

# A note on the identifiability of the conditional expectation for the mixtures of neural networks

Jürgen Franke\*  
Jean-Pierre Stockis  
Joseph Tadjuidje Kamgaing

*University of Kaiserslautern*

January 10, 2007

## Abstract

We consider a generalized mixture of nonlinear AR models, a hidden Markov model for which the autoregressive functions are single layer feedforward neural networks. The non trivial problem of identifiability, which is usually postulated for hidden Markov models, is addressed here.

**Keywords:** Mixture Models; Neural networks; Identifiability.

---

\*University of Kaiserslautern, Department of Mathematics, Erwin-Schroedinger-Str., 67663 Kaiserslautern, Germany. E-mail address: franke@mathematik.uni-kl.de

The work was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the priority research program 1114 *Mathematical Methods of Time Series and Digital Image Analysis*, the center of excellence *Dependable Adaptive Systems and Mathematical Modeling* funded by the state of Rhineland-Palatinate as well as the "Graduiertenkolleg Mathematik und Praxis" and by the Fraunhofer ITWM.

# 1 Introduction

Neural networks provide a flexible tool for learning an unknown mapping from input space to output space in the presence of noise. Neural networks have been proven to have nice asymptotic properties under some general assumptions (see, e.g, White (1984) or Franke et al. (2001) and some of the references therein).

We assume in this paper that our time series data obey the following hidden Markov model

$$X_t = \sum_{k=1}^K S_{tk} (m_k(X_{t-1}, \dots, X_{t-p}) + \epsilon_t) \quad (1.1)$$

with

$$S_{tk} = \begin{cases} 1 & \text{for } Q_t = k \\ 0 & \text{otherwise} \end{cases} \quad (1.2)$$

$\{Q_t\}$  is a hidden Markov chain with values in  $\{1, \dots, K\}$ ,  $K \in \mathbb{N}$ , and with transition probability matrix  $A$ .  $m_k$ ,  $k = 1, \dots, K$ , are unknown real valued functions. The random errors  $\epsilon_t$  are i.i.d. with mean zero and support  $\mathbb{R}$ .

The process  $\{X_t\}$  passes through different dynamics depending on  $\{Q_t\}$ . This type of mixture of nonparametric autoregressive models plays in practice an important role, for example in econometrics (see, e.g. Franke et al. (2004), Gouriéroux (1997)) where it is often not realistic to assume that the process is driven by only one trend function throughout time.

Now, a natural way of estimating this type of mixture models is to replace in equation (1.1) the unknown  $m_k(x)$  by

$$m_k(x) = \nu_{0k} + \sum_{h=1}^H \nu_{hk} \psi(\langle \alpha_{hk}, x \rangle + \beta_{hk}), \quad k = 1, \dots, K \quad (1.3)$$

where  $m_k(x)$  is a one layer feedforward neural network with  $H$  hidden neurons,  $\langle, \rangle$  is the classical scalar product on  $\mathbb{R}^p$  and  $\psi$  the activation functions, taken here, for sake of simplicity to be the logistic function.

Compared to a single neural network this framework has the advantage of mimicking well the structure of the data and also it allows for the detection of the change points in the dynamics of the models.

Considering the asymptotic behavior of the parameter estimates, it is well known that an important theoretical problem with neural networks is the unidentifiability of the parameters (for a definition see, e.g. Hwang and Ding (1997)). If identifiability is not satisfied, asymptotic properties like consistency and asymptotic normality of the parameter estimates are not guaranteed any more and resampling techniques like bootstrap will not work properly (see, Franke and Neumann (2000)). In this paper we are providing sufficient conditions for identifiability (up to the family of transformation generated by (5) in Hwang and Ding (1997)); in the remainder of this paper for conciseness we shall simply write identifiability.

To achieve this, we will mainly generalize works for simple neural networks of Hwang and Ding (1997, p. 749) and Farinas et al. (2004).

## 2 Results

Let  $S_t = (S_{t1}, \dots, S_{tK})'$  and assume

$$\mathbb{P}(S_t | S_{t-1}, X_{t-1}, X_{t-2}, \dots) = \mathbb{P}(S_t | S_{t-1}). \quad (2.1)$$

Now we are focusing on the identifiability of

$$\mathbb{E}(X_t | S_{t-1}, X_{t-1}, \dots, X_{t-p}). \quad (2.2)$$

While it is clear that the expectation needs to be conditioned on the  $X_{t-1}, \dots, X_{t-p}$ , at first sight it may be surprising to condition also on  $S_{t-1}$ .

Notice first that without  $S_{t-1}$ , terms of the form  $\mathbb{P}(S_t | S_{t-1}, X_{t-1}, X_{t-2}, \dots)$  would appear in the conditional expectation, terms which are typically not tractable in practice. Moreover, we condition on  $S_{t-1}$  and not on  $S_t$  and although we do not know  $S_{t-1}$  either, it is possible to get a good estimation given  $X_{t-1}, X_{t-2}, \dots$ . Finally, from a theoretical point of view, we can consider  $S_{t-1}$  as an exogenous variable as in Farinas et al (2004).

Let us now focus on the conditional expectation,

$$\mathbb{E}(X_t | S_{t-1}, X_{t-1}, \dots, X_{t-p}) = \sum_{k=1}^K m_k(X_{t-1}, \dots, X_{t-p}) \mathbb{P}(S_{t,k} = 1 | S_{t-1}); \quad (2.3)$$

if  $S_{t-1} = (1, 0, \dots, 0)'$  then,

$$\mathbb{E}(X_t | S_{t-1}, X_{t-1}, \dots, X_{t-p}) = \sum_{k=1}^K a_{1k} m_k(X_{t-1}, \dots, X_{t-p}).$$

Similarly, if  $S_{t-1}$  is the  $j$ th unit vector,

$$\mathbb{E}(X_t | S_{t-1}, X_{t-1}, \dots, X_{t-p}) = \sum_{k=1}^K a_{jk} m_k(X_{t-1}, \dots, X_{t-p})$$

and taking  $m = (m_1(X_{t-1}, \dots, X_{t-p}), \dots, m_K(X_{t-1}, \dots, X_{t-p}))'$  one can hence write

$$\mathbb{E}(X_t | S_{t-1}, X_{t-1}, \dots, X_{t-p}) = \langle S_{t-1}, Am \rangle \quad (2.4)$$

where  $\langle, \rangle$  is the classical scalar product on  $\mathbb{R}^K$  and recall that  $A$  is the transition probability matrix of the hidden process.

To achieve the announced results we shall need the following assumptions

**A. 2.1**  $\forall k, m_k$  is not redundant (i.e. there exist no other networks with fewer hidden neurons that represent exactly the same relationship function)

**A. 2.2** The transition probability matrix  $A$  of the hidden process  $\{Q_t\}$  is invertible.

**A. 2.3**  $(\alpha_{hk}, \beta_{hk}) \neq (\alpha_{uj}, \beta_{uj}) \forall h \neq u, h, u \in \{1, \dots, H\}, k, j \in \{1, \dots, K\}$

**Proposition 1** Under assumptions **A.2.1**, **A.2.2** and **A.2.3** the identifiability of the parameters of our model is achieved.

**Proof:** In the following proof, we will focus on the case  $m : \mathbb{R} \rightarrow \mathbb{R}$ . The multidimensional case can easily be reduced to the one dimensional case in a similar way as in the paper by Hwang and Ding (1997).

The conditions of Theorem 2.3(a) of Hwang and Ding (1997) and thus the identifiability of the parameters are satisfied if

$$\mathbb{E}(X_t | X_{t-1} = x, S_{t-1}) = 0, \forall x \in \mathbb{R}, \forall S_{t-1} \Rightarrow \nu_{0k} = \nu_{hk} = 0, \forall h, k$$

To show this implication, it suffices to show that

$$A(m_1(x), \dots, m_K(x))' = 0, \forall x \in \mathbb{R} \Rightarrow \nu_{0k} = \nu_{hk} = 0, \forall h, k$$

Now consider assumption **A.2.2** and

$$A(m_1(x), \dots, m_K(x))' = 0, \forall x \in \mathbb{R}$$

it follows that

$$\nu_{0k} + \sum_{h=1}^H \nu_{hk} \psi(\alpha_{hk}x + b_{hk}) = 0, \forall k \in \{1, \dots, K\}, \forall x \in \mathbb{R}$$

and hence  $\nu_{0k} = \nu_{hk} = 0, \forall h, k$  (see Lemma 2.7 of Hwang and Ding (1997)). ■

### 3 Conclusion

In conclusion, we can notice that the only condition on our driving hidden Markov chain  $\{Q_t\}$  for ensuring the identifiability of the parameters of our mixture model is a weak condition, namely the invertibility of the transition probability matrix. Furthermore, this result can obviously be extended to a large variety of parametric models with switching regimes driven by a hidden Markov process.

The identifiability of the parameters of our model is an important step toward the proof of asymptotic results for the estimators, e.g. their consistency. Asymptotic inference and practical results based on our model are beyond the scope of this paper and will be analyzed in further research.

## References

- [1] Fariñas, M.S., Pedreira, C.E., Medeiros, M.C. (2004), Local Global Neural Networks: A New Approach for Nonlinear time Series modeling, *J. Amer. Statist. Assoc.* **99**, 1092-1107.
- [2] Franke, J., Härdle, W., Hafner, Ch. (2004), *Statistics of Financial Markets* (Springer, Berlin-Heidelberg).
- [3] Franke, J., Diagne M. (2001), Estimating Market risk with neural networks, *Statistics and Decisions*, **19**, 331-366.
- [4] Franke, J., Neumann H. M. (2000), Bootstrapping Neural Networks, *Neural Computation*, **12**, 1929-1949.
- [5] Gouriéroux, Ch. (1997), *ARCH Model and Financial Application* (Springer, New York).
- [6] Hwang, J.T.G., Ding, A.A. (1997), Prediction Intervals for Artificial Neural networks, *J. Amer. Statist. Assoc.* **92**, 748-757.
- [7] Pötscher, B.M., Prucha, I.R. (1997), *Dynamic Nonlinear Econometric Models: Asymptotic Theory* (Springer, Germany).
- [8] White, H. (1984), *Asymptotic Theory for Econometricians* (Academic Press, New York).