Testing for Neglected Nonlinearity Using Artificial Neural Networks with Many Randomized Hidden Unit Activations*

Tae-Hwy Lee†, Zhou Xi‡, and Ru Zhang§
Department of Economics
University of California, Riverside

August 2012

Abstract

This paper makes a simple but previously neglected point with regard to an empirical application of the test of White (1989) and Lee, White and Granger (LWG, 1993), for neglected nonlinearity in conditional mean, using the feedforward single layer artificial neural network (ANN). Because the activation parameters in the hidden layer are not identified under the null hypothesis of linearity, LWG suggested to activate the ANN hidden units based on the randomly generated activation parameters. Their Monte Carlo experiments demonstrated the excellence performance (good size and power), even if LWG considered a fairly small number (10 or 20) of random hidden unit activations. However, in this paper we note that the good size and power of Monte Carlo experiments are the average frequencies of rejecting the null hypothsis over multiple replications of the data generating process. The average over many simulations in Monte Carlo smooths out the randomness of the activations. In an empirical study, unlike in a Monte Carlo study, multiple realizations of the data are not possible or available. In this case, the ANN test is sensitive to the randomly generated activation parameters. One solution is the use of Bonferroni bounds as suggested in LWG (1993), which however still exhibit some excessive dependence on the random activations (as shown in this paper). Another solution can be to integrate the test statistic over the nuisance parameter space, for which however, bootstrap or simulation should be used to obtain the null distribution of the integrated statistic. In this paper, we consider a much simpler solution that is shown to work very well. That is, we simply increase the number of randomized hidden unit activations to a (very) large number (e.g., 1000). We show that using *many* randomly generated activation parameters can robustify the performance of the ANN test when it is applied to a real empirical data. This robustification is reliable and useful in practice, and can be achieved at no cost as increasing the number of random activations is almost costless given today's computer technology.

Keywords: Many Activations. Randomized Nuisance Parameters. Boferroni Bounds. Principal Components.

JEL Classification Codes: C1, C4, C5

*This paper is dedicated to an extraordinary scholar and teacher, Professor Halbert White, who created the original idea of this paper. We thank a referee for many useful comments.

†Corresponding author. Department of Economics, University of California, Riverside, CA 92521, USA. E-mail: tae.lee@ucr.edu

‡Department of Economics, University of California, Riverside, CA 92521, USA. E-mail: zhou.xi@email.ucr.edu.

§Department of Economics, University of California, Riverside, CA 92521, USA. E-mail: ru.zhang@email.ucr.edu.

1 Introduction

This paper revisits the test of White (1989) and Lee, White and Granger (LWG, 1993), for neglected nonlinearity in conditional mean using the feed-forward single layer artificial neural network (ANN). The advantage to use ANN model to test nonlinearity is that the ANN model inherits the flexibility as a universal approximator of unknown functional form. The ANN test is designed to use the predictive ability of the ANN hidden layer activations, which may be neglected in linear models. Because the estimation of the ANN model is often difficult and the activation parameters in the hidden layer are not identified under the null hypothesis of linearity, LWG suggested to activate the ANN hidden units based on the randomly generated neural network activation parameters. LWG considered only a *small* set of random activation parameters (limited by the computing power two decades ago). Nevertheless, their Monte Carlo experiment demonstrated the excellent performance of the ANN test in size and power. The ANN test has been cited in numerous papers as a benchmark method in the literature on testing neglected nonlinearity.

However, in this paper, we note that the size and power of any Monte Carlo experiments are the empirical *average* frequencies of rejecting the null hypothesis, when the null hypothesis is true (size) or when the null is not true (power), over many Monte Carlo replications of the data generating process (DGP). Unlike in a Monte Carlo study where the data are replicated multiple times, an empirical study has only one realized sample. When the ANN test is applied to one realized sample, its performance is largely affected by the randomly generated activation parameters. Applying the test to a particular real data amounts to one single Monte Carlo replication. In this paper we show that a small set of random activation parameters will make the performance of the ANN test quite random. This was not noticed in LWG (1993) and any other papers that have studied the ANN test, perhaps because most of these studies compare the performance in Monte Carlo where the performance is measured in average rejection over many replications. We will show that, when a real data is tested by the ANN test, a small number of random activations makes the ANN test quite unstable and sensitive to the random activations. Interestingly, however, we will also show that increasing the number of the randomly generated activation parameters can robustify the performance of the ANN test when it is applied to a single real data set. This robustification is important and useful in practice, which can be achieved at no cost as increasing the number of random activations is almost costless given the computer technology available today.

The rest of the paper is organized as follows. Section 2 reviews the ANN

test with randomized hidden unit activations. In Section 3, we examine the ANN test with Monte Carlo to confirm the LWG's results on the excellent size and power of the randomly activated ANN test. In Section 4, for each simulated series, we point out a problem of the randomized ANN test when the number of randomized activations is small and then show that this problem can be easily resolved by simply increasing it to a very large number of randomized activations. In Section 5 we repeat what we have done in Section 4 using actual economic data. Section 6 concludes.

2 The ANN Test

The linear-augmented single hidden-layer feedforward ANN model has the following architecture:

$$y_t = f(\mathbf{x}_t, \theta) + \varepsilon_t := \mathbf{x}_t'\alpha + \sum_{j=1}^{q} \beta_j \psi\left(\mathbf{x}_t'\gamma_j\right) + \varepsilon_t, \tag{1}$$

where $t = 1, \ldots, n$, $\mathbf{x}_t = (x_{1t}, \ldots x_{k,t})'$, $\theta = \left(\alpha', \beta', \gamma_1', \ldots, \gamma_q'\right)'$, $\alpha = (\alpha_1, \ldots, \alpha_k)'$, $\beta = (\beta_1, \ldots, \beta_q)'$, and $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{j,k})'$ for $j = 1, \ldots, q$, and $\psi(\cdot)$ is an activation function.[1] An example of the activation function is the logistic function $\psi(z) = (1 + \exp(z))^{-1}$. α is a column vector of connection strength from the input layer to the output layer; γ_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 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.

Hornik, Stinchcombe and White (1989, 1990) show that neural network model in (1) is a nonlinear flexible functional form being capable of representing arbitrarily accurate approximations to any mappings. White (1990) and White and Wooldridge (1991) show that these approximations are learnable (i.e., consistently estimable) by proper control of the growth of network complexity q as network experience accumulates (i.e., the sample size n grows). While they give theoretical results for controlling the growth rate of q as a function of n, the proper rate depends critically on the dependence properties of $(y_t \ \mathbf{x}_t')'$ which makes a choice of the growth rate for network complexity not immediately

[1] '$a := b$' means that a is defined by b, while '$a =: b$' means that b is defined by a.

obvious. As a referee pointed out it would be desirable if we could give some guidance on the adaptive choice of q. Unfortunately, to date there is no unified theory on this rate. See Chen (2007, p. 5575) for more discussion. While White (1990, p. 538) gave some guidelines on the choice of q for different n and different dependence properties, he recommended to use cross-validation to choose q in estimating ANN models in practice. This paper, however, deals with testing for neglected nonlinearity in a linear model without having to estimate the nonlinear ANN model. The main purpose of this paper is about the choice of q in using the ANN model for testing whether β_j's are all zero. As we do this with the randomization of γ_j's and then we take a small number of their principal components, a very large q, even larger than n, may be used and may be more desirable as examined in Sections 4 and 5.

To test whether the process y_t is linear in mean conditional on $\mathbf{x}_t$, we consider the following null and alternative hypotheses

$$H_0 : \Pr[\mathrm{E}(y_t|\mathbf{x}_t) = \mathbf{x}_t'\alpha^*] = 1 \quad \text{for some } \alpha^* \in \mathbb{R}^k$$
$$H_1 : \Pr[\mathrm{E}(y_t|\mathbf{x}_t) = \mathbf{x}_t'\alpha] < 1 \quad \text{for all } \alpha \in \mathbb{R}^k$$

When the null hypothesis is rejected, a linear model is said to suffer from neglected nonlinearity. White (1989) and LWG (1993) developed a test for neglected nonlinearity likely to have power against a range of alternatives based on ANN models. See also Teräsvirta *et al* (1993) and Teräsvirta (1996) on the neural network test and its comparison with other specification tests. The neural network test is based on the activations of 'phantom' hidden units $\psi(\mathbf{x}_t'\gamma_j),\ j = 1, \ldots, q$. That is,

$$H_0 : E[\psi(\mathbf{x}_t'\gamma_j)\,\varepsilon_t] = 0, \qquad j = 1, \ldots, q, \tag{2}$$

or

$$E(\Psi_t\varepsilon_t) = 0, \tag{3}$$

where $\Psi_t := (\psi(\mathbf{x}_t'\gamma_1), \ldots, \psi(\mathbf{x}_t'\gamma_q))'$ is a phantom hidden unit activation vector and ε_t is the error term from the two layer *affine* network $y_t = \mathbf{x}_t'\alpha + \varepsilon_t$ (with $q = 0$). Evidence of correlation of ε_t with Ψ_t is evidence against the null hypothesis that y_t is linear in conditional mean. If correlation exists, augmenting the linear network by including an additional hidden unit with activations Ψ_t would permit an improvement in network performance. Thus the tests are based on the sample correlation of affine network errors with phantom hidden unit activations,

$$n^{-1}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t = n^{-1}\sum_{t=1}^{n}\Psi_t(y_t - \mathbf{x}_t'\hat{\alpha}), \tag{4}$$

where $\hat{\alpha}$ is least squares estimator of α. Under suitable regularity conditions it follows from a 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-squared statistic can be formed as

$$\left(n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t\right)'\hat{W}_n^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t\right)\xrightarrow{d}\chi_q^2. \tag{5}$$

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 general approximator. To alleviate the estimation errors of the ANN, it is useful to note that, for given values of γ_j's, the ANN is linear in $\mathbf{x}$ and the activation functions Ψ and therefore (α',β') can be estimated from the linear regression once $(\gamma_1,\ldots,\gamma_q)$ have been given. The LWG's (1993) approach is to use a set of randomly generated $(\gamma_1,\ldots,\gamma_q)$. The additional *hidden* unit activation functions $\Psi_t(\gamma_1,\ldots,\gamma_q)$ are hidden (or phantom) because they do not exist under the null hypothesis. The $(\gamma_1,\ldots,\gamma_q)$ are randomly generated in testing because they are nuisance parameters not identified under the null hypothesis.

This approach is shown to have excellent size and power properties from Monte Carlo simulation and has been used in many subsequent nonlinear testing papers as a benchmark method in comparison. However, it is not noted in the literature that the LWG's excellent performance even with a small number $(q=10,20)$ of the randomized phantom activations is in terms of the Monte Carlo size and power. The good size and power in Monte Carlo experiments are the average frequencies of rejecting the null hypothesis over multiple replications of the data generating process (DGP). The averaging in Monte Carlo smooths out the randomness of the test result in each replication. However, in an empirical application, unlike in a Monte Carlo study, multiple realizations of the data are not possible or available. In this case, the ANN test is sensitive to the randomly generated activation parameters and its performance is generally unstable. When applying to real data, this randomness problem resulted from using different sets of randomized activation parameters $(\gamma_1,\ldots,\gamma_q)$ may lead to inconsistent conclusions.

One solution is the use of Bonferroni bounds of the p-values of the test statistics that are computed from m randomizations of the activation parameters $\left(\gamma_1^{(i)},\ldots,\gamma_q^{(i)}\right)_{i=1}^{m}$, as suggested in LWG (1993). However, the Bonferroni bounds still exhibit dependence on the randomized activations when q is small (as shown later in Table 3 of Section 5).

Another solution is to integrate the test statistic over the nuisance parameter space of $(\gamma_1, \ldots, \gamma_q)$. However, this approach requires bootstrap or simulation to obtain the null distribution of the integrated statistic (more on this in Section 4).

In this paper, we show a much simpler solution. That is to increase the number of randomized hidden unit activations to a (very) large number (e.g., 1000). We show that 'many' randomly generated activation parameters can robustify the performance of the ANN test when it is applied to a real empirical data. It also makes the Bonferroni bounds tighter (as shown in Section 5). We will demonstrate this in the remaining sections of the paper in Monte Carlo and in empirical applications. While this proposal may sound trivial, no previous papers have noted this problem. It is partly because all studies were able to show the excellent performance via Monte Carlo simulations with a small q and also because it was difficult to compute the singular value decomposition of a $q \times q$ matrix for a large q (to compute the principal components). It was 1989 when LWG (1993) conducted their Monte Carlo on an IBM 286 PC. The set of randomly selected parameters $(\gamma_1, \ldots, \gamma_q)$ should be large enough so that it can be dense and make the ANN an universal approximator. A large set of γ's (i.e., large q) enables $\sum_{j=1}^{q} \beta_j \psi(\mathbf{x}_t'\gamma_j)$ to capture the maximal nonlinear structure. We will show that the proposal of increasing q in fact provides a practically useful, powerful, and cheap solution to the randomness of random activations. The robustification is stable and reliable, and thus enables the ANN test to be employed in autopilot in its applications.

A large number q of random activation parameters $(\gamma_1, \ldots, \gamma_q)$ will make the activation functions $\psi(\mathbf{x}_t'\gamma_j)$ collinear with each other over $j = 1, \ldots, q$ and with $\mathbf{x}_t$. Thus LWG (1993) conducted a test on $q^* < q$ principal components of Ψ_t not collinear with $\mathbf{x}_t$, denoted Ψ_t^*. The key to the success with the large number of randomized network activations is the regularization of the network performance by principal components for dimensionality reduction. The ANN test takes two steps, randomization and regularization.

Then LWG employed the asymptotically equivalent test statistic (under conditional homoskedasticity) which avoids explicit computation of $\hat{W}_n$

$$T_n(q, q^* \mid \gamma_1, \ldots, \gamma_q) := nR^2 \xrightarrow{d} \chi^2_{q^*}, \tag{6}$$

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. In this paper, while we consider two values of q (small and very large), we fix $q^* = 3$ to simplify our presentation. Different values of q^* do not affect

the conclusions of this paper. Therefore the test statistic will be henceforth denoted as $T_n(q, 3 \mid \gamma_1, \ldots, \gamma_q) =: T_n(q \mid \gamma_1, \ldots, \gamma_q)$ or simply $T_n(q)$.

In Section 3, we conduct a Monte Carlo to show the ANN test has good size and power even with a small $q = 20$. The size and power from Monte Carlo do not tell the problem discussed above from using a small q. To see the problem, we conduct a different Monte Carlo, in Section 4. Only one realization (to mimic an empirical study) of $\{y_t\}_{t=1}^{n=200}$ which is linear in mean is generated, for which the ANN statistic $T_n\left(q \mid \gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)$ and its p-value P_i are computed from m different randomly generated activation parameters $\left(\gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)_{i=1}^{m}$. We show that the ANN statistic with a small number ($q = 20$) of randomized phantom activations exhibits large variation over $i = 1, \ldots, m$, while it becomes stable with a very large number ($q = 1000$) of randomized phantom activations. Hence, we can improve and robustify the ANN test by simply increasing q (say, from 20 to 1000). Section 5 demonstrates this with the five US monthly economics time series. In practice, we suggest to choose q as large as possible provided the computational ability permits. This is because a larger q will stabilize the p-values. Since we take the principle components of the activation functions, we can allow q to be even larger than the number of observations n. In our simulations and empirical experiments, for a moderately large data (with n around 200), choosing q to be 1000 leads to good results.

3 Small q vs. Large q in Monte Carlo Size and Power

The purpose of this section is to confirm the result of LWG (1993) that Monte Carlo studies will show excellent performance of the ANN test in terms of size and power, computed from 1000 Monte Carlo replications. To generate data we use the following DGPs, all of which have been used in the related literature. Two blocks of DGP are considered in this section: the first block has DGPs using the univariate autoregressive time series of y_t with one lagged endogenous input y_{t-1}; the second block includes cross-sectional networks with two exogenous inputs x_{1t} and x_{2t} which follow a bivariate normal distribution. To see the sensitivity of the test statistic under conditional heteroskedasticity, we also consider ARCH(1) and GARCH(1,1) processes for AR in Block 1. 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. All the error terms ε_t below are i.i.d. $N(0, 4)$. $1(\cdot)$ is an

indicator function which takes one if its argument is true and zero otherwise. The index $t = 1, \ldots, n$ with $n = 200$ being the sample size.

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 $\text{sgn}(y_{t-1}) = 1\left(y_{t-1} > 0\right) - 1\left(y_{t-1} < 0\right).$

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):

Assume x_{1t}, x_{2t} follow a bivariate normal distribution of $N\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\right)$ with $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1$, and $\rho = 0$ or 0.7. We have the following three cases:

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

2. Interaction

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

3. Squared

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

In the simulations of the ANN test, LWG chose q equal to 10 or 20 and q^* equal to 2 or 3 in different DGPs, and the sample size of 50, 100, and 200. Moreover, they dropped the first largest principle component of $\psi(\mathbf{x}_t'\gamma_j)$ to avoid the multicollinearity problem. In our paper, for the simulation results, we tried both the case with dropping the first principle component and without dropping the first principle component, the results were similar. So we keep the original LWG method to drop the first principal component for the LWG test in this paper. The information set is $\mathbf{x}_t = y_{t-1}$ for Block 1 and $\mathbf{x}_t = (x_{t1}$ $x_{t2})'$ for Block 2.

In practice, we need to generate γ's carefully so that $\mathbf{x}_t'\gamma_j$ is within a suitable range. If γ's are chosen to be too small then activation functions ψ's are approximately linear in $\mathbf{x}$, and we want to avoid this situation since they can not capture much nonlinearity; if γ's are too large the activation functions ψ's will take values close to 0 or 1 (their minimum or maximum values), and we want to avoid this situation as well. The logistic squasher $\psi(\mathbf{x}'\gamma_j) = [1 + \exp(-\mathbf{x}'\gamma_j)]^{-1}$ is used with γ_j being generated from the uniform distribution on $[-2, 2]$ and y_t, $\mathbf{x}_t$ being rescaled onto $[0,\ 1]$.

Bierens (1990) suggested an alternative randomization method for obtaining a χ^2 limiting distribution. Following theorem 4 in Bierens (1990) and applying to our context, suppose γ_0 is a point in the q-dimension Γ space. Let $\hat{\gamma} = \arg\max_{\gamma\in\Gamma} \hat{T}(\gamma)$, where $\hat{T}(\gamma)$ is a consistent estimator of the statistic in equation (5). For some real numbers $\lambda > 0$ and $\rho \in (0, 1)$, let $\tilde{\gamma} = \gamma_0$ if $\hat{T}(\hat{\gamma}) - \hat{T}(\gamma_0) \leq \lambda n^{\rho}$, otherwise $\tilde{\gamma} = \hat{\gamma}$. Then under H_0, $\hat{T}(\tilde{\gamma})$ has a χ^2 distribution. However, this result has some drawbacks. Firstly, the choice of $\tilde{\gamma}$ may be sensitive to the real numbers λ and ρ. Secondly and more importantly, the choice of $\tilde{\gamma}$ depends on a q-dimensional maximization problem. If we choose q to be too small, say 3, then the activation functions may not perform well as a universal approximator. If we choose q to be moderately large like 10, then it will be very difficult to find the global maximum. Although theorem 5 in Bierens (1990) is more practical, it still requires the chosen sequence to be dense in the Γ space and the required number of γ's in the chosen sequence will explode exponentially as q increases. This motivates us to use the principle components of the activation functions rather than the activation functions

themselves in our statistics, and just simply generating a large number of γ's randomly from uniform distribution.

In generating γ_j randomly from the uniform distribution on $[-2, 2]$, we did it in two different ways in our Monte Carlo experiment, namely by newly generating them for each replication or by fixing one same set of randomly generated γ_j for all replications. To compare the test results using randomized and fixed hidden units across replications, we report the Monte Carlo results using two methods to generate the γ's in Panels A and B of Table 1.

Table 1A reports the size and power for the ANN test with $q = 20$ and $q = 1000$ using uniformly randomized generated hidden units across replications. The numbers in the tables are the rejection frequencies under the null hypothesis at 5% and 10% levels. It is seen that both $T_n(20)$ and $T_n(1000)$ have good size. The power for both are similar. Hence, $T_n(q)$ with small q and large q behaves equally well in size and power.

Figure 1 shows the Monte Carlo distribution of the test statistic $T_n(q)$ from the 1000 Monte Carlo replications with the sample size $n = 200$. The three figures in the left panel are for $T_n(20)$, and the three figures in the right panel are for $T_n(1000)$. The solid line shows the asymptotic distribution, χ^2_3. All three DGPs in Figure 1 are linear in mean. Figure 1 confirms the size result of Table 1, showing that both $T_n(20)$ and $T_n(1000)$, despite the very different numbers of phantom activations, have the finite sample distributions very close to the asymptotic χ^2_3 distribution. These findings hold for all three DGPs under the null – AR, Linear($\rho = 0$), and Linear($\rho = 0.7$), that are linear in mean.

Table 1B repeats Table 1A using fixed hidden unit activations. In Table 1B, we generate γ_j from $U[-2, 2]$ and fix it across all 1000 replications. The results are similar to those in Table 1A – the size and power of $T_n(20)$ and $T_n(1000)$ are equally good. From Tables 1A, 1B we see that both randomly generated and fixed γ's provide good size. For power, when γ's are fixed, we see increasing power as we increase q from 20 to 1000 for Block 1. But for Block 2, the performance is similar. In general, fixed γ's can not beat randomly generated γ's in terms of power.

We also examine the possible effect of the conditional heteroskedasticity on the test. The AR in Block 1 is modified to have conditionally heteroskedastic errors as follows:

$$\text{AR-ARCH}: \quad y_t = 0.6y_{t-1} + \varepsilon_t, \quad h_t^2 = E(\varepsilon_t^2|y_{t-1}) = 0.9 + 0.1\varepsilon_{t-1}^2 \tag{7}$$

$$\text{AR-GARCH}: \quad y_t = 0.6y_{t-1} + \varepsilon_t, \quad h_t^2 = E(\varepsilon_t^2|y_{t-1}) = 0.1 + 0.1\varepsilon_{t-1}^2 + 0.8h_{t-1}^2 \tag{8}$$

In the cases when the errors are conditionally heteroskedastic, the test statistic

in (6) is not valid. We use the test statistic in equation (5) with Ψ_t replaced by Ψ_t^* and a corresponding consistent covariance matrix used. The test statistic has a valid asymptotic distribution of $\chi^2_{q^*}$. Table 1C reports the size of the test statistic, which is very close to the nominal size. The good size and good power of the randomized ANN tests under conditional homoskedasticity presented in Table 1A and Table 1B are not affected under conditional heteroskedasticity when the heteroskedasticity-robust statistics are employed as shown in Table 1C.

Table 1 and Figure 1 are in line with the known results in the literature showing outstanding properties of the ANN test even using a very small number of randomized hidden activations. These results do not show any difference in $T_n(20)$ and $T_n(1000)$, and thus they do not reveal some hidden problem of using a small number of randomized hidden activations.

In the next two sections, we show apparent difference in $T_n(20)$ and $T_n(1000)$. The main finding is that the ANN test with a small q, say $T_n(20)$, is not reliable to use in practice as it exhibits substantial variation to the random activations, while the ANN test with a large q, say $T_n(1000)$, is quite robust to the randomized activations as the large number of random activation is more dense in the nonlinear function space and thus reduces the variation of the statistic substantially.

To demonstrate the advantage of increasing q, we first conduct a Monte Carlo experiment again, in Section 4, but with only 5 replications for each DGP (rather than taking average over 1000 replications). We next apply $T_n(20)$ and $T_n(1000)$ to five monthly economic time series in Section 5 to show the advantage of $T_n(1000)$ over $T_n(20)$.

4 Small q vs. Large q in Sensitivity to Randomized Hidden Unit Activations

The simulation results reported in LWG (1993) and also in the previous section, show that the LWG has proper size and good power. However there is a hidden problem of the ANN test with small q. That is when q is small, the statistic and the corresponding p-value are sensitive to the randomized hidden unit activations.

Consider a sample $\{y_t\}_{t=1}^{n=200}$ for which the ANN statistic $T_n\left(q \mid \gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)$ and its p-value P_i are computed from m different randomly generated activation parameters $\left(\gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)_{i=1}^{m}$. Even if we use one same sample, it is possible that we sometimes get a small statistic and fail to reject the null for some

i, while other times we get a statistic large enough to reject the null for other i. Thus we may draw contradictory conclusions because of this sensitivity. As a result, the ANN test with small q can not be applied to empirical data and we need a solution to this problem.

We can deal with this problem in the following three ways. One approach is Teräsvirta, Lin and Granger (1993), who use a Taylor series expansion of the ANN function $f(\mathbf{x}_t, \theta)$ in (1) to write it into a parametric nonlinear approximation, and compare the estimated model with a linear model by the Wald test or LR test. The second approach is to generate $(\gamma_1, \ldots, \gamma_q)$ randomly from their parameter space Γ and integrate the statistic over Γ with a certain weight function $\phi(\gamma_1, \ldots, \gamma_q)$. This is to take a weighted average ANN statistic over the nuisance parameter space. The asymptotic theory has been established. But implementing this will require either the tabulation of the asymptotic distribution via simulation as it involves the integration of the Gaussian process or the use of bootstrap. Bierens (1982), Bierens (1990), Bierens and Ploberger (1997), and Härdle and Mammen (1993) take the statistics integrated over the nuisance parameter space. Corradi and Swanson (2002) use this method to test for nonlinear Granger-causality in out of sample. Alternative to taking the average of the statistic over nuisance parameter space Γ, Rossi and Inoue (2012) take the maximum of the statistic over Γ and Hansen and Timmermann (2011) take the minimum p-value over Γ. Their methods are in essence the same because of the one-to-one mapping between the statistic and the p-value. The asymptotic distributions of these statistics are integrals of Brownian motion. To obtain the correct critical value we need to either use bootstrap or follow the conditional p-value approach of Hansen (1996). Both methods are not easy to use so we turn to seek a simple and practical solution to the nuisance parameter problem.

This paper considers an obvious approach, the third approach, which is to increase q to a very large number. To compare how the ANN test works for small q and large q, we simulate a sample $\{y_t\}_{t=1}^{n=200}$ using DGP "Linear" in Block 2 with x_1 and x_2 following a bivariate normal distribution with correlation $\rho = 0.7$. Then we generate $m = 100$ different randomly generated activation parameters $\left(\gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)_{i=1}^{m=100}$, with which the ANN test statistic $T_n\left(q \mid \gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)$ and its p-value P_i are computed. We plot the histogram of the p-values and statistics with $q = 20$ or 1000 in Figure 3.

When $q = 20$, the p-values range from 0.0806 to 0.6719 for $i = 1, \ldots, m = 100$ (Figure 2a). We observe three of the 100 p-values are less than 0.10, which means in these three cases we incorrectly reject the null hypothesis

at 10% level. When we increase q to 1000 the p-values range from 0.2784 to 0.4567, all above the 10% level (Figure 2b). From these experiments we conjecture that if q is large enough, the p-value will be concentrated to a small area or even converge to a point. The sample variances of the p-values are 0.0255 and 0.0013 for $q = 20, 1000$ respectively. We also plot histograms of the m test statistics $\left\{T_n\left(q \mid \gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)\right\}_{i=1}^{m=100}$ with $q = 20$ (Figure 2c) and $q = 1000$ (Figure 2d). Since there is one-to-one mapping between the test statistic and the p-value, we shall see the similar pattern in the test statistic when q increases.

Table 2 reports the range and standard deviation (SD) of the p-values of $T_n(q)$ for $m = 100$ randomized hidden unit activations. For each DGP, we report the results for 5 replications. For each replication, we conduct testing with m randomized hidden unit activations. Comparing the range and SD of the p-values for $q = 20$ and $q = 1000$, we find that when q increases the range of p-value gets tighter and SD gets smaller, which makes the test outcome more stable over the m randomizations of γ_j's. When the DGP has an ARCH error tighter range and smaller SD are also found across all 5 replications as q increases from 20 to 1000. The results for AR-GARCH is not reported here since it is similar to the AR-ARCH case. Hence, increasing q makes the randomized ANN test more stable as well even under conditional heteroskedasticity.

Both Figure 2 and Table 2 show that increasing q is a good solution to the problem caused by randomizing the activation parameters. While the ANN statistic with a small number ($q = 20$) of randomized phantom activations exhibits large variation over $i = 1, \ldots, m$, it becomes stable with $q = 1000$. We can robustify the ANN test and reduce its sensitivity to the randomization of γ's by simply increasing q.

5 Small q vs. Large q in Applications

In this section, we compare $T_n(20)$ and $T_n(1000)$ for the same five monthly US economic time series used in LWG (1993) with updated time period from 1990:1 to 2011:12 with $n = 264$. The five series are US/Japan exchange rate (EX); US three-month T-bill interest rate (INT); US M2 money stock (M2); US personal income (PI), and US unemployment rate (UNE). We have made the same transformation as in LWG (p. 287), by taking logarithms and/or the first differencing, to ensure stationarity.

For each $\{y_t\}_{t=1}^{n=264}$ of these five series, we fit a linear AR(1) model un-

der H_0, so that the ANN has one input $\mathbf{x}_t = y_{t-1}$. The ANN statistic $T_n\left(q \mid \gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)$ and its p-value P_i are computed from m randomly generated activation parameters $\left(\gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)_{i=1}^{m}$. Table 3 reports the p-values $\{P_i\}$ with $i = 1, \ldots, m = 20$. Table 3 also reports the Hochberg's (1988) Bonferroni bound $HB(m)$ and the Simple Bonferroni bound $SB(m)$, both to be defined below, computed using the first m p-values (with $m = 5, 20$). Figure 3 presents the histograms of the p-values $\{P_i\}_{i=1}^{m=100}$.

For exchange rate and unemployment rate data, both the $T_n(20)$ and $T_n(1000)$ give consistent results among 20 times of tests. So with both $q = 20, 1000$, the null hypothesis of linearity is not rejected for exchange rate in all 20 p-values, but it is clearly rejected for unemployment rate by the ANN test using all $m = 20$ randomized hidden unit activations. However, for personal income PI, using $T_n(20)$ will give 2 times of failure of rejection in 20 randomized neural network activations, while using $T_n\,(1000)$ test, we reject the linearity using all 20 randomizations. For the M2 series, using $T_n(20)$ and $T_n(1000)$ will give us contradicting conclusions, as $T_n(20)$ rejects the null hypothesis 8 times out of 20 and $T_n(1000)$ rejects linearity in all 20 statistics. In this case, using $T_n(1000)$ yields more reliable result. For the interest rate INT, both $T_n(20)$ and $T_n(1000)$ give some uncertainty in the results in the sense that there are 3 or 4 times of failure of rejection out of the total 20 randomized activations. To examine this case further, we further increase q. The results (not shown in the table) show that when q increased to 2000, we can get 19 times of rejection out of 20. For INT, if $q = 1000$, some p-values are greater than 10% and some are even greater than 20%. But if $q = 2000$, all p-values are below 10% except one that is only slightly above it.

Table 3 reports the p-values for $T_n(20)$ and $T_n(1000)$ with $m = 20$ different randomly generated hidden unit activation parameters $\left(\gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)_{i=1}^{m}$. A low p-value suggests a rejection of the null hypothesis of linearity in conditional mean. Since the tests may not give consistent results over the different randomized activations, we use Bonferroni bounds on the p-value as a reference value. Let $\{P_1, ..., P_m\}$ be the p-values of m different randomized activations, and let $\left\{P_{(1)}, ..., P_{(m)}\right\}$ denote the ordered p-values from the smallest to the largest. Then the Bonferroni inequality leads to rejection of the null hypothesis at level α if $P_{(1)} \leq \alpha/m$, so we call $SB(m) := mP_{(1)}$ the Simple Bonferroni bound. One disadvantage of the Simple Bonferroni bound is that it is too conservative when m is large. Hochberg (1988) modified the rejection rule to reject the null hypothesis if there exists an i such that $P_{(i)} \leq \alpha/(m - i + 1)$, $i = 1, ..., m$. We call $HB(m) := \min_{i=1,...,m}(m - i + 1)P_{(i)}$ the Hochberg Bonferroni bound. In

Table 3, reported are $SB(m)$ and $HB(m)$ with $m = 5, 20$.

A disadvantage of the Simple Bonferroni bound is that it could be larger than 1, especially when m is large. The Simple Bonferroni bound is more sensitive to q than the Hochberg Bonferroni bound. Comparing Bonferroni bounds over $q = 20, 1000$, the Hochberg Bonferroni bounds for $m = 5$ and $m = 20$ are close for $T_n(1000)$, but the difference between the two bounds $HB(5)$ and $HB(20)$ is larger for $T_n(20)$. Hence, increasing the number of the randomized hidden activations not only makes the ANN test more robust but also the Bonferroni bounds tighter. From the formula $HB(m) := \min_{i=1,\ldots,m}(m - i + 1)P_{(i)}$, it is easy to see that, when q is large, the Hochberg Bonferroni bound tend to be the maximum p-value $HB(m) \approx P_{(m)}$ since the p-values tend to be concentrated to a small region as discussed in the previous section (Table 2 and Figure 2). However, when q is small, the Hochberg Bonferroni bound may give inconsistent conclusion according to different values of m. For instance, for the money stock M2 series, for $T_n(20)$, we do not reject linearity when $m = 5$ at 10% level, yet we reject linearity when $m = 20$ at 10% level. And in this case, we can reject linearity using $T_n(1000)$ with both $m = 5$ and $m = 20$. Thus the Hochberg Bonferroni bound is preferred to the Simple Bonferroni bound. Moreover, if we use $T_n(1000)$ instead $T_n(20)$, we can take a smaller value of m and get reliable conclusion.

The reported numbers in the last part of Table 3 are the rejection frequency in these $m = 20$ p-values that are less than 0.10 (at 10% level), $REJ = \frac{1}{m}\sum_{i=1}^{m=20} 1(P_i \leq 0.10)$. To compare the rejection frequency using different approaches, we compare the rejection frequency using the Bonferroni approach to the False Discovery Rate (FDR) of Storey (2003) and Benjamini and Hochberg (1995). The results are reported in the last two rows of Table 3. REJ-B is the rejection frequency using the Bonferroni approach, where $REJ\text{-B} = \sum_{i=1}^{m=20} 1(P_i \leq \frac{0.10}{20})$. REJ-FDR is the rejection frequency using FDR, where $REJ\text{-FDR} = \sum_{i=1}^{m=20} 1(P_{(i)} \leq \frac{0.10i}{20})$. Note that, using the Bonferroni approach we get fewer times of rejection for interest rate for both $q = 20$ and $q = 1000$, while for individual test, we can reject most of the time. This problem can be solved if we use FDR. The results show that FDR can improve the power of the test for all the series. Storey (2003) pointed out that the positive False Discovery Rate (pFDR) could improve the power of FDR when the number of tests is large. In our study we find the rejection frequency of pFDR depends heavily on the choice of tuning parameter. As we get good power for our data with FDR, we do not report the results with pFDR here.

In addition to Table 3 for which 20 p-values (with $m = 20$) are used, we also experiment this with $m = 100$ random draws of the hidden unit activations

and 100 p-values are presented in Figure 3. For all five economic time series, the p-values tend to get concentrated at a narrow region or even converge to a single value when $q = 1000$ compared with $q = 20$. For M2 data, the p-values of $T_n(20)$ range widely from 0 to 1 and close to 1 for around 40 times among the 100 p-values, while all the 100 p-values of $T_n(1000)$ are near zero. For personal income, when $q = 1000$, we can get rejection among all 100 times of tests while when $q = 20$, we cannot reject for around 10 times of tests. For interest rate INT as REJ becomes $\frac{19}{20}$ when we experiment it with $q = 2000$ (not shown). These results clearly indicate that choosing a large $q = 1000$ can give more stable conclusion compared with choosing a small $q = 20$.

6 Conclusions

In this paper, we revisit the ANN-based test statistics for neglected nonlinearity in conditional mean. The ANN test has a set of nuisance parameters that are not identified under the null hypothesis. As the nuisance parameters are identified only under the alternative, the alternative ANN model can be estimated to form a Wald-type test statistic. However, the estimation of the ANN models are known to be difficult and the estimated models are often contaminated by large estimation errors. To avoid the estimation of the ANN models, LWG (1993) suggested a noble test in a Lagrange multiplier (LM) test framework for which the ANN model under the alternative hypothesis needs not be estimated. As suggested in LWG (1993), in constructing an LM test, the unidentified nuisance parameters under the null hypothesis can be randomly generated from their parameter space. LWG show excellent performance of the ANN test when a small number of hidden activations is based on the randomly generated nuisance parameters.

It has not been noted in the literature that the ANN test is sensitive to the number of the randomized activations. We demonstrate this sensitivity problem and propose a simple solution. We examine how the performance of the ANN test can be improved by simply increasing the number of randomized hidden unit activations. This paper shows that the benefit of increasing it is substantial. This robustification is reliable and does not require either the use of Bonferroni bounds or the integration of the test statistic over the nuisance parameter. We provide a practically useful insight to make the ANN test reliably applicable in applied work. As increasing the number of random activations is almost costless, the ANN test based on "many" randomized hidden unit neural network activations can be easily included in a diagnostics toolbox for applied research.

References

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

Benjamini, Y. and Hochberg, Y. (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing ". *Journal of the Royal Statistical Society Series B* 57(1): 289-300.

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

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

Bierens, H. and Ploberger, W. (1997): "Asymptotic Theory of Integrated Conditional Moment Test". *Ecomometrica* 65: 1129-1151.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models", *Handbook of Econometrics*, Vol. 6B, Chapter 76, Elsevier B.V.

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

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

Hansen, P. R. and Timmermann, A. (2011): "Choice of Sample Split in Out-of-Sample Forecast Evaluation". EUI and UCSD, Working Paper.

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

Hochberg, Y. (1988): "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika* 75, 800-802.

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

Hornik, K., Stinchcombe, M., and White, H. (1990): "Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks," *Neural Networks* 3, 551-560.

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

Rossi, B. and Inoue, A. (2012): "Out-of-Sample Forecast Tests Robust to the Choice of Window Size". *Journal of Business and Economic Statistics* 30(3): 432-453.

Storey, J. (2003): "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value". *The Annals of Statistics* 31(6): 2013-2035.

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

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

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

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. (1990): "Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings," *Neural Networks* 3, 535-549.

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

White, H. and Wooldridge, J.M. (1991): "Some Results for Sieve Estimation with Dependent Observations," in W. Barnett, J. Powell and G. Tauchen, eds., *Nonparametric and Semi-Parametric Methods in Econometrics and Statistics*. New York: Cambridge University Press, 459-493.

Table 1. Monte Carlo: Size and Power of the ANN Test

Panel A. Using Randomized Hidden Unit Activations

	$q = 20$		$q = 1000$	
	5%	10%	5%	10%
AR	0.037	0.085	0.041	0.089
TAR	0.263	0.391	0.268	0.374
SGN	0.812	0.901	0.835	0.910
NAR	0.079	0.162	0.099	0.178
MRS	0.179	0.273	0.197	0.284
Linear($\rho = 0$)	0.047	0.105	0.049	0.097
Linear($\rho = 0.7$)	0.045	0.097	0.058	0.106
Interaction($\rho = 0$)	0.112	0.183	0.082	0.141
Interaction($\rho = 0.7$)	0.244	0.252	0.261	0.369
Squared($\rho = 0$)	0.191	0.297	0.186	0.272
Squared($\rho = 0.7$)	0.346	0.375	0.238	0.352

Panel B. Using Fixed Hidden Unit Activations

	$q = 20$		$q = 1000$	
	5%	10%	5%	10%
AR	0.055	0.095	0.047	0.103
TAR	0.197	0.285	0.318	0.402
SGN	0.838	0.846	0.908	0.915
NAR	0.081	0.115	0.151	0.184
MRS	0.134	0.166	0.201	0.254
Linear($\rho = 0$)	0.055	0.115	0.043	0.100
Linear($\rho = 0.7$)	0.040	0.101	0.035	0.093
Interaction($\rho = 0$)	0.131	0.215	0.068	0.132
Interaction($\rho = 0.7$)	0.190	0.280	0.213	0.334
Squared($\rho = 0$)	0.130	0.230	0.160	0.265
Squared($\rho = 0.7$)	0.193	0.284	0.245	0.367

Panel C. Size with Conditional Heteroskedasticity

	$q = 20$		$q = 1000$	
	5%	10%	5%	10%
AR-ARCH	0.048	0.112	0.058	0.107
AR-GARCH	0.058	0.122	0.055	0.113
AR	0.051	0.100	0.049	0.116

Notes: Sample size is $n = 200$. Reported values are the rejection frequencies of the $T_n(q)$ tests out of the total 1000 Monte Carlo replications, at 5% and 10% levels. In Panel A and Panel C, the hidden unit activations γ's are randomly generated for each replication. In Panel B, the hidden unit activations are fixed to be one random draw from $U[-2,\ 2]$ for all replications. The ANN test statistic in (6) is used in Panel A and Panel B, while the heteroskedasticity robust statistic of the form in (5) with the principal components is used in Panel C.

Table 2. P-values of $T_n(q)$ with $m = 100$ Randomizations of q Hidden Unit Activations

Panel A. Block 1

	Range		SD	
	q=20	q=1000	q=20	q=1000
AR	0.5540	0.0391	0.1315	0.0063
	0.4582	0.0814	0.1030	0.0139
	0.2468	0.0382	0.0686	0.0081
	0.6711	0.0898	0.1435	0.0182
	0.4973	0.1188	0.1353	0.0262
TAR	0.7319	0.0448	0.1691	0.0076
	0.8275	0.2084	0.2567	0.0469
	0.4369	0.0291	0.0943	0.0061
	0.5807	0.1165	0.1389	0.0237
	0.0485	0.0006	0.0092	0.0001
SGN	0.0010	0.0002	0.0002	0.0000
	0.1525	0.0186	0.0262	0.0043
	0.0280	0.0017	0.0062	0.0003
	0.0004	0.0000	0.0000	0.0000
	0.1791	0.0199	0.0361	0.0037
NAR	0.8330	0.2214	0.2415	0.0424
	0.4501	0.0294	0.0812	0.0058
	0.8446	0.0595	0.1806	0.0121
	0.5092	0.1478	0.1435	0.0270
	0.0869	0.0218	0.0198	0.0049
MRS	0.8610	0.2738	0.2090	0.0557
	0.9992	0.1667	0.3300	0.0298
	0.2210	0.0337	0.0529	0.0072
	0.3194	0.0247	0.0557	0.0049
	0.7001	0.2378	0.2062	0.0447
AR-ARCH	0.2754	0.1164	0.0784	0.0219
	0.5257	0.0696	0.1998	0.0155
	0.6207	0.0188	0.1695	0.0034
	0.3866	0.0560	0.1281	0.0121
	0.1530	0.0046	0.0424	0.0000

Table 2. (Continued).

Panel B. Block 2

	Range		SD	
	q=20	q=1000	q=20	q=1000
Linear	0.3846	0.0855	0.0781	0.0149
($\rho = 0$)	0.9793	0.5644	0.2943	0.1345
	0.9389	0.1879	0.2807	0.0332
	0.7879	0.1093	0.2289	0.0212
	0.6020	0.0755	0.1425	0.0179
Linear	0.2470	0.0373	0.0490	0.0065
($\rho = 0.7$)	0.4526	0.0120	0.0616	0.0027
	0.9684	0.3642	0.3033	0.0819
	0.3680	0.0322	0.0500	0.0067
	0.5981	0.1580	0.1640	0.0361
Intersection	0.5982	0.1626	0.1646	0.0361
($\rho = 0$)	0.8872	0.2460	0.2114	0.0510
	0.9198	0.1641	0.2285	0.0284
	0.4399	0.0914	0.0995	0.0215
	0.8509	0.1895	0.2234	0.0332
Intersection	0.5417	0.0558	0.0819	0.0101
($\rho = 0.7$)	0.4461	0.0606	0.0943	0.0116
	0.9064	0.4364	0.2627	0.0889
	0.4144	0.1015	0.0911	0.0213
	0.1813	0.0205	0.0308	0.0042
Squared	0.5354	0.0481	0.1253	0.0095
($\rho = 0$)	0.9339	0.4809	0.2538	0.1140
	0.8627	0.4690	0.2661	0.1000
	0.7223	0.1195	0.1503	0.0259
	0.8681	0.0881	0.2040	0.0196
Squared	0.3645	0.0391	0.0632	0.0094
($\rho = 0.7$)	0.6779	0.1191	0.1435	0.0249
	0.9652	0.3963	0.2997	0.0933
	0.3984	0.0475	0.0592	0.0088
	0.3805	0.0556	0.0663	0.0110

Note: Sample size $n = 200$. The p-values of $T_n(q)$ are computed for simulated data of each DGP from five replications. The statistic $T_n(q)$ is computed with $m = 100$ random draws of $\left\{\gamma_j^{(i)}\right\}_{i=1,\ldots,m}$ from $U[-2,\ 2]$. The table reports the range and standard deviation (SD) of the m p-values in each of 5 replications with $q = 20, 1000$.

Table 3. Empirical Analysis: P-values, Bonferroni Bounds, and Rejection Frequencies

	EX		INT		M2	
$q =$	20	1000	20	1000	20	1000
$i = 1$	0.1192	0.7126	0.0023	0.0156	0.3034	0.0034
$i = 2$	0.1752	0.7164	0.8501	0.0096	0.0576	0.0029
$i = 3$	0.4740	0.7926	0.0665	0.0032	0.0229	0.0033
$i = 4$	0.8045	0.7726	0.1198	0.0752	1.0000	0.0108
$i = 5$	0.1565	0.7589	0.0064	0.0024	1.0000	0.0041
$i = 6$	0.4497	0.8244	0.0034	0.2006	1.0000	0.0023
$i = 7$	0.5505	0.7688	0.0125	0.0595	0.2030	0.0030
$i = 8$	0.5022	0.7575	0.0608	0.0535	0.0049	0.0036
$i = 9$	0.4750	0.7676	0.0258	0.1487	0.0049	0.0101
$i = 10$	0.4628	0.7587	0.0407	0.0115	1.0000	0.0019
$i = 11$	0.4800	0.7097	0.0121	0.0013	1.0000	0.0023
$i = 12$	0.2813	0.8057	0.1246	0.0146	0.3971	0.0115
$i = 13$	0.4717	0.6834	0.0003	0.1007	1.0000	0.0028
$i = 14$	0.4730	0.6988	0.0217	0.0537	0.0495	0.0025
$i = 15$	0.5196	0.7630	0.0090	0.0015	0.8678	0.0029
$i = 16$	0.1573	0.7742	0.4018	0.0332	0.0033	0.0181
$i = 17$	0.5109	0.8010	0.0044	0.0191	0.0351	0.0067
$i = 18$	0.5241	0.7831	0.0102	0.0955	0.9987	0.0037
$i = 19$	0.4386	0.7584	0.0017	0.0333	0.0378	0.0051
$i = 20$	0.4380	0.7807	0.0247	0.0354	0.1197	0.0042
$HB(5)$	0.5256	0.7926	0.0115	0.0120	0.1145	0.0082
$HB(20)$	0.8045	0.8244	0.0060	0.0260	0.0660	0.0181
$SB(5)$	0.5960	3.5630	0.0115	0.0120	0.1145	0.0145
$SB(20)$	2.3840	13.6680	0.0060	0.0260	0.0660	0.0380
REJ	$\frac{0}{20}$	$\frac{0}{20}$	$\frac{16}{20}$	$\frac{17}{20}$	$\frac{8}{20}$	$\frac{20}{20}$
REJ-B	$\frac{0}{20}$	$\frac{0}{20}$	$\frac{5}{20}$	$\frac{4}{20}$	$\frac{3}{20}$	$\frac{14}{20}$
REJ-FDR	$\frac{0}{20}$	$\frac{0}{20}$	$\frac{16}{20}$	$\frac{16}{20}$	$\frac{3}{20}$	$\frac{20}{20}$

Table 3. (Continued).

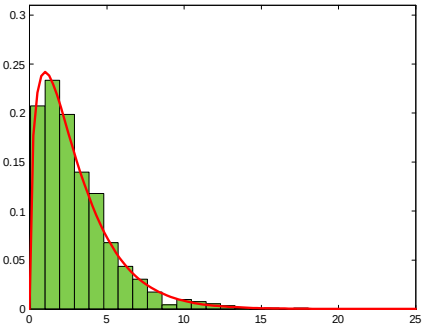
	PI		UNE	
$q =$	20	1000	20	1000
$i = 1$	0.0000	0.0000	0.0011	0.0006
$i = 2$	0.0000	0.0000	0.0003	0.0006
$i = 3$	0.0000	0.0000	0.0024	0.0005
$i = 4$	0.0361	0.0000	0.0005	0.0008
$i = 5$	0.0000	0.0000	0.0003	0.0008
$i = 6$	0.0000	0.0000	0.0005	0.0006
$i = 7$	0.0000	0.0000	0.0004	0.0004
$i = 8$	0.0017	0.0000	0.0004	0.0008
$i = 9$	0.0000	0.0000	0.0006	0.0008
$i = 10$	0.0000	0.0000	0.0059	0.0007
$i = 11$	0.0000	0.0000	0.0002	0.0005
$i = 12$	0.0000	0.0000	0.0003	0.0007
$i = 13$	0.0000	0.0000	0.0013	0.0006
$i = 14$	0.0000	0.0000	0.0003	0.0006
$i = 15$	0.0000	0.0000	0.0003	0.0007
$i = 16$	0.0000	0.0000	0.0003	0.0006
$i = 17$	0.0000	0.0000	0.0003	0.0006
$i = 18$	0.3395	0.0000	0.0005	0.0007
$i = 19$	0.2982	0.0000	0.0004	0.0006
$i = 20$	0.0000	0.0000	0.0002	0.0007
$HB(5)$	0.0000	0.0000	0.0012	0.0008
$HB(20)$	0.0000	0.0000	0.0030	0.0008
$SB(5)$	0.0000	0.0000	0.0015	0.0023
$SB(20)$	0.0000	0.0000	0.0040	0.0089
REJ	$\frac{18}{20}$	$\frac{20}{20}$	$\frac{20}{20}$	$\frac{20}{20}$
REJ-B	$\frac{17}{20}$	$\frac{20}{20}$	$\frac{19}{20}$	$\frac{20}{20}$
REJ-FDR	$\frac{18}{20}$	$\frac{20}{20}$	$\frac{20}{20}$	$\frac{20}{20}$

Notes: Data range from 1990:1-2011:12, monthly. EX: US/Japan exchange rate. INT: US three-month T-bill interest rate. M2: US M2 money stock. PI: US personal income. UNE: US unemployment rate. We use AR(1) as a model under the null hypothesis in each case. The 20 rows ($i = 1, \ldots, 20$) show the 20 sets of p-values $\{P_i\}_{i=1}^{m}$ of the ANN(q) test statistics with $q = 20$ or 1000. $HB(m) = \min_{i=1,\ldots,m}(m - i + 1) \times P_{(i)}$ is the Hochberg's Bonferroni bound computed the first m p-values $\{P_i\}_{i=1}^{m}$ with $m = 5$ or 20. $HB(5)$ in the Hochberg Bonferroni bound computed using the first 5 p-values $\{P_i\}_{i=1}^{m=5}$. $SB(m) = mP_{(1)}$ is the Simple Bonferoni bound computed the first m p-values. $P_{(i)}$ is the ith smallest (ordered

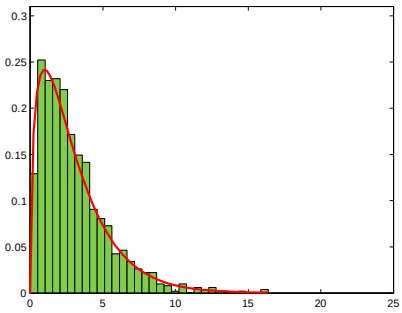
from the smallest to the largest) p-value among the m p-values. The reported numbers in the last three rows are the rejection frequency in these $m = 20$ p-values that are less than 0.10 (at 10% level), $REJ = \frac{1}{m}\sum_{i=1}^{m=20} 1(P_i \leq 0.10)$, REJ-B $= \sum_{i=1}^{m=20} 1(P_i \leq \frac{0.10}{20})$, and REJ-FDR $= \sum_{i=1}^{m=20} 1(P_{(i)} \leq \frac{0.10i}{20})$.

Figure 1. Monte Carlo Distribution of $T_n(q)$ under H_0

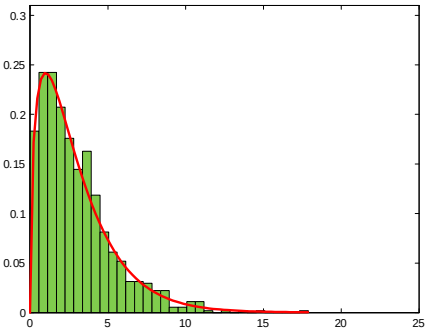
(a) $q = 20$. DGP: AR



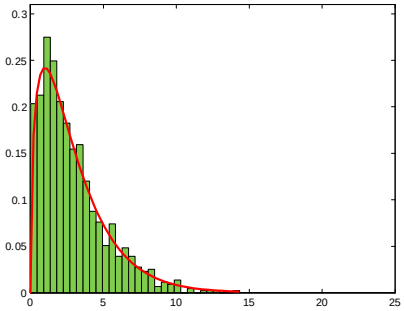
(b) $q = 1000$. DGP: AR



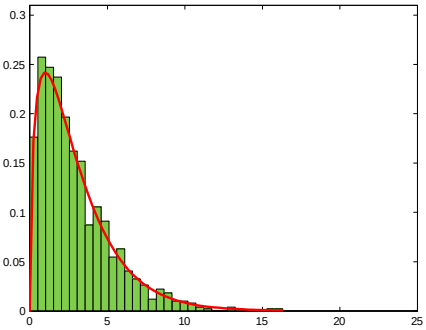
(c) $q = 20$. DGP: Linear($\rho = 0$)



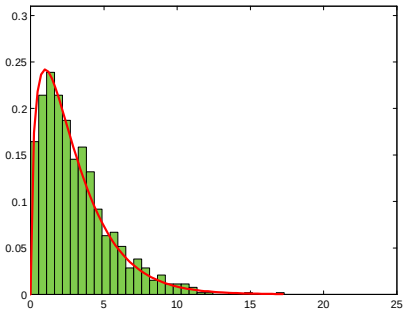
(d) $q = 1000$. DGP: Linear($\rho = 0$)



(e) $q = 20$. DGP: Linear($\rho = 0.7$)



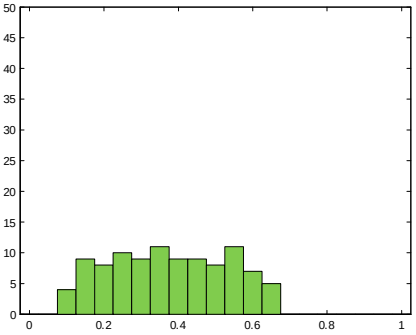
(f) $q = 1000$. DGP: Linear($\rho = 0.7$)



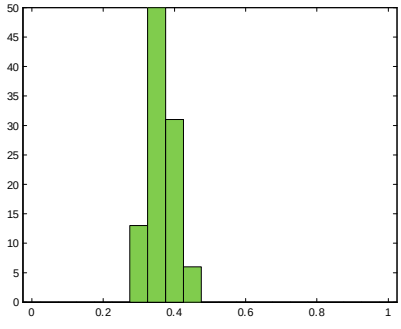
Note: The histograms are the Monte Carlo distribution of the test statistic $T_n(q)$ from the 1000 Monte Carlo replications with the sample size $n = 200$. The three figures in the left panel are for $T_n(20)$, and the three figures in the right panel are for $T_n(1000)$. The solid line is the χ^2_3 density. All DGPs here are linear in mean.

Figure 2. P-values of $T_n(q)$ under H_0 with $m = 100$ Randomizations of q Hidden Units

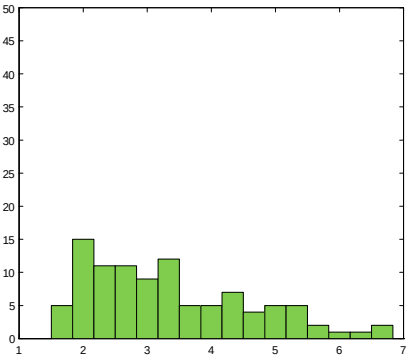
(a) $q = 20$. P-values



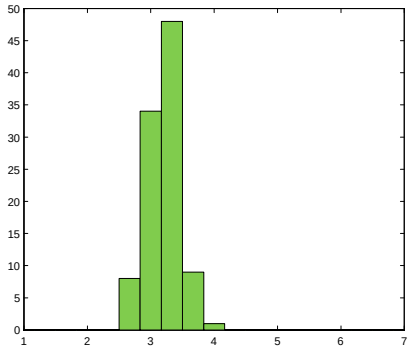
(b) $q = 1000$. P-values



(c) $q = 20$. Test Statistics



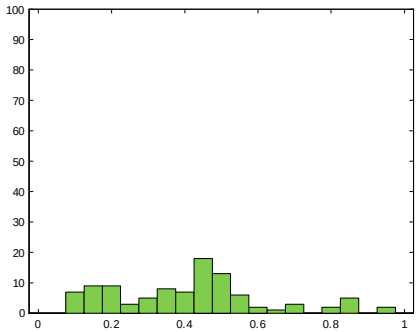
(d) $q = 1000$. Test Statistics



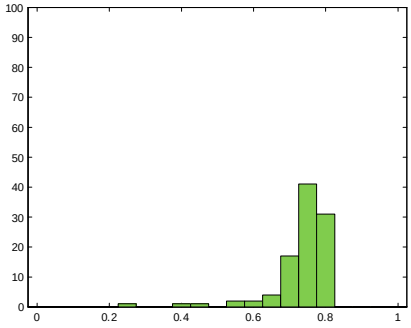
Note: Sample size $n = 200$. The p-values and the test statistics $T_n(q)$ are computed for a simulated data from one replication of DGP, "Linear" with $\rho = 0.7$. For the same data, the statistic $T_n(q)$ is computed with $m = 100$ random draws of $\left\{\gamma_j^{(i)}\right\}_{i=1,\ldots,m}$ from $U[-2,\ 2]$. The figures are frequency histograms of the m p-values and the m statistics. The top panels report the p-values and the bottom panels report the statistics. The left panels are for $q = 20$ and the right panels are for $q = 1000$.

Figure 3. Empirical Applications with $m = 100$ Randomized Hidden Unit Activations

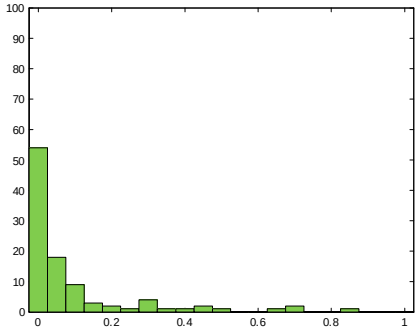
(a) P-values of EX with $q = 20$



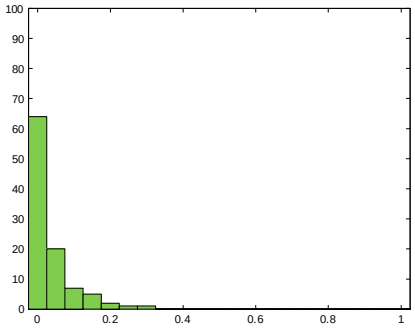
(b) P-values of EX with $q = 1000$



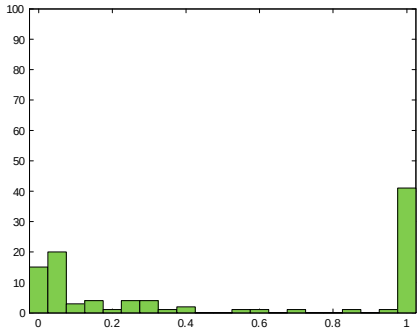
(c) P-values of INT with $q = 20$



(d) P-values of INT with $q = 1000$



(e) P-values of M2 with $q = 20$



(f) P-values of M2 with $q = 1000$

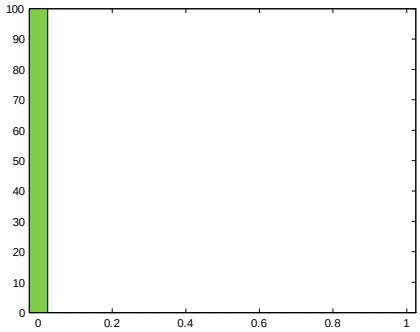
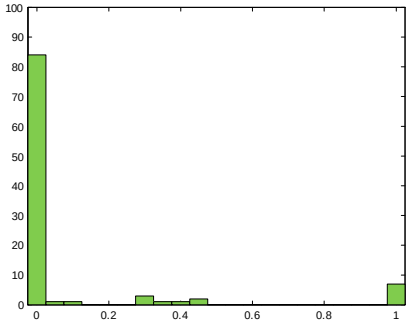
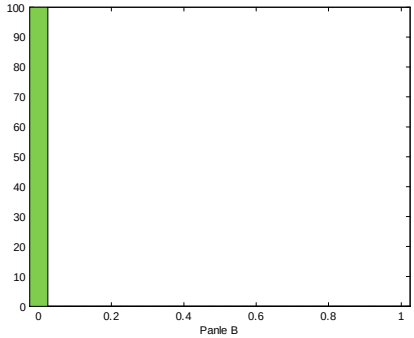


Figure 3 (Continued).

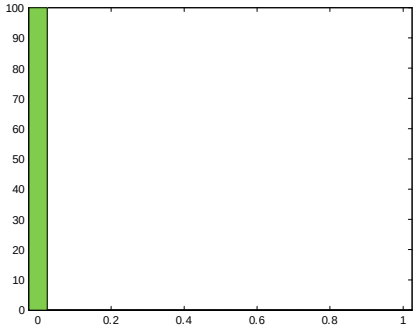
(g) P-values of PI with $q = 20$



(h) P-values of PI with $q = 1000$



(i) P-values of UNE with $q = 20$



(j) P-values of UNE with $q = 1000$