Identifiability of Recurrent Neural Networks
Author(s): A. A. Al-Falou and D. Trummer
Source: *Econometric Theory*, Vol. 19, No. 5 (Oct., 2003), pp. 812-828
Published by: Cambridge University Press
Stable URL: http://www.jstor.org/stable/3533436
Accessed: 19-01-2018 16:02 UTC

Econometric Theory, **19**, 2003, 812–828. Printed in the United States of America.
DOI: 10.1017/S0266466603195059

IDENTIFIABILITY OF RECURRENT NEURAL NETWORKS

A.A. AL-FALOU AND D. TRUMMER
Vienna University of Technology

We examine the identifiability of a nonlinear state space system under general assumptions. The discrete time evolution of the state is generated by a recurrent Elman network. For a large set of Elman networks we determine the class of observationally equivalent minimal systems, i.e., minimal systems that exhibit the same input-output behavior.

1. INTRODUCTION

In recent years neural networks have attracted increasingly more interest in the statistical analysis of time series. In spite of an extensive literature on practical applications (see, e.g., Granger, Lee, and White, 1993; Draisma and Franses, 1997; Richards and Baker, 1999; Tkacz, 2001; Gencay and Garcia, 2000; and references therein) theoreticians have only recently begun work on the underlying mathematics of neural networks in the structural and statistical analysis (see Albertini and Sontag, 1993; Albertini and Dai Pra, 1995; Albertini, 1993). The final aim of this research is to build a solid mathematical theory similar to the well-established mathematics of linear time series analysis (see, e.g., Brockwell and Davis, 1991; Hannan and Deistler, 1988).

In this paper we shall investigate an important question within structure theory, namely, the properties of observationally equivalent recurrent neural nets, i.e., neural nets that exhibit the same input-output behavior. This is part of a still incomplete theory of identification of recurrent neural nets. The theory of identification is important to understand further research on consistency theory and efficient estimation algorithms (Trapletti, Leisch, and Hornik, 1998; Trapletti, Leisch, and Hornik, 1999; Leisch, Trapletti, and Hornik, 1999).

Despite the pioneering work in Albertini and Sontag (1993) and Albertini and Dai Pra (1995), the conditions needed for the characterization of the set of observationally equivalent systems have not been completely analyzed yet. Albertini and Dai Pra (1995) restrict their systems such that they obtain results

The authors are grateful for discussions with M. Deistler and D. Bauer. A.A. Al-Falou was funded by the ERNSI network within the European Union program Training and Mobility of Researchers (TMR). D. Trummer was funded by the FWF (Austrian Science Fund). We thank an anonymous referee for helpful comments and suggestions. Address correspondence to: A.A. Al-Falou, Vienna University of Technology, Argentinierstr. 8, 1040 Wien, Austria; e-mail: Al-Falou@t-online.de.

  0266-4666/03 $12.00

similar to linear state space systems. In their case, the class of observationally equivalent systems is characterized by transformations via a permutation matrix including sign changes. We show that the assumptions on the nonlinearity can be weakened to identify a larger set of recurrent Elman networks. One consequence is that the Heaviside function (which is a classical nonsmooth transfer function in neural nets) plays an exceptional role. We will show that it is in some sense (via the admissibility of the nonlinearity) related to the function $1/x$ that is the only rational function for which the class of observationally equivalent systems is no longer obtained by transformations via a permutation matrix.

2. BASIC NOTATION AND DEFINITIONS

We shall begin this section with the notation of a general Elman recurrent neural network (see also Elman, 1989, 1991). Then we need to specify the sigmoid function of the network. Our aim is to keep the sigmoid function as general as to include previous definitions (Albertini and Sontag, 1993; Albertini and Dai Pra, 1995; Dörfler and Deistler, 1998) and also to include the Heaviside function, which is of practical importance. In a third step we need to ensure that the inputs generate sufficient excitations as to separate two observationally nonequivalent systems.

We consider a recurrent neural network of the form

$$\begin{aligned} \underline{x}_{t+1} &= \underline{\sigma}(A\underline{x}_t + B\underline{u}_t), \\ \underline{y}_t &= C\underline{x}_t, \end{aligned} \tag{1}$$

where

$$\underline{x}_t \in \mathbb{R}^n, \qquad \underline{u}_t \in \mathbb{R}^m, \qquad \text{and} \quad \underline{y}_t \in \mathbb{R}^p.$$

The vector $\underline{x}_t$ denotes the state of the system at time t; $\underline{u}_t$ and $\underline{y}_t$ are the input and the output of the system, respectively. The system is initialized by $\underline{x}_0$. The nonlinear function $\sigma : \mathbb{R} \to \mathbb{R}$, which is assumed to be a sigmoid odd function, acts componentwise on the vector $\underline{x} = (x_1, \ldots, x_n)$, i.e.,

$$\underline{\sigma}(x_1, \ldots, x_n)' = (\sigma(x_1), \ldots, \sigma(x_n))'. \tag{2}$$

In addition to the assumption that σ is odd we shall need the following independence property. Let D be $\mathbb{R}$ without the finite set of poles of g.

DEFINITION 2.1. *We assume that $g : D \to \mathbb{R}$ is a rational odd or even function on $\mathbb{R}$. An odd function $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies the g-independence property (g-IP) if for any $N \in \mathbb{N}$ and any set*

$$(a_1, b_1), \ldots, (a_N, b_N) \in \mathbb{R}^2 \tag{3}$$

such that $a_i g(b_i) \neq \pm a_j g(b_j)$ *for* $i \neq j$ *and* $b_i \neq 0$ *the functions*

$$\xi \to \sigma(a_i + \xi\, b_i)$$

are linearly independent.

Remark 2.2.

(i) In the previous definition we could allow for general non-odd and non-even g. However in this case, we could always find a pair $(a_1, b_1), (a_2, b_2)$, $a_1 = -a_2$ and $b_1 = -b_2$ that satisfies (3) but for which $\sigma(a_1 + b_1\xi) = -\sigma(a_2 + b_2\xi)$. For such a non-odd and non-even g the set of sigmoid functions satisfying the g-IP property would be empty. Therefore it is reasonable to restrict our definition to odd or even g.

(ii) We shall briefly discuss two important cases of g-IP. First, the case $g \equiv 1$ generalizes previous definitions of the independence property and future results on identifiability (see Albertini and Sontag, 1993; Albertini and Dai Pra, 1995; Dörfler and Deistler, 1998). The case $g \equiv 1$ includes many commonly used sigmoids such as $\sigma = \tanh$ or the logistic function.

(iii) Unfortunately, the previous definitions of the independence property have excluded the Heaviside function. The Heaviside (or step) function, which is often used in neural network modeling, is defined by

$$H(x) = 1 \quad \text{for } x > 0, \qquad H(0) = 0, \qquad \text{and} \quad H(x) = -1 \quad \text{for } x < 0.$$

It is clear that the Heaviside function does not obey the previous notion of the independence property. However, it is straightforward to show that the Heaviside function obeys the g-IP if $g(x) = \text{constant}/x$.

In the next section on identification we need to ensure that not all inputs have the same effect on different components of $\underline{\sigma}$. Such would be the case if two rows of B were identical. This leads to the notion of admissibility. Throughout this paper we shall assume that the matrix B is admissible.

DEFINITION 2.3. *The parameter matrix* $B \in \mathbb{R}^{n\times m}$ *in (1) is called admissible if* $\underline{B}_i \neq 0$ *for all* i *and*

$$\underline{B}_i \neq \pm \underline{B}_j \quad \text{for } i \neq j,$$

where $\underline{B}_i'$ *denotes the* i*th row vector of the matrix* B.

Our definition of row vectors is based on the view that a vector $\underline{v}$ is a column vector by default. Hence, we write $\underline{B}_i'$ for the row vectors of a matrix B throughout this paper.

DEFINITION 2.4. *The parameter matrix* $B \in \mathbb{R}^{n\times m}$ *in (1) is strongly admissible if* $\underline{B}_i \neq \alpha \underline{B}_j$ *for* $i \neq j$ *and* $\alpha \in \mathbb{R}$.

The matrices A, B, and C are the parameters of the system. We relegate the theory of estimation of these real-valued parameters (and $\underline{x}_0$) from a set of input-

output data $\{(\underline{u}_t, \underline{y}_t); t = 0, \ldots, T\}$ to a future paper. Here we shall focus on the structural properties of the system only.

We combine the parameters of the system in (1) into $(\Sigma_n, \underline{x}_0)$ with $\Sigma_n = (A, B, C) \in \mathbb{R}^{n\times n+n\times m+p\times n}$. Then we can define an input-output map parametrized by $(\Sigma_n, \underline{x}_0)$ that maps an input sequence $\{\underline{u}_t\}_t$ to an output sequence $\{\underline{y}_t\}_t$, i.e.,

$$\lambda_{\Sigma_n, x_0}: (\mathbb{R}^m)^{\mathbb{N}} \to (\mathbb{R}^p)^{\mathbb{N}}, \quad \text{with } \lambda_{\Sigma_n, x_0}(\{\underline{u}_t\}_t) = \{\underline{y}_t\}_t.$$

DEFINITION 2.5. *Two systems* $(\Sigma_n, \underline{x}_0)$ *and* $(\tilde{\Sigma}_{\tilde{n}}, \tilde{\underline{x}}_0)$ *are called observationally equivalent if* $\lambda_{\Sigma, \underline{x}_0} \equiv \lambda_{\tilde{\Sigma}_{\tilde{n}}, \tilde{\underline{x}}_0}$.

Similar to linear systems we can restrict observationally equivalent systems to those with minimal dimension.

DEFINITION 2.6. *A system* Σ_n *is called minimal if for all* $\underline{x}_0$ *and all systems* $(\tilde{\Sigma}_{\tilde{n}}, \tilde{\underline{x}}_0)$ *that are observationally equivalent to* $(\Sigma_n, \underline{x}_0)$ *it holds that* $\tilde{n} \geq n$.

The particular assumptions on the nonlinear function $\underline{\sigma}$ as in (2) imply that $\underline{\sigma}$ commutes with permutations. Because σ is odd we can extend this invariance to general permutations with sign changes, i.e.,

$$G\underline{\sigma}(\underline{x}) = \underline{\sigma}(G\underline{x}), \tag{4}$$

where $G \in \mathcal{G}_n$, the group generated by permutations and sign changes (which is of size $2^n n!$). This intrinsic symmetry of $\underline{\sigma}$ yields for any system $(A, B, C, \underline{x}_0)$ a number of $2^n n!$ intrinsically observationally equivalent systems of the form

$$\tilde{A} = GAG^{-1}, \qquad \tilde{B} = GB, \qquad \tilde{C} = CG^{-1}, \qquad \text{and} \quad \tilde{\underline{x}}_0 = G\underline{x}_0, \tag{5}$$

where $G \in \mathcal{G}_n$. In any case, we cannot identify a system beyond the intrinsically equivalent systems. This needs to be taken into account in an appropriate definition of identifiability of recurrent neural networks.

DEFINITION 2.7. *A system* $(\Sigma_n, \underline{x}_0)$ *is called identifiable if there are no other observationally equivalent systems but the intrinsically equivalent systems in (5). A subset* $\theta \subset \mathbb{R}^{n\times n+n\times m+p\times n} \times \mathbb{R}^n$ *is called identifiable if every element in* θ *is identifiable.*

3. SOME BASIC RESULTS

In the following discussion we shall introduce some results on the system theoretic properties of recurrent neural networks obtained by Albertini and Dai Pra (1995). Most results and their proofs can be found in Albertini and Dai Pra (1995). However, to keep this paper self-contained we shall restate the proofs briefly. Similar to the case of linear state space systems identifiability of recurrent neural networks is related to minimality (see Definition 2.6) and observ-

ability of the network. Albertini, Dai Pra, and Sontag were the first to realize the consequences of minimality in the structure theory of neural nets. These consequences will become clear in the proofs of Theorems 4.4 and 4.5. Roughly, minimality ensures that the proofs work for general matrices C if the statement can be shown for C with no zero column vectors. In preparation of these proofs we need to elaborate further on the characterization of minimality.

A *coordinate subspace* $V \subseteq \mathbb{R}^n$ is a subspace generated by elements of the canonical basis $\{\underline{e}_1, \ldots, \underline{e}_n\}$. The reason for defining coordinate subspaces lies in the fact that $\forall\ \underline{z}, \hat{\underline{z}} \in \mathbb{R}^n$

$$\underline{z} - \hat{\underline{z}} \in V \Leftrightarrow \forall \underline{u} \in \mathbb{R}^m \qquad \underline{\sigma}(\underline{z} + B\underline{u}) - \underline{\sigma}(\hat{\underline{z}} + B\underline{u}) \in V.$$

DEFINITION 3.1. *Let for a given system $O_C(A,C)$ be the largest coordinate subspace such that $O_C(A,C) \subseteq \ker C$ and $A\ O_C(A,C) \subseteq O_C(A,C)$.*

The next condition for minimality of a system is due to Albertini and Dai Pra (1995).

THEOREM 3.2. *A system Σ is minimal $\Leftrightarrow O_C(A,C) = 0$.*

Proof. It is straightforward to show that $O_C(A,C) \neq 0$ implies that the system is not minimal (see Albertini and Dai Pra, 1995; Dörfler and Deistler, 1998). The other direction follows from Theorem 4.4 in Section 4, i.e., part (ii) of the proof and equation (13). ■

The previous theorem yields an equivalent characterization of $O_C(A,C)$ that we shall use in the proof of the main theorem in the next section. Let I_C be the set of indices i, such that $\underline{c}_i = 0$ ($\underline{c}_i :=$ ith column of C). We begin with the coordinate subspace V spanned by $\underline{c}_i, i \in I_C$ which is obviously contained in $\ker C$. Then we recursively drop indices for which $AV \subseteq V$ does not hold:

$$\begin{aligned} J_0 &:= \{1, \ldots, n\} \backslash I_C\,, \\ J_{d+1} &:= \{i: \quad \exists j \in J_d \text{ such that } A_{ij} \neq 0\} \cup J_0, \\ J &:= \cup_d J_d, \\ O_C(A,C) &= span\{\underline{e}_i: \quad i \notin J\}. \end{aligned} \tag{6}$$

Another aspect in structure theory is the notion of observability, which we shall introduce here for the sake of completeness. In contrast to minimality the characterization of observable systems depends on the properties of the sigmoid function, e.g., the independence property. We shall not need observability in the proofs of Theorems 4.4 and 4.5. In fact, we can use the results on identifiability in Theorem 4.4 to describe observability in Theorem 3.6, which follows.

DEFINITION 3.3. *For a given system Σ_n, two states $\underline{x}_0, \hat{\underline{x}}_0$ are called indistinguishable if $\lambda_{\Sigma_n, \underline{x}_0} = \lambda_{\Sigma_n, \hat{\underline{x}}_0}$ holds.*

DEFINITION 3.4. *The system Σ_n is said to be observable if there are no two different indistinguishable states.*

LEMMA 3.5. *An observable system is minimal.*

Proof. Assume that the system were not minimal, i.e., $O_C(A,C) \neq 0$. Find $\underline{x}_0, \hat{\underline{x}}_0$ such that $\underline{x}_0 - \hat{\underline{x}}_0 \in O_C(A,C)$. By definition of $O_C(A,C)$ $\underline{x}_0$ and $\hat{\underline{x}}_0$ are indistinguishable for the system, which is a contradiction. ■

The following theorem is stated similarly in Albertini and Dai Pra (1995), but we give a shorter proof here that uses Theorem 4.4.

THEOREM 3.6. *An admissible system Σ satisfying g-IP is observable* $\Leftrightarrow$ $(O_C(A,C) = 0$ *and* $\ker A \cap \ker C = 0)$.

Proof. For a system $(\Sigma_n, \underline{x}_0)$ assume $O_C(A,C) = 0$ and $\ker A \cap \ker C = 0$. Consider an observationally equivalent minimal system $(\Sigma_n, \hat{\underline{x}}_0)$. From Theorem 4.4 it is clear that $B = GB$ and $\hat{\underline{x}}_0 = G\underline{x}_0$, which by admissibility of B implies $G = Identity$. Thus $\underline{x}_0 = \hat{\underline{x}}_0$ and the system is observable.

If on the other hand $(\Sigma_n, \underline{x}_0)$ is observable, then by Lemma 3.5 $O_C(A,C) = 0$. It remains to show that $\ker A \cap \ker C = 0$. If $\ker A \cap \ker C \neq 0$ would hold, we could find $\underline{x}_0$, $G\hat{\underline{x}}_0 \in \ker A \cap \ker C$, which by Theorem 4.4 has to fulfill $\underline{x}_0 - G\hat{\underline{x}}_0 = 0$. ■

Remark 3.7. Obviously observability implies minimality. In Theorem 4.4, where observationally equivalent systems are considered, we assume minimality only because the second condition of observability is needed for $\underline{x}_0$ only. This shall become clear in the proof of the following theorem.

4. IDENTIFICATION OF RECURRENT NEURAL NETWORKS

In this section we present the main results of this paper in Theorem 4.4 and Theorem 4.5. The first theorem generalizes previous results on identifiability for a more general class of sigmoids that satisfy the so-called g-IP property. In this general class of sigmoids, g can be any rational function; we identify $g(x) = \text{constant}/x$ as a singular case. The most prominent representative of this singular case is the Heaviside function. In preparation of the theorems we need the following technical results.

LEMMA 4.1.

(i) We assume that the matrix $B \in \mathbb{R}^{n\times m}$ is admissible. Then there exist a $\delta_0 > 0$ and $\underline{u}_0 \in \mathbb{R}^m$ such that for every $\underline{u} \in \mathcal{S} = \{\underline{u}_0 + \delta\, \underline{v};\ \|\underline{v}\| = 1,\ 0 \leq \delta \leq \delta_0\}$ we have

$$\underline{B}_i' \underline{u} \neq \pm \underline{B}_j' \underline{u}, \quad \textit{for all } i \neq j,$$

$$\underline{B}_i' \underline{u} \neq 0, \quad \textit{for all } i,$$

$$\underline{B}_i' \underline{u} \neq d_s, \quad \textit{for all } i \textit{ and } s = 1, \ldots, r, (r \in \mathbb{N}), \tag{7}$$

where $d_1, \ldots, d_r$ are given constants and $\underline{B}_i'$ denotes the ith row vector of B. Note that these properties hold for $\alpha \underline{u}_0$ $(1 - \delta_0 \leq \alpha \leq 1 + \delta_0)$ in particular.

(ii) We assume that the matrix $B \in \mathbb{R}^{n \times m}$ is strongly admissible. Then there exist $\underline{u}_0, \underline{u}_1 \in \mathbb{R}^m$, $\delta_0 > 0$ such that for every $\underline{u} \in \mathcal{S} = \{\underline{u}_0 + \delta \underline{v}; \|\underline{v}\| = 1, 0 \leq \delta \leq \delta_0\}$ we have

$$\underline{B}_i' \underline{u}_1 \neq 0, \quad \textit{for all } i,$$

$$\frac{1}{\underline{B}_i' \underline{u}_1} \underline{B}_i' \underline{u} \neq \pm \frac{1}{\underline{B}_j' \underline{u}_1} \underline{B}_j' \underline{u}, \quad \textit{for all } i \neq j,$$

$$\underline{B}_i' \underline{u} \neq 0, \quad \textit{for all } i. \tag{8}$$

Proof.

(i) If B is admissible then the vectors $\underline{B}_i \pm \underline{B}_j$ and $\underline{B}_i$ are nonzero. Thus, the orthogonal spaces $\{\underline{B}_i + \underline{B}_j\}^\perp$, $\{\underline{B}_i - \underline{B}_j\}^\perp$, and $\{\underline{B}_i\}^\perp$ are hyperplanes through the origin with dimension $m - 1$ (for $i \neq j$). Similarly, the spaces $\{\underline{x}; \underline{B}_i' \underline{x} = d_s\}$ are $m - 1$-dimensional hyperplanes. Because

$$\mathcal{B} = \bigcup_{i \neq j} (\{\underline{B}_i + \underline{B}_j\}^\perp \cup \{\underline{B}_i - \underline{B}_j\}^\perp) \cup \bigcup_i \{\underline{B}_i\}^\perp \cup \bigcup_{i,s} \{\underline{x}; \underline{B}_i' \underline{x} = d_s\}$$

is a finite union of $m - 1$-dimensional hyperplanes we can find a $\underline{u}_0$ in the complement of $\mathcal{B}$ and a sufficiently small m-dimensional sphere of radius δ_0 around $\underline{u}_0$ that remains in the complement of $\mathcal{B}$.

(ii) As in (i) we can choose a $\underline{u}_1$ that lies in the complement of the union of $m - 1$ dimensional hyperplanes $\bigcup_i \{\underline{B}_i\}^\perp$. Thus, $\underline{B}_i' \underline{u}_1 \neq 0$ for all i. Because B is strongly admissible, all vectors $1/(\underline{B}_i' \underline{u}_1)\, \underline{B}_i' \pm 1/(\underline{B}_j' \underline{u}_1)\, \underline{B}_j'$ are nonzero. As before, we can find a set $\mathcal{S}$ in the complement of

$$\mathcal{B} = \bigcup_{i \neq j} (\{1/(\underline{B}_i' \underline{u}_1)\, \underline{B}_i' + 1/(\underline{B}_j' \underline{u}_1)\, \underline{B}_j'\}^\perp \cup \{1/(\underline{B}_i' \underline{u}_1)\, \underline{B}_i' - 1/(\underline{B}_j' \underline{u}_1)\, \underline{B}_j'\}^\perp) \cup \bigcup_i \{\underline{B}_i\}^\perp$$

that yields the statement in (8). ■

LEMMA 4.2.

(i) We assume that $B \in \mathbb{R}^{n \times m}$ is admissible and g is a rational function on $\mathbb{R}$ with $g(x) \neq \text{constant}/x$. Let $d_1, \ldots, d_r$ denote the poles and zeros of g. Then there exist a $\delta_1 > 0$, a proper interval I_1, and $\underline{u}_0 \in \mathbb{R}^m$ such that for every $\underline{u} \in \mathcal{S} = \{\underline{u}_0 + \delta \underline{v}; \|\underline{v}\| = 1, 0 \leq \delta \leq \delta_1\}$ we have

$$g(\underline{B}_i'(\alpha \underline{u}_0))\, \underline{B}_i' \underline{u} \neq \pm g(\underline{B}_j'(\alpha \underline{u}_0))\, \underline{B}_j' \underline{u}, \quad \textit{for all } i \neq j,\ \alpha \in I_1,$$

$$\underline{B}_i' \underline{u}_0 \neq 0, \quad \textit{for all } i,$$

$$\underline{B}_i'(\alpha \underline{u}_0) \neq d_s, \quad \textit{for all } i,\ \alpha \in I_1 \text{ and } s = 1, \ldots, r. \tag{9}$$

(ii) We assume that $B \in \mathbb{R}^{n \times m}$ *is strongly admissible and* $g(x) = \text{constant}/x$. *Then there exist a* $\delta_1 > 0$ *and* $\underline{u}_0, \underline{u}_1 \in \mathbb{R}^m$ *such that for every* $\underline{u} \in \mathcal{S} = \{\underline{u}_0 + \delta\, \underline{v};\ \|\underline{v}\| = 1,\ 0 \leq \delta \leq \delta_1\}$ *we have*

$$\underline{B}_i' \underline{u}_1 \neq 0, \quad \textit{for all } i,$$

$$\underline{B}_i' \underline{u} \neq 0, \quad \textit{for all } i,$$

$$g(\underline{B}_i' \underline{u}_1)\, \underline{B}_i' \underline{u} \neq \pm g(\underline{B}_j' \underline{u}_1)\, \underline{B}_j' \underline{u}, \quad \textit{for all } i \neq j. \tag{10}$$

Proof.

(i) We know that we have a $\underline{u}_0$ and an interval $(1 - \varepsilon, 1 + \varepsilon)$ such that $\alpha\, \underline{u}_0$ satisfies (7) in Lemma 4.1. In particular, $\underline{B}_i'(\alpha \underline{u}_0)$ is not equal to any poles or zeros of g, nonzero, and $\underline{B}_i'(\alpha \underline{u}_0) \neq \pm \underline{B}_j'(\alpha \underline{u}_0)$. Thus, the equality

$$(\underline{B}_i' \underline{u}_0)\, g(\underline{B}_i'(\alpha \underline{u}_0)) = \pm (\underline{B}_j' \underline{u}_0)\, g(\underline{B}_j'(\alpha \underline{u}_0))$$

can hold only for all $\alpha \in (1 - \varepsilon, 1 + \varepsilon)$ if $g(x) = \text{constant}/x$, which has been excluded by assumption. The last equality can hold only for finitely many α, and hence we can find an interval $I_1 \subset [1 - \varepsilon, 1 + \varepsilon]$ such that

$$|(\underline{B}_i' \underline{u}_0)\, g(\underline{B}_i'(\alpha \underline{u}_0)) \pm (\underline{B}_j' \underline{u}_0)\, g(\underline{B}_j'(\alpha \underline{u}_0))| > \text{constant} > 0 \quad \text{for all } i, j,\ \alpha \in I_1.$$

In turn, this means that all hyperplanes, defined by the equations

$$(\mp g(\underline{B}_j'(\alpha \underline{u}_0))\underline{B}_j - g(\underline{B}_i'(\alpha \underline{u}_0))\underline{B}_i)\underline{x} = (\underline{B}_i' \underline{u}_0)\, g(\underline{B}_i'(\alpha \underline{u}_0)) \pm (\underline{B}_j' \underline{u}_0)\, g(\underline{B}_j'(\alpha \underline{u}_0)),$$

are bound away from the origin, uniformly for all $\alpha \in I_1$. Hence, we can choose the radius δ_1 of a sphere around the origin sufficiently small that it never intersects with the preceding hyperplanes. In other words,

$$(\mp g(\underline{B}_j'(\alpha \underline{u}_0))\underline{B}_j - g(\underline{B}_i'(\alpha \underline{u}_0))\underline{B}_i)(\delta_1 \underline{v})$$

$$\neq (\underline{B}_i' \underline{u}_0)\, g(\underline{B}_i'(\alpha \underline{u}_0)) \pm (\underline{B}_j' \underline{u}_0)\, g(\underline{B}_j'(\alpha \underline{u}_0))$$

for all $\underline{v}$ with $\|\underline{v}\| \leq 1$ and all $\alpha \in I_1$.

(ii) is a restatement of Lemma 4.1(ii). ■

LEMMA 4.3. *We assume that* $\sigma(a_1 + b_1 \xi), \ldots, \sigma(a_n + b_n \xi)$ *are* n g-IP *sigmoid functions on* $\mathbb{R}$ $(a_i g(b_i) \neq \pm a_j g(b_j))$ *and* $\underline{\sigma}(\xi) = (\sigma(a_1 + b_1 \xi), \ldots, \sigma(a_n + b_n \xi))'$. *Then we can find* $\xi_1, \ldots, \xi_n \in \mathbb{R}$ *such that* $\underline{\sigma}(\xi_1), \ldots, \underline{\sigma}(\xi_n)$ *are* n *linearly independent vectors.*

Proof. If there were no such $\xi_1, \ldots, \xi_n$ then $\mathcal{C} = \overline{\text{span}\{\underline{\sigma}(\xi); \xi \in \mathbb{R}\}} \neq \mathbb{R}^n$. Thus, we could choose a $\underline{c} \in \mathcal{C}^{\perp}$, $\underline{c} \neq 0$ such that $\underline{c}' \underline{\sigma}(\xi) = 0$ for all ξ, which is a contradiction to the g-IP assumption. ■

THEOREM 4.4. *Let $\Sigma_n = (A, B, C)$ be a minimal system with admissible B that satisfies g-IP (with $g(x) \neq$ constant/x). Let $\tilde{\Sigma}_n = (\tilde{A}, \tilde{B}, \tilde{C})$ be a minimal system. Then $(\Sigma_n, \underline{x}_0)$ and $(\tilde{\Sigma}_n, \underline{\tilde{x}}_0)$ are observationally equivalent if and only if*

$$\tilde{A} = GAG', \qquad \tilde{B} = GB, \qquad \tilde{C} = CG',$$

$$\text{and} \quad \underline{x}_0 - G'\underline{\tilde{x}}_0 \in \ker(C) \cap \ker(A), \tag{11}$$

where $G \in \mathcal{G}_n$ is a permutation with sign changes.

Proof. In what follows we use $\underline{A}_i'$ to denote the ith row vector of A, $\underline{B}_i'$ to denote the ith row vector of B, and $\underline{c}_i$ to denote the ith column vector of C. The analog notation applies to the tilde system. The proof is decomposed into two parts. First, we assume that all column vectors of C are nonzero. In the second part, we shall use minimality to show the theorem for general C.

(i) We assume that all column vectors of C are nonzero.

Observationally equivalence of $\Sigma_n = (A, B, C)$ and $\tilde{\Sigma}_n = (\tilde{A}, \tilde{B}, \tilde{C})$ yields at time t

$$\underline{y}_t = \underline{\tilde{y}}_t \quad \text{for all } \underline{u} \tag{12}$$

or equivalently,

$$\begin{aligned} &\underline{c}_1 \sigma(\underline{A}_1' \underline{x}_{t-1} + \underline{B}_1' \underline{u}) + \cdots + \underline{c}_n \sigma(\underline{A}_n' \underline{x}_{t-1} + \underline{B}_n' \underline{u}) \\ &\quad = \tilde{\underline{c}}_1 \sigma(\underline{\tilde{A}}_1' \underline{\tilde{x}}_{t-1} + \underline{\tilde{B}}_1' \underline{u}) + \cdots + \tilde{\underline{c}}_n \sigma(\underline{\tilde{A}}_n' \underline{\tilde{x}}_{t-1} + \underline{\tilde{B}}_n' \underline{u}) \end{aligned} \tag{13}$$

for all $\underline{u} \in \mathbb{R}^m$. Because B is admissible we have a sphere $\mathcal{S}$ as in Lemma 4.2(i) whose elements satisfy (9). In equation (13) we set $\underline{u} = \underline{z}_0 \eta + \alpha \underline{u}_0 \xi$, where $\alpha \in I_1$ and $\underline{z}_0 \in \mathcal{S}$ (I_1 and $\underline{u}_0$ as in Lemma 4.2(i)). Thus, (13) becomes

$$\begin{aligned} &\underline{c}_1 \sigma(a_1 + b_1 \xi) + \cdots + \underline{c}_n \sigma(a_n + b_n \xi) \\ &\quad = \tilde{\underline{c}}_1 \sigma(\tilde{a}_1 + \tilde{b}_1 \xi) + \cdots + \tilde{\underline{c}}_n \sigma(\tilde{a}_n + \tilde{b}_n \xi), \end{aligned} \tag{14}$$

where

$$\begin{aligned} &a_i = \underline{A}_i' \underline{x}_{t-1} + \underline{B}_i' \underline{z}_0 \eta, \qquad b_i = \alpha \underline{B}_i' \underline{u}_0, \qquad \tilde{a}_i = \underline{\tilde{A}}_i' \underline{\tilde{x}}_{t-1} + \underline{\tilde{B}}_i' \underline{z}_0 \eta, \\ &\tilde{b}_i = \alpha \underline{\tilde{B}}_i' \underline{u}_0, \qquad i = 1, \ldots, n. \end{aligned} \tag{15}$$

Because $\underline{z}_0 \in \mathcal{S}$ we know from (9) that $g(b_i)\, \underline{B}_i' \underline{z}_0 \neq \pm g(b_j)\, \underline{B}_j' \underline{z}_0$, for $i \neq j$, $\alpha \in I_1$, and $\underline{z}_0 \in \mathcal{S}$. In turn, this means that the functions

$$a_i g(b_i) = \underline{A}_i' \underline{x}_{t-1} g(b_i) + \underline{B}_i' \underline{z}_0 g(b_i) \eta, \qquad i = 1, \ldots, n \tag{16}$$

have different slopes in η, and thus we can find a nontrivial interval $I_2 \subset \mathbb{R}$ in which for each $\alpha \in I_1$, $\underline{z}_0 \in \mathcal{S}$

$$a_i g(b_i) \neq (\pm 1) a_j g(b_j) \tag{17}$$

for all $\eta \in I_2$ and $i \neq j$. Hence, by Definition 2.1 of g-IP the functions $\sigma(a_1 + b_1\xi),\ldots,\sigma(a_n + b_n\xi)$ in (14) are linearly independent, and again by Definition 2.1 equality on both sides of (14) can hold only if for every $\sigma(a_i + b_i\xi)$ on the left-hand side of (14) there is one corresponding $\sigma(\tilde{a}_j + \tilde{b}_j\xi)$ on the right-hand side with

$$a_i g(b_i) = (\pm 1)\tilde{a}_j g(\tilde{b}_j) \quad \text{and} \quad \underline{c}_i = \pm\tilde{\underline{c}}_j \tag{18}$$

for $\alpha \in I_1$, $\underline{z}_0 \in \mathcal{S}$, and $\eta \in I_2$. Note that the sign of $\pm\tilde{\underline{c}}_j$ and (± 1) are not correlated. The latter stems from the definition of g-IP in Definition 2.1.

In principle the assignment of j to a given i in the last equation could depend on the parameters $\eta, \alpha, \underline{z}_0$ because $a_i, b_i, \tilde{a}_j, \tilde{b}_j$ depend on them too (see equation (15)). In this case we would have $j = j(\eta, \alpha, \underline{z}_0)$ for a given i on the left-hand side of (18). First, we show that j is independent of η for a given α, $\underline{z}_0$. For a given i we can find a j_0 and a nontrivial sequence of $\eta_n \in I_2$ (as there are only finitely many j) such that (18) holds for this j_0 and all $(\eta_n)_n$. If (18) holds for i, j_0, and a nontrivial sequence $(\eta_n)_n$ then it must also hold for all $\eta \in I_2$ (because the expressions $a_i g(b_i)$, $\tilde{a}_{j_0} g(\tilde{b}_{j_0})$ are first-order polynomials in η). Hence, the j on the right-hand side of (18) is independent of η. By exactly the same reasoning we can show that j is independent of $\alpha \in I_1$ for a given $\underline{z}_0$ (because the expressions $a_i g(b_i)$, $\tilde{a}_{j_0} g(\tilde{b}_{j_0})$ are rational functions in α).

Inserting for the coefficients a_i, b_i, $\tilde{a}_j$, and $\tilde{b}_j$ in (15) we have

$$(\underline{A}_i'\underline{x}_{t-1} + \underline{B}_i'\underline{z}_0\eta)g(\alpha\underline{B}_i'\underline{u}_0) = (\pm 1)(\tilde{\underline{A}}_j'\tilde{\underline{x}}_{t-1} + \tilde{\underline{B}}_j'\underline{z}_0\eta)g(\alpha\tilde{\underline{B}}_j'\underline{u}_0). \tag{19}$$

Separation of the coefficients in η and inserting for $\underline{z}_0$ yields

$$g(\alpha\underline{B}_i'\underline{u}_0)\underline{B}_i'(\underline{u}_0 + \delta\underline{v}) = (\pm 1)g(\alpha\tilde{\underline{B}}_j'\underline{u}_0)\tilde{\underline{B}}_j'(\underline{u}_0 + \delta\underline{v}). \tag{20}$$

At this point we need to remember that the assignment of j to a given i might depend on δ and $\underline{v}$. First, we set $\delta = 0$. Because $\underline{B}_i'\underline{u}_0$ is nonzero and $g(x) \neq$ constant/x we obtain from

$$g(\alpha\underline{B}_i'\underline{u}_0)\underline{B}_i'\underline{u}_0 = (\pm 1)g(\alpha\tilde{\underline{B}}_{j(0)}'\underline{u}_0)\tilde{\underline{B}}_{j(0)}'\underline{u}_0 \tag{21}$$

that $\underline{B}_i'\underline{u}_0 = \pm\tilde{\underline{B}}_{j(0)}'\underline{u}_0$. We fix another $\underline{v} \neq 0$. Because there are only finitely many permutations we can find a sequence $\delta(\underline{v})_n \to 0$ such that the assignment of $j = j(\underline{v})$ remains identical for a particular i, i.e.,

$$g(\alpha\underline{B}_i'\underline{u}_0)\underline{B}_i'(\underline{u}_0 + \delta(\underline{v})_n\underline{v}) = (\pm 1)g(\alpha\tilde{\underline{B}}_{j(\underline{v})}'\underline{u}_0)\tilde{\underline{B}}_{j(\underline{v})}'(\underline{u}_0 + \delta(\underline{v})_n\underline{v}). \tag{22}$$

As before we can conclude that $\underline{B}_i'\underline{u}_0 = \pm\tilde{\underline{B}}_{j(\underline{v})}'\underline{u}_0$. On the other hand, we know that $\underline{B}_i'\underline{u}_0 \neq \pm\underline{B}_k'\underline{u}_0$ for $i \neq k$ (see Lemma 4.2), and thus, $j(\underline{v}) = j(0)$ (otherwise we would assign $\underline{B}_i'\underline{u}_0 = \pm\tilde{\underline{B}}_{j(0)}'\underline{u}_0 = \pm\underline{B}_k'\underline{u}_0$, $i \neq k$). With (22) we can also conclude that $\delta(\underline{v})_n = \delta_n$. In summary, we have that the j in equation (20) is independent of $\underline{v}$ for a sequence δ_n and that $\underline{B}_i'\underline{u}_0 = \pm\tilde{\underline{B}}_j'\underline{u}_0$. Because the vector $\underline{v}$ is any vector with $\|\underline{v}\| = 1$ equation (20) yields

$$\underline{B}_i = \pm \underline{\tilde{B}}_j. \tag{23}$$

Inserting this into (19) yields

$$\underline{A}'_i \underline{x}_{t-1} = \pm \underline{\tilde{A}}'_j \underline{\tilde{x}}_{t-1}. \tag{24}$$

With (18) we can summarize the previous equations as

$$B = G'\tilde{B}, \qquad A\underline{\tilde{x}}_{t-1} = G'\tilde{A}\underline{\tilde{x}}_{t-1}, \qquad C = \tilde{C}G. \tag{25}$$

Inserting the preceding expression into $\underline{x}_t = \underline{\sigma}(A\underline{\tilde{x}}_{t-1} + B\underline{u}_t)$ we also obtain

$$\underline{x}_t = G'\underline{\tilde{x}}_t. \tag{26}$$

Note that $\underline{c}_i = \underline{\tilde{c}}_j$ yields $C = \tilde{C}G$ because we permute the column vectors of C, whereas a permutation of row vectors $\underline{B}'_i = \underline{\tilde{B}}'_j$ yields $B = G'\tilde{B}$. The matrix G in (25) and (26) is independent of t. Otherwise, we contradict the assumption that B is admissible.

It remains to show (25) for a set of linearly independent vectors $\underline{x}_t$. We recall that we can choose a control $\underline{u} = \underline{z}_0\, \eta + \alpha\, \underline{u}_0\, \xi$ such that the resulting sigmoids $\sigma(a_1 + b_1\xi), \ldots, \sigma(a_n + b_n\xi)$ in (14) are linearly independent. Note, that $\underline{x}_t$ in (26) still depends on ξ. We can now use Lemma 4.3 to obtain a set of linear independent $\{\underline{x}_t(\xi_1), \ldots, \underline{x}_t(\xi_n)\}$ and let the following $\underline{x}_t$ be any vector from this set. Then we can repeat exactly the same argument as before beginning with

$$\underline{y}_{t+1} = \underline{\tilde{y}}_{t+1} \quad \text{for all } \underline{u}.$$

By the same reasoning as before we arrive at equation (25) with $\underline{x}_{t-1}$ replaced by $\underline{x}_t$; i.e., $A\underline{x}_t = G'\tilde{A}\underline{\tilde{x}}_t = G'\tilde{A}G\underline{x}_t$. Note that the other quantities are the same as before. However, this time we can choose $\underline{x}_t$ from a set of linearly independent vectors, and hence we obtain from (25)

$$A = G'\tilde{A}G. \tag{27}$$

With equations (25) and (27) we have proved the theorem in the case where all column vectors of C are nonzero.

(ii) In the second part of this proof we use a method similar to Albertini and Dai Pra's work (1995), i.e., use minimality to show the statement of the first part for general C. We can now turn to the general case, where we assume without loss of generality that $\underline{c}_{k+1} = \cdots = \underline{c}_n = 0$ and the remaining column vectors of C are nonzero. Furthermore, we assume without loss of generality that the set of indices J in (6) is given by

$$J_0 = \{1, \ldots, k\}, \qquad J_d = \{s_{d-1} + 1, \ldots, s_d\}, \qquad \text{and} \quad J = \cup_{d=1}^{r} J_d. \tag{28}$$

Otherwise, we can permute the columns of A such that the last equation holds. We recall the significance of the sets $J_d = \{s_{d-1} + 1, \ldots, s_d\}$ and J_{d+1}. For each $i = s_d + 1, \ldots, s_{d+1}$ (i.e., $i \in J_{d+1}$) there is a row vector in $\underline{A}'_{s_{d-1}+1}, \ldots, \underline{A}'_{s_d}$ whose ith component is nonzero.

As before in (12), observationally equivalence implies that at time t

$$\underline{y}_t = \tilde{\underline{y}}_t \quad \text{for all } \underline{u}, \tag{29}$$

or equivalently,

$$\begin{aligned} &\underline{c}_1 \sigma(\underline{A}_1' \underline{x}_{t-1} + \underline{B}_1' \underline{u}) + \cdots + \underline{c}_k \sigma(\underline{A}_k' \underline{x}_{t-1} + \underline{B}_k' \underline{u}) \\ &\quad = \tilde{\underline{c}}_1 \sigma(\tilde{\underline{A}}_1' \tilde{\underline{x}}_{t-1} + \tilde{\underline{B}}_1' \underline{u}) + \cdots + \tilde{\underline{c}}_n \sigma(\tilde{\underline{A}}_n' \tilde{\underline{x}}_{t-1} + \tilde{\underline{B}}_n' \underline{u}) \end{aligned} \tag{30}$$

for all $\underline{u} \in \mathbb{R}^m$.

For an appropriate choice of $\underline{u}$ ($\underline{u} = \underline{u}(\xi_1) = \underline{z}_0 \eta + \alpha \underline{u}_1 \xi_1$ as in (14) and (15)) we can ensure linearly independence of the sigmoid functions on the left-hand side. Analogously to (18), (22), and (23) we can use independence and Lemma 4.2(i) to arrive at

$$\begin{aligned} \underline{B}_1 &= \pm \tilde{\underline{B}}_{j_1}, \ldots, \underline{B}_k = \pm \tilde{\underline{B}}_{j_k}, \\ \underline{A}_1 \underline{x}_{t-1} &= \pm \tilde{\underline{A}}_{j_1}' \tilde{\underline{x}}_{t-1}, \ldots, \underline{A}_k \underline{x}_{t-1} = \pm \tilde{\underline{A}}_{j_k}' \tilde{\underline{x}}_{t-1}, \\ \underline{c}_1 &= \pm \tilde{\underline{c}}_{j_1}, \ldots, \underline{c}_k = \pm \tilde{\underline{c}}_{j_k}, \qquad \text{and} \\ \sigma(a_1 + b_1 \xi_1) &= \pm \sigma(\tilde{a}_{j_1} + \tilde{b}_{j_1} \xi_1), \ldots, \sigma(a_k + b_k \xi_1) = \pm \sigma(\tilde{a}_{j_k} + \tilde{b}_{j_k} \xi_1), \end{aligned} \tag{31}$$

the latter being equivalent with $x_{t,1} = \pm \tilde{x}_{t,j_1}, \ldots, x_{t,k} = \pm \tilde{x}_{t,j_k}$ ($x_{t,i}$ is the ith entry of $\underline{x}_t$). The remaining column vectors of $\tilde{C}$ must be zero. At this point two or more σ on the right-hand side of (30) could equal one σ on the left-hand side. However, in this case we would not have enough σ to match all left-hand side σ in the following.

To keep the following discussion simple we define new vectors $\hat{\underline{x}}_{t-1} = S \tilde{\underline{x}}_{t-1}$, $\hat{\underline{x}}_t = S \tilde{\underline{x}}_t$, where $S \in \mathcal{G}_n$ is a permutation with sign changes such that

$$\hat{x}_{t,1} = \tilde{x}_{t,j_1}, \ldots, \hat{x}_{t,k} = \tilde{x}_{t,j_k}. \tag{32}$$

Additionally, we define $\hat{A} = S \tilde{A} S'$, $\hat{B} = S \tilde{B}$, and $\hat{C} = \tilde{C} S'$. Then (31) can be rewritten as

$$\underline{B}_i = \hat{\underline{B}}_i, \qquad \underline{A}_i' \underline{x}_{t-1} = \hat{\underline{A}}_i' \hat{\underline{x}}_{t-1}, \qquad \sigma(a_i + b_i \xi_1) = \sigma(\hat{a}_i + \hat{b}_i \xi_1), \qquad x_{t,i} = \hat{x}_{t,i} \tag{33}$$

for $i = 1, \ldots, k$, where $\hat{a}_i$ and $\hat{b}_i$ are defined as in (15) through $\hat{A}$ and $\hat{B}$. Note that in general, for $k < n$, we have several $S \in \mathcal{G}_n$ such that (32) and (33) hold. The full vectors $\underline{x}_t$ and $\hat{\underline{x}}_t$ are of the form

$$\begin{aligned} \underline{x}_t &= (\sigma(a_1 + b_1 \xi_1), \ldots, \sigma(a_k + b_k \xi_1), \sigma(a_{k+1} + b_{k+1} \xi_1), \ldots, \sigma(a_n + b_n \xi_1))', \\ \hat{\underline{x}}_t &= (\sigma(a_1 + b_1 \xi_1), \ldots, \sigma(a_k + b_k \xi_1), \sigma(\hat{a}_{k+1} + \hat{b}_{k+1} \xi_1), \ldots, \sigma(\hat{a}_n + \hat{b}_n \xi_1))'. \end{aligned} \tag{34}$$

Note that the first k entries of $\underline{x}$ and $\hat{\underline{x}}$ are equal according to (33).

We use minimality to identify the remaining sigmoids recursively, as sketched in Figure 1. We begin with the set of indices $J_1 = \{k+1,\dots,s_1\}$. Note that the definition of J_1 implies that for each $i = k+1,\dots,s_1$ at least one of the vectors $\underline{A}_1,\dots,\underline{A}_k$ has a nonzero ith component.

We shall show that $x_{t,i} = \hat{x}_{t,i}$, $\underline{B}_i = \underline{\hat{B}}_i$, and $\underline{A}_i'\underline{x}_{t-1} = \underline{\hat{A}}_i'\underline{\hat{x}}_{t-1}$ for $i = k+1,\dots,s_1$. Note that in the previous step we have begun at an arbitrary time t with two arbitrary $\underline{x}_{t-1}, \underline{\hat{x}}_{t-1}$ and the equation of observationally equivalence $\underline{y}_t = \underline{\hat{y}}_t$ has brought us to equation (33). In particular we have obtained $\underline{A}_i'\underline{x}_{t-1} = \underline{\hat{A}}_i'\underline{\hat{x}}_{t-1}$ for $i = 1,\dots,k$. If we increment the time by one, we begin from

$$\underline{y}_{t+1} = \underline{\hat{y}}_{t+1} \quad \text{for all } \underline{u} \tag{35}$$

and arrive by the same reasoning at

$$\underline{A}_i'\underline{x}_t = \underline{\hat{A}}_i'\underline{\hat{x}}_t \quad \text{for } i = 1,\dots,k. \tag{36}$$

In the last equation we can insert (34) for $\underline{x}_t, \underline{\hat{x}}_t$

$$\begin{aligned}\underline{A}_i'&(\sigma(a_1 + b_1\xi_1),\dots,\sigma(a_k + b_k\xi_1),\sigma(a_{k+1} + b_{k+1}\xi_1),\dots,\sigma(a_n + b_n\xi_1))' \\ &= \underline{\hat{A}}_i'(\sigma(a_1 + b_1\xi_1),\dots,\sigma(a_k + b_k\xi_1), \\ &\qquad \sigma(\hat{a}_{k+1} + \hat{b}_{k+1}\xi_1),\dots,\sigma(\hat{a}_n + \hat{b}_n\xi_1))',\end{aligned} \tag{37}$$

with $i = 1,\dots,k$. We can use these equations to identify $\sigma(a_{k+1} + b_{k+1}\xi_1),\dots,\sigma(a_{s_1} + b_{s_1}\xi_1)$ on the left-hand side with some corresponding sigmoids on the right-hand side of (37). This works because we know from the definition of J_1 that for each $i = k+1,\dots,s_1$ there is a row vector $\underline{A}_j'$ in (37) that has a nonzero component. In other words each $\sigma(a_{k+1} + b_{k+1}\xi_1),\dots,\sigma(a_{k+s_1} + b_{k+s_1}\xi_1)$ occurs in at least one of the equations in (37). Wherever such a sigmoid occurs we can use g-IP to identify it with a corre-

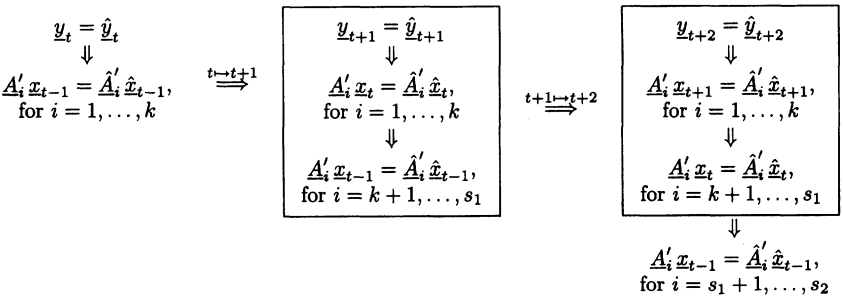

FIGURE 1. Sketch of the proof for the first three time steps. The first conclusions in the next time step $t+2$ are obtained from the current time $t+1$ shifted by one (framed boxes). Only the last conclusion $\underline{A}_i'\underline{x}_{t-1} = \underline{\hat{A}}_i'\underline{\hat{x}}_{t-1}$ is newly derived in each step.

sponding sigmoid on the right-hand side of (37). As before we can then redefine $S \in \mathcal{G}_n$ to obtain similarly to (33)

$$\underline{B}_i = \hat{\underline{B}}_i, \qquad \underline{A}_i' \underline{x}_{t-1} = \hat{\underline{A}}_i' \hat{\underline{x}}_{t-1}, \qquad \sigma(a_i + b_i \xi_1) = \sigma(\hat{a}_i + \hat{b}_i \xi_1), \qquad x_{t,i} = \hat{x}_{t,i} \tag{38}$$

for $i = k+1, \ldots, s_1$, and

$$A_{i,j} = \hat{A}_{i,j} \quad \text{for } i,j = 1, \ldots, k, \tag{39}$$

where $A_{i,j}$ is the jth entry of the vector $\underline{A}_i$. We can summarize as follows: the equation

$$\underline{y}_{t+1} = \hat{\underline{y}}_{t+1} \tag{40}$$

yields

$$\underline{B}_i = \hat{\underline{B}}_i, \qquad \underline{A}_i' \underline{x}_{t-1} = \hat{\underline{A}}_i' \hat{\underline{x}}_{t-1}, \qquad \sigma(a_i + b_i \xi_1) = \sigma(\hat{a}_i + \hat{b}_i \xi_1), \qquad x_{t,i} = \hat{x}_{t,i} \tag{41}$$

for $i = 1, \ldots, s_1$ and in particular

$$\underline{A}_i' \underline{x}_{t-1} = \hat{\underline{A}}_i' \hat{\underline{x}}_{t-1} \quad \text{for } i = k+1, \ldots, s_1. \tag{42}$$

As before the choice of t and $\underline{x}_{t-1}, \hat{\underline{x}}_{t-1}$ is arbitrary, and hence we can increment the time by one to generate from a set of equations (see also Figure 1)

$$\underline{y}_{t+2} = \hat{\underline{y}}_{t+2} \tag{43}$$

the set of equations

$$\underline{A}_i' \underline{x}_t = \hat{\underline{A}}_i' \hat{\underline{x}}_t \quad \text{for } i = k+1, \ldots, s_1. \tag{44}$$

Here we are in the same situation as in (36) except that the index i runs further to s_1. We can now consider the set $J_2 = \{s_1 + 1, \ldots, s_2\}$ and, by the definition of J_2, we can use the last equation to identify the sigmoids $\sigma(a_{s_1+1} + b_{s_1+1}\xi_1), \ldots, \sigma(a_{s_2} + b_{s_2}\xi_1)$ in $\underline{x}_t$.

We can repeat the whole procedure until, after a finite number of steps, we arrive at (41) for all $i = 1, \ldots, n$. Because $\hat{\underline{x}}_{t-1} = \hat{\underline{x}}_{t-1} = S\tilde{\underline{x}}_{t-1}$, $A = \hat{A} = S\tilde{A}S'$, $B = \hat{B} = S\tilde{B}$, and $C = \hat{C} = \tilde{C}S'$ for an $S \in \mathcal{G}_n$ we obtain the statement of the theorem with $G = S'$. Formally, this part needs to be proved by induction (see also Albertini and Dai Pra, 1995). However, we think that the induction step follows evidently from Figure 1 and the preceding discussion. Apart from more indices, a formal induction would not generate further insight.

(iii) In the last part of the proof we need to relate $\underline{x}_0$ and $\tilde{\underline{x}}_0$. Observationally equality in $t = 0$, i.e.,

$$\underline{y}_0 = C\underline{x}_0 = \tilde{\underline{y}}_0 = CG'\underline{x}_0 \tag{45}$$

yields $\underline{x}_0 - G'\tilde{\underline{x}}_0 \in \ker(C)$. Furthermore, we have the equation $A\underline{x}_0 = G'\tilde{A}\tilde{\underline{x}}_0$ from part (i) or (ii) (see (25) or (41) for $i = 1,\ldots,n$). Inserting for $\tilde{A}$ yields $\underline{x}_0 - G'\tilde{\underline{x}}_0 \in \ker(A)$. ■

THEOREM 4.5. *We assume that we have a strongly admissible, g-*IP *(with $g(x) = \text{constant}/x$) minimal system that is parametrized by $\Sigma_n = (A,B,C)$ and a minimal system $\tilde{\Sigma}_n = (\tilde{A},\tilde{B},\tilde{C})$. If (Σ_n, x_0) and $(\tilde{\Sigma}_n, \tilde{x}_0)$ are observationally equivalent then*

$$\tilde{A} = G\Lambda AG', \qquad \tilde{B} = G\Lambda B, \qquad \tilde{C} = CG',$$

$$\text{and} \quad \underline{x}_0 - G'\tilde{\underline{x}}_0 \in \ker(C) \cap \ker(A), \tag{46}$$

where $G \in \mathcal{G}_n$ is a permutation with sign changes and Λ is a diagonal matrix with strictly positive diagonal elements.

Proof. A careful investigation of the previous proof of Theorem 4.4 reveals that all arguments remain the same except the step from equation (20) to (23). At this point the previous $G\tilde{B}$ turns into $G\Lambda\tilde{B}$ and the equation $A\underline{x}_t = G\tilde{A}\tilde{\underline{x}}_t$ into $A\underline{x}_t = G\Lambda\tilde{A}\tilde{\underline{x}}_t$. The $\underline{\tilde{x}}_t$ and $\tilde{C}$ do not change. Noting that $\ker(\Lambda A) = \ker(A)$ we obtain the statement of the theorem. In the following discussion we shall present the steps that deviate from the proof of Theorem 4.4. All other conclusions can be copied literally from the proof of Theorem 4.4.

Similarly to part (i) in the proof of Theorem 4.4 we insert $\underline{u} = \underline{z}_0\,\eta + \underline{u}_1\,\xi$, where $\underline{z}_0 \in \mathcal{S}$ ($\underline{u}_1$ and $\mathcal{S}$ as in Lemma 4.2(ii)). We arrive at equation (14) with

$$a_i = \underline{A}_i'\underline{x}_{t-1} + \underline{B}_i'\underline{z}_0\eta, \qquad b_i = \underline{B}_i'\underline{u}_1, \qquad \tilde{a}_i = \tilde{\underline{A}}_i'\tilde{\underline{x}}_{t-1} + \tilde{\underline{B}}_i'\underline{z}_0\eta,$$

$$\tilde{b}_i = \tilde{\underline{B}}_i'\underline{u}_1, \qquad i = 1,\ldots,n. \tag{47}$$

Again, by Lemma 4.2(ii) the functions $a_i g(b_i) = \tilde{\underline{A}}_i'\tilde{\underline{x}}_{t-1} g(b_i) + \underline{B}_i'\underline{z}_0 g(b_i)\eta$ have different slopes in η, and thus we can find a nontrivial interval $I_2 \subset \mathbb{R}$ in which $a_i g(b_i) \neq (\pm 1) a_j g(b_j)$ for $i \neq j$. As before, we can separate the functions on the left-hand side in (14), and g-IP yields

$$(\underline{A}_i'\underline{x}_{t-1} + \underline{B}_i'\underline{z}_0\eta)g(\underline{B}_i'\underline{u}_1) = (\pm 1)(\tilde{\underline{A}}_j'\tilde{\underline{x}}_{t-1} + \tilde{\underline{B}}_j'\underline{z}_0\eta)g(\tilde{\underline{B}}_j'\underline{u}_1)$$

$$\text{and} \quad \underline{c}_i = \pm\tilde{\underline{c}}_j. \tag{48}$$

Note that the sign of (± 1) (resulting from the definition of g-IP) and $\pm\tilde{\underline{c}}_j$ are uncorrelated. First, we separate coefficients in η. Then it can be shown (exactly as in part (i) of the proof of Theorem 4.4) that j is independent of η, $\underline{v}$, and δ. With $g(x) = \text{constant}/x$ we obtain

$$\underline{B}_i/(\underline{B}_i'\underline{u}_1)) = (\pm 1)\tilde{\underline{B}}_j/(\tilde{\underline{B}}_j'\underline{u}_1), \tag{49}$$

which is satisfied for $\underline{B}_i = \pm\lambda\underline{\tilde{B}}_j$ for any $\lambda > 0$. Inserting this expression in (48) yields

$$B = G'\Lambda\tilde{B}, \qquad A'\underline{x}_{t-1} = G'\Lambda\tilde{A}'\underline{\tilde{x}}_{t-1}, \qquad C = \tilde{C}G, \tag{50}$$

where G is a permutation with sign changes and Λ a diagonal matrix with positive diagonal entries. From here onward the arguments are identical with the proof of Theorem 4.4. ■

The last theorem and Remark 2.2 yield the following corollary.

COROLLARY 4.6. *We assume that the sigmoid function in the Elman network is the Heaviside function. Furthermore, let* $\Sigma_n = (A, B, C)$ *be a strongly admissible, minimal system and* $\tilde{\Sigma}_n = (\tilde{A}, \tilde{B}, \tilde{C})$ *a minimal system. Then* (Σ_n, x_0) *and* $(\tilde{\Sigma}_n, \tilde{x}_0)$ *are observationally equivalent if and only if*

$$\tilde{A} = G\Lambda AG', \qquad \tilde{B} = G\Lambda B, \qquad \tilde{C} = CG',$$

$$\textit{and} \quad \underline{x}_0 - G'\underline{\tilde{x}}_0 \in \ker(C) \cap \ker(A), \tag{51}$$

where $G \in \mathcal{G}_n$ *is a permutation with sign changes and* Λ *is a diagonal matrix with strictly positive diagonal elements.*

5. DISCUSSION

We have shown that previous results on identifiability can be extended to a general class of Elman neural nets satisfying the g-IP property. Second, within these neural nets we have singled out the case $g = \text{constant}/x$ with different properties of identification. In this respect, Heaviside-like functions play an exceptional role in the identification of neural nets. Intuitively we could have anticipated this result qualitatively because the Heaviside function maps infinitely many inputs to a finite set of output states.

We think that Theorem 4.4 could have been extended to an even broader class of neural nets if we allowed for more general g in the g-IP. For instance, we could have considered g where $xg(x)$ is continuous in 0 (to obtain a statement analogous to Lemma 4.2). However, to keep the proof of Theorem 4.4 simple we have restricted our analysis to rational g. Furthermore, our results can be extended to mixed neural networks and the time continuous case (see also Albertini and Sontag, 1993).

REFERENCES

Albertini, F. (1993) *Controllability of Discrete-Time Nonlinear Systems and Some Related Topics.* PhD thesis, New Brunswick, New Jersey.

Albertini, F. & P. Dai Pra (1995) Recurrent neural networks: Identification and other system theoretic properties. *Neural Network Systems Techniques and Applications* 3, 1–41.

Albertini, F. & E. Sontag (1993) For neural networks, function determines form. *Neural Networks* 6, 975–990.

Brockwell, P.J. & R.A. Davis (1991) *Time Series: Theory and Methods*. Berlin: Springer.

Dörfler, M. & M. Deistler (1998) A structure theory for identification of recurrent neural networks, part 1. In H.J.C.Huijberts, H.Nijmeijer, A.J. Van der Schaft, & J.M.A. Scherpen (eds.), *Proceedings of the 4th IFAC Nonlinear Control Systems Design Symposium*, pp. 459–464. Amsterdam: Elsevier.

Draisma, G. & P.H. Franses (1997) Recognizing changing seasonal patterns using artificial neural networks. *Journal of Econometrics* 81, 273–280.

Elman, J.L. (1989) Structured representations and connectionist models. In G. Olson & E. Smith (eds.), *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, pp. 17–25. Hillsdale, NJ: Erlbaum.

Elman, J.L. (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7, 195–226.

Gencay, R. & R. Garcia (2000) Pricing and hedging derivative securities with neural networks and a homogeneity hint. *Journal of Econometrics* 94, 93–115.

Granger, C.W.J., T.-H. Lee, & H. White (1993) Testing for neglected nonlinearity in time series models. *Journal of Econometrics* 56, 269–290.

Hannan, H.J. & M. Deistler (1988) *The Statistical Theory of Linear Systems*. New York: Wiley.

Leisch, F., A. Trapletti, & K. Hornik (1999) Stationarity and stability of autoregressive neural network processes. In M.S. Kearns, S.A. Solla, & D.A. Cohn (eds.), *Advances in Neural Information Processing Systems*, vol. 11, pp. 267–273. Cambridge, MA: MIT Press.

Richards, C.E. & B.D. Baker (1999) A comparison of conventional linear regression methods and neural networks for forecasting educational spending. *Economics of Education Review* 18, 405–415.

Tkacz, G. (2001) Neural network forecasting of Canadian gdp growth. *International Journal of Forecasting* 17, 57–69.

Trapletti, A., F. Leisch, & K. Hornik (1998) Stationarity and Integrated Autoregressive Neural Network Processes. Working paper, Sonderforschungsbereich 10.

Trapletti, A., F. Leisch, & K. Hornik (1999) On the Ergodicity and Stationarity of the ARMA(1,1) Recurrent Neural Network Process. Working paper 24, Sonderforschungsbereich 10.