Neurocomputing 74 (2011) 2552–2565

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Testing correct model specification using extreme learning machines

Jin Seo Cho [a,*], Halbert White [b]

[a] School of Economics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Republic of Korea
[b] Department of Economics, University of California, San Diego, 9500 Gilman Drive #0508, La Jolla, CA 92093-0508, United States

ARTICLE INFO

Available online 12 May 2011

Keywords:
Artificial neural networks
Gaussian process
Functional regression
Extreme learning machines

ABSTRACT

Testing the correct model specification hypothesis for artificial neural network (ANN) models of the conditional mean is not standard. The traditional Wald, Lagrange multiplier, and quasi-likelihood ratio statistics weakly converge to functions of Gaussian processes, rather than to convenient chi-squared distributions. Also, their large-sample null distributions are problem dependent, limiting applicability. We overcome this challenge by applying functional regression methods of Cho et al. [8] to extreme learning machines (ELM). The Wald ELM (WELM) test statistic proposed here is easy to compute and has a large-sample standard chi-squared distribution under the null hypothesis of correct specification. We provide associated theory for time-series data and affirm our theory with some Monte Carlo experiments.

1. Introduction

Artificial neural networks (ANNs) are extensively used to approximate stochastic relationships of unknown form. Their appeal is based on their universal approximation properties (see, e.g., [20,21,31,37]). Nevertheless, ANNs are often difficult to apply, as they may require estimating a large number of unknown parameters (network weights). Consequently, ANNs may suffer from overfitting, and their predictive power can be poor for this reason.

This limitation motivates the search for parsimonious ANN models. For this, one can first train a small ANN network and then train a larger network. One can then test whether or not the fit of the larger network improves to a statistically significant degree. If it does, one has evidence that the parsimonious model suffers from misspecification, in the sense that the smaller model's errors contain approximation error as well as pure prediction error. If not, one has evidence that the larger network is unnecessary, as the smaller network's errors are mostly pure prediction error.

There are many test statistics in the literature that can be applied for this purpose; their variety is due to the fact that this is a non-standard testing problem. Specifically, there are unidentified parameters under the null of correct model specification (see [11,12]); this causes standard test statistics to behave in non-standard ways. For example, the Wald test examined by Bierens [4] and Hansen [18] for cross-section and time-series data, respectively, does not follow a standard chi-squared distribution under the null of correct specification. Instead, it weakly converges to a function of a Gaussian process, and its limiting null distribution is problem dependent. As another example, the quasi-likelihood ratio (QLR) test statistic designed to overcome the twofold identification problem pointed out by Cho et al. [9] also does not follow a standard distribution under the null. As White [35] and Kuan and White [23] note, it can be a challenging task to construct test statistics in such a way that they follow standard distributions under the null and, at the same time, have non-negligible power to detect misspecification.

The goal of the current study is motivated by this observation. We seek a statistical test for the correct model specification hypothesis, whose application is straightforward and that has a standard asymptotic null distribution. This test can then be used to help construct parsimonious ANN models.

We achieve our goal by combining the theory of functional regression with that of extreme learning machines (ELM). Cho et al. [8] study functional regression, in which functional data are regressed against known deterministic functions. These authors develop a statistic to test whether or not the population mean of the functional data is a constant function. Conveniently, the statistic follows a standard limiting chi-squared distribution under the null. Nevertheless, the required integrations make computing the test statistic quite demanding. This limits applicability, except when the observed functional data have a fairly simple structure. Here, we avoid this difficulty by applying extreme learning machines (ELM) to generate functional data for the test. Applying ELM methods proposed by Huang et al. [22] and by White [42] (QuickNet) indeed enables our Wald ELM (*WELM*) test to be computed very conveniently.

The plan of this paper is as follows. We first introduce and heuristically motivate the WELM test in Sections 2 and 3. The

* Corresponding author.
E-mail addresses: jinseocho@yonsei.ac.kr (J.S. Cho), hwhite@ucsd.edu (H. White).
URLS: http://web.yonsei.ac.kr/jinseocho (J.S. Cho), http://weber.ucsd.edu/~hwhite/ (H. White).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.11.031

WELM test is analyzed in Section 4 under a set of formal regularity conditions. There, we examine large-sample properties of the WELM statistic under the null and alternative hypotheses. We report the results of some Monte Carlo experiments in Section 5; mathematical proofs are gathered into the Appendix.

2. The data generating process (DGP) and ANN models

We suppose the data to be analyzed are weakly dependent time-series data:

Assumption A1 (*DGP*). Let $(\Omega,\mathcal{F},\mathbb{P})$ be a complete probability space, and let $k\in\mathbb{N}$. Let $\{(Y_t,\mathbf{X}_t')':\Omega\to\mathbb{R}^{1+k}:t=1,2,\ldots\}$ be a strictly stationary and absolutely regular process with mixing coefficients β_τ such that for some $\rho>1$, $\sum_{\tau=1}^{\infty}\tau^{2\rho/(\rho-1)}\beta_\tau<\infty$.

Here, Y_t and $\mathbf{X}_t$ are serially dependent target and explanatory variables, respectively. For convenience, we permit $\mathbf{X}_t$ to include a constant element. It may also include Y_{t-1},Y_{t-2}, etc. The mixing coefficients β_τ measure the dependence of the stochastic process as

$$\beta_\tau\equiv\sup_{s\in\mathbb{N}}E\left[\sup_{A\in\mathcal{F}_{s+\tau}^{\infty}}|P(A|\mathcal{F}_{-\infty}^{s})-P(A)|\right],$$

where $\mathcal{F}_s^t$ is the σ-field (information set) generated by $(Y_s,\mathbf{X}_s'),\ldots,(Y_t,\mathbf{X}_t')$, and Assumption A1 implies that β_τ is of the size $-2\rho/(\rho-1)$. That is, for some $\varepsilon>0$, $\beta_\tau=O(\tau^{-2\rho/(\rho-1)-\varepsilon})$. The dependence allowed here is less than that assumed in the previous literature, such as Hansen [18,19], Cho and White [7,10], Cho et al. [8,9]. On the other hand, this stronger condition enables us not only to consistently estimate the asymptotic covariance matrix of our estimator, despite the dependence, but also to ensure applicability of the functional central limit (FCLT) theorem, both of which are needed in deriving the asymptotic null distribution of our test statistic.

We suppose that the researcher's interest lies in estimating $E[Y_t|\mathbf{X}_t]$, the mean-squared-error optimal predictor of Y_t given $\mathbf{X}_t$ (Conditional quantile-based prediction may also be of interest, as in [38]; here we focus only on the conditional mean.). Suppose further that the researcher approximates $E[Y_t|\mathbf{X}_t]$ using a function Φ as

$$E[Y_t|\mathbf{X}_t]\approx\Phi(\mathbf{X}_t,\boldsymbol{\theta}_*),$$

where $\boldsymbol{\theta}_*$ is a suitably chosen parameter vector. For example, $\Phi(\mathbf{X}_t,\boldsymbol{\theta}_*)$ could be the output of a hidden layer feedforward network (e.g., as in [23]) with network weights $\boldsymbol{\theta}_*$. We call Φ the researcher's "specification" for $E[Y_t|\mathbf{X}_t]$. The question we consider here is whether for some $\boldsymbol{\theta}_*$ the approximation is exact (correct specification), so there is no possible way to improve the prediction, or if the approximation using Φ is not perfect (misspecification), implying that prediction improvements are possible. If the former is true, we can avoid suboptimal predictions resulting from unnecessarily elaborating the prediction method. If the latter, we can avoid suboptimal predictions resulting from an overly simple prediction method.

To address this issue we impose the following assumption.

Assumption A2 (*Model*). Let $\boldsymbol{\Theta}\subset\mathbb{R}^d$, $d\in\mathbb{N}$, be a non-empty compact convex set and suppose that for each $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $\Phi(\cdot,\boldsymbol{\theta}):\mathbb{R}^k\mapsto\mathbb{R}$ is a measurable function such that for each $\omega\in F\in\mathcal{F}$ with $\mathbb{P}(F)=1$, $\Phi(\mathbf{X}_t(\omega),\cdot):\boldsymbol{\Theta}\mapsto\mathbb{R}$ is twice continuously differentiable. Let $\boldsymbol{\Lambda}\subset\mathbb{R}$ and $\boldsymbol{\Delta}\subset\mathbb{R}^k$ be non-empty compact convex sets with $0\in\mathrm{int}(\boldsymbol{\Lambda})$, and suppose that $\Psi:\mathbb{R}\mapsto\mathbb{R}$ is a non-polynomial analytic function.

We can now formally state the null hypothesis of correct specification of Φ and the alternative of misspecification as

$$\mathcal{H}_0:\text{For some }\boldsymbol{\theta}_*\in\boldsymbol{\Theta},\quad\mathbb{P}[E(Y_t|\mathbf{X}_t)=\Phi(\mathbf{X}_t,\boldsymbol{\theta}_*)]=1$$

versus

$$\mathcal{H}_1:\text{For all }\boldsymbol{\theta}\in\boldsymbol{\Theta},\quad\mathbb{P}[E(Y_t|\mathbf{X}_t)=\Phi(\mathbf{X}_t,\boldsymbol{\theta})]<1.$$

The *augmented* specification $f(\mathbf{X}_t;\boldsymbol{\theta},\lambda,\boldsymbol{\delta})\coloneqq\Phi(\mathbf{X}_t,\boldsymbol{\theta})+\lambda\Psi(\mathbf{X}_t'\boldsymbol{\delta})$ is the original specification, Φ, augmented by the contribution from an additional hidden unit with activation function Ψ, input-to-hidden weights $\boldsymbol{\delta}$, and hidden-to-output weight λ. The augmented specification generates the (augmented) *model* $\mathcal{M}$ as

$$\mathcal{M}\coloneqq\{f(\cdot;\boldsymbol{\theta},\lambda,\boldsymbol{\delta}):(\boldsymbol{\theta},\lambda,\boldsymbol{\delta})\in\boldsymbol{\Theta}\times\boldsymbol{\Lambda}\times\boldsymbol{\Delta}\}.$$

Thus, the *null* model $\mathcal{M}_0\coloneqq\{\Phi(\cdot,\boldsymbol{\theta}):\boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ obtains when $\lambda=0$ and is nested in $\mathcal{M}$. When Φ is correctly specified, $E[Y_t|\mathbf{X}_t]\in\mathcal{M}_0$, that is, $\mathcal{H}_0$ holds. Otherwise, $\mathcal{H}_1$ holds. The advantage to specifying $\mathcal{M}$ as we have is that, as Stinchcombe and White [31] show, if Ψ is *generically comprehensively revealing* (GCR), then whenever $\mathcal{H}_1$ holds, the augmented specification f is guaranteed to provide a better prediction than the original specification Φ, revealing the presence of arbitrary misspecification of Φ. That is, we just need to check whether adding a single suitably chosen (i.e., GCR) hidden unit to the original specification can improve prediction performance. By assuming that Ψ is non-polynomial analytic, we ensure that it is GCR.

There are many admissible choices for Ψ, and each has its own merits. For example, White [35] considers the logistic cumulative distribution function (CDF) for Ψ; Bierens [4] examines the correct model specification assumption by letting $\Psi=\exp$; and Candés [6] analyzes ridgelet functions. In addition, the familiar trigonometric functions will also work.

Note also that Φ is only mildly restricted. It can be any feedforward network with any finite number of hidden units and weights and any sufficiently smooth activation functions. A particularly simple but important case is that of a linear input–output network, $\Phi(\mathbf{X}_t,\boldsymbol{\theta})=\mathbf{X}_t'\boldsymbol{\theta}$, as this form is widely used for making predictions. White [35] and Lee et al. [24], among others, explicitly test the linearity hypothesis using various analytic functions Ψ.

The literature provides many testing procedures suited to our present goal. For example, the goodness-of-fit test examined by Delgado and Stute [13] can be used to test $\mathcal{H}_0$. Nevertheless, to maintain a tight focus in what follows, we limit attention to tests utilizing the universal approximation feature of ANNs.

First, Wald and Lagrange multiplier (LM) test statistics are specifically examined by Bierens [4] and Hansen [18]. Their approach mainly focuses on the coefficient λ. The null model $\mathcal{M}_0$ is generated if λ_* is zero, where λ_* is the probability limit of the nonlinear least-squares (NLS) estimator, say $\hat{\lambda}_n$ Thus, a diagnostic test for correct specification can be constructed using the standard Wald or LM statistics, testing $\lambda_*=0$. But when $\lambda_*=0$, the associated optimal prediction parameter $\boldsymbol{\delta}_*$ is not identified, where $\boldsymbol{\delta}_*$ is the probability limit of the NLS estimator, say $\hat{\boldsymbol{\delta}}_n$. The so-called Davies problem [11,12] of nuisance parameters identified only under the alternative hypothesis is present, and the asymptotic null distributions of the test statistics are different from the standard chi-squared distribution. Accordingly, a great deal of effort has focused on determining these test statistics' asymptotic null distribution. In most cases, this is a function of a Gaussian process defined on $\boldsymbol{\Lambda}$, but the asymptotic null distribution is different for each model $\mathcal{M}$. This is also true of Neyman's [27–29] $C(\alpha)$ test.

Second, White [35] and Lee et al. [24] take a different approach. They avoid the problem of nuisance parameters identified only under the alternative by randomly selecting input-to-hidden weights $\boldsymbol{\delta}_j$, $j=1,\ldots,p$, and estimating the hidden-to-output weights by simple least-squares methods. This random selection and estimation approach is exactly that known as extreme learning machines (ELM) in the analysis of Huang et al. [22] and also