Chapter 3
Testing for Neglected Nonlinearity Using Regularized Artificial Neural Networks

Tae-Hwy Lee, Zhou Xi, and Ru Zhang

Abstract The artificial neural network (ANN) test of Lee et al. (Journal of Econometrics 56, 269–290, 1993) uses the ability of the ANN activation functions in the hidden layer to detect neglected functional misspecification. As the estimation of the ANN model is often quite difficult, LWG suggested activate the ANN hidden units based on randomly drawn activation parameters. To be robust to the random activations, a large number of activations is desirable. This leads to a situation for which regularization of the dimensionality is needed by techniques such as principal component analysis (PCA), Lasso, Pretest, partial least squares (PLS), among others. However, some regularization methods can lead to selection bias in testing if the dimensionality reduction is conducted by supervising the relationship between the ANN hidden layer activations of inputs and the output variable. This paper demonstrates that while these supervised regularization methods such as Lasso, Pretest, PLS, may be useful for forecasting, they may not be used for testing because the supervised regularization would create the post-sample inference or post-selection inference (PoSI) problem. Our Monte Carlo simulation shows that the PoSI problem is especially severe with PLS and Pretest while it seems relatively mild or even negligible with Lasso. This paper also demonstrates that the use of unsupervised regularization does not lead to the PoSI problem. Lee et al. (Journal of Econometrics 56, 269–290, 1993) suggested a regularization by principal components, which is a unsupervised regularization. While the supervised regularizations may be useful in forecasting, regularization should not be supervised in inference.

Keywords Randomized ANN activations • Dimension reduction • Supervised regularization • Unsupervised regularization • PCA • Lasso • PLS • Pretest • PoSI problem

T.-H. Lee (✉) • Z. Xi • R. Zhang
Department of Economics, University of California, Riverside, CA 92521, USA
e-mail: taelee@ucr.edu; zhou.xi@email.ucr.edu; ru.zhang@email.ucr.edu

J. Ma and M. Wohar (eds.), *Recent Advances in Estimating Nonlinear Models: With Applications in Economics and Finance*, DOI 10.1007/978-1-4614-8060-0_3,

3.1 Introduction

In this paper we explore the issues in testing for functional forms, especially for neglected nonlinearity in parametric linear models. Many papers have appeared in the recent literature which deal with the issues of how to carry out various specification tests in parametric regression models. To construct the tests, various methods are used to estimate the alternative models. For example, Fan and Li (1996), Li and Wang (1998), Zheng (1996), and Bradley and McClelland (1996) use local constant kernel regression; Hjellvik, Yao, and Tjøstheim (1998) and Tjøstheim (1999) use local polynomial kernel regression; Cai, Fan, and Yao (2000) and Matsuda (1999) use nonparametric functional coefficient models; Poggi and Portier (1997) use a functional autoregressive model; White (1989), Lee, White, and Granger (1993), Teräsvirta, Lin, and Granger (1993), Granger and Teräsvirta (1993), Teräsvirta (1996), and Corradi and Swanson (2002) use neural network models; Eubank and Spiegelman (1990) use spline regression; Hong and White (1995) use series regression; Stengos and Sun (2001) use wavelet methods; and Hamilton (2001) uses a parametric flexible regression model.

There are also many papers which compare different approaches in testing for linearity. For example, Lee, White, and Granger (1993), Teräsvirta, Lin, and Granger (1993), Teräsvirta (1996), and Lee (2001) examine the neural network test and many other tests. Dahl (2002) and Dahl and González-Rivera (2003) study Hamilton's (2001) test and compare it with various tests including the neural network test. Blake and Kapetanios (2000, 2003) extend the neural network test using a radial basis function for the neural network activation function instead of using the typical logistic function used in Lee, White, and Granger (1993).[1] Lee and Ullah (2001, 2003) examine the tests of Li and Wang (1998), Zheng (1996), Ullah (1985), Cai, Fan, and Yao (2000), Härdle and Mammen (1993), and Aït-Sahalia, Bickel and Stoker (2001). Fan and Li (2001) compare the tests of Li and Wang (1998), Zheng (1996), and Bierens (1990). Whang (2000) generalizes the Kolmogorov–Smirnov and Cramer-von Mises tests to the regression framework and compare them with the tests of Härdle and Mammen (1993) and Bierens and Ploberger (1997). Hjellvik and Tjøstheim (1995, 1996) propose tests based on nonparametric estimates of conditional mean and variances and compare them with a number of tests such as the bispectrum test and the BDS test.

This paper further investigates the artificial neural network (ANN) test. The ANN test is a conditional moment test whose null hypothesis consists of conditional moment conditions that hold if the linear model is correctly specified for the conditional mean. The ANN test differs from other tests by the choice of the 'test function' that is chosen to be the ANN's hidden layer activations. It can be checked for their correlation with the residuals from the linear regression model. The advantage to use an ANN model to test nonlinearity is that the ANN model

[1] For radial basis functions, see (e.g.) Campbell, Lo and Mackinlay (1997, p. 517).

inherits the flexibility as a universal approximator of unknown functional form. Hornik et al. (1989) show that neural network is a nonlinear flexible functional form being capable of approximating any Borel measurable function to any desired level of accuracy provided sufficiently many hidden units are available.

We consider an augmented single hidden layer feedforward neural network model in which network output y_t is determined given input $\mathbf{x}_t$ as

$$y_t = \mathbf{x}_t'\boldsymbol{\alpha} + \sum_{j=1}^{q} \beta_j \Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right) + u_t, \tag{3.1}$$

where $t = 1, \ldots, T$, $\mathbf{x}_t = (x_{1,t}, \ldots x_{N,t})'$, $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_q'\right)'$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$, and $\boldsymbol{\gamma}_j = \left(\gamma_{j,1}, \ldots, \gamma_{j,N}\right)'$ for $j = 1, \ldots, q$, and $\Psi(\cdot)$ is an activation function. An example of the activation function is the logistic function $\Psi(z) = (1 + \exp(z))^{-1}$. $\boldsymbol{\alpha}$ is a conformable column vector of connection strength from the input layer to the output layer; $\boldsymbol{\gamma}_j$ is a conformable column vector of connection strength from the input layer to the hidden units, $j = 1, \ldots, q$; β_j is a (scalar) connection strength from the hidden unit j to the output unit, $j = 1, \ldots, q$; and Ψ is a squashing function (e.g., the logistic squasher) or a radial basis function. Input units $\mathbf{x}$ send signals to intermediate hidden units, then each of the hidden unit produces an activation Ψ that then sends signals toward the output unit. The integer q denotes the number of hidden units added to the affine (linear) network. When $q = 0$, we have a two-layer affine network $y_t = \mathbf{x}_t'\boldsymbol{\alpha} + u_t$.

It is well known that the ANN models are generally hard to estimate and suffer from possibly large estimation errors which can adversely affect their ability as a universal approximator. To alleviate the estimation errors of an ANN model, it is useful to note that, for given values of $\boldsymbol{\gamma}_j$'s, the ANN is linear in $\mathbf{x}$ and the activation function Ψ and therefore $\left(\boldsymbol{\alpha}', \boldsymbol{\beta}'\right)$ can be estimated from linear regression once $\left(\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_q'\right)$ are estimated or given. As suggested in Lee, White and Granger (1993), a set of $\boldsymbol{\gamma}$'s can be randomly generated. In this paper, we will generate a *large* set of $\boldsymbol{\gamma}$'s such that $\sum_{j=1}^{q} \beta_j \Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$ can capture the maximal nonlinear structure. The LWG statistic is designed to detect neglected nonlinearity in the linear model by checking for correlation between the residual from a linear model and the additional hidden activation functions with randomly generated $\boldsymbol{\gamma}$'s. The additional hidden activation functions are hidden (or phantom) because they do not exist under the null hypothesis. The $\boldsymbol{\gamma}$'s are randomly generated in testing because they are not identified under the null hypothesis. The set of randomly selected $\boldsymbol{\gamma}$'s should be large enough so that it can be dense and make the ANN a universal approximator.

While the architecture of the ANN model makes a universal approximator, it involves a very large number of parameters. Kock and Teräsvirta (2011) consider regularizing the complexity of an ANN model and demonstrate that the regularization of the large dimension is crucial in using ANN models for out-of-sample forecasting. This motivates us to consider regularizing the ANN for testing for

neglected nonlinearity. In fact, Lee et al. (1993) uses a (unsupervised) regularization method, namely the principal component analysis (PCA), for the randomly activated test functions. Kock and Teräsvirta (2011) consider two (supervised) regularization approaches. They insightfully notice that the supervised regularizations will result in the size distortion in inference, and they use these approaches only for forecasting.

One supervised regularization approach considered by Kock and Teräsvirta (2011) to select a small q^* from a large q number of $\boldsymbol{\gamma}$'s is the simple-to-general algorithm, e.g., the QuickNet algorithm of White (2006), that adds one $\boldsymbol{\gamma}$ and one activation function at a time to the ANN. The QuickNet expands starting from 0 activation to q^* activations until the additional hidden unit activation is not found to improve the network capability. The second supervised regularization approach considered by Kock and Teräsvirta (2011) is the general-to-simple approach. This approach, from a variable-selection perspective, reduces the number of activations from an initial large number q (say, 1,000) to a smaller number q^* by penalizing the complexity of the ANN model. The penalized regression methods include the smoothly clipped absolute deviation penalty (SCAD) (Fan and Li 2001), adaptive Lasso (Zou 2006), adaptive elastic net (Zou and Zhang 2009), the bridge estimator (Huang, Horowitz and Ma 2008), among others. In the case where q is larger than the degrees of freedom, the marginal bridge estimator (Huang, Horowitz and Ma 2008) or the sure independence screening (SIS) (Fan and Lv 2008) may be used to reduce q below the degrees of freedom and then apply these estimation methods.

The third approach is to follow Lee et al. (1993) to compute the q^* principal components of the q additional hidden activation functions. Since the activation functions using randomly generated $\boldsymbol{\gamma}$'s may be collinear with each other and with $\mathbf{x}_t$, LWG used principal components of the q additional hidden activation functions. Unlike the above two supervised approaches, the principal components are not supervised for the output y.

The purpose of this paper is to examine the effect of various regularization on the ANN test for neglected nonlinearity when the ANN is activated based on a large number of random activation parameters. We learn two points. First, when we consider the Lasso, the partial least square (PLS) method, the Pretest method, and a method combining Lasso with principal components, these supervised regularization methods bring size-distortion and the ANN test suffers from the post-sample inference or post-selection inference (PoSI) problem.[2] Secondly, when we use the PCA as used in Lee et al. (1993), this unsupervised regularization of the dimension reduction does not bring the PoSI problem, works really well for a large q, and the asymptotic $\chi^2(q^*)$ distribution does well in approximating the finite sample distribution of the ANN test statistic. To sum, while the supervised regularizations are useful in forecasting as studied by Bai and Ng (2008), Bair, Hastie, Paul, and Tibshirani (2006), Inoue and Kilian (2008), Huang and Lee (2010), Hillebrand, Huang, Lee, and Li (2011), Kock and Teräsvirta (2011), and Kock (2011), this paper shows that regularization should not be supervised in inference.

[2]See Pötscher and Leeb (2009) and Berk, Brown, Buja, Zhang and Zhao (2011).

Our Monte Carlo simulation shows that the PoSI problem is especially severe with PLS and Pretest while it seems relatively mild or even negligible with Lasso. This paper also demonstrates that the use of unsupervised regularization by principal components does not lead to the PoSI problem.

The plan of the paper is as follows. In Section 3.2 we review the ANN test. Section 3.3 introduces various regularizations in two types, unsupervised and supervised. Section 3.4 presents the simulation results which demonstrate the PoSI problem of supervised methods. Section 3.5 concludes.

3.2 Testing for Neglected Nonlinearity Using ANN

Consider $Z_t = (y_t \ \mathbf{x}_t')'$, where y_t is a scalar and $\mathbf{x}_t$ may contain a constant and lagged values of y_t. Consider the regression model

$$y_t = m(\mathbf{x}_t) + \varepsilon_t, \tag{3.2}$$

where $m(x_t) \equiv E\,(y_t|\mathbf{x}_t)$ is the true but unknown regression function and ε_t is the error term such that $E(\varepsilon_t|\mathbf{x}_t) = 0$ by construction. To test for a parametric model $g(\mathbf{x}_t, \boldsymbol{\theta})$ we consider

$$H_0 : m(\mathbf{x}_t) = g(\mathbf{x}_t, \boldsymbol{\theta}^*) \text{ for some } \boldsymbol{\theta}^*, \tag{3.3}$$

$$H_1 : m(\mathbf{x}_t) \neq g(\mathbf{x}_t, \boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta}. \tag{3.4}$$

In particular, if we are to test for neglected nonlinearity in the regression models, set $g(\mathbf{x}_t, \boldsymbol{\theta}) = \mathbf{x}_t'\boldsymbol{\alpha}$, $\boldsymbol{\alpha} \subset \boldsymbol{\theta}$. Then under H_0, the process $\{y_t\}$ is linear in mean conditional on $\mathbf{x}_t$, i.e.,

$$H_0 : m(\mathbf{x}_t) = \mathbf{x}_t'\boldsymbol{\alpha}^* \text{ a.e. for some } \boldsymbol{\alpha}^*. \tag{3.5}$$

The alternative of interest is the negation of the null hypothesis, that is,

$$H_1 : m(\mathbf{x}_t) \neq \mathbf{x}_t'\boldsymbol{\alpha} \text{ on a set with positive measure for all } \boldsymbol{\alpha}. \tag{3.6}$$

When the alternative is true, a linear model is said to suffer from "neglected nonlinearity" (Lee, White, and Granger 1993).

If a linear model is capable of an exact representation of the unknown function $m(\mathbf{x}_t)$, then there exists a vector $\boldsymbol{\alpha}^*$ such that (3.5) holds, which implies

$$E\left(\varepsilon_t^*|\mathbf{x}_t\right) = 0 \text{ a.e.}, \tag{3.7}$$

where $\varepsilon_t^* = y_t - \mathbf{x}_t'\boldsymbol{\alpha}^*$. By the law of iterated expectations ε_t^* is uncorrelated with any measurable functions of $\mathbf{x}_t$, say $h(\mathbf{x}_t)$. That is,

$$E\left[h(\mathbf{x}_t)\varepsilon_t^*\right] = 0. \tag{3.8}$$

Depending on how we choose the "test function" $h(\cdot)$, various specification tests may be obtained. The specification tests based on these moment conditions, the so-called the conditional moment tests, have been studied by Newey (1985), Tauchen (1985), White (1987, 1994), Bierens (1982, 1990), Lee et al. (1993), Bierens and Ploberger (1997), and Stinchcombe and White (1998), among others. The ANN test exploits (3.8) with the test function $h(\cdot)$ being chosen as the neural network hidden unit activation functions.

Lee et al. (1993) considered the test of "linearity in conditional mean" using the ANN model. To test whether the process y_t is linear in mean conditional on $\mathbf{x}_t$, they used the following null and alternative hypothesis:

$$H_0 : \Pr\left[\mathrm{E}(y_t|\mathbf{x}_t) = \mathbf{x}_t'\boldsymbol{\alpha}^*\right] = 1 \quad \text{for some } \boldsymbol{\alpha}^*$$
$$H_1 : \Pr\left[\mathrm{E}(y_t|\mathbf{x}_t) = \mathbf{x}_t'\boldsymbol{\alpha}\right] < 1 \quad \text{for all} \quad \boldsymbol{\alpha}.$$

The procedure to construct the LWG test statistic is as follows. Under the null hypothesis that y_t is linear in conditional mean, we first estimate a linear model of y_t on $\mathbf{x}_t$, then if any nonlinearity is neglected in the OLS regression, it will be captured by the residual term $\hat{u}_t$. Since the ANN model inherits the flexibility as a universal approximator of unknown functional form, we can apply an ANN function to approximate any possible types of nonlinearity in the residual term $\hat{u}_t$.

The neural network test is based on a test function $h(\mathbf{x}_t)$ chosen as the activations of "phantom" hidden units $\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$, $j = 1, \ldots, q$, where $\boldsymbol{\gamma}_j$ are randomly generated column vectors independent of $\mathbf{x}_t$. $\boldsymbol{\gamma}_j$'s are not identified under the null hypothesis of linearity, cf. Davies (1977, 1987), Andrews and Ploberger (1994), and Hansen (1996). That is,

$$E\left[\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)\varepsilon_t^*\right] = 0 \quad j = 1, \ldots, q, \tag{3.9}$$

under H_0, so that

$$E\left(\Psi_t\varepsilon_t^*\right) = 0, \tag{3.10}$$

where

$$\Psi_t = \left(\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_1\right), \ldots, \psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_q\right)\right)' \tag{3.11}$$

is a phantom hidden unit activation vector. Evidence of correlation of ε_t^* with Ψ_t is evidenced against the null hypothesis that y_t is linear in mean. If correlation exists, augmenting the linear network by including an additional hidden unit with activations $\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$ would permit an improvement in network performance. Thus the tests are based on sample correlation of affine network errors with phantom hidden unit activations,

$$n^{-1}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t = n^{-1}\sum_{t=1}^{n}\Psi_t\left(y_t - \mathbf{x}_t'\hat{\boldsymbol{\alpha}}\right), \tag{3.12}$$

where $\hat{\varepsilon}_t = y_t - \mathbf{x}_t'\hat{\boldsymbol{\alpha}}$ are estimated by OLS. Under suitable regularity conditions it follows from the central limit theorem that $n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t \xrightarrow{d} N(0, W^*)$ as $n \to \infty$, and if one has a consistent estimator for its asymptotic covariance matrix, say $\hat{W}_n$, then an asymptotic chi-square statistic can be formed as

$$\left(n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t\right)' \hat{W}_n^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t\right) \xrightarrow{d} \chi^2(q). \tag{3.13}$$

Construct the following auxiliary regression:

$$\hat{u}_t = \mathbf{x}_t'\boldsymbol{\alpha} + \sum_{j=1}^{q}\beta_j\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right) + v_t,$$

where $t = 1, \ldots, T, \mathbf{x}_t = (x_{1,t}, \ldots x_{N,t})', \boldsymbol{\theta} = \left(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_q'\right)', \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)', \boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$, and $\boldsymbol{\gamma}_j = \left(\gamma_{j,1}, \ldots, \gamma_{j,N}\right)'$ for $j = 1, \ldots, q$, and $\psi(\cdot)$ is an activation function. LWG chose the logistic function $\psi(z) = (1 + \exp(z))^{-1}$ as the activation function. If there is nonlinearity remained in the residual, we expect the goodness of fit for the auxiliary regression is high. However, one problem to estimate the auxiliary regression is that, when q is large, there may exist multicollinearity between $\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$ and $\mathbf{x}_t$ and among $\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$ themselves. LWG suggested to choose q^* principal components of q activation functions $\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$, with $q^* < q$, and then use these q^* principal components to run the auxiliary regression. Under the null hypothesis that the sequence y_t is linear conditional on $\mathbf{x}_t$, the goodness of fit in the auxiliary regression will be low. Lee et al. (1993) constructed an LM-type test statistic which has an asymptotic $\chi^2(q^*)$ distribution under the null hypothesis. In their simulations, LWG chose q equal to 10 or 20 and q^* equal to 2 or 3 in different data generating processes (DGP), and the sample size 50, 100, or 200. Moreover, they dropped the first principal component of Ψ_t to avoid the multicollinearity problem. In this paper, we have tried the ANN test both with and without dropping the first principal component, the results do not change much. Thus we keep the original LWG method with dropping the first principal component for the ANN test in this paper.

In practice, we need to generate $\boldsymbol{\gamma}$'s carefully so that $\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$ is within a suitable range. If $\boldsymbol{\gamma}$'s are chosen to be too small, then activation functions $\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$ are approximately linear in $\mathbf{x}$. We want to avoid this situation since they cannot capture much nonlinearity. If $\boldsymbol{\gamma}$'s are too large, the activation functions $\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$ take values close to ± 1 (their maximum or minimum values), and we want to avoid this situation as well. In our study, for different $\mathbf{x}$'s we generate $\boldsymbol{\gamma}$'s from uniform distributions with different supports so that the activation functions are neither too small or too large.

3.3 Regularizing the ANN Test

As discussed above, Lee et al. (1993) regularized the large number of the network activation functions using principal components in order to avoid possible collinearity problem. The $q^* < q$ principal components are used out of q activations. We note that the principal components make its variance largest, yet may not necessarily be the ones that best explain the residuals from the OLS regression, $\hat{u}_t$. In other words, these principal components are not "supervised" for y_t and thus for $\hat{u}_t$. The regularization may be supervised so that the activations that are uncorrelated with $\hat{u}_t$ can be dropped and the activations that are correlated with $\hat{u}_t$ can be selected to increase the power of the test. Such regularization methods include the Lasso method, the PLS method, the Pretest method, and the PCA-first-and-then-Lasso method. We first review the PCA method in the next subsection, and then other regularization methods in the following subsections.

3.3.1 Unsupervised Regularization of the ANN Test Using PCA

Lee et al. (1993) found that the elements of Ψ_t in (3.11) tend to be collinear with $\mathbf{x}_t$ and with themselves and computation of $\hat{W}_n$ can be tedious. Thus they conducted a test on $q^* < q$ principal components of Ψ_t not collinear with $\mathbf{x}_t$, denoted Ψ_t^*, and employ the equivalent test statistic (under conditional homoskedasticity) that avoids explicit computation of $\hat{W}_n$, denoted T_n^{PCA}

$$T_n^{\text{PCA}} \equiv nR^2 \xrightarrow{d} \chi^2(q^*), \tag{3.14}$$

where R^2 is uncentered squared multiple correlation from a standard linear regression of $\hat{\varepsilon}_t$ on Ψ_t^* and $\mathbf{x}_t$. This test is to determine whether or not there exists some advantage to be gained by adding hidden units to the affine network.

It should be noted that the asymptotic equivalence of (3.13) and (3.14) holds under the conditional homoskedasticity, $E(\varepsilon_t^*|x_t) = \sigma^2$. Under the presence of conditional heteroskedasticity such as ARCH, T_n^{PCA} will not be $\chi^2(q^*)$ distributed. To resolve the problem in that case, we can either use (3.13) with $\hat{W}_n$ being estimated robust to the conditional heteroskedasticity (White 1980, Andrews 1991) or use (3.13) with the empirical null distribution of the statistic computed by a bootstrap procedure that is robust to the conditional heteroskedasticity (Wu 1986, Liu 1988).

3.3.2 Supervised Regularization of the ANN Test Using Lasso

The Lasso method is a shrinkage method which can be used as a selector of the activation functions for the ANN test. We use a penalized regression for the auxiliary

model where the coefficients of $\Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$ are shrunken to zero if it is smaller than a particular value. The Lasso problem can be written as

$$\hat{\beta}^{\text{Lasso}} = \arg\min_{\beta} \left\{ \sum_{t=1}^{T} \left(\hat{u}_t - \mathbf{x}_t'\boldsymbol{\alpha} - \sum_{j=1}^{q} \beta_j \Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right) \right)^2 + \lambda \sum_{j=1}^{q} |\beta_j| \right\}.$$

The Lasso method uses the L^1-penalty term $|\beta_j|$, and it has the sparsity property such that some of the β_j's that are small will be shrunken to zero, yet it does not have oracle property according to Fan and Li (2001) in the sense that it will give biased estimates of β_j even when sample size is large. The Lasso method is easier to implement than some other methods that has the oracle property. Since the activation functions are selected according to its explanation power to $\hat{u}_t$, the Lasso is a supervised regularization. The tuning parameter λ determines the number of activation functions selected. To get the test statistic using the Lasso method, we will do the auxiliary regression of $\hat{u}_t$ on the q^* selected activation functions Ψ^* (denoting q^*-vector of Lasso-selected activations) and get $T_n^{\text{Lasso}} = nR^2_{\text{Lasso}}$. We choose λ such that $q^* = 3$. In Section 3.4, we will examine if it has the asymptotic $\chi^2(q^*)$ distribution or if it is subjected to the PoSI problem due to the supervision in regularizing the dimension from q to q^*.

3.3.3 *Supervised Regularization of the ANN Test Using PLS*

Like PCA, the PLS method constructs variables using linear combinations of activation functions. Yet like Lasso, it is supervised using information about $\hat{u}_t$. The algorithm of the PLS method used in this test is described as follows:

1. Standardize each $\Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$ to zero mean and unit variance. Set $\tilde{u}_t^{(0)} = \bar{\hat{u}}_t\boldsymbol{\iota}$, $\Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)^{(0)} = \Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$, for $j = 1, \ldots, q$, where $\boldsymbol{\iota} = (1, ..., 1)'$.
2. For $m = 1, \ldots, q$,
   (a) Construct the linear combination, $z_m = \sum_{j=1}^{q} \omega_m \Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)^{(m-1)}$, where the weight is equal to the covariance between $\Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)^{(m-1)}$ and $\hat{u}_t$: $\omega_m = \text{cov}\left(\Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)^{(m-1)}, \hat{u}_t\right)$.
   (b) Regress $\hat{u}_t$ on z_m, and get the coefficient: $\hat{\theta}_m = \text{cov}(z_m, \hat{u}_t)/\text{var}(z_m)$.
   (c) Update $\tilde{u}_t^{(m)}$ by $\tilde{u}_t^{(m)} = \tilde{u}_t^{(m-1)} + \hat{\theta}_m z_m$.
   (d) Update $\Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)^{(m)}$ by orthogonalizing each $\Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)^{(m-1)}$ with respect to z_m: $\Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)^{(m)} = \Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)^{(m-1)} - \left[\text{cov}(\Psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)^{(m-1)}, z_m)/\text{var}(z_m)\right] z_m$, $j = 1, \ldots, q$.
3. The fitted value of residual terms by PLS is given by $\tilde{u}_t^{(m)}$ and the selected linear combinations of activation functions are given by z_m.

In this test, we select the first q^* largest z_m and then do auxiliary regression of $\hat{u}_t$ on z_m to get the test statistic $T_n^{\text{PLS}} = nR^2_{\text{PLS}}$. In Section 3.4, we will examine if it has the asymptotic $\chi^2(q^*)$ distribution or if it is subjected to the PoSI problem due to the supervision in regularizing the dimension from q to q^*.

3.3.4 *Supervised Regularization of the ANN Test Using Pretests*

The PCA shrinkage includes all the information of the activation vector Ψ_t, including those that are irrelevant to explain the residuals from the linear regression. We may consider to make further shrinkage from the principal components. In this section, we consider the Pretest method on the principal components, as implemented by Inoue and Kilian (2008). We first get $k = 20$ principal components from the q activation vector Ψ_t and then regress the residual from the OLS regression on these k principal components. Then we choose $q^* = 3$ principal components corresponding to the coefficients with the highest absolute t-values. Then the test statistic for this Pretest method is equal to $T_n^{\text{Pretest}} = nR^2_{\text{Pretest}}$. Similarly, we will examine if it has the asymptotic $\chi^2(q^*)$ distribution or if it is subjected to the PoSI problem due to the supervision in regularizing the dimension from q to q^*, in Section 3.4.

The Pretest method described here is essentially the "PCA-first-and-then-Pretest." In the next subsection, we will consider the "PCA-first-and-then-Lasso."

3.3.5 *Supervised Regularization of the ANN Test Using PCA-First-and-Then-Lasso*

Instead of using Pretest to supervise the original ANN test, we also use the Lasso method to supervise the principal components. In this subsection, we combine the PCA and the Lasso method. We first get a relatively larger number of k (e.g., 100, 50, 10 or 5) principal components from the q-vector Ψ_t of activation functions and then use the Lasso method to shrink them except for the $q^* = 3$ principal components. In this way, we can select the principal components that best fits the residuals from the OLS regression and increase the power of the test. We then do the auxiliary regression using the selected q^* principal components and get the test statistic $T_n^{\text{PCA}-\text{lasso}} = nR^2_{\text{PCA}-\text{lasso}}$. In Section 3.4, we will examine if the ANN test using this method of "PCA-first-and-then-Lasso" can still follow the asymptotic $\chi^2(q^*)$ distribution or if it is subjected to the PoSI problem due to the supervision in regularizing the dimension from q to q^*.

3.3.6 The PoSI Problem

Regularized methods of estimation have been developed intensively in the past 20 years. Examples includes the Bridge estimator of Frank and Friedman (1993), the least absolute selection and shrinkage (Lasso) estimator of Tibshirani (1996), the least angle regression (LARS) of Efron, Hastie, Johnstone, Tibshirani (2004), the SCAD estimator of Fan and Li (2001), and the traditional hard-thresholding Pretest methods. It is tempting to use these supervised regularization in reducing the large number of randomized ANN activations. However, as noted in Leeb and Pötscher (2003, 2005, 2006, 2008), Pötscher and Leeb (2009), Berk et al. (2011), and others, subset-searches like the Lasso shrinkage method suffer from the post sample inference (PoSI) problem. See also Hoover (2012) on a related issue of size distortion resulted from model-search. In Section 3.4, we show that PLS, Pretest, PCA-first-and-then-Lasso will cause the PoSI problem that the distribution under the null hypothesis is different from the $\chi^2(q^*)$ distribution cf. Leeb and Pötscher (2008).

To illustrate the PoSI problem, we take the Lasso supervision as an example. When using the Lasso method to select the activation functions, we are actually making selection between the following two models:

$$\mathbf{M}_0 : \mathbf{Y} = \mathbf{X}_0'\boldsymbol{\beta}_0 + \boldsymbol{\upsilon}_1$$

versus

$$\mathbf{M}_1 : \mathbf{Y} = \mathbf{X}_0'\boldsymbol{\beta}_0 + \mathbf{X}_1'\boldsymbol{\beta}_1 + \boldsymbol{\upsilon}_2,$$

where Y is the residual term $\hat{u}_t$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are vectors of parameters, $\mathbf{X}_0$ and $\mathbf{X}_1$ are partitions of the activation function Ψ_t and $\boldsymbol{\upsilon}_1$, $\boldsymbol{\upsilon}_2$ are the error terms. If the Lasso method shrinks $\boldsymbol{\beta}_1$ to 0, then we use model $\mathbf{M}_0$ to test the null hypothesis $\mathbf{H}_0 : \boldsymbol{\beta}_0 = 0$, and we denote the corresponding LM test statistic by $T_{n,\mathbf{M}_0}$; if the Lasso method does not shrink $\boldsymbol{\beta}_1$ to 0, we pick up model $\mathbf{M}_1$ and obtain the test statistic $T_{n,\mathbf{M}_1}$. Let $\mathbf{M}$ be the model selected, therefore the test statistic accounting for model selection is:

$$T = T_{n,\mathbf{M}_0} \times \mathbf{1}_{(\mathbf{M}=\mathbf{M}_0)} + T_{n,\mathbf{M}_1} \times \mathbf{1}_{(\mathbf{M}=\mathbf{M}_1)},$$

where $\mathbf{1}(\cdot)$ is the indicator function.

If $\mathbf{M}_0$ is the true model, we know $T_{n,\mathbf{M}_0}$ follows a $\chi^2(q_0)$ distribution with q_0 equal to $\dim \boldsymbol{\beta}_0$; on the other hand, if $\mathbf{M}_1$ is the true model, $T_{n,\mathbf{M}_1}$ has a $\chi^2(q_1)$ distribution with q_1 equal to $\dim \boldsymbol{\beta}_0 + \dim \boldsymbol{\beta}_1$. In both cases, we know the exact distribution and can find the critical value. However, since we randomly draw $\boldsymbol{\gamma}_j$'s and randomly activate $\psi\left(\mathbf{x}_t'\boldsymbol{\gamma}_j\right)$, $j = 1, \ldots, q$, many elements in the activation vector Ψ_t can be highly collinear and as a result the Lasso method may not distinguish the two models. Hence, even if $\mathbf{M}_0$ is the true model the Lasso supervision may include some incorrect activation functions, and the distribution of the test statistic can be a mixture of two χ^2 distributions with different degrees

of freedom. To make things worse, as every time we randomly generate different sets of Ψ_t, we cannot compute the probability of choosing $\mathbf{M}_0$ or $\mathbf{M}_1$ as the true model. This means that we cannot obtain the exact distribution of the test statistic and the usual $\chi^2_{q^*}$ critical value is invalid. This will be shown via simulation in the next section. As will be shown, the PoSI problem is especially severe with PLS and Pretest while it seems relatively mild or even negligible with Lasso.

3.4 Monte Carlo

3.4.1 DGPs and Simulation Design

To generate data we use the following DGPs, all of which have been used in the related literature. There are two blocks. All the error terms ε_t below are i.i.d. $N(0, 2^2)$. Two blocks of DGP are considered. The first block has DGPs using the univariate series of y_t, and the second block introduces two external variables x_{1t} and x_{2t} which follow a bivariate normal distribution. All DGPs below fulfil the conditions for the investigated testing procedures. For those regularity conditions and moment conditions, see White (1994, Chapter 9) for the ANN tests.

Block 1 (Time-series data generating processes)

1. Autoregressive (AR)

$$y_t = 0.6y_{t-1} + \varepsilon_t$$

2. Threshold autoregressive (TAR)

$$y_t = \begin{cases} 0.9y_{t-1} + \varepsilon_t & \text{if } |y_{t-1}| \leq 1 \\ -0.3y_{t-1} + \varepsilon_t & \text{otherwise} \end{cases}$$

3. Sign autoregressive (SGN)

$$y_t = \text{sgn}(y_{t-1}) + \varepsilon_t$$

where

$$\text{sgn}(y_{t-1}) = \begin{cases} 1 & \text{if } y_{t-1} > 0 \\ 0 & \text{if } y_{t-1} = 0 \\ -1 & \text{otherwise} \end{cases}$$

4. Nonlinear autoregressive (NAR)

$$y_t = \frac{0.7|y_{t-1}|}{|y_{t-1}| + 2} + \varepsilon_t$$

5. Markov regime-switching (MRS)

$$y_t = \begin{cases} 0.6y_{t-1} + \varepsilon_t & \text{if } S_t = 0 \\ -0.5y_{t-1} + \varepsilon_t & \text{if } S_t = 1 \end{cases}$$

where S_t follows a two-state Markov chain with transition probabilities $\Pr(S_t = 1|S_{t-1} = 0) = \Pr(S_t = 0|S_{t-1} = 1) = 0.3$.

Block 2 (Cross-sectional data generating processes)

This block includes DGPs similar to those in Zheng (1996). Assume x_{1t}, x_{2t} follow a bivariate normal distribution of $N(0, 0, 1, 1, \rho)$ where the correlation $\rho = 0$ or 0.7. We have the following three cases:

1. Linear

$$y_t = 1 + x_{1t} + x_{2t} + \varepsilon_t$$

2. Cross-Product

$$y_t = 1 + x_{1t} + x_{2t} + 0.2x_{1t}x_{2t} + \varepsilon_t$$

3. Squared

$$y_t = 1 + x_{1t} + x_{2t} + 0.2x_{2t}^2 + \varepsilon_t$$

For the simulations, the information set are $\mathbf{x}_t = y_{t-1}$ for Block 1 and $\mathbf{x}_t = (x_{t1}\ x_{t2})'$ for Block 2. The logistic squasher $\psi = [1 + \exp(-\mathbf{x}'\boldsymbol{\gamma})]^{-1}$ is used with $\boldsymbol{\gamma}$ being generated randomly from a uniform distribution on an interval depending on the data range. The number of additional hidden units to the affine network $q = 200$ is used. We set $q^* = 3$ for all regularization methods for simplicity.

3.4.2 Results

Tables 3.1 and 3.2 report the size and power for ANN test with $q = 200$ using various regularization methods (PCA, Lasso, PLS, Pretest, "PCA-first-and-then-Lasso"). The numbers in the tables are the rejection frequencies of the null hypothesis at 5% and 10% levels. The sample size n is equal to 200. We use 1,000 Monte Carlo replications. As demonstrated in Lee et al. (1993) and Lee, Xi and Zhang (2012), the ANN test with PCA, that is an unsupervised regularization, exhibits good size under null hypothesis from observing the rows for AR, Linear $(\rho = 0)$, Linear $(\rho = 0.7)$. It also exhibits good power against a variety of nonlinear structures. In Figure 3.1 we plot the histograms of the test statistic

Table 3.1 Size and power of LWG, Lasso, PLS, and Pretest (with $q = 200$)

	PCA		Lasso		PLS		Pretest	
	5%	10%	5%	10%	5%	10%	5%	10%
AR	0.047	0.102	0.054	0.098	0.064	0.127	0.733	0.869
TAR	0.243	0.373	0.248	0.354	0.375	0.510	0.930	0.976
SGN	0.841	0.914	0.735	0.829	0.849	0.917	0.991	0.998
NAR	0.104	0.183	0.086	0.238	0.135	0.243	0.764	0.892
MRS	0.167	0.259	0.164	0.344	0.181	0.283	0.926	0.974
Linear ($\rho = 0$)	0.043	0.088	0.052	0.112	0.192	0.341	0.726	0.880
Linear ($\rho = 0.7$)	0.043	0.091	0.057	0.129	0.113	0.190	0.728	0.878
Cross product ($\rho = 0$)	0.075	0.126	0.216	0.364	0.370	0.517	0.806	0.919
Cross product ($\rho = 0.7$)	0.240	0.362	0.320	0.456	0.288	0.434	0.839	0.936
Squared ($\rho = 0$)	0.178	0.277	0.219	0.303	0.503	0.675	0.856	0.937
Squared ($\rho = 0.7$)	0.220	0.341	0.267	0.384	0.344	0.496	0.854	0.938

Notes: Sample size $n = 200$. $q = 200$. "Pretest" denotes "PCA-first-and-then-Pretest." $k = 20$ is used for the Pretest method

Table 3.2 Size and power of PCA-first-and-then-Lasso with $k = 100, 50, 10, 5$

	$k = 100$		$k = 50$		$k = 10$		$k = 5$	
	5%	10%	5%	10%	5%	10%	5%	10%
AR	0.085	0.158	0.078	0.142	0.048	0.087	0.041	0.080
TAR	0.126	0.204	0.125	0.206	0.146	0.222	0.135	0.212
SGN	0.204	0.262	0.226	0.287	0.352	0.401	0.628	0.700
NAR	0.089	0.161	0.096	0.165	0.064	0.110	0.124	0.226
MRS	0.190	0.267	0.186	0.280	0.136	0.215	0.135	0.203
Linear ($\rho = 0$)	0.096	0.183	0.067	0.121	0.052	0.086	0.047	0.108
Linear ($\rho = 0.7$)	0.097	0.178	0.065	0.117	0.045	0.080	0.046	0.094
Cross product ($\rho = 0$)	0.109	0.183	0.096	0.154	0.089	0.160	0.163	0.251
Cross product ($\rho = 0.7$)	0.114	0.199	0.100	0.172	0.092	0.161	0.216	0.328
Squared ($\rho = 0$)	0.108	0.187	0.078	0.168	0.148	0.227	0.203	0.309
Squared ($\rho = 0.7$)	0.110	0.196	0.082	0.139	0.134	0.204	0.227	0.352

Notes: Sample size is $n = 200$. $q = 200$

under the null hypothesis. The solid line is the probability density function of χ^2_3 distribution. In all three cases of AR and Linear, the finite sample distribution (histogram) of the test statistic is very close to its asymptotic χ^2_3 distribution, which means the unsupervised ANN test with PCA has good size not only in 5% and 10% levels but also across the entire distribution. This demonstrates that use of unsupervised regularization for the ANN test does not lead to the PoSI problem.

In contrast, it seems that the use of supervised regularization for the ANN test does lead to the PoSI problem to some different extent depending on different method. Looking at the size in Table 3.1, we may see only slight over-rejections

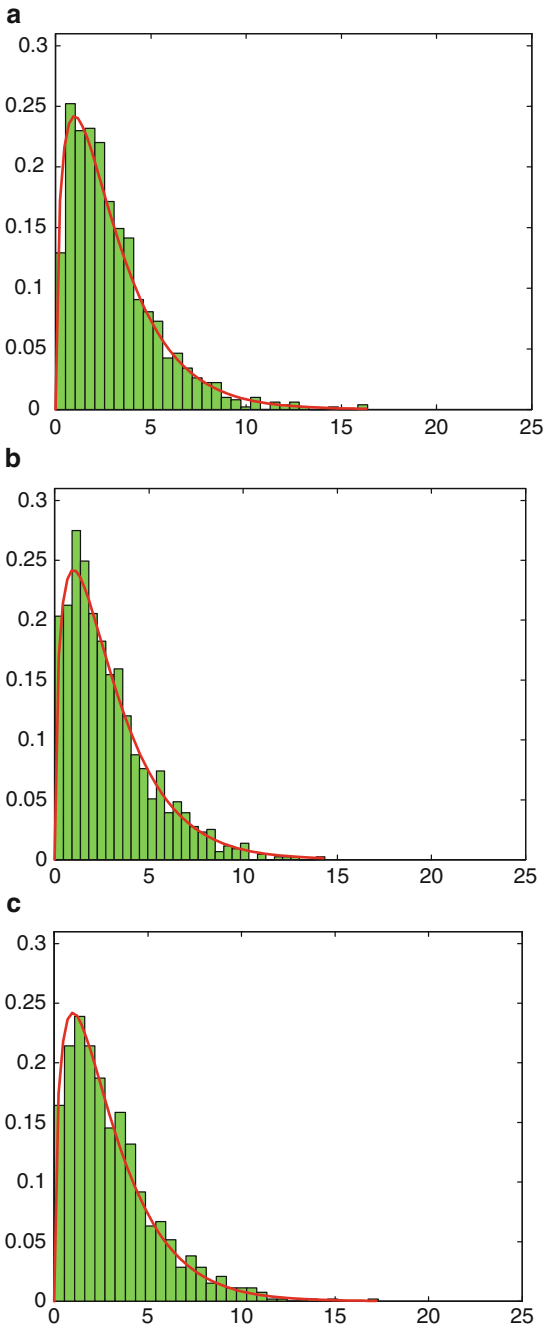

Fig. 3.1 Distribution of T_n^{PCA} under H_0. (**a**) DGP: AR compared to χ_3^2. (**b**) DGP: Linear ($\rho = 0$) compared to χ_3^2. (**c**) DGP: Linear ($\rho = 0.7$) compared to χ_3^2

at 10% level for Linear ($\rho = 0.7$).[3] While the power of the supervised ANN test using Lasso is quite similar to those of the unsupervised ANN test with PCA in Block 1, it is higher in Block 2. Because Table 3.1 presents only the 5% and 10% quantiles in the right tail (i.e., 95% and 90% quantiles) of the null distribution of the statistic, the results of the tables do not show the difference between PCA and Lasso. However, comparing Figures 3.1 and 3.2 for the entire distribution can tell some apparent difference especially in the left tail and to some lesser degree in the middle of the null distribution (but not in the right tail as shown in the tables). From Figure 3.2, we can see that the Lasso method suffers from the PoSI problem in the sense that the distributions of the test statistic diverge from the theoretical asymptotic χ^2_3 distribution. This can be more clearly seen in the AR case in Block 1. But for the cross-sectional cases in Block 2, the histograms of the test statistics are still close to the χ^2_3 distribution, although they are not as good as the ones in Figure 3.1. Hence, it seems that the PoSI problem is relative mild or even negligible with Lasso.

For the size of the supervised ANN test using PLS, we observe from Table 3.1 typical over-rejections at 5% and 10% levels in all three linear cases. This clearly shows that the PoSI problem is severe for the PLS supervision, which leads to power much higher than those of the unsupervised ANN test with PCA. In Figure 3.3, we can see the histograms of test statistics shift out of the χ^2_3 distribution, which again implies the PoSI problem.

For the PCA-first-and-then-Pretest method (in short, the Pretest method), the PoSI problem is most obvious. Table 3.1 shows the test results for $k = 20$, we can see that even the size under 5% and 10% is close to 1. We also tried different values of k, and the results are similar, so we do not report them in the table. Figure 3.4 shows the distribution of test statistic for Pretest method with $k = 20$, which diverge heavily from the χ^2_3 distribution. Finally, to show how different degrees of supervised regularization lead to different degrees of PoSI problem, we experiment the supervised ANN test using the PCA-first-and-then-Lasso with different values of k, the number of the principal components selected by PCA in the first step of the method. The PCA-first-and-then-Lasso method has two steps. The first step is to compute principal components of the $q = 200$ randomly activated neural network hidden units. Among them we select the first k principal components. Then in the second step we select $q^* = 3$ of the k principal components. We consider $k = 3, 5, 10, 20, 50, 100, 200$. When $k = q = 200$, this method is the same as Lasso (as presented in Figure 3.2), for which there is no role of the first-step in the PCA-first-and-then-Lasso as no principal components are used. When $k = q^* = 3$, this method is the same as PCA (as presented in Figure 3.1), for which there is no

[3] At 5% level, since the p-value is Bernoulli distributed with success probability of 0.05, the standard error of the p-value from the 1,000 Monte Carlo replication is $\sqrt{(0.05 \times 0.95)/1000} \approx 0.0069$. The 95% confidence interval is $0.05 \pm 1.96 \times 0.0069 = (0.0365, 0.0635)$. At 10% level, the standard error of the p-value is $\sqrt{(0.1 \times 0.9)/1000} = 0.0095$, and the 95 % confidence interval is $0.10 \pm 1.96 \times 0.0095 = (0.0814, 0.1186)$.

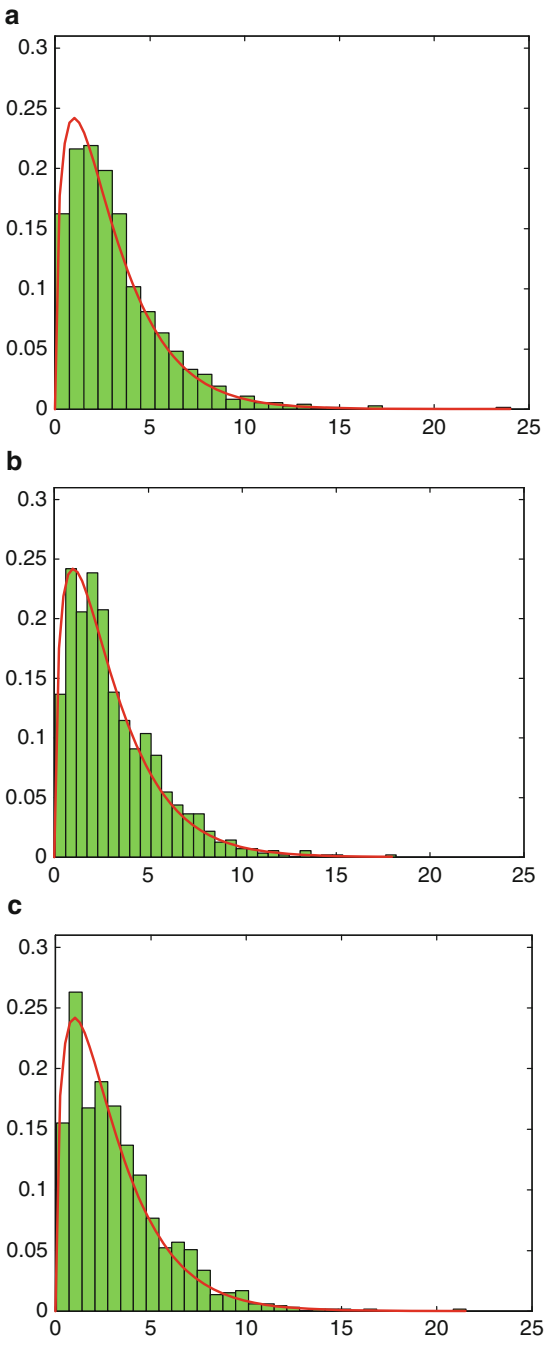

Fig. 3.2 Distribution of T_n^{Lasso} under H_0. (**a**) DGP: AR compared to χ_3^2. (**b**) DGP: Linear ($\rho = 0$) compared to χ_3^2. (**c**) DGP: Linear ($\rho = 0.7$) compared to χ_3^2

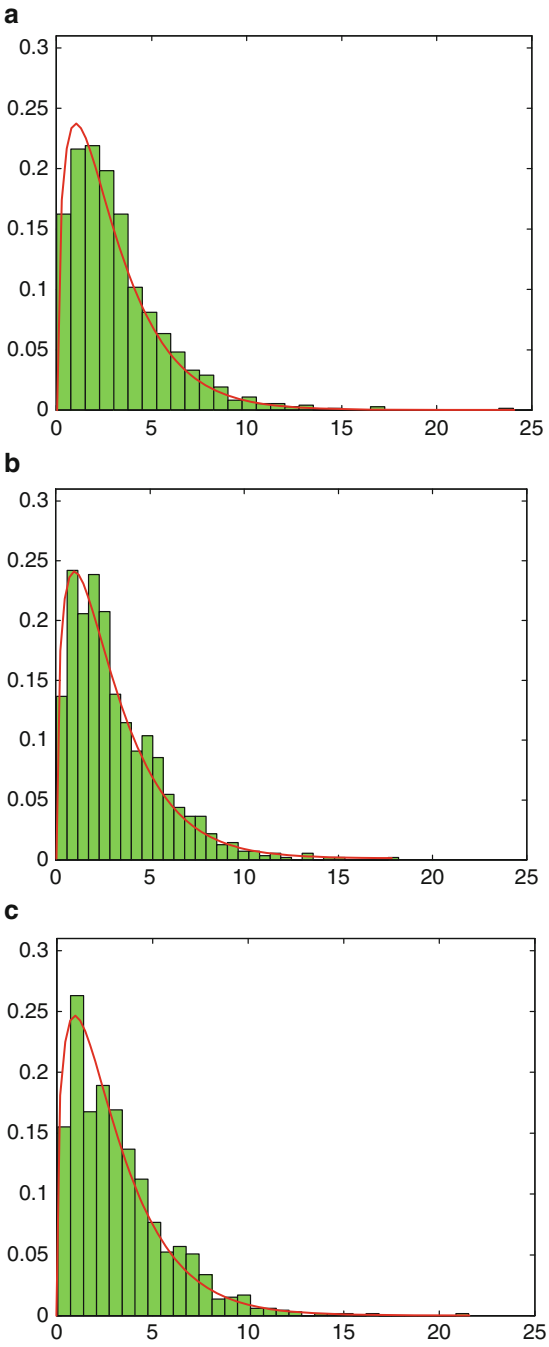

Fig. 3.3 Distribution of T_n^{PLS} under H_0. (**a**) DGP: AR compared to χ_3^2. (**b**) DGP: Linear ($\rho = 0$) compared to χ_3^2. (**c**) DGP: Linear ($\rho = 0.7$) compared to χ_3^2

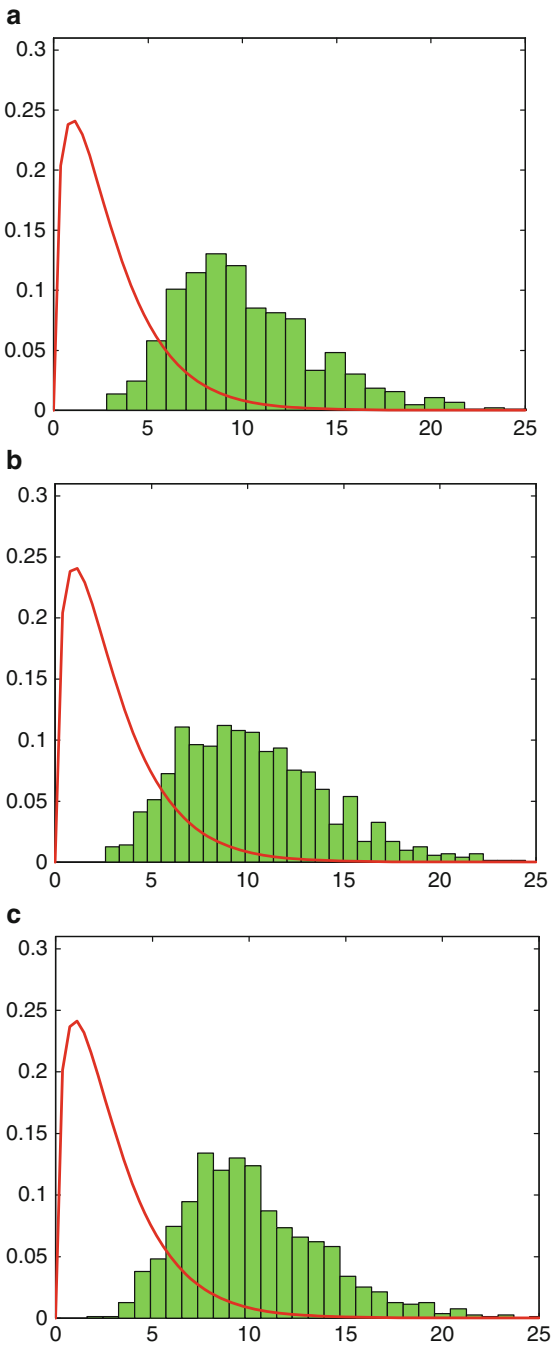

Fig. 3.4 Distribution of T_n^{Pretest} (PCA-first-and-then-Pretest) under H_0. (**a**) DGP: AR compared to χ^2_3. (**b**) DGP: Linear ($\rho = 0$) compared to χ^2_3. (**c**) DGP: Linear ($\rho = 0.7$) compared to χ^2_3

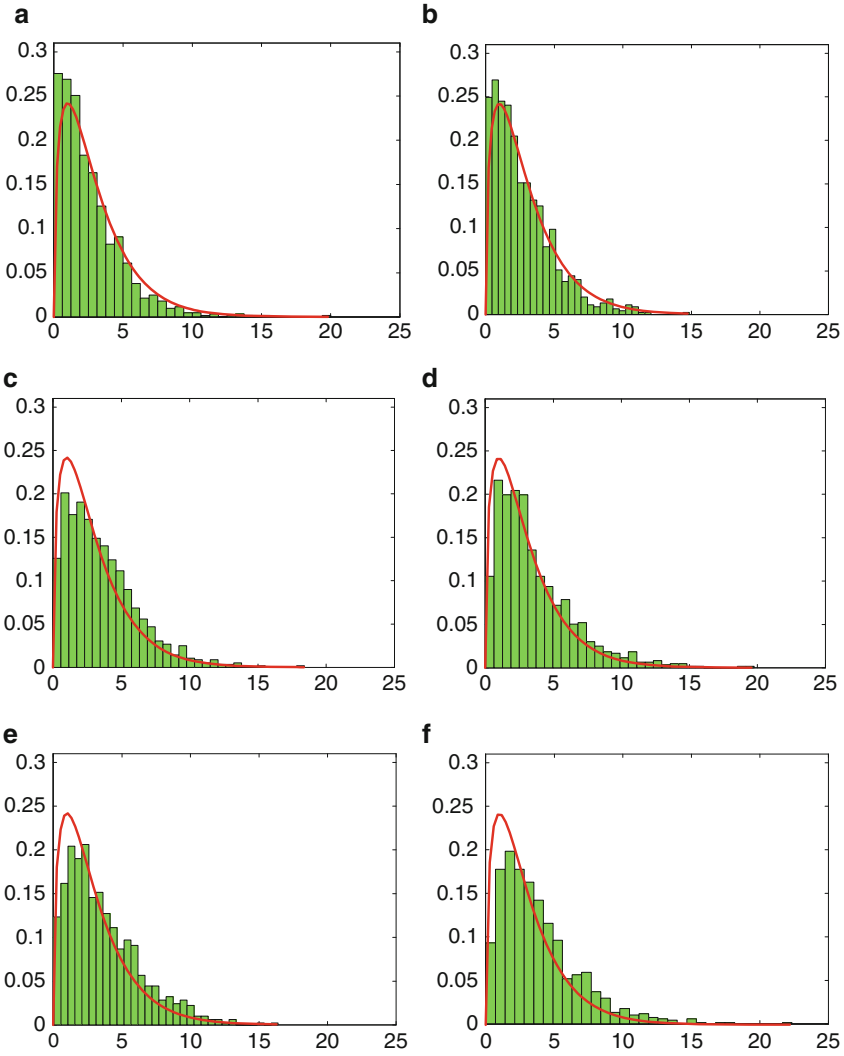

Fig. 3.5 Distribution of $T_n^{\text{PCA}-\text{Lasso}}$ (PCA-first-and-then-Lasso) under H_0. (**a**) DGP: AR compared to χ_3^2 ($k = 5$). (**b**) DGP: Linear ($\rho = 0.7$) compared to χ_3^2 ($k = 5$). (**c**) DGP: AR compared to χ_3^2 ($k = 50$). (**d**) DGP: Linear ($\rho = 0.7$) compared to χ_3^2 ($k = 50$). (**e**) DGP: AR compared to χ_3^2 ($k = 100$). (**f**) DGP: Linear ($\rho = 0.7$) compared to χ_3^2 ($k = 100$)

role of the second-step in the PCA-first-and-then-Lasso as no Lasso is used. If k is very small, for example $k = 5$ (as presented in Figure 3.5a, b), this method is similar to the unsupervised ANN test with PCA. In the other extreme, if k is very large, say $k = 100$ (as presented in Figure 3.5e, f), then the LASSO will play a very important role but PCA will have little effect on the test. Table 3.2 shows the

size and power of this method using different values of $k = 5, 10, 20, 100$. Let us first look at the size. The test behaves reasonably good when k is equal to 5 because when k is small, this test is close to the unsupervised ANN test with PCA and therefore suffers little from the PoSI problem. But when k increases to 50 and 100, we can see the over-rejection from the PoSI problem becomes more severe. The PoSI problem can be found in Figure 3.5, where we draw the histograms of test statistics for different k. For $k = 5$, the histograms are very close to the χ^2_3 distribution. But as k increases to 50 and 100, the histograms gradually shift to the right which indicates over-rejection.

When it comes to the power, the supervised ANN test using the PCA-first-and-then-Lasso method does very badly especially when k is large. Table 3.2 shows that the power for $k = 50$ and $k = 100$ are substantially lower than the power for $k = 5$ in all cases except for MRS. When comparing with the unsupervised ANN test with PCA, this test shows inferior power in most cases. The reason for this lowered power is ascribe to how the Lasso works. In the LWG test, we choose the second to the fourth principal components which account for a large fraction of the variance of Ψ_t, so that they contain a lot of information and therefore can help detect the nonlinearity. But the Lasso will keep principal components with larger coefficients in the regression. Hence those principal components with large coefficients but maybe with less information can be kept; those ones with small coefficients but maybe with more information are dropped. That may be why the PCA-first-and-then-Lasso method performs poorly in power. When we increase k, it is more likely that the Lasso may pick up unimportant principal components and will reduce the power even more. On the other hand, if we set $k = q^*$, the Lasso to PCA test is essentially the LWG's original ANN test, and this explains the increasing power when k is very small.

3.5 Conclusions

In this paper, we applied the ANN model to test neglected nonlinearity in conditional mean. The ANN test uses the residuals from a linear model and check for their correlation with the ANN's hidden unit activation functions. We generated a large number of activation functions based on the randomly drawn activation parameters. The large number of the activation functions is necessary to get good approximation of an unknown nonlinear functional form. Then in order to avoid the collinearity problem, we apply different regularization methods to select a moderate number of activation functions. One regularization method suggested by Lee et al. (1993) is the PCA, which is unsupervised. In this paper, we consider four supervised regularization methods to select a subset of many activation functions. We show that the use of supervised regularization such as Lasso, PLS, Pretest would lead to the PoSI problem, while the PCA does not lead to such problem.

A way of avoiding the PoSI problem is to conduct the simultaneous inference for all possible submodels under consideration which will make the resulting PoSI valid but conservative, by using a Bonferroni-type bound as used by Lee et al. (1993) for PCA. As Leeb and Pötscher (2008) noted, finding the distribution of post-selection estimates is hard and perhaps impossible. Pötscher and Leeb (2009) show that the distribution of regularized estimators by Lasso, SCAD, and Pretest is highly non-normal (non chi-squared in our testing setup of this paper). Nevertheless, a valid PoSI is possible via simultaneous inference as studied by Berk, Brown, Buja, Zhang and Zhao (2011). Whether/how the simultaneous inference may be applied for Lasso, Pretest, PLS requires further research.

We note that the PoSI "problem" (for inference) is not necessarily a problem (for forecasting). Knowing the PoSI problem could provide valuable information. The question is what for. The answer is that the PoSI problem can be a measure of the possible gain by supervision, and therefore it will be useful information for forecasting. The over-rejection in inference due to the PoSI problem of the various supervised regularization methods shows that the null distribution of the test statistic based on the regularized (selected) randomized ANN activations can be shifted towards the right tail, especially when the Pretest method is in use. While it is a serious problem in inference, it may be a valuable information for forecasting. The degree of the PoSI problem can be translated into a measure of supervision in the regularization, i.e., a measure of the information contents for the forecast target from the variables (predictors) selected through the supervision. However, the results from Table 3.2 for the PCA-first-and-then-Lasso method indicates that this may not be a straightforward matter because it is shown that more supervision does not necessarily increase the power of the ANN test. It remains to be studied that it might be possible that the more supervised regularization can lead to poor forecasting performance of the ANN model. Hence, it will be interesting to examine whether the different degrees of the PoSI problem among the different regularization methods may be carried over to different degrees of improvement in forecasting ability of the ANN model. We leave this in our research agenda.

References

Aït-Sahalia, Y., P. J. Bickel and T. M. Stoker (2001), "Goodness-of-fit Tests for Kernel Regression with an Application to Option Implied Volatilities," *Journal of Econometrics* 105(2), 363–412.

Andrews, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica* 59(3), 817–858.

Andrews, D. W. K. and W. Ploberger (1994), "Optimal Tests when a Nuisance Parameter is Present Only under the Alternative," *Econometrica* 62, 1383–1414.

Bai, J. and Ng, S. (2008), "Forecasting Economic Time Series Using Targeted Predictors", *Journal of Econometrics* 146, 304–317.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006), "Prediction by Supervised Principal Components," *Journal of the American Statistical Association* 101(473), 119–137.

Berk, R., L. Brown, A. Buja, K. Zhang and L. Zhao (2011), "Valid Post-Selection Inference," The Wharton School, University of Pennsylvania, Working Paper. Submitted to *the Annals of Statistics*.

Bierens, H. J. (1982), "Consistent Model Specification Tests," *Journal of Ecomometrics* 20, 105–134.

Bierens, H. J. (1990), "A Consistent Conditional Moment Test of Functional Form," *Econometrica* 58, 1443–1458.

Bierens, H. J. and W. Ploberger (1997), "Asymptotic Theory of Integrated Conditional Moment Tests," *Econometrica* 65, 1129–1151.

Blake, A. P. and G. Kapetanios (2000), "A Radial Basis Function Artificial Neural Network Test for ARCH," *Economics Letters* 69, 15–23.

Blake, A. P. and G. Kapetanios (2003), "A Radial Based Function Artificial Neural Network Test for Neglected Nonlinearity," *Econometrics Journal* 6, 356–372.

Bradley, E., H., Trevor j. Iain and T., Robert (2004), *Annals of Statistics* 32(2), 407–499.

Bradley, R. and R. McClelland (1996), "A Kernel Test for Neglected Nonlinearity," *Studies in Nonlinear Dynamics and Econometrics* 1(2), 119–130.

Cai, Z., J. Fan and Q. Yao (2000), "Functional-Coefficient Regression Models for Nonlinear Time Series," *Journal of the American Statistical Association* 95(451), 941–956.

Campbell, J. Y., A. W. Lo and A. C. Mackinlay (1997), *The Economometrics of Financial Markets,* Princeton University Press.

Corradi, V. and N. R. Swanson (2002), "A Consistent Test for Nonlinear Out of Sample Predictive Accuracy," *Journal of Econometrics* 110, 353–381.

Dahl, C. M. (2002), "An Investigation of Tests for Linearity and the Accuracy of Likelihood Based Inference using Random Fields," *Econometrics Journal* 1, 1–25.

Dahl, C. M. and G. Gonzalez-Rivera (2003), "Testing for Neglected Nonlinearity in Regression Models: A Collection of New Tests," *Journal of Econometrics* 114(1), 141–164.

Davies, R. B. (1977), "Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative," *Biometrika* 64, 247–254.

Davies, R. B. (1987), "Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative," *Biometrika* 74, 33–43.

Eubank, R. L. and C. H. Spiegelman (1990), "Testing the Goodness of Fit of a Linear Model Via Nonparametric Regression Techniques," *Journal of the American Statistical Association* 85(410), 387–392.

Fan, Y. and Q, Li (1996), "Consistent Model Specification Tests: Omitted Variables and Semi-parametric Functional Forms," *Econometrica* 64, 865–890.

Fan, J. and J, Lv (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society Series B* 70, 849–911.

Fan, J. and R. Li (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association* 96, 1348–1360.

Frank, I.E., and J.H. Friedman (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–148.

Granger, C. W. J. and T. Teräsvirta (1993), *Modelling Nonlinear Economic Relationships,* Oxford University Press: New York.

Hamilton, J. D. (2001), "A Parametric Approach to Flexible Nonlinear Inference," *Econometrica* 69(3): 537–573.

Hansen, B. E. (1996), "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis," *Ecomometrica* 64: 413–430.

Härdle, W. and E. Mammen (1993), "Comparing Nonparametric versus Parametric Regression Fits," *Annals of Statistics* 21: 1926–1947.

Hillebrand, E., H. Huang, T.-H. Lee, and Canlin Li (2011), "Using the Yield Curve in Forecasting Output Growth and Inflation", Manuscript. Aarhus University and UC Riverside.

Hjellvik, V. and D. Tjøstheim (1995), "Nonparametric Tests of Linearity for Time Series," *Biometrika* 82(2), 351–368.

Hjellvik, V. and D. Tjøstheim (1996), "Nonparametric Statistics for Testing of Linearity and Serial Independence," *Journal of Nonparametric Statistics* 6, 223–251.

Hjellvik, V., Q. Yao and D. Tjøstheim (1998), "Linearity Testing Using Local Polynomial Approximation," *Journal of Statistical Planning and Inference* 68, 295–321.

Hong, Y. and H. White (1995), "Consistent Specification Testing via Nonparametric Series Regression," *Econometrica* 63, 1133–1160.

Hoover, K.D. (2012), "The Role of Hypothesis Testing in the Molding of Econometric Models," Duke University, Center for the History of Political Economy (CHOPE) Working Paper No. 2012–03.

Hornik, K., M. Stinchcombe, and H. White (1989), "Multi-Layer Feedforward Networks Are Universal Approximators," *Neural Network* 2, 359–366.

Huang, J., J. Horowitz and Ma, S. (2008), "Asymptotic Properties of Bridge Estimators in Sparse High-dimensional Regression Models," *Annals of Statistics* 36, 587–613.

Huang, H. and T.-H. Lee (2010), "To Combine Forecasts or To Combine Information?" *Econometric Reviews* 29, 534–570.

Inoue, A. and Kilian, L. (2008), "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation," *Journal of the American Statistical Association* 103(482), 511–522.

Kock, A. B. (2011), "Forecasting with Universal Approximators and a Learning Algorithm,"*Journal of Time Series Econometrics* 3, 1–30.

Kock, A. B. and T. Teräsvirta (2011), "Forecasting Macroeconomic Variables using Neural Network Models and Three Automated Model Selection Techniques," CREATES Research Paper 27.

Lee, T.-H. (2001), "Neural Network Test and Nonparametric Kernel Test for Neglected Nonlinearity in Regression Models," *Studies in Nonlinear Dynamics and Econometrics* 4(4), 169–182.

Lee, T.-H. and A. Ullah (2001), "Nonparametric Bootstrap Tests for Neglected Nonlinearity in Time Series Regression Models," *Journal of Nonparametric Statistics* 13, 425–451.

Lee, T.-H. and A. Ullah (2003), "Nonparametric Bootstrap Specification Testing in Econometric Model," *Computer-Aided Econometrics*, Chapter 15, edited by David Giles, Marcel Dekker, New York, pp. 451–477.

Lee, T.-H., H. White and C. W. J. Granger (1993), "Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests," *Journal of Econometrics* 56, 269–290.

Lee, T.-H., Z. Xi and R. Zhang (2012), "Testing for Neglected Nonlinearity Using Artificial Neural Networks with Many Random Hidden Unit Activations," UCR, manuscript.

Leeb, H. and B.M. Pötscher (2003), "The finite-sample distribution of post model-selection estimators and uniform versus nonuniform approximations," *Econometric Theory* 19, 100–142.

Leeb, H. and B.M. Pötscher (2005), "Model selection and inference: Facts and fiction," *Econometric Theory* 21, 21–59.

Leeb, H. and B.M. Pötscher (2006), "Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results," *Econometric Theory* 22, 21–59.

Leeb, H. and B.M. Pötscher (2008), "Can One Estimate The Unconditional Distribution Of Post-Model-Selection Estimators?," *Econometric Theory* 24(2), 338–376.

Li, Q. and S. Wang (1998), "A Simple Consistent Bootstrap Test for a Parametric Regression Function," *Journal of Econometrics* 87, 145–165.

Liu, R. Y. (1988), "Bootstrap Procedures under Some Non-iid Models," *Annals of Statistics* 16: 1697–1708.

Matsuda, Y. (1999), "A Test of Linearity Against Functional-Coefficient Autoregressive Models," *Communications in Statistics, Theory and Method* 28(11), 2539–2551.

Newey, W. K. (1985), "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica* 53, 1047–1070.

Poggi, J. M. and B. Portier (1997), "A Test of Linearity for Functional Autoregressive Models," *Journal of Time Series Analysis* 18(6), 615–639.

Pötscher, B.M. and H. Leeb, (2009), "On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding," *Journal of Multivariate Analysis* 100(9), 2065–2082.

Stengos, T. and Y. Sun (2001), "Consistent Model Specification Test for a Regression Function Based on Nonparametric Wavelet Estimation," *Econometric Reviews* 20(1), 41–60.

Stinchcombe, M. and H. White (1998), "Consistent Specification Testing with Nuisance Parameters Present only under the Alternative," *Econometric Theory* 14, 295–325.

Tauchen, G. (1985), "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics* 30, 415–443.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of Royal Statistical Society, B,* 58, 267–288.

Teräsvirta, T. (1996), "Power Properties of Linearity Tests for Time Series," *Studies in Nonlinear Dynamics and Econometrics* 1(1), 3–10.

Teräsvirta, Timo, C.-F. Lin and C. W. J. Granger (1993), "Power of the Neural Network Linearity Test," *Journal of Time Series Analysis* 14(2), 209–220.

Tjøstheim, D. (1999), "Nonparametric Specification Procedures for Time Series," in S. Ghosh (ed.), *Asymptotics, Nonparametrics, and Time Series: A Tribute to M.L. Puri*, Marcel Dekker.

Ullah, A. (1985), "Specification Analysis of Econometric Models," *Journal of Quantitative Economics* 1: 187–209.

Whang, Y.-J. (2000), "Consistent Bootstrap Tests of Parametric Regression Functions," *Journal of Econometrics* 98, 27–46.

White, H. (1980), "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48, 817–838.

White, H. (1987), "Specification Testing in Dynamic Models," T.F. Bewley (ed.), *Advances in Econometrics, Fifth World Congress*, Vol 1, New York: Cambridge University Press, 1–58.

White, H. (1989), "An Additional Hidden Unit Tests for Neglected Nonlinearity in Multilayer Feedforward Networks," *Proceedings of the International Joint Conference on Neural Networks*, Washington, DC. (IEEE Press, New York, NY), II: 451–455.

White, H. (1994), *Estimation, Inference, and Specification Analysis*, Cambridge University Press.

White, H. (2006), "Approximate Nonlinear Forecasting Methods," *Handbook of Economic Forecasting* 1, 459–512.

Wu, C. F. J. (1986), "Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis," *Annals of Statistics* 14: 1261–1350.

Zheng, J. X. (1996), "A Consistent Test of Functional Form via Nonparametric Estimation Techniques," *Journal of Econometrics* 75, 263–289.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association* 101(476), 1418–1429.

Zou, H. and H. Zhang (2009), "On The Adaptive Elastic-Net with a Diverging Number of Parameters, " *The Annals of Statistics* 37(4), 1773–1751.