A simple block bootstrap for asymptotically normal out-of-sample test statistics

Gray Calhoun*
Iowa State University

2015-04-21

Abstract

This paper proposes an improved block bootstrap method for out-of-sample statistics. Previous block bootstrap methods for these statistics have centered the bootstrap out-of-sample average on the observed out-of-sample average, which can cause the distribution to be miscentered under the null — these papers have used either a short out-of-sample period or an adjustment to the model parameter estimators under the bootstrap to correct this centering problem. Our approach centers the bootstrap replications correctly under the null while continuing to use the standard formulas to estimate the model parameters under the bootstrap, while allowing the out-of-sample period to remain large. The resulting approach is computationally more efficient, easier to program, and more widely applicable.

Keywords: Forecast Evaluation, Martingale Difference Sequence, Model Selection, Family-Wise Error Rate; Multiple Testing; Bootstrap; Reality Check

JEL Classification Numbers: C22, C53

*Economics Department; Iowa State University; Ames, IA 50011. Telephone: (515) 294-6271. Email: «gcalhoun@iastate.edu». Web: «http://gray.clhn.org». If you find an error in this paper, please let me know by email or by opening a new issue at «https://git.ece.iastate.edu/gcalhoun/oosbootstrap/issues».
I'd like to thank Helle Bunzel, Todd Clark, Michael McCracken, Elie Tamer, Ken West and two anonymous referees for helpful comments and discussions. I'd also like to thank Amit Goyal for providing computer code and data for his 2008 RFS paper with Ivo Welch (Goyal and Welch, 2008).

1 Introduction

This paper develops a block bootstrap method that can be used to consistently estimate the distributions of asymptotically normal out-of-sample (OOS) test statistics. We propose the "obvious" approach of drawing a large number of bootstrap samples from the full dataset — using the Moving Blocks, Circular Block, or Stationary bootstraps proposed by Kunsch (1989), Liu and Singh (1992), and Politis and Romano (1992, 1994), (which we will define shortly) — and then calculating the OOS statistic of interest for each bootstrap sample. We show that this approach is valid under conditions similar to West's (1996) and McCracken's (2000); i.e. when the OOS statistic itself is asymptotically normal.

The block bootstraps mentioned in the previous paragraph are all nonparametric techniques: each of these bootstraps draws J blocks of length ℓ at random from the original dataset, and assembles them into a new bootstrap time-series. If $\ell \to \infty$ as $T \to \infty$, the blocks capture the serial dependence in the original data without any additional effort by the researcher. (Under the right weak-dependence assumptions and other conditions on the DGP, obviously.) These methods differ slightly in how they conduct this random sampling. For the *Moving Blocks Bootstrap* developed by Kunsch (1989) and Liu and Singh (1992), ℓ is set by the researcher, and each block of ℓ consecutive observations is equally likely to be chosen. The same principle applies for Politis and Romano's (1992) *Circular Block Bootstrap*, but now the bootstrap is allowed to "wrap around" the endpoints of the original time series and choose, for example, the block with indices $T-1, T, 1, 2, \dots, \ell-2$.[1] Politis and Romano's (1994) *Stationary Bootstrap* extends the Circular Block Bootstrap by drawing the block length independently for each block from the geometric distribution.[2]

Although the nonparametric aspect of these block bootstraps has led to their popularity in many areas of time-series econometrics, they have been relatively unpopular in the OOS testing literature. This is due to several factors. Although the first papers developing the theoretical properties of these statistics, Diebold and Mariano (1995) and West (1996), prove asymptotic normality, subsequent papers show that asymptotic normality tends to hold only under restrictive conditions and fails otherwise. (See Clark and McCracken, 2001, and McCracken, 2007, in particular.) Consequently, most papers focus on misspecification tests for nested models, where it is natural to impose that a restricted benchmark model holds under the null hypothesis and to use that restricted model to generate the bootstrap samples, as in Kilian (1999) and Clark and McCracken (2005). However, Giacomini and White (2006), Clark and West (2006, 2007), and Calhoun (2015) have proposed OOS test statistics that are asymptotically normal under

[1] This modification ensures that the mean of the distribution induced by the bootstrap always equals the sample mean.

[2] This additional source of randomization produces a strictly stationary bootstrap sequence. It also reduces the efficiency of the Stationary Bootstrap relative to the other block bootstraps, but by less than was originally thought. See Nordman (2009) for a discussion of this issue.

general conditions, so it is worth exploring whether block bootstraps can be applied to these new statistics. This is especially true since researchers will often want to allow the benchmark model to be misspecified under the null hypothesis in forecasting applications and when comparing several models, which is straightforward with block bootstraps but more difficult with parametric bootstraps. (See in particular White, 2000, Hansen, 2005, and Romano and Wolf, 2005.)

Previous treatments of these bootstraps have focused on restricted or recentered bootstraps, but ours seems to be the first to study the theoretical properties of a standard bootstrap applied to the entire dataset. White (2000) and Hansen (2005), for example, require the out-of-sample period to be very small relative to the total sample size to remove the effects of estimating the unknown parameters of the forecasting models. Corradi and Swanson (2007) propose a different bootstrap procedure that adds a recentering term to the parameter estimates and the OOS average; these adjustments can be somewhat awkward to implement and can add to the computation time, which reduces some of the block bootstrap's advantages. Moreover, Corradi and Swanson's (2007) procedure is designed for M-estimators, and it is not obvious how to extend it to, for example, GMM. In our paper, in contrast, we show that standard nonparametric block bootstraps are consistent without modification and derive the correct centering term to ensure consistency. Although this paper presents results for M-estimators, like Corradi and Swanson's (2007), the bootstrap and mathematical arguments are standard and apply to other nonlinear estimation strategies as well, including GMM.

The next section presents our theoretical results and further explains the statistics that we cover in this paper. Section 3 presents an empirical illustration of our approach based on Calhoun's (2015) mixed-window OOS statistic, and Section 4 presents a Monte Carlo experiment that studies the bootstrap's finite sample properties. Finally, Section 5 concludes.

2 The Bootstrap for Out-of-Sample Statistics

We'll develop our theoretical results in a fairly general framework. Let y_{t+1} be a target variable of interest — a variable that is being predicted — and let x_t be a vector of other variables that are potentially informative about y_{t+1} — these are our predictors. The forecast $\hat{y}_{t+1}$ depends on the variables x_t and an estimated parameter $\hat{\beta}_t$. In the research project that we're trying to model, we're interested in a function of these variables and parameters, and the OOS average of that function is our test statistic.

In symbols, we're interested in statistics of the form

$$\bar{f} = \tfrac{1}{\sqrt{P}} \sum_{t=R}^{T-1} f(y_{t+1}, x_t, \hat{\beta}_{1t}, \dots, \hat{\beta}_{kt}) \equiv \tfrac{1}{\sqrt{P}} \sum_{t=R}^{T-1} f_t(\hat{\beta}_{1t}, \dots, \hat{\beta}_{kt}),$$

where each $\hat{\beta}_{kt}$ corresponds to a different forecasting model. To make the notation cleaner, we'll define $f_t(\beta_1, \dots, \beta_k) \equiv f(y_{t+1}, x_t, \beta_1, \dots, \beta_k)$. We're also going to assume that (y_{t+1}, x_t) is strictly stationary to simplify our presentation. One could derive the same results under the marginally weaker assumption that certain functions of these variables are weakly stationary.

The coefficients are updated each period to mimic a true OOS forecasting exercise. Using standard terminology, the estimator $\hat{\beta}_{it}$ is defined as

$$\hat{\beta}_{it} = \begin{cases} \arg\min_\beta \sum_{s=1}^{t-1} q_i(y_{s+1}, x_s, \beta) & \text{recursive window} \\ \arg\min_\beta \sum_{s=t-R+1}^{t-1} q_i(y_{s+1}, x_s, \beta) & \text{rolling window} \\ \arg\min_\beta \sum_{s=1}^{R-1} q_i(y_{s+1}, x_s, \beta) & \text{fixed window,} \end{cases} \tag{1}$$

and, as before, to make the notation cleaner, define $q_{is}(\beta) \equiv q_i(y_{s+1}, x_s, \beta)$. Obviously, for this to be a reasonable estimation approach q_{is} will need to satisfy standard assumptions that we'll discuss soon. The implicit assumption is that the researcher is interested in conducting inference on $\mathrm{E} f_t(\beta_{10}, \dots, \beta_{k0})$, where

$$\beta_{i0} = \arg\min_\beta \mathrm{E}\, q_i(y_{s+1}, x_s, \beta)$$

is the pseudotrue equivalent of $\hat{\beta}_{it}$,

We know from West (1996) that, under appropriate assumptions, $\sqrt{P}(\bar{f} - \mathrm{E} f_t(\beta_0))$ is asymptotically normal with mean zero. The key insight in our paper is that we can match this result in the bootstrap, but we need to be careful about the exact centering term. In particular, we should expect

$$\sqrt{P}(\bar{f}^* - \mathrm{E}^* f_t^*(\beta_0^*))$$

to have the same asymptotic distribution and to give reliable confidence intervals, etc. where a * denotes a quantity under the bootstrap distribution and

$$\bar{f}^* = \tfrac{1}{P} \sum_{t=R}^{T-1} f(y_{t+1}^*, x_t^*, \hat{\beta}_{1t}^*, \dots, \hat{\beta}_{kt}^*) \equiv \tfrac{1}{P} \sum_{t=R}^{T-1} f_t^*(\hat{\beta}_{1t}^*, \dots, \hat{\beta}_{kt}^*) \tag{2}$$

where $\hat{\beta}_{it}^*$ is estimated exactly the same way as $\hat{\beta}_t$:

$$\hat{\beta}_{it}^* = \begin{cases} \arg\min_\beta \sum_{s=1}^{t-1} q_{is}^*(\beta) & \text{recursive window} \\ \arg\min_\beta \sum_{s=t-R}^{t-1} q_{is}^*(\beta) & \text{rolling window} \\ \arg\min_\beta \sum_{s=1}^{R-1} q_{is}^*(\beta) & \text{fixed window} \end{cases} \tag{3}$$

and $q_{is}^*(\beta) \equiv q_i(y_{s+1}^*, x_s^*, \beta)$. For the circular and stationary block bootstraps,

$$\mathrm{E}^* f_t^*(\beta_1, \dots, \beta_k) = \tfrac{1}{T-1} \sum_{t=1}^{T-1} f_t(\beta_1, \dots, \beta_k)$$

and

$$\beta_{i0}^* = \arg\min_\beta \sum_{s=1}^{T-1} q_i(y_{s+1}, x_s, \beta).$$

For the moving blocks bootstrap, a slight correction is necessary since observations at the ends of the sample are less likely to be selected, but the same equations hold approximately. In any of these three bootstraps, $\mathrm{E}^* f_t^*(\beta_0^*)$ does not incorporate the $\hat{\beta}_t$ terms.

In general, let $\to^{p^*}$ and $\to^{d^*}$ refer to convergence in probability or distribution conditional on the observed data. We will present the required theoretical assumptions first, then present our results.

Assumption 1. *The estimators $\hat{\beta}_{it}$ and $\hat{\beta}_{it}^*$ are estimated as defined in Equations* (1) *and* (3). *Moreover each $\beta_{i0} = \arg\min_\beta \mathrm{E}\, q_{is}(\beta)$ is uniquely identified and the vector $(\beta_{10}, \dots, \beta_{k0})$ is an element of a compact set Θ.*

For the next result, let $\nabla h(\beta)$ and $\nabla^2 h(\beta)$ refer to the first and second derivative of the function h. If β is a K-vector, $\nabla h(\beta)$ will be $K \times 1$ and $\nabla^2 h(\beta)$ will be $K \times K$. Also let $\nabla_i h(\beta)$ refer to the ith element of $\nabla h(\beta)$ and $\nabla_{ij}^2 h(\beta)$ to the (i, j) element of $\nabla^2 h(\beta)$.

This assumption imposes standard moment and smoothness conditions on the underlying functions $f_t(\cdot)$ and $q_{it}(\cdot)$. It is quite likely that these are stronger than necessary and could be weakened to smoothness conditions on $\mathrm{E} f_t(\cdot)$ and $\mathrm{E} q_{it}(\cdot)$, as in McCracken (2000), but we leave that for future work.[3]

Assumption 2. *The functions $f_t(\beta_1, \dots, \beta_k)$ and $q_{it}(\beta)$ are almost surely twice continuously differentiable in an open neighborhood N of $(\beta_{10}, \dots, \beta_{k0})$ and $\mathrm{E}\, \nabla^2 q_{it}(\beta)$ is positive definite uniformly in N. There also exists a sequence of random variables m_t such that $\sup_{\beta \in N} |\nabla_i^2 q_{jt}(\beta)| \le m_t$, $\sup_{\beta \in N} |\nabla_{ij}^2 f_t(\beta_1, \dots, \beta_k)| \le m_t$, $\sup_{\beta \in N} |\nabla_i q_{jt}(\beta)| \le m_t$, and $\sup_{\beta \in N} |\nabla_i f_t(\beta_1, \dots, \beta_k)| \le m_t$ almost surely and $\mathrm{E}\, m_t^r$ is uniformly finite, with $r > 2$ defined further in Assumption 3.*

The next assumptions handle weak dependence and stationarity. These assumptions are weaker than are typically used in this literature because of advances in the underlying CLT and bootstrap theory used. For Assumption 3, define

$$g_t(\beta_0) = (f_t(\beta_0), \nabla q_{1t}(\beta_1), \dots, \nabla q_{kt}(\beta_i))'.$$

Assumption 3. *The stochastic process $(g_t(\beta_0), vec(\nabla g_t(\beta_0)))$ is weakly stationary. Moreover, (y_{t+1}, x_t) is strong-mixing of size $-r/(r-2)$ or uniform mixing of size $-r/(2r-2)$ with $r > 2$.*

[3] Extending our results in this way would be equivalent to extending de Jong and Davidson (2000) in the same way, which appears feasible but nontrivial.

The next assumption limits the practical applicability of these results, but is difficult to relax in general. There are OOS test statistics that satisfy this condition (Giacomini and White, 2006, Clark and West, 2006, 2007, and Calhoun, 2015) but many do not.

Assumption 4. *The asymptotic variance matrix of $\bar{f}$ is uniformly positive definite.*

Finally, we make standard assumptions on the size of the in-sample and out-of-sample sizes and on the block length of the bootstrap.

Assumption 5. *$R, P \to \infty$ as $T \to \infty$. The bootstrap sequence $(y_2^*, x_1^*), \dots, (y_T^*, x_{T-1}^*)$ is constructed using a moving blocks, circular blocks, or stationary bootstrap with block lengths drawn from the geometric distribution. The (expected) block length ℓ satisfies $\ell \to \infty$ and $\ell / T \to 0$.*

Then the main result is quite simple: the bootstrap distribution is consistent for the asymptotic distribution of the statistic and the bootstrap variance is consistent for the asymptotic variance of the original OOS statistic.

Theorem 1. *Under Assumptions 1–5,* $\mathrm{var}(\bar{f}) / \mathrm{var}^*(\bar{f}^*) \to^p 1$ *and*

$$\Pr\Big[\sup_x \Big| \Pr^*[\sqrt{P}(\bar{f}^* - \mathrm{E}^* f_t^*) \leq x] - \Pr[\sqrt{P}(\bar{f} - \mathrm{E} f_t) \leq x] \Big| > \epsilon \Big] \to 0 \tag{4}$$

for all $\epsilon > 0$.

A few quick remarks follow.

Remark 1. Typically this result will be used to test the null hypothesis $\mathrm{E} f_t = 0$. To generate critical values for this test, researchers need to calculate $\mathrm{E}^* f_t^*$ so that it can be removed from the bootstrapped OOS statistic. We can use the bootstrap average to approximate $\mathrm{E}^* f_t^*$ as one would expect:

$$\mathrm{E}^* f_t^* \approx \tfrac{1}{n} \sum_{i=1}^{n} \bar{f}_i^*$$

where there are n bootstrap replications and $\bar{f}_i^*$ represents the ith realization of the bootstrapped statistic.

Remark 2. Often the bootstrap is more accurate for studentized statistics than for the corresponding sample mean. One can certainly estimate the asymptotic variance in our setting and apply the bootstrap to the studentized OOS statistic. But there are other options as well: first, one can use the bootstrap to estimate the variance of the OOS statistic, then use a double bootstrap to normalize the bootstrap statistic. Obviously, this may be computationally impractical. Another approach is to partially studentize the statistics by dividing by the naive estimator of the standard deviation, which may reduce some of the effects of the variance.

Remark 3. As mentioned earlier, this bootstrap procedure can be especially useful when comparing multiple forecasting models, which will be part of our empirical example. White (2000), Hansen (2005) and Romano and Wolf (2005) are substantial contributions to this literature and Romano et al. (2008) review aspects of this literature as well.

Remark 4. The issue of choosing the block length is clearly very important but is beyond the scope of this paper. For some guidance, see Politis and White (2004), Romano and Wolf (2006), and Patton et al. (2009).

Related to Remark 4, economic theory will sometimes imply that f_t should be a martingale difference sequence, at least under the null hypothesis of interest. Under this stronger null hypothesis, researchers can avoid choosing the block length and the procedure is simplified somewhat: the bootstrap is consistent with a block length 1. Theorem 2 formalizes this result.

Theorem 2. *Suppose that Assumptions 1–5 hold and also assume that $f_t - \mathrm{E} f_t$ is an MDS and that the i.i.d. bootstrap is used instead of the block bootstraps of Theorem 1. Then* $\mathrm{var}(\bar{f})/\mathrm{var}^*(\bar{f}^*) \to^p 1$ *and*

$$\Pr\Big[\sup_x \Big|\Pr^*[\sqrt{P}(\bar{f}^* - \mathrm{E}^* f_t^*) \leq x] - \Pr[\sqrt{P}(\bar{f} - \mathrm{E} f_t) \leq x]\Big| > \epsilon\Big] \to 0 \tag{5}$$

for all $\epsilon > 0$.

The proof is a straightforward modification of that of Theorem 1 and is omitted. It is important to recognize that this result allows for other forms of serial dependence, as long as the MDS property holds.

3 Empirical Illustration

This section demonstrates the use of the bootstrap by revisiting Goyal and Welch's (2008) study of excess stock returns. Goyal and Welch argue that many variables thought to predict excess returns (measured as the difference between the yearly log return of the S&P 500 index and the T-bill interest rate) on the basis of in-sample evidence fail to do so out-of-sample. To show this, Goyal and Welch look at the forecasting performance of models using a lag of the variable of interest, and show that these models do not significantly outperform the excess return's recursive sample mean.

We will conduct the same analysis here, but using the asymptotically normal MDS test proposed by Calhoun (2015). The benchmark model is the excess return's sample mean (as in the original) and the alternative models are of the form

$$\text{excess return}_t = \alpha_0 + \alpha_1 \text{ predictor}_{t-1} + \varepsilon_t. \tag{6}$$

The predictors used are listed in the left column of Table 1 (see Goyal and Welch, 2008, for a detailed description of the variables). The data set is annual data beginning in 1927 and ending in 2009, and the rolling window uses 10 observations.

To implement Calhoun's (2015) statistic, we estimate α_0 and α_1 for each predictor using OLS with a 10-year rolling window to produce forecasts $\hat{y}_{it}$. We also use the sample mean calculated with a *recursive* window as the benchmark forecast, $\hat{y}_{0t}$. The OOS statistic is based on the adjusted difference in squared-error between these forecasts for each predictor,

$$\bar{f}_i = \tfrac{1}{P}\sum_{t=R}^{T-1}\big[(y_{t+1} - \hat{y}_{0,t+1})^2 - (y_{t+1} - \hat{y}_{i,t+1})^2 + (\hat{y}_{0,t+1} - \hat{y}_{i,t+1})^2\big].$$

Calhoun (2015) shows that this statistic remains asymptotically normal as $T \to \infty$ as long as the window length of the rolling window stays fixed and that $\mathrm{E}\,\bar{f}_i$ is asymptotically normal with mean zero under the null hypothesis that $y_{t+1} - \mathrm{E}\,y_{t+1}$ is an MDS with respect to the information set generated by each of the predictors considered by Goyal and Welch (2008) and listed in Table 1.

We will test against one-sided alternatives, so the test rejects for large values of $\bar{f}_i$. (i.e. when the regressor in the alternative model has predictive power.) We use Hansen's (2005) test of *Superior Predictive Ability* (SPA) to account for multiplicity. Hansen's test uses studentized statistics,[4] so we use the variance formula derived by Calhoun (2015) (let $\hat{\sigma}_i^2$ denote the estimated variance for $\sqrt{P}\bar{f}_i$), and proceeds in several steps:

- First, all of the statistics with $\sqrt{P}\bar{f}_i/\hat{\sigma}_i$ less than $-\sqrt{2\log\log P}$ are removed and set aside. These statistics are far enough from their alternatives that they can be treated as if the were known to be true. Keeping them in the analysis, as originally proposed by White (2000), will make the overall test unnecessarily conservative.

  Let S be the set of the indices of statistics remaining after this step, so

  $$S = \{i \mid \sqrt{P}\bar{f}_i/\hat{\sigma}^i > -\sqrt{2\log\log P}\}.$$

  In our application, none of the statistics are removed by this first step.

- Second, calculate the $1-\alpha$ quantile of

  $$\max_{i\in S}\sqrt{P}(\bar{f}_i^* - \mathrm{E}^* f_{it}^*)/\hat{\sigma}_i^*$$

  with the bootstrap. Call the value of this quantile $\hat{c}$. In this application, α is 0.1 and $\hat{c} = 2.67$. (Based on 599 replications with i.i.d. sampling, as suggested by Theorem 2.)

[4]We actually depart a little from Hansen's (2005) SPA test, in that we studentize each of the bootstrapped statistics as well. Hansen recommends normalizing each $\bar{f}_i^*$ with its population standard deviation under the distribution induced by the bootstrap; this is a shortcut that can save computation time, but is not necessary here.

	value	naive	SPA	ours
long term rate	1.56	sig.	sig.	
book to market	1.41	sig.	sig.	
dividend yield	1.27		sig.	
dividend price ratio	0.95			
net equity	0.70			
dividend payout ratio	0.64			
treasury bill	0.53			
stock variance	0.50			
default return spread	0.16			
default yield spread	0.09			
inflation	−0.09			
term spread	−0.43			
earnings price ratio	−0.56			
long term yield	−0.74			

Table 1: Results from OOS comparison of equity premium prediction models; the benchmark is the recursive sample mean of the equity premium and each alternative model is a constant and single lag of the variable listed in the "predictor" column. The dataset begins in 1927 and ends in 2009 and is annual data. The "value" column lists the value of this paper's OOS statistic, the "naive" column indicates whether the statistic is significant at standard critical values, the "SPA" column indicates significance using the SPA bootstrap (incorrectly) to account for the number of models, and the "corrected" column indicates significance using the critical values generated by our bootstrap that account for the number of models using Hansen's (2005) SPA algorithm correctly. See Section 3 for details.

- Last, compare the individual test statistics to $\hat{c}$. If any $\sqrt{P}(\bar{f}_i^* - \mathrm{E}^* f_{it}^*)/\hat{\sigma}_i^* > \hat{c}$, the MDS null hypothesis is rejected. Moreover, the weaker null hypothesis that $y_{t+1} - \mathrm{E}\, y_{t+1}$ is an MDS with respect to the ith predictor alone is also rejected.

Table 1 presents the results of our analysis.[5] The column "value" gives the value of the test statistic for each model, while the "naive" and "corrected" columns indicate whether the statistic is greater than the standard size-10% critical value (1.28) and the critical value estimated by the bootstrap (2.67). The "SPA" column indicates whether the statistic is greater than the critical value produced by misapplication of Hansen's (2005) SPA algorithm (1.26) — i.e. bootstrapping the $\hat{f}_{it}$ values to generate a critical value. This is similar to the procedure proposed by Hansen, but Hansen's (2005) proposal is for a very different setting where R is large and P is small.

Two predictors are significant at the naive critical value, the long term interest rate and the book to market ratio, and a third at the the misused SPA critical value, the divi-

[5]This statistical analysis was conducted in R (R Development Core Team, 2011) using the xtable (Dahl, 2009, version 1.6-0), and dbframe (Calhoun, 2010, version 0.2.7) packages.

dend yield. But none are significant after correcting for both parameter estimation error and data snooping with our proposed approach. This suggests that the error introduced by a misapplied bootstrap has as large of an effect on inference as neglecting to control for multiple comparisons.

4 Monte Carlo results

The Monte Carlo simulations we present are aimed at addressing a simple question: if a standard block bootstrap works well in this setting, why has no one used it? Does the block bootstrap work at all? Obviously, this is just a preliminary first step in understanding the finite-sample behavior of these statistics, and I plan to add several other designs to future versions of this paper.

For now, we will use a very simple Data Generating Process, originally proposed by West (1996). In this example, the data are generated by the following system of equations:

$$\begin{aligned} y_t &= \gamma_0 + \gamma_1 w_{1t} + \gamma_2 w_{2t} + v_t \\ w_{it} &= z_{it} + v_t \\ (v_t, z_{1t}, z_{t2}) &\sim \text{i.i.d. } N(0, I_3). \end{aligned}$$

The two competing forecasting models are $y_t = \alpha_i + \beta_i w_{it} + u_{it}$ and the coefficients are estimated by Instrumental Variables using z_{it} as the instrument. The OOS test statistic is just the difference in squared loss associated with the forecasts, and the null is that the expected squared loss is the same. This estimation strategy — IV instead of OLS— is appropriate when one wants to use the forecast performance of the models as a proxy for other aspects of their specification. One could, of course, estimate the models with OLS instead of IV, but that would not say anything about the structural models underlying the regressions.

This design has several interesting features. First, the IV estimators are not M-estimators but are GMM estimators, and it is not trivial to derive Corradi and Swanson's (2007) correction term for them, so we omit their bootstrap. Even though our theoretical results assume that M-estimators are used for the forecast, inspection of the mathematical proofs show that the basic strategy used would apply equally well to GMM estimators, and so we should expect our results to apply here as well. Moreover, since we are proposing a simple block bootstrap, it is easy to implement regardless of the actual statistic.

Second, parameter estimation error has a substantial effect in this setting. In West's (1996) simulations, the naive OOS test statistic that ignores this source of error can have rejection probabilities of up to 50% for a test with nominal size 5%. The size distortions shrink if P is quite small relative to R. In our setting, we would expect the same behavior for the naive OOS bootstrap that only samples from $\hat{f}_R, \dots \hat{f}_{T-1}$.

T	P	naive bootstrap	our bootstrap
300	50	24.1	7.5
	100	34.6	7.2
	200	51.2	7.3
	250	55.3	7.1
500	50	19.7	8.8
	150	32.6	7.8
	350	50.1	7.9
	450	58.5	8.2

Table 2: Results of the Monte Carlo experiment, based on 2000 simulations from the DGPdescribed in Section 4 with 499 bootstrap replications each. The "naive bootstrap" column lists the actual size using resamples of the observed out-of-sample loss to produce the bootstrap critical values and the column labeled "our bootstrap" uses the method proposed in this paper.

For the specific simulation parameters, we run 2000 simulations each with 499 bootstrap replications. We run simulations with 300 observations and 500 observations and consider several splits between R and P. Since the observations are independent, we do a simple i.i.d. bootstrap. (This is equivalent to using a block length of 1.) All tests are two-sided with $\alpha = 10\%$, and all of the simulations were conducted in Julia version 0.3.6 (described in Bezanson et al., 2012, and Bezanson et al., 2014).

The simulation results are presented in both a table (Table 2) and in a dot chart (Figure 1); the table lets us read the individual values cleanly and the chart makes it easy to spot patterns. We can see immediately that the test based on the correct bootstrap is slightly undersized, but appears to not overreject for any of the parametrizations we consider. The test based on naively bootstrapping the realized out-of-sample values, on the other hand, is seriously deficient and overrejects considerably; almost 60% at worst and almost 20% at best. As we expect, the overrejection probability is smaller when P is very small relative to the total sample size, but increases as P gets larger.

These results are extremely preliminary and incomplete. For future versions of the paper we plan to include the following as well:

- Comparison to West's (1996) critical values that do not use the bootstrap.
- Compare studentized statistics (which we don't cover) with unstudentized.
- Use the DGP and statistics from the empirical section to make sure that the bootstrap works in that setting and to compare to other bootstrap methods.
- Check FWE control in multiple testing.

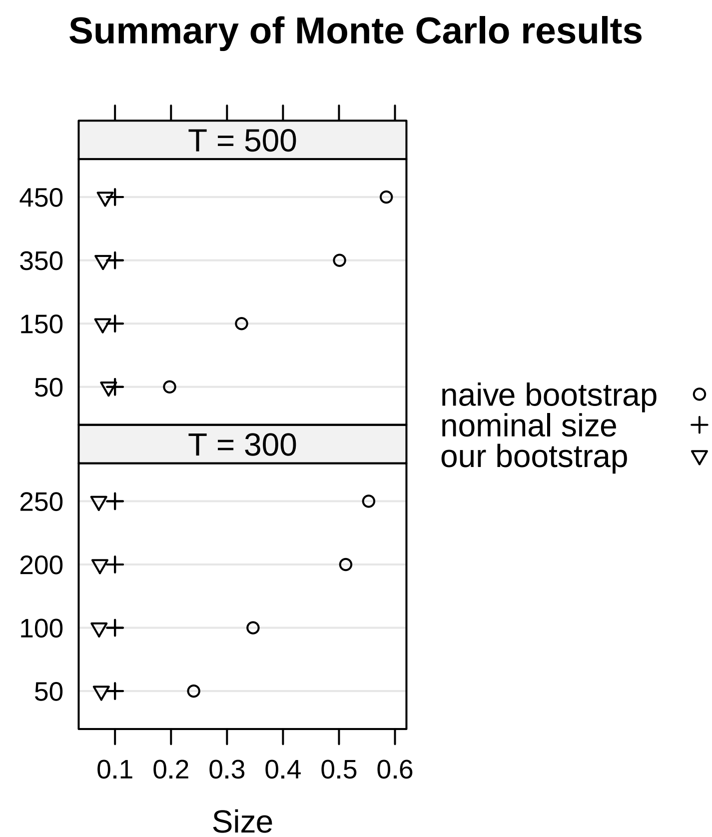

Figure 1: Results of the Monte Carlo experiment, based on 2000 simulations from the DGPdescribed in Section 4 with 499 bootstrap replications each. The circles labeled "naive bootstrap" plot the actual size using resamples of the observed out-of-sample loss to produce the bootstrap critical values and the points labeled "our bootstrap" use the method proposed in this paper. The tests' nominal size is plotted for reference.

5 Conclusion

This paper establishes that standard block bootstraps can be used to consistently estimate the distribution of asymptotically normal OOS statistics. We also show how the bootstrap can be used to correct for multiple testing in empirical applications, along the lines of White's (2000) Reality Check and provide simulation evidence on the performance of our approximation in finite samples.

Appendix: Additional mathematical results

We will prove our results under the simplifying assumption that there is a single model and a single sequence of M-estimators $\hat{\beta}_t$. Since we are assuming non-degeneracy of the models, this assumption does not appreciably change our arguments. This also implies that we will drop the i index for the estimators $\hat{\beta}_{it}$, estimation criteria $q_{it}(\beta)$ etc.

We will also present proofs for the recursive window; the fixed and rolling window have similar proofs but are less complicated.

To make the mathematical results in this appendix clearer, we will introduce the following additional notation:

- $f_t = f_t(\beta_0)$ and $f_t^* = f_t^*(\beta_0^*)$
- $F_t(\beta) = \nabla f_t(\beta)$ and $F_t^*(\beta) = \nabla f_t^*(\beta)$,
- $F_t = F_t(\beta_0)$ and $F_t^* = F_t^*(\beta_1^*)$,
- $F = \mathrm{E}\, F_t$ and $F^* = \mathrm{E}^*\, F_t^*$,
- $h_t(\beta) = \nabla q_t(\beta)$ and $h_t^*(\beta) = \nabla q_t^*(\beta)$,
- $h_t = h_t(\beta_0)$ and $h_t^* = h_t^*(\beta_0^*)$.

Where it is feasible, we will reuse notation from West (1996) and West and McCracken (1998).

Also define

$$S_{ff} = \sum_{j=-\infty}^{\infty} \mathrm{E}\, f_t f_{t-j}'$$

$$S_{fh} = \sum_{j=-\infty}^{\infty} \mathrm{E}\, f_t h_{t-j}'$$

$$S_{hh} = \sum_{j=-\infty}^{\infty} \mathrm{E}\, h_t h_{t-j}',$$

$\pi = \lim P/R$, and

$$\lambda_{fh} = \begin{cases} 1 - \pi^{-1} \ln(1+\pi) & \text{recursive window, } \pi \in (0,\infty) \\ \pi/2 & \text{rolling window, } \pi \leq 1 \\ 1 - \pi/2 & \text{rolling window, } \pi > 1 \\ 0 & \text{fixed window,} \end{cases}$$

$$\lambda_{hh} = \begin{cases} 2\lambda_{fh} & \text{recursive window} \\ \pi - \pi^2/3 & \text{rolling window, } \pi \leq 1 \\ \pi - 1/3\pi & \text{rolling window, } \pi > 1 \\ \pi & \text{fixed window.} \end{cases}$$

Also define $u_1, \dots, u_J$ to be the first period of each block of the circular bootstrap, and, for each $j = 1, \dots, J$, define the σ-fields

$$\mathscr{H}_j = \sigma(u_1, \dots, u_j)$$

and

$$\mathscr{H}_j^* = \sigma(u_1, \dots, u_j; y_1, \dots, y_T; x_1, \dots, x_T).$$

Also let $l = T - J\ell$ be the number of elements in the last block.

Proof of Theorem 1

The proof proceeds in several steps. First, we prove via a Taylor expansion (as in West, 1996) that

$$\Pr{}^*[\sqrt{P}(\bar{f}^* - \mathrm{E}^* f_t^*) \le x] \to^p \Phi(x/\sigma) \tag{7}$$

where Φ is the CDF of the standard normal and σ is a known constant. Similar arguments directly following West's imply that

$$\Pr[\sqrt{P}(\bar{f} - \mathrm{E} f_t) \le x] \to^p \Phi(x/\sigma) \tag{8}$$

under our assumptions, so

$$\Pr{}^*[\sqrt{P}(\bar{f}^* - \mathrm{E}^* f_t^*) \le x] \to^p \Pr[\sqrt{P}(\bar{f} - \mathrm{E} f_t) \le x]. \tag{9}$$

Moreover, the assumed moment conditions ensure that the variance of $\sqrt{P}\bar{f}^*$ under the bootstrap distribution converges to the variance of $\sqrt{P}\bar{f}$. Finally, a standard argument attributed to Polyà ensures that (4) follows from (9). (See the proof of Theorem 1 in Calhoun, 2014, for example, for an explicit statement of these final steps.)

For (7), begin by expanding $f_t^*(\hat{\beta}_t^*)$ around β_0^* to get

$$\begin{aligned}\sqrt{P}(\bar{f}^* - \mathrm{E}^* f_t^*) &= \tfrac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (f_t^* - \mathrm{E}^* f_t^*) + \tfrac{1}{\sqrt{P}} \sum_{t=R}^{T-1} F_t^* \cdot (\hat{\beta}_t^* - \beta_0^*) + \tfrac{1}{\sqrt{P}} \sum_{t=R}^{T-1} w_t^* \\ &= \tfrac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (f_t^* - \mathrm{E}^* f_t^*) + F^* B^* \tfrac{1}{\sqrt{P}} \sum_{t=1}^{T-1} a_t h_t^* + o_{p^*}(1)\end{aligned}$$

where (similar to West, 1996)

$$w_t = \tfrac{1}{2}(\hat{\beta}_t^* - \beta_0^*)' \nabla^2 f_t^*(b_t^*)(\hat{\beta}_t^* - \beta_0^*),$$

$$a_t = \begin{cases} \sum_{s=\max(R-1,t)}^{T-1} 1/s & \text{recursive window} \\ \min(\frac{t}{R-1}, \frac{T-t}{R-1}, 1) & \text{rolling window} \\ \frac{P}{R-1} 1\{t < R-1\} & \text{fixed window,} \end{cases} \tag{10}$$

and each b_t^* lies between $\hat{\beta}_t^*$ and β_0^*. The second equality holds because $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1} w_t^* = o_{p^*}(1)$ and

$$\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1} F_t^* \cdot (\hat{\beta}_t^* - \beta_0^*) = F^* B^* \tfrac{1}{\sqrt{P}}\sum_{t=1}^{T-1} a_t h_t^* + o_{p^*}(1),$$

both from Lemma A.4.

By Lemma A.5,

$$\tfrac{1}{\sqrt{P}}\sum_{t=1}^{T-1}\begin{pmatrix}(f_t^* - \mathrm{E}^* f_t^*)1\{t \geq R\} \\ a_t h_t^*\end{pmatrix} \to^{d^*} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}S_{ff} & \lambda_{fh} S_{fh} \\ \lambda_{fh} S_{fh}' & \lambda_{hh} S_{hh}\end{pmatrix}\right) \tag{11}$$

and $F^* \to^p F$ and $B^* \to^p B$ by Lemma A.1, so

$$\Pr{}^*[\sqrt{P}(\bar{f}^* - \mathrm{E}^* f_t^*) < x] \to^p \Phi(x/\sigma) \tag{12}$$

for all x, with

$$\sigma^2 = S_{ff} + \lambda_{fh}(FBS_{fh} + S_{fh}' B' F') + \lambda_{hh} FBS_{hh} B' F'.$$

Normality for the original OOS average follows essentially the same argument (this is essentially the same argument as in West, 1996; the Lemmas referenced above establish intermediate results for the original OOS statistic under our assumptions as well as the bootstrapped statistic), so

$$\Pr[\sqrt{P}(\bar{f} - \mathrm{E} f_t) < x] \to \Phi(x/\sigma) \tag{13}$$

for all x. As discussed above, this completes the proof. □

A Supporting Results

Lemma A.1. *Under the conditions of Theorem 1, $\beta_0^* \to^p \beta_0$, $B^* \to^p B$, $F^* \to^p F$.*

Proof of Lemma A.1. We'll present proofs of these results for the circular block bootstrap; the proofs for the moving blocks and stationary bootstraps are similar. For β_0^*, by definition $\beta_0^* = \arg\min_\beta \sum_{s=2}^T q_s(\beta)$. Our smoothness and moment conditions ensure that $\sum_{s=2}^T q_s(\beta)$ obeys a uniform LLN and converges in probability to $\mathrm{E}\, q_s(\beta)$ for all $\beta \in \Theta$. Then consistency of β_0^* follows from, for example, Theorem 2.1 of Newey and McFadden (1994).

For F^*, we have

$$\Pr[|F^* - F| > \epsilon] \leq \Pr[|F^* - F| 1\{\beta_0^* \in N\} > \epsilon] + \Pr[\beta_0^* \notin N].$$

The second probability converges to zero by consistency of β_0^*. Now $F^* - F = F^* - F(\beta_0^*) + F(\beta_0^*) - F$, and $F^* - F(\beta_0^*) \to^p 0$ by the uniform LLN. Choose Δ so that $|\beta_1 - \beta_2| < \Delta$ implies that $|F(\beta_1) - F(\beta_2)| < \epsilon$. Then

$$\Pr[|F(\beta_0^*) - F| > \epsilon] \leq \Pr[|\beta_0^* - \beta_0| > \Delta]$$

which converges to zero by the first part of this Lemma. The proof for B^* is similar. □

Lemma A.2. *Under the conditions of Theorem 1,*

$$\max_{t=R,\dots,T-1} |\hat{\beta}_t - \beta_0| \to^p 0 \tag{14}$$

$$\max_{t=R,\dots,T-1} |\hat{\beta}_t^* - \beta_0^*| \to^{p^*} 0 \tag{15}$$

$$\max_{t=R,\dots,T-1} \Big| -\tfrac{1}{t-1} \sum_{s=1}^{t-1} \nabla h_s(b_t) - B^{-1} \Big| \to^p 0 \tag{16}$$

and

$$\max_{t=R,\dots,T-1} \Big| -\tfrac{1}{t-1} \sum_{s=1}^{t-1} \nabla h_s^*(b_t^*) - B^{*-1} \Big| \to^p 0 \tag{17}$$

where each b_t is any array a.s. between $\hat{\beta}_t$ and β_0 and each b_t^* is any array a.s. between $\hat{\beta}_t^*$ and β_0^*.

The proof of (16) follows from standard arguments for M-estimators and is also omitted.

Proof of (15). First, assume $t \to \infty$ as $T \to \infty$. We have $\Pr^*[|\hat{\beta}_t^* - \beta_0^*| > \epsilon] \to^p 0$ if $\Pr[|\hat{\beta}_t^* - \beta_0^*| > \epsilon] \to 0$. To prove this second convergence, we will first establish that

$$\sup_{\beta \in N} \tfrac{1}{t-1} \sum_{s=1}^{t-1} (q_s^*(\beta) - \mathrm{E}^* q_s^*(\beta)) \to^p 0. \tag{18}$$

Pointwise convergence holds from the LLN (Calhoun, 2014) and stochastic equicontinuity of this function is implied by our moment and smoothness conditions, so (18) holds by standard arguments. Then given uniform convergence and identification, $\Pr[|\hat{\beta}_t^* - \beta_0^*| > \epsilon] \to 0$ follows.

Then extending this result to

$$\Pr[\max_{t=R,\dots,T-1} |\hat{\beta}_t^* - \beta_0^*| > \epsilon] \to 0$$

follows the same argument as used in Calhoun's (2014) FCLT. □

Proof of (17). First, observe that for any Δ

$$\Pr^*\Big[\sup_{t=R,\dots,T-1} \Big| -\tfrac{1}{t} \sum_{s=1}^{t} \nabla h_s^*(b_t^*) - B^{*-1} \Big| > \Delta\Big] \tag{19}$$

$$\leq \Pr^*\Big[\sup_{t=R,\dots,T-1} \Big| -\tfrac{1}{t} \sum_{s=1}^{t} \nabla h_s^* - B^{*-1} \Big| 1\{\beta_0^* \in N\} > \Delta\Big] \tag{20}$$

$$+ \Pr\Big[\sup_{t=R,\dots,T-1} \Big| -\tfrac{1}{t} \sum_{s=1}^{t} (\nabla h_s^*(b_t^*) - \nabla h_s^*) \Big| 1\{\beta_0^* \in N,\ \hat{\beta}_t^* \in N\} > \Delta\Big] \tag{21}$$

$$+ \Pr[\beta_0^* \notin N] + \Pr[\hat{\beta}_t^* \notin N \text{ for some } t = R, \dots, T-1] \tag{22}$$

The last two probabilities converges to zero by Lemma A.3 and by (15). Moreover, just as in the proof of Theorem 1, $(1/t)\sum_{s=1}^{t} \nabla h_s^*$ can be re-expressed as the sum of a uniformly integrable MDS that obeys a uniform LLN, so the first probability on the rhs of (19) converges to zero. Finally, since $\nabla h_s(\beta)$ is continuous uniformly in N, we can choose δ so that $|\beta_1 - \beta_2| < \delta$ implies that $|\nabla h_s(\beta_1) - \nabla h_s(\beta_2)| < \Delta$. Then

$$\Pr^*\Big[\sup_{t=R,\dots,T-1} \Big| -\tfrac{1}{t}\sum_{s=1}^{t}(\nabla h_s^*(b_t^*) - \nabla h_s^*)\Big| 1\{\beta_0^* \in N,\ \hat{\beta}_t^* \in N\} > \Delta\Big] \leq$$
$$\Pr^*[\sup_{t=R,\dots,T-1} |b_t^* - \beta_0^*| > \delta \text{ and } \beta_0^* \in N, \text{ and } \hat{\beta}_t^* \in N \text{ for all } t = R,\dots,T-1]$$

which again converges to zero in probability.

Now choose Δ so that

$$\Big| -\tfrac{1}{t}\sum_{s=1}^{t} \nabla h_s^*(b_t) - B^{*-1}\Big| < \Delta$$

implies that

$$\Big|\Big[-\tfrac{1}{t}\sum_{s=1}^{t} \nabla h_s^*(b_t)\Big]^{-1} - B^*\Big| < \epsilon.$$

Then

$$\Pr^*\Big[\sup_{t=R,\dots,T-1} \Big|\Big[-\tfrac{1}{t}\sum_{s=1}^{t} \nabla h_s^*(b_t)\Big]^{-1} - B^*\Big| > \epsilon\Big] \leq$$
$$\Pr^*\Big[\sup_{t=R,\dots,T-1} \Big| -\tfrac{1}{t}\sum_{s=1}^{t} \nabla h_s^*(b_t) - B^{*-1}\Big| > \Delta\Big] \to^{p^*} 0,$$

completing the proof. □

Lemma A.3. *If $a \in [0, 1/2)$ and the conditions of Theorem 1 hold, then*

$$\max_{t=R,\dots,T-1} \Big|(t-1)^{a-1}\sum_{s=1}^{t-1} h_s\Big| \to^p 0 \tag{23}$$

$$\max_{t=R,\dots,T-1} \Big|(t-1)^{a-1}\sum_{s=1}^{t-1} h_s^*\Big| \to^{p^*} 0 \tag{24}$$

$$\max_{t=R,\dots,T-1} (t-1)^{a}|\hat{\beta}_t - \beta_0| \to^p 0 \tag{25}$$

and

$$\max_{t=R,\dots,T-1} (t-1)^{a}|\hat{\beta}_t^* - \beta_0^*| \to^{p^*} 0. \tag{26}$$

The proofs of (23) and (25) follow the same arguments as West (1996) with minor tweaks as in Calhoun (2015, Lemma A.2) and are omitted. Note that (25) and (26) are refinements of (14) and (15); (14) and (15) establish basic consistency results using standard arguments, and these results are used heavily in the other proofs, but (25) and (26) strengthen those results by adding rate of convergence conditions.

Proof of (24). We will present this proof under the assumption that h_t is univariate to reduce the notational clutter. Otherwise the argument holds element-by-element.

Let δ be a positive number less than $1/2 - a$ and define $H_i^* = \sum_{t=K_{i-1}+1}^{K_i} h_t^*/t^{1-\delta}$, so

$$\max_{t=R,\dots,T-1} \Big| \tfrac{1}{t-1} \sum_{s=1}^{t-1} h_t^* \Big| \leq R^{-\delta} \max_{j=j_R^*,\dots,J} \Big| \sum_{i=1}^{j} H_i^* \Big| + R^{-\delta} \max_{t=R,\dots,T-1} \Big| \sum_{s=K_{j_{t-1}^*}+1}^{t-1} h_s^*/(t-1)^{1-\delta} \Big|,$$

where j_s^* is defined to be the index of the block containing observation s of the bootstrap sequence. (So, for example, $j_1^* = 1$.) Now observe that $\{H_i^*, \mathscr{H}_i^*\}$ is a martingale difference sequence, so the maximal inequality implies that

$$\Pr^* \Big[\max_{j=j_R^*,\dots,J} \Big| \sum_{i=1}^{j} H_i^* \Big| > \epsilon \Big] \leq (1/\epsilon^2) \sum_{i=1}^{J} \mathrm{E}^*(H_i^{*2} \mid \mathscr{H}_{i-1}^*).$$

By definition

$$\begin{aligned}\mathrm{E}^*(H_i^{*2} \mid \mathscr{H}_{i-1}^*) &= \tfrac{1}{T-1} \sum_{u=0}^{T-2} \Big[\sum_{t=1}^{\ell} h_{u+t}(\beta_0^*)/(K_{i-1}+t)^{1-\delta} \Big]^2 \\ &= \tfrac{1}{T-1} \sum_{u=0}^{T-2} \Big[\sum_{t=1}^{\ell} (h_{u+t} + h_{u+t}(\beta_0^*) - h_{u+t})/(K_{i-1}+t)^{1-\delta} \Big]^2.\end{aligned}$$

Since $R^{-\delta} \to 0$, to prove (24) it suffices to show that

$$\tfrac{1}{T-1} \sum_{u=0}^{T-2} \sum_{i=1}^{J} \Big[\sum_{t=1}^{\ell_i} h_{u+t}/(K_{i-1}+t)^{1-a-\delta} \Big]^2 = O_p(1) \tag{27}$$

$$\tfrac{1}{T-1} \sum_{u=0}^{T-2} \sum_{i=1}^{J} \sum_{t=1}^{\ell_i} \big[(h_{u+t}(\beta_0^*) - h_{u+t})/(K_{i-1}+t)^{1-a-\delta} \big]^2 = O_p(1) \tag{28}$$

and

$$\max_{t=R,\dots,T-1} \Big| \sum_{s=K_{j_{t-1}^*}+1}^{t-1} h_s^*/(t-1)^{1-a-\delta} \Big| = O_{p^*}(1). \tag{29}$$

As in Calhoun (2015), our assumptions ensure that h_t is an L_2-mixingale of size $-1/2$. And if c_t and ζ_j denote its mixingale constants and coefficients, $h_t/t^{1-a-\delta}$ is also an L_2-mixingale of size $-1/2$ with constants $c_t/t^{1-a-\delta}$.

For (29), we have

$$\begin{aligned}
&\mathrm{E}^*\Big(\max_{t=R,\dots,T-1}\Big|\sum_{s=K_{j_t^*-1}+1}^{t} h_s^*/(t-1)^{1-a-\delta}\Big|\Big)^2 \\
&\qquad\le \mathrm{E}^*\sum_{i=1}^{J}\max_{t=K_{i-1}+1,\dots,K_i}\Big|\sum_{s=K_{i-1}+1}^{t} h_s^*/(t-1)^{1-a-\delta}\Big|^2 \\
&\qquad= O_p\Big(\tfrac{1}{T-1}\sum_{i=1}^{J}\sum_{u=0}^{T-2}\max_{t=K_{i-1}+1,\dots,K_i}\Big|\sum_{s=K_{i-1}+1}^{t} h_{s+u}/(t-1)^{1-a-\delta}\Big|^2 \\
&\qquad\quad+\tfrac{1}{T-1}\sum_{i=1}^{J}\sum_{u=0}^{T-2}\max_{t=K_{i-1}+1,\dots,K_i}\Big|\sum_{s=K_{i-1}+1}^{t}(h_{s+u}(\beta_0^*)-h_{s+u})/(t-1)^{1-a-\delta}\Big|^2\Big).
\end{aligned}$$

McLeish's (1975) maximal inequality for mixingales implies that

$$\mathrm{E}\max_{t=K_{i-1}+1,\dots,K_i}\Big|\sum_{s=K_{i-1}+1}^{t} h_{s+u}/(t-1)^{1-a-\delta}\Big|^2 \le \mathrm{E}\Big|\sum_{s=K_{i-1}+1}^{K_i} h_{s+u}/(s-1)^{1-a-\delta}\Big|^2.$$

Moreover,

$$\begin{aligned}
&\tfrac{1}{T-1}\sum_{i=1}^{J}\sum_{u=0}^{T-2}\max_{t=K_{i-1}+1,\dots,K_i}\Big|\sum_{s=K_{i-1}+1}^{t}(h_{s+u}(\beta_0^*)-h_{s+u})/(t-1)^{1-a-\delta}\Big|^2 \\
&\qquad\qquad\le \tfrac{1}{T-1}\sum_{i=1}^{J}\sum_{u=0}^{T-2}\sum_{s=K_{i-1}+1}^{K_i}\big[(h_{s+u}(\beta_0^*)-h_{s+u})/(t-1)^{1-a-\delta}\big]^2,
\end{aligned}$$

so the net result is that (29) holds whenever (27) and (28) do.

We'll prove (27) first. Using McLeish's (1975) maximal inequality (again) implies that

$$\begin{aligned}
&\mathrm{E}\Big|\tfrac{1}{T-1}\sum_{u=0}^{T-2}\sum_{i=1}^{J}\Big[\sum_{t=1}^{\ell} h_{u+t}/(K_{i-1}+t)^{1-a-\delta}\Big]^2\Big| \\
&\qquad= \tfrac{1}{T-1}\sum_{u=0}^{T-2}\sum_{i=1}^{J}\mathrm{E}\Big[\sum_{t=1}^{\ell} h_{u+t}(\beta_0)/(K_{i-1}+t)^{1-a-\delta}\Big]^2 \\
&\qquad= O\big(\tfrac{1}{T-1}\big)\sum_{u=0}^{T-2}\sum_{i=1}^{J}\sum_{t=1}^{\ell}(K_{i-1}+t)^{2a+2\delta-2} \\
&\qquad= O(1)\sum_{t=1}^{T-1} t^{2a+2\delta-2}.
\end{aligned}$$

Since δ was chosen to ensure that $2a+2\delta-2<-1$, this last summation is finite as required.

For (28), expanding $h_{u+t}(\beta_0^*)$ around β_0 gives

$$\begin{aligned}
&\tfrac{1}{T-1}\sum_{u=0}^{T-2}\sum_{i=1}^{J}\Big(\sum_{t=1}^{\ell}(h_{u+t}(\beta_0^*)-h_{u+t}(\beta_0))/(K_{i-1}+t)^{1-a-\delta}\Big)^2 \\
&\quad= \tfrac{1}{T-1}\sum_{u=0}^{T-2}\sum_{i=1}^{J}\Big(\sum_{t=1}^{\ell}\nabla h_{u+t}(b_{u+t})\cdot(\beta_0^*-\beta_0)/(K_{i-1}+t)^{1-a-\delta}\Big)^2 \\
&\quad= (\beta_0^*-\beta_0)'\Big[\tfrac{1}{T-1}\sum_{u=0}^{T-2}\sum_{i=1}^{J}\sum_{s,t=1}^{\ell}\big(\tfrac{1}{(K_{i-1}+s)(K_{i-1}+t)}\big)^{1-a-\delta}\nabla h_{u+t}(b_{u+t})\nabla h_{u+s}(b_{u+s})'\Big](\beta_0^*-\beta_0) \\
&\quad= O_p\big(\tfrac{1}{T^2}\big)\sum_{u=0}^{T-2}\sum_{i=1}^{J}\sum_{s,t=1}^{\ell}\big(\tfrac{1}{(K_{i-1}+s)(K_{i-1}+t)}\big)^{1-a-\delta}\nabla h_{u+t}(b_{u+t})'\nabla h_{u+s}(b_{u+s})
\end{aligned}$$

where each b_{u+t} lies between β_0^* and β_0 almost surely, and so

$$\begin{aligned}
&\Pr\Big[\tfrac{1}{T-1}\sum_{u=0}^{T-2}\sum_{i=1}^{J}\Big(\sum_{t=1}^{\ell}(h_{u+t}(\beta_0^*)-h_{u+t})/(K_{i-1}+t)^{1-a-\delta}\Big)^2>\epsilon\Big] \\
&\quad\le \Pr\Big[\tfrac{1}{T^2}\sum_{u=0}^{T-2}\sum_{i=1}^{J}\sum_{s,t=1}^{\ell}\Big|\big(\tfrac{1}{(K_{i-1}+s)(K_{i-1}+t)}\big)^{1-a-\delta}\nabla h_{u+t}(b_{u+t})'\nabla h_{u+s}(b_{u+s})\Big|1\{\beta_0^*\in N\}>\epsilon\Big] \\
&\qquad+\Pr[\beta_0^*\notin N].
\end{aligned}$$

The second probability, $\Pr[\beta_0^*\notin N]$, converges to zero by Lemma A.1. For the first, we have

$$\begin{aligned}
&\mathrm{E}\,\tfrac{1}{T^2}\sum_{u=0}^{T-2}\sum_{i=1}^{J}\sum_{s,t=1}^{\ell}\Big|\big(\tfrac{1}{(K_{i-1}+s)(K_{i-1}+t)}\big)^{1-a-\delta}\nabla h_{u+t}(b_{u+t})'\nabla h_{u+s}(b_{u+s})\Big|1\{\beta_0^*\in N\} \\
&\quad\le \tfrac{1}{T^2}\sum_{u=0}^{T-2}\sum_{i=1}^{J}\sum_{s,t=1}^{\ell}\mathrm{E}\sup_{\beta\in N}\Big|\big(\tfrac{1}{(K_{i-1}+s)(K_{i-1}+t)}\big)^{1-a-\delta}\nabla h_{u+t}(\beta)'\nabla h_{u+s}(\beta)\Big| \\
&\quad\le O(\tfrac{1}{T^2})\sum_{u=0}^{T-2}\sum_{i=1}^{J}\sum_{s,t=1}^{\ell}\big(\tfrac{1}{(K_{i-1}+s)(K_{i-1}+t)}\big)^{1-a-\delta}\,\mathrm{E}\,|m_{u+t}m_{u+s}| \\
&\quad\le O(\tfrac{1}{T^2})\sum_{u=0}^{T-2}\sum_{i=1}^{J}\sum_{s=1}^{\ell}(K_{i-1}+s)^{2a+2\delta-2}\,\mathrm{E}\,m_{u+s}^2
\end{aligned}$$

where the second inequality holds by assumption and the third follows from repeated application of the Cauchy-Schwarz inequality. Since $\mathrm{E}\,m_{u+s}^2$ is bounded, the large summation is $O(T)$ and this final term converges to zero, completing the proof. □

Proof of (26). Expanding $h_t^*(\hat\beta_t^*)$ around β_0^* gives

$$\hat\beta_t^*-\beta_0^* = \Big(-\sum_{s=1}^{t-1}\nabla h_s^*(b_s^*)\Big)^{-1}\sum_{s=1}^{t-1}h_s^*/(t-1)$$

with each b_s^* a.s. between $\hat\beta_t^*$ and β_0^*. Then

$$\max_{t=R,\dots,T-1}(t-1)^a|\hat\beta_t-\beta_0| \le \max_{t,u=R,\dots,T-1}\Big|\Big[\Big(\sum_{s=1}^{t-1}\nabla h_s^*(b_s^*)\Big)^{-1}-B^*\Big](t-1)^{a-1}\sum_{s=1}^{t-1}h_s^*\Big| + \max_{t=R,\dots,T-1}\Big|B^*(t-1)^{a-1}\sum_{s=1}^{t-1}h_s^*\Big| \quad (30)$$

and both terms converge to zero in (conditional) probability by the previous arguments. □

Lemma A.4. *Under the conditions of Theorem 1,*

$$\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\hat\beta_t^*-\beta_0^*)'\nabla^2 f_{it}^*(b_{it}^*)(\hat\beta_t^*-\beta_0^*) \to^{p^*} 0 \quad (31)$$

and

$$\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}F_t^*\cdot(\hat\beta_t^*-\beta_0^*) = F^*B^*\tfrac{1}{\sqrt{P}}\sum_{t=1}^{T-1}a_t h_t^* + o_{p^*}(1). \quad (32)$$

Proof of (31). We have

$$\begin{aligned}\Pr\Big[\Big|\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\hat\beta_t^*-\beta_0^*)'\nabla^2 f_{it}^*(b_{it}^*)(\hat\beta_t^*-\beta_0^*)\Big|>\epsilon\Big] \\ \le \Pr\Big[1\{\beta_0^*\in N,\hat\beta_t^*\in N \text{ for all } t\}\Big|\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\hat\beta_t^*-\beta_0^*)'\nabla^2 f_{it}^*(b_{it}^*)(\hat\beta_t^*-\beta_0^*)\Big|>\epsilon\Big] \\ +\Pr[\beta_0^*\notin N]+\Pr[\hat\beta_t^*\notin N \text{ for some } t=R,\dots,T-1\}\end{aligned}$$

The second two probabilities on the rhs converge to zero by Lemma A.3 and the random variable inside the first probability is bounded by

$$\begin{aligned}1\{\beta_0^*\in N,\hat\beta_t^*\in N \text{ for all } t\}\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\hat\beta_t^*-\beta_0^*)'\nabla^2 f_{it}^*(b_{it}^*)(\hat\beta_t^*-\beta_0^*) \\ \le \Big(\sup_{t=R,\dots,T-1}|P^{1/4}(\hat\beta_t^*-\beta_0^*)|^2\Big)\tfrac{1}{P}\sum_{t=R}^{T-1}\nabla^2 f_{it}^*(b_{it}^*)1\{\beta_0^*\in N,\hat\beta_t^*\in N\}.\end{aligned}$$

The summation is $O_p(1)$ by assumption and the supremum converges to zero by using Lemma A.3 again. □

Proof of (32). For (32), we have the upper bound

$$\Big|\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\big(F_t^*\cdot(\hat\beta_t^*-\beta_0^*)-F^*B^*a_t h_t^*\big)\Big| \le$$

$$\sup_{s=R,\dots,T-1}|\hat\beta_s^*-\beta_0^*|\,\Big|\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(F_t^*-F^*)\Big|+\Big|F^*\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\big((\hat\beta_t^*-\beta_0^*)-B^*a_t h_t^*\big)\Big|.$$

The first term converges in conditional probability to zero by Lemma A.3.

For the second, expand each $\sum_{s=1}^t \nabla q_s^*(\hat\beta_t^*)$ around β_0^* to get

$$\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\hat\beta_t^*-\beta_0^*)=\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\Big[\tfrac{1}{t}\sum_{s=1}^{t}\nabla^2 q_s^*(b_t^*)\Big]^{-1}\tfrac{1}{t}\sum_{s=1}^{t}h_s^*$$

where b_t^* is between $\hat\beta_t^*$ and β_0^*. Then

$$\Big|\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\big((\hat\beta_t^*-\beta_0^*)-B^*a_t h_t^*\big)\Big| \le \sup_{t=R,\dots,T-1}\Big\|\Big[\tfrac{1}{t}\sum_{s=1}^{t}\nabla^2 q_s^*(b_t)\Big]^{-1}-B^*\Big\|\Big|\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\tfrac{1}{t}\sum_{s=1}^{t}h_s^*\Big|$$

$$= O_{p^*}(1)\sup_{t=R,\dots,T-1}\Big\|\Big[\tfrac{1}{t}\sum_{s=1}^{t}\nabla^2 q_s^*(b_t)\Big]^{-1}-B^*\Big\|$$

by Lemma A.3 and the supremum converges to zero in probability by Lemma A.3 as well. □

Lemma A.5. *Under the conditions of Theorem 1,* (11) *holds.*

Proof. We will use arguments very similar to Calhoun (2014). Define

$$\zeta_{st}^*=\gamma_1'(f_t^*-\mathrm{E}^* f_t^*)+a_s\gamma_2' h_t^*$$

where γ_1 and γ_2 are arbitrary nonzero vectors, and also define

$$z_j^*=\tfrac{1}{\sqrt{P}}\sum_{s=(j-1)\ell+1}^{j\ell}\zeta_s^*$$

and

$$v^{*2}=J\,\mathrm{var}^*(z_j^*)$$

where $\gamma=(\gamma_1',\gamma_2')'$. By construction, $\mathrm{E}^* h_t^*=0$ almost surely, so $\mathrm{E}(z_j\mid \mathscr{H}_{j-1}^*)=0$ almost surely and $\{z_j^*,\mathscr{H}_j^*\}$ is a martingale difference sequence.

From the MDS property, we have

$$\sum_{j=1}^{J} z_j^*/\sqrt{v^*}\to^d N(0,1)$$

as long as the following properties hold:

$$\sum_{j=1}^{J} \mathrm{E}^*(z_j^{*2} 1\{z_j^{*2} > \epsilon\} \mid \mathscr{H}_{j-1}^*) \to^p 0 \tag{33}$$

and

$$\Pr^* \Big[\Big| \sum_{j=1}^{J} z_j^{*2}/v^{*2} - 1 \Big| > \epsilon \Big] \to^p 0. \tag{34}$$

For (34), we have the usual bound

$$\Pr^* \Big[\Big| \sum_{j=1}^{J} z_j^{*2}/v^{*2} - 1 \Big| > \varepsilon \Big] \leq \Pr^* \Big[1\{\beta_0^* \in N\} \Big| \sum_{j=1}^{J} z_j^{*2}/v^{*2} - 1 \Big| > \epsilon \Big] + \Pr^*[\beta_0^* \notin N]$$

and we can rewrite the summation in the first term as

$$1\{\beta_0^* \in N\} \Big(\sum_{j=1}^{J} z_j^{*2}/v^{*2} - 1 \Big) = \sum_{j=1}^{J} \Big(1\{\beta_0^* \in N\} z_j^{*2}/v^{*2} - \mathrm{E}(1\{\beta_0^* \in N\} z_j^{*2}/v^{*2} \mid \mathscr{H}_{j-1}^*) \Big).$$

This term is the sum of a uniformly integrable martingale difference sequence and satisfies the LLN (i.e. Davidson's, 1994, Theorem 19.7), and so it converges in (conditional) probability to zero. The second term converges in probability to zero by consistency of β_0^* (Lemma A.1).

Similarly, (33) holds if

$$1\{\beta_0^* \in N\} \sum_{j=1}^{J} \mathrm{E}^*(z_j^{*2} 1\{z_j^{*2} > \epsilon\} \mid \mathscr{H}_{j-1}^*) \to 0,$$

which holds by uniform integrability of $1\{\beta_0^* \in N\} z_j^{*2}$.

Finally, since the variance of the bootstrapped statistic can be rewritten as a HAC variance estimator,

$$v^{*2} \to^p \gamma_1' S_{ff} \gamma_1 + 2\lambda_{fh}(\gamma_2' S_{fh}' \gamma_1) + \lambda_{hh} \gamma_2' S_{hh} \gamma_2$$

holds by Theorem 2.2 of de Jong and Davidson (2000), using West's (1996) arguments to handle the a_s terms. □

References

J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman. Julia: A fast dynamic language for technical computing, September 2012. arXiv:1209.5145.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing, November 2014. arXiv:1411.1607.

G. Calhoun. *dbframe: An R to SQL interface*, 2010.

G. Calhoun. Block bootstrap consistency under weak assumptions. Working Paper 11017, Iowa State University, 2014.

G. Calhoun. An asymptotically normal out-of-sample test based on mixed estimation windows. Available at http://www.econ.iastate.edu/~gcalhoun, 2015.

T. E. Clark and M. W. McCracken. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110, Nov. 2001.

T. E. Clark and M. W. McCracken. Evaluating direct multistep forecasts. *Econometric Reviews*, 24(4):369, 2005.

T. E. Clark and K. D. West. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1):155–186, 2006.

T. E. Clark and K. D. West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311, May 2007.

V. Corradi and N. R. Swanson. Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. *International Economic Review*, 48: 67–109, 2007.

D. B. Dahl. *xtable: Export tables to LaTeX or HTML*, 2009. URL http://CRAN.R-project.org/package=xtable. R package version 1.5-6.

R. M. de Jong and J. Davidson. Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, 68(2):407–423, 2000.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263, 1995.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74(6): 1545–1578, 2006.

A. Goyal and I. Welch. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508, 2008.

P. R. Hansen. A test for superior predictive ability. *Journal of Business and Economic Statistics*, 23(4):365–380, 2005.

L. Kilian. Exchange rates and monetary fundamentals: What do we learn from long-horizon regressions? *Journal of Applied Econometrics*, 14(5):491–510, 1999.

H. R. Kunsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

R. Y. Liu and K. Singh. Moving blocks Jackknife and Bootsrap capture weak dependence. In R. LePage and L. Billard, editors, *Exploring the Limits of Bootstrap*, pages 225–248. John Wiley, New York, 1992.

M. W. McCracken. Robust out-of-sample inference. *Journal of Econometrics*, 99(2): 195–223, 2000.

M. W. McCracken. Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics*, 140(2):719–752, Oct. 2007.

D. McLeish. A maximal inequality and dependent strong laws. *The Annals of Probability*, 3(5):829–839, Oct. 1975.

W. K. Newey and D. L. McFadden. Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

D. Nordman. A note on the stationary bootstrap's variance. *The Annals of Statistics*, 37 (1):359–370, 2009.

A. J. Patton, D. N. Politis, and H. White. Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. *Econometric Reviews*, 28(4): 372–375, 2009.

D. N. Politis and J. P. Romano. A circular block resampling procedure for stationary data. In R. Page and R. LePage, editors, *Exploring the Limits of Bootstrap*, pages 263–270. John Wiley, New York, 1992.

D. N. Politis and J. P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994.

D. N. Politis and H. White. Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, 23(1):53–70, 2004.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL http://www.R-project.org/.

J. P. Romano and M. Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

J. P. Romano and M. Wolf. Improved nonparametric confidence intervals in time series regressions. *Journal of Nonparametric Statistics*, 18(2):199–214, 2006.

J. P. Romano, A. M. Shaikh, and M. Wolf. Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2):404–447, 2008.

K. D. West. Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–1084, Sept. 1996.

K. D. West and M. W. McCracken. Regression-based tests of predictive ability. *International Economic Review*, 39(4):817–840, 1998.

H. White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.