Choice of Sample Split in Out-of-Sample Forecast Evaluation*

Peter Reinhard Hansen
European University Institute, Stanford University and CREATES

Allan Timmermann
UCSD and CREATES

February 7, 2012

Abstract

Out-of-sample tests of forecast performance depend on how a given data set is split into estimation and evaluation periods, yet no guidance exists on how to choose the split point. Empirical forecast evaluation results can therefore be difficult to interpret, particularly when several values of the split point might have been considered. When the sample split is viewed as a choice variable, rather than being fixed *ex ante*, we show that very large size distortions can occur for conventional tests of predictive accuracy. Spurious rejections are most likely to occur with a short evaluation sample, while conversely the power of forecast evaluation tests is strongest with long out-of-sample periods. To deal with size distortions, we propose a test statistic that is robust to the effect of considering multiple sample split points. Empirical applications to predictability of stock returns and inflation demonstrate that out-of-sample forecast evaluation results can critically depend on how the sample split is determined.

Keywords: Out-of-sample forecast evaluation; data mining; recursive estimation; predictability of stock returns; inflation forecasting.

JEL Classification: C12, C53, G17.

*Valuable comments were received from Frank Diebold, Jim Stock, and seminar participants at University of Pennsylvania, the Triangle Econometrics Seminar, UC Riverside, the UCSD conference in Honor of Halbert White, and the NBER/NSF Summer Institute 2011.

1 Introduction

Statistical tests of a model's forecast performance are commonly conducted by splitting a given data set into an in-sample period, used for initial parameter estimation and model selection, and an out-of-sample period, used to evaluate forecast performance. Empirical evidence based on out-of-sample forecast performance is generally considered more trustworthy than evidence based on in-sample performance which can be more sensitive to outliers and data mining (White (2000*b*)). Out-of-sample forecasts also better reflect the information available to the forecaster in "real time" (Diebold & Rudebusch (1991)). This has led many researchers to regard out-of-sample performance as the "ultimate test of a forecasting model" (Stock & Watson (2007, p. 571)).[1]

This paper focuses on a dimension of the forecast evaluation problem that has so far received little attention. When presenting out-of-sample evidence, the sample split defining the beginning of the evaluation period is a choice variable, yet there are no broadly accepted guidelines for how to select the sample split.[2] Instead, researchers have adopted a variety of practical approaches. One approach is to choose the initial estimation sample to have a minimum length and use the remaining sample for forecast evaluation. For example, Marcellino, Stock & Watson (2006) and Pesaran, Pick & Timmermann (2011) use the first 20 years of data, when available, to estimate forecasting models for a variety of macroeconomic variables. Another common approach is to do the reverse and reserve a certain sample length, e.g., 10 or 20 years of observations, for the out-of-sample period (Inoue & Kilian (2008)). Alternatively, researchers such as Welch & Goyal (2008) and Rapach, Strauss & Zhou (2010) consider multiple out-of-sample periods and report the significance of forecasting performance for each. Ultimately, however, these approaches all depend on ad-hoc choices of the individual split points.

The absence of guidance on how to select the split point that separates the in-sample and out-of-sample periods, raises several questions. First, a 'data-mining' issue arises because multiple split points might have been considered and the reported values could be those

[1] For excellent reviews of the forecast evaluation problem, see West (2006) and Clark & McCracken (2012).

[2] See, e.g., Welch & Goyal (2008, p.1464), "It is not clear how to choose the periods over which a regression model is estimated and subsequently evaluated." Stock & Watson (2007, p. 571) recommend "Pick a date near the end of the sample, estimate your forecasting model using data up to that date, then use that estimated model to make a forecast."

that most favor a given model. Even if individual researchers consider only a single split point, the community of researchers could collectively have examined a range of split points, thereby influencing individual researchers' choice.[3] When compared to test statistics that assume a single (predetermined) split point, results that are optimized in this manner can lead to size distortions and may ameliorate the tendency of out-of-sample tests of predictive accuracy to underreject (Inoue & Kilian (2004) and Clark & West (2007)). It is therefore important to investigate how large such size distortions are, how they depend on the split point–whether they are largest if the split point is at the beginning, middle or end of the sample–and how they depend on the dimension of the prediction model under study.

A second question is related to how the choice of sample split trades off the effect of estimation error on forecast precision versus the power of the test as determined by the number of observations in the out-of-sample period. Given the generally weak power of out-of-sample forecast evaluation tests, it is important to choose the sample split to generate the highest achievable power. This will help direct the power in a way that maximizes the probability of correctly finding predictability. We find that power is maximized if the sample split falls relatively early in the sample so as to obtain the longest available out-of-sample evaluation period.

A third issue is how one can construct a test that is robust to sample split mining. To address this point, we propose a minimum p-value approach that accounts for search across different split points while allowing for heteroskedasticity across the distribution of critical values associated with different split points. The approach yields conservative inference in the sense that it is robust to search across all possible sample split points, which from an inferential perspective represents the 'worst case' scenario. Another possibility is to construct a joint test for out-of-sample predictability at multiple split points, but this leaves aside the issue of how best to determine these multiple split points.

The main contributions of our paper are the following. First, using a simple theoretical setup, we show how predictive accuracy tests such as those proposed by McCracken (2007) and Clark & McCracken (2001, 2005) are affected when researchers optimize or "mine" over the sample split point. The rejection rate tends to be highest if the split point is chosen at the beginning or end of the sample. We quantify the effect of such mining over the

[3] Rules of thumb such as using the first 10 or 20 years of data for estimation or forecast evaluation purposes are clearly designed to reduce the arbitrariness of how the split point is selected.

sample split on the probability of rejecting the null of no predictability. Rejection rates are found to be far higher than the nominal critical levels. For example, tests of predictive accuracy for a model with one additional parameter conducted at the nominal 5% level, but conducted at all split points between 10% and 90% of the sample, reject 15% of the time, i.e., three times as often as they should. Similar inflation in rejection rates are seen at other critical levels, although they grow even larger as the dimension of the prediction model increases (for a fixed benchmark). Second, we extend the results in McCracken (2007) and Clark & McCracken (2001, 2005) in many ways. We derive results under weaker assumptions and provide simpler expressions for the limit distributions. The latter mimic those found in asymptotic results for quasi maximum likelihood analysis. In particular, we show that expressions involving stochastic integrals can be reduced to simple convolutions of chi-squared random variables. This greatly simplifies calculation of critical values for the test statistics. Third, we propose a test statistic that is robust to mining over the sample split point. In situations where the "optimal" sample split is used, our test shows that in order to achieve, say, a five percent rejection rate, test statistics corresponding to a far smaller nominal critical level, such as one percent or less, should be used. Fourth, we derive analytical results for the asymptotic power of the tests which add insight on existing simulation-based results in the literature. We characterize power as a function of the split point and show how this gets maximized if the split point is chosen to fall at the beginning of the sample. Fourth and finally, we provide empirical illustrations for US stock returns and inflation that illustrate the importance of accounting for sample split mining when conducting inference about predictive performance.

Our analysis is related to a large literature on the effect of data mining arising from search over model specifications. When the best model is selected from a larger universe of competing models, its predictive accuracy cannot be compared with conventional critical values. Rather, the effect of model specification search must be taken into account. To this end, White (2000*b*) proposed a bootstrap reality check that facilitates calculation of adjusted critical values for the single best model and Hansen (2005) proposed various refinements to this approach; see also Politis & Romano (1995). This literature considers mining across model specifications, but takes the sample split point as given. Instead the forecast model is kept constant in our analysis, and any mining is confined to the sample split. This makes a material difference and introduces some unique aspects in our analysis.

The temporal dependence in forecast performance measured across different sample splits is very different from the cross-sectional dependencies observed in the forecasting performance measured across different model specifications. While the evaluation samples are identical in the bootstrap reality check literature, they are only partially overlapping when different sample splits are considered. Moreover, the recursive updating scheme for the parameter estimates of the forecast model introduces a common source of heteroskedasticity and persistence across different sample splits.

In a paper written independently and concurrently with our work, Rossi & Inoue (2011) study the effect of "mining" over the length of the estimation window in out-of-sample forecast evaluations. While the topic of their paper is closely related to ours there are important differences, which we discuss in details in Section 4.

The outline of the paper is as follows. Section 2 introduces the theory through linear regression models, while the power of out-of-sample tests is addressed in Section 3. A test that is robust to mining over the split point is proposed in Section 4, and Section 5 presents empirical applications to forecasts of U.S. stock returns and U.S. inflation. Section 6 concludes.

2 Theory

We focus on the common case where forecasts are produced by linear models estimated through recursive least squares and forecast accuracy is evaluated using mean squared error (MSE) loss. Other estimation schemes such as a rolling window or a fixed window could be considered and would embody slightly different trade-offs. However, in a stationary environment, recursive estimation based on an expanding data window makes most efficient use of the data.

Our analysis uses a regression setup that is first illustrated through a simple example which then is extended to more general regression models.

2.1 A Simple Illustrative Example

Consider the simple regression model that includes only a constant:

$$y_t = \beta + \varepsilon_t, \qquad \varepsilon_t \sim (0, \sigma_\varepsilon^2). \tag{1}$$

Suppose that β is estimated recursively by least squares, so that $\hat{\beta}_t = \frac{1}{t}\sum_{s=1}^{t} y_s$. The prediction of y_{t+1} given information at time t is then given by

$$\hat{y}_{t+1|t} = \hat{\beta}_t.$$

The least squares forecast is compared to a simple benchmark forecast

$$\hat{y}^b_{t+1|t} = 0.$$

This can be interpreted as the regression-based forecast under the assumption that $\beta = 0$, so that no regression parameters need to be estimated.

For purposes of out-of-sample forecast evaluation, the sample is divided into two parts. A fraction, $\rho \in (0,1)$, of the sample is reserved for initial parameter estimation while the remaining fraction, $1-\rho$, is used for evaluation. Thus, for a given sample size, n, the initial estimation period is $t = 1, \ldots, n_\rho$ and the (out-of-sample) evaluation period is $n_\rho + 1, \ldots, n$, where $n_\rho = \lfloor \rho n \rfloor$ is the integer part of ρn.

Forecasts are evaluated by means of their out-of-sample MSE-values measured relative to those of the benchmark forecasts:

$$D_n(\rho) = \sum_{t=n_\rho+1}^{n} (y_t - \hat{y}^b_{t|t-1})^2 - (y_t - \hat{y}_{t|t-1})^2. \tag{2}$$

Given a consistent estimator of σ^2_ε such as $\hat{\sigma}^2_\varepsilon = [(1-\rho)n]^{-1}\sum_{t=n_\rho+1}^{n}(y_t - \hat{y}_{t|t-1})^2$, under the null hypothesis, $H_0 : \beta = 0$, it can be shown that

$$T_n(\rho) = \frac{D_n(\rho)}{\hat{\sigma}^2_\varepsilon} \xrightarrow{d} 2\int_\rho^1 u^{-1}B(u)\mathrm{d}B(u) - \int_\rho^1 u^{-2}B(u)^2\mathrm{d}u, \tag{3}$$

where $B(u)$ is a standard Brownian motion, see McCracken (2007). The right hand side of (3) characterizes the limit distribution of the test statistic, and we denote the corresponding CDF by $F_{\rho,1}(x)$. Later we will introduce similar distributions deduced from multivariate Brownian motions, which explains the second subscript of F. For a given value of ρ, $T_n(\rho)$ can be compared to the critical values tabulated in McCracken (2007, table 4). Alternatively, the p-value can be computed directly by

$$p(\rho) = 1 - F_{\rho,1}(t), \qquad \text{where } t = T_n(\rho).$$

Since $T_n(\rho) \xrightarrow{d} F_{\rho,1}$ and $F_{\rho,1}(t)$ is continuous, it follows that the asymptotic distribution of $p(\rho)$ is the uniform distribution on $[0,1]$.

One contribution of this paper is to show that the expression in (3) can be greatly simplified. As we shall see, the limit distribution in (3) is simply given by $\sqrt{1-\rho}(Z_1^2 - Z_2^2) + \log\rho$, where Z_1 and Z_2 are independent standard normal random variables.

2.1.1 Mining over the Sample Split Point: Actual Type I Error Rate

Since the choice of ρ is somewhat arbitrary, a researcher may have computed p-values for several values of ρ. Such practices raise the danger of a subtle bias affecting predictive accuracy tests which are only valid provided that ρ is predetermined and not selected after observing the data. In particular, it suggests treating the sample split point as a choice variable which could depend on the observed data.

Suppose that the sample split point, n_ρ, is used as a choice parameter, and the reported p-value is in fact the smallest p-value obtained over a range of sample splits, such as

$$p_{\min} \equiv \min_{\underline{\rho}\leq\rho\leq\bar{\rho}} p(\rho), \qquad \text{with} \quad 0 < \underline{\rho} \leq \rho < \bar{\rho} < 1.$$

Clearly this is no longer a valid p-value, because the basic requirement of a p-value, $\Pr(p_{\min} \leq \alpha) \leq \alpha$, does not hold for the smallest p-value which represents a "worst case" scenario.[4] Note that we bound the range of admissible values of ρ away from both zero and one. Excluding a proportion of the data at the beginning and end of the sample is common practice and ensures that the distribution of the out-of-sample forecast errors is well behaved.

To illustrate this point, Figure 1 plots the limit distribution of $p_{\min}$ as a function of the nominal critical level, α. The distribution is shown over its full support along with a close-up of the lower range of the distribution that is relevant for testing at conventional significance levels. The extent to which the CDF is above the 45 degree line reveals the over-rejections arising from the search over possible split points. For example, the CDF of $p_{\min}$ is about 15% when evaluated at a 5% critical level, which tells us that there is a 15% probability that the *smallest* p-value, $\min_{0.1\leq\rho\leq0.9}\{p(\rho)\}$, is less than 5%. The figure clearly shows how sensitive out-of-sample predictive inference can be to mining over the sample split point.

It turns out that this mining is most sensitive to sample splits occurring towards the end of the sample. For example, we find $\min_{0.8\leq\rho\leq0.9} p(\rho) \leq 0.05$ with a probability that exceeds 10%. Even a relatively modest mining over split points towards the end of the

[4] For simplicity, the notation suppresses the dependence of $p_{\min}$ on $\underline{\rho}$ and $\bar{\rho}$.

sample can result is substantial over-rejection. To see this, Figure 2 shows the location of the smallest p-value, as defined by

$$\left\{ \rho_{\min} : p(\rho_{\min}) = \min_{10\% \leq \rho \leq 90\%} p(\rho) \right\}.$$

The location of the smallest p-value, $\rho_{\min}$, is a random variable with support on the interval $[0.1, 0.9]$. The histograms in Figure 2 reveal that under the null hypothesis the smallest p-value is more likely to be located late in the sample (i.e., between 80% and 90% of the data). The three other panels of Figure 2 show the location of $\rho_{\min}$ under the local alternatives, $\beta = c\frac{\sigma_\varepsilon}{\sqrt{n}}$, with $c = 2$, $c = 3$, and $c = 4$. As the value of c approaches zero, the histogram under the local alternative approaches that of the null hypothesis. For more distant local alternatives such as $c = 5$, it is very unlikely that the smallest p-value is found late in the sample.

These findings suggest, first, that conventional tests of predictive accuracy that assume a fixed and pre-determined value of ρ can substantially over-reject the null of no predictive improvement over the benchmark when in fact ρ is chosen to maximize predictive performance. Second, spurious rejection of the null hypothesis is most likely to be found with a sample split that leaves a relatively small proportion of the sample for out-of-sample evaluation. Conversely, true rejections of a false null hypothesis are more likely to produce a small p-value if the sample split occurs relatively early in the sample.

These are important considerations. It is quite common to use a short evaluation sample. However, our analysis suggests that short forecast evaluation samples are associated with a higher chance of spurious rejection.

2.2 General Case

Next, consider the general case in which the benchmark model has k regressors, $X_{1t} \in \mathbb{R}^k$, whereas the alternative forecast model is based on a larger regression with $k+q$ regressors, $X_t = (X_{1t}', X_{2t}')' \in \mathbb{R}^{k+q}$, which nests the benchmark model.[5] Forecasts could be computed multiple steps ahead. Letting $h \geq 1$ denote the forecast horizon, the benchmark model's regression-based forecast is now given by

$$\hat{y}_{t+h|t}^b = \tilde{\beta}_{1,t}' X_{1t}, \tag{4}$$

[5] West (1996) considers the non-nested case.

with

$$\tilde{\beta}_{1,t} = \left(\sum_{s=1}^{t} X_{1,s-h} X_{1,s-h}' \right)^{-1} \sum_{s=1}^{t} X_{1,s-h} y_s,$$

while the alternative forecast is

$$\hat{y}_{t+h|t} = \hat{\beta}_{1,t}' X_{1t} + \hat{\beta}_{2,t}' X_{2t}, \tag{5}$$

where $\hat{\beta}_t = (\hat{\beta}_{1,t}', \hat{\beta}_{2,t}')'$ is the least squares estimator from regressing y_s on $(X_{1,s-h}', X_{2,s-h}')'$, for $s = 1, \ldots, t$. For simplicity, we suppress the horizon subscript, h, on the least squares estimators.

The test statistic takes the same form as in our earlier example,

$$T_n(\rho) = \frac{\sum_{t=n_\rho+1}^{n} (y_t - \hat{y}_{t|t-h}^b)^2 - (y_t - \hat{y}_{t|t-h})^2}{\hat{\sigma}_\varepsilon^2}, \tag{6}$$

but its asymptotic distribution is now given from a convolution of q independent random variables, $2\int_\rho^1 u^{-1} B(u) \mathrm{d}B(u) - \int_\rho^1 u^{-2} B(u)^2 \mathrm{d}u$, as we make precise below in Theorem 1.

The asymptotic distribution is derived under assumptions that enable us to utilize the results for near-epoch dependent (NED) processes established by De Jong & Davidson (2000). We also formulate mixing assumptions (similar to those made in Clark & McCracken (2005)) that enable us to utilize results in Hansen (1992). The results in Hansen (1992) are more general than those established in De Jong & Davidson (2000) in ways that are relevant for our analysis of the split-mining robust test in Section 4.

In the assumptions below we consider the process, $V_t = (y_t, X_{t-h}')'$, and let $\mathcal{V}_t$ be some auxiliary process that defines the filtration $\mathcal{F}_{t-m}^{t+m} = \sigma(\mathcal{V}_{t-m}, \ldots, \mathcal{V}_{t+m})$.

Assumption 1 *The matrix,* $\Sigma_{vv} = \mathrm{E}(V_t V_t')$, *is positive definite and does not depend on* t, *and* $\mathrm{var}[n^{-1/2} \sum_{t=1}^{\lfloor un \rfloor} \mathrm{vech}(V_t V_t' - \Sigma_{vv})]$ *exists for all* $u \in [0, 1]$.

The first part of the assumption ensures that the population predictive regression coefficients do not depend on t while the second part, in conjunction with Assumption 2 stated below, ensures that we can establish the desired limit results.

Assumption 2 *For some* $r > 2$, (i) $\|V_t\|_{2r}$ *is bounded uniformly in* t; (ii) $\left\|V_t - E(V_t|\mathcal{F}_{t-m}^{t+m})\right\|_4 \leq d_t \nu(m)$, *where* $\nu(m) = O(n^{-1/2-\epsilon})$ *for some* $\epsilon > 0$ *and* d_t *is a uniformly bounded sequence of constants;* (iii) $\mathcal{V}_t$ *is either* α*-mixing of size* $-r/(r-2)$, *or* ϕ*-mixing of size* $-r/(2(r-1))$.

Assumption 2 establishes V_t as an L_4-NED process of size $-\frac{1}{2}$ on $\mathcal{V}_t$, where the latter sets limits on the "memory" of V_t. The advantage of formulating our assumptions in terms of NED processes is that the dependence properties carries over to higher moments of the process. Specifically, $\text{vech}(V_tV_t')$ will be L_2-NED of size $-\frac{1}{2}$ on $\mathcal{V}_t$, and key stochastic integrals that show up in our limit results are derived from the properties of $\text{vech}(V_tV_t')$.

It is convenient to express the block structure of Σ_{vv} in the following ways

$$\Sigma_{vv} = \begin{pmatrix} \Sigma_{yy} & \bullet \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \qquad \text{with} \quad \Sigma_{xx} = \begin{pmatrix} \Sigma_{11} & \bullet \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where the blocks in Σ_{xx} refer to X_{1t} and X_{2t}, respectively. Similarly, define the "error" term from the large model

$$\varepsilon_t = y_t - \Sigma_{yx}\Sigma_{xx}^{-1}X_{t-h},$$

and the auxiliary variable

$$Z_t = X_{2t} - \Sigma_{21}\Sigma_{11}^{-1}X_{1t},$$

so that Z_t is constructed to be the part of X_{2t} that is orthogonal to X_{1t}.

Further, define the population objects, $\sigma_\varepsilon^2 = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ and $\Sigma_{zz} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. It follows that $\sigma_\varepsilon^2 > 0$ and that Σ_{zz} is positive definite, because Σ_{vv} is positive definite. Finally, define

$$W_n(u) := \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor un \rfloor} Z_{t-h}\varepsilon_t, \tag{7}$$

which is a CADLAG on the unit interval that maps into $\mathbb{R}^q$. The space of such functions is denoted $\mathbb{D}_{[0,1]}^q$. Two important matrices in our asymptotic analysis are

$$\Omega := \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{s,t=1}^{n} Z_{s-h}\varepsilon_s\varepsilon_t Z_{t-h}' \quad \text{and} \quad \Sigma = \sigma_\varepsilon^2\Sigma_{zz},$$

where the former is the long-run variance of $\{Z_{t-h}\varepsilon_t\}$. From Assumption 1 it follows that both Ω and Σ are well defined and positive definite.

We shall make use of the following mixing assumption:

Assumption 2' *For some $r > \beta > 2$, $w_t = Z_{t-h}\varepsilon_t$ is an α-mixing sequence with mixing coefficients of size $r\beta/(r-\beta)$ and $\sup_t \mathrm{E}|w_t^r| < C < \infty$.*

We then have the following theorem:

Theorem 1 *Given Assumptions 1 and 2, or Assumptions 1 and 2', we have*

$$W_n(u) \Rightarrow W(u) = \Omega^{1/2}B(u),$$

where $B(u)$ is a standard q-dimensional Brownian motion.

This result shows that a functional central limit theorem applies to that part of the score from the "large" prediction model that differentiates it from the nested benchmark model. The result is needed for hypothesis tests that rely on the relative accuracy of the two models.

The next assumption is a mild additional requirement that is easy to verify if the prediction errors are unpredictable in the sense that $\mathrm{E}(\varepsilon_{t+j}|\varepsilon_t, Z_t, \varepsilon_{t-1}, Z_{t-1}, \ldots) = 0$ for $j \geq h$.

Assumption 3 $\mathrm{cov}(Z_{t-h}\varepsilon_t, Z_{s-h}\varepsilon_s) = 0$ *for* $|s-t| \geq h$.

This assumption requires a mild form of unpredictability of the h-step-ahead forecast errors. Without it there would be an asymptotic bias term in the limit distribution given below.

We can now present the limit distribution of the test statistic $T_n(\rho)$ for the general case.

Theorem 2 *Suppose Assumptions 1, 2 and 3 or 1, 2' and 3 hold and* $\hat{\sigma}_\varepsilon^2 \xrightarrow{p} \sigma_\varepsilon^2$. *Under the null hypothesis,* $H_0 : \beta_2 = 0$, *we have*

$$T_n(\rho) \xrightarrow{d} \sum_{j=1}^{q} \lambda_j \left[2\int_\rho^1 u^{-1} B_j(u) \mathrm{d}B_j(u) - \int_\rho^1 u^{-2} B_j(u)^2 \mathrm{d}u \right],$$

where $\lambda_1, \ldots, \lambda_q$ *are the eigenvalues of* $\Sigma^{-1}\Omega$, *and* $B_j(u), j = 1, ..., q$, *are independent standard Brownian motion processes.*

The limit distribution of the test statistic in Theorem 2 can also be expressed as

$$2\int_\rho^1 u^{-1} B'(u) \Lambda \mathrm{d}B(u) - \int_\rho^1 u^{-2} B'(u) \Lambda B(u) \mathrm{d}u, \tag{8}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$, and we denote the CDF of this distribution by $F_{\rho,\Lambda}$. The standard Brownian motion, B, that appears in Theorem 2 and equation (8) characterizes the limit distribution. This Brownian motion need not be identical to that used in Theorem 1. In fact, one is a simple orthonormal rotation of the other; see the proof for details.

The expression for the limit distribution in Theorem 2 involves two types of random variables. The first term is the stochastic integral, $\int_\rho^1 u^{-1} B'(u) \Lambda \mathrm{d}B(u)$, that arises from the recursive estimation scheme. The second term, $-\int_\rho^1 u^{-2} B'(u) \Lambda B(u) \mathrm{d}u$, is a non-positive random variable that characterizes the prediction loss induced by the estimation error, which

arises from the additional parameters in the larger model. Stated somewhat informally, prediction errors map into $\mathrm{d}B(u)$ and parameter estimation errors map into $B(u)$. In the recursive estimation scheme, prediction errors influence parameter estimates in subsequent predictions.

Our expression for the asymptotic distribution in Theorem 2 is simpler than that derived in Clark & McCracken (2005). For instance, our expression simplifies the nuisance parameters to a diagonal matrix, Λ, as opposed to a full $q \times q$ matrix. Moreover, it is quite intuitive that the "weights", $\lambda_1, \ldots, \lambda_q$, that appear in the diagonal matrix, Λ, are given as eigenvalues of $\Sigma^{-1}\Omega$, because the two matrices play a similar role to that of the two types of information matrices that can be computed in quasi maximum likelihood analysis, see White (1994).

λ_1,..., λ_q, can be consistently estimated as the eigenvalues of $\hat{\Sigma}^{-1}\hat{\Omega}$, where

$$\hat{\Sigma} = \hat{\sigma}_\varepsilon^2 \frac{1}{n} \sum_{t=1}^{n} \hat{Z}_{t-h}\hat{Z}'_{t-h}, \qquad \hat{\Omega} = \sum_i k(\tfrac{i}{b_n})\hat{\Gamma}_i.$$

Here $k(\cdot)$ is a kernel function, e.g., the Parzen kernel, b_n is a bandwidth parameter, and

$$\hat{\Gamma}_j = \frac{1}{n} \sum_{t=1}^{n} \hat{Z}_{t-h}\hat{Z}'_{t-h-j}\hat{\varepsilon}_t\hat{\varepsilon}_{t-j},$$

with $\hat{Z}_t = X_{2t} - \sum_{s=1}^{t} X_{2s}X'_{1s}(\sum_{s=1}^{t} X_{1s}X'_{1s})^{-1}X_{1t}$ and $\hat{\varepsilon}_t = y_t - \hat{\beta}'_{t-h}X_{t-h}$. In the absence of autocorrelation in $Z_{t-h}\varepsilon_t$, one can use the estimate $\hat{\Omega} = \frac{1}{n}\sum_{t=1}^{n} \hat{Z}_{t-1}\hat{Z}'_{t-1}\hat{\varepsilon}_t^2$. This situation may apply when $h = 1$. In the homoskedastic case, $\sigma_\varepsilon^2 = E[\varepsilon_t^2|Z_{t-h}] = E[\varepsilon_t^2]$, $\Lambda = I_{q\times q}$, we can simplify the notation $F_{\rho,\Lambda}$ to $F_{\rho,q}$. This is consistent with the notation used in our simplified (univariate and homoskedastic) example. The homoskedastic result is well known in the literature, see McCracken (2007).

2.3 Simplification of Stochastic Integrals

Generating critical values for the distribution of $2\int_\rho^1 u^{-1}B\mathrm{d}B - \int_\rho^1 u^{-2}B^2\mathrm{d}u$ has so far proven computationally burdensome because it involves both a discretization of the underlying Brownian motion and drawing a large number of simulations. McCracken (2007) presents a table with critical values based on a 5,000-point discretization of the Brownian motion and 10,000 repetitions. This design makes the first decimal point in the critical values somewhat accurate. The analytical result in the next Theorem provide a major simplification of the asymptotic distribution.

Theorem 3 *Let $B(u)$ be a standard Brownian motion and $\rho \in (0,1)$. Then*

$$2\int_\rho^1 u^{-1}B(u)\mathrm{d}B(u) - \int_\rho^1 u^{-2}B(u)^2\mathrm{d}u = B^2(1) - \rho^{-1}B^2(\rho) + \log\rho. \tag{9}$$

The derivation of Theorem 3 can be illustrated using Ito calculus. Consider $F_t = \frac{1}{t}B_t^2 - \log t$, for $t > 0$ so that

$$\partial F_t/\partial B_t = \tfrac{2}{t}B_t, \quad \partial^2 F_t/(\partial B_t)^2 = \tfrac{2}{t}, \quad \text{and} \quad \partial F_t/\partial t = -\left(\tfrac{1}{t^2}B_t^2 + \tfrac{1}{t}\right).$$

Then by Ito calculus we have

$$\mathrm{d}F_t = \left[\tfrac{\partial F_t}{\partial t} + \tfrac{1}{2}\tfrac{\partial^2 F_t}{(\partial B_t)^2}\right]\mathrm{d}t + \tfrac{\partial F_t}{\partial B_t}\mathrm{d}B_t = -\tfrac{1}{t^2}B_t^2\mathrm{d}t + \tfrac{2}{t}B_t\mathrm{d}B_t,$$

so that

$$\int_\rho^1 \tfrac{2}{t}B_t\mathrm{d}B_t - \int_\rho^1 \tfrac{1}{t^2}B_t^2\mathrm{d}t = \int_\rho^1 \mathrm{d}F_t = F_1 - F_\rho = B_1^2 - B_\rho^2/\rho + \log\rho.$$

A more detailed proof of Theorem 3 is provided in the Appendix.

Theorem 3 establishes that the limit distribution is given as a very simple transformation of two random variables. Apart from the constant, $\log\rho$, the distribution is simply the difference between two (dependent) χ_1^2-distributed random variables, as we next show:

Corollary 1 *Let Z_1 and Z_2 be independently distributed, $Z_i \sim N(0,1)$, $i = 1,2$. Then the distribution in Theorem 3 is given by*

$$\sqrt{1-\rho}(Z_1^2 - Z_2^2) + \log\rho.$$

Because the distribution is expressed in terms of two independent χ^2-distributed random variables, in the homoskedastic case where $\lambda_1 = \cdots = \lambda_q = 1$ it is possible to obtain relatively simple closed form expressions for the distribution in Theorem (2):

Corollary 2 *The density of $\sum_{j=1}^q \left[2\int_\rho^1 u^{-1}B_j(u)\mathrm{d}B_j(u) - \int_\rho^1 u^{-2}B_j(u)^2\mathrm{d}u\right]$ is given by*

$$f_1(x) = \tfrac{1}{2\pi\sqrt{1-\rho}}\mathrm{K}_0(\tfrac{|x-\log\rho|}{2\sqrt{1-\rho}}),$$

for $q = 1$, where $\mathrm{K}_0(x) = \int_0^\infty \frac{\cos(xt)}{\sqrt{1+t^2}}\mathrm{d}t$ is the modified Bessel function of the second kind. For $q = 2$ we have

$$f_2(x) = \tfrac{1}{4\sqrt{1-\rho}}\exp\left(-\tfrac{|x-2\log\rho|}{2\sqrt{1-\rho}}\right),$$

which is the noncentral Laplace distribution.

The densities for $q = 3, 4, 5, \ldots$ can be obtained based on those stated in Corollary 2.

When $q = 2$, we obtain an analytical expression for the CDF from the Laplace distribution:

$$F_{\rho,2}(x) = \begin{cases} \frac{1}{2}\exp\left(\frac{x/2-\log\rho}{\sqrt{1-\rho}}\right) & x < \log\rho \\ 1 - \frac{1}{2}\exp\left(\frac{-x/2+\log\rho}{\sqrt{1-\rho}}\right) & x \geq \log\rho \end{cases}.$$

The associated critical values are therefore given from the quantile function

$$F^{-1}_{\rho,2}(p) = \begin{cases} 2[\log\rho + \sqrt{1-\rho}\log(2p)] & p < 0.5, \\ 2[\log\rho - \sqrt{1-\rho}\log(2(1-p))] & p \geq 0.5. \end{cases}$$

In the present context we reject the null for large values of the test statistic, so for $\alpha \leq 0.5$ the critical value, c_2^α, is found by setting $p = 1 - \alpha$. Hence,

$$c_2^\alpha = 2[\log\rho - \sqrt{1-\rho}\log(2\alpha)], \qquad \alpha \leq 0.5.$$

These results greatly simplify calculation of critical values for the limiting distribution of the test statistics. We next make use of them to illustrate the rejection rates induced by mining over the sample split.

Table 1 compares the exact critical values to those provided by McCracken (2007) for different values of ρ between 0.33 and 0.90 or, equivalently for $\pi = P/R$ between 0.1 and 2, using the notation in McCracken. To save space, we only show results for $q = 2$ and consider three levels of α, namely $\alpha = 0.90, 0.95$ and 0.99. The two sets of critical values are generally close and practical inference is unlikely to be overturned by the differences. However, our approach makes it far more convenient to compute critical values outside the cases tabulated by McCracken, particularly in cases where q is large.

2.4 Rejection Rates Induced by Mining over the Sample Split

When the sample is divided so that a predetermined fraction, ρ, is reserved for initial estimation of model parameters, and the remaining fraction, $1 - \rho$, is left for out-of-sample evaluation, we obtain the $T_n(\rho)$-statistic in (6). This statistic can be used to test the null hypothesis, $\beta_2 = 0$, by simply comparing it to the critical values from $F_{\rho,\Lambda}$. For instance, if $c_\alpha(\rho)$ is the $1 - \alpha$ quantile of $F_{\rho,\Lambda}$, i.e., $c_\alpha(\rho) = F^{-1}_{\rho,\Lambda}(1-\alpha)$, it follows that

$$\lim_{n\to\infty} \Pr(T_n(\rho) > c_\alpha(\rho)) = \alpha.$$

Suppose instead that the out-of-sample test statistic, T_ρ, is computed over a range of split points, $\underline{\rho} \leq \rho \leq \bar{\rho}$, in order to find a split point where the alternative is most favored by

the data. This corresponds to mining over the sample split, and the inference problem becomes similar to the situation where one tests for structural change with an unknown change point, see, e.g., Andrews (1993).

To explore the importance of such mining over the sample split for the actual rejection rates, we compute how often the test based on the asymptotic critical values in McCracken (2007) would reject the null of no predictability.

Table 2 presents the actual rejection rates based on the asymptotic critical values in McCracken (2007) for $\alpha = 0.01, 0.05, 0.10, 0.20$, using $q = 1, ..., 5$ additional predictor variables in the alternative model. These numbers are computed as the proportion of paths, $u \in [\underline{\rho}, \bar{\rho}]$ with $\underline{\rho} = 1 - \bar{\rho} = 0.1$, for which at least one rejection of the null occurs at the nominal α level. The computations are based on $N = 10,000$ simulations (simulated paths) and a discretization of the underlying Brownian motion, $B(u) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor un \rfloor} z_i$, with $n = 10,000$ and $z_i \sim \text{iid}N(0,1)$.

The results are very strong. The inflation in the rejection rate from 5% to 15% reported earlier with one additional regressor ($q = 1$) increases to nearly 22% as q rises from one to five. Similar results hold no matter which critical level the test is conducted at. For example, at the $\alpha = 1\%$ nominal level, mining over the sample split point leads to rejection rates between 3.7% and 5.5%, both far larger than the nominal critical level. When the test is conducted at the $\alpha = 10\%$ nominal level, the test that mines over split points actually rejects between 25% and 38% of the time for values of q between one and five, while for $\alpha = 20\%$, rejection rates above 60% are observed for the larger models.

2.5 A ρ-Invariant Asymptotically Pivotal Test Statistic

The limit distribution of $T_n(\rho)$ motivates the simple transformation,

$$S_n(\rho) = \frac{T_n(\rho) - q \log \rho}{\sqrt{1-\rho}}, \tag{10}$$

which defines a test statistic that has a ρ-invariant limit distribution in the homoskedastic case.

Corollary 3 *Suppose that the assumptions of Theorem 2 hold and that $\Lambda = I$. Then the limit distribution of $S_n(\rho)$ in (10) is given by*

$$\xi_1 - \xi_2,$$

where ξ_1, ξ_2 are independent χ^2-distributed with q degrees of freedom.

Note that the limit distribution of S_n does not depend on any nuisance parameters so that $S_n(\rho)$ is asymptotically pivotal. The fact that the limit distribution does not depend on ρ in this case is convenient. Unlike in the case with $T_n(\rho)$, it is not necessary to tabulate critical values for different values of ρ. However, the homoskedasticity required for $\Lambda = I$ is unrealistic in most empirical applications. The dependence on ρ could still be removed asymptotically using the definition $S_n(\rho) = (T_n(\rho) - \text{tr}\{\hat{\Sigma}^{-1}\hat{\Omega}\}\log\rho)/\sqrt{1-\rho}$, but the limit distribution would still depend on $\lambda_1, \ldots, \lambda_q$. Consequently, in most practical situations the effort required to make a test based on $S_n(\rho)$ would be identical to that using a test based on $T_n(\rho)$.

3 Power of the Test

The scope for size distortions in conventional tests of predictive accuracy is only one issue that arises when considering the sample split for forecast evaluation purposes, with the power of the test also mattering. Earlier we found that the risk of spuriously rejecting the null due to sample split mining is highest when the sample split occurs towards the end of the sample. This section shows that, in contrast, the power of the predictive accuracy test is highest when the sample split occurs early in the sample.

Specifically, under a local alternative hypothesis we have the following result:

Theorem 4 *Suppose that Assumptions 1-3 hold, and consider the local alternative $\beta_{n,2} = \frac{c}{\sqrt{n}}a'$, where $a \in \mathbb{R}^q$ with $a'\Sigma_{zz}a = \sigma_\varepsilon^{-2}$. Then*

$$\begin{aligned} T_n(\rho) \quad &\xrightarrow{d} \quad c^2(1-\rho) + 2\frac{c}{\sigma_\varepsilon}a'\Omega^{1/2}Q'\left[B(1) - B(\rho)\right] \\ &\quad + \sum_{j=1}^{q}\lambda_j\left[B_j^2(1) - \rho^{-1}B_j^2(\rho) + \log\rho\right], \end{aligned}$$

where the matrix Q and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_q)$ are obtained from $Q'\Lambda Q = \Omega^{1/2}\Sigma^{-1}\Omega^{1/2}$.

This Theorem establishes the analytical theory that underlies the simulation results presented in Clark & McCracken (2001, tables 4 and 5).

For a given sample size and a particular alternative of interest, e.g., $\beta_{n,2} = b$, the theorem yields an asymptotic approximation to the finite sample distribution. To this end, simply set $a = \frac{1}{\kappa}b$, where $\kappa^2 = \sigma_\varepsilon^2 b'\Sigma_{zz}b$ and $c = \kappa\sqrt{n}$, so that $a'\Sigma_{zz}a = \sigma_\varepsilon^{-2}$ and $b = \frac{c}{\sqrt{n}}a$.

Insight about the power of the test and its dependence on ρ can be gained by considering the asymptotically pivotal quantity $S_n(\rho) = (T_n(\rho) - q \log \rho)/\sqrt{1-\rho}$ in the homoskedastic case $\Lambda = I$. In this case its limit distribution does not depend on ρ under the null hypothesis, nor does it depend on any other nuisance parameters. Under the alternative hypothesis, the non-centrality parameter associated with $S_n(\rho)$, which is key for the power of the test, is given by $c^2\sqrt{1-\rho}$. Thus, the non-centrality parameter is strictly decreasing in ρ, which strongly suggests that the power is decreasing in ρ. However the power of the test is is also influenced by a random term as is evident from Theorem 4. In the univariate case this term is proportional to cU where $U = (B(1) - B(\rho))/\sqrt{1-\rho} \sim N(0,1)$. While its distribution is ρ-invariant, it is not independent of the Brownian motion that defines the null distribution, so its impact on the power of the test is not entirely clear.

3.1 Local Power in the Illustrative Example

In our illustrative example from Section 2.1, $a'\Sigma_{zz}a = \sigma_\varepsilon^{-2}$ with $\Sigma_{zz} = 1$ implies that $a = \sigma_\varepsilon$, so a local alternative takes the form

$$\beta = \frac{c}{\sqrt{n}}\sigma_\varepsilon,$$

and the limit distribution is given by

$$T_n(\rho) \xrightarrow{d} B^2(1) - \rho^{-1}B^2(\rho) + \log\rho + c^2(1-\rho) + 2c\left[B(1) - B(\rho)\right].$$

How the power depends on the split point can be illustrated by the distribution of the p-value, defined by $p(\rho) = 1 - F_{\rho,1}(T_n(\rho))$, under the local alternative. Figure 3 presents the power of the test as a function of ρ for four local alternatives, $c = 1$, $c = 2$, $c = 3$, and $c = 4$ based on a test conducted at the nominal 5%-level. The power is decreasing in ρ which makes it difficult to justify using a late sample split with this test.

Empirical studies tend to use a relatively large estimation period, i.e., a large ρ. This is precisely the range where one is most likely to find spurious rejections of the null hypothesis. In fact, the power of the $T_n(\rho)$ test provides a strong argument for adopting a smaller (initial) estimation sample, i.e., a small value of ρ.

4 A Split-Mining Robust Test

The results in Table 2 demonstrate that mining over the start of the out-of-sample period can substantially raise the rejection rate when its effects are ignored. A question that naturally arises from this finding is how to design a suitable test that is robust to sample split mining in the sense that it will correctly reject at the stipulated rate even if such mining took place.

To address this, suppose we want to guard ourselves against mining over the range $\rho \in [\underline{\rho}, \bar{\rho}]$. One possibility is to consider the maximum value of $T_n(\rho)$ across a range of split points. However, $\max_{\rho\in[\underline{\rho},\bar{\rho}]} T_n(\rho)$ is ill-suited for this purpose, because the marginal distribution of $T_n(\rho)$ varies a great deal with ρ, both in terms of scale and location. The implication is that critical values for max-$T_n(\rho)$ will be disproportionately influenced by certain ranges for ρ, and distribute power unevenly over different values of ρ in an arbitrary manner.

These observations suggest redefining the test statistic so as to make its limit distribution less sensitive to ρ. For instance, we could consider $S_n(\rho) = (T_n(\rho) - q\log\rho)/\sqrt{1-\rho}$ whose limit distribution is invariant to ρ in the homoskedastic case as shown in Corollary 3, but unfortunately not in the heteroskedastic case. Instead we pursue a method, well known in the literature on multiple testing, that combines individual p-values.

Specifically, we first map the test statistics for each of the sample split points into nominal p-values, $p(\rho) = 1 - F_{\rho,\Lambda}(T_n(\rho))$. Next, the smallest p-value is computed:

$$p_{\min} = \min_{\rho\in[\underline{\rho},\bar{\rho}]} p(\rho).$$

Because each of the p-values, $p(\rho)$, is asymptotically uniformly distributed on the unit interval, the resulting test statistic is constructed from test statistics with similar properties, see, e.g., Westfall & Young (1993). The limit distribution of $p_{\min}$ will clearly not be uniformly distributed and so cannot be interpreted as a valid p-value, but should instead be viewed as a test statistic, whose distribution we seek. To this end, let B denote a q-dimensional standard Brownian motion and for $u \in (0,1)$ define

$$G(u) = B(1)'\Lambda B(1) - u^{-1}B(u)'\Lambda B(u) + \log u.$$

To establish the asymptotic properties of $p_{\min}$ we will need a stronger convergence result than that used earlier to derive the distribution of $T_n(\rho)$ for a fixed value of ρ. Specifically,

we need that

$$T_n(u) \Rightarrow G(u), \qquad \text{on } \mathbb{D}_{[\underline{\rho},\bar{\rho}]}. \tag{11}$$

The stronger result holds under mixing assumptions, see Hansen (1992), but has not been established under near-epoch assumptions. It is worth noting that the near-epoch conditions are the weakest set of assumptions needed for the functional central limit theorem and the (point-wise) convergence to the stochastic integral, see De Jong & Davidson (2000), so it may be redundant to state (11) as an additional assumption in the near-epoch setting.

Theorem 5 *Given Assumptions 1-3 and (11), or Assumptions 1, 2' and 3,* $p_{\min}$ *converges in distribution, and the cdf of the limit distribution is given by*

$$F(\alpha) = \Pr\{\sup_{\underline{\rho},\leq u\leq \bar{\rho}} [G(u) - c_\alpha(u)] \geq 0\}, \qquad \alpha \in [0,1],$$

where $G(u)$ *is given above and*

$$c_\alpha(u) = F^{-1}_{u,\Lambda}(1-\alpha).$$

Using this result, we can numerically compute the p-value adjusted for sample split mining by sorting the $p_{\min}$-values for a large number of sample paths and choosing the α-quantile of this (ranked) distribution.

Table 3 shows how nominal p-values map into p-values adjusted for any split-mining. For example, suppose a critical level of $\alpha = 5\%$ is desired and that $q = 1$. Then the smallest p-value computed using the McCracken (2007) test statistic for all possible split points $\rho \in [0, 1, 0.9]$ should fall below 1.3% for the out-of-sample evidence to be significant at the 5% level. This drops further to 1.1% when $q = 2$ and to a value below 0.1% (the smallest p-value considered in our calculations) for values of $q \geq 3$. Similarly, with a nominal rejection level of 10%, the smallest p-value (computed across all admissible sample splits) would have to fall below 2.9% when $q = 1$ and below 2% when $q = 5$. Clearly, mining over the sample split brings the adjusted critical values much further out in the tail of the distribution.

The robust test that we propose is related to the literature on multiple hypotheses testing. Each sample split results in a hypothesis test, with the special circumstance that it is the same hypothesis that gets tested at every sample split. The proposed test procedure seeks to control the familywise error rate. Combining p-values, rather than test statistics with distinct limit distributions, creates a degree of balance across hypothesis tests.

In a related paper, Rossi & Inoue (2011) consider methods for out-of-sample forecast evaluation that are robust to data snooping over the length of the estimation window and accounts for parameter instability. The first version of their paper was written concurrently and independently of the results in the present paper. The analysis in the first version of their paper mainly focused on the case with a rolling estimation window. However, in the latest version of their paper they also consider encompassing tests for the comparison of nested models. Under the recursive estimation scheme, the fraction of the sample used for the (initial) window length is identical to the choice of sample split, ρ, which is the focus of our paper. Despite the similarities in this special case, their approach is substantially different from ours.

First, their theoretical setup is based on high-level assumptions that must be verified for the problem at hand (see Rossi & Inoue (2011, appendix A) for a wide range of situations). These assumptions enable Rossi & Inoue (2011) to cover a lot of ground with the same framework, at the expense of shedding little light on the exact properties of the limit distribution, such as its intricate dependence on ρ. In contrast, we cover less ground but offer detailed analytical results for the limit distribution. Our results cast important light on issues such as where the smallest p-value is most likely to be found under the null and alternative hypothesis. Second, Rossi and Inoue provide finite-sample simulation results to illustrate the power of their test, whereas we have analytical power results. Third, they construct robust test procedures using an approach where a range of test statistics (based on different window sizes) are combined by either taking the supremum or the average. Instead, we combine statistics whose location and scale is insensitive to ρ, which makes them better suited for comparison. In the homoskedastic case, the test statistic $S_n(\rho)$ is well suited for this purpose, because its limit distribution does not depend on ρ. An alternative, and our preferred approach, is to combine the individual p-values, which allows for the case with heteroskedasticity. Specifically, we propose a minimum p-value test which makes the test statistics corresponding to different sample splits more comparable. The empirical findings in Rossi & Inoue (2011) are consistent with ours, however, and confirm that data snooping over the choice of estimation window can lead to significant size distortions.

5 Empirical Examples

This section provides empirical illustrations of the methods and results discussed previously. We consider two forecasting questions that have attracted considerable empirical interest in economics and finance, namely whether stock returns are predictable and whether inflation forecasts can be improved by using broad summary measures of the state of the economy in the form of common factors.

5.1 Predictability of U.S. stock returns

It is a long-standing issue whether returns on a broad U.S. stock market portfolio can be predicted using simple regression models, see, e.g., Keim & Stambaugh (1986), Campbell & Shiller (1988), Fama & French (1988), and Campbell & Yogo (2006). While these studies were concerned with in-sample predictability, papers such as Pesaran & Timmermann (1995), Campbell & Thompson (2008), Welch & Goyal (2008), Johannes, Korteweg & Polson (2009), and Rapach et al. (2010) study return predictability in an out-of-sample context. For example, in their analysis of return predictability covering the period 1947-2005, Rapach et al. (2010) use three different out-of-sample periods, namely 1965-2005, 1976-2005, and 2000-2005. This corresponds to using the last 70%, 50% and 10% of the sample, respectively, for out-of-sample forecast evaluation.

Welch & Goyal (2008) find that so-called prevailing mean forecasts generated by a constant equity premium model

$$y_{t+1} = \beta_1 + \varepsilon_{t+1},$$

lead to lower out-of-sample MSE-values than univariate forecasts from a range of prediction models of the form

$$y_{t+1} = \beta_1 + \beta_2 x_t + \varepsilon_{t+1}.$$

We focus on models where x_t is the default spread, measured as the difference between the yield on BAA-rated corporate bonds versus that on AAA-rated corporate bonds or the dividend yield, measured as dividends paid over the preceding 12-month period divided by the current stock price. Our data consist of monthly observations on stock returns on the S&P500 index and the corresponding default spread over the period 1926:01−2010:12, a total of 1020 observations. Setting $\underline{\rho} = 1 - \bar{\rho} = 0.1$, our initial estimation sample uses

102 observations and so the beginning of the various forecast evaluation periods runs from 1934:07 through 2002:05. The end point of the out-of-sample period is always 2010:12.

The top left window in Figure 4 shows how the $T_n(\rho)$-statistic evolves over the forecast evaluation period.[6] The minimum value obtained for $T_n(\rho)$ is -6.77, while its maximum is 2.01. Due to the partial overlap in both estimation and forecast evaluation windows, the test statistic evolves relatively smoothly and is quite persistent, although the effect of occasional return outliers is also clear from the plot.

The $p(\rho)$-values associated with the $T_n(\rho)$ statistics computed for different values of ρ are plotted in the bottom left window of Figure 4. There is little evidence of return predictability when the out-of-sample period begins after the mid-seventies. However, once the forecast evaluation period is expanded backwards to include the early seventies, evidence of predictability grows stronger. This is consistent with the finding by Pesaran & Timmermann (1995) and Welch & Goyal (2008) that return predictability was particularly high after the first oil shock in the seventies. For out-of-sample start dates running from the early fifties to the early seventies, p-values below 5-10% are consistently found. In contrast, had the start date for the out-of-sample period been chosen either before or after this period, then forecast evaluation tests, conducted at conventional critical levels, would have failed to reject the null of no return predictability.

Such sensitivity of the empirical results to the choice of ρ highlights the need to have a test that is robust to how the start of the out-of-sample period is determined. In fact, the smallest p-value, selected across the entire out-of-sample period $\rho \in [0.1, 0.9]$ is 0.034. Table 3 suggests that this corresponds to a split-mining adjusted p-value that exceeds 10%. Hence, the evidence of time-varying return predictability from the default spread is not statistically significant at conventional levels. We therefore cannot conclude that the lagged default spread model generates more precise out-of-sample forecasts of stock returns than a constant equity premium model, at least not in a way that is robust to how the beginning of the out-of-sample period is chosen.

We next consider a return forecasting model that uses the lagged dividend yield as the predictor variable. Using the same sample as above, for this model the maximum value of $T_n(\rho)$, plotted in the top right window in Figure 4, is 3.57 while the smallest p-value falls

[6] We use a Newey-West HAC estimator with four lags to estimate the variance of the residuals from the forecast model, $\hat{\sigma}^2_\varepsilon$.

below 0.001 which, according to Table 3, means that out-of-sample predictability from this model is robust to mining over the sample split. Interestingly, for this model, predictability is strongest when ρ lies either at the beginning or at the end of the sample, with the p-value reaching a value of 0.01 when the evaluation sample starts in the mid-thirties, then reaching even lower levels when the split point occurs in the late 1990s or subsequently.

5.2 Inflation Forecasts

Simple autoregressive prediction models have been found to perform well for many macroeconomic variables capturing wages, prices and inflation (Marcellino et al. (2006) and Pesaran et al. (2011)). However, as illustrated by the many studies using factor-augmented vector autoregressions and other factor-based forecasting models, it is also of interest to see whether the information contained in common factors, extracted from large-dimensional data, can help improve forecasting performance.

To address this issue, we consider out-of-sample predictability of U.S. inflation measured by the monthly log first-difference in the consumer price index (CPI) captured by the CPIAUSCL series. Our benchmark is a simple autoregressive specification with two lags:

$$y_{t+1} = \beta_0 + \sum_{i=1}^{2} \beta_{yi} y_{t+1-i} + \varepsilon_{y,t+1}, \tag{12}$$

where $y_{t+1} = \log(CPI_{t+1}/CPI)$ is the monthly growth rate in the consumer price index.

The alternative forecasting model adds four common factors to the AR(2) specification in (12):[7]

$$y_{t+1} = \beta_0 + \sum_{i=1}^{2} \beta_{yi} y_{t+1-i} + \sum_{i=1}^{4} \beta_{fi} \hat{f}_{it} + \varepsilon_{y,t+1}. \tag{13}$$

Here $\hat{f}_{it}$ is the i-th principal component (factor) extracted from a set of 131 economic variables. Data on these 131 variables is taken from Ludvigson & Ng (2009) and run from 1960 through 2007. We extract factors recursively from this data, initially using the first ten years of the data so the first point of factor construction is 1969:12. Setting $\underline{\rho} = 1 - \bar{\rho} = 0.1$, the start of the out-of-sample evaluation period runs from mid-1973 through early 2004.

The top left window in Figure 5 shows the $T_n(\rho)$-statistic for different values of ρ. This rises throughout most of the sample from -23 to a terminal value just above zero. The

[7] The empirical results are not sensitive to the number of autoregressive lags in the benchmark model or to the number of factors included in the extended model.

associated $p(\rho)$-values are shown in the bottom left window of Figure 5. These start close to one but drop significantly after the change in the Federal Reserve monetary policy in 1979. Between 1980 and 1982, the $p(\rho)$ plot declines sharply to values below 0.10, before oscillating for much of the rest of the sample, with an overall minimum p-value of 0.023. Hence, in this example a researcher starting the forecast evaluation period after 1979 and ignoring mining over the sample split might well conclude that the additional information from the four factors helped improve on the autoregressive model's forecasting performance. Unless the researcher had reasons, ex ante, for considering only specific values of ρ, this conclusion could be misleading since the split-mining adjusted test statistic is not significant. In fact, the global minimum p-value of 0.018 is not significant at the 5% level when compared against the split-mining adjusted p-values in Table 3.

Given the significant changes in monetary policy from 1979-1982, a structural break in the data generating process is a natural concern when interpreting these results. To address this issue, we therefore undertake an analysis that discards data prior to 1983. The results from this analysis are shown in the right windows of Figure 5. For this sample the minimum p-value occurs early in the sample and is 0.035. This is insignificant at the 10% critical level when compared against the adjusted p-values in Table 3.

6 Conclusion

Choice of the sample split used to divide data into in-sample estimation and out-of-sample evaluation periods affects out-of-sample forecast evaluation tests in fundamental ways, yet has received little attention in the forecasting literature. As a consequence, this choice variable is often selected without regard to the properties of the predictive accuracy test or the possible size distortions that result when the sample split is chosen to most favor the forecast model under consideration.

When multiple split points are considered and, in particular, when researchers−individually or collectively−may have mined over the sample split point, forecast evaluation tests can be grossly over-sized, leading to spurious evidence of predictability. In fact, the nominal rejection rates can be grossly inflated as a result of such mining over the split point, and the danger of spurious rejection induced by search over the split point tends to be associated with short evaluation windows, corresponding to starting the out-of-sample period late in

the sample. Conversely, power is highest when the forecast evaluation window begins early, corresponding to a long out-of-sample period.

References

Andrews, D. W. K. (1993), 'Test for parameter instability and structural change with unknown change point', *Econometrica* **61**, 821–856.

Campbell, J. & Shiller, R. (1988), 'Stock prices, earnings and expected dividents', *Journal of Finance* **46**, 661–676.

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

Campbell, J. Y. & Yogo, M. (2006), 'Efficient tests of stock return predictability', *Journal of Financial Economics* **81**, 27–60.

Clark, T. E. & McCracken, M. W. (2001), 'Tests of equal forecast accuracy and encompassing for nested models', *Journal of Econometrics* **105**, 85–110.

Clark, T. E. & McCracken, M. W. (2005), 'Evaluating direct multi-step forecasts', *Econometric Reviews* **24**, 369–404.

Clark, T. E. & West, K. D. (2007), 'Approximately normal tests for equal predictive accuracy in nested models', *Journal of Econometrics* **127**, 291–311.

Clark, T. & McCracken, M. (2012), Advances in forecast evaluation, *in* G. Elliott & A. Timmermann, eds, 'Handbook of Economic Forecasting vol. 2', North-Holland, Amsterdam, p. forthcoming.

De Jong, R. M. & Davidson, J. (2000), 'The functional central limit theorem and convergence to stochastic integrals I: Weakly dependent processes', *Econometric Theory* **16**, 621–642.

Diebold, F. X. & Rudebusch, G. (1991), 'Forecasting output with the composite leading index: A real-time analysis', *Journal of American Statistical Association* **86**, 603–610.

Fama, E. F. & French, K. R. (1988), 'Dividend yields and expected stock returns', *Journal of Financial Economics* **22**, 3–25.

Hansen, B. (1992), 'Convergence to stochastic integrals for dependent heterogeneous processes', *Econometric Theory* **8**, 489–500.

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

Inoue, A. & Kilian, L. (2004), 'In-sample or out-of-sample tests of predictability: Which one should we use?', *Econometrics Reviews* **23**, 371–402.

Inoue, A. & Kilian, L. (2008), 'How useful is bagging in forecasting economic time series? a case study of u.s. consumer price inflation', *Journal of the American Statistical Association* **103**, 511–522.

Johannes, M., Korteweg, A. & Polson, N. (2009), 'Sequential learning, predictive regressions, and optimal portfolio returns', *Mimeo, Columbia University* .

Keim, D. & Stambaugh, R. (1986), 'Predicting returns in the stock and bond markets', *Journal of Financial Economics* **17**, 357–390.

Ludvigson, S. & Ng, S. (2009), 'Macro factors in bond risk premia', *Review of Financial Studies* **22**, 5027–5067.

Marcellino, M., Stock, J. H. & Watson, M. W. (2006), 'A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series', *Journal of Econometrics* **135**, 499–526.

McCracken, M. W. (2007), 'Asymptotics for out-of-sample tests of granger causality', *Journal of Econometrics* **140**, 719–752.

Pesaran, M. H., Pick, A. & Timmermann, A. (2011), 'Variable selection, estimation and inference for multi-period forecasting problems', *Journal of Econometrics* **164**, 173–187.

Pesaran, M. H. & Timmermann, A. (1995), 'Predictability of stock returns: Robustness and economic significance', *Journal of Finance* **50**, 1201–1228.

Politis, D. N. & Romano, J. P. (1995), 'Bias-corrected nonparametric spectral estimation', *Journal of time series analysis* **16**, 67–103.

Rapach, D. E., Strauss, J. K. & Zhou, G. (2010), 'Out-of-sample equity premium prediction: Combination forecasts and links to the real economy', *Review of Financial Studies* **23**, 821–862.

Rossi, B. & Inoue, A. (2011), 'Out-of-sample forecast tests robust to the window size choice', *working paper, Duke University* .

Stock, J. & Watson, M. (2007), *Introduction to Econometrics. 2nd Edition*, AddisonWesley.

Welch, I. & Goyal, A. (2008), 'A comprehensive look at the empirical performance of equity premium prediction', *The Review of Financial Studies* pp. 1455–1508.

West, K. (2006), Advances in forecast evaluation, *in* G. Elliott & A. Timmermann, eds, 'Handbook of Economic Forecasting', Vol. 1, North-Holland, Amsterdam, pp. 99–134.

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

Westfall, P. H. & Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments*, Wiley, New York.

White, H. (1994), *Estimation, Inference and Specification Analysis*, Cambridge University Press, Cambridge.

White, H. (2000*a*), *Asymptotic Theory for Econometricians*, revised edn, Academic Press, San Diego.

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

Wooldridge, J. M. & White, H. (1988), 'Some invariance principles and central limit theorems for dependent heterogeneous processes', *Econometric Theory* **4**, 210–230.

Appendix of Proofs

A.1 Derivations related to the simple example in Section 2.1

Suppose that $\beta = c\sigma_\varepsilon/\sqrt{n}$. Then, from (1)-(2), we have

$$\begin{aligned} D_n(\rho) &= \sum_{t=n_\rho+1}^{n} (y_t - \hat{y}^b_{t|t-1})^2 - (y_t - \hat{y}_{t|t-1})^2 \\ &= \sum_{t=n_\rho+1}^{n} (y_t - \beta + \beta)^2 - [y_t - \beta - (\hat{\beta}_{t-1} - \beta)]^2 \\ &= \sum_{t=n_\rho+1}^{n} (\varepsilon_t + \beta)^2 - \left(\varepsilon_t - \tfrac{1}{t-1}\sum_{s=1}^{t-1}\varepsilon_s\right)^2 \\ &= \sum_{t=n_\rho+1}^{n} \beta^2 + 2\beta\varepsilon_t - \left(\tfrac{1}{t-1}\sum_{s=1}^{t-1}\varepsilon_s\right)^2 + 2\left(\tfrac{1}{t-1}\sum_{s=1}^{t-1}\varepsilon_s\right)\varepsilon_t. \end{aligned}$$

Now define

$$W_n(u) = \frac{1}{\sqrt{n}} \sum_{s=1}^{\lfloor un \rfloor} \varepsilon_s, \qquad u \in [0,1].$$

By Donsker's Theorem

$$W_n(u) \Rightarrow \sigma_\varepsilon B(u),$$

where $B(u)$ is a standard Brownian motion. Hence,

$$\begin{aligned}
\sum_{t=n_\rho+1}^{n} \left(\tfrac{1}{t-1} \sum_{s=1}^{t-1} \varepsilon_s \right)^2 &= \frac{1}{n} \sum_{t=n_\rho+1}^{n} \left(\tfrac{n}{t-1} W_n(\tfrac{t-1}{n}) \right)^2 \\
&\xrightarrow{d} \sigma_\varepsilon^2 \int_\rho^1 u^{-2} B(u)^2 \mathrm{d}u. \\
\sum_{t=n_\rho+1}^{n} \left(\tfrac{1}{t-1} \sum_{s=1}^{t-1} \varepsilon_s \right) \varepsilon_t &= \sum_{t=n_\rho+1}^{n} \tfrac{n}{t-1} W_n(\tfrac{t-1}{n}) \left[W_n(\tfrac{t}{n}) - W_n(\tfrac{t-1}{n}) \right] \\
&\xrightarrow{d} \sigma_\varepsilon^2 \int_\rho^1 u^{-1} B(u) \mathrm{d}B(u). \\
\sum_{t=n_\rho+1}^{n} \beta^2 + 2\beta\varepsilon_t &= (n-n_\rho)\frac{\sigma_\varepsilon^2 c^2}{n} + 2\frac{\sigma_\varepsilon c}{\sqrt{n}} \sum_{t=n_\rho+1}^{n} \varepsilon_t \\
&= c^2\sigma_\varepsilon^2 (1 - \frac{n_\rho}{n}) + 2c\sigma_\varepsilon \left[W_n(1) - W_n(\tfrac{n_\rho}{n}) \right] \\
&\xrightarrow{d} \sigma_\varepsilon^2 \left\{ c^2(1-\rho) + 2c \left[B(1) - B(\rho) \right] \right\}.
\end{aligned}$$

A.2 Proof of Theorem 1

By Assumption 1 it follows that $\mathrm{E}(Z_{t-h}\varepsilon_t) = 0$ and that Ω is well defined. Under the mixing assumptions (Assumptions 1 and 2') the result follows from Wooldridge & White (1988, corollary 4.2), see also Hansen (1992).

Under the near-epoch dependence assumptions (Assumptions 1 and 2), we can adapt results in De Jong & Davidson (2000) to our framework. These assumptions are the weakest known; see also White (2000*a*, theorems 7.30 and 7.45) who adapt their results to a setting with global covariance stationary mixing processes.

Define $\mathcal{U}_t = \mathrm{vech}(V_t V_t' - \Sigma_{vv})$ and consider $X_{nt} = \omega' \mathcal{U}_t / \sqrt{n}$ for some arbitrary vector ω, so that $\omega' \Psi \omega = 1$, where $\Psi = \mathrm{var}[n^{-1/2} \sum_{t=1}^{n} \mathrm{vech}(V_t V_t' - \Sigma_{vv})]$, which is well defined under Assumption 1. We verify the conditions in De Jong & Davidson (2000, Assumption 1) for X_{nt}. Their assumption has four parts, (a)-(d). Since X_t is L_4-NED of size $-\frac{1}{2}$ on $\mathcal{V}_t$, it follows that X_{nt} is L_2-NED of the same size on $\mathcal{V}_t$ where we can set $d_{nt} = d_t/\sqrt{n}$. This proves the first part of (c) and part (a) follows directly from $\mathrm{E}(\mathcal{U}_t) = 0$ and $\omega' \Psi \omega = 1$. Part (b) follows with $c_{nt} = n^{-1/2}$ and the last part of (c) follows because $d_{nt}/c_{nt} = d_t$ is assumed to be uniformly bounded. The last condition, part (d), is trivial when $c_{nt} = n^{-1/2}$.

As a corollary to De Jong & Davidson (2000, Theorem 4.1) we have that $\mathcal{W}_n(u) = n^{-1/2} \sum_{t=1}^{\lfloor un \rfloor} \mathcal{U}_t \Rightarrow \mathcal{W}(u)$, where $\mathcal{W}(u)$ is a Brownian motion with covariance matrix Ψ.

From this it also follows that

$$\sup_{u\in(0,1]}\left|\frac{1}{n}\sum_{t=1}^{\lfloor un\rfloor}V_tV_t'-u\Sigma_{vv}\right|=o_p(1), \tag{A.1}$$

which we will use in the proofs below. Moreover, De Jong & Davidson (2000, Theorem 4.1) establishes the joint convergence

$$\left(\mathcal{W}_n(u),\sum_{t=1}^{n}\mathcal{W}_n(\tfrac{t-1}{n})[\mathcal{W}_n(\tfrac{t}{n})-\mathcal{W}_n(\tfrac{t-1}{n})]-A_n\right)\Rightarrow\left(\mathcal{W}(u),\int_0^1\mathcal{W}(u)\mathrm{d}\mathcal{W}(u)'\right),$$

where $A_n=\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t-1}\mathrm{E}\mathcal{U}_s\mathcal{U}_t'$.

Define the matrices

$$L=(0_{q\times1},-\Sigma_{21}\Sigma_{11}^{-1},I_{q\times q})\text{ and }R=(1,-\Sigma_{xx}^{-1}\Sigma_{xy}).$$

Then it is easy to verify that $L\Sigma_{vv}R'=0$ and

$$Z_{t-h}\varepsilon_t=LV_tV_t'R'=L(V_tV_t'-\Sigma_{vv})R',$$

so that the convergence results involving $\{Z_{t-h}\varepsilon_t\}$ follow from those for $V_tV_t'-\Sigma_{vv}$. Thus we only need to express the asymptotic bias term and the variance of the Brownian motion.

Let $U_{nt}=Z_{t-h}\varepsilon_t/\sqrt{n}$, $W_n(u)=\sum_{t=1}^{\lfloor un\rfloor}U_{nt}$, and write $\int_0^s W\mathrm{d}W'$ as short for $\int_0^s W(u)\mathrm{d}W(u)'$. Theorem 1 now follows as a special case of the following theorem:

Theorem A.1 *Given Assumptions 1-2 we have $W_n\Rightarrow W$, and if in addition Assumption 3 holds, we have*

$$\left(W_n,\sum_{t=1}^{n}\sum_{s=1}^{t-h}U_{ns}U_{nt}'\right)\Rightarrow\left(W,\int_0^1 W\mathrm{d}W'\right).$$

Proof. From De Jong & Davidson (2000, Theorem 4.1) it follows that

$$\left(W_n,\sum_{t=1}^{n}\sum_{s=1}^{t-1}U_{ns}U_{nt}'-A_n\right)\Rightarrow\left(W,\int_0^1 W\mathrm{d}W'\right),$$

where $A_n=\sum_{t=1}^{n}\sum_{s=1}^{t-1}\mathrm{E}U_{ns}U_{nt}'$. Moreover, $\sum_{t=1}^{n}\sum_{s=1}^{t-1}U_{ns}U_{nt}'-\sum_{t=1}^{n}\sum_{s=1}^{t-h}U_{ns}U_{nt}'=\sum_{t=1}^{n}\sum_{j=1}^{h-1}U_{n,t-j}U_{nt}'$, where

$$\sum_{t=1}^{n}\sum_{j=1}^{h-1}(U_{n,t-j}U_{nt}'-\mathrm{E}U_{n,t-j}U_{nt}')=o_p(1).$$

By Assumption 3 it follows that $\mathrm{E}U_{ns}U_{nt}'=0$ for $|s-t|\geq h$, so that $A_n=\sum_{t=1}^{n}\sum_{j=1}^{h-1}\mathrm{E}U_{n,t-j}U_{nt}'$, and the result follows. ■

For h-step-ahead forecasts, we expect non-zero autocorrelations up to order $h-1$. These autocorrelations do not, however, affect the asymptotic distribution due to the construction of the empirical stochastic integral, $\sum_{t=1}^{n}\sum_{s=1}^{t-h}U_{ns}U_{nt}'=\int W_n(\frac{t-h}{n})\mathrm{d}W_n(\frac{t}{n})'$, where the first term is evaluated at $\frac{t-h}{n}$ rather than $\frac{t-1}{n}$.

A.3 Proof of Theorem 2

The proof of Theorem 2 follows from the proof of Theorem 4 by imposing the null hypothesis, i.e., by setting $c = 0$.

A.4 Proof of Theorem 3

Theorem 3 can be proved using the following simple result:

Lemma A.1 *If* $b_t = b_{t-1} + \varepsilon_t$, *then* $2b_{t-1}\varepsilon_t = b_t^2 - b_{t-1}^2 - \varepsilon_t^2$.

Proof.

$$\begin{aligned} b_{t-1}\varepsilon_t &= (b_t - \varepsilon_t)\varepsilon_t = b_t(b_t - b_{t-1}) - \varepsilon_t^2 = b_t^2 - b_t b_{t-1} - \varepsilon_t^2 \\ &= b_t^2 - (b_{t-1} + \varepsilon_t)b_{t-1} - \varepsilon_t^2 = b_t^2 - b_{t-1}^2 - b_{t-1}\varepsilon_t - \varepsilon_t^2. \end{aligned}$$

Rearranging the terms, we get the result. ■

Proof. Define $b_{n,t} = B(\frac{t}{n})$ and $\varepsilon_{n,t} = b_{n,t} - b_{n,t-1}$. Our stochastic integrals are given as the probability limits of

$$2\sum_{t=\rho n}^{n} \tfrac{n}{t} b_{n,t-1}\varepsilon_{n,t} - \frac{1}{n}\sum_{t=\rho n}^{n} \left(\tfrac{n}{t}\right)^2 b_{n,t}^2.$$

Throughout we assume that ρn is an integer to simplify notation. From Lemma A.1 we have

$$2\sum_{t=\rho n}^{n} \tfrac{n}{t} b_{n,t-1}\varepsilon_{n,t} = \sum_{t=\rho n}^{n} \tfrac{n}{t}(b_{n,t}^2 - b_{n,t-1}^2) - \sum_{t=\rho n}^{n} \tfrac{n}{t}\varepsilon_{n,t}^2,$$

and one can verify that

$$\sum_{t=\rho n}^{n} \tfrac{n}{t}\varepsilon_{n,t}^2 \xrightarrow{p} -\log\rho,$$

using that $\mathrm{E}\left(\sum_{t=\rho n}^{n} \frac{n}{t}\varepsilon_{n,t}^2\right) = \sum_{t=\rho n}^{n} \frac{n}{t}\mathrm{E}\left(\varepsilon_{n,t}^2\right) = \sum_{t=\rho n}^{n} \frac{n}{t}\frac{1}{n}$ and

$$\frac{1}{n}\sum_{t=\rho n}^{n} \tfrac{n}{t} \xrightarrow{d} \int_{\rho}^{1} \frac{1}{u}\mathrm{d}u = \log 1 - \log\rho.$$

Next, consider

$$\begin{aligned} \sum_{t=n_\rho+1}^{n} \tfrac{n}{t}(b_{n,t}^2 - b_{n,t-1}^2) &= b_{n,n}^2 + n\sum_{t=n_\rho+1}^{n-1} \left(\tfrac{1}{t} - \tfrac{1}{t+1}\right) b_{n,t}^2 - \tfrac{n}{n_\rho} b_{n,\rho n}^2 \\ &= b_{n,n}^2 + \frac{1}{n}\sum_{t=n_\rho+1}^{n-1} \tfrac{n^2}{t^2+t} b_{n,t}^2 - (\rho + O(n^{-1}))^{-1} b_{n,\rho n}^2, \end{aligned}$$

where the first and last terms equal $B(1)^2$ and $-\rho^{-1}B^2(\rho)$, respectively. Since

$$\frac{1}{n}\sum_{t=n_\rho+1}^{n-1} \tfrac{n^2}{t^2+t} b_{n,t}^2 - \frac{1}{n}\sum_{t=n_\rho+1}^{n} \left(\tfrac{n}{t}\right)^2 b_{n,t}^2 = o_p(1),$$

the result follows. ■

A.5 Proof of Corollary 1

Proof. Let $U = \frac{B(1)-B(\rho)}{\sqrt{1-\rho}}$ and $V = \frac{B(\rho)}{\sqrt{\rho}}$ so that $B(1) = \sqrt{1-\rho}U + \sqrt{\rho}V$, and note that U and V are independent standard Gaussian random variables.

The distribution we seek is that of $W = \left(\sqrt{1-\rho}U + \sqrt{\rho}V\right)^2 - V^2 + \log\rho$, where $U, V \sim$ iid$N(0,1)$, which can be expressed in the quadratic from:

$$W = \begin{pmatrix} U \\ V \end{pmatrix}' \begin{pmatrix} 1-\rho & \sqrt{\rho(1-\rho)} \\ \sqrt{\rho(1-\rho)} & \rho - 1 \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix} + \log\rho.$$

Since a real symmetric matrix, A, can be decomposed into $A = Q'\Lambda Q$ where $Q'Q = I$ and Λ is a diagonal matrix with the eigenvalues of A in the diagonal, we find that

$$W = Z' \begin{pmatrix} \sqrt{1-\rho} & 0 \\ 0 & -\sqrt{1-\rho} \end{pmatrix} Z + \log\rho,$$

where $Z \sim N_2(0, I)$. Here Z is a simply rotation of $(U, V)'$, given by $Z = Q(U, V)'$, where

$$Q = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{1+\sqrt{1-\rho}} & \sqrt{1-\sqrt{1-\rho}} \\ -\sqrt{1-\sqrt{1-\rho}} & \sqrt{1+\sqrt{1-\rho}} \end{pmatrix}.$$

It follows that $W = \sqrt{1-\rho}(Z_1^2 - Z_2^2) + \log\rho$, which proves the result. ■

A.6 Proof of Corollary 2

Proof. Let Z_{1i}, Z_{2i}, $i = 1, \ldots, q$ be i.i.d. $N(0,1)$, so that $X = \sum_{i=1}^q Z_{1,i}^2$ and $Y = \sum_{i=1}^q Z_{2,i}^2$ are both χ_q^2-distributed and independent. The distribution we seek is given by the convolution,

$$\sum_{i=1}^{q} \left[\sqrt{1-\rho}(Z_{1,i}^2 - Z_{2,i}^2) + \log\rho\right] = \sqrt{1-\rho}(X - Y) + q\log\rho,$$

so we seek the distribution of $S = X - Y$ where X and Y are independent χ_q^2-distributed random variables. The density of a χ_q^2 is

$$\psi(u) = 1_{\{u\geq 0\}} \frac{1}{2^{q/2}\Gamma(\frac{q}{2})} u^{q/2-1} e^{-u/2},$$

and we seek the convolution of X and $-Y$

$$\begin{aligned} \int 1_{\{u\geq 0\}}\psi(u) 1_{\{u-s\geq 0\}}\psi(u-s)\mathrm{d}u &= \int_{0\vee s}^{\infty} \psi(u)\psi(u-s)\mathrm{d}u, \\ &= \int_{0\vee s}^{\infty} \frac{1}{2^{q/2}\Gamma(\frac{q}{2})} u^{q/2-1} e^{-u/2} \frac{1}{2^{q/2}\Gamma(\frac{q}{2})} (u-s)^{q/2-1} e^{-(u-s)/2}\mathrm{d}u \\ &= \frac{1}{2^q\Gamma(\frac{q}{2})\Gamma(\frac{q}{2})} e^{s/2} \int_{0\vee s}^{\infty} (u(u-s))^{q/2-1} e^{-u}\mathrm{d}u. \end{aligned}$$

For $s < 0$ the density is $2^{-q}\Gamma(\frac{q}{2})^{-2}e^{s/2}\int_0^\infty (u(u-s))^{q/2-1}\,e^{-u}\mathrm{d}u$, and by taking advantage of the symmetry about zero, we obtain the expression

$$\frac{1}{2^q\Gamma(\frac{q}{2})\Gamma(\frac{q}{2})}e^{-|s|/2}\int_0^\infty (u(u+|s|))^{q/2-1}\,e^{-u}\mathrm{d}u.$$

When $q = 1$ this simplifies to $f_1(s) = \frac{1}{2\pi}\mathrm{B}_0(\frac{|s|}{2})$ where $\mathrm{B}_k(x)$ denotes the modified Bessel function of the second kind. For $q = 2$ we have the simpler expression $f_2(x) = \frac{1}{4}e^{-\frac{|s|}{2}}$ which is the Laplace distribution with scale parameter 2. ■

A.7 Proof of Theorem 4

To prove Theorem 4, we first establish two lemmas.

Lemma A.2 *The loss differential* $(y_t - \hat{y}^b_{t|t-h})^2 - (y_t - \hat{y}_{t|t-h})^2$ *equals*

$$\begin{aligned}
&\beta_2'Z_{t-h}Z_{t-h}'\beta_2 + 2\beta_2'Z_{t-h}\varepsilon_t - 2\beta_2'Z_{t-h}X_{1,t-h}'(\tilde{\beta}_{1,t-h} - \delta)\\
&+2(\hat{\beta}_{2,t-h} - \beta_2)'Z_{t-h}\varepsilon_t - (\hat{\beta}_{2,t-h} - \beta_2)'Z_{t-h}Z_{t-h}'(\hat{\beta}_{2,t-h} - \beta_2)\\
&-2(\hat{\beta}_{2,t-h} - \beta_2)'Z_{t-h}X_{1,t-h}'(\tilde{\beta}_{1,t-h} - \delta)\\
&-\zeta_{t-h}^2 + 2\zeta_{t-h}\left[\varepsilon_t - X_{1,t-h}'(\tilde{\beta}_{1,t-h} - \delta) - Z_{t-h}'(\hat{\beta}_{2,t-h} - \beta_2)\right],
\end{aligned}$$

where $\zeta_t = \hat{\beta}_{2,t}'(\Sigma_{21}\Sigma_{11}^{-1} - M_{21,t}M_{11,t}^{-1})X_{1,t}$ *with* $M_{ij,t} = \sum_{s=1}^t X_{i,s}X_{j,s}'$ *for* $i, j = 1, 2$.

Proof. For the benchmark forecast in (4) we have

$$\tilde{\beta}_{1,t}'X_{1,t} = \delta X_{1,t} + \beta_2'Z_t + (\tilde{\beta}_{1,t} - \delta)'X_{1,t} - \beta_2'Z_t,$$

where the true model assumes that $y_{t+h} = \delta'X_{1,t} + \beta_2'Z_t + \varepsilon_{t+h}$. Hence the forecast error from the benchmark model takes the form

$$y_{t+h} - \tilde{\beta}_{1,t}'X_{1,t} = \varepsilon_{t+h} - (\tilde{\beta}_{1,t} - \delta)'X_{1,t} + \beta_2'Z_t.$$

Similarly, for the alternative forecast in (5) we have

$$\begin{aligned}
\hat{\beta}_t'X_t &= \hat{\beta}_{1,t}'X_{1,t} + \hat{\beta}_{2,t}'X_{2,t}\\
&= (\hat{\beta}_{1,t}' + \hat{\beta}_{2,t}'M_{21,t}M_{11,t}^{-1})X_{1,t} + \hat{\beta}_{2,t}'(X_{2,t} - M_{21,t}M_{11,t}^{-1}X_{1,t})\\
&= \tilde{\beta}_{1,t}'X_{1,t} + \hat{\beta}_{2,t}'(X_{2,t} - M_{21,t}M_{11,t}^{-1}X_{1,t})\\
&= \tilde{\beta}_{1,t}'X_{1,t} + \hat{\beta}_{2,t}'(X_{2,t} - \Sigma_{21}\Sigma_{11}^{-1}X_{1,t}) + \hat{\beta}_{2,t}'(\Sigma_{21}\Sigma_{11}^{-1} - M_{21,t}M_{11,t}^{-1})X_{1,t}\\
&= \delta'X_{1,t} + \beta_2'Z_t + (\tilde{\beta}_{1,t} - \delta)'X_{1,t} + (\hat{\beta}_{2,t} - \beta_2)'Z_t + \zeta_t
\end{aligned}$$

so that

$$y_{t+h} - \hat{\beta}_t'X_t = \varepsilon_{t+h} - (\tilde{\beta}_{1,t} - \delta)'X_{1,t} - (\hat{\beta}_{2,t} - \beta_2)'Z_t + \zeta_t.$$

Next, consider the loss differential, which from equations (4) to (5) is given by

$$\begin{aligned}
& (y_t - \hat{y}^b_{t|t-h})^2 - (y_t - \hat{y}_{t|t-h})^2 \\
= \; & (y_t - \tilde{\beta}'_{1,t-h} X_{1,t-h})^2 - (y_t - \hat{\beta}'_{t-h} X_{t-h})^2 \\
= \; & (\varepsilon_t - (\tilde{\beta}_{1,t-h} - \delta)' X_{1,t-h} + \beta_2' Z_{t-h})^2 \\
& - \Big(\varepsilon_t - (\tilde{\beta}_{1,t-h} - \delta)' X_{1,t-h} - (\hat{\beta}_{2,t-h} - \beta_2)' Z_{t-h} + \zeta_{t-h}\Big)^2 .
\end{aligned}$$

The result now follows by multiplying out. ■

Lemma A.3 *With $\beta_2 = \frac{c}{\sqrt{n}} v$ for some $v \in \mathbb{R}^q$ and given Assumptions 1-3 we have,*

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} \beta_2' Z_{t-h} Z'_{t-h} \beta_2 \xrightarrow{p} (1-\rho) c^2 v' \Sigma_{zz} v \tag{A.2}$$

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} \beta_2' Z_{t-h} \varepsilon_t \xrightarrow{d} c v' \left[W(1) - W(\rho) \right] \tag{A.3}$$

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} (\hat{\beta}_{2,t-h} - \beta_2)' Z_{t-h} \varepsilon_t \xrightarrow{d} \int_\rho^1 \frac{1}{u} W(u)' \Sigma_{zz}^{-1} \mathrm{d}W(u), \tag{A.4}$$

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} (\hat{\beta}_{2,t-h} - \beta_2)' Z_{t-h} Z'_{t-h} (\hat{\beta}_{2,t-h} - \beta_2) \xrightarrow{d} \int_\rho^1 \frac{1}{u^2} W(u)' \Sigma_{zz}^{-1} W(u) \mathrm{d}u \tag{A.5}$$

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} \beta_2' Z_{t-h} X'_{1,t-h} (\tilde{\beta}_{1,t-h} - \delta) \xrightarrow{p} 0 \tag{A.6}$$

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} (\hat{\beta}_{2,t-h} - \beta_2)' Z_{t-h} X'_{1,t-h} (\tilde{\beta}_{1,t-h} - \delta) \xrightarrow{p} 0 \tag{A.7}$$

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} \zeta_{t-h}^2 \xrightarrow{p} 0 \tag{A.8}$$

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} \zeta_{t-h} \varepsilon_t \xrightarrow{p} 0 \tag{A.9}$$

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} \zeta_{t-h} X'_{1,t-h} (\tilde{\beta}_{1,t-h} - \delta) \xrightarrow{p} 0 \tag{A.10}$$

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} \zeta_{t-h} Z'_{t-h} (\hat{\beta}_{2,t-h} - \beta_2) \xrightarrow{p} 0 \tag{A.11}$$

Proof. To simplify notation, introduce

$$\Sigma_n(\rho) = \frac{1}{n} \sum_{t=1}^{\lfloor \rho n \rfloor} Z_{t-h} Z'_{t-h},$$

so that $Z_{t-h}Z_{t-h}' = n\left[\Sigma_n(\frac{t}{n}) - \Sigma_n(\frac{t-1}{n})\right]$ and

$$\hat{\beta}_{2,t} - \beta_2 = \frac{1}{\sqrt{n}}\Sigma_n^{-1}(\tfrac{t}{n})W_n(\tfrac{t}{n}).$$

The result for the first term, (A.2),

$$\sum_{\lfloor \rho n \rfloor + 1}^{n} \beta_2' Z_{t-h}Z_{t-h}'\beta_2 = c^2 v' \left[\Sigma_n(1) - \Sigma_n(\rho)\right] v,$$

follows from (A.1). Similarly, (A.3) follows by,

$$\beta_2' \sum_{\lfloor \rho n \rfloor + 1}^{n} Z_{t-h}\varepsilon_t = cv' \left[W_n(1) - W_n(\rho)\right],$$

and Theorem A.1. Next,

$$\begin{aligned}
\sum_{\lfloor \rho n \rfloor + 1}^{n} (\hat{\beta}_{2,t-h} - \beta_2)' Z_{t-h}\varepsilon_t &= \sum_{t=\lfloor \rho n \rfloor + 1}^{n} W_n(\tfrac{t-h}{n})'\Sigma_n^{-1}(\tfrac{t}{n})\left[W_n(\tfrac{t}{n}) - W_n(\tfrac{t-1}{n})\right] \\
&= \sum_{t=\lfloor \rho n \rfloor + 1}^{n} W_n(\tfrac{t-h}{n})'\frac{1}{u}\Sigma_{zz}^{-1}\left[W_n(\tfrac{t}{n}) - W_n(\tfrac{t-1}{n})\right] + o_p(1),
\end{aligned}$$

where again we used (A.1). From Theorem A.1, $\int_\rho^1 W_n(u)\mathrm{d}W_n(u)' \xrightarrow{d} \int_\rho^1 W(u)\mathrm{d}W(u)'$, so

$$\begin{aligned}
\int_\rho^1 W_n(u)'\Sigma_{zz}^{-1}\mathrm{d}W_n(u) &= \int_\rho^1 \mathrm{tr}\left\{\mathrm{d}W_n(u)'\Sigma_{zz}^{-1}W_n(u)\right\} \\
&= \mathrm{tr}\left\{\Sigma_{zz}^{-1}\int_\rho^1 W_n(u)\mathrm{d}W_n(u)'\right\} \\
&\xrightarrow{d} \mathrm{tr}\left\{\Sigma_{zz}^{-1}\int_\rho^1 W\mathrm{d}W'\right\} = \int_\rho^1 W'\Sigma_{zz}^{-1}\mathrm{d}W.
\end{aligned}$$

Since $\rho > 0$, it follows that $\int_\rho^1 \frac{n}{\lfloor un \rfloor} W_n(u)'\Sigma_{zz}^{-1}\mathrm{d}W_n(u) \xrightarrow{d} \int_\rho^1 \frac{1}{u}W'\Sigma_{zz}^{-1}\mathrm{d}W$, proving (A.4).

The last non-vanishing term in (A.5) is given by:

$$\begin{aligned}
&\frac{1}{n}\sum_{t=\lfloor \rho n \rfloor + 1}^{n} W_n(\tfrac{t-h}{n})'\Sigma_n^{-1}(\tfrac{t}{n})Z_{t-h}Z_{t-h}'\Sigma_n^{-1}(\tfrac{t}{n})W_n(\tfrac{t-h}{n}) \\
= &\frac{1}{n}\sum_{t=\lfloor \rho n \rfloor + 1}^{n} W_n(\tfrac{t-h}{n})'\Sigma_n^{-1}(\tfrac{t}{n})\Sigma_{zz}\Sigma_n^{-1}(\tfrac{t}{n})W_n(\tfrac{t-h}{n}) \\
&+\frac{1}{n}\sum_{t=\lfloor \rho n \rfloor + 1}^{n} W_n(\tfrac{t-h}{n})'\Sigma_n^{-1}(\tfrac{t}{n})\left(Z_{t-h}Z_{t-h}' - \Sigma_{zz}\right)\Sigma_n^{-1}(\tfrac{t}{n})W_n(\tfrac{t-h}{n}).
\end{aligned}$$

The final term in this expression is $O_p(n^{-1/2})$ because with $\mathcal{V}_n(u) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor un \rfloor} \text{vec}(Z_{t-h} Z_{t-h}' - \Sigma_{zz})$, and continuous g we have

$$(W_n, \mathcal{V}_n, \int g(W_n) \mathrm{d}\mathcal{V}_n) \Rightarrow (W, \mathcal{V}, \int g(W) \mathrm{d}\mathcal{V}),$$

so that

$$\sum_{t=\lfloor \rho n \rfloor + 1}^{n} W_n(\tfrac{t-h}{n})' \Sigma_n^{-1}(\tfrac{t}{n}) \tfrac{Z_{t-h} Z_{t-h}' - \Sigma}{\sqrt{n}} \Sigma_n^{-1}(\tfrac{t}{n}) W_n(\tfrac{t-h}{n}) \xrightarrow{d} \int_\rho^1 \tfrac{1}{u^2} \text{vec}(\Sigma_{zz}^{-1})' (\Sigma_{zz}^{-1} \otimes W(u) W(u)') \mathrm{d}\mathcal{V}(u),$$

where we used $\text{tr}\{ABCD\} = \text{vec}(D')'(C' \otimes A)\text{vec}(B)$. The first term in (A.5) is given by

$$\begin{aligned}
& \frac{1}{n} \sum_{t=\lfloor \rho n \rfloor + 1}^{n} W_n(\tfrac{t-h}{n})' \Sigma_n^{-1}(\tfrac{t}{n}) \Sigma_{zz} \Sigma_n^{-1}(\tfrac{t}{n}) W_n(\tfrac{t-h}{n}) \\
= & \int_\rho^1 W_n(u)' \Sigma_n^{-1}(u) \Sigma_{zz} \Sigma_n^{-1}(u) W_n(u) \mathrm{d}u \\
= & \int_\rho^1 u^{-2} W_n(u)' \Sigma_{zz}^{-1} W_n(u) \mathrm{d}u + o_p(1) \\
\xrightarrow{d} & \int_\rho^1 u^{-2} W(u)' \Sigma_{zz}^{-1} W(u) \mathrm{d}u.
\end{aligned}$$

Next consider the terms involving ζ_t and/or $Z_{t-h} X_{1,t-h}'$. First, note that for $\rho > 0$, as $n \to \infty$, $\sup_{\rho n < t \leq n} \left| \tilde{\beta}_{1,t-h} - \delta \right| = o_p(n^{-1/2})$ and $\sup_{\rho n < t \leq n} \left| \hat{\beta}_{2,t-h} - \beta_2 \right| = o_p(n^{-1/2})$ so that

$$\left| \sum_{\lfloor \rho n \rfloor + 1}^{n} c v' \tfrac{Z_{t-h} X_{1,t-h}'}{n} n^{1/2} (\tilde{\beta}_{1,t-h} - \delta) \right| \leq \left| \frac{1}{n} c v' \sum_{\lfloor \rho n \rfloor + 1}^{n} Z_{t-h} X_{1,t-h}' \right| n^{1/2} \sup_{\rho n < t \leq n} \left| \tilde{\beta}_{1,t-h} - \delta \right| = o_p(1).$$

Similarly, $\sum_{\lfloor \rho n \rfloor + 1}^{n} n^{1/2} (\hat{\beta}_{2,t-h} - \beta_2)' \tfrac{Z_{t-h} X_{1,t-h}'}{n} n^{1/2} (\tilde{\beta}_{1,t-h} - \delta) = o_p(1)$ from which (A.6) and (A.7) follow. Next recall that $\zeta_t = \hat{\beta}_{2,t}' (\Sigma_{21} \Sigma_{11}^{-1} - M_{21,t} M_{11,t}^{-1}) X_{1,t-h}$ and for any fixed $\gamma > 0$, we have by (A.1) that $\sup_{t \geq \gamma n} \left| M_{21,t} M_{11,t}^{-1} - \Sigma_{21} \Sigma_{11}^{-1} \right| = o_p(1)$. For $\beta_2 = O(n^{-1/2})$, we have $\sup_{\rho n < t \leq n} \left| \hat{\beta}_{2,t-h} - \beta_2 \right| = O_p(n^{-1/2})$ so that

$$\begin{aligned}
\left| \sum_t \zeta_{t-h}^2 \right| &\leq n^{1/2} \sup_t |\hat{\beta}_{2,t}'| \sup_t \left| \Sigma_{21} \Sigma_{11}^{-1} - M_{21,t} M_{11,t}^{-1} \right| \left| \frac{1}{n} \sum_t X_{1,t} X_{1,t}' \right| \\
&\quad \times \sup_t \left| \Sigma_{21} \Sigma_{11}^{-1} - M_{21,t} M_{11,t}^{-1} \right|' n^{1/2} \sup_t |\hat{\beta}_{2,t}| = o_p(1), \\
\left| \sum_t \zeta_{t-h} \varepsilon_t \right| &\leq n^{1/2} \sup_t |\hat{\beta}_{2,t}|' \sup_t \left| \Sigma_{21} \Sigma_{11}^{-1} - M_{21,t} M_{11,t}^{-1} \right| \left| n^{-1/2} \sum_t X_{1,t-h} \varepsilon_t \right| = o_p(1).
\end{aligned}$$

This proves (A.8) and (A.9). Finally, the absolute value of the last two terms, (A.10) and (A.11), are bounded by

$$\begin{aligned} n^{1/2}\sup_t|\hat{\beta}_{2,t}|'\sup_t\left|\Sigma_{21}\Sigma_{11}^{-1}-M_{21,t}M_{11,t}^{-1}\right|\left|\sum_t\tfrac{X_{1,t}X_{1,t}'}{n}\right|n^{1/2}\sup_t\left|\tilde{\beta}_{1,t}-\delta\right| &= o_p(1), \\ n^{1/2}\sup_t|\hat{\beta}_{2,t}|'\sup_t\left|\Sigma_{21}\Sigma_{11}^{-1}-M_{21,t}M_{11,t}^{-1}\right|\left|\sum_t\tfrac{X_{1,t}Z_t'}{n}\right|n^{1/2}\sup_t\left|\hat{\beta}_{2,t}-\beta_2\right| &= o_p(1), \end{aligned}$$

which completes the proof. ■

From the decomposition in Lemma A.2 and the limit results in Lemma A.3 we are now ready to derive the asymptotic properties of $D_n(\rho)$ and $T_n(\rho)$. From Lemmas A.2 and A.3 it follows that

$$\begin{aligned} T_n(\rho)=\frac{D_n(\rho)}{\hat{\sigma}_\varepsilon^2} \xrightarrow{d} & c^2(1-\rho)\tfrac{v'\Sigma_{zz}v}{\sigma_\varepsilon^2}+\tfrac{2c}{\sigma_\varepsilon^2}v'\Omega^{1/2}\left[B(1)-B(\rho)\right] \\ & +2\int_\rho^1 u^{-1}B(u)'\Omega^{1/2}\Sigma^{-1}\Omega^{1/2}\mathrm{d}B(u) \\ & -\int_\rho^1 u^{-2}B(u)'\Omega^{1/2}\Sigma^{-1}\Omega^{1/2}B(u)\mathrm{d}u, \end{aligned}$$

where we have used the fact that $\Sigma=\sigma_\varepsilon^2\Sigma_{zz}$ so that $\Sigma_{zz}^{-1}/\sigma_\varepsilon^2=\Sigma^{-1}$. Now decompose $\Omega^{1/2}\Sigma^{-1}\Omega^{1/2}=Q'\Lambda Q$, where $\Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_q)$ is a diagonal matrix with eigenvalues of $\Omega^{1/2}\Sigma^{-1}\Omega^{1/2}$ that coincide with the eigenvalues of $\Omega\Sigma^{-1}$, and $Q'Q=I$. It follows that $\tilde{B}(u)=QB(u)$ is a standard (q-dimensional) Brownian motion when $B(u)$ is. Hence,

$$\begin{aligned} T_n(\rho)=\frac{D_n(\rho)}{\hat{\sigma}_\varepsilon^2} \xrightarrow{d} & c^2(1-\rho)\tfrac{v'\Sigma_{zz}v}{\sigma_\varepsilon^2}+\tfrac{2c}{\sigma_\varepsilon^2}v'\Omega^{1/2}Q'\left[\tilde{B}(1)-\tilde{B}(\rho)\right] \\ & +2\int_\rho^1 u^{-1}\tilde{B}(u)'\Lambda\mathrm{d}\tilde{B}(u)-\int_\rho^1 u^{-2}\tilde{B}(u)'\Lambda\tilde{B}(u)\mathrm{d}u, \end{aligned}$$

from which Theorem 4 follows. □

A.8 Proof of Theorem 5

Proof. It follows from the definition of $G(u)$ that the path of critical values, $c_\alpha(u)$ is continuous in u because $F_{u,\Lambda}(x)$ is continuous in (u,x) on $[\underline{\rho},\bar{\rho}]\times\mathbb{R}$. So $c_\alpha(u)\in\mathbb{D}_{[\underline{\rho},\bar{\rho}]}$. Hence, by the continuous mapping theorem and (8) (which is implied by the mixing assumptions, and assumed under the near-epoch assumptions), the smallest p-value over the range of split points, $[\underline{\rho},\bar{\rho}]$, converges in distribution and the CDF of the limit distribution is given by

$$\begin{aligned} \Pr\{p_{[\underline{\rho},\bar{\rho}]}\leq\alpha\} &= \Pr\{G(u)\geq c_\alpha(u)\text{ for some }u\in[\underline{\rho},\bar{\rho}]\} \\ &= \Pr\{\sup_{\underline{\rho}\leq u\leq\bar{\rho}}\left[G(u)-c_\alpha(u)\right]\geq 0\}. \end{aligned}$$

■

McCracken Critical values versus exact critical values

π	0.1	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
ρ	0.909	0.833	0.714	0.625	0.556	0.500	0.455	0.417	0.385	0.357	0.333
$\alpha = 0.99$	1.996	2.691	3.426	3.907	4.129	4.200	4.362	4.304	4.309	4.278	4.250
	2.168	2.830	3.509	3.851	4.040	4.146	4.202	4.225	4.227	4.214	4.191
$\alpha = 0.95$	1.184	1.453	1.733	1.891	1.820	1.802	1.819	1.752	1.734	1.692	1.706
	1.198	1.515	1.789	1.880	1.895	1.870	1.824	1.766	1.702	1.633	1.563
$\alpha = 0.90$	0.794	0.912	1.029	1.077	1.008	0.880	0.785	0.697	0.666	0.587	0.506
	0.780	0.949	1.048	1.031	0.970	0.890	0.800	0.708	0.614	0.522	0.431

Table 1: This table compares the critical values in McCracken (2007), which uses Monte Carlo simulation to evaluate stochastic integrals, to the exact critical values obtained from the CDF of the non-central Laplace distribution. For each critical value (α) the first row shows the McCracken critical values, while the second line shows the exact critical values. All calculations assume q = 2 additional predictor variables.

Type I error rate induced by split point mining

	Nominal level			
q	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	0.4475	0.2582	0.1482	0.0373
2	0.5252	0.3118	0.1723	0.0448
3	0.5701	0.3382	0.1979	0.0546
4	0.6032	0.3611	0.211	0.0528
5	0.6157	0.3795	0.2195	0.0549

Table 2: This table shows the actual rejection rate for different nominal critical levels, (α) and different values of the dimension (q) by which the alternative model exceeds the benchmark. Simulations are conducted under the null model with $\underline{\rho} = 1 - \bar{\rho} = 0.1$. and use a discretization with $n = 10,000$ and $N = 10,000$ simulations.

Split-adjusted Critical values for the minimum p-value

	critical values:			
q	$\alpha = 20\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 1\%$
1	0.073	0.029	0.013	0.001
2	0.059	0.024	0.011	0.001
3	0.05	0.021	0.001	0.001
4	0.046	0.02	0.001	0.001
5	0.044	0.02	0.001	0.001

Table 3: This table shows the split-mining adjusted critical values at which the minimum p-value, $p_{[\underline{\rho},\bar{\rho}]}$, is significant when $\underline{\rho} = 1 - \bar{\rho} = 0.1$. The critical values for the minimum p-value are given for $q = 1, \ldots, 5$ and four significance levels, $\alpha = 0.20$, 0.10, 0.05, and 0.01 and use a discretization with $n = 10,000$ and $N = 10,000$ simulated series.

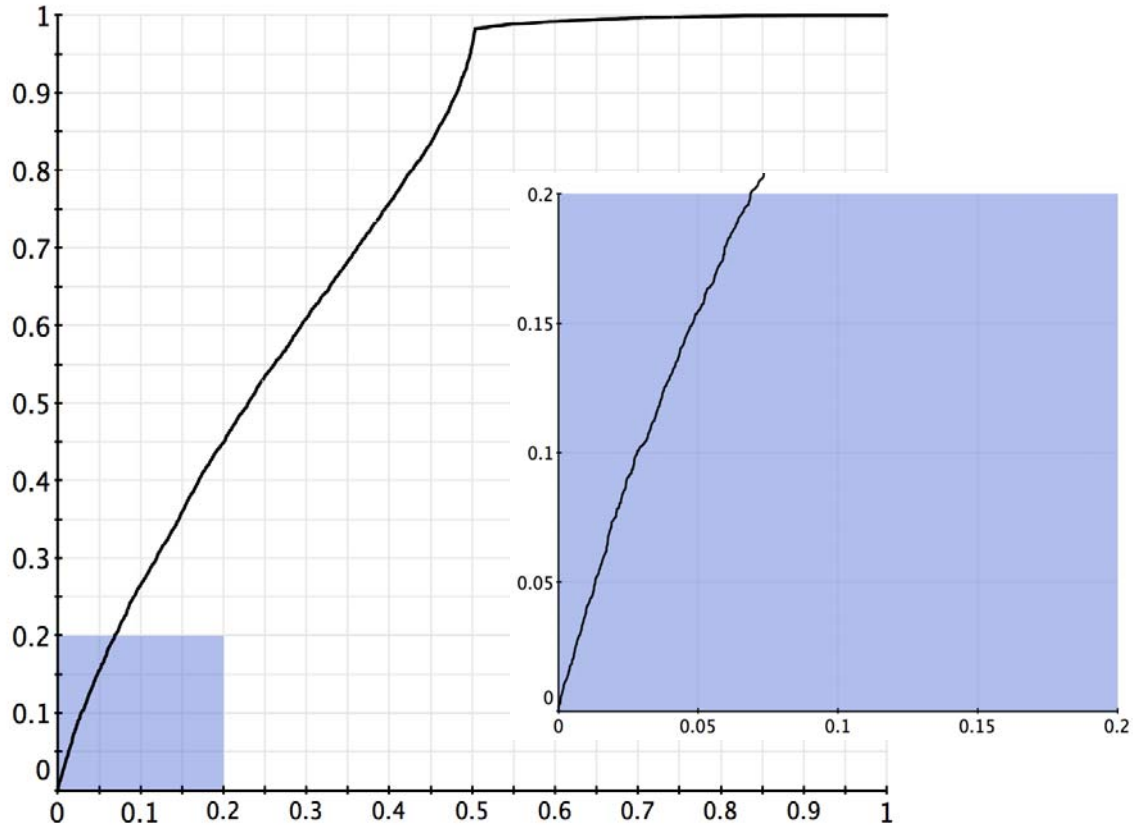

Figure 1: Plot of the CDF for the minimum p-value ($p_{\min}$) as a function of the nominal critical level (α) with one predictor added to the benchmark model (univariate and homoskedastic case).

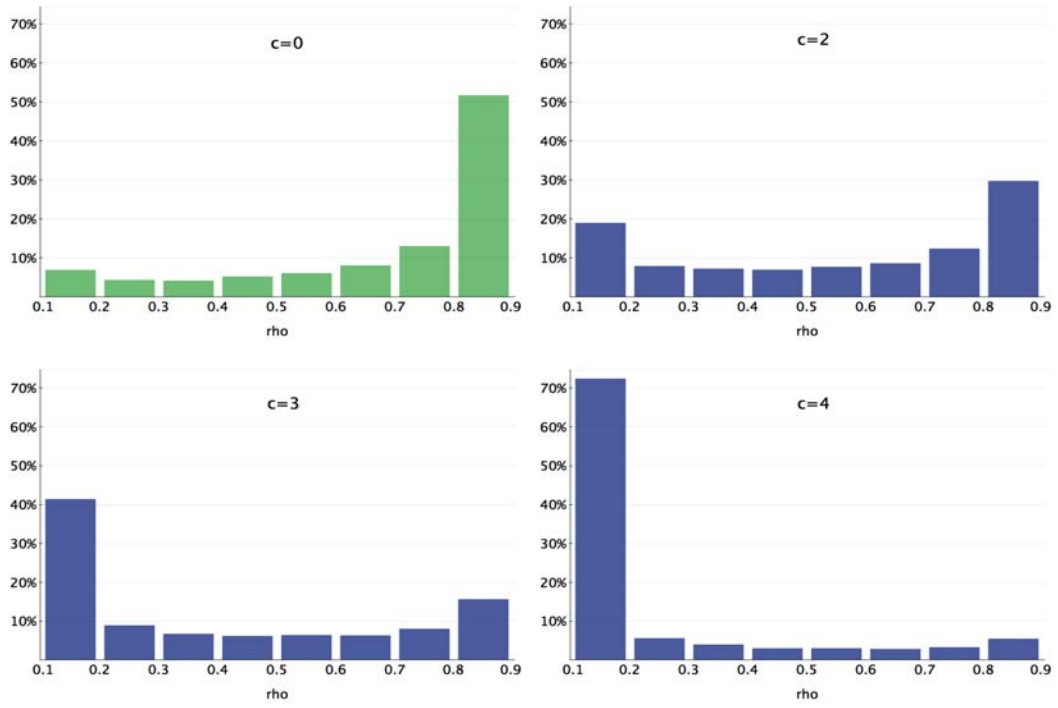

Figure 2: Histograms of the location of the smallest p-value ($p_{\min}$) under the null hypothesis ($c = 0$) and three local alternatives. Under the null hypothesis, the smallest p-value, $\min_{\underline{\rho} \leq r \leq \bar{\rho}} p_r$, is most likely to be located towards the end of the sample, while under the alternative ($c > 0$) the smallest p-value is more likely to be located early in the sample if c is large or late in the sample if c is small.

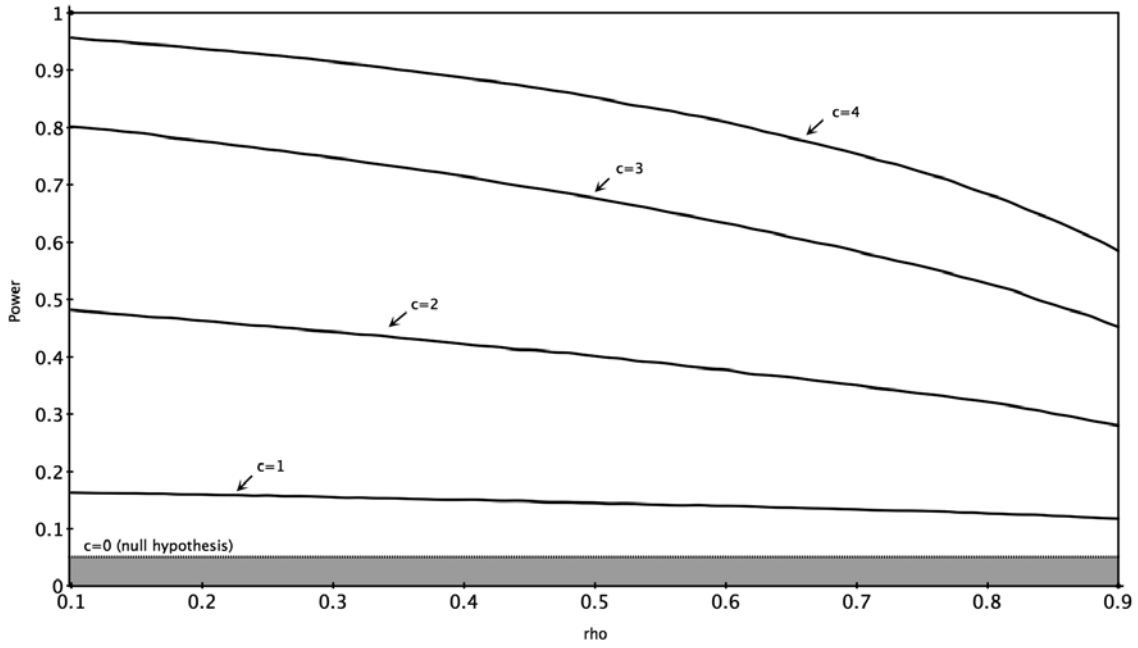

Figure 3: Power of the test under four local alternatives, $c = 1$, $c = 2$, $c = 3$, and $c = 4$ as a function of the sample split point, ρ, assuming that $q = 1$, $\Lambda = 1$, and $h = 1$.

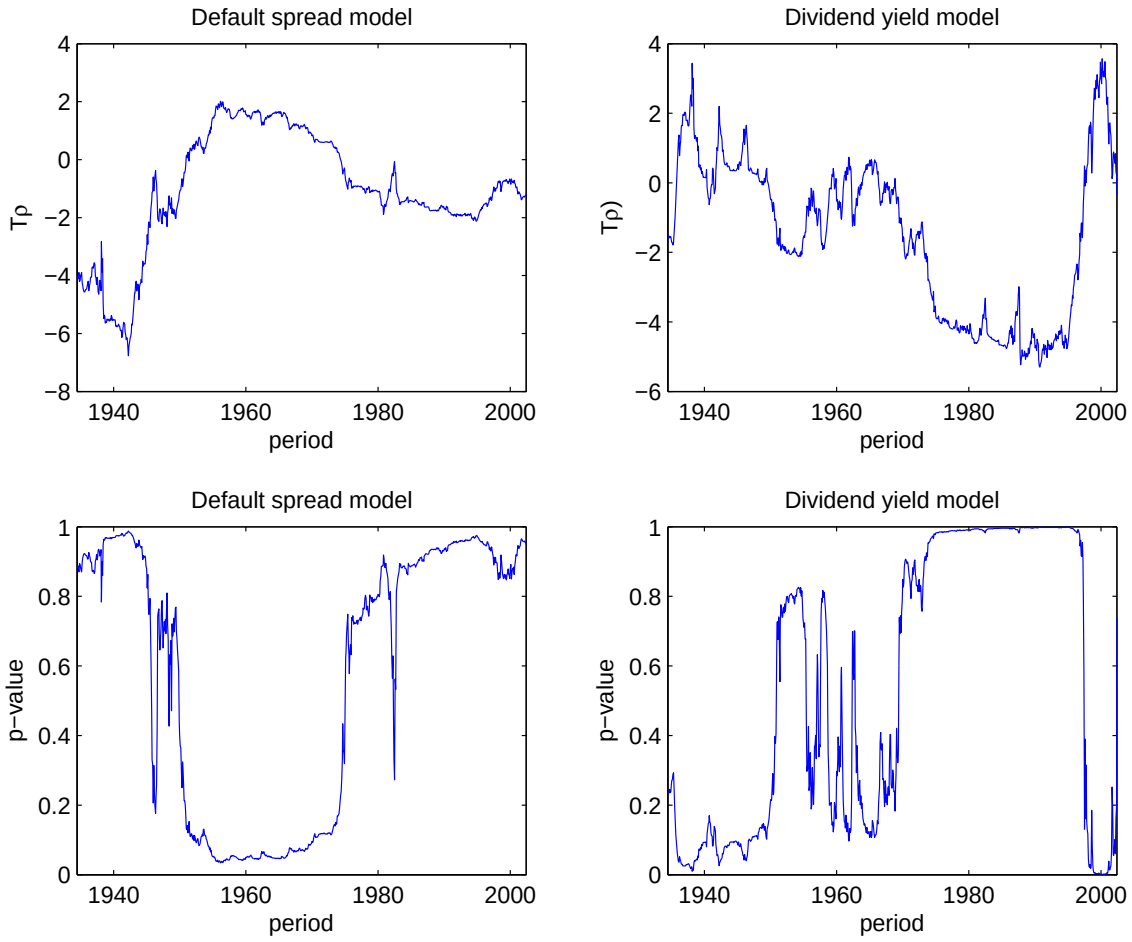

Figure 4: Values of the $T_n(\rho)$ statistic and $p(\rho)$-values for different choices of the sample split point, ρ. Values are based on the U.S. stock return prediction model that uses the default spread (left windows) or the dividend yield (right windows) as a predictor variable.

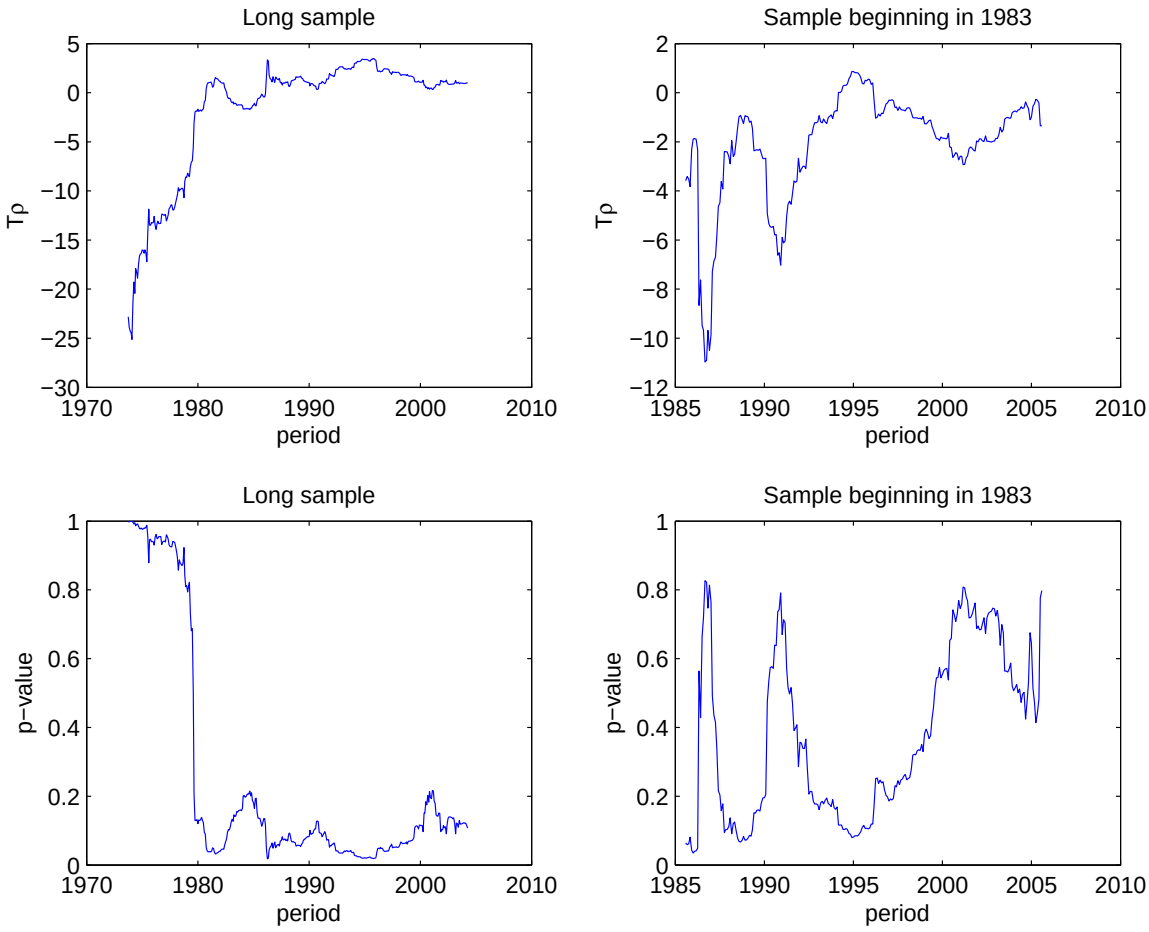

Figure 5: Values of the $T_n(\rho)$ statistic and $p(\rho)$-values for different choices of the sample split point, ρ. The plots are based on the U.S. inflation prediction model that uses four common factors as additional predictor variables on top of two autoregressive lags.