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

Arnaud Dufays[a] and Morvan Nongni[b]

[a,b] *Département d'économique, Université Laval*

Version: January 31, 2018

Abstract We propose a new volatility process in which the parameters vary over time according to an artificial neural network (ANN). We prove the process stationarity as well as the global identification of the parameters. Since ANNs require economic series as input variables to operate, we develop a shrinkage approach to select which explanatory variables are relevant to forecast the volatility. We also show that our process encompasses standard GARCH models and can be viewed as an extension of GARCH-X processes. Empirically, the proposed model favourably compares with other flexible processes in terms of in-sample fit and out-of-sample predictions. Extensions to others standard models are promising.

...

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

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

1 Introduction

Due to its crucial importance in risk management and derivatives (stocks, options, etc.) pricing, a numerous amount of research has been devoted to modelling volatilities of financial returns. This includes the Generalized Autoregressive Conditional Heteroskedastic (GARCH) model (Engle, 1982; Bollerslev, 1986) in which the volatility evolves over time deterministically according to past volatilities and past financial log-returns, the stochastic volatility (SV) model (Taylor, 1986), and many other alternatives. However, the majority of existing models typically exhibits fixed parameters, ignoring the impact of changes in the financial system which may lead to bad forecasts and decisions (e.g. Dieobold (1986), Lamoureux and Lastrapes (1990), Hillebrand (2005)).

To take economic changes into account, the research community has been proposing models that allow the parameters to vary over time. At least, two leading processes are currently competing, namely the Markov-switching (MS) process and the time-varying parameter (TVP) model. The former assumes fixed parameters over a period window and detects abrupt changes of these parameters at specific times (e.g., Francq et al. (2001), Haas et al. (2004), Bauwens et al. (2010)). On the contrary, the TVP process makes each parameter evolving at each time period. To alter the parameters at each period, a dynamic for them is hypothesized. A popular choice is an autoregressive (AR) process since it leads to a fair balance between parsimony and flexibility of the process[1] (e.g.,Kim et al. (1999), Krolzig (2013)). This paper proposes a class of TVP models in which the parameters evolve according to artificial neural networks (ANNs) instead of the standard AR process.

The artificial neural network GARCH (ANN-GARCH) model is based on an ANN function for the parameter dynamics of the GARCH volatility equation instead of assuming an AR process for the parameter dynamics like in standard TVP models. The use of ANNs is not arbitrary. In fact, they exhibit the appealing property of being universal approximators (see, e.g., Bishop (2006)). Put simply, this means that if a true dynamic of the parameters

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

exists and is deterministic then an ANN can arbitrarily closely approximate this function. Consequently, we believe that ANNs are natural choices for the parameter dynamics. Moreover, since any TVP application could potentially be revisited, the applicability of the new approach is potentially large.

In addition of being universal approximators, ANNs exhibit two other advantages compared to the (stochastic) AR process; namely i) inference tractability and ii) the interpretation of the parameter dynamics. Regarding the inference tractability, since ANNs are deterministic functions of the input variables, one can directly evaluate the likelihood function which is the keystone ingredient for inferring model parameters. To contrast with current 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 generated numerous papers just focusing on the TVP model estimation (see, e.g., Broto and Ruiz (2004) for a review) and it remains one of the vivid TVP research lines. Additionally, ANNs require to choose input variables (e.g. explanatory variables such as past volatilities, Book-to-market ratio, previous trading volumes in the market or other financial factors) for modeling the financial log-return volatilities through the parameter dynamics. These variables are of principal interests for an economist since it creates a network of crucial variables that Granger cause the output value of the ANNs. This direct interpretation is lacking in the current TVP models.

With regard to the TVP literature, closely related processes are different kinds of smooth transition models. The latter process has been firstly proposed in Luukkonen et al. (1988) and it consists of a TVP model in which the 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 (see, e.g., 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.

This paper deviates in many aspects from the smooth transition models. All these related papers stick to a one-layer ANN with logistic functions while in the ANN literature, it has been highlighted that deeper ANNs generally produce better estimates. Moreover, these papers assume that all the 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, Peluso et al. (2016) or Dufays et al. (2016)). Regarding the ANN input variables, the smooth transition papers are limited to at most three variables while we propose to go far beyond these values. In fact, the model shrinks irrelevant parameters toward zero such that it can potentially include a lot of explanatory variables. The shrinkage method is another difference with the current smooth transition literature as most of the papers are developed in the frequentist paradigm (one exception is Deschamps (2008)). We propose a sound Bayesian counterpart including model comparisons, efficient estimations and statistical properties of the developed process. An empirical exercise based on US financial returns highlights that the ANN-GARCH model is generally superior to standardly used processes in terms of marginal likelihood.

The paper is organized as follows. Section 2 presents the ANN-GARCH process and studies its statistical properties (conditions for weak and strong stationnarities). Section 3 is devoted to the model identification conditions. Estimation of the model is considered in Section 4. We illustrate the algorithm on simulated data in Section 5, and empirical examples are considered in section 6. Finally, Section 7 concludes the paper and all technical proofs for theoretical results are exposed in the Appendix.

2 Model definition and theoretical results

2.1 Model specification

We consider the 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 $\alpha + \beta < 1$. As stationary conditions are assumed, let us denote the unconditional variance by $\bar{\omega} = \omega/(1-\alpha-\beta)$. Then the GARCH(1,1) dynamic can be reframed as

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

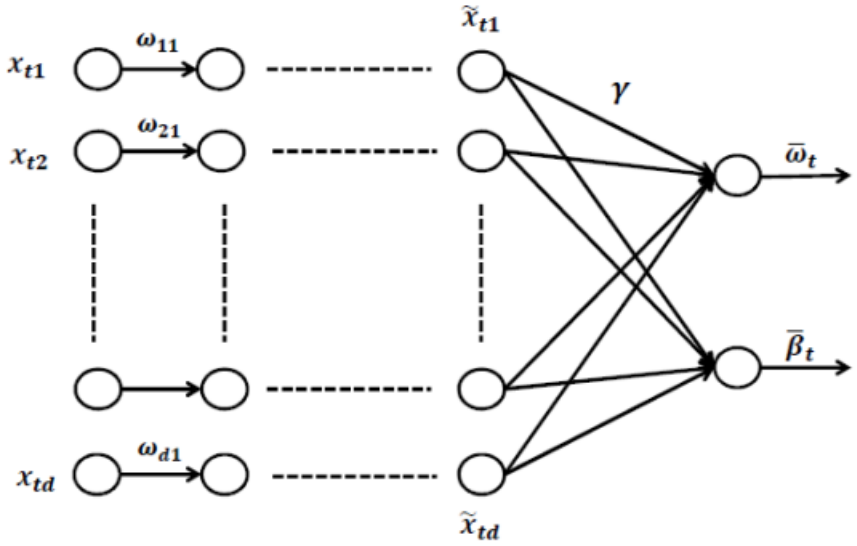
in which $\phi = \alpha + \beta$ stands for the persistence of the variance dynamic. We use an artificial neural network (ANN) to model the dynamics of the unconditional variance as well as the persistence parameter. Analogously to Equation (1), the ANN-GARCH(1,1) model is specified as

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

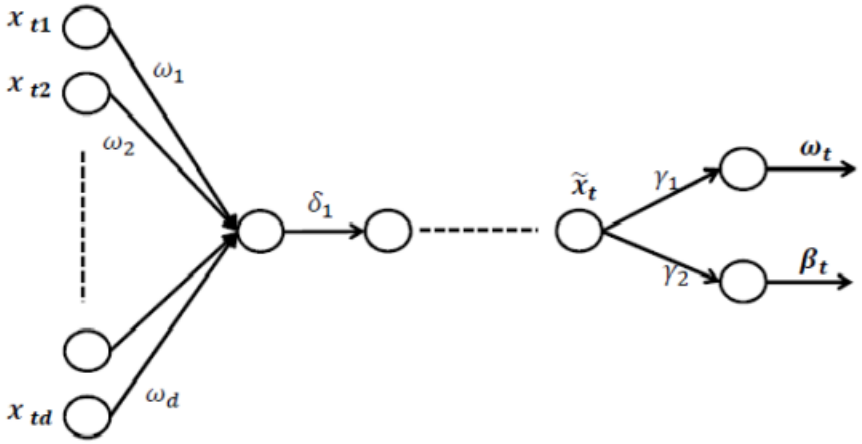
where $\phi_t = \alpha + (1-\alpha)\tilde{\beta}_t$. The outputs of the ANN are $\{\bar{\omega}_t, \tilde{\beta}_t\}$.

2.2 ANN structure

Many neural networks topologies have been 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. These two ANNs exhibit the enjoyable property of being globally



(a) Multiple factor architecture



(b) One factor architecture

Figure 1 – Structures of the two ANN used in the ANN-GARCH process.

identified as it will be clear from theorem 3.

The two structures are comprised of an input layer, one or more hidden layers and one output layer. These structures assume a fixed number of hidden nodes in each hidden layer and logistic functions, i.e. $f(x) = (1 + exp(-x))^{-1}$, as non-linear activation functions. The number of nodes in each hidden layer is equal to the number of input variables for the multiple factors architecture (Figure 1-(a)). In the one factor architecture (Figure 1-(b)), each hidden layer exhibits a unique node.

2.2.1 The multiple factor structure

The multiple factor architecture network (see Figure 1-a) consists of the specification of (a) d *input nodes*, labeled $1, \cdots, d$, (b) ℓ *hidden layers*, labeled $1, \cdots, \ell$, (c) two *outputs nodes*, (d) an array $\{\omega_{ij}\}$ of weights for connections between input nodes and the first hidden nodes on one hand and between hidden nodes of different hidden layers on the other hand (e) an array $\{\gamma_{ji}\}$ of weights for connections from the last hidden nodes to the output nodes, and (f) an array $\{b_{ij}\}$ of biases for hidden nodes. The ω_{ij} are indexed by $i \in \{1, \cdots, d\}$ $j \in \{1, \cdots, \ell\}$ so that ω_{ij} is the weight of the connection from the $i-$th node of the $(j-1)$th layer to the $i-$th node of the $j-$th hidden layer. The weights γ_{ki}, $i \in \{1, \cdots, d\}$ $k \in \{1, 2\}$ connects the $i-$th node of the last ($\ell-$th) hidden layer to the $j-$th output node (Note: γ_{10} and γ_{20} are the bias for the two output nodes). The parameters b_{ji}, $i \in \{1, \cdots, d\}$ $j \in \{1, \cdots, \ell\}$ stands for the bias for the $i-$th node of the $j-$th hidden layer.

Given explanatory variables $\boldsymbol{x}_t = (x_{t1}, \cdots, x_{td})' \in \mathbb{R}^{d \times 1}$, the output $H_1(x_{ti})$ of the $i-$th node of first hidden layer is given by $H_1(x_{ti}) = f(\omega_{i1} x_{ti} + b_{i1})$. Similarly, the net output of the $i-$th node of the second hidden layer is $H_2(x_{ti}) = f(\omega_{i2} H_1(x_{ti}) + b_{i2})$. Recursively, the output of the $i-$th node of the $k-$th hidden layer depend on the output of the $i-$th node of the $(k-1)$th hidden layer as follows $H_k(x_{ti}) = f(\omega_{ik} H_{k-1}(x_{ti}) + b_{ik})$, for $k \in \{2, \cdots, \ell\}$. The net $i-$th input of the output nodes, named $\tilde{x}_{ti}$ (see Figure 1-a) correspond to the output of the $i-$th node of the last ($\ell-$th) hidden layer of the network and is equal to $\tilde{x}_{ti} = f(\omega_{i\ell} H_{\ell-1}(x_{ti}) + b_{i\ell})$, for $k \in \{2, \cdots, \ell\}$. Given the hidden output of the last nodes, i.e. $\tilde{\boldsymbol{x}}_t$, the GARCH parameters are computed as

$$\begin{aligned} \bar{\omega}_t &= \ln f(\boldsymbol{\gamma}_1' \tilde{\boldsymbol{x}}_t + \gamma_{10}) = \ln f\left(\sum_{i=1}^{d} \gamma_{1i} \tilde{x}_{ti} + \gamma_{10}\right), \\ \tilde{\beta}_t &= f(\boldsymbol{\gamma}_2' \tilde{\boldsymbol{x}}_t + \gamma_{20}) = f\left(\sum_{i=1}^{d} \gamma_{2i} \tilde{x}_{ti} + \gamma_{20}\right). \end{aligned} \tag{3}$$

These final transformations make the parameters $\bar{\omega}_t$ lying in $\Re_+$ and $\tilde{\beta}_t$ lying in $]0, 1[$. Given

this structure for the network, the parameters vector of the model is $\boldsymbol{\theta} = (\omega_{ij}, b_{ij}, \gamma_{ki}, \gamma_{10}, \gamma_{20}, \alpha)$ where $i \in \{1, \cdots, d\}$, $j \in \{1, \cdots, \ell\}$ and $k \in \{1, 2\}$.

2.2.2 The one factor structure

Focusing on the one factor architecture network (see Figure 1-b), the GARCH parameters are exactly defined as in Equation (3). However as one hidden input value is related to the final node, this unique input both controls the dynamics of the unconditional variance parameter and of the persistence parameter. Mathematically, it reads as

$$\begin{aligned} \bar{\omega}_t &= \ln f(\gamma_1 \tilde{x}_t + \gamma_{10}), \\ \tilde{\beta}_t &= f(\gamma_2 \tilde{x}_t + \gamma_{20}), \end{aligned} \tag{4}$$

where γ_1, γ_2 are the weights of the links between the node of the last hidden layer and those of the output layer, and γ_{10}, γ_{20} are the bias of the the output nodes. The hidden value $\tilde{x}_t$ is the node output of the last hidden layer (see figure 1-b). When the number of hidden layer is equal to $\ell = 1$, the hidden value $\tilde{x}_t$ is directly related to the explanatory variables since $\tilde{x}_t = f(\sum_{i=1}^{d} \omega_i x_{ti} + b_1)$. Let us denote $H_\ell(\boldsymbol{x}_t)$, the output of the $l-$th hidden layer of the network. Then the first hidden layer is given by $H_1(\boldsymbol{x}_t) = f(\sum_{i=1}^{d} \omega_i x_{ti} + b_1)$. Recursively, when $\ell \geq 2$ hidden layers, we have $\tilde{x}_t = f(\delta_{\ell-1} H_{\ell-1}(\boldsymbol{x}_t) + b_\ell)$ where $\boldsymbol{x}_t = (x_{t1}, \cdots, x_{td})$ is the input vector and ω_i is the weight for the connection from the $i-$th input node to the single node of the first hidden layer. The parameter vector $\{\delta_j\}$, for $j = 1, \cdots, \ell-1$ holds only for a network with more than one hidden layer and δ_j is the weight for the link between the node of $j-$th hidden layer and the one of the $(j+1)$th hidden layer. With the one factor structure as network structure, the parameter vector of the model is $\boldsymbol{\theta} = (\omega_i, \delta_j, b_r, \gamma_k, \gamma_{10}, \gamma_{20}, \alpha)$ where $i \in \{1, \cdots, d\}$, $j \in \{1, \cdots, \ell - 1\}$, $r \in \{1, \cdots, \ell\}$ and $k \in \{1, 2\}$.

2.3 Model properties

Stationarity and ergodicity are enjoyable properties for a stochastic process. Following Medeiros and Veiga (2009), the next theorem presents a necessary and sufficient log-moment condition for the strict stationarity and ergodicity of the ANN-GARCH model.

Theorem 1. *Let $y_t \in \mathbb{R}$ follows an ANN-GARCH process as in (2). Assuming $0 < \alpha < 1$, the process $\boldsymbol{u}_t = (y_t, \sigma_t^2)'$ is strictly stationary and ergodic if and only if*

$$\mathbb{E}\left\{\ln\left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2\right]\right\} < 0, \quad \forall t \in Z. \tag{5}$$

Proof. See Appendix A. □

Corollary 1. *Assuming $\alpha \in]0,1[$, the ANN-GARCH process is 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, given the fact that $0 < \alpha < 1$ and $\tilde{\beta}_t$ is bounded between 0 and 1, then $\alpha < (1-\alpha)\tilde{\beta}_t + \alpha < 1$. □

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

Theorem 2. *Let $y_t \in \mathbb{R}$ follows an ANN-GARCH process as in (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*

$$\mathbb{E}\left\{\left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2\right]^k\right\} < 1. \tag{6}$$

Proof. See Appendix A. □

Note that the inequality (6) is always true for $k = 1$. In fact, when $k = 1$, the inequality simplifies into

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

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 ANN-GARCH process always exhibits a well-defined unconditional variance.

3 Identification of the model

Identification of the ANN-GARCH(1,1) 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 (see for example Sussmann (1992), Hwang and Ding (1997)). To prove the identification of the ANN-GARCH parameters, we follow Hwang and Ding (1997) and rely on two concepts called reducibility and minimality.

The ANN-GARCH(1,1) model is identifiable if the associated ANN is identifiable. Putting differently, the ANN-GARCH(1,1) process is identifiable if it does not exist two sets of parameters such that, given the same (non-constant) inputs, produces an identical output. First of all, 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 the permutation or interchangeability of two hidden nodes of the hidden layer in the network. The two structures considered here have been chosen such that permutations of nodes are not possible and the sign equivalence property of the logistic function may not 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 ANN-GARCH model is identifiable and minimal.

Definition 1. The 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.

In fact, reducible ANN contains irrelevant layers. This means that, given the same input, the ANN output can be exactly obtained from another ANN with fewer hidden nodes or layers. Consequently irreducibility is 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 called 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 one 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 make the following set of restrictions on the ANN parameter sets:

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

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

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

Assumption 2. *Parameters of the one factor factors 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 ANN-GARCH model is globally identifiable and minimal if*

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

Proof. See Appendix A. □

4 Shrinkage priors and estimation

4.1 Selection of the artificial neural network

Given the structures in figure 1, the problem of network selection is reduced to the choice of inputs variables and the number of hidden layers. We use shrinkage priors to assess the predictive power of inputs variables and then discriminate between them. Once the inputs are selected, we rely on a comparison criterion, the marginal log-likelihood (MLL) which penalizes the addition of extra parameters to select the optimal number of hidden layers.

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, a technique called SMC sampler (see Del Moral et al. (2006)). This method exhibits an important advantage compared to the standard MCMC approach when it comes to estimate ANN models. In fact, ANNs are recursive applications of non-linear functions and consequently the posterior distribution of the ANN parameters are bound to exhibit a highly complex shape. Moreover, to select the ANN input variables, we use shrinkage priors which also lead to multi-modal posterior distributions. As shown by Jasra et al. (2007) or Herbst and Schorfheide (2014), MCMC based on one single Markov-chain are less appropriate than SMC methods to simulate multi-modal distributions.

SMC sampler also exhibits some disadvantages, the most important of them being the number of user-specified parameters to tune in order to run the algorithm. To deal with this issue, we rely on the Time and Tempered (TNT) algorithm (see Dufays (2016)), i.e. a variant of the Sequential Monte Carlo sampler (Del Moral et al. (2006)), that automates the choice of the SMC parameters. The algorithm sequentially iterates by producing realizations from the prior distribution to the posterior distribution by combining many importance sampling and MCMCs. Let us denote by θ the model parameters when the ANN structure is fixed. Then, the ANN-GARCH model is estimated by Algorithm 1.

4.3 MCMC sampler

The SMC algorithm presented in algorithm 1 requires an MCMC sampler that targets the posterior distribution $\pi_\phi(\theta|y_{1:T})$. The sampler uses a Metropolis-type algorithm to update random blocks of the ANN-GARCH parameters. Note that M particles evolve simultaneously in the SMC sampler. To rejuvenate one particle during the MCMC sampler, say θ^i, for $i = 1, ..., M$, we sample randomly from one of the ten proposal distributions presented in Dufays (2016) and accept or reject the candidate according to a Metropolis-Hastings

Algorithm 1 SMC algorithm

Sample M particles from the prior distribution : $\{\theta^i\}_{i=1}^M$
Set the tempered function $\phi = 0$, the normalized weights $W_i = 1/M \ \forall i \in [1, M]$ and $ESS = M$
while $\phi < 1$ **do**
A - Correction step :
Find $\tilde{\phi} > \phi$ such that $M/(\sum_{i=1}^N \tilde{w}_i^2) = 0.95ESS$ where $\tilde{w}_i \propto W_i[f(y_{1:T}|\theta^i)]^{\tilde{\phi}-\phi}$
$\forall i \in [1, M]$; Set $w_i = W_i[f(y_{1:T}|\theta^i)]^{\tilde{\phi}-\phi}$
$\forall i \in [1, M]$; Compute the normalized weights $W_i = w_i / \sum_{j=1}^N w_j$ and $ESS = M/(\sum_{i=1}^N W_i^2)$
Set $\phi = \min(\tilde{\phi}, 1)$
B - Re-sample step if $ESS < 0.75M$
Re-sample the particles by stratified sampling (Carpenter et al. (1999))
C - MCMC step with targeted distribution : $\pi_\phi(\theta|y_{1:T}) \propto [f(y_{1:T}|\theta)]^\phi f(\theta)$
for $j = 1$ to N **do**
Apply N iterations of the MCMC sampler given in Section 4.3 on the M particles.
end for
end while

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 jth MCMC iteration, updating the ith particle θ^i is done as follows

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

   - Standard mutation:

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

     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:

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

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

$M-1$ remaining particles.

- Firefly mutation:

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

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 θ' for the parameters θ^i:

- DREAM proposal:
  - If standard move was selected:

$$\theta' = \theta^i + F(\delta, d)(\sum_{g=1}^{\delta} \theta^{r_1(g)} - \sum_{h=1}^{\delta} \theta^{r_2(h)}) + \zeta, \tag{7}$$

where $i \neq r_1(g)$, $r_2(h)$; $r_1(.)$ and $r_2(.)$ stand for 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 $\zeta \sim N(0, \eta_x^2 I)$; $\delta \sim U[1, 3]$, $F(\delta, d) = 2.38/\sqrt{2\delta d}$.

  - Otherwise:

$$\theta' = \theta^i + Z_{\text{Dir}} F(\delta = 1, d)(\theta^{\text{Mut}} - \theta^{r_1}) + \zeta, \tag{8}$$

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:

$$\theta' = \theta^{\text{Mut}} + Z_{\text{Stretch}}(\theta^i - \theta^{\text{Mut}}), \tag{9}$$

in which $Z_{\text{Stretch}} \sim Z_{\text{S}}$ and Z_{S} is a random variable whose the cumulative density function is given by $F_{Z_S}(x) = \frac{\sqrt{a_S x}-1}{a_S-1}$ with $a_S = 2.5$.

- Walk move:

$$\theta' = \theta^i + Z_{RW}(\theta^i - x^{\text{Mut}}), \tag{10}$$

in which $Z_{RW} \sim Z_W$ and Z_W is a random variable whose the cumulative density function is 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 θ' according to the Metropolis-Hastings ratio,

$$\min\{\frac{q(\theta')\pi_\phi(\theta'|y_{1:T}, \tau, \boldsymbol{\rho})}{\pi_\phi(\theta|y_{1:T}, \tau, \boldsymbol{\rho})}, 1\}, \tag{11}$$

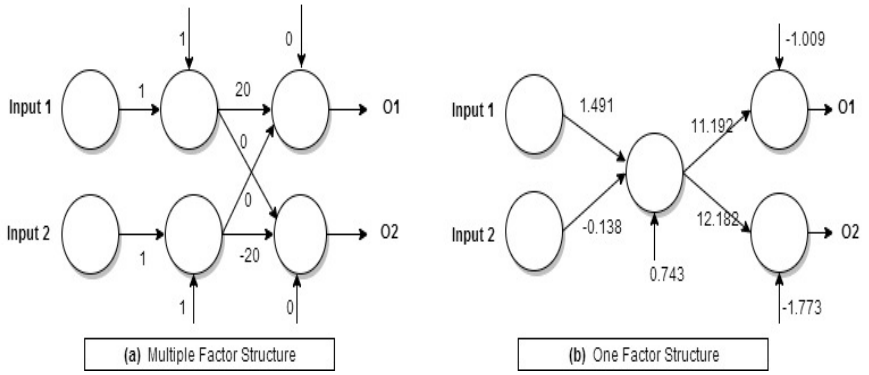
in which $q(\theta') = 1$ if the DREAM proposal has been chosen, $q(\theta') = |1 + Z_{\text{W}}|^{K_1-1}$ and $q(\theta') = |Z_{\text{S}}|^{K_1-1}$ if the Walk move or stretch move were respectively used.

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

5 Simulations

In this section, we simulate two series of 4605 observations from the ANN-GARCH model and investigate our estimation strategy for selecting the number of layers and input variables. The first series is simulated from an ANN-GARCH process with a one factor ANN structure exhibiting one hidden layer while the second data generating process (DGP), that generated the second series, uses a multiple factor ANN with two hidden layers. The ANN inputs consist in two variables derived from the S&P 500 series used in the empirical exercise (see Section 6.2). In particular, we consider the lagged realization of the S&P 500 as first input

Table 1 – Parameters values of the simulated DGPs

$y_t = \sigma_t \eta_t \quad \text{with } \eta_t \sim IID(0,1)$ $\sigma_t^2 = \bar{\omega}_t + \phi_t(\sigma_{t-1}^2 - \bar{\omega}_t) + \alpha(y_{t-1}^2 - \sigma_{t-1}^2), \quad \text{where } \phi_t = \alpha + (1-\alpha)\tilde{\beta}_t$
Parameters values for DGP with Multiple Factor Structure for the ANN: $\theta = (\omega, \alpha = 0.05)$ $\omega = (11.1926, 12.1825, -1.009, -1.7736, 0.7438, 1.4912, -0.1383)$
Parameters values for DGP with One Factor Structure for the ANN: $\theta = (\omega, \alpha = 0.05)$ $\omega = (1, 1, 1, 1, 0, 0, 20, 0, 0, -20)$


and the empirical variance over the last 250 days as second input. Table 1 documents the parameters values of the two DGPs. Using these values, we generated one series per DGP. Figure 2 shows the graph of the generated time series and the corresponding instantaneous conditional variance and persistence.

As expected, the two series are stationary and reproduce some stylized facts observed in financial time series such as the 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 to that, the instantaneous conditional variance ω_t also varies over time. These parameter dynamics highlight that the model can produce a large range of variations from smooth ones typically observed in Spline-GARCH processes or stochastic volatility models to abrupt and erratic behaviours generated by Change-point and Markov-Switching processes.

We start our simulation study by empirically assessing the identification of the model

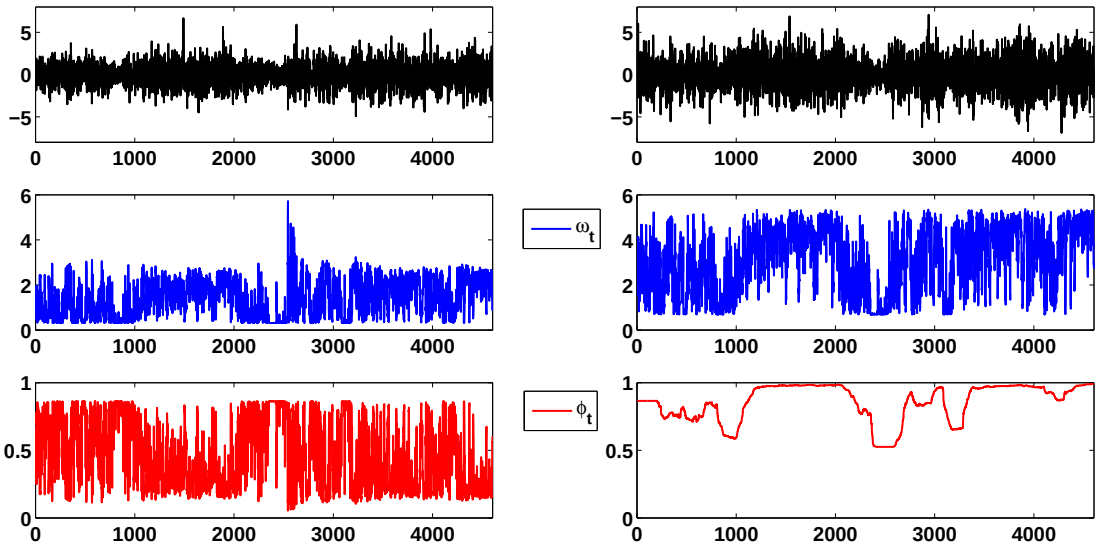


Figure 2 – Simulated series from the ANN-GARCH model (left) with the one factor structure (right) with the multiple factor structure

parameters. To do so, we estimate the ANN parameters on the simulated conditional variances by minimising the mean squared errors. Then, we compare our estimates with the true parameters. Table 2 displays the estimated parameters for the two series. We observe that the estimates are almost identical to the true parameters.

Table 2 – Estimation results of the simulated series

Multiple factor structure			One factor structure		
	True Par.	Est. Par.		True Par.	Est. Par.
1	1.00	1.00	1	11.1926	11.192
2	1.00	1.00	2	12.1825	12.1818
3	1.00	0.9998	3	-1.0091	-1.0091
4	1.00	0.9997	4	-1.7736	-1.7735
5	0.00	0.00	5	0.7438	0.7438
6	0.00	0.00	6	1.4912	1.4913
7	20.0	19.9965	7	-0.1383	-0.1383
8	0.00	0.00	8	0.05	0.05
9	0.00	0.00			
10	-20.0	-19.9956			
11	0.05	0.05			

We now focus on the estimation of 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 assumes five ANN inputs that are displayed in Figure 5 (see Appendix B). Among these five inputs, only the first two S&P 500 series are used to simulate the log-returns. In addition to that, as highlighted in Table 2, the first ANN inputs only impacts on the unconditional parameter $\bar{\omega}_t$ while the second input only triggers the persistence parameter. Table 9 highlights the correlation between the five inputs. We observe that the correlation between the true and the spurious inputs can be quite high. Table 3 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, regarding the multiple factor structure, Input 1 only triggers the unconditional variance parameter while the second input makes varying the persistence parameter as set in the DGP.

Table 3 – Spike probabilities for Inputs selection

Spike probabilities			
	Multiple Factor Structure		One Factor Structure
	Output 1	Output 2	
Inputs 1	0	0,998	0
Inputs 2	0,9895	0	0
Inputs 3	0,9937	0,9947	0,9205
Inputs 4	0,9825	0,8422	0,8872
Inputs 5	0,9932	0,9822	0,959

Input significant if probability $>$ 0,5

Once the explanatory variables have been selected, we find the best ANN number of layers using the marginal likelihood criterion. Practically, we estimate several ANN-GARCH models without shrinkage priors that differ in the number of hidden layers and then, we select the process that exhibits the highest marginal log-likelihood (MLL). Table 4 provides the MLLs for the different models. In both cases, we select the correct number of layers.

Table 4 – Marginal log-likelihood (MLL) for ANN-GARCH with different hidden layers in the network

Multiple Factor Structure		One Factor Structure	
Nbr of hidden layers	MLL	Nbr of hidden layers	MLL
0	-8812,7	1	**-7133,9**
1	**-8790**	2	-7133,9
2	-8797,3	3	-7140,3

6 Empirical results

We now examine the performance of the ANN-GARCH model to model financial series. Section 6.1 discusses the financial data used in the empirical exercise while Section 6.2 focuses on the ANN-GARCH performance compared to standard flexible models.

6.1 Data

We consider five US bank log-returns spanning from May 1999 to August 2017. The number of observations amounts to 4605 observations.[2] In this empirical study, we focus on Bank of America (BAC), Citigroup (C), Morgan Stanley (MS), Goldman Sach (GS) and JP Morgan (JPM). We also consider the S&P 500 index and the shiller index as eligible inputs for the ANN. Summary statistics of the series are presented in Table 5.

Estimating models with neural networks generally require to select inputs variables. For each series, we use shrinkage priors to discriminate between nine input variables. As inputs, we consider lagged log-returns and empirical variances over the h previous periods. In particular, these inputs consist in 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 lagged of the most relevant principal component (PCA) of the other bank log-returns and its empirical variance over 22 periods and eventually, the first lag of the Shiller index.

[2] The financial variables used were downloaded on yahoo finance.

Table 5 – Daily log-returns: Descriptive statistics.

Series	Mean	Std. dev	Skewness	Kurtosis
BAC	0	2.994	-0.326	28.56
C	0	3.162	-0.527	42.81
MS	0	3.179	0.302	14.34
GS	0	2.406	0.264	15.87
WFC	0	2.446	1.288	48.28
SP500	0	1.224	-0.193	11.25
Shiller Index	0	25.54	-0.398	2.200

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

6.2 Estimation results

To find out the relevant input variable, each model estimation is carried out with these nine inputs, two hidden layers for the ANN structures and we use shrinkage priors to identify the series that have predictive power on the conditional variance. Table 6 documents, for each log-returns and for the two ANN structures, the selected inputs. Interestingly, similar variables are selected over the bank returns.

Table 6 – ANN inputs selection using Shrinkage priors

Predictors	Multiple Factor Structure					One 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	√		√	√	√		√	√	√	
SP500 lag. ret.								√		
SP500 lag. sq. ret. 22								√		
PCA lag. ret.	√		√	√	√				√	√
PCA lag. sq. ret. 22										
Shiller Index	√					√				

Once the inputs selected, we need to find out 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 ANN-GARCH models with a number of hidden layers ranging from 0 to 2 (which corresponds to $L = 1$ to $L = 3$).

To begin with, Table 7 documents the MLLs of the ANN-GARCH models with different number of hidden layer. The best model exhibits one hidden layer for the BAC series and no hidden layer for the other series. The fact that the best model for BAC series exhibits two hidden layers suggest that BAC volatility dynamics may be more complex than others bank volatility dynamics.

Table 7 – US banks daily log-return. Mmarginal log-likelihoods (MLLs) of the ANN-GARCH models with different number of hidden layers. The highest value is bolded

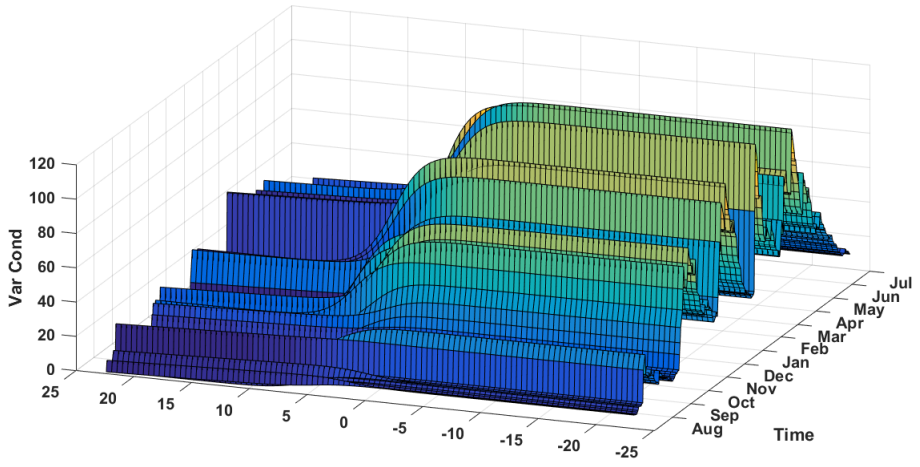
	Number of Hidden Layers		
Series	L =1	L = 2	L = 3
	One factor structure		
BAC	-9309.6	**-9302.4**	-9302.7
C	**-9463.2**	-9464.3	-9467.4
MS	**-9462.1**	-9463.1	-9465.9
GS	**-9214.0**	-9220.5	-9223.1
JPM	**-10220**	-10224	-10229
	Multiple Factor Structure		
BAC	-9302.6	**-9273.4**	-9293.4
C	**-9439.8**	-9463.2	-9469.8
MS	**-9418.7**	-9439.7	-9450.5
GS	**-9194.6**	-9207.2	-9221.1
JPM	**-10205**	-10226	-10230

We observe that multiple factor ANNs always deliver higher MLLs than the one factor structure. This indicates evidence in favor of this complex structure. Additionnally, MLLs rapidly decrease when more hidden layers are considered. As the best model for all the returns is an ANN-GARCH process with a multiple factor structure, we focus on this specification to compare the ANN-GARCH model to existing alternatives as well as to investigate the ANN impacts 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 undercover the inputs influence on the variance. As an example,

Figure 4 investigates the effect of past log-returns on the conditional variance for Goldman Sachs during the financial crisis. In particular, we focus on Goldman Sachs and given the conditional variance generated by the maximum á 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 4 exhibits substantial asymmetries to negatives returns emphasizing that a kind of leverage effect is captured by the lagged log-return input.

Figure 3 – Goldman Sachs variance during the crisis



6.3 Comparison with existing models

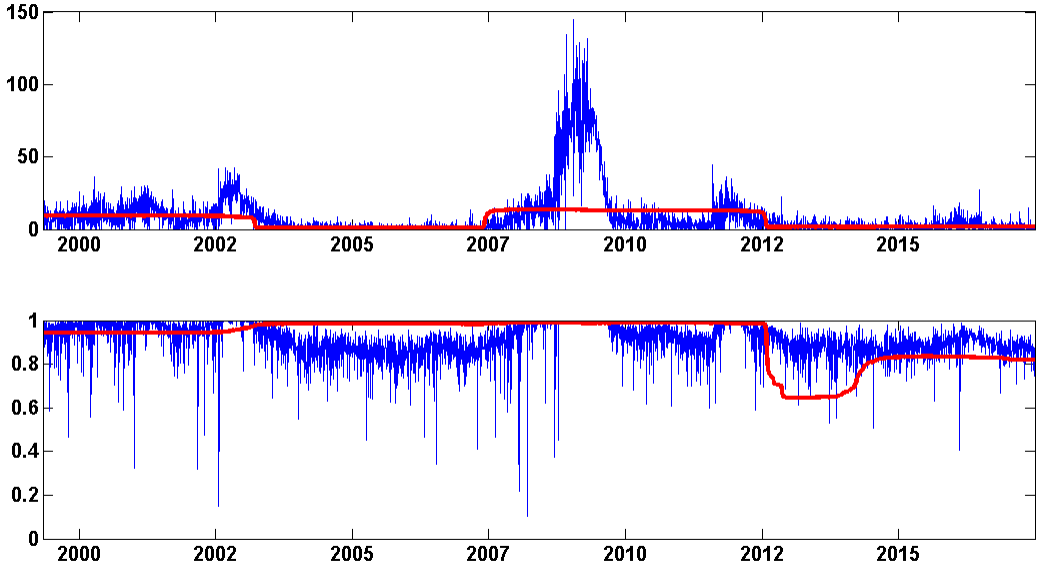
So far, our empirical study has focused on the ANN-GARCH process. We now compare our in-sample results with three alternatives, namely the standard GARCH(1,1) model, the GARCH-X process and the CP-GARCH(1,1) model. The model specifications are provided in Appendix B. Regarding the GARCH-X process, we focus on explanatory variables that lie in the positive support to insure positive conditional variance. However, to improve its performance we consider all the possible combinations of the explanatory variables and we report the best in-sample specification. Table 8 documents the MLLs of these alternative models with respect to the MLL values of the ANN-GARCH process. The ANN-GARCH process strongly dominates the other processes on four out of five series.

Table 8 – Log marginal likelihoods (MLL_s) for estimated models

	MLL			
Series	GARCH	GARCHX	CP-GARCH	ANN-GARCH
BAC	-9356.9	-9334.8	**-9268.7**	-9273.4
C	-9498.4	-9485.9	-9468.8	**-9439.8**
MS	-9482.0	-9467.1	-9470.4	**-9418.7**
GS	-9255.0	-9242.8	-9323.3	**-9194.6**
JPM	-10242	-10231	-10227	**-10205**

As a final analysis, Figure 4 shows the posterior means of the unconditional variance parameter (i.e. $\bar{\omega}_t$) and the persistent parameter (i.e. ϕ_t) obtained from the Goldman Sachs log-returns estimation. These dynamics are in relation with the posterior means of the CP-GARCH (local) unconditional variance (i.e. $\frac{\omega_{s_t}}{1-\alpha_{s_t}-\beta_{s_t}}$) and the CP-GARCH persistence parameter (i.e. $\alpha_{s_t}+\beta_{s_t}$). The CP-GARCH dynamics track well the ANN-GARCH dynamics and act as a smooth approximation of the ANN-GARCH dynamics. It emphasizes the ANN-GARCH ability of capturing highly volatile time-varying parameter dynamics, a feature that is difficult to obtain with a CP-GARCH process.

Figure 4 – Goldman Sachs variance during the crisis



7 Conclusion

In this paper, we propose a new time-varying parameter volatility model, called the artificial neural network GARCH (ANN-GARCH) process, in which parameters are driven by a (deep) neural network. We derive the stationarity conditions as well as the conditions for the existence of moments of the proposed model. Conditions for model identification are also discussed. Using ANNs makes the model highly flexible and allows for the possibility of including additional explanatory variables. In addition to that, the likelihood function of the model is tractable without resorting to any filtering procedure, a nice feature that simplifies the parameter estimation. Regarding the estimation, a Sequential Monte Carlo sampler is proposed 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 algorithm.

The paper ends with an empirical study using five US bank log-returns series that span from May 1999 to August 2017. On one hand, we show that in-sample comparisons with respect to several standard models (i.e. GARCH, GARCH-X and Change-point (CP) GARCH) are in favour of the ANN-GARCH model. On the other hand, the empirical exercise highlights that the ANN-GARCH process can produce more flexible parameter dynamics than the CP-GARCH model.

References

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

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.

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

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

Das, S. and Suganthan, P. (2011). Differential evolution: A survey of the state-of-the-art. *Evolutionary Computation, IEEE Transactions on*, 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.

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

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

Dufays, A., Rombouts, J. V., et al. (2016). Sparse change-point har models for realized variance. Technical report.

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.

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. (2001). Conditional heteroskedasticity driven by hidden markov chains. *Journal of Time Series Analysis*, 22:197–220.

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

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.

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

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.

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

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(01):117–161.

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

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

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. (1986). *Modelling financial time series*. (Chichester: John Wiley).

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:

$$\sigma_t^2 = (1-\alpha)\bar{\omega}_t(1-\tilde{\beta}_t) + \sum_{k=1}^{t-1}\prod_{j=0}^{k-1}(1-\alpha)\bar{\omega}_{t-k}(1-\tilde{\beta}_{t-k})c_{t-1-j} + \prod_{j=0}^{t-1}c_{t-1-j}\sigma_0^2 \qquad (12)$$

$\bar{\omega}_t > 0 \quad \forall t$, $\alpha < 1$ and $\tilde{\beta}_t$ is bounded between 0 and 1 for all t. Therefore, $\sigma_0^2 > 0$ with probability one 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 that, 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. By taking this information and considering the fact 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. For the rest of the proof, see Medeiros and Veiga (2009) and Nelson (1990). Finally, we use the fact that the product of strictly stationary and ergodic series 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:

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

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

$$\mathbf{u}_t = \mathbf{a}_t + \mathbf{C}_{t-1}\mathbf{u}_t - 1 \tag{13}$$

where $\mathbf{a}_t \in \mathbb{R}$ is a vector of specific elements given by $a_{t,j} = \left[(1-\alpha)\bar{\omega}_t(1-\tilde{\beta}_t)\right]^j$, $j = 0, 1, ..., k$ and $\mathbf{C}_{t-1}$ is a $k \times k$ upper triangular matrix with diagonal elements given by

$$\mathbf{C}_{t-1,jj} = \left[(1-\alpha)\tilde{\beta}_t + \alpha\eta_{t-1}^2)\right]^j \quad j = 0, ..., k \tag{14}$$

Let note $\mathscr{I}_{t-2}$ the set of all information available up to time $t-2$. Considering the fact that $a_{t,j}$ is bounded, the conditional expectation of $\mathbf{u}_t$ in equation (13) satisfy 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 multiple factor structure for the neural network

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

(b) Case $\ell = 2$: 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 layer is minimal, we need to show that for all $i = 1, \cdots, d$ $\tilde{x}_{ti}$ cannot be obtained with an $\ell = 1$th hidden layer neural network. To put it another way, consider α_1 and α_2 two reels, we have to show that

$$\alpha_1 f(\omega_{i1}x_{ti} + b_{i1}) + \alpha_2 f(\omega_{i2}f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2}) = 0 \tag{15}$$

if and only if $\alpha_1 = \alpha_2 = 0$. Suppose that (15) hold. We prove that $\alpha_1 = \alpha_2 = 0$.

By studying the decay rate of the terms on the left side of (15) at $x_{ti} \longrightarrow -\infty$, we have that $f(\omega_{i1}x_{ti} + b_{i1}) \longrightarrow 0$. It then follow that

$$\alpha_2 f(b_{i2}) = 0 \tag{16}$$

It follow from equation (16) that $\alpha_2 = 0$ since $f(b_{i2}) \neq 0$ for reasonable values of b_{i2}.

Again, letting $x_{ti} \longrightarrow +\infty$ in (15) imply that

$$\alpha_1 + \alpha_2 f(\omega_{i2} + b_{i2}) = 0 \tag{17}$$

It follow from (17) that $\alpha_1 = -\alpha_2 f(\omega_{i2} + b_{i2}) = 0$ since $\alpha_2 = 0$ and $f(\omega_{i2} + b_{i2}) \neq 0$

(c) Case $\ell = 3$: 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 procedure as for the case $\ell = 2$, let α_1, α_2 and α_3 be three reels and suppose that

$$\alpha_1 f(\omega_{i1}x_{ti} + b_{i1}) + \alpha_2 f\left(\omega_{i2} f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2}\right) + \alpha_3 f\left(\omega_{i3} f(\omega_{i2} f(\omega_{i1}x_{ti} + b_{i1}) + b_{i2}) + b_{i3}\right) = 0 \tag{18}$$

We prove that $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Letting $x_{ti} \longrightarrow +\infty$, we have $f(\omega_{i1}x_{ti} + b_{i1}) \longrightarrow 1$ and equation (18) become

$$\alpha_1 + \alpha_2 f\left(\omega_{i2} + b_{i2}\right) + \alpha_3 f\left(\omega_{i3} f(\omega_{i2} + b_{i2}) + b_{i3}\right) = 0 \tag{19}$$

Under restrictions (R.1) and (R-2), Lemma 2.7 in Hwang and Ding (1997) implied that equation (19) 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 $\ell = 1$: A typical $\ell = 1$ hidden layers artificial neural network with multiple

factors structure give rise to the outputs

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

then O_1 is the first output of the network while O_2 is the second output. Let note $\boldsymbol{W}$ the vector of parameters of the network. Suppose that $\tilde{\boldsymbol{W}}$ is another vector of parameters such that

$$f\left(\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i1} x_{ti} + b_{i1})\right) = f\left(\tilde{\gamma}_{k0} + \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i1} x_{ti} + \tilde{b}_{i1})\right) \quad k \in \{1, 2\} \tag{20}$$

since the function f is a strictly increasing function, equation (20) is equivalent to

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i1} x_{ti} + b_{i1}) = \tilde{\gamma}_{k0} + \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i1} x_{ti} + \tilde{b}_{i1}) \quad k \in \{1, 2\} \tag{21}$$

To prove global identifiability of the model, we need to show that, under restrictions (R.1) and (R-2), (22) is satisfied if and only if $\boldsymbol{W} = \tilde{\boldsymbol{W}}$.

Equation (22) can be rewritten as:

$$\theta_{k0} + \sum_{i=1}^{2d} \theta_{ik} f(\phi_i(\bar{x}_{ti})) = 0 \quad k \in \{1, 2\} \tag{22}$$

where $\theta_{ik} = \gamma_{ik}$ for $i = 1, \cdots, d$, $\theta_{ik} = -\tilde{\gamma}_{i-d,k}$ for $i = d+1, \cdots, 2d$, $\phi_i = \phi_i(x_{ti}) = \omega_{i1} x_{ti} + b_{i1}$ for $i = 1, \cdots, d$ and $\phi_i = \phi_{i-d}(x_{t,i-d}) = \tilde{\omega}_{i-d,1} x_{t,i-d} + \tilde{b}_{i-d,1}$ for $i = d+1, \cdots, 2d$.

Lemma 2.7 in Hwang and Ding (1997) implies that if ϕ_{i1} and ϕ_{i2} are not sign-equivalent (that means $|\phi_{i1}| \neq |\phi_{i2}|$), $i1 \in \{1, \cdots, 2d\}$, $i2 \in \{1, \cdots, 2d\}$, equation (22) is satisfied if and only if θ_{k0} and θ_{ik} vanish jointly for every $i \in \{1, \cdots, 2d\}$ and $k \in \{1, 2\}$. But Restriction (R-2) of Assumption 1 preclude that possibility. Therefore ϕ_{i1} and ϕ_{i2} must be sign-equivalent. But Restriction (R-1) of Assumption 1 preclude that possibility. Hence, the only way for equation (22) to be satisfied is that $\gamma_{k0} = \tilde{\gamma}_{ik}$, $\gamma_{ik} = \tilde{\gamma}_{ik}$, $\omega_{i1} = \tilde{\omega}_{i1}$ and $b_{i1} = \tilde{b}_{i1}$

(b) Case $\ell = 2$: A typical $\ell = 2$ hidden layers artificial neural network with multiple

factors structure give 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, let note $\boldsymbol{W}$ the vector of parameters of the network. Suppose that $\tilde{\boldsymbol{W}}$ is another vector of parameters such that

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i2} f(\omega_{i1} x_{ti} + b_{i1}) + b_{i2}) = \tilde{\gamma}_{k0} + \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i2} f(\tilde{\omega}_{i1} x_{ti} + \tilde{b}_{i1}) + \tilde{b}_{i2}) \tag{23}$$

To prove global identifiability of the model, we need to show that, under restrictions (R.1) and (R-2), (23) is satisfied if and only if $\boldsymbol{W} = \tilde{\boldsymbol{W}}$. The expression $f(\omega_{i1} x_{ti} + b_{i1})$ is a monotone transformation of an identified function. so it is identified. In fact,

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

Then note $\tilde{x}_{ti} = f(\omega_{i1} x_{ti} + b_{i1}) = f(\tilde{\omega}_{i1} x_{ti} + \tilde{b}_{i1})$. Equation (15) became:

$$\gamma_{k0} + \sum_{i=1}^{d} \gamma_{ik} f(\omega_{i2} \tilde{x}_{ti} + b_{i2}) = \tilde{\gamma}_{k0} + \sum_{i=1}^{d} \tilde{\gamma}_{ik} f(\tilde{\omega}_{i2} \tilde{x}_{ti} + \tilde{b}_{i2}) \tag{24}$$

At this point of the proof, equation (16) is similar to equation (22) of case (a). We then use the same arguments as in case (a) to conclude the global identification.

(c) Case $\ell = 3$: A tipical $\ell = 3$ hidden layers artificial neural network with multiple factors structure give 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 (25)

$$\gamma_{k0}+\sum_{i=1}^{d}\gamma_{ik}f(\omega_{i3}f(\omega_{i2}f(\omega_{i1}x_{ti}+b_{i1})+b_{i2})+b_{i3}) = \tilde{\gamma}_{k0}+\sum_{i=1}^{d}\tilde{\gamma}_{ik}f(\tilde{\omega}_{i3}f(\tilde{\omega}_{i2}f(\tilde{\omega}_{i1}x_{ti}+\tilde{b}_{i1})+\tilde{b}_{i2})+\tilde{b}_{i3}) \quad k \in \{1,2\} \tag{25}$$

is satisfied if and only if $\boldsymbol{W} = \tilde{\boldsymbol{W}}$. 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. Then equation (25) became

$$\gamma_{k0} + \sum_{i=1}^{d}\gamma_{ik}f(\omega_{i3}\tilde{x}_{ti} + b_{i3}) = \tilde{\gamma}_{k0} + \sum_{i=1}^{d}\tilde{\gamma}_{ik}f(\tilde{\omega}_{i3}\tilde{x}_{ti} + \tilde{b}_{i3}) \quad k \in \{1,2\} \tag{26}$$

Then by using equation (26) and following the same procedure as in case (a), it is easy to show that (25) is satisfied if and only if all the parameters in the left are equal to those in the right side.

Proof of minimality with one factor structure for the neural network

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

(b) Case $\ell = 2$: With $\ell = 2$ hidden layers, $\tilde{x}_t = f\left(\delta_1 f(\sum_{i=1}^{d}\omega_i x_{ti} + b_1) + b_2\right)$, $i \in \{1, \cdots, d\}$. As for the case of multiple factor structure, the network with $\ell = 2$ hidden layer is minimal if under restrictions (R.3) to (R.5) of Assumption 2, $\tilde{x}_t$ cannot be obtained with an $\ell = 1$th hidden layer neural network. So considering α_1 and α_2 two reals, we have to show that

$$\alpha_1 f\left(\sum_{i=1}^{d}\omega_i x_{ti} + b_1\right) + \alpha_2 f\left(\delta_1 f\left(\sum_{i=1}^{d}\omega_i x_{ti} + b_1\right) + b_2\right) = 0 \tag{27}$$

if and only if $\alpha_1 = \alpha_2 = 0$. Suppose that (15) hold. We prove that $\alpha_1 = \alpha_2 = 0$.

By studying the decay rate of the terms on the left side of (27) at $x_{ti} \longrightarrow -\infty$, we have that $f\left(\sum_{i=1}^{d}\omega_i x_{ti} + b_1\right) \longrightarrow 0$. It then follow that

$$\alpha_2 f(b_2) = 0 \tag{28}$$

It follow from equation (28) that $\alpha_2 = 0$ since $f(b_{i2}) \neq 0$ for reals values of b_2.

Again, letting $x_{ti} \longrightarrow +\infty$ in (27) imply that

$$\alpha_1 + \alpha_2 f(\delta_1 + b_2) = 0 \tag{29}$$

It follow from (29) that $\alpha_1 = -\alpha_2 f(\delta_1 + b_2) = 0$ since $\alpha_2 = 0$ and $f(\delta_1 + b_2) \neq 0$

(c) Case $\ell = 3$: 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 for the case $\ell = 2$, let α_1, α_2 and α_3 be three reals and suppose that

$$\alpha_1 f\left(\sum_{i=1}^{d} \omega_i x_{ti} + b_1\right) + \alpha_2 f\left(\delta_1 f\left(\sum_{i=1}^{d} \omega_i x_{ti} + b_1\right) + b_2\right) + \alpha_3 f\left(\delta_2 f\left(\delta_1 f\left(\sum_{i=1}^{d} \omega_i x_{ti} + b_1\right) + b_2\right) + b_3\right) = 0 \tag{30}$$

We prove that $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Letting $x_{ti} \longrightarrow +\infty$, we have $f\left(\sum_{i=1}^{d} \omega_i x_{ti} + b_1\right) \longrightarrow 1$ and equation (30) become

$$\alpha_1 + \alpha_2 f\left(\delta_1 + b_2\right) + \alpha_3 f\left(\delta_2 f(\delta_1 + b_2) + b_3\right) = 0 \tag{31}$$

Under restrictions (R.3) to (R.5), Lemma 2.7 in Hwang and Ding (1997) implied that equation (31) holds if and only if $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

Proof of global identification with one factor structure for the neural network

The outputs of a typical $\ell = 1$ hidden layers artificial neural network with one factor structure are written as:

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

note $\boldsymbol{W}$ the vector of parameters of the network. Suppose that $\tilde{\boldsymbol{W}}$ is another vector of

parameters such that

$$f\left(\gamma_{k0}+\gamma_k f(\sum_{i=1}^{d}\omega_i x_{ti}+b_1)\right)=f\left(\tilde{\gamma}_{k0}+\tilde{\gamma}_k f(\sum_{i=1}^{d}\tilde{\omega}_i x_{ti}+\tilde{b}_1)\right)\quad k\in\{1,2\} \tag{32}$$

since the function f is a strictly increasing function, equation (32) is equivalent to

$$\gamma_{k0}+\gamma_k f(\sum_{i=1}^{d}\omega_i x_{ti}+b_1)=\tilde{\gamma}_{k0}+\tilde{\gamma}_k f(\sum_{i=1}^{d}\tilde{\omega}_i x_{ti}+\tilde{b}_1)\quad k\in\{1,2\} \tag{33}$$

To prove global identifiability of the model, we need to show that, under restrictions (R.3) to (R.5) of Assumption 2, (33) is satisfied if and only if $\boldsymbol{W}=\tilde{\boldsymbol{W}}$. This equality follows by applying the same steps as in the case of global identification of the multiple factor structure. This also stands for the case $\ell=2$ and $\ell=3$ as well.

□

B Others tables and figures

Figure 5 – ANNs inputs used in the simulation to test the model selection strategy

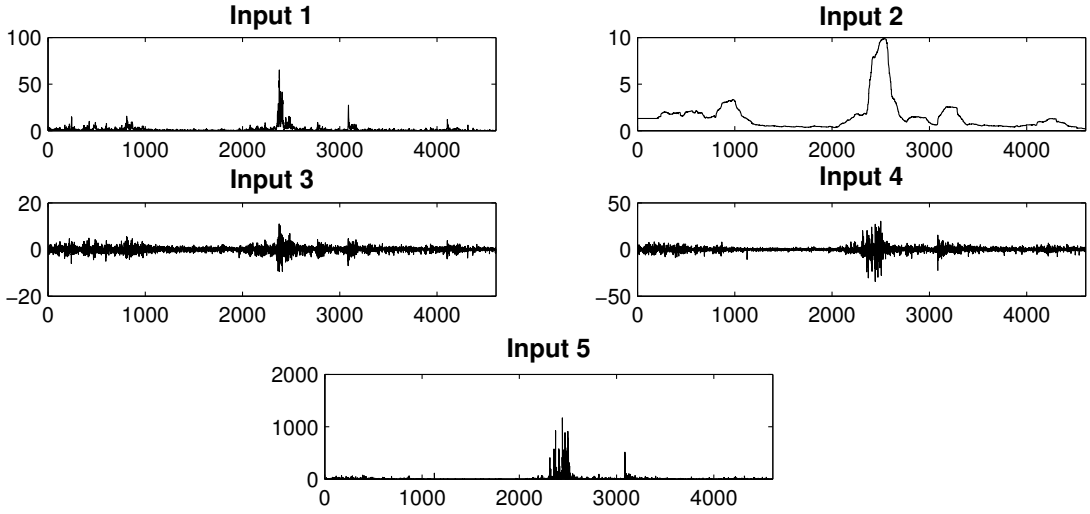


Table 9 – Correlation matrix of inputs

	Input 1	**Input 2**	**Input 3**	**Input 4**	**Input 5**
Input 1	1	0.382	0.043	0.023	0.313
Input 2		1	0.006	0.0003	0.343
Input 3			1	0.662	-0.057
Input 4				1	-0.062
Input 5					1

Table 10 – Specifications of alternative models.

	$y_t = \sigma_t \eta_t \quad \text{with } \eta_t \sim IID(0,1)$
GARCH(1,1)	$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2$
CPGARCH	$\sigma_t^2 = \omega_{s_t} + \alpha_{s_t} y_{t-1}^2 + \beta_{s_t} \sigma_{t-1}^2$
GARCHX	$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2 + \boldsymbol{\omega}' \boldsymbol{X}_{t-1}$

Notes: s_t is an integer random variable taking values in [1,K+1] and $\boldsymbol{X}_t$ is a vector of observed covariates which are assumed satonnary and are squared to ensure that the conditional variance σ_t^2 is always positive.