quad \text{with } \eta_t \sim IID(0,1), \\ \sigma_t^2 &= \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned}$$

with $\omega > 0$, $\alpha > 0$ and $\alpha + \beta < 1$. Since we assume the standard stationary conditions, let us denote the unconditional variance as $\bar{\omega} = \omega/(1-\alpha-\beta)$. We can therefore reframe the GARCH model as

$$\sigma_t^2 = \bar{\omega} + \phi(\sigma_{t-1}^2 - \bar{\omega}) + \alpha(y_{t-1}^2 - \sigma_{t-1}^2), \tag{1}$$

in which $\phi = \alpha + \beta$ is the persistence of the variance dynamic. We use an ANN to model the dynamics of the unconditional variance as well as the persistence parameter. Analogously to Equation (1), the TVP_{ANN}-GARCH(1,1) model is specified as

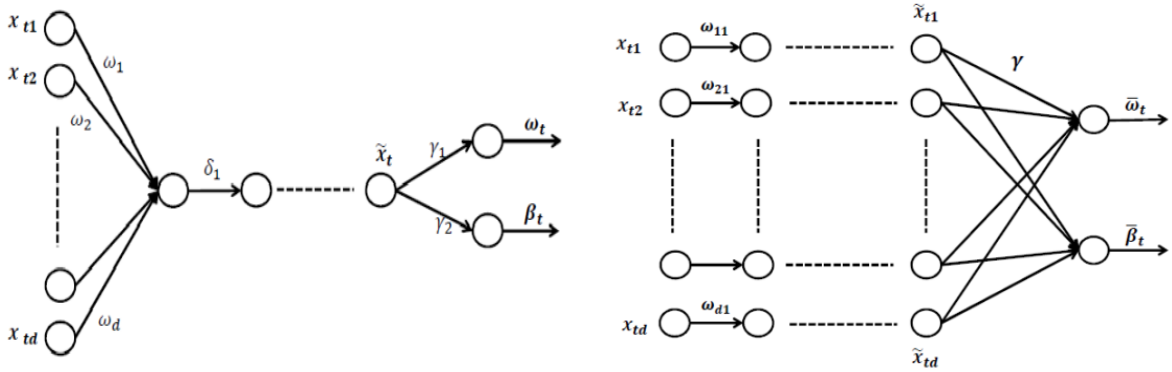
$$\begin{aligned} y_t &= \sigma_t \eta_t \quad \text{with } \eta_t \sim IID(0,1), \\ \sigma_t^2 &= \bar{\omega}_t + \phi_t(\sigma_{t-1}^2 - \bar{\omega}_t) + \alpha(y_{t-1}^2 - \sigma_{t-1}^2), \end{aligned} \tag{2}$$

where $\phi_t = \alpha + (1-\alpha)\tilde{\beta}_t$. The outputs of the ANN are $\{\bar{\omega}_t, \tilde{\beta}_t\}$. By doing so, the long term variance as well as the persistence are now time-varying. This is in line with the theoretical work of Engle and Rangel (2008) and with recent GARCH literature that found evidence of dynamic parameters (e.g. He and Maheu, 2010; Bauwens et al., 2013; Engle et al., 2013).

2.1 ANN structure

Many neural network topologies are proposed in the machine learning literature (e.g. Demuth et al. (2014)). In this paper, we consider two simple neural network architectures depicted in Figure 1 that exhibit the desirable property of being globally identified (see Theorem 3

below). As shown by Figure (1a), the single-factor structure with ℓ hidden layers, denoted by TVP$_{\text{ANN}(\ell,1)}$-GARCH, exhibits one node in each layer. This is a parsimonious ANN specification that captures the most relevant factor of the explanatory variables. The second ANN architecture with ℓ hidden layers, detailed in Figure (1b) and called TVP$_{\text{ANN}(\ell,X)}$-GARCH, involves more parameters but also implies more flexibility because multiple factors can activate the outputs. We use logistic functions (i.e., $f(x) = (1 + exp(-x))^{-1}$) as non-linear activation functions. Given explanatory variables $\boldsymbol{x}_t = (x_{t1}, \cdots, x_{td})' \in \mathbb{R}^{d\times 1}$, Table 1 shows the mathematical structure of the two ANNs.



(a) Single-factor structure

(b) Multiple-factor structure

Figure 1 – Structures of the two ANNs used in the TVP$_{\text{ANN}}$-GARCH process.

The ANN outputs are the time-varying parameters of the GARCH volatility Equation (2) such that the long-term variance $\bar{\omega}_t$ evolves over $]0, K[$ and the persistence $\tilde{\beta}_t$ lies in $]0, 1[$. In the empirical exercise, the upper bound is set to 100. For easing the notations, the model parameters are denoted by $\boldsymbol{\theta}$ whatever the network structures. In particular, when the TVP$_{\text{ANN}(\ell,X)}$-GARCH model is estimated, the parameter set referred to $\boldsymbol{\theta} = (\omega_{ij}, b_{ij}, \gamma_{ki}, \gamma_{10}, \gamma_{20}, \alpha)$, where $i \in \{1, \cdots, d\}$, $j \in \{1, \cdots, \ell\}$ and $k \in \{1, 2\}$. When we consider the TVP$_{\text{ANN}(\ell,1)}$-GARCH process, it refers to $\boldsymbol{\theta} = (\omega_i, \delta_j, b_r, \gamma_k, \gamma_{10}, \gamma_{20}, \alpha)$, where $i \in \{1, \cdots, d\}$, $j \in \{1, \cdots, \ell - 1\}$, $r \in \{1, \cdots, \ell\}$ and $k \in \{1, 2\}$.

Table 1 – Structure of the two artificial neural networks.
The subscripts i and j are used for the input dimension and the number of layers, respectively. They belong to the following sets: $\forall i \in \{1, \ldots, d\}$ and $\forall j \in \{2, \ldots, \ell\}$.

$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH (Single-factor structure)	
Output	$\bar{\omega}_t = Kf(\gamma_1 H_\ell(\boldsymbol{x}_t) + \gamma_{10}), \quad \tilde{\beta}_t = f(\gamma_2 H_\ell(\boldsymbol{x}_t) + \gamma_{20})$
Layers	$\begin{cases} H_j(\boldsymbol{x}_t) = f(\delta_j H_{j-1}(\boldsymbol{x}_t) + b_j), \forall j \in \{2, \ldots, \ell\}, \\ H_1(\boldsymbol{x}_t) = f(\sum_{i=1}^{d} \omega_i x_{ti} + b_1) \end{cases}$
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH (Multiple-factor structure)	
Output	$\bar{\omega}_t = Kf(\boldsymbol{\gamma}_1' H_\ell(\boldsymbol{x}_t) + \gamma_{10}), \quad \tilde{\beta}_t = f(\boldsymbol{\gamma}_2' H_\ell(\boldsymbol{x}_t) + \gamma_{20})$
Layers	$\begin{cases} H_\ell(\boldsymbol{x}_t) = (H_\ell(x_{t1}), H_\ell(x_{t2}), \ldots, H_\ell(x_{td}))' \\ H_j(x_{ti}) = f(\omega_{ij} H_{j-1}(x_{ti}) + b_{ij}), \forall j \in \{2, \ldots, \ell\}, \\ H_1(x_{ti}) = x_{ti} \end{cases}$

2.2 Related processes

Logistic functions have been used in volatility models such as in the ANN-GARCH (e.g., Donaldson and Kamstra, 1997) and in the smooth transition (ST) GARCH processes (e.g., Hagerud et al., 1997). The difference between these two types of processes is tenuous because they mainly differ in their dynamic interpretation. An ANN-GARCH model will be used to capture nonlinearities in the conditional variance dynamic. On the contrary, a smooth transition process is interpreted as a time-varying parameter (TVP) GARCH model in which the model parameters evolve according to a logistic function. In this section, we highlight how the TVP_{ANN}-GARCH model differs from those processes.

Donaldson and Kamstra (1997) introduce the ANN-GARCH model in which an ANN is added to the GARCH equation. In its simplest form, their ANN-GARCH model and the subsequent works of Caulet and Peguin-Feissolle (2000), Lanne and Saikkonen (2005) and Miazhynskaia et al. (2008) can be specified as

$$\begin{aligned} y_t &= \sigma_t \epsilon_t, \\ \sigma_t^2 &= \omega_t + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned} \tag{3}$$

in which $\omega_t = \omega + \text{ANN}(\mathbf{x_t}, \gamma)$. Equation (3) highlights that the ANN-GARCH model is also a TVP-GARCH process. The papers on ANN-GARCH models use a single-layer neural network with logistic functions. They mainly differ by their number of neurons and by using inputs variables such as the lagged shocks (i.e., ϵ_{t-i} with $i > 0$) and/or the lagged conditional variance. In particular, Donaldson and Kamstra (1997) use several powers of the lagged returns as inputs and a linear combination of logistic functions (i.e., up to five neurons). To limit the computational burden, the ANN parameters are not estimated but drawn from a Uniform distribution. Caulet and Peguin-Feissolle (2000) propose a test of non-linearity in the GARCH equation and use the lagged return as an input variable. Miazhynskaia et al. (2008) and the most simple specification of Lanne and Saikkonen (2005) suggest the lagged conditional variance for feeding the ANN. By choosing the appropriate inputs, the TVP_{ANN}-GARCH model nests all these frameworks. In addition, it extends these studies by at least three aspects. First, it makes the parameters $\alpha + \beta$ time-varying. Second, we consider and show how to select among a large amount of ANN inputs. Finally, our ANN structure is deeper because it can exhibit multiple hidden layers.

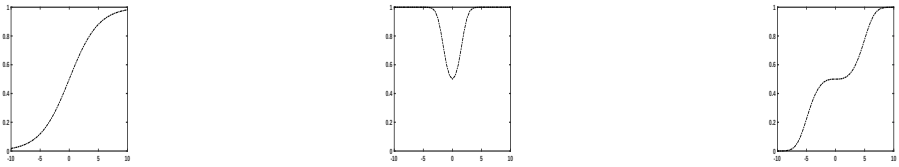
In the smooth transition literature, the logistic function is not viewed as an activation function in an ANN but as a function which makes the model parameter smoothly varying from one state to another. This is because a logistic function is monotonic between 0 and 1 (i.e., the two states). The standard ST-GARCH(1,1) model is specified as

$$y_t = \sigma_t \epsilon_t, \tag{4}$$

$$\sigma_t^2 = \omega_1 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + (\omega_2 + \alpha_2 y_{t-1}^2 + \beta_2 \sigma_{t-1}^2) f(x_t, \gamma, c), \tag{5}$$

in which x_t is an explanatory variable and $f(x_t, \gamma, c) = \Big(1 + \exp\Big[-\gamma(x_t - c)\Big]\Big)^{-1}$ with $\gamma > 0$. Focusing on the intercept, the ST-GARCH process is a TVP-GARCH model with a dynamic given by $\omega_t = \omega_1 + \omega_2 f(x_t, \gamma, c)$. The parameter ω_t smoothly evolves between ω_1 and ω_2 according to a non-linear transformation of x_t. For a GARCH(1,1) equation, the models

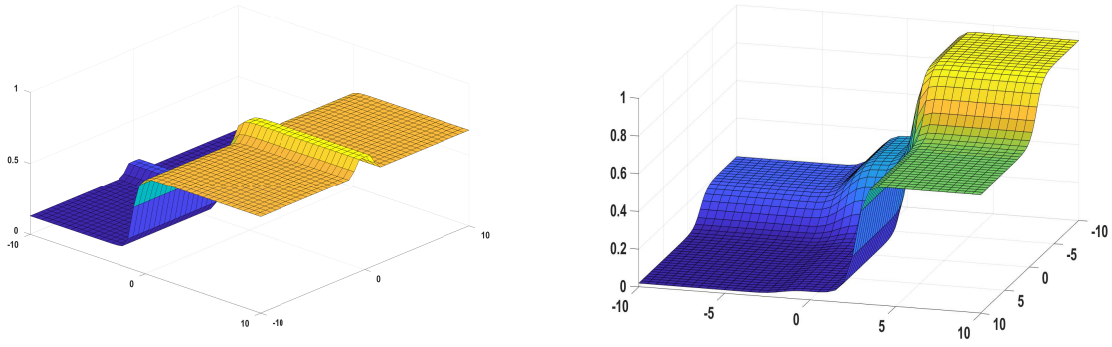
of Hagerud et al. (1997), González-Rivera (1998) and Anderson et al. (1999) are particular cases of the process given by Equations (4)-(5). Medeiros and Veiga (2009) extend the ST-GARCH process to multiple states while Kılıç (2011) include smooth transition dynamics in a fractionally integrated GARCH (IGARCH) equation. All these references use the lagged returns (or the shocks when the conditional mean is different from zero) as x_t. By considering multiple states, the parameters in Medeiros and Veiga (2009) can smoothly vary between more than two extreme values which makes their framework very appealing. Interestingly, the number of states can also be generated by ex-ante transforming the explanatory variable. Figure 2 illustrates the logistic function with an explanatory variable at different powers. Note that the explanatory variable used in Medeiros and Veiga (2009) is the lagged returns. Therefore, one could consider the squared lagged returns (or the cube) to generate additional states without resorting to multiple logistic functions as they do. Our framework allows for combinations of multiple explanatory variables and by doing so, the TVP$_{\text{ANN}}$-GARCH model can create multiple states as well (see Figure 3 for two examples). It is worth mentioning that we also differ from Medeiros and Veiga (2009) by disentangling the dynamics of the time-varying parameters, by proposing a Bayesian estimation, a testing procedure and by allowing for deeper ANN structures.



(a) $f(x, \gamma = 0.4, c = 0)$ (b) $f(x^2, \gamma = 0.4, c = 0)$ (c) $f(x^3, \gamma = 0.01, c = 0)$

Figure 2 – Logistic function given by $f(x^q, \gamma, c) = \Big(1 + \exp\Big[-\gamma(x^q - c)\Big]\Big)^{-1}$ with respect to x.

As a final reference, Dueker et al. (2011) develop the contemporaneous-threshold (C)



(a) Squared input: y^2 (b) Cube input: y^3

Figure 3 – Examples of outputs given by the multiple factor structure with one hidden layer and two inputs variables (x and y^q). The output functions are displayed with respect to x and y.

ST-GARCH model. The process stands for a 2-component GARCH model in which the probability is time-varying according to a smooth transition function depending on the lagged variance and the lagged return. The model is given by

$$y_t = \sigma_t \epsilon_t, \tag{6}$$

$$\sigma_t^2 = G(y_{t-1}^2, \sigma_{t-1}^2)\sigma_{1,t}^2 + (1 - G(y_{t-1}^2, \sigma_{t-1}^2))\sigma_{2,t}^2, \tag{7}$$

$$\sigma_{i,t}^2 = \omega_i + \alpha_i y_{t-1}^2 + \beta_i \sigma_{t-1}^2, \tag{8}$$

for $i = 1, 2$ and in which $G(y_{t-1}^2, \sigma_{t-1}^2)$ is the conditional cumulative distribution of y_t given the information set up to $t-1$. Expanding the conditional variance equation, the model becomes a ST-GARCH process with another smooth transition function since it leads to

$$\sigma_t^2 = G(y_{t-1}^2, \sigma_{t-1}^2)[\omega_1 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2] + (1 - G(y_{t-1}^2, \sigma_{t-1}^2))[\omega_2 + \alpha_2 y_{t-1}^2 + \beta_2 \sigma_{t-1}^2].$$

The C-ST-GARCH model of Dueker et al. (2011) highlights the tight link between the ST-GARCH and the component GARCH models with time-varying weights such as in Bauwens and Storti (2009). Like any ST-GARCH process, the TVP_{ANN}-GARCH model also exhibits

a 2-component GARCH representation that is given by

$$\sigma_t^2 = (1-\tilde{\beta}_t)\sigma_{1,t}^2 + \tilde{\beta}_t\sigma_{2,t}^2, \tag{9}$$

$$\sigma_{1,t}^2 = \bar{\omega}_t(1-\alpha) + \alpha y_{t-1}^2, \tag{10}$$

$$\sigma_{2,t}^2 = \alpha y_{t-1}^2 + (1-\alpha)\sigma_{t-1}^2. \tag{11}$$

The conditional variance of the TVP$_{\text{ANN}}$-GARCH model is driven by an ARCH model with time-varying long-term variance and an IGARCH process.[2] From this representation, we see that the persistence of the conditional variance is generated by the persistence in the inputs of $\bar{\omega}_t$ and by the time-varying weight $\tilde{\beta}_t$. In fact, when the probability is close to one, the TVP$_{\text{ANN}}$-GARCH model becomes highly persistent.

2.3 Model properties

Stationarity and ergodicity are useful properties for a stochastic process. Over the paper, we make the following assumption:

(H.1) The process $\{x_t\} \equiv \{x_{t1}, \ldots, x_{td}\}$ is jointly strictly stationary and ergodic.

Following Medeiros and Veiga (2009), Theorem 1 presents a necessary and sufficient log-moment condition for the strict stationarity and ergodicity of the TVP$_{\text{ANN}}$-GARCH model.

Theorem 1. *Let $y_t \in \mathbb{R}$ follow an TVP$_{ANN}$-GARCH process as in Equation (2). Assuming*

[2] Note that any weakly stationary GARCH(1,1) model can be re-framed into a 2-component GARCH process with an ARCH and an IGARCH dynamics. To see this, we consider a GARCH(1,1) specification, in which $\bar{\omega} = \frac{\omega}{1-\alpha-\beta}$ and $\beta = (1-\alpha)q$,

$$\begin{aligned}\sigma_t^2 &= \bar{\omega}(1-\alpha-\beta) + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2, \\ &= (1-q)[\bar{\omega}(1-\alpha) + \alpha y_{t-1}^2] + q[\alpha y_{t-1}^2 + (1-\alpha)\sigma_{t-1}^2].\end{aligned}$$

We find the 2-component GARCH model given by,

$$\begin{aligned}\sigma_{1,t}^2 &= \bar{\omega}(1-\alpha) + \alpha y_{t-1}^2, \\ \sigma_{2,t}^2 &= \alpha y_{t-1}^2 + (1-\alpha)\sigma_{t-1}^2.\end{aligned}$$

$0 < \alpha < 1$, *the process* $\boldsymbol{u}_t = (y_t, \sigma_t^2)'$ *is strictly stationary and ergodic if and only if*

$$\mathbb{E}\left\{\ln\left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2\right]\right\} < 0, \quad \forall t \in Z. \tag{12}$$

Proof. See Appendix A. □

Corollary 1. *Assuming* $\alpha \in]0,1[$, *the* TVP_{ANN}*-GARCH process is strictly stationary and ergodic.*

Proof. Using Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}\left\{\ln\left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2\right]\right\} &< \ln\left\{\mathbb{E}\left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2\right]\right\}, \\
&= \ln\left((1-\alpha)\tilde{\beta}_t + \alpha\right), \\
&< 0,
\end{aligned}$$

where the last inequality is obtained by observing that $0 < \alpha < 1$ and $\tilde{\beta}_t$ is bounded between 0 and 1. □

The following theorem presents the necessary conditions for the existence of moments of the TVP_{ANN}-GARCH process.

Theorem 2. *Let* $y_t \in \mathbb{R}$ *follow an* TVP_{ANN}*-GARCH process as in Equation (2) and* $\mathbb{E}[\eta_t^{2k}] < \infty$, *for* $k = 1, 2, 3, \ldots$. *Assuming* $\alpha \in]0,1[$ *and that the moments of order up to* $n = k-1$ *exist, the* $2k$*-th-order moment of* y_t *exists if*

$$\mathbb{E}\left\{\left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2\right]^k\right\} < 1. \tag{13}$$

Proof. See Appendix A. □

Note that inequality (13) always holds for $k = 1$. In fact, when $k = 1$, the inequality simplifies to

$$\mathbb{E}\left\{\left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2\right]\right\} = (1-\alpha)\tilde{\beta}_t + \alpha.$$

As $\tilde{\beta}_t$ is bounded between 0 and 1 for all t, it implies that $0 < \tilde{\beta}_t < 1$ and that $\alpha < (1-\alpha)\tilde{\beta}_t + \alpha < 1$. Consequently, the TVP$_{\text{ANN}}$-GARCH process always exhibits a well-defined unconditional variance.

2.4 About the persistence of the TVP$_{\text{ANN}}$-GARCH model

The TVP$_{\text{ANN}}$-GARCH conditional variance is specified as,

$$y_t = \sigma_t \eta_t, \tag{14}$$

$$\sigma_t^2 = \bar{\omega}_t(\boldsymbol{x}_t)(1-\phi_t(\boldsymbol{x}_t)) + \alpha y_{t-1}^2 + (\phi_t(\boldsymbol{x}_t)-\alpha)\sigma_{t-1}^2, \tag{15}$$

in which we highlight that the time-varying parameters depend on the explanatory variables $\boldsymbol{x}_t \in \Re^d$ and are driven by an ANN. Let us assume that $\boldsymbol{x}_t = (z_{t1}, z_{t2})'$ where z_{t1} and z_{t2} are independent stationary processes with no serial dependence and that $\bar{\omega}_t(\boldsymbol{x}_t) = \bar{\omega}_t(z_{t1})$, $\phi_t(\boldsymbol{x}_t) = \phi_t(z_{t2})$ such that $\bar{\omega}_t(z_{t1}) = \omega_1 \mathbb{1}_{\{z_{t1} \leq \bar{z}_1\}} + \omega_2 \mathbb{1}_{\{z_{t1} > \bar{z}_1\}}$ and $\phi_t(z_{t2}) = \phi_1 \mathbb{1}_{\{z_{t2} \leq \bar{z}_2\}} + \phi_2 \mathbb{1}_{\{z_{t2} > \bar{z}_2\}}$. As emphasized by Figures 2-3, our ANN structure can produce this type of two-state dynamics. Setting $P[z_{ti} \leq \bar{z}_i] = p_i$ for $i = 1, 2$, the TVP$_{\text{ANN}}$-GARCH model boils down to a mixture GARCH process with four different states. The state space and the mixture weights are given by $\chi = \{\bar{\omega}_1(1-\phi_1), \bar{\omega}_1(1-\phi_2), \bar{\omega}_2(1-\phi_1), \bar{\omega}_2(1-\phi_2)\}$ and $\pi = \{p_1 p_2, p_1(1-p_2), (1-p_1)p_2, (1-p_1)(1-p_2)\}$, respectively. Applying the results of Francq and Zakoian (2008) and assuming $\mathbb{E}(\eta_t^4) < \infty$ as well as $\mathbb{E}(y_t^4) < \infty$,[3] we find that

$$\mathbb{E}(y_t^2) = \bar{\omega}_1 p_1 + (1-p_1)\bar{\omega}_2, \tag{16}$$

$$\rho_1 \equiv \text{Corr}(y_t^2, y_{t-1}^2) = \rho_1^{\text{G}} \frac{(1+\alpha^2-\phi_1^2)(1+\alpha[\phi_2(1-p_2)+\phi_1 p_2] - (1-p_2)\phi_2^2 - p_2\phi_1^2)}{(1+\alpha\phi_1-\phi_1^2)(1+\alpha^2-(1-p_2)\phi_2^2-p_2\phi_1^2)},$$

$$\rho_i = (\phi_1 p_2 + (1-p_2)\phi_2)^{i-1}\rho_1, \text{ for } i > 1, \tag{17}$$

[3] In this particular case, the assumption $\mathbb{E}(y_t^4) < \infty$ is satisfied if

$$p_2\phi_1^2 + (1-p_2)\phi_2^2 + (\kappa-1)\alpha^2 < 1$$

where $\kappa = \mathbb{E}(\eta_t^4)$. The later inequality is a direct consequence of Theorem 2.

in which ρ_1^{G} denotes the first autocorrelation of a GARCH(1,1) model with parameters given by $\{\bar{\omega}_1, \alpha, \phi_1\}$. As illustrated by Equations (16)-(17), the time-varying parameters act differently on the variance dynamic. In this example, the parameter $\bar{\omega}(\boldsymbol{x}_t)$ controls the long-term variance while the parameter $\phi_t(\boldsymbol{x}_t)$ impacts on the persistence of the conditional variance. Additionally and as documented by Equation (17), the persistence of the TVP$_{\text{ANN}}$-GARCH model can be higher than the persistence of a standard GARCH model given by ϕ_1 (i.e., when $\phi_2 > \phi_1$). Of course, this orthogonal interpretation of the model parameters does not perfectly hold when the explanatory variables exhibit serial dependence. In such a case, the impacts of the two parameters are intertwined.

3 Identification of the model

Model identification is a relevant issue in statistics. From a frequentist perspective, identification is a necessary condition for consistency results. In the Bayesian paradigm, model identification helps for simulating the posterior distribution of the parameters. In fact, if a model is not identified, the posterior distribution exhibits multiple global modes which lead to complex multi-modal posterior distributions to simulate (see, e.g., Fruhwirth-Schnatter (2004)). Consequently, identification of the TVP$_{\text{ANN}}$-GARCH model is desirable for both paradigms. Identification of the TVP$_{\text{ANN}}$-GARCH parameters is directly related to the ANN identifiability. In fact, identification of the ANN parameters has always been a theoretical issue in the ANN literature (for example, see Sussmann (1992) and Hwang and Ding (1997)). To prove the identification of the TVP$_{\text{ANN}}$-GARCH parameters, we follow Hwang and Ding (1997) and rely on two concepts called "reducibility" and "minimality," which we define below.

The TVP$_{\text{ANN}}$-GARCH model is identifiable if the associated ANN is identifiable. Put another way, the TVP$_{\text{ANN}}$-GARCH process is identifiable if there does not exist two sets of parameters that, given the same (non-constant) inputs, produce an identical output. First, we need to preclude some obvious transformations of the ANN structure that cause identification issues. These transformations are related to the sign equivalence of the logistic

activation function and to the permutations of two hidden nodes of the network's hidden layer. The two structures we use have been chosen such that permutations of nodes are not possible and that the sign equivalence property of the logistic function cannot lead to identification issues. Consequently, the problem of model identification is only related to the concepts of minimality (or redundancy) and reducibility of the associated ANN. We will discuss these two concepts and establish the conditions that guarantee that the TVP_{ANN}-GARCH model is identifiable and minimal.

Definition 1. The TVP_{ANN}-GARCH model is minimal (or nonredundant) if the input-output map of the corresponding ℓ hidden layers' neural network is irreducible and cannot be obtained from a neural network with $k < \ell$ hidden layers.

By definition, a reducible ANN contains irrelevant nodes. This means that, given the same input, the ANN output can be obtained from another ANN with fewer hidden nodes. Irreducibility is therefore a necessary condition for minimality to hold. ANN irreducibility can be achieved by listing transformations under which the ANN can be reduced and by imposing parameter conditions to avoid these transformations. Hereafter, an ANN will be considered irreducible if none of these transformations can occur.

Definition 2. The multiple-factor ANN is *reducible* if one of the following conditions holds:

(i) $\gamma_{ij} = 0 \quad$ for some $\quad i = 1, 2 \quad$ *and* $\quad j = 1, \cdots, d,$

(ii) $\omega_{ij} \leq 0 \quad$ for some $\quad i = 1, \cdots, d \quad$ *and* $\quad j = 1, \cdots, \ell.$

Definition 3. The single-factor factor ANN is *reducible* if one of the following conditions holds:

(i) $\gamma_i = 0 \quad$ for some $\quad i = 1, 2,$

(ii) $\omega_i = 0 \quad$ for some $\quad i = 1, \cdots, d,$

(iii) $\delta_i \leq 0 \quad$ for some $\quad i = 1, \cdots, \ell.$

Considering these conditions, we constrain our ANN parameters as follows:

Assumption 1. *Parameters of the multiple-factor ANN satisfy the conditions*

(R.1) $\omega_{ij} > 0 \quad i = 1, \cdots, d \quad \textit{and} \quad j = 1, \cdots, \ell,$

(R.2) $\gamma_{ij} \neq 0 \quad i = 1, 2 \quad \textit{and} \quad j = 1, \cdots, \ell.$

Assumption 2. *Parameters of the single-factor ANN satisfy the conditions*

(R.3) $\omega_i \neq 0 \quad i = 1, \cdots, d,$

(R.4) $\delta_i > 0 \quad i = 1, \cdots, \ell,$

(R.5) $\gamma_i \neq 0 \quad i = 1, 2.$

Now we can state the model identification theorem:

Theorem 3. *The TVP_{ANN}-GARCH model is globally identifiable and minimal if*

- *the ANN structure is a multiple-factor model that has $\ell \in \{1, 2, 3\}$ hidden layers and satisfies Assumption 1.*
- *the ANN structure is a single-factor model that has $\ell \in \{1, 2, 3\}$ hidden layers and satisfies Assumption 2.*

Proof. See Appendix A. □

4 Shrinkage priors and estimation

Given the ANN structures in Figure 1, the problem of network selection is reduced to the choice of input variables and the number of hidden layers. A standard model selection strategy consists of estimating all the models and of choosing the best one according to the marginal likelihood (see Miazhynskaia et al., 2008, for an example with ANN-GARCH models). Note, however, that we explore three model dimensions that are the ANN structure, the number of hidden layers and the number of inputs, implying $2 \times 3 \times 2^d$ models. For instance, assuming $d = 9$, as in our empirical application, leads to 3072 models (i.e., $2 \times 3 \times 2^9 = 3072$). To avoid the proliferation of models, we rely on shrinkage priors to explore the model space generated by the inputs (see also Ročková and George, 2014, in which shrinkage

priors are used for variable selection). Once the inputs are selected, we estimate the six remaining models and discriminate their fit using the marginal likelihood.

Concretely, our strategy consists of estimating the TVP_{ANN}-GARCH process with all the potentially significant explanatory variables, in conjunction with an ANN with three hidden layers. Using the penalty imposed by the shrinkage prior to each model parameter, we assess which explanatory variables exhibit the strongest predictive performance for modeling the financial volatility through the ANN. Then, we rely on this subset of explanatory variables to fix the number of hidden layers. To do so, we estimate the same TVP_{ANN}-GARCH model given one, two and three hidden layers, and then select the best number of layers (and the ANN structure) based on the MLL. In the next subsection, we provide more detail on the shrinkage priors.

4.1 Shrinkage priors and prior elicitation

Since many explanatory variables could be used to model financial volatility, an exhaustive search by model selection would be too cumbersome. To avoid this prohibitive procedure, we use shrinkage priors to determine which variables to include (see Bhadra et al., 2019, for a recent review on shrinkage priors). To do so, we note that when the parameter ω_i (resp. γ_{ki} with $k \in [1, 2]$) for $i \in [1, d]$ of the single-factor structure (resp. of the multiple-factor structure) is set to zero, the variable $\{x_{ti}\}_{t=1}^{T}$ does not influence the conditional variance σ_t^2 at any time. Consequently, we apply our shrinkage priors only to these parameters. For ease of discussion, we focus on one of the shrinkage priors, which we refer to λ, in the following analysis.

We apply shrinkage prior to solve a small dimensional variable selection problem because the parameter dimension is much smaller than the sample size. Therefore, we could choose any of the popular Bayesian shrinkage priors such as the spike and slab (S&S) priors of Ishwaran and Rao (2005) or of Ročková and George (2018), the Bayesian lasso of Park and Casella (2008), the horseshoe prior of Carvalho et al. (2010), the Normal-Gamma of Griffin

and Brown (2010), or the Dirichlet-Laplace prior of Bhattacharya et al. (2015). In this paper, we use the S&S prior to shrink the parameters because its two-component mixture eases the decision of selecting a variable. In particular, we use a S&S prior based on uniform distributions and introduced by Dufays and Rombouts (2020). The shrinkage prior penalizes the log-likelihood function similarly to standard information criteria and consists of a two uniform component distribution called $2MU(a, b, P)$. It is specified as

$$\lambda|a, b, P \sim qU[\frac{-a}{2}, \frac{a}{2}] + (1-q)U[\frac{-b}{2}, \frac{b}{2}], \tag{18}$$

$$q = \frac{a(1-e^P)}{be^P + a(1-e^P)}, \tag{19}$$

where $\mathrm{U}[\frac{-a}{2}, \frac{a}{2}]$ denotes a uniform distribution with bounds that depend on a and the parameter q stands for the weight of the mixture. As with the continuous S&S prior (see Ishwaran and Rao (2005)), we assume a narrow component by setting a to a small value and a wide component by making b very large. The hyper-parameter P represents a penalty on the log-likelihood function if the parameter λ is in the region $[\frac{-b}{2}, \frac{b}{2}]\backslash[\frac{-a}{2}, \frac{a}{2}]$.

Let us relate the shrinkage prior to the popular model selection procedures. A standard approach for variable selection relies on an information criterion and consists in maximizing a penalized likelihood function given by

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} \ \ln f(y_{1:T}|\boldsymbol{\theta}) + P(\boldsymbol{\theta}), \tag{20}$$

in which $\boldsymbol{\theta} \in \Re^d$ stands for the model parameters and $\ln f(y_{1:T}|\boldsymbol{\theta})$ denotes the log-likelihood function. Widely used penalty functions are the Akaike information criterion (i.e., $P(\boldsymbol{\theta}) = -\sum_{i=1}^{d} 2\mathbb{1}_{\{\theta_i \neq 0\}}$) and the Bayesian information criterion (BIC, $P(\boldsymbol{\theta}) = -\sum_{i=1}^{d} \frac{1}{2} \ln T \mathbb{1}_{\{\theta_i \neq 0\}}$) in which d is the number of model parameters and T stands for the sample size. These are hard-thresholding penalty functions (i.e., L_0) and so, are difficult to maximize although they enjoy excellent theoretical properties. By introducing the Lasso penalty function, Tibshirani (1994) offers new perspectives to handle variable selection in high-dimensional settings as

the penalty function becomes continuous (besides at zero) and convex.

As highlighted in Park and Casella (2008), Equation (20) is equivalent to finding the maximum a posteriori (MAP) of a Bayesian posterior distribution with prior distributions given by $\exp(P(\boldsymbol{\theta}))$. Assuming that all the model parameters are driven by a 2MU distribution (i.e., $\theta_i \sim 2MU(a, b, P)$), the MAP is given by,[4]

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \text{argmax}_{\boldsymbol{\theta}} \quad \ln f(y_{1:T}|\boldsymbol{\theta}) + \sum_{i=1}^{p} \log\Big(q\frac{1}{a}\mathbb{1}_{\{\boldsymbol{\theta}_i \in \underbrace{[-\frac{a}{2}, \frac{a}{2}]}_{\text{Spike}}\}} + (1-q)\frac{1}{b}\mathbb{1}_{\{\boldsymbol{\theta}_i \in \underbrace{[-\frac{b}{2}, \frac{b}{2}]}_{\text{Slab}}\}}\Big), \\ &\propto \quad \ln f(y_{1:T}|\boldsymbol{\theta}) + \sum_{i=1}^{p} P\mathbb{1}_{\{\boldsymbol{\theta}_i \in \text{Slab}\setminus\text{Spike}\}}.\end{aligned}$$

The 2MU prior distributions is a Bayesian equivalence to the standard information criteria as the MAP solves a penalized function with L_0 type of penalty functions. When the MAP of a model parameter lies in the slab support, the log-likelihood function is penalized by P.

The 2MU prior depends on three hyper-parameters: a, b and P. We follow the procedure in Dufays and Rombouts (2018) and Dufays and Rombouts (2020) to set them. In particular, the narrow bound, a, is different for each model parameter and is fixed at 0.1 times the standard deviation of the model parameter when it is estimated with a diffuse prior (i.e., a normal distribution with an expected value of 0 and variance of 10). The bound of the wide component, b, is set to a large value, 50. To set P, we should ideally use the MLL. Indeed, to test if the explanatory variable $\{x_{it}\}$ is relevant, a standard approach would compare the posterior probabilities of the model with and without $\{x_{it}\}$, and would include the variable if, for instance, the posterior probability of the model exhibiting the variable is at least 95%. Unfortunately, the MLL of the TVP_{ANN}-GARCH process is not analytically available. Therefore, we rely on the BIC to approximate it. Given a 95% acceptance threshold and the BIC approximation, the penalty parameter is set to $P = \ln \frac{0.05}{0.95T}$ (see Dufays

[4] using the fact that $q = \frac{a(1-\exp(P))}{a(1-\exp(P))+b\exp(b)}$,

and Rombouts (2018) for theoretical results and additional discussion about this choice). In the empirical examples, when the posterior probability that a model parameter lies within the wide component is greater than 50%, we consider that the related explanatory variable is relevant for modeling the conditional variance of the TVP_{ANN}-GARCH process.

While shrinkage priors are used to select the explanatory variables, other prior distributions are also required for the remaining model parameters. Table 2 summarizes our choices. Note that γ_{kj} and ω_i are driven by a 2MU(a,b,P) when we search for which explanatory variables to include in the TVP_{ANN}-GARCH process. When the selection has been made and the subset of explanatory variables included in the ANN is fixed, these parameters are distributed as standard normal distributions to obtain the number of hidden layers via the marginal likelihood.

4.2 Model estimation

We simulate the posterior distribution of the parameters by combining Markov-chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) methods. This technique is called the "SMC sampler" (see Del Moral et al. (2006)) and it has an important advantage over the standard MCMC approach when it comes to estimating ANN models. ANNs are recursive applications of non-linear functions and consequently, the posterior distribution of the ANN parameters is bound to exhibit a highly complex shape. Moreover, to select the ANN input variables, we use shrinkage priors, which leads to multi-modal posterior distributions. As shown by Jasra et al. (2007) and Herbst and Schorfheide (2014), MCMC methods based on a single Markov chain are less appropriate than SMC methods for simulating multi-modal distributions.

The SMC sampler also has some disadvantages, the most significant being the number of user-specified parameters to tune in order to run the algorithm. To address this issue, we rely on the tempered and time (TNT) algorithm (see Dufays (2016)), a variant of the SMC

sampler (Del Moral et al. (2006)) that automates the choice of SMC parameters and that is adapted to online estimations (i.e., updating the posterior distribution when new data are released without re-launching the algorithm). This latter feature is desirable for backtesting a model as we do in section 6.3. The algorithm sequentially iterates by producing realizations from the prior distribution to the posterior distribution by combining many importance sampling and MCMCs. The TVP_{ANN}-GARCH model is estimated by Algorithm 1. In particular, choosing $\underline{t} \leq T$, the algorithm starts by drawing from the posterior distribution of the model parameters given observations up to time $\underline{t}$ (i.e., the observation before the first required prediction). Then the sampler adapts the correction step to iterate over time by adding new observations one by one.

Algorithm 1 Online SMC algorithm

Sample M particles from the prior distribution: $\{\boldsymbol{\theta}^i\}_{i=1}^M$.
Set the tempered function $\phi = 0$ and t^* to the first time period of interest $t^* = \underline{t}$.
Set the normalized weights $W_i = 1/M \ \forall i \in [1, M]$ and $ESS = M$.
while $\phi < 1$ and $t^* < T$ **do**
  A - Correction step:
  if $\phi < 1$ **then**
    Find $\tilde{\phi} > \phi$ such that $M/(\sum_{i=1}^N \tilde{w}_i^2) = 0.95ESS$, where $\tilde{w}_i \propto W_i[f(y_{1:t^*}|\boldsymbol{\theta}^i)]^{\tilde{\phi}-\phi}$
    Set $\phi = \min(\tilde{\phi}, 1)$
    $\forall i \in [1, M]$; Set $w_i = W_i[f(y_{1:t^*}|\boldsymbol{\theta}^i)]^{\tilde{\phi}-\phi}$
  else
    Set $t^* = t^* + 1$ and $\forall i \in [1, M]$; Compute $\tilde{w}_i \propto W_i \frac{f(y_{1:t^*}|\boldsymbol{\theta}^i)}{f(y_{1:t^*-1}|\boldsymbol{\theta}^i)}$
  end if
  $\forall i \in [1, M]$; Compute the normalized weights $W_i = w_i / \sum_{j=1}^N w_j$ and $ESS = M/(\sum_{i=1}^N W_i^2)$
  B - Resample step if $ESS < 0.75M$
  Resample the particles by stratified sampling (Carpenter et al. (1999))
  C - MCMC step with targeted distribution: $\pi_\phi(\boldsymbol{\theta}|y_{1:t^*}) \propto [f(y_{1:t^*}|\boldsymbol{\theta})]^\phi f(\boldsymbol{\theta})$
  Apply N iterations of the MCMC sampler on the M particles (see Appendix D).
end while

At each SMC iteration, an MCMC algorithm is run to rejuvenate the particles. Typically, a standard random walk is used to update the particles in a Metropolis step and the variance of the proposal distribution is adapted at each SMC iteration using the current particles. However, as our posterior distribution is likely multi-modal, we prefer a more sophisticated

proposal distribution that is based on differential evolution updates. The complete Metropolis move is developed in Appendix D.

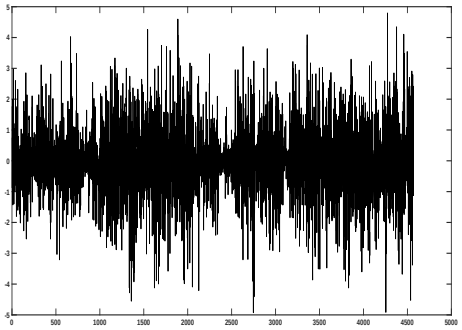
5 Simulations

In this section, we simulate two series of 4,605 observations from the TVP_{ANN}-GARCH model and investigate our estimation strategy for selecting the number of layers and input variables.[5] The first series is simulated from an TVP_{ANN}-GARCH process with a single-factor ANN structure exhibiting one hidden layer. The second series is generated by a data-generating process (DGP) that uses a multiple-factor ANN with two hidden layers. The ANN inputs consist of two variables derived from the S&P 500 series used in the empirical exercise (see Section 6.2). In particular, we consider the empirical variance of the S&P 500 over the previous five days as the first input, and the empirical variance over the previous 200 days as second input. Table 3 shows the parameter values of the two DGPs. Using these values, we generate one series per DGP. Figure 4 shows the generated time series as well as the corresponding long-term variance ($\bar{\omega}_t$) and persistence (ϕ_t).

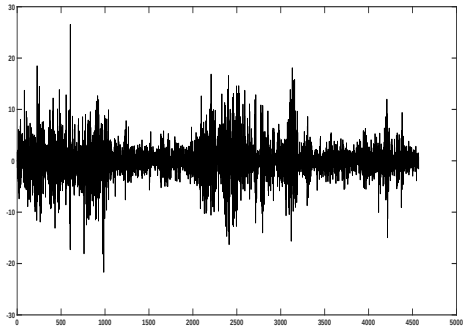
As expected, the two series are stationary and consistent with some stylized facts observed in financial time series, such as volatility clustering. The persistence parameter, ϕ_t, either smoothly varies over time or rapidly switches depending on the type of explanatory variables used as ANN inputs. In addition, the long-term variance $\bar{\omega}_t$ also varies over time. These parameter dynamics show that the model can produce a wide range of variations, from smooth variations typically observed in Spline-GARCH processes (see Engle and Rangel (2008)) or stochastic volatility models (e.g. Ghysels et al. (1996)) to abrupt and erratic behaviors generated by change-point and Markov-switching processes (e.g. Gray (1996) or He and Maheu (2009)).

We start our simulation study by empirically assessing the identification of the model parameters. To do so, we estimate the ANN parameters on the simulated conditional vari-

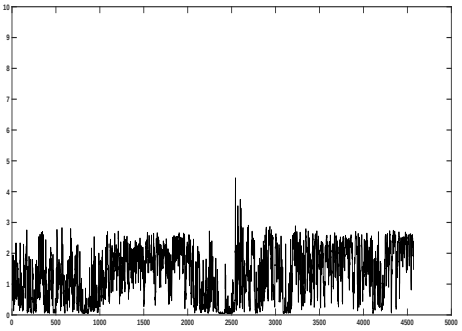
[5] The number of observations comes from the empirical exercise.



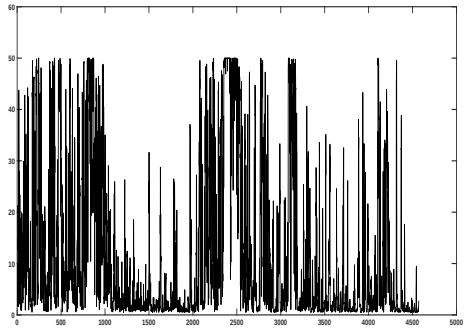
(a) 1 factor: Simulated returns



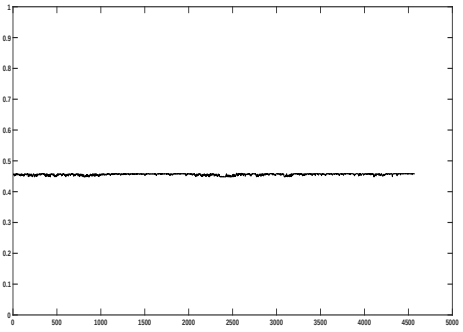
(b) Multiple factors: Simulated returns



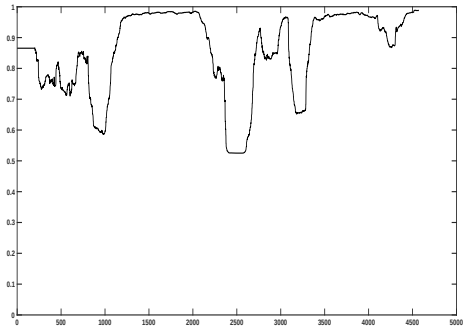
(c) 1 factor: $\bar{\omega}_t$



(d) Multiple factors: $\bar{\omega}_t$



(e) 1 factor: ϕ_t



(f) Multiple factors: ϕ_t

Figure 4 – Simulated series from the TVP_{ANN}-GARCH model with the single-factor structure (left) and with the multiple-factor structure (right). The first row shows the simulated returns, the second row shows the unconditional variance ($\bar{\omega}_t$), and the third row shows the persistence parameter (ϕ_t).

ances by minimizing the mean-squared errors.[6] Then, we compare our estimates to the true

[6] We try many different starting points, all leading to the same parameter values.

parameters. Table 4 displays the estimated parameters for the two series. We observe that the estimates are almost identical to the true parameters.

We now focus on estimating the parameters directly from the simulated log-returns to assess if our model selection strategy practically works. To do so, we use the SMC algorithm developed in Section 4 and assume five ANN inputs that are displayed in Figure 6 (see Appendix B). Among these five inputs, only the first two S&P 500 series are used to simulate the log-returns. In addition, as highlighted in Table 4, the first ANN input only impacts the long-term variance ($\bar{\omega}_t$), while the second input only triggers the persistence parameter. Table 13 shows the correlation between the five inputs. We observe that the correlation between true and spurious inputs can be relatively high. Table 5 provides the posterior probabilities of the parameters lying in the narrow uniform component of the 2MU distribution. The first two inputs have been perfectly detected as relevant explanatory variables. Moreover, in the multiple factor structure, Input 1 only triggers the unconditional variance parameter, while the second input varies the persistence parameter as set out in the DGP.

Once the explanatory variables have been selected, we find the best number of layers for the ANN using the marginal likelihood criterion. Practically, we estimate several TVP_{ANN}-GARCH models without shrinkage priors that differ in the number of hidden layers; then, we select the process that exhibits the highest MLL. Table 6 provides the MLLs for the different models. In both cases, we select the correct number of layers.

6 Empirical results

We now examine how the TVP_{ANN}-GARCH model performs when modeling a financial series. Section 6.1 discusses the financial data used in the empirical exercise, while Sections 6.2 and 6.3 discuss the TVP_{ANN}-GARCH in-sample results and compare the model's performance to that of standard flexible models. For all these empirical applications, we assume that the TVP_{ANN}-GARCH innovation follows a Normal distribution, that is $\eta_t \sim N(0, 1)$.

6.1 Data

We consider demeaned log-returns from five US bank stocks (Bank of America (BAC), Citigroup (C), Morgan Stanley (MS), Goldman Sachs (GS) and JP Morgan (JPM)), over the May 1999 to August 2017 period.[7] This data set consists of 4,605 observations. For explanatory variables, we use several transformations of five returns that we define below. We also consider the S&P 500 index and the Shiller index (due to the subprime crisis) as eligible inputs. Summary statistics of the series are presented in Table 7.

Estimating models with neural networks generally requires to select input variables. For each series, we use shrinkage priors to choose between nine input variables. As inputs, we consider lagged log-returns and empirical variances over the h previous periods. In particular, these inputs consist of the first lag of the dependent variable, the empirical variance of the dependent variable over 1, 22 and 100 periods, the first lag of the S&P 500 log-returns and its empirical variance over 22 periods, the first lag of the most relevant principal component (PC) of the other bank log-returns and its empirical variance over 22 periods, and the first lag of the Shiller index.

We motivate the set of potential inputs using a partial equilibrium argument. Let us denote p asset log-returns at time t by $\boldsymbol{y}_t = (y_{1,t}, y_{2,t}, \ldots, y_{p,t})'$ and the market log-return by r_t. Following Engle (2016), we consider the multivariate model given by,

$$\boldsymbol{z}_t | F_{t-1} = \begin{pmatrix} \boldsymbol{y}_t \\ r_t \end{pmatrix} | F_{t-1} \quad \sim \quad N(\mathbf{0}, \mathbf{H}_t), \tag{21}$$

in which the variance-covariance matrix H_t is measurable with respect to F_{t-1}, the information set adapted to the past of $\boldsymbol{z}_t \equiv (\boldsymbol{y}_t', r_t)'$. As shown in Engle (2016) and in Maheu and Shamsi (2019), the conditional expectation $E(\boldsymbol{y}_t | x_t, F_{t-1})$ is equal to the capital asset pricing model when the covariance matrix $\mathbf{H}_t$ is constant over time. Because the covariance matrix is empirically time-varying, Engle (2016) defines the dynamic conditional beta which

[7] The financial variables were downloaded from Yahoo! Finance.

directly depends on the conditional covariance matrix $\mathbf{H}_t$. Bali et al. (2017) show that the conditional beta improves return predictability with respect to the constant beta. Assuming that the returns follow a scalar BEKK model (e.g. Engle and Kroner, 1995), the conditional variance-covariance matrix is given by,

$$\mathbf{H}_{\mathrm{t}} = \mathbf{C}'\mathbf{C} + \mathbf{A}\boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}'\mathbf{A}' + \mathbf{B}\mathbf{H}_{t-1}\mathbf{B}', \tag{22}$$

where $\mathbf{C}'\mathbf{C}$ is a $(p+1)\times(p+1)$ is an upper triangular matrix such that $\mathbf{C}'\mathbf{C}$ is assumed to be semi-positive definite. From Equations (21)-(22), we can derive the marginal variance of $y_{i,t}$ with respect to F_{t-1} for $i = 1, \ldots, p$. Denoting $\boldsymbol{z}_{t,\neq i} = (y_{1,t}, \ldots, y_{i-1,t}, y_{i+1,t}, y_{p,t}, r_t)'$, it leads to

$$H_{y_i,t} \quad = \quad \omega(\boldsymbol{z}_{t-1,\neq i}, \ldots, \boldsymbol{z}_{1,\neq i}) + a_{ii}^2 y_{i,t-1}^2 + b_{ii}^2 H_{y_i,t-1}, \tag{23}$$

in which $\omega(\boldsymbol{z}_{t-1,\neq i}, \ldots, \boldsymbol{z}_{1,\neq i})$ highlights that the term non-linearly depends on $\boldsymbol{z}_{t-1,\neq i}, \ldots, \boldsymbol{z}_{1,\neq i}$ and a_{ii}, b_{ii} and $H_{y_i,t}$ denote the ith element of the diagonal of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{H}_{\mathrm{t}}$, respectively. The conditional variance given by Equation (23) shows that the GARCH parameters vary with past log-returns and with the lagged market returns. Also, other dynamics than the BEKK one would lead to other non-linear GARCH equations. Consequently, using ANNs to approximate the non-linear dynamic of the GARCH parameters (such as we do with our $\mathrm{TVP}_{\mathrm{ANN}}$-GARCH(1,1) model) allows to avoid the specification of the multivariate conditional variance-covariance matrix.

6.2 Estimation results

To determine the relevant input variable, each model estimation is carried out with these nine inputs, two hidden layers for the ANN structures, and shrinkage priors to identify the series that have predictive power on the conditional variance. Table 8 documents the selected

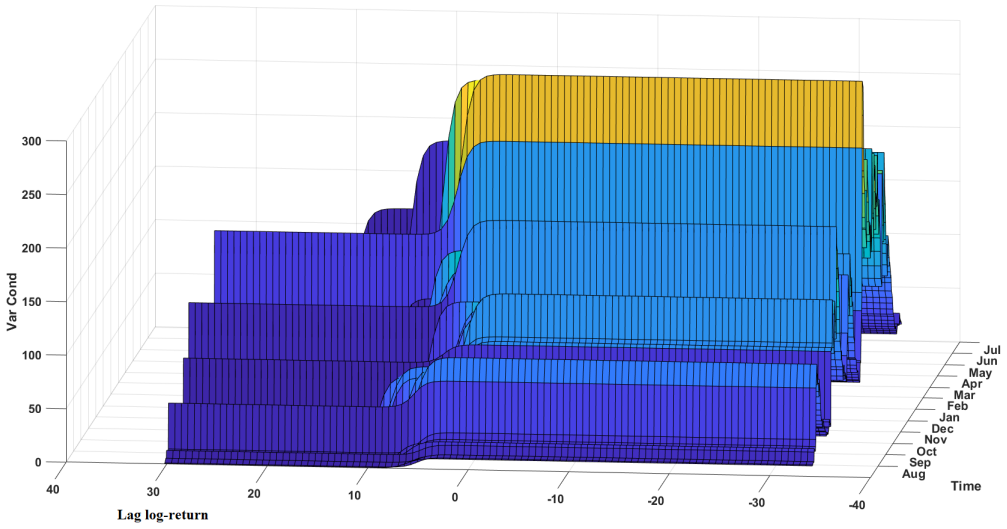
inputs, for each log-return and ANN structure. Interestingly, similar variables are selected for the bank returns.

Once the inputs are selected, we need to determine the optimal number of hidden layers. To do so, we rely on the MLL to select the optimal number of hidden layers as emphasized in the simulated exercise. For each bank, we consider TVP_{ANN}-GARCH models with 0 to 2 hidden layers (which corresponds to $\ell = 1$ to $\ell = 3$). Table 9 documents the MLLs of the TVP_{ANN}-GARCH models with different number of hidden layers. The best models exhibit a multiple factor structure and besides JP Morgan, all have one hidden layer. This suggests that ANN with hidden layers are relevant for generating flexible conditional variances.

We observe that multiple-factor ANNs always deliver higher MLLs than the single-factor structure, suggesting that multiple-factor ANNs may be superior. Additionally, MLLs rapidly decrease when more hidden layers are considered. As the best model for all the returns is a TVP_{ANN}-GARCH process with a multiple-factor structure, we will focus on this specification to investigate the ANN's impact on the conditional variance.

The selected inputs for the ANN can be large and the multiple-factor structure does not readily highlight how the conditional variance varies with respect to each inputs. We suggest one possible method to uncover the influence that inputs have on the conditional variance. As an example, Figure 5 investigates the effect of past log-returns on the conditional variance of Bank of America during the financial crisis. In particular, given the conditional variance generated by the maximum a posteriori parameters, we consider different values of the lagged log-returns and look at the impact on the conditional variance the day after. Interestingly, Figure 5 exhibits substantial asymmetries to negatives returns, suggesting that the lagged log-return input captures a kind of leverage effect. In addition to that, compared to standard leverage models like the GJR-GARCH process (see Glosten et al. (1993a)) in which the asymmetry is fixed around 0, the asymmetry already stands out for small positive log-returns highlighting that the threshold of asymmetry in standard models should be estimated (see, e.g., Caporin and McAleer (2006)).

Figure 5 – BAC: Sensitivity of conditional variances during the financial crisis (from July 2008 to July 2009) with respect to lag log-returns.



6.3 Comparison with existing models

As benchmarks, we consider the standard GARCH(1,1), the GJR-GARCH(1,1), the FC-GARCH processes with two components and the Markov-switching GARCH model with two regimes, see Appendix C for the model specifications. For a fair comparison with the other models, we only use past observations of the dependent variable as inputs for the TVP_{ANN}-GARCH model. By doing so, we can produce long-term forecasts without making assumptions about the future values of exogenous variables. In particular, we consider the past return, the past squared return and the empirical variance of the returns over 5, 22, 44 and 88 days (to mimic the HAR model of Corsi, 2009, and beyond). As explained in section 4, we first select the relevant inputs with the shrinkage prior and we then rely on the marginal log-likelihood for choosing the optimal number of layers. Regarding the FC-GARCH model of Medeiros and Veiga (2009), we use the past log-return for the state variable as suggested in their paper.

6.3.1 In-sample performance

In-sample fits are compared using the marginal log-likelihood (MLL). For each series, Table 10 documents the MLLs of all the competitors but the MS-GARCH model as the R package of Ardia et al. (2016) does not provide an estimate of it. The TVP_{ANN}-GARCH process always dominates the GARCH and the GJR-GARCH model. It also outperforms the FC-GARCH process five times out of the six series and is equivalent to it for the remaining series. The differences in MLLs are substantial when our process dominate since they range from 13 to 64 (i.e., GS and BAC series, respectively). According to Kass and Raftery (1995), these are strong evidence in favor of the TVP_{ANN}-GARCH model. Interestingly, the TVP_{ANN}-GARCH model with exogenous inputs given in Table 9 provides sizable fit improvements as their corresponding MLLs are always higher than the best model documented in Table 10.[8]

Table 10 – **In-sample comparison - marginal log-likelihoods**
This table reports the log-marginal likelihood for several volatility models. Bold values denote the best model according to the criterion.

Series	GARCH	GJR-GARCH	FC-GARCH	$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH
S&P 500	-6399	-6382	-6289	-6299	**-6256**
BAC	-9355	-9360	-9333	-9286	**-9269**
C	-9496	-9498	-9461	**-9436**	-9446
MS	-9479	-9483	-9470	**-9444**	-9449
GS	-9254	-9244	-9214	-9205	**-9201**
JPM	-10241	-10243	**-10208**	**-10208**	-10216

6.3.2 Out-of-sample performance

Regarding the forecasting exercise, we take two loss functions to gauge the $h-$step-ahead forecast performance: the root mean squared forecast error (RMSFE) and the cumulative log-predictive likelihood (CLPL). They are respectively given by

$$RMSFE(h) = \sqrt{\frac{1}{(T-h+1)-\underline{t}}\sum_{t=\underline{t}}^{T-h}(\frac{1}{h}[\sum_{j=1}^{h}y_{t+j}^2-\sum_{j=1}^{h}\hat{\sigma}_{t+j}^2])^2}, \tag{24}$$

[8] The MLL differences range from 10 to 32 in favor of the models in Section 6.2.

$$CLPL(h) \quad = \quad \sum_{t=\underline{t}}^{T-h} \log f(y_{t+h}|y_{1:t}), \tag{25}$$

where $f(y_{t+h}|y_{1:t}) = \int f(y_{t+1:t+h}|y_{1:t}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|y_{1:t}) d\boldsymbol{\theta} dy_{t+1:t+h-1}$ with $y_{t+1:t} = \{\emptyset\}$, $\hat{\sigma}^2_{t+j}$ is the median of the posterior distribution of $\sigma^2_{t+j}|y_{1:t}$ and $\underline{t}+1$ denotes the start of the out-of-sample forecast period. The predictions are computed via Monte Carlo integration. For instance, the predictive density is approximated by $f(y_{t+h}|y_{1:t}) \approx \frac{1}{M} \sum_{i=1}^{M} f(y_{t+h}|y^{(i)}_{t+1:t+h-1}, \boldsymbol{\theta}^{(i)}, y_{1:t})$ where M are the SMC particles, $\{y^{(i)}_{t+1:t+h-1}\}_{i=1}^{M}$ and $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{M}$ are draws from the posterior distributions $y_{1:t+h-1}|\boldsymbol{\theta}$ and $\boldsymbol{\theta}|y_{1:t}$, respectively.

Table 11 provides the RMSFEs and compares predictive performance with respect to the GARCH(1,1) model using the test of Diebold and Mariano (1995). Note that the TVP_{ANN}-GARCH process delivers accurate one-step-ahead forecast but is outperformed by the MS-GJR-GARCH process for longer horizons. Nevertheless, the TVP_{ANN}-GARCH model remains the second-best process because it dominates the FC-GARCH process and the two standard GARCH processes in 33 out of 36 predictions. In addition, the $\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH process significantly outperforms the GARCH model at short horizons (up to 5 days) for all but one series. Table 12 documents the CLPL and provides a statistical comparison with respect to the GARCH(1,1) model. Again, at short horizons, the TVP_{ANN}-GARCH process delivers most of the time the best predictive densities and repeatedly dominates the GARCH model significantly. While the MS-GJR-GARCH model is highly competitive, it turns out that the TVP_{ANN}-GARCH model still produces the best predictions for the JPM series at any horizon and at short horizons for all the other series but BAC.

These forecast results have been obtained without exogenous variables and without searching for the best transformation of the past log-returns to feed the ANNs. Consequently, we see these forecast results as promising and we believe that the TVP_{ANN}-GARCH process could dominate the MS-GARCH process uniformly if adequate exogenous variables were used.

Table 11 – **Forecasting exercise - Root mean squared forecast errors**
This table reports the root mean squared forecast errors defined in Equation (24) of several volatility models. The out-of-sample period covers approximately three years (i.e., $3 \times [22 \times 12] = 792$ observations) and spans from May 12, 2014 to June 30, 2017. Bold values denote the best RMSFE. Using the test of Diebold and Mariano (1995), a star indicates a significant improvement at a level of 10% with respect to the GARCH(1,1) model while two stars are used for a 5% significance level.

Horizon	1	2	5	10	15	20
			S&P 500			
GARCH	1.34	1.07	0.87	0.76	0.72	0.67
GJR-GARCH	1.33	1.05	0.87	0.76	0.72	0.68
FC-GARCH	1.28*	0.99**	0.82**	0.73**	0.69**	0.65*
MS-GARCH	1.33	1.03**	0.74**	0.55*	0.46*	0.40*
MS-GJR-GARCH	1.27**	**0.93****	**0.63****	**0.47***	**0.41***	**0.37***
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	1.29	1.01	0.83*	0.72**	0.67	0.63
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	**1.26****	0.97**	0.81**	0.71**	0.67**	0.62**
			BAC			
GARCH	5.72	4.50	3.36	2.77	2.55	2.43
GJR-GARCH	5.69**	4.47**	3.32**	2.71**	2.48*	2.37
FC-GARCH	**5.62****	4.39*	3.24**	2.65*	2.43	2.33
MS-GARCH	5.66	3.99**	2.71**	2.04**	1.72**	1.53**
MS-GJR-GARCH	5.63	**3.95****	**2.64****	**1.92****	**1.61****	**1.44****
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	5.72	4.51	3.39	2.79	2.58	2.48
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	5.62**	4.39*	3.21**	2.55**	2.28*	2.15*
			C			
GARCH	5.97	4.68	3.52	2.86	2.60	2.44
GJR-GARCH	5.94**	4.63**	3.45**	2.78**	2.51*	2.35
FC-GARCH	5.87*	4.55*	3.38**	2.73**	2.47*	2.34
MS-GARCH	6.00	4.25*	**2.60****	**1.87****	1.70*	1.64*
MS-GJR-GARCH	5.91	**4.19****	2.82**	1.99**	**1.61****	**1.40****
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	**5.80****	4.45**	3.28**	2.68	2.52	2.51
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	5.83**	4.49**	3.33*	2.75	2.58	2.56
			MS			
GARCH	4.06	3.03	2.16	1.74	1.56	1.46
GJR-GARCH	4.06	3.03	2.15	1.73	1.54	1.43
FC-GARCH	4.09	3.05	2.19	1.79	1.62	1.54
MS-GARCH	4.14	2.75**	1.87*	1.54	1.37	1.29
MS-GJR-GARCH	4.22	**2.74****	**1.82****	**1.50***	**1.35**	**1.28**
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	**4.03***	2.97*	2.10	1.75	1.66	1.67
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	4.09	3.05	2.15	1.70	1.50	1.39
			GS			
GARCH	4.32	3.35	2.43	1.98	1.82	1.72
GJR-GARCH	4.28**	3.30*	2.39	1.94	1.77	1.67
FC-GARCH	4.25*	3.28	2.37	1.93	1.77	1.66
MS-GARCH	4.34	**3.07****	2.08**	1.59**	1.39**	1.29*
MS-GJR-GARCH	4.25	3.12**	**2.08****	**1.52****	**1.30****	**1.17****
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	4.22**	3.24**	2.33**	1.92	1.83	1.81
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	**4.22****	3.23**	2.29**	1.82*	1.64	1.53
			JPM			
GARCH	6.66	5.01	3.69	3.02	2.71	2.53
GJR-GARCH	6.64*	4.98*	3.65*	2.97*	2.65	2.47
FC-GARCH	6.65	5.00	3.69	3.03	2.73	2.56
MS-GARCH	6.74	**4.50****	2.89**	2.35*	2.18	2.09
MS-GJR-GARCH	6.74	4.50**	**2.77****	**2.16****	**1.95***	**1.86**
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	**6.56****	4.88**	3.54**	2.90	2.65	2.57
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	6.60**	4.92**	3.55**	2.82**	2.45*	2.24*

Table 12 – **Forecasting exercise - Cumulative log-predictive likelihood**
This table reports the cumulative log-predictive likelihood defined in Equation (25) of several volatility models. The out-of-sample period covers approximately three years (i.e., $3 \times [22 \times 12] = 792$ observations) and spans from May 12, 2014 to June 30, 2017. Bold values denote the best CLPL. Using the test of Diebold and Mariano (1995), a star indicates a significant improvement at a level of 10% with respect to the GARCH(1,1) model while two stars are used for a 5% significance level.

Horizon	1	2	5	10	15	20
			S&P 500			
GARCH	-880.8	-897.7	-931.5	-936.3	-943.8	-936.1
GJR-GARCH	-900.1	-913.5	-958.8	-966.3	-982.2	-969.5
FC-GARCH	-853.9**	-864.6**	-908.8**	**-926.4****	-936.5**	-932.8**
MS-GARCH	-881.0**	-894.7**	-927.8**	-933.5**	-942.8**	-938.1**
MS-GJR-GARCH	-859.5**	-868.6**	-912.5**	-927.3**	**-935.1****	**-930.9****
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	-859.3**	-870.3**	-920.0**	-959.7	-999.9	-1025.6
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	**-840.5****	**-849.9****	**-905.0****	-931.3	-948.4	-948.8
			BAC			
GARCH	-1508.7	-1518.6	-1533.2	-1527.7	-1519.5	-1514.4
GJR-GARCH	-1509.1**	-1519.2	-1532.8	-1531.9	-1527.3	-1527.2
FC-GARCH	**-1500.6****	-1511.2**	-1526.8**	-1526.2	-1524.1	-1516.2
MS-GARCH	-1503.3**	-1510.3**	-1522.6**	-1519.7**	**-1514.3**	-1511.6
MS-GJR-GARCH	-1502.0**	**-1508.5****	**-1520.3****	**-1518.6****	-1516.7	-1515.7
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	-1511.4	-1524.1	-1524.8**	-1519.5**	-1514.7	**-1508.5**
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	-1509.1**	-1520.9	-1541.7	-1546.6	-1564.4	-1567.0
			C			
GARCH	-1434.1	-1450.9	-1471.4	-1466.7	-1455.6	-1449.4
GJR-GARCH	-1434.4	-1448.7	-1468.8	-1467.5	-1460.4	-1462.4
FC-GARCH	-1421.8**	-1433.4**	-1456.6**	-1457.4**	**-1453.1**	**-1448.2**
MS-GARCH	-1428.2**	-1445.5	-1466.8	-1473.5	-1475.9	-1470.0
MS-GJR-GARCH	-1422.8**	-1434.3**	**-1453.7****	**-1457.0****	-1455.0	-1460.5
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	**-1420.5****	**-1430.5****	-1457.9	-1475.7	-1489.4	-1503.7
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	-1424.1**	-1436.0**	-1464.7	-1474.1	-1473.8	-1476.8
			MS			
GARCH	-1384.5	-1390.8	-1399.6	-1392.1	-1384.7	**-1381.9**
GJR-GARCH	-1387.2	-1393.6	-1399.4	**-1391.6**	**-1384.4****	-1382.3
FC-GARCH	-1384.3**	-1389.0**	**-1398.2**	-1392.1	-1386.3	-1384.1
MS-GARCH	-1387.5	-1398.2	-1410.9	-1403.3	-1404.5	-1397.9
MS-GJR-GARCH	-1384.0	-1389.0	-1404.3	-1402.0	-1402.1	-1398.5
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	**-1371.7****	**-1380.2****	-1406.8	-1426.6	-1448.7	-1477.7
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	-1379.8**	-1383.9**	-1398.6**	-1397.8	-1396.9	-1393.8
			GS			
GARCH	-1331.3	-1347.0	-1374.5	-1372.5	-1362.6	-1353.2
GJR-GARCH	-1341.6	-1362.1	-1399.0	-1408.0	-1403.0	-1400.5
FC-GARCH	-1321.8**	-1336.0**	-1360.2**	-1363.2**	**-1356.8****	**-1349.2**
MS-GARCH	-1331.1	-1348.6	-1361.0	-1363.1	-1363.4	-1362.9
MS-GJR-GARCH	-1325.2**	-1336.0**	**-1354.3****	**-1360.1**	-1358.3	-1357.2
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	**-1310.1****	**-1324.2****	-1360.5**	-1378.2	-1383.4	-1388.1
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	-1314.6**	-1328.6**	-1360.7**	-1381.8	-1395.7	-1405.1
			JPM			
GARCH	-1508.4	-1518.4	-1537.9	-1533.3	-1521.2	-1510.6
GJR-GARCH	-1516.1	-1525.5	-1546.4	-1550.2	-1536.4	-1527.8
FC-GARCH	-1503.3**	-1510.8**	-1527.3**	-1525.8**	-1517.4**	-1510.3
MS-GARCH	-1505.1	-1516.8	-1545.0	-1542.2	-1539.6	-1540.7
MS-GJR-GARCH	-1500.6**	-1509.5	-1533.2	-1530.3	-1530.8	-1531.6
$\text{TVP}_{\text{ANN}(\ell,1)}$-GARCH	**-1492.4****	**-1500.2****	**-1523.3****	-1533.9	-1537.3	-1540.3
$\text{TVP}_{\text{ANN}(\ell,X)}$-GARCH	-1499.5**	-1507.0**	-1524.9**	**-1521.5****	**-1512.8****	**-1507.1**

Table 2 – Prior distributions for the TVP_{ANN}-GARCH parameters

Parameter	**Multiple-factor structure** $\forall i \in [1,d]$, $\forall j \in [1,l]$ and $k \in [1,2]$	**Single-factor structure**
Weight	$\ln \omega_{ij} \sim N(0,10)$	$\omega_i \sim N(0,10)$
Bias	$b_{ij} \sim N(0,10)$	$b_j \sim N(0,10)$
Output weight	$\gamma_{kj} \sim N(0,10)$	$\gamma_k \sim N(0,10)$
Output bias	$\gamma_{k0} \sim N(0,10)$	$\gamma_{k0} \sim N(0,10)$
		$\ln \delta_r \sim N(0,10)$, $\forall r \in [1,l-1]$
GARCH parameter	$\ln \frac{\alpha}{1-\alpha} \sim N(0,10)$	

N(a,b) represents a normal distribution with expectation a and variance b.

Table 3 – Parameter values of the simulated DGPs

TVP_{ANN}-GARCH model

$$y_t = \sigma_t \eta_t \quad \text{with } \eta_t \sim N(0,1)$$
$$\sigma_t^2 = \bar{\omega}_t + \phi_t(\sigma_{t-1}^2 - \bar{\omega}_t) + \alpha(y_{t-1}^2 - \sigma_{t-1}^2), \quad \text{where } \phi_t = \alpha + (1-\alpha)\tilde{\beta}_t$$

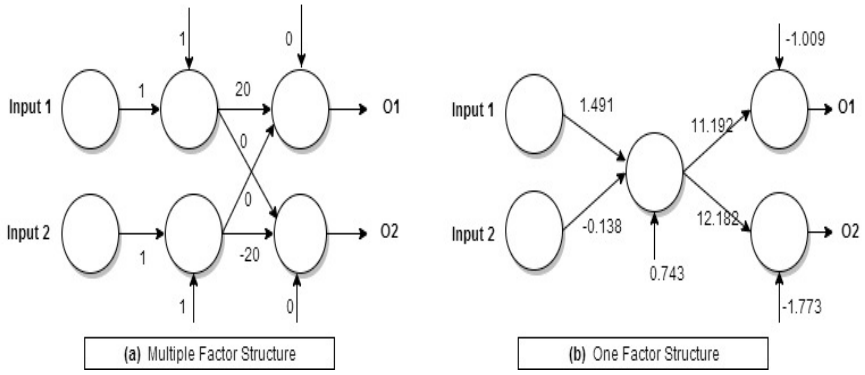
Parameter values for DGP with Single-factor structure for the ANN

$$\boldsymbol{\theta} = (\omega_{1:7}, \alpha = 0.05)$$
$$\omega_{1:7} = (7.31, 0.08, 0.24, 0.24, 0.15, -0.64, 0.04)$$

Parameter values for DGP with multiple-factor structure for the ANN

$$\boldsymbol{\theta} = (\omega_{1:10}, \alpha = 0.05)$$
$$\omega_{1:10} = (1, 1, 1, 1, 0, 0, 20, 0, 0, -20)$$

Graphical illustration



Notes: $\omega_{1:p} = (\omega_1, \cdots, \omega_p)$. Input 1 is the empirical variance of the S&P 500 over the previous five days and input 2 is the empirical variance of the S&P 500 over the previous 200 days.

Table 4 – Parameter estimates of the simulated series given the conditional variance.

Multiple-factor structure			Single-factor structure		
	True Par	Est Par		True Par	Est Par
ω_1	1	1.00	ω_1	7.31	7.31
ω_2	1	1.00	ω_2	0.08	0.08
ω_3	1	1.00	ω_3	0.24	0.24
ω_4	1	1.00	ω_4	0.24	0.24
ω_5	0	0.00	ω_5	0.15	0.15
ω_6	0	-0.00	ω_6	-0.64	-0.64
ω_7	20	20.00	ω_7	0.04	0.04
ω_8	0	0.00	α	0.05	0.05
ω_9	0	0.00			
ω_{10}	-20	-20.00			
α	0.05	0.05			

Estimation of the ANN parameters given the true conditional variance. True parameters and estimated parameters are denoted by True Par and Est Par, respectively.

Table 5 – Posterior probability of the spike components

	Multiple-factor structure		Single-factor structure
	Output 1	Output 2	
Input 1	**0**	0.99	**0**
Input 2	0.99	**0**	**0.01**
Input 3	0.97	0.99	0.95
Input 4	0.99	0.95	0.96
Input 5	0.99	0.97	0.93

Posterior probabilities of the parameters related to the explanatory variables of being sampled from the 2MU narrow component. We include inputs if the probability is less than 0.5.

Table 6 – Marginal log-likelihood (MLL) of the TVP_{ANN}-GARCH process exhibiting different hidden layers

Multiple-factor structure		Single-factor structure	
No. of hidden layers	MLL	No. of hidden layers	MLL
1	-11097.47	1	**-6394.53**
2	**-11074.16**	2	-6396.05
3	-11082.79	3	-6398.37

Table 7 – Daily demeaned log-returns: Descriptive statistics

Series	Std. dev	Skewness	Kurtosis
BAC	2.994	-0.326	28.56
C	3.162	-0.527	42.81
MS	3.179	0.302	14.34
GS	2.406	0.264	15.87
WFC	2.446	1.288	48.28
S&P 500	1.224	-0.193	11.25

Notes: The table reports the sample standard deviation, the sample skewness, and the sample kurtosis for each series.

Table 8 – ANN input selection using shrinkage priors

Predictor	Multiple-factor structure					Single-factor structure				
	BAC	C	MS	GS	JPM	BAC	C	MS	GS	JPM
Lag. ret	√	√	√		√	√	√	√	√	√
Lag sq. ret 1	√	√				√				
Lag sq. ret 22	√	√				√				
Lag sq. ret. 100	√	√	√	√	√	√	√	√	√	√
S&P 500 lag. ret.	√	√	√	√		√	√	√	√	√
S&P 500 lag. sq. ret. 22										
PC lag. ret.	√	√	√	√	√	√		√	√	
PC lag. sq. ret. 22	√			√		√				√
Shiller Index			√	√		√				

A sign √ indicates a posterior probability of the 2MU spike component that is less than 0.5. In such a case, we include the explanatory variable.

Table 9 – Marginal log-likelihoods (MLLs) of the TVP_{ANN}-GARCH model exhibiting a different number of hidden layers

	Number of Hidden Layers		
Series	ℓ =1	$\ell = 2$	$\ell = 3$
	Single-factor structure		
BAC	-9264.26	-9265.42	-9269.65
C	-9428.68	-9432.30	-9436.44
MS	-9430.55	-9432.50	-9433.74
GS	-9203.36	-9207.86	-9211.90
JPM	-10201.05	-10204.38	-10209.18
	Multiple-factor structure		
BAC	-9306.29	**-9250.60**	-9266.98
C	-9453.23	**-9426.13**	-9439.78
MS	-9469.87	**-9417.38**	-9432.14
GS	-9259.55	**-9168.57**	-9180.66
JPM	**-10198.31**	-10201.11	-10211.51

For each US bank daily log-returns, the highest MLL is bolded.

7 Conclusion

In this paper, we propose a new time-varying parameter volatility model, called the time-varying parameter GARCH process using artificial neural networks (i.e., TVP_{ANN}-GARCH), in which parameters are driven by a (deep) neural network. We derive the stationarity conditions of the process as well as the conditions for the existence of unconditional moments. Parameter identification is also discussed. Using ANNs makes the variance dynamic highly flexible and allows for the possibility of including additional explanatory variables. Also, the likelihood function of the model is tractable without resorting to any filtering procedure, which simplifies the parameter estimation. We propose a Sequential Monte Carlo sampler to simulate the posterior distribution of the model parameters and we use shrinkage priors to select the relevant inputs of the ANNs. Once the inputs are chosen, we rely on the marginal likelihood for selecting the number of layers of the ANN. A detailed simulation study illustrates the performance of the procedure.

The paper ends with an empirical study using the log-returns of five US bank stocks and of the S&P 500 financial index. These series span from May 1999 to August 2017. In-sample comparisons show that the TVP_{ANN}-GARCH model performs better than several flexible models (GARCH, GJR-GARCH and FC-GARCH). The TVP_{ANN}-GARCH process also produces superior predictions than those models in terms of root mean squared forecast errors and log-predictive densities. In addition, the TVP_{ANN}-GARCH process compares favorably with the MS-GJR-GARCH model for short-term forecasts (up to five days). These results are promising because the TVP_{ANN}-GARCH predictions exhibit room of improvements by changing the ANN inputs.

We see this work as an encouraging step toward using artificial neural networks to model the dynamics of some model parameters. Several related research avenues can be pursued. For instance, one can investigate another type of ANN (such as a recurrent network) or extend this time-varying framework to any other standard model with fixed parameters.

References

Anderson, H. M., Nam, K., and Vahid, F. (1999). Asymmetric nonlinear smooth transition garch models. In *Nonlinear time series analysis of economic and financial data*, pages 191–207. Springer.

Ardia, D., Bluteau, K., Boudt, K., Catania, L., and Trottier, D.-A. (2016). Markov-switching garch models in r: The msgarch package. *Journal of Statistical Software, Forthcoming*.

Bali, T. G., Engle, R. F., and Tang, Y. (2017). Dynamic conditional beta is alive and well in the cross section of daily stock returns. *Management Science*, 63(11):3760–3779.

Bauwens, L., Dufays, A., and Rombouts, J. (2013). Marginal likelihood for markov switching and change-point garch models. *Journal of Econometrics*, 178 (3):508–522.

Bauwens, L., Preminger, A., and Rombouts, J. (2010). Theory and inference for a Markov-switching GARCH model. *Econometrics Journal*, 13:218–244.

Bauwens, L. and Storti, G. (2009). A component garch model with time varying weights. *Studies in Nonlinear Dynamics and Econometrics*, 13:2, Article 1.

Bhadra, A., Datta, J., Polson, N. G., Willard, B., et al. (2019). Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

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

Broto, C. and Ruiz, E. (2004). Estimation methods for stochastic volatility models: A survey. *Journal of Economic Surveys*, 18(5):613–649.

Caporin, M. and McAleer, M. (2006). Dynamic asymmetric garch. *Journal of Financial Econometrics*, 4(3):385–412.

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

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Caulet, R. and Peguin-Feissolle, A. (2000). Un test d'hétéroscédasticité conditionnelle inspiré de la modélisation en termes de réseaux neuronaux artificiels. *Annales d'Economie et de Statistique*, pages 177–197.

Chan, J. C. and Eisenstat, E. (2018). Bayesian model comparison for time-varying parameter vars with stochastic volatility. *Journal of Applied Econometrics*, 33(4):509–532.

Christen, J. A. and Fox, C. (2010). A general purpose sampling algorithm for continuous distributions (the t-walk). *Bayesian Analysis*, 5(2):263–281.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Das, S. and Suganthan, P. (2011). Differential evolution: A survey of the state-of-the-art. *EEE Transactions on Evolutionary Computation*, 15(1):4–31.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *The Royal Statistical Society: Series B(Statistical Methodology)*, 68:411–436.

Demuth, H. B., Beale, M. H., De Jess, O., and Hagan, M. T. (2014). *Neural network design*. Martin Hagan.

Deschamps, P. J. (2008). Comparing smooth transition and Markov switching autoregressive models of US unemployment. *Journal of Applied Econometrics*, 23(4):435–462.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

Dieobold, F. X. (1986). Modeling the persistence of conditional variances: A comment. *Econometric Reviews*, 5(1):51–56.

Donaldson, R. G. and Kamstra, M. (1997). An artificial neural network-garch model for international stock return volatility. *Journal of Empirical Finance*, 4(1):17–46.

Dueker, M. J., Psaradakis, Z., Sola, M., and Spagnolo, F. (2011). Contemporaneous-threshold smooth transition garch models. *Studies in Nonlinear Dynamics & Econometrics*, 15(2).

Dufays, A. (2016). Evolutionary sequential Monte Carlo for change-point models. *Econometrics*, 4(1).

Dufays, A. and Rombouts, J. (2020). Relevant parameter changes in structural break models. *forthcoming in Journal of Econometrics*.

Dufays, A. and Rombouts, J. V. (2018). Sparse change-point HAR models for realized variance. *Econometric Reviews*, pages 1–24.

Engle, R. and Kroner, F. (1995). Multivariate simultaneous generalized arch. *Econometric Theory*, 11:122–150.

Engle, R. and Rangel, J. (2008). The Spline-GARCH model for low-frequency volatility and its global macroeconomic causes. *Review of Financial Studies*, 21:1187–1222.

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

Engle, R. F. (2016). Dynamic conditional beta. *Journal of Financial Econometrics*, 14(4):643–667.

Engle, R. F., Ghysels, E., and Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics*, 95(3):776–797.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2013). emcee: The MCMC hammer. *PASP*, 125:306–312.

Francq, C., Roussignol, M., and Zakoian, J.-M. (2001). Conditional heteroskedasticity driven by hidden Markov chains. *Journal of Time Series Analysis*, 22:197–220.

Francq, C. and Zakoian, J.-M. (2008). Deriving the autocovariances of powers of markov-switching garch models, with applications to statistical inference. *Computational Statistics & Data Analysis*, 52(6):3027–3046.

Fruhwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and markov-switching models using bridge sampling techniques. *Econometrics Journal*, 7:143–167.

Ghysels, E., Harvey, A., and Renault, E. (1996). Stochastic volatility. In Maddala, G. and Rao, C., editors, *Handbook of Statistics*, pages 119–191. Elsevier Science, Amsterdam.

Glosten, L., Jagannathan, R., and Runkle, D. (1993a). On the relation between expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48:1779–1801.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993b). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, 48(5):1779–1801.

González-Rivera, G. (1998). Smooth-transition garch models. *Studies in Nonlinear Dynamics & Econometrics*, 3(2).

Gray, S.-F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42:27–62.

Griffin, J. and Brown, P. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5:171–188.

Haas, M., Mittnik, S., and Paolella, M. (2004a). A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics*, 2:493–530.

Haas, M., Mittnik, S., and Paolella, M. S. (2004b). A new approach to markov-switching garch models. *Journal of financial econometrics*, 2(4):493–530.

Hagerud, G. E. et al. (1997). *A smooth transition ARCH model for asset returns*. Stockholm School of Economics, the Economic Research Inst.

He, Z. and Maheu, J. M. (2009). Real time detection of structural breaks in garch models. *Computational Statistics & Data Analysis, forthcoming*.

He, Z. and Maheu, J. M. (2010). Real time detection of structural breaks in garch models. *Computational Statistics & Data Analysis*, 54(11):2628–2640.

Herbst, E. and Schorfheide, F. (2014). Sequential Monte Carlo sampling for DSGE models. *Journal of Applied Econometrics*, 29(7):1073–1098.

Hillebrand, E. (2005). Neglecting parameter changes in GARCH models. *Journal of Econometrics*, 129(1):121–138.

Hwang, J. G. and Ding, A. A. (1997). Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, 92(438):748–757.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Jasra, A., Stephens, D. A., and Holmes, C. C. (2007). On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kılıç, R. (2011). Long memory and nonlinearity in conditional variances: A smooth transition figarch model. *Journal of Empirical Finance*, 18(2):368–378.

Kim, C.-J., Nelson, C. R., et al. (1999). State-space models with regime switching: Classical and Gibbs-sampling approaches with applications. *MIT Press Books*, 1.

Krolzig, H.-M. (2013). *Markov-switching vector autoregressions: Modelling, statistical inference, and application to business cycle analysis*, volume 454. Springer Science & Business Media.

Lamoureux, C. G. and Lastrapes, W. D. (1990). Persistence in variance, structural change, and the GARCH model. *Journal of Business & Economic Statistics*, 8(2):225–234.

Lanne, M. and Saikkonen, P. (2005). Non-linear garch models for highly persistent volatility. *The Econometrics Journal*, 8(2):251–276.

Luukkonen, R., Saikkonen, P., and Terasvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika*, 75(3):491–499.

Maheu, J. M. and Shamsi, A. (2019). Nonparametric dynamic conditional beta. *Journal of Financial Econometrics*.

McAleer, M. and Medeiros, M. C. (2008). A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. *Journal of Econometrics*, 147(1):104–119.

Medeiros, M. C. and Veiga, Á. (2005). A flexible coefficient smooth transition time series model. *IEEE Transactions on Neural Networks*, 16(1):97–113.

Medeiros, M. C. and Veiga, A. (2009). Modeling multiple regimes in financial volatility with a flexible coefficient garch (1, 1) model. *Econometric Theory*, 25(1):117–161.

Miazhynskaia, T., Frühwirth-Schnatter, S., and Dorffner, G. (2008). Neural network models for conditional distribution under bayesian analysis. *Neural computation*, 20(2):504–522.

Nelson, D. B. (1990). Stationarity and persistence in the GARCH (l, l) model. *Econometric Theory*, 6(3):318–334.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Peluso, S., Chib, S., and Mira, A. (2016). Semiparametric multivariate and multiple change-point modelling. *Working Paper*.

Ročková, V. and George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Scharth, M. and Medeiros, M. C. (2009). Asymmetric effects and long memory in the volatility of Dow Jones stocks. *International Journal of Forecasting*, 25(2):304–327.

Sussmann, H. J. (1992). Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural networks*, 5(4):589–593.

Taylor, S. J. (2008). *Modelling financial time series*. world scientific.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., and Higdon, D. (2009). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptative randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulations*, 10:271–288.

A Proofs

Proof of Theorem 1. We can write the conditional variance σ_t^2 in (2) as:

$$\sigma_t^2 = (1-\alpha)\bar{\omega}_t(1-\tilde{\beta}_t) + \left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2\right]\sigma_{t-1}^2.$$

Consider $c_{t-1} = (1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2)$ so that

$$\sigma_t^2 = (1-\alpha)\bar{\omega}_t(1-\tilde{\beta}_t) + c_{t-1}\sigma_{t-1}^2.$$

Rearranging this last equation, σ_t^2 can be written as:

$$\sigma_t^2 = (1-\alpha)\bar{\omega}_t(1-\tilde{\beta}_t) + \sum_{k=1}^{t-1}\prod_{j=0}^{k-1}(1-\alpha)\bar{\omega}_{t-k}(1-\tilde{\beta}_{t-k})c_{t-1-j} + \prod_{j=0}^{t-1}c_{t-1-j}\sigma_0^2 \qquad (26)$$

where $\bar{\omega}_t > 0 \quad \forall t$, $\alpha < 1$ and $\tilde{\beta}_t$ is bounded between 0 and 1 for all t. Therefore, $\sigma_0^2 > 0$ with probability 1, and there is a positive and finite constant M such that $(1-\alpha)\bar{\omega}_t(1-\tilde{\beta}_t) > M \quad \forall t$. Because of this, we have

$$\sigma_t^2 \geq M\left[\underset{1\leq k\leq t-1}{Sup}\prod_{j=0}^{k}c_{t-1-j}\right].$$

Since $\eta_t \sim IID(0,1)$, it follows that $\{\eta_t\}$ is strictly stationary and ergodic. Using this information and considering that $\tilde{\beta}_t$ is bounded, it is easy to show that the sequence $\{c_t\}$ is strictly stationary and ergodic with $\mathbb{E}\left[|c_t|^{1+\delta}\right] < \infty \quad \forall t$, and for any arbitrary δ close to zero. We then follow steps identical to those of the proof of Theorem 2 in Nelson (1990) to conclude about the strict stationarity and ergodicity of σ_t^2. Finally, as $y_t = \sigma_t\eta_t$, we use the

fact that the product of strictly stationary and ergodic processes is also strictly stationary and ergodic to conclude that y_t is strictly stationary and ergodic. □

Proof of Theorem 2. We have that $y_t^{2k} = \sigma_t^{2k}\eta_t^{2k}$. Since $\mathbb{E}[\eta_t^{2k}] < \infty$ by assumption, it follows that $\mathbb{E}[y_t^{2k}] < \infty$ if and only if $\mathbb{E}[\sigma_t^{2k}] < \infty$. The binomial expansion of $\sigma_t^{2k} = \left[(1-\alpha)\bar{\omega}_t(1-\tilde{\beta}_t) + \left((1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2\right)\sigma_{t-1}^2\right]^k$ is given by:

$$\sigma_t^{2k} = \sum_{p=0}^{k} \binom{k}{p} \left[(1-\alpha)\bar{\omega}_t(1-\tilde{\beta}_t)\right]^p \left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2)\right]^{k-p} \sigma_{t-1}^{2(k-p)}.$$

As in Medeiros and Veiga (2009), let $\mathbf{u}_t = \left[\sigma_t^{2k}, \sigma_t^{2(k-1)}, ..., \sigma_t^2\right]$. Then,

$$\mathbf{u}_t = \mathbf{a}_t + \mathbf{C}_{t-1}\mathbf{u}_{t-1} \tag{27}$$

where $\mathbf{a}_t \in \mathbb{R}^k$ is a vector of specific elements given by $a_{t,j} = \left[(1-\alpha)\bar{\omega}_t(1-\tilde{\beta}_t)\right]^j$, $j = 1, \ldots, k$ and $\mathbf{C}_{t-1}$ is a $k \times k$ upper triangular matrix with diagonal elements given by

$$\mathbf{C}_{t-1,jj} = \left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2)\right]^j \quad j = 1, \ldots, k \tag{28}$$

Let us denote $\mathscr{I}_{t-2}$ as the set of all information available up to time $t-2$. Considering the fact that $a_{t,j}$ is bounded (with the assumption that $\bar{\omega}_t$ is bounded), the conditional expectation of $\mathbf{u}_t$ in equation (27) satisfies the inequality $\mathbb{E}\left[\mathbf{u}_t|\mathscr{I}_{t-2}\right] \leq cte + \mathbb{E}\left[\mathbf{C}_{t-1}|\mathscr{I}_{t-2}\right]\mathbf{u}_{t-1}$.

One can therefore use the same arguments as Medeiros and Veiga (2009) and the iterated expectations formula to conclude the proof.

□

Proof of Theorem 3.

Proof of minimality with a multiple-factor structure for the neural network

(a) Case where $\ell = 1$: Trivial under Assumption 1.

(b) Case where $\ell = 2$: With $\ell = 2$ hidden layers, $\tilde{x}_{ti} = f(\omega_{i2}f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2})$, $i \in \{1, \cdots, d\}$. To prove that under (R.1) and (R.2), the network with $\ell = 2$ hidden layers is minimal, we need to show that for all $i \in \{1, \cdots, d\}$, $\tilde{x}_{ti}$ cannot be obtained with a neural network containing $\ell = 1$ hidden layers. Put another way, consider two real numbers, α_1 and

α_2. We have to show that

$$\alpha_1 f(\omega_{i1}x_{ti} + b_{i1}) + \alpha_2 f(\omega_{i2} f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2}) = 0 \tag{29}$$

if and only if $\alpha_1 = \alpha_2 = 0$. Suppose that equation (29) holds. We prove that $\alpha_1 = \alpha_2 = 0$.

Set $\bar{X}_{ti}^{(1)} = \omega_{i1}x_{ti} + b_{i1}$ and $\bar{X}_{ti}^{(2)} = \omega_{i2} f(\bar{X}_{ti}^{(1)}) + b_{i2}$. Since f is strictly increasing, $|\bar{X}_{ti}^{(1)}(\omega) - f(\bar{X}_{ti}^{(1)}(\omega))| \neq 0 \ \forall \omega \in \Omega$ where Ω is the sampling space of the explanatory variables x_{ti} ($i \in \{1, \cdots, d\}$) which are assumed to be strictly stationary and ergodic. Equation (29) can be written as

$$\alpha_1 f(\bar{X}_{ti}^{(1)}) + \alpha_2 f(\bar{X}_{ti}^{(2)}) = 0. \tag{30}$$

The nonlinearity of f, together with the fact that $\bar{X}_{ti}^{(2)}$ is a linear transformation of $f(\bar{X}_{ti}^{(1)})$ which is different to $\bar{X}_{ti}^{(1)}$ for all the sampling space allow to conclude that under restrictions (R.1) and (R.2), $\bar{X}_{ti}^{(1)}$ and $\bar{X}_{ti}^{(2)}$ will be different for almost every points of the sampling space i.e., $P(\bar{X}_{ti}^{(1)}(\omega) = \bar{X}_{ti}^{(2)}(\omega)) \neq 1 \ \forall t$ and $\forall \omega \in \Omega$. Consequently, $f(\bar{X}_{ti}^{(1)}) \neq f(\bar{X}_{ti}^{(2)})$ for all t and $\alpha_1 f(\bar{X}_{ti}^{(1)}) + \alpha_2 f(\bar{X}_{ti}^{(2)}) = 0$ if and only if $\alpha_1 = \alpha_2 = 0$.

(c) Case where $\ell = 3$: With $\ell = 3$ hidden layers, $\tilde{x}_{ti} = f(\omega_{i3} f(\omega_{i2} f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2}) + b_{i3})$. Using the same approach as when $\ell = 2$, let α_1, α_2 and α_3 be three real numbers and suppose that

$$\alpha_1 f(\omega_{i1}x_{ti} + b_{i1}) + \alpha_2 f\left(\omega_{i2} f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2}\right) + \alpha_3 f\left(\omega_{i3} f(\omega_{i2} f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2}) + b_{i3}\right) = 0. \tag{31}$$

We prove that $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

Again, let $\bar{X}_{ti}^{(1)} = \omega_{i1}x_{ti} + b_{i1}$, $\bar{X}_{ti}^{(2)} = \omega_{i2} f(\bar{X}_{ti}^{(1)}) + b_{i2}$ and $\bar{X}_{ti}^{(3)} = \omega_{i3} f(\bar{X}_{ti}^{(2)}) + b_{i3}$. Note that $\bar{X}_{ti}^{(1)}$, $\bar{X}_{ti}^{(2)}$ and $\bar{X}_{ti}^{(3)}$ are strictly stationary and ergodic since explanatory variables x_{ti}, $i \in \{1, \cdots, d\}$ are strictly stationary and ergodic. Equation (31) can be written as

$$\alpha_1 f(\bar{X}_{ti}^{(1)}) + \alpha_2 f(\bar{X}_{ti}^{(2)}) + \alpha_3 f(\bar{X}_{ti}^{(3)}) = 0. \tag{32}$$

With restrictions (R.1) and (R.2) in mind and considering the fact that f is strictly increasing, it is easy to see that $\bar{X}_{ti}^{(1)} \neq \bar{X}_{ti}^{(2)}$ and $\bar{X}_{ti}^{(2)} \neq \bar{X}_{ti}^{(3)}$. Moreover, the logistic function f is

nonlinear and cannot be approximated by a polynomial function of degree lower than two. Consequently, $\bar{X}_{ti}^{(3)} \neq \bar{X}_{ti}^{(1)}$.

The fact that $\bar{X}_{ti}^{(1)} \neq \bar{X}_{ti}^{(2)} \neq \bar{X}_{ti}^{(3)}$ implies that $f(\bar{X}_{ti}^{(1)}) \neq f(\bar{X}_{ti}^{(2)}) \neq f(\bar{X}_{ti}^{(3)})$ for all t since f is an strictly increasing function (the logistic function). These results, combined with the linearity of equation (32) and the fact that f is nonlinear imply that equation (32) holds if and only if $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

Proof of global identification with multiple-factor structure for the neural network

(a) Case where $\ell = 1$: A typical artificial neural network with $\ell = 1$ hidden layers and a multiple-factor structure gives rise to the outputs

$$O_k = f\left(\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i1} x_{ti} + b_{i1})\right) \quad k \in \{1, 2\}$$

where O_1 is the first output of the network and O_2 is the second output. Let us use $\boldsymbol{\theta}$ to denote the vector of parameters of the network. Suppose that $\tilde{\boldsymbol{\theta}}$ is another vector of parameters such that

$$f\left(\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i1} x_{ti} + b_{i1})\right) = f\left(\tilde{\gamma}_{k0} + \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i1} x_{ti} + \tilde{b}_{i1})\right), \quad k \in \{1, 2\}. \tag{33}$$

Since the function f is a strictly increasing function, equation (33) is equivalent to

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i1} x_{ti} + b_{i1}) - \tilde{\gamma}_{k0} - \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i1} x_{ti} + \tilde{b}_{i1}) = 0, \qquad k \in \{1, 2\}. \tag{34}$$

To prove global identifiability of the model, we need to show that, under restrictions (R.1) and (R.2), equation (34) is satisfied if and only if $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

Equation (34) can be rewritten as

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f((\boldsymbol{\beta}_i)^t \boldsymbol{x}_t + b_{i1}) - \tilde{\gamma}_{k0} - \sum_{i=1}^{d} \tilde{\gamma}_{ik} f((\tilde{\boldsymbol{\beta}}_i)^t \boldsymbol{x}_t + \tilde{b}_{i1}) = 0 \quad k \in \{1, 2\} \tag{35}$$

so that $\left\{(\boldsymbol{\beta}_i)^t, (\tilde{\boldsymbol{\beta}}_i)^t, b_{i1}, \tilde{b}_{i1}\right\}_{i=1}^{d}$ are different for index $i's$. In particular, we have that $\boldsymbol{x}_t = (x_{t1}, \cdots, x_{td})^t$ and $\boldsymbol{\beta}_i = (0, \cdots, \omega_{i1}, \cdots, 0)^t$ is a $(d \times 1)$ vector with ω_{i1} at the ith position

and 0 else where. Similarly, we fix $\tilde{\boldsymbol{\beta}}_i = (0, \cdots, \omega_{i1}, \cdots, 0)^t$.

Following Hwang and Ding (1997) (proof of Theorem 2.3), we reduce $\boldsymbol{x}_t$ in (35) to a one dimension variable. Let $\boldsymbol{V}$ be a unit vector such that the projection of distinct $\boldsymbol{\beta}_i$ and $\tilde{\boldsymbol{\beta}}_i$ are distinct. Note that the vector $\boldsymbol{V}$ exists because the sets $\{\boldsymbol{\beta}_i\}$ and $\{\tilde{\boldsymbol{\beta}}_i\}$ have only finite many points. Let $\mu_i = (\boldsymbol{\beta}_i)^t\boldsymbol{V}$ and $\tilde{\mu}_i = (\tilde{\boldsymbol{\beta}}_i)^t\boldsymbol{V}$ be the projections. Then replacing $\boldsymbol{x}_t$ in (35) by $y_t\boldsymbol{V}$ with $y_t \in \mathbb{R}$ lead to

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\mu_i y_t + b_{i1}) - \tilde{\gamma}_{k0} - \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\mu}_i y_t + \tilde{b}_{i1}) = 0, \quad k \in \{1, 2\}. \tag{36}$$

Let $\phi_j = \mu_j y_t + b_{j1}$ for $j = 1, \cdots, d$ and $\phi_j = \tilde{\mu}_{j-d} y_t + \tilde{b}_{j-d,1}$ for $j = d+1, \cdots, 2d$. Lemma 2.7 in Hwang and Ding (1997) implies that if ϕ_{j1} and ϕ_{j2} are not sign-equivalent for each $j_1 \in \{1, \cdots, 2d\}$, $j_2 \in \{1, \cdots, 2d\}$ (i.e., $|\phi_{j_1}| \neq |\phi_{j_2}|$), equation (36) is satisfied if and only if γ_{k0}, $\tilde{\gamma}_{k0}$, γ_{ik} and $\tilde{\gamma}_{ik}$ vanish jointly for every $i \in \{1, \cdots, d\}$ and $k \in \{1, 2\}$. But Restriction (R.2) of Assumption 1 precludes that possibility. Therefore, ϕ_{j_1} and ϕ_{j_2} must be sign-equivalent. But Restriction (R.1) of Assumption 1 precludes that possibility. Hence, equation (36) only holds if $\gamma_{k0} = \tilde{\gamma}_{k0}$, $\gamma_{ik} = \tilde{\gamma}_{ik}$, $\omega_{i1} = \tilde{\omega}_{i1}$ and $b_{i1} = \tilde{b}_{i1}$.

(b) Case where $\ell = 2$: A typical artificial neural network with $\ell = 2$ hidden layers and a multiple-factor structure gives rise to the outputs

$$O_k = f\left(\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i2} f(\omega_{i1} x_{ti} + b_{i1}) + b_{i2})\right), \quad k \in \{1, 2\}.$$

Once again, we use $\boldsymbol{\theta}$ to represent the vector of parameters of the network. Suppose that $\tilde{\boldsymbol{\theta}}$ is another vector of parameters such that

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i2} f(\omega_{i1} x_{ti} + b_{i1}) + b_{i2}) = \tilde{\gamma}_{k0} + \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i2} f(\tilde{\omega}_{i1} x_{ti} + \tilde{b}_{i1}) + \tilde{b}_{i2}). \tag{37}$$

To prove global identifiability of the model, we need to show that, under restrictions (R.1) and (R.2), equation (37) is satisfied if and only if $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. The expression $f(\omega_{i1} x_{ti} + b_{i1})$ is a monotone transformation of an identified function, so it is identified. In fact,

$$f(\omega_{i1} x_{ti} + b_{i1}) = f(\tilde{\omega}_{i1} x_{ti} + \tilde{b}_{i1}) \Leftrightarrow \omega_{i1} x_{ti} + b_{i1} = \tilde{\omega}_{i1} x_{ti} + \tilde{b}_{i1} \Leftrightarrow \omega_{i1} = \tilde{\omega}_{i1} \quad \text{and} \quad b_{i1} = \tilde{b}_{i1}.$$

Now, Suppose that $\tilde{x}_{ti} = f(\omega_{i1}x_{ti} + b_{i1}) = f(\tilde{\omega}_{i1}x_{ti} + \tilde{b}_{i1})$ (the case when $f(\omega_{i1}x_{ti} + b_{i1}) \neq f(\tilde{\omega}_{i1}x_{ti} + \tilde{b}_{i1})$ is dealt as in case (a)). Equation (37) becomes

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i2}\tilde{x}_{ti} + b_{i2}) = \tilde{\gamma}_{k0} + \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i2}\tilde{x}_{ti} + \tilde{b}_{i2}) \tag{38}$$

At this point in the proof, equation (38) is similar to equation (34) of case (a). We then use the same arguments as in case (a) to conclude about the global identification.

(c) Case where $\ell = 3$: A typical artificial neural network with $\ell = 3$ hidden layers and a multiple-factor structure gives rise to the outputs

$$O_k = f\left(\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i3} f(\omega_{i2} f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2}) + b_{i3})\right), \quad k \in \{1, 2\}.$$

To prove global identifiability of the model, we need to show that under restrictions (R.1) and (R.2), equation (39)

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i3} f(\omega_{i2} f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2}) + b_{i3}) = \tilde{\gamma}_{k0} + \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i3} f(\tilde{\omega}_{i2} f(\tilde{\omega}_{i1}x_{ti} + \tilde{b}_{i1}) + \tilde{b}_{i2}) + \tilde{b}_{i3}) \quad k \in \{1, 2\} \tag{39}$$

is satisfied if and only if $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. Using the same techniques as in case (b), it is easy to show that $\tilde{x}_{ti} = f(\omega_{i2} f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2})$ is identified. Equation (39) becomes

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i3}\tilde{x}_{ti} + b_{i3}) = \tilde{\gamma}_{k0} + \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i3}\tilde{x}_{ti} + \tilde{b}_{i3}) \quad k \in \{1, 2\}. \tag{40}$$

Then, using equation (40) and following the same procedure as in case (a), it is easy to show that equation (39) is satisfied if and only if all the parameters on the left side are equal to those on the right side.

Proof of minimality with single-factor structure for the neural network

(a) Case where $\ell = 1$: Trivial under Assumption 2.

(b) Case where $\ell = 2$: With $\ell = 2$ hidden layers, $\tilde{x}_t = f\left(\delta_1 f(\sum_{i=1}^{d} \omega_i x_{ti} + b_1) + b_2\right)$, $i \in \{1, \cdots, d\}$. Using the same approach as with the multiple-factor structure, the network with $\ell = 2$ hidden layers is minimal if under restrictions (R.3) to (R.5) of Assumption 2, $\tilde{x}_t$ cannot

be obtained with a neural network of $\ell = 1$ hidden layers. Considering two real numbers α_1 and α_2, we have to show that

$$\alpha_1 f\left(\sum_{i=1}^{d} \omega_i x_{ti} + b_1\right) + \alpha_2 f\left(\delta_1 f\left(\sum_{i=1}^{d} \omega_i x_{ti} + b_1\right) + b_2\right) = 0 \tag{41}$$

if and only if $\alpha_1 = \alpha_2 = 0$. Suppose that equation (41) holds. We prove that $\alpha_1 = \alpha_2 = 0$.

By setting $\bar{X}_t^{(1)} = \sum_{i=1}^{d} \omega_i x_{ti} + b_1$ and $\bar{X}_t^{(2)} = \delta_1 f(\bar{X}_t^{(1)}) + b_2$, equation (41) can be written as

$$\alpha_1 f(\bar{X}_t^{(1)}) + \alpha_2 f(\bar{X}_t^{(2)}) = 0. \tag{42}$$

Then using restrictions (R.3) to (R.5) and applying the same steps as in case (b) of the proof of minimality with multiple factor structure, it follows that equation (42) holds if and only if $\alpha_1 = \alpha_2 = 0$.

(c) Case where $\ell = 3$: With $\ell = 3$ hidden layers, $\tilde{x}_t = f(\delta_2 f\left(\delta_1 f(\sum_{i=1}^{d} \omega_i x_{ti} + b_1) + b_2\right) + b_3)$. Using the same procedure as when $\ell = 2$, let α_1, α_2 and α_3 be three real numbers and suppose that

$$\alpha_1 f\left(\sum_{i=1}^{d} \omega_i x_{ti} + b_1\right) + \alpha_2 f\left(\delta_1 f\left(\sum_{i=1}^{d} \omega_i x_{ti} + b_1\right) + b_2\right) + \alpha_3 f\left(\delta_2 f\left(\delta_1 f\left(\sum_{i=1}^{d} \omega_i x_{ti} + b_1\right) + b_2\right) + b_3\right) = 0. \tag{43}$$

We prove that $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Consider $\bar{X}_t^{(1)} = \sum_{i=1}^{d} \omega_i x_{ti} + b_1$, $\bar{X}_t^{(2)} = \delta_1 f(\bar{X}_t^{(1)}) + b_2$ and $\bar{X}_t^{(3)} = \delta_2 f(\bar{X}_t^{(2)}) + b_3$, then equation (43) can be written as

$$\alpha_1 f(\bar{X}_t^{(1)}) + \alpha_2 f(\bar{X}_t^{(2)}) + \alpha_3 f(\bar{X}_t^{(3)}) = 0. \tag{44}$$

Then by using restrictions (R.3) to (R.5) and applying the same steps as in case (c) of the proof of minimality with multiple factor structure, it follows that equation (44) holds if and only if $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

Proof of global identification with single-factor structure for the neural network

The outputs of a typical artificial neural network with $\ell = 1$ hidden layers and a single-

factor structure are written as

$$O_k = f\left(\gamma_{k0} + \gamma_k f(\sum_{i=1}^{d} \omega_i x_{ti} + b_1)\right) \quad k \in \{1, 2\}.$$

We use $\boldsymbol{\theta}$ as the vector of parameters of the network. Suppose that $\tilde{\boldsymbol{\theta}}$ is another vector of parameters such that

$$f\left(\gamma_{k0} + \gamma_k f(\sum_{i=1}^{d} \omega_i x_{ti} + b_1)\right) = f\left(\tilde{\gamma}_{k0} + \tilde{\gamma}_k f(\sum_{i=1}^{d} \tilde{\omega}_i x_{ti} + \tilde{b}_1)\right) \quad k \in \{1, 2\}. \tag{45}$$

Since the function f is strictly increasing, equation (45) is equivalent to

$$\gamma_{k0} + \gamma_k f(\sum_{i=1}^{d} \omega_i x_{ti} + b_1) = \tilde{\gamma}_{k0} + \tilde{\gamma}_k f(\sum_{i=1}^{d} \tilde{\omega}_i x_{ti} + \tilde{b}_1) \quad k \in \{1, 2\}. \tag{46}$$

To prove the global identifiability of the model, we need to show that under restrictions (R.3) to (R.5) of Assumption 2, equation (46) is satisfied if and only if $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. This equality follows by applying the same steps as in the case of global identification of the multiple-factor structure. This also holds when $\ell = 2$ and $\ell = 3$.

□

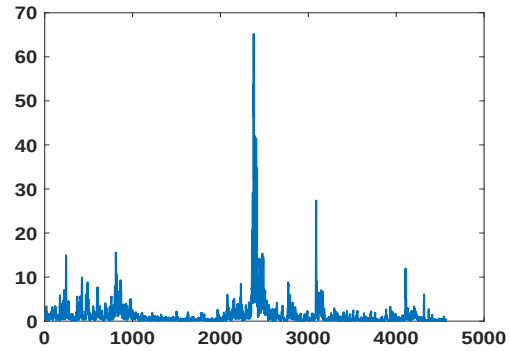
B Others tables and figures

C Specifications of alternative models

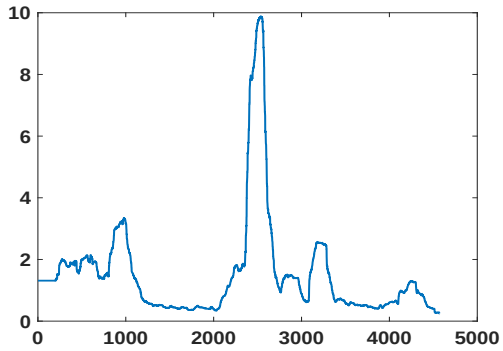
The TVP_{ANN}-GARCH process is compared to four volatility models: the standard GARCH process, the GJR-GARCH process , the MS-GARCH (and MS-GJR-GARCH) process and the FC-GARCH process. In this Section, we detail the specification of these alternatives.

- **The GARCH and the GJR-GARCH model**:

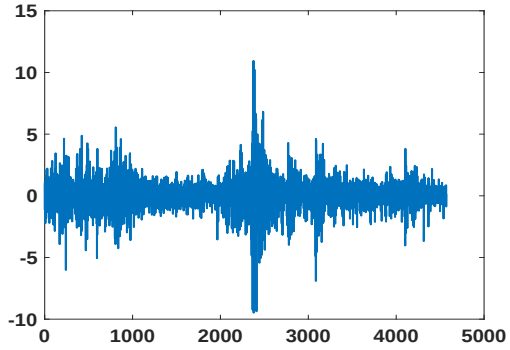
  The GJR-GARCH model proposed by Glosten et al. (1993b) allows the conditional variance of the process $\{y_t\}$ to respond differently to the past positive and negative



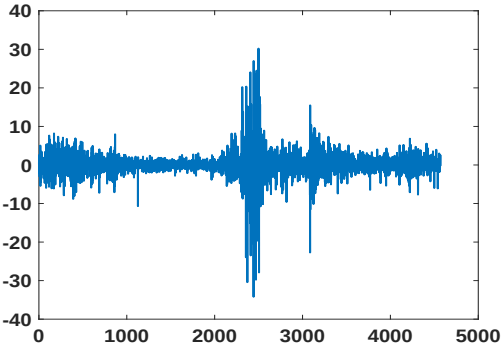
(a) Input 1



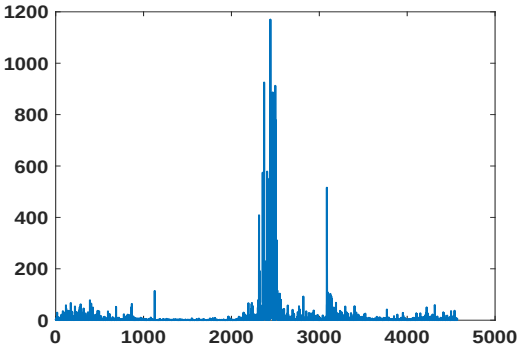
(b) Input 2



(c) Input 3



(d) Input 4



(e) Input 5

Figure 6 – Plots of ANN inputs used in the simulation to test the model selection strategy.

innovations. The GJR-GARCH(1,1) is specified as,

$$\begin{aligned} y_t &= \sigma_t \epsilon_t, \\ \sigma_t^2 &= \omega + \left(\alpha + \gamma \mathbb{1}_{\{y_{t-1}<0\}}\right) y_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned} \tag{47}$$

Table 13 – Correlation matrix of inputs

	Input 1	**Input 2**	**Input 3**	**Input 4**	**Input 5**
Input 1	1	0.38	0.04	0.02	0.31
Input 2		1	0.01	0.00	0.34
Input 3			1	0.66	-0.06
Input 4				1	-0.06
Input 5					1

Notes: Input 1 is the empirical variance of S&P 500 returns over the previous 5 days, Input 2 is the empirical variance of the S&P 500 over the previous 200 days, Input 3 is the series of lagged realizations of the S&P 500 returns, Input 4 is the series of lagged realizations of Bank of America returns, and Input 5 is the square root of lagged realizations of Bank of America returns.

where ϵ_t is driven by a standardized normal distribution and $\mathbb{1}_{\{.\}}$ denotes the indicator function. The GARCH(1,1) model is a particular case of the GJR process with $\gamma = 0$. We estimate the GJR-GARCH model using a SMC sampler algorithm and we set the following prior distributions: $\log \omega \sim N(0,2)$, $\log \frac{\beta-0.4}{1-\beta} \sim N(0,2)$, $\log \frac{\alpha}{0.3-\alpha} \sim N(0,2)$ and $\log \frac{\gamma}{0.3-\gamma} \sim N(0,2)$ in order to have $\alpha \in (0, 0.3)$, $\gamma \in (0, 0.3)$ and $\beta \in (0.4, 1)$.

- **The MS-GARCH and MS-GJR-GARCH models**:

  The MS-GJR-GARCH model that we consider is based on the specification introduced by Haas et al. (2004b). Let $\{S_t\}$ be a discrete time unobserved first-order ergodic and homogeneous Markov chain with finite state space $\{1, \cdots, K\}$ and transition matrix $P = (p_{ij})_{i,j=1}^{K}$ where $p_{ij} = P(S_t = j \mid S_{t-1} = i)$ for $i, j = 1, \cdots, K$ (i.e probability of a transition from state $S_{t-1} = i$ to state $S_t = j$). The MS-GJR-GARCH process is given by

$$\begin{aligned} y_t &= \sigma_{t,S_t} \epsilon_t, \\ \sigma_{t,j}^2 &= \omega_j + \left(\alpha_j + \gamma_j \mathbb{1}_{\{y_{t-1}<0\}}\right) y_{t-1}^2 + \beta_j \sigma_{t-1,j}^2, \quad j = 1, \ldots, K, \end{aligned} \tag{48}$$

  where ϵ_t follows a standard normal distribution and $\sigma_{t,j}^2$ is the conditional variance of $\{y_t\}$ in regime j. The parameter γ_j controls the degree of asymmetry in the conditional variance response to the past shock in regime j. Our MS-GJR-GARCH model

is estimated via the R package of Ardia et al. (2016). We assume two regimes (i.e., $K = 2$). The MS-GARCH model is a particular case of the MS-GJR-GARCH process with $\gamma_j = 0$ for $j = 1, \ldots, K$.

- **The FC-GARCH model**:

  The Flexible Coefficient GARCH, or FC-GARCH, process was developed by Medeiros and Veiga (2009) as a generalization of the GARCH and logistic Smooth Transition GARCH (ST-GARCH) formulations. The return process $\{y_t\}$ is said to follow a first-order Flexible Coefficient GARCH model with $m = H + 1$ limiting regimes (i.e FC-GARCH(m,1,1)) if it is specified as

$$\begin{aligned} y_t =& \sigma_t \varepsilon_t, \\ \sigma_t^2 =& G\left(\mathbf{w}_t; \boldsymbol{\psi}\right) = \omega_0 + \beta_0 h_{t-1} + \alpha_0 y_{t-1}^2, \\ &+ \sum_{i=1}^{H} \left[\omega_i + \beta_i \sigma_{t-1}^2 + \alpha_i y_{t-1}^2\right] f\left(s_t; \gamma_i, c_i\right), \quad t = 1, \ldots, T, \end{aligned} \tag{49}$$

  where $\varepsilon_t \sim N(0,1)$, σ_t^2 stands for the conditional variance at time t and $f\left(s_t; \gamma_i, c_i\right)$, $i = 1, \ldots, H$ is the logistic function defined as

$$f\left(s_t; \gamma_i, c_i\right) = \frac{1}{1 + e^{-\gamma_i (s_t - c_i)}}.$$

  The variable s_t is called the transition variable and must be observable at time $t-1$. The parameters γ_i, $i = 1, \ldots, H$ are the slope parameters and determine the speed of transition between two limiting regimes (when $\gamma_i \to \infty$, the logistic function becomes a step function). We set $m = 2$ limiting regimes and we use the past log-return y_{t-1} for the transition variable s_t as suggested in Medeiros and Veiga (2009). We estimate the FC-GARCH(2,1,1) model using a SMC sampler algorithm and we set the following prior distributions: $\log \omega_j \sim N(0,2)$, $\log \frac{\beta_0 - 0.4}{1 - \beta_0} \sim N(0,2)$, $\log \frac{\beta_1 + 1}{1 - \beta_1} \sim N(0,2)$, $\log \frac{\alpha_0}{0.3 - \alpha_0} \sim N(0,2)$, $\log \frac{\alpha_1 + 0.6}{0.3 - \alpha_1} \sim N(0,2)$ in order to have $\alpha_0 \in (0, 0.3)$, $\alpha_1 \in (-0.6, 0.3)$ and $\beta_0 \in (0.4, 1)$, $\beta_1 \in (-1, 1)$. The priors of the parameters of the logistic function

are given by $\log \gamma_1 \sim N(0,2)$ and $c_1 \sim N(0,2)$.

D MCMC sampler

The SMC algorithm presented in algorithm 1 requires an MCMC sampler that targets the posterior distribution, $\pi_\phi(\boldsymbol{\theta}|y_{1:T})$. The sampler uses a Metropolis-type algorithm to update random blocks of the TVP_{ANN}-GARCH parameters. Note that M particles evolve simultaneously in the SMC sampler. To rejuvenate one particle during the MCMC sampler, say $\boldsymbol{\theta}^i$, for $i = 1, ..., M$, we sample randomly from one of the 10 proposal distributions presented in Dufays (2016) and accept or reject the candidate according to a Metropolis-Hastings acceptance ratio. The model-free proposal distributions are adapted from the differential evolution (DE) optimization literature (for a review, see Das and Suganthan (2011)). In particular, at the j-th MCMC iteration, the i-th particle $\boldsymbol{\theta}^i$ is updated as follows:

1. Choose uniformly and apply one of the three types of mutation:

   - Standard mutation:

   $$\boldsymbol{\theta}^{\text{Mut}} = \boldsymbol{\theta}^{r_1} + F_{DE}(\boldsymbol{\theta}^{r_2} - \boldsymbol{\theta}^{r_3}),$$

   in which F_{DE} is a fixed constant and r_1,r_2,r_3 are taken without replacement in the $M-1$ remaining particles.

   - Trigonometric mutation:

   $$\boldsymbol{\theta}^{\text{Mut}} = \sum_{i=1}^{3} \boldsymbol{\theta}^{r_i}/3 + (p_2 - p_1)(\boldsymbol{\theta}^{r_1} - \boldsymbol{\theta}^{r_2}) + (p_3 - p_2)(\boldsymbol{\theta}^{r_2} - \boldsymbol{\theta}^{r_3}) + (p_1 - p_3)(\boldsymbol{\theta}^{r_3} - \boldsymbol{\theta}^{r_1}),$$

   in which $p_i \propto f(y_{1:t}|\boldsymbol{\theta}^{r_i})f(\boldsymbol{\theta}^{r_i})$ for $i \in [1,3]$ are probabilities such that $\sum_{i=1}^{3} p_i = 1$ and the index r_i, for $i \in [1,3]$, is taken without replacement in the $M-1$ remaining particles.

- Firefly mutation:

$$\boldsymbol{\theta}^{\text{Mut}} = \boldsymbol{\theta}^{r_1} + F_{FF}(\boldsymbol{\theta}^{r_1} - \boldsymbol{\theta}^{r_2}),$$

where F_{FF} is a chosen constant and r_1,r_2 are taken without replacement in the $M-1$ remaining particles.

2. Select with equal probability one of the three different moves and propose a new candidate $\boldsymbol{\theta}'$ for the parameters $\boldsymbol{\theta}^i$:

- DREAM proposal:
  - If a standard move was selected:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}^i + F(\delta, d)(\sum_{g=1}^{\delta} \boldsymbol{\theta}^{r_1(g)} - \sum_{h=1}^{\delta} \boldsymbol{\theta}^{r_2(h)}) + \zeta, \tag{50}$$

where $i \neq r_1(g)$, $r_2(h)$; $r_1(.)$ and $r_2(.)$ denote random integers uniformly distributed on the support $[1, M]_{-i}$ and it is required that $r_1(g) \neq r_2(h)$ when $g = h$ and $\zeta \sim N(0, \eta_x^2 I)$; $\delta \sim U[1,3]$, $F(\delta, d) = 2.38/\sqrt{2\delta d}$ with d, the dimension of $\boldsymbol{\theta}$.

  - Otherwise:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}^i + Z_{\text{Dir}} F(\delta = 1, d)(\boldsymbol{\theta}^{\text{Mut}} - \boldsymbol{\theta}^{r_1}) + \zeta, \tag{51}$$

in which $\zeta \sim N(0, \eta_x^2 I)$; $\delta \sim U[1,3]$, $F(\delta, d) = 2.38/\sqrt{2\delta d}$, $Z_{\text{Dir}} = 1$ with probability 0.5 and -1 otherwise.

- Stretch move proposal:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}^{\text{Mut}} + Z_{\text{Stretch}}(\boldsymbol{\theta}^i - \boldsymbol{\theta}^{\text{Mut}}), \tag{52}$$

in which $Z_{\text{Stretch}} \sim Z_{\text{S}}$ and Z_{S} is a random variable with a cumulative density

function given by $F_{Z_S}(x) = \frac{\sqrt{a_S x}-1}{a_S-1}$ with $a_S = 2.5$.

- Walk move:

$$\boldsymbol{\theta}' \quad = \quad \boldsymbol{\theta}^i + Z_{RW}(\boldsymbol{\theta}^i - x^{\text{Mut}}), \tag{53}$$

in which $Z_{RW} \sim Z_W$ and Z_W is a random variable with a cumulative density function given by $F_{Z_W}(x) = 1 - \frac{(a_W+1)^{1/2}-(x+1)^{1/2}}{(a_W+1)^{1/2}-(a_W+1)^{-1/2}}$, with $a_W = 2$.

3. Accept or reject the candidate $\boldsymbol{\theta}'$ according to the Metropolis-Hastings ratio,

$$\min\{\frac{q(\boldsymbol{\theta}')\pi_\phi(\boldsymbol{\theta}'|y_{1:T})}{\pi_\phi(\boldsymbol{\theta}|y_{1:T})}, 1\}, \tag{54}$$

in which $q(\boldsymbol{\theta}') = 1$ if the DREAM proposal has been chosen, $q(\boldsymbol{\theta}') = |1 + Z_{\text{W}}|^{d-1}$ if the walk move has been chosen, or $q(\boldsymbol{\theta}') = |Z_{\text{S}}|^{d-1}$ if the stretch move was chosen.

The three types of moves have been proposed independently in the MCMC literature (see Vrugt et al. (2009) for the DREAM algorithm, and Christen and Fox (2010) and Foreman-Mackey et al. (2013) for the walk and the stretch moves).