3-30-2014

Out-of-sample comparisons of overfit models

Gray Calhoun
Iowa State University, gcalhoun@iastate.edu

Out-of-sample comparisons of overfit models

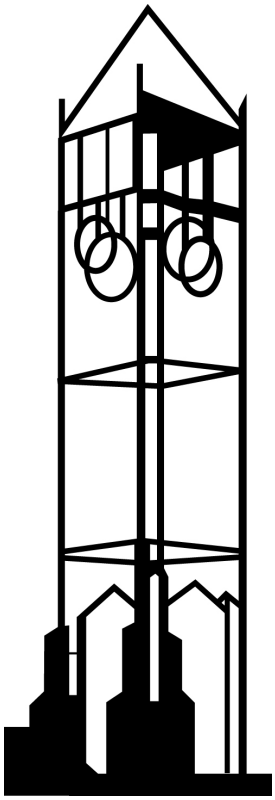
Abstract
This paper uses dimension asymptotics to study why overfit linear regression models should be compared out-of-sample; we let the number of predictors used by the larger model increase with the number of observations so that their ratio remains uniformly positive. Our analysis gives a theoretical motivation for using out-of-sample (OOS) comparisons: the DMW OOS test allows a forecaster to conduct inference about the expected future accuracy of his or her models when one or both is overfit. We show analytically and through Monte Carlo that standard full-sample test statistics can not test hypotheses about this performance. Our paper also shows that popular test and training sample sizes may give misleading results if researchers are concerned about overfit. We show that P 2 /T must converge to zero for theDMW test to give valid inference about the expected forecast accuracy, otherwise the test measures the accuracy of the estimates constructed using only the training sample. In empirical research, P is typically much larger than this. Our simulations indicate that using large values of P with the DMW test gives undersized tests with low power, so this practice may favor simple benchmark models too much.

Keywords
generalization error, forecasting, model selection, t-test, dimension asymptotics

Disciplines
Economics

Out-Of-Sample Comparisons of Overfit Models

Gray Calhoun



Working Paper No. 11002
March 2014

IOWA STATE UNIVERSITY
Department of Economics
Ames, Iowa, 50011-1070

Out-of-Sample Comparisons of Overfit Models

Gray Calhoun*
Iowa State University

March 30, 2014

Abstract

This paper uses dimension asymptotics to study why overfit linear regression models should be compared out-of-sample; we let the number of predictors used by the larger model increase with the number of observations so that their ratio remains uniformly positive. Our analysis gives a theoretical motivation for using out-of-sample (OOS) comparisons: the DMW OOS test allows a forecaster to conduct inference about the expected future accuracy of his or her models when one or both is overfit. We show analytically and through Monte Carlo that standard full-sample test statistics can not test hypotheses about this performance. Our paper also shows that popular test and training sample sizes may give misleading results if researchers are concerned about overfit. We show that P^2/T must converge to zero for the DMW test to give valid inference about the expected forecast accuracy, otherwise the test measures the accuracy of the estimates constructed using only the training sample. In empirical research, P is typically much larger than this. Our simulations indicate that using large values of P with the DMW test gives undersized tests with low power, so this practice may favor simple benchmark models too much.
JEL Classification: C12, C22, C52, C53
Keywords: Generalization Error, Forecasting, Model Selection, t-test, Dimension Asymptotics

*email: gcalhoun@iastate.edu. I would like to thank Julian Betts, Helle Bunzel, Stéphane Bonhomme, Marjorie Flavin, Nir Jaimovich, Lutz Kilian, Ivana Komunjer, Michael McCracken, Seth Pruitt, Ross Starr, Jim Stock, Yixiao Sun, Allan Timmermann, Hal White, several anonymous referees, participants at the 2010 Midwest Economics Association Annual Meetings, the 2010 International Symposium on Forecasting, the 2010 Joint Statistical Meetings, the 2011 NBER-NSF Time-Series Conference, and in many seminars at UCSD, and especially Graham Elliott for their valuable suggestions, feedback and advice in writing this paper. I would 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

Consider two sequences of length P of prediction errors, the result of forecasting the same variable with two different estimated models. Both models are estimated with R observations, collectively called the *estimation window*, and are used to forecast an additional P observations, called the *test sample*. There are T observations in all, and $R+P=T$. This paper introduces a new limit theory for statistics constructed from these prediction errors designed to approximate the behavior of the statistics when one of the models is overfit. In doing so, we provide a theoretical justification for forecasters to use OOS instead of in-sample comparisons: the DMW OOS test[1] allows a forecaster to conduct inference about the expected future accuracy of his or her models when one or both is overfit. We show analytically and through Monte Carlo simulations that standard full-sample test statistics can not test hypotheses about this performance.

Our paper also shows that popular test and training sample sizes may give misleading results if researchers are concerned about overfit. We show that P^2/T must converge to zero for the DMW test to give valid inference about the expected forecast accuracy, otherwise the test measures the accuracy of the estimates constructed using only the training sample. In empirical research, P is typically much larger than this. Our simulations indicate that using large values of P with the DMW test gives undersized tests with low power, so this practice may favor simple benchmark models too much. Existing corrections, proposed by Clark and McCracken (2001, 2005), McCracken (2007) and Clark and West (2006, 2007), seem to overcorrect for this problem, though, and reject too often when the benchmark model is more accurate.

Although OOS comparisons have been popular in Macroeconomics and Finance since Meese and Rogoff's (1983) seminal study of exchange rate models, it has been unclear from a theoretical perspective whether or not the statistics are useful. Empirical researchers often cite "overfit" or "instability" as reasons for using OOS comparisons, as in Stock and Watson (2003), but neither term is precisely defined or formalized. Compounding this problem, the asymptotic distributions of these statistics are derived under conditions that rule out either instability or overfit and allow a researcher to use a conventional in-sample comparison—a variation of the F-test, for example. As Inoue and Kilian (2004) argue, the statistics themselves are designed to test hypotheses that can be tested by these in-sample statistics. For example, Diebold and Mariano (1995) and West (1996) derive the limiting distributions of many popular OOS test statistics under conditions that would justify these full-sample tests. Much of the subsequent research by McCracken (2000, 2007), Chao et al. (2001), Clark and McCracken (2001, 2005), Corradi and Swanson (2002, 2004), Clark and West (2006, 2007), Anatolyev (2007), and others relaxes several of Diebold and Mariano's and West's assumptions, but maintains the stationarity and dependence conditions that permit in-sample comparisons (see West,

[1] In this paper, we will refer to the basic OOS t-test studied by Diebold and Mariano (1995) and West (1996) as the DMW test.

2006, for a review of this literature).[2] Giacomini and White (2006) and Giacomini and Rossi (2009, 2010) are exceptions. Instead of focusing on hypotheses that can be tested by in-sample comparisons, Giacomini and White (2006) derive an OOS test for the null hypothesis that the difference between two models' OOS forecasting performance is unpredictable, a martingale difference sequence (MDS); Giacomini and Rossi (2009) test whether the OOS forecasting performance of a model suffers from a breakdown relative to its in-sample performance; and Giacomini and Rossi (2010) test whether the forecasting performance is stable. However, those papers focus on a particular OOS estimation strategy and do not address why OOS comparisons might be useful as a general strategy.

Since in-sample and OOS statistics require similar assumptions and test similar hypotheses, one might expect that they would give similar results. They do not. In-sample analyses tend to support more complicated theoretical models and OOS analyses support simple benchmarks, as seen in Meese and Rogoff (1983), Stock and Watson (2003), and Goyal and Welch (2008). Since these different approaches strongly influence the outcome of research, it is important to know when each is appropriate. The explanations in favor of OOS comparisons claim that they should be more robust to unmodeled instability (Clark and McCracken, 2005; Giacomini and White, 2006; Giacomini and Rossi, 2009, 2010) or to overfit (McCracken, 1998; Clark, 2004). Both explanations presume that the in-sample comparison is invalid and the OOS comparisons are more reliable. Of course, as Inoue and Kilian (2004, 2006) point out, both in-sample and OOS methods could be valid, but the OOS methods could have lower power.

In this paper, we study the "overfit" possibility and leave "instability" to future research. This paper uses dimension asymptotics to study the behavior of OOS comparisons when at least one of the models is overfit—the number of its regressors increases with the number of observations so that their ratio remains positive. Although overfit is sometimes used to describe the situation where a forecaster chooses from many different models, i.e. *data-mining* or *data-snooping*, we view these as separate issues. Procedures that account for the presence of many models have been, and are continuing to be, developed (see, for example, White, 2000; Hansen, 2005; Romano and Wolf, 2005; Hsu et al., 2010; Clark and McCracken, 2012b), but it is unclear whether those procedures should themselves use in-sample or OOS comparisons. Understanding the difference between in-sample and OOS comparisons in the context of a simple comparison between two models is necessary before resolving any new issues that arise with multiple comparisons. Moreover, the empirical research that motivates this paper uses *pseudo* OOS comparisons and not true OOS comparisons. Even if a true OOS comparison could account for some forms of data-snooping better than White's (2000) BRC or its extensions, in-sample and pseudo OOS comparisons would both be affected by the data-snooping, a point also made by Inoue and Kilian (2004).

[2]Like us, Anatolyev (2007) allows the number of regressors to increase with T. But in that paper, the number of regressors increases slowly enough that the OLS coefficients are consistent and asymptotically normal.

We focus on linear regression models estimated with a fixed window for simplicity, but our basic conclusions should be true for other models and estimation strategies as well. Under this asymptotic theory, where the number of regressors K increases with T so that the limit of K/T is positive and less than one, the OLS coefficient estimator is no longer consistent or asymptotically normal (Huber, 1973) and has positive variance in the limit. We show that, even so, the usual OOS average is asymptotically normal and can consistently estimate the difference between the models' generalization error, the expected loss in the future conditional on the on the data available in period T.[3] Under these asymptotics, the generalization error does not converge to the expected performance of the pseudotrue models, so existing in-sample and OOS comparisons measure different quantities and should be expected to give different results for reasons beyond simple size and power comparisons. Under our limit theory, the model that is closer to the true DGP in population can forecast worse. In such a situation, a standard in-sample comparison would correctly reject the null hypothesis that the benchmark is true, and an OOS comparison would correctly fail to reject the null that the benchmark is more accurate.[4] Also note that, in this situation, a model that performs well in-sample can perform badly out-of-sample even if there are no structural breaks or other forms of instability. Researchers have argued that a breakdown of in-sample forecasting ability indicates a structural break (see Bossaerts and Hillion, 1999, and Stock and Watson, 2003, among others), but we show that this breakdown can be caused by overfit as well.

It is important to realize that we are not *advocating* linear regression for highly overparametrized models. If K is very large relative to T, researchers will often want to use some form of shrinkage forecast, for example the LASSO (Tibshirani, 1996) or a factor model (Stock and Watson, 2002; Bai and Ng, 2002). Our main focus is practical settings where K is *moderately* large and it is not clear whether overfit will dominate the results. In these settings it is crucial to understand the behavior of different evaluation criteria when K is large, in case overfit does turn out to be a concern, and to have reliable methods for estimating these overfit models' performance. This is the goal of this paper.

Our theoretical results partially justify OOS comparisons when researchers want to choose a model for forecasting. Although there has been little emphasis on hypothesis testing in this setting, testing is usually appropriate: there is usually a familiar benchmark model in place, and the cost of incorrectly abandoning the benchmark for a less accurate alternative model is higher than the cost of incorrectly failing to switch to a more accurate alternative. We show that the DMW test lets the forecaster control the probability of the first error, just as with conventional hypothesis testing.

But we also identify new practical limitations for applying the DMW test to overfit

[3] See, for example, Hastie et al. (2008) for a discussion of generalization error.

[4] In a pair of papers similar to ours, Clark and McCracken (2012a,b) study in-sample and OOS tests that the larger model has nonzero coefficients that are too close to zero to expect the model to forecast more accurately. Like this paper, they argue that the larger model can be true but less accurate. However, they focus on an aspect of the DGP that makes this phenomenon likely, while we focus on the coefficient estimates that produce less accurate forecasts. Moreover, the implications of our asymptotic theories are different and their papers do not provide reasons to do OOS comparisons.

models. Since the models' coefficients are imprecisely estimated in the limit, the test sample must be small enough that the model estimated over the training sample is similar to the one that will be estimated over the full sample. In particular, $P/T \to 0$ is required for consistent estimation of the difference in the two models' performance, and $P^2/T \to 0$ is required for valid confidence intervals and inference. For larger P, the OOS comparisons remain asymptotically normal, but are centered on the forecasting performance associated with the period R estimates. In practice, researchers typically use large values of P, so these studies may be too pessimistic about their models' future accuracy if they use the DMW test. Section 3.1 lays out the asymptotic behavior of the DMW test under this limit theory.

Popular in-sample tests and model selection criteria, like the Wald test and the AIC, do not help forecasters in this setting. We show that these statistics do not select the more accurate model when choosing between overfit models. For many DGPs the Wald test will choose the larger model over a small benchmark with probability much greater than its nominal size, regardless of which model is more accurate, and the AIC behaves similarly. The BIC, however, chooses the benchmark model with probability approaching one even when the *alternative* model is more accurate.[5] This result holds even though modifications of the F-test are valid under this asymptotic theory, as shown by Boos and Brownie (1995), Akritas and Arnold (2000), Akritas and Papadatos (2004), Calhoun (2011), and Anatolyev (2012), among others.[6] Moreover, under this asymptotic theory, many recent OOS test statistics, such as those derived by Clark and McCracken (2001, 2005), McCracken (2007), and Clark and West (2006, 2007) should have the same problems as in-sample tests.[7] These tests are also designed to reject the benchmark when the alternative model is true, and so they may reject too often when the benchmark is misspecified but more accurate. Obviously, since the distribution of these statistics converges to the normal when $P/T \to 0$ (with the number of regressors fixed), these statistics behave like the DMW test when P is small, but should overreject the benchmark when P is large. Section 3.2 presents our theoretical results for full-sample statistics and Section 4 presents Monte Carlo evidence to support these claims.

Finally, this paper introduces a new method of proof for OOS statistics. We use a coupling argument (Berbee's Lemma, 1979) to show that sequences of OOS loss behave

[5] Our result holds for a broad class of full-sample statistics, but there may be other potential statistics that mimic the OOS test and remain valid. Exploring such statistics is left for future research.

[6] Also see Efron (1986, 2004) for a discussion of naive in-sample loss comparisons.

[7] Our theoretical results apply directly to McCracken's (2007) OOS t-test, since it simply proposes more liberal critical values for the same test statistic that we study. Since Clark and West (2006, 2007) use a finite length estimation window, our asymptotics are incompatible with theirs and prevent us from studying their test directly, as well as Giacomini and White's (2006) and other tests based on Giacomini and White's (2006) asymptotics. But Clark and West's (2006; 2007) test can be viewed as a stochastic adjustment to the critical values of the usual OOS-t test, so our conclusions should apply informally as well. Specifically, we show that the DMW test rejects with probability equal to nominal size when the estimated benchmark model is expected to be more accurate, so more liberal critical values result in overrejection.

like mixingales when the underlying series are absolutely regular, even if the forecasts depend on non-convergent estimators. Moreover, we also show that transformations of these processes behave like mixingales after they are appropriately recentered, so many arguments used to prove asymptotic results for Near Epoch Dependent (NED) functions of mixing processes can be used for these OOS processes with only slight modification.

The rest of the paper proceeds as follows. Section 2 introduces our notation and assumptions. Section 3 gives the main theoretical results for the DMW OOS test, shows that standard in-sample tests reject the benchmark model too often when it is misspecified but more accurate than the alternative, and shows that standard model selection methods can run into similar problems. Section 3 also presents an example DGP that illustrates these results. Section 4 presents a Monte Carlo study supporting our theoretical results. Section 5 applies the OOS statistic to Goyal and Welch's (2008) dataset for equity premium prediction, and Section 6 concludes. Proofs and supporting results are listed in the Appendix.

2 Setup and assumptions

The first part of this section will describe the environment in detail and set up our models and notation. The second part lists the assumptions underlying our theoretical results.

2.1 Notation and forecasting environment

We assume the following forecasting environment. There are two competing linear models that give forecasts for the target, y_{t+h}:

$$y_{t+h} = x_{1t}'\theta_1 + \varepsilon_{1,t+h}, \quad \text{and} \quad y_{t+h} = x_{2t}'\theta_2 + \varepsilon_{2,t+h};$$

$t = 1, \ldots, T-h$, h is the forecast horizon and the variables y_t, x_{1t}, and x_{2t} are all known in period t. The coefficients θ_1 and θ_2 minimize the population Mean Square Error, so

$$\theta_i = \arg\min_{\theta} \sum_{t=1}^{T-h} \mathrm{E}(y_{t+h} - x_{it}'\theta)^2,$$

making $\varepsilon_{i,t+h}$ uncorrelated with $x_{i,t}$; $\varepsilon_{i,t+h}$ can exhibit serial correlation so both of the linear models may be misspecified. Let $\mathscr{F}_t = \sigma(y_1, x_1, \ldots, y_t, x_t)$ be the information set available in period t, with x_t the vector of all stochastic elements of x_{1t} and x_{2t} after removing duplicates, and let E_t and var_t denote the conditional mean and variance given $\mathscr{F}_t$. The first model uses K_1 regressors, and the second uses K_2. Without loss of generality, assume that $K_1 \leq K_2$. At least one of the models is overfit, which we represent asymptotically by letting K_2 grow with T quickly enough that $\lim K_2/T$ is positive; K_1 may grow with T as well. Since the models change with T, a stochastic array underlies all of our asymptotic theory, but we suppress that notation to simplify the presentation.

In the settings we are interested in, a forecaster observes the data (y_t, x_t) for periods 1 through T and divides these observations into an estimation sample of the first R observations and a test sample of the remaining P observations. The forecaster then compares the models' performance over the test sample, which entails constructing two sequences of forecasts with a fixed-window estimation strategy,

$$\hat{y}_{i,t+h} = x_{it}'\hat{\theta}_{it}, \qquad \text{for } i = 1,2;\ t = R+1,\ldots,T-h,$$

where

$$\hat{\theta}_{it} = \Big(\sum_{s=1}^{R-h} x_{is}x_{is}'\Big)^{-1}\sum_{s=1}^{R-h} x_{is}y_{s+h}, \qquad \text{for } i = 1,2;\ t = R+1,\ldots,T-h.^{8}$$

The models are then compared by their forecast performance over the test sample. There are many statistics that have been considered in the literature, but we focus on perhaps the most natural, the DMW OOS-t test (Diebold and Mariano, 1995; West, 1996).[9] This statistic is based on the difference in the models' loss over the test sample, $\bar{D}_R$, defined as

$$\bar{D}_R \equiv P^{-1}\sum_{t=R+1}^{T-h} D_t$$

where

$$D_t = L(y_{t+h} - x_{1t}'\hat{\theta}_{1t}) - L(y_{t+h} - x_{2t}'\hat{\theta}_{2t}),$$

and L is a known loss function. The OOS-t test is defined as $\sqrt{P}\bar{D}_R/\hat{\sigma}$, where $\hat{\sigma}^2$ is an estimator of the asymptotic variance of $\bar{D}_R$. (Possibly a Heteroskedasticity- and Autocorrelation-Consistent, or HAC, estimator.)

Statistics like OOS-t have a long history in empirical economics because they capture an intuitive idea of model fit: that a good model should be able to forecast well on new data. Meese and Rogoff (1983) use such a statistic to study exchange rate models and find that none of the models that existed at the time of their study outperform a random walk. This finding has been remarkably durable and has spawned an enormous literature; see Mark (1995), Kilian and Taylor (2003), Cheung et al. (2005), Engel and West (2005), Rossi (2005), and Bacchetta et al. (2010), among many others. Research in financial markets has found a similar pattern, e.g. Bossaerts and Hillion (1999), Goyal and Welch (2008, 2003), and Timmermann (2008).

Most theoretical research on these statistics, such as Diebold and Mariano (1995), West (1996), and McCracken (2007), has focused on using the OOS-t statistic to test hypotheses about the pseudotrue values θ_1 and θ_2. In particular, that research focuses on testing the null hypothesis that

$$\mathrm{E}\,L(y_{t+h} - x_{1t}'\theta_1) = \mathrm{E}\,L(y_{t+h} - x_{2t}'\theta_2).$$

[8] It may not be clear why we are using the index t in $\hat{\theta}_{it}$, since $\hat{\theta}_{it} = \hat{\theta}_{iR}$ almost surely for all $t \le T-h$. But $\hat{\theta}_{it}$ will be defined for $t > T-h$ soon and will not equal $\hat{\theta}_{iR}$ for those values of t.

[9] The core insights of our paper apply to other OOS statistics as well.

But the population quantities in this equation do not determine which model is more accurate in practice. The models' accuracy will also depend on the specific estimates of θ_1 and θ_2 used to produce the forecasts.

When the forecaster will use one of the models to make a number of predictions (call it Q) in the future, the quantity of interest becomes

$$\mathrm{E}_T \bar{D}_T = Q^{-1} \sum_{t=T+1}^{T+Q} \mathrm{E}_T D_t,$$

where D_t is defined as before,

$$D_t = L(y_{t+h} - x_{1t}'\hat{\theta}_{1t}) - L(y_{t+h} - x_{2t}'\hat{\theta}_{2t}),$$

but now uses the full-sample estimates of the models' parameters,

$$\hat{\theta}_{it} = \Big(\sum_{s=1}^{T-h} x_{is}x_{is}'\Big)^{-1} \sum_{s=1}^{T-h} x_{is}y_{s+h}, \qquad \text{for } i = 1,2;\ t = T+1,\ldots,T+Q.$$

If $\mathrm{E}_T \bar{D}_T$ is positive, the second model is expected to forecast better than the first over the next Q periods, and if $\mathrm{E}_T \bar{D}_T$ is negative then the first model is better. We use a conditional expectation because the coefficient estimates in $\bar{D}_T$ are stochastic but known in period T, and their values will determine the performance of the two models.

Under conventional fixed-K asymptotic theory, $\mathrm{E}_T \bar{D}_T$ would converge in probability to the difference in the expected loss associated with the pseudotrue models,[10]

$$\mathrm{E}\, L(y_{t+h} - x_{1t}'\theta_1) - \mathrm{E}\, L(y_{t+h} - x_{2t}'\theta_2). \tag{1}$$

But if K_2 increases with T these quantities can have different limits. For a simple example, assume squared-error loss, let $x_{1,t} = 1$ for all t, and let $(y_{t+h}, x_{2,t})$ be i.i.d. $N(0,\Sigma)$. Then the difference between $\mathrm{E}_T \bar{D}_T$ and the in quantity (1) is

$$\begin{aligned}
\mathrm{E}_T \bar{D}_T - (\mathrm{E}\, L(y_{t+h} - x_{1t}'\theta_1) - \mathrm{E}\, L(y_{t+h} - x_{2t}'\theta_2)) & \\
&= (\mathrm{E}_T(y_{T+h+1} - \hat{\theta}_{1,T})^2 - \mathrm{E}_T(y_{T+h+1} - x_{T+1}'\hat{\theta}_{2,T})^2) \\
&\quad - (\mathrm{E}\, y_{T+h+1}^2 - \mathrm{E}(y_{T+h+1} - x_{T+1}'\theta_2)^2) \\
&= (\hat{\theta}_{2,t} - \theta_2)' \operatorname{var}(x_{2,t}) (\hat{\theta}_{2,t} - \theta_2) + o_p(1).
\end{aligned}$$

This last term has expectation equal to $\operatorname{var}(y_T)\frac{K_2}{T-K_2-1}$ and would converge to zero in probability if K_2 were fixed, but does not when $\lim K_2/T > 0$. In Section 3.1 we show that the OOS-t statistic can estimate $\mathrm{E}_T \bar{D}_T$ under our increasing K asymptotics and does not estimate the expected loss associated with the pseudotrue coefficients.

[10] This statement is subject to the usual assumptions: some form of stationarity, bounded moments, and weak dependence.

The conditional expectation $\mathrm{E}_T \bar{D}_T$ has been studied heavily in cross-sectional settings with independent observations. In such a setting, $\mathrm{E}_T \bar{D}_T$ is equal to the difference in the models' *generalization error*, which has been used widely as a measure of model accuracy in the machine learning literature (see Hastie et al., 2008, for further discussion). Moreover, with i.i.d. observations, the expectation of $\mathrm{E}_T \bar{D}_T$ equals Akaike's (1969) Final Prediction Error (FPE). Both generalization error and FPE are defined by a model's performance on a new, independent, data set, but, for lack of a better term, we will call $\mathrm{E}_T \bar{D}_T$ the "difference in generalization error" for the rest of the paper with hopefully no risk of confusion.

Finally, define the following notation. The l_v-norm for vectors in $\mathbb{R}^p$ (with p arbitrary) is denoted $|\cdot|_v$, and the L_v-norm for L_v-integrable random variables is $\|\cdot\|_v$. The functions $\lambda_i(\cdot)$ take a square-matrix argument and return its ith eigenvalue (with $\lambda_i(A) \leq \lambda_{i+1}(A)$ for any matrix A). All limits are taken as $T \to \infty$ unless stated otherwise.

2.2 Assumptions

The next conditions are assumed to hold throughout the paper. The first assumption controls the dependence of the underlying random array. The second lays out the details of our asymptotic approximation. The third assumption controls the smoothness of the loss function and bounds the moments of the difference in the models' performance; the fourth assumption describes the behavior of the estimation and test windows. And the last assumption describes the kernel used to estimate the OOS average's asymptotic variance.

Assumption 1. *The random array $\{y_t, x_t\}$ is stationary and absolutely regular with coefficients β_j of size $-\rho/(\rho-2)$; ρ is greater than two and discussed further in Assumption 3.*

This assumption is a standard condition on the dependence of the underlying stochastic array. The only novelty is that we use absolute regularity instead of strong or uniform mixing as our weak dependence condition; absolute regularity admits a particular coupling argument, *Berbee's Lemma* (Berbee, 1979, reproduced in this paper as Lemma A.1 for reference) that is unavailable for strong mixing sequences. Absolute regularity implies uniform mixing but is more restrictive than strong mixing, so this assumption is not unduly strong. For a detailed discussion of these weak dependence conditions, please see Davidson (1994) or Doukhan (1994).

Our strict stationarity assumption is also somewhat stronger than is typically used; West (1996) and McCracken (2007), for example, present results assuming covariance stationarity of the loss associated with the pseudotrue models. We need to make a stronger assumption because we will need to prove asymptotic results when the $\hat{\theta}_{it}$ remain random—so we would need covariance stationarity to hold for almost all estimates of θ_i and not just for the pseudotrue value. The only way to guarantee that condition is to assume strict stationarity for the underlying stochastic processes.

The next assumption describes our asymptotic experiment.

Assumption 2. *The number of regressors for each model, K_1 and K_2, are less than R and $(K_2 - K_0)/T$ is uniformly positive; K_0 is the number of regressors shared by the two models ($(K_1 - K_0)/T$ may be uniformly positive as well, but is not required to be).*

The variance of y_{t+h} given $\mathscr{F}_t$ is uniformly positive and finite and all of the eigenvalues of the covariance matrix of x_t are uniformly positive and finite as well. Moreover,

$$\lambda_{\max}(X_{iS}' X_{iS}) = O_{L_3}(S), \tag{2}$$

$$\lambda_{\max}((X_{iS}' X_{iS})^{-1}) = O_{L_3}(1/S), \tag{3}$$

$$\lambda_{\max}\Big(\mathrm{E}\big(\sum_{s,t=U}^{V-h} \varepsilon_{i,s+h}\varepsilon_{i,t+h}x_{is}x_{it}' \,\big|\, x_{i1},\ldots,x_{i,U-1}; \sum_{s=U}^{V-h} x_{is}x_{is}'; x_{i,V-h+1},\ldots,x_{i,T-h}\big)\Big) = O_{L_3}(\max(V-U, K_i)), \tag{4}$$

and

$$\operatorname{tr}\mathrm{E}\big(\sum_{s,t=U}^{V-h} \varepsilon_{i,s+h}\varepsilon_{i,t+h}x_{is}x_{it}' \,\big|\, x_{i1},\ldots,x_{i,U-1}; \sum_{s=U}^{V-h} x_{is}x_{is}'; x_{i,V-h+1},\ldots,x_{i,T-h}\big)\Big) = O_{L_3}((V-U)\times K_i) \tag{5}$$

for large enough T, where $S = R,\ldots,T$, $1 \le U \le V-h \le T-h$, $i = 1, 2$,

$$X_{iS} \equiv [x_{i1} \quad \ldots \quad x_{i,S-h}]' \qquad \textit{and} \qquad \boldsymbol{\varepsilon}_{iS} = (\varepsilon_{i,1+h},\ldots,\varepsilon_{i,S})'.$$

Additionally, the Euclidean norms of the pseudotrue coefficients, θ_1 and θ_2, satisfy $|\theta_1|_2 = O(1)$ and $|\theta_2|_2 = O(1)$.

The assumption on K_1 and K_2 is crucial to the paper; we assume that the model complexity grows with T fast enough to break consistency. This assumption is how we derive an asymptotic concept of "overfit."

The assumption that y_{t+h} and x_t have positive and finite variance is straightforward. The conditions on the eigenvalues are technical and control the behavior of the OLS estimator as the number of regressors gets large—the third and fourth assumptions are nonstandard but can be easily verified under, for example, independence. Section 3.3 contains such an example. The restrictions on the pseudotrue coefficients ensure that the regression model doesn't dominate the variance of y_{t+h} in the limit.

The next assumption establishes moment conditions for the OOS loss process and smoothness conditions for the loss function itself. The moment conditions are standard and apply to D_t, and the smoothness conditions are relatively weak.

Assumption 3. *The loss function L is continuous, has finite left and right derivatives, and $L(0) = 0$. There is a constant B_L and a function L' that bounds the left and right derivative of L at every point such that $\|D_t\|_\rho \leq B_L$; $\|D_t^*\|_\rho \leq B_L$ for all t, where*

$$D_t^* = L(y^* - x_1^{*\prime}\hat{\theta}_{1t}) - L(y^* - x_2^{*\prime}\hat{\theta}_{2t}) \tag{6}$$

and (y^*, x_1^*, x_2^*) equals (y_t, x_{1t}, x_{2t}) in distribution but is independent of $\mathscr{F}_T$ (ρ is defined in Assumption 1); and

$$\|L'(y^* - x_i^{*\prime}(\alpha\hat{\theta}_{iR} + (1-\alpha)\hat{\theta}_{iT}))\|_2 \leq B_L \tag{7}$$

for any $\alpha \in [0,1]$.

The differentiability condition in Assumption 3 is weak and allows the loss function itself to be non-differentiable; for example, absolute error and many asymmetric loss functions satisfy this assumption. The assumption makes use of both $(y_{t+h}, x_{1t}, x_{2t})$ and (y^*, x_1^*, x_2^*) because the period t observations can be dependent on $\hat{\theta}_{iT}$ in complicated ways. When the underlying observations are independent these assumptions can simplify considerably.

The next assumption controls the growth of the test and training samples.

Assumption 4. *(a) $P, R, Q \to \infty$ as $T \to \infty$. (b) $P^2/T \to 0$ and $P/Q \to 0$ as $T \to \infty$.*

The requirements that P and R grow with T are common. Parts of the assumption are new, in particular the requirement that $P^2/T \to 0$. See Lemma 2 for a discussion of its implications. In practical terms, this assumption requires that the test sample be large and that the training sample be much larger, by enough that including or excluding the test sample does not affect the estimates of θ_1 or θ_2 very much.

A final assumption restricts the class of variance estimators we will consider. We use the same class of estimators studied by de Jong and Davidson (2000) (their class $\mathscr{K}$); see their paper for further discussion.

Assumption 5. *W is a kernel from $\mathbb{R}$ to $[-1,1]$ such that $W(0) = 1$, $W(x) = W(-x)$ for all x,*

$$\int_{-\infty}^{\infty} |W(x)|dx < \infty, \quad \int_{-\infty}^{\infty} |\psi(x)|dx < \infty \tag{8}$$

with

$$\psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(z)e^{ixz}dz, \tag{9}$$

and $W(\cdot)$ is continuous at zero and all but a finite number of points.

These assumptions are broadly similar to those existing in the literature, with some differences in our assumptions relating the estimated coefficients to future values of the DGP. Section 3.3 contains an extended example that shows how these assumptions are satisfied in a simple setting.

3 Theoretical results

This section lays out our theoretical results. The first subsection presents results for the DMW OOS-t test; we show that it is asymptotically normal when one or both forecasting models is overfit. (i.e., under our increasing-K asymptotic approximation.) We also present a new limitation on these statistics—OOS comparisons heavily penalize overfit models unless the size of the test sample is a very small proportion of the total sample size, which will not be practical in much applied research. The second subsection presents results for full-sample statistics and shows that widely used test statistics and model selection criteria are misleading when choosing a model for forecasting. In contrast to OOS comparisons, these full-sample criteria choose overfit models too often, even when they are less accurate than simple benchmark models. These results are somewhat abstract, so the third subsection works through an example DGP in detail.

3.1 Asymptotic normality of the DMW test

This section has two main conceptual results. The first, Lemma 1, shows that the OOS average, $\bar{D}_R$, is asymptotically normal as the size of the test sample grows, even when the models are overfit. But $\bar{D}_R$ is centered at $\mathrm{E}_R \bar{D}_R$, the difference in the generalization error of the models estimated over the *training sample*, which is not the quantity of interest to forecasters. Forecasters will generally want to estimate or test hypotheses about the difference in the generalization error of the models estimated with the full sample, $\mathrm{E}_T \bar{D}_T$. Our second result, Lemma 2, shows that these quantities are approximately equal only when the test sample is very small relative to the total sample size. In particular, we show that $\bar{D}_R$ is a consistent estimator of $\mathrm{E}_T \bar{D}_T$ when $P/T \to 0$ and is asymptotically normal with mean $\mathrm{E}_T \bar{D}_T$ when $P^2/T \to 0$. After establishing these Lemmas, we then show that the OOS-t test is asymptotically standard normal and can be used to test hypotheses about $\mathrm{E}_T \bar{D}_T$, which requires the two Lemmas as well as a consistent estimator of the variance of the OOS average.

In the first result, we show that $\sqrt{P}(\bar{D}_R - \mathrm{E}_R \bar{D}_R)$ is asymptotically normal as $P \to \infty$. This application of the CLT is complicated by a hidden source of dependence—the training sample estimators $\hat{\theta}_{iR}$. Under conventional asymptotic theory, we would replace each $\hat{\theta}_{iR}$ with its pseudotrue value θ_i and apply the CLT to $L(y_{t+h} - x_{it}'\theta_i)$, with some complications potentially arising from the replacement (as in West, 1996, Clark and McCracken, 2001, or McCracken, 2007, for example). But we can not make that replacement here, because our asymptotic approximation prevents $x_{it}'\hat{\theta}_{iR}$ from converging to $x_{it}'\theta_i$.

Instead, we show that $D_t - \mathrm{E}_R D_t$ is an L_2-mixingale that satisfies the CLT. Mixingales satisfy a weak-dependence condition similar to MDSes,[11] but they have some limitations.

[11] An array $Z_{n,t}$ and an increasing sequence of σ-fields $\mathscr{G}_{n,t}$ is an *L_2-mixingale of size* $-1/2$ if there is an array of constants $\{c_{n,t}\}$ and a sequence of constants $\zeta_l = O(l^{-1/2-\delta})$ for some $\delta > 0$ such that

$$\|\mathrm{E}(Z_{n,t} \mid \mathscr{G}_{n,t-l})\|_2 \leq c_{n,t}\zeta_l \qquad \text{and} \qquad \|Z_{n,t} - \mathrm{E}(Z_{n,t} \mid \mathscr{G}_{n,t+l})\|_2 \leq c_{n,t}\zeta_{l+1}.$$

Transformations of mixingales are typically not themselves mixingales, which means that CLTs for mixingale processes require additional assumptions to hold beyond the mixingale property.[12] This is in contrast to Near Epoch Dependent (NED) processes, which retain the NED property after transformations. (See chapter 17 of Davidson, 1994, for further discussion of these properties.) But the OOS loss has more structure than most mixingales and behaves like an NED process in important respects. (The key result here is Lemma A4 in the Appendix.) Lemma 1 presents the CLT that we use later, and Section 3.3 works through an i.i.d. example in detail. For i.i.d. observations, the OOS loss is an MDS and is easier to work with.

Lemma 1. *If Assumptions 1–3 hold then* $\{D_t - \mathrm{E}_R D_t, \mathscr{F}_t\}$ *is an* L_2*-mixingale of size* $-1/2$*. Moreover*

$$\sqrt{P}(\bar{D}_R - \mathrm{E}_R \bar{D}_R)/\sigma \to^d N(0,1) \tag{10}$$

as $P \to \infty$ *if* σ^2 *is uniformly almost surely positive, where* $\sigma^2 = \mathrm{var}_R(\sqrt{P}\bar{D}_R)$ *(which is equal to* $P\,\mathrm{E}_R(\bar{D}_R - \mathrm{E}_R \bar{D}_R)^2$*).*

It may be helpful to compare Lemma 1 to the method of proof in Giacomini and White (2006). Giacomini and White (2006) show that the OOS average is asymptotically normal when the forecasts are estimated with a fixed length rolling window. In that case, each $\hat{\theta}_{it}$ depends on only the most recent R observations and, since R is fixed in their theory, the forecast errors $y_{t+h} - x_{it}'\hat{\theta}_{it}$ are themselves mixing processes. Transformations of their forecast errors are still obviously mixing processes and obey the CLT.

In our paper, R is not fixed and the forecast errors are not a convenient weakly-dependent process, since the estimation error in $\hat{\theta}_{it}$ introduces strong dependence. Consequently, transformations of the forecast errors are not weakly dependent either. But this additional dependence has a special form and can be removed by subtracting the conditional mean; specifically $g(y_{t+h} - x_{it}'\hat{\theta}_{it})$ is not weakly dependent but $g(y_{t+h} - x_{it}'\hat{\theta}_{it}) - \mathrm{E}_R\, g(y_{t+h} - x_{it}'\hat{\theta}_{it})$ is. Assumptions 1–3 allow us to show directly that $D_t - \mathrm{E}_R D_t$ and (crucially) $D_t^2 - \mathrm{E}_R D_t^2$ are both weakly dependent mixingales, ensuring Lemma 1.

The next Lemma connects $E_R\bar{D}_R$ to $\mathrm{E}_T \bar{D}_T$, the difference in the models' generalization error. Under conventional asymptotics, these quantities are generally close. But they are not for overfit models and the models will tend to forecast better when estimated over the full sample than over the training sample. Consequently, OOS comparisons will penalize overfit models too much unless the test sample is small relative to the total data set.

Lemma 2. *Under Assumptions 1–4a,*

$$\mathrm{E}_T \bar{D}_T - \mathrm{E}_R \bar{D}_R = O_p(\sqrt{P/T}) + o_p(P^{-1/2}) + o_p(Q^{-1/2}). \tag{11}$$

Mixingales were introduced and developed by McLeish (1974, 1975a,b, 1977).

[12] See de Jong (1997) for an illustration. We will borrow heavily from his results in our proofs.

We can view $\mathrm{E}_T \bar{D}_T - \mathrm{E}_R \bar{D}_R$ as noise introduced by approximating the performance of the full-sample estimates with that of the partial-sample estimates. The key term on the RHS of (11) in practice is $O_p(\sqrt{P/T})$—unless P/T is small, this noise dominates the OOS average, making it an inconsistent estimator of $\mathrm{E}_T \bar{D}_T$. The $o_p(P^{-1/2})$ term is completely unrestrictive in our applications because we will multiply the OOS average by at most $\sqrt{P}$ (for the CLT). And the $o_p(Q^{-1/2})$ term has some implications for interpreting these results: we can only measure the average performance of the forecasting models for an extended period in the future, long enough that the future observations are essentially independent of the current information. For $Q = 1$, for example, the dependence between y_{T+h+1} and the current information set $\mathscr{F}_T$ may be very strong.

Finally, we can use Lemmas 1 and 2 to show that the DMW test is asymptotically normal and centered at $\mathrm{E}_T \bar{D}_T$, as long as Assumption 4b holds.

Theorem 1. *Suppose that Assumptions 1–5 hold (including 4b), that $\gamma \to \infty$ and $\gamma/P \to 0$ as $T \to \infty$, and that σ^2 is uniformly a.s. positive. Define $\hat{\sigma}^2$ to be the usual OOS HAC estimator of the asymptotic variance of $\bar{D}_R$,*

$$\hat{\sigma}^2 \equiv P^{-1} \sum_{s,t=R+1}^{T-h} (D_t - \bar{D}_R)(D_s - \bar{D}_R) W((t-s)/\gamma). \tag{12}$$

Then

$$\sqrt{P}(\bar{D}_R - \mathrm{E}_T \bar{D}_T)/\hat{\sigma} \to^d N(0,1). \tag{13}$$

The requirement that σ^2 be uniformly a.s. positive is not restrictive under our asymptotic theory. Since the models' coefficients are estimated with uncertainty in the limit, the two models give different forecasts even if they both nest the DGP. This is intuitively similar to Giacomini and White's (2006) rolling-window result, but comes from different asymptotic theory; Giacomini and White keep the variance of the OOS average positive by letting $P \to \infty$ with R fixed; in this paper, $R \to \infty$ but the variance remains positive since $K \to \infty$ (quickly) as well.

If K_1 and K_2 were finite, West's (1996) and McCracken's (2007) asymptotic theories would apply and the OOS-t test would remain normal as long as $P/T \to 0$, even if $\sigma^2 \to 0$. Under that asymptotic approximation, the variance will converge to zero whenever the true DGP is nested in both of the forecasting models. The same principles may also apply if K_1 and K_2 grow slowly, with $K_2/T \to 0$. It is likely that a second order expansion along the lines of McCracken (2007) would lead to asymptotic normality for that intermediate case, but we leave that issue to future research.

Finally, Theorem 2 establishes that the DMW test can be use to construct confidence intervals for $\mathrm{E}_T \bar{D}_T$ and test hypotheses about $\mathrm{E}_T \bar{D}_T$. In particular, forecasters will often want to test the null hypothesis that $\mathrm{E}_T \bar{D}_T \leq 0$, meaning that the benchmark model is expected to be more accurate than the alternative model in the future.

Theorem 2. *Suppose that the conditions of Theorem 1 hold. Then each of the usual Gaussian confidence intervals,*

$$[\bar{D}_R - z_{\alpha/2}\,\hat{\sigma}/\sqrt{P}, \bar{D}_R + z_{\alpha/2}\hat{\sigma}/\sqrt{P}], \tag{14}$$

$$[\bar{D}_R - z_{\alpha}\,\hat{\sigma}/\sqrt{P}, +\infty), \tag{15}$$

and

$$(-\infty, \bar{D}_R + z_{\alpha}\,\hat{\sigma}/\sqrt{P}], \tag{16}$$

contains $\mathrm{E}_T\,\bar{D}_T$ *with probability* α *in the limit, with* z_α *the* $1-\alpha$ *quantile of the standard normal distribution. If, in addition,* $\lim \Pr[\mathrm{E}_T\,\bar{D}_T \leq 0]$ *is positive then*

$$\lim \Pr[P^{1/2}\bar{D}_R/\hat{\sigma} > z_\alpha \mid \mathrm{E}_T\,\bar{D}_T \leq 0] \leq \alpha. \tag{17}$$

The results for confidence intervals in Theorem 2 follow immediately from Theorem 1, but (17) requires additional proof.

3.2 Failure of in-sample statistics for forecast evaluation

In this subsection we look at the behavior of some full-sample statistics under the same asymptotic theory as before. We show that these statistics, which include common tests such as the Wald test as well as model selection criteria such as the AIC, do not measure the models' generalization error and consequently do not indicate which model will be more accurate in the future. Some of these statistics will tend to choose the larger model regardless of which model will be more accurate in the future, whereas others tend to choose the smaller model. This is a different issue than whether or not full-sample tests are valid for testing hypotheses about the pseudotrue coefficients of the models; as Calhoun (2011) and Anatolyev (2012) demonstrate, variations of the Wald test can be valid for those hypotheses under increasing-K asymptotics.[13] To simplify the presentation and the intuition we only derive results for nested models and for MSE loss, but the conclusions hold much more generally.[14]

The full-sample statistics we study in this paper share a common property. They choose the alternative model when the distance between a subset of their coefficient estimates and the origin exceeds a threshold, and choose the benchmark otherwise.

[13] Anatolyev (2012) shows that the Wald test is invalid in general and gives an adjustment that corrects the critical values; he also shows that the F-test is asymptotically valid under certain conditions on the distribution of the regressors. Calhoun (2011) shows that the F-test is asymptotically invalid without Anatolyev's constraint, even under homoskedasticity, and provides a correction that gives valid tests. Both papers only consider independent observations.

[14] In this section, we define $M_1 = [I_{K_1}\ 0_{K_1 \times (K_2 - K_1)}]$ and $M_2 = [0_{(K_2-K_1)\times K_2}\ I_{K_2-K_1}]$ as the selection matrices that return the first K_1 and the last $K_2 - K_1$ elements of θ_2, and assume that the regressors are ordered so that the first K_1 elements of $x_{2,t}$ correspond to $x_{1,t}$.

For the Wald test, for example, this threshold is chosen so that the test has correct size when those coefficients are zero in population. For small models that can be consistently estimated, these coefficient estimates are close to their true values and so this criterion can be a reasonable proxy for the relative accuracy of the larger model.

But for overfit models, the estimates will typically be far from both their pseudotrue values and also from zero. In that case, the Wald test and the AIC will both tend to choose the larger model even when it is less accurate than the smaller benchmark model. This phenomenon is driven by the dimensionality of the alternative model, since there are more potential values of the coefficient estimator that are far from zero when it has many elements. The threshold for the Wald test and the AIC is set by construction to be a bounded distance from the origin, and every other statistic that shares this property has the same behavior and can reject the benchmark model with high probability, even when it is more accurate. This behavior is formalized in Theorem 3.

Theorem 3. *Suppose Assumptions 1–4 hold and let Λ be a model selection statistic that takes the values zero or one; $\Lambda = 0$ indicates that the benchmark model is chosen and $\Lambda = 1$ indicates the alternative. Moreover, assume that $L(e) = e^2$ and that there exist a deterministic scalar c and a sequence of possibly random matrices V_T such that*

$$\limsup_{T\to\infty} \Pr[\Lambda \geq 1\{\hat{\theta}_{2T}' M_2' V_T M_2 \hat{\theta}_{2T} > c\}] \to 0 \tag{18}$$

where the eigenvalues of V_T are uniformly bounded in probability, $\operatorname{plim} \operatorname{rank}(V_T)/T > 0$*, and neither V_T nor c depend on θ or the value of $\hat{\theta}_{2T}$.*

Then there exist DGPs satisfying these assumptions such that

$$\liminf_{T\to\infty} \mathrm{E}(\Lambda \mid \mathrm{E}_T \bar{D}_T \leq 0) \geq 1/2. \tag{19}$$

As we state above, (19) means that, with high probability, the benchmark model is rejected even when it is more accurate. In general, δ can be made arbitrarily small by increasing the region enclosed by $1\{\hat{\theta}_{2T}' M_2' V_T M_2 \hat{\theta}_{2T}\}$ and DGPs can be chosen to put the limiting quantity in (19) arbitrarily close to 1 while still satisfying the assumptions of the Theorem. The DGPs and statistics used in Theorem 3 are simple and common. They include correctly specified linear models with normal and homoskedastic errors; and the statistics that meet the restrictions on Λ include the F-test and AIC.

For example, the F-statistic for the null hypothesis that the benchmark model is correctly specified is known to have the form

$$F = \frac{\hat{\theta}_{2T}' M_2' (M_2 (X_{2T}' X_{2T})^{-1} M_2')^{-1} M_2 \hat{\theta}_{2T}}{s^2 (K_2 - K_1)}$$

where s^2 is the usual estimator of the variance of the regression error. Then let c be any number greater than 1 and define

$$V_T = M_2' (M_2 (\tfrac{1}{s^2(K_2-K_1)} X_{2T}' X_{2T})^{-1} M_2')^{-1} M_2'.$$

If Assumption 2 holds and s^2 is consistent then V_T has uniformly bounded eigenvalues and has rank $K_2 - K_1$, so it satisfies (18). Robust variations of the Wald test can obviously satisfy (18) for similar reasons, and the AIC for nested linear models is equivalent to using the F-test with the critical value $(e^{2(K_2-K_1)/T} - 1) \cdot (T - K_2)/(K_2 - K_1)$, which converges to a finite limit, so the AIC satisfies (18) as well.

For statistics that don't satisfy (18), the behavior can be quite different. The BIC, for example, can also be written in terms of the F-statistic and is equivalent to using the critical value $(T\, e^{2(K_2-K_1)/T} - 1) \cdot (T - K_2)/(K_2 - K_1)$. This critical value diverges as $T \to \infty$, ensuring that (18) fails for any c. For this statistic, we have the opposite problem as before: when the alternative model is *more* accurate, the coefficient estimates of the larger model are contained in a bounded region of the parameter space. Since the acceptance region of the BIC grows, it eventually contains *any* bounded region of the parameter space. For large enough T, the BIC will always choose the *smaller model*, even when the larger model is more accurate.

This behavior is formalized in Theorem 4.

Theorem 4. *Suppose Assumptions 1–4 hold, let $L(e) = e^2$, and let Λ be a model selection statistic as in Theorem 3. Also assume that, for any finite scalar c,*

$$\Pr[\Lambda \leq 1\{\hat{\theta}_{2T}' M_2' J M_2 \hat{\theta}_{2T} > c\}] \to 1 \quad \text{as } T \to \infty. \tag{20}$$

Then there exist DGPs satisfying these assumptions such that

$$\mathrm{E}(\Lambda \mid \mathrm{E}_T \bar{D}_T \geq 0) \to^p 0. \tag{21}$$

The condition 20 requires that the acceptance region of Λ eventually contains any finite cylinder centered at the origin. Again, (21) implies that statistics like the BIC will always choose the smaller model for some DGPs, even when the larger model will give more accurate forecasts. Both models may be overfit, in that both K_1/T and K_2/T may both be positive in the limit; the key is that $(K_2 - K_1)/T$ is also positive in the limit.

It is important to remember that previous research, such as Calhoun (2011) and Anatolyev (2012), does not predict these results. When Λ represents a test statistic, the test may have correct size for the null hypothesis that the additional coefficients on the larger model are zero. The results in this subsection are being driven by the full-sample statistics' behavior when the smaller model is misspecified but more accurate.

Any statistic that uses the distance of the models' estimated coefficients from a set point (the origin being the most common) is poorly suited for choosing between overfit forecasting models. These models only forecast well when their coefficient estimates are close to their pseudotrue values, which can be far from the origin or any other prespecified point. Depending on the statistic, it can be biased towards choosing the larger model or the smaller model. Formal in-sample *tests* will likely be biased towards the larger model, as we show for the F-test and Wald test in Theorem 3.

3.3 An extended simple example

This subsection illustrates the previous theoretical results with a concrete example. Let $L(e) = e^2$ and $h = 1$. Suppose that the benchmark is nested in the alternative and $(x_{2t}, \varepsilon_{2,t+1}) \sim i.i.d.\ N(0, I)$, and let $y_{t+1} = x_{2t}'\theta_2 + \varepsilon_{2,t+1}$ be the DGP. Also assume that $K_2/T \to c_2$ and $K_1/T \to c_1 < c_2$.

The first part of this subsection shows how the assumptions in Section 2.2 are satisfied and the second part demonstrates asymptotic normality of the DMW OOS-t test. The third part explicitly shows that $\mathrm{E}_R \bar{D}_R$ converges to $\mathrm{E}_T \bar{D}_T$ only when the test sample is small. And the last part demonstrates that the F-test does not indicate which model will be more accurate in the future, even in this simple example.

Fulfillment of Assumptions 1–5

We will go through the assumptions one by one. The first is a condition on the dependence and moments of the process. Since the DGP is i.i.d. Normal, these conditions are satisfied trivially.

Assumption 2 deals with the design matrix. In this example, we directly assume that K_2 and K_1 grow at the correct rates so the the assumption is satisfied. The variance of y_{t+1} given $\mathcal{F}_t$ is simply the unconditional variance of $\varepsilon_{2,t+1}$, which is 1; and the eigenvalues of $\mathrm{E}\, x_{2t}x_{2t}$ all equal 1 as well; and so they are all uniformly positive and finite as required.

Assumption 2 also requires the eigenvalues of $S^{-1}X_{iS}'X_{iS}$ to be positive and finite, and for the largest eigenvalue of $S^{-1}X_{iS}'X_{iS}$ to be bounded in L_3. These results follow from developments in random matrix theory. Geman (1980) establishes that the largest eigenvalue of $X_{iS}'X_{iS}$ is of order $S + K_i$ and Johnstone (2001) shows that it converges in distribution to the Tracy-Widom law of order 1 (Tracy and Widom, 1996), which has finite 3rd moments. Silverstein et al. (1985) and Baker et al. (1998) prove similar results for the smallest eigenvalue. Assumption 2 also requires the largest eigenvalue of $(X_{iS}'X_{iS})^{-1}$ be bounded in L_3; this should follow from a similar argument, but we are unaware of any papers that explicitly derive moments for this eigenvalue.

For the conditional expectation of $\sum_{s,t=U}^{V-1} \varepsilon_{i,t+1}\varepsilon_{i,s+1}x_{i,s}x_{i,t}'$, independence implies that

$$\mathrm{E}\Big(\sum_{s,t=U}^{V-1} \varepsilon_{i,s+1}\varepsilon_{i,t+1}x_{is}x_{it}' \,\Big|\, x_{i1},\ldots,x_{i,U-1}; \sum_{s=U}^{V-1} x_{is}x_{is}'; x_{i,V},\ldots,x_{i,T-1}\Big) = \sum_{s=U}^{V-1} x_{is}x_{is}'.$$

The largest eigenvalue of this last matrix is of order $K_i + V - U$, as discussed, and it has at most $V - U$ nonzero eigenvalues, ensuring that both (4) and (5) hold.

Assumption 3 restricts the loss function and the realized OOS loss. In this example, we have

$$D_t^* =^d D_t = (e_{2,t+1} + x_{2t}'\theta_2 - x_{1t}'\hat{\theta}_{1t})^2 - (e_{2,t+1} + x_{2t}'(\theta_2 - \hat{\theta}_{2t}))^2$$

which has bounded ρth moments by construction. The loss function is differentiable and $L'(e) = 2e$, so (7) holds as well.

Finally, Assumption 4 deals with the choice of test and training sample and holds automatically. And Assumption 5 requires the kernel of the HAC variance estimator to be continuous at zero; it is straightforward to construct an estimator that satisfies this condition. However, in this section we will use the estimator

$$\hat{\sigma}^2 = \frac{1}{P}\sum_{t=R+1}^{T-1}(D_t - \bar{D}_R)^2$$

which does not satisfy Assumption 5 but nevertheless is consistent for the conditional variance of $\bar{D}_R$ because the underlying observations are independent.

Asymptotic normality of the OOS average, Lemma 1

This subsection walks through Lemma 1's CLT. Since the observations in this example are i.i.d., we do not need to use mixingale theory to derive the results, but the OOS process is still affected by the estimation error in $\hat{\theta}_{1t}$ and $\hat{\theta}_{2t}$ and has a high degree of dependence. This dependence makes the OOS process an MDS, and MDS asymptotic theory replaces mixingale theory in this example.

By construction, we have

$$D_t = \begin{cases} 2\varepsilon_{t+1}(x_{2t}'\hat{\theta}_{2R} - x_{1t}'\hat{\theta}_{1R}) + (x_{2t}'\theta_2 - x_{1t}'\hat{\theta}_{1R})^2 - (x_{2t}'\theta_2 - x_{2t}'\hat{\theta}_{2R})^2 & \text{if } t < T \\ 2\varepsilon_{t+1}(x_{2t}'\hat{\theta}_{2T} - x_{1t}'\hat{\theta}_{1T}) + (x_{2t}'\theta_2 - x_{1t}'\hat{\theta}_{1T})^2 - (x_{2t}'\theta_2 - x_{2t}'\hat{\theta}_{2T})^2 & \text{if } t > T \end{cases}$$

and so we can explicitly derive the components of Lemma 1. First,

$$\begin{aligned} D_t - \mathrm{E}_R D_t = 2\varepsilon_{t+1}(x_{2t}'\hat{\theta}_{2R} - x_{1t}'\hat{\theta}_{1R}) &+ \left\{(x_{2t}'\theta_2 - x_{1t}'\hat{\theta}_{1R})^2 - \mathrm{E}_R(x_{2t}'\theta_2 - x_{1t}'\hat{\theta}_{1R})^2\right\} \\ &- \left\{(x_{2t}'\theta_2 - x_{2t}'\hat{\theta}_{2R})^2 - \mathrm{E}_R(x_{2t}'\theta_2 - x_{2t}'\hat{\theta}_{2R})^2\right\} \end{aligned}$$

for $t < T$. Since the underlying observations are i.i.d., $\mathrm{E}_{t-1} D_t = \mathrm{E}_R D_t$ a.s. and consequently $\{D_t - \mathrm{E}_R D_t, \mathscr{F}_t; t = R+1, \ldots, T-1\}$ is an MDS.

Although transformations of $D_t - \mathrm{E}_R D_t$ will obviously not be MDSes in general, it should be clear that transformations of D_t will be MDSes after subtracting their conditional mean; i.e.

$$\{g(D_t) - \mathrm{E}_R\, g(D_t); \mathscr{F}_t; t = R+1, \ldots, T-1\}$$

is an MDS as long as $g(D_t)$ has finite mean, but $g(D_t - \mathrm{E}_R D_t)$ is not. This MDS result holds because x_t and y_{t+1} are independent of $\hat{\theta}_{1R}$ and $\hat{\theta}_{2R}$, so

$$\begin{aligned} \mathrm{E}_R\, g(D_t) &= \int g\big((x_1'\hat{\theta}_{1R})^2 - (x_2'\hat{\theta}_{2R})^2 + 2y(x'\hat{\theta}_{2R} - x'\hat{\theta}_{1R})\big) f(y,x)\, dx\, dy \\ &= \mathrm{E}_{t-1}\, g(D_t) \end{aligned}$$

a.s., where f is the density of (y_{t+1}, x_{2t}) and x_1 denotes the first K_1 elements of x. This result parallels our results for mixingales in the general case.

As a result, $\bar{D}_R$ and $(1/P)\sum_{t=R+1}^{T-1} D_t^2$ both obey LLNs: $\bar{D}_R \to^p \mathrm{E}_R \bar{D}_R$ and

$$(1/P)\sum_{t=R+1}^{T-1} D_t^2 \to^p \mathrm{E}_R D_T^2.$$

Moreover, these convergence results imply that $\hat{\sigma}^2 - \mathrm{var}_R \sqrt{P}\bar{D}_R \to^p 0$. And so

$$\sqrt{P}\bar{D}_R/\hat{\sigma} \to N(0,1)$$

by the MDS CLT.

Convergence of $\mathrm{E}_R \bar{D}_R$ to $\mathrm{E}_T \bar{D}_T$, Lemma 2

This subsection works through Lemma 2. In this example, the difference between $\mathrm{E}_R \bar{D}_R$ and $\mathrm{E}_T \bar{D}_T$ equals

$$\begin{aligned}
\mathrm{E}_R \bar{D}_R - \mathrm{E}_T \bar{D}_T &= \left\{(\hat{\theta}_{1R} - \theta_1)'(\hat{\theta}_{1R} - \theta_1) - (\hat{\theta}_{2R} - \theta_2)'(\hat{\theta}_{2R} - \theta_2)\right\} \\
&\quad - \left\{(\hat{\theta}_{1T} - \theta_1)'(\hat{\theta}_{1T} - \theta_1) - (\hat{\theta}_{2T} - \theta_2)'(\hat{\theta}_{2T} - \theta_2)\right\} \\
&= \boldsymbol{\varepsilon}_T' \left\{\tilde{X}_{1R}(X_{1R}'X_{1R})^{-2}\tilde{X}_{1R}' - X_{1T}(X_{1T}'X_{1T})^{-2}X_{1T}' \right. \\
&\quad \left. - \tilde{X}_{2R}(X_{2R}'X_{2R})^{-2}\tilde{X}_{2R}' + X_{2T}(X_{2T}'X_{2T})^{-2}X_{2T}'\right\}\boldsymbol{\varepsilon}_T.
\end{aligned}$$

with $\tilde{X}_{iR}$ the $T \times K_i$ matrix $[X_{iR}' \; 0]'$. This last term is a quadratic form and the regressors and errors are assumed to be normal, so we can find the rate that the difference converges to zero in probability by calculating its first two moments.

The mean difference is

$$\begin{aligned}
\mathrm{E}(\mathrm{E}_R \bar{D}_R - \mathrm{E}_T \bar{D}_T) &= O\Big(\max_i \mathrm{E}\,\mathrm{tr}\,\{\tilde{X}_{iR}(X_{iR}'X_{iR})^{-2}\tilde{X}_{iR} - X_{iT}(X_{iT}'X_{iT})^{-2}X_{iT}\}\Big) \\
&= O\Big(\max_i \mathrm{tr}\,\{\,\mathrm{E}(X_{1R}'X_{1R})^{-1} - \mathrm{E}(X_{1T}'X_{1T})^{-1}\}\Big) \\
&= O(K_1 P/(R-K_1)(T-K_1)) \\
&= O(P/T)
\end{aligned}$$

The first equality follows from the expectation of a quadratic form, the second from routine manipulations of the trace operator, and the third from the moments of the inverse Wishart distribution.

Similarly, the variance of the difference is

$$\begin{aligned}
&\mathrm{var}(\mathrm{E}_R\bar{D}_R - \mathrm{E}_T\bar{D}_T) \\
&\quad = O\Big(\max_i \mathrm{E}\,[\,\mathrm{tr}((X_{iR}'X_{iR})^{-1} - (X_{iT}'X_{iT})^{-1})]^2 + 2\,\mathrm{tr}\,\mathrm{E}\,((X_{iR}'X_{iR})^{-2} - (X_{iT}'X_{iT})^{-2}) \\
&\qquad - [\,\mathrm{tr}\,\mathrm{E}((X_{iR}'X_{iR})^{-1} - (X_{iT}'X_{iT})^{-1})]^2\Big) \\
&\quad = O(P/T)^2.
\end{aligned}$$

So, in this example, $P^{1/2}(\mathrm{E}_R \bar{D}_R - \mathrm{E}_T \bar{D}_T) \to^p 0$ if $P^3/T^2 \to 0$, which is slightly weaker than the general requirement that $P^2/T \to 0$. If $\lim P^3/T^2 > 0$ the OOS average still obeys the MDS CLT, but it is not centered correctly at $\mathrm{E}_T \bar{D}_T$.

Behavior of the F-test

This part illustrates the behavior of full-sample statistics in our simple example. To simplify the presentation even more, assume that the full-sample design matrix is orthogonal in-sample, not just in population, so $X_{2T}'X_{2T} = T \cdot I_{K_2 \times K_2}$, and again let M_1 and M_2 be the selection matrices for the first K_1 and the last $K_2 - K_1$ elements of θ_2. In this example,

$$M_2\hat{\theta}_{2T} \sim N(M_2\theta_2, (1/T)\, I_{K_2-K_1})$$

and, conditional on $\mathrm{E}_T \bar{D}_T = 0$, the density of $M_2\hat{\theta}_{2T}$ concentrates uniformly on the surface of the sphere centered at $M_2\theta_2$ and passing through the origin:

$$(M_2\hat{\theta}_{2T} - M_2\theta_2)'(M_2\hat{\theta}_{2T} - M_2\theta_2) = \theta_2' M_2' M_2 \theta_2.$$

To see intuitively that it must pass through the origin, note that the two models will give identical forecasts when $\hat{\theta}_{2T} = 0$ since the regressors are orthogonal. (Identical forecasts obviously have the same MSE.)

This region is a cylinder in the original space, $\mathbb{R}^{K_2}$. We can also represent the null $\mathrm{E}_T \bar{D}_T \leq 0$ by conditioning on $\mathrm{E}_T \bar{D}_T = -d$ for a fixed positive constant c. In that case, the density of $M_2\hat{\theta}_{2T}$ concentrates on the sphere

$$(M_2\hat{\theta}_{2T} - M_2\theta_2)'(M_2\hat{\theta}_{2T} - M_2\theta_2) = \theta_2' M_2' M_2 \theta_2 + d$$

which has the same center but larger radius.

For the F-test we have, as before,

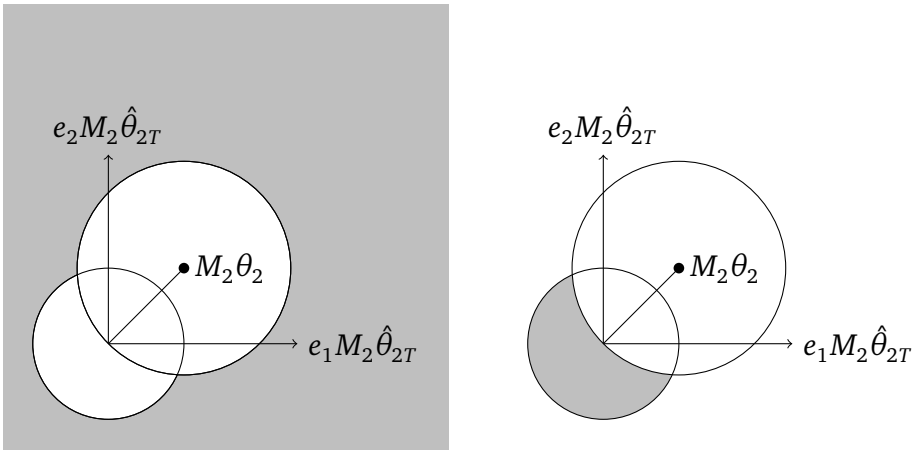
$$F = \tfrac{T}{s^2(K_2-K_1)}\hat{\theta}_{2T}' M_2' M_2 \hat{\theta}_{2T}.$$

We know that $\sqrt{T}(F-1)$ is asymptotically normal since its numerator is Chi-squared with $K_2 - K_1$ degrees of freedom and obeys a CLT. The test accepts if

$$\tfrac{T}{s^2(K_2-K_1)}\hat{\theta}_{2T}' M_2' M_2 \hat{\theta}_{2T} \leq 1 + \delta/\sqrt{T}, \tag{22}$$

where δ is chosen to determine the size of the test. In other words, the test accepts if $M_2\hat{\theta}_{2T}$ falls in the sphere centered at the origin with radius $s((K_2-K_1)/T)^{1/2} + O_p(1/\sqrt{T})$.

This is perhaps best illustrated with a picture. Figure 1 plots these quantities when $K_2 - K_1 = 2$. The circle centered at the origin plots the threshold for the F-test; when $M_2\hat{\theta}_{iT}$ falls outside this circle, the F-test rejects. The circle centered at $M_2\theta_2$ plots the set of points for which $\mathrm{E}_T \bar{D}_T = 0$. The shaded region in Figure 1 (a) plots the rejection region given $\mathrm{E}_T \bar{D}_T \leq 0$ and the shaded region in Figure 1 (b) plots the acceptance region given $\mathrm{E}_T \bar{D}_T \leq 0$.



(a) (b)

Figure 1: Graphs indicating the rejection region and region of equal generalization error for the models discussed in Section 3.2. The smaller model has no estimated parameters and the larger has two coefficients. The shaded regions in Figures (a) and (b) show the rejection region and the acceptance region respectively of the full-sample test given $\mathrm{E}_T \bar{D}_T \leq 0$. Here $e_1 = (1,0)$ and $e_2 = (0,1)$ are the first and second selection vectors, so the horizontal axes represent changes in the first unique element of $\hat{\theta}_{2T}$ and the vertical axes represent changes in the second unique element.

In the picture, the true coefficients θ_2 satisfy

$$\theta_2' M_2' M_2 \theta_2 > \mathrm{var}(\varepsilon_t)^{1/2}((K_2 - K_1)/T)^{1/2}$$

so the center of the conditional distribution of $\hat{\theta}_{2T}$ given $\mathrm{E}_T \bar{D}_T \leq 0$ is outside the acceptance region of the F-test. When this is the case, the conditional probability that $\hat{\theta}_{2T}$ falls in the rejection region is less than $1/2$, since $M_2\hat{\theta}_{2T}$ is uniformly distributed on the cylinder

$$(M_2\hat{\theta}_{2T} - M_2\theta_2)'(M_2\hat{\theta}_{2T} - M_2\theta_2) = \theta_2' M_2' M_2 \theta_2 + d$$

for some positive d.

The AIC behaves essentially the same way. For the BIC, the radius of the cylinder centered at the origin increases with T and eventually encompasses the cylinder centered at $M_2\theta_2$, so it never selects the alternative once T is large enough.

4 Monte Carlo

This section presents two simulations that investigate the accuracy of our theory in small samples. We do several Monte Carlo exercises. The first looks at whether our theoretical results for the OOS average are accurate: whether or not the OOS average

is approximately normal, and whether it is centered on $\mathrm{E}_R \bar{D}_R$, $\mathrm{E}_T \bar{D}_T$, or somewhere else entirely. The second issue is whether or not the OOS-t test is *useful* for conducting inference about $\mathrm{E}_T \bar{D}_T$. Our theoretical results suggest that it may not be, because we require $P^2/T \to 0$ for inference about $\mathrm{E}_T \bar{D}_T$ to be valid. A highly related issue is whether other statistics are useful for conducting inference about $\mathrm{E}_T \bar{D}_T$—again, our theoretical results suggest that they are not.

We use the same DGP for all of these simulations, and it is described in the next subsection. Results are presented in the subsection after that. Simulations were conducted in R (R Development Core Team, 2010) using the *MASS* (Venables and Ripley, 2002), *Matrix* (Bates and Maechler, 2013), and *rlecuyer* (Sevcikova and Rossini, 2012) and graphs are produced using *Lattice* (Sarkar, 2010). In this paper, we present results for the fixed window, but recursive window results are available in a separate Appendix and are similar.

4.1 Setup

The Monte Carlo experiment is intentionally very simple so that we can isolate the influence of the models' complexity. In particular, we do not include some features that are common in forecasting environments—serial dependence, heteroskedasticity, and complicated DGPs. The DGP we use is given by the equation

$$y_t = x_t'\theta + \varepsilon_t, \quad \varepsilon_t \sim N(0,1), \quad t = 1, \ldots, T. \tag{23}$$

The first element of x_t is 1 and the remaining $K_2 - 1$ elements are independent Standard Normal. The benchmark model is

$$y_{1t} = \sum_{j=1}^{K_1} x_{jt}\theta_j + \varepsilon_t \tag{24}$$

and the alternative model is the DGP (23). We let (K_1, K_2) equal either $(2, 3)$ or $(T/20, T/10)$ to study our theory in its intended application as well as for more parsimonious models. We let T equal 100, 250, or 500. We also vary θ, and do so giving the benchmark and the alternative model comparable weight in predicting y_t. Specifically, we set

$$\theta_j = \begin{cases} \frac{c}{\sqrt{K_1}} & j = 1, \ldots, K_1 \\ \frac{c}{\sqrt{K_2 - K_1}} & j = K_1 + 1, \ldots, K_2 \end{cases}$$

with c equal to zero or one. When c is one, we're more likely to draw values of X and Y that make the estimated larger model more accurate than the benchmark, and when c is zero we're unlikely to draw such values of X and Y. For all of the studies, $L(e) = e^2$.

To study the accuracy of our theoretical approximations, we first estimate the coverage probabilities of OOS confidence intervals for $\mathrm{E}_R \bar{D}_R$ and $\mathrm{E}_T \bar{D}_T$. For each draw of X and

Y, we construct the one-sided OOS interval defined in Theorem 2:

$$[\bar{D}_R - 1.28\hat{\sigma}, \infty) \quad \text{with} \quad \hat{\sigma}^2 = \frac{1}{P}\sum_{t=R+1}^{T}(D_t - \bar{D}_R)^2$$

for $P = 10, \ldots, 2T/3$; we then calculate the percentage of simulations where these intervals contain $\mathrm{E}_R\,\bar{D}_R$ and the percentage that contain $\mathrm{E}_T\,\bar{D}_T$. Since the data are i.i.d., both of these quantities are easy to calculate (see Section 3.3). For these calculations, we draw 2000 samples for each combination of the design parameters.

Our second set of results studies whether these different OOS and in-sample statistics are valid for testing the null hypothesis

$$H_0: \quad E_T\bar{D}_T \leq 0,$$

namely that the benchmark model is expected to be more accurate in the future than the alternative. Informally, we are interested in whether the conditional probability

$$\Pr[\text{test rejects} \mid E_T\bar{D}_T \leq 0] \leq \alpha$$

where α is the nominal size of the test. If this inequality does not hold, the test statistic is rejecting the benchmark model too often.

We look at four different statistics—the full-sample F-test, the DMW t-test, the OOS t-test using McCracken's (2007) critical values,[15] and Clark and West's (2006, 2007) Gaussian out-of-sample statistic.[16] For the F-test, we simply test whether the coefficients on the larger model are nonzero. For the out-of-sample tests, we conduct a one-sided test of out-of-sample performance for every value of P as before. For each of these simulations, we discard draws from the DGP that violate the null hypothesis $E_T\bar{D}_T \leq 0$ and then use the remaining samples to calculate the conditional probability that each test rejects. The simulations end when 2000 draws have been retained for each choice of the design parameters.

4.2 Results

We discuss results for the confidence intervals first. Figures 2 and 3 show the coverage probability of these intervals as a function of P for each combination of T, K_1 and K_2, and c. The nominal coverage is 90% and is represented with a gray horizontal line. Each panel displays the coverage for a different choice of design parameters.

Figure 2 gives the results for $\mathrm{E}_R\,\bar{D}_R$. The actual coverage is very close to the nominal coverage except when P is very small. The poor behavior for small P is unsurprising, as

[15] These critical values are not published for $K_2 - K_1 > 10$, so we do not report them for $K_2 = T/10$.

[16] Clark and West (2006, 2007) derive their statistic using the rolling window estimation scheme. Here we use the same statistic, but with a fixed window scheme. A supplemental appendix presents results for their statistic, using the recursive window.

it simply means that the CLT is a poor approximation when the test sample is small. The interval for $T = 100$, $K = T/10$, and $c = 0$ is the worst, with actual coverage near 85% for most values of P/T. But the others are much closer to nominal coverage, even for the parsimonious models ($K_1 = 1$ and $K_2 = 3$) where one might expect the theoretical results to break down.

Figure 3 gives the results for $\mathrm{E}_T \bar{D}_T$. In columns 2 and 4—the overfit models—the coverage is near nominal coverage for moderately small values of P. As P increases to $2T/3$, the coverage increases above nominal size; near 100% for some DGPs. With the parsimonious model, the coverage is near nominal coverage for all P for the one-sided interval with $c = 1$, but only for moderately small P when $c = 0$.

The behavior for $K_2 = T/10$ is exactly what our theory predicts. When P^2/T is small, the coverage is close to nominal levels. The behavior as P increases, combined with the results for $\mathrm{E}_R \bar{D}_R$, indicate that typically $\mathrm{E}_T \bar{D}_T \geq E_R \bar{D}_R$. Since

$$E_T \bar{D}_T = E_R \bar{D}_R + (E_T \bar{D}_T - E_R \bar{D}_R),$$

and the interval is approximately centered at $\mathrm{E}_R \bar{D}_R$, the difference $\mathrm{E}_T \bar{D}_T - E_R \bar{D}_R$ adds a substantial positive quantity when P^2/T is not near zero, increasing the coverage of the one-sided interval.

We present the size simulations next, in Figures 1–7. For the OOS tests we plot each OOS test's conditional rejection probability,

$$\Pr[\text{test rejects} \mid E_T \bar{D}_T \leq 0],$$

for each combination of T, K_1, K_2, and c as a function of P. The F-test does not depend on P, so we calculate and tabulate it's conditional rejection probability as a single value for each combination of design parameters.

Table 1 summarizes the simulation results for the F-test. For $c = 0$, the estimated conditional rejection probability is almost exactly equal to the test's nominal size (10%), which is unsurprising. For $c = 0$, the F-test is exact; moreover, the larger model will almost always be less accurate than the smaller one, so conditioning on $\mathrm{E}_T \bar{D}_T \leq 0$ is almost unrestrictive. When c increases, though, the F-test overrejects badly—rejecting at roughly 50% when $c = 1$ for the parsimonious model and from 70% to 100% for the overfit model. This agrees with our Section 3.2 results and matches results seen in empirical practice: the F-test rejects the benchmark with very high probability, even though it is, by construction, more accurate than the alternative model.

Figure 4 presents the size estimates for the DMW OOS-t test. Again, different panels display results for different combinations of the design parameters. Each graph plots the rejection probability against P/T. For $K/T = 10$, the rejection probability falls as P/T increases, from near nominal size when P/T is small to zero when P/T is near $2/3$. Moreover, the rejection probability falls faster when T is large, as our theory predicts. When $K = 3$, the rejection probability stays closer to nominal size, but falls with P/T for $c = 0$, under-rejecting by about 5pp when $P/T = 2/3$, and rises with P/T for $c = 1$,

overrejecting by about 10pp when $P/T = 2/3$. For small P, the rejection probability is near 10% for all simulations (the farthest is $K = T/10$, $c = 0$, where the rejection probability is about 5%; the other simulations are much closer).

We observe the following patterns. The DMW test has close to nominal size when P is small for every combination of design parameters. In most cases, the rejection probability decreases as P/T increases—the exception is for $K_2 = 3$ and $c = 1$. For the large-K simulations, the rejection probability drops to zero for most of the simulations as P/T increases. The rejection probability increases with c, but the rejection probability still is near nominal probability for small P with $c = 1$.

Clark and West's (2006, 2007) statistic, presented in Figure 5, behaves quite differently. For $c = 0$ the test is correctly sized for both the overfit and parsimonious studies, as we saw for the F-test. When $c = 1$, the rejection probability increases rapidly with P/T. For $K_2 = 3$, the rejection probability is near 10% when P is small but about 40% when $P/T = 2/3$. For $K = T/10$, the rejection probability is even higher and increases with T as well, from a maximum over 50% when $T = 100$ to a maximum of nearly 100% when $T = 1000$.

Results using McCracken's (2007) critical values are presented in Figure 6 and are similar to those using Clark and West's test. For $c = 0$ the rejection probability is nearly the test's nominal size. For $c = 1$, the rejection probability increases with P/T, from close to the nominal size when P/T is small to over 25% when $P/T = 2/3$. Note that all of the simulations use the parsimonious model. McCracken's statistic overrejects here by slightly less than Clark and West's, but still by a substantial amount. Note that we are unable to plot results for McCracken's statistic when $K/T = 10$, which is where the breakdown in Clark and West's test is most pronounced.

Since the DMW test tends to have low rejection probability, the test's power is a concern. Figure 7 power results for the DMW test, simulating from (24) with $c = 1$ or 2 subject to the constraint that $\mathrm{E}_T \bar{D}_T > 0$.[17] Since the other test statistics greatly overreject, we do not present their power. For $c = 1$, the power is never greater than nominal size and decreases to zero as P/T increases for the overfit model. For $c = 2$ the power is better, increasing with P/T at first stretch and then decreasing as P/T grows beyond approximately 1/4 for the overfit model. Larger values of T give a higher peak and greater power overall, but the power still falls to nearly zero if P/T is too large (approximately 2/3 in our simulations). The power with the parsimonious model is typically quite low but greater than nominal size for $c = 2$.

Both sets of simulations support our theoretical results. The first simulation confirms that the DMW OOS t-test is centered at $\mathrm{E}_R \bar{D}_R$ for all choices of P and R and is centered on $\mathrm{E}_T \bar{D}_T$ only when P is small. The second simulation confirms that the DMW test has correct size for the null hypothesis that $\mathrm{E}_T \bar{D}_T \leq 0$ when P is small and that tests designed to test whether the benchmark is true, like the F-test and Clark and West's (2006, 2007) and McCracken's (2007) OOS tests can reject by much more than their

[17] Draws of X and Y with $\mathrm{E}_T \bar{D}_T > 0$ are very rare when $c = 0$, so we do not present results for that value of c.

nominal size when testing the null $\mathrm{E}_T \bar{D}_T \leq 0$. Moreover, these simulations demonstrate that the restriction that P be small is binding in practice, as the DMW test under-rejects and has very low power when P is too large.

5 Empirical analysis of equity premium predictability

This section presents a study of equity premium predictability based on Goyal and Welch (2008). Goyal and Welch conduct an out of sample analysis of 18 different variables thought, based on previous research, to predict the equity premium (calculated as the difference between the return on the S&P 500 index and the T-bill rate).[18] Their analysis has two notable features: Goyal and Welch compare different predictors across the same time periods and frequency as much as possible, making the accuracy of these predictors directly comparable; many of the papers that originally proposed these predictors used different time periods, methods, or observational frequencies so their results were not directly comparable. And Goyal and Welch compare these predictors out-of-sample; many of the original studies found in-sample evidence of equity premium predictability but did not look at out-of-sample evidence.

Goyal and Welch consider many different models; most of them are very simple—regression onto a constant and a single stochastic predictor—but some are more complicated. They find that essentially none of the models outperform a simple benchmark, the prevailing mean of the equity premium, and conclude that these variables do not predict the equity premium. A large literature has subsequently sought to explain or rebut these results, including several responses to Goyal and Welch's original working paper published in the same special issue of *The Review of Financial Studies* as Goyal and Welch (2008).[19]

Goyal and Welch (2008), as well as Bossaerts and Hillion (1999), Lettau and Van Nieuwerburgh (2008), and many other authors, propose that instability could explain the OOS failure of these models. However, we have shown in this paper that overfit can also explain this pattern, significant in-sample results that do not hold up out of sample. In this section, we explore the extent to which overfit is a potential concern in this data set and estimate the expected forecasting performance of the largest model they consider, a model with 13 regressors, using 81 observations (annual data from 1928 to 2009). The predictors are listed in Table 2; please see Goyal and Welch's original paper for detailed information about these variables.[20]

[18] Goyal and Welch (2008) builds on previous research by Bossaerts and Hillion (1999) and Goyal and Welch (2003).

[19] Those papers are Campbell and Thompson (2008), Cochrane (2008), Boudoukh et al. (2008), and Lettau and Van Nieuwerburgh (2008).

[20] Table 2 only lists the variables used in Goyal and Welch's (2008) "kitchen sink" model. Some of the variables that they use are excluded from this model either because the series are too short or because the variables are linear combinations of other variables.

Goyal and Welch (2008) focus primarily on univariate regression models of the form

$$r_{t+1} = \beta_0 + \beta_1 x_{it} + \varepsilon_{t+1}$$

using OLS with a recursive window, where r_{t+1} is the equity premium and x_{it} is one of the predictors listed in Table 2. This focus is consistent with much of the rest of the literature. But it is not clear that this approach—using a large number of restricted models—is any more reliable than using a single, large model. Very few papers in the equity-premium predictability literature explicitly account for the multiplicity of model comparisons[21] and most widely-used statistics for multiple comparisons are derived under the assumption that there is a finite number of hypotheses or models (as is the case in Sullivan et al., 1999, White, 2000, Hansen, 2005, and Lehmann and Romano, 2005), so there is little theoretical evidence that they are any more reliable than a regression model that encompasses all of the smaller models. Moreover, since Goyal and Welch (2008) (along with most of the rest of the literature) want to interpret these univariate regressions as meaningful statements about the true relationships between the equity premium and the regressors, omitted variable bias is a serious potential issue.

For example, Table 3 presents two full-sample coefficient estimates for each of the 12 predictors we consider.[22] The first column is the estimator of the coefficient on the variable in the full regression,

$$r_{t+1} = \beta_0 + \sum_{i=1}^{12} \beta_i x_{it} + \varepsilon_{t+1},$$

and the third column is the p-value associated with a two-sided t-test that the coefficient is zero in population. The second column is the estimator of the coefficient from the univariate regression

$$r_{t+1} = \alpha_i + \gamma_i x_{it} + \varepsilon_{t+1},$$

and the fourth column is the p-value associated with its t-test.[23] Some of the coefficient estimates agree, but some do not. The coefficient estimate for the *default yield spread*, for example, is substantially and significantly negative in the full model (−7.79 with p-value 0.042), but is near zero in the univariate regression (0.79 with p-value 0.720). This pair of results implies that the the default yield spread contains information about the equity premium but is also correlated with other poor predictors, and that this additional correlation adds noise to the univariate regression and masks the true relationship. Similarly, *net equity expansion* is significant at 10% in the univariate regression but not in the full model (p-values of 0.060 and 0.626 respectively), which is likely attributable to omitted variable bias. So there is merit to studying a large model that includes all of the variables.

[21] Rapach and Wohar (2006), Rapach and Zhou (2012), and Calhoun (2013) are exceptions.

[22] For all of our full sample results, we studentize the regressors to make it easier to compare coefficient estimates and we express the equity premium in basis points.

[23] All of the standard errors were calculated using a Newey-West kernel with two lags.

Results for the full sample estimates of these models are in Table 3.[24,25] The last rows of the table list measures of the full-model fit. The p-value for the test of full model fit is very small (less than 0.01), indicating that some of the coefficients are nonzero in population and at least one of these predictors is correlated with the equity premium.[26] As we argue throughout the paper, this result does not imply that the model will forecast well.

To determine whether the full model can forecast well, we use the DMW OOS-t test to compare it to a sample mean benchmark,

$$r_{t+1} = \mu + \varepsilon_{1,t+1}.$$

Both models are estimated by OLS using the fixed-window scheme. We also present results for a restricted model proposed by Campbell and Thompson (2008) that imposes that $\hat{r}_{t+1}$ be non-negative for each forecast. To study the effect of the training sample size on the DMW statistic, we calculate the one-sided confidence interval for $\mathrm{E}_T \bar{D}_T$ given by (15) corresponding to the null and alternative hypotheses

$$H_0: \quad E_T \bar{D}_T \leq 0 \qquad H_A: \quad E_T \bar{D}_T > 0$$

using the fixed-window scheme for each value of R between 20 and $T-10$. The standard deviation is estimated using a Newey-West estimator with $\lfloor P^{1/4} \rfloor$ lags. For small values of R, the OOS average is expected to underestimate the performance of the larger model relative to the smaller, but this may not hold in this particular dataset. For the OOS results, we express the equity premium in percentage points.

Figure 8 plots the OLS results and Figure 9 imposes Campbell and Thompson's (2008) restriction. The solid line in each figure shows the OOS average, $\bar{D}_R$, and the shaded region indicates the 95% one-sided confidence interval implied by the DMW test. Negative numbers indicate that the full model has higher out-of-sample loss. We can see that the same patterns hold for both models: the performance difference decreases as R grows, but the full model is never more accurate. We also see that the performance difference decreases suddenly over the period $R = 29$ to $R = 34$ (corresponding to the years 1956–1961). Figure 10 plots the accuracy of the individual forecasts (only for the linear models) and shows that this change is the result of a sudden improvement in

[24] Calculations in this section are done in R (R Development Core Team, 2010) using the *lmtest* (Zeileis and Hothorn, 2002) and *sandwich* (Zeileis, 2004) packages.

[25] We compare the test statistic to critical values from the F-distribution, which have been shown to be more reliable than Chi-squared critical values when there are many regressors (Anatolyev, 2012; Calhoun, 2011).

[26] We have done additional analysis to try to identify which predictors were correlated with the equity premium, but none of the individual regressors were significant after correcting for multiplicity. The Bonferroni correction, for example, suggests that an individual p-value would need to be less than $0.10/12 \approx 0.0083$ for its corresponding coefficient to be significant at the 10% level, but the smallest p-value is 0.035. One can improve on the Bonferroni correction, of course, but a comprehensive analysis is beyond the scope of this paper.

the full model. This change may indicate instability in the underlying relationship, as proposed by Goyal and Welch (2008).

In summary, we fail to reject the null that the benchmark prevailing mean model is more accurate than the full model including all of Goyal and Welch's (2008) predictors. This result is consistent with Goyal and Welch's original analysis. Unlike Goyal and Welch, we attribute this result, at least in part, to parameter uncertainty—the full sample results indicate that there is a true predictive relationship between some of these variables and the equity premium and the larger model could predict better than the benchmark with enough data.[27] These also indicate that combination or shrinkage estimators of the full model have the potential to significantly improve on the benchmark.[28]

6 Conclusion

This paper gives a theoretical motivation for using OOS comparisons: the DMW OOS test allows a forecaster to conduct inference about the expected future accuracy of his or her models when one or both is overfit. We show analytically and through Monte Carlo that standard full-sample test statistics can not test hypotheses about this performance.

Our paper also shows that popular test and training sample sizes may give misleading results if researchers are concerned about overfit. We show that P^2/T must converge to zero for the DMW test to give valid inference about the expected forecast accuracy, otherwise the test measures the accuracy of the estimates constructed using only the training sample. In empirical research, P is typically much larger than this. Our simulations indicate that using large values of P with the DMW test gives undersized tests with low power, so this practice may favor simple benchmark models too much. Existing corrections, proposed by Clark and McCracken (2001, 2005), McCracken (2007) and Clark and West (2006, 2007), seem to correct too much, though, and reject too often when the benchmark model is more accurate.

More work remains. The requirement that P^2/T converge to zero is limiting, as it implies that in typical macroeconomic datasets, only a handful of observations should be used for testing. This requirement can be relaxed only slightly; $P = O(T^{1/2})$ is required for the OOS test to have nontrivial power in general, but there are loss functions and DGPs for which some relaxation is possible. This constraint could be mitigated by extending our results to cross-validation or other resampling strategies, or by constructing full-sample statistics that allow inference about $\mathrm{E}_T \bar{D}_T$. It would also be useful to extend our results to other forecasting models and to explore how stationarity could be relaxed, but such extensions are less important than improving the available statistics.

[27] Bacchetta et al. (2010) make a similar point about exchange rate models, but see also Chinn (2010) and Giannone (2010).

[28] See Rapach and Zhou (2012) for a recent review of this literature.

Appendix: mathematical details

Supporting results

The results in this paper rely heavily on a coupling argument for absolutely regular sequences, Berbee's Lemma (Berbee, 1979). Many of the results of this paper (Lemma 1 and Theorem 1) use modifications of existing results for NED functions of mixing processes by de Jong (1997) and de Jong and Davidson (2000); this coupling argument is used to explicitly derive inequalities that arise naturally for NED processes. Lemma A3 establishes these inequalities, which are based on a proposition of Merlevède and Peligrad (2002).

We present Merlevède and Peligrad's (2002) statement of Berbee's Lemma for the reader's reference. In the following Lemma, $\beta(X,Y)$ is the coefficient of absolute regularity:

$$\beta(X,Y) = \sup_{A\in\sigma(Y)} \mathrm{E}|\Pr(A \mid \sigma(X)) - \Pr(A)|. \tag{25}$$

Lemma A1.

> *Let X and Y be random variables defined on a probability space $(\Omega,\mathscr{T},\Pr)$ with values in a Polish space S. Let $\sigma(X)$ be a σ-algebra generated by X and let U be a random variable uniformly distributed on $[0,1]$ independent of (X,Y). Then there exists a random variable Y^* measurable with respect to $\sigma(X)\vee\sigma(Y)\vee\sigma(U)$, independent of X and distributed as Y, and such that $\Pr(Y\neq Y^*)=\beta(X,Y)$.*
> *(Merlevède and Peligrad, 2002)*

The advantage of this result over coupling arguments that use other forms of weak dependence is that the difference between the original variable, Y, and the new variable, Y^*, does not depend on their dimension. Similar results for strong mixing sequences depend on the dimension of Y, which makes them unsuitable for this paper.

Lemma A2. *Suppose that X and X^* are L_p-bounded random variables, with $p>2$, that satisfy $\Pr[X\neq X^*]=c$. Then*

$$\|X-X^*\|_2 \leq 2^{1/p}(\|X\|_p + \|X^*\|_p)c^{(p-2)/2p} \tag{26}$$

The proof is virtually identical to the proof of Proposition 2.3 in Merlevède and Peligrad (2002) and is omitted.

Lemma A3. *Suppose Assumptions 1–3 hold. Then, for any T, s, t, and u with $s<t\leq u$, there exist random variables $D_t^*,\ldots,D_u^*$ such that*

$$P[(D_t^*,\ldots,D_u^*)\neq(D_t,\ldots,D_u)] \leq \beta_{t-s} \tag{27}$$

and

$$\mathrm{E}(\phi(D_t^*,\ldots,D_u^*)\mid\mathscr{F}_s)=\int\phi(D_t,\ldots,D_u)f(\mathbf{x},\mathbf{y})\,d\mathbf{x}\,d\mathbf{y} \tag{28}$$

almost surely for all measurable functions ϕ such that the expectations are finite, where

$$\mathbf{x}=(x_t',\ldots,x_u')',\qquad \mathbf{y}=(y_{t+h},\ldots,y_{u+h})',$$

and f is the joint density of $(\mathbf{x},\mathbf{y})$. Moreover,

$$\|D_v^*-D_v\|_2\le 2^{1+1/\rho}B_L\beta_{t-s}^{(\rho-2)/2\rho},\qquad v=t,\ldots,u. \tag{29}$$

Proof. The proof follows as a consequence of Lemmas A1 and A2. Let $l=u-t$. For any fixed values of l and T, the sequence of vectors

$$V_t=(y_{t+h},x_t',\ldots,y_{t+l+h},x_{t+l}')$$

is absolutely regular of size $\rho/(\rho-2)$. Berbee's Lemma implies that there is a random vector V^* that is independent of $\mathscr{F}_s$, equal to V_t in distribution, and satisfies

$$\Pr[V^*\neq V_t]=\beta_{t-s}.$$

Now define

$$D_v^*=\begin{cases}L(y_{v+h}^*-x_{1v}^{*\prime}\hat\theta_{1R})-L(y_{v+h}^*-x_{2v}^{*\prime}\hat\theta_{2R}) & v\le T\\ L(y_{v+h}^*-x_{1v}^{*\prime}\hat\theta_{1T})-L(y_{v+h}^*-x_{2v}^{*\prime}\hat\theta_{2T}) & v>T\end{cases}$$

with y_{v+h}^* and x_{iv}^* denoting the elements of V^* corresponding to y_{v+h} and x_{iv} in V_t. Equations (27) and (28) are satisfied by construction, and (29) follows from Lemma A2. □

Lemma A4. *Suppose Assumptions 1–3 hold. Let b_T be a sequence of integers such that $b_T\to\infty$ and $b_T=o(P)$ and define*

$$Z_i=P^{-1/2}\sum_{s=R+(i-1)b_T+1}^{R+ib_T}[D_s-\mathrm{E}_R\,D_s]. \tag{30}$$

Then

$$\sum_{i=1}^{\lfloor P/b_T\rfloor}(\mathrm{E}_R\,Z_i^2-\mathrm{E}_{R+(i-1)b_T}\,Z_i^2)\to^p 0\quad \text{as } P\to\infty. \tag{31}$$

If the assumptions of Theorem 1 also hold, b_T is restricted further so that $b_T\equiv\lfloor\gamma/\delta\rfloor$ for some positive scalar δ, and we define

$$\eta_\delta(x)\equiv\delta^{-1}(2\pi)^{-1/2}e^{-x^2/2\delta^2}, \tag{32}$$

then

$$\sum_{t=-P+R+1}^{2P+R} (Z_{1t}Z_{2t} - \mathrm{E}_R\, Z_{1t}Z_{2t}) \to^p 0. \tag{33}$$

where

$$Z_{1t} = (P\gamma)^{-1/2} \sum_{l=\max(R+1-t,-b_T)}^{\min(R+P-t,b_T)} (D_{t+l} - \mathrm{E}_R\, D_{t+l}) W(l/\gamma), \tag{34}$$

and

$$Z_{2t} = (P\gamma)^{-1/2} \sum_{j=\max(R+1-t,-b_T)}^{\min(R+P-t,b_T)} (D_{t+j} - \mathrm{E}_R\, D_{t+j}) \eta_\delta(j/\gamma). \tag{35}$$

Proof. The first result, (31), follows a similar argument to Lemma 5 of de Jong (1997) and (33) to Lemma A.4 of de Jong and Davidson (2000). Since these arguments are similar and our modification is the same for both, we'll just present the more complicated version, (33).

Note that $\{Z_{1t}^2 P\gamma_T/b_T\}$ and $\{Z_{2t}^2 P\gamma_T/b_T\}$ are uniformly integrable. As in de Jong and Davidson (2000, Lemma A.4), we can assume that there is a constant C such that Z_{1t} and Z_{2t} are bounded in absolute value by $C\sqrt{b_T/P\gamma_T}$; uniform integrability ensures that the difference between the unbounded random variables and these truncated versions is negligible for large enough values of C.

Let $r = \lfloor 3P/2b_T \rfloor$ and rewrite the summation as

$$\begin{aligned}
\sum_{t=-P+R+1}^{2P+R} (Z_{1t}Z_{2t} - \mathrm{E}_R\, Z_{1t}Z_{2t}) = &\sum_{i=1}^{r} \sum_{t=(2i-2)b_T-P+R+1}^{(2i-1)b_T-P+R} (Z_{1t}Z_{2t} - \mathrm{E}_R\, Z_{1t}Z_{2t}) \\
&+ \sum_{i=1}^{r} \sum_{t=(2i-1)b_T-P+R+1}^{2ib_T-P+R} (Z_{1t}Z_{2t} - \mathrm{E}_R\, Z_{1t}Z_{2t}) \\
&+ \sum_{t=rb_T-P+R+1}^{2P+R} (Z_{1t}Z_{2t} - \mathrm{E}_R\, Z_{1t}Z_{2t}) \\
&\equiv \sum_{i=1}^{r} (U_i - \mathrm{E}_R\, U_i) + \sum_{i=1}^{r} (U_i' - \mathrm{E}_R\, U_i') + o_{L_1}(1).
\end{aligned}$$

The proof then holds if we can show that both U_i and U_i' obey LLNs. We'll do so by proving that $\{U_i - \mathrm{E}_R\, U_i, \mathscr{F}_{(2i-1)b_T+R-P}\}$ and $\{U_i' - \mathrm{E}_R\, U_i', \mathscr{F}_{(2i)b_T+R-P}\}$ are L_2-mixingales of size $-1/2$ and using the bound $\mathrm{E}(\sum_{i=1}^{r}(U_i - \mathrm{E}_R\, U_i))^2 = O(\sum_{i=1}^{r} c_i^2)$ where the c_i are the mixingale magnitude indices (McLeish, 1975a).

For non-negative m, we have

$$U_i - \mathrm{E}_R\, U_i \in \mathscr{F}_{(2i+2m-1)b_T+R-P},$$

establishing half of the mixingale result trivially. Now fix i and $m > 0$ and use Lemma A3 to define D^*_{ts} for each $t = (2i-2)b_T - P + R + 1, \ldots, (2i-1)b_T - P + R$ and $s = \max(t - b_T, R+1), \ldots, \min(t + b_T, T - h)$ such that

$$\mathrm{E}_R\, D^*_{ts} = \mathrm{E}_{(2i-2m-1)b_T+R-P}\, D^*_{ts} \quad a.s.$$

and

$$\|D^*_{ts} - D_s\|_2 \leq 2^{(1+\rho)/\rho} B_L\, \beta^{(\rho-2)/2\rho}_{s-(2i-2m-1)b_T+P}.$$

Also define

$$Z^*_{1t} = (P\gamma_T)^{-1/2} \sum_{l=\max(R+1-t,-b_T)}^{\min(R+P-t,b_T)} (D^*_{t,t+l} - \mathrm{E}_R\, D^*_{t,t+l})\; W(l/\gamma_T),$$

and

$$Z^*_{2t} = (P\gamma_T)^{-1/2} \sum_{j=\max(R+1-t,-b_T)}^{\min(R+P-t,b_T)} (D^*_{t,t+l} - \mathrm{E}_R\, D^*_{t,t+l})\; \eta_\delta(j/\gamma_T).$$

Now, we have the inequalities

$$\begin{aligned}
&\|\mathrm{E}(U_i - \mathrm{E}_R\, U_i \mid \mathscr{F}_{(2i-2m-1)b_T+R-P})\|_2 \\
&\qquad \leq \sum_{t=(2i-2)b_T-P+R+1}^{(2i-1)b_T-P+R} \|\mathrm{E}(Z_{1t}Z_{2t} \mid \mathscr{F}_{(2i-2m-1)b_T+R-P}) - \mathrm{E}_R\, Z_{1t}Z_{2t}\|_2 \\
&\qquad = \sum_{t=(2i-2)b_T-P+R+1}^{(2i-1)b_T-P+R} \|\mathrm{E}(Z_{1t}Z_{2t} \mid \mathscr{F}_{(2i-2m-1)b_T+R-P}) \\
&\qquad\quad - \mathrm{E}(Z^*_{1t}Z^*_{2t} \mid \mathscr{F}_{(2i-2m-1)b_T+R-P}) \\
&\qquad\quad + \mathrm{E}(Z^*_{1t}Z^*_{2t} \mid \mathscr{F}_{(2i-2m-1)b_T+R-P}) - \mathrm{E}_R\, Z_{1t}Z_{2t}\|_2 \\
&\qquad \leq 2 \sum_{t=(2i-2)b_T-P+R+1}^{(2i-1)b_T-P+R} \|Z_{1t}Z_{2t} - Z^*_{1t}Z^*_{2t}\|_2 \\
&\qquad \leq 2 \sum_{t=(2i-2)b_T-P+R+1}^{(2i-1)b_T-P+R} (\|Z_{1t} - Z^*_{1t}\|_2 \|Z_{2t}\|_\infty + \|Z_{2t} - Z^*_{2t}\|_2 \|Z^*_{1t}\|_\infty) \\
&\qquad \leq \frac{2C b_T^{1/2}}{(P\gamma_T)^{1/2}} \sum_{t=(2i-2)b_T-P+R+1}^{(2i-1)b_T-P+R} (\|Z_{1t} - Z^*_{1t}\|_2 + \|Z_{2t} - Z^*_{2t}\|_2).
\end{aligned}$$

And we can finish the proof with the following inequalities:

$$\begin{aligned}
&\frac{2C b_T^{1/2}}{(P\gamma_T)^{1/2}} \sum_{t=(2i-2)b_T-P+R+1}^{(2i-1)b_T-P+R} \|Z_{1t} - Z_{1t}^*\|_2 \\
&\qquad \le \frac{4C b_T^{1/2}}{P\gamma_T} \sum_{t=(2i-2)b_T-P+R+1}^{(2i-1)b_T-P+R} \sum_{l=\max(R+1-t,-b_T)}^{\min(R+P-t,b_T)} \|D_{t+l} - D_{t,t+l}^*\|_2 W(l/\gamma_T) \\
&\qquad \le O\Big(\frac{b_T^{1/2}}{P\gamma_T}\Big) \sum_{t=(2i-2)b_T-P+R+1}^{(2i-1)b_T-P+R} \sum_{l=\max(R+1-t,-b_T)}^{\min(R+P-t,b_T)} \beta_{t+l-(2i-2m-1)b_T+P}^{(\rho-2)/2\rho} \\
&\qquad = O\Big(\frac{b_T^{1/2}}{P\gamma_T}\Big) O(b_T^{3/2-u} m^{-1/2-u})
\end{aligned}$$

for some positive u. The same argument holds for Z_{2t}. As a result,

$$E\Big(\sum_{i=1}^{r}(U_i - \mathrm{E}_R U_i)\Big)^2 = o\Big(\sum_{i=1}^{r} b_T^2/P\gamma_T\Big) = o(b_T/\gamma_T) \to 0,$$

as required. □

Lemma A5. *Suppose the conditions of Theorem 1 hold. Then*

$$\hat{\sigma}^2 - P^{-1} \sum_{s,t=R+1}^{T-h} (D_s - \mathrm{E}_R D_s)(D_t - \mathrm{E}_R D_t) W((t-s)/\gamma) \to^{L_1} 0. \tag{36}$$

Proof. It follows from simple algebra that

$$\begin{aligned}
\Big|\hat{\sigma}^2 - P^{-1} \sum_{s,t=R+1}^{T-h} (D_s - \mathrm{E}_R D_s)(D_t - \mathrm{E}_R D_t) W((t-s)/\gamma)\Big| \le \\
P^{-1} \sum_{s,t=R+1}^{T-h} |(D_s - \mathrm{E}_R D_s)(\mathrm{E}_R D_t - \bar{D}_R)| W((t-s)/\gamma) \\
+ P^{-1} \sum_{s,t=R+1}^{T-h} |(D_s - \bar{D}_R)(\mathrm{E}_R D_t - \bar{D}_R)| W((t-s)/\gamma) + o_p(1).
\end{aligned}$$

We'll prove that these two sums are $o_p(1)$; uniform integrability then implies convergence in L_1. The arguments for each are almost identical, so we'll only present the first.

Applying the Cauchy-Schwarz inequality twice and simplifying gives the upper bound

$$\begin{aligned}
&P^{-1} \sum_{s,t=R+1}^{T-h} |(D_s - \mathrm{E}_R D_s)(\mathrm{E}_R D_t - \bar{D}_R)| W((t-s)/\gamma) \\
&\qquad \le O(1)\Big[P^{-1} \sum_{s=R+1}^{T-h} (D_s - \mathrm{E}_R D_s)^2\Big]^{1/2} \Big[P^{-1} \sum_{s=R+1}^{T-h} (\mathrm{E}_R D_s - \bar{D}_R)^2\Big]^{1/2}.
\end{aligned}$$

Now, $P^{-1}\sum_{s=R+1}^{T-h}(D_s - \mathrm{E}_R D_s)^2 = O_p(1)$, and it suffices to prove that

$$P^{-1}\sum_{s=R+1}^{T-h}(\mathrm{E}_R D_s - \bar{D}_R)^2 = o_p(1).$$

Observe that

$$\begin{aligned}P^{-1}\sum_{s=R+1}^{T-h}(\mathrm{E}_R D_s - \bar{D}_R)^2 &= O_p\big(P^{-1}\sum_{s=R+1}^{T-h}(\mathrm{E}_R D_s - \mathrm{E}_R \bar{D}_R)^2\big) + O_p(\bar{D}_R - \mathrm{E}_R \bar{D}_R)^2 \\ &= O_p\big(P^{-1}\sum_{s=R+1}^{T-h}(\mathrm{E}_R D_s)^2 - (\mathrm{E}_R \bar{D}_R)^2\big) + o_p(1),\end{aligned}$$

with the second term $o_p(1)$ by Lemma 1 and Davidson's (1993) mixingale LLN.

Now define D_s^*, $s = R+1, \ldots, T-h$, as in Lemma A3 so that $\mathrm{E}_{s-1} D_s^* = \mathrm{E}_R D_s^*$ almost surely. Note that we also have the equality $\mathrm{E}_R D_s^* = \mathrm{E}_R D_{R+1}^*$ almost surely for all $s \geq R+1$, and so

$$P^{-1}\sum_{s=R+1}^{T-h}(\mathrm{E}_R D_s^*)^2 = \big(P^{-1}\sum_{s=R+1}^{T-h}\mathrm{E}_R D_s^*\big)^2 \quad \text{a.s.}$$

Consequently,

$$\begin{aligned}&P^{-1}\sum_{s=R+1}^{T-h}(\mathrm{E}_R D_s)^2 - (\mathrm{E}_R \bar{D}_R)^2 \\ &\qquad = P^{-1}\sum_{s=R+1}^{T-h}[(\mathrm{E}_R D_s)^2 - (\mathrm{E}_R D_s^*)^2] + \big(P^{-1}\sum_{s=R+1}^{T-h}\mathrm{E}_R D_s^*\big)^2 - (\mathrm{E}_R \bar{D}_R)^2 \quad \text{a.s.} \\ &\qquad = O_p\big(P^{-1}\sum_{s=R+1}^{T-h}[(\mathrm{E}_R D_s)^2 - (\mathrm{E}_R D_s^*)^2]\big) + O_p\big(P^{-1}\sum_{s=R+1}^{T-h}\mathrm{E}_R(D_s - D_s^*)\big).\end{aligned}$$

Finally,

$$\begin{aligned}\big\|P^{-1}\sum_{s=R+1}^{T-h}[(\mathrm{E}_R D_s)^2 - (\mathrm{E}_R D_s^*)^2]\big\|_1 &\leq P^{-1}\sum_{s=R+1}^{T-h}\|\mathrm{E}_R(D_s - D_s^*)\|_2\|\mathrm{E}_R(D_s + D_s^*)\|_2 \\ &\leq (4B_L/P)\sum_{s=R+1}^{T-h}\|\mathrm{E}_R(D_s - D_s^*)\|_2,\end{aligned}$$

and this last term vanishes as in the proof of Lemma A4, completing the proof. □

Proof of Lemma 1

We will start by proving that $\{D_t - \mathrm{E}_R D_t, \mathscr{F}_t\}$ is an L_2-mixingale of size $-1/2$; D_t is $\mathscr{F}_t$-measurable and so it suffices to prove that

$$\|\mathrm{E}_{t-l} D_t - \mathrm{E}_R D_t\|_2 \leq 2^{2+1/\rho} B_L \beta_l^{(\rho-2)/2\rho} \tag{37}$$

for $l = 1, \ldots, t-R$, since $\beta_l^{(\rho-2)/2\rho} = O(l^{-1/2-\delta})$ for some $\delta > 0$ by assumption. Fix R, t and l and use Lemma A3 to define D_t^* so that

$$\mathrm{E}_R D_t^* = \mathrm{E}_{t-l} D_t^* \quad a.s.$$

and

$$\|D_t - D_t^*\|_2 \leq 2^{(1+\rho)/\rho} B_L \beta_l^{(\rho-2)/2\rho}.$$

Then

$$\begin{aligned}\|\mathrm{E}_{t-l} D_t - \mathrm{E}_R D_t\|_2 &\leq \|\mathrm{E}_{t-l} D_t - \mathrm{E}_{t-l} D_t^*\|_2 + \|\mathrm{E}_{t-l} D_t^* - \mathrm{E}_R D_t\|_2 \\ &\leq 2\|D_t - D_t^*\|_2 \\ &\leq 2^{2+1/\rho}\ B_L\ \beta_l^{(\rho-2)/2\rho}.\end{aligned}$$

□

Asymptotic normality follows from (37) using a modification of de Jong's (1997) CLTs for mixingale and NED arrays. Define

$$Z_i = P^{-1/2} \sum_{s=R+(i-1)b_T+1}^{R+ib_T} [D_s - \mathrm{E}_R D_s]$$

where b_T is a sequence that satisfies $b_T \leq P$, $b_T \to \infty$, and $b_T/P \to 0$. The same arguments used in de Jong's (1997) Theorem 1 show that

$$\sum_{i=1}^{\lfloor P/b_T \rfloor} Z_i = P^{-1/2} \sum_{s=R+1}^{T-h} (D_t - \mathrm{E}_R D_t) + o_p(1).$$

and

$$\sum_{i=1}^{\lfloor P/b_T \rfloor} Z_i = \sum_{i=1}^{\lfloor P/b_T \rfloor} (Z_i - E_{R+(i-1)b_T} Z_i) + o_p(1).$$

Note that $\{Z_i - E_{R+(i-1)b_T} Z_i, \mathscr{F}_{R+ib_T}\}_i$ is an MDS by construction, so Hall and Heyde's (1980) Theorem 3.2 and Corollary 3.1 ensure that $\sigma^{-1} \sum_{i=1}^{\lfloor P/b_T \rfloor} Z_i \to^d N(0,1)$ as long as

$$\sigma^2 - \sum_{i=1}^{\lfloor P/b_T \rfloor} \mathrm{E}_R Z_i^2 \to^p 0,$$

and

$$\sum_{i=1}^{\lfloor P/b_T \rfloor} \mathrm{E}_R Z_i^2 - \sum_{i=1}^{\lfloor P/b_T \rfloor} \mathrm{E}_{R+(i-1)b_T} Z_i^2 \to^p 0.^{29}$$

The first equation holds as in de Jong (1997) (see the proof of his Theorem 2); the second is ensured by Lemma A4. □

[29] Note that $\sigma^2 \in \mathscr{F}_t$ for all $t \geq R$, so Hall and Heyde's condition (3.21) is unnecessary—see the remarks after their result.

Proof of Lemma 2

Equation (11) holds if we show

$$\mathrm{E}_R \bar{D}_R = \mathrm{E}(L(y^* - x_1^{*\prime}\hat{\theta}_{1R}) - L(y^* - x_2^{*\prime}\hat{\theta}_{2R}) \mid \hat{\theta}_R) + o_p(P^{-1/2}), \tag{38}$$

$$\mathrm{E}_T \bar{D}_T = \mathrm{E}(L(y^* - x_1^{*\prime}\hat{\theta}_{1T}) - L(y^* - x_2^{*\prime}\hat{\theta}_{2T}) \mid \hat{\theta}_T) + o_p(Q^{-1/2}), \tag{39}$$

and

$$\mathrm{E}(L(y^* - x_i^{*\prime}\hat{\theta}_{iR}) \mid \hat{\theta}_R) - \mathrm{E}(L(y^* - x_i^{*\prime}\hat{\theta}_{iT}) \mid \hat{\theta}_T) = O_p(\sqrt{P/T}), \tag{40}$$

where $\hat{\theta}_R = (\hat{\theta}_{1R}, \hat{\theta}_{2R})$, $\hat{\theta}_T = (\hat{\theta}_{1T}, \hat{\theta}_{2T})$, and y^*, x_1^* and x_2^* are random variables drawn from the joint distribution of $(y_{t+h}, x_{1t}, x_{2t})$ independently of $\mathscr{F}_T$.

Proof of (38) *and* (39). For (38), define D_t^* for each $t = R+1, R+2, \ldots, T-h$ so that

$$\|D_t^* - D_t\|_2 \leq 2^{(1+\rho)/\rho} B_L \beta_{t-R}^{(\rho-2)/2\rho}$$

and

$$\mathrm{E}_R D_t^* = \mathrm{E}(L(y^* - x_1^{*\prime}\hat{\theta}_{1R}) - L(y^* - x_2^{*\prime}\hat{\theta}_{2R}) \mid \hat{\theta}_R) \qquad a.s.$$

Lemma A3 ensures that these D_t^* exist. Now,

$$\begin{aligned}
\left\| \mathrm{E}_R \bar{D}_R - \mathrm{E}\left(P^{-1} \sum_{t=R+1}^{T-h} D_t^* \mid \hat{\theta}_R\right) \right\|_2 &= \left\| \mathrm{E}_R \left(\bar{D}_R - P^{-1} \sum_{t=R+1}^{T-h} D_t^*\right) \right\|_2 \\
&\leq P^{-1} \sum_{t=R+1}^{T-h} \|D_t - D_t^*\|_2 \\
&= O(P^{-1}) \sum_{t=R+1}^{T-h} \beta_{t-R}^{(\rho-2)/2\rho}
\end{aligned}$$

and this last term is $o(P^{-1/2})$ by assumption. Essentially the same argument proves (39) as well. □

Proof of (40). Assumption 3 and the definition of the OLS estimator ensure that

$$\begin{aligned}
\|L(y^* - x_i^{*\prime}\hat{\theta}_{iR}) - L(y^* - x_i^{*\prime}\hat{\theta}_{iT})\|_1 &\leq B_L \|x_i^{*\prime}(\hat{\theta}_{iR} - \hat{\theta}_{iT})\|_2 \\
&\leq B_L \, \|x_i^{*\prime}[(X_{iT}'X_{iT})^{-1} - (X_{iR}'X_{iR})^{-1}]X_{iR}'\varepsilon_{iR}\|_2 \\
&\quad + B_L \, \|x_i^{*\prime}(X_{iT}'X_{iT})^{-1}[X_{iT}'\varepsilon_{iT} - X_{iR}'\varepsilon_{iR}]\|_2.
\end{aligned}$$

To simplify notation, define $V = (X_{iT}'X_{iT})^{-1} - (X_{iR}'X_{iR})^{-1}$ and $W = X_{iR}'\varepsilon_{iR}\varepsilon_{iR}'X_{iR}$. The first term in the upper bound satisfies

$$\begin{aligned}
\|x_i^{*\prime}VX_{iR}'\varepsilon_{iR}\|_2^2 &\leq \lambda_{K_i}(\mathrm{E}\, x_i^* x_i^{*\prime}) \cdot \mathrm{E}\,\mathrm{tr}\,(V^2 W) \\
&= O(1) \cdot \mathrm{E}\,\mathrm{tr}\{V^2\, \mathrm{E}(W \mid X_{iR}'X_{iR}, x_{i,R-h+1}x_{i,R-h+1}', \ldots, x_{i,T-h}x_{i,T-h}')\}
\end{aligned}$$

by assumption. Observe that

$$\begin{aligned}\mathrm{E}\,\mathrm{tr}\{V^2\,\mathrm{E}(W \mid X_{iR}'X_{iR}, x_{i,R-h+1}x_{i,R-h+1}', \ldots, x_{i,T-h}x_{i,T-h}')\} \\ \leq \Big\|\sum_{i=1}^{K_i}\lambda_i^2(V)\Big\|_{3/2}\|\lambda_{K_i}(\mathrm{E}(W \mid X_{iR}'X_{iR}, x_{i,R-h+1}x_{i,R-h+1}', \ldots, x_{i,T-h}x_{i,T-h}'))\|_3.\end{aligned}$$

The second term in this product is $O(R)$ by Assumption 2. To bound the first term, note that $(X_{iT}'X_{iT})^{-1} - (X_{iR}'X_{iR})^{-1}$ has rank P and each of its nonzero eigenvalues is bounded in absolute value by the eigenvalues of $(X_{iR}'X_{iR})^{-1}$. The eigenvalues of $(X_{iR}'X_{iR})^{-1}$ are $O_{L_3}(1/R)$ by Assumption 2, so

$$\Big\|\sum_{j=1}^{K_i}\lambda_j^2(V)\Big\|_{3/2} \leq \Big\|\sum_{j=K_i-P+1}^{K_i}\lambda_j^2(V)\Big\|_{3/2} = O(P/R^2).$$

Consequently,

$$\mathrm{E}\,\mathrm{tr}\left\{[(X_{iT}'X_{iT})^{-1} - (X_{iR}'X_{iR})^{-1}]^2X_{iR}'\varepsilon_{iR}\varepsilon_{iR}X_{iR}\right\} = O(P/R)$$

and so

$$\|x_i^{*\prime}[(X_{iT}'X_{iT})^{-1} - (X_{iR}'X_{iR})^{-1}]X_{iR}'\varepsilon_{iR}\|_2 = O(\sqrt{P/R}).$$

A similar argument proves that

$$\|x_i^{*\prime}(X_{iR}'X_{iR})^{-1}[X_{iT}'\varepsilon_{iT} - X_{iR}\varepsilon_{iR}]\|_2 = O(\sqrt{P/R}),$$

completing the proof. □

Proof of Theorem 1

We can rewrite the centered OOS-t statistic as

$$\sqrt{P}(\bar{D}_R - \mathrm{E}_T\,\bar{D}_T)/\hat{\sigma} = \frac{\sigma}{\hat{\sigma}}\left(\sqrt{P}(\bar{D}_R - \mathrm{E}_R\,\bar{D}_R)/\sigma + \sqrt{P}(\mathrm{E}_R\,\bar{D}_R - \mathrm{E}_T\,\bar{D}_T)/\sigma\right)$$

so Lemma 1 and 2 ensure that this term is asymptotically standard normal as long as $\sigma/\hat{\sigma} \to^p 1$. Since σ is almost surely positive, this convergence is equivalent to $\sigma^2 - \hat{\sigma}^2 \to^p 0$.

The proof that $\sigma^2 - \hat{\sigma}^2 \to^p 0$ follows de Jong and Davidson's (2000) Theorem 2.1 closely. We start by defining similar quantities to theirs, borrowing their notation when possible to make the similarities apparent. Let $b_T \equiv \lfloor \gamma/\delta \rfloor$, define $\eta_\delta(x)$ as in Equation

(32), and define the following terms as in de Jong and Davidson (2000):

$$\sigma^2_{0,\delta} \equiv P^{-1} \sum_{s,t=R+1}^{T-h} (D_t - \mathrm{E}_R D_t)(D_s - \mathrm{E}_R D_s) W((t-s)/\gamma),$$

$$\begin{aligned}\sigma^2_{1,\delta} \equiv & \sum_{t=-P+R+1}^{2P+R} (P\gamma)^{-1/2} \sum_{l=\max(1-t,-P)}^{\min(P-t,P)} (D_{t+l+R} - \mathrm{E}_R D_{t+l+R}) W(l/\gamma) \\ & \times (P\gamma)^{-1/2} \sum_{j=1-t}^{P-t} (D_{t+j+R} - \mathrm{E}_R D_{t+j+R}) \eta_\delta(j/\gamma),\end{aligned}$$

$$\begin{aligned}\sigma^2_{2,\delta} \equiv & \sum_{t=-P+R+1}^{2P+R} (P\gamma)^{-1/2} \sum_{l=\max(R+1-t,-b_T)}^{\min(R+P-t,b_T)} (D_{t+l} - \mathrm{E}_R D_{t+l}) W(l/\gamma) \\ & \times (P\gamma)^{-1/2} \sum_{j=1-t}^{P-t} (D_{t+j+R} - \mathrm{E}_R D_{t+j+R}) \eta_\delta(j/\gamma),\end{aligned}$$

$$\begin{aligned}\sigma^2_{3,\delta} \equiv & \sum_{t=-P+R+1}^{2P+R} (P\gamma)^{-1/2} \sum_{l=\max(R+1-t,-b_T)}^{\min(R+P-t,b_T)} (D_{t+l} - \mathrm{E}_R D_{t+l}) W(l/\gamma) \\ & \times (P\gamma)^{-1/2} \sum_{j=\max(R+1-t,-b_T)}^{\min(R+P-t,b_T)} (D_{t+j} - \mathrm{E}_R D_{t+j}) \eta_\delta(j/\gamma).\end{aligned}$$

These definitions give the inequalities

$$\begin{aligned}\|\hat{\sigma}^2 - \sigma^2\|_1 \leq & \|\hat{\sigma}^2 - \sigma^2_{0,\delta}\|_1 + \|\sigma^2_{0,\delta} - \sigma^2_{1,\delta}\|_1 + \|\sigma^2_{1,\delta} - \sigma^2_{2,\delta}\|_1 + \|\sigma^2_{2,\delta} - \sigma^2_{3,\delta}\|_1 \\ & + \|\sigma^2_{3,\delta} - \mathrm{E}_R \sigma^2_{3,\delta}\|_1 + \|\mathrm{E}_R \sigma^2_{2,\delta} - \mathrm{E}_R \sigma^2_{3,\delta}\|_1 + \|\mathrm{E}_R \sigma^2_{1,\delta} - \mathrm{E}_R \sigma^2_{2,\delta}\|_1 \\ & + \|\mathrm{E}_R \sigma^2_{0,\delta} - \mathrm{E}_R \sigma^2_{1,\delta}\|_1 + \|\mathrm{E}_R \sigma^2_{0,\delta} - \sigma^2\|_1 \\ \leq & \|\hat{\sigma}^2 - \sigma^2_{0,\delta}\|_1 + 2(\|\sigma^2_{0,\delta} - \sigma^2_{1,\delta}\|_1 + \|\sigma^2_{1,\delta} - \sigma^2_{2,\delta}\|_1 + \|\sigma^2_{2,\delta} - \sigma^2_{3,\delta}\|_1) \\ & + \|\sigma^2_{3,\delta} - \mathrm{E}_R \sigma^2_{3,\delta}\|_1 + \|\mathrm{E}_R \sigma^2_{0,\delta} - \sigma^2\|_1.\end{aligned}$$

De Jong and Davidson (2000) prove that

$$\lim_{\delta \to 0} \limsup_{T \to \infty} \|\sigma^2_{0,\delta} - \sigma^2_{1,\delta}\|_1 = 0,$$

$$\lim_{\delta \to 0} \limsup_{T \to \infty} \|\sigma^2_{1,\delta} - \sigma^2_{2,\delta}\|_1 = 0,$$

$$\lim_{\delta \to 0} \limsup_{T \to \infty} \|\sigma^2_{2,\delta} - \sigma^2_{3,\delta}\|_1 = 0,$$

and

$$\lim \|\mathrm{E}_R \sigma^2_{0,\delta} - \sigma^2\|_1 = 0.$$

Their proofs of these four convergence results use the fact that NED functions of mixing processes are also mixingale processes and do not use any other properties specific to NED processes, so their results hold here as well. We do need to modify their proofs that

$$\|\sigma^2_{3,\delta} - \mathrm{E}_R\, \sigma^2_{3,\delta}\|_1 \to 0$$

and

$$\|\hat{\sigma}^2 - \sigma^2_{0,\delta}\|_1 \to 0$$

for all positive δ, though, since those proofs exploit NED properties. These results are presented as Lemmas A4 and A5 respectively. □

Proof of Theorem 2

Coverage of the confidence intervals is immediate from Theorem 1, so we will present a proof of (17). By Lemma 2, we know that $\mathrm{E}_R\, \bar{D}_R - \mathrm{E}_T\, \bar{D}_T \to^p 0$, so

$$\Pr[\mathrm{E}_R\, \bar{D}_R \leq 0] - \Pr[\mathrm{E}_T\, \bar{D}_T \leq 0] \to 0.$$

Now we can use Lemma A3 to define D^*_t as in the proofs of Lemma 1 and 2 so that $\mathrm{E}_R\, D^*_t = \mathrm{E}_{t-l}\, D^*_t$ a.s. and

$$\|D_t - D^*_t\|_2 \leq 2^{(1+\rho)/\rho} B_L\, \beta_l^{(\rho-2)/2\rho}$$

and define $\bar{D}^*_R = P^{-1} \sum_{t=R+1}^{T-h} D^*_t$. Then

$$\|\sqrt{P}\bar{D}_R - \sqrt{P}\bar{D}^*_R\|_2 \leq P^{-1/2} \sum_{t=R+1}^{T-h} \|\bar{D}_R - \bar{D}^*_R\|_2 = o(1),$$

and, consequently, we have convergence in probability of the following vectors:

$$(\sqrt{P}\bar{D}^*_R/\sigma,\ \mathrm{E}_R\, \bar{D}_R) - (\sqrt{P}\bar{D}_R/\hat{\sigma},\ \mathrm{E}_T\, \bar{D}_T) \to^p 0,$$

where we are implicitly using consistency of $\hat{\sigma}^2$ for σ^2. More importantly, this implies convergence in distribution of these vectors, so

$$\Pr[\sqrt{P}\bar{D}^*_R/\sigma > z_\alpha \text{ and } \mathrm{E}_R\, \bar{D}_R \leq 0] - \Pr[\sqrt{P}\bar{D}_R/\hat{\sigma} > z_\alpha \text{ and } \mathrm{E}_T\, \bar{D}_T \leq 0] \to 0.$$

For large enough T, both $\Pr[\mathrm{E}_R\, \bar{D}_R \leq 0]$ and $\Pr[\mathrm{E}_T\, \bar{D}_T \leq 0]$ are positive (the second by assumption, the first by convergence to the second) so (for these T)

$$\Pr[\sqrt{P}\bar{D}^*_R/\sigma > z_\alpha \mid \mathrm{E}_R\, \bar{D}_R \leq 0] = \frac{\Pr[\sqrt{P}\bar{D}^*_R/\sigma > z_\alpha \text{ and } \mathrm{E}_R\, \bar{D}_R \leq 0]}{\Pr[\mathrm{E}_R\, \bar{D}_R \leq 0]} \tag{41}$$

and

$$\Pr[\sqrt{P}\bar{D}_R/\hat{\sigma} > z_\alpha \mid \mathrm{E}_T\,\bar{D}_T \leq 0] = \frac{\Pr[\sqrt{P}\bar{D}_R/\hat{\sigma} > z_\alpha \text{ and } \mathrm{E}_T\,\bar{D}_T \leq 0]}{\Pr[\mathrm{E}_T\,\bar{D}_T \leq 0]} \tag{42}$$

almost surely. Since the terms on the RHS of Equations (41) and (42) converge to the same limit in probability, we have

$$\Pr[\sqrt{P}\bar{D}_R/\hat{\sigma} > z_\alpha \mid \mathrm{E}_T\,\bar{D}_T \leq 0] - \Pr[\sqrt{P}\bar{D}_R^*/\sigma > z_\alpha \mid \mathrm{E}_R\,\bar{D}_R \leq 0] \to^p 0$$

and it suffices to show that

$$\operatorname{plim} \Pr[\sqrt{P}\bar{D}_R^*/\sigma > z_\alpha \mid \mathrm{E}_R\,\bar{D}_R \leq 0] \leq \alpha.$$

To establish this inequality, we have

$$\begin{aligned}\Pr[\sqrt{P}\bar{D}_R^*/\sigma > z_\alpha \mid \mathrm{E}_R\,\bar{D}_R \leq 0] &\leq \Pr[\sqrt{P}(\bar{D}_R^* - \mathrm{E}_R\,\bar{D}_R)/\sigma > z_\alpha \mid \mathrm{E}_R\,\bar{D}_R \leq 0] \\ &\leq \mathrm{E}\left(\Pr[\sqrt{P}(\bar{D}_R^* - \mathrm{E}_R\,\bar{D}_R)/\sigma > z_\alpha \mid \mathscr{F}_R] \mid \mathrm{E}_R\,\bar{D}_R \leq 0\right).\end{aligned}$$

By construction,

$$\Pr[\sqrt{P}(\bar{D}_R^* - \mathrm{E}_R\,\bar{D}_R)/\sigma > z_\alpha \mid \mathscr{F}_R] \to \alpha$$

since generated pseudo OOS observations in each D_t^* are independent of $\mathscr{F}_R$, completing the proof. □

Proof of Theorem 3

Let $(\varepsilon_{t+h}, x_t) \sim i.i.d.\ N(0, I)$ and let $\theta_1 = 0$. For any $d \geq 0$, the event $\mathrm{E}_T\,\bar{D}_T = -d$ implies that

$$\mathrm{E}_T(y_{T+h+1} - x_{1T+1}'\hat{\theta}_{1T})^2 = \mathrm{E}_T(y_{T+h+1} - x_{2T+1}'\hat{\theta}_{2T})^2 - d \quad a.s.$$

which can be expressed as

$$d + \theta_2' M_2' M_2 \theta_2 + \hat{\theta}_{1T}'\hat{\theta}_{1T} - \hat{\theta}_{2T}' M_1' M_1 \hat{\theta}_{2T} = (\hat{\theta}_{2T} - \theta_2)' M_2' M_2 (\hat{\theta}_{2T} - \theta_2) \quad a.s. \tag{43}$$

$M_2\hat{\theta}_{2T}$ is normally distributed conditional on $\hat{\theta}_{1T}$ and $M_1\hat{\theta}_{2T}$, and is distributed on the surface of the sphere defined by (43) conditional on $\hat{\theta}_{1T}$, $M_1\hat{\theta}_{2T}$, and the event $\mathrm{E}_T\,\bar{D}_T = -d$.

Since $M_2\hat{\theta}_{2T}$ is normal, this conditional distribution is invariant to reflection across the axes defined by the eigenvectors of its covariance matrix. So when θ_2 lies outside the cylinder that contains the acceptance region of Λ, i.e. when $\theta_2' M_2' V_T M_2 \theta_2 > c$, we have[30]

$$\begin{aligned}\Pr[\hat{\theta}_{2T}' M_2' V_T M_2 \hat{\theta}_{2T} \leq c \mid \mathscr{G}] &= \mathrm{E}\left(\Pr[\hat{\theta}_{2T}' M_2' V_T M_2 \hat{\theta}_{2T} \leq c \mid \mathscr{G},\ \hat{\theta}_{1T},\ M_1\hat{\theta}_{2T}] \mid \mathscr{G}\right) \\ &< 1/2\end{aligned}$$

[30] To keep the notation in these equations manageable, define the information set $\mathscr{G} = \sigma(\mathrm{E}_T\,\bar{D}_T \leq 0,\ \theta_2' M_2' V_T M_2 \theta_2 > c)$.

since the inner conditional probability is less than $1/2$.

Now let $\delta > 0$ be an arbitrary but small constant, and choose θ_2 far enough from the origin to ensure that

$$\Pr[\,\theta_2' M_2' V_T M_2 \theta_2 \le c\,] \le \delta$$

for large enough T. Then

$$\begin{aligned}
\mathrm{E}(\Lambda \mid \mathrm{E}_T \bar{D}_T \le 0) &\ge \Pr[\hat{\theta}_{2T}' M_2' V_T M_2 \hat{\theta}_{2T} > c \mid \mathrm{E}_T \bar{D}_T \le 0] \\
&= \Pr[\hat{\theta}_{2T}' M_2' V_T M_2 \hat{\theta}_{2T} > c \ \text{ and } \ \theta_2' M_2' V_T M_2 \theta_2 > c \mid \mathrm{E}_T \bar{D}_T \le 0] \\
&\quad + \Pr[\hat{\theta}_{2T}' V_T M_2 \hat{\theta}_{2T} > c \ \text{ and } \ \theta_2' M_2' V_T M_2 \theta_2 \le c \mid \mathrm{E}_T \bar{D}_T \le 0].
\end{aligned}$$

The second term is nonnegative by design. By conditioning on $\theta_2' M_2' V_T M_2 \theta_2 > c$ our earlier argument shows that the first term is greater than or equal to $(1-\delta)/2$. Since δ is arbitrarily small, this completes the proof. □

Proof of Theorem 4

As in the proof of Theorem 3, let $(\varepsilon_{t+h}, x_t) \sim i.i.d.\ N(0, I)$ and let $\theta_1 = 0$. The event $\mathrm{E}_T \bar{D}_T \ge 0$ implies that

$$\mathrm{E}_T(y_{T+h+1} - x_{1T+1}' \hat{\theta}_{1T})^2 \ge \mathrm{E}_T(y_{T+h+1} - x_{2T+1}' \hat{\theta}_{2T})^2 \quad a.s.$$

which can be expressed as

$$\theta_2' M_2 M_2' \theta_2 + \hat{\theta}_{1T}' \hat{\theta}_{1T} - \hat{\theta}_{2T}' M_1' M_1 \hat{\theta}_{2T} \ge (\hat{\theta}_{2T} - \theta_2)' M_2' M_2 (\hat{\theta}_{2T} - \theta_2) \quad a.s. \tag{44}$$

$M_2 \hat{\theta}_{2T}$ is normally distributed conditional on $\hat{\theta}_{1T}$ and $M_1 \hat{\theta}_{2T}$, and is a.s. contained in the sphere defined by (44) conditional on $\hat{\theta}_{1T}$, $M_1 \hat{\theta}_{2T}$, and the event $\mathrm{E}_T \bar{D}_T \ge 0$.

Let δ be a small arbitrary constant, define θ^* as

$$\begin{aligned}
&\theta^* = \arg\max_{\theta} \theta' M_2' M_2 \theta \\
s.t. \quad &\theta_2' M_2' M_2 \theta_2 + \hat{\theta}_{1T}' \hat{\theta}_{1T} - \hat{\theta}_{2T}' M_1' M_1 \hat{\theta}_{2T} = (\theta - \theta_2)' M_2' M_2 (\theta - \theta_2)
\end{aligned}$$

(θ^* is stochastic and depends on $\hat{\theta}_{1T}$ and $M_1 \hat{\theta}_{2T}$) and choose c so that

$$\Pr[\theta^{*\prime} M_2 M_2 \theta^* > c] < \delta.$$

Then, we have for large enough T

$$\begin{aligned}
\mathrm{E}(\Lambda \mid \mathrm{E}_T \bar{D}_T \ge 0) &\le \Pr[\hat{\theta}_{2T}' M_2' M_2 \hat{\theta}_{2T} > c \mid \mathrm{E}_T \bar{D}_T \ge 0) \\
&\le \delta
\end{aligned}$$

Since δ is arbitrarily close to zero, this completes the proof.

□

References

H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247, 1969.

M. Akritas and S. Arnold. Asymptotics for analysis of variance when the number of levels is large. *Journal of the American Statistical Association*, 95:212–226, Mar. 2000.

M. Akritas and N. Papadatos. Heteroscedastic one-way ANOVA and lack-of-fit tests. *Journal of the American Statistical Association*, 99(466):368–382, June 2004.

S. Anatolyev. Inference about predictive ability when there are many predictors. Working Paper, 2007.

S. Anatolyev. Inference in regression models with many predictors. *Journal of Econometrics*, 2012. forthcoming.

P. Bacchetta, E. van Wincoop, and T. Beutler. Can parameter instability explain the Meese-Rogoff puzzle? In L. Reichlin and K. D. West, editors, *NBER International Seminar on Macroeconomics 2009*, pages 125–173. University of Chicago Press, 2010.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

T. Baker, P. Forrester, and P. Pearce. Random matrix ensembles with an effective extensive external charge. *Journal of Physics A: Mathematical and General*, 31(29):6087–6101, 1998.

D. Bates and M. Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2013. R package version 1.0-14, available at http://cran.r-project.org/package=Matrix.

H. C. P. Berbee. *Random walks with stationary increments and renewal theory*. Mathematical Center Tracts. Mathematisch Centrum, Amsterdam, 1979.

D. Boos and C. Brownie. ANOVA and rank tests when the number of treatments is large. *Statistics & Probability Letters*, 23:183–191, May 1995.

P. Bossaerts and P. Hillion. Implementing statistical criteria to select return forecasting models: what do we learn? *Review of Financial Studies*, 12(2):405–428, 1999.

J. Boudoukh, M. Richardson, and R. F. Whitelaw. The myth of long-horizon predictability. *Review of Financial Studies*, 21(4):1577–1605, 2008.

G. Calhoun. Hypothesis testing for linear regression when k/n is large. *Journal of Econometrics*, 165(2):163–174, 2011.

G. Calhoun. An asymptotically normal out-of-sample test of equal predictive accuracy for nested models. Unpublished mauscript, 2013.

J. Y. Campbell and S. B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531, 2008.

J. C. Chao, V. Corradi, and N. R. Swanson. An out of sample test for Granger causality. *Macroeconomic Dynamics*, 5(4):598–620, 2001.

Y. Cheung, M. D. Chinn, and A. G. Pascual. Empirical exchange rate models of the nineties: Are any fit to survive? *Journal of International Money and Finance*, 24(7): 1150–1175, Nov. 2005.

M. D. Chinn. Comment on "Can parameter instability explain the Meese-Rogoff puzzle?". In L. Reichlin and K. D. West, editors, *NBER International Seminar on Macroeconomics 2009*, pages 174–179. University of Chicago Press, 2010.

T. E. Clark. Can out-of-sample forecast comparisons help prevent overfitting? *Journal of Forecasting*, 23(2):115–139, 2004.

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 M. W. McCracken. In-sample tests of predictive ability: a new approach. *Journal of Econometrics*, 170(1):1–14, 2012a.

T. E. Clark and M. W. McCracken. Reality checks and nested forecast model comparisons. *Journal of Business and Economic Statistics*, 30(1):53–66, 2012b.

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.

J. H. Cochrane. The Dog That Did Not Bark: A Defense of Return Predictability. *Review of Financial Studies*, 21(4):1533–1575, 2008.

V. Corradi and N. R. Swanson. A consistent test for nonlinear out of sample predictive accuracy. *Journal of Econometrics*, 110(2):353–381, Oct. 2002.

V. Corradi and N. R. Swanson. Some recent developments in predictive accuracy testing with nested models and (generic) nonlinear alternatives. *International Journal of Forecasting*, 20(2):185–199, 2004.

J. Davidson. An L_1-Convergence theorem for heterogeneous mixingale arrays with trending moments. *Statistics & Probability Letters*, 16(4):301–304, Mar. 1993.

J. Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Advanced Texts in Econometrics. Oxford University Press, 1994.

R. M. de Jong. Central limit theorems for dependent heterogeneous random variables. *Econometric Theory*, 13(3):353–367, 1997.

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.

P. Doukhan. *Mixing: properties and examples*. Springer New York, 1994.

B. Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81:461–470, 1986.

B. Efron. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, Sept. 2004.

C. Engel and K. D. West. Exchange rates and fundamentals. *Journal of Political Economy*, 113(3):485–517, June 2005.

S. Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8(2):252–261, 1980.

R. Giacomini and B. Rossi. Detecting and predicting forecast breakdowns. *Review of Economic Studies*, 76(2):669–705, 2009.

R. Giacomini and B. Rossi. Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620, 2010.

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

D. Giannone. Comment on "Can parameter instability explain the Meese-Rogoff puzzle?". In L. Reichlin and K. D. West, editors, *NBER International Seminar on Macroeconomics 2009*, pages 180–190. University of Chicago Press, 2010.

A. Goyal and I. Welch. Predicting the equity premium with dividend ratios. *Management Science*, 49(5):639–654, 2003.

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. Hall and C. Heyde. *Martingale limit theory and its application*. Academic Press, 1980.

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

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2nd edition, 2008.

P.-H. Hsu, Y.-C. Hsu, and C.-M. Kuan. Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance*, 17(3):471–484, 2010.

P. Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

A. Inoue and L. Kilian. In-sample or out-of-sample tests of predictability: which one should we use? *Econometric Reviews*, 23(4):371–402, 2004.

A. Inoue and L. Kilian. On the selection of forecasting models. *Journal of Econometrics*, 130(2):273–306, Feb. 2006.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

L. Kilian and M. P. Taylor. Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics*, 60(1):85–107, May 2003.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer texts in statistics. Springer Verlag, 3rd edition, 2005.

M. Lettau and S. Van Nieuwerburgh. Reconciling the return predictability evidence. *Review of Financial Studies*, 21(4):1607–1652, 2008.

N. C. Mark. Exchange rates and fundamentals: Evidence on long-horizon predictability. *American Economic Review*, 85(1):201–218, 1995.

M. W. McCracken. Data mining and out-of-sample inference. Manuscript, Louisiana State University, 1998.

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. Dependent central limit theorems and invariance principles. *The Annals of Probability*, 2(4):620–628, Aug. 1974.

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

D. McLeish. Invariance principles for dependent variables. *Probability Theory and Related Fields*, 32:165–178, 1975b.

D. McLeish. On the invariance principle for nonstationary mixingales. *The Annals of Probability*, 5(4):616–621, Aug. 1977.

R. A. Meese and K. Rogoff. Empirical exchange rate models of the seventies: do they fit out of sample? *Journal of International Economics*, 14(1-2):3–24, Feb. 1983.

F. Merlevède and M. Peligrad. On the coupling of dependent random variables and applications. In H. Dehling, T. Mikosch, and M. Sørensen, editors, *Empirical Process Techniques for Dependent Data*, pages 171–193. Birkhäuser, 2002.

W. K. Newey and K. D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, May 1987.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, http://www.r-project.org, Vienna, Austria, 2010.

D. Rapach and G. Zhou. Forecasting stock returns. *Handbook of Economic Forecasting*, 2: 327–384, 2012.

D. E. Rapach and M. E. Wohar. In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. *Journal of Empirical Finance*, 13(2):231–247, 2006.

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

B. Rossi. Testing long-horizon predictive ability with high persistence, and the Meese-Rogoff puzzle. *International Economic Review*, 46(1):61–92, 2005.

D. Sarkar. *lattice: Lattice Graphics*, 2010. R package version 0.18-5.

H. Sevcikova and T. Rossini. *rlecuyer: R interface to RNG with multiple streams*, 2012. R package version 0.3-3, available at http://CRAN.R-project.org/package=rlecuyer.

J. W. Silverstein et al. The smallest eigenvalue of a large dimensional wishart matrix. *The Annals of Probability*, 13(4):1364–1368, 1985.

J. H. Stock and M. W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, 20(2):147–162, 2002.

J. H. Stock and M. W. Watson. Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature*, 41(3):788–829, 2003.

R. Sullivan, A. Timmermann, and H. White. Data-snooping, technical trading rule performance, and the bootstrap. *The journal of Finance*, 54(5):1647–1691, 1999.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

A. Timmermann. Elusive return predictability. *International Journal of Forecasting*, 24 (1):1–18, 2008.

C. A. Tracy and H. Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727–754, 1996.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 4th edition, 2002.

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

K. D. West. Forecast evaluation. In G. Elliott, C. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, pages 99–134. Elsevier, 2006.

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

A. Zeileis. Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004.

A. Zeileis and T. Hothorn. Diagnostic checking in regression relationships. *R News*, 2(3): 7–10, 2002.

			Conditional rejection probability (size)	
K_1	K_2	T	$\|\theta_2\|_2 = 0$	$\|\theta_2\|_2 = 1$
2	3	100	0.104	0.485
		250	0.102	0.528
		500	0.096	0.503
$T/50$	$T/10$	100	0.106	0.701
		250	0.108	0.904
		500	0.104	0.989

Table 1: Simulated rejection probabilities for the F-test under the null hypothesis that the benchmark model will forecast better, $\mathrm{E}_T \bar{D}_T \leq 0$. Nominal size is 10% and values greater than 10% indicate that the test rejects the benchmark model too often. See Section 4.1 for a discussion of the simulation design.

Category	Variable
Stock market variables	Dividend to price ratio (log)
	Earnings to price ratio (log)
	Stock market variance
	Book to market ratio
	Net equity expansion
	Percent equity issuing
Interest rate variables	Treasury Bill rate (3 month)
	Long term yield
	Long term rate
	Default return spread
	Default yield spread
	Inflation rate

Table 2: Variables used to predict the equity premium (Section 5). Please see Goyal and Welch's original paper (Goyal and Welch, 2008) for a detailed description of each variable.

	Coefficient estimates		p-values	
	Full	Univariate	Full	Univariate
Stock market variance	4.61	0.36	0.081	0.870
Book to market ratio	4.46	4.81	0.251	0.027
Dividend to price ratio (log)	4.40	3.91	0.400	0.073
Long term rate	3.78	2.91	0.035	0.184
Long term yield	3.23	−1.25	0.438	0.569
Earnings to price ratio (log)	0.79	3.50	0.849	0.109
Inflation rate	0.63	0.69	0.696	0.754
Default return spread	0.53	−1.74	0.795	0.429
Net equity expansion	−1.88	−4.10	0.626	0.060
Treasury Bill rate	−4.38	−2.47	0.316	0.260
Percent equity issuing	−5.56	−5.10	0.040	0.019
Default yield spread	−7.79	0.79	0.042	0.720
Robust F-statistic (12/69 df)	3.37		0.001	
R^2	0.23			
Adjusted R^2	0.10			

Table 3: Coefficient estimates and model fit for the univariate and full models described in Section 5. The "Full" column lists the coefficient estimates from regressing the equity premium on all of the variables listed; the "Univariate" column lists the coefficient estimates from a univariate regression of the equity premium on the variable alone. All of the regressions include a constant, but its estimate is not listed. The equity premium is expressed in basis points and all of the regressors are studentized. These models are estimated using U.S. annual data from 1928–2009 (81 observations). The standard errors and robust F-statistic are calculated using a Newey-West HAC estimator with two lags, implemented in the *sandwich* and *lmtest* R packages (Newey and West, 1987; Zeileis and Hothorn, 2002; Zeileis, 2004).

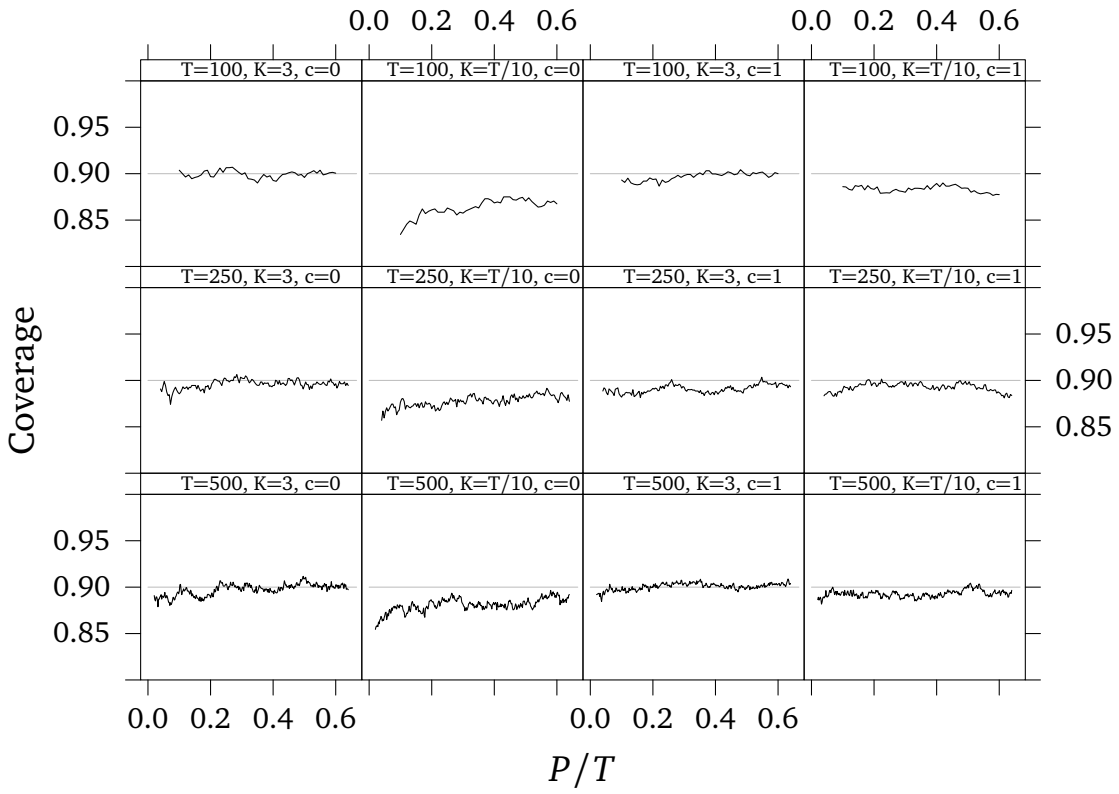


Figure 2: Simulated coverage of $E_R \bar{D}_R$ at 90% confidence using a one-sided interval based on the DMW OOS-t test, plotted as a function of the fraction of observations used in the test sample, P/T. The solid horizontal line denotes the intervals' nominal coverage.

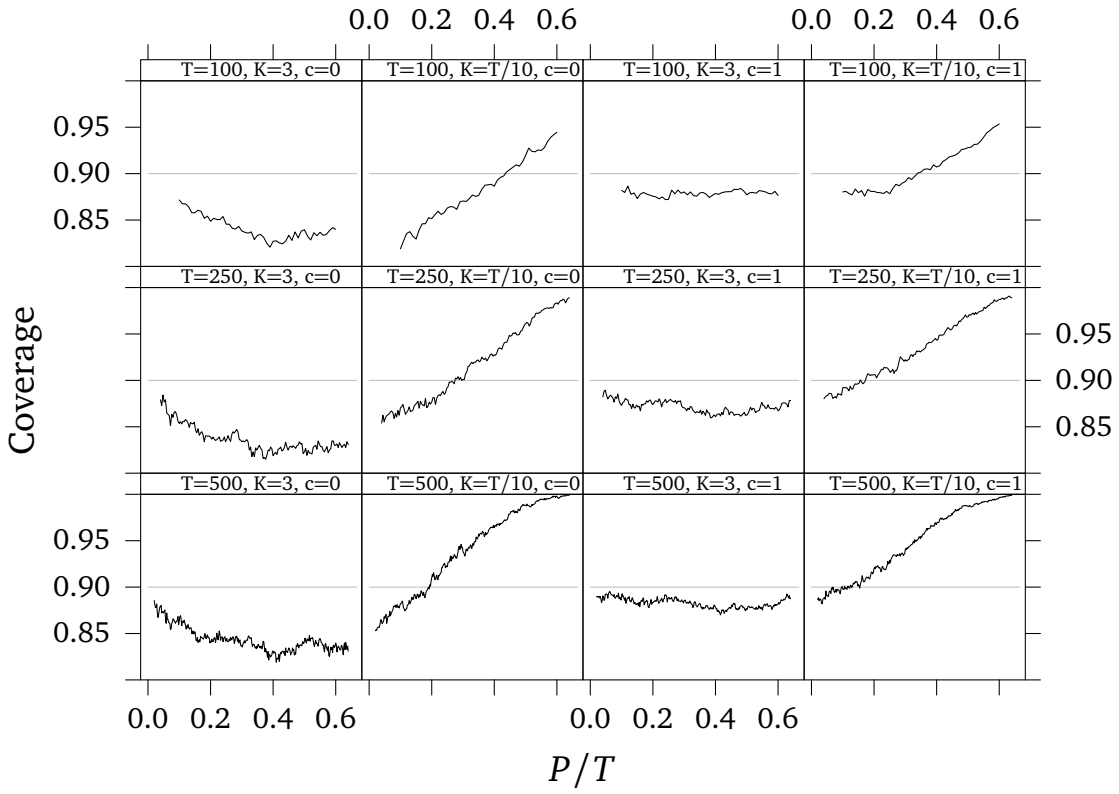


Figure 3: Simulated coverage of $\mathrm{E}_T\,\bar{D}_T$ at 90% confidence using a one-sided interval based on the DMW OOS-t test, plotted as a function of the fraction of observations used in the test sample, P/T. The solid horizontal line denotes the intervals' nominal coverage.

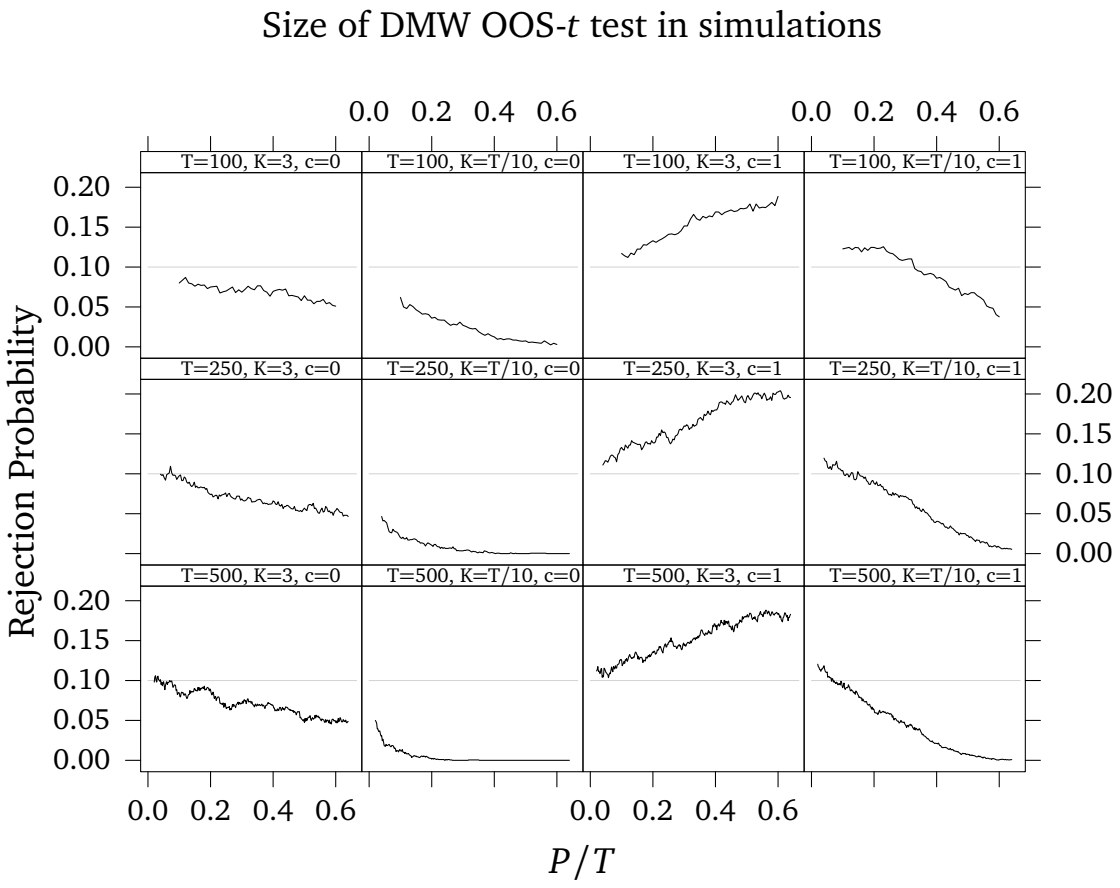


Figure 4: Simulated rejection probabilities for the DMW OOS-t test under the null hypothesis that the benchmark model will forecast better, $\mathrm{E}_T \bar{D}_T \leq 0$. The nominal size is 10% and is marked with a solid horizontal line. Values greater than 10% indicate that the test rejects the benchmark model too often. See Section 4.1 for a discussion of the simulation design.

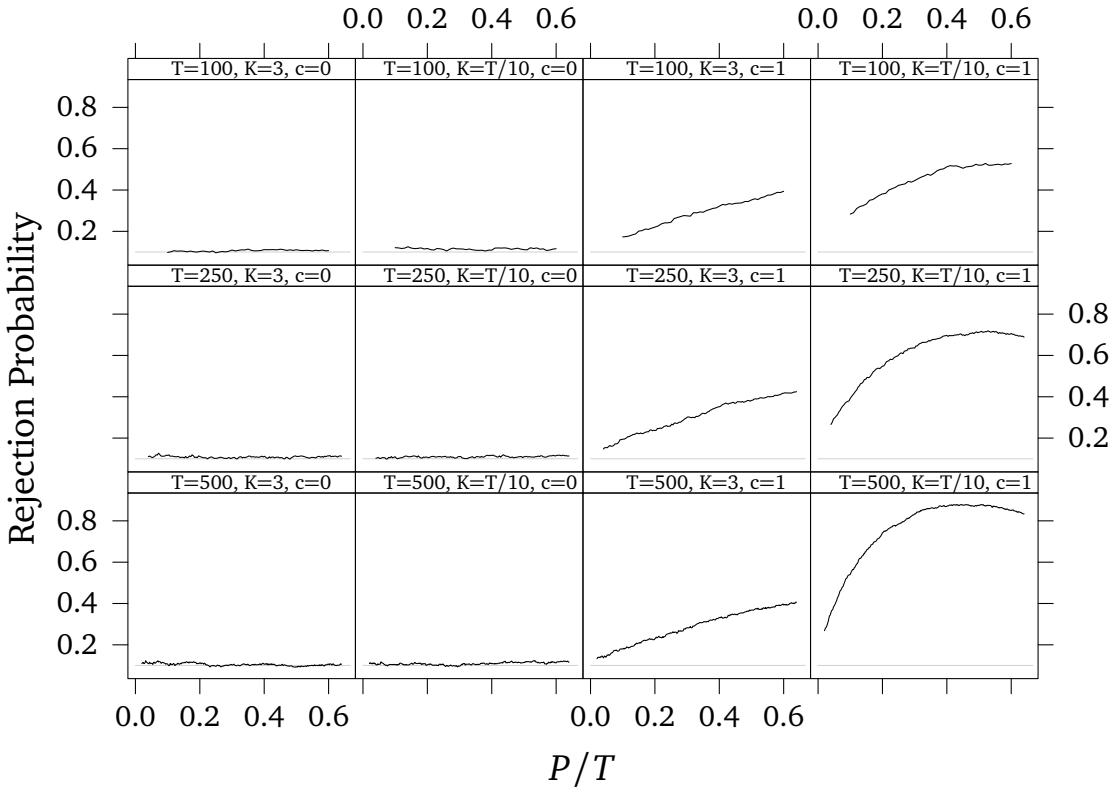


Figure 5: Simulated rejection probabilities for Clark and West's (2006, 2007) OOS test statistic under the null hypothesis that the benchmark will forecast better, $\mathrm{E}_T \bar{D}_T \leq 0$. The nominal size is 10% and is marked with a solid horizontal line. Values greater than 10% indicate that the test rejects the benchmark model too often. See Section 4.1 for a discussion of the simulation design.

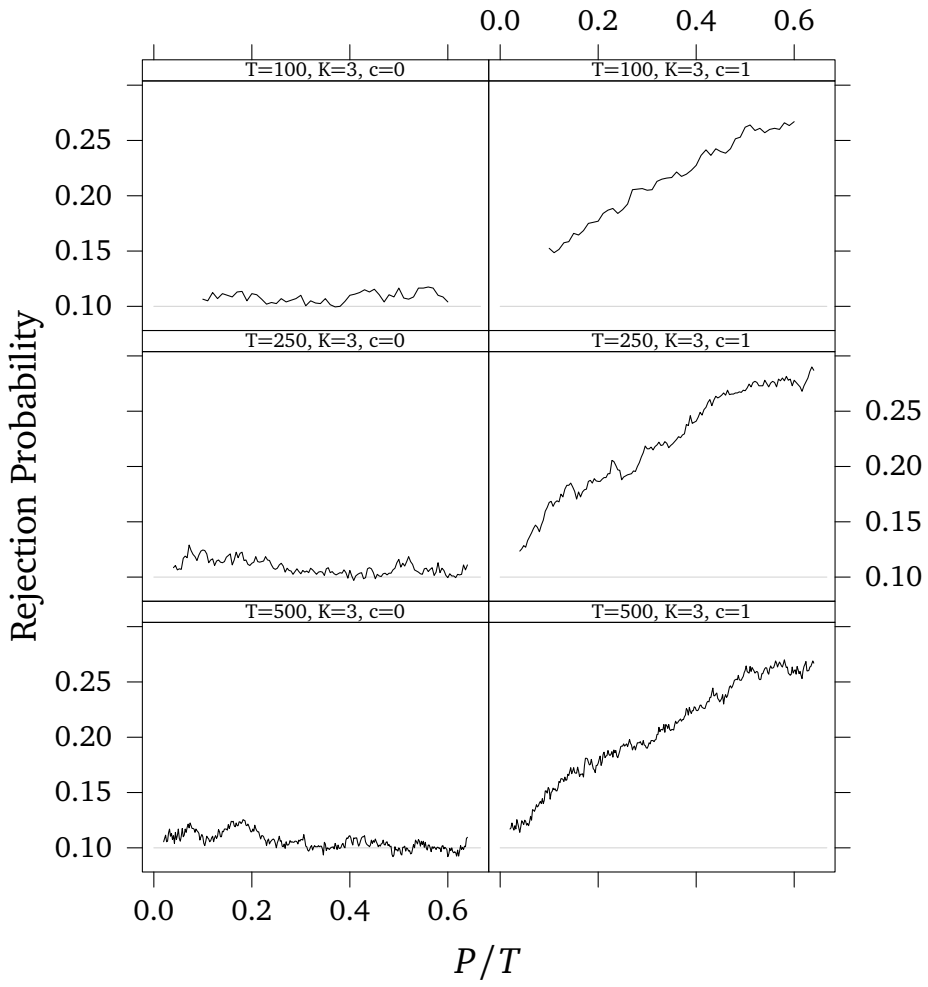


Figure 6: Simulated rejection probabilities for McCracken's (2007) OOS-t test under the null hypothesis that the benchmark model is more accurate, $\mathrm{E}_T \bar{D}_T \leq 0$. Nominal size is 10% and is marked with a solid horizontal line. Values greater than 10% indicate that the test rejects the benchmark model too often. See Section 4.1 for a discussion of the simulation design.

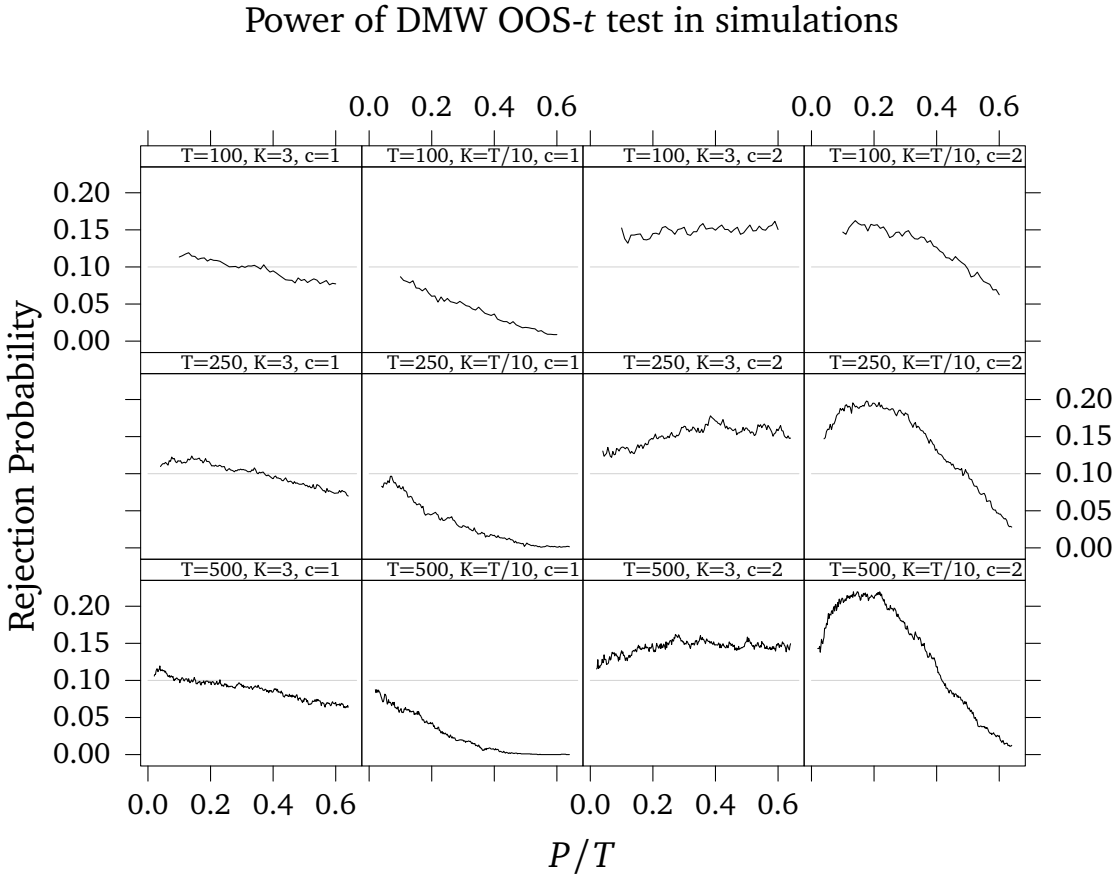


Figure 7: Simulated rejection probabilities for the DMW OOS-t test under the alternative that the benchmark model is less accurate, $\mathrm{E}_T \bar{D}_T > 0$. Nominal size is 10% and is marked with a solid horizontal line. Values greater than 10% indicate that the test rejects the benchmark model too often. See Section 4.1 for a discussion of the simulation design.

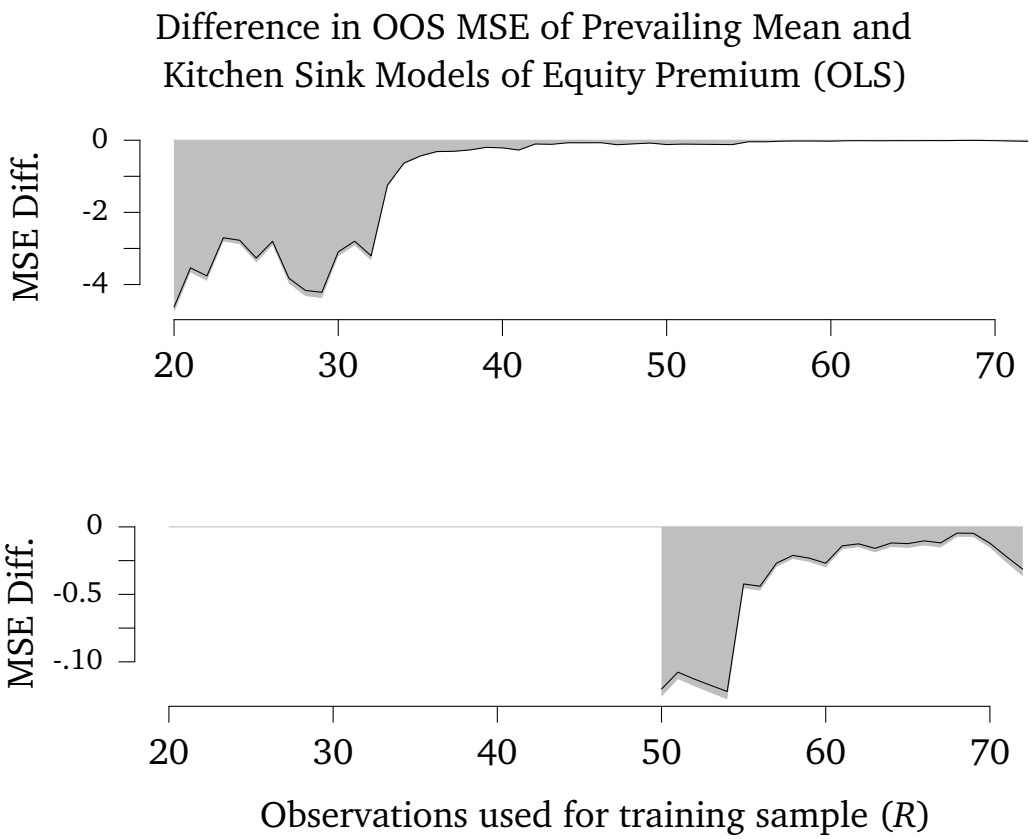


Figure 8: OOS difference in the MSE of the prevailing mean benchmark and the kitchen sink model as a function of the test sample size, R. Both models forecast the equity premium using annual data from 1928–2008. The solid line gives the OOS average, and the shaded region indicates the one-sided 95% confidence interval implied by the DMW test. The bottom panel is a detailed view of the top panel for $R \geq 50$.

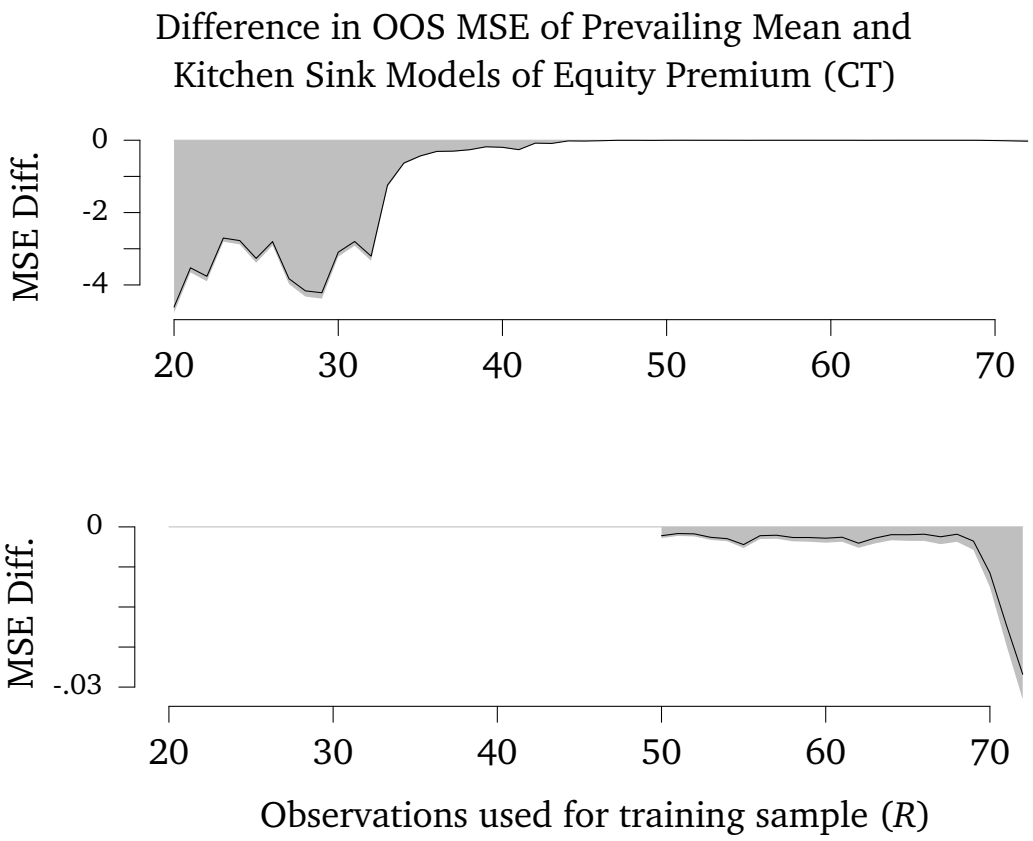


Figure 9: OOS difference in the MSE of the prevailing mean benchmark and the kitchen sink model as a function of the test sample size, R. Both models forecast the equity premium using annual data from 1928–2008. The solid line gives the OOS average, and the shaded region indicates the one-sided 95% confidence interval implied by the DMW test. The bottom panel is a detailed view of the top panel for $R \geq 50$.

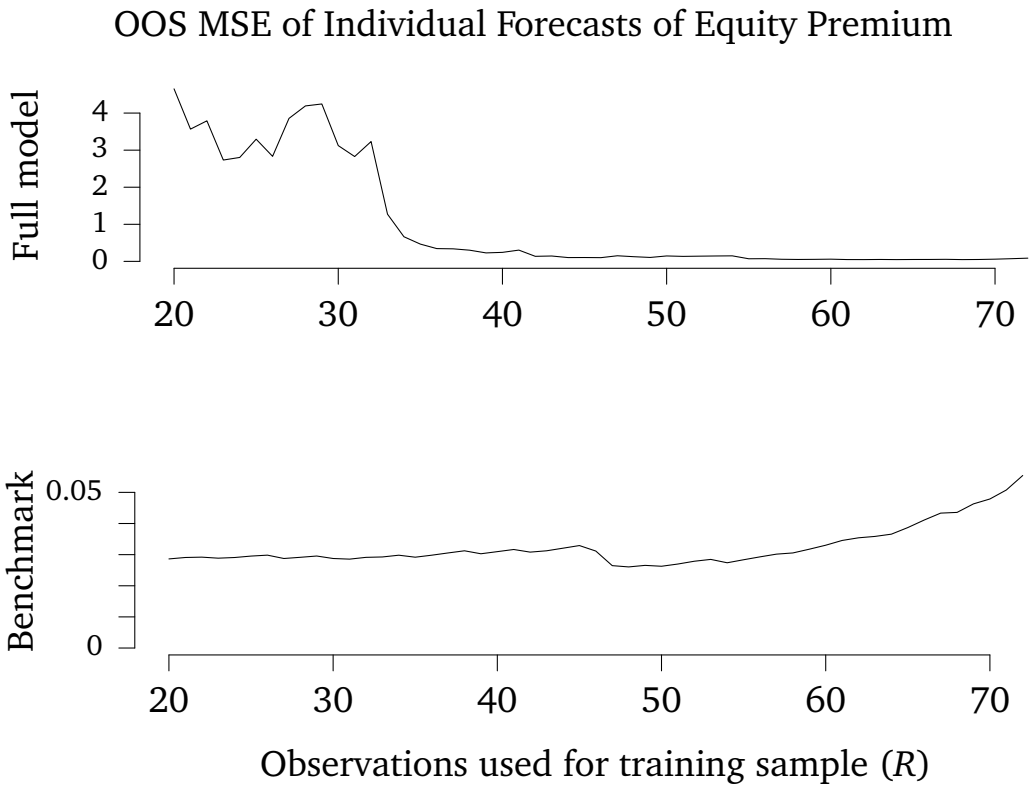


Figure 10: OOS MSE of the Prevailing Mean (PM) and Kitchen Sink (KS) models for equity premium prediction as a function of the size of the training sample, R. Please note that the vertical scales are different in the two plots.