DATA-DRIVEN MODEL EVALUATION:
A TEST FOR REVEALED PERFORMANCE

JEFFREY S. RACINE AND CHRISTOPHER F. PARMETER

Abstract. When comparing two competing approximate models, the one having smallest ‘ex-
pected true error’ is closest to the data generating process (according to the specified loss function)
and is therefore to be preferred. In this paper we consider a data-driven method of testing whether
two competing approximate models, are equivalent in terms of their expected true error (i.e., their
expected performance on unseen data drawn from the same data generating process). The pro-
posed test is quite flexible with regards to the types of models that can be compared. Moreover,
our method overcomes two of the drawbacks associated with current approaches that are popular
and dominant among practitioners in time-series setting, namely, their reliance on only one split of
the data and the need to have a sufficiently large hold-out sample in order for the test to possess
adequate power. Finite-sample performance and an illustrative application is considered.

1. Introduction

Having estimated a parametric model in the course of applied data analysis, one ought naturally
test for model adequacy (i.e., for correct specification). When the parametric model is rejected by
the data, practitioners often turn to more flexible methods, for example, nonparametric models.
But there is no guarantee that the nonparametric model one has adopted will perform any better
than the parametric model that has been deemed inadequate, even though the nonparametric model
may indeed exhibit an apparent marked improvement in (within-sample) fit according to a variety
of metrics.\(^1\)

This is widely appreciated in the time-series literature where out-of-sample predictive perform-
ance is an overriding concern.\(^2\) By way of example, Medeiros, Teräsvirta & Rech (2006) consider

\(^{1}\)Alternatively, as White (2000, page 1097) discusses, resorting to extensive specification searches runs the risk that
the observed good performance of a model is not the result of superior fit but rather luck and he labels such practices
‘data snooping’.

\(^{2}\)Corradi & Swanson (2002, page 356) underscore the importance of this issue when they discuss “...whether simple
linear models (e.g. ARIMA models) provide forecasts which are (at least) as accurate as more sophisticated nonlinear
models. If this were shown to be the case, there would be no point in using nonlinear models for out-of-sample
prediction, even if the linear models could be shown to be incorrectly specified, say based on the application of
in-sample nonlinearity tests...” (our italics).
using autoregressive neural network models (AR-NN) to model financial time-series. However, having rejected linearity, fitted an AR-NN model, and conducted a rigorous post-mortem analysis of each model’s ability to predict stock returns, Medeiros et al. (2006, page 69) conclude that the “NN modelling strategy […] is not any better than a linear model with a constant composition of variables. A nonlinear model cannot therefore be expected to do better than a linear one.” See also Racine (2001) for an alternative example.

Indeed, there is no guarantee that a parametric model that passes a test for model adequacy will perform better than a nonparametric model as it is known that overspecified parametric models may perform worse than alternative specifications, including nonparametric ones. However, focusing instead on out-of-sample predictive ability may provide the applied researcher with a potential avenue for discriminating among such models. Though a literature that advocates in-sample predictive evaluation in time-series settings has recently emerged (see Inoue & Kilian 2004), this runs against the tide of a large body of literature that convincingly argues for the use of sample-splitting mechanisms whereby one splits the full sample into two sub-samples, then uses one sub-sample for estimation and the other to guide predictive evaluation; see Corradi & Swanson (2007) and the references therein.

Out-of-sample predictive performance has become the metric of choice for time-series researchers; see Diebold & Mariano (1995), West (1996), West & McCracken (1998) and McCracken (2000), among others. However, as will be demonstrated, there remains scope for improving these methods. In this paper we show how, through judicious use of an appropriate resampling mechanism, the proposed approach provides an appealing alternative to popular time-series tests for predictive accuracy by overcoming what we regard as limitations associated with such tests, namely, the reliance on a single split of the data and the subsequent size of the hold-out sample.

In this paper we take the view that fitted statistical models are approximations, a perspective that differs from that of consistent model selection which posits a finite-dimensional ‘true model’. That is, in this paper we are not interested in tests that hypothesize one model being the ‘true model’. Rather, our goal is instead to test whether one approximate model’s expected performance is better than another on data drawn from the same DGP according to a pre-specified loss function such as square or absolute error loss. The loss function is provided by the user hence the method suggested herein is quite general.4

Our approach is firmly embedded in the statistics literature dealing with apparent versus true error estimation; for a detailed overview of ‘apparent’, ‘true’, and ‘excess’ error, we direct the reader to Efron (1982, Chapter 7). In effect, within-sample measures of fit gauge ‘apparent error’ which

---

3See Hansen (2005, pgs.62-63) for an eloquent discussion of this issue.

4This allows us to address how much more accurate one method is compared to another on average with respect to the chosen loss function. Indeed, this is in direct agreement with Goodwin (2007, page 334): “[…] when comparing the accuracy of forecasting methods […] The interesting questions are, how much more accurate is method A than method B, and is this difference of practical significance?” Our approach will allow a researcher to tackle both of these questions in a simple and easily implementable framework, though we take a broader view by considering ‘forecasting’ in time-series ones.
will be more optimistic than ‘true error’, sometimes strikingly so, since a model is selected to fit the
data best. For a given loss function, $\ell(u)$, one might compute the expected loss, $n^{-1} \sum_{i=1}^{n} \ell(\hat{u}_i)$, which provides an estimate of the apparent error arising from the modelling process. But all such within-sample measures are fallible which is why they cannot be recommended as guides for model selection; for example, $R^2$ does not take into account model complexity, adjusted $R^2$ measures are not defined for many nonparametric methods, whereas information-based measures such as AIC can be biased if the sequence of competing (parametric) models is nonnested; see Ye (1998) and Shen & Ye (2002).

The approach we advocate involves constructing the distribution function of a model’s true error and testing whether the expected true error is statistically smaller for one model than another. This will be accomplished by leveraging repeated splits of the data rather than just one as is commonly done and by computing the estimated loss for the hold-out data for each split. At the end of the day we will conclude that one model has statistically smaller estimated true error than another and therefore is expected to be closer to the true DGP hence is preferred though both models are, at best, approximations.

The basic idea is, of course, not new and involves splitting the data into two independent samples of size $n_1$ and $n_2$, fitting models on the first $n_1$ observations, then evaluating the models on the remaining $n_2 = n - n_1$ observations using, by way of example, average square prediction error (ASPE) (we know the outcomes for the evaluation data, hence this delivers an estimate of true error).\(^5\) However, one might mistakenly favor one model when the estimate of true error is lower but this in fact simply reflects a particular division of the data into two independent subsets which may not be representative of the DGP, i.e., this can be overly influenced by which data points end up in each of the two samples. To overcome this limitation, one might consider repeating this process a large number of times, say $S = 10,000$ times, each time refitting the models on the ‘training’ data (the $n_1$ observations) and evaluating on the independent ‘evaluation’ data (the $n_2 = n - n_1$ hold-out observations). This repeated sample-splitting experiment will thereby produce two vectors of length $S$ which represent draws from the distribution of actual ASPEs for each model.\(^6\) These two vectors of draws can then be used to discriminate between the two models.\(^7\) For what follows we consider a simple (paired) test of differences in means for the two distributions, but also consider simple graphical tools that will help reveal stochastic dominance relationships, if present. Given that the test is a test of whether the data at hand reveals that the predictive performance of one

\(^{5}\)Readers familiar with Diebold & Mariano’s (1995) test for predictive accuracy will immediately recognize this strategy.

\(^{6}\)Clearly, for (strictly) stationary dependent processes we cannot use sample splitting directly, however, we can use resampling methods that are appropriate for such processes. When each resample is the outcome of an appropriate bootstrap methodology it will mimic dependence present in the original series and can itself be split; see Politis & Romano (1992) by way of example.

\(^{7}\)For instance, we might perform a test of the hypothesis that the mean ASPE (‘expected true error’) for the $S = 10,000$ splits is equal (less than or) for two models against a one-sided alternative (greater than) in order to maximize power. Or, one could test for stochastic dominance of one distribution over the other.
econometric model is statistically different from that of another, we dub the test the ‘RP’ test to denote ‘revealed performance’.

The statistics literature on cross-validated estimation of excess error is a well-studied field (‘expected excess error’ is the expected amount by which the true error exceeds the apparent error). However, this literature deals with model specification within a class of models (i.e., which predictor variables should be used, whether or not to conduct logarithmic transformations on the dependent variable and so forth) and proceeds by minimizing excess error. Our purpose here is substantively different, and is perhaps most closely related to the literature on non-nested model testing (see Davidson & MacKinnon 2002). Unlike this literature, however, we are asking an inherently different question that is not the subject of interest in the non-nested literature, namely, whether the expected true error associated with one model differs significantly from that for another model, whether nested or not.

The rest of this paper proceeds as follows. Section 2 outlines our basic approach and defines the framework for our proposed test. Section 3 conducts several simulation exercises to assess the performance of the proposed approach when the DGP is known. Section 4 presents several empirical examples. Section 5 presents some concluding remarks.

2. Methodology

The method we describe here is closest in spirit to the original application of cross-validation in which the data set is randomly divided into two halves, the first of which is used for model fitting and the second for cross-validation where the regression model fitted to the first half of the data is used to predict the second half. The more common modern variant in which one leaves out one data point at a time, fits the model to the remaining points, then takes the average of the prediction errors (each point being left out once) yielding a cross-validated measure of true error, has been widely studied, and we direct the interested reader to Stone (1974), Geisser (1975), and Wahba & Wold (1975) for detailed descriptions of this method. It is noteworthy that White (2000, page 1108) argues that cross-validation represents a “more sophisticated use of ‘hold-out’ data” and indicates that this “is a fruitful area for further research”. Our approach indeed supports this claim as we demonstrate that cross-validation can lead to dramatic power improvements over existing, single-split techniques commonly used in the applied times-series literature.

In our regression problem the data consists of pairs \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\) where \(X_i\) is a \(1 \times p\) vector of predictor variables and \(Y_i\) is a real-valued response variable. For our purposes we presume that \(Z_i = (X_i, Y_i)\) represent random draws from a (strictly) stationary ergodic process with unknown distribution function \(F\) defined on \(\mathcal{H} = \mathbb{R}^{p+1}\),

\[
Z_1, Z_2, \ldots, Z_{n_1} \sim F.
\]

We observe \(Z_1 = z_1, Z_2 = z_2, \ldots, Z_{n_1} = z_{n_1}\) and for what follows we let \(Z^{n_1} = (Z_1, Z_2, \ldots, Z_{n_1})\) and \(z^{n_1} = (z_1, z_2, \ldots, z_{n_1})\). Having observed \(Z^{n_1} = z^{n_1}\) we fit a regression model which will be used
to predict some ‘future’ values of the response variable, which we denote

\[ \hat{g}_{z^n_1}(x^{n_2}) , \]

where the superscript \( n_2 \) indicates a new set of observations, \( z^{n_2} = (z_{n_1+1}, z_{n_1+2}, \ldots, z_n) \), which are distinct from \( z^{n_1} = (z_1, z_2, \ldots, z_n) \) where \( n_2 = n - n_1 \). By way of example, simple linear regression would provide \( \hat{g}_{z^n_1}(x^{n_2}) = x^{n_2} \hat{\beta}_{n_1} \) where \( \hat{\beta}_{n_1} = (x^{n_1}T x^{n_1})^{-1}x^{n_1}T y^{n_1} \), \( T \) denotes transpose, and \( y^{n_1} = (y_1, y_2, \ldots, y_{n_1}). \)

We are interested in estimating a quantity known as ‘expected true error’ (Efron 1982, page 51).\(^8\) Following Efron (1982), we first define the ‘true error’ to be

\[ E_{n_2,F}[\ell(Y^{n_2} - \hat{g}_{Z^n_1}(X^{n_2}))] , \]

where \( \ell(\cdot) \) denotes a loss function specified by the researcher satisfying regularity conditions given in Assumption 2.2 below. The notation \( E_{n_2,F} \) indicates expectation over the new point(s)

\[ Z_{n_1+1}, Z_{n_1+2}, \ldots, Z_n \sim F, \]

independent of \( Z_1, Z_2, \ldots, Z_{n_1} \), the variables which determine \( \hat{g}_{Z^n_1}(\cdot) \) in (3) (we refer to \( Z^{n_1} \) as the ‘training set’, terminology borrowed from the literature on statistical discriminant analysis). Next, we define ‘expected true error,’

\[ E (E_{n_2,F}[\ell(\cdot)]) , \]

the expectation over all potential regression surfaces \( \hat{g}_{Z^n_1}(\cdot) \), for the selected loss function \( \ell(\cdot) \). When comparing two approximate models, the model possessing lower ‘expected true error’ will lie closest to the true DGP given the loss function \( \ell(\cdot) \) and would therefore be preferred in applied settings.

A realization of ‘true error’ (3) based upon the observed \( z^{n_2} = (z_{n_1+1}, z_{n_1+2}, \ldots, z_n) \), is given by

\[ \frac{1}{n_2} \sum_{i=n_1+1}^{n_2} \ell(y_i - \hat{g}_{z^n_1}(x_i)) , \]

an average prediction error which, for square error loss, we denote ASPE (‘average square prediction error’).

Were we given \( S \) such splits of the data, we could then construct the empirical distribution function (EDF) of (6). Given two competing models and each model’s respective EDF of realized true error, we can then use the respective EDFs to determine whether one model has statistically significantly lower expected true error than another. Note that here we have transformed the problem into a (paired) two-sample problem where we wish to test for equivalence of expected true error defined in (5) based upon two vectors of realizations of true error defined in (6).\(^9\) Thus, the procedure we consider is strictly data-driven and nonparametric in nature.

---

\(^8\)Efron (1982, page 51) considers estimation of ‘expected excess error’, while we instead consider estimation of ‘expected true error’.

\(^9\)The pairing will be seen to arise from potential correlation between model predictions among competing models.
2.1. The Empirical Distribution of True Error. Suppose we arbitrarily denote one approximate model ‘Model A’ and the other ‘Model B’. For the sake of concreteness, let us presume one was interested in comparing, say, a nonparametric kernel regression model (‘Model A’) to a parametric regression model (‘Model B’). In a time-series context, it might appear that there is only one possible split of the data, \( \{ z_i \}_{i=1}^n \) and \( \{ z_i \}_{i=t_1}^n \), and this one split underlies many tests for predictive accuracy (or forecast equality) such as Diebold & Mariano’s (1995) test. But, there is nothing to preclude conducting repeated resampling with time-series data; we just need to use an appropriate resampling methodology!

In the context of time-series prediction (‘forecasting’), resampling methods are widely used. For instance, Corradi & Swanson (2002) propose a consistent test for nonlinear predictive accuracy for nested models where interest lies in testing whether the null model can outperform the nesting alternative model based upon “real-time forecasts” (i.e., one-step recursive forecasts for period \( t+1, t+2 \) and so on) and one split of the data. Corradi & Swanson (2004) examine finite-sample properties of their (2002) test where critical values are based on application of the block bootstrap. Corradi & Swanson (2004, Tables 1 and 2) employ manually set fixed block lengths and they note that the value of the test statistic(s) under consideration and the resulting power properties vary dramatically as the block length is changed. There is no reason to require the user to set block lengths manually, however, just as there is no reason to require users to manually specify bandwidths for kernel estimation; automatic methods possessing desirable statistical properties have recently been proposed and are available for use.

In what follows, we shall exploit recent advances in time-series resampling methodology, and use geometric (‘stationary’) block bootstrapping to generate a bootstrap replication of the series of size \( n \) which then can itself be split into two samples of size \( n_1 \) and \( n_2 \) thereby preserving the dependence structure present in the full sample. That is, in a (univariate) time-series setting, we deploy a time-series bootstrap based on automatic block length selection where we resample from, say, \( z = \{ y_i \}_{i=1}^n \). By way of illustration, we elect to use the method of Politis & Romano (1992) for which Politis & White (2004) recently proposed a fully automatic method for choosing the block length that has excellent finite-sample properties.\(^{10}\) This bootstrap preserves the underlying dependence structure by resampling the data in blocks of random length, where the lengths are derived from a geometric distribution, hence the name. See both Davison & Hinkley (1997, pp.401-408) and Lahiri (2003, sections 2.7.2 and 3.3) for more on the theoretical underpinnings underlying the geometric bootstrap.\(^{11}\) In a time-series setting, where the data represent draws from a (strictly) stationary ergodic time-series process, we proceed as follows:

\(^{10}\)See Patton, Politis & White (2008) for a correction to several of the results in Politis & White (2004).

\(^{11}\)Our choice of the stationary block bootstrap is for expositional purposes. In practice we recommend that the user employ a bootstrap appropriate for the type of dependence apparent in the data. For example, additional types of bootstraps are the Markov conditional bootstrap Horowitz (2004), the circular bootstrap Politis & Romano (1992) and the sieve bootstrap Bühlmann (1997); see Lahiri (2003) for an up-to-date and detailed coverage of available block resampling schemes. One can easily implement a variety of block bootstrap procedures by using the tsboot() command available in the boot package (Canty & Ripley 2008) in R (R Development Core Team 2008).
(i) Apply the stationary bootstrap to resample from \( z = \{y_i\}_{i=1}^n \) and call these \( z^* = \{y_{i}^*\}_{i=1}^n \).  
(ii) Let the first \( n_1 \) of the resampled observations form a training sample, \( z^{n_1}_* = \{y_{i}^*\}_{i=1}^{n_1} \) and the remaining \( n_2 = n - n_1 \) observations form an evaluation sample, i.e., \( z^{n_2}_* = \{y_{i}^*_i\}_{i=n_1+1}^n \).  
(iii) Holding the degree of smoothing\(^{12}\) of the nonparametric Model A and the functional form of the parametric Model B fixed (i.e., at that for the full sample), fit each model on the training observations \((z^{n_1}_*)\) and then generate predictions for the \( n_2 \) evaluation observations.  
(iv) Compute the ASPE of each model which we denote \( \text{ASPE}^A = n_2^{-1} \sum_{i=n_1+1}^n (y_i^* - \hat{g}^A_{z^{n_1}_*}(y_i^*{,\ldots}))^2 \) and \( \text{ASPE}^B = n_2^{-1} \sum_{i=n_1+1}^n (y_i^* - \hat{g}^B_{z^{n_1}_*}(y_i^*{,\ldots}))^2 \).  
(v) Repeat this a large number of times, say, \( S = 10,000 \), yielding \( S \) draws, \{\text{ASPE}^A_s, \text{ASPE}^B_s\}_{s=1}^S\}.  

We refer to the respective empirical distributions as \( \hat{F}^A_S \) and \( \hat{F}^B_S \) where each places mass \( 1/S \) at \( \text{ASPE}^A_s \) and \( \text{ASPE}^B_s \).

We can now proceed to use \( \hat{F}^A_S \) and \( \hat{F}^B_S \) to discriminate between models. At this stage we point out that the choice \( S = 1 \) is typically used to discriminate among time-series models, i.e., one split only of the data is the norm. By way of example, the popular time-series test for predictive accuracy of Diebold & Mariano (1995) is based on only one split, hence attention has shifted to determining how large \( n_2 \) need be (e.g., see Ashley 2003). One might, however, be worried about basing inference on only one split, mistakenly favoring one model over another when this simply reflects a particular division of the data into two independent subsets that may not be representative of the underlying DGP, i.e., this can be overly influenced by which data points end up in each of the two samples.

However, by instead basing inference on \( S \gg 1 \) (i.e., averaging over a large number of such splits), we can control for mistakes arising from divisions of the data that are not representative of the DGP. In fact, it will be seen that the power of our test increases with \( S \), which is obvious in hindsight. Furthermore, by averaging over a large number of splits, we can base inference on much smaller evaluation sample sizes (i.e., \( n_2 \)) thereby taking maximal advantage of the estimation data which would be particularly advantageous in time-series settings. Ashley (2003) clearly illustrates this dilemma in the \( S = 1 \) time-series context by highlighting that one may need \( n_2 \) to be quite large in order for such tests to have power; the dilemma is that for a time-series of fixed length \( n \),

\(^{12}\) A ‘scaling factor’ refers to the unknown constant \( c \) in the formula for the optimal bandwidth, \( h_{opt} = cn^{−1/(4+p)} \). Cross-validation can be thought of as a method for estimating the unknown constant \( c \), where \( c \) is independent of the sample size \( n \). This constant can then be rescaled for samples of differing size drawn from the same DGP thereby ensuring that the same degree of smoothing is applied to the full sample and the subsample (see Racine 1993). The rationale for so doing is as follows. Think of estimating a univariate density function where the data represent independent draws from the \( N(0,1) \) distribution. The optimal bandwidth in this case is known to be \( h_{opt} = 1.059n^{−1/5} \). If \( n = 200 \) then \( h_{opt} = 0.3670 \) while if \( n = 100 \) then \( h_{opt} = 0.4215 \). Cross-validation delivers an estimate of \( h_{opt} \) for a sample of size \( n \), i.e. \( \hat{h} = cn^{−1/5} \), while it can be shown that \( (\hat{h} - h_{opt})/h_{opt} \to 1 \) asymptotically (see Stone 1984). If you don’t rescale the cross-validated bandwidth for subsamples of size \( n_1 < n \) (i.e., adjust \( \hat{h} \) when \( n \) falls to \( n_1 \)) then you are in fact doing a different amount of smoothing on subsamples of size \( n_1 < n \) (i.e., \( h = 0.3670 \) will undersmooth when \( n_1 < 200 \), so the estimate based on \( h = 0.3670 \) and \( n_1 < 200 \) will be overly variable). But, by using \( c \) corresponding to the cross-validated bandwidth for the full sample, \( \hat{h} \), we can ensure that the same degree of smoothing is applied to the subsamples of size \( n_1 < n \) and the full sample of size \( n \). This keeps the baseline nonparametric model fixed at that for the full sample, in the same way that we hold the functional form of the parametric model fixed at that for the full sample.
increasing $n_2 = n - n_1$ means that the models are less efficiently estimated since they are based on fewer observations. Our approach will be seen to effectively overcome this limitation.

2.2. Validity of the Bootstrap. We now consider conditions that justify our use of the bootstrap for obtaining valid approximations to the unknown loss distributions for two competing approximate models, which we denote $F^A$ and $F^B$, respectively. For what follows we leverage Lemma A.3 and Theorem 2.3 in White (2000) to establish consistency of our bootstrap approach. The conditions required (for competing parametric models) involve assumptions on the data, parameter estimates, behavior of the bootstrap, properties of the loss function, and some additional regularity conditions. For concreteness, we focus our theoretical arguments on the case where the competing models are both of parametric form (but potentially nonlinear). Extensions to semiparametric and nonparametric estimators are easily handled with (minor) modifications to the requisite assumptions listed below. As the conditions we impose involve theoretical arguments described in three separate papers, we shall outline each set of assumptions in turn and cite sources accordingly.

We begin with an assumption given in Politis & Romano (1994) that is required to demonstrate consistency of the stationary bootstrap under a range of settings. For what follows we have $\ell_s = \ell(u_s)$ where the index of $s$ follows from the context. $\beta^*$ represents an unknown parameter vector.

**Assumption 2.1.**

(i) Let $q$ denote the probability of the geometric distribution used for the stationary bootstrap ($q$ is equivalent to one over the block length). Assume that $q \to 0$ and that $nq \to \infty$ as $n \to \infty$.

(ii) Let $Z_1, Z_2, \ldots,$ be a strictly stationary process with $E|Z_i|^{6+\eta} < \infty$ for some $\eta > 0$.

(iii) Let $\{Z_n\}$ be $\alpha$-mixing with $\alpha_Z(k) = O(k^{-r})$ for some $r > 3(6+\eta)/\eta$.

Assumption 2.1(i) establishes the rate at which the block length in the stationary bootstrap can grow. Assumptions 2.1(ii) and 2.1(iii) are required to ensure the data behaves in a manner consistent with the theoretical arguments of both Politis & Romano (1994) and White (2000). Of course, in cross-section settings these conditions are automatically satisfied. Note that Assumption 2.1 is the same as that used by Politis & Romano (1994) for much of their theoretical work in this area (see Theorems 2-4, Politis & Romano 1994).

Additionally, we require assumptions 1-4 in West (1996). We restate these and label them jointly as Assumption 2.2.

**Assumption 2.2.**

(i) Let the loss function be measurable and second order continuously differentiable at $\beta^* \equiv \text{plim} \hat{\beta}$ where $\hat{\beta}$ is defined below. Additionally, the matrix of second order derivatives is dominated by $m_n$ where $E|m_n| < D$ for $D < \infty$.

(ii) Let the parameter estimates be linear combinations of orthogonality conditions used to identify the response. More formally we have that (for the parametric regression model $y_j = X_j^T \beta + \epsilon_j$, $j = 1, \ldots, n$) $\hat{\beta} - \beta^* = B(n)H(n)$ where $B(n)$ is $(k \times q)$ and $H(n)$ is $(q \times 1)$
We mention that Theorem 2.3 follows under the condition that the objective function used for invoke his Theorem 2.3 (which follows immediately from Lemma A.3) directly to achieve the result. Given that Assumptions 2.1, 2.2 and 2.3 are identical to those in White (2000), we can

Proof. Under Assumptions 2.1, 2.2 and 2.3, the stationary bootstrap estimates of the distributional laws \( F^A \) and \( F^B \), denoted \( \hat{F}^A_S \) and \( \hat{F}^B_S \), converge in probability to \( F^A \) and \( F^B \).

Assumption 2.2(i) ensures that the loss function is well behaved in a neighborhood of a specified parameter value. Essentially, the loss function evaluated at the prediction errors needs to be bounded and satisfy certain moment conditions in order to use White’s (2000) bootstrap theory. As noted by West (1996), Assumption 2.2(ii) does not assume that either \( \varepsilon \) or \( X\varepsilon \) is serially uncorrelated. Assumption 2.2(iii) is used to pin down the behavior of the mean of the losses for a particular model by suitable application of a law of large numbers applicable to mixing processes (see Section 3.4, White 2001). Assumption 2.2(iv) is needed to invoke asymptotic arguments related to either the estimation sample size (\( n_1 \)) or the evaluation sample size (\( n_2 \)).

In order to invoke either Lemma A.3 or Theorem 2.3 of White (2000) we need two additional conditions. In White (2000) these are Assumption A.5 and Assumption C. We state them here for convenience.

Assumption 2.3.

(i) Let the spectral density of \( [\ell_i - E\ell_i]'h'_jB' \) where \( \ell_i = \ell(y_i - \hat{y}_i) \), at frequency zero, multiplied by a scale factor, be positive definite.

(ii) Let the parameter estimates (\( \hat{\beta} \)) obey a law of iterated logarithm.

Assumption 2.3(ii) is required to bound a pseudo-studentized term involving \( \hat{\beta} \) in White’s (2000) Theorem 2.3.

These conditions are sufficient to establish that the bootstrap distribution of any (parametric) candidate model’s evaluation sample loss is consistent for the distribution of expected true error, which we now state formally.

Theorem 2.1. Under Assumptions 2.1, 2.2 and 2.3, the stationary bootstrap estimates of the distributional laws \( F^A \) and \( F^B \), denoted \( \hat{F}^A_S \) and \( \hat{F}^B_S \), converge in probability to \( F^A \) and \( F^B \).

Proof. Given that Assumptions 2.1, 2.2 and 2.3 are identical to those in White (2000), we can invoke his Theorem 2.3 (which follows immediately from Lemma A.3) directly to achieve the result. We mention that Theorem 2.3 follows under the condition that the objective function used for
estimation and loss function are equivalent. This is not a deterrent as Corradi & Swanson (2007, Proposition 1, page 77) generalize the results of White (2000) for the case where the loss function differs from the objective function used to obtain the parameter estimates. In our work, and certainly for most applications, they are identical. □

Theorem 2.1 allows us to therefore implement a variety of two-sample tests to assess revealed performance (pairwise) across a set of candidate models. Of course, we need to address the possibility that the realizations defined in (6) are correlated for Model A and Model B (i.e., that there may exist pairwise correlation of the realizations underlying \( \hat{F}^A_S \) and \( \hat{F}^B_S \), thus one’s testing procedure must accommodate potential pairwise correlation. But such tests are widely available to sanctioners, two popular examples being the paired \( t \)-test and the paired Mann-Whitney-Wilcoxon tests; see also Mehta & Gurland (1969) for an alternative to the paired \( t \)-test.

When considering two-sample tests of it would be convenient to be able to inform the users choice of \( n_2 \) and guide their choice of tests. It is known that a sufficient sample size for the sampling distribution of the paired Mann-Whitney-Wilcoxon test to be approximated by the normal curve is \( n_2 \geq 10 \) regardless of the distributions \( F^A \) and \( F^B \), while the matched-pairs \( t \)-test is strictly speaking valid only when \( F^A \) and \( F^B \) are normally distributed though it is known that the \( t \)-test is quite robust to departures from normality. For the simulations we consider there is no qualitative difference between rejection frequencies based upon the paired Mann-Whitney-Wilcoxon and the \( t \)-test, so for what follows we consider the popular paired \( t \)-test for equality in means to assess whether one distribution dominates the other (i.e., test equality (less than or equal) of means against the alternative hypothesis that the true difference in means is greater than zero). Full results for both tests beyond those reported here are available upon request.

Formally, we state the null and alternative as

\[
H_0 : E \left( E_{n_2,F^A} [\ell(\cdot)] \right) - E \left( E_{n_2,F^B} [\ell(\cdot)] \right) \leq 0
\]

and

\[
H_A : E \left( E_{n_2,F^A} [\ell(\cdot)] \right) - E \left( E_{n_2,F^B} [\ell(\cdot)] \right) > 0
\]

which arises directly from our notation in (5).

This is, of course, not the only test available to practitioners. One might prefer, say, the Mann-Whitney-Wilcoxon test (i.e., test equality (less than or equal) of locations against the alternative hypothesis that the true location shift is greater than zero); see Bauer (1972). Or perhaps one might undertake a more sophisticated test for, say, first-order stochastic dominance (e.g., Davidson & Duclos 2000). We argue that this is not needed in the present context and a simple test for equality of locations and summary plots of the vectors of ASPEs is more than sufficient for our purposes. Indeed, one of the appealing aspects of the proposed approach lies in its simplicity, though nothing would preclude the practitioner from considering additional tests in this setting as they will all be based on \( \hat{F}^A_S \) and \( \hat{F}^B_S \) which have been pre-computed and are consistent given Theorem 2.1.
We now proceed to some Monte Carlo simulations designed to assess the finite-sample performance of the proposed method.

3. Monte Carlo Simulations

3.1. Finite-Sample Performance: Time-Series Data. The time-series literature dealing with predictive accuracy and forecasting is quite vast, and we make no claims at surveying this literature here.\(^\text{13}\) Early work on forecast model comparison by Ashley, Granger & Schmalensee (1980) and Granger & Newbold (1986) generated broad interest in this topic. However, only recently have formal tests that directly relate to forecast accuracy and predictive ability surfaced. Most notably the available tests include Diebold & Mariano (1995) (the ‘DM’ test) and the size corrected counterpart of Harvey, Leybourne & Newbold (1997) (the ‘MDM’ test) along with those proposed by Swanson & White (1997), Ashley (1998), Harvey, Leybourne & Newbold (1998), West & McCracken (1998), Harvey & Newbold (2000), Corradi & Swanson (2002), van Dijk & Franses (2003), Hyndman & Koehler (2006) and Clark & West (2007), among others. Given the popularity of Diebold & Mariano’s (1995) test\(^\text{14}\) we perform a simple Monte Carlo simulation similar to that presented in Section 3 but with stationary time-series models as opposed to cross-section ones.

We generate data from an AR(2) model given by

\[
y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \varepsilon_t,
\]

where \(\rho_1 = 0.9\) for all simulations, \(\rho_2\) varies from 0 (an AR(1) model) to -0.8 in increments of 0.4 and \(\varepsilon_t\) is \(N(0,1)\). For all simulations we conduct \(M = 1,000\) Monte Carlo replications using \(S = 10,000\) sample splits for our revealed performance approach. We use sample sizes of \(n = 200\) and \(n = 400\) holding out the last \(n_2 = 5, 10, 25,\) or 50 observations of each resample with which to forecast.\(^\text{15}\) When \(\rho_2 = 0\) we can determine the extent to which our test predicts (worse than or) equivalent accuracy of the forecasts, while when \(\rho_2 \neq 0\) we can assess how often our method determines that an AR(2) model predicts better than an AR(1) when indeed it should. We also compare the AR(1) to MA(1) and MA(2) specifications by way of comparison. We compare our results with the DM and MDM test, noting that the DM test has a tendency to over-reject hence our inclusion of the size-corrected MDM results. We report empirical rejection frequencies at the 5% level in Tables 1 and 2.

First, a comparison of the DM and MDM results indicate that indeed the DM test suffers from rather substantial upwards size distortions. We direct the interested reader to Harvey et al. (1998)\(^\text{16}\)

\(^{13}\)See the review by De Gooijer & Hyndman (2006) for a thorough and up-to-date survey and bibliography on the subject.

\(^{14}\)This paper was named one of the 10 best papers in the 20\textsuperscript{th} anniversary edition of the *Journal of Business & Economic Statistics*, while a Social Science Citations Index search, conducted on October 5, 2011, yielded 811 citations.

\(^{15}\)For each Monte Carlo simulation, the initial data generated is passed through the automatic block length selection mechanism of Politis & White (2004) to determine the optimal block length. This block length is then used for each of the \(S\) splits of the data.

\(^{16}\)See also Harvey & Newbold (2000), Meade (2002) and van Dijk & Franses (2003).
Table 1. Each entry represents rejection frequencies for a one-sided test at the 5% level of the hypothesis that the AR(1) model has predictive accuracy better than or equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy.

\[ n = 200, \ n_2 = 10, \ S = 10,000. \]

Table 2. Each entry represents rejection frequencies for a one-sided test at the 5% level of the hypothesis that the AR(1) model has predictive accuracy better than or equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy.

who report on the formal size and power properties of the DM and MDM tests. Comparing Table 1 \( (n = 200) \) with 2 \( (n = 400) \) we see that the upwards size distortion (i.e., \( \rho_2 = 0 \), test AR(1) versus AR(2)) falls as \( n \) increases. In fact, size distortions present in the DM test are so large as to all but dismiss it from contention, at least for the simulation at hand.\(^{17}\) The size corrected MDM test performs well, at least from the perspective of correcting for size distortions. All three approaches increase in power as \( n_2 \) increases as expected, however, the RP test rejection frequencies approach one more quickly as \( |\rho_2| \) increases suggesting that smaller holdout samples are required in order to discriminate between models in terms of their predictive accuracy. This feature of our approach overcomes one known drawback of the MDM and related tests, namely, the need to have a sufficiently large hold-out sample in order for the test to have power. Lastly, comparing the RP and MDM test one will see immediately that as \( |\rho_2| \) increases our rejection frequencies approach one at a faster rate for the RP than for the MDM test, indicating that this approach is successful at detecting gains in predictive accuracy outside of an iid setting.

We note that the ability to choose one’s loss function when using our approach may be appealing to practitioners. For instance, if interest lies in penalizing more heavily over or underprediction, the use of asymmetric loss functions may be of interest (LINEX for example, Chang & Hung 2007). See Efron (1983, 1986) for more on issues related to prediction rules and apparent error in relation to cross-validation and bootstrapping.

4. Empirical Illustration

4.1. Application to Interest Rate Data. In this section we consider modelling the U.S. federal funds interest rate. The data is taken from Davidson & MacKinnon (2004, page 601). This data is a time-series of monthly observations on the interest rate from January 1955 to December 2001 (a

\(^{17}\)To see this note that when \( \rho_2 = 0 \) (i.e., the true model is an AR(1)) the DM test will select an overspecified AR(2) model a high percentage of the time, unlike the size-corrected MDM test. Furthermore, as \( n_2 \) increases the rejection frequency falls when \( \rho_2 = 0 \) while that for the MDM test remains stable.
total of \( n = 564 \) observations). Figures 1 and 2 display the raw and first differenced interest rate series as well as auto-covariance and partial autocorrelation functions for 20 lags.

Before proceeding we note that we reject stationarity at the 5% level for the raw series, hence we first difference our data at which point we no longer reject stationarity. Next, we note that the large spike of our differenced data for the first autocorrelation (Figure 2, lower right panel) suggests that a MA(1) process may be present. However, the presence of positive and significant autocorrelations past lag 1 and the regular pattern in the distant autocorrelations suggests that a more complex DGP may be present. Also, the partial autocorrelations (Figure 2, lower left panel), which have a positive and significant spike for lag 1 and a negative and significant spike for lag 2 would rule out the use of an AR(1) model but could be consistent with an AR(2) model.\footnote{The positive and significant partial autocorrelations at lag 8, 13, and 26 are difficult to interpret.}

Fitting the best ARIMA process to the first differenced data suggests that an MA(1) process is appropriate.\footnote{This was done using the the entire dataset with the auto.arima() function in the forecast package (Hyndman 2008) in R (R Development Core Team 2008).} For our empirical exercise we use a holdout sample of 12 observations, which corresponds to one year for the undifferenced data, for our proposed test. The automatic block length selection of Politis & White (2004) suggests we use a random block length of 4 when resampling.

Before providing our graphical evidence we compare our 12 1-step ahead forecasts from an AR(1), AR(2), MA(1), ARMA(1,1), ARMA(2,1) and an ARMA(2,2) model with the MA(1) being our baseline model. The \( P \)-values of our test, along with those of the DM and the MDM test appear in Table 3. While all three tests suggest that we cannot reject the null of equal forecast accuracy for the AR(1) and ARMA(1,1) against the MA(1) at all conventional significance levels, both the DM and MDM tests reject the null hypothesis of equal forecast accuracy for the ARMA(2,1) and ARMA(2,2) at the 1% level of significance against the MA(1) and reject the AR(2) against the MA(1) at the 10% level. Our test reveals that we cannot reject the null of equal revealed performance (forecast accuracy in this setting) for all five of the alternative models relative to the MA(1) process.

<table>
<thead>
<tr>
<th>Model</th>
<th>DM</th>
<th>MDM</th>
<th>MDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>1.000</td>
<td>0.9560</td>
<td>0.4185</td>
</tr>
<tr>
<td>AR(2)</td>
<td>0.0324</td>
<td>0.0520</td>
<td>0.4240</td>
</tr>
<tr>
<td>ARMA(1,1)</td>
<td>1.000</td>
<td>0.9998</td>
<td>0.4344</td>
</tr>
<tr>
<td>ARMA(2,1)</td>
<td>1.570e-04</td>
<td>0.0027</td>
<td>0.4239</td>
</tr>
<tr>
<td>ARMA(2,2)</td>
<td>6.885e-05</td>
<td>0.0019</td>
<td>0.4762</td>
</tr>
</tbody>
</table>

Our boxplot and ECDF comparison plots appear in Figures 3 and 4. One must study the plots carefully to detect visual discrepancies in both the ECDFs and the boxplots for all six models. The AR(2) has the lowest upper quartile for APSE of the six models but overall these six models
appear to have distributions of expected true errors that appear equal over 10,000 bootstrap simulations. Thus, our RP test results confirm the visual evidence.

5. Conclusion

In this paper we propose a general methodology for assessing the predictive performance of competing approximate models based on resampling techniques. The approach involves taking repeated hold-out samples (appropriately constructed) from the data at hand to create an estimate of the expected true error of a model and then using this as the basis for a test. A model possessing lower expected true error than another is closer to the underlying DGP according to the specified loss function and is therefore to be preferred. Our approach allows practitioners to compare a broad range of modelling alternatives and is not limited to the regression-based examples provided herein. The approach can be used to determine whether or not a more flexible model offers any gains in terms of expected performance than a less complex model and provides an alternative avenue for direct comparison of parametric and nonparametric regression surfaces (e.g., Härdle & Marron 1990, Härdle & Mammen 1993).

We present both simulated and empirical evidence underscoring the utility of the proposed method in dependent data settings. Our simulation results indicate that, relative to popular time-series tests, our RP test is capable of delivering substantial gains when assessing predictive accuracy. Thus, our method can be used as an auxiliary tool for assessing the accuracy of a selected model thereby enhancing any insights one might otherwise glean from empirical exercises.

We leave rigorous analysis on optimal selection of the hold-out-sample size and its impact on the resulting test statistic for future research. One could also trivially extend our testing idea to include formal tests of stochastic dominance as opposed to the visual arguments advocated in the paper, though we leave this an an exercise for the interested reader.


**Figure 1.** Time plot, auto-covariance and partial autocorrelation plots for the federal funds interest rate.

**Figure 2.** Time plot, auto-covariance and partial autocorrelation plots for the first differenced federal funds interest rate time-series.
Figure 3. Boxplots of the ASPE for the $S = 10,000$ splits of the first differenced federal funds rate time-series data. Median values for each model appear in the subtitle below the figure.

Figure 4. Empirical distribution functions of the ASPE for the $S = 10,000$ splits of the first differenced federal funds rate time-series data. Mean values for each model appear in the subtitle below the figure.
References


URL: http://www.robyndman.com/Rlibrary/forecast/
R Development Core Team (2008), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: http://www.R-project.org