On the Stationarity of Autoregressive Neural Network Models

Friedrich Leisch
Adrian Trapletti
Kurt Hornik

Report No. 21
October 1998
October 1998

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

Papers published in this report series are preliminary versions of journal articles and not for quotations.

This paper was accepted for publication in:

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 Stationarity of Autoregressive Neural Network Models

Friedrich Leisch¹, Adrian Trapletti² & Kurt Hornik¹

¹ Institut für Statistik
Technische Universität Wien
Wiedner Hauptstraße 8–10 / 1071
A-1040 Wien, Austria
firstname.lastname@cit.wien.ac.at

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

Abstract

We analyze the asymptotic behavior of autoregressive neural network (AR-NN) processes using techniques from Markov chains and non-linear time series analysis. It is shown that standard AR-NNs without shortcut connections are asymptotically stationary. If linear shortcut connections are allowed, only the shortcut weights determine whether the overall system is stationary, hence standard conditions for linear AR processes can be used.

1 Introduction

In this paper we consider the popular class of nonlinear autoregressive processes driven by additive noise, which are defined by stochastic difference equations of form

\[ \xi_t = g(\xi_{t-1}, \ldots, \xi_{t-p}, \theta) + \epsilon_t \]  

where \( \epsilon_t \) is an iid. noise process. If \( g(\cdot, \theta) \) is a feedforward neural network with parameter ("weight") vector \( \theta \), we call Equation 1 an autoregressive neural network process of order \( p \), short AR-NN(\( p \)) in the following.

AR-NNs are a natural generalization of the classic linear autoregressive AR(\( p \)) process

\[ \xi_t = a_1\xi_{t-1} + \cdots + a_p\xi_{t-p} + \epsilon_t. \]  

See, e.g., Brockwell & Davis (1987) for a comprehensive introduction into AR and ARMA (autoregressive moving average) models.

One of the most central questions in linear time series theory is the stationarity of the model, i.e., whether the probabilistic structure of the series is constant over time or at least asymptotically constant (when not started in equilibrium). Surprisingly, this question has not gained much interest in the NN literature, especially there are—up to our knowledge—no results giving conditions for the stationarity of AR-NN models. There are results on the stationarity of Hopfield nets (Wang & Sheng, 1996), but these nets cannot be used to estimate conditional expectations for time series prediction.
The rest of this paper is organized as follows: In Section 2 we recall some results from time series analysis and Markov chain theory defining the relationship between a time series and its associated Markov chain. In Section 3 we use these results to establish that standard AR-NN models without shortcut connections are stationary. We also give conditions for AR-NN models with shortcut connections to be stationary. Section 4 examines the NN modeling of an important class of non-stationary time series, namely integrated series. Finally we illustrate the results with examples in Section 5. All proofs are deferred to the appendix.

2 Some Time Series and Markov Chain Theory

2.1 Stationarity

Let $\xi_t$ denote a time series generated by a (possibly nonlinear) autoregressive process as defined in (1). If $IE \xi_t = 0$, then $g$ equals the conditional expectation $IE(\xi_t|\xi_{t-1}, \ldots, \xi_{t-p})$ and $g(\xi_{t-1}, \ldots, \xi_{t-p})$ is the best prediction for $\xi_t$ in the mean square sense.  

If we are interested in the long term properties of the series, we may ask whether certain features such as mean or variance change over time or remain constant. The time series is called weakly stationary if $IE \xi_t = \mu$ and $cov(\xi_t, \xi_{t+h}) = \gamma_h$, $\forall t$, i.e., mean and covariances do not depend on the time $t$. A stronger criterion is that the whole distribution (and not only mean and variance) of the process does not depend on the time, in this case the series is called strictly stationary. Strong stationarity implies weak stationarity if the second moments of the series exist. For details see standard time series textbooks such as Brockwell & Davis (1987).

If $\xi_t$ is strictly stationary, then $IP(\xi_t \in A) = \pi(A), \forall t$ and $\pi(\cdot)$ is called the stationary distribution of the series. Obviously the series can only be stationary from the beginning if it is started with the stationary distribution such that $\xi_0 \sim \pi$. If it is not started with $\pi$, e.g., because $\xi_0$ is a constant, then we call the series asymptotically stationary if it converges to its stationary distribution:

$$\lim_{t \to \infty} IP(\xi_t \in A) = \pi(A)$$

2.2 Time Series as Markov Chains

Using the notation

$$x_{t-1} = (\xi_{t-1}, \ldots, \xi_{t-p})'$$

$G(x_{t-1}) = (g(x_{t-1}, \xi_{t-1}, \ldots, \xi_{t-p+1})'$

$$c_t = (c_t, 0, \ldots, 0)'$$

we can write scalar autoregressive models of order $p$ as such as (1) or (2) as a first order vector model

$$x_t = G(x_{t-1}) + c_t$$

with $x_t, c_t \in \mathbb{R}^p$ (e.g., Chan & Tong, 1985). If we write

$$p^n(x, A) = IP\{x_{t+n} \in A|x_t = x\}$$

$$p(x, A) = p^1(x, A)$$

2
for the probability of going from point \( x \) to set \( A \in \mathcal{B} \) in \( n \) steps, then \( \{x_t\} \) with \( p(x, A) \) forms a Markov chain with state space \((\mathbb{R}^p, \mathcal{B}, \lambda)\), where \( \mathcal{B} \) are the Borel sets on \( \mathbb{R}^p \) and \( \lambda \) is the usual Lebesgue measure.

The Markov chain \( \{x_t\} \) is called \textit{\( \varphi \)-irreducible}, if for some \( \sigma\)-finite measure \( \varphi \) on \((\mathbb{R}^p, \mathcal{B}, \lambda)\)
\[
\forall x \in \mathbb{R}^p : \sum_{n=1}^{\infty} p^n(x, A) > 0
\]
whenever \( \varphi(A) > 0 \). This means essentially, that all parts of the state space can be reached by the Markov chain irrespective of the starting point. Another important property of Markov chains is \textit{aperiodicity}, which loosely speaking means that there are no (infinitely often repeated) cycles. See, e.g., Tong (1990) for details.

The Markov chain \( \{x_t\} \) is called \textit{geometrically ergodic}, if there exists a probability measure \( \pi(A) \) on \((\mathbb{R}^p, \mathcal{B}, \lambda)\) and a \( \rho > 1 \) such that
\[
\forall x \in \mathbb{R}^p : \lim_{n \to \infty} \rho^n |p^n(x, \cdot) - \pi(\cdot)| = 0
\]
where \(|\cdot|\) denotes the total variation. Then \( \pi \) satisfies the invariance equation
\[
\pi(A) = \int p(x, A) \pi(dx), \quad \forall A \in \mathcal{B}
\]

There is a close relationship between a time series and its associated Markov chain. If the Markov chain is geometrically ergodic, then its distribution will converge to \( \pi \) and the time series is asymptotically stationary. If the time series is started with distribution \( \pi \), i.e., \( x_0 \sim \pi \), then the series \( \{\xi_t\} \) is strictly stationary.

3 Stationarity of AR-NN Models

We now apply the concepts defined in Section 2 to the case where \( g \) is defined by a neural network. Let \( x \) denote a \( p \)-dimensional input vector, then we consider the following standard network architectures:

**Single hidden layer perceptrons:**
\[
g(x) = g_0 + \sum_i \beta_i \sigma(a_i + a'_i x)
\]
(7)

where \( a_i, \beta_i \) and \( g_0 \) are scalar weights, \( a_i \) are \( p \)-dimensional weight vectors, and \( \sigma(\cdot) \) is a bounded sigmoid function such as \( \tanh(\cdot) \).

**Single hidden layer perceptrons with shortcut connections:**
\[
g(x) = g_0 + c x + \sum_i \beta_i \sigma(a_i + a'_i x)
\]
(8)

where \( c \) is an additional weight vector for shortcut connections between inputs and output. In this case we define the characteristic polynomial \( c(z) \) associated with the linear shortcuts as
\[
c(z) = 1 - c_1 z - c_2 z^2 - \ldots - c_p z^p, \quad z \in \mathbb{C}.
\]
Radial basis function networks:

\[ g(x) = \gamma_0 + \sum_i \beta_i \phi(a_i | x - m_i |) \]  

(9)

where \( m_i \) are center vectors and \( \phi(\cdot \cdot \cdot) \) is one of the usual bounded radial basis functions such as \( \phi(x) = \exp(-x^2) \).

Lemma 1 Let \( \{ x_t \} \) be defined by (6), let \( 1E \{ \xi_t \} < \infty \) and let the PDF of \( \xi_t \) be positive everywhere in \( \mathbb{R} \). Then if \( g \) is defined by any of (7), (8) or (9), the Markov chain \( \{ x_t \} \) is \( \phi \)-irreducible and aperiodic.

Lemma 1 basically says that the state space of the Markov chain, i.e., the points that can be reached, cannot be reduced depending on the starting point. An example for a reducible Markov chain would be a series that is always positive if only \( x_0 > 0 \) (and negative otherwise). This cannot happen in the AR-NN(p) case due to the unbounded additive noise term.

Theorem 1 Let \( \{ \xi_t \} \) be defined by (1), \( \{ x_t \} \) by (6), further let \( 1E \{ \xi_t \} < \infty \) and the PDF of \( \xi_t \) be positive everywhere in \( \mathbb{R} \). Then

1. If \( g \) is a network without linear shortcuts as defined in (7) and (9), then \( \{ x_t \} \) is geometrically ergodic and \( \{ \xi_t \} \) is asymptotically stationary.

2. If \( g \) is a network with linear shortcuts as defined in (8) and additionally \( c(z) \neq 0, \forall z \in \mathbb{C} : |z| \leq 1 \), then \( \{ x_t \} \) is geometrically ergodic and \( \{ \xi_t \} \) is asymptotically stationary.

The time series \( \{ \xi_t \} \) remains stationary if we allow for more than one hidden layer (→ multi layer perceptron, MLP) or non-linear output units, as long as the overall mapping has bounded range. An MLP with shortcut connections combines a (possibly non-stationary) linear AR(p) process with a non-linear stationary NN part. Thus, the NN part can be used to model non-linear fluctuations around a linear process like a random walk.

The only part of the network that controls whether the overall process is stationary are the linear shortcut connections (if present). If there are no shortcuts, then the process is always stationary. With shortcuts, the usual test for stability of a linear system applies.

4 Integrated Models

An important method in classic time series analysis is to first transform a non-stationary series into a stationary one and then model the remainder by a stationary process. The probably most popular models of this kind are autoregressive integrated moving average (ARIMA) models, which can be transformed into stationary ARIMA processes by simple differencing.

Let \( \Delta^k \) denote the \( k \)-th order difference operator

\[ \Delta \xi_t = \xi_t - \xi_{t-1} \]  

(10)

\[ \Delta^2 \xi_t = \Delta(\Delta \xi_{t-1}) = \xi_t - 2\xi_{t-1} + \xi_{t-2} \]  

(11)

\[ \cdots \]

\[ \Delta^k \xi_t = \Delta(\Delta \cdots (\Delta \xi_t)) = \sum_{n=0}^{k} (-1)^n \binom{k}{n} \xi_{t-n} \]  

(12)
with $\Delta^1 = \Delta$. E.g., a standard random walk $\xi_t = \xi_{t-1} + \epsilon_t$ is non-stationary because of the growing variance, but can be transformed into the iid (and hence stationary) noise process $\epsilon_t$ by taking first differences.

If a time series is non-stationary, but can be transformed into a stationary series by taking $k$-th differences, we call the series integrated of order $k$. Standard MLPs or RBFs without shortcuts are asymptotically stationary. It is therefore important to take care that these networks are only used to model stationary processes. Of course the network can be trained to mimic a non-stationary process on a finite time interval, but the out-of-sample or prediction performance will be poor, because the network inherently cannot capture some important features of the process (see the example in Section 5.1). One way to overcome this problem is to first transform the process into a stationary series (e.g., by differencing an integrated series) and train the network on the transformed series (Chng et al., 1996).

As differencing is a linear operation, this transformation can also be easily incorporated into the network by choosing the shortcut connections and weights from input to hidden units accordingly. Assume we want to model an integrated series of integration order $k$, such that

$$\Delta^k \xi_t = g(\Delta^k \xi_{t-1}, \ldots, \Delta^k \xi_{t-p}) + \epsilon_t$$

where $\Delta^k \xi_t$ is stationary. By (12) this is equivalent to

$$\xi_t = \sum_{n=1}^{k} (-1)^{n-1} \binom{k}{n} \xi_{t-n} + g(\Delta^k \xi_{t-1}, \ldots, \Delta^k \xi_{t-p}) + \epsilon_t$$

which (for $p > k$) can be modeled by an MLP with shortcut connections as defined by (8) where the shortcut weight vector $c$ is fixed to

$$c = \left( \binom{k}{1}, \ldots, (-1)^{p-1} \binom{k}{p} \right)' = 0 \text{ for } n > k$$

and $\tilde{g}$ is such that $\tilde{g}(\xi_{t-1}, \ldots, \xi_{t-p-k}) = g(\Delta^k x_{t-1})$. This is always possible and can basically be obtained by adding $c$ to all weights between input and first hidden layer of $g$.

An AR-NN(p) can model integrated series up to integration order $p$. If the order of integration is known, the shortcut weights can either be fixed, or the differenced series is used as input. If the order is unknown, we can also train the complete network including the shortcut connections and implicitly estimate the order of integration. After training the final model can be checked for stationarity by looking at the characteristic roots of the polynomial defined by the shortcut connections.

### 4.1 Fractional Integration

Up to now we have only considered integrated series with positive integer order of integration, i.e., $k \in \mathbb{N}$. In the last years models with fractional integration order became very popular (again). Series with integration order of $0.5 < k < 1$
\[ \tilde{z} + 1 - \tilde{z} = \tilde{z} \]

Figures 1 and 2 show a simple random walk.

\[ z \]

\[ \text{Figure 1: Training set for a simple random walk (top) and residuals of a standard} \]

\[ \text{exponential distribution (bottom).} \]

\[ \text{Figure 2: MLp without shortcuts (bottom).} \]
As another example we consider the process defined by

$$y(t) = (y(t-1) + \gamma) \mod 2$$

Random Walk with Nonlinearity

\[ \text{Figure 2: Test set for a simple random walk (top) and residuals of a standard} \]

\[ \text{Figure 3: Test set for a simple random walk (top) and residuals of a standard} \]
\[
\begin{align*}
    h(x) &= \sigma(x) \ast (1 - \sigma(x)) \ast (1 - 2 \ast \sigma(x)) \\
    \sigma(x) &= (1 + e^{-x})^{-1}
\end{align*}
\]

where \(\sigma(\cdot)\) is the usual sigmoid function, \(h(\cdot)\) is the second derivative of \(\sigma(\cdot)\) and \(\epsilon_t\) is iid. standard normal. The non-linearity was chosen such that it can be learned without too much difficulty by an MLP, yet is strongly non-linear (local minimum and maximum). Without \(h\), this example would be a standard random walk.

The non-linearity \(h\) (see Figure 3) adds fluctuations around zero, i.e., small positive values are transformed into negative values and vice versa. Figure 2 shows a typical sample path for this process. For larger values \(|k_t| > 50\) the series looks like a Brownian motion, for smaller values it exhibits strong non-linear fluctuations due to the influence of \(h\).

We trained an MLP (with and without shortcut) using one input \((\xi_{t-1})\), 6 hidden nodes and one output node \((\xi_t)\) on this training sample of size 10000. Obviously an example with only one input is rather small, but it has the advantage that the estimate of the conditional expectation \(IE(\xi_t | k_{t-1})\) can easily be plotted, hence the example is ideal for demonstration purposes.

Figure 5 shows the outputs of the two trained networks. Both capture the non-linearity quite well, of course only the model with shortcut connections captures the Brownian motion part \((f(x) = x)\). The weight of the shortcut
connection after training was 0.9985, which is quite close to the true value of 1.

6 Summary

We have shown that AR-NN models using standard NN architectures without shortcuts are asymptotically stationary. If linear shortcuts between inputs and outputs are included—which many popular software packages have already implemented—then only the weights of the shortcut connections determine if the overall system is stationary. It is also possible to model many integrated time series by this kind of networks.

The asymptotic behavior of AR-NNs is especially important for predictions over larger intervals of time or when using the network to generate artificial time series. E.g., one might train the network on an available sample and then use the trained network afterwards—driven by artificial noise from a random number generator—to generate new data with similar properties than the training sample. The asymptotic stationarity guarantees that the AR-NN model cannot show "explosive" behavior or growing variance with time.

Acknowledgement

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

Appendix: Mathematical Proofs

Proof of Lemma 1

It can easily be shown that \( \{ x_t \} \) is \( \varphi \)-irreducible if the support of the probability density function (PDF) of \( q_t \) is the whole real line, i.e., the PDF is positive everywhere in \( \mathbb{R} \) (Chan & Tong, 1985). In this case every non-null \( p \)-dimensional hypercube is reached in \( p \) steps with positive probability (and hence every non-null Borel set \( A \)).

A necessary and sufficient condition for \( \{ x_t \} \) to be aperiodic is that there exists a set \( A \) and positive integer \( n \) such that \( p^n(x,A) > 0 \) and \( p^{n+1}(x,A) > 0 \) for all \( x \in A \) (Tong, 1990, p. 455). In our case this is true for all \( n \) due to the unbounded additive noise.

Proof of Theorem 1

We use the following result from nonlinear time series theory:

Theorem 2 (Chan & Tong 1985) Let \( \{ x_t \} \) be defined by (1), (6) and let \( G \) be compact, i.e. preserve compact sets. If \( G \) can be decomposed as \( G = G_b + G_d \) and

1. \( G_d(\cdot) \) is of bounded range

2. \( G_b(\cdot) \) is continuous and homogeneous, i.e., \( G_b(\alpha x) = \alpha G_b(x) \)
3. the origin is a fixed point of $G_h$ and $G_h$ is uniform asymptotically stable $4$. $|\xi| < \infty$ and the PDF of $\xi$ is positive everywhere in $\mathbb{R}$ then $\{x_t\}$ is geometrically ergodic.

The noise process $\xi$ fulfills condition 4 by assumption (e.g., Gaussian noise). Clearly all networks are continuous compact functions. Standard MLPs without shortcut connections and RBFs have a bounded range, hence $G_h \equiv 0$ and $G \equiv G_d$, and the series $\{\xi_t\}$ is asymptotically stationary.

If we allow for linear shortcut connections between the input and the outputs, we get

$$G_h = \epsilon x$$

$$G_d = \gamma + \sum_i \beta_i \sigma(a_i + d_i x)$$

i.e., $G_h$ is the linear shortcut part of the network, and $G_d$ is a standard MLP without shortcut connections. Clearly, $G_h$ is continuous, homogeneous and has the origin as a fixed point. Hence, the series $\{\xi_t\}$ is asymptotically stationary if $G_h$ is asymptotically stable, i.e., when all characteristic roots of $G_h$ have a magnitude less than unity. Obviously the same is true for RBFs with shortcut connections.

Note that the model reduces to a standard linear AR($p$) model if $G_d \equiv 0$.

References


