Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 415–429
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Variable selection in neural network regression models with dependent data: a subsampling approach☆

Michele La Rocca*, Cira Perna

Department of Economics and Statistics, University of Salerno, via Ponte Don Melillo, Fisciano, Salerno 84084, Italy

Received 1 March 2003; received in revised form 1 January 2004; accepted 15 January 2004

Abstract

The problem of variable selection in neural network regression models with dependent data is considered. In this framework, a test procedure based on the introduction of a measure for the variable relevance to the model is discussed. The main difficulty in using this procedure is related to the asymptotic distribution of the test statistic which is not one of the familiar tabulated distributions. Moreover, it depends on matrices which are very difficult to estimate because of their complex structure. To overcome these analytical issues and to get a consistent approximation for the sampling distribution of the statistic involved, a subsampling scheme is proposed. The procedure, which takes explicitly into account the dependence structure of the data, will be justified from an asymptotic point of view and evaluated in finite samples by a small Monte Carlo study.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Artificial neural networks; Dependent data; Subsampling

1. Introduction

Over the past decade artificial neural networks have become increasingly widely used in a variety of statistical problems but they are not yet supported by the rich collection of specification and diagnostic tests usually employed in statistical and econometric

☆ The paper has been supported by MURST and CNR funds.
* Corresponding author. Tel.: +39-089-962200; fax: +39-089-962049.
E-mail address: larocca@unisa.it (M. La Rocca).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.01.004

modelling. In a regression framework, a key issue is variable selection, to avoid omission of relevant variables or inclusion of irrelevant ones. In addition, when dealing with dependent data a lag selection procedure is also needed. Due to the black-box nature of the network, this problem cannot be faced focusing on single weights since it is difficult, if not impossible, to interpret them. Testing whether the weights are significantly different from zero could be misleading since a given approximation accuracy can be obtained with different network topologies. Moreover, this approach can be inadequate in testing the overall significance of an explanatory variable. Hence, we focus on statistical procedures based on the introduction of a measure for the variable relevance to the model, which can be used as a basis for a formal statistical test.

The aim of the paper is twofold. Firstly, we suggest a test procedure to select a proper set of input variables and we derive its asymptotic distribution by extending some results available for the iid framework to the case of dependent data. Secondly, we propose to use a subsampling scheme, which takes explicitly into account the dependence structure of the data, to get an alternative approximation for the sampling distribution of the test statistic. The use of a resampling scheme is necessary since the asymptotic distribution of the test statistic is not one of the familiar tabulated distributions and it depends on matrices which are difficult to estimate because of their complex structure. The proposed subsampling test procedure will be justified from an asymptotic point of view, by proving its consistency, and evaluated in finite samples by a small Monte Carlo study.

The paper is organized as follows. In Section 2 we describe the structure of the data generating process and the neural network model employed. In Section 3 we discuss the relevance measure approach to variable selection in neural network regression models with dependent data. In Section 4 we derive the asymptotic distribution of a proper class of test statistics. In Section 5 we introduce a subsampling scheme to get an approximation for its sampling distribution and we prove a consistency result. In Section 6 in order to evaluate the performance of the proposed procedure for finite samples, we discuss the results of a small Monte Carlo study. Some concluding remarks will close the paper.

2. The data generating process and the neural network model

Let $\{Y_t\}$, $t \in \{1,\dots,T\}$, a time series modeled as $Y_t = g(\mathbf{X}_t) + \varepsilon_t$, where $g(\cdot)$ is a continuously differentiable function defined on a compact subset $\mathscr{X}$ of $\mathbb{R}^d$ and $\mathbf{X}_t = (X_{1t},\dots,X_{dt})'$ is a vector of d random variables possibly including explanatory variables, lagged explanatory variables and lagged values of Y_t.

We will assume that the following framework holds.

Assumption A (*Data generating process*). (i) $\mathbb{E}(\varepsilon_t|\mathbf{X}_t) = 0, \ \forall t$.

(ii) Let $\mathbf{Z}_t = \{(Y_t, \mathbf{X}_t')\}'$ with $\mathbf{X}_t$ bounded and $\|Y_t\|_p < \infty$ where $\|\cdot\|_p = (\mathbb{E}|\cdot|^p)^{1/p}$. $\mathbf{Z}_t$ is a stationary, α-mixing sequence on a complete probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with $\alpha(m)$ of size $-p/(p-2)$ $p \geqslant 4$.

Download English Version:

https://daneshyari.com/en/article/10328176

Download Persian Version:

https://daneshyari.com/article/10328176

Daneshyari.com