Econometrica, Vol. 68, No. 5 (September, 2000), 1097–1126

A REALITY CHECK FOR DATA SNOOPING

BY HALBERT WHITE[1]

Data snooping occurs when a given set of data is used more than once for purposes of inference or model selection. When such data reuse occurs, there is always the possibility that any satisfactory results obtained may simply be due to chance rather than to any merit inherent in the method yielding the results. This problem is practically unavoidable in the analysis of time-series data, as typically only a single history measuring a given phenomenon of interest is available for analysis. It is widely acknowledged by empirical researchers that data snooping is a dangerous practice to be avoided, but in fact it is endemic. The main problem has been a lack of sufficiently simple practical methods capable of assessing the potential dangers of data snooping in a given situation. Our purpose here is to provide such methods by specifying a straightforward procedure for testing the null hypothesis that the best model encountered in a specification search has no predictive superiority over a given benchmark model. This permits data snooping to be undertaken with some degree of confidence that one will not mistake results that could have been generated by chance for genuinely good results.

KEYWORDS: Data mining, multiple hypothesis testing, bootstrap, forecast evaluation, model selection, prediction.

1. INTRODUCTION

WHENEVER A "GOOD" FORECASTING MODEL is obtained by an extensive specification search, there is always the danger that the observed good performance results not from actual forecasting ability, but is instead just luck. Even when no exploitable forecasting relation exists, looking long enough and hard enough at a given set of data will often reveal one or more forecasting models that look good, but are in fact useless.

This is analogous to the fact that if one sequentially flips a sufficiently large number of coins, a coin that always comes up heads can emerge with high likelihood. More colorfully, it is like running the newsletter scam: One selects a large number of individuals to receive a free copy of a stock market newsletter; to half the group one predicts the market will go up next week; to the other, that the market will go down. The next week, one sends the free newsletter only to those who received the correct prediction; again, half are told the market will go up and half down. The process is repeated ad libitum. After several months

[1] The author is grateful to the editor, three anonymous referees, Paul Churchland, Frank Diebold, Dimitris Politis, Ryan Sullivan, and Joseph Yukich for helpful comments, and to Douglas Stone of Nicholas Applegate Capital Management for helping to focus my attention on this topic. All errors are the author's responsibility. Support for this research was provided by NeuralNet R&D Associates and QuantMetrics R&D Associates, LLC. Computer implementations of the methods described in this paper are covered by U.S. Patent 5,893,069.

there can still be a rather large group who have received perfect predictions, and who might pay for such ''good'' forecasts.

Also problematic is the mutual fund or investment advisory service that includes past performance information as part of their solicitation. Is the past performance the result of skill or luck?

These are all examples of ''data snooping.'' Concern with this issue has a noble history. In a remarkable paper appearing in the first volume of *Econometrica*, Cowles (1933) used simulations to study whether investment advisory services performed better than chance, relative to the market. More recently, resulting biases and associated ill effects from data snooping were brought to the attention of a wide audience and well documented by Lo and MacKinley (1990). Because of these difficulties, it is widely acknowledged that data snooping is a dangerous practice to be avoided; but researchers still routinely data snoop. There is often no other choice for the analysis of time-series data, as typically only a single history for a given phenomenon of interest is available.

Data snooping is also known as data mining. Although data mining has recently acquired positive connotations as a means of extracting valuable relationships from masses of data, the negative connotations arising from the ease with which naive practitioners may mistake the spurious for the substantive are more familiar to econometricians and statisticians. Leamer (1978, 1983) has been a leader in pointing out these dangers, proposing methods for evaluating the fragility of the relationships obtained by data mining. Other relevant work is that of Mayer (1980), Miller (1981), Cox (1982), Lovell (1983), Pötscher (1991), Dufour, Ghysels, and Hall (1994), Chatfield (1995), Kabaila (1995), and Hoover and Perez (1998). Each examines issues of model selection in the context of specification searches, with specific attention to issues of inference. Recently, computer scientists have become concerned with the potential adverse effects of data mining. An informative consideration of problems of model selection and inference from this perspective is that of Jensen and Cohen (2000).

Nevertheless, none of these studies provides a rigorously founded, generally applicable method for testing the null hypothesis that the best model encountered during a specification search has no predictive superiority over a benchmark model. The purpose of this paper is to provide just such a method. This permits data snooping/mining to be undertaken with some degree of confidence that one will not mistake results that could have been generated by chance for genuinely good results.

Our null hypothesis is formulated as a multiple hypothesis, the intersection of l one-sided hypotheses, where l is the number of models considered. As such, bounds on the p-value for tests of the null can be constructed using the Bonferroni inequality (e.g. Savin (1980)) and its improvements via the union-intersection principle (Roy (1953)) or other methods (e.g. Hochberg (1988), Hommel (1989)). Resampling-based methods for implementing such tests are treated by Westfall and Young (1993). Nevertheless, as Hand (1998, p. 115) points out, ''these [multiple comparison approaches] were not designed for the sheer numbers of candidate patterns generated by data mining. This is an area

that would benefit from some careful thought.'' Thus, our goal is a method that does not rely on such bounds, but that directly delivers, at least asymptotically, appropriate *p*-values.

In taking this approach, we seek to control the simultaneous rate of error under the null hypothesis. As pointed out by a referee, one may alternatively wish to control the average rate of error (i.e., the frequency at which we find ''better'' models). Which is preferred can be a matter of taste; Miller (1981, Chapter 1) provides further discussion. Because our interest here focuses on selecting and using an apparently best model, rather than just asking whether or not a model better than the benchmark may exist, we adopt the more stringent approach of controlling the simultaneous rate of error. Nevertheless, the results presented here are also relevant for controlling average error rates, if this is desired.

2. THEORY

2.a *The Basic Framework*

We build on recent work of Diebold and Mariano (1995) and West (1996) regarding testing hypotheses about predictive ability. Our usage and notation will be similar.

Predictions are to be made for n periods, indexed from R through T, so that $T = R + n - 1$. The predictions are made for a given forecast horizon, τ. The first forecast is based on the estimator $\hat{\beta}_R$, formed using observations 1 through R, the next based on the estimator $\hat{\beta}_{R+1}$, and so forth, with the final forecast based on the estimator $\hat{\beta}_T$.

We test a hypothesis about an $l \times 1$ vector of moments, $E(f^*)$, where $f^* \equiv f(Z, \beta^*)$ is an $l \times 1$ vector with elements $f_k^* \equiv f_k(Z, \beta^*)$, for a random vector Z and parameters $\beta^* \equiv \text{plim}\, \hat{\beta}_T$. Typically, Z will consist of vectors of dependent variables, say Y, and predictor variables, say X. Our test is based on the $l \times 1$ statistic

$$\bar{f} \equiv n^{-1} \sum_{t=R}^{T} \hat{f}_{t+\tau},$$

where $\hat{f}_{t+\tau} \equiv f(Z_{t+\tau}, \hat{\beta}_t)$, and the observed data are generated by $\{Z_t\}$, a stationary strong (α-) mixing sequence having marginal distributions identical to that of Z, with the predictor variables of $Z_{t+\tau}$ available at time t. For suitable choice of f, the condition

$$H_o\colon\ E(f^*) \leq 0$$

will express the null hypothesis of no predictive superiority over a benchmark model.

Although we follow West (1996) in formulating our hypotheses in terms of β^*, it is not obvious that β^* is necessarily the most relevant parameter value for

finite samples, as a referee points out. Instead, the estimators $\hat{\beta}_t$ are the parameter values directly relevant for constructing forecasts, so alternative approaches worth consideration would be, for example, to test the hypothesis that $\lim_{n\to\infty} E(\bar{f}) \leq 0$, that $\lim_{n\to\infty} E(\bar{f} \mid Z_1, \ldots, Z_R) \leq 0$, or that $E(f_{t+\tau} \mid Z_1, \ldots, Z_t) \leq 0$. We leave these possibilities to subsequent research.

Some examples will illustrate leading cases of interest. For simplicity, set $\tau = 1$. For now, take $l = 1$.

Example 2.1: To test whether a particular set of variables has predictive power superior to that of some benchmark regression model in terms of (negative) mean squared error, take

$$\hat{f}_{t+1} = -\left(y_{t+1} - X'_{1,t+1}\hat{\beta}_{1,t}\right)^2 + \left(y_{t+1} - X'_{0,t+1}\hat{\beta}_{0,t}\right)^2,$$

where y_{t+1} is a scalar dependent variable, $\hat{\beta}_{1,t}$ is the OLS estimator based on $\{(y_s, X_{1,s}), s = 1, \ldots, t\}$ (using regressors X_1), and $\hat{\beta}_{0,t}$ is the OLS estimator based on $\{(y_s, X_{0,s}), s = 1, \ldots, t\}$, using benchmark regressors X_0. Here $\hat{\beta}_t \equiv (\hat{\beta}'_{0,t}, \hat{\beta}'_{1,t})'$. Note that the regression models need not be nested.

Example 2.2: To test whether a financial market trading strategy yields returns superior to a benchmark strategy take

$$\hat{f}_{t+1} = \log[1 + y_{t+1}S_1(X_{1,t+1}, \beta_1^*)] - \log[1 + y_{t+1}S_0(X_{0,t+1}, \beta_0^*)].$$

Here y_{t+1} represents per period returns and S_0 and S_1 are "signal" functions that convert indicators ($X_{0,t+1}$ and $X_{1,t+1}$) and given parameters (β_0^* and β_1^*) into market positions. The signal functions are step functions, with three permissible values: 1 (long), 0 (neutral), and -1 (short). As is common in examining trading strategies (e.g., Brock, Lakonishok, and LeBaron (1992)), the parameters of the systems are set a priori and do not require estimation. We are thus in Diebold and Mariano's (1995) framework. It is plausible that estimated parameters can be accommodated in the presence of step functions or other discontinuities, but we leave such cases aside here. The first log term represents returns from strategy one, while the second represents returns from the benchmark strategy. An important special case is $S_0 = 1$, the buy and hold strategy.

Example 2.3: To test generally whether a given model is superior to a benchmark, take

$$\hat{f}_{t+1} = \log L_1\left(y_{t+1}, X_{1,t+1}, \hat{\beta}_{1,t}\right) - \log L_0\left(y_{t+1}, X_{0,t+1}, \hat{\beta}_{0,t}\right),$$

where $\log L_k(y_{t+1}, X_{k,t+1}, \hat{\beta}_{k,t})$ is the predictive log-likelihood for predictive model k, based on the quasi-maximum likelihood estimator (QMLE) $\hat{\beta}_{k,t}$, $k = 0, 1$. The first example is a special case.

Not only do we have $E(f^*) \leq 0$ under the null of no predictive superiority, but the moment function also serves as a model selection criterion. Thus we can search over $l \geq 1$ specifications by assigning one moment condition/model

selection criterion to each model. To illustrate, for the third example the $l \times 1$ vector $\hat{f}_{t+1}$ now has components

$$\hat{f}_{k,t+1} = \log L_k\left(y_{t+1}, X_{k,t+1}, \hat{\beta}_{k,t}\right) - \log L_0\left(y_{t+1}, X_{0,t+1}, \hat{\beta}_{0,t}\right) \quad (k = 1, \ldots, l).$$

We select the model with the best model selection criterion value, so the appropriate null is that the best model is no better than the benchmark. Formally,

$$H_o\colon \max_{k=1,\ldots,l} E(f_k^*) \leq 0.$$

The alternative is that the best model is superior to the benchmark.

A complexity penalty to enforce model parsimony is easily incorporated; for example, to apply the Akaike Information Criterion, subtract $p_k - p_0$ from the above expression for $\hat{f}_{k,t+1}$, where $p_k (p_0)$ is the number of parameters in the kth (0th) model. We thus select the model with the best (penalized) predictive log-likelihood.

The null hypothesis H_o is a multiple hypothesis, the intersection of the one-sided individual hypotheses $E(f_k^*) \leq 0$, $k = 1, \ldots, l$. As discussed in the introduction, our goal is a method that does not rely on bounds, such as Bonferroni or its improvements, but that directly delivers, at least asymptotically, appropriate p-values.

2.b *Basic Theory*

We can provide such a method whenever $\bar{f}$, appropriately standardized, has a continuous limiting distribution. West's (1996) Main Theorem 4.1 gives convenient regularity conditions (reproduced in the Appendix as Assumption A) which ensure that

$$n^{1/2}(\bar{f} - E(f^*)) \Rightarrow N(0, \Omega),$$

where $\Rightarrow$ denotes convergence in distribution as $T \to \infty$, and Ω $(l \times l)$ is

$$\Omega = \lim_{T \to \infty} \operatorname{var}\left[n^{-1/2} \sum_{t=R}^{T} f(Z_{t+\tau}, \beta^*) \right],$$

provided that either $F \equiv E[(\partial/\partial\beta) f(Z, \beta^*)] = 0$ or $n/R \to 0$ as $T \to \infty$. When neither of these conditions holds, West's Theorem 4.1(b) establishes the same conclusion, but with a more complex expression for Ω. For Examples 2.1 and 2.3, $F = 0$ is readily verified. In Example 2.2, there are no estimated parameters, so F plays no role.

From this, West obtains standard asymptotic chi-squared statistics $n\bar{f}'\hat{\Omega}^{-1}\bar{f}$ for testing the null hypothesis $E(f^*) = 0$, where $\hat{\Omega}$ is a consistent estimator for Ω. In sharp contrast, our interest in the null hypothesis $E(f^*) \leq 0$ leads naturally to tests based on $\max_{k=1,\ldots,l} \bar{f}_k$. Methods applicable to testing $E(f^*)$

$= 0$ follow straightforwardly from our results; nevertheless, for succinctness we focus here strictly on testing $E(f^*) \leq 0$.

Our first result establishes that selecting the model with the best predictive model selection criterion does indeed identify the best model when there is one.

PROPOSITION 2.1: *Suppose that* $n^{1/2}(\bar{f} - E(f^*)) \Rightarrow N(0, \Omega)$ *for* Ω *positive semi-definite* (*e.g. Assumption A of the Appendix holds*). (*a*) *If* $E(f_k^*) > 0$ *for some* $1 \leq k \leq l$, *then for any* $0 \leq c < E(f_k^*)$, $P[\bar{f}_k > c] \to 1$ *as* $T \to \infty$. (*b*) *If* $l > 1$ *and* $E(f_1^*) > E(f_k^*)$, *for all* $k = 2, \ldots, l$, *then* $P[\bar{f}_1 > \bar{f}_k$ *for all* $k = 2, \ldots, l] \to 1$ *as* $T \to \infty$.

Part (a) says that if some model (e.g., the best model) beats the benchmark, then this is eventually revealed by a positive estimated relative performance. When $l = 1$, this result is analogous to a model selection result of Rivers and Vuong (1991), for a nonpredictive setting. It is also analogous to a model selection result of Kloek (1972) for $l \geq 1$, again in a nonpredictive setting. Part (b) says that the best model eventually has the best estimated performance relative to the benchmark, with probability approaching one.

A test of H_o for the predictive model selection criterion follows from the following proposition.

PROPOSITION 2.2: *Suppose that* $n^{1/2}(\bar{f} - E(f^*)) \Rightarrow N(0, \Omega)$ *for* Ω *positive semi-definite* (*e.g. Assumption A holds*). *Then as* $T \to \infty$

$$\max_{k=1,\ldots,l} n^{1/2}\left\{\bar{f}_k - E(f_k^*)\right\} \Rightarrow V_l \equiv \max_{k=1,\ldots,l}\{Z_k\}$$

and

$$\min_{k=1,\ldots,l} n^{1/2}\left\{\bar{f}_k - E(f_k^*)\right\} \Rightarrow W_l \equiv \min_{k=1,\ldots,l}\{Z_k\},$$

where Z *is an* $l \times 1$ *vector with components* Z_k, $k = 1, \ldots, l$, *distributed as* $N(0, \Omega)$.

Given asymptotic normality, the conclusion holds regardless of whether the null or the alternative is true. We enforce the null for testing by using the fact that the element of the null least favorable to the alternative is that $E(f_k^*) = 0$ for all k. The behavior of the predictive model selection criterion for the best model, say

$$\overline{V}_l \equiv \max_{k=1,\ldots,l} n^{1/2}\bar{f}_k,$$

is thus known under the element of the null least favorable to the alternative, approximately, for large T, permitting construction of asymptotic p-values. By enforcing the null hypothesis in this way, we obtain the critical value for the test in a manner akin to inverting a confidence interval for $\max_k E(f_k^*)$. Any method for obtaining (a consistent estimate of) a p-value for H_o: $E(f^*) \leq 0$ in the context of a specification search we call a ''Reality Check,'' as this provides an

objective measure of the extent to which apparently good results accord with the sampling variation relevant for the search.

The challenge to implementing the Reality Check is that the desired distribution, that of the extreme value of a vector of correlated normals for the general case, is not known. An analytic approach to the Reality Check is not feasible.

Nevertheless, there are at least two ways to obtain the desired p-values. The first is Monte Carlo simulation. For this, compute a consistent estimator of Ω, say $\hat{\Omega}$. For example, one can use the block resampling estimator of Politis and Romano (1994a) or the block subsampling estimator of Politis and Romano (1994c). Then one samples a large number of times from $N(0, \hat{\Omega})$ and obtains the desired p-value from the distribution of the extremes of $N(0, \hat{\Omega})$. We call this the ''Monte Carlo Reality Check'' p-value.

To appreciate the computations needed for the Monte Carlo approach, consider the addition of one more model (say model l) to the existing collection. First, we compute the new elements of the estimate $\hat{\Omega}$, its lth row, $\hat{\Omega}_l = (\hat{\Omega}_{l1}, \ldots, \hat{\Omega}_{ll})$. For concreteness, suppose we manipulate $[\hat{f}_{k,t+\tau}, k = 1, \ldots, l; t = 1, \ldots, T]$ to obtain

$$\hat{\Omega}_{lk} = \hat{\gamma}_{lk0} + \sum_{s=1}^{T} w_{Ts}(\hat{\gamma}_{kls} + \hat{\gamma}_{lks}) \qquad (k = 1, \ldots, l),$$

where w_{Ts}, $s = 1, \ldots, T$ are suitable weights and $\hat{\gamma}_{kls} \equiv (T - s)^{-1} \sum_{t=s+1}^{T} \hat{f}_{k,t+\tau} \hat{f}_{l,t+\tau-s}$.

Next, we draw independent $l \times 1$ random variables $Z_i \sim N(0, \hat{\Omega})$, $i = 1, \ldots, N$. For this, compute the Cholesky decomposition of $\hat{\Omega}$, say $\hat{C}$ (so $\hat{C}\hat{C}' = \hat{\Omega}$), and form $Z_i = \hat{C}\eta_i^l$, where η_i^l is l-variate standard normal ($N(0, I_l)$). Finally, compute the Monte Carlo Reality Check p-value from the order statistics of $\zeta_{i,l} \equiv \max_{k=1,\ldots,l} Z_{i,k}$ where $Z_i = (Z_{i1}, \ldots, Z_{il})'$.

The computational demands of constructing $\zeta_{i,l}$ can be reduced by noting that $\hat{C}$ is a triangular matrix whose lth row depends only on $\hat{\Omega}_l$ and the preceding $l - 1$ rows of $\hat{C}$. Thus, by storing $\hat{\Omega}_l$, $\hat{C}$, and $(\eta_i^{l\prime}, \zeta_{i,l})$, $i = 1, \ldots, N$, at each stage ($l = 1, 2, \ldots$), one can construct $\zeta_{i,l}$ at the next stage as $\zeta_{i,l} = \max(\zeta_{i,l-1}, \hat{C}_l \eta_i^l)$, where $\hat{C}_l$ is the $(l \times 1)$ lth row of $\hat{C}$, and η_i^l is formed recursively as $\eta_i^l = (\eta_i^{l-1\prime}, \eta_{i,l})'$, with $\eta_{i,l}$ independently drawn as (scalar) unit normal.

To summarize, obtaining the Monte Carlo Reality Check p-value requires storage and manipulation of $[\hat{f}_{k,t+\tau}]$, $\hat{\Omega}_l$, $\hat{C}$, and $(\eta_i^{l\prime}, \zeta_{i,l})$, $i = 1, \ldots, N$. These storage and manipulation requirements increase with the square of l. Also, if one is to account for the data-snooping efforts of others, their $[\hat{f}_{k,t+\tau}]$ matrix is required.

A second approach relies on the bootstrap, using a resampled version of $\bar{f}$ to deliver the ''Bootstrap Reality Check'' p-value for testing H_o. For suitably chosen random indexes $\theta(t)$, the resampled statistic is computed as

$$\bar{f}^* \equiv n^{-1} \sum_{t=R}^{T} \hat{f}^*_{t+\tau}, \qquad \hat{f}^*_{t+\tau} \equiv f\left(Z_{\theta(t)+\tau}, \hat{\beta}_{\theta(t)}\right) \qquad (t = R, \ldots, T).$$

To handle time-series data, we require a resampling procedure applicable to dependent processes. The moving blocks method of Kuensch (1989) and Liu and Singh (1992) is one such procedure. It works by constructing a resample from fixed length blocks of observations where the starting index for each block is drawn randomly. A block length of one gives the standard bootstrap, whereas larger block lengths accommodate increasing dependence. A more sophisticated version of this approach is the tapered block bootstrap of Paparoditis and Politis (2000). Although any of these methods can be validly applied, for analytic simplicity and concreteness we apply and analyze the stationary bootstrap of Politis and Romano (1994a,b) (henceforth, P & R). This procedure is analogous to the moving blocks bootstrap, but, instead of using blocks of fixed length (b, say) one uses blocks of random length, distributed according to the geometric distribution with mean block length b. As P & R show, this procedure delivers valid bootstrap approximations for means of α-mixing processes, provided b increases appropriately with n.

To implement the stationary bootstrap, P & R propose the following algorithm for obtaining the $\theta(t)$'s. Start by selecting a smoothing parameter $q = 1/b = q_n$, $0 < q_n \leq 1$, $q_n \to 0$, $n q_n \to \infty$ as $n \to \infty$, and proceed as follows: (i) Set $t = R$. Draw $\theta(R)$ at random, independently and uniformly from $\{R, \ldots, T\}$. (ii) Increment t. If $t > T$, stop. Otherwise, draw a standard uniform random variable U (supported on $[0, 1]$) independently of all other random variables. (a) If $U < q$, draw $\theta(t)$ at random, independently and uniformly from $\{R, \ldots, T\}$; (b) if $U \geq q$, set $\theta(t) = \theta(t-1) + 1$; if $\theta(t) > T$, reset to $\theta(t) = R$. (iii) Repeat (ii). As P & R show, this delivers blocks of random length, distributed according to the geometric distribution with mean block length $1/q$.

When β^* appears instead of $\hat{\beta}_{\theta(t)}$ in the definition of $\bar{f}^*$, as it does in Diebold and Mariano's (1995) setup, P & R's (1994a) Theorem 2 applies immediately to establish that under appropriate conditions (see Assumption B of the Appendix), the distribution, conditional on $\{Z_{R+\tau}, \ldots, Z_{T+\tau}\}$, of $n^{1/2}(\bar{f}^* - \bar{f})$ converges, as n increases, to that of $n^{1/2}(\bar{f} - E(f^*))$.

Thus, by repeatedly drawing realizations of $n^{1/2}(\bar{f}^* - \bar{f})$, we can build up an estimate of the desired distribution $N(0, \Omega)$. The Bootstrap Reality Check p-value for the predictive model selection statistic, $\bar{V}_l$, can then immediately be obtained from the quantiles of

$$\bar{V}_l^* \equiv \max_{k=1,\ldots,l} n^{1/2}\left(\bar{f}_k^* - \bar{f}_k\right).$$

When $\hat{\beta}_{\theta(t)}$ appears in $\bar{f}^*$, careful argument under mild additional regularity conditions delivers the same conclusion. It suffices that $\hat{\beta}_T$ obeys a law of the iterated logarithm, a refinement of the central limit theorem. With mild additional regularity (see Sin and White (1996) or Altissimo and Corradi (1996)) one can readily verify the following.

ASSUMPTION C: *Let B and H be as defined in Assumption A.2 of the Appendix, and let $G \equiv B[\lim_{T\to\infty} \mathrm{var}(T^{1/2} H(t))]B'$. For all λ $(k \times 1)$, $\lambda'\lambda = 1$,*

$$P\left[\limsup_T T^{1/2} |\lambda'\left(\hat{\beta}_T - \beta^*\right)| / \{\lambda' G \lambda \log\log(\lambda' G \lambda) T\}^{1/2} = 1\right] = 1.$$

Our main result can now be stated as follows:

THEOREM 2.3: *Suppose either*: (i) *Assumptions A and B hold and there are no estimated parameters*; *or* (ii) *Assumptions A–C hold, and either*: (a) $F = 0$ *and* $q_n = cn^{-\gamma}$ *for constants* $c > 0$, $0 < \gamma < 1$ *such that* $(n^{\gamma+\varepsilon}/R)\log\log R \to 0$ *as* $T \to \infty$ *for some* $\varepsilon > 0$; *or* (b) $(n/R)\log\log R \to 0$ *as* $T \to \infty$. *Then for* $\bar{f}^*$ *computed using P&R's stationary bootstrap,*

$$\rho\Big(\mathbf{L}\Big[n^{1/2}\big(\bar{f}^* - \bar{f}\big) \mid Z_1, \ldots, Z_{T+\tau}\Big], \mathbf{L}[n^{1/2}(\bar{f} - E(f^*))]\Big) \xrightarrow{p} 0,$$

as $T \to \infty$, *where* ρ *is any metric metrizing convergence in distribution and* $\mathbf{L}[\cdot]$ *denotes the probability law of the indicated random vector*.

Observe that the condition $(n/R)\log\log R \to 0$ appearing in (ii.b) is slightly stronger than $n/R \to 0$ appearing in West's Theorem 4.1(a). West does not require a condition linking n and R when $F = 0$; our condition in (ii.a), necessitated by the bootstrap, is nevertheless weaker than that of (ii.b), as $\gamma < 1$. It is an appealing and somewhat remarkable feature of this result that the coefficient estimates $\hat{\beta}_t$ do not have to be recomputed under the resampling. A key role in ensuring this is played by the requirements of (ii). When neither condition holds, the conclusion still holds, provided that $\bar{f}^*$ is modified to include a term estimating $-n^{-1}\sum_{t=R}^{T} \nabla f_{t+\tau}(\beta^*)(\hat{\beta}_t - \beta^*)$, such as $-n^{-1}\sum_{t=R}^{T} \nabla f^*_{t+\tau}(\hat{\beta}_T)(\hat{\beta}^*_t - \hat{\beta}_T)$. We omit further discussion of this situation for the sake of brevity.

Although the absence of the need to recompute $\hat{\beta}_t$ is quite convenient, it is possible that a method in which $\hat{\beta}_t$ is recomputed as part of the resampling could yield an improvement in the resulting approximations, as a referee points out. We leave this possibility to further research.

COROLLARY 2.4: *Under the conditions of Theorem* 2.3, *we have that as* $T \to \infty$

$$\rho\Big(\mathbf{L}\Big[\overline{V}_l^* \mid Z_1, \ldots, Z_{T+\tau}\Big], \mathbf{L}\Big[\max_{k=1,\ldots,l} n^{1/2}\big(\bar{f}_k - E(f_k^*)\big)\Big]\Big) \xrightarrow{p} 0$$

and

$$\rho\Big(\mathbf{L}\Big[\overline{W}_l^* \mid Z_1, \ldots, Z_{T+\tau}\Big], \mathbf{L}\Big[\min_{k=1,\ldots,l} n^{1/2}\big(\bar{f}_k - E(f_k^*)\big)\Big]\Big) \xrightarrow{p} 0,$$

where $\overline{W}_l^* \equiv \min_{k=1,\ldots,l} n^{1/2}(\bar{f}_k^* - \bar{f}_k)$.

Thus, by comparing $\overline{V}_l$ to the quantiles of a large sample of realizations of $\overline{V}_l^*$, we can compute a Bootstrap Reality Check p-value for testing that the best model has no predictive superiority relative to the benchmark.

When Ω is singular, we can partition $n^{1/2}(\bar{f} - E(f^*))$ as

$$\boldsymbol{Z}_n = (\boldsymbol{Z}_{1n}', \boldsymbol{Z}_{2n}')' = n^{1/2}(\bar{f} - E(f^*))$$

such that $Z_{2n} - A Z_{1n} \xrightarrow{p} 0$, where Z_{1n} is $l_1 \times 1$, $l_1 = l - l_2$, $l_2 \neq 0$, Z_{2n} is $l_2 \times 1$, and A is a finite $l_2 \times l_2$ matrix. Then Ω has the form

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{11} A' \\ A\Omega_{11} & A\Omega_{11} A' \end{bmatrix},$$

where $Z_{1n} \Rightarrow N(0, \Omega_{11})$. Corollary 2.4 continues to hold because, e.g., $\max_{k=1,\ldots,l} n^{1/2}(\bar{f}_k - E(f_k^*)) = \max[Z_n] = \max[(Z_{1n}', Z_{2n}')'] = \max[(Z_{1n}', (A Z_{1n} + (Z_{2n} - A Z_{1n}))'] = v(Z_{1n}, (Z_{2n} - A Z_{1n}))$, say, which is a continuous function of its arguments. Straightforward arguments parallel to those of Theorem 2.3 and Corollary 2.4 show that the probability law of $\bar{V}_l^* = v(Z_{1n}^*, (Z_{2n}^* - A Z_{1n}^*))$ coincides with that of $v(Z_{1n}, (Z_{2n} - A Z_{1n}))$.

The test's level can be driven to zero at the same time the power approaches one, as our test statistic diverges at rate $n^{1/2}$ under the alternative:

PROPOSITION 2.5: *Suppose that* $n^{1/2}(\bar{f}_1 - E(f_1^*)) \Rightarrow N(0, \omega_{11})$ *for* $\omega_{11} \geq 0$ (*e.g. Assumption* A.1(a) *or* A.1(b) *of the Appendix holds*), *and suppose that* $E(f_1^*) > 0$ *and, if* $l > 1$, $E(f_1^*) > E(f_k^*)$, *for all* $k = 2, \ldots, l$.

Then for any $0 < c < E(f_1^*)$, $P[\bar{V}_l > n^{1/2} c] \to 1$ *as* $T \to \infty$.

2.c *Extensions and Variations*

We now discuss some of the simpler extensions of the preceding results.

First, we let the model selection criterion be a function of a vector of averages. Examples are the prediction sample R^2 for evaluating forecasts or the prediction sample Sharpe ratio for evaluating investment strategies.

In this case we seek to test the null hypothesis

$$H_o\colon \max_{k=1,\ldots,l} g(E[h_k^*]) \leq g(E[h_0^*]),$$

where g maps U ($\subset \Re^m$) to $\Re$, with the random m-vector $h_k^* \equiv h_k(Z, \beta^*)$, $k = 0, \ldots, l$. We require that g be continuously differentiable on U, such that its Jacobian, Dg, is nonzero at $E[h_k^*] \in U$, $k = 0, \ldots, l$.

Relevant sample statistics are $\bar{f}_k \equiv g(\bar{h}_k) - g(\bar{h}_0)$, where $\bar{h}_0$ and $\bar{h}_k$ are $m \times 1$ vectors of averages computed over the prediction sample for the benchmark model and the kth specification respectively, i.e., $\bar{h}_k \equiv n^{-1} \sum_{t=R}^{T} \hat{h}_{k,t+\tau}$, $\hat{h}_{k,t+\tau} \equiv h_k(Z_{t+\tau}, \hat{\beta}_t)$, $k = 0, \ldots, l$. Relevant bootstrapped values are, for $k = 0, \ldots, l$, $\bar{f}_k^* \equiv g(\bar{h}_k^*) - g(\bar{h}_0^*)$, with $\bar{h}_k^* \equiv n^{-1} \sum_{t=R}^{T} \hat{h}_{k,t+\tau}^*$, where $\hat{h}_{k,t+\tau}^* \equiv h_k(Z_{\theta(t)+\tau}, \hat{\beta}_{\theta(t)})$, $t = R, \ldots, T$.

Let $\bar{f}$ be the $l \times 1$ vector with elements $\bar{f}_k$, let $\bar{f}^*$ be the $l \times 1$ vector with elements $\bar{f}_k^*$, and let μ^* be the $l \times 1$ vector with elements $\mu_k^* \equiv g(E[h_k^*]) - g(E[h_0^*])$, $k = 1, \ldots, l$. Under asymptotic normality, application of the mean value theorem gives

$$n^{1/2}(\bar{f} - \mu^*) \Rightarrow N(0, \Omega),$$

for suitably redefined Ω. A version of Proposition 2.2 now holds with $E(f_k^*)$ replaced by μ_k^* and F replaced by $H \equiv E(Dh(Z, \beta^*))$, where Dh is the Jacobian of h with respect to β. To state analogs of previous results, we modify Assumption A.

ASSUMPTION A: *Assumption A holds for h replacing f.*

COROLLARY 2.6: *Let* $g: U \to \Re$ $(U \subset \Re^m)$ *be continuously differentiable such that the Jacobian of g, Dg, has full row rank one at* $E[h_k^*] \in U$, $k = 0, \ldots, l$. *Suppose either*: (i) *Assumptions A′ and B hold and there are no estimated parameters*; *or* (ii) *Assumptions A′, B, and C hold, and either*: (a) $H = 0$ *and* $q_n = cn^{-\gamma}$ *for constants* $c > 0$, $0 < \gamma < 1$ *such that* $(n^{\gamma+\varepsilon}/R)\log\log R \to 0$ *as* $T \to \infty$ *for some* $\varepsilon > 0$; *or* (b) $(n/R)\log\log R \to 0$ *as* $T \to \infty$. *Then for* $\bar{f}^*$ *computed using P&R's stationary bootstrap, as* $T \to \infty$

$$\rho\Big(\mathbf{L}\Big[n^{1/2}\big(\bar{f}^* - \bar{f}\big) \mid Z_1, \ldots, Z_{T+\tau}\Big], \mathbf{L}[n^{1/2}(\bar{f} - \mu^*)]\Big) \xrightarrow{p} 0.$$

Using the original definitions of $\overline{V}_l^*$ and $\overline{W}_l^*$ in terms of $\bar{f}_k$ and $\bar{f}_k^*$ gives the following corollary.

COROLLARY 2.7: *Under the conditions of Corollary* 2.6, *we have that as* $T \to \infty$,

$$\rho\Big(\mathbf{L}\Big[\overline{V}_l^* \mid Z_1, \ldots, Z_{T+\tau}\Big], \mathbf{L}\Big[\max_{k=1,\ldots,l} n^{1/2}\big(\bar{f}_k - \mu_k^*\big)\Big]\Big) \xrightarrow{p} 0$$

and

$$\rho\Big(\mathbf{L}\Big[\overline{W}_l^* \mid Z_1, \ldots, Z_{T+\tau}\Big], \mathbf{L}\Big[\min_{k=1,\ldots,l} n^{1/2}\big(\bar{f}_k - \mu_k^*\big)\Big]\Big) \xrightarrow{p} 0.$$

As before, the test can be performed by comparing $\overline{V}_l$ to the order statistics of $\overline{V}_{l,i}^*$. Again, the test statistic diverges to infinity at rate $n^{1/2}$ under the alternative.

PROPOSITION 2.8: *Let* $\bar{f}$, μ^*, *and* Ω *be as defined above. Suppose that* $n^{1/2}(\bar{f}_1 - \mu_1^*) \Rightarrow N(0, \omega_{11})$ *for* $\omega_{11} \geq 0$, *and suppose that* $\mu_1^* > 0$ *and, if* $l > 1$, $\mu_1^* > \mu_k^*$ *for all* $k = 2, \ldots, l$.

Then for any $0 < c < \mu_1^*$, $P[\overline{V}_l > n^{1/2}c] \to 1$ *as* $T \to \infty$.

Throughout, we have assumed that $\hat{\beta}_t$ is updated with each new observation. It is easily proven that less frequent updates do not invalidate our results. The key condition is the asymptotic normality of $n^{1/2}(\bar{f} - \mu^*)$, which holds with less frequent updates, as West (1994) discusses.

Indeed, the estimated parameters need not be updated at all. If the in-sample estimate $\hat{\beta}_R$ is applied to all out-of-sample observations, the proofs simplify significantly. (Also, inferences may then be drawn conditional on $\hat{\beta}_R$, which only entails application of part (i) of Theorem 2.3 or Corollary 2.6.) Application of an

in-sample estimate to a ''hold-out'' or ''test'' dataset is common practice in cross-section modeling. It is easily proven that the Monte Carlo and Bootstrap Reality Check methods apply directly. For example, one can test whether a neural network of apparently optimal complexity (as determined from the hold-out set) provides a true improvement over a simpler benchmark, e.g., a zero hidden unit model. Applications to stratified cross-section data require replacing stationary α-mixing with suitable controlled heterogeneity assumptions for independent not identically distributed data. Results of Gonçalves and White (1999) establishing the validity of the P&R's stationary bootstrap for heterogenous near epoch dependent functions of mixing processes, analogous to results of Fitzenberger (1997) for the moving blocks bootstrap with heterogeneous α-mixing processes, suggest that this should be straightforward.

Cross-validation (Stone (1974, 1977)) represents a more sophisticated use of ''hold-out'' data. It is plausible that our methods may support testing that the best cross-validated model is no better than a given benchmark. A rigorous analysis is beyond our current scope, but is a fruitful area for further research.

Our results assume that $\hat{\beta}_t$ always uses all available data. In applications, ''rolling'' or ''moving'' window estimates are often used. These construct $\hat{\beta}_t$ from a finite length window of the most recent observations. The use of rolling/moving windows also has no adverse impact. Our results apply immediately, because the parameter estimate is now a function of a finite history of a mixing process, which is itself just another mixing process, indexed by t. The estimation aspect of the analysis thus disappears.

Typically, rolling/moving window estimates are used to handle nonexplosively nonstationary data. The results of Gonçalves and White (1999) again suggest that it should be straightforward to relax the stationarity assumption to one of controlled heterogeneity. In the rolling window case, there is again no necessity of dealing explicitly with estimation aspects of the problem.

A different type of nonstationarity important for economic modeling is that arising in the context of cointegrated processes (Engle and Granger (1987)). Recent work of Corradi, Swanson, and Olivetti (1998) shows how the present methods extend to model selection for cointegrating relationships.

Our use of the bootstrap has been solely to obtain useful approximations to the asymptotic distribution of our test statistics. As our statistics are nonpivotal, we can make no claims as to their possible higher order approximation properties, as can often be done for pivotal statistics. Nor does there appear to be any way to obtain even an asymptotically pivotal statistic for the extreme value statistics of interest here. Nevertheless, recentering and rescaling may afford improvements. We leave investigation of this issue to subsequent research.

3. IMPLEMENTING THE BOOTSTRAP REALITY CHECK

We now discuss step-by-step implementation of the Bootstrap Reality Check, demonstrating its simplicity and convenience. As we show, the Bootstrap Reality Check is especially well-suited for recursive specification searches of the sort typically undertaken in practice.

We suppose that set of n prediction observations, $t = R, \ldots, T$, is given, and that the performance/selection criterion has been decided upon. We also assume that a method for generating a collection of l model specifications has been specified.

We next specify the number of resamples, N, and the smoothing parameter, $q = q_n$. As N determines the accuracy of the p-values estimated, this should be a moderately large number, say 500 or 1000. The time required for resampling increases linearly with N on a serial computer, but resampling can proceed simultaneously on a parallel computer. Dependence in $\{Z_t\}$ is accomodated by q; the more dependence, the smaller q should be. If $\{f^*_{t+\tau}\}$ is a martingale difference sequence (at least under the null), then set q to 1. More generally, q can be determined in a data dependent manner, e.g., in a manner analogous to that analyzed by Hall, Horowitz, and Jing (1995). For concreteness and simplicity, suppose a satisfactory value for q is specified a priori, say $q = .1$.

Next we apply P&R's stationary bootstrap to generate N sets of random observation indexes of length $n, \{\theta_i(t), t = R, \ldots, T\}$, $i = 1, \ldots, N$. These indexes are generated once and for all at the outset.

Significantly, the only information required to generate the resampling indexes is R, T, q, N, an agreed upon random number generator (RNG), and the RNG seed. As we discuss further below, this enables the Bootstrap Reality Check to be carried out by researchers at separate locations and at separate times. Further, researchers do not need to share the $n \times l$ matrix of data $[\hat{f}_{k,t+\tau}]$, which might easily be unavailable. Only the scalars $\bar{V}_l, \bar{V}^*_{l,i}$, $i = 1, \ldots, N$ are required. The data storage and manipulation requirements for this are proportional to l, compared to the l^2 requirements for the Monte Carlo method.

The specification search can be conveniently done in a simple recursive manner. First, compute parameter estimates and performance values for the benchmark model, say $\hat{h}_{0,t+1} \equiv -(y_{t+1} - X'_{0,t+1}\hat{\beta}_{0,t})^2$ $(t = R, \ldots, T)$. Then compute parameter estimates and performance values for the first model, $\hat{h}_{1,t+1} \equiv -(y_{t+1} - X'_{1,t+1}\hat{\beta}_{1,t})^2$. From these form $\hat{f}_{1,t+1} = \hat{h}_{1,t+1} - \hat{h}_{0,t+1}$ and $\bar{f}_1 = n^{-1}\sum_{t=R}^{T} \hat{f}_{1,t+1}$. Using the P&R indexes we also form $\bar{f}^*_{1,i} = n^{-1}\sum_{t=R}^{T} \hat{f}_{1,\theta_i(t)+1}$, $i = 1, \ldots, N$. Now set $\bar{V}_1 = n^{1/2}\bar{f}_1$, $\bar{V}^*_{1,i} = n^{1/2}(\bar{f}^*_{1,i} - \bar{f}_1)$, $i = 1, \ldots, N$. Inferences for the first model result by comparing the sample value of $\bar{V}_1$ to the percentiles of $\bar{V}^*_{1,i}$.

For the second model, compute $\hat{h}_{2,t+1} \equiv -(y_{t+1} - X^{*\prime}_{2,t+1}\hat{\beta}_{2,t})^2$, form $\hat{f}_{2,t+1} = \hat{h}_{2,t+1} - \hat{h}_{0,t+1}$, $\bar{f}_2 = n^{-1}\sum_{t=R}^{T} \hat{f}_{2,t+1}$, and $\bar{f}^*_{2,i} = n^{-1}\sum_{t=R}^{T} \hat{f}_{2,\theta_i(t)+1}$. Then set

$$\bar{V}_2 = \max\left\{n^{1/2}\bar{f}_2, \bar{V}_1\right\}, \quad \text{and}$$

$$\bar{V}^*_{2,i} = \max\left\{n^{1/2}\left(\bar{f}^*_{2,i} - \bar{f}_2\right), \bar{V}^*_{1,i}\right\} \qquad (i = 1, \ldots, N).$$

To test whether the better of the two models beats the benchmark, compare the sample value of $\bar{V}_2$ to the percentiles of $\bar{V}^*_{2,i}$.

Proceed recursively in this manner for $k = 3, \ldots, l$, testing whether the best of the k models analyzed so far beats the benchmark by comparing the sample value of

$$\overline{V}_k = \max\left\{n^{1/2}\bar{f}_k, \overline{V}_{k-1}\right\}$$

to the percentiles of

$$\overline{V}^*_{k,i} = \max\left\{n^{1/2}\left(\bar{f}^*_{k,i} - \bar{f}_k\right), \overline{V}^*_{k-1,i}\right\} \qquad (i = 1, \ldots, N).$$

Specifically, denote the sorted values of $\overline{V}^*_{l,i}$ (the order statistics) as $\overline{V}^*_{l,(1)}$, $\overline{V}^*_{l,(2)}, \ldots, \overline{V}^*_{l,(N)}$. Find M such that $\overline{V}^*_{l,(M)} \leq \overline{V}_l < \overline{V}^*_{l,(M+1)}$. Then a simple version of the Bootstrap Reality Check p-value is

$$P_{RC} = 1 - M/N.$$

This value can be refined by interpolation or by fitting a suitable density tail model to the order statistics and obtaining the Bootstrap Reality Check p-value from the fitted model.

The recursions given for $\overline{V}_k$ and $\overline{V}^*_{k,i}$ make it clear that to continue a specification search using the Bootstrap Reality Check, it suffices to know $\overline{V}_{l-1}$, $\overline{V}^*_{l-1,i}$, $i = 1, \ldots, N$, and the P&R indexes $\theta_i(t)$. For the latter, knowledge of R, T, q, N, the RNG, and the RNG seed suffice. Knowing or storing $[\hat{f}_{k,t+\tau}]$ for $k < l$ is unnecessary, nor do we need to compute or store $\hat{\Omega}_l$, $\hat{C}$, and $(\eta^{l\prime}_i, \zeta_{i,l})$, $i = 1, \ldots, N$. This demonstrates not only a computational advantage for the Bootstrap Reality Check over the Monte Carlo version, but also the possibility for researchers at different locations or at different times to further understanding of the phenomenon modeled without needing to know the specifications tested by their collaborators or competitors. Some cooperation is nevertheless required, as R, T, q, N, the RNG, the RNG seed, and $\overline{V}_{l-1}, \overline{V}^*_{l-1,i}$, $i = 1, \ldots, N$ must still be shared, along with the data and the specification and estimation method for the benchmark model.

Subsequent specification searches can potentially contribute to understanding in two different ways. First, a better specification may be discovered; second, the p-values associated with the current best may change. The first possibility is precisely the direction in which the hope for scientific advances lies; this is what motivates economists and others to continually revisit the available data. It might be thought, however, that danger lies in the second direction: might not the p-values for the current best model erode to insignificance as the search continues, casting into doubt a model that actually represents a useful understanding?

The present theory ensures that when testing a finite number of specifications, the Reality Check p-value of a truly best model declines to zero as T grows. Nevertheless, when theory does not provide strong constraints on the number of plausible specifications, it is natural to consider what happens when l grows with

T. Even then, it is plausible that the p-value of a truly best model can still tend to zero, provided that the complexity of the collection of specifications tested is properly controlled.

The basis for this claim is that the statistic of interest, $\bar{V}_l$, is asymptotically the extreme of a Gaussian process with mean zero under the null. When the complexity (e.g., metric entropy or Vapnick-Chervonenkis dimension) of the collection of specifications is properly controlled, the extremes satisfy strong probability inequalities *uniformly* over the collection (e.g., Talagrand (1994)). These imply that the test statistic is bounded in probability under the null, so the critical value for a fixed level of test is bounded. Under the alternative, our statistic still diverges, so the power can still increase to unity, even as the level approaches zero.

Precisely this effect operates in testing for a shift in the coefficients of a regression model at an unknown point, as, e.g., in Andrews (1993). For this, one examines a growing number of models (indexed by the breakpoint) as the sample size increases. Nevertheless, power does not erode, but increases with the sample size. A rigorous treatment for our context is beyond our present scope, but these heuristics strongly suggest that a ''real'' relationship need not be buried by properly controlled data snooping. Our illustrative examples (Section 4) provide some empirical evidence on this issue.

4. AN ILLUSTRATIVE EXAMPLE

We illustrate the Reality Check by applying it to forecasting daily returns of the S&P 500. Index one day ahead ($\tau = 1$). We have a sample of daily returns from March 29, 1988 through May 31, 1994. We select $R = 803$ and $T = 1560$ to yield $n = 758$, covering the period June 3, 1991 through May 31, 1994. Daily returns are $y_t = (p_t - p_{t-1})/p_{t-1}$, where p_t is the closing price of the S&P 500 Index on trading day t.

Figure 1 plots the S&P 500 closing price and returns. The market generally trended upward, although there was a substantial pullback and retracement from day 600 (August 10, 1990) to day 725 (February 7, 1991). Somewhat higher returns volatility occurs in the first half of the period than in the last. This is nevertheless consistent with martingale difference (therefore unforecastable) excess returns, as the simple efficient markets hypothesis implies.

To see if excess returns are forecastable, we consider a collection of linear models that use ''technical'' indicators of the sort used by commodity traders, as these are easily calculated from prices and there is some recent evidence that certain such indicators may have predictive ability (Brock, Lakonishok, and LeBaron (1992)) in a period preceding that analyzed here. Altogether, we use 29 different indicators and construct forecasts using linear models including a constant and exactly three predictors chosen from the 29 available. We examine all $l = {}_{29}C_3 = 3{,}654$ models. Our benchmark model ($k = 0$) contains only a constant, embodying the simple efficient markets hypothesis.

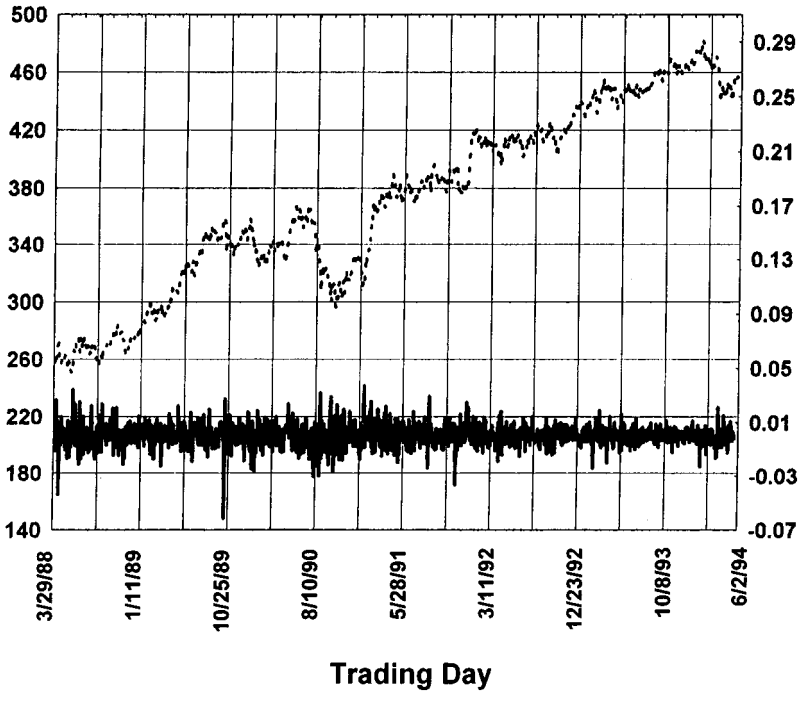

FIGURE 1.—S&P500 close and daily returns.

Notes: The finely dashed line represents the daily close of the S&P500 cash index; values can be read from the left-hand axis. The solid line represents daily returns for the S&P500 cash index, as measured from the previous day's closing price; values can be read from the right-hand axis.

The indicators consist of lagged returns ($Z_{t,1} = y_{t-1}$), a collection of "momentum" measures ($Z_{t,2}, \ldots, Z_{t,11}$), a collection of "local trend" measures ($Z_{t,12}, \ldots, Z_{t,15}$), a collection of "relative strength indexes" ($Z_{t,16}, \ldots, Z_{t,19}$), and a collection of "moving average oscillators" ($Z_{t,20}, \ldots, Z_{t,29}$).

The momentum measures are $Z_{t,j} = (p_{t-1} - p_{t-1-j})/p_{t-1-j}$, $j = 2, \ldots, 11$. The local trends ($Z_{t,12}, \ldots, Z_{t,15}$) are the slopes from regressing the price on a constant and a time trend for the previous five, ten, fifteen, and twenty days. The relative strength indexes ($Z_{t,16}, \ldots, Z_{t,19}$) are the percentages of the previous five, ten, fifteen, or twenty days that returns were positive. Each moving average oscillator is the difference between a simple moving average of the closing price over the previous q_1 days and that over the previous q_2 days, where $q_1 < q_2$, for $q_1 = 1, 5, 10, 15$, and $q_2 = 5, 10, 15, 20$. The ten possible combinations of q_1 and q_2 yield indicators ($Z_{t,20}, \ldots, Z_{t,29}$).

For each model, we compute OLS estimates for $R = 803$ through $T = 1560$. Using a version of recursive least squares (Ljung (1987, Ch. 11)) dramatically speeds computation.

We first consider the (negative) mean squared prediction error performance measure $\hat{f}_{k,t+1} = -(y_{t+1} - X^*_{k,t+1}\hat{\beta}_{1,t})^2 + (y_{t+1} - X'_{0,t+1}\hat{\beta}_{0,t})^2$, where $X_{k,t+1}$

contains a constant and three of the Z_t's, and $X_{0,t+1}$ contains a constant only. We also consider directional accuracy,

$$\hat{f}_{k,t+1} = 1\left[y_{t+1} \cdot X'_{k,t+1} \hat{\beta}_{1,t} > 0 \right] - 1\left[y_{t+1} \cdot X'_{0,t+1} \hat{\beta}_{0,t} > 0 \right],$$

where $1[\cdot]$ is the indicator function. The average of $\hat{f}_{k,t+1}$ here is the difference between the rate that specification k correctly predicts the market direction and that of a naive predictor based on average previous behavior.

Because of its computational convenience, we apply the Bootstrap Reality Check, specifying $N = 500$ and $q = .5$ for P & R's stationary bootstrap. Given the apparent lack of correlation in the regression errors, this should easily provide sufficient smoothing. In fact, Sullivan, Timmerman, and White (1999) find little sensitivity to the choice of q in a related context.

Note that Corollary 2.4 does not immediately apply to the directional accuracy case, due to the nonsmoothness of the indicator function and the presence of estimated parameters. Nevertheless, reasoning similar to that used in establishing the asymptotic normality of the least absolute deviations estimator should plausibly ensure that the conditions of Proposition 2.2 hold, so that results analogous to Corollary 2.4 (and its extension to the case in which $F \neq 0$) can be established under similar conditions. Supporting evidence is provided by Monte Carlo experiments reported in Sullivan and White (1999), where, for the case of directional accuracy with estimated parameters, the Bootstrap Reality Check delivers quite good approximations to the desired limiting distribution—better, in fact, than for the mean squared prediction error case. This gives us some assurance that the directional accuracy case is appropriate here as an illustration.

Examining the numerical results presented in Table I, we see that we fail to reject the null that the prediction mse-best model does not beat the efficient markets benchmark. This is not surprising, but without the Reality Check, there would be no way to tell whether or not we should be surprised by the observed superiority of the mse-best model.

TABLE I

REALITY CHECK RESULTS: PREDICTION MEAN SQUARED ERROR PERFORMANCE

Best predictor variables: $Z_{t,5}$, $Z_{t,13}$, $Z_{t,25}$	Best Experiment	Benchmark
RMSE	.006373	.006410
Difference in Prediction Mean Squared Error:	.4791E-06	
Bootstrap Reality Check *p*-value:	.3674	
Naive *p*-value:	.1068	

Notes: The "Difference in Prediction Mean Squared Error" is the largest difference in candidate model performance relative to the benchmark across all experiments, measured as the difference in (negative) prediction mean squared error between the candidate model for a given experiment and that of the benchmark model. The "Bootstrap Reality Check *p*-value" is that corresponding to the best model found. The "Naive *p*-value" is the Bootstrap Reality Check *p*-value computed by treating the best model as if it were the only model considered.

Conducting inference without properly accounting for the specification search can be extremely misleading. We call such a *p*-value a ''naive'' *p*-value. Applying the bootstrap to the best specification alone yields a naive *p*-value estimate of .1068, which might be considered borderline significant. The difference between the naive *p*-value and that of the Reality Check gives a direct estimate of the data-mining bias, which is seen to be fairly substantial here.

Our results lend themselves to graphical presentation, revealing several interesting features. Figure 2 shows how the Reality Check *p*-values evolve. The order of experiments is arbitrary, so only the numbers on the extreme right ultimately matter. Nevertheless, the evolution of the performance measures and the *p*-value for the best performance observed so far exhibit noteworthy features.

Specifically, we see that the *p*-value drops each time a new best performance is observed, consistent with the occurrence of a new tail event. Otherwise, the *p*-value creeps up, consistent with taking proper account of data re-use. This movement is quite gradual, and becomes even more so as the experiments

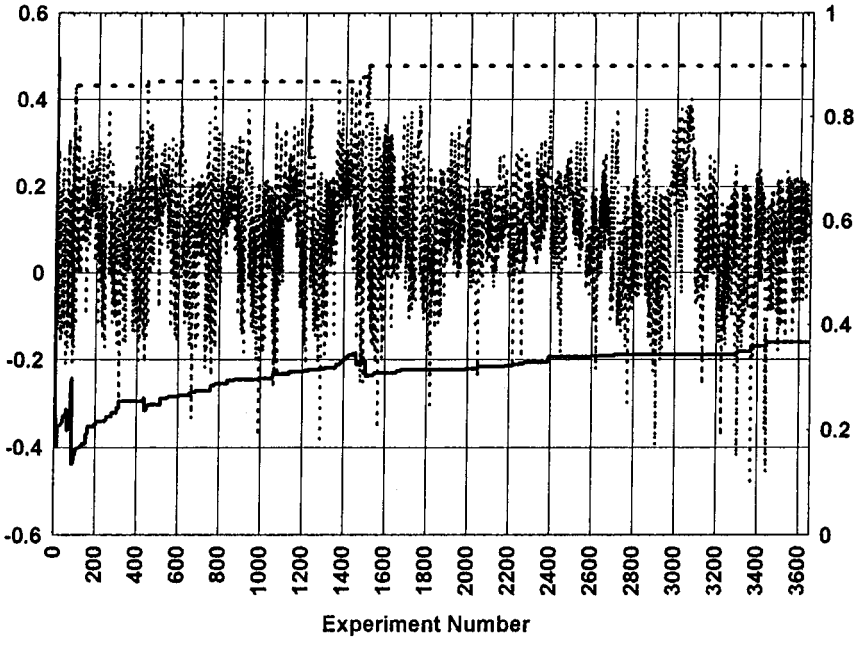

FIGURE 2.—S&P500 MSE experiments.

Notes: The finely dashed line represents candidate model performance relative to the benchmark, measured as the difference in (negative) prediction mean squared error between the candidate model for a given experiment and that of the benchmark model. The coarsely dashed line represents the best relative performance encountered as of the given experiment number. The values for both of these can be read from the left-hand axis. The solid line represents the Bootstrap Reality Check *p*-value for the best model encountered as of the given experiment number. The *p*-value can be read from the right-hand axis.

proceed. In fact, the *p*-value stays flat for modest stretches, due to the relatively high correlation among the forecasts. This illustrates that consideration of even a large number of models need not lead to dramatic erosion of the Reality Check *p*-value.

The indicators optimizing directional accuracy differ from those optimizing prediction mse. While there is an impressive gain in directional accuracy achieved by the best model, as seen in the numerical results of Table II, this is not statistically significant. This example dramatically illustrates the dangers of data mining. The naive *p*-value is .0036! Anyone relying on this would be seriously misled. Viewing the intermediate results in Figure 3, we observe features similar to those already seen in the prediction mse experiments, reinforcing our earlier observations.

Although use of the naive *p*-value is potentially dangerous, it does have value. Specifically, if the naive *p*-value is large, there is no need to compute the Reality Check *p*-value, as this can only be larger than the naive *p*-value. But if the naive *p*-value is small, one can then compute the Reality Check *p*-value in order to accurately assess the evidence against the null.

5. SUMMARY AND CONCLUDING REMARKS

Data snooping occurs when a given set of data is used more than once for purposes of inference or model selection. When such data reuse occurs, there is always the possibility that any satisfactory results obtained may simply be due to chance rather than to any merit inherent in the method yielding the results. Our new procedure, the Reality Check, provides simple and straightforward procedures for testing the null that the best model encountered in a specification search has no predictive superiority over a given benchmark model, permitting account to be taken of the effects of data snooping.

Many fascinating research topics remain. These include permitting the number of specifications tested to increase with the sample size, application of the method to the results of cross-validation, and the use of recentering, rescaling,

TABLE II

REALITY CHECK RESULTS: DIRECTIONAL ACCURACY PERFORMANCE

Best predictor variables: $Z_{t,13}$, $Z_{t,14}$, $Z_{t,26}$	Best Experiment	Benchmark
Percent Correct	54.7493	50.7916
Difference in Prediction Directional Accuracy:	.0396	
Bootstrap Reality Check *p*-value:	.2040	
Naive *p*-value:	.0036	

Notes: The "Difference in Prediction Directional Accuracy" is the largest difference in candidate model performance relative to the benchmark across all experiments, measured as the difference in the proportion of correct predicted direction between the candidate model for a given experiment and that of the benchmark model. The "Bootstrap Reality Check *p*-value" is that corresponding to the best model found. The "Naive *p*-value" is the Bootstrap Reality Check *p*-value computed by treating the best model as if it were the only model considered.

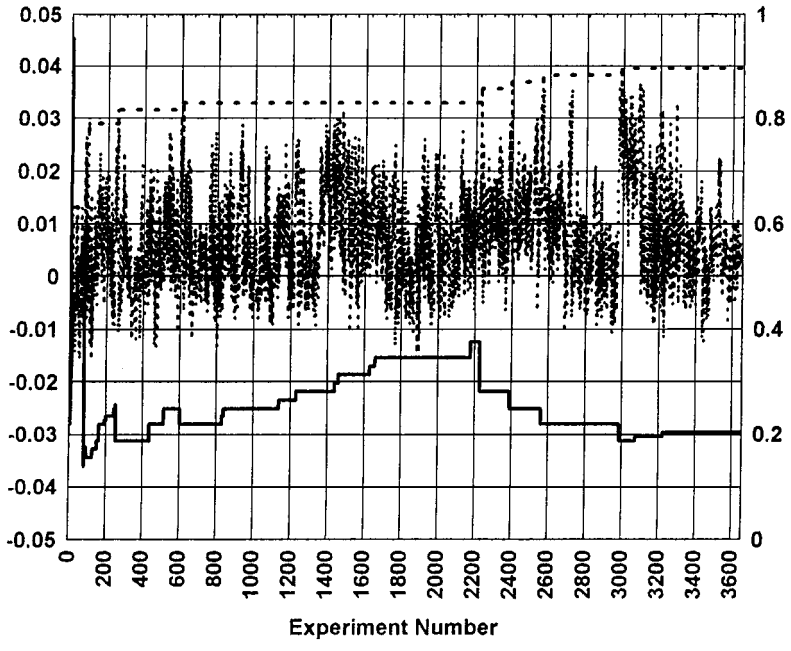

FIGURE 3.—S&P500 direction experiments.

Notes: The finely dashed line represents candidate model performance relative to the benchmark, measured as the difference in the proportion of correct predicted direction between the candidate model for a given experiment and that of the benchmark model. The coarsely dashed line represents the best relative performance encountered as of the given experiment number. The values for both of these can be read from the left-hand axis. The solid line represents the Bootstrap Reality Check *p*-value for the best model encountered as of the given experiment number. The *p*-value can be read from the right-hand axis.

or other modifications to achieve improvements in sampling distribution approximations.

Simulation studies of the finite sample properties of both the Monte Carlo and the Bootstrap versions of the Reality Check are a top priority. A first step in this direction is Sullivan and White (1999), in which we find that the tests typically (though not always) appear conservative, that test performance is relatively insensitive to the choice of the bootstrap smoothing parameter q, and that there is much better agreement between actual and bootstrapped critical values when the performance measure has fewer extreme outlying values.

Finally, and of particular significance for economics, finance, and other domains where our scientific world-view has been shaped by studies in which data reuse has been the unavoidable standard practice, there is now the opportunity for a re-assessment of that world-view, taking into account the effects of data reuse. Do we really know what we think we know? That is, will currently accepted theories withstand the challenges posed by a quantitative accounting of the effects of data snooping? A start in this direction is made by

studies of technical trading rules (Sullivan, Timmermann, and White (1999)) and calendar effects (Sullivan, Timmermann, and White (1998)) in the asset markets.

Those of us who study phenomena generated once and for all by a system outside our control lack the inferential luxuries afforded to the experimental sciences. Nevertheless, through the application of such methods as described here, we need no longer necessarily suffer the poverty enforced by our previous ignorance of the quantitative effects of data reuse.

Dept. of Economics, University of California, San Diego, and QuantMetrics R&D Associates, LLC, 6540 Lusk Blvd., Suite C-157, San Diego, CA 92121, U.S.A.; halwhite@earthlink.net

Manuscript received June, 1997; final revision received July, 1999.

MATHEMATICAL APPENDIX

In what follows, the notation corresponds to that of the text unless otherwise noted. For convenience, we reproduce West's (1996) assumptions with this notation.

ASSUMPTION A:

A.1: *In some open neighborhood N around β^*, and with probability one*: (*a*) $f_t(\beta)$ *is measurable and twice continuously differentiable with respect to* β; (*b*) *let* f_{it} *be the ith element of* f_t; *for* $i = 1, \ldots, l$ *there is a constant* $D < \infty$ *such that for all* t, $\sup_{\beta \in N} |\partial^2 f_{it}(\beta)/\partial\beta\partial\beta'| < m_t$ *for a measurable* m_t, *for which* $Em_t < D$.

A.2: *The estimate* $\hat{\beta}_t$ *satisfies* $\hat{\beta}_t - \beta^* = B(t)H(t)$, *where* $B(t)$ *is* $(k \times q)$ *and* $H(t)$ *is* $(q \times 1)$, *with* (*a*) $B(t) \xrightarrow{a.s.} B$, B *a matrix of rank* k; (*b*) $H(t) = t^{-1}\sum_{s=1}^{t} h_s(\beta^*)$ *for a* $(q \times 1)$ *orthogonality condition* $h_S(\beta^*)$; (*c*) $Eh_s(\beta^*) = 0$.

Let

$$f_t^* \equiv f_t(\beta^*), \qquad f_{t\beta}^* \equiv \frac{\partial f_t}{\partial \beta}(\beta^*), \qquad F \equiv Ef_{t\beta}^*.$$

A.3: (a) *For some* $d > 1$, $\sup_t E\|[\text{vec}(f_{t\beta}^*)', f_t^{*\prime}, h_t^{*\prime}]'\|^{4d} < \infty$, *where* $\|\cdot\|$ *denotes Euclidean norm*. (b) $[\text{vec}(f_{t\beta}^* - F)', (f_t^* - Ef_t^{*\prime}), h_t^{*\prime}]'$ *is strong mixing, with mixing coefficients of size* $-3d/(d-1)$. (c) $[\text{vec}(f_{t\beta}^*)', f_t^{*\prime}, h_t^{*\prime}]'$ *is covariance stationary.* (d) *Let* $\Gamma_{ff}(j) = E(f_t^* - Ef_t^*)(f_{t-j}' - Ef_t^*)'$, $S_{ff} = \sum_{j=-\infty}^{\infty}\Gamma_{ff}(j)$. *Then* S_{ff} *is p.d.*

A.4: $R, n \to \infty$ *as* $T \to \infty$, *and* $\lim_{T\to\infty}(n/R) = \pi$, $0 \le \pi, \le \infty$; $\pi = \infty \Leftrightarrow \lim_{T\to\infty}(R/n) = 0$.

A.5: *Either*: (a) $\pi = 0$ *or* $F = 0$; *or* (b) S *is positive definite, where* (*West* (1996, *pp.* 1071–1072))

$$S \equiv \begin{bmatrix} S_{ff} & S_{fh}B' \\ BS_{hf} & BS_{hh}B' \end{bmatrix}.$$

We let P denote the probability measure governing the behavior of the time series $\{Z_t\}$.

PROOF OF PROPOSITION 2.1: We first prove (b). Suppose first that Ω is positive definite, so that for all k, $S_k'\Omega S_k > 0$, where S_k is an $l \times 1$ vector with 1 in the first element, -1 in the kth element and zeroes elsewhere. Let $A_k = [\bar{f}_1 > \bar{f}_k]$. We seek to show $P[\bigcap_{k=2}^{l} A_k] \to 1$ or equivalently that $P[\bigcup_{k=2}^{l} A_k^c] \to 0$. As $P[\bigcup_{k=2}^{l} A_k^c] \leq \sum_{k=2}^{l} P[A_k]$, it suffices that for any $\varepsilon > 0$, $\max_{2 \leq k \leq l} P[A_k^c] < \varepsilon/l$ for all T sufficiently large. Now

$$P[A_k^c] = P[\bar{f}_1 - \bar{f}_k \leq 0] = P[n^{1/2}(\bar{f}_1 - E(f_1^*) - [\bar{f}_k - E(f_k^*)])/S_k'\Omega S_k \leq n^{1/2}(E(f_k^*) - E(f_1^*))/S_k'\Omega S_k].$$

By the assumed asymptotic normality, we have the unit asymptotic normality of $Z_k \equiv n^{1/2}(\bar{f}_1 - E(f_1^*) - [\bar{f}_k - E(f_k^*)])/S_k'\Omega S_k$, so that

$$\begin{aligned} P[A_k^c] &= \Phi(z_k) + P[Z_k \leq z_k] - \Phi(z_k) \\ &\leq \Phi(z_k) + \sup_z |P[Z_k \leq z] - \Phi(z)|, \end{aligned}$$

where $z_k \equiv n^{1/2}(E(f_k^*) - E(f_1^*))/S_k'\Omega S_k$. Because $E(f_1^*) > E(f_k^*)$ and $S_k'\Omega S_k < \infty$ we have $z_k \to -\infty$ as $T \to \infty$ and we can pick T sufficiently large that $\Phi(z_k) < \varepsilon/2l$, uniformly in k. Polya's theorem (e.g. Rao (1973, p. 118)) applies given the continuity of Φ to ensure that for T sufficiently large $\sup_z |P[Z_t \leq z] - \Phi(z)| < \varepsilon/2l$. Hence for all k we have $P[A_k^c] < \varepsilon/l$ for all T sufficiently large, and the first result follows. Replacing A_k with $A_k' = [\bar{f}_1 > c]$ and arguing analogously gives (a).

Now suppose that Ω is positive semi-definite, such that for one or more values of k, $S_k'\Omega S_k = 0$. Then, redefining Z_k to be $Z_k \equiv n^{1/2}(\bar{f}_1 - E(f_1^*) - [\bar{f}_k - E(f_k^*)])$, we have $Z_k \xrightarrow{p} 0$, so that

$$\begin{aligned} P[A_k^c] &= P[\bar{f}_1 - \bar{f}_k \leq 0] \\ &= P[n^{1/2}(\bar{f}_1 - E(f_1^*) - [\bar{f}_k - E(f_k^*)]) \leq n^{1/2}(E(f_k^*) - E(f_1^*))] \\ &= P[Z_k \leq z_k], \end{aligned}$$

where now $z_k \equiv n^{1/2}(E(f_k^*) - E(f_1^*))$. Because $E(f_1^*) > E(f_k^*)$ we have $z_k < -\delta$ for any $\delta > 0$ and all T sufficiently large. It follows from $Z_k \xrightarrow{p} 0$ that for all T sufficiently large we have for suitable choice of δ that $P[Z_k \leq z_k] \leq P[Z_k \leq -\delta] < \varepsilon/2l$, uniformly in k. The results now follow as before. *Q.E.D.*

PROOF OF PROPOSITION 2.2: By assumption, $n^{1/2}(\bar{f} - E(f)) \Rightarrow N(0, \Omega)$. As the maximum or minimum of a vector is a continuous function of the elements of the vector, the results claimed follow immediately from the Continuous Mapping Theorem (Billingsley (1968, Theorem 2.2)). *Q.E.D.*

The proof of our main result (Theorem 2.3) uses the following result of Politis (1999).

LEMMA A.1: *Let* $\{X_{nt}^*\}$ *be obtained by P&R's stationary bootstrap applied to random variables* $\{X_1, \ldots, X_n\}$ *using smoothing parameter* q_n, *and let* $\alpha_n^*(k)$ *denote the* α*-mixing coefficients for* $\{X_{nt}^*\}$ *under the bootstrap probability conditional on* $\{X_1, \ldots, X_n\}$. *Then*: (i) $\alpha_n^*(k) = n(1 - q_n)^k$ *for all* k *sufficiently large*; *and* (ii) *if* $q_n = cn^{-\gamma}$ *for some constants* $c > 0$ *and* $0 < \gamma < 1$, *then* $\alpha_n^*(k) \leq n \exp(-ckn^{-\gamma})$ *for all* $k \geq n^{\gamma}$.

PROOF: (i) The finite Markov chain $\{X_{nt}^*\}$ has transition probability $P^*[X_{n,t+1}^* = x \mid X_{n,t}^* = x_i] = q_n/n$ for $x \in \{x_1, \ldots, x_i\} \cup \{x_{i+2}, \ldots, x_n\}$ and $= 1 - q_n + q_n/n$ for $x = x_{i+1}$, where P^* denotes bootstrap probability conditional on $\{X_1, \ldots, X_n\}$. For all n sufficiently large, the minimum transition probability is q_n/n. As the Markov chain has n states, Billingsley (1995, Example 27.6) implies $\alpha_n^*(k) = n(1 - nq_n/n)^k = n(1 - q_n)^k$. (ii) Substituting $q_n = cn^{-\gamma}$ gives $\alpha_n^*(k) = n(1 - cn^{-\gamma})^k = n(1 - cn^{-\gamma})^{(n^{\gamma}/c)ck/n^{\gamma}} \leq n \exp(-ckn^{-\gamma})$. *Q.E.D.*

Next, we provide a statement of a version of P&R's (1994a) Theorem 2.

THEOREM A.2: *Let $X_1, X_2, \ldots$ be a strictly stationary process, with $E|X_1|^{6+\varepsilon} < \infty$ for some $\varepsilon > 0$, and let $\mu \equiv E(X_1)$ and $\bar{X}_n \equiv n^{-1}\sum_{t=1}^{n} X_t$. Suppose that $\{X_t\}$ is α-mixing with $\alpha(k) = O(k^{-r})$ for some $r > 3(6+\varepsilon)/\varepsilon$. Then $\sigma_\infty^2 \equiv \lim_{n\to\infty} \mathrm{var}(n^{1/2}\bar{X}_n)$ is finite. Moreover, if $\sigma_\infty > 0$, then*

$$\sup_x |P\{n^{1/2}(\bar{X}_n - \mu) \le x\} - \Phi(x/\sigma_\infty)| \to 0,$$

where Φ is the standard normal cumulative distribution function. Assume that $q_n \to 0$ and $nq_n \to \infty$ as $n \to \infty$. Then for $\{X_t^*\}$ obtained by P&R's stationary bootstrap

$$\sup_x |P\{n^{1/2}(\bar{X}_n^* - \bar{X}_n) \le x \mid X_1, \ldots, X_n\} - P\{n^{1/2}(\bar{X}_n - \mu) \le x\}| \xrightarrow{p} 0,$$

where $\bar{X}_n^* \equiv n^{-1}\sum_{t=1}^{n} X_t^*$.

Now we can state our next assumption:

ASSUMPTION B: *The conditions of Theorem A.2 hold for each element of f_t^*.*

Note that Assumption B ensures that the conditions of Theorem A.2 hold for $X_t = \lambda' f_t^*$ with $\sigma_\infty > 0$ for every λ, $\lambda'\lambda = 1$, given the positive definiteness of S, thus justifying the use of the Cramer-Wold device.

Our next lemma provides a convenient approach to establishing the validity of bootstrap methods in general situations similar to ours. Similar methods have been used by Liu and Singh (1992) and Politis and Romano (1992), but to the best of our knowledge, a formal statement of this approach has not previously been given.

LEMMA A.3: *Let $(\Omega, \mathbf{F}, P)$ be a complete probability space and for each $\omega \in \Omega$ let $(\Lambda, \mathbf{G}, Q_\omega)$ be a complete probability space. For $m, n = 1, 2, \ldots$ and each $\omega \in \Omega$ define*

$$T_{m,n}(\cdot, \omega) = S_{m,n}(\cdot, \omega) + X_{m,n}(\cdot, \omega) + Y_n(\omega),$$

where $S_{m,n}(\cdot, \omega): \Lambda \to \Re$ and $X_{m,n}(\cdot, \omega): \Lambda \to \Re$ are measurable-$\mathbf{G}$. Suppose also that for each $\lambda \in \Lambda$, $S_{m,n}(\lambda, \cdot): \Omega \to \Re$ and $X_{m,n}(\lambda, \cdot): \Omega \to \Re$ are measurable-$\mathbf{F}$. Let $Y_n: \Omega \to \Re$ be measurable-$\mathbf{F}$ such that $Y_n = o_P(1)$.

Suppose there exist random variables $Z_n(\cdot, \omega)$ on $(\Lambda, \mathbf{G}, Q_\omega)$ such that for each $\lambda \in \Lambda$, $Z_n(\lambda, \cdot): \Omega \to \Re$ is measurable-$\mathbf{F}$ with $P[C_n] \to 1$ as $n \to \infty$, for

$$C_n \equiv \{\omega \mid S_{m,n}(\cdot, \omega) \Rightarrow_{Q_\omega} Z_n(\cdot, \omega) \text{ as } m \to \infty\},$$

where $\Rightarrow_{Q_\omega}$ denotes convergence in distribution under the measure Q_ω, with

$$\sup_{z \in \Re} |F_n(z, \cdot) - F(z)| = o_P(1),$$

where $F_n(z, \omega) \equiv Q_\omega[Z_n(\cdot, \omega) \le z]$ for some cumulative distribution function F, continuous on $\Re$. Suppose further that $P[D_n] \to 1$ as $n \to \infty$, for

$$D_n \equiv \{\omega \mid X_{m,n}(\cdot, \omega) \to_{Q_\omega} 0 \text{ as } m \to \infty\},$$

where $\to_{Q_\omega}$ denotes convergence in probability under Q_ω.

Let $m = m_n \to \infty$ as $n \to \infty$. Then for all $\varepsilon > 0$,

$$P\{\omega | \sup_{z \in \Re} |Q_\omega[T_{m,n}(\cdot, \omega) \le z] - F(z)| > \varepsilon\} \to 0 \text{ as } n \to \infty.$$

PROOF: The asymptotic equivalence lemma (e.g., White (1984, Lemma 4.7)) ensures that when $S_{m,n}(\cdot, \omega) \Rightarrow_{Q_\omega} Z_n(\cdot, \omega)$ and $X_{m,n}(\cdot, \omega) = o_{Q_\omega}(1)$, it follows that $S_{m,n}(\cdot, \omega) + X_{m,n}(\cdot, \omega) \Rightarrow_{Q_\omega} Z_n(\cdot, \omega)$, which holds for all ω in $C_n \cap D_n$, $P[C_n \cap D_n] \to 1$. It thus suffices to prove the result for $T_{m,n}(\cdot, \omega) = S_{m,n}(\cdot, \omega) + Y_n(\omega)$.

For notational convenience, we suppress the dependence of m_n on n, writing $m = m_n$ throughout. Pick $\varepsilon > 0$, and for δ to be chosen below define

$$A_{n,\delta} \equiv \{\omega \mid |Y_n(\omega)| > \delta\},$$
$$B_{n,\delta} \equiv \{\omega | \sup_z |F_n(z,\omega) - F(z)| > \delta\}.$$

Because $Y_n = o_p(1)$ and $\sup_z |F_n(\omega, z) - F(z)| = o_p(1)$, we can choose n sufficiently large that $P[A_{n,\delta}] < \varepsilon/3$, $P[B_{n,\delta}] < \varepsilon/3$, and $P[C_n^c] < \varepsilon/3$.

For $\omega \in K_n \equiv A_{n,\delta}^c \cap B_{n,\delta}^c \cap C_n (\subset A_{n,\delta}^c)$ we have $|Y_n(\omega)| \le \delta$. This and $S_{m,n}(\cdot,\omega) \le z - \delta$ imply $T_{m,n}(\cdot,\omega) \le z$, so that for $\omega \in K_n$

$$Q_\omega[S_{m,n}(\cdot,\omega) \le z - \delta] \le Q_\omega[T_{m,n}(\cdot,\omega) \le z].$$

Similarly, $|Y_n(\omega)| \le \delta$ and $T_{m,n}(\cdot,\omega) \le z$ imply $S_{m,n}(\cdot,\omega) \le z + \delta$, so that for $\omega \in K_n$

$$Q_\omega[T_{m,n}(\cdot,\omega) \le z] \le Q_\omega[S_{m,n}(\cdot,\omega) \le z + \delta].$$

Subtracting $Q_\omega[S_{m,n}(\cdot,\omega) \le z]$ from these inequalities for $\omega \in K_n$ gives

$$\begin{aligned} &Q_\omega[S_{m,n}(\cdot,\omega) \le z - \delta] - Q_\omega[S_{m,n}(\cdot,\omega) \le z] \\ &\quad \le Q_\omega[T_{m,n}(\cdot,\omega) \le z] - Q_\omega[S_{m,n}(\cdot,\omega) \le z] \\ &\quad \le Q_\omega[S_{m,n}(\cdot,\omega) \le z + \delta] - Q_\omega[S_{m,n}(\cdot,\omega) \le z]. \end{aligned}$$

We argue explicitly using the second inequality; an analogous argument applies to the first. From the triangle inequality applied to the last expression (which is nonnegative) we have

$$\begin{aligned} &Q_\omega[T_{m,n}(\cdot,\omega) \le z] - Q_\omega[S_{m,n}(\cdot,\omega) \le z] \\ &\quad \le |Q_\omega[S_{m,n}(\cdot,\omega) \le z + \delta] - F_n(z+\delta,\omega)| \\ &\qquad + |Q_\omega[S_{m,n}(\cdot,\omega) \le z] - F_n(z,\omega)| \\ &\qquad + |F_n(z+\delta,\omega) - F(z+\delta)| + |F_n(z,\omega) - F(z)| \\ &\qquad + |F(z+\delta) - F(z)|. \end{aligned}$$

For $\omega \in K_n$ $(\subset C_n)$ we can choose n (hence m) sufficiently large that each of the first two terms is bounded by $\varepsilon/7$, uniformly in z by Polya's theorem, given the continuity of $F_n(\cdot,\omega)$ for n sufficiently large ensured by the uniform convergence of F_n to F and the continuity of F. For n sufficiently large, the next two terms are each bounded by $\varepsilon/7$, uniformly in z for $\omega \in K_n$ $(\subset B_{n,\delta}^c)$. The continuity of F (uniformly) on $\Re$ ensures that we can pick δ_ε sufficiently small that for n sufficiently large, the last term is bounded by $\varepsilon/7$, uniformly in z, so that for $\omega \in K_n$ we have

$$Q_\omega[T_{m,n}(\cdot,\omega) \le z] - Q_\omega[S_{m,n}(\cdot,\omega) \le z] \le 5\varepsilon/7,$$

uniformly in z. Analogous argument for the lower bound with $\omega \in K_n$ gives

$$-5\varepsilon/7 \le Q_\omega[T_{m,n}(\cdot,\omega) \le z] - Q_\omega[S_{m,n}(\cdot,\omega) \le z],$$

so that uniformly in z

$$|Q_\omega[T_{m,n}(\cdot,\omega) \le z] - Q_\omega[S_{m,n}(\cdot,\omega) \le z]| \le \varepsilon/7.$$

By the triangle inequality we have

$$\begin{aligned} &\sup_{z \in \Re} |Q_\omega[T_{m,n}(\cdot,\omega) \le z] - F(z)| \\ &\quad \le \sup_{z \in \Re} |Q_\omega[T_{m,n}(\cdot,\omega) \le z] - Q_\omega[S_{m,n}(\cdot,\omega) \le z]| \\ &\qquad + \sup_{z \in \Re} |Q_\omega[S_{m,n}(\cdot,\omega) \le z] - F_n(z,\omega)| \\ &\qquad + \sup_{z \in \Re} |F_n(z,\omega) - F(z)| \\ &\quad \le 5\varepsilon/7 + \varepsilon/7 + \varepsilon/7 = \varepsilon \end{aligned}$$

for all n sufficiently large and $\omega \in K_n$, which ensures that the second term is bounded by $\varepsilon/7$ ($K_n \subset C_n$) and that the final term is also bounded by $\varepsilon/7$ ($K_n \subset B_{n,\delta}^c$). Thus K_n implies

$$L_n \equiv \{\omega | \sup_{z \in \Re} | Q_\omega[T_{m,n}(\cdot, \omega) \leq z] - F(z) | \leq \varepsilon\},$$

so that $P[L_n^c] \leq P[K_n^c] \leq P[A_{n,\delta}] + P[B_{n,\delta}] + P[C_n^c] \leq \varepsilon$ for all n sufficiently large. But ε is arbitrary, and the result follows. *Q.E.D.*

PROOF OF THEOREM 2.3: We prove only the result for (ii). That for (i) is immediate. We denote $f_{t+\tau}^{**} \equiv f(Z_{\theta(t)+\tau}, \beta^*)$. Adding and subtracting appropriately gives

$$\begin{aligned} n^{1/2}(\bar{f}^* - \bar{f}) &= n^{-1/2} \sum_{t=R}^{T} \hat{f}_{t+\tau}^* - \hat{f}_{t+\tau} \\ &= n^{-1/2} \sum_{t=R}^{T} [f_{t+\tau}^{**} - f_{t+\tau}^*] - n^{-1/2} \sum_{t=R}^{T} [\hat{f}_{t+\tau} - f_{t+\tau}^*] \\ &\quad + n^{-1/2} \sum_{t=R}^{T} [\hat{f}_{t+\tau}^* - f_{t+\tau}^{**}] \\ &\equiv \zeta_{1n} + \zeta_{2n} + \zeta_{3n}, \end{aligned}$$

with obvious definitions. Under Assumption B, Theorem A.2 ensures that ζ_{1n} obeys the conditions imposed on $S_{m,n}$ in Lemma A.3 with $m = n$ and $F(z) = \Phi(z/\sigma_\infty)$. Applying West (1996, p. 1081) to ζ_{2n} ensures that $\zeta_{2n} \to 0$ a.s., hence in probability-P, satisfying the conditions imposed on Y_n in Lemma A.3. The result follows from Lemma A.3 if $P[\zeta_{3n} = o_Q(1)] \to 1$ as n increases, where Q is the probability distribution induced by the stationary bootstrap (conditional on $Z_1, \ldots, Z_{T+\tau}$), so that ζ_{3n} satisfies the conditions imposed on $X_{n,m}$ in Lemma A.3 with $m = n$. For notational convenience, we suppress the dependence of Q on ω.

By a mean value expansion, we have

$$\zeta_{3n} = n^{-1/2} \sum_{t=R}^{T} \nabla f_{t+\tau}^{**} \cdot (\hat{\beta}_t^* - \beta^*) + n^{-1/2} \sum_{t=R}^{T} w_{t+\tau}^*,$$

where $\nabla f_{t+\tau}^{**} \equiv \nabla f(Z_{\theta(t)+\tau}, \beta^*)$ and w_{t+t}^* is the vector with elements

$$w_{j,t+\tau}^* = (\hat{\beta}_t^* - \beta^*)'[\partial^2 f_{jt+\tau}(\bar{\beta}_{(j),t}^*)/\partial\beta\partial\beta'](\hat{\beta}_t^* - \beta^*),$$

with $\bar{\beta}_{(j),t}^*$ a mean value lying between $\hat{\beta}_t^*$ and β^*. Routine arguments deliver $n^{-1/2}\sum_{t=R}^{T} w_{t+\tau}^* = o_Q(1)$ with probability-P approaching one. It remains to show that the first term of ζ_{3n} vanishes in probability-Q with probability-P approaching one.

To proceed, we write $Z_{t-R+1}^* = X_{t-R+1}^* \delta_{t-R+1}^* \equiv \nabla f_{t+\tau}^{**} \cdot (\hat{\beta}_t^* - \beta^*)$, with $\delta_{t-R+1}^* \equiv (\hat{\beta}_t^* - \beta^*)$. By Chebyshev's inequality

$$Q\left[|n^{-1/2} \sum_{t=1}^{n} (Z_t^* - E_Q(Z_t^*))| \geq \varepsilon\right] \leq \varepsilon^{-2} \operatorname{var}_Q\left[n^{-1/2} \sum_{t=1}^{n} (Z_t^* - E_Q(Z_t^*))\right],$$

where E_Q and var_Q are the expectation and variance induced by probability measure Q.

We now show that $\operatorname{var}_Q[n^{-1/2}\sum_{t=1}^{n}(Z_t^* - E_Q(Z_t^*))] \to 0$. By Proposition 3.1 of P & R (1994a) and Lemma A.1, $\{Z_t^*\}$ is stationary and α-mixing. Standard inequalities for α-mixing processes (e.g.,

White (1984, Lemma 6.15)) ensure that

$$\begin{aligned}
&\mathrm{var}_Q\left[n^{-1/2}\sum_{t=1}^{n}(\boldsymbol{Z}_t^* - E_Q(\boldsymbol{Z}_t^*))\right] \\
&\quad = \mathrm{var}_Q(\boldsymbol{Z}_1^*) + 2\sum_{\tau=1}^{n-1}(1-\tau/n)\mathrm{cov}_Q(\boldsymbol{Z}_1^*, \boldsymbol{Z}_{1+\tau}^*) \\
&\quad \le \mathrm{var}_Q(\boldsymbol{Z}_1^*) + 2(2^{1/2}+1)[\mathrm{var}_Q(\boldsymbol{Z}_1^*)]^{1/2}\|\boldsymbol{Z}_1^* - E(\boldsymbol{Z}_1^*)\|_{Q,r} \\
&\qquad \times \sum_{\tau=1}^{n-1}(1-\tau/n)\alpha_n^*(\tau)^{1/2-1/r},
\end{aligned}$$

where $\|\boldsymbol{Z}\|_{Q,r} \equiv (E_Q|\boldsymbol{Z}|^r)^{1/r}$ for some $r > 2$, and we make use of stationarity in writing the equality. Now $\|\boldsymbol{Z}_1^* - E(\boldsymbol{Z}_1^*)\|_{Q,r} \le 2\|\boldsymbol{Z}_1^*\|_{Q,r}$ and $[\mathrm{var}_Q(\boldsymbol{Z}_1^*)]^{1/2} = \|\boldsymbol{Z}_1^* - E(\boldsymbol{Z}_1^*)\|_{Q,2} \le \|\boldsymbol{Z}_1^*\|_{Q,2} \le \|\boldsymbol{Z}_1^*\|_{Q,r}$ (by Jensen's inequality). Thus,

$$\begin{aligned}
&\mathrm{var}_Q\left[n^{-1/2}\sum_{t=1}^{n}(\boldsymbol{Z}_t^* - E_Q(\boldsymbol{Z}_t^*))\right] \\
&\quad \le (\|\boldsymbol{Z}_1^*\|_{Q,r})^2\left(1 + 4(2^{1/2}+1)\sum_{\tau=1}^{n-1}(1-\tau/n)\alpha_n^*(\tau)^{1/2-1/r}\right).
\end{aligned}$$

By Minkowski's inequality, $\|\boldsymbol{Z}_1^*\|_{Q,r} = \|\Sigma_j \boldsymbol{X}_{j1}^*\delta_{j1}^*\|_{Q,r} \le \Sigma_j\|\boldsymbol{X}_{j1}^*\delta_{j1}^*\|_{Q,r}$, where $\boldsymbol{X}'_{j1}$ is the jth component of $\nabla f_{R+\tau}^{**}$ and δ_{j1}^* is the jth component of $(\hat{\beta}_R^* - \beta^*) = (\hat{\beta}_\theta - \beta^*)$ for some randomly chosen θ, $R \le \theta \le T$. Assumption C (law of iterated logarithm) ensures that for all t sufficiently large (almost all t, a.a. t) $t^{1/2}|\hat{\beta}_{jt} - \beta_j^*|/\sigma_j(\log\log t\sigma_j)^{1/2} \le 1$ a.s.-P, where σ_j is the jth diagonal element of G. Thus, $|\hat{\beta}_{jt} - \beta_j^*| \le \sigma_j(\log\log R\sigma_j)^{1/2}/R^{1/2}$ for all $t \ge R$, a.a. R, a.s.-P, so that $|\delta_{j1}^*| \le \bar{\sigma}(\log\log R\bar{\sigma})^{1/2}/R^{1/2}$, a.a. R, a.s.-P, where $\bar{\sigma} \equiv \max_j \sigma_j$. Thus, $\|\boldsymbol{Z}_1^*\|_{Q,r} \le [\bar{\sigma}(\log\log R\bar{\sigma})^{1/2}/R^{1/2}]\Sigma_j\|\boldsymbol{X}_{j1}^*\|_{Q,r}$, a.a. R, a.s.-P. By Assumption A.3, $\|\boldsymbol{X}_{j1}^*\|_{Q,r} \le \Delta < \infty$ for all j, a.s.-P (with $r = 4d$, $d > 1$), so that $\|\boldsymbol{Z}_1^*\|_{Q,r} \le k\Delta\bar{\sigma}(\log\log R\bar{\sigma})^{1/2}/R^{1/2}$, a.a. R, a.s.-P.

Because $q_n = cn^{-\gamma}$, we have $\alpha_n^*(\tau) \le n\exp(-cn^{\varepsilon})$ for $\tau \ge n^{\gamma+\varepsilon}$, $\varepsilon > 0$. Then

$$\begin{aligned}
&\sum_{\tau=1}^{n-1}(1-\tau/n)\alpha_n^*(\tau)^{1/2-1/r} \\
&\quad \le \sum_{\tau=1}^{n^{\gamma+\varepsilon}-1}(1-\tau/n) + \sum_{\tau=n^{\gamma+\varepsilon}}^{n-1}(1-\tau/n)[n\exp(-cn^{\varepsilon})]^{1/2-1/r} \\
&\quad \le (n^{\gamma+\varepsilon}-1) + (n-n^{\gamma+\varepsilon})n^{1/2-1/r}\exp(-c(1/2-1/r)n^{\varepsilon}).
\end{aligned}$$

For all n sufficiently large $(n-n^{\gamma+\varepsilon})n^{1/2-1/r}\exp(-c(1/2-1/r)n^{\varepsilon}) \le 1$, so that

$$\sum_{\tau=1}^{n-1}(1-\tau/n)\alpha_n^*(\tau)^{1/2-1/r} \le n^{\gamma+\varepsilon}.$$

Collecting together the foregoing inequalities gives that for a.a. R, a.s.-P,

$$\mathrm{var}_Q\left[n^{-1/2}\sum_{t=1}^{n}(\boldsymbol{Z}_t^* - E_Q(\boldsymbol{Z}_t^*))\right] \le (k^2\Delta^2\bar{\sigma}^2(\log\log R\bar{\sigma})/R)(1 + 4(2^{1/2}+1)n^{\gamma+\varepsilon}).$$

By assumption, $(n^{\gamma+\varepsilon}/R)(\log\log R) \to 0$ as $T \to \infty$, ensuring that the variance on the left converges to zero as was to be shown. It now follows that, a.s.-P,

$$n^{-1/2} \sum_{t=1}^{n} (\, Z_t^* - E_Q(\, Z_t^*)) = o_Q(1).$$

But $n^{-1/2}\Sigma_{t=1}^{n} E_Q(\, Z_t^*) = n^{-1/2}\Sigma_{t=R}^{T} \nabla f_{t+\tau}^*(\, \hat{\beta}_t - \beta^*)$ and the desired result follows if

$$n^{-1/2} \sum_{t=R}^{T} \nabla f_{t+\tau}^*(\, \hat{\beta}_t - \beta^*) = o_P(1).$$

By West (1996, proof of (a), p. 1051),

$$n^{-1/2} \sum_{t=R}^{T} \nabla f_{t+\tau}^*(\, \hat{\beta}_t - \beta^*) = FB[\, n^{-1/2}\Sigma H(t)] + o_P(1).$$

Because $n^{-1/2}\Sigma H(t) = O_P(1)$ (a consequence of West's Lemma 4.1(a) and the Chebyshev inequality), the desired result follows immediately when $F = 0$, and the proof of (ii.a) is complete.

When $F \neq 0$ we use stronger conditions on n and R to reach the desired conclusion for (ii.b). Elementary inequalities give

$$\begin{aligned}
&|n^{-1/2} \sum_{t=R}^{T} \nabla f_{t+\tau}^*(\, \hat{\beta}_t - \beta^*)| \\
&\quad \leq n^{-1/2} \sum_{t=R}^{T} \sum_{j} |\nabla f_{j,\,t+\tau}^*|\, |\, \hat{\beta}_{jt} - \beta^*| \\
&\quad \leq n^{-1/2} \sum_{t=R}^{T} \sum_{j} |\nabla f_{j,\,t+\tau}^*| \sigma_j\, (\log\log R\sigma_j)^{1/2}/R^{1/2} \\
&\quad \leq \left(n^{-1} \sum_{t=R}^{T} \sum_{j} |\nabla f_{j,\,t+\tau}^*| \right) \bar{\sigma}\, (n/R)^{1/2}\, (\log\log R\bar{\sigma})^{1/2},
\end{aligned}$$

where $\bar{\sigma} \equiv \max_j \sigma_j$, and the second inequality follows by application of the law of the iterated logarithm. It follows from Assumption A.3 that

$$n^{-1} \sum_{t=R}^{T} \sum_{j} |\nabla f_{j,\,t+\tau}^*| = O_P(1)$$

by application of the law of large numbers for mixing processes (e.g., White (1984, Corollary 3.48)). The result now follows as $(n/R)\,(\log\log R) = o(1)$ trivially ensures $(n/R)^{1/2}\,(\log\log R\bar{\sigma})^{1/2} = o(1)$. *Q.E.D.*

PROOF OF PROPOSITION 2.4: Immediate from Theorem 2.3 and the Continuous Mapping Theorem. *Q.E.D.*

PROOF OF PROPOSITION 2.5: By definition

$$P[\bar{V}_l > n^{1/2}c] = P[\max_{k=1,\ldots,l} n^{1/2}\bar{f}_k > n^{1/2}c].$$

But

$$\begin{aligned}
P[\max_{k=1,\ldots,l} n^{1/2}\bar{f}_k > n^{1/2}c] &\geq P[\, n^{1/2}\bar{f}_1 > n^{1/2}c] \\
&= P[\, n^{1/2}(\, \bar{f}_1 - E(\, f_1^*))/\omega_{11} > n^{1/2}(c - E(\, f_1^*))/\omega_{11}],
\end{aligned}$$

where ω_{11} is the 1,1 element of Ω. Hence

$$P[\bar{V}_l > n^{1/2}c] \geq (1 - \Phi(z_n)) - \sup_z |P[Z_n \leq z] - \Phi(z)|,$$

where $z_n = n^{1/2}[c - E(f_1^*)]/\omega_{11} < 0$ and $Z_n = n^{1/2}(\bar{f}_1 - E(f_1^*))/\omega_{11}$. As in the proof of Proposition 2.1, we can choose T sufficiently large that $\Phi(z_n) < \varepsilon/2$ as well as $\sup_z |P[Z_n \leq z] - \Phi(z)| < \varepsilon/2$, so that for all T sufficiently large $P[\bar{V}_l > n^{1/2}c] \geq 1 - \varepsilon$. *Q.E.D.*

PROOF OF COROLLARY 2.6: Let $\mathbf{g}: \times_{k=0}^{l} U \to \Re^{l+1}$ be defined as $\mathbf{g}(h) = (g(h_0), g(h_1), \ldots, g(h_l))'$ for $h_0, \ldots, h_l \in U$. Let $\bar{h}$ be the $(l+1)m \times 1$ vector $\bar{h} = (\bar{h}_0, \bar{h}_1, \ldots, \bar{h}_l)'$ and let $E(h^*) = (E(h_0^*), E(h_1^*), \ldots, E(h_l^*))'$. A mean value expansion gives

$$\begin{aligned} n^{1/2}(\mathbf{g}(\bar{h}) - \mathbf{g}(E(h^*))) &= D\ddot{\mathbf{g}}\, n^{1/2}(\bar{h} - E(h^*)) \\ &= D\mathbf{g}^* n^{1/2}(\bar{h} - E(h^*)) + (D\ddot{\mathbf{g}} - D\mathbf{g}^*) n^{1/2}(\bar{h} - E(h^*)), \end{aligned}$$

where $D\ddot{\mathbf{g}}$ is the $(l+1) \times m(l+1)$ Jacobian matrix structured such that we obtain

$$n^{1/2}(g(\bar{h}_k) - g(E(h_k^*))) = D\ddot{g}\, n^{1/2}(\bar{h}_k - E(h_k^*))$$

in the kth row, where $D\ddot{g}$ is evaluated at a mean value lying between $\bar{h}_k$ and $E(h_k^*)$. $D\mathbf{g}^*$ is structured analogously but with elements evaluated at the appropriate components of $E(h^*)$.

It follows from Theorem 4.1 of West that $n^{1/2}(\bar{h} - E(h_k^*))$ is $O_P(1)$, while this and the assumed continuity of $D\mathbf{g}$ ensures that $D\ddot{\mathbf{g}} - D\mathbf{g}^* = o_P(1)$. Consequently,

$$n^{1/2}(\mathbf{g}(\bar{h}) - \mathbf{g}(E(h^*))) = D\mathbf{g}^* n^{1/2}(\bar{h} - E(h)) + o_P(1).$$

It follows from the asymptotic equivalence lemma (e.g. Lemma 4.7 of White (1984)) and (e.g.) Corollary 4.24 of White (1984) that

$$n^{1/2}(\mathbf{g}(\bar{h}) - \mathbf{g}(E(h^*))) \Rightarrow N(0, D\mathbf{g}^* \Omega D\mathbf{g}^{*\prime}),$$

given that $n^{1/2}(\bar{h} - E(h^*)) \Rightarrow N(0, \Omega)$ as ensured by West (1996, Theorem 4.1).

The results now follow by arguments identical to those for Proposition 2.2.

PROOF OF COROLLARY 2.7: Identical to that of Corollary 2.4, mutatis mutandis. *Q.E.D.*

PROOF OF COROLLARY 2.8: Identical to that of Proposition 2.5, mutatis mutandis. *Q.E.D.*

REFERENCES

ALTISSIMO, F., AND V. CORRADI (1996): "A LIL for m-Estimators and Applications to Hypothesis Testing with Nuisance Parameters," University of Pennsylvania Department of Economics Discussion Paper.

ANDREWS, D. W. K. (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.

BILLINGSLEY, P. (1968): *Convergence of Probability Measures*. New York: Wiley.

——— (1995): *Probability and Measure*, Third Edition. New York: Wiley.

BROCK, W., J. LAKONISHOK, AND B. LEBARON (1992): "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns," *Journal of Finance*, 47, 1731–1764.

CHATFIELD, C. (1995): "Model Uncertainty, Data Mining and Statistical Inference," *Journal of the Royal Statistical Society, Series A*, 158, 419–466.

CORRADI, V., N. R. SWANSON, AND C. OLIVETTI (1998): "Predictive Ability With Cointegrated Variables," Pennsylvania State University Department of Economics Discussion Paper.

COWLES, A. (1933): "Can Stock Market Forecasters Forecast?" *Econometrica*, 1, 309–324.

COX, D. R. (1982): "Statistical Significance Tests," *British Journal of Clinical Pharmacology*, 14, 325–331.

DIEBOLD, F., AND R. MARIANO (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–265.

DUFOUR, J.-M., E. GHYSELS, AND A. HALL (1994): "Generalized Predictive Tests and Structural Change Analysis in Econometrics," *International Economic Review*, 35, 199–229.

ENGLE, R., AND C. W. J. GRANGER (1987): "Co-integration and Error Correction: Representation, Estimation and Testing," *Econometrica*, 55, 251–276.

FITZENBERGER, B. (1997): "The Moving Blocks Bootstrap and Robust Inference for Linear Least Squares and Quantile Regressions," *Journal of Econometrics*, 82, 235–287.

GONÇALVES, S., AND H. WHITE (1999): "The Bootstrap of the Mean for Dependent Heterogeneous Arrays," UCSD Department of Economics Discussion Paper.

HALL, P., J. HOROWITZ, AND B.-Y. JING (1995): "On Blocking Rules for the Bootstrap with Dependent Data," *Biometrika*, 82, 561–574.

HAND, D. (1998): "Data Mining: Statistics and More?" *The American Statistician*, 52, 112–118.

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

HOMMEL, G. (1989): "A Comparison of Two Modified Bonferroni Procedures," *Biometrika*, 76, 625–625.

HOOVER, K. D., AND S. J. PEREZ (1998): "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search," UC Davis Department of Economics Discussion Paper.

JENSEN, D., AND P. COHEN (2000): "Multiple Comparisons in Induction Algorithms," *Machine Learning*, 38, 309–338.

KABAILA, P. (1995): "The Effect of Model Selection on Confidence Regions and Prediction Regions," *Econometric Theory*, 11, 537–549.

KLOEK, T. (1972): "Note on a Large-Sample Result in Specification Analysis," *Econometrica*, 43, 933–936.

KUENSCH, H. R. (1989): "The Jackknife and Bootstrap for General Stationary Observations," *Annals of Statistics*, 17, 1217–1241.

LEAMER, E. (1978): *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

——— (1983): "Let's Take the Con out of Econometrics," *American Economic Review*, 73, 31–43.

LIU, R. Y., AND K. SINGH (1992): "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence," in *Exploring the Limits of Bootstrap*, ed. by R. Lepage and L. Billiard. New York: Wiley, pp. 225–248.

LJUNG, L. (1987): *System Identification: Theory for the User*. Englewood Cliffs: Prentice-Hall.

LO, A., AND C. MACKINLEY (1990): "Data Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies*, 3, 431–468.

LOVELL, M. C. (1983): "Data Mining," *Review of Economics and Statistics*, 45, 1–12.

MAYER, T. (1980): "Economics as a Hard Science: Realistic Goal or Wishful Thinking?" *Economic Inquiry*, 18, 165–178.

MILLER, JR., R. G. (1981): *Simultaneous Statistical Inference*. New York: Springer-Verlag.

PAPARODITIS, E., AND D. POLITIS (2000): "Tapered Block Bootstrap," U.C. San Diego Dept. of Mathematics Discussion Paper.

POLITIS, D. (1999): Personal Communication.

POLITIS, D., AND J. ROMANO (1992): "A General Resampling Scheme for Triangular Arrays of α-mixing Random Variables with Application to the Problem of Spectral Density Estimation," *Annals of Statistics*, 20, 1985–2007.

——— (1994a): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

——— (1994b): "Limit Theorems for Weakly Dependent Hilbert Space Valued Random Variables with Application to the Stationary Bootstrap," *Statistica Sinica*, 4, 461–476.

——— (1994c): "Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions," *Annals of Statistics*, 22, 2031–2050.

PÖTSCHER, B. (1991): "Effects of Model Selection on Inference," *Econometric Theory*, 7, 163–185.

RAO, C. R. (1973): *Linear Statistical Inference and its Applications*. New York: Wiley.

RIVERS, D., AND Q. VUONG (1991): "Model Selection Tests for Nonlinear Dynamic Models," University of Southern California Department of Economics Discussion Paper.

ROY, S. N. (1953): "On a Heuristic Method of Test Construction and its Uses in Multivariate Analysis," *Annals of Mathematical Statistics*, 24, 220–239.

SAVIN, N. E. (1980): "The Bonferroni and the Scheffé Multiple Comparison Procedures," *Review of Economic Studies*, 48, 255–273.

SIN, C.-Y., AND H. WHITE (1996): "Information Criteria for Selecting Possibly Misspecified Parametric Models," *Journal of Econometrics*, 71, 207–225.

STONE, M. (1974): "Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion)," *Journal of the Royal Statistical Society*, Series B, 36, 111–147.

——— (1977): "Cross-Validation: A Review," *Mathematics of Operations Research and Statistics*, 9, 127–140.

SULLIVAN, R., A. TIMMERMANN, AND H. WHITE (1998): "Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns," UC San Diego Department of Economics Discussion Paper 98-31.

——— (1999): "Data Snooping, Technical Trading Rule Performance, and the Bootstrap," *Journal of Finance*, 54, 1647–1692.

SULLIVAN, R., AND H. WHITE (1999): "Finite Sample Properties of the Bootstrap Reality Check for Data-Snooping: A Monte Carlo Assessment," QRDA, LLC Technical Report, San Diego.

TALAGRAND, M. (1994): "Sharper Bounds for Gaussian and Empirical Processes," *Annals of Probability*, 22, 28–76.

WEST, K. (1994): "Asymptotic Inference About Predictive Ability," University of Wisconsin Department of Economics Discussion Paper.

——— (1996): "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084.

WESTFALL, P. H., AND S. S. YOUNG (1993): *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: Wiley.

WHITE, H. (1984): *Asymptotic Theory for Econometricians*. Orlando: Academic Press.