Working Paper Series

On the Ergodicity and Stationarity of the ARMA(1,1) Recurrent Neural Network Process

Adrian Trapletti
Friedrich Leisch
Kurt Hornik

Working Paper No. 37
May 1999

Working Paper Series

May 1999

SFB
'Adaptive Information Systems and Modelling in Economics and Management Science'

Vienna University of Economics
and Business Administration
Augasse 2–6, 1090 Wien, Austria

in cooperation with
University of Vienna
Vienna University of Technology

http://www.wu-wien.ac.at/am

This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modelling in Economics and Management Science').

On the Ergodicity and Stationarity of the ARMA(1,1) Recurrent Neural Network Process

Adrian Trapletti
Institut für Unternehmensführung
Wirtschaftsuniversität Wien
Augasse 2–6, A-1090 Vienna, Austria
adrian.trapletti@wu-wien.ac.at

Friedrich Leisch
Institut für Statistik
Technische Universität Wien
Wiedner Hauptstr. 8–10
A-1040 Vienna, Austria
friedrich.leisch@ci.tuwien.ac.at

Kurt Hornik
Institut für Statistik
Technische Universität Wien
Wiedner Hauptstr. 8–10
A-1040 Vienna, Austria
kurt.hornik@ci.tuwien.ac.at

Abstract

In this note we consider the autoregressive moving average recurrent neural network ARMA-NN$(1,1)$ process. We show that in contrast to the pure autoregressive process simple ARMA-NN processes exist which are not irreducible. We prove that the controllability of the linear part of the process is sufficient for irreducibility. For the irreducible process essentially the shortcut weight corresponding to the autoregressive part determines whether the overall process is ergodic and stationary.

1 Introduction

We consider the *autoregressive moving average recurrent neural network ARMA-NN(*$1,1$*)* process, which is defined by stochastic difference equations of the form

$$y_t = h(y_{t-1}, \varepsilon_{t-1}, \theta) + \varepsilon_t. \tag{1}$$

$h(y_{t-1}, \varepsilon_{t-1}, \theta)$ denotes a multi layer perceptron (MLP) with input y_{t-1}, feedback ε_{t-1}, and weight (parameter) vector θ. ε_t is a sequence of independently and identically distributed random variables and represents the noise process.

The output of the MLP with q hidden units and activation function G is given by

$$h(y_{t-1}, \varepsilon_{t-1}, \theta) = \psi_1 y_{t-1} + \psi_2 \varepsilon_{t-1} + \nu + \sum_{i=1}^{q} \beta_i G(\phi_{i,1} y_{t-1} + \phi_{i,2} \varepsilon_{t-1} + \mu_i). \tag{2}$$

The weight vectors are the shortcut connections $\psi = (\psi_1, \psi_2)'$, the hidden to output unit weights $\beta = (\nu, \beta_1, \ldots, \beta_q)'$, and the input to hidden unit weights $\phi = (\phi_{1,1}, \ldots, \phi_{q,2}, \mu_1, \ldots, \mu_q)'$ collected together in the network weight vector θ.

The ARMA-NN$(1,1)$, short ARMA-NN, can be seen as an interesting subclass of a more general recurrent neural network model, since the feedback ε_{t-1} essentially represents the difference between the observed value y_{t-1} and the network output $\hat{y}_{t-1} = h(y_{t-2}, \varepsilon_{t-2}, \theta)$ (see, e.g., [1]).

On the other hand the ARMA-NN is a natural generalization of the linear autoregressive moving average ARMA$(1,1)$ process (see, e.g., [2] for a comprehensive introduction into classical time series analysis). Here, $h(\cdot)$ is a linear function of its arguments. Furthermore, the non-linear autoregressive feedforward neural network AR-NN(1) process is a special case of the ARMA-NN with no feedback connection $\psi_2 = \phi_{i,2} \equiv 0$ (see, e.g., [3] for a definition of AR-NNs).

The ergodicity and stationarity of a process is of particular importance for statistics in time series analysis. For such processes a single realization displays the whole probability law of the data generation process (DGP). Recent results exist for AR-NN processes [3, 4], but there are — up to our knowledge — no results giving conditions for the ergodicity and stationarity of the ARMA-NN process. We extend these results for the ARMA-NN process. In particular, we show that controllability of the linear part of the process implies the fundamental property of irreducibility. Concerning the irreducible process essentially the shortcut weight corresponding to the autoregressive part determines whether the overall process is ergodic and stationary.

The remainder of this note is organized as follows: Section 2 introduces some relevant concepts from Markov chain theory and discusses irreducibility of the ARMA-NN process. Section 3 presents sufficient conditions for the ergodicity and stationarity of the ARMA-NN process. All proofs are deferred to the appendix.

2 Irreducibility

To start we put the ARMA-NN in state space form

$$x_t = \Psi x_{t-1} + F(x_{t-1}) + \Sigma \varepsilon_t, \tag{3}$$

where $x_t = (y_t, \varepsilon_t)'$, $\Sigma = (1,1)'$, $\Psi = \begin{bmatrix} \psi_1 & \psi_2 \\ 0 & 0 \end{bmatrix}$, $F(x_{t-1}) = (g(y_{t-1}, \varepsilon_{t-1}, \beta, \phi), 0)'$, and $g(y_{t-1}, \varepsilon_{t-1}, \beta, \phi) = \nu + \sum_{i=1}^{q} \beta_i G(\phi_{i,1} y_{t-1} + \phi_{i,2} \varepsilon_{t-1} + \mu_i)$. Then $\{x_t\}$ is a *Markov chain* with state space $\mathsf{X} \subseteq \mathbb{R}^2$ equipped with the usual Borel σ-field $\mathcal{B}$ and Lebesgue measure λ. [5] give a comprehensive introduction to Markov chains.

The concept of irreducibility is fundamental when considering the structure of a Markov chain: all parts of the state space can be reached by the Markov chain irrespective of the starting point. If $P^t(x, \mathcal{A})$ denotes the transition probability from the state x to the set $\mathcal{A} \in \mathcal{B}$ in t steps, a chain is called *irreducible* if $\sum_{t=1}^{\infty} P^t(x, \mathcal{A}) > 0$ for all $x \in \mathsf{X}$ whenever $\lambda(\mathcal{A}) > 0$.

In contrast to pure autoregressive processes, irreducibility does not hold almost automatically for the ARMA-NN. Suppose, e.g., that in (3) the linear part vanishes. The non-linear part of neural networks is usually bounded, $\|F(\cdot)\| < M < \infty$, where $\|\cdot\|$ denotes Euclidean norm. Then, the set $\mathcal{C} = \{(x, y) \in \mathbb{R}^2, |x - y| \geq M\}$ can never be reached by the Markov chain. However, less trivial examples exist: Consider the following ARMA-NN

$$y_t = \tanh(2y_{t-1} - 2e_{t-1}) + \varepsilon_t. \tag{4}$$

By recursively substituting into equation (4) we get

$$y_t = \tanh(2 \tanh(\dots 2 \tanh(2y_0 - 2e_0) \dots)) + e_t. \tag{5}$$

Now, as $t \to \infty$, the first term on the right hand side of the equation converges to a fixpoint of the function $\tanh(2x)$. $\tanh(2x)$ has three fixpoints: $-0.957\dots$, 0.0, and $0.957\dots$.

Consequently, three different classes of solutions exist for the difference equation (4) corresponding to different initial values $y_0 - e_0$. For example, negative values of $y_0 - e_0$ imply that the first term converges to the negative fixpoint. Each solution is asymptotically stationary. But the process $\{y_t\}$ is not ergodic due to the reducibility of the corresponding Markov chain $\{x_t\}$.

Irreducibility is closely related to the concept of forward accessibility from control theory (see, e.g., [5], Section 7). Equation (3) may be interpreted as a control system driven by the control sequence $\{\varepsilon_t\}$

$$x_t = F_t(x_0, \varepsilon_1, \dots, \varepsilon_t), \tag{6}$$

where the definition of $F_t(\cdot)$ follows inductively from (3). Define $\mathcal{A}_+^t(x)$ as the set of all states which are accessible from x at time t, i.e., $\mathcal{A}_+^0(x) := \{x\}$ and $\mathcal{A}_+^t(x) := \{F_t(x_0, \varepsilon_1, \dots, \varepsilon_t),\ \varepsilon_i \in \mathcal{O}\}$, where the control set $\mathcal{O}$ is an open set in $\mathbb{R}$. Then the control system F_t is called *forward accessible* if the set $\bigcup_{t=0}^{\infty} \mathcal{A}_+^t(x)$ has non-empty interior for each $x \in \mathsf{X}$. Hence, forward accessibility means that the set of reachable states is not concentrated in some lower dimensional subset of X. This property together with an additional assumption on the noise process is essential to ensure that the corresponding Markov process is irreducible.

To state forward accessibility for the ARMA-NN, we consider the following activation functions.

Assumption 1. $G \in C^\infty$ is a bounded, non-constant, and asymptotically constant function.

Note, that most activation functions such as the logistic function $G(x) = (1 + \exp(-x))^{-1}$ and the $\tanh(\cdot)$ squasher satisfy Assumption 1.

Now, a sufficient condition for forward accessibility can be obtained by rewriting the control system (3) as

$$\begin{aligned} x_t &= \Psi x_{t-1} + F(x_{t-1}) + \Sigma\varepsilon_t \\ &= \Psi^2 x_{t-2} + \Psi F(x_{t-2}) + F(x_{t-1}) + \Psi\Sigma\varepsilon_{t-1} + \Sigma\varepsilon_t \\ &= \Psi^2 x_{t-2} + \Psi F(x_{t-2}) + F(x_{t-1}) + (\psi_1 + \psi_2, 0)'\varepsilon_{t-1} + (1,1)'\varepsilon_t. \end{aligned} \tag{7}$$

If both control terms are "active", then the motion of the control system is not concentrated in some lower dimensional subset of X. For the linear system (7) ($F \equiv 0$) this implies that every point of the state space can be reached irrespective of the starting point for some control values ε_{t-1} and ε_t. The linear control system F_t is then called *controllable*.

Assumption 2. The linear part of (3) is controllable, i.e., $\psi_1 + \psi_2 \neq 0$.

Assumption 2 implies that the linear part of (2) does not vanish, i.e., $(\psi_1, \psi_2) \neq 0$.

Proposition 1. *Under Assumptions 1 and 2 the control model* (3) *is forward accessible.*

Under a suitable condition on the distribution of the noise process ε_t, irreducibility of the corresponding Markov chain immediately follows.

Proposition 2. *Suppose the distribution of ε_t is absolutely continuous with respect to the Lebesgue measure λ and the probability density function $\gamma(\cdot)$ of ε_t is positive everywhere in $\mathbb{R}$ and lower semi-continuous. Then, under the conditions in Proposition 1 the Markov chain* (3) *is irreducible on the state space* $(\mathbb{R}^2, \mathcal{B})$.

For example, Gaussian white noise fulfils the conditions of Proposition 2.

Controllability of the linear system is not a necessary condition. However, if the linear part of the system is not controllable, then it is not possible to establish forward accessibility

and irreducibility without considering additional restrictive assumptions on the non-linear part of (3). Furthermore, the state space has to be restricted to some open subset of $\mathbb{R}^2$. Here, we do not further exploit this aspect.

3 Ergodicity and Stationarity

We are interested in conditions on the weights, for which $\{x_t\}$ is a (strictly) stationary process. This problem is closely related to the ergodicity of the process. A Markov chain $\{x_t\}$ is called geometrically ergodic if there exists a probability measure π on $(\mathsf{X}, \mathcal{B})$ and a constant $\varrho > 1$ such that

$$\lim_{t\to\infty} \varrho^t \|P^t(x,\cdot) - \pi(\cdot)\| = 0 \tag{8}$$

for each $x \in \mathsf{X}$ and $\|\cdot\|$ denotes the total variation norm. Then, the distribution of $\{x_t\}$ converges to π and $\{x_t\}$ is asymptotically stationary. If $\{x_t\}$ is started either with initial distribution π or in the infinite past, then $\{x_t\}$ is strictly stationary. Clearly, the same holds for the associated process y_t.

Recall from the results on AR-NNs and on ARMAs that the non-linear part and the feedback term have no influence on the stability of y_t. Thus, it is not surprising that the following result holds.

Proposition 3. *Suppose the Markov chain $\{x_t\}$ of the ARMA-NN$(1,1)$ process satisfies the conditions of Proposition 2 and $\mathsf{E}|\varepsilon_t| < \infty$. Then, a sufficient condition for the geometric ergodicity of the Markov chain $\{x_t\}$ is that $|\psi_1| < 1$.*

It should be noted that Proposition 3 states sufficient conditions for the ergodicity. For example, if $\psi_1 = 1$, then the long term behaviour of the ARMA-NN process is essentially determined by the "state-dependent intercept", i.e., the non-linear part and the intercept of the process: driftless processes exhibit random walk behaviour; a drift towards $+\infty$ or $-\infty$ results in a transient process; a state dependent drift towards the "centre" of the state space gives an ergodic and asymptotically stationary solution (see [4] for an analysis in the context of AR-NNs).

4 Discussion and Conclusions

In this note we have studied several classical concepts from Markov chain theory, control theory, and non-linear time series analysis in the context of the ARMA-NN$(1,1)$ process.

We have discussed the fundamental property of irreducibility for the ARMA-NN and its close relationship to the concept of forward accessibility of the corresponding control system. We have shown that in contrast to the pure autoregressive process simple ARMA-NN exist which are not forward accessible. It turns out that these processes can exhibit troublesome properties for the statistical analysis, e.g., asymptotic stationarity but not ergodicity. The implications are quite dramatic: for example, in general it is not possible to recover the DGP from the observed data.

Furthermore, we have proved that controllability of the linear part of the ARMA-NN control system implies forward accessibility. Although this condition is only sufficient but not necessary, it is a natural condition: it is not necessary to further restrict the non-linear part of the ARMA-NN; the two dimensional system has the whole $\mathbb{R}^2$ as state space and not only some open subset of it. If the support of the distribution of the noise process is sufficiently large, then the associated Markov chain is also irreducible.

Concerning the irreducible ARMA-NN, our results show that essentially the shortcut weight corresponding to the autoregressive part determines whether the overall process

is ergodic and stationary. Hence, the standard condition from linear time series analysis can be used. Ergodicity of the ARMA-NN process has wide ranging consequences for statistics. For example, it implies that the stationary ARMA-NN is also strong mixing with mixing coefficients vanishing exponentially fast. It also implies that an integrated (in the sense of classical time series analysis) standardized ARMA-NN process "converges" to a Wiener process. It can also be shown that — under suitable regularity conditions — it is possible to consistently estimate the weights of the ARMA-NN model from a realization (path) of the ARMA-NN process. See [4] for a rigorous analysis of these connections in the context of AR-NNs.

For practical applications where the ARMA-NN model is estimated from an observed time series this study provides several useful results: In the ideal situation, where the data is in fact generated by the irreducible ARMA-NN process, the estimated weights are not too far from the true weights (consistency). The estimated weights can be used to draw indirect conclusions about the statistical nature of the observed data: If the estimated sum of the shortcut weights is sufficiently different from zero, then the DGP is not likely to come from an irreducible model. If the estimated shortcut weight corresponding to the autoregressive part is in absolute value less than one, then the DGP is also ergodic and stationary. On the other hand, if one of these conditions is violated, then the model is likely to be misspecified and the estimation results should be interpreted with greatest care.

The reported research is currently extended towards the general and possibly multivariate ARMA-NN(p, q) process. The ARMA-NN(p, q) class is considerably more complex than the ARMA-NN($1, 1$). But it seems that under conditions similar to the ARMA-NN($1, 1$) case, results for the general process will be obtained.

Acknowledgments

This piece of research was partly supported by the Austrian Science Foundation (FWF) under SFB#010 ("Adaptive Information Systems and Modelling in Economics and Management Science").

Appendix: Mathematical Proofs

Proof of Proposition 1. We first state a useful Lemma. Related Lemmas are Lemma 2.5–2.7 in [6].

Lemma 1. *If Assumption 1 holds, then for any positive integer k and any scalars β_0, β_i, μ_i, and $\phi_i \neq 0$ ($i = 1, \ldots, k$), the condition*

$$\beta_0 + \sum_{i=1}^{k} \beta_i G'(\phi_i x + \mu_i) = 0, \quad \forall x \in \mathbb{R} \tag{9}$$

implies that $\beta_0 = 0$.

Proof of Lemma 1. Considering the limes $x \to \infty$ of (9) yields the result. □

To continue we use Proposition 7.1.4 from [5]. The generalized controllability matrix is $C^2_{x_0} = \begin{bmatrix} c & 1 \\ 0 & 1 \end{bmatrix}$, where $c = \psi_1 + \psi_2 + \sum_{i=1}^{q} \beta_i(\phi_{i,1} + \phi_{i,2}) G'((\phi_{i,1} + \phi_{i,2})\varepsilon_1 + \phi_{i,1}\hat{y}_1 + \mu_i)$. Because the non-singularity of $C^2_{x_0}$ is equivalent with $c \neq 0$, it is sufficient to show that for all $\hat{y}_1$, there exists $\varepsilon_1 \in \mathcal{O}$, such that $c \neq 0$. Set $\mathcal{O} \equiv \mathbb{R}$ and choose any $\hat{y}_1$.

Then, Assumption 2 implies that $c \neq 0$ for at least one $\varepsilon_1 \in \mathcal{O}$ (Lemma 1). □

Proof of Proposition 2. We use Proposition 7.2.5 and Theorem 7.2.6 from [5].

We show that $x^* = (0, 0)$ is a globally attracting state in the sense of [5] for the control system (3). The second component of x_t can trivially reach the point 0 in one step irrespective of the

starting point. The iterated (from time $t = 0$ to $t = 2$) first component can be written as $y_2 = \ldots + (\psi_1 + \psi_2)\varepsilon_1 + g(\ldots + \varepsilon_1, \varepsilon_1, \beta, \phi)$, where all terms which are functions of the starting point or the second component are omitted. Now, the first component can reach 0 in two steps irrespective of the second component and the starting point, because $\psi_1 + \psi_2 \neq 0$ and $g(\cdot)$ is bounded and continuous. By the same argumentation every state in $\mathbb{R}^2$ can be reached in two steps, the state space is connected, and, hence, the Markov chain is also aperiodic ([5], Proposition 7.3.4). □

Proof of Proposition 3. Geometric ergodicity can be proved using the drift criterion of [5], Theorem 15.0.1. The proof follows identically to [4], Theorem 1, because the absolute value of the largest eigenvalue of Ψ is less than one. □

References

[1] J. Connor, L. E. Atlas, and D. R. Martin. Recurrent networks and NARMA modeling. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 301–308. Morgan Kaufmann Publishers, Inc., 1992.

[2] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer Verlag, New York, 2nd edition, 1991.

[3] F. Leisch, A. Trapletti, and K. Hornik. Stationarity and stability of autoregressive neural network processes. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, USA, 1999. To appear.

[4] A. Trapletti, F. Leisch, and K. Hornik. Stationary and integrated autoregressive neural network processes. Working Paper No. 24, SFB#010 ("Adaptive Information Systems and Modelling in Economics and Managment Science"), Augasse 2-6, A-1090 Vienna, Austria, 1998. http://www.wu-wien.ac.at/am/workpap.html.

[5] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Verlag, London, 1993.

[6] J. T. G. Hwang and A. A. Ding. Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, 92:748–757, 1997.