

# STATISTICAL INFERENCE, THE BOOTSTRAP, AND NEURAL NETWORK MODELING WITH APPLICATION TO FOREIGN EXCHANGE RATES

HALBERT WHITE AND JEFF RACINE

**ABSTRACT.** In this paper we propose tests for individual and joint irrelevance of network inputs. Such tests can be used to determine whether an input or group of inputs “belong” in a particular model, thus permitting valid statistical inference based on estimated feedforward neural network models. The approaches employ well known statistical resampling techniques. We conduct a small Monte Carlo Experiment showing that our tests have reasonable level and power behavior, and we apply our methods to examine whether there are predictable regularities in foreign exchange rates. We find that exchange rates do appear to contain information that is exploitable for enhanced point prediction, but the nature of the predictive relations evolves through time.

## 1. INTRODUCTION

To date, most applications of feedforward artificial neural networks (ANNs) have been concerned with the estimation of relationships, especially relationships between input and target variables of interest. Nevertheless, artificial neural networks have a broader utility that has yet to be fully appreciated by neural network practitioners, but which has the potential to significantly enhance scientific understanding of empirical phenomena subject to neural network modeling. In particular, the estimates obtained from network learning can serve as a basis for formal statistical inference, making possible statistical tests of specific scientific hypotheses of interest. Because of the ability of artificial neural networks to extract complex nonlinear and interactive effects, the alternatives against which such tests can have power may extend usefully beyond those within reach of more traditional methods, such as linear regression modeling.

As a concrete example, artificial neural networks can be used to diagnose the occurrence of heart attacks (Baxt (1992)). As Baxt notes, some of the inputs that the network appears to make use of are not traditionally recognized as being useful diagnostics (rales, jugular venous distension, and syncope). The question arises as to whether these inputs truly are useful diagnostics, or whether they only appear to be useful as a result of random sampling variation. Formally, one can address this issue by conducting a statistical test of the null hypothesis that the diagnostic utility of a given

symptom is nil, versus the alternative that the symptom does have diagnostic utility. See Baxt & White (1995).

As another example, it is of interest in international finance to know if foreign exchange markets are efficient in the sense that past movements of the price of a given currency have no predictive value in forecasting future movements of the currency price. This issue can be addressed by conducting a statistical test of the null hypothesis that past currency price movements do not influence future currency price movements against the alternative that they do. We examine this issue here by applying our methods to exchange rates in G7 countries.

In standard parametric modeling, such null hypotheses can usually be cast as specific restrictions on the model parameters (network weights). Known asymptotic properties (*i.e.* properties in large samples) of the parameter estimator can then be exploited to assess the extent to which the parameter estimates do or do not accord with the restrictions specified by the null hypothesis. The null hypothesis can then be rejected or not, in a way that controls the (large sample) probability of Type I error, the error of wrongly rejecting the null.

White (1989) gives statistics that can be used to test hypotheses of interest regarding optimal network weights, for example that a given input or set of inputs is irrelevant in the sense that the optimal weights on connections from the set of inputs in question to hidden units in the next layer are zero. We shall refer to such procedures as “weight-based,” because they focus on properties of network weights that have implications for the properties of the network mapping. The properties of the network weights tested may be either necessary or sufficient for the network behavior of interest. To the extent that the hypothesis on the optimal weights is not both necessary and sufficient for the network behavior of interest, inference regarding network behavior will be incomplete.

An alternative to weight-based procedures is to posit and test hypotheses regarding network behavior of interest directly. In the case of the irrelevant inputs hypothesis, one can specify and test the hypothesis that the partial derivatives with respect to the inputs in question are zero for the optimal network. In this paper we investigate the formulation and testing of precisely this network-based approach to the hypothesis of irrelevant inputs. As will be clear from our discussion, our network-based approach can be similarly applied to other hypotheses of interest regarding network behavior.

A challenge facing implementation of this direct approach is that the statistics forming the basis for our tests have somewhat complicated large sample distributions. Because these distributions

are not to be found in standard tables, we must provide suitable methods for computing the large sample critical values necessary to performing tests of controlled level (probability of type I error). Because of its straightforwardness and broad applicability, we adopt Efron's (1983) bootstrap method for calculating the required critical values. Although computationally intensive, the bootstrap is feasible for inference using artificial neural networks given a modern PC computing environment.

An additional benefit of the bootstrap is that in many applications it provides approximations to the sampling distribution of the test statistic of interest that are considerably more accurate than the analytically obtained large sample approximation (see Hall (1992)). Formal investigation of this additional benefit of the bootstrap is beyond the scope of this work. We have a full agenda just to establish the asymptotic validity of the bootstrap procedures of interest to us. Nevertheless, we anticipate that certain of our bootstrap procedures may well afford such superior finite sample approximations; we reserve formal investigation of that issue to subsequent work.

Succinctly put, our goal here will therefore be to establish the validity of practical bootstrap procedures for testing a leading hypothesis of interest, specifically that of null effect for a given set of inputs. These procedures should facilitate the practice of statistical inference using artificial neural networks, with the attendant potential for enhanced scientific understanding of empirical phenomena.

This paper proceeds as follows. Section 2 sets forth the hypothesis of null effect for the networks of interest, introduces statistics appropriate for testing this hypothesis, obtains their asymptotic distribution, and justifies use of the bootstrap to approximate this distribution. Section 3 contains a discussion of the consequences of using model selection procedures to determine optimal model complexity before testing the irrelevant input hypothesis. Provided that complexity determination is suitably carried out, no adverse effects need arise. Section 4 discusses details of implementation relied on in Sections 5 and 6. Section 5 describes some simulations examining the behavior of our procedures, while Section 6 presents an application to testing the irrelevant input hypothesis in the context of foreign exchange spot rate movements. Section 7 contains some concluding remarks. Proofs of our mathematical results are gathered into the Mathematical Appendix.

2.1. **Network Modeling and the Hypothesis of Interest.** We will be concerned in this paper with data generated according to the following conditions.

**Assumption A.1:** Observed data are the realization of a sequence  $\{Z_t = (Y_t, X_t^T)^T\}$  of independent identically distributed (iid) random  $(k + 1) \times 1$  vectors,  $k \in \mathbb{N}$ .

The random variables  $Y_t$  represent targets; their relationship to the variables  $X_t$  is of primary interest. When  $E(Y_t) < \infty$ , the conditional expectation of  $Y_t$  given  $X_t$  exists, denoted  $g(X_t) = E(Y_t|X_t)$ . Defining  $\epsilon_t \equiv Y_t - g(X_t)$ , we can also write

$$Y_t = g(X_t) + \epsilon_t.$$

The unknown function  $g$  embodies the systematic part of the stochastic relation between  $Y_t$  and  $X_t$ . The error  $\epsilon_t$  is “noise”, with the property that  $E(\epsilon_t|X_t) = 0$  by construction. As in White (1989), we let  $\nu$  denote the joint distribution of  $Z_t$ , and we let  $\mu$  denote the joint distribution of  $X_t$ . Thus,  $\mu(A)$  gives the probability of observing  $X_t$  to lie in any (Borel) subset  $A$  of  $\mathbb{R}^k$ . We denote the smallest Borel set  $A$  such that  $\mu(A) = 1$  as  $\text{supp } \mu$ , the “support” of  $\mu$ . The set  $\text{supp } \mu$  is “where the action is”:  $X_t$  is observed to lie outside of  $\text{supp } \mu$  with probability zero. Because  $g$  is unknown, we approximate it using a family of known functions. Of particular interest to us are the output functions of hidden layer feedforward networks. We begin by assuming that these functions have the following properties:

**Assumption A.2:** (a) Network output is given by a function  $f : \mathbb{R}^k \times W \rightarrow \mathbb{R}$ , where  $W$  is a compact subset of  $\mathbb{R}^p$ ,  $p \in I$ , such that for each  $x$  in  $\text{supp } \mu$ ,  $f(x, \cdot) : W \rightarrow \mathbb{R}$  is continuous and for each  $w$  in  $W$ ,  $f(\cdot, w) : \mathbb{R}^k \rightarrow \mathbb{R}$  is measurable.

Network weights are here restricted to lie in a compact set  $W$  of finite dimension  $p$ ;  $p$  is thus the total number of weights. This requirement for network output functions is satisfied for single hidden layer feedforward networks of the form

$$f(x, w) = w_{00} + \sum_{j=1}^h w_{0j} \psi(\tilde{x}^T w_{1j}),$$

where  $w \equiv (w_{00}, w_{01}, \dots, w_{0h}, w_{11}^T, \dots, w_{1h}^T)^T$ ,  $\tilde{x} = (1, x^T)^T$ , whenever the hidden unit activation function  $\psi$  is continuous.

As in White (1989) and White (1994), we suppose that learning is conducted in such a way as to solve the optimization problem

$$(1) \quad \min_{w \in W} n^{-1} \sum_{t=1}^n \pi(Y_t, f(X_t, w)),$$

where  $\pi$  is a suitably chosen function. For example, least squares learning occurs when

$$\pi(Y_t, f(X_t, w)) = (Y_t - f(X_t, w))^2.$$

Many other useful possibilities are permitted by viewing learning in this way. We impose the following requirement on  $\pi$ .

**Assumption A.3:** (a) The function  $\pi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  is such that  $\pi(y, \cdot) : \mathbb{R} \rightarrow \mathbb{R}$  is continuous for each  $y \in \mathbb{R}$ ,  $\pi(\cdot, a) : \mathbb{R} \rightarrow \mathbb{R}$  is measurable for each  $a \in \mathbb{R}$ , and there exists a measurable function  $D : \mathbb{R}^{k+1} \rightarrow \mathbb{R}^+$  such that

$$\sup_{w \in W} |\pi(y, f(x, w))| \leq D(z)$$

and  $E(D(Z_t)) < \infty$ , i.e.  $|\pi(Y_t, f(X_t, w))|$  is dominated by a function integrable with respect to  $v$ .

For the case of least squares learning with single hidden layer feedforward networks, this condition holds provided that  $E(Z_t^T Z_t) < \infty$  and  $|\psi(b)| \leq \Delta_1 |b| + \Delta_2$  for some finite constants  $\Delta_1, \Delta_2$  and all  $b \in \mathbb{R}$ .

It follows from Theorem 1 of White (1989, p. 457) that a weight vector  $\hat{w}_n$  solving (1) exists and converges almost surely to  $w^*$ , which solves

$$(2) \quad \min_{w \in W} \lambda(w) \equiv \int \pi(y, f(x, w)) dv(z),$$

provided that this solution is unique. (Note that Assumptions A.1-A.3 ensure the existence of this integral.) In the absence of appropriate restrictions, the permutability of hidden units and the admissibility of sign-flip operations generally results in a non-unique  $w^*$ , as there are numerous distinct weight vectors yielding identical network outputs. Sussmann (1992), Chen, Lu & Hecht-Nielsen (1993) and Coetzee & Stonick (1995) provide conditions sufficient to ensure uniqueness of  $w^*$  in a suitable  $W$  for specific network configurations. To avoid bogging down in technicalities, we assume such restrictions have been imposed:

**Assumption A.4:** The problem (2) has a unique solution vector  $w^*$  interior to  $W$ .

In the case of least squares learning with a single hidden layer feedforward network, this requires a unique solution to the problem

$$\min_{w \in W} E([Y_t - f(X_t, w)]^2)$$

or to the equivalent problem

$$(3) \quad \min_{w \in W} E([g(X_t) - f(X_t, w)]^2),$$

due to the facts that  $E(\epsilon|X_t) = 0$  and  $E(\epsilon_t^2) < \infty$  does not depend on  $w$ . In this case, we see that  $w^*$  indexes a mean squared error-optimal approximation  $f(\cdot, w^*)$  to the unknown function of interest  $g$ .

We are now in a position to discuss the hypotheses of interest to us. Ideally, we are interested in whether certain elements of  $X_t$  have any effect on  $Y_t$ ; in the least squares case this can be formulated as

$$(4) \quad \partial g(x)/\partial x_i = 0 \text{ for all } x \in \text{supp } \mu, i \in I_0,$$

where  $I_0$  is a set of indexes specifying the inputs whose relevance is at issue. However,  $g$  is unknown to us, so we cannot directly investigate (4). Instead, we will settle for investigating the validity of the hypothesis

$$(5) \quad \partial f(x, w^*)/\partial x_i = 0 \text{ for all } x \in \text{supp } \mu, i \in I_0,$$

which is accessible to us, given that  $f$  is known and  $w^*$  can be closely approximated.

There is a crucial difference between (4) and (5) that is important to bear in mind. Equation (4) expresses the hypothesis that the inputs specified by  $I_0$  do not enter into the relation between  $Y_t$  and  $X_t$  embodied by  $g$ . In contrast, equation (5) expresses the hypothesis that the inputs specified by  $I_0$  do not enter into the approximation to  $g$  provided by the network using weights  $w^*$ , which are optimal in the sense of (3). The former is thus a pure statement about the empirical phenomenon of interest, while the latter is a statement about the relationship of our network model to the phenomenon of interest. If our network model delivers a sufficiently good approximation, then there may be little or no difference between the two; but the presence of such a difference can be ignored only at the peril of drawing unsupported inferences.

In fact, we see that the function  $g$  does not enter equation (2) explicitly, but the preceding discussion justifies testing the hypothesis (5) regardless of whether an explicit  $g$  is available or not. In general, then, we interpret a test of (5) as testing the hypothesis that the inputs specified by  $I_0$  do not enter the approximation to the target itself ( $Y_t$ ) provided by the network using weights  $w^*$  (i.e.,  $f(X_t, w^*)$ ), which are optimal in the sense of (2). See Chen & White (1999) for a treatment of inference about  $g$  itself using artificial neural networks.

**2.2. Tests Based on Partial Derivatives.** A number of equivalent ways of formulating (5) are available to us. We first consider the quantity

$$m^* = \sum_{i \in I_0} \int f_i(x, w^*)^2 d\mu(x),$$

where  $f_i(\cdot, w)$  denotes  $\partial f(\cdot, w)/\partial x_i$ . Equation (5) is true if and only if  $m^* = 0$ .

Two obstacles prevent us from calculating  $m^*$  directly. The first is that  $w^*$  is unknown; however,  $\hat{w}_n$  consistently estimates  $w^*$ , so we may replace  $w^*$  with  $\hat{w}_n$ . The second is that  $\mu$ , the distribution of  $X_t$ , is unknown; however,  $\mu$  is consistently estimated by the empirical distribution  $\hat{\mu}_n$  which places mass  $n^{-1}$  at every realization observed in a training sample of size  $n$ . A computable statistic is therefore

$$\begin{aligned} \hat{m}_n &= n^{-1} \sum_{t=1}^n \sum_{i \in I_0} f_i(X_t, \hat{w}_n)^2 \\ &= \int \sum_{i \in I_0} f_i(x, \hat{w}_n)^2 d\hat{\mu}_n(x). \end{aligned}$$

Even when  $m^*$  is truly zero, we will not observe  $\hat{m}_n = 0$  because of random sampling variation. Nevertheless, we should expect  $\hat{m}_n$  to be close to zero when (5) is true; values far from zero are incompatible with (5). To determine how far from zero is too far for compatibility with (5), we can make use of the sampling distribution of  $\hat{m}_n$ . The following result provides this distribution approximately for large training samples. We let  $\nabla$  denote the gradient operator with respect to weights  $w$ . Absolute values  $|\cdot|$  are understood to be taken element by element.

**Theorem 2.1:** *Let  $\{X_t\}$  be an iid sequence of random  $k \times 1$  vectors with distribution  $\mu$ , and let  $\{\hat{w}_n\}$  be a sequence of random vectors taking values in the compact set  $W \subset \mathbb{R}^p$  such that*

$$\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, C^*) \text{ as } n \rightarrow \infty,$$

where  $w^*$  is a finite vector interior to  $W$ ,  $C^*$  is a finite positive semi-definite covariance matrix,  $N(0, C^*)$  denotes the normal distribution with mean zero and covariance matrix  $C^*$ , and  $\xrightarrow{d}$  denotes convergence in distribution.

Let  $m : \mathbb{R}^k \times W \rightarrow \mathbb{R}$  be a function such that for each  $x$  in  $\text{supp } \mu$ ,  $m(x, \cdot) : W \rightarrow \mathbb{R}$  is continuously differentiable of order 2, and for each  $w \in W$ ,  $m(\cdot, w) : \mathbb{R} \rightarrow \mathbb{R}$  is measurable;  $m(x, w^*) = 0$  and  $\nabla m(x, w^*) = 0$  for all  $x$  in  $\text{supp } \mu$ ; and  $\{\nabla^2 m(X_t, w)\}$  obeys the weak uniform law of large numbers, i.e.  $\sup_{w \in W} |\sum_{i=1}^n [\nabla^2 m(X_t, w) - E(\nabla^2 m(X_t, w))]| = o_p(1)$ .

Then

$$\sum_{t=1}^n m(X_t, \hat{w}_n) = n(\hat{w}_n - w^*)^T M^* (\hat{w}_n - w^*) + o_p(1)$$

and

$$n(\hat{w}_n - w^*)^T M^* (\hat{w}_n - w^*) \xrightarrow{d} N_2(0, C^*; M^*),$$

where  $N_2(0, C^*; M^*)$  is the mixture of independent  $\chi^2$ s defined by White (1994, Lemma 8.2), with  $M^* = E(\nabla^2 m(X_t, w^*)/2)$ , so that also

$$\sum_{t=1}^n m(X_t, \hat{w}_n) \xrightarrow{d} N_2(0, C^*; M^*). \quad \square$$

In our application,  $m(x, w)$  corresponds to  $\sum_{i \in I_0} f_i(x, w)^2$ .

In this result, the requirement that  $m(x, w^*) = 0$  for all  $x$  in  $\text{supp } \mu$  can be regarded as the null hypothesis. We can therefore apply this result to obtain the large sample distribution of a considerable range of test statistics of interest under a suitable null hypothesis. Theorem 2 of White (1989) provides conditions ensuring that  $\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, C^*)$ . It suffices to impose the following further conditions, where we write

$$l(Z_t, w) \equiv \pi(Y_t, f(X_t, w)).$$

**Assumption A.2:** (b) For each  $x$  in  $\text{supp } \mu$ ,  $f(x, \cdot)$  is continuously differentiable of order 2 on  $W$ .

**Assumption A.3:** (b) For each  $y$  in  $\mathbb{R}$ ,  $\pi(y, \cdot)$  is continuously differentiable of order 2 on  $\mathbb{R}$ .

**Assumption A.5:** (a) The elements of  $|\nabla l(Z_t, w)|$  and  $|\nabla^2 l(Z_t, w)|$  are dominated by functions integrable with respect to  $v$ ; (b) The elements of  $|\nabla l(Z_t, w) \nabla l(Z_t, w)^T|$  are dominated by functions integrable with respect to  $v$ .

**Assumption A.6:** (a)  $E(\nabla l(Z_t, w^*)^T \nabla l(Z_t, w^*)) < \infty$ ; (b)  $B^* \equiv E[\nabla l(Z_t, w^*) \nabla l(Z_t, w^*)^T]$  is nonsingular; (c)  $A^* \equiv E[\nabla^2 l(Z_t, w^*)]$  is nonsingular.

With Assumptions A.1-A.6, we have  $\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, C^*)$  with  $C^* = A^{*-1} B^* A^{*-1}$ .

We apply Theorem 2.1 by taking  $m(x, w) = \sum_{i \in I_0} (f_i(x, w))^2$ . To ensure that  $m$  is twice differentiable in  $w$ , we must strengthen A.2 with

**Assumption A.2:** (c) For each  $w$  in  $W$ ,  $f(\cdot, w)$  is continuously differentiable on an open set containing  $\text{supp } \mu$  with partial derivatives  $f_i(\cdot, w) \equiv \partial f(\cdot, w) / \partial x_i, i = 1, \dots, k$ ; and (d) for each  $x$  in  $\text{supp } \mu$ ,  $f_i(x, \cdot)$  is continuously differentiable of order 2 on  $W, i = 1, \dots, k$ .

With  $m(x, w) = \sum_{i \in I_0} (f_i(x, w))^2$ , we have  $\nabla m(x, w^*) = 2 \sum_{i \in I_0} \nabla f_i(x, w^*) f_i(x, w^*) = 0$  for all  $x \in \text{supp } \mu$  as required, under (5). To ensure that  $\{\nabla^2 m(X_t, w)\}$  obeys the weak law of large numbers it suffices to impose

**Assumption A.7:** For  $i \in I_0$ , the elements of  $|\nabla f_i(X_i, w) \nabla f_i(X_t, w)^T + \nabla^2 f_i(X_t, w) f_i(X_t, w)|$  are dominated by functions integrable with respect to  $\mu$ .

For the single hidden layer feedforward network case, we have

$$f_i(x, w) = \sum_{j=1}^h w_{0j} w_{1j} \psi(\tilde{x}^T w_{1j}),$$

where  $\psi'$  denotes the derivative of  $\psi$ . A simple condition that ensures that A.7 holds when  $E(X_t^T X_t) < \infty$  is to require that the first three derivatives of  $\psi$  exist and are bounded. As  $\psi$  is under our control, we can easily guarantee this holds by choosing  $\psi$  to be an activation function like the logistic or hyperbolic tangent.

We can now state a corollary to Theorem 2.1 describing the distribution of a statistic suitable for testing (5).

**Corollary 2.2:** *Suppose Assumptions A.1-A.5(a), A.6 and A.7 hold. Then if (5) holds we have*

$$n\hat{m}_n = \sum_{t=1}^n \sum_{i \in I_0} (f_i(X_t, \hat{w}_n))^2 = n(\hat{w}_n - w^*)^T M^* (\hat{w}_n - w^*) + o_p(1),$$

and  $n\hat{m}_n \rightarrow N_2(0, C^*; M^*)$  where  $C^* = A^{*-1} B^* A^{*-1}$

$$\text{and } M^* = E \left( \sum_{i \in I_0} [\nabla f_i(X_t, w^*) \nabla f_i(X_t, w^*)^T + \nabla^2 f_i(X_t, w^*) f_i(X_t, w^*)] \right). \quad \square$$

Observe that when (5) holds,  $M^*$  simplifies to

$$M_0^* = E \left( \sum_{i \in I_0} \nabla f_i(X_t, w^*) \nabla f_i(X_t, w^*)^T \right).$$

We define  $M^*$  as we have, because the difference between  $M^*$  and  $M_0^*$  becomes relevant under the alternative.

A challenge in applying this corollary is that the asymptotic  $\chi^2$  mixture distribution  $N_2(0, C^*; M^*)$  is not one of the familiar tabulated distributions, so that computing a critical value appropriate for testing (5) is not an entirely simple matter.

Several options are available to us. Because of the straightforwardness of the required computations, we elect to make use of Efron's (1983) bootstrap method. The idea underlying the bootstrap is appealingly simple. Suppose we are interested in a statistic that can be written as a Von Mises functional of the empirical distribution, *i.e.* as  $T(\hat{v}_n)$ . The relevant statistic in our application is  $T(\hat{v}_n) = M^{*\frac{1}{2}}\hat{w}_n$ , where  $M^{*\frac{1}{2}}M^{*\frac{1}{2}} = M^*$ . This is a functional of  $\hat{v}_n$ , as  $\hat{w}_n$  is obtained by solving the problem (1),

$$\min_{w \in W} n^{-1} \sum_{t=1}^n \pi(Y_t, f(X_t, w)) = \int \pi(y, f(x, w)) d\hat{v}_n(z).$$

Because  $\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, C^*)$  we know

$$\sqrt{n}(T(\hat{v}_n) - T(v)) = M^{*\frac{1}{2}}\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, M^{*\frac{1}{2}}C^*M^{*\frac{1}{2}}),$$

where  $T(v) = M^{*\frac{1}{2}}w^*$ , viewing  $w^*$  as the solution to the problem (2),

$$\min_{w \in W} \lambda(w) \equiv \int \pi(y, f(w, w)) dv(z).$$

Efron's (1983) bootstrap exploits the properties of the statistics

$$\hat{b}_n^* = \sqrt{n}(T(\hat{v}_n^*) - T(\hat{v}_n)) = M^{*\frac{1}{2}}\sqrt{n}(\hat{w}_n^* - \hat{w}_n),$$

where  $\hat{v}_n^*$  is the "re-sampled" empirical distribution obtained by drawing an iid sample of size  $n$  (*i.e.* with replacement) from the original empirical distribution  $\hat{v}_n$ , and  $\hat{w}_n^*$  is the corresponding vector of network weights. Under general conditions, it can be established that

$$\hat{b}_n^* \xrightarrow{d} N(0, M^{*\frac{1}{2}}C^*M^{*\frac{1}{2}})$$

almost surely (*i.e.*, for almost every sample from  $v$  and re-sample from  $\hat{v}_n$ ). Consequently,  $\hat{b}_n^*$  has the same distribution asymptotically as our statistic of interest. What's more, we can compute as many realizations of  $\hat{b}_n^*$  as we wish, permitting us to build up an estimate of the distribution of our statistic under the null hypothesis. The same is true for any statistic that is a continuous function of  $\hat{b}_n^*$ , such as  $\hat{b}_n^{*T} \hat{b}_n^*$ , which has the same limiting distribution as our test statistic:

$$n(\hat{w}_n - w^*)^T M^* (\hat{w}_n - w^*)$$

and

$$\sum_{t=1}^n \sum_{i \in I_0} f_i(X_t, \hat{w}_n)^2.$$

This gives us the desired procedure for obtaining the needed asymptotic distribution.

Once we have obtained a sufficiently detailed estimate of the distribution of  $\hat{b}_n^{*T} \hat{b}_n^*$  we can use this distribution to conduct a test of the hypothesis (5) of approximate level  $\alpha$ , using a one-sided  $(1-\alpha)\%$  acceptance region: if the original test statistics falls beyond the  $(1-\alpha)$  percentile of  $\hat{b}_n^{*T} \hat{b}_n^*$ , then reject (5).

Implementing the bootstrap thus involves the following steps:

1) Using the original sample, solve (1) to get weights  $\hat{w}_n$ ;

Draw a sample  $\{Z_1^*, \dots, Z_n^*\}$  with replacement from the original sample and compute re-sampled weights  $\hat{w}_n^*$  by solving (1) with  $Z_t^*$  replacing  $Z_t$ ;

Compute bootstrap statistics  $\hat{b}_n^{*T} \hat{b}_n^*$ , where

$$\hat{b}_n^* = M^{*\frac{1}{2}} \sqrt{n}(\hat{w}_n^* - \hat{w}_n);$$

Replicate steps 2) and 3) many times, say  $N = 100$  or  $N = 1,000$  times;

Using the statistics  $\hat{b}_n^{*T} \hat{b}_n^*$  created in step 4), compute the one-sided  $(1-\alpha)\%$  acceptance region  $(-\infty, c_\alpha)$ , where  $c_\alpha$  is the  $(1-\alpha)$  percentile of the size  $N$  sample of  $\hat{b}_n^{*T} \hat{b}_n^*$ . Reject (5) if the original test statistic  $n\hat{m}_n = \sum_{t=1}^n \sum_{i \in I_0} f_i(X_t, \hat{w}_n)^2$  exceeds  $c_\alpha$ ; otherwise, fail to reject.

As it stands, the procedure just given cannot be implemented, as step 3) requires knowledge of  $M^*$ . This matrix can be replaced with a consistent estimator; several possibilities are available. Observe, however, that computation of an estimate of  $M^*$  can be quite complicated, as (mixed) derivatives of at least second order are involved. Fortunately, computation of an estimate of  $M^*$  can be avoided entirely by working with a re-sampled version of the statistic  $\sum_{t=1}^n m(X_t, \hat{w}_n)$ . Consider

a two term Taylor expansion

$$\begin{aligned}
\sum_{t=1}^n m(X_t^*, \hat{w}_n^*) &= \sum_{t=1}^n m(X_t^*, \hat{w}_n) + \sum_{t=1}^n \nabla^T m(X_t^*, \hat{w}_n) (\hat{w}_n^* - \hat{w}_n) \\
&\quad + n(\hat{w}_n^* - \hat{w}_n)^T [n^{-1} \sum_{t=1}^n \nabla^2 m(X_t^*, \bar{w}_n^*)] (\hat{w}_n^* - \hat{w}_n) / 2 \\
&= \sum_{t=1}^n m(X_t^*, \hat{w}_n) + \sum_{t=1}^n \nabla^T m(X_t^*, \hat{w}_n) (\hat{w}_n^* - \hat{w}_n) + \hat{b}_n^{*T} \hat{b}_n^* + o_{P^*}(1).
\end{aligned}$$

In the first equation,  $[n^{-1} \sum_{t=1}^n \nabla^2 m(X_t^*, \bar{w}_n^*)]$  denotes the Hessian, each row of which is evaluated at a different mean value lying between  $\hat{w}_n^*$  and  $\hat{w}_n$ . In the second equation, the quadratic form is replaced by  $\hat{b}_n^{*T} \hat{b}_n^*$ , leaving a remainder that vanishes in probability. We write  $o_{P^*}(1)$  to denote the fact that the probability measure is a joint measure reflecting both the original random sampling and the re-sampling.

From this it follows that step 3) can be replaced by

3') In place of  $\hat{b}_n^{*T} \hat{b}_n^*$ , compute the bootstrap statistic

$$\hat{\mathcal{B}}_n^* \equiv \sum_{t=1}^n m(X_t^*, \hat{w}_n^*) - \sum_{t=1}^n m(X_t^*, \hat{w}_n) - \sum_{t=1}^n \nabla^T m(X_t^*, \hat{w}_n) (\hat{w}_n^* - \hat{w}_n)$$

All other steps remain the same, with  $\hat{\mathcal{B}}_n^*$  replacing  $\hat{b}_n^{*T} \hat{b}_n^*$ . The two terms appearing after the resampled version of the original statistic can be thought of as a bias-removing recentering. No longer is it necessary to perform computations involving  $\nabla^2 m$ ; instead, we must deal only with the more tractable  $\nabla m$ . In fact, a further simplification is possible. A two term Taylor expansion of  $\sum_{t=1}^n m(X_t, \hat{w}_n^*)$  parallel to that above (observe that  $X_t$  appears in place of  $X_t^*$ ) shows that we also have

$$\begin{aligned}
\sum_{t=1}^n m(X_t, \hat{w}_n^*) - \sum_{t=1}^n m(X_t, \hat{w}_n) - \sum_{t=1}^n \nabla^T m(X_t, \hat{w}_n) (\hat{w}_n^* - \hat{w}_n) \\
= \hat{b}_n^{*T} \hat{b}_n^* + o_{P^*}(1).
\end{aligned}$$

Step 3) can therefore be replaced by

3'') In place of  $\hat{b}_n^{*T} \hat{b}_n^*$ , compute the bootstrap statistic

$$\bar{\mathcal{B}}_n^* \equiv \sum_{t=1}^n m(X_t, \hat{w}_n^*) - \sum_{t=1}^n m(X_t, \hat{w}_n) - \sum_{t=1}^n \nabla^T m(X_t, \hat{w}_n) (\hat{w}_n^* - \hat{w}_n).$$

Note the difference:  $X_t$  replaces  $X_t^*$ , so that  $\sum_{t=1}^n m(X_t, \hat{w}_n)$  and  $\sum_{t=1}^n \nabla^T m(X_t, \hat{w}_n)$  need not be recomputed with each re-sample. We have the following result.

**Theorem 2.3:** *Suppose Assumptions A.1 -A.7 hold. Then with probability 1,*

$$\hat{\mathcal{B}}_n^* \xrightarrow{d} N_2(0, C^*; M^*) \text{ and } \bar{\mathcal{B}}_n^* \xrightarrow{d} N_2(0, C^*; M^*).$$

*i.e. the bootstrap “works”.* □

Note that we do not assume that (5) holds in stating Theorem 2.3. If (5) does indeed hold, then the primary consequence is that  $M^* = M_0^*$ , so that the bootstrapped distribution is precisely the null distribution we seek. Tests based on this distribution thus have proper size. On the other hand, if (5) is false, then we still have the bootstrapped distribution  $N_2(0, C^*; M^*)$  given in the theorem, but now  $M^* \neq M_0^*$ . This is no longer the null distribution, but because the alternative is true there is no effect on size. There is, instead, a potential effect on the power of the resulting test, analogous to the effect of using a covariance matrix estimator consistent under the null but not under the alternative in performing a classical Lagrange Multiplier test in a parametric model (*e.g.* Engle (1982)). Nevertheless, because the bootstrap distribution gives a critical value  $c_\alpha$  bounded in probability under the alternative and because  $n\hat{m}_n \rightarrow \infty$  *a.s.* under the alternative, the test based on  $c_\alpha$  still is consistent, *i.e.* has power approaching unity as  $n \rightarrow \infty$ .

If it were desired to obtain a statistic based on  $M_0^*$  under the alternative, one could compute re-sampled statistics

$$n(\hat{w}_n^* - \hat{w}_n)^T \hat{M}_{0n}(\hat{w}_n^* - \hat{w}_n),$$

where

$$\hat{M}_{0n} = n^{-1} \sum_{t=1}^n \left[ \sum_{i \in I_0} \nabla f_i(X_t, \hat{w}_n) \nabla f_i(X_t, \hat{w}_n)^T \right].$$

**2.3. Tests Based on Comparing Networks.** An alternative method for investigating the relevance of the set of inputs indexed by  $I_0$  is to compare the outputs of two suitable networks, one trained with the inputs in question included and the other trained with the inputs in question excluded. The idea is that if the inputs in question are truly irrelevant, then the network outputs will be the same for all  $x$  in  $\text{supp } \mu$ , while if the inputs in question are truly relevant then the network outputs will differ, at least for  $x$  in some set having positive probability.

Care must be taken, however, to ensure that the relevance of the inputs in question is the only reason for divergence between the network outputs, as opposed, for example, to reasons involving

the relative flexibility of the two networks. This is the reason for the qualifier “suitable” in the preceding paragraph. The next result provides appropriate structure.

**Proposition 2.4:** (a) Let  $f_1 : \mathbb{R}^k \times W_1 \rightarrow \mathbb{R}$  and  $f_2 : \mathbb{R}^k \times W_2 \rightarrow \mathbb{R}$  satisfy Assumptions A.2(a) and (c), where  $W_1$  and  $W_2$  are compact subsets of  $\mathbb{R}^{p_1}$  and  $\mathbb{R}^{p_2}$  respectively,  $p_1, p_2 \in \mathbb{N}$ .

(b) Suppose that for every  $w_1$  in  $W_1$  and  $x$  in  $\text{supp } \mu$  we have  $\partial f_1(x, w_1)/\partial x_i = 0, i \in I_0$ , while there exists a non-empty subset  $W_2^0$  of  $W_2$  such that for each  $w_2$  in  $W_2^0$  there exists a Borel subset  $A(w_2)$  of  $\text{supp } \mu$  with  $\mu(A(w_2)) > 0$  such that for all  $x$  in  $A(w_2)$  we have  $\partial f_2(x, w_2)/\partial x_i \neq 0$  for some  $i \in I_0$ .

(c) Suppose further that for every  $w_2$  in  $W_2$  such that  $\partial f_2(x, w_2)/\partial x_i = 0$  for all  $x$  in  $\text{supp } \mu$  and  $i \in I_0$  there exists  $w_1$  in  $W_1$  such that  $f_1(x, w_1) = f_2(x, w_2)$  for all  $x$  in  $\text{supp } \mu$ .

(d) Let Assumptions A.3 and A.4 hold for  $\pi$  and  $f_1$  and for  $\pi$  and  $f_2$  respectively, with unique minimizers  $w_1^*$  and  $w_2^*$  respectively. Suppose that there exists a non-empty subset  $W_2^1$  of  $W_2$  such that for all  $w_2 \in W_2^1$

$$\lambda_2(w_2) \equiv \int \pi(y, f_2(x, w_2)) dv(z) \leq \lambda_1(w_1^*) = \int \pi(y, f_1(x, w_1^*)) dv(z).$$

Further, if the subset of  $W_2^1$  such that  $\lambda_2(w_2) = \lambda_1(w_1^*)$  is non-empty, then this subset contains a further subset  $W_2^{10}$  such that for all  $w_2$  in  $W_2^{10}$  we have  $\partial f_2(x, w_2)/\partial x_i = 0$  for all  $x$  in  $\text{supp } \mu$   $i \in I_0$ .

Then

$$(6) \quad \partial f_2(x, w_2^*)/\partial x_i = 0 \text{ for all } x \in \text{supp } \mu, i \in I_0$$

if and only if

$$f_1(x, w_1^*) = f_2(x, w_2^*) \text{ for all } x \in \text{supp } \mu. \quad \square$$

Thus, for suitable  $f_1$  and  $f_2$ , equality of optimal network outputs is equivalent to the irrelevance of the inputs indexed by  $I_0$  in the optimal network 2. Although the conditions for suitability take some space to state, the requirements are straightforward. Condition (a) simply says that both networks must be as well behaved as the single network previously considered. Condition (b) says that network 1 always ignores the inputs in question while network 2 may respond to them. Condition (c) says that whenever network 2 does ignore the inputs in question, then network 1 is capable of producing an output identical to that of network 2. Finally, Condition (d) says that

there is always a set of weights for network 2 at least as good as the best set of weights for network 1, and if performance is identical, then network 2 can achieve this level of performance by ignoring the inputs in question. Thus, network 2 is in this sense at least as flexible as network 1. For brevity, whenever two functions  $f_1$  and  $f_2$  satisfy all these conditions we formally designate them “suitable.”

Comparison of suitable output functions for the purpose of testing the relevance of particular inputs can thus be based on the quantity

$$m^* = \int (f_1(x, w_1^*) - f_2(x, w_2^*))^2 d\mu(x).$$

Replacing  $w_1^*, w_2^*$  and  $\mu$  with consistent estimators as before gives

$$\hat{m}_n = n^{-1} \sum_{t=1}^n (f_1(X_t, \hat{w}_{1n}) - f_2(X_t, \hat{w}_{2n}))^2,$$

where  $\hat{w}_{1n}$  and  $\hat{w}_{2n}$  are solutions of (1) with  $f_1$  and  $f_2$  replacing  $f$ , respectively.

The analysis of this choice for  $\hat{m}_n$  parallels that for the previous case exactly, and Theorem 2.1 applies, now with  $m(x, w) = (f_1(x, w_1) - f_2(x, w_2))^2$ , setting  $w = (w_1^T, w_2^T)^T$ ,  $p = p_1 + p_2$ .

To apply Theorem 2.1 we must ensure the asymptotic normality of  $\sqrt{n}(\hat{w}_n - w^*)$ , with  $\hat{w}_n = (\hat{w}_{1n}^T, \hat{w}_{2n}^T)^T$ . The following condition suffices.

**Assumption B.1:** Assumption A.1 holds, and Assumptions A.2(a,b), A.3(a,b). A.4 - A.6 hold for suitable network output functions  $f_1$  and  $f_2$  respectively.

The analog of Assumption A.2(d) is that  $f_1$  and  $f_2$  are twice differentiable in the weights, but this is already imposed in A.2(b). It remains only to ensure the uniform law of large numbers holds for  $\{\nabla^2 m(X_t, w)\}$ . We impose

**Assumption B.2:** The elements of  $|\nabla_1 f_1(X_t, w) \nabla_1 f_1(X_t, w_1)^T + \nabla_1^2 f_1(X_t, w_1)(f_1(X_t, w_1) - f_2(X_t, w_2))|$ ,  $|\nabla_1 f_1(X_t, w_1) \nabla_2 f_2(X_t, w_2)^T|$  and  $|\nabla_2 f_2(X_t, w_2) \nabla_2 f_2(X_t, w_2)^T + \nabla_2^2 f_2(X_t, w_2)(f_1(X_t, w_1) - f_2(X_t, w_2))|$  are dominated by functions integrable with respect to  $\mu$ .

The large sample distribution of  $n\hat{m}_n$  is given by our next result.

**Corollary 2.5:** *Suppose Assumptions B.1 and B.2 hold. Then if (6) holds we have  $n\hat{m}_n = \sum_{t=1}^n (f_1(X_t, \hat{w}_{1n}) - f_2(X_t, \hat{w}_{2n}))^2 = n(\hat{w}_n - w^*)^T M^*(\hat{w}_n - w^*) + o_P(1)$  and  $n\hat{m}_n \xrightarrow{d} N_2(0, C^*; M^*)$ , where*

$$C^* = \begin{bmatrix} A_1^{*-1} B_1^* A_1^{*-1} & A_1^{*-1} B_{12}^* A_2^{*-1} \\ A_2^{*-1} B_{21}^* A_1^{*-1} & A_2^{*-1} B_2^* A_2^{*-1} \end{bmatrix},$$

with  $B_{12}^* = E(\nabla_1 l_1(Z_t, w_1^*) \nabla_2 l_2(Z_t, w_2^*)^T)$  and with  $M^*$  as given in the proof of Corollary 2.5 (See *Mathematical Appendix*).  $\square$

Again we have a simpler form  $M_0^*$  for  $M^*$  under (6).

Reasoning entirely parallel to that for the previous case supports the use of the bootstrap. Implementation involves the following:

- 1) Using the original sample, solve (1) for  $f_1$  and  $f_2$  to get weights  $\hat{w}_{1n}$  and  $\hat{w}_{2n}$ ;
- 2) Draw a sample  $\{Z_1^*, \dots, Z_n^*\}$  with replacement from the original sample, and compute re-sampled weights  $\hat{w}_n^* = (\hat{w}_{1n}^{*T}, \hat{w}_{2n}^{*T})^T$  by solving (1) for  $f_1$  and  $f_2$  with  $Z_t^*$  replacing  $Z_t$ ;
- 3) Compute bootstrap statistics

$$\hat{A}_n^* \equiv \sum_{t=1}^n m(X_t^*, \hat{w}_t^*) - \sum_{t=1}^n m(X_t^*, \hat{w}_n) - \sum_{t=1}^n \nabla^T m(X_t^*, \hat{w}_n) (\hat{w}_n^* - \hat{w}_n)$$

or

$$\tilde{A}_n^* \equiv \sum_{t=1}^n m(X_t, \hat{w}_n^*) - \sum_{t=1}^n m(X_t, \hat{w}_n) - \sum_{t=1}^n \nabla^T m(X_t, \hat{w}_n) (\hat{w}_n^* - \hat{w}_n),$$

where  $m(x, w) \equiv (f_1(x, w_1) - f_2(x, w_2))^2$ ;

- 4) Replicate steps 2) and 3) many times, say  $N = 100$  or  $N = 1,000$  times;
- 5) Using the statistics  $\hat{A}_n^*$  or  $\tilde{A}_n^*$  created in step 4) compute a one-sided  $(1-\alpha)\%$  acceptance region  $(-\infty, c_\alpha)$ , where  $c_\alpha$  is the  $(1-\alpha)$  percentile of the size  $N$  sample of  $\hat{A}_n^*$  or  $\tilde{A}_n^*$ .
- 6) Reject (6) if the original test statistic  $n\hat{\eta}_n$  exceeds  $c_\alpha$ ; otherwise, fail to reject. Use of this bootstrap procedure is justified by the following result:

**Theorem 2.6:** *Suppose Assumptions B.1 and B.2 hold. Then with probability 1*

$$\hat{A}_n^* \xrightarrow{d} N_2(0, C^*; M^*) \text{ and } \tilde{A}_n^* \xrightarrow{d} N_2(0, C^*; M^*)$$

where  $C^*$  and  $M^*$  are as in Corollary 2.5. Thus, the bootstrap works.  $\square$

The comments regarding  $M_0^*$  and  $M^*$  following Theorem 2.3 apply here as well.

### 3. EFFECTS OF PRIOR MODEL SELECTION

In the preceding section, the network architectures  $f$ ,  $f_1$ , and  $f_2$  forming the basis for our test procedures were assumed to be given. In practice, however, the appropriate complexity for the network is usually unknown, for example, how many hidden units to include in how many hidden

layers. Provided that they are properly chosen and implemented, data-driven complexity determination procedures can be used to arrive at appropriate network complexity without disturbing either the properties of the statistic or those of the bootstrap. Such procedures as penalized log-likelihood complexity determination (Sin & White (1996)) or cross-validation methods (Plutowski, Sakata & White (1994)) can be shown to have the needed properties under conditions compatible with those of the previous section.

In order to leave the procedures of the previous section undisturbed, the complexity determination procedure essentially must operate in such a way that a desired architecture is selected with probability approaching one as  $n$  increases. To formalize this, suppose there is a finite set  $F$  of network output functions  $f(\cdot, \cdot; \gamma) : \mathbb{R}^k \times W_\gamma \rightarrow \mathbb{R}$ , indexed by integers  $\gamma$ . Without loss of generality, we may suppose that each weight space  $W_\gamma$  is a subset of some “master” space  $W$  of weights. Associated with each such output function is a measure of network complexity  $C_\gamma$ . Network architectures may have the same complexity  $C_\gamma$  without being identical. We write

$$F \equiv \{f(\cdot, \cdot; \gamma), \gamma \in \Gamma\},$$

where  $\Gamma$  is an index set,  $\Gamma = \{1, \dots, \bar{\gamma}\}$ , say, and the functions in  $F$  are understood to satisfy regularity conditions such as Assumption A.2, etc., appropriate to the context.

Let our measure of network performance now be represented as

$$\lambda(w, \gamma) = \int \pi(y, f(x, w; \gamma)) dv(z).$$

The parsimony principle dictates that we choose the smallest complexity network that optimizes this performance. This involves selecting  $w^*$  and  $\gamma^*$  such that

$$\lambda(w^*, \gamma^*) \leq \lambda(w, \gamma) \text{ for all } w \in W_\gamma, \gamma \in \Gamma,$$

$$C_{\gamma^*} \leq C_\gamma \text{ for all } \gamma \in \Gamma.$$

In the typically rare situation that a unique  $\gamma^*$  is not found, but there is instead more than one choice for  $\gamma$  that satisfies the above condition, we assume the existence of some final tie-breaking procedure that results in selection of a unique  $\gamma^*$ , so that the pair  $(w^*, \gamma^*)$  can be taken to be unique. The null hypothesis corresponding to (5) is now

$$f_i(x, w^*; \gamma^*) = 0 \text{ for all } x \in \text{supp } \mu, i \in I_0;$$

and that corresponding to (6) becomes

$$f_1(x, w_1^*; \gamma_1^*) = f_2(x, w_2^*; \gamma_2^*) \text{ for all } x \in \text{supp } \mu,$$

where  $f_1(\cdot, \cdot; \gamma_1^*)$ ,  $f_2(\cdot, \cdot; \gamma_2^*)$  are suitable network output functions.

If the optimal network architecture were known, we would base our test procedures on  $\hat{m}_n(\gamma^*)$ , where now

$$\hat{m}_n(\gamma) = n^{-1} \sum_{t=1}^n m(X_t, \hat{w}_n(\gamma); \gamma),$$

and  $\hat{w}_n(\gamma)$  solves

$$\min_{w \in W_\gamma} n^{-1} \sum_{t=1}^n \pi(Y_t, f(X_t, w; \gamma)).$$

Because  $\gamma^*$  is unknown, we must work with  $\hat{\gamma}_n$ , say, that indexes the network selected by the data-driven complexity selection mechanism. This leads to consideration of the statistic  $\hat{m}_n(\hat{\gamma}_n)$ . If  $P[\hat{\gamma}_n = \gamma^*] \rightarrow 1$  as  $n \rightarrow \infty$ , then it follows that

$$n\hat{m}_n(\gamma^*) - n\hat{m}_n(\hat{\gamma}_n) = o_P(1).$$

To see this, note that  $\hat{\gamma}_n = \gamma^*$  implies  $n\hat{m}_n(\gamma^*) = n\hat{m}_n(\hat{\gamma}_n)$ . Consequently  $|n\hat{m}_n(\gamma^*) - n\hat{m}_n(\hat{\gamma}_n)| > \epsilon$  implies  $\hat{\gamma}_n \neq \gamma^*$  for any  $\epsilon > 0$ , so that

$$P[|n\hat{m}_n(\gamma^*) - n\hat{m}_n(\hat{\gamma}_n)| > \epsilon] \leq P[\hat{\gamma}_n \neq \gamma^*] \rightarrow 0.$$

The asymptotic distribution of  $n\hat{m}_n(\gamma^*)$  is thus identical to that of  $n\hat{m}_n(\hat{\gamma}_n)$ . We can approximate the latter distribution simply by applying the results of the previous section with the architecture fixed at that indexed by  $\hat{\gamma}_n$ . This gives us a bootstrap test procedure that uses the data-determined network complexity, but has the same level (probability of wrongly rejecting the null hypothesis) asymptotically as if the optimal network complexity were known.

The key requirement, then, is that the complexity selection procedure deliver  $P[\hat{\gamma}_n = \gamma^*] \rightarrow 1$  as  $n \rightarrow \infty$  (*cf* Pötscher (1991)). In our foreign exchange application we select network complexity using the Schwarz Information Criterion (SIC) (Schwarz (1978)). This has the required property.

#### 4. IMPLEMENTATION DETAILS

For all subsequent analysis, we implement normalized versions of  $n\hat{m}_n$  and  $\hat{\mathcal{B}}_n^*$  of Section 2, where, following standard practice, the normalization is  $n^{-1}$ . To compute a test statistic and its  $(1 - \alpha)$  percentile under the null,  $c_\alpha$ , we proceed as follows:

- (1) All computations are double precision. We use the conjugate gradient search algorithm for training, where the fractional precision used for training is  $1.19209e - 07$  (floating point machine precision) and the search takes place over the transformed target rescaled to lie in  $[0, 1]$ . When training, we attempt to avoid the presence of local minima. In particular, whenever training terminates, training is immediately restarted at the point at which the algorithm claims to have found a minimum. As well, multiple restarts of the training algorithm from random initial weights are conducted where appropriate.
- (2) Network complexity (number of hidden units) is determined via SIC, is allowed to range from 1 through 10 hidden units, and is based on 10 restarts from random initial weights for each architecture. We use adaptive training which increases the number of restarts when increasing network complexity does not result in a better within-sample fit. If this occurs then the number of restarts is incremented by 10 and search is restarted with this new (higher) number of restarts. This is repeated until the within-sample-fit improves. For a given model, candidate weights for each architecture are saved based on these 10 restarts yielding a candidate SIC function. Out of concern that the weights for a particular architecture may be those for a local rather than a global minimum, 10 additional passes over the candidate SIC function are then conducted, each based on 10 restarts from random initial weights but *including* the previously saved weights with the best weights again being saved. The resulting weights minimizing the SIC function are then used to compute the test statistic  $\hat{m}_n$ .
- (3) We next draw a resample *with replacement* from  $\{X_t, Y_t\}$  which we denote  $\{X_t^*, Y_t^*\}$ . We estimate the network weights  $\hat{w}_n^*$  for this resample where the search algorithm is started from *initial values*  $\hat{w}_n$ . If  $\hat{w}_n$  indeed represents a global optimum, this ensures that the weights  $\hat{w}_n^*$  will be somewhat close to  $\hat{w}_n$ . Using the resampled weights  $\hat{w}_n^*$  and the original weights  $\hat{w}_n$  we then compute  $n^{-1}\hat{\mathcal{B}}_n^*$ .
- (4) We repeat this procedure 1,000 times obtaining 1,000 values of  $n^{-1}\hat{\mathcal{B}}_n^*$ , one for each resample. Using these 1,000 values generated from the *null distribution* of  $n\hat{m}_n$ , we obtain

the  $(1 - \alpha)$  percentile,  $c_\alpha$  and we also obtain the empirical  $p$ -value. Finally, we reject  $H_0$  if  $n\hat{m}_n > c_\alpha$ , otherwise we fail to reject.

## 5. MONTE CARLO

We implement a modest Monte Carlo experiment designed to gauge the finite-sample performance of our test procedure. We consider a simple linear data generating process (DGP) given by

$$\begin{aligned}
 (7) \quad y_t &= E[y_t | x_{t1}, x_{t2}] + \epsilon_t \\
 &= \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \epsilon_t.
 \end{aligned}$$

The inputs  $x_{t1}$  and  $x_{t2}$  are drawn from independent  $U[0, 1]$  distributions,  $\beta_0 = 1.0$ ,  $\beta_1 = 1.0$ ,  $\beta_2 \in [0, 1]$ , and  $\epsilon_t \sim iid N(0, \sigma^2)$  with  $\sigma = 0.1$ . We set the sample size at  $n = 100$ .

We consider our proposed test based on the feedforward network outlined in Section 2 with the network architecture automatically selected using the SIC criterion as outlined in Section 4. We also implement the standard regression  $t$ -test of significance based upon the correctly specified parametric model (Equation (7)) which shall serve as a benchmark.

At this point we are interested in two related issues: i) does the proposed test have correct level? ii) does the power of the test increase as the distance between the true value of  $\beta_2$  and its hypothesized value  $H_0 : \beta_2 = 0$  grows, and how does its power compare to that of the benchmark test?

When the data are generated under the null ( $\beta_2 = 0$ ) we can examine the test's level, while we can examine the test's power by permitting the parameter  $\beta_2$  to take on non-zero values. We repeat each experiment allowing  $\beta_2$  to be incremented by 0.01 [0.01, 0.02, ... 0.20] resulting in a total of 20 Monte Carlo experiments, one for each value of  $\beta_2$ . For each of the 20 Monte Carlo power experiments, 1,000 draws from the DGP are obtained and the test described in Section 2 is conducted for each draw.

We summarize the 20 Monte Carlo power experiments in the form of the power curve for our test which appears in Figure 1 (values of the empirical rejection frequencies appear in Appendix 11.). For each value of  $\beta_2$ , we compute the empirical rejection frequency for the test with a nominal level of  $\alpha = 0.05$ , and the power curve plots the rejection frequency versus the value of  $\beta_2$ . As the test is symmetric, we reflect the power curve on  $[-0.2, 0]$  to generate a traditional power curve.

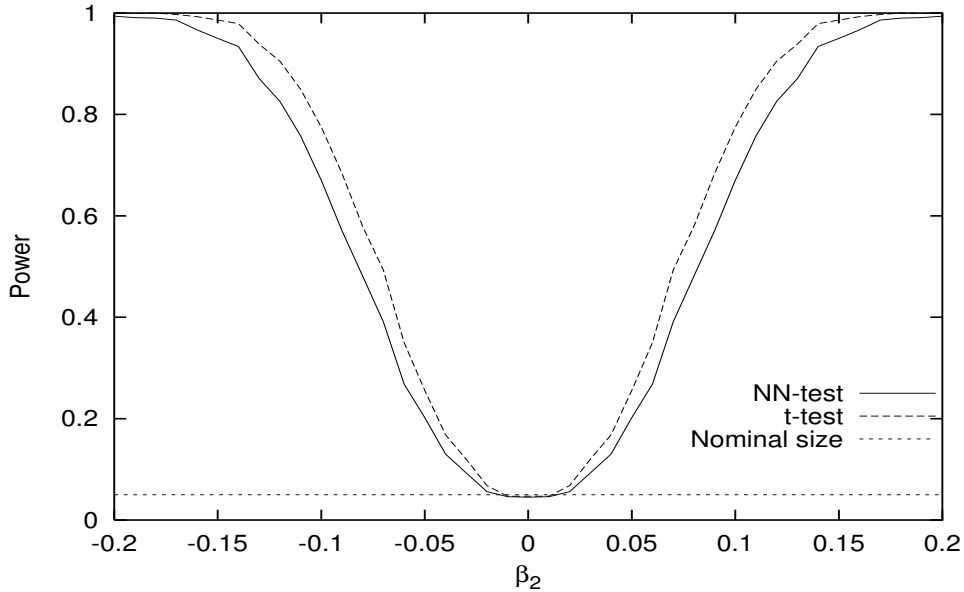


FIGURE 1. Power curve of the test procedure for the DGP given in Equation (7). The test is conducted with a nominal level of  $\alpha = 0.05$ . The uppermost curve is the power curve for a  $t$ -test for a correctly specified parametric model, the curve directly below that is the power curve for the proposed test, and the dashed horizontal line is the nominal level of the test ( $\alpha = 0.05$ ).

We are pleased with the behavior of our test in this small sample setting given that, unlike the  $t$ -test which is based on the correct parametric model, we assume that *nothing* is known about the functional form of the data generating process and we use a *data-driven* method of architecture selection. Of course, the test's power is expected to be lower than that based on correctly guessing the functional form of the DGP (the  $t$ -test based upon the correctly specified linear model) which we observe. However, the loss with respect to not knowing the DGP is surprisingly small. With respect to the test's performance we note first that the test has correct level and second that power increases as the distance between the true value of  $\beta_2$  and its hypothesized value  $H_0 : \beta_2 = 0$  grows. In summary, this modest Monte Carlo experiment indicates that the test is behaving properly.

Given that the test is well-behaved in a small-sample setting, we now turn our attention to an application involving the prediction of exchange rates.

## 6. APPLICATION: MARKET EFFICIENCY AND THE UNPREDICTABILITY OF EXCHANGE RATES

The market efficiency hypothesis applied to foreign exchange rates is typically interpreted to mean that there is no information contained in past percentage changes that is relevant for the prediction of future percentage changes. The linear unpredictability of exchange rates has a long

history beginning with the early work on efficient markets by Fama (1965) and Cootner (1964), while more recent approaches to modeling unpredictability include Diebold & Nason (1990), Kuan & Liu (1995), Moody & Wu (1998), and the volumes by Dunis & Zhou (1998), Abu-Mostafa, LeBaron, Lo & Weigend (2000), Dacorogna, Gencay, Muller, Pictet & Olsen (2001, in press) and the references therein. The literature on the unpredictability of exchange rates is vast, and it is fair to say that results are mixed and somewhat inconclusive. We briefly review a few of the relevant findings from this literature before proceeding to our application.

In their frequently cited article, Diebold & Nason (1990) compare out-of-sample predictions of locally weighted regression (Cleveland, Devlin & Grosse (1988)) versus a parametric random walk specification for forecasting weekly exchange rate series, and they conclude “Our findings bode poorly for recent conjectures that exchange rates contain nonlinearities exploitable for enhanced point prediction”. On the other hand, Kuan & Liu (1995) use feedforward and recurrent neural networks to forecast daily exchange rates, and they find that “for certain exchange rate series, some selected network models have significant market timing ability and/or significantly lower out-of-sample mean squared prediction error relative to the random walk model”. More recently, Moody & Wu (1998) use a state space approach to model high frequency exchange rate data (roughly 4 quotes per minute) and conclude that “trends may in fact exist in noisy FX data. The conventional random walk models of efficient market theory do not explain the correlation structures that are present in the high frequency data. Further studies are required.”

In contrast to approaches that compare out-of-sample predictions to those generated by the random walk model, we shall focus on directly testing for unpredictability. Our approach will be similar in spirit to that of LeBaron (1998) who conducts tests of significance on models inspired by technical trading rules. We address the sensitivity of our results to changes in data frequencies by sampling at daily, weekly, and monthly frequencies (we avoid high frequency data at this point simply due to the vast amounts of data involved), while we address the sensitivity of our results to the estimation period by considering a variety of periods of analysis.

We obtain historical data for nominal daily dollar spot exchange rates for the G7 countries from the Federal Reserve Board of St. Louis. All data are measured in units of foreign currency per US dollar. Each daily series begins on 1/4/71 and runs through 1/18/01 with the exception of France,

Germany, and Italy which run through 12/29/98<sup>1</sup>. This daily data set was used to construct the weekly and monthly series in the following manner; the weekly series were constructed from the Friday close unless there was no data available, in which case the previous day for which data exists was used; the monthly series were constructed from the Friday close occurring every four weeks unless there was no data available, in which case the previous day for which data exists was used.

Having created the weekly and monthly series from the daily data for each country, we then consider a variety of periods of analysis. We first consider the full data set, and we then move to shorter periods in order to examine periods of analysis which may be more homogeneous than the entire sample. For the shorter periods we consider the three sub-decades and we also consider the six five-year sub-periods. For each of the six countries this gives us ten data sets for each of the daily, weekly, and monthly data frequencies. The data are available from the authors upon request.

Following Diebold & Nason (1990), we construct percentage exchange rate changes ( $\Delta \log S_t$ ) thereby avoiding potential problems associated with estimation of non-stationary regression functions and highly collinear inputs. We note that the transformed exchange rate series  $\Delta \log S_t$  may not in fact be an *iid* series; however, our resampling procedure will in fact deliver the desired sampling distribution under the weaker assumption, plausible here, that the gradient of the network output function with respect to the parameters multiplied by the errors constitute a martingale difference sequence.

Again following Diebold & Nason (1990), we consider lag structures of one, three, and five lags (their findings were unaffected by using different lag structures). For a given country, we therefore estimate three nonlinear models:

$$\begin{aligned} \text{Model 1: } & \Delta \log S_t = f(\Delta \log S_{t-1}, w_1) + \epsilon_1, \\ (8) \quad \text{Model 2: } & \Delta \log S_t = f(\Delta \log S_{t-1}, \Delta \log S_{t-2}, \Delta \log S_{t-3}, w_2) + \epsilon_2, \\ \text{Model 3: } & \Delta \log S_t = f(\Delta \log S_{t-1}, \Delta \log S_{t-2}, \Delta \log S_{t-3}, \Delta \log S_{t-4}, \Delta \log S_{t-5}, w_3) + \epsilon_3. \end{aligned}$$

We conduct our joint test for each of the three models, six countries, ten time periods, and three data frequencies requiring the estimation of 540 separate models, while the test was applied to each

---

<sup>1</sup>“Upon the introduction of the euro on January 1, 1999, we discontinued posting [daily] dollar exchange rates against the ecu and the currencies of the eleven countries participating in the European Economic and Monetary Union,” *Bulletin, Federal Reserve Board of St. Louis*.

model exactly as described in Section 4. Results<sup>2</sup> are summarized in tables<sup>3</sup> 1 through 10. We present results for the most recent decade in Table 1 while results for the other nine periods can be found in Appendix B.

Before proceeding to a discussion of the results, we briefly review the role of Type I errors in a testing framework. A Type I error is said to occur when we conduct an hypothesis test with level  $\alpha$ , the null hypothesis is in fact true, yet we incorrectly reject the null and conclude that the alternative is supported by the empirical evidence. For the application that we consider, if the null hypothesis is true then there is no information contained in past percentage changes that is relevant for the prediction of future percentage changes. If we considered testing this hypothesis at, say, the  $\alpha = 0.05$  (5%) level and *if* the null were in fact true, then we would expect to incorrectly reject the null five times out of one hundred, on average, thereby incorrectly concluding that the alternative hypothesis of predictability is supported by the data when in fact it is not.

To account for the fact that we are testing hypotheses with multiple models, we can perform a computation that provides an upper bound on the  $p$ -value for the joint null hypothesis that in *none of a particular group of models* do the lagged changes matter. The computation is based on the Bonferroni inequality for multiple hypotheses. Let  $P_1, \dots, P_m$  be  $p$ -values corresponding to a group of  $m$  test statistics, and let  $P_{(1)}, \dots, P_{(m)}$  be the ordered  $p$ -values. The Bonferroni inequality leads to rejection of the joint null at the  $\alpha$  level if  $P_{(1)} \leq \alpha/m$ , so the Bonferroni  $p$ -value bound is  $\alpha = mP_{(1)}$ . A disadvantage of this simple procedure is that it is based only on the smallest  $p$ -value,  $P_{(1)}$ , and may thus be too loose a bound. A tighter bound can be obtained in a simple way using Hochberg's (1988) modification of the Bonferroni method. Hochberg's inequality implies rejecting the joint null at the  $\alpha$  level if there exists a  $j$  such that  $P_{(j)} \leq \alpha/(m - j + 1)$ ,  $j = 1, \dots, m$ . This gives an improved Hochberg-Bonferroni (H-B) bound of  $\alpha = \min_{j=1, \dots, m} (m - j + 1)P_{(j)}$ . For each period we calculate the H-B  $p$ -value bound with groupings (1) by country and model, (2) by country and data frequency, (3) by country, model, and data frequency, (4) by model, (5) by data frequency, and (6) by model and data frequency. This permits us to specify precisely the groupings within which we reject or fail to reject the corresponding joint null hypothesis.

---

<sup>2</sup>We report the empirical  $p$ -values which represent the smallest level of significance at which one would *just* reject the null hypothesis. The smaller the  $p$ -value the stronger is the sample evidence against the validity of the null hypothesis. Conventional levels of significance are 10% (0.10), 5% (0.05), and 1% (0.01). Therefore, if the empirical  $p$ -value were, say, 0.02 then we would reject the null hypothesis at the 10% and 5% levels, but *not* at the 1% level. Small entries (say,  $p < 0.10$ ) would support predictability of exchange rates, while large entries (say,  $p > 0.10$ ) would support the random walk hypothesis.

<sup>3</sup>Updated tables can be found at [http://nonlin.bsn.usf.edu/papers/wr\\_tables.pdf](http://nonlin.bsn.usf.edu/papers/wr_tables.pdf)

TABLE 1. Empirical  $p$ -values for the period 1990-2000. The number of SIC-optimal hidden units appear in parentheses. Individual test  $p$ -values lie in cells bordered with Model and Frequency. H-B bounds for all models or frequencies for each country lie in rows and columns so denoted. The entry directly below the country name is the H-B bound for that country across all models and data frequencies. The last three rows are the H-B bounds for cross-country groupings.

Country	Model	Data Frequency			H-B Bounds	
		Daily	Weekly	Monthly		
Canada	1	0.290 (1)	0.323 (1)	0.175 (1)	0.323	
	2	0.342 (1)	0.163 (1)	0.362 (1)	0.362	
	0.471	3	0.471 (1)	0.208 (1)	0.264 (1)	0.471
	H-B Bounds		0.471	0.323	0.362	
France	1	0.073 (1)	0.136 (1)	0.066 (1)	0.136	
	2	0.393 (1)	0.043 (1)	0.189 (1)	0.129	
	0.009	3	0.001 (1)	0.196 (1)	0.255 (1)	0.003
	H-B Bounds		0.003	0.129	0.198	
Germany	1	0.144 (1)	0.249 (1)	0.262 (1)	0.262	
	2	0.581 (1)	0.266 (1)	0.254 (1)	0.532	
	0.581	3	0.251 (1)	0.557 (1)	0.264 (1)	0.528
	H-B Bounds		0.432	0.532	0.264	
Italy	1	0.717 (1)	0.219 (1)	0.034 (1)	0.102	
	2	0.004 (2)	0.023 (3)	0.001 (2)	0.003	
	0.007	3	0.036 (2)	0.001 (2)	0.001 (1)	0.002
	H-B Bounds		0.012	0.003	0.002	
Japan	1	0.266 (1)	0.142 (1)	0.231 (1)	0.266	
	2	0.251 (1)	0.034 (1)	0.261 (2)	0.102	
	0.008	3	0.001 (1)	0.001 (3)	0.027 (2)	0.002
	H-B Bounds		0.003	0.003	0.081	
England	1	0.282 (1)	0.001 (2)	0.073 (1)	0.003	
	2	0.043 (1)	0.372 (2)	0.175 (1)	0.129	
	0.008	3	0.205 (1)	0.001 (2)	0.207 (2)	0.003
	H-B Bounds		0.129	0.002	0.207	
H-B Bounds		Model 1	Model 2	Model 3		
		0.018	0.018	0.013		
All Countries		Daily	Weekly	Monthly		
		0.017	0.015	0.017		
		Overall				
		0.047				

Note: Because we use  $N=1,000$  bootstrap resamples, we are not able to resolve empirical  $p$ -values less than 0.001. For cases in which our statistic lies beyond the greatest resampled value, we report the conservative  $p$ -value of 0.001. H-B bounds are calculated using this conservative value.

An examination of Table 1 reveals that for the period 1990-2000 there appears to be information contained in past percentage changes that is relevant for the prediction of future percentage changes

for three out of the six countries, Italy, Japan, and England. Further, the apparent predictability for Italy, Japan and England is not just a data mining artifact, as evidenced by the small  $p$ -value bounds obtained by grouping over all countries.

Similar results appear in Tables 2 through 10 in Appendix B, in that in each case we see overall  $p$ -value bounds that are uniformly small. Despite this common finding, we nevertheless see a fair degree of heterogeneity among the  $p$ -value bounds for the different countries. None of the individual countries have uniformly small bounds across all the different sub-periods. We see similar heterogeneity across the different models and data frequencies, although the daily models do exhibit some uniformity: the highest  $p$ -value bound across all countries and time periods is 0.108. We interpret this heterogeneity as evidence that foreign exchange market structure is evolving through time.

In summary, then, we have evidence that there does appear to be information in past percentage changes, particularly for Japan over the last ten years and possibly for France over the last five years (see Table 10). Nevertheless, the evolution of foreign exchange market structure implies that the particular forms of these predictive relationships can be expected to have limited life spans.

## 7. CONCLUDING REMARKS

In this paper we have investigated the use of bootstrap methods for inference using artificial neural networks, with particular emphasis on testing the null hypothesis that a given set of inputs has no effect. We apply our methods to test for predictive power in foreign exchange rates and find that exchange rates do appear to contain information that is exploitable for enhanced point prediction, but the nature of the predictive relations evolves through time. Our results suggest that bootstrap-based inference can be a valuable addition to the toolkits of neural network practitioners, with the consequent potential to significantly enhance scientific understanding of empirical phenomena subject to neural network modeling.

In closing, we wish to highlight an important distinction between hypothesis testing and input selection. These remarks are intended to minimize misuse of the methods given here and to encourage use of alternative more appropriate procedures for input selection. Throughout, we have focused solely on use of statistical methods for testing the hypothesis that a particular set of inputs is irrelevant in the context of a particular network-approximated relationship. Once this hypothesis has been tested and an inference drawn, our procedures have fulfilled their purpose; we are done.

There is no necessary justification for further use of whatever inferences may have been drawn for subsequent modeling using the same data; indeed, such use can be quite misleading.

In particular, it is tempting to use our statistics for input selection in the following way: test a given set of inputs for irrelevance; drop variables found to be irrelevant and keep the rest; then retrain the network and test another set of variables for irrelevance; drop the irrelevant variables, retrain, and proceed in the same way, until no further variables are found to be irrelevant. This amounts to a form of backward stepwise regression. We recommend against this procedure. The reason is that the performance of such procedures depends critically on the level chosen for the tests, and the level of the tests used in the second and subsequent rounds of dropping variables must take into account the results of the statistical tests conducted in prior rounds. Properly choosing and keeping track of the level are quite complicated to do, and the danger of neglecting to do so is that the wrong variables can be kept and/or dropped. See Judge, Hill, Griffiths, Lütkepohl & Lee (1988, pp. 832-835) for a discussion of such “pre-test” estimation methods.

## REFERENCES

- Abu-Mostafa, Y. S., LeBaron, B., Lo, A. W. & Weigend, A. S., eds (2000), *Computational Finance 1999*, Cambridge: MIT Press.
- Baxt, W. (1992), 'Analysis of the clinical variables driving decision in an artificial neural network trained to identify the presence of myocardial infarction', *Annals of Emergency Medicine* **21**, 1439–1444.
- Baxt, W. & White, H. (1995), 'Bootstrapping confidence intervals for clinical input variable effects in a network trained to identify the presence of acute myocardial infarction', *Neural Computation* **7**, 624–638.
- Chen, A. M., Lu, H. M. & Hecht-Nielsen, R. (1993), 'On the geometry of feedforward neural network error surfaces', *Neural Computation* **5**, 910–927.
- Chen, X. & White, H. (1999), 'Improved rates and asymptotic normality for nonparametric neural network estimators', *IEEE Transactions on Information Theory* **45**, 682–691.
- Cleveland, W., Devlin, S. & Grosse, E. (1988), 'Regression by local fitting: Methods, properties, and computational algorithms', *Journal of Econometrics* **37**, 87–114.
- Coetzee, F. M. & Stonick, V. L. (1995), 'Topology and geometry of single hidden layer network', *Neural Computation* **7**, 672–705.
- Cootner, P. (1964), *The Random Character of Stock Market Prices*, Cambridge: MIT Press.
- Dacorogna, M., Gencay, R., Muller, U., Pictet, O. & Olsen, R. (2001, in press), *An Introduction to High-Frequency Finance*, New York: Academic Press.
- Diebold, F. X. & Nason, J. A. (1990), 'Nonparametric exchange rate prediction', *Journal of International Economics* **28**, 315–332.
- Dunis, C. & Zhou, B., eds (1998), *Nonlinear Modeling of High Frequency Financial Time Series*, New York: Wiley.
- Efron, B. (1983), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics.
- Engle, R. (1982), 'A general approach to lagrange multiplier model diagnostics', *Journal of Econometrics* **20**, 83–104.
- Fama, E. F. (1965), 'The behavior of stock market prices', *Journal of Business* **38**, 34–105.
- Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.
- Hochberg, Y. (1988), 'A sharper bonferroni procedure for multiple tests of significance', *Biometrika* **75**, 800–802.
- Jennrich, R. (1969), 'Asymptotic properties of nonlinear least squares estimators', *Annals of Mathematical Statistics* **40**, 633–643.
- Judge, G., Hill, R., Griffiths, W., Lütkepohl, H. & Lee, T. C. (1988), *Introduction to the Theory and Practice of Econometrics*, New York: Wiley.
- Kuan, C.-M. & Liu, T. (1995), 'Forecasting exchange rates using feedforward and recurrent neural networks', *Journal of Applied Econometrics* **10**(4), 347–364.
- LeBaron, B. (1998), Technical trading rules and regime shifts in foreign exchange, in E. Acar & S. Satchell, eds, 'Advanced Trading Rules', Massachusetts: Butterworth-Heinemann, pp. 5–40.

- Moody, J. & Wu, L. (1998), High frequency foreign exchange rates: Price behavior analysis and 'true price' models, in C. Dunis & B. Zhou, eds, 'Nonlinear Modeling of High Frequency Financial Time Series', New York: Wiley, pp. 23–47.
- Plutowski, M., Sakata, S. & White, H. (1994), 'Cross-validation estimates integrated mean squared error', *Advances in Neural Information Processing Systems* **6**, 391–398.
- Pötscher, B. M. (1991), 'Effects of model selection on inference', *Econometric Theory* **7**, 163–185.
- Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**, 461–464.
- Sin, C. Y. & White, H. (1996), 'Information criteria for selecting possibly misspecified parametric models', *Journal of Econometrics* **71**, 207–225.
- Sussmann, H. J. (1992), 'Uniqueness of the weights for minimal feedforward nets with a given input-output map', *Neural Networks* **5**, 589–593.
- White, H. (1989), 'Learning in artificial neural networks: A statistical perspective', *Neural Computation* **1**, 425–464.
- White, H. (1994), *Estimation, Inference, and Specification Analysis*, New York: Cambridge University Press.
- White, H. (2000), *Asymptotic Theory for Econometricians (revised edition)*, New York: Academic Press.

**Proof of Theorem 2.1:** A two-term mean value expansion around  $w^*$  gives

$$\begin{aligned} \sum_{t=1}^n m(X_t, \hat{w}_n) &= \sum_{t=1}^n m(X_t, w^*) + \sum_{t=1}^n \nabla^T m(X_t, w^*) (\hat{w}_n - w^*) \\ &\quad + n(\hat{w}_n - w^*)^T (n^{-1} \sum_{t=1}^n \nabla^2 m(X_t, \bar{w}_n)) (\hat{w}_n - w^*) / 2. \end{aligned}$$

By hypothesis,  $m(x, w^*) = 0$  and  $\nabla m(x, w^*) = 0$  for  $x$  in  $\text{supp } \mu$ . Using these facts and adding and subtracting terms appropriately gives

$$\begin{aligned} \sum_{t=1}^n m(X_t, \hat{w}_n) &= n(\hat{w}_n - w^*)^T E(\nabla^2 m(X_t, w^*)) (\hat{w}_n - w^*) / 2 \\ &\quad + n(\hat{w}_n - w^*)^T [n^{-1} \sum_{t=1}^n \nabla^2 m(X_t, \bar{w}_n) \\ &\quad - E(\nabla^2 m(X_t, w^*))] (\hat{w}_n - w^*) / 2. \end{aligned}$$

Because  $\bar{w}_n = w^* + o_P(1)$  and  $\{\nabla^2 m(x_t, w)\}$  obeys the weak uniform law of large numbers, we have  $n^{-1} \sum_{t=1}^n \nabla^2 m(X_t, \bar{w}_n) - E(\nabla^2 m(X_t, w^*)) = o_P(1)$  by Corollary 3.8 of White (1994). As  $\sqrt{n}(\hat{w}_n - w^*) = O_p(1)$  by hypothesis, the final term above is  $o_P(1)$ . It follows that  $\sum_{t=1}^n m(X_t, \hat{w}_n)$  has the same asymptotic distribution as

$$n(\hat{w}_n - w^*)^T E(\nabla^2 m(X_t, w^*)) (\hat{w}_n - w^*) / 2$$

by Lemma 4.7 of White (2000). From Theorem 8.6 of White (1994), this has the mixture of  $\chi^2$ 's distribution asymptotically, denoted  $N_2(0, C^*; M^*)$ , with  $M^* = E(\nabla^2 m(X_t, w^*)) / 2$ .

**Proof of Corollary 2.2:** We verify the conditions of Theorem 2.1. Assumption A.1 ensures that  $\{X_t\}$  is iid. Assumptions A.1 - A.6 guarantee that  $\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, C^*)$  by Theorem 2 of White (1989), where  $w^*$  is interior to  $W$  by Assumption A.4. Theorem 2 of White (1989) specifies that  $C^* = A^{*-1} B^* A^{*-1}$ .

Taking  $m(x, w) = \sum_{i \in I_0} f_i(x, w)^2$ , we have  $m(x, \cdot)$  continuously differentiable of order 2 by A.2(d) and  $m(\cdot, w)$  measurable given that  $f_i(\cdot, w)$  is continuous by A.2(c). Given (5) we have  $m(x, w^*) = \sum_{i \in I_0} f_i(x, w^*)^2$  and  $\nabla m(x, w^*) = 2 \sum_{i \in I_0} \nabla f_i(x, w^*) = 0$ . Assumptions A.1 and A.7 guarantee that  $\{\nabla^2 m(X_t, w)\}$  obeys the weak uniform law of large numbers by Jennrich (1969, theorem 2). The result follows.  $\square$

**Proof of Theorem 2.3:** We give the proof only for  $\tilde{\mathcal{B}}_n^*$ ; that for  $\hat{\mathcal{B}}_n^*$  is identical.

A two term Taylor expansion around  $\hat{w}_n$  gives

$$\begin{aligned} \sum_{t=1}^n m(X_t, \hat{w}_n^*) &= \sum_{t=1}^n m(X_t, \hat{w}_n) + \sum_{t=1}^n \nabla^T m(X_t, \hat{w}_n) (\hat{w}_n^* - \hat{w}_n) \\ &\quad + (\hat{w}_n^* - \hat{w}_n)^T [n^{-1} \sum_{t=1}^n \nabla^2 m(X_t, \bar{w}_n^*)] (\hat{w}_n^* - \hat{w}_n) / 2. \end{aligned}$$

Interiority of  $w^*$  (A.4) ensures that  $\hat{w}_n$  is interior to  $W$  for all  $n$  sufficiently large a.s., as is required to apply the Taylor theorem. Hence

$$\begin{aligned} \sum_{t=1}^n m(X_t, \hat{w}_n^*) - \sum_{t=1}^n m(X_t, \hat{w}_n) - \sum_{t=1}^n \nabla^T m(X_t, \hat{w}_n) (\hat{w}_n^* - \hat{w}_n) \\ = n(\hat{w}_n^* - \hat{w}_n)^T M^* (\hat{w}_n^* - \hat{w}_n) \\ + n(\hat{w}_n^* - \hat{w}_n)^T [n^{-1} \sum_{t=1}^n \nabla^2 m(x_t, \bar{w}_n^*) / 2 - M^*] (\hat{w}_n^* - \hat{w}_n). \end{aligned} \quad (\text{A.1})$$

The desired result follows from Lemma 4.7 of White (2000) by showing that  $\hat{b}_n^{*T} \hat{b}_n^* \equiv (\hat{w}_n^* - \hat{w}_n)^T M^* (\hat{w}_n^* - \hat{w}_n) \xrightarrow{d} N_2(0, C^*; M^*)$  and that the second term on the right of (A.1) vanishes in probability.

If  $\sqrt{n}(\hat{w}_n^* - \hat{w}_n) \xrightarrow{d} N(0, C^*)$  it follows from Theorem 8.6 of White (1994) that  $\hat{b}_n^{*T} \hat{b}_n^* \xrightarrow{d} N_2(0, C^*; M^*)$ . This also implies  $\hat{w}_n^* - \hat{w}_n = o_P(1)$ . Because  $\{\nabla^2 m(X_t, w)\}$  obeys the weak uniform law of large numbers given A.7, it follows from Corollary 3.8 of White (1994) that

$$n^{-1} \sum_{t=1}^n \nabla^2 m(X_t, \bar{w}_n^*) = M^* + o_P(1).$$

This, together with  $\sqrt{n}(\hat{w}_n^* - \hat{w}_n) \xrightarrow{d} N(0, C^*)$  implies that the second term on the right side of (A.1) is  $o_P(1)$  by Corollary 2.36 of White (2000).

Thus, the desired result follows, provided  $\sqrt{n}(\hat{w}_n^* - \hat{w}_n) \xrightarrow{d} N(0, C^*)$ . Now it follows from theorems 1 and 2 of White (1989) that except for re-sampled sequences having probability zero

$$\hat{C}_n^{-\frac{1}{2}} \sqrt{n}(\hat{w}_n^* - \hat{w}_n) \xrightarrow{d} N(0, 1)$$

where  $\hat{C}_n = \hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1}$ ,

$$\begin{aligned}\hat{A}_n &\equiv \int \nabla^2 l(z, \hat{w}_n) d\hat{v}_n(z) \\ \hat{B}_n &\equiv \int \nabla l(z, \hat{w}_n) \nabla l(z, \hat{w}_n)^T d\hat{v}_n(z).\end{aligned}$$

Because  $\hat{w}_n = w^* + o_P(1)$  and the (weak) uniform law of large numbers holds for  $\{\nabla^2 l(Z_t, w)\}$  and  $\{\nabla l(Z_t, w) \nabla l(Z_t, w)^T\}$  given Assumptions A.1 and A.5, we have  $\hat{A}_n = A^* + o_P(1)$  and  $\hat{B}_n = B^* + o_P(1)$  so that  $\hat{C}_n = C^* + o_P(1)$ . It follows by Corollary 4.24 of White (2000) that  $\sqrt{n}(\hat{w}_n^* - \hat{w}_n) \xrightarrow{d} N(0, C^*)$  and the proof is complete.  $\square$

**Proof of Proposition 2.4:** First suppose that  $\partial f_2(x, w_2^*)/\partial x_i = 0$  for all  $x \in \text{supp } \mu, i \in I_0$ . Then by Condition 2.4(c) there exists  $w_1^\dagger$  (say) in  $W_1$  such that  $f_1(x, w_1^\dagger) = f_2(x, w_2^*)$  for all  $x \in \text{supp } \mu$ , and it follows that  $\lambda_1(w_1^\dagger) = \lambda_2(w_2^*)$ . By definition of  $w_1^*$ ,  $\lambda_1(w_1^\dagger) \geq \lambda_1(w_1^*)$ . Suppose  $\lambda_1(w_1^\dagger) > \lambda_1(w_1^*)$ . Then  $\lambda_2(w_2^*) > \lambda_1(w_1^*)$ , which, in view of Condition 2.4(d), violates the property that  $w_2^*$  minimizes  $\lambda_2(w)$ . Consequently,  $\lambda_1(w_1^\dagger) = \lambda_1(w_1^*)$ . But the uniqueness of  $w_1^*$  requires  $w_1^\dagger = w_1^*$ . It follows that  $f_1(x, w_1^*) = f_1(x, w_1^\dagger) = f_2(x, w_2^*)$  for all  $x \in \text{supp } \mu$ .

Next suppose that  $f_1(x, w_1^{*0}) = f_2(x, w_2^*)$  for all  $x \in \text{supp } \mu$ . To obtain a contradiction, we assume it is false that  $\partial f_2(x, w_2^*)/\partial x_i = 0$  for all  $x \in \text{supp } \mu, i \in I_0$ . It follows that  $w_2^*$  belongs to  $W_2^0$  of Condition 2.4(b). Because  $f_1(x, w_1^*) = f_2(x, w_2^*)$  we have  $\lambda_1(w_1^*) = \lambda_2(w_2^*)$ . But then Condition 2.4(d) requires either that  $w_2^*$  belongs to  $W_2^{10}$  which contradicts  $w_2^* \in W_2^0$ , or that there exists  $w_2^\dagger \in W_2^{10}$  with  $\lambda_2(w_2^*) = \lambda_2(w_2^\dagger)$ , which contradicts the uniqueness of  $w_2^*$ . The proof is complete.  $\square$

**Proof of Corollary 2.5:** Again we verify the conditions of Theorem 2.1. Assumption A.1 ensures that  $\{X_t\}$  is iid. Assumptions A.1-A.4, A.5(a) and A.6 imposed in B.1 ensure that

$$\begin{aligned}\sqrt{n}(\hat{w}_{1n} - w_1^*) &= -A_1^{*-1} n^{-\frac{1}{2}} \sum_{t=1}^n \nabla_1 l_{1t}^* + o_P(1) \\ \sqrt{n}(\hat{w}_{2n} - w_2^*) &= -A_2^{*-1} n^{-\frac{1}{2}} \sum_{t=1}^n \nabla_2 l_{2t}^* + o_P(1)\end{aligned}$$

by Theorem 6.2 of White (1994). Putting  $\hat{w}_n = (\hat{w}_{1n}^T, \hat{w}_{2n}^T)^T$ ,  $w^* = (w_1^{*T}, w_2^{*T})^T$ , and

$$A^{*-1} = \begin{bmatrix} A_1^{*-1} & 0 \\ 0 & A_2^{*-1} \end{bmatrix}$$

we have

$$\sqrt{n}(\hat{w}_n - w^*) = -A^{*-1}n^{-\frac{1}{2}} \sum_{t=1}^n \nabla l_t^* + o_P(1).$$

It follows from the Lindberg-Levy central limit theorem (*e.g.* White, 2000, Theorem 5.2) that  $\lambda^T A^{*-1} n^{-\frac{1}{2}} \sum_{t=1}^n \nabla l_t^* \xrightarrow{d} N(0, \lambda^T C^* \lambda)$  for all  $\lambda$  such that  $\lambda^T \lambda = 1$  and  $\lambda^T C^* \lambda > 0$ , where  $C^* = A^{*-1} B^* A^{*-1}$ ,

$$B^* = \begin{bmatrix} B_1^* & B_{12}^* \\ B_{21}^* & B_2^* \end{bmatrix},$$

$$B_1^* = E(\nabla_1 l_t^* \nabla_1 l_{1t}^{*T}),$$

$$B_{12}^* = E(\nabla_1 l_{1t}^* \nabla_2 l_{2t}^{*T}),$$

$$B_2^* = E(\nabla_2 l_{2t}^* \nabla_2 l_{2t}^{*T}).$$

For  $\lambda$  such that  $\lambda^T \lambda = 1$  and  $\lambda^T C^* \lambda = 0$ , we have  $\lambda^T A^{*-1} n^{-\frac{1}{2}} \sum_{t=1}^n \nabla l_t^* \rightarrow 0$  from Chebyshev's inequality. Thus  $A^{*-1} n^{-\frac{1}{2}} \sum_{t=1}^n \nabla l_t^* \xrightarrow{d} N(0, C^*)$ ,  $C^*$  symmetric and positive semi-definite. It then follows from Lemma 4.7 of White (2000) that

$$\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, C^*).$$

Taking  $m(x, w) = [f_2(x, w_1) - f_2(x, w_2)]^2$  we have  $m(x, \cdot)$  continuously differentiable of order 2 by A.2(b) and  $m(\cdot, w) = 0$  measurable by A.2(a). Given (6) we have  $m(x, w^*) = 0$  from Proposition 2.4; also  $\nabla m(x, w^*) = 2(\nabla_1 f_1(x, w_1^*)^T, -\nabla_2 f_2(x, w_2^*)^T)^T (f_1(x, w_1^*) - f_2(x, w_2^*)) = 0$  by Proposition 2.4. Assumptions A.1 and B.2 guarantee that  $\{\nabla^2 m(X_t, w)\}$  obeys the weak uniform law of large numbers by Jennrich (1969, theorem 2). The result follows with

$$M^* = \begin{bmatrix} M_1^* & M_{12}^* \\ M_{21}^* & M_2^* \end{bmatrix},$$

$$M_1^* = E[\nabla_1 f_2(X_t, w_1^*) \nabla_1 f_1(X_t, w_1^*)^T + \nabla_1^2 f_1(X_t, w_1^*) (f_1(X_t, w_1^*) - f_2(X_t, w_2^*))],$$

$$M_{12}^* = -E[\nabla_1 f_1(X_t, w_1^*) \nabla_2 f_2(X_t, w_2^*)],$$

$$M_2^* = E[\nabla_2 f_2(X_t, w_1^*) \nabla_2 f_2(X_t, w_2^*)^T - \nabla_2^2 f_2(X_t, w_2^*) (f_1(X_t, w_1^*) - f_2(X_t, w_2^*))].$$

**Proof of Theorem 2.6:** The proof is identical to that of Theorem 2.3; the result follows, provided  $\sqrt{n}(\hat{w}_n^* - \hat{w}_n) \xrightarrow{d} N(0, C^*)$  *a.s.*, where  $\hat{w}_n^* = (\hat{w}_{1n}^{*T}, \hat{w}_{2n}^{*T})^T$ . Argument identical to that of

the proof of Corollary 2.5 establishes that with  $\mathcal{Z}_n = \sqrt{n}(\hat{w}_n^* - \hat{w}_n)$

$$E(i\tau^T \mathcal{Z}_n) - \exp(-\tau^T \hat{C}_n \tau / 2) \rightarrow 0$$

for any  $p \times 1$  vector  $\tau$ , where  $i \equiv \sqrt{-1}$  and  $\hat{C}_n = \hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1}$ ,

$$\hat{A}_n = \begin{bmatrix} \hat{A}_{1n}^{-1} & 0 \\ 0 & \hat{A}_{2n}^{-1} \end{bmatrix}$$

$$\hat{B}_n = \begin{bmatrix} \hat{B}_{1n} & \hat{B}_{12n} \\ \hat{B}_{21n} & \hat{B}_{2n} \end{bmatrix}$$

$$\hat{A}_{1n} = n^{-1} \sum_{t=1}^n \nabla_1^2 l_1(Z_t, \hat{w}_{1n}),$$

$$\hat{A}_{2n} = n^{-1} \sum_{t=1}^n \nabla_2^2 l_2(Z_t, \hat{w}_{2n}),$$

$$\hat{B}_n = n^{-1} \sum_{t=1}^n \nabla_1 l_1(Z_t, \hat{w}_{1n}) \nabla_1 l_1(Z_t, \hat{w}_{1n})^T,$$

$$\hat{B}_{12n} = n^{-1} \sum_{t=1}^n \nabla_1 l_1(Z_t, \hat{w}_{1n}) \nabla_2 l_2(Z_t, \hat{w}_{2n})^T,$$

$$\hat{B}_{2n} = n^{-1} \sum_{t=1}^n \nabla_2 l_2(Z_t, \hat{w}_{2n})^T.$$

It follows from Assumption A.5 that  $\hat{A}_{1n} = A^* + o_{as}(1)$ , and  $\hat{B}_n = B^* + o_{as}(1)$ , so that  $\hat{C}_n = C^* + o_{as}(1)$ . It follows that

$$\exp(-\tau^T \hat{C}_n \tau / 2) - \exp(-\tau^T C^* \tau / 2) \rightarrow 0 \text{ a.s.}$$

so that  $E(i\tau^T \mathcal{Z}_n) \rightarrow \exp(-\tau^T C^* \tau / 2)$  a.s. This limit is the characteristic function of  $N(0, C^*)$ , so that by Theorem 4.17 of White (2000)  $\sqrt{n}(\hat{w}_n^* - \hat{w}_n) \xrightarrow{d} N(0, C^*)$  a.s. as required, and the proof is complete.  $\square$

TABLE 2. Empirical  $p$ -values for the period 1971-2000. The number of SIC-optimal hidden units appear in parentheses.

Country	Model	Data Frequency				
		Daily	Weekly	Monthly		
Canada 0.027	1	0.464 (1)	0.305 (1)	0.310 (1)	0.464	
	2	0.192 (2)	0.161 (1)	0.428 (1)	0.384	
	3	0.292 (2)	0.003 (1)	0.156 (1)	0.009	
	H-B Bounds		0.464	0.009	0.428	
France 0.009	1	1.000 (1)	0.001 (1)	0.260 (1)	0.003	
	2	1.000 (1)	0.306 (2)	0.003 (1)	0.009	
	3	0.148 (1)	0.024 (2)	0.122 (1)	0.072	
	H-B Bounds		0.444	0.003	0.009	
Germany 0.007	1	0.001 (1)	0.242 (1)	0.165 (1)	0.003	
	2	0.001 (1)	0.001 (1)	0.339 (1)	0.002	
	3	0.163 (1)	0.138 (1)	0.099 (1)	0.163	
	H-B Bounds		0.002	0.003	0.297	
Italy 0.007	1	0.001 (1)	0.002 (1)	0.096 (1)	0.003	
	2	0.003 (1)	0.081 (1)	0.046 (1)	0.009	
	3	0.010 (3)	0.001 (2)	0.001 (2)	0.002	
	H-B Bounds		0.003	0.003	0.003	
Japan 0.009	1	0.161 (1)	0.078 (1)	0.158 (1)	0.161	
	2	0.223 (1)	0.246 (1)	0.007 (2)	0.021	
	3	0.272 (1)	0.013 (2)	0.001 (1)	0.003	
	H-B Bounds		0.272	0.039	0.003	
England 0.008	1	0.066 (1)	0.001 (1)	1.000 (1)	0.003	
	2	0.077 (1)	0.145 (1)	0.059 (1)	0.145	
	3	0.493 (1)	0.001 (2)	0.005 (1)	0.003	
	H-B Bounds		0.154	0.002	0.015	
H-B Bounds All Countries		Model 1	Model 2	Model 3		
		0.015	0.017	0.015		
		Daily	Weekly	Monthly		
		0.016	0.014	0.017		
		Overall				
		0.045				

TABLE 3. Empirical  $p$ -values for the period 1970-1980. The number of SIC-optimal hidden units appear in parentheses.

Country	Model	Data Frequency				
		Daily	Weekly	Monthly		
Canada	1	0.230 (1)	0.033 (1)	0.309 (1)	0.099	
	2	0.001 (1)	0.322 (1)	0.215 (2)	0.003	
	0.009	3	0.486 (2)	0.239 (1)	0.096 (1)	0.288
	H-B Bounds		0.003	0.099	0.288	
France	1	1.000 (1)	0.001 (1)	0.001 (1)	0.002	
	2	0.017 (1)	0.377 (2)	0.164 (2)	0.051	
	0.008	3	0.228 (1)	0.330 (3)	0.006 (2)	0.018
	H-B Bounds		0.051	0.003	0.003	
Germany	1	0.248 (2)	0.368 (3)	0.036 (2)	0.108	
	2	0.252 (2)	0.073 (3)	0.149 (1)	0.219	
	0.045	3	0.277 (2)	0.216 (2)	0.005 (1)	0.015
	H-B Bounds		0.277	0.219	0.015	
Italy	1	0.208 (2)	0.365 (2)	0.179 (1)	0.365	
	2	0.332 (3)	0.033 (2)	0.215 (2)	0.099	
	0.027	3	0.083 (5)	0.124 (4)	0.003 (1)	0.009
	H-B Bounds		0.249	0.099	0.009	
Japan	1	0.017 (2)	0.084 (2)	0.194 (1)	0.051	
	2	0.143 (2)	0.007 (2)	0.007 (2)	0.014	
	0.009	3	0.023 (5)	0.001 (2)	0.245 (2)	0.003
	H-B Bounds		0.046	0.003	0.021	
England	1	0.210 (2)	0.362 (1)	0.257 (2)	0.362	
	2	0.301 (1)	0.325 (1)	0.361 (1)	0.361	
	0.362	3	0.079 (3)	0.330 (1)	0.211 (1)	0.237
	H-B Bounds		0.237	0.362	0.361	
H-B Bounds		Model 1	Model 2	Model 3		
All		Daily	Weekly	Monthly		
Countries		0.018	0.017	0.018		
		Overall				
		0.051				

TABLE 4. Empirical  $p$ -values for the period 1980-1990. The number of SIC-optimal hidden units appear in parentheses.

Country	Model	Data Frequency			
		Daily	Weekly	Monthly	
Canada	1	0.649 (1)	0.270 (1)	0.263 (1)	0.540
	2	0.349 (2)	0.176 (1)	0.306 (1)	0.349
	3	0.458 (1)	0.469 (1)	0.035 (1)	0.105
H-B Bounds		0.649	0.469	0.105	
France	1	0.283 (2)	0.270 (1)	0.212 (1)	0.283
	2	0.394 (1)	0.618 (1)	0.168 (1)	0.504
	3	0.009 (1)	0.293 (1)	0.317 (1)	0.027
H-B Bounds		0.027	0.586	0.317	
Germany	1	0.257 (1)	0.323 (1)	0.200 (1)	0.323
	2	0.287 (1)	0.405 (1)	0.168 (1)	0.405
	3	0.337 (1)	0.576 (1)	0.281 (1)	0.576
H-B Bounds		0.337	0.576	0.281	
Italy	1	0.566 (1)	0.348 (1)	0.192 (1)	0.566
	2	0.001 (1)	0.001 (1)	0.232 (1)	0.002
	3	0.262 (1)	0.215 (1)	0.241 (1)	0.262
H-B Bounds		0.003	0.003	0.241	
Japan	1	0.358 (1)	0.291 (1)	0.249 (1)	0.358
	2	0.298 (1)	0.283 (1)	0.117 (1)	0.298
	3	0.517 (1)	0.229 (1)	0.120 (1)	0.360
H-B Bounds		0.517	0.291	0.240	
England	1	0.383 (1)	0.342 (2)	0.619 (1)	0.619
	2	0.372 (1)	0.147 (1)	0.120 (1)	0.294
	3	0.289 (1)	0.259 (1)	0.191 (2)	0.289
H-B Bounds		0.383	0.342	0.360	
		Model 1	Model 2	Model 3	
H-B Bounds		0.649	0.017	0.162	
All		Daily	Weekly	Monthly	
Countries		0.018	0.018	0.619	
		Overall			
		0.053			

TABLE 5. Empirical  $p$ -values for the period 1971-1975. The number of SIC-optimal hidden units appear in parentheses.

Country	Model	Data Frequency			
		Daily	Weekly	Monthly	
Canada 0.434	1	0.092 (2)	0.205 (1)	0.402 (2)	0.276
	2	0.294 (1)	0.318 (1)	0.212 (2)	0.318
	3	0.434 (1)	0.284 (1)	0.163 (2)	0.434
	H-B Bounds		0.276	0.318	0.402
France 0.008	1	0.012 (1)	0.044 (3)	0.195 (2)	0.036
	2	0.001 (2)	0.137 (3)	0.271 (4)	0.003
	3	0.021 (2)	0.001 (3)	0.191 (4)	0.003
	H-B Bounds		0.003	0.003	0.271
Germany 0.108	1	0.173 (2)	0.319 (3)	0.277 (2)	0.319
	2	0.018 (3)	0.238 (2)	0.219 (1)	0.054
	3	0.272 (2)	0.019 (2)	0.012 (1)	0.036
	H-B Bounds		0.054	0.057	0.036
Italy 0.440	1	0.529 (2)	0.360 (2)	0.051 (2)	0.153
	2	0.327 (2)	0.222 (2)	0.174 (4)	0.327
	3	0.055 (2)	0.256 (2)	0.541 (3)	0.165
	H-B Bounds		0.165	0.360	0.153
Japan 0.008	1	0.205 (2)	0.156 (3)	0.340 (1)	0.340
	2	0.001 (3)	0.002 (2)	0.266 (3)	0.003
	3	0.002 (2)	0.001 (2)	0.073 (2)	0.003
	H-B Bounds		0.003	0.003	0.219
England 0.009	1	1.000 (1)	0.332 (1)	0.294 (2)	0.664
	2	0.434 (1)	0.004 (2)	0.202 (2)	0.012
	3	0.349 (1)	0.001 (1)	0.211 (5)	0.003
	H-B Bounds		0.868	0.003	0.294
H-B Bounds		Model 1	Model 2	Model 3	
		0.216	0.017	0.016	
All Countries		Daily	Weekly	Monthly	
		0.017	0.016	0.216	
		Overall			
		0.050			

TABLE 6. Empirical  $p$ -values for the period 1975-1980. The number of SIC-optimal hidden units appear in parentheses.

Country	Model	Data Frequency				
		Daily	Weekly	Monthly		
Canada	1	0.260 (1)	0.051 (1)	0.281 (1)	0.153	
	2	0.011 (1)	0.413 (1)	0.291 (1)	0.033	
	0.099	3	0.237 (1)	0.361 (1)	0.267 (2)	0.361
	H-B Bounds		0.033	0.153	0.291	
France	1	0.390 (2)	0.355 (2)	0.481 (1)	0.481	
	2	0.167 (3)	0.163 (2)	0.245 (1)	0.245	
	0.144	3	0.016 (2)	0.036 (2)	0.063 (1)	0.048
	H-B Bounds		0.048	0.108	0.189	
Germany	1	0.252 (1)	0.262 (1)	0.018 (1)	0.054	
	2	0.008 (1)	0.020 (2)	0.125 (1)	0.024	
	0.009	3	0.001 (1)	0.004 (1)	0.008 (1)	0.003
	H-B Bounds		0.003	0.012	0.024	
Italy	1	0.017 (3)	0.001 (1)	0.607 (1)	0.003	
	2	0.147 (6)	0.124 (1)	0.597 (1)	0.294	
	0.008	3	0.001 (4)	0.037 (2)	0.004 (1)	0.003
	H-B Bounds		0.003	0.003	0.012	
Japan	1	0.077 (2)	0.266 (1)	0.150 (1)	0.231	
	2	0.457 (2)	0.318 (1)	0.342 (3)	0.457	
	0.144	3	0.016 (2)	0.033 (2)	0.052 (3)	0.048
	H-B Bounds		0.048	0.099	0.156	
England	1	0.319 (2)	0.285 (1)	0.169 (1)	0.319	
	2	0.351 (1)	0.230 (1)	0.353 (1)	0.353	
	0.353	3	0.278 (2)	0.303 (1)	0.200 (1)	0.303
	H-B Bounds		0.351	0.303	0.353	
H-B Bounds		Model 1	Model 2	Model 3		
		0.018	0.144	0.017		
All Countries		Daily	Weekly	Monthly		
		0.017	0.018	0.072		
		Overall				
		0.052				

TABLE 7. Empirical  $p$ -values for the period 1980-1985. The number of SIC-optimal hidden units appear in parentheses.

Country	Model	Data Frequency				
		Daily	Weekly	Monthly		
Canada	1	0.331 (1)	0.412 (1)	0.075 (1)	0.225	
	2	0.230 (1)	0.292 (1)	0.019 (1)	0.057	
	0.117	3	0.013 (1)	0.292 (1)	0.134 (2)	0.039
	H-B Bounds		0.039	0.412	0.057	
France	1	0.063 (1)	0.289 (1)	0.288 (1)	0.189	
	2	0.055 (1)	0.465 (1)	0.207 (2)	0.165	
	0.009	3	0.004 (1)	0.001 (2)	0.025 (1)	0.003
	H-B Bounds		0.012	0.003	0.075	
Germany	1	0.317 (1)	0.618 (1)	0.251 (1)	0.618	
	2	0.173 (1)	0.005 (1)	0.261 (2)	0.015	
	0.009	3	0.350 (1)	0.001 (1)	0.409 (2)	0.003
	H-B Bounds		0.350	0.003	0.409	
Italy	1	0.002 (1)	0.263 (1)	0.296 (1)	0.006	
	2	0.003 (1)	0.293 (1)	0.112 (2)	0.009	
	0.018	3	0.004 (1)	0.320 (1)	0.191 (3)	0.012
	H-B Bounds		0.004	0.320	0.296	
Japan	1	0.510 (1)	0.321 (1)	0.014 (1)	0.042	
	2	0.132 (1)	0.337 (1)	0.100 (1)	0.264	
	0.126	3	0.242 (1)	0.152 (1)	0.503 (3)	0.456
	H-B Bounds		0.396	0.337	0.042	
England	1	0.211 (1)	0.560 (1)	0.264 (1)	0.528	
	2	0.473 (1)	0.132 (2)	0.045 (1)	0.135	
	0.189	3	0.021 (1)	0.241 (1)	0.128 (1)	0.063
	H-B Bounds		0.063	0.396	0.135	
H-B Bounds		Model 1	Model 2	Model 3		
		0.036	0.054	0.017		
All Countries		Daily	Weekly	Monthly		
		0.036	0.017	0.252		
		Overall				
		0.053				

TABLE 8. Empirical  $p$ -values for the period 1985-1990. The number of SIC-optimal hidden units appear in parentheses.

Country	Model	Data Frequency				
		Daily	Weekly	Monthly		
Canada	1	0.368 (1)	0.288 (1)	0.199 (1)	0.368	
	2	0.200 (1)	0.422 (1)	0.274 (1)	0.422	
	0.045	3	0.237 (1)	0.005 (2)	0.316 (1)	0.015
	H-B Bounds		0.368	0.015	0.316	
France	1	0.448 (1)	1.000 (1)	0.321 (1)	0.896	
	2	0.095 (1)	0.248 (2)	0.269 (1)	0.269	
	0.009	3	0.344 (1)	0.006 (1)	0.001 (1)	0.003
	H-B Bounds		0.285	0.018	0.003	
Germany	1	0.453 (1)	0.223 (1)	0.133 (1)	0.399	
	2	0.282 (1)	0.315 (1)	0.203 (1)	0.315	
	0.045	3	0.292 (1)	0.005 (1)	0.213 (1)	0.015
	H-B Bounds		0.453	0.015	0.213	
Italy	1	0.287 (1)	0.001 (1)	0.262 (1)	0.003	
	2	0.462 (1)	0.012 (1)	0.453 (1)	0.036	
	0.008	3	0.005 (1)	0.001 (1)	0.439 (2)	0.003
	H-B Bounds		0.015	0.002	0.453	
Japan	1	0.001 (1)	0.253 (1)	0.238 (1)	0.003	
	2	0.397 (1)	0.130 (1)	0.120 (1)	0.260	
	0.009	3	0.620 (1)	0.112 (2)	0.344 (3)	0.336
	H-B Bounds		0.003	0.253	0.344	
England	1	0.299 (1)	0.001 (1)	0.181 (1)	0.003	
	2	0.443 (1)	0.001 (1)	0.326 (1)	0.003	
	0.008	3	0.393 (1)	0.025 (2)	0.123 (1)	0.075
	H-B Bounds		0.443	0.002	0.326	
H-B Bounds		Model 1	Model 2	Model 3		
		0.016	0.018	0.017		
All Countries		Daily	Weekly	Monthly		
		0.018	0.015	0.018		
		Overall				
		0.049				

TABLE 9. Empirical  $p$ -values for the period 1990-1995. The number of SIC-optimal hidden units appear in parentheses.

Country	Model	Data Frequency				
		Daily	Weekly	Monthly		
Canada	1	0.430 (1)	0.384 (1)	0.219 (1)	0.430	
	2	0.425 (1)	0.069 (1)	0.143 (2)	0.207	
	0.445	3	0.357 (1)	0.063 (1)	0.445 (1)	0.189
	H-B Bounds		0.430	0.138	0.429	
France	1	0.212 (1)	0.217 (1)	0.179 (1)	0.217	
	2	0.003 (1)	0.008 (1)	0.038 (2)	0.009	
	0.027	3	0.011 (1)	0.156 (1)	0.178 (1)	0.033
	H-B Bounds		0.009	0.024	0.114	
Germany	1	1.000 (1)	0.233 (1)	0.103 (1)	0.309	
	2	0.272 (1)	0.199 (1)	0.397 (1)	0.397	
	0.009	3	0.001 (1)	0.133 (1)	0.338 (1)	0.003
	H-B Bounds		0.003	0.233	0.309	
Italy	1	0.304 (1)	0.118 (1)	0.065 (1)	0.195	
	2	0.146 (2)	0.270 (2)	0.543 (2)	0.438	
	0.009	3	0.008 (2)	0.001 (1)	0.046 (1)	0.003
	H-B Bounds		0.024	0.003	0.130	
Japan	1	0.293 (1)	0.374 (1)	0.228 (1)	0.374	
	2	0.203 (1)	0.222 (1)	0.242 (1)	0.242	
	0.009	3	0.016 (1)	0.001 (1)	0.221 (1)	0.003
	H-B Bounds		0.048	0.003	0.242	
England	1	0.485 (1)	0.198 (2)	0.073 (1)	0.219	
	2	0.262 (1)	0.002 (2)	0.004 (1)	0.006	
	0.009	3	0.080 (1)	0.111 (2)	0.001 (1)	0.003
	H-B Bounds		0.240	0.006	0.003	
H-B Bounds		Model 1	Model 2	Model 3		
All		Daily	Weekly	Monthly		
Countries		0.018	0.017	0.018		
		Overall				
		0.051				

TABLE 10. Empirical  $p$ -values for the period 1995-2000. The number of SIC-optimal hidden units appear in parentheses.

Country	Model	Data Frequency				
		Daily	Weekly	Monthly		
Canada	1	0.369 (1)	0.530 (1)	0.310 (1)	0.530	
	2	0.522 (1)	0.041 (1)	0.428 (1)	0.123	
	0.153	3	0.465 (1)	0.017 (1)	0.466 (1)	0.051
	H-B Bounds		0.522	0.051	0.466	
France	1	0.378 (1)	0.165 (1)	0.238 (1)	0.378	
	2	0.080 (1)	0.255 (1)	0.224 (1)	0.240	
	0.009	3	0.001 (1)	0.252 (1)	0.404 (1)	0.003
	H-B Bounds		0.003	0.255	0.404	
Germany	1	0.081 (1)	0.430 (1)	0.382 (1)	0.243	
	2	0.254 (1)	0.228 (1)	0.080 (1)	0.240	
	0.430	3	0.182 (1)	0.077 (1)	0.237 (1)	0.231
	H-B Bounds		0.243	0.231	0.240	
Italy	1	1.000 (1)	0.175 (1)	0.276 (1)	0.525	
	2	0.370 (1)	0.277 (1)	0.074 (2)	0.222	
	0.666	3	0.293 (1)	0.280 (1)	0.091 (2)	0.273
	H-B Bounds		0.740	0.280	0.182	
Japan	1	0.301 (1)	0.128 (1)	0.077 (1)	0.231	
	2	0.001 (1)	0.104 (1)	0.278 (1)	0.003	
	0.008	3	0.161 (2)	0.452 (2)	0.001 (1)	0.003
	H-B Bounds		0.003	0.256	0.003	
England	1	1.000 (1)	1.000 (1)	0.201 (1)	0.603	
	2	0.158 (1)	0.312 (1)	0.455 (1)	0.455	
	1.000	3	0.503 (1)	0.139 (1)	0.287 (1)	0.417
	H-B Bounds		0.474	0.417	0.455	
H-B Bounds		Model 1	Model 2	Model 3		
		1.000	0.018	0.017		
All Countries		Daily	Weekly	Monthly		
		0.017	0.306	0.018		
		Overall				
		0.052				

Table 11 presents the empirical rejection frequencies for the Monte Carlo simulation reported in Section 6 which are plotted in Figure 1. The test was conducted with a nominal level of  $\alpha = 0.05$ . Both tests have actual level which does not differ significantly from the nominal level at all conventional levels of significance.

TABLE 11. Level and Power

$\beta_2$	NN	$t$ -test
0.00	0.045	0.047
0.01	0.046	0.047
0.02	0.056	0.068
0.03	0.093	0.120
0.04	0.130	0.168
0.05	0.202	0.256
0.06	0.268	0.350
0.07	0.391	0.493
0.08	0.481	0.579
0.09	0.571	0.684
0.10	0.670	0.775
0.11	0.758	0.850
0.12	0.826	0.905
0.13	0.871	0.939
0.14	0.934	0.979
0.15	0.950	0.986
0.16	0.967	0.993
0.17	0.986	0.997
0.18	0.990	1.000
0.19	0.991	0.999
0.20	0.994	0.999

HALBERT WHITE, DEPARTMENT OF ECONOMICS 0508, UNIVERSITY OF CALIFORNIA, SAN DIEGO, 9500 GILLMAN DRIVE, LA JOLLA, CA, USA 92093-0508. JEFF RACINE, DEPARTMENT OF ECONOMICS, BSN3403, UNIVERSITY OF SOUTH FLORIDA, 4202 WEST FOWLER AVENUE, TAMPA, FL, USA 33620.