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

Peter Reinhard Hansen
Stanford University and CREATES

Allan Timmermann
UCSD and CREATES

February 23, 2011

**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. We show that very large size distortions can occur, more than tripling the rejection rates of conventional tests of predictive accuracy, when the sample split is viewed as a choice variable, rather than being fixed ex ante. To deal with this issue, we propose a test statistic that is robust to the effect of mining over the start of the out-of-sample period. Moreover, we show that the power of out-of-sample forecast evaluation tests is strongest when the sample split occurs early in the sample. 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.

# 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 forecasting 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 (e.g., White (2000*b*)). Out-of-sample forecasts also better reflect the information available to the forecaster in "real time" (Diebold & Rudebusch (1991).)

This paper focuses on a dimension of the forecast evaluation problem that has so far received little, if any, attention. When presenting out-of-sample evidence, the sample split defining the beginning of the evaluation period is a choice variable, yet there seems to be no broadly accepted guidelines for how to select the sample split. 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, Stock & Watson (1999) use the first 10 years of data to estimate forecasting models for U.S. inflation while, in their forecasts of US stock returns, Welch & Goyal (2008) use 20 years of monthly observations as the initial estimation sample and the remainder for out-of-sample evaluation. Another approach is to do the reverse and reserve a certain sample length, e.g., 20 years of observations, for the out-of-sample period, as in Inoue & Kilian (2007). Alternatively, researchers such as Rapach, Strauss & Zhou (2010) use multiple out-of-sample forecast samples and report the significance of forecasting performance across all samples. 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 dividing the in-sample and out-of-sample periods raises several questions. First, a 'data-mining' issue arises because researchers could have considered several split points and simply report results for the best choice. 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 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 issue is whether a test statistic that is robust to sample split mining can be derived. 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 all possible sample split points having been considered 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.

A third question is related to how the choice of split point 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 (Inoue & Kilian (2004)), 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.

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 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, tend to 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 grows (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. 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 in this context that 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 end 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.

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). Sullivan, Timmermann & White (1999) show that such adjustments can make a big difference in the context of inference on the ability of technical trading rules to generate excess profits in financial trading. 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 nature of the temporal dependence in forecast performance measured across different sample splits is 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.

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

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

We focus on the common case where forecasts are produced from recursively estimated regression models using least squares and forecast accuracy is evaluated using mean squared errors, (e.g., Diebold & Rudebusch (1991), Inoue & Kilian (2008), Patton & Timmermann (2007), and Stock & Watson (2002).) 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.

## 2.1 A Simple Illustrative Example

Consider the simple regression model that only includes a constant:

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

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

$$\hat{y}_{t+1|t} = \beta_t. \tag{2}$$

The least squares forecast is compared to a simple benchmark forecast

$$\hat{y}^b_{t+1|t} = 0. \tag{3}$$

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, $\lambda \in (0, 1)$, of the sample is reserved for initial estimation while the remaining fraction, $(1-\lambda)$ is used for evaluation. Thus, for a given sample size, $n$, the initial estimation period is $t = 1, \ldots, n_\lambda$ and the (out-of-sample) evaluation period is $n_\lambda + 1, \ldots, n$, where $n_\lambda = \lfloor \lambda n \rfloor$ is the integer part of $\lambda n$.

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

$$D_n(\lambda) = \sum_{t=n_\lambda+1}^{n} (y_t - \hat{y}^b_{t|t-1})^2 - (y_t - \hat{y}_{t|t-1})^2. \tag{4}$$

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

$$T_n(\lambda) = \frac{D_n(\lambda)}{\hat{\sigma}^2_\varepsilon} \xrightarrow{d} 2\int_\lambda^1 u^{-1}B(u)\mathrm{d}B(u) - \int_\lambda^1 u^{-2}B(u)^2\mathrm{d}u, \tag{5}$$

where $B(u)$ is a standard Brownian motion, see McCracken (2007). The CDF for the test statistic in Eq. (5) is denoted by $\Psi_{\lambda,1}(t)$. For a given value of $\lambda$, $T_n(\lambda)$ can be computed and compared to the critical values tabulated in McCracken (2007, table 4). Alternatively, the $p$-value can be computed directly by

$$p_\lambda = 1 - \Psi_{\lambda,1}(t), \qquad \text{where } t = T_n(\lambda). \tag{6}$$

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

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

Since the choice of $\lambda$ is somewhat arbitrary, a researcher may have computed $p$-values for several values of $\lambda$. Even if individual researchers consider only a single value of $\lambda$, the community of researchers could collectively have computed $p$-values for a range of $\lambda$s and this could influence an individual researcher's choice of $\lambda$. Such practices raise the danger of a subtle bias affecting predictive accuracy tests which are only valid provided that $\lambda$ 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.

If the sample split point, $n_\lambda$, is being used as a choice parameter, and the reported $p$-value is in fact the smallest $p$-value obtained over a range of $\lambda$s, $\lambda \in (\gamma, 1-\gamma)$, such as

$$p_{\lambda_{\min}} \equiv \min_{\gamma \le \lambda \le 1-\gamma} p_\lambda, \qquad 0 < \gamma < \frac{1}{2},$$

then it is no longer a valid $p$-value, because the basic requirement of a $p$-value, $\Pr(p_{\lambda_{\min}} \le \alpha) \le \alpha$, does not hold for the smallest $p$-value.[1] Note that we bound the range of admissible

[1] For simplicity the notation suppresses the dependence of $p_{\lambda_{\min}}$ on $\gamma$.

values of $\lambda$ away from both zero and one by assuming that $\lambda \in (\gamma, 1-\gamma)$. Excluding a proportion of the sample at the beginning and end of the data is common practice from the theory on structural breaks and ensures that the distribution of the out-of-sample forecast errors is well behaved.

Figure 1 plots the limit distribution of $p_{\lambda_{\min}}$ as a function of the nominal critical level, $\alpha$. The distribution is shown over its full support in the left panel, and the right panel shows 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 instance the CDF of $p_{\lambda_{\min}}$ is about 14% when evaluated at a 5% critical level, which tells us that there is a 14% probability that the *smallest* $p$-value, $\min_{0.1\leq\lambda\leq0.9}\{p_\lambda\}$, is less than 5%. The figure clearly shows how sensitive out-of-sample predictive inference can be to mining over the split point.

It turns out that this mining is most sensitive to sample splits occurring towards the end of the sample. For instance we find $\min_{0.8\leq\lambda\leq0.9} p_\lambda \leq 0.05$ with a probability that exceeds 10%. Hence, even a relatively modest mining over split points towards the end of the sample can result is substantial over-rejection. To see this, Figure 2 shows the location of the smallest $p$-value, as defined by

$$\left\{\lambda_{\min} : p_{\lambda_{\min}} = \min_{\gamma\leq\lambda\leq1-\gamma} p_\lambda\right\}.$$

The location of the smallest $p$-value, $\lambda_{\min}$, is a random variable with support on the interval $[\gamma, 1-\gamma]$. The histograms in Figure 2 reveal that under the null hypothesis (left panel) the smallest $p$-value is more likely to be located late in the sample (i.e., between 80% and 90% of the data), whereas under the alternative hypothesis the smallest $p$-value is more likely to be found early in the sample. The right panel of Figure 2 shows the location of $\lambda_{\min}$ under the local alternative, $\beta = c\frac{\sigma_\varepsilon}{\sqrt{n}}$, with $c = 3$. For more distant local alternatives such as $c = 5$, the difference becomes more pronounced. As the value of $c$ approaches zero, the histogram under the local alternative approaches that of the null hypothesis.

These findings suggest, first, that conventional tests of predictive accuracy that assume a fixed and pre-determined value of $\lambda$ can substantially over-reject the null of no predictive improvement over the benchmark when in fact $\lambda$ 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.

### 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 is based on a larger regression model with $k+q$ regressors, $X_t = (X_{1t}', X_{2t}')' \in \mathbb{R}^{k+q}$. Forecasts could be computed multiple steps ahead, so the benchmark model's regression-based forecast is now given by

$$\hat{y}^b_{t+h|t} = \tilde{\beta}_{1,t}' X_{1t}, \tag{7}$$

with

$$\tilde{\beta}_{1,t} = \left( \sum_{s=1}^{t} X_{1,s-h} X_{1,s-h}' \right)^{-1} \sum_{s=1}^{t} X_{1,s-h} y_s, \tag{8}$$

while the alternative forecast is

$$\hat{y}_{t+h|t} = \hat{\beta}_{1,t}' X_{1t} + \hat{\beta}_{2,t}' X_{2t}, \tag{9}$$

where $\hat{\beta}_t = (\hat{\beta}_{1,t}', \hat{\beta}_{2,t}')'$ is the least squares estimator obtained by 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,

$$T_n(\lambda) = \frac{\sum_{t=n_\lambda+1}^{n} (y_t - \hat{y}^b_{t|t-h})^2 - (y_t - \hat{y}_{t|t-h})^2}{\hat{\sigma}^2_\varepsilon}, \tag{10}$$

but its asymptotic distribution is now given from a convolution of $q$ independent random variables, $2\int_\lambda^1 u^{-1} B(u) \mathrm{d}B(u) - \int_\lambda^1 u^{-2} B(u)^2 \mathrm{d}u$, 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). 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** (i) $\|V_t\|_{2r}$ *is bounded uniformly in* $t$ *with* $r > 2$; (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* $\alpha$*-mixing of size* $-r/(r-2)$, *or* $\phi$*-mixing of size* $-r/(2(r-1))$.

The assumption establishes $V_t$ as an $L_4$-NED process of size $-\frac{1}{2}$ on $\mathcal{V}_t$, where the latter is an auxiliary process that 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. We have, in particular that $\text{vech}(V_tV_t')$ is $L_2$-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 $\text{vech}(V_tV_t')$.

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

This assumption in conjunction with Assumption 1, ensures that we can establish the desired limit results. Moreover, the assumption ensures that the population regression coefficients, in our predictive regressions, do not depend on $t$.

It is convenient to express the block structure of $\Xi$ in the following ways

$$\Xi = \begin{pmatrix} \Xi_{00} & \bullet \\ \Xi_{z0} & \Xi_{zz} \end{pmatrix} = \begin{pmatrix} \Xi_{00} & \bullet & \bullet \\ \Xi_{10} & \Xi_{11} & \bullet \\ \Xi_{20} & \Xi_{21} & \Xi_{22} \end{pmatrix}.$$

Similarly, define the "error" term

$$\varepsilon_t = y_t - \Xi_{0z}\Xi_{zz}^{-1}X_{t-h},$$

and the auxiliary variable

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

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

Next, we introduce the population objects, $\sigma_\varepsilon^2 = \Xi_{00} - \Xi_{0z}\Xi_{zz}^{-1}\Xi_{z0}$ and $\Sigma = \Xi_{22} - \Xi_{21}\Xi_{11}^{-1}\Xi_{12}$. It follows that $\sigma_\varepsilon^2 > 0$ and that $\Sigma$ is positive definite, because $\Xi$ is positive definite. Finally, define

$$W_n(u) := \frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor un \rfloor} Z_{t-h}\varepsilon_t, \tag{11}$$

which is a CADLAG on the unit interval that maps into $\mathbb{R}^q$. The space of such functions is denoted $\mathbb{D}_{[0,1]}^q$. Let $\Omega$ denote some positive definite matrix. We then have the following theorem:

**Theorem 1** *Given Assumptions 1-2 we have*

$$W_n(u) \Rightarrow W(u),$$

*where* $W(u) \equiv \sigma_\varepsilon \Omega^{1/2} B(u)$ *and* $B(u)$ *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 use the relative accuracy of the two models. Not surprisingly, $\Omega$ will be defined from the long-run variance of $Z_{t-h}\varepsilon_t$ apart from a scaling factor, $\sigma_\varepsilon^2$.

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

This assumption translates into the $h$-step-ahead forecast errors, being "unpredictable" at the time the prediction is made. Without this assumption there would be an asymptotic bias term in the limit distribution given below.

We are now ready to present the limit distribution of the test statistic in the general case.

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

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

*where* $\rho_1, \ldots, \rho_q$ *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 in the form

$$\Psi_{\lambda,\Lambda} = 2 \int_\lambda^1 u^{-1} B'(u) \Lambda \mathrm{d}B(u) - \int_\lambda^1 u^{-2} B'(u) \Lambda B(u) \mathrm{d}u, \tag{12}$$

where $\Lambda = \mathrm{diag}(\rho_1, \ldots, \rho_q)$. $B$ denotes a standard Brownian motion like that in Theorem 1, with one being obtainable from the other through a simple rotation.

Our expression for the asymptotic distribution in Theorem 2 is a great deal simpler than that derived in Clark & McCracken (2005). For instance, our expression simplifies the nuisance parameters to a diagonal matrix, $\Lambda$, as opposed to a full $q \times q$ matrix. Moreover, it is quite intuitive that the "weights", $\rho_1, \ldots, \rho_q$, that appear in the diagonal matrix, $\Lambda$, 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 e.g. White (1994). Our result is also derived under weaker assumptions than those in Clark & McCracken (2005). Our assumptions are an adaptation of those in De Jong & Davidson (2000), which are the weakest known assumptions that ensures convergence to stochastic integrals, such as $\int_0^\lambda B(u)\mathrm{d}B(u)$. Clark & McCracken (2005) use an adaptation of the assumptions in Hansen (1992) which are stronger than those in De Jong & Davidson (2000).

The values of $\rho_1$,..., $\rho_q$ can be estimated as the eigenvalues of $\hat{\Sigma}^{-1}\hat{\Omega}$, where

$$\hat{\Sigma} = \frac{1}{n}\sum_{t=1}^{n} \hat{Z}_{t-h}\hat{Z}'_{t-h}, \qquad \hat{\Omega} = \frac{1}{\hat{\sigma}_\varepsilon^2}\sum_i k(\tfrac{i}{b_n})\hat{\Gamma}_i, \tag{13}$$

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

$$\hat{\Gamma}_j = \frac{1}{n}\sum_{t=1}^{n} \hat{Z}_{t-h}\hat{Z}'_{t-h-j}\hat{\varepsilon}_t\hat{\varepsilon}_{t-j}, \tag{14}$$

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 $Z_{t-h}\varepsilon_t$, which may be applicable when $h = 1$, one can use the estimate $\hat{\Omega} = \frac{1}{\hat{\sigma}_\varepsilon^2}\frac{1}{n}\sum_{t=1}^n \hat{Z}_{t-1}\hat{Z}'_{t-1}\hat{\varepsilon}_t^2$.

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 $\Psi_{\lambda,\Lambda}$ to $\Psi_{\lambda,q}$. 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).

### 2.2.1 Rejection Rates Induced by Mining over the Sample Split

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

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

Suppose instead that the out-of-sample test statistic, $T_\lambda$, is computed over a range of split points, $\gamma \leq \lambda \leq 1-\gamma$, 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 1 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, $i$, for which at least one rejection of the null occurs at the $\alpha$ 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. For example, with one additional regressor ($q = 1$), a test for no predictability that would reject 5% of the time if conducted for a fixed sample split, rejects three times as often as a result of mining over the sample split point, namely 14.8% of the time. Moreover, this rejection rate 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% critical 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% critical 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.

## 3 Power of the Test

The scope for size distortions of 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 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 start of the forecast evaluation period occurs early in the sample.

Under the alternative hypothesis we have the following result:

**Theorem 3** *Suppose that Assumptions 1-3 hold, and consider the local alternative* $\beta_2 =$

$c\frac{\sigma_\varepsilon}{\sqrt{n}}a'$, *where* $a \in \mathbb{R}^q$ *with* $a'\Sigma a = 1$. *Then*

$$\begin{aligned} T_n(\lambda) \quad &\xrightarrow{d} \quad c^2(1-\lambda) + 2ca'\Omega^{1/2}Q' \left[B(1) - B(\lambda)\right] \\ &\quad +2\int_\lambda^1 u^{-1}B(u)'\Lambda \mathrm{d}B(u) - \int_\lambda^1 u^{-2}B(u)'\Lambda B(u)\mathrm{d}u, \end{aligned}$$

*where the matrix* $Q$ *is obtained from* $Q'\Lambda Q = \Omega^{1/2}\Sigma^{-1}\Omega^{1/2}$.

This Theorem provides the theoretical explanation for the simulation results in Clark & McCracken (2001).

### 3.1 Local Power in the Illustrative Example

In our illustrative example from Section 2.1 a local alternative takes the form

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

and so the limit distribution is given by

$$T_n(\lambda) \xrightarrow{d} 2\int_\lambda^1 u^{-1}B\mathrm{d}B - \int_\lambda^1 u^{-2}B^2\mathrm{d}u + c^2(1-\lambda) + 2c\left[B(1) - B(\lambda)\right]. \tag{15}$$

The power depends on the split point, which can be illustrated by the distribution of the $p$-value under the local alternative. Recall that the $p$-value is defined by $p_\lambda = 1 - \Psi_{\lambda,1}(T_n(\lambda))$. Figure 3 presents the distribution of $p_\lambda$ for two local alternatives, $c = 1$ and $c = 2$, and three sample split ratios, $\lambda = 0.25$, $\lambda = 0.50$, and $\lambda = 0.75$. The two upper panels set $c = 1$ while the lower panels set $c = 2$. The right panels zoom in on the lower left corner of the left panels. If a 5% critical value is used, the upper panels ($c = 1$) show that the power of the test will be about 16%, 14%, and 13% for $\lambda = 0.25$, $\lambda = 0.50$ and $\lambda = 0.75$, respectively. For $c = 2$ (lower panels) the power is 45%, 39%, and 33% for $\lambda = 0.25$, $\lambda = 0.50$ and $\lambda = 0.75$, respectively. Hence, the power is substantially higher with $\lambda = 0.25$ than $\lambda = 0.75$.

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

While this finding is in line with that of Inoue & Kilian (2004), it raises important questions concerning the appropriateness of testing the null hypothesis $\beta_0 = 0$ using the test statistic $T_n(\lambda)$. Under a recursive estimation scheme, a short initial estimation sample

is associated with greater estimation errors and hence will tend to drag down forecasting performance, particularly at the beginning of the sample. However, it also results in a longer out-of-sample evaluation window and the concomitant higher power. A long initial estimation sample reduces the effect of estimation error on the initial forecasts, but also lowers the power due to the shorter evaluation sample. The trade-off between these effects is complicated by the highly persistent nature of parameter estimation errors when a recursive estimation scheme is used to generate forecasts. Further discussion of this point is beyond the scope of the present paper.

# 4 A Split-Mining Robust Test

The results in Table 1 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 whether a test can be designed 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 $\lambda \in [\gamma, 1-\gamma]$. One possibility is to consider the maximum value of $T_n(\lambda)$ across a range of split points. However, $\max_{\lambda\in[\gamma,1-\gamma]} T_n(\lambda)$ is ill-suited for this purpose, because the marginal distribution of $T_n(\lambda)$ varies a great deal with $\lambda$. Because of the resulting heteroskedasticity across different $\lambda-$values, the max-$T_n(\lambda)$ statistic implicitly favors certain values of $\lambda$.

Instead, we propose to first translate the test statistics for each of the sample split points into nominal $p$-values, $p_\lambda = 1 - \Psi_{\lambda,\Lambda}(T_n(\lambda))$. In a second step, the smallest $p$-value is computed:

$$p_{\lambda_{\min}} = \min_{\lambda\in[\gamma,1-\gamma]} p_\lambda.$$

Because each of the $p$-values, $p_\lambda$, are uniformly distributed on the unit interval (asymptotically) the resulting test statistic is constructed from test statistics with similar properties. The limit distribution of $p_{[\gamma,1-\gamma]}$ will clearly not be uniformly distributed, so the smallest $p$-value, $p_{\lambda_{\min}}$, cannot be interpreted as a valid $p$-value, but should instead be viewed as a test statistic, whose distribution we seek. To this end, for $u \in (0,1)$ define

$$G(u) = 2\int_u^1 s^{-1}B'(s)\Lambda \mathrm{d}B(s) - \int_u^1 s^{-2}B'(s)\Lambda B(s)\mathrm{d}s.$$

Furthermore, we make the following assumption:

**Assumption 4** *For* $0 < \gamma \leq 1/2$

$$T_n(u) \Rightarrow G(u) \qquad \text{on } \mathbb{D}_{[\gamma,1-\gamma]}.$$

Assumption 4 requires a joint convergence that is stronger than the point-wise result established earlier. A closely related result, which appears in the literature on unit roots, is $\sum_{t=1}^{\lfloor nu \rfloor} \sum_{s=1}^{t-1} \varepsilon_s \varepsilon_t \Rightarrow \sigma_\varepsilon^2 \int_0^u B(s) \mathrm{d}B(s)$, $u \in [0,1]$. This joint convergence is known to hold under several sets of assumptions. However, the joint convergence has not been established with near-epoch assumptions that are the weakest set of assumptions needed for the functional central limit theorem and the (point-wise) convergence to the stochastic integral, such as $\sum_{t=1}^{\lfloor nu \rfloor} \sum_{s=1}^{t-1} \varepsilon_s \varepsilon_t \xrightarrow{d} \sigma_\varepsilon^2 \int_0^u B(s) \mathrm{d}B(s)$ for a particular value of $u$, see De Jong & Davidson (2000).[2] Hence, Assumption 4 may turn out to be redundant in this context.

**Theorem 4** *Given Assumption 1-4,* $p_{\lambda_{\min}}$ *converges in distribution, and the cdf of the limit distribution is given by*

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

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

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

Using this result we can compute the $p$-value adjusted for sample split mining by sorting the $p_{\lambda_{\min}}$-values for all paths and choosing the $\alpha$-quantile of this (ranked) distribution.

Table 2 shows how nominal $p$-values translate 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

[2] Joint convergence requires an extension of the pointwise convergence, $T_n(u) \xrightarrow{d} G(u)$, to the joint convergence, e.g., by the use of stochastic uniform equicontinuity. Stochastic uniform equicontinuity requires

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \Pr\left( \sup_{|u-u'|<\delta} \left| T_n(u) - T_n(u') \right| > \epsilon \right) = 0,$$

where $T_n(u) = D_n(u)/\hat{\sigma}_\varepsilon^2$ with $D_n(u) = \sum_{t=n_u+1}^{n} (y_t - \hat{y}_{t|t-h}^b)^2 - (y_t - \hat{y}_{t|t-h})^2$. Assuming that $u < u'$, for the simplest case

$$\begin{aligned} D_n(u) - D_n(u') &= \sum_{t=n_u+1}^{n_{u'}} (y_t - \hat{y}_{t|t-h}^b)^2 - (y_t - \hat{y}_{t|t-h})^2 \\ &= \sum_{t=n_u+1}^{n_{u'}} \left( \beta^2 + 2\beta\varepsilon_t - \left( \tfrac{1}{t-1} \sum_{s=1}^{t-1} \varepsilon_s \right)^2 + 2 \left( \tfrac{1}{t-1} \sum_{s=1}^{t-1} \varepsilon_s \right) \varepsilon_t \right). \end{aligned}$$

$p$-value computed using the McCracken (2007) test statistic at all possible split points $\lambda \in [0, 1, 0.9]$ 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.

In related work, Rossi & Inoue (2011) construct a method for out-of-sample forecast evaluation that pertains to the case where the parameters of the forecast model are estimated under a rolling window. Their approach is robust to data snooping across the length of the estimation window, but does not consider the effect of using different sample splits, as the length of the out-of-sample evaluation period is fixed in their analysis. Instead they construct a test statistic as the supremum over different window sizes of a conventional forecast performance test.

### 4.1 A Simple Robustness Check

Researchers may be aware of the problem arising if multiple values for the sample split, $\lambda$, have been considered and so may only look at a single value of $\lambda$, although their choice may be influenced by what other researchers have done. For such researchers the previous approach could be too conservative. If all researchers could agree ex ante on a common split ratio, $\lambda^*$ say, and all reported $p_{\lambda^*}$, it would eliminate the problems arising from mining over split points.

One possible suggestion is to always report the $p$-value computed at $\lambda = 0.50$. In specific applications there might be good arguments for using a different sample split, yet in such cases it would still be beneficial to report $p_{0.5}$ in conjunction with the "preferred" value of $\lambda$. For instance if both values are significant it offers some protection against the criticism that the split point was selected through split mining because, when $n$ is large,

$$\Pr(p_\lambda \leq \alpha, p_{0.5} \leq \alpha) \leq \Pr(p_{0.5} \leq \alpha) \approx \alpha.$$

# 5 Empirical Examples

This section provides empirical illustrations of the methods and results discussed in the previous sections. We consider two forecasting questions that have attracted considerable empirical interest in economics and finance, namely whether the corporate default spread helps predict stock returns and whether inflation forecasts can be improved by using broad summary measures on 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 forecast combinations spanning quarterly returns over 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

$$y_{t+1} = \beta_0 + \varepsilon_{t+1}, \tag{16}$$

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

$$y_{t+1} = \beta_0 + \beta_1 x_t + \varepsilon_{t+1}. \tag{17}$$

We focus on models where $x_t$ is the default spread, measured as the difference between the yield on AAA-rated corporate bonds versus that on BAA-rated corporate bonds. Our data consist of monthly observations on stock returns on the S&P500 index and the corresponding yield spread over the period from 1926:01 to 2008:12 (a total of 996 observations). Setting $\gamma = 0.1$, our initial estimation sample uses one hundred observations and so the beginning of the various forecast evaluation periods runs from 1934:05 through 2000:04. The end point of the out-of-sample period is always 2008:12.

The top window in Figure 4 shows how the $T_n(\lambda)$-statistic evolves over the forecast evaluation period.[3] The minimum value obtained for $T_n(\lambda)$ is $-6.79$, while its maximum is 2.18. Due to the partial overlap in both estimation and forecast evaluation windows, as expected, the test statistic evolves relatively smoothly and is quite persistent, although the effect of occasional return outliers is also clear from the plot. Towards the end of the sample (where $\lambda$ is close to 0.90), the test statistic shows a mild upward drift.

The $p_\lambda$-values associated with the $T_n(\lambda)$ statistics computed for different values of $\lambda$ are plotted in the bottom 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.

The sensitivity of the empirical results to the choice of $\lambda$ 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 $\lambda \in [0.1, 0.9]$ is 0.03. Table 2 suggests that this corresponds to a split-mining adjusted $p-$value that exceeds 10%. Hence, the evidence of time-varying return predictability from the yield spread is not statistically significant at conventional levels. We cannot therefore 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 the effect of mining over the beginning of the out-of-sample period.

To illustrate that some forecasting models are in fact robust to mining over the sample selection split, we also considered a return forecasting model that uses the lagged dividend yield as the predictor variable. Using the same sample as above, for this model we found that the maximum value of $T_n(\lambda)$ was 5.27 and the smallest $p$-value fell below 0.001 which,

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

according to Table 2, means that out-of-sample predictability from this model is robust to mining over the sample split. Interestingly, for this model, predictability is concentrated towards the very end of the sample, i.e., from the late nineties and onwards, and does not seem to be present for other subsamples, consistent with an alternative explanation related to structural breaks in the forecast model.

### 5.2 Inflation Forecasts

Simple autoregressive prediction models have been found to perform well for many macroeconomic variables capturing wages, prices and inflation (Marcellino, Stock & Watson (2006) and Pesaran, Pick & Timmermann (2010)). 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:

$$y_{t+1} = \beta_0 + \sum_{i=1}^{2} \beta_{yi} y_{t+1-i} + \varepsilon_{y,t+1}, \tag{18}$$

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 Eq. (18):

$$y_{t+1} = \beta_0 + \sum_{i=1}^{2} \beta_{yi} y_{t+1-i} + \sum_{i=1}^{4} \beta_{fi} \hat{f}_{it} + \varepsilon_{y,t+1}. \tag{19}$$

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 (2007) 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 $\gamma = 0.1$, the out-of-sample forecasting period runs from mid-1973 through early 2004.

The top window in Figure 5 shows the $T_n(\lambda)$-statistic for different values of $\lambda$. This rises throughout most of the sample from around -23 to a terminal value just above zero. The associated $p_\lambda$-values are shown in the bottom 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 $p_\lambda$ plot declines sharply to values below 0.10, before oscillating for much of the rest of the sample, with an overall minimum $p$-value is 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 $\lambda$, this conclusion could be misleading since the split-mining adjusted test statistic is not significant. In fact, the globally minimum $p$-value of 0.023 is not even significant at the 10% level when compared against the split-mining adjusted $p$-values in Table 2.

# 6 Conclusion

Choice of the sample split used to divide data into an in-sample estimation period and an out-of-sample evaluation period 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 may−individually or collectively−have mined over the split point, forecast evaluation tests can be grossly oversized, leading to spurious evidence of predictability. In fact, the nominal rejection rates can be more than tripled as a result of such mining over the split point, and the danger of spurious rejection tends to be highest when a short evaluation window is used, i.e., when the out-of-sample period begins late in the sample. Conversely, power is highest when the out-of-sample period is as long as possible and so the evaluation window begins early.

Two empirical applications show that choice of sample split can have important consequences in practice for conclusions on whether economic time-series are predictable. Variations in U.S. stock returns do not appear to be predictable by means of the lagged default spread, nor does U.S. consumer price inflation appear to be predictable by means of common factors in a way that is robust to how the start of the out-of-sample period is selected.

# 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.

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. (2007), 'How useful is bagging in forecasting economic time series? a case study of u.s. CPI inflation', *Journal of the American Statistical Association* . forthcoming.

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 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. (2007), 'The empirical risk-return relation: A factor analysis approach', *Journal of Financial Economics* **83**, 171–222.

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.

Patton, A. & Timmermann, A. (2007), 'Testing forecast optimality under unknown loss', *Journal of American Statistical Association* **102**, 1172–1184.

Pesaran, M. H., Pick, A. & Timmermann, A. (2010), 'Variable selection, estimation and inference for multi-period forecasting problems', *working paper* .

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. H. & Watson, M. W. (1999), 'Forecasting inflation', *Journal of Monetary Economics* **44**, 293–335.

Stock, J. H. & Watson, M. W. (2002), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**, 1167–1179.

Sullivan, R., Timmermann, A. & White, H. (1999), 'Data-snooping, technical trading rules, and the bootstrap.', *Journal of Finance* **54**, 1647–1692.

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

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.

# Appendix of Proofs

## A.1 Derivations related to the simple example

Suppose that $\beta = c\sigma_\varepsilon/\sqrt{n}$. Then, from Equations (1)-(4), we have

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

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,

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

## A.2 Proof of Theorem 1

Our assumptions follow De Jong & Davidson (2000), adapted 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 $\mathcal{U}_t = \mathrm{vech}(V_tV_t' - \Xi)$ and consider $X_{nt} = \omega'\mathcal{U}_t/\sqrt{n}$ for some arbitrary vector $\omega$, so that $\omega'\Psi\omega = 1$, where $\Psi = \mathrm{var}[n^{-1/2}\sum_{t=1}^{n}\mathrm{vech}(V_tV_t' - \Xi)]$, which is well defined under Assumption 2. 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 $\Psi$. From this it also follows that

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

which we will use in our 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 $A_n = \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t-1}\mathrm{E}\mathcal{U}_s\mathcal{U}_t'$.

Now define the matrices

$$L = (0_{q\times 1}, -\Xi_{21}\Xi_{11}^{-1}, I_{q\times q}) \text{ and } R = (1, -\Xi_{zz}^{-1}\Xi_{z0}).$$

Then it is easy to verify that $L\Xi R' = 0$ and

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

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

Define $U_{nt} = Z_{t-h}\varepsilon_t/\sqrt{n}$, $W_n(u) = \sum_{t=1}^{\lfloor un\rfloor} U_{nt}$, 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 1-3, 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 $h$-period ahead forecasts, we should 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 3 by imposing the null hypothesis, i.e., by setting $c = 0$.

## A.4 Proof of Theorem 3

To prove Theorem 3, we first establish two lemmas.

**Lemma A.1** *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* $\zeta_t = \hat{\beta}_{2,t}'(\Xi_{21}\Xi_{11}^{-1} - M_{21,t}M_{11,t}^{-1})X_{1,t}$.

**Proof.** For the benchmark forecast in Eq. (7) 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 Eq. (9) we have

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

where $M_{21,t} = \sum_{s=1}^{t-1} X_{2,s} X_{1,s}'$ and $M_{11,t} = \sum_{s=1}^{t-1} X_{1,s} X_{1,s}'$, so that

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

Consider next the loss differential, which from equations (7) to (9) is given by

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

The result now follows by multiplying out. ■

**Lemma A.2** *With* $\beta_2 = c\frac{\sigma_\varepsilon}{\sqrt{n}} v$ *for some* $v \in \mathbb{R}^q$ *and given Assumptions 1-3 we have,*

$$\sum_{\lfloor \lambda n \rfloor + 1}^{n} \beta_2' Z_{t-h} Z_{t-h}' \beta_2 \quad \xrightarrow{p} \quad (1-\lambda) c^2 \sigma_\varepsilon^2 v' \Sigma v \tag{A.2}$$

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

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

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

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

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

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

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

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

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

**Proof.** To simplify notation, introduce

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

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),

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

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

$$\beta_2' \sum_{\lfloor \lambda n \rfloor + 1}^{n} Z_{t-h} \varepsilon_t = c\sigma_\varepsilon v' \left[ W_n(1) - W_n(\lambda) \right],$$

and Theorem A.1. Next,

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

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

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

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

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

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

The last term in this expression is $O_p(n^{-1/2})$ because with $\mathcal{V}_n(u)=\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor un\rfloor}\mathrm{vec}(Z_{t-h}Z_{t-h}'-\Sigma)$, 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

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

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

$$\frac{1}{n}\sum_{t=\lfloor\lambda n\rfloor+1}^{n} W_n(\tfrac{t-h}{n})'\Sigma_n^{-1}(\tfrac{t}{n})\Sigma\Sigma_n^{-1}(\tfrac{t}{n})W_n(\tfrac{t-h}{n})$$

$$
\begin{aligned}
&= \int_\lambda^1 W_n(u)'\Sigma_n^{-1}(u)\Sigma\Sigma_n^{-1}(u)W_n(u)\mathrm{d}u \\
&= \int_\lambda^1 u^{-2}W_n(u)'\Sigma^{-1}W_n(u)\mathrm{d}u + o_p(1) \\
&\xrightarrow{d} \int_\lambda^1 u^{-2}W(u)'\Sigma^{-1}W(u)\mathrm{d}u.
\end{aligned}
$$

Next consider the terms involving $\zeta_t$ and/or $Z_{t-h}X_{1,t-h}'$. First

$$
\begin{aligned}
\sum_{\lfloor\lambda n\rfloor+1}^{n} c\sigma_\varepsilon v'\tfrac{Z_{t-h}X_{1,t-h}'}{n}n^{1/2}(\tilde\beta_{1,t-h}-\delta) &\approx o_p(1)O_p(1). \\
\sum_{\lfloor\lambda n\rfloor+1}^{n} n^{1/2}(\hat\beta_{2,t-h}-\beta_2)'\tfrac{Z_{t-h}X_{1,t-h}'}{n}n^{1/2}(\tilde\beta_{1,t-h}-\delta) &\approx O_p(1)o_p(1)O_p(1),
\end{aligned}
$$

from which equations (A.6) and (A.7) follow. Next recall that $\zeta_t = \hat\beta_{2,t}'(\Xi_{21}\Xi_{11}^{-1}-M_{21,t}M_{11,t}^{-1})X_{1,t-h}$ and for any fixed $\gamma>0$, we have by (A.1) that $\sup_{t\geq\gamma n}\left|M_{21,t}M_{11,t}^{-1}-\Xi_{21}\Xi_{11}^{-1}\right| = o_p(1)$, so both expressions vanish in the limit. Next,

$$
\begin{aligned}
\sum_{\lfloor\lambda n\rfloor+1-h}^{n-h}\zeta_t^2 &= \sum_{\lfloor\lambda n\rfloor+1-h}^{n-h} n^{1/2}\hat\beta_{2,t}'(\Xi_{21}\Xi_{11}^{-1}-M_{21,t}M_{11,t}^{-1})\tfrac{X_{1,t-h}X_{1,t-h}'}{n}(\Xi_{21}\Xi_{11}^{-1}-M_{21,t}M_{11,t}^{-1})'n^{1/2}\hat\beta_{2,t} \\
&= O_p(1)o_p(1)o_p(1)o_p(1)o_p(1)O_p(1), \\
\sum_{\lfloor\lambda n\rfloor+1}^{n}\zeta_{t-h}\varepsilon_t &= \sum_{\lfloor\lambda n\rfloor+1}^{n} n^{1/2}\hat\beta_{2,t-h}'(\Xi_{21}\Xi_{11}^{-1}-M_{21,t-h}M_{11,t-h}^{-1})n^{-1/2}X_{1,t-h}\varepsilon_t \\
&= O_p(1)o_p(1)O_p(1), \\
&\quad \sum_{\lfloor\lambda n\rfloor+1}^{n} n^{1/2}\hat\beta_{2,t-h}'(\Xi_{21}\Xi_{11}^{-1}-M_{21,t-h}M_{11,t-h}^{-1})\tfrac{X_{1,t-h}X_{1,t-h}'}{n}n^{-1/2}(\tilde\beta_{1,t-h}-\delta) \\
&= O_p(1)o_p(1)o_p(1)O_p(1), \\
&\quad \sum_{\lfloor\lambda n\rfloor+1}^{n} n^{1/2}\hat\beta_{2,t-h}'(\Xi_{21}\Xi_{11}^{-1}-M_{21,t-h}M_{11,t-h}^{-1})\tfrac{X_{1,t-h}Z_{t-h}'}{n}n^{-1/2}(\hat\beta_{2,t-h}-\beta_2) \\
&= O_p(1)o_p(1)o_p(1)O_p(1).
\end{aligned}
$$

All of these terms vanish in the limit, proving (A.8) - (A.11). ■

From the decomposition in Lemma A.1 and the limit results in Lemma A.2 we are now ready to derive the asymptotic properties of $D_n(\lambda)$ and $T_n(\lambda)$. From Lemmas A.1 and A.2 it follows that

$$
\begin{aligned}
T_n(\lambda) = \frac{D_n(\lambda)}{\hat\sigma_\varepsilon^2} &\xrightarrow{d} c^2(1-\lambda)v'\Sigma v + 2cv'\Omega^{1/2}\left[B(1)-B(\lambda)\right] \\
&\quad +2\int_\lambda^1 u^{-1}B(u)'\Omega^{1/2}\Sigma^{-1}\Omega^{1/2}\mathrm{d}B(u)
\end{aligned}
$$

$$- \int_{\lambda}^{1} u^{-2} B(u)' \Omega^{1/2} \Sigma^{-1} \Omega^{1/2} B(u) \mathrm{d}u.$$

Now decompose $\Omega^{1/2}\Sigma^{-1}\Omega^{1/2} = Q'\Lambda Q$, where $\Lambda = \mathrm{diag}(\rho_1, \ldots, \rho_q)$ 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,

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

from which Theorem 3 follows. □

## A.5 Proof of Theorem 4

**Proof.** From the definition of $G(u)$ (defined through Assumptions 1-3), it follows that the path of critical values, $c_\alpha(u)$ is continuous in $u$ (because $\Psi_{u,\Lambda}(x)$ is continuous in $(u, x)$ on $[\gamma, 1-\gamma] \times \mathbb{R}$), so $c_\alpha(u) \in \mathbb{D}_{[\gamma,1-\gamma]}$. Hence, by the continuous mapping theorem and Assumption 4, the smallest $p$-value over the range of split points, $[\gamma, 1-\gamma]$, converges in distribution and the CDF of the limit distribution is given by

$$\begin{aligned} \Pr\{p_{[\gamma,1-\gamma]} \leq \alpha\} &= \Pr\{G(u) \geq c_\alpha(u) \text{ for some } u \in [\gamma, 1-\gamma]\} \\ &= \Pr\{\sup_{\gamma \leq u \leq 1-\gamma} [G(u) - c_\alpha(u)] \geq 0\}. \end{aligned}$$

■

**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 1: This table shows the actual rejection rate for different nominal critical levels ($\alpha$)) and different dimensions ($q$) of the alternative model relative to the benchmark. Simulations are conducted under the null model with $\gamma = 0.1$. 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 2: This table shows the split-mining adjusted critical values at which the minimum $p$-value, $p_{[\gamma,1-\gamma]}$, is significant when $\gamma = 0.1$. 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 0.01and use a discretization with $n = 10,000$ and $N = 10,000$ simulated series).

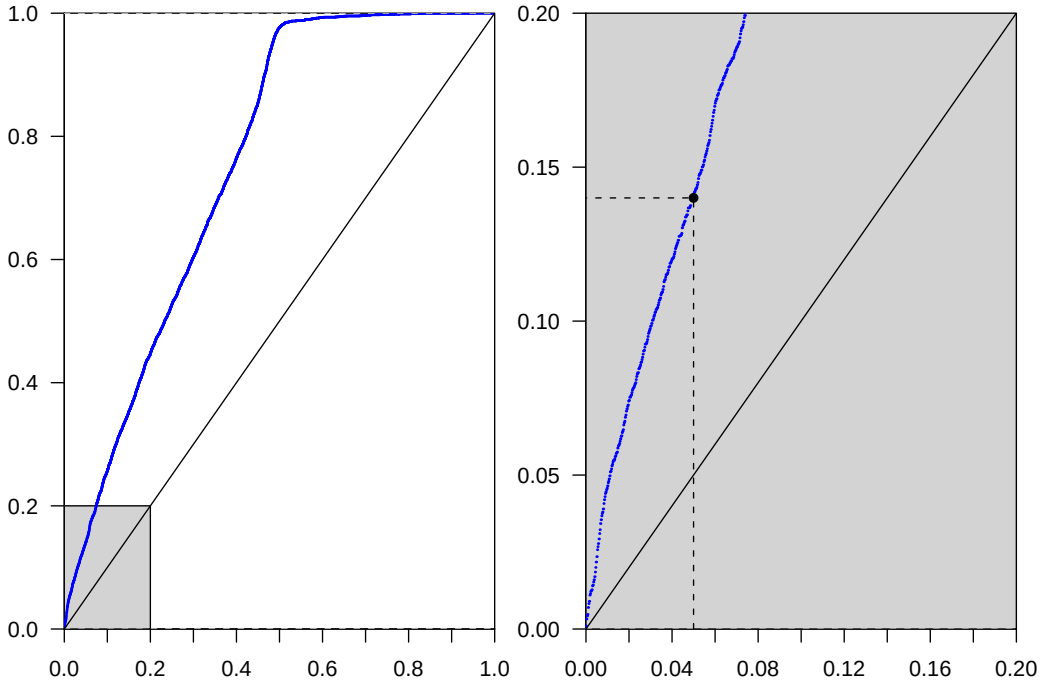


Figure 1: The CDF of the minimum $p$-value, $p_{[\gamma,1-\gamma]}$, for $\gamma = 0.1$.

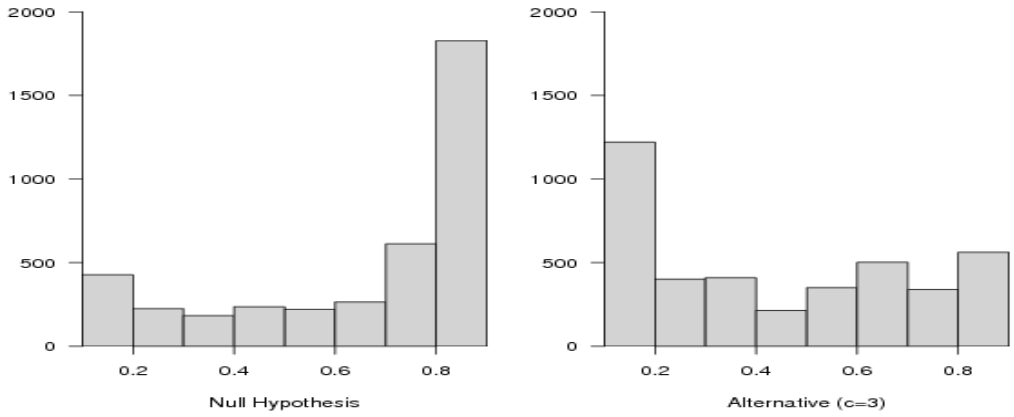


Figure 2: Histograms of the location of the smallest $p$-value under the null hypothesis and the alternative. Under the null hypothesis, the smallest $p$-value, $\min_{\gamma \leq s \leq 1-\gamma} p_s$, is most likely to be located towards the end of the sample, while under the alternative the smallest $p$-value is more likely to be located early in the sample.

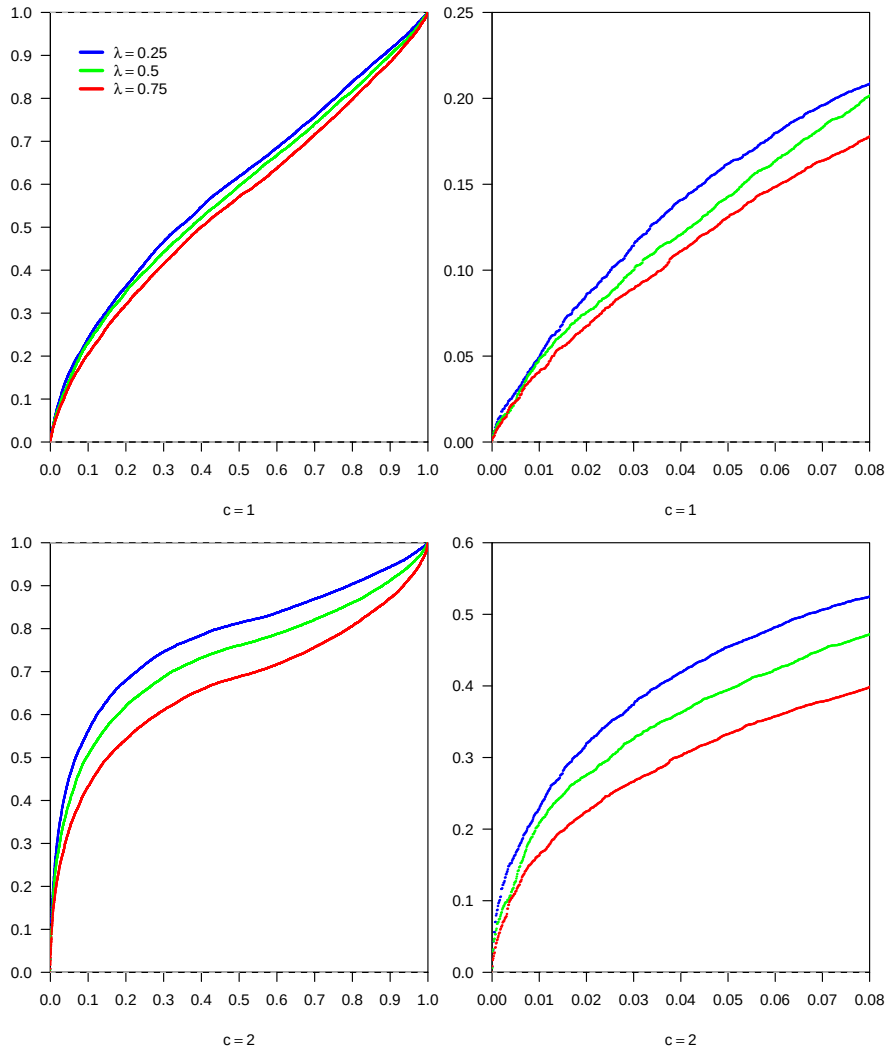


Figure 3: Distribution of $p$-values under the local alternatives $c = 1$ and $c = 2$ for $\lambda = 0.25$, 0.50 and 0.75, when $q = 1$, $\Lambda = 1$, and $h = 1$. Note that power is largest for $\lambda = 0.25$.

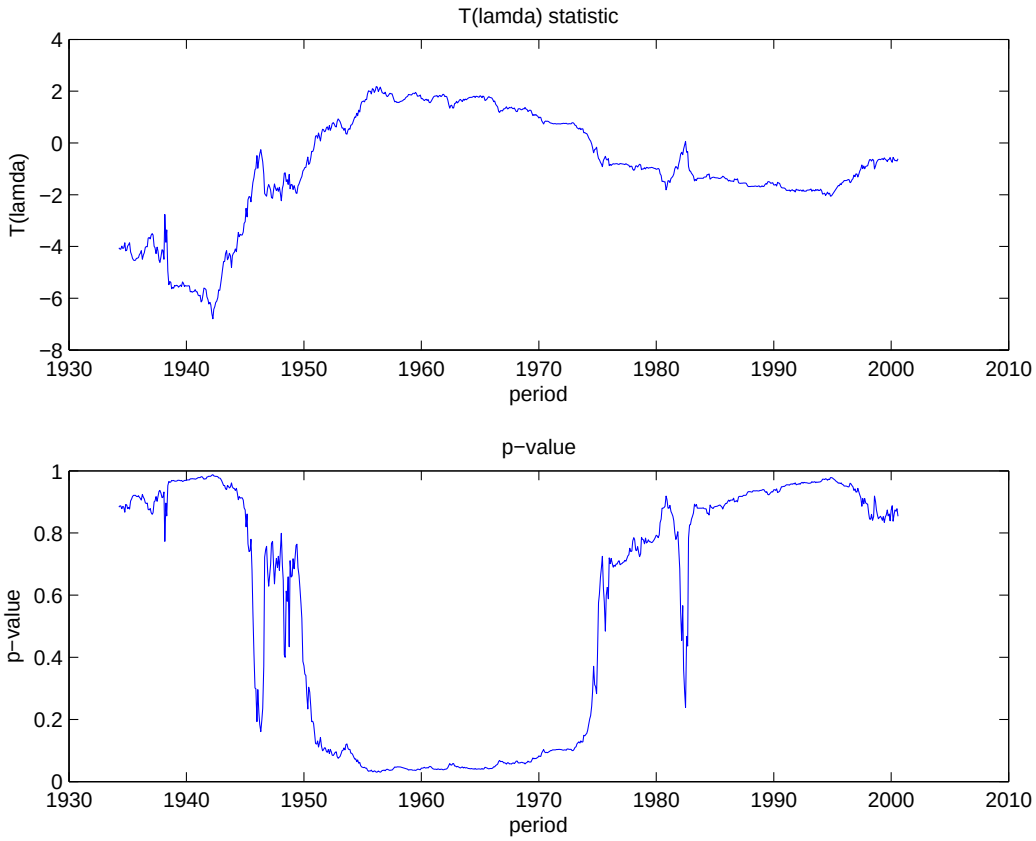


Figure 4: Values of the $T_\lambda(n)$ statistic and $p_\lambda$-values for different choices of the sample split point, $\lambda$. Values are based on the U.S. stock return prediction model that uses the default spread as a predictor variable.

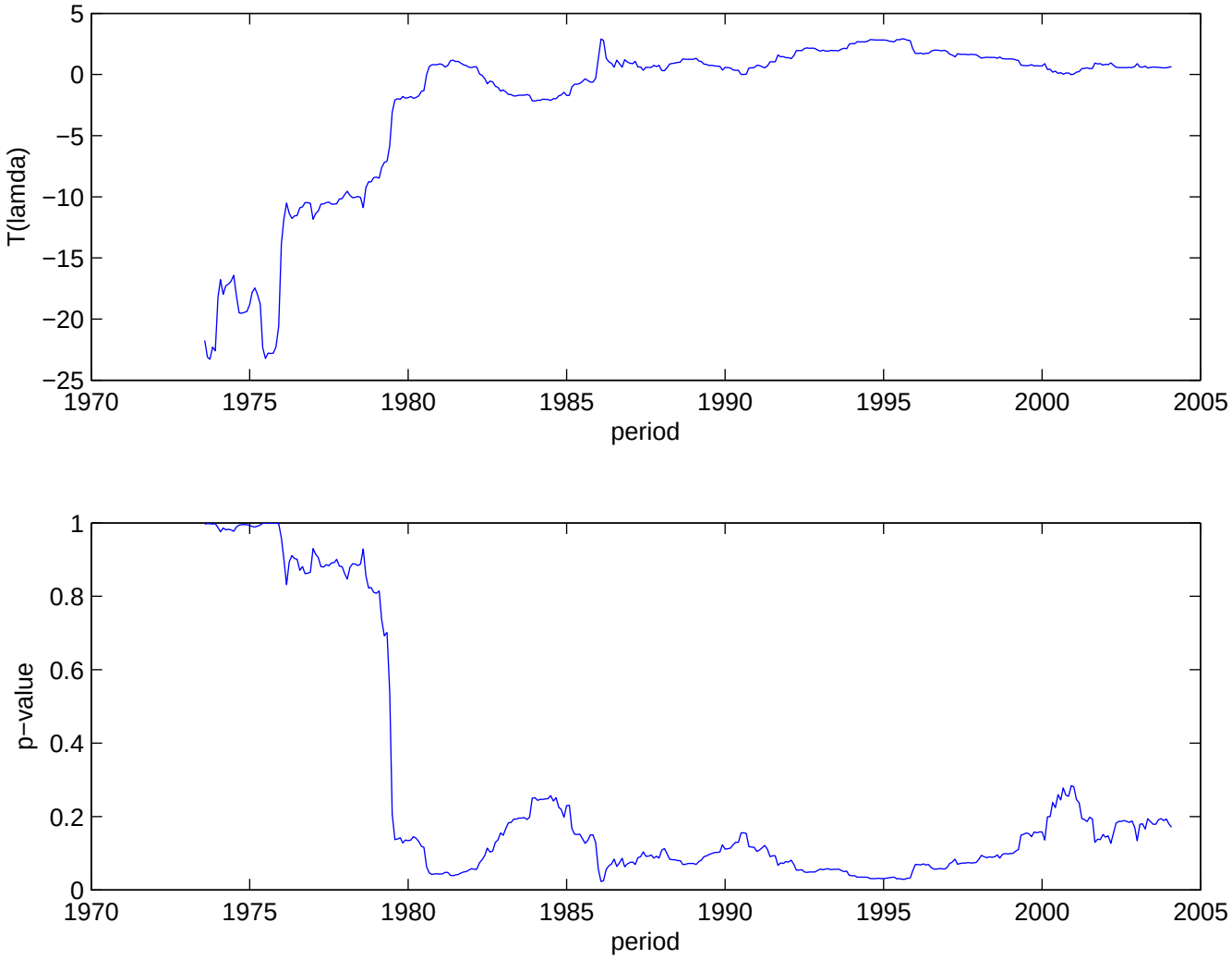


Figure 5: Values of the $T_\lambda(n)$ statistic and $p_\lambda$-values for different choices of the sample split point, $\lambda$. The plots are based on the U.S. inflation prediction model that uses four common factors as additional predictor variables.