Specification Tests Robust to Multiple Instabilities

Lukas Hoesch*

Job Market Paper

This version: November 20, 2020

[Click here for most recent version]

Abstract

I develop a hypothesis test for model evaluation which is robust to time-variation in parameters. The proposed method can be applied in-sample and out-of-sample to any economic model based on moment conditions. In-sample, the test selects between two nested model specifications in the presence of parameter instabilities. Out-of-sample, the test can be used to evaluate the performance of model or judgmental forecasts robust to time-variation. The key feature of the proposed test is that it is particularly powerful in the presence of multiple shifts in parameters without imposing a specific form of time-variation. Further, the test statistic provides narrative evidence on which parts of the sample drive the rejection of the null hypothesis. Simulations show that the test is accurately sized in finite samples and is more powerful than tests assuming constant coefficients or a single break if the data-generating process exhibits multiple shifts in parameters. Using the proposed test, I document the presence of short-horizon predictability in the U.S. equity premium during the postwar period. I find evidence of predictability for a large set of variables once time-variation is taken into account. The test further provides evidence of heterogeneity in the location of predictability episodes across variables. The findings explain why traditional tests often fail to uncover predictability in the full sample and why studies that split the sample at different dates often arrive at conflicting results regarding the predictive ability of a wide class of variables.

JEL classification: C22, C32, C52, C53, C58

Keywords: Specification Tests, Instabilities, Forecasting, Predictability of Stock Returns

*Universitat Pompeu Fabra and Barcelona Graduate School of Economics. Address: Department of Economics and Business, Universitat Pompeu Fabra, Carrer Ramon Trias Fargas 25-27, Mercé Rodoreda Building, 08005 Barcelona, Spain. Email: lukas.hoesch@upf.edu. Website: www.lukashoesch.com.
I am grateful for the support, advice and encouragement I have received from Barbara Rossi. I would also like to thank Geert Mesters for his support and valuable comments. Further, I would like to thank Christian Brownlees, Bjarni G. Einarsson, Kirill Evdokimov, Katharina Janezic, Adam Lee, Katerina Petrova, Francesco Ravazzolo, Tatevik Sekhposyan, André Souza, as well as participants at the 40th International Symposium on Forecasting, the 8th SIdE Workshop for PhD students in Econometrics and Empirical Economics, the 7th Barcelona GSE PhD Jamboree and the Econometrics seminars at Universitat Pompeu Fabra for their feedback and helpful comments. I gratefully acknowledge financial support from the 7th Economics Job Market Best Paper Award (UniCredit Foundation) and the Best Student Presentation Award of the International Institute of Forecasters sponsored by Amazon.

1 Introduction

Instabilities in models of economic and financial time series are widespread. For example, when forecasting U.S. stock returns, financial ratios might contain useful information during some periods, but have no predictive ability during other periods (Farmer et al., 2019; Chinco et al., 2019). Similarly, when estimating a structural model of real economic activity, the absence of financial frictions in economic models might be unproblematic if the estimation sample covers "normal" periods, but might be a crucial omission during financial crises (Christiano et al., 2018). Instabilities are often implicitly acknowledged by conducting sensitivity checks over different subsamples and are sometimes explicitly addressed by testing for structural breaks in model parameters. However, they are commonly ignored when evaluating the specification of a model by means of hypothesis tests. Recently, a growing literature has raised concerns about this practice, arguing that in unstable environments, traditional specification tests have low power and may give conflicting results depending on the subsample considered (Rossi, 2013, 2020). This issue is particularly relevant when estimation samples span long time periods which cover different policy regimes, making it likely that model parameters undergo more than one shift. In such an environment, researchers face an econometric problem: "How can we take multiple instabilities into account when evaluating economic models or their forecasts?"

In this paper, I provide a general approach to test whether a parameter should be included in a model robust to instabilities. The proposed hypothesis test can be applied in-sample to select between two nested specifications of an economic model in the presence of parameter instabilities[1] or out-of-sample to evaluate the forecasting performance of model or judgmental forecasts robust to time-variation.[2] The main advantage of the proposed test is that it is particularly powerful in the presence of multiple shifts in parameters without imposing a specific form of time variation. At the same time, the test is accurately sized in finite-samples and has high power even when model parameters only undergo one shift or are constant. This makes the test particularly useful when the researcher faces uncertainty about whether and how parameters change over time. The test is simple to compute, can be efficiently implemented by a dynamic programming algorithm provided in the paper and the test statistic path can be plotted to provide narrative evidence on which parts of the sample drive the rejection of the null hypothesis.

[1] Such tests are widely used in the macroeconomic and financial literature and many studies document evidence of instabilities. For example, Rossi (2006) evaluates whether exchange rates are random walks and finds instabilities in the parameters of interest. Similarly, Rossi (2013) finds instabilities when evaluating predictive models of inflation. Welch and Goyal (2007) and Timmermann (2008) evaluate a wide set of predictive models for US stock returns and document that predictive ability is time-varying.

[2] There is a large literature evaluating out-of-sample forecast performance by means of specification tests; Clark and McCracken (2013) provide an overview. Rossi and Sekhposyan (2016) and Rossi (2020) document that out-of-sample specification tests are affected by instabilities and discuss how to take instabilities into account when evaluating forecasts.

The proposed test is a joint hypothesis test for *both* parameter instability *and* a constant non-zero value of the parameter. In particular, the null hypothesis of the test specifies that the parameter, which can potentially be time-varying, has a zero value *at every point in time* throughout the sample. The test rejects against alternatives in which the parameter has a non-zero value *at some point in time* over the sample. Therefore, the test detects departures from the null hypothesis even when they only occur over short periods of the sample. This makes the test more powerful than traditional hypothesis tests (such as t-tests, Wald or LM tests) which are based on the full sample and fail to reject the null hypothesis if instabilities "average out" over the sample. The joint null hypothesis also distinguishes the test from tests of multiple structural breaks which are designed to detect parameter instability *only* and do not reject against constant alternatives.

The novel test statistic is intuitive and flexible. The test statistic jointly considers all possible splits of the sample at K splitting points into a sequence of consecutive blocks of variable lengths. For each block, a statistic is computed which evaluates whether the data inside each block supports a rejection of the null hypothesis. The test rejects if the combined information from all possible splits supports the alternative hypothesis. This allows the test to achieve high power in the presence of multiple shifts. The test can be constructed based on a set of moment conditions involving the parameters of interest using both a Lagrange-Multiplier (LM) form and a Wald form. The LM form *imposes* the null hypothesis that the parameter is zero at every point in time. In contrast, the Wald form *estimates* the entire parameter vector for each considered block by a partial-sample Generalized Method of Moments (GMM) estimator. The test statistic can be computed for a fixed number of splits or by specifying an upper bound of splits to take into account. Regardless of the number of splits taken into account, the test statistic can be implemented efficiently by a dynamic programming algorithm provided in the paper.

Finally, the test is widely applicable. In particular, the test can be applied to any economic model which is described by a set of moment conditions. Moment conditions can be derived from many reduced form models and structural models such as linear regressions, vector autoregressions, structural equations identified using instrumental variables or even dynamic stochastic general equilibrium models. Alternatively, the test can be used to evaluate out-of-sample forecasting performance by applying it to a moment condition describing a sequence of out-of-sample forecast errors. These forecast errors can be obtained either from a forecasting model whose parameters are estimated using a recursive scheme or from model-free forecasts such as survey or judgmental forecasts. Applications include testing in the presence of instabilities for forecast unbiasedness, rationality, efficiency or forecast encompassing.

This paper makes three contributions to the literature.

First, I provide an instability-robust hypothesis test for a general class of models, explicitly taking multiple discrete shifts in parameters into account. In contrast to structural break tests which test parameter stability only, the procedure *jointly* tests parameter stability and a linear hypothesis on the parameter vector specified by the researcher. Contrary to tests which assume a single break in parameters and only indicate the location of the largest shift, the path of the proposed test statistic can be plotted to provide narrative evidence on which periods of the sample are driving the rejection of the null hypothesis. I derive the limiting distribution of the test statistic which is a function of independent Brownian Motions and tabulate its critical values.

Second, I investigate the finite sample performance of the proposed test across a series of data-generating processes and compare its performance to that of other tests from the literature. The simulations illustrate that asymptotically, traditional hypothesis tests using the full sample and tests of structural change, such as the $UD\max$ test of Bai and Perron (1998), have no power against some of the relevant alternatives. In contrast, the proposed test exhibits significant and monotonic power for these alternatives. The simulations further show that the proposed test is accurately sized across a variety of sample sizes and various forms of serial correlation. Finally, I compare the finite sample power of the proposed procedure to that of traditional specification tests based on the full sample and the QLR_T^* specification test imposing one break by Rossi (2005). I find that the proposed test yields substantially larger finite-sample power if the data-generating process exhibits multiple shifts in parameters. Further, if the data-generating process exhibits one shift or constant coefficients, the power loss compared to existing tests is small. Thus, researchers can use the proposed test without prior knowledge of whether and how parameters vary over time.

Third, I use the proposed test to document the presence of local short-horizon predictability in the U.S. equity premium during the 1946-2019 period using a set of financial variables considered by Welch and Goyal (2007). Recently, various studies have provided theoretical and empirical evidence that predictability is concentrated in short-lived periods, so-called "pockets of predictability" (Timmermann, 2008).[3] This form of predictability is particularly difficult to detect using traditional specification tests (Rossi, 2020) and previous efforts are based on repeated tests in overlapping samples of the data, leading to issues associated with multiple testing (Farmer et al., 2019). In contrast, the test proposed in this paper explicitly takes the search across multiple subsamples into account, thereby avoiding the multiple testing problem. Hence, the test can be used to detect predictability even in the presence of "predictability pockets". I find that one-month-ahead excess market returns are predictable from a larger set of variables than typically found in the literature once mul-

[3] Timmermann (2008) notes that "[...] there appear to be pockets in time where there is modest evidence of local predictability; [...] the identity of the best forecasting method can be expected to vary over time, and there are likely to be periods of model breakdown where no approach seems to work".

tiple shifts in predictability are taken into account. In contrast to traditional predictability tests, the conclusions from the proposed test are invariant to starting the sample after the 1951 Treasury Accord Act. Furthermore, the paths of the test statistics provide evidence of heterogeneity in the location of predictability episodes across predictors. The findings explain why traditional tests often fail to uncover predictability in the full sample and why studies that split the sample at different dates often arrive at conflicting results regarding the predictive ability of a wide class of variables.

Literature. Several papers have proposed specification tests robust to instabilities. However, these focus either on the case of a single break or on a different class of alternatives than the one considered in this paper.

A related method to the one proposed in this paper which also builds on moment conditions and tests a joint hypothesis is developed in Rossi (2005) which considers optimal tests for the case of a single break in parameters. In contrast, the test statistic considered in this paper allows the researcher to consider an unknown number of shifts in parameters up to a specifiable upper bound and nests the case of a single break. This makes the test more powerful in the presence of multiple shifts in parameters while retaining comparable power if the parameter shifts only once. In addition, the path of the proposed test statistic provides narrative evidence on which periods of the sample are driving the rejection of the hypothesis whereas a test imposing one break indicates the location of the largest shift in the parameter vector.

A different strand of the literature designs hypothesis tests which are robust against particular alternatives. These include tests for predictability in threshold models (Gonzalo and Pitarakis, 2012, 2017), tests of relative forecasting performance under Markov-switching alternatives (Odendahl et al., 2020), automated model-selection in the presence of instabilities (Castle et al., 2012) and real-time detection of predictability regimes (Harvey et al., 2020). The advantage of the test which is proposed in this paper is that it remains agnostic about the specific process driving the changes in parameters and thus offers a general approach that can be used if the researcher has no prior knowledge on whether and how parameters vary over time.

Finally, there is a large literature on testing for multiple structural changes. A seminal contribution is Bai and Perron (1998) who proposed sup F tests in a class of linear regression models. Structural break tests of multiple changes have since been extended to more general classes of models (Sowell, 1996; Perron and Qu, 2006; Qu and Perron, 2007; Elliott and Müller, 2006). A related literature provides structural change tests for predictive regression models; Pitarakis (2017) and Georgiev et al. (2018) are two examples of recent contributions. In contrast to the approach presented in this paper which tests a joint hypothesis, structural change tests focus on the null hypothesis of parameter stability only and therefore do not

have power against some of the alternatives considered in this paper.

OUTLINE. Section 2 discusses the hypotheses of interest and proposes the test statistic. Sections 3 and 4 discuss the relevant asymptotic theory to conduct in-sample and out-of-sample inference, respectively. Section 5 provides a guide for implementing the test. Section 6 explores the finite sample performance of the proposed test by means of extensive Monte Carlo simulations. Section 7 applies the test to study the predictability of the U.S. equity premium robust to instabilities. Section 8 concludes.

2 Specification Tests Robust to Multiple Instabilities

This section formalizes the testing problem of parameter inclusion under instabilities and introduces the test statistic.

2.1 Model and Hypotheses

Consider a model indexed by a v-dimensional parameter vector θ_t for $t = 1, \ldots, T$. Assume the parameter vector partitions $\theta_t = (\beta_t', \delta)$ where β_t is $(p \times 1)$ and δ is $(q \times 1)$. Further, assume the model satisfies the following m-dimensional moment condition

$$\mathbb{E}\left[f(z_t, \beta_t, \delta)\right] = 0 \tag{1}$$

where z_t is an r-dimensional random vector of data and $f : \mathbb{R}^r \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^m$.

Based on the moment condition in (1), the researcher wants to test whether the sequence of *possibly* time-varying parameters β_t should be included when modeling the data. The relevant hypotheses are

$$H_0 : \ \beta_t = 0 \quad \forall\, t \qquad \text{vs.} \qquad H_A : \ \beta_t \neq 0 \quad \text{for some} \ \ t \geq 1 \tag{2}$$

Under the null hypothesis β_t is zero *at any point in the sample* and thus can be excluded from the model. Under the alternative, β_t is different from zero *at some point in the sample* and should be included in the model.[4] The test of parameter inclusion is thus required to be sensitive to situations in which β_t departs from zero only during short periods over the sample.

It will be useful to express the null hypothesis as an intersection. Let β_0 denote the true parameter value under the null hypothesis that β_t is constant in which case $\theta_0 = (\beta_0, \delta_0)$.

[4]Note that by redefining β_t one can similarly test any linear hypothesis on the parameter.

Then, we can express the null hypothesis in (2) as

$$H_0 : \beta_t \in B^1 \cap B^2 \text{ with } B^1 \equiv \left\{\beta_t \in \mathbb{R}^p : \beta_t = \beta_0 \; \forall \; t\right\}, \; B^2 \equiv \left\{\beta_0 \in \mathbb{R}^p : \beta_0 = 0\right\} \tag{3}$$

Denote the first part of this null hypothesis imposing a constant parameter vector by $H_0^{(1)} : \beta_t \in B^1$ and the second part imposing a zero value by $H_0^{(2)} : \beta_t \in B^2$.

It is important to note that the hypotheses in (2) are different from those of a test assuming constant parameters. In particular, if the researcher misspecifies the moment condition in (1) by assuming that θ is constant and tests the hypothesis $H_0 : \beta = 0$ against the alternative $H_A : \beta \neq 0$, the test will reject if the constant parameter β is different from zero rather than β_t being zero. Depending on the form of time-variation in β_t, it can be the case that $\beta = 0$ but that $\beta_t \neq 0$ for some t. In that case instabilities "average out" over the sample and a test of the hypotheses in (2) will reject while a test assuming constant parameters will not.[5]

One might be tempted to think that a test of the hypotheses in (2) is simply a test of parameter stability. However, as (3) shows, the relevant null hypothesis is a *joint hypothesis* which simultaneously imposes (i) a constant coefficient vector and (ii) a zero value on the constant coefficient vector. In contrast, the null hypothesis of a test for parameter stability only focuses on the first part, that is $H_0^{(1)}$. This implies that there are data-generating processes for which a test of parameter stability does not reject while the test considered in this paper does. In particular, this would be the case when the true parameter vector is constant, but different from zero.

The test statistic proposed in this paper detects violations of the null hypothesis in equation (2) by combining information obtained from partitioning the sample into a series of discrete consecutive blocks. In particular, the test searches over all possible combinations of $K+1$ discrete segments and evaluates whether there is evidence to reject the null hypothesis. While the test does not impose an alternative of a particular form, it has particularly high power when the parameter undergoes a series of discrete changes. Specifically, assume that β_t has a "baseline value" of β_A but that it undergoes a series of K discrete changes where the value of β_t departs from β_A. Collect these changepoints expressed as a fraction of the sample in a K-dimensional vector $\lambda^K := (\lambda_1, \ldots, \lambda_K)$ where $\lambda_j \in (0, 1)$ and $\lambda_j > \lambda_{j-1}$. Further, collect the magnitudes of these changes in a K-dimensional vector β_Δ. Under this assumption, the alternative hypothesis in (2) can be expressed as

$$H_A : \beta_t = \beta_A \; + \; \sum_{j=1}^{K} \mathbb{1}\left([\lambda_j T] < t \leq [\lambda_{j+1} T]\right) \cdot \beta_{\Delta,j} \qquad t = 1, \ldots, T \tag{4}$$

where $\beta_{\Delta,j}$ denotes the j-th element of β_Δ, $\lambda_0 \equiv 0$, $\lambda_{K+1} \equiv 1$ and $[\cdot]$ is the integer part

[5] For a detailed explanation of this argument see the discussion in Rossi (2005) and Rossi (2013).

operator.[6] The proposed test statistic searches across all possible values of λ^K to best approximate the path of β_t and evaluates for each candidate λ^K whether the null hypothesis $\beta_t = 0 \; \forall t$ can be rejected. If the path of β_t takes the form described in (4), there is a value of K and λ^K for which the approximation will be exact. This makes the test particularly powerful in the presence of discrete changes.

Tests of the hypothesis considered above have many applications in empirical work. At the end of this section, I provide various examples of problems which have been studied in the macroeconomic and financial literature and illustrate how they fit into the testing framework.

2.2 Test Statistics

This section introduces a class of test statistics which can be used to test the null hypothesis defined in (2). As previewed before, the main idea of the tests is simple: To detect departures from the null hypothesis, the test statistic jointly considers all possible splits of the sample at K splitting points into a sequence of $K+1$ consecutive blocks of variable length. For each block, a statistic is computed which evaluates whether the data inside each block supports a rejection of the null hypothesis. If there is a sample split for which the sum of statistics computed on each block supports the alternative, the test rejects. In what follows, I first describe how to construct the test statistic for a fixed number of sample splits K. Consecutively, I discuss how to robustify the test against the choice of K.

Test with a fixed number of splits K

Assume the researcher wants to test parameter inclusion robust to time-variation according to the hypotheses in (2) on a sample $t = T_0, \dots, T$ where T_0 denotes the first observation and is typically set to 1. Let λ^K denote a K-dimensional vector of splitting points $\lambda^K := (\lambda_1, \dots, \lambda_K)$ where $\lambda_j \in (0,1)$ and $\lambda_j > \lambda_{j-1}$. Each value of λ^K implies a different partition of the sample into a sequence of $K+1$ consecutive blocks where block j spans data from $t = [\lambda_{j-1}T] + 1, \dots, [\lambda_j T]$ with $T_0 \equiv [\lambda_0 T] + 1$ and $\lambda_{K+1} \equiv 1$.

For simplicity of exposition, I first abstract from the choice of the number of splits and assume that K is a known value. The proposed test statistic for testing the null hypothesis in (2) using K sample splits takes the following form.

$$\begin{aligned} \sup \Phi_T(K) := \sup_{\lambda^K \in \Lambda_\epsilon} \; & \sum_{j=1}^{K+1} \Phi_{T,j}(\lambda_{j-1}, \lambda_j) \\ \Lambda_\epsilon \equiv \Big\{ \lambda_j : \; & \lambda_j \in (\lambda_0 + \epsilon, \lambda_{K+1} - \epsilon), \; \lambda_j > \lambda_{j-1} + \epsilon, \;\; j = 1, \dots, K \Big\} \end{aligned} \tag{5}$$

[6] For example, when $\beta_t = 0$ for $t = 1, \dots, [T/2]$ and $\beta_t = \beta_\Delta$ for $t = [T/2]+1, \dots, T$, then $\beta_A = 0$, $\lambda_1 = 1/2$, $K = 1$ and β_Δ is a scalar. In contrast, when β_t is constant at β_A for $t = 1, \dots, T$, $K = 0$.

For a given sample split λ^K, the test statistic is simply the sum of $K+1$ statistics $\Phi_{T,j}(\lambda_{j-1}, \lambda_j)$ computed on each block of the data. The $\sup_{\lambda^K}$ part of the test statistic searches over all possible combinations of K splitting points for the choice of λ^K which maximizes this sum. The value at the optimal choice of λ^K which yields the maximum value for the sum term is the final value of the test statistic.

Note that the search of λ^K is restricted to a set Λ_ϵ defined by a trimming parameter $\epsilon \in (0, 1)$ which imposes that each of the blocks contains at least $[\epsilon T]$ observations. This parameter is set by the researcher prior to conducting the test and its choice depends on the stochastic properties of the data.[7] The simulations presented in Section 6 provide guidance on the trade-offs of choosing a lower or higher value of ϵ. In most applications, a choice of $\epsilon = 0.05$ or $\epsilon = 0.1$ is sufficient.

Choice for $\Phi_{T,j}(\cdot, \cdot)$

The test statistic above crucially depends on the choice for $\Phi_{T,j}(\cdot, \cdot)$. In this paper, I consider two forms, a Lagrange-Multiplier (LM) statistic and a Wald statistic. The LM form imposes the value of β_t to be zero in each block while the Wald form estimates β_t in each block. In both cases, the test statistic builds on partial sums of the moment condition defined in equation (1), evaluated at estimates of δ.

The tests proposed in this paper can be conducted based on moment conditions formulated in-sample or out-of-sample. These two cases differ in the portion of the sample on which the test statistic is constructed as well as the estimation scheme which is used to estimate δ. In the in-sample case, the test is conducted on the full sample setting $T_0 = 1$ and δ is estimated based on the moment condition in (1). The resulting estimate is denoted $\hat{\delta}$. In contrast, in the out-of-sample case the test is conducted on an out-of-sample portion of the data setting $T_0 = R$ where $R \gg 1$. Here, δ is the parameter of a forecasting model identified by a separate moment condition which is estimated using a recursive scheme on an in-sample portion of the data, yielding a sequence of estimators of δ denoted $\{\hat{\delta}_t\}_{t=R}^T$. The in-sample case is discussed in more detail in Section 3 while the out-of-sample case is discussed in Section 4 of the paper.

Lagrange-Multiplier Form. If the test is implemented using the Lagrange-Multiplier

[7] The use of trimming parameters is standard in tests for structural breaks, see e.g. Andrews (1993) or Bai and Perron (1998).

form, the test statistic is constructed using the following choice for $\Phi_{T,j}(\cdot,\cdot)$.

$$\begin{aligned}\Phi^{LM}_{T,j}(\lambda_{j-1},\lambda_j) &:= \hat{\mathcal{F}}'_{T,j} \times \hat{\Omega}_{T,j} \times \hat{\mathcal{F}}_{T,j} \\ \hat{\mathcal{F}}_{T,j} &:= (T-T_0+1)^{-1/2}\ \hat{\Sigma}_{ff}^{-1/2} \sum_{t=[\lambda_{j-1}T]+1}^{[\lambda_j T]} f(z_t,\tilde{\theta}_t)\end{aligned} \tag{6}$$

Note that $f(\cdot,\cdot)$ is the moment function which was defined in equation (1), $\hat{\Sigma}_{ff}$ is a consistent estimator of the long-run variance of the sample moments under the null hypothesis and $\hat{\Omega}_{T,j}$ is a consistent estimator of the long-run variance of $\hat{\mathcal{F}}_{T,j}$. Formulas to compute these estimators are given in Section 5 of this paper. $\tilde{\theta}_t$ is a restricted generalized method of moments (GMM) estimator of θ_t that imposes the joint null hypothesis defined in (2) which restricts $\beta_t = \beta_0 = 0\ \forall\ t$ while leaving δ unspecified.

The difference between testing in-sample and out-of-sample using the statistic above lies in how the estimate of δ, and consequently $\tilde{\theta}_t$ is obtained. In the out-of-sample case, the restricted estimator is formed as $\tilde{\theta}_t := (0_{p\times 1}, \hat{\delta}_t)$ where $\{\hat{\delta}_t\}_{t=T_0}^{T}$ is a sequence of estimates of δ which is obtained via a recursive estimation scheme. Section 4 discusses in detail how to obtain these estimates. In contrast, in the in-sample case $\tilde{\theta}_t = \tilde{\theta}\ \forall t$ where $\tilde{\theta}$ is a constant GMM estimator which is defined as follows.

$$\begin{aligned}\tilde{\theta} &:= \arg\max_{\theta\,\in\,\Theta}\ \hat{Q}_T(\theta) \qquad\qquad \hat{Q}_T(\theta) \equiv \hat{F}_T(\theta)'\ W_T\ \hat{F}_T(\theta) \\ \hat{F}_T(\theta) &\equiv (T-T_0+1)^{-1} \sum_{t=T_0}^{T} f(z_t,\theta) \\ \text{subject to}\quad & A\,\theta = 0_{p\times 1} \qquad\qquad A = \begin{bmatrix} I_{p\times p} & 0_{p\times q}\end{bmatrix}\end{aligned} \tag{7}$$

$\hat{F}_T(\theta)$ is the sample analogue of the moment condition defined in (1) and W_T is a positive definite weighting matrix.

Wald Form. If the test is implemented using the Wald form, the test statistic is constructed using the following choice for $\Phi_{T,j}(\cdot,\cdot)$.

$$\Phi^{W}_{T,j}(\lambda_{j-1},\lambda_j) := (T-T_0+1)\ \left[\ \hat{\beta}_j(\lambda_{j-1},\lambda_j)\ \right]'\ \times\ \hat{\Omega}^{-1}_{T,j}\ \times\ \left[\ \hat{\beta}_j(\lambda_{j-1},\lambda_j)\ \right] \tag{8}$$

The difference between conducting the test in-sample and out-of-sample lies in how the estimate of $\hat{\beta}_j(\cdot,\cdot)$ is obtained. In the in-sample case, $\hat{\beta}_j(\lambda_{j-1},\lambda_j) := A\ \hat{\theta}_j(\lambda_{j-1},\lambda_j)$ where $A \equiv \begin{bmatrix} I_{p\times p} & 0_{p\times q}\end{bmatrix}$ and $\hat{\theta}_j(\lambda_{j-1},\lambda_j)$ is defined as the following GMM estimator which assumes

the parameter θ has a constant value in block j.

$$\begin{aligned} \hat{\theta}_j &:= \arg\max_{\theta \in \Theta} \hat{Q}_{T,j}(\theta) \qquad \hat{Q}_{T,j}(\theta) \equiv \hat{F}_{T,j}(\theta)' \, W_T \, \hat{F}_{T,j}(\theta) \\ \hat{F}_{T,j}(\theta) &\equiv \big([\lambda_j T] - [\lambda_{j-1} T]\big)^{-1} \sum_{t=[\lambda_{j-1}T]+1}^{[\lambda_j T]} f(z_t, \theta) \end{aligned} \tag{9}$$

where $\hat{F}$ is a partial-sample analogue of the moment condition defined in (1) and W_T is a positive definite weighting matrix.
In contrast, in the out-of-sample case $\hat{\beta}_j$ is obtained using a similar GMM estimator.

$$\begin{aligned} \hat{\beta}_j &:= \arg\max_{\beta \in B} \hat{Q}_{T,j}(\beta) \qquad \hat{Q}_{T,j}(\beta) \equiv \hat{F}_{T,j}(\beta)' \, W_T \, \hat{F}_{T,j}(\beta) \\ \hat{F}_{T,j}(\beta) &\equiv \big([\lambda_j T] - [\lambda_{j-1} T]\big)^{-1} \sum_{t=[\lambda_{j-1}T]+1}^{[\lambda_j T]} f(z_t, \beta, \hat{\delta}_t) \end{aligned} \tag{10}$$

In contrast to the in-sample case, the estimator does not define an estimate of δ, but evaluates the partial-sample moment at the given sequence of parameter estimates, $\{\hat{\delta}_t\}_{t=T_0}^{T}$.

In a given application, both the LM and the Wald form of $\Phi_{T,j}(\cdot,\cdot)$ can be used to construct the $\sup \Phi_T(K)$ test statistic defined above. However, their performance may differ in finite samples depending on the application considered. Generally, the LM form of the test statistic is computationally efficient as it only requires computing an estimate of θ under the null hypothesis. In contrast, the Wald form of the test statistic requires re-estimating θ for each block and every sample partition considered.[8] Because of its general computational benefits, the rest of the paper will mainly focus on the LM form of the test statistic.

To implement the LM and Wald statistics, one requires the variance estimators $\hat{\Sigma}_{ff}$ and $\hat{\Omega}_{T,j}$. Formulas to compute these estimators are given in Section 5. Computation of the $\sup_\lambda^K$ operator in (5) can be achieved efficiently by means of a dynamic programming algorithm which is also provided in Section 5. The algorithm computes the LM form in $\mathcal{O}(T^2)$ operations computing the sum term in equation (5).

Test for an unknown number of splits

The discussion in the previous section has made a simplifying assumption, namely that the researcher wants to conduct the test with a specific number of splits K in mind. However, in practice it is unlikely that the researcher has prior information about the appropriate choice of K. To circumvent this problem, this section presents a robustified version of the test statistic which abstracts from the choice of K by combining information from computing

[8] For linear models, this estimator can be implemented efficiently by a recursion.

the test statistic for different values for K, starting at one and stopping at some pre-defined ceiling, $\bar{K}$.[9] The proposed test statistic for testing the joint null hypothesis in (2) considering up to $\bar{K}$ splits of the sample has the following form.

$$D\sup\Phi_T(\bar{K}) := \max_{1\leq k\leq \bar{K}} \left\{ \sup\Phi_T(k)/k \right\} \tag{11}$$

where $\sup\Phi_T$ is the test statistic defined in equation (5).

To compute this test statistic, one computes the $\sup\Phi_T(k)$ statistic for different choices of $k = 1, \ldots, \bar{K}$. The resulting values are then weighted by the number of shifts used to compute them. The maximal value of this reweighted series of test statistics is the final value of the test statistic.

In practice, a choice of $\bar{K}$ as low as five or ten is often sufficient in applications. In fact, simulations show that the choice of $\bar{K}$ has little impact on size and power of the test beyond these values.[10] In general, the admissible values of $\bar{K}$ are bounded above by the choice of the trimming parameter ϵ which is used to compute the $\sup\Phi_T$ statistic since it imposes a minimum number of observations for each segment and therefore implicitly defines an upper bound on K. For instance when $\epsilon = .1$, the maximum number of admissible splits which can be considered is $\bar{K} = 10$.

2.3 Examples

Tests which are robust to heterogeneity in parameters over time have many applications in empirical work. In the following, I provide two examples of testing problems which have been widely studied in the macroeconomic and financial literature and illustrate how they fit into the testing framework.

Predictive Regressions. Consider the following model

$$y_{t+h} = \delta + X_t'\,\beta + \eta_{t+h} \qquad\qquad t = 1, \ldots, T \tag{12}$$

where y_{t+h} is a scalar series to be predicted in-sample at horizon h, X_t is a $(p \times 1)$ vector of predictors which are suspected to have a time-varying relationship with y_{t+h}, δ is the constant of the regression and η_{t+h} is a sequence of unforecastable errors. A large literature in macroeconomics and finance studies tests of the hypothesis $H_0 : \beta = 0$ in the model above e.g. to determine the predictive ability of financial variables (Pitarakis and Gonzalo, 2019) or to evaluate model specifications for variables such as inflation (Rossi, 2005). However, in

[9] A similar method to robustify tests to the choice of K was first proposed in the context of structural break tests in Bai and Perron (1998).

[10] The simulation results are available on request.

recent years many studies have documented that predictive ability in these models, which is captured by β, is time-varying and that tests based on the full sample may fail to reject in the presence of predictability (Welch and Goyal, 2007; Timmermann, 2008; Rossi, 2013).

To implement the test robust to instabilities in the parameters, replace β in equation (12) by β_t and note that the resulting model implies the following moment condition.

$$\mathbb{E}\left[f(z_t, \beta_t, \delta)\right] = 0 \qquad f(z_t, \beta_t, \delta) \equiv \begin{bmatrix} X_t \cdot (y_{t+h} - \delta - X_t'\beta_t) \\ 1 \cdot (y_{t+h} - \delta - X_t'\beta_t) \end{bmatrix}$$

where $z_t = (y_{t+h}, X_t')'$ and $\theta_t = (\beta_t', \delta)'$.

To conduct the $D\sup\Phi_T(K)$ test based on the LM statistic, compute the restricted GMM estimator $\tilde{\theta}$ via the estimator in (7) which imposes the null hypothesis $\beta_t = 0\ \forall t$. Set $T_0 = 1$ and use the variance estimator in (15) or (17) to compute $\hat{\Sigma}_{ff}$ and use (18) to compute $\hat{\Omega}_{T,j}$. Then, compute $D\sup\Phi_T(\bar{K})$ defined in (11) by means of the dynamic programming algorithm described in Section 5. Reject the null hypothesis of no predictive ability if the computed value of the test statistic is larger than the appropriate critical value reported in Section 3.

Out-Of-Sample Predictive Ability. West and McCracken (1998) proposed a framework to evaluate out-of-sample predictive ability by testing the null hypothesis $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ in the following linear model

$$v_{t+h}(\hat{\delta}_t) = \hat{\xi}(z_t, \hat{\delta}_t)'\ \beta + \eta_{t+h}, \qquad t = R, \ldots, T \tag{13}$$

where $v_{t+h}(\hat{\delta}_t)$ is a sequence of forecast errors derived from a parametric forecasting model which depend on the sequence of estimated parameters of the forecasting model $\hat{\delta}_t$. By appropriately choosing the function $\hat{\xi}(z_t, \hat{\delta}_t)$, the framework includes popular test for forecast evaluation such as tests of forecast unbiasedness, forecast rationality or forecast encompassing. Many studies provide empirical evidence that out-of-sample predictive ability is time-varying e.g. for forecast rationality of private sector forecasts or for forecast encompassing tests evaluating the information-advantage of Federal Reserve forecasts (Rossi and Sekhposyan, 2016; Hoesch et al., 2020).

To implement the proposed test robust to heterogeneity in β, note that for time-varying β_t, the model above implies the following moment condition.

$$\mathbb{E}\left[f(z_t, \beta_t, \hat{\delta}_t)\right] = 0 \qquad f(z_t, \beta_t, \hat{\delta}_t) \equiv \hat{\xi}(z_t, \hat{\delta}_t) \cdot \left(v_{t+h}(\hat{\delta}_t) - \hat{\xi}(z_t, \hat{\delta}_t)'\beta_t\right) \tag{14}$$

where $z_t = v_{t+h}(\hat{\delta}_t)$.

For the purpose of this example, focus on a test of forecast unbiasedness which sets

$\hat{\xi}(z_t, \hat{\delta}_t) = 1$. West and McCracken (1998) showed that in this case parameter estimation error in δ_t is asymptotically irrelevant (for more details see Section 4) so that we can use the simple formulas for the variance estimators reported above. To conduct the $D\sup\Phi_T(K)$ test based on the LM statistic, set $\tilde{\theta}_t = (0'_{p\times 1}, \hat{\delta}_t)$ and $T_0 = R$ and use the variance estimator in (15) or (17) to compute $\hat{\Sigma}_{ff}$ and use (18) to compute $\hat{\Omega}_{T,j}$. Then, compute $D\sup\Phi_T(\bar{K})$ defined in (11) by means of the dynamic programming algorithm described in Section 5. The null hypothesis of forecast unbiasedness can be rejected if the computed value of the test statistic is larger than the appropriate critical value reported in Section 4.

3 In-Sample Inference

This section describes the relevant asymptotic theory to conduct in-sample inference. In the in-sample case, the $\sup\Phi_T(K)$ test statistic defined in equation (5) and the $D\sup\Phi_T(\bar{K})$ test statistic in equation (11) are constructed setting $T_0 = 1$ and evaluating $\Phi_T(\cdot,\cdot)$ at the relevant GMM estimators defined in equations (7) and (9) for the LM and Wald case, respectively. This section presents and discusses a set of regularity assumptions which are sufficient to obtain weak convergence of the test statistics under the null hypothesis to a function of Brownian Motions. This result is established in the main theorem of this section.

NOTATION. I introduce some notational conventions that are required for this section and used throughout the rest of the paper. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space on which all of the random elements are defined. Unless specified otherwise, all limits are taken as the sample size $T \to \infty$. The symbol $\xrightarrow{p}$ denotes convergence in probability and $\xrightarrow{d}$ denotes convergence in distribution. Next, $\Rightarrow$ denotes weak convergence for sequences of measurable random elements of a space of bounded Euclidean-valued cadlag functions on the product space $D[0,1]^m$ as defined in Phillips and Durlauf (1986) where each component space $D[0,1]$ is equipped with the Skorohod metric. $\|\cdot\|$ denotes the Euclidean norm of a vector or matrix.

The following regularity assumptions are sufficient to obtain weak convergence of the test statistics under the null hypothesis to the limiting distribution characterized in the theorem below.

ASSUMPTION 3.1 (Regularity conditions): *Assume the following regularity conditions hold.*

(i) $\{z_t\}$ *is strong mixing with strong mixing coefficients* $\{\alpha(n)\}, \sum_{n=1}^{\infty}\alpha(n)^{1-2/\gamma} < \infty$ *with* $\gamma > 2$.

(ii) $\{z_t\}$ *is weakly stationary. In addition* $\mathbb{E}[f(z_t,\theta_0)] = 0$ *for all* $t = 1,\ldots,T$ *and* $T = 1,2,\ldots$ *and the individual elements of* $f(z_t,\theta_0)$ *have the finite absolute moment* $\mathbb{E}\left[|f^{(i)}(z_t,\theta_0)|^{\gamma}\right] < \infty$ *for* $i = 1,\ldots,m$ *and* $\gamma > 2$.

(iii) $\Sigma_{ff} \equiv \lim_{T\to\infty} \mathbb{E}\left[T^{-1}\left\{\sum_{t=1}^{T} f(z_t,\theta_0)\right\}\left\{\sum_{t=1}^{T} f(z_t,\theta_0)\right\}'\right] \in \mathbb{R}^{m\times m}$ *is positive definite.*

(iv) $f(z,\theta)$ *is continuously partially differentiable in* θ *in a neighborhood of* θ_0 *for every* $\theta_0 \in \Theta^*$ *where* Θ^* *is some convex or open set that contains* Θ*. The functions* $f(z,\theta)$ *and* $\nabla_\theta f(z,\theta) \equiv \partial f(z,\theta)/\partial\theta$ *are measurable functions of* z *for each* $\theta \in \Theta$ *and* $\mathbb{E}\left[\sup_{\theta\in\Theta^*}\|\nabla_\theta f(z_t,\theta)\|\right] < \infty$*.* $\mathbb{E}[f(z_t,\theta_0)'f(z_t,\theta_0)] < \infty$*, and* $\sup_{\theta\in\Theta}\|f(z_t,\theta)\| < \infty$ *for all* $t = 1,\ldots,T$ *and* $T = 1,2,\ldots$*. Each element of* $f(z_t,\theta_0)$ *is uniformly square integrable, for all* $t = 1,\ldots,T$ *and* $T = 1,2,\ldots$*.*

(v) The parameter space Θ *is a compact subset of* $\mathbb{R}^v$*.*

(vi) $\lim_{T\to\infty}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f(z_t,\theta)\right] = 0$*, only when* $\theta = \theta_0$

(vii) The sequence of positive definite weighting matrices $W_T \to_p \Sigma_{ff}^{-1}$*.*

(viii) $M \equiv \lim_{T\to\infty}\mathbb{E}\left[T^{-1}\sum_{t=1}^{T}\frac{\partial f(z_t,\theta)}{\partial\theta'}|_{\theta=\theta_0}\right] \in \mathbb{R}^{m\times v}$ *has full column rank.*

I now discuss Assumption 3.1. Assumptions 3.1.(i) and 3.1.(ii) are asymptotic weak dependence and stationarity conditions on the data which are typical of those found in other literature on nonlinear dynamic models and are closest to the conditions given in Sowell (1996) or Rossi (2005).[11]. Together with Assumption 3.1.(iii) which assumes positive definiteness of the long-run variance of the sample moments, the assumptions are sufficient to obtain weak convergence of the partial sample moments to Brownian Motions using the multivariate functional central limit theorem of Phillips and Durlauf (1986). Assumption 3.1.(iv) are standard smooothness and boundedness conditions on the sample moment function under the null hypothesis $f(z,\theta)$. An analogue of this assumption is used in Sowell (1996). Together with Assumption 3.1.(v) which assumes a compact parameter space and Assumption 3.1.(i), the conditions ensure uniform convergence of the GMM objective function via the generic uniform law of large numbers of Andrews (1987). Together with Assumption 3.1.(vi) which assumes identification under the null hypothesis, the conditions are sufficient to obtain consistency of the GMM estimators used to construct the test statistic. Assumption 3.1.(vii) restricts the choice of weighting matrices used to construct the GMM estimators by requiring that an efficient GMM estimator is used. Finally, Assumption 3.1.(viii) which requires the gradient of the sample moment to have full rank ensures that the test statistic has a well-defined asymptotic variance.

The following theorem establishes the asymptotic distribution of the test statistic under the null hypothesis.

THEOREM 3.1 (Limiting distribution for in-sample tests): *Assume that the regularity con-*

[11] The assumptions, in particular the weak stationarity condition, are stronger than necessary and the results presented in this paper are expected to hold if the assumptions are relaxed to the near-epoch-dependence case in Andrews (1993).

ditions in Assumption 3.1 hold. Under the null hypothesis defined in (3), it holds that

$$\sup \Phi_T(K) \;\Rightarrow\; \sup_{\lambda^K \in \Lambda_\epsilon} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j) - \mathcal{B}_p(\lambda_{j-1})\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

$$D\sup \Phi_T(\bar{K}) \Rightarrow \max_{1\le k\le \bar{K}} (1/k) \sup_{\lambda^K \in \Lambda_\epsilon} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j) - \mathcal{B}_p(\lambda_{j-1})\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

$$\Lambda_\epsilon \equiv \Big\{ \lambda_j : \;\; \lambda_j \in (\epsilon, 1-\epsilon),\; \lambda_j > \lambda_{j-1} + \epsilon,\;\; j = 1, \ldots, K \Big\}$$

where $\lambda_0 \equiv 0, \lambda_{K+1} \equiv 1$ *and* $\mathcal{B}_p(\cdot)$ *is a* $(p \times 1)$ *vector of independent standard Brownian motions on* $[0, 1]$.

The proof of this theorem is reported in Appendix C.

One can show that the limiting distribution of $\sup \Phi_T(K)$ is equivalent to

$$\sup_{\lambda \in \Lambda_\epsilon} \sum_{i=1}^{K} \frac{\|\lambda_{i+1}\, \mathcal{B}_p(\lambda_i) - \lambda_i\, \mathcal{B}_p(\lambda_{i+1})\|^2}{\lambda_i \lambda_{i+1}(\lambda_{i+1} - \lambda_i)} \;+\; \mathcal{B}_p(1)'\, \mathcal{B}_p(1)$$

where the first term is the same as $(1/Kp)$ times the limiting distribution of the sup F test for parameter stability of Bai and Perron (1998) and depends on the number of splitting points K. The second component reflects the additional restrictions on β and does not depend on K. It is equivalent to the χ^2 distribution with p degrees of freedom which is the limting distribution of a standard LM test conducted on the full sample. Further, note that for $K = 1$, the limiting distribution reduces to the limiting distribution of the QLR^*_T model selection test robust to instabilities which was proposed in Rossi (2005) and imposes one break.

Critical values of the test statistics can be obtained by directly simulating the limiting distributions in Theorem 3.1 using a dynamic programming algorithm analog to the one provided in Section 5. Table 1 reports a selection of critical values for the $D\sup \Phi_t(K)$ test for commonly used significance levels and trimming parameters for $p = 1, 2$. The critical values were obtained by simulating the asymptotic distributions based on 10,000 Monte Carlo replicatons and an approximation length of $N = 3600$ for the Brownian Motions. Details and a full tabulation of the critical values for a wide array of values for p, ϵ and significance levels, α, are provided in Appendix A.

4 Out-Of-Sample Inference

This section describes the relevant asymptotic theory to conduct out-of-sample inference. The limiting distributions derived in this section apply when the test statistic proposed

Table 1: Selected critical values for $D\sup\Phi_T$ tests

	$\epsilon = .05$			$\epsilon = .1$			$\epsilon = .15$		
p	10%	5%	1%	10%	5%	1%	10%	5%	1%
1	10.12	11.57	15.20	9.39	10.99	14.54	8.84	10.48	14.17
2	14.01	15.79	19.96	13.30	15.06	19.20	12.80	14.65	18.62

This table reports simulated quantiles of the limiting distributions of the $D\sup\Phi_T$ tests. The critical values were obtained based on 10,000 Monte-Carlo replications and an approximation length of $N = 3600$ observations for the partial sums to simulate the Brownian Motions. Appendix A provides the full table.

in Section 2 are used in conjunction with a moment condition formulated on out-of-sample forecast errors. I start by discussing the forecasting environment, in particular how to obtain the sequence of estimates $\{\hat{\delta}_t\}_{t=T_0}^T$ which are used to construct the test statistics proposed in Section 2. I then present and discuss the required regularity conditions and derive the limiting distribution of the proposed test statistics.

Assume the available sample is of size $T + h$ and that the data $z_t = (y'_{t+h}, x'_t)$ includes a random variable y_t to be predicted h steps ahead as well as a vector of predictors, x_t. The sample is divided into an in-sample portion of length R and an out-of-sample portion of size P such that $R + P = T + h$. Given the sample split, forecasts of y_{t+h} for $t = R, \ldots, T$ are generated using parametric models of the form $y_{t+h} = g(x_t, \delta) + u_{t+h}$ for a known function $g(\cdot, \cdot)$ and an unknown q-dimensional parameter vector δ.

The parameters of the forecasting model are estimated based on a d-dimensional vector of moment equations $\mathbb{E}[h(z_t, \delta)] = 0$. This allows for a variety of estimation methods such as (nonlinear) least squares, maximum likelihood or generalized methods of moments. The parameters are of the forecasting model are assumed to be estimated using a recursive scheme where the parameter vector is estimated at each $t = R, \ldots, T$ using all available information which yields a sequence of parameter estimates $\{\hat{\delta}_t\}_{t=R}^T$. The predictors and parameter estimates are used to generate forecasts $\hat{y}_{t+h} = \hat{g}_{t+h}(x_t, \hat{\delta}_t)$ for $t = R, \ldots, T$. which are used in turn to construct a series of forecast errors, $\hat{v}_{t+h} = y_{t+h} - \hat{y}_{t+h}$. Figure 1 illustrates the out-of-sample forecasting environment described above.

The following regularity assumptions are sufficient to obtain weak convergence of the test statistics under the null hypothesis to the limiting distribution characterized in the main theorem of this section.
Let $\xi(z_{t+h}, \theta) \equiv \left[f(z_{t+h}, \beta, \delta)', h_t(\delta)'\right]'$ be an $(m + d \times 1)$ vector stacking the moment functions. Further, let $F \equiv \mathbb{E}\left[\frac{\partial f(z_t,\theta)}{\partial \delta}|_{\theta=\theta_0}\right] \in \mathbb{R}^{m\times q}$. The following regularity conditions are assumed to hold under the null hypothesis.

ASSUMPTION 4.1 (Regularity conditions): *Assume the following regularity conditions hold.*

(i) *Assume that* $h < \infty$ *and that* K *is fixed while* $R \to \infty$, $T \to \infty$ *and* $\lim_{T\to\infty} R/T = \rho \in (0,1)$.

(ii) *The estimate* $\hat{\delta}_t$ *satisfies* $\hat{\delta}_t - \delta_0 = B_t H_t$ *where* B_t *is a* $(q \times d)$ *matrix which satisfies* $B_t \xrightarrow{as} B$ *where* B *has rank* q *and* H_t *is* $(d \times 1)$ *with* $H_t = t^{-1}\sum_{r=1}^{t} h(z_r, \delta_0)$ *(recursive estimation scheme) for a* $(d \times 1)$ *moment condition* $h(z_r, \delta)$.

(iii) *For some* $p > \beta > 2$, $\xi(z_{t+h}, \theta_0)$ *is zero mean, strong mixing with mixing coefficients* α_m *of size* $-p\beta/(p-\beta)$ *and it holds that* $\sup_{t\geq 1}\|\xi(z_{t+h}, \theta_0)\|_p = C < \infty$.

(iv) $\{z_{t+h}\}$ *is weakly stationary. In addition,* $\mathbb{E}\left[\xi(z_{t+h}, \theta_0)\right] = 0$ *for all* $t = 1, \ldots, T$ *and* $T = 1, 2, \ldots$.

(v) $\Sigma \equiv \lim_{T\to\infty} \mathbb{E}\left[T^{-1}\left\{\sum_{t=1}^{T}\xi(z_{t+h}, \theta_0)\right\}\left\{\sum_{t=1}^{T}\xi(z_{t+h}, \theta_0)\right\}'\right] \in \mathbb{R}^{(m+d)\times(m+d)}$ *is positive definite.*

(vi) $f(z, \theta)$ *is continuously partially differentiable in* θ *in a neighborhood of* θ_0 *for every* $\theta_0 \in \Theta^*$ *where* Θ^* *is some convex or open set that contains* Θ. *The functions* $f(z, \theta)$ *and* $\nabla_\theta f(z, \theta) \equiv \partial f(z, \theta)/\partial\theta$ *are measurable functions of* z *for each* $\theta \in \Theta$ *and* $\mathbb{E}\left[\sup_{\theta\in\Theta^*}\|\nabla_\theta f(z_t, \theta)\|\right] < \infty$. $\mathbb{E}[f(z_t, \theta_0)' f(z_t, \theta_0)] < \infty$, *and* $\sup_{\theta\in\Theta}\|f(z_t, \theta)\| < \infty$ *for all* $t = 1, \ldots, T$ *and* $T = 1, 2, \ldots$. *Each element of* $f(z_t, \theta_0)$ *is uniformly square integrable, for all* $t = 1, \ldots, T$ *and* $T = 1, 2, \ldots$.

(vii) *The parameter space* Θ *is a compact subset of* $\mathbb{R}^v$.

(viii) $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f(z_t, \theta)\right] = 0$, *only when* $\theta = \theta_0$

(ix) *The sequence of positive definite weighting matrices* $W_T \to_p \Sigma_{ff}^{-1}$.

(x) $M \equiv \lim_{T\to\infty} \mathbb{E}\left[T^{-1}\sum_{t=1}^{T}\frac{\partial f(z_t,\theta)}{\partial\theta'}|_{\theta=\theta_0}\right] \in \mathbb{R}^{m\times v}$ *has full column rank.*

(xi) $\lim_{T\to\infty}\sup_{r,s\in(0,1),s>r>\rho} T^{-1/2}\sum_{t=[rT]+1}^{[sT]}\left(\nabla_\delta f_t(\theta_0, \delta_0) - F\right)BH_t = o_p(1)$

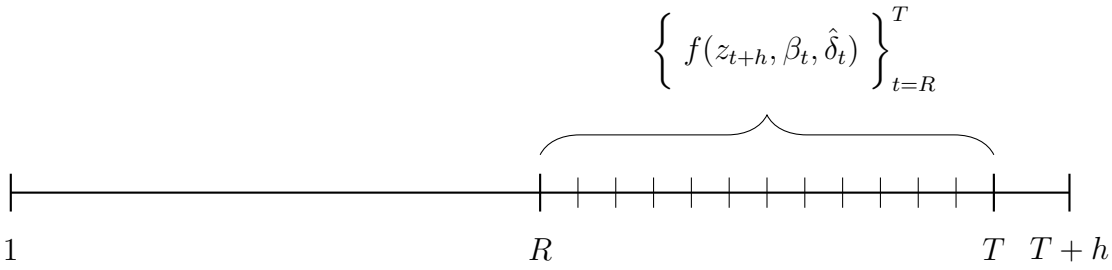
(xii) $\lim_{T\to\infty}\sup_{r,s\in(0,1),s>r>\rho} T^{-1/2} F\sum_{t=[rT]+1}^{[sT]}(B_t - B)H_t = o_p(1)$

(xiii) $\lim_{T\to\infty}\sup_{r,s\in(0,1),s>r>\rho} T^{-1/2}\sum_{t=[rT]+1}^{[sT]}\left(\nabla_\delta f_t(\theta_0, \delta_0) - F\right)(B_t - B)H_t = o_p(1)$

(xiv) $\lim_{T\to\infty}\sup_{r,s\in(0,1),s>r>\rho}\left[T^{-1}\sum_{t=[rT]+1}^{[sT]}\left(\nabla_\theta f_t(\theta_0, \delta_0) - M\right)\right] = o_p(1)$

I now discuss Assumption 4.1. Assumption 4.1.(i) defines the relevant asymptotic experiment as one where both the size of the in-sample and out-of-sample portions diverge to infinity while the size of the in-sample and out-of-sample portions remains a fixed proportion of the total sample size. Assumption 4.1.(ii) is a regularity assumption on the sequence of parameter estimates of the forecasting model. It allows for a variety of estimation methods such as (nonlinear) least squares, maximum likelihood or generalized methods of moments. In addition, the assumption describes the recursive parameter estimation scheme. This assumption is typical of the literature on forecast evaluation, see West and McCracken (1998) or Rossi and Sekhposyan (2016). Assumptions 4.1.(iii) and 4.1.(iv) are weak dependence assumptions equivalent to the ones discussed in the previous section. Together with Assumption 4.1.(v) which assumes positive definiteness of the long-run variance, the

Figure 1: Illustration of the out-of-sample forecasting environment



The figure shows how the data sample is partitioned into an in-sample and out-of-sample portion at R. A forecasting model specified by the researcher is used to generate a sequence of parameter estimates $\{\hat{\delta}_t\}$ using data from the in-sample portion of the sample up to a specified point in time. The forecasting model is used to obtain a sequence of forecast errors for the out-of-sample portion of the data which enters the moment condition. The test builds on the sample moments in the out-of-sample portion of the data $t = R, \ldots, T$.

assumptions are sufficient to obtain weak convergence of the partial sample moments to Brownian Motions using the multivariate functional central limit theorem of Phillips and Durlauf (1986). Assumption 4.1.(vi) are standard smoothness and boundedness condition on the sample moment function under the null hypothesis $f(z, \theta)$. An analogue of this assumption is used in Sowell (1996). Together with assumption 4.1.(vii) which assumes a compact parameter space and Assumption 4.1.(viii) which assumes identification under the null hypothesis, the conditions are sufficient to yield consistency of the GMM estimator used to construct the test statistic. Assumption 4.1.(ix) restricts the choice of weighting matrices used to construct the GMM estimators by requiring that an efficient GMM estimator is used. Assumption 4.1.(x) ensures that the test statistic has a well-defined asymptotic variance. Finally, Assumptions 4.1.(xi)-4.1.(xiv) are boundedness conditions which guarantee that the remainder of a mean-value expansion of the sample moments around δ is asymptotically negligible.

Before I derive the limiting distribution, I provide some intuition on why the limiting distribution in the out-of-sample case differs from the limiting distribution in the in-sample case which was derived in Section 3. The crucial difference to the in-sample case is that the moment functions $f(z_t, \cdot, \cdot)$ depend on the sequence of estimated parameters $\{\hat{\delta}_t\}_{t=R}^{T}$ which were obtained from a separate moment condition. This makes it necessary to take parameter estimation error in δ explicitly into account when evaluating the limiting distribution of the test statistics proposed in Section 2 (see West (1996) for a similar argument).

Under the regularity conditions provided in Assumption 4.1, the following mean-value approximation of the partial sample moments evaluated at $\{\hat{\delta}_t\}_{t=T}^{T}$ holds.

LEMMA 4.1 (OOS Mean-Value Approximation): *Under the regularity conditions in Assumption 4.1 and the null hypothesis defined in (3), for any $r, s \in [0,1]$ with $s > r > \rho$ it holds that*

$$P^{-1/2} \sum_{t=[rT]+1}^{[sT]} f(z_{t+h}, \beta_0, \hat{\delta}_t) = (T/P)^{1/2} \left\{ \frac{1}{\sqrt{T}} \sum_{t=R}^{[sT]} f(z_{t+h}, \beta_0, \delta_0) - \frac{1}{\sqrt{T}} \sum_{t=R}^{[rT]} f(z_{t+h}, \beta_0, \delta_0) \right\}$$
$$+ (T/P)^{1/2} \, FB \left\{ \frac{1}{\sqrt{T}} \sum_{t=R}^{[sT]} H_t(\delta_0) - \frac{1}{\sqrt{T}} \sum_{t=R}^{[rT]} H_t(\delta_0) \right\} + o_{p,rs}(1)$$

where H_t, B are as defined in Assumption 4.1.(ii) and $x_t(r,s) = o_{p,rs}(1)$ denotes that $\sup_{r,s\in[0,1],\ s>r>\rho} \|x_t(r,s)\| = o_p(1)$.

The proof of the Lemma is reported in Appendix C.2.

The expansion in Lemma 4.1 decomposes the partial sample moment into two terms. The first term on the right hand side represents uncertainty that is present even if δ_0 is known. The second part reflects uncertainty about δ_0 originating from estimating the parameter of the forecasting model δ based on the moment function $\mathbb{E}\left[h(z_t, \theta)\right] = 0$. This is in contrast to the in-sample case where δ is estimated using the same moment condition which is used to construct the test statistic, $\mathbb{E}\left[f(z_t, \theta_t)\right] = 0$. Further note that whether parameter estimation error in δ needs to be taken into account crucially depends on F having a non-zero value.

The following theorem establishes the asymptotic distribution of the test statistic under the null hypothesis.

THEOREM 4.1 (OOS Inference): *Assume that the regularity conditions in Assumption 4.1 hold. Under the null hypothesis defined in (3), it holds that*

$$\sup \Phi_T(K) \ \Rightarrow \ \sup_{\lambda^K \in \Lambda_{\epsilon,\rho}} \ \sum_{j=1}^{K+1} \Phi_j(\lambda_{j-1}, \lambda_j)$$

$$D \sup \Phi_T(\bar{K}) \ \Rightarrow \ \max_{1\le k\le \bar{K}} (1/k) \ \sup_{\lambda^K \in \Lambda_{\epsilon,\rho}} \ \sum_{j=1}^{K+1} \Phi_j(\lambda_{j-1}, \lambda_j)$$

$$\Lambda_{\epsilon,\rho} = \Big\{ \lambda_j, \ j = 1, \ldots, K : \ \lambda_j \in (\rho + \epsilon, 1 - \epsilon), \ \lambda_j > \lambda_{j-1} + \epsilon \Big\}, \quad \lambda_0 \equiv \rho, \ \lambda_{K+1} \equiv 1$$

where

$$\Phi_j(\lambda_{j-1},\lambda_j) \equiv \left[\mathcal{B}_m\left(\int_0^{\lambda_j}\omega(u,\lambda_{j-1},\lambda_j)\omega(u,\lambda_{j-1},\lambda_j)'\,du\right)\right]' \times \left\{\int_0^{\lambda_j}\omega(u,\lambda_{j-1},\lambda_j)\,\omega(u,\lambda_{j-1},\lambda_j)'\,du\right\}^{-1} \times \left[\mathcal{B}_m\left(\int_0^{\lambda_j}\omega(u,\lambda_{j-1},\lambda_j)\omega(u,\lambda_{j-1},\lambda_j)'\,du\right)\right]$$

with

$$\omega(u,r,s) \equiv M'\,\Sigma_{ff}^{-1}\,(1-\rho)^{-1/2}\begin{bmatrix} I_m & FB\end{bmatrix} \times \left\{\left[\Omega(u,s)^{1/2}-\Omega(u,r)^{1/2}\right]\mathbb{1}(u\le r) + \Omega(u,s)^{1/2}\,\mathbb{1}(r<u\le s)\right\}\Sigma^{1/2}$$

and where $\Omega(s,\tau)$ *is as defined as*

$$\Omega(s,\tau)^{1/2} \equiv \begin{pmatrix} \mathbb{1}(s\le\rho)\cdot I_m & 0_{m\times d} \\ 0_{d\times m} & \left\{[\ln\tau-\ln\rho]\,\mathbb{1}(s\le\rho)+[\ln(\tau)-\ln(s)]\,\mathbb{1}(\rho<s\le\tau)\right\}\cdot I_d\end{pmatrix}$$

The proof of this theorem can be found in Appendix C.2.

Note that in the general case the limiting distribution in Theorem 4.1 depend on nuisance parameters of the data-generating process and has to be simulated for each application.[12]

Forecast unbiasedness, Efficiency tests and survey forecasts

The general result presented in Theorem 4.1 above simplifies considerably in two cases that are of great interest to practitioners.[13]

The first case applies when parameter estimation error in $\hat{\delta}$ is irrelevant i.e. if it holds that $F = 0$. This case is particularly of interest when the tests are used to evaluate the out-of-sample predictive ability of survey forecasts where the model which generated the forecasts is not available and thus the correction for parameter estimation error can not be applied. Relevant examples of such forecasts are survey and judgemental forecasts produced by central banks such as the Greenbook projections produced by the Federal Reserve Board or private sector forecasts such as the Survey of Professional Forecasters (SPF) or the Blue-

[12] Critical values for this limiting distribution can be simulated by a dynamic programming algorithm similar to the one discussed in Section 5 of this paper where Φ_j is simulated using a modification of the algorithm described Rossi and Sekhposyan (2016).

[13] These special cases were also considered in Rossi and Sekhposyan (2016).

Chip Economic Indicators (BCEI).

The second case applies when the parameter estimation error is asymptotically negligible. This case was discussed in West and McCracken (1998) Corollary 5 for the case of regression-based tests of out-of-sample predictive ability based on the full-sample and is also considered in Rossi and Sekhposyan (2016). In such cases, a special condition holds which considerably simplifies the asymptotic distributions ot the proposed test statistic. The condition is given in 4.1 below. This condition is satisfied in many applications of interest to empirical researchers, particularly tests for forecast unbiasedness and efficiency under general conditions as well as several other tests under more specific assumptions. These cases are discussed in West and McCracken (1998).

The limiting distribution for the special cases is provided in the following Corollary of Theorem 4.1.

COROLLARY 4.1 (OOS Inference in Special Cases): *If (a) $F = 0$, that is parameter estimation error is irrelevant, or (b) the following condition holds*

$$\Sigma_{ff} = -\frac{1}{2}(FB\Sigma_{hf} + \Sigma_{fh}B'F') = FB\Sigma_{hh}B'F'$$

then, the result of Theorem 4.1 simplifies to

$$\sup \Phi_T(K) \Rightarrow \sup_{\lambda \in \Lambda_{\epsilon,\rho}} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j - \rho) - \mathcal{B}_p(\lambda_{j-1} - \rho)\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

$$D\sup \Phi_T(\bar{K}) \Rightarrow \max_{1 \le k \le \bar{K}} (1/k) \sup_{\lambda \in \Lambda_{\epsilon,\rho}} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j - \rho) - \mathcal{B}_p(\lambda_{j-1} - \rho)\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

The proof of this corollary can be found in Appendix C.2.

Note the similarities between the limiting distribution in the in-sample case which was discussed in Section 3 and the limiting distribution in the special cases provided above. In particular, the limiting distribution of the in-sample case is obtained when setting $\rho = 0$. Under the special cases, the critical values do not depend on the data-generating process and can be tabulated.[14] In the case where the tests are applied to survey or judgemental forecasts, the only sample available to researchers is $t = R, \ldots, T$. In particular, the researcher cannot specify a value of ρ as the length of the in-sample portion is unknown. In these cases, critical values for the proposed tests can be obtained by setting $\rho = 0$ in the limiting distribution above and the critical values provided in Section 6 can be used to conduct the test.

[14] A table of these critical values is available upon request.

5 Implementation

This section gives detailed instructions on how to implement the tests.

5.1 Variance estimators

To implement the test statistics defined in (5) and (11), we require the estimators $\hat{\Sigma}_{ff}$ and $\hat{\Omega}_{T,j}$ that appear in the formulas of the LM and Wald statistics.

Computation of $\hat{\Sigma}$ depends on whether there is serial correlation in the moment conditions. When $f(z_t, \theta_0)$ consists of mean-zero uncorrelated random variables, a consistent estimator is given by

$$\hat{\Sigma}_{ff} = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \left[f(z_t, \tilde{\theta}) - \bar{f}_T(\tilde{\theta}) \right] \left[f(z_t, \tilde{\theta}) - \bar{f}_T(\tilde{\theta}) \right]' \tag{15}$$

$$\bar{f}_T(\tilde{\theta}) \equiv \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} f(z_t, \tilde{\theta}) \tag{16}$$

where $\bar{f}_T(\tilde{\theta})$ is the mean of the sample moments. Alternatively, if $f(z_t, \theta_0)$ consists of mean-zero but serially correlated random variables, then a consistent estimator is given by a kernel-based HAC estimator such as

$$\begin{aligned} \hat{\Sigma}_{ff} = &\sum_{l=0}^{T-1} \left\{ \kappa(l/q_T) \frac{1}{T - T_0 + 1} \sum_{t=l+T_0}^{T} \left(f(z_t, \tilde{\theta}) - \bar{f}_T(\tilde{\theta}) \right) \left(f(z_{t-l}, \tilde{\theta}) - \bar{f}_T(\tilde{\theta}) \right)' \right\} \\ &+ \sum_{l=1}^{T-1} \left\{ \kappa(l/q_T) \frac{1}{T - T_0 + 1} \sum_{t=l+T_0}^{T} \left(f(z_{t-l}, \tilde{\theta}) - \bar{f}_T(\tilde{\theta}) \right) \left(f(z_t, \tilde{\theta}) - \bar{f}_T(\tilde{\theta}) \right)' \right\} \end{aligned} \tag{17}$$

where $\kappa(\cdot)$ is a kernel and q_T a bandwidth parameter which can depend on the data. A kernel choice that guarantees that the estimator $\hat{\Sigma}_{ff}$ is positive semi-definite is the Quadratic Spectral Kernel discussed in Newey and West (1987).

Next, consider the estimator $\hat{\Omega}_{j,T}$. This estimator crucially depends on whether the test is conducted in-sample or out-of-sample. In the case where the test is conducted in-sample,

$\hat{\Omega}_{T,j}$ can be computed from simple formulas. Specifically, $\hat{\Omega}_{j,T}$ can be computed as

$$\begin{aligned}
\hat{\Omega}_{j,T} &= (\lambda_j - \lambda_{j-1})^{-1}\, \hat{C}'(\hat{C}\hat{C}')^{-1}\hat{C} \\
\hat{C} &\equiv \bar{M}_\beta'(I_m - \bar{P}_\delta) \\
\bar{P}_\delta &\equiv \bar{M}_\delta(\bar{M}_\delta'\bar{M}_\delta)^{-1}\bar{M}_\delta' \\
\bar{M} &= \hat{\Sigma}_{ff}^{-1/2}\, \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \frac{\partial f_t(z_t, \theta_t)}{\partial \theta'}\Big|_{\theta_t = \tilde{\theta}}
\end{aligned} \tag{18}$$

where $\bar{M}_\beta$ and $\bar{M}_\delta$ are obtained from partitioning $\bar{M} = (\bar{M}_\beta, \bar{M}_\delta)$. $\tilde{\theta}$ is the restricted GMM estimator defined in (7).

5.2 Dynamic Programming Algorithm

This section discusses how to compute the test statistics described above via an efficient dynamic programming algorithm. First, note that *given a fixed vector of sample splits,* λ, the computation of the $\Phi_{T,j}$ parts of the test statistic defined in equation (6) is straightforward. It simply requires computing the restricted GMM estimator defined in equation (7) and computing the components of equation (6) via the estimators provided in the previous section.

However, to compute the $\sup \Phi_T$ and $D \sup \Phi_T$ test statistics which allow for an *unknown vector of sample splits*, one has to compute the sup operator over $\lambda \in \Lambda_\epsilon$. This is computationally challenging as it involves computing a series of test statistics $\{\Phi_T(\lambda_{j-1}, \lambda_j)\}_{j=0}^{K+1}$ for *every possible partition of the sample into* K *segments*, respecting the minimal segement length implicitly defined by the trimmming parameter, ϵ. In principle, a grid search procedure could be used, but with $K > 2$ this becomes quickly infeasible as it involves the computation of $\Phi_T(\cdot)$ of order $\mathcal{O}(T^K)$.

To solve the computational problem, I employ a dynamic programming algorithm which efficiently computes the $\sup \Phi_T$ and $D \sup \Phi_T$ statistics in $\mathcal{O}(T^2)$ operations, regardless of the value of K. The algorithm is based on the early work of Hawkins (1976) and extensions by Bai and Perron (1998, 2003) and Qu and Perron (2007).[15]

The basic idea of the algorithm is as follows. For any given number of changes, K, the $\sup \Phi_T$ test statistic in equation (5) is given by the sum $\Phi_{T,j}(\cdot)$ statistics for $j = 1, \ldots K+1$, which are associated with a specific partition of the sample defined by $\lambda = (\lambda_1, \ldots, \lambda_K)$. The problem of computing the sup over all possible values of $\lambda \in \Lambda_\epsilon$ can therefore be transformed

[15] The dynamic programming algorithm of Bai and Perron (1998, 2003) computes the sum of squared residuals (SSR) of a linear regression model for every possible partition of the sample into $K+1$ regimes. Qu and Perron (2007) extend this algorithm to compute QMLE estimates assuming a linear pseudo-model with gaussian errors. In contrast, I modify the dynamic programming algorithm to directly compute the sup-Φ_T test statistics which are functions of sums of partial sample moments as well as the restricted GMM estimator.

into several steps which are described in the following algorithm.

ALGORITHM 5.1 (Computation of $\sup \Phi_T(K)$ and $D \sup \Phi_T(\bar{K})$ tests): *The* $\sup \Phi_T(K)$ *test is computed by implementing Steps 1 and 2 of the following algorithm. The* $D \sup \Phi_T$ *test is computed by implementing Steps 1-3, setting* $K = \bar{K}$ *for the first two steps.*

Step 1: *Compute and store all possible segments of the test statistic* $\Phi_{T,j}(T_m, T_n) := \Phi_{T,j}([\lambda_m T], [\lambda_n T])$ *for* $T_m, T_n \in t = 1, \ldots, T$ *which satisfy* $T_m > T_n$ *and* $\lambda_r := [\lambda_r T]$. *In the case of the LM test,* $\Phi_{T,j}$ *is as defined in equation (6).*

Step 2: *Recursively maximize the sum of k+1 of these partitions for* $k = 1, 2, \ldots, K$ *using the following Bellman equation.*

$$\Phi_T(\{\lambda_{k,T}\}) = \max_{kh \leq T_j \leq T-h} \left[\Phi_T(\{\lambda_{k-1,T_j}\}) + \Phi_T(T_j + 1, T)\right] \tag{19}$$

where $\Phi_T(\{\lambda_{k,T_j}\})$ *denotes the value of the* $\sup \Phi_T$ *statistic associated with an optimal partition based on k changes and using observations* $t = 1, \ldots, T_j$ *and* $h = [\epsilon T]$ *is the minimum segment length implied by the trimming parameter. The recursion is initialized with* $\Phi_T(\{\lambda_{0,T_j}\}) \equiv \Phi_T(T_j + 1, T)$.

Step 3: *Carrying out Steps 1-2 with* $K = \bar{K}$ *yields a series of test statistics under k changes,* $\left\{\Phi_T(\{\lambda_{k,T}\})\right\}_{k=1}^{\bar{K}}$. *Then, compute* $D \sup \Phi_T(\bar{K})$ *as in equation (11) by first normalizing each element of the series by dividing by the respective k and computing* $D \sup \Phi_T(\bar{K})$ *as the maximum element of the normalized series.*

An efficient software implementation of the steps above is described in Appendix A. Note that the computation of all possible seqments in Step 1 requires computing less than $T(T+1)/2$ times the Φ_T test statistics and is therefore of order $\mathcal{O}(T^2)$.[16]

As stated previously, the test presented in section 2 can be conducted based on a Lagrange-Multiplier form (Φ_T^{LM}) or a Wald-form (Φ_T^W). If all coefficients of the model are to be tested (i.e. $\theta = \beta$ is the full parameter vector), both the LM and Wald tests can be computed via the algorithm described above. However, when the test is carried out on a subvector of θ, the algorithm above needs to be modified to compute the Wald test. One has to augment the steps above with another layer, conditioning on estimates $\hat{\delta}$ under the alternative hypothesis and iterating until convergence.[17] Therefore, the LM test has con-

[16] Depending on the value of the trimming parameter, ϵ, substantially less than $T(T+1)/2$ computations are needed. This is because one only has to consider segments which have a minimum length of $h = [\epsilon T]$ observations. Further, for specific models, additional computational simplifications are possible by using an updating rule to compute Φ_T.

[17] A similar strategy to compute Wald tests has been used in the structural break literature, see e.g. the algorithms in Qu and Perron (2007) and Bai and Perron (2003). Details on the modified algorithm are available from the author on request.

siderable computational benefits over the Wald versions as it only requires the computation of the restricted GMM estimator under the null hypothesis.

6 Simulation Studies

To investigate the finite sample performance of the specification tests proposed in this paper, I conduct a series of Monte-Carlo experiments using several data-generating processes. The goal of the simulation exercises is three-fold: First, they illustrate the difference between the proposed model specification test, tests of multiple structural breaks and traditional hypothesis tests assuming a constant parameter. Second, they assess the finite-sample properties of the proposed tests for a variety of data-generating processes. Third, they study under which conditions allowing for multiple shifts in the coefficient vector leads to power gains relative to tests imposing one break. In what follows, I describe each simulation exercise, define the respective data-generating processes and discuss the simulation results.

6.1 Asymptotic Power Illustration

Before studying the finite-sample approximation quality of the limiting distributions, I conduct an asymptotic power exercise to illustrate why it is useful to jointly test the hypotheses in (3) rather than relying on traditional hypothesis tests or traditional structural break tests. Recall from the discussion in Section 2 that traditional tests are designed to test *either* $H_0^{(1)}$ or $H_0^{(2)}$ and therefore might not reliably detect departures originating in the other part of null hypothesis. In contrast, the tests proposed in this paper, *jointly* test $H_0^{(1)}$ and $H_0^{(2)}$ and reject against any combination of these hypotheses. To illustrate this argument, I conduct a simulation exercise using the following simple data-generating process:

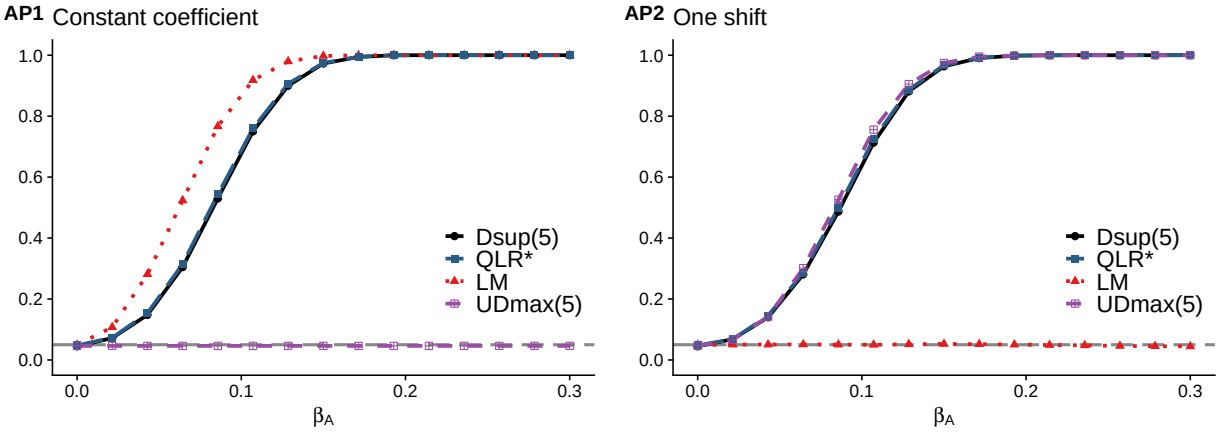
$$y_{t+1} = \beta_{t,T}\, x_t + \eta_t \qquad\qquad x_t, \eta_t \sim iid\ \ \mathcal{N}(0,1) \tag{20}$$

I simulate data from this model using a large sample size of $T = 1,000$ and inspect rejection rates of (i) a full-sample LM test, (ii) Bai and Perron (1998)'s UDmax(5) test, (iii) Rossi (2005)'s QLR_T^* test and (iv) the $D\sup(5)$ test proposed in this paper against two stylized data generating-processes capturing a departure from $H_0^{(1)}$ and $H_0^{(2)}$, respectively.[18] The next two paragraphs briefly describe the two considered designs and discuss the simulated power curves which are reported in Figure 2.

Design AP1. The first design has $\beta_{t,T} = \beta_A\ \forall\ t$ i.e. the parameter which is tested is not time-varying. Figure 2 panel AP1 shows the "asymptotic" power of the tests as a function of β_A. It illustrates that when the parameter is not time-varying, the full-sample parameter

[18] All tests are conducted at 95% level and a trimming parameter of $\epsilon = 0.05$ is used for tests (ii)-(iv).

Figure 2: Asymptotic Power Illustration



The figure shows simulated rejection rates for tests of the null hypothesis $H_0 : \beta = 0$ in model (20) with 5% nominal size. $D\sup(5)$ denotes the proposed instability-robust test with a maximum of $\bar{K} = 5$ shifts in the parameter vector. LM denotes a traditional full-sample Lagrange-Multiplier test. QLR^* denotes Rossi (2005)'s QLR^*_T test which imposes one break. 'UDmax(5)' denotes Bai and Perron (1998)'s UDmax test with a maximum of 5 breaks. Rejection rates are based on 5,000 replications and $T = 1,000$ observations.

test (LM) is the most powerful test among the four tests considered. In contrast, the test for a structural break (UDmax(5)) has a flat power function equal to the nominal size 5%. The test robust to multiple instabilities proposed in this paper ($Dsup(5)$) exhibits high power against this alternative.

Design AP2. The second design has $\beta_{t,T} = \beta_A \mathbb{1}(t \leq T/2) - \beta_A \mathbb{1}(t > T/2)$ i.e. there is a single shift in the parameter which is tested. Figure 2 panel AP2 shows the "asymptotic power" of the tests as a function of β_A. It illustrates that this shift is not detected by a traditional hypothesis test (LM) which has a flat power function equal to the nominal size 5%. In contrast, the structural break test ($UDmax$) and the instability-robust tests (QLR^* and $D\sup(5)$) exhibit substantial power against this alternative. Finally, it is important to note that while the QLR^* test is optimal for the case of one break, both the structural break test and the proposed $D\sup(5)$ test exhibit virtually the same power.

6.2 Finite-sample size

Next, I assess the quality of the finite-sample approximation of the limiting distribution of the test. I first focus on the finite-sample size of the proposed test procedure. Size control in finite-samples is an important feature of any test procedure since a researcher choosing a particular significance level α expects the test to reject only in $(1 - \alpha)\%$ of cases when the null hypothesis is true. I study a data-generating process resembling a linear regression which predicts a scalar series, y_t, with past values of a predictor, x_t, and a control variable,

w_t correlated with the predictor. Both the prediction error and the predictor variable admit serial correlation in the form of an AR(1) process. This class of models has received considerable attention in the predictability literature, both in macroeconomics and finance (see the reviews in Pitarakis and Gonzalo (2019) and Rossi (2013)). The data-generating process is defined as follows

$$y_t = \beta_{t,T}\, x_t + \delta\, w_t \; + \eta_t \qquad \eta_t = \phi_\eta\, \eta_{t-1} + \zeta_t \tag{21}$$

$$x_t = \phi_x\, x_{t-1} + \xi_t \qquad w_t = \rho_{xw}\, x_t + \iota_t \tag{22}$$

where ζ_t, ξ_t, ι_t are independent and $iid\ \mathcal{N}(0,1)$ and x_0, η_0 are drawn from the unconditional distribution of the respective AR process.

To assess the finite-sample size of the tests, I simulate 10,000 samples from the model above, imposing the null hypothesis $\beta_{t,T} = 0$, and compute rejection rates for (i) the traditional LM test, (ii) Rossi (2005)'s QLR^*_T test and (iii) the $D\sup(5)$ test proposed in this paper. All tests are conducted on the β subvector while leaving δ unspecified. I simulate specifications with sample sizes $T \in \{125, 250, 500, 1000\}$ and various degrees of serial correlation in the predictor and prediction errors, $\phi_\eta, \phi_x \in \{0, 0.25, 0.5\}$ and consider tests both with and without HAC correction.[19] In all specifications, the correlation between the predictor and control variable is fixed at $\rho_{xw} = 0.25$.

Results. Table 2 reports the results regarding the empirical size of the model specification tests for the model in equation (21) for 5% nominal size and a trimming parameter of $\epsilon = 0.05$. First, consider the case in which no serial correlation is present in the data ($\phi_x = 0, \phi_\eta = 0$). When constructed using a heteroskedasticity-robust variance estimator (upper-left corner of Panel A), we observe that the proposed $D\sup$ test allowing for up to $\bar{K} = 5$ shifts exhibits good size control with finite-sample size being virtually identical to the size of Rossi (2005)' QLR_T test which imposes one break. In comparison to the traditional LM test, both instability-robust tests are slightly undersized in small samples, but size quickly converges to the nominal level as T grows. Using a variance estimator with HAC correction as described above yields similar size results; only in small samples are the $D\sup(5)$ and QLR^*_T more conservative than without the HAC correction (upper-left corner of Panel B). Next, consider the case with serial correlation in the predictor and/or prediction error. If serial correlation is ignored and the tests is constructed using a heteroskedasticity-robust variance estimator, size control crucially depends on the structure and amount of serial correlation in the data. The table illustrates that when serial correlation is present only in the predictor variable or the prediction errors (second and third row/column of Panel

[19] The HAC correction is based on AR(1) approximaiton using Andrews (1991) data-dependent method and a Quadratic-Spectral kernel.

A), finite-sample size is barely affected. However, when serial correlation is present in both model components, all three tests become oversized with rejection rates growing both with ϕ_x and ϕ_η. In that case, the instability-robust tests have significantly larger size-distortions than the LM tests with the proposed $D\sup(5)$ test exhibiting mildly worse size control than the QLR^* test. However, when the serial correlation is acknowledged and the test is constructed using the variance estimator with HAC correction provided in Section 2, the $D\sup(5)$ test recovers good size control and nominal size is close to 5% in medium to large samples. Finally, it is interesting to assess the robustness of these findings to choosing a larger trimming parameter. As Table B.1 in the appendix shows, increasing the trimming parameter to $\epsilon = 0.1$ yields nearly identical results.

To conclude, the simulation exercise shows that, when constructed using the appropriate variance estimator, the $D\sup(K)$ test exhibits good size control across a variety of specifications with finite-sample size comparable to that of Rossi (2005)'s QLR^*_T test. This implies that it is possible to allow for more than one break under the alternative without incurring a penalty with respect to the finite-sample size of the testing procedure.

6.3 Finite-sample power

Finally, I examine the finite-sample power of the proposed test. Understanding the power properties of (correctly sized) testing procedures is important as a researcher ideally would like to choose the testing procedure that maximises the chances of correctly detecting a departure from the null hypothesis based on an available data sample. As power properties typically depend on the data-generating process considered, a careful analysis of finite-sample rejection rates helps to understand under which conditions testing procedures should be used. To assess finite-sample power of the proposed tests, I focus on a class of alternatives where $\beta_{t,T}$ exhibits local departures from the null hypothesis, β_0. Specifically, I focus on alternatives in which the coefficient has a value of zero for the majority of the sample, but there are multiple short episodes during which the coefficient departs from zero. This type of data-generating process has received considerable attention in the predictability literature in recent years (see e.g. the "pockets of predictability hypothesis" in Timmermann (2008) and Farmer et al. (2019) and empirical studies Gonzalo and Pitarakis (2012, 2017) and Rossi (2020).

As in the size simulations discussed in the previous subsection, I employ the following data-generating process

$$y_t = \beta_{t,T}\, x_t + \delta\, w_t \; + \eta_t \qquad \eta_t = \phi_\eta\, \eta_{t-1} + \zeta_t \tag{23}$$

$$x_t = \phi_x\, x_{t-1} + \xi_t \qquad w_t = \rho_{xw}\, x_t + \iota_t \tag{24}$$

where $\zeta_t, \xi_t, \iota_t \sim iid\ \mathcal{N}(0,1)$ and x_0, η_0 are drawn from the unconditional distribution of the

Table 2: Finite-sample size for $\epsilon = 0.05$. Nominal size 5%

		$\phi_\eta = 0$			$\phi_\eta = 0.25$			$\phi_\eta = 0.5$		
T	ϕ_x	Dsup(5)	LM	QLR^*	Dsup(5)	LM	QLR^*	Dsup(5)	LM	QLR^*
Panel A: Heteroskedasticity-robust										
125	0.00	0.043	0.052	0.040	0.047	0.051	0.044	0.057	0.051	0.048
250	0.00	0.042	0.046	0.040	0.043	0.047	0.041	0.049	0.047	0.042
500	0.00	0.044	0.049	0.042	0.045	0.049	0.044	0.051	0.048	0.045
1,000	0.00	0.047	0.050	0.046	0.047	0.050	0.048	0.049	0.052	0.050
125	0.25	0.043	0.051	0.040	0.074	0.065	0.064	0.133	0.081	0.102
250	0.25	0.046	0.045	0.042	0.074	0.061	0.068	0.130	0.078	0.108
500	0.25	0.043	0.051	0.044	0.080	0.066	0.075	0.138	0.079	0.120
1,000	0.25	0.049	0.051	0.046	0.083	0.064	0.078	0.143	0.081	0.128
125	0.50	0.054	0.051	0.045	0.126	0.080	0.101	0.265	0.118	0.194
250	0.50	0.049	0.048	0.044	0.130	0.080	0.110	0.289	0.121	0.224
500	0.50	0.048	0.048	0.045	0.135	0.080	0.119	0.310	0.119	0.247
1,000	0.50	0.048	0.053	0.050	0.141	0.085	0.125	0.330	0.125	0.263
Panel B: HAC - AR(1) approximation, QS kernel, Andrews (1991) bandwidth										
125	0.00	0.032	0.047	0.030	0.031	0.048	0.032	0.036	0.046	0.035
250	0.00	0.036	0.044	0.034	0.036	0.045	0.035	0.039	0.044	0.036
500	0.00	0.043	0.049	0.041	0.041	0.049	0.042	0.046	0.048	0.042
1,000	0.00	0.045	0.049	0.044	0.045	0.050	0.046	0.046	0.051	0.048
125	0.25	0.030	0.046	0.030	0.034	0.052	0.035	0.039	0.053	0.038
250	0.25	0.037	0.044	0.036	0.041	0.050	0.041	0.045	0.052	0.043
500	0.25	0.040	0.049	0.042	0.048	0.054	0.049	0.054	0.054	0.048
1,000	0.25	0.047	0.050	0.045	0.054	0.053	0.052	0.057	0.055	0.054
125	0.50	0.033	0.047	0.034	0.038	0.053	0.039	0.035	0.053	0.036
250	0.50	0.040	0.045	0.038	0.046	0.052	0.046	0.045	0.053	0.044
500	0.50	0.042	0.048	0.040	0.051	0.054	0.048	0.050	0.052	0.049
1,000	0.50	0.047	0.052	0.046	0.055	0.056	0.056	0.055	0.054	0.053

The table reports simulated finite-sample size for tests of the null hypothesis $H_0 : \beta = 0$ with 5% nominal size. $D\sup(5)$ denotes the proposed instability-robust test with a maximum of $\bar{K} = 5$ breaks, LM denotes a traditional LM test and QLR^* denotes Rossi (2005)'s instability-robust test imposing one break. Rejection rates are based on 10,000 replications from the model in equation (21) using a sample of T observations where the serial correlation of the predictor and prediction error is controlled by ϕ_x, ϕ_η, respectively.

respective AR process.

To assess the finite-sample power of the tests, I simulate 5,000 samples from the model above and compute rejection rates for (i) the traditional LM test, (ii) Rossi (2005)'s QLR_T^* test and (iii) the $D\,\sup(5)$ test proposed in this paper. All tests are conducted on the β subvector while leaving δ unspecified. The instability-robust tests use a trimming parameter of $\epsilon = 0.05$. I focus on a sample size $T = 400$ to ensure all tests have a similar size and raw power can be compared between the tests. As in the size simulations, the correlation between the predictor and control variable is fixed at $\rho_{xw} = 0.25$. For clarity of exposition, the main text presents results for the case without serial correlation and using a heteroskedasticity-robust variance estimator. Appendix B reports additional results for $\phi_x = 0.5, \phi_\eta = 0.5$ and a variance estimator with HAC correction, based on an AR(1) approximation using Andrews (1991) data-dependent method and a Quadratic-Spectral kernel.

I consider power curves based on three designs for the time-varying coefficient vector, $\beta_{t,T}$ which are denoted P1 - P3. Figure 3 Panel D illustrates the three designs which differ in the number and location of the local departures from the null hypothesis $\beta_{t,T} = 0$ over the sample as well as the sign of the shifts. The width of the predictability pockets is fixed at 5% of the sample size i.e. each pocket has a duration of 20 observations. The magnitude of the shifts is uniform over the pockets and scaled by a scalar parameter β_A, where $\beta_A = 0$ implies no predictability at any point in time. In what follows, I briefly discuss the considered designs and the corresponding power curves in Figure 3.

Design P1. Figure 3 Panel P1 shows power curves for the predictability process P1 which features two predictability pockets of opposite signs, located at one-third and three-fourths of the sample, respectively. The simulation illustrates the need for predictability tests that take instabilities into account. It is evident that the traditional LM test exhibits no power against alternatives of this form; the rejection rate of the LM test stays constant at nominal size 5%, regardless of the magnitude of the shift, β_A. In contrast, the predictability tests that allow for instabilities have power in detecting predictability of this form. Furhter, we note that the power of the $D\,\sup$ test uniformly dominates that of Rossi (2005)'s test, showing that allowing for multiple shifts in the coefficient vector under the alternative leads to power gains in finite-samples.

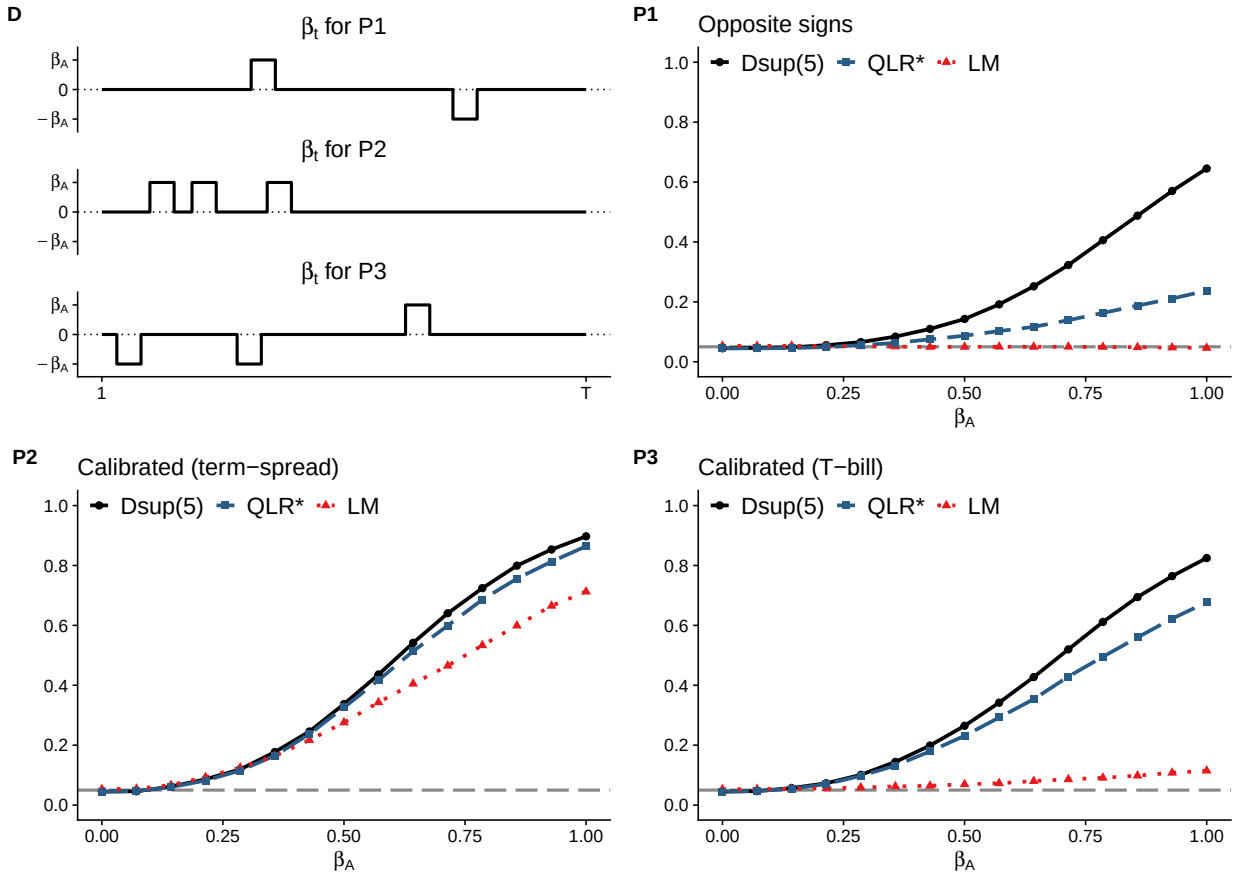
Designs P2 & P3. The bottom two panels in Figure 3 show power comparisons for the predictability proccesses P2 and P3, respectively. These processes are calibrated to empirical results from Farmer et al. (2019) who conduct non-parametric regressions to study the presence of predictability pockets in various commonly considered predictors of the equity premium. When evaluating power of the proposed test in the presence of predictability pockets, there are many choices of processes for $\beta_{t,T}$ that differ with respect to number

of the pockets, their location and duration as well sign and magnitude of the implied coefficient shifts. Calibrating $\beta_{t,T}$ allows to assess the test's performance under conditions that could be encountered in "real-world" empirical examples and therefore provides a good benchmark for assessing the power of the test. I calibrate the processes P2 and P3 to the empirical findings of Farmer et al. (2019) by matching the number of simulated pockets, the relative location of the pockets over the samples as well as the sign of the implied shifts to the results reported by the authors in predicting excess returns from (i) the term spread using daily data (P2) and (ii) the T-bill rate using monthly data (P3).[20] For both processes, the power of the proposed D sup test uniformly dominates the power of the test imposing one break. This confirms that allowing multiple shifts in the coefficient vector under the alternative leads to power gains in empirically relevant scenarios. Further, as with P1, the traditional LM test shows substantially lower power for P2 and almost no power for P3 which features both negative and positive pockets, reiterating the need for predictability tests that explicitly take instabilities into account.

Sensitivity Checks. I assess the sensitivity of the conclusions drawn above to variations in the specification of the considered tests. Figure B.1 reports finite-sample power curves for the same data-generating process without serial correlation ($\phi_x = 0, \phi_\eta = 0$), but using variance estimator with a HAC correction is used. All conclusions regarding relative power of the tests discussed above remain unchanged. Finally, I inspect power in the case where both $\phi_x = 0.5$ and $\phi_\eta = 0.5$. Figure B.2 shows that the power of all of the three test decreases considerably relative to the case without serial correlation, but that the proposed D sup test remains the most powerful among the three tests considered.

[20] Specifically, Farmer et al. (2019) find 3 pockets which have "less than a 5% chance of being spurious" for predicting from the term-spread using daily data. These pockets are located at 12%, 21% and 37% of the sample with all pockets exhibiting positve coefficient shifts. For predicting using the T-bill rate in monthly data, using the same criterion, they find 3 pockets which are located at 6%, 30% and 65% of the sample with the first two pockets exhibiting negative shifts of the coefficient and the third pocket having a positive shift.

Figure 3: Finite-sample power for $\epsilon = 0.05$



The figure shows simulated rejection rates for tests of the null hypothesis $H_0 : \beta = 0$ against different designs for the alternative, $\beta_{t,T}$, denoted P1 - P3. Panel D illustrates the different designs for $\beta_{t,T}$. Power curves are reported for increasing size of the shifts, β_A, under the alternative. All tests are conducted at $\alpha = 5\%$ significance level. The solid black line denotes the proposed $D\sup(5)$ instability-robust test with a maximum of $\bar{K} = 5$ breaks. The blue shaded line denotes Rossi (2005)'s QLR_T^* test imposing one break and the red dotted line denotes a traditional LM test. Rejection rates are based on 5,000 replications for a sample of $T = 400$ observations from the model in equation (23) where the serial correlation of the predictor and prediction error $\phi_x = 0, \phi_\eta = 0$, respectively.

7 Local Stock Return Predictability

Are stock returns predictable by financial valuation ratios or term-structure variables? This question is at the center of an important research agenda in finance and has been analyzed by a large array of seminal studies.[21] However, despite a significant volume of research being devoted to this question, the predictability debate has not yet reached a consensus. For example, while some studies find evidence of predictability by valuation ratios or consumption ratios (Lettau and Ludvigson, 2001), other studies find that these results are unstable and crucially depend on the stochastic properties of the predictors or the sample period studied (Campbell and Yogo, 2006). Welch and Goyal (2007) came to the conlcusion that "[...] the literature has yet to find some variable that has meaningful and robust empirical equity premium forecasting power [...]" (p. 1505).

One explanation for the difficulty of establishing a consensus is that predictability can vary over time. For example, Pesaran and Timmermann (1995) find that the ability of various economic variables to predict stock returns changes with the volatility of returns and Rapach and Wohar (2006) provide evidence of parameter instability in predictive regressions. Recently, some studies have presented evidence that predictability is a local phenomenon and is concentrated in short subsamples of the data. Timmermann (2008) concludes that the "[...] empirical findings suggest that most of the time stock returns are not predictable, but there appear to be pockets in time where there is modest evidence of local predictability." Pesaran and Timmermann (2000) note that oil prices were an important predictor for stock prices during the 1970s but that their importance subsequently vanished. Similarly, Gonzalo and Pitarakis (2012, 2017) and Henkel et al. (2011) find that predictability is linked to measures of business cycle conditions, leading to short episodes of significant predictability. Further theoretical support for local predictability is given in Timmermann (2008) who argues that investors' successful search for good forecasting models itself might generate "pockets of predictability" i.e. short-lived periods of significant predictability that are followed by long periods without predictability. In a recent study, Farmer et al. (2019) argue that such pockets of return predictability are consistent with an asset pricing model featuring incomplete learning and provide ample empirical evidence in support of the predictability pockets hypothesis using non-parametric regressions.

Empirical support for or against predictability crucially relies on specification tests in predictive regressions. However, as argued throughout this paper, when predictability varies over time, traditional hypothesis tests may have low or no power against potentially important alternatives. Simply testing for predictability at each point in time or over a collection of various subsamples does not offer a good alternative as it suffers from a multiple testing

[21] For early research on this topic see for example Fama and French (1988) and Campbell and Shiller (1988). Ang and Bekaert (2007), Welch and Goyal (2007) and Timmermann (2008), Cochrane (2008) provide more recent discussions.

problem affecting size and power of the tests. The test proposed in this paper, however, is robust to multiple shifts in magnitude and signs of the parameter vector and can be applied to a general class of models. This makes it a good choice to investigate the hypothesis of local predictability in return prediction.

This paper is not the first to address the issue of robustifying inference in return prediction to episodic predictability. For example, Gonzalo and Pitarakis (2012, 2017) analyze return predictability in a threshold model that links time variation in predictability to the state of the economy (for example a variable measuring business cycle fluctuations). Henkel et al. (2011) follow a similar approach. In general, however, predictability might be linked to a variety of features of the economic environment that are not necessarily tied to the business cycle. The advantage of using the test proposed in this paper over existing approaches is that the researcher does not need to condition predictability on a set of known variables measuring the state of the economy. Rather, the test can be used as a first step to establish whether there is evidence of local predictability before the researcher comes up with a hypothesis about potential driving forces of the predictability process.

7.1 Data

The issue of stock return predictability has been analyzed using a large variety of specifications. To keep the exposition in this paper compact and to ensure comparability with the literature, I focus on the most commonly employed specification which predicts monthly US stock market excess returns over the post-war period 1946-2019 using the set of financial variables considered in Welch and Goyal (2007).[22] In the following, I give details on the construction of the excess return series and the predictor variables.

Equity Premium. The dependent variable is constructed based on a CRSP (Center for Research in Security Prices) dataset that is widely used in the literature. Specifically, I construct the excess return on the US stock market (equity premium) as the difference between a measure of the US stock market log return and a risk-free log return. The US stock market return is measured by the value-weighted S&P 500 total stock market return including dividends. The risk-free rate is the three-month T-bill rate from FRED. This measure of the equity premium is widely used in the literature e.g. recently by Welch and Goyal (2007) and Kostakis et al. (2015). For robustness, I also present results using an alternative measure of the excess return where the US stock market return is the value-weighted CRSP stock market return including dividends for NYSE, AMEX and NASDAQ and the risk-free rate is proxied by a 1-month Treasury bill rate from Ibbotson and Associates Inc. This alternative measure of the equity premium has been used recently in Gonzalo and Pitarakis (2012) and

[22] Studies of excess return prediction often restrict the sample to the post-war period, see e.g. Gonzalo and Pitarakis (2012).

Gonzalo and Pitarakis (2017). Figure B.3 Panel A shows the equity premium series used in the empirical analysis

PREDICTOR VARIABLES. The source of the predictor data is an updated version of the monthly dataset used in Welch and Goyal (2007).[23] This predictor dataset has been considered in numerous studies in the literature and has become a benchmark in the predictability literature. I focus on the same set of predictors recently considered in Kostakis et al. (2015), namely the *dividend-payout ratio*, defined as the difference between the log of dividends and the log of earnings, the *earnings-price ratio*, defined as the difference between the log of earnings and the log of stock prices, the *long-term yield*, defined as the long-term US government bond yield from Ibbotson's Stocks, Bonds Bills and Inflation Yearbook, the *T-bill rate* which after 1934 is the 3-month T-bill rate from FRED and before is extracted from the NBER Macrohistory database, the *term-spread* which is the difference between the long-term yield and the T-bill rate, the *dividend-price ratio*, defined as the log of dividends over stock prices, the *dividend-yield*, defined as the log of dividends over lagged prices, the *default yield spread*, defined as the difference between the BAA and AAA-rated corporate bond yields taken from FRED, the *book-to-market ratio* which is the ratio of book value to market value for the DJIA, the *net equity expansion*, defined as the ratio of the twelve month moving sum of net equity issues by NYSE listed stocks divided by the total end-of-year market capitalization of these stocks and the *inflation rate* calculated from the Consumer Price Index of the Bureau of Labor Statistics. Figure B.3 panels B-L show the predictor series used in the empirical analysis.

7.2 Predictability Tests Robust to Instabilities

As is customary in the literature, I study the individual predictive ability of each of the financial variables using the following univariate predictive model

$$r^e_{t+1} = \alpha + \beta\, x_t + \eta_{t+1} \tag{25}$$

where r^e_{t+1} denotes the one-month-ahead excess return and x_t is the considered predictor. Predictability studies typically conduct tests of the hypothesis $H_0 : \beta = 0$ over the full sample or specific subsamples. However, as illustrated in the previous sections, these tests are not robust to the presence of instabilities and might therefore fail to detect locally ocurring predictability. The test proposed in this paper considers the same null hypothesis, but explicitly takes local predictability into account. Using the proposed test, I revisit the specification above and re-evaluate the predictability of the set of predictors described in the

[23] The dataset updated until December 2019 was downloaded from Amit Goyal's website and at the time of writing this paper could be found at http://www.hec.unil.ch/agoyal/.

previous subsection. As in the simulation studies discussed in Section 6, I use a $D\sup(5)$ test with 5% trimming. All tests are conducted based on the HAC variance estimators reported in Section 2 based on an AR(1) approximation with Andrews (1991) automatic bandwidth selection procedure and a Quadratic Spectral kernel

Results. Table 3 reports the predictability tests robust to instabilities for each of the potential predictors. To facilitate comparison with the literature, the left panel reports the full-sample least squares estimates ($\hat{\beta}_{OLS}$), the R^2 of the full-sample regression (in percentage points) as well as the traditional predictability tests using a t-ratio with HAC correction for the full-sample (t^{HAC}) and a subsample starting in 1952 (t^{HAC}_{1952}). The right panel reports the results from the instability-robust $D\sup(5)$ predictability tests proposed in Section 2 for the same subsamples.

I first dicuss the traditional inference approach using test statistics constructed over the full-sample. Note that standard least-squares inference indicates that only few of the financial variables have significant predictive ability for the equity premium at 5% level, namely the T-bill rate, the dividend-price ratio, the dividend-yield and the inflation rate. Further, when considering a slightly different subsample starting in January 1952[24] (t^{HAC}_{1952}), the significance of the dividend-price ratio and the dividend-yield disappears and only the T-bill rate remains significant at 5% level. The results are in line with findings from traditional predictability tests reported in Kostakis et al. (2015) and, more generally, match the conclusion of previous studies documenting that predictive ability of valuation ratios crucially depends on the subsample considered (see e.g. the discussion in Campbell and Yogo (2006) or Welch and Goyal (2007)).[25]

The picture changes considerably when looking at the instability-robust tests. The $D\sup(5)$ test conducted on the 1946 - 2019 sample shows additional significant (local) predictability at 5% level for the dividend-payout ratio, the earnings-price ratio, the default yield spread and net equity expansion; only the long-term yield, the term-spread and the book-to-market ratio are not significant at 5% level. In addition, contrary to the traditional tests, the findings from the $D\sup(5)$ test are robust to moving to the post-1952 subsample.

The test results support the hypothesis that predictability from financial variables is a local phenomenon and therefore difficult to detect with traditional tests which do not take instabilities into account. The findings generalize those found in earlier studies which document evidence of episodic predictability for specific variables (e.g. Gonzalo and Pitarakis

[24]The post-1952 subsample is often considered in predictability studies since term structure variables are thought to be more informative after the passing of the 1952 Treasury Accord which separated government debt management from monetary policy.

[25]In particular, the tests results in Gonzalo and Pitarakis (2012) also indicate vanishing excess return predictability for the dividend-yield and Kostakis et al. (2015) find the same variables to be significant in the post-1952 period when using traditional tests. Finally, all regressions have low explanatory power with R^2 below 1%, a feature documented e.g. in Welch and Goyal (2007).

Table 3: Predictability Tests for the Equity Premium

			Traditional		Robust	
Predictor	$\hat{\beta}_{OLS}$	R^2 (%)	t^{HAC}	t^{HAC}_{1952}	$D\sup(5)$	$D\sup(5)_{1952}$
Dividend payout ratio	0.002	0.03	0.30	0.45	13.72**	14.42**
Earnings-price ratio	0.005	0.28	1.10	0.61	13.86**	12.24**
Long-term yield	-0.086	0.34	-1.62	-1.52	8.57	8.50
T-bill rate	-0.110	0.66	-2.39**	-2.40**	12.07**	11.50*
Term spread	0.192	0.38	1.70*	1.89*	8.40	9.30
Dividend-price ratio	0.006	0.41	1.98**	1.41	16.01***	14.86**
Dividend-yield	0.006	0.46	2.08**	1.53	15.68***	14.85**
Default yield spread	0.154	0.03	0.30	0.36	13.89**	13.54**
Book-to-market ratio	0.004	0.06	0.63	0.29	6.55	5.76
Net equity expansion	-0.041	0.04	-0.37	-0.45	21.43***	21.62***
Inflation rate	-0.915	0.96	-2.58**	-1.85*	17.46***	11.03*

The table presents the results of conducting predictability tests of the null hypothesis $\beta = 0$ for the post-war sample 1946-2019 in model (25) using the S&P 500 Equity Premium. The left panel reports the full-sample least squares estimates, $\hat{\beta}_{OLS}$, the R^2 of the full-sample regression (in percentage points) as well as the traditional predictability tests using a t-ratio with HAC correction for the full-sample, t^{HAC}, and a subsample starting in 1952, t^{HAC}_{1952}. The right panel reports the results from the instability-robust $D\sup LM$ model-specification tests with a maximum of $\bar{K} = 5$ shifts and trimming parameter set at $\epsilon = 0.05$ for the same subsamples. For all test statistics, the stars denote a rejection the null hypothesis of no predictability at significance levels 1% (***), 5% (**), and 10% (*), respectively.

(2012) document episodic predictability for the dividend-yield.). Further, they explain why studies that split the sample at different dates have often come to conflicting conclusions regarding the predictive ability of a wide class of predictors and highlight the need for robustifying inference to local instabilities.

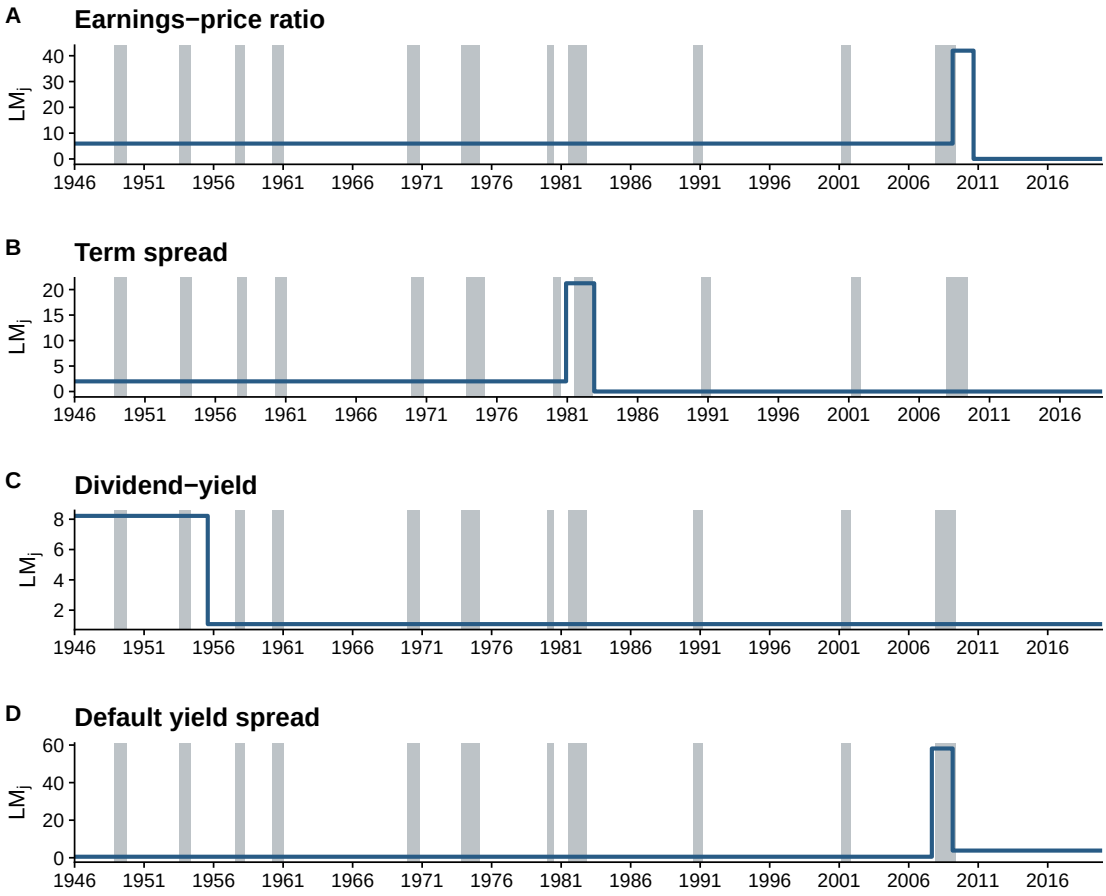
Robustness. To assess the robustness of the results documented above, I repeat the predictability tests using a different measure of the equity premium based on CRSP stock returns including dividends for NYSE, AMEX and NASDAW with a 1-month treasury bill rate as the risk-free rate (see data section above). Table B.2 reports the test results from this alternative measure of the equity premium. All conclusions discussed in this section are robust to this change. I also address a potential concern regarding the persistence of the considered predictor variables. The predictors typically studied with predictive models such as the one in equation (25) are often highly persistent and recently, studies have started to model these predictors as near-unit-root predictors when conducting predictability tests (see e.g. the discussion in Pitarakis and Gonzalo (2019) and the IVX approach proposed by Kostakis et al. (2015)). While it is in principle possible to apply the proposed test to an IVX moment condition, I focus here on exploring the robustness of the results discussed above by using first-differenced values of the predictors, $\Delta x_t = x_t - x_{t-1}$. Table

B.3 presents the results from the predictability tests using first-differenced predicors. In comparison to the results discussed above, the evidence of predictability indeed disappears for the earnings-price ratio, the dividend-price ratio, the dividend-yield and the inflation rate, all variables for which there is evidence of near-unit-root behaviour, highlighting the importance of carefully assessing the stochastic properties of each predictor before applying predictability tests. However, for the remaining predictors the main conclusions discussed above continue to hold in the first-differenced model. Most importantly, the traditional tests still give conflicting evidence when considering the two subsamples while the proposed instability-robust tests provide stable inference.

Predictability Paths. The evidence discussed above that predictability from financial variables is a local phenomenon, raises an interesting question: Is there heterogeneity in the location of predictability for different predictors? And if yes, how does predictability evolve over the sample? Is predictability really concentrated in short "predictability pockets" such as hypothesized by recent studies or are there larger episodes of predictability? While the tests proposed in this paper do not allow to draw precise inference on which periods over the sample are significant or not at a given significance level,[26] the components of the $D\sup(K)$ test statistic do provide a narrative view on how predictability might evolve over the sample. Figure 4 shows the evolution of the $\Phi_{T,j}$ components of the $D\sup(K)$ test over the sample. In contrast to the previous section, I consider a larger upper ceiling of shifts, $\bar{K} = 10$ and a lower trimming parameter, $\epsilon = .02$. I adopt this specification of the tests since Farmer et al. (2019) who assess predictability paths using a non-parametric testing procedure based on multiple t-tests report evidence of particularly short-lived predictability episodes. Simulation studies available on request show that the tests still exhibits good size control and has the same power properties against the data-generating processes discussed in Section 6 when using a sample of the size available here. To choose the number of shifts, I adopt the BIC criterion discussed in Bai and Perron (2006) and only report predictability paths for the variables for which the criterion detects at least one shift. The figure provides evidence that predictability is indeed concentrated in different periods over the sample, depending on the predictor considered. Specifically, while the predictive ability of the dividend-yield seems concentrated in the pre-1956 period, the earnings-price ratio and the default yield spread predictors to have an episode of large predictive ability during the Great Recession period. Finally, the predictive ability of the term spread variable seems concentrated in a brief period during the early 1980s.

[26] Since the proposed tests are joint hypothesis tests, it is not straightforward to extend the methods to a sequential procedure in the spirit of Bai and Perron (1998)'s $F_T(k+1|k)$ test as the researcher would need to add an intermediate step that tests whether the rejection occurred due to constant predictability or the presence of an additional shift in the coefficient vector. However, developing repeated testing procedures that correct for the multiple testing at each stage is an interesting avenue for future research.

Figure 4: Predictability Paths



The graph shows the $\Phi_{T,j}$ components of the $D\sup LM(10)$ model-specification tests for the S&P 500 Equity Premium based on the model in equation (25) with trimming parameter $\epsilon = .02$. The number of coefficient shifts is selected via the BIC criterion discussed in Bai and Perron (2006). Gray vertical bars denote NBER recessions.

8 Conclusion

This paper develops a general approach to test whether a parameter should be included in an economic model robust to time-variation in parameters. The hypothesis test can be used to evaluate any economic model described by a set of moment conditions in-sample or out-of-sample. In-sample, the test selects between two nested model specifications in the presence of parameter instabilities. Out-of-sample, the test can be used to evaluate the performance of model forecasts or model-free forecasts such as survey or judgmental forecasts robust to time-variation. The key feature of the proposed test is that it is particularly powerful in the presence of multiple breaks in parameters without imposing a specific form of time-variation. Further, the test statistic provides narrative evidence on which parts of the sample drive the rejection of the null hypothesis.

The approach jointly tests for both parameter instability and a constant non-zero value of the parameter. This allows the test to detect departures from the null hypothesis, even when they only occur over short periods of the sample and makes the test more powerful than traditional hypothesis tests which are based on the full sample. The test statistic jointly considers all possible partitions of the sample up to an upper bound of $\bar{K}$ splits to evaluate whether there is evidence to reject the null hypothesis. It can be constructed based on a Lagrange-Multiplier or a Wald form and can be efficiently implemented via a dynamic programming algorithm provided in the paper.

Extensive Monte-Carlo simulations show that the proposed test is accurately sized in finite samples and is more powerful than tests assuming constant coefficients or a single break if the data-generating process exhibits multiple breaks in parameters. At the same time, the test has high power when model parameters only undergo one shift or are constant. This makes the test particularly useful when the researcher faces uncertainty about whether and how parameters change over time.

The empirical study uses the test to document the presence of local short-horizon predictability in the US equity premium during the 1946-2019 period from a set of financial variables considered in Welch and Goyal (2007). There is significant predictive ability with respect to one-month ahead excess market returns for a large set of predictors, once time-variation is taken into account. In contrast to traditional predictability tests based on the full sample, the conclusions from the proposed test are invariant to changes in the considered sample. Furthermore, the test provides evidence of heterogeneity in the location of predictability episodes across variables. The findings explain why traditional tests often fail to uncover predictability in the full sample and why studies that split the sample at different dates often arrive at conflicting results regarding the predictive ability of a wide class of variables.

References

Andrews, D. (1991). Heteroskedasticity and autocorrelation consistent covariant matrix estimation. *Econometrica*, 59(3):817–858.

Andrews, D. W. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica: Journal of the Econometric Society*, pages 1465–1471.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, pages 821–856.

Ang, A. and Bekaert, G. (2007). Stock return predictability: Is it there? *The Review of Financial Studies*, 20(3):651–707.

Bai, J. (1999). Likelihood ratio tests for multiple structural changes. *Journal of Econometrics*, 91(2):299–323.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

Bai, J. and Perron, P. (2006). Multiple structural change models: a simulation analysis. *Econometric theory and practice: Frontiers of analysis and applied research*, 1:212–237.

Campbell, J. Y. and Shiller, R. J. (1988). Stock prices, earnings, and expected dividends. *The Journal of Finance*, 43(3):661–676.

Campbell, J. Y. and Yogo, M. (2006). Efficient tests of stock return predictability. *Journal of Financial Economics*, 81(1):27–60.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2012). Model selection when there are multiple breaks. *Journal of Econometrics*, 169(2):239–246.

Cavaliere, G. (2005). Unit root tests under time-varying variances. *Econometric Reviews*, 23(3):259–292.

Chinco, A., Clark-Joseph, A. D., and Ye, M. (2019). Sparse signals in the cross-section of returns. *The Journal of Finance*, 74(1):449–492.

Christiano, L. J., Eichenbaum, M. S., and Trabandt, M. (2018). On dsge models. *Journal of Economic Perspectives*, 32(3):113–40.

Clark, T. and McCracken, M. (2013). Chapter 20 - advances in forecast evaluation. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2 of *Handbook of Economic Forecasting*, pages 1107 – 1201. Elsevier.

Cochrane, J. H. (2008). The dog that did not bark: A defense of return predictability. *The Review of Financial Studies*, 21(4):1533–1575.

Dagum, L. and Menon, R. (1998). Openmp: an industry standard api for shared-memory programming. *IEEE computational science and engineering*, 5(1):46–55.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Elliott, G. and Müller, U. K. (2006). Efficient tests for general persistent time variation in regression coefficients. *The Review of Economic Studies*, 73(4):907–940.

Fama, E. F. and French, K. R. (1988). Dividend yields and expected stock returns. *Journal of financial economics*, 22(1):3–25.

Farmer, L., Schmidt, L., and Timmermann, A. (2019). Pockets of predictability. *Available at SSRN 3152386*.

Georgiev, I., Harvey, D. I., Leybourne, S. J., and Taylor, A. R. (2018). Testing for parameter instability in predictive regression models. *Journal of Econometrics*, 204(1):101–118.

Gonzalo, J. and Pitarakis, J.-Y. (2012). Regime-specific predictability in predictive regressions. *Journal of Business & Economic Statistics*, 30(2):229–241.

Gonzalo, J. and Pitarakis, J.-Y. (2017). Inferring the predictability induced by a persistent regressor in a predictive threshold model. *Journal of Business & Economic Statistics*, 35(2):202–217.

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

Harvey, D. I., Leybourne, S. J., Sollis, R., and Taylor, A. R. (2020). Real-time detection of regimes of predictability in the u.s. equity premium*. *Journal of Applied Econometrics*.

Hawkins, D. M. (1976). Point estimation of the parameters of piecewise regression models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25(1):51–57.

Henkel, S. J., Martin, J. S., and Nardari, F. (2011). Time-varying short-horizon predictability. *Journal of Financial Economics*, 99(3):560–580.

Hoesch, L., Rossi, B., and Sekhposyan, T. (2020). Has the information channel of monetary policy disappeared? revisiting the empirical evidence.

Kostakis, A., Magdalinos, T., and Stamatogiannis, M. P. (2015). Robust econometric inference for stock return predictability. *The Review of Financial Studies*, 28(5):1506–1553.

Lettau, M. and Ludvigson, S. (2001). Consumption, aggregate wealth, and expected stock returns. *the Journal of Finance*, 56(3):815–849.

McLeish, D. (1975). Invariance principles for dependent variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(3):165–178.

Newey, W. K. and McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111 – 2245. Elsevier.

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

Odendahl, F., Rossi, B., and Sekhposyan, T. (2020). Comparing forecast performance with state dependence. *Working Paper*.

Perron, P. and Qu, Z. (2006). Estimating restricted structural change models. *Journal of Econometrics*, 134(2):373–399.

Pesaran, M. H. and Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance*, 50(4):1201–1228.

Pesaran, M. H. and Timmermann, A. (2000). A recursive modelling approach to predicting uk stock returns. *The Economic Journal*, 110(460):159–191.

Phillips, P. C. and Durlauf, S. N. (1986). Multiple time series regression with integrated processes. *The Review of Economic Studies*, 53(4):473–495.

Pitarakis, J.-Y. (2017). A simple approach for diagnosing instabilities in predictive regressions. *Oxford Bulletin of Economics and Statistics*, 79(5):851–874.

Pitarakis, J.-Y. and Gonzalo, J. (2019). Predictive regressions. *Oxford Research Encyclopedias (Economics and Finance)*.

Qu, Z. and Perron, P. (2007). Estimating and testing structural changes in multivariate regressions. *Econometrica*, 75(2):459–502.

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

Rossi, B. (2005). Optimal tests for nested model selection with underlying parameter instability. *Econometric theory*, 21(5):962–990.

Rossi, B. (2006). Are exchange rates really random walks? some evidence robust to parameter instability. *Macroeconomic dynamics*, 10(1):20–38.

Rossi, B. (2013). Advances in forecasting under instability. In *Handbook of economic forecasting*, volume 2, pages 1203–1324. Elsevier.

Rossi, B. (2020). Forecasting in the presence of instabilities: How do we know whether models predict well and how to improve them.

Rossi, B. and Sekhposyan, T. (2016). Forecast rationality tests in the presence of instabilities, with applications to federal reserve and survey forecasts. *Journal of Applied Econometrics*, 31(3):507–532.

Sanderson, C. and Curtin, R. (2016). Armadillo: a template-based c++ library for linear algebra. *Journal of Open Source Software*, 1(2):26.

Sowell, F. (1996). Optimal tests for parameter instability in the generalized method of moments framework. *Econometrica: Journal of the Econometric Society*, pages 1085–1107.

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

Welch, I. and Goyal, A. (2007). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

West, K. and McCracken, M. (1998). Regression-based tests of predictive ability. *International Economic Review*, 39(4):817–40.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, pages 1067–1084.

Appendix A Critical Values & Implementation

This section provides additional details on the implementation of the tests. The first section reports the critical values for the tests and describes the simulation procedure used to obtain them. The second section describes a software package accompanying the paper available from the author.

A.1 Asymptotic Critical Values

Critical values of the test statistics can be obtained by directly simulating the limiting distributions listed in Theorem 3.1 using a dynamic programming algorithm.[27] The table below reports critical values for the predictability tests discussed in the paper. The significance levels considered in the tables are 10%, 5%, 2.5% and 1%. The critical values were obtained by simulating the asymptotic distributions based on 10,000 Monte Carlo replicatons and an approximation length of $N = 3600$ for the Brownian Motions.[28]

A.2 Software Implementation

Accompanying the paper, the author makes available a software package that can be used to conveniently use the tests for applied work and for simulating the critical values reported in Table A.1. The code is mainly written in C++ using the Armadillo C++ linear algebra library (Sanderson and Curtin, 2016) and offers an R interface provided in the form of an R package using the Rcpp library (Eddelbuettel and François, 2011). The routines carrying out the dynamic programming algorithm have been parallelized using OpenMP application programming interface (Dagum and Menon, 1998). The software package can be obtained from the author on request.

[27] Alternatively, one could obtain the critical values from an approximation strategy for functions of Brownian motions discussed in Bai (1999).

[28] Simulations of critical values were carried out on the Amazon Web Services Elastic Compute Cloud (AWS EC2) using a *c5.xlarge* instance type (4 vCPUs, 8 GB memory) running Amazon Linux 2.

Table A.1: Asymptotic Critical Values for $\sup \Phi_T(K)$ and $D \sup \Phi_T(\bar{K})$ tests

ϵ	p	α	$D\sup(5)$	$\sup(1)$	$\sup(2)$	$\sup(3)$	$\sup(4)$	$\sup(5)$
0.05	1	0.100	10.115	9.644	17.542	23.162	28.471	33.017
0.05	1	0.050	11.566	11.293	19.593	25.721	30.974	35.638
0.05	1	0.025	13.180	13.048	21.542	27.837	33.124	38.112
0.05	1	0.010	15.196	15.155	24.182	30.330	36.257	41.303
0.05	2	0.100	14.008	13.810	23.920	31.926	39.282	46.025
0.05	2	0.050	15.789	15.647	26.053	34.516	42.134	49.096
0.05	2	0.025	17.468	17.361	28.146	36.901	44.720	52.020
0.05	2	0.010	19.953	19.915	30.832	39.694	47.824	55.510
0.10	1	0.100	9.386	9.131	15.605	20.002	23.664	26.368
0.10	1	0.050	10.985	10.813	17.555	22.215	26.101	28.879
0.10	1	0.025	12.555	12.462	19.423	24.315	28.143	31.150
0.10	1	0.010	14.528	14.528	21.754	26.888	31.063	34.225
0.10	2	0.100	13.301	13.185	21.621	28.170	33.857	38.278
0.10	2	0.050	15.054	15.007	24.040	30.691	36.529	41.426
0.10	2	0.025	16.703	16.690	25.863	33.087	39.379	44.537
0.10	2	0.010	19.189	19.162	28.168	36.201	42.386	47.949
0.15	1	0.100	8.833	8.670	14.007	17.367	19.426	18.791
0.15	1	0.050	10.474	10.335	16.057	19.726	21.857	21.349
0.15	1	0.025	12.113	12.075	18.036	21.777	23.950	23.486
0.15	1	0.010	14.163	14.163	20.391	24.511	27.025	26.211
0.15	2	0.100	12.800	12.734	19.997	25.229	28.851	29.077
0.15	2	0.050	14.651	14.627	22.291	27.870	31.651	31.824
0.15	2	0.025	16.262	16.230	24.231	30.194	34.195	34.773
0.15	2	0.010	18.613	18.581	26.794	32.879	37.592	37.952

This table reports simulated quantiles of the limiting distributions of the $\sup \Phi_T(K)$ and $D \sup \Phi_T(\bar{K})$ tests. The critical values were obtained based on 10,000 Monte-Carlo replications and an approximation length of $N = 3600$ observations for the partial sums to simulate the Brownian Motions.

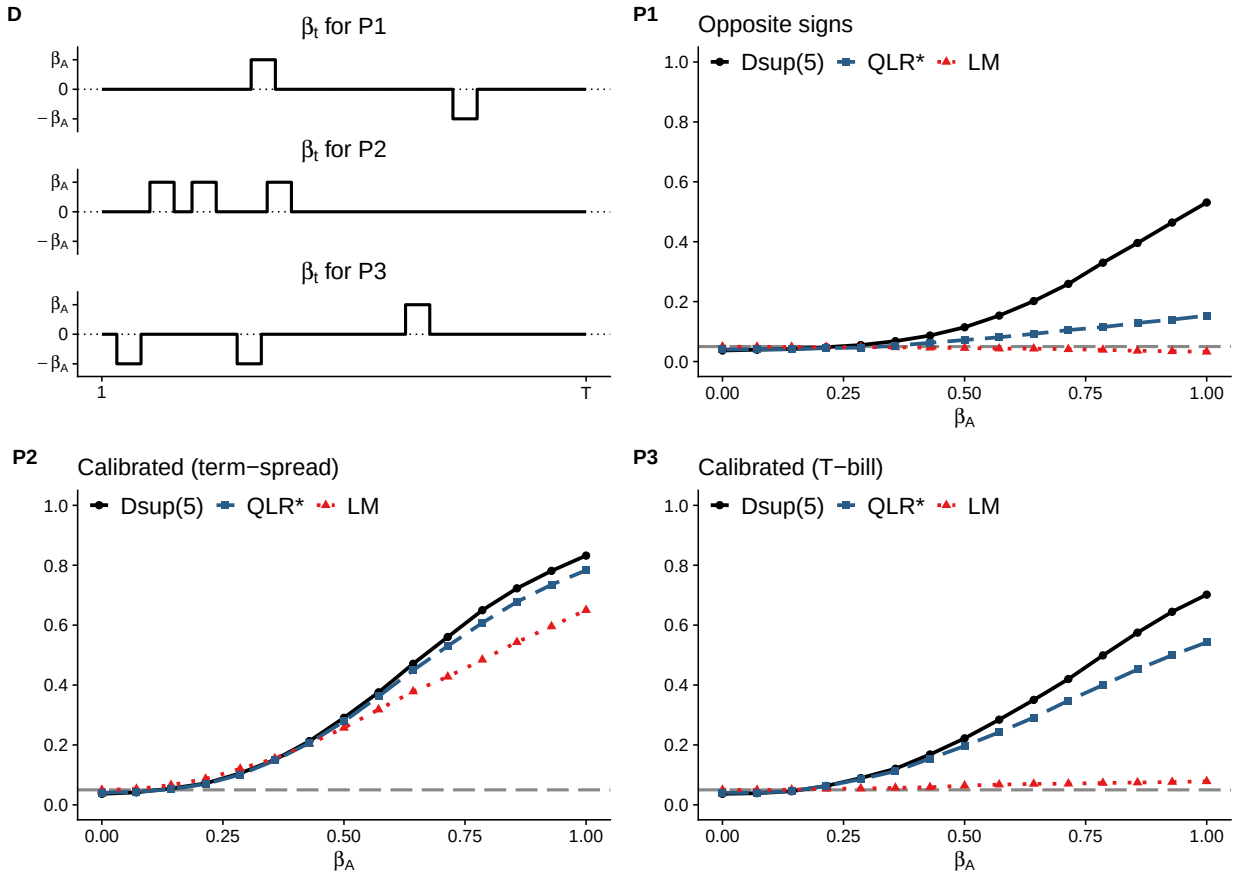
Appendix B Additional Tables & Figures

Table B.1: Finite-sample size for $\epsilon = 0.1$. Nominal size 5%

		$\phi_\eta = 0$			$\phi_\eta = 0.25$			$\phi_\eta = 0.5$		
T	ϕ_x	Dsup(5)	LM	QLR^*	Dsup(5)	LM	QLR^*	Dsup(5)	LM	QLR^*
Panel A: Heteroskedasticity-robust										
125	0.00	0.032	0.052	0.034	0.032	0.051	0.035	0.036	0.051	0.035
250	0.00	0.037	0.046	0.039	0.040	0.047	0.038	0.040	0.047	0.040
500	0.00	0.042	0.049	0.042	0.042	0.049	0.040	0.042	0.048	0.044
1,000	0.00	0.045	0.050	0.045	0.045	0.050	0.045	0.046	0.052	0.048
125	0.25	0.031	0.051	0.034	0.054	0.065	0.054	0.086	0.081	0.084
250	0.25	0.038	0.045	0.041	0.066	0.061	0.065	0.105	0.078	0.104
500	0.25	0.041	0.051	0.042	0.071	0.066	0.071	0.114	0.079	0.113
1,000	0.25	0.046	0.051	0.046	0.078	0.064	0.076	0.122	0.081	0.118
125	0.50	0.036	0.051	0.037	0.085	0.080	0.084	0.176	0.118	0.172
250	0.50	0.043	0.048	0.041	0.105	0.080	0.102	0.227	0.121	0.210
500	0.50	0.043	0.048	0.043	0.110	0.080	0.111	0.244	0.119	0.230
1,000	0.50	0.046	0.053	0.047	0.120	0.085	0.117	0.260	0.125	0.240
Panel B: HAC - AR(1) approximation, QS kernel, Andrews (1991) bandwidth										
125	0.00	0.024	0.047	0.025	0.022	0.048	0.025	0.024	0.046	0.025
250	0.00	0.033	0.044	0.032	0.033	0.045	0.032	0.032	0.044	0.032
500	0.00	0.040	0.049	0.041	0.040	0.049	0.040	0.039	0.048	0.041
1,000	0.00	0.044	0.049	0.043	0.044	0.050	0.045	0.042	0.051	0.046
125	0.25	0.023	0.046	0.024	0.026	0.052	0.028	0.027	0.053	0.028
250	0.25	0.034	0.044	0.034	0.037	0.050	0.039	0.038	0.052	0.039
500	0.25	0.039	0.049	0.040	0.044	0.054	0.046	0.045	0.054	0.045
1,000	0.25	0.044	0.050	0.044	0.051	0.053	0.050	0.051	0.055	0.052
125	0.50	0.024	0.047	0.028	0.028	0.053	0.031	0.025	0.053	0.027
250	0.50	0.037	0.045	0.037	0.040	0.052	0.044	0.037	0.053	0.039
500	0.50	0.039	0.048	0.039	0.045	0.054	0.047	0.043	0.052	0.045
1,000	0.50	0.045	0.052	0.044	0.051	0.056	0.053	0.050	0.054	0.052

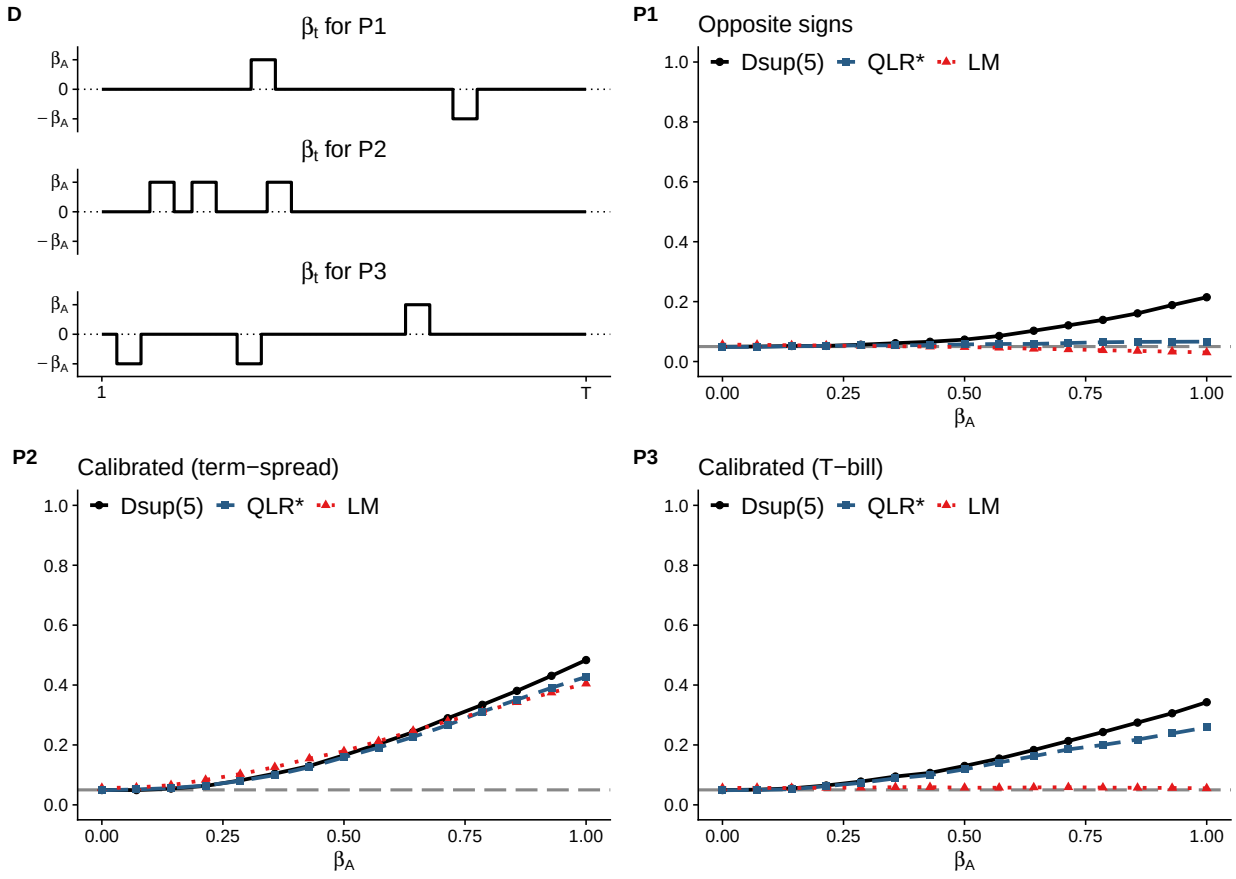
The table reports simulated finite-sample size for tests of the null hypothesis $H_0 : \beta = 0$ with 5% nominal size. $D\sup(5)$ denotes the proposed instability-robust test with a maximum of $\bar{K} = 5$ breaks, LM denotes a traditional LM test and QLR^* denotes Rossi (2005)'s instability-robust test imposing one break. Rejection rates are based on 10,000 replications from the model in equation (21) using a sample of T observations where the serial correlation of the predictor and prediction error is controlled by ϕ_x, ϕ_η, respectively.

Figure B.1: Finite-sample power for $\epsilon = 0.05$ (HAC correction)



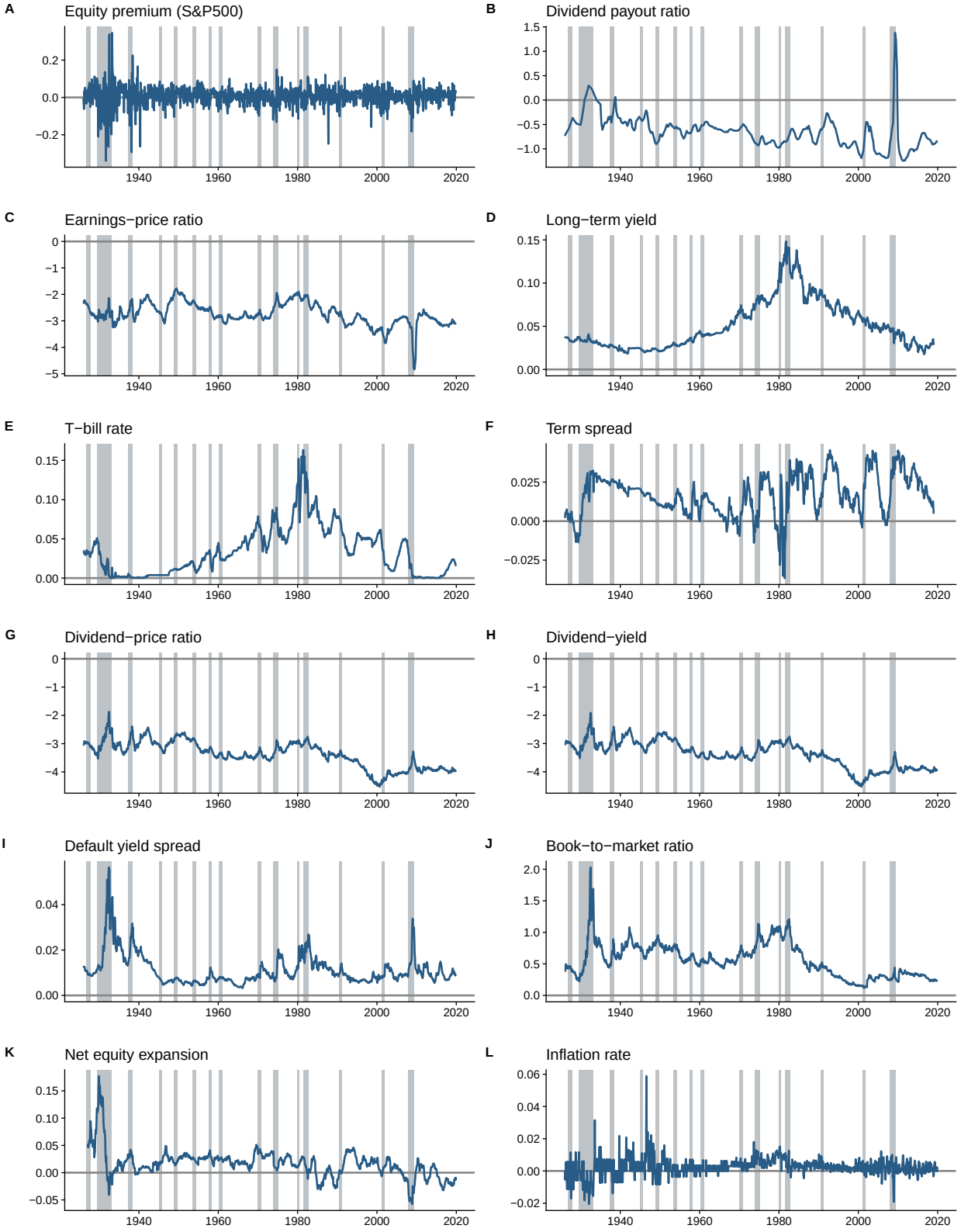
The figure shows simulated rejection rates for tests of the null hypothesis $H_0 : \beta = 0$ against different designs for the alternative, $\beta_{t,T}$, denoted P1 - P3. Panel D illustrates the different designs for $\beta_{t,T}$. Power curves are reported for increasing size of the shifts, β_A, under the alternative. All tests are conducted at $\alpha = 5\%$ significance level. The solid black line denotes the proposed $D\sup(5)$ instability-robust test with a maximum of $\bar{K} = 5$ breaks. The blue shaded line denotes Rossi (2005)'s QLR^*_T test imposing one break and the red dotted line denotes a traditional LM test. Rejection rates are based on 5,000 replications for a sample of $T = 400$ observations from the model in equation (23) where the serial correlation of the predictor and prediction error $\phi_x = 0, \phi_\eta = 0$, respectively.

Figure B.2: Finite-sample power for $\epsilon = 0.05$ (HAC correction)



The figure shows simulated rejection rates for tests of the null hypothesis $H_0 : \beta = 0$ against different designs for the alternative, $\beta_{t,T}$, denoted P1 - P3. Panel D illustrates the different designs for $\beta_{t,T}$. Power curves are reported for increasing size of the shifts, β_A, under the alternative. All tests are conducted at $\alpha = 5\%$ significance level. The solid black line denotes the proposed $D\sup(5)$ instability-robust test with a maximum of $\bar{K} = 5$ breaks. The blue shaded line denotes Rossi (2005)'s QLR^*_T test imposing one break and the red dotted line denotes a traditional LM test. Rejection rates are based on 5,000 replications for a sample of $T = 400$ observations from the model in equation (23) where the serial correlation of the predictor and prediction error $\phi_x = 0.5, \phi_\eta = 0.5$, respectively.

Figure B.3: Data used in Equity Premium prediction



The figure shows the equity premium series y_{t+1} in panel A and the raw predictor data used in the empirical application, x_t in panels B to L. Shaded gray bars denote recessions as measured by the NBER indicator.

Table B.2: Predictability Tests for the Equity Premium

			Traditional		Robust	
Predictor	$\hat{\beta}_{OLS}$	R^2 (%)	t^{HAC}	t^{HAC}_{1952}	$D\sup(5)$	$D\sup(5)_{1952}$
Dividend payout ratio	0.003	0.05	0.43	0.58	14.58**	15.47***
Earnings-price ratio	0.005	0.24	1.02	0.57	11.80**	10.71*
Long-term yield	-0.086	0.32	-1.57	-1.48	7.25	7.18
T-bill rate	-0.112	0.65	-2.22**	-2.23**	11.29*	10.66*
Term spread	0.205	0.41	1.70*	1.87*	8.09	8.71
Dividend-price ratio	0.006	0.42	1.90*	1.42	13.83**	12.62**
Dividend-yield	0.007	0.49	2.06**	1.61	13.60**	12.58**
Default yield spread	0.271	0.07	0.49	0.54	12.71**	12.40**
Book-to-market ratio	0.005	0.08	0.74	0.43	5.66	5.72
Net equity expansion	-0.038	0.03	-0.32	-0.39	18.86***	19.12***
Inflation rate	-0.940	0.97	-2.67***	-1.89*	17.37***	10.86*

The table presents the results of conducting predictability tests of the null hypothesis $\beta = 0$ for the post-war sample 1946-2019 in model (25) using the CRSP Equity Premium. The left panel reports the full-sample least squares estimates, $\hat{\beta}_{OLS}$, the R^2 of the full-sample regression (in percentage points) as well as the traditional predictability tests using a t-ratio with HAC correction for the full-sample, t^{HAC}, and a subsample starting in 1952, t^{HAC}_{1952}. The right panel reports the results from the instability-robust $D\sup LM$ model-specification tests with a maximum of $\bar{K} = 5$ shifts and trimming parameter set at $\epsilon = 0.05$ for the same subsamples. For all test statistics, the stars denote a rejection the null hypothesis of no predictability at significance levels 1% (***), 5% (**), and 10% (*), respectively.

Table B.3: Predictability Tests for the Equity Premium (First differences)

			Traditional		Robust	
Predictor	$\hat{\beta}_{OLS}$	R^2 (%)	t^{HAC}	t^{HAC}_{1952}	$D\sup(5)$	$D\sup(5)_{1952}$
Dividend payout ratio	-0.061	0.45	-2.14**	-2.29**	20.21***	20.00***
Earnings-price ratio	0.018	0.07	0.88	0.96	4.87	4.72
Long-term yield	-1.590	1.04	-3.02***	-2.92***	11.89**	12.86**
T-bill rate	-1.102	1.06	-3.30***	-3.26***	19.37***	17.81***
Term spread	0.350	0.11	0.96	1.00	11.37*	10.42*
Dividend-price ratio	-0.037	0.14	-0.91	-1.04	9.11	9.65
Dividend-yield	0.027	0.07	0.81	0.72	3.65	3.82
Default yield spread	0.261	0.00	0.15	0.19	13.50**	13.32**
Book-to-market ratio	-0.069	0.20	-1.33	-1.64	7.69	7.90
Net equity expansion	-0.674	0.35	-1.52	-1.68*	11.25*	10.82*
Inflation rate	-0.265	0.08	-1.22	-1.98**	3.14	4.87

The table presents the results of conducting predictability tests of the null hypothesis $\beta = 0$ for the post-war sample 1946-2019 in model (25) using the S&P 500 Equity Premium where all variables are transformed to first-differences. The left panel reports the full-sample least squares estimates, $\hat{\beta}_{OLS}$, the R^2 of the full-sample regression (in percentage points) as well as the traditional predictability tests using a t-ratio with HAC correction for the full-sample, t^{HAC}, and a subsample starting in 1952, t^{HAC}_{1952}. The right panel reports the results from the instability-robust $D\sup LM$ model-specification tests with a maximum of $\bar{K} = 5$ shifts and trimming parameter set at $\epsilon = 0.05$ for the same subsamples. For all test statistics, the stars denote a rejection the null hypothesis of no predictability at significance levels 1% (***), 5% (**), and 10% (*), respectively.

Appendix C Mathematical Derivations

NOTATION. Before presenting the derivations, recall some notational conventions that are used throughout the rest of the appendix. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space on which all of the random elements are defined. Unless specified otherwise, all limits are taken as the sample size $T \to \infty$. The symbol $\xrightarrow{p}$ denotes convergence in probability and $\xrightarrow{d}$ denotes convergence in distribution. Next, $\Rightarrow$ denotes weak convergence for sequences of measurable random elements of a space of bounded Euclidean-valued cadlag functions on the product space $D[0,1]^T$ as defined in Phillips and Durlauf (1986) where each component space $D[0,1]$ is equipped with the Skorohod metric. $\|\cdot\|$ denotes the Euclidean norm of a vector of matrix and $[\cdot]$ is the integer part operator. For notational simplicity, I say that $x_t(r,s) = o_{p,rs}(1)$ if it holds that $\sup_{r,s\in[0,1],s>r\rho} \|x_t(r,s)\| = o_p(1)$.

C.1 In-Sample Inference

THEOREM 3.1 (Limiting distribution for in-sample tests): *Assume that the regularity conditions in Assumption 3.1 hold. Under the null hypothesis defined in (3), it holds that*

$$\sup \Phi_T(K) \;\Rightarrow\; \sup_{\lambda^K \in \Lambda_\epsilon} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j) - \mathcal{B}_p(\lambda_{j-1})\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

$$D \sup \Phi_T(\bar{K}) \Rightarrow \max_{1\le k\le \bar{K}} (1/k) \sup_{\lambda^K \in \Lambda_\epsilon} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j) - \mathcal{B}_p(\lambda_{j-1})\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

$$\Lambda_\epsilon \equiv \Big\{ \lambda_j : \;\; \lambda_j \in (\epsilon, 1-\epsilon),\; \lambda_j > \lambda_{j-1} + \epsilon,\;\; j = 1, \ldots, K \Big\}$$

where $\lambda_0 \equiv 0, \lambda_{K+1} \equiv 1$ *and* $\mathcal{B}_p(\cdot)$ *is a* $(p \times 1)$ *vector of independent standard Brownian motions on* $[0,1]$.

Proof. Note that under the null hypothesis in (3), it holds that $\theta_t = \theta_0 = (0_{p\times 1}, \delta)\;\; \forall\; t$. Therefore, under the null hypothesis the function defining the moment condition $f(z_t, \theta_t) = f(z_t, \beta_t, \delta)$ can be written as a function of a constant parameter $f(z_t, \theta)$. This notation will be used in the subsequent derivations.

To prove the weak convergence results stated in the theorem, I start by showing that the partial sample moments satisfy the following invariance principle under the null hypothesis.

$$T^{-1/2}\, W_T^{1/2} \sum_{t=1}^{[sT]} f(z_t, \theta_0) \;\Rightarrow\; \mathcal{B}_m(s) \qquad s \in (0,1]$$

Recall the following result from Corollary 2.2. of Phillips and Durlauf (1986) which gener-

alizes the stationary version of the univariate invariance principle by McLeish (1975).

> Corollary 2.2, Phillips and Durlauf (1986): *Let* $\{u_t\}_{t=1}^{\infty}$ *be a weakly stationary sequence of random* $n \times 1$ *vectors satisfying* $\mathbb{E}[u_t] = 0 \ \forall\, t$. *If (a)* $\mathbb{E}|u_{i1}|^{\beta} < \infty$ $(i = 1, ..., n)$ *for some* $2 \leq \beta < \infty$ *and (b) either* $\sum_{n=1}^{\infty} \varphi_n^{1-1/\beta} < \infty$ *or,* $\beta > 2$ *and* $\sum_{n=1}^{\infty} \alpha_n^{1-2/\beta} < \infty$, *then*
>
> $$\Sigma = \lim_{T\to\infty} \mathbb{E}\left[T^{-1} S_T S_T'\right] = \mathbb{E}[u_1 u_1'] + \sum_{k=2}^{\infty} \{\mathbb{E}[u_1 u_k'] + \mathbb{E}[u_k u_1']\}$$
>
> *where* $S_t = \sum_{j=1}^{[Tt]} u_j$. *If* Σ *is positive definite, then* $X_T(t) = \frac{1}{\sqrt{T}} \Sigma^{-1/2} S_{[Tt]} \Rightarrow W(t)$ *as* $T \to \infty$.

Choose $u_t := f(z_t, \theta_0)$ and verify the conditions of the Corollary. The first requirement and condition (a) follow from Assumption 3.1.(ii) and Assumption 3.1.(iv). Condition (b) follows from 3.1.(i). The last requirements follows from Assumption 3.1.(iii). Applying the Corollary, using Assumption 3.1.(vii) and Slutskys Theorem, it follows that

$$T^{-1/2} W_T^{1/2} \sum_{t=1}^{[sT]} f(z_t, \theta_0) \ \Rightarrow \ \mathcal{B}_m(s) \tag{C.1}$$

Lagrange-Multiplier Form, Φ_T^{LM}

The Lagrange-Multiplier form builds on the restricted GMM estimator defined in (7). I start by proving that this estimator is consistent under the null hypothesis in (3) i.e. that $\tilde{\theta} \xrightarrow{p} \theta_0$. Recall from equation (7) the definition of $\tilde{\theta}$.

$$\begin{aligned} \tilde{\theta} &:= \arg\max_{\theta\in\Theta} \ \hat{Q}_T(\theta) \qquad \text{subject to} \ \ A\tilde{\theta} = 0 \\ \hat{Q}_T(\theta) &:= \hat{F}_T(\theta)' \, W_T \, \hat{F}_T(\theta) \\ \hat{F}_T(\theta) &\equiv \frac{1}{T} \sum_{t=1}^{T} f(z_t, \theta) \end{aligned}$$

where $T_0 = 1$ and $A \equiv \begin{bmatrix} I_{p\times p} & 0_{p\times q} \end{bmatrix}$.

To prove consistency of $\tilde{\theta}$, I first show consistency of the unrestricted estimator $\hat{\theta}$ which is defined as the estimator above, but ignores the constraint $A\tilde{\theta} = 0$. Define the limiting objective function $Q_0(\theta) \equiv \mathbb{E}[F_T(\theta)]' \ \Sigma_{ff}^{-1} \ \mathbb{E}[F_T(\theta)]$ and apply Theorem 2.1 of Newey and McFadden (1994) to show $\hat{\theta} \xrightarrow{p} \theta_0$. The theorem requires that (i) $Q_0(\theta)$ is uniquely maximized at θ_0; (ii) Θ is compact; (iii) $Q_0(\theta)$ is continuous and (iv) $\hat{Q}_T(\theta)$ converges uniformly in probability to $Q_0(\theta)$. Requirement (i) is satisfied by the identification assumption in

3.1.(vi) and positive definiteness of Σ_{ff} in 3.1.(iii). Requirement (ii) is satisfied by 3.1.(v). Requirement (iii) is satisfied by 3.1.(iv). The uniform convergence requirement in (iv) follows from verifying Assumptions A1, B1, and A5 in Andrews (1987) and applying the main theorem. Assumption A1 of Andrews (1987) follows from Assumption 3.1.(v), Assumption B1 follows from Assumption 3.1.(i) and Assumption A5 follows from 3.1.(v) and 3.1.(iv). Having shown that all requirements are satisfied, we apply Theorem 2.1 of Newey and McFadden (1994) and get $\hat{\theta} \xrightarrow{p} \theta_0$. Consistency of the restricted estimator $\tilde{\theta} \to \theta_0$ then follows from the argument of Theorem 9.1 of Newey and McFadden (1994).

Next, I derive a preliminary asymptotic result characterizing the limiting distribution of the normalized partial sample moment for any block of the sample with $t = [rT] + 1, \ldots, [sT]$, $r, s \in [0, 1]$ and $s > r$. Start again from the constrained GMM estimator $\tilde{\theta}$ defined in equation (7). Define the following Lagrangian for $\tilde{\theta}$:

$$\begin{aligned} \tilde{\theta} &= \arg\max_{\theta \in \Theta} \mathcal{L}_T(\theta, \mu) & \mathcal{L}_T(\theta, \mu) &= \frac{1}{2} F_T(\theta)' W_T F_T(\theta) + a(\theta)' \mu_T \\ a(\theta) &:= A\,\theta = \beta & A &= \begin{bmatrix} I_p & 0_{p \times q} \end{bmatrix} \end{aligned}$$

where μ_T is a $(p \times 1)$ vector of Lagrangian multipliers which will be non-zero if the constraints are binding. The first-order conditions of this optimization problem are

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \sqrt{T}\, \nabla_\theta F_T(\tilde{\theta})' W_T F_T(\tilde{\theta}) - \nabla_\theta a(\tilde{\theta})' \sqrt{T} \tilde{\mu}_T \\ a(\tilde{\theta}) \end{bmatrix} \tag{C.2}$$

An element-by-element mean value expansion of $f(z_t, \theta)$ around θ_0, evaluated at $\tilde{\theta}$ yields

$$f(z_t, \tilde{\theta}) = f(z_t, \theta_0) + \frac{\partial f(z_t, \bar{\theta})}{\partial \theta} (\tilde{\theta} - \theta_0)$$

where $\bar{\theta} = [\bar{\theta}^{(1)}, \ldots, \bar{\theta}^{(v)}]'$ and $\bar{\theta}^{(i)} = \alpha^{(i)} \tilde{\theta}^{(i)} + (1 - \alpha^{(i)}) \theta_0^{(i)}$ for some $\alpha^{(i)} \in [0, 1]$ and each $t = 1, \ldots, T$ and $i = 1, \ldots, k$. Summing these terms from 1 to T, dividing by T and pre-multiplying by $\sqrt{T}$ gives

$$\sqrt{T}\, F_T(\tilde{\theta}) = \sqrt{T} F_T(\theta_0) + \nabla_\theta F_T(\bar{\theta})\, \sqrt{T}(\tilde{\theta} - \theta_0) \tag{C.3}$$

A similar mean-value expansion of $a(\theta)$ about θ_0, evaluated at $\tilde{\theta}$ gives

$$\sqrt{T}\, a(\tilde{\theta}) = \sqrt{T} a(\theta_0) + A\, \sqrt{T}(\tilde{\theta} - \theta_0) \tag{C.4}$$

Substituting the expansions into the first order conditions and rearranging yields

$$\begin{bmatrix} -\sqrt{T}\,\nabla_\theta F_T(\bar{\theta})' W_T F_T(\theta_0) \\ -\sqrt{T} a(\theta_0) \end{bmatrix} = \begin{bmatrix} \nabla_\theta F_T(\tilde{\theta})' W_T \nabla_\theta F_T(\bar{\theta}) & A' \\ A & 0 \end{bmatrix} \begin{bmatrix} \sqrt{T}(\tilde{\theta} - \theta_0) \\ \sqrt{T}\tilde{\mu}_T \end{bmatrix} \tag{C.5}$$

Using the consistency result proved above that $\tilde{\theta} \xrightarrow{p} \theta_0$ and uniform convergence of $\nabla_\theta F_T(\theta)$ following from the assumptions of the theorem, we get

$$\begin{bmatrix} -M'\Sigma^{-1/2}\,\Sigma_{ff}^{-1/2}\,\sqrt{T}F_T(\theta_0) \\ -\sqrt{T}a(\theta_0) \end{bmatrix} = \begin{bmatrix} D & A' \\ A & 0 \end{bmatrix} \begin{bmatrix} \sqrt{T}(\tilde{\theta} - \theta_0) \\ \sqrt{T}\tilde{\mu}_T \end{bmatrix} + o_p \tag{C.6}$$

where $D \equiv M'\Sigma^{-1}M = \bar{M}'\bar{M}$.
From the formula for inverses of block matrices, we have

$$\begin{bmatrix} D & A' \\ A & 0 \end{bmatrix}^{-1} = \begin{bmatrix} D^{-1/2}(I-P)D^{-1/2} & D^{-1}A'(AD^{-1}A')^{-1} \\ (AD^{-1}A')^{-1}AD^{-1} & -(AD^{-1}A')^{-1} \end{bmatrix} \tag{C.7}$$

where $P \equiv D^{-1/2}A'(AD^{-1}A')^{-1}AD^{-1/2}$ is an $m \times m$ idempotent matrix of rank q. Solving for $\sqrt{T}(\tilde{\theta} - \theta_0)$ using the formula for the block-inverse and re-arranging, we get

$$\begin{aligned} \sqrt{T}(\tilde{\theta} - \theta_0) = &- D^{-1/2}(I-P)D^{-1/2}\bar{M}'\,\Sigma_{ff}^{-1/2}\,\sqrt{T}F_T(\theta_0) \\ &- D^{-1}A'(AD^{-1}A')^{-1}\sqrt{T}a(\theta_0) + o_p \end{aligned} \tag{C.8}$$

Next, we calculate an alternative form for $f(z_t, \tilde{\theta})$. An element-by-element mean-value expansion of $f(z_t, \tilde{\theta})$ around θ_0 gives

$$\sqrt{T}f(z_t, \tilde{\theta}) = \sqrt{T}f(z_t, \theta_0) + \nabla_\theta f(z_t, \bar{\theta})\,\sqrt{T}(\tilde{\theta} - \theta_0)$$

where again $\bar{\theta} = [\bar{\theta}^{(1)}, \ldots, \bar{\theta}^{(v)}]'$ and $\bar{\theta}^{(i)} = \alpha^{(i)}\tilde{\theta}^{(i)} + (1 - \alpha^{(i)})\theta_0^{(i)}$ for some $\alpha^{(i)} \in [0,1]$ and each $t = 1, \ldots, T$ and $i = 1, \ldots, k$.

Take any $r, s \in [0,1]$ with $s > r$. Summing the expansion above between $[rT] + 1$ and $[sT]$,

multiplying by $\frac{1}{\sqrt{T}} W_T^{1/2}$ and using the expression for $\sqrt{T}(\tilde{\theta} - \theta_0)$ (C.8), we get

$$\begin{aligned}\frac{1}{\sqrt{T}}\, W_T^{1/2} \sum_{t=[rT]+1}^{[sT]} f(z_t, \tilde{\theta}) = \frac{1}{\sqrt{T}}\, W_T^{1/2} \sum_{t=[rT]+1}^{[sT]} f(z_t, \theta_0) \\ - W_T^{1/2}\, (1/T) \sum_{t=[rT]+1}^{[sT]} \nabla_\theta f_t(\bar{\theta}) D^{-1/2}(I - P) D^{-1/2} \bar{M}' \\ \times \sqrt{T}\, \Sigma_{ff}^{-1/2}\, F_T(\theta_0) \\ - W_T^{1/2}\, (1/T) \sum_{t=[rT]+1}^{[sT]} \nabla_\theta f_t(\bar{\theta}) D^{-1} A' (A D^{-1} A')^{-1} \sqrt{T} a(\theta_0) + o_p\end{aligned} \tag{C.9}$$

where under the null hypothesis $a(\theta_0) = 0_{p\times 1}$ so that the third term disappears.

To derive the limiting distribution, we inspect the convergence of each component of the sum above. First, note that since $\tilde{\theta} \xrightarrow{p} \theta_0$, it follows that $\bar{\theta} \xrightarrow{p} \theta_0$ and under the assumptions of the theorem we have that

$$(1/T) \sum_{t=[rT]+1}^{[sT]} \nabla_\theta f_t(\bar{\theta}) \xrightarrow{p} (s - r) \cdot M \tag{C.10}$$

Further, one can show that

$$\bar{M} D^{-1/2}(I - P) D^{1/2} \bar{M}' = \bar{P}_\delta \tag{C.11}$$

where $\bar{M}$ and $\bar{P}_\delta$ are as defined in the main text of Section 3.

Inspect the first term of the expression in (C.9). Rewriting the partial sum as a difference of two partial sums, applying the result proved in C.1 above as well as the continous mapping theorem, we get

$$\begin{aligned}\frac{1}{\sqrt{T}}\, W_T^{1/2} \sum_{t=[rT]+1}^{[sT]} f(z_t, \theta_0) = \frac{1}{\sqrt{T}}\, W_T^{1/2} \sum_{t=1}^{[sT]} f(z_t, \theta_0) - \frac{1}{\sqrt{T}}\, W_T^{1/2} \sum_{t=1}^{[rT]} f(z_t, \theta_0) \\ \Rightarrow \mathcal{B}_m(s) - \mathcal{B}_m(r)\end{aligned} \tag{C.12}$$

Next, inspect the second term of the sum in (C.9). By the same result in (C.1), we have that $\sqrt{T}\, \Sigma_{ff}^{-1/2} F_T(\theta_0) \Rightarrow \mathcal{B}_m(1)$. Then, using the results in equations (C.10) and (C.11) as

well as the continous mapping theorem, we have that

$$
\begin{aligned}
& W_T^{1/2}\ (1/T) \sum_{t=[rT]+1}^{[sT]} \nabla_\theta f_t(\bar{\theta}) D^{-1/2}(I-P)D^{-1/2}\bar{M}'\ \sqrt{T}\ \Sigma_{ff}^{-1/2}\ F_T(\theta_0) \\
& \Rightarrow (s-r)\ \bar{P}_\delta\, \mathcal{B}_m(1)
\end{aligned}
\tag{C.13}
$$

Using the two convergence results in (C.12) and (C.13), the continous mapping theorem and regrouping terms, we get

$$
T^{-1/2}\ W_T^{1/2} \sum_{t=[rT]+1}^{[sT]} f(z_t, \tilde{\theta}) \ \Rightarrow\ \mathcal{Z}(r,s) \tag{C.14}
$$

$$
\mathcal{Z}(r,s) \equiv \bar{P}_\delta\big[\,\mathcal{B}\,\mathcal{B}_m(s) - \mathcal{B}\,\mathcal{B}_m(r)\big] + (I_m - \bar{P}_\delta)\big[\,\mathcal{B}_m(s) - \mathcal{B}_m(r)\big] \tag{C.15}
$$

for any $r, s \in [0,1]$ with $s > r$ where $\mathcal{B}\,\mathcal{B}_m(s) := \mathcal{B}_m(l) - l\,\mathcal{B}_m(1)$ denotes a $m \times 1$ vector of independent Brownian bridges for $l \in [0,1]$.

Finally, I derive the limiting distribution of the $\sup \Phi_T^{LM}(K)$ test statistic. To characaterize the limiting distribution, I follow the strategy employed in Sowell (1996) by deriving a continous functional[29] mapping from $D[0,1]^m$ to $\mathbb{R}$ which defines the test statistic when applied to the normalized partial sum of the sample moments between $t = [rT]+1, \ldots, [sT]$ for some $r, s \in [0,1], s > r$ in (C.14). The same continous functional is then applied to the limiting stochastic process $\mathcal{Z}(r,s)$ defined in (C.14) to characterize the limiting distribution of the test statistic under the null hypothesis which follows from the continous mapping theorem.

Consider the following functional defining the $\sup \Phi_T^{LM}(K)$ test statistic for given K and given $\lambda^K \in \Lambda_\epsilon$ where $\lambda^K \equiv (\lambda_1, \ldots, \lambda_K)$, $\lambda_0 \equiv 0$ and $\lambda_{K+1} \equiv 1$ where the consistent variance estimators $\hat{\Sigma}_{ff}$ and $\hat{\Omega}_{T,j}$ have been replaced by their limits.

$$
\begin{aligned}
\sup \Phi_T^{LM}(K) &:= \sup_{\lambda^K \in \Lambda_\epsilon} \sum_{j=1}^{K+1} \mathcal{F}_{T,j}(\lambda_{j-1}, \lambda_j)'\ \ \Omega_{T,j}(\lambda_{j-1}, \lambda_j)\ \ \ \mathcal{F}_{T,j}(\lambda_{j-1}, \lambda_j) \\
\mathcal{F}_{j,T}(\lambda_{j-1}, \lambda_j) &:= \bar{M}_\beta'(I_m - \bar{P}_\delta)\ \ \times \frac{1}{\sqrt{T}}\, \Sigma_{ff}^{-1/2} \sum_{t=[\lambda_{j-1}T]+1}^{[\lambda_j T]} f(z_t, \tilde{\theta}) \\
\Omega_{j,T}(\lambda_{j-1}, \lambda_j) &:= (\lambda_j - \lambda_{j-1})^{-1}\ \big[\bar{M}_\beta'\ (I_m - \bar{P}_\delta)\ \bar{M}_\beta\big]^{-1}
\end{aligned}
$$

Apply this functional to the limiting stochastic process $\mathcal{Z}(r,s)$ defined in (C.14) to characterize the limiting distribution of the test statistic. The limiting stochastic process is defined

[29] Continuous with respect to the uniform metric.

by the mapping

$$\sup \Phi^{LM}(K) := \sup_{\lambda^K \in \Lambda_\epsilon} \sum_{j=1}^{K+1} A_j(\lambda)' \, V_j(\lambda)^{-1} \, A_j(\lambda)$$
$$A_j(\lambda) := \bar{M}_\beta'(I_m - \bar{P}_\delta) \, Z(\lambda_{j-1}, \lambda_j)$$
$$V_j(\lambda) := (\lambda_j - \lambda_{j-1}) \, \bar{M}_\beta' \, (I_m - \bar{P}_\delta) \, \bar{M}_\beta$$

Consider first $A_j(\lambda)$. Using the properties of the projection matrix, $\bar{P}_\delta$, we have

$$\begin{aligned} A_j(\lambda) &= \bar{M}_\beta'(I_m - \bar{P}_\delta) \;\; \mathcal{Z}(\lambda_{j-1}, \lambda_j) \\ &= \bar{M}_\beta'(I_m - \bar{P}_\delta) \left\{ \bar{P}_\delta \big[\mathcal{B}\mathcal{B}_m(\lambda_j) - \mathcal{B}\mathcal{B}_m(\lambda_{j-1}) \big] + (I_m - \bar{P}_\delta) \big[\mathcal{B}_m(\lambda_j) - \mathcal{B}_m(\lambda_{j-1}) \big] \right\} \\ &= \bar{M}_\beta'(I_m - \bar{P}_\delta) \, \big[\mathcal{B}_m(\lambda_j) - \mathcal{B}_m(\lambda_{j-1}) \big] \end{aligned}$$

Define $C := \bar{M}_\beta'(I_m - \bar{P}_\delta) \in \mathbb{R}^{p \times m}$. Using the result above, we have that

$$\begin{aligned} A_j(\lambda)' \, V_j(\lambda)^{-1} \, A_j(\lambda) &= \big[C \;\; \mathcal{Z}(\lambda_{j-1}, \lambda_j) \big]' \times (\lambda_j - \lambda_{j-1})^{-1} \, (CC')^{-1} \;\; \times \big[C \;\; \mathcal{Z}(\lambda_{j-1}, \lambda_j) \big]' \\ &= \left\{ (\lambda_j - \lambda_{j-1})^{-1/2} \, (CC')^{-1/2} C \, \big[\mathcal{B}_m(\lambda_j) - \mathcal{B}_m(\lambda_{j-1}) \big] \right\}' \times \\ &\quad \left\{ (\lambda_j - \lambda_{j-1})^{-1/2} \, (CC')^{-1/2} C \, \big[\mathcal{B}_m(\lambda_j) - \mathcal{B}_m(\lambda_{j-1}) \big] \right\} \end{aligned}$$

where the last step follows since $(CC')^{-1}$ is a square, symmetric and positive-semidefinite matrix and therefore has a matrix square root $(CC')^{-1} = (CC')^{-1/2} \, (CC')^{-1/2}$ (Newey and McFadden, 1994, Lemma 9.6).
Next, it is easy to verify that $(CC')^{-1/2}C$ is an orthonormal $p \times m$ matrix so that it holds that $\big\{(CC')^{-1/2}C\big\} \;\; \big\{(CC')^{-1/2}C\big\}' = I_p$. Since $(CC')^{-1/2}C$ is orthonormal, $(CC')^{-1/2}C \, \mathcal{B}_m(s)$ has the same distribution as $\mathcal{B}_p(s)$ and we have

$$\begin{aligned} A_j(\lambda)' \, V_j(\lambda)^{-1} \, A_j(\lambda) &= [B_p(\lambda_j) - B_p(\lambda_{j-1})]' \;\; (\lambda_j - \lambda_{j-1})^{-1} \;\; [B_p(\lambda_j) - B_p(\lambda_{j-1})] \\ &= \frac{\|B_p(\lambda_j) - B_p(\lambda_{j-1})\|^2}{\lambda_j - \lambda_{j-1}} \end{aligned}$$

and therefore

$$\sup \Phi^{LM}(K) = \sup_{\lambda^K \in \Lambda_\epsilon} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j) - \mathcal{B}_p(\lambda_{j-1})\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

so that in conclusion we have shown that

$$\sup \Phi_T^{LM}(K) \;\Rightarrow\; \sup_{\lambda^K \in \Lambda_\epsilon} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j) - \mathcal{B}_p(\lambda_{j-1})\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

The limiting distribution of the $D \sup \Phi_T^{LM}$ statistic follows then from the continuity of the max operator in (11) and the continous mapping theorem. This concludes the proof of Theorem 3.1 for the Lagrange-Multiplier form.

Wald Form

The proof for the Wald form follows the same strategy as above, first showing consistency of $\hat{\beta}_j \xrightarrow{p} \theta_0$ under the null hypothesis, then deriving the limiting stochastic process of $\sqrt{T}(\hat{\beta}_j - \beta_0)$ based on the estimator defined in (9) and finally applying a continous mapping to form the test statistic and to characterize its limiting distribution. The full proof is available on request.

□

C.2 Out-Of-Sample Inference

Lemma 4.1 (OOS Mean-Value Approximation): *Under the regularity conditions in Assumption 4.1 and the null hypothesis defined in (3), for any $r, s \in [0,1]$ with $s > r > \rho$ it holds that*

$$P^{-1/2} \sum_{t=[rT]+1}^{[sT]} f(z_{t+h}, \beta_0, \hat{\delta}_t) = (T/P)^{1/2} \left\{ \frac{1}{\sqrt{T}} \sum_{t=R}^{[sT]} f(z_{t+h}, \beta_0, \delta_0) - \frac{1}{\sqrt{T}} \sum_{t=R}^{[rT]} f(z_{t+h}, \beta_0, \delta_0) \right\}$$
$$+ (T/P)^{1/2}\, FB \left\{ \frac{1}{\sqrt{T}} \sum_{t=R}^{[sT]} H_t(\delta_0) - \frac{1}{\sqrt{T}} \sum_{t=R}^{[rT]} H_t(\delta_0) \right\} + o_{p,rs}(1)$$

where H_t, B are as defined in Assumption 4.1.(ii) and $x_t(r,s) = o_{p,rs}(1)$ denotes that $\sup_{r,s\in[0,1],\, s>r>\rho} \|x_t(r,s)\| = o_p(1)$.

Proof. A second-order element-by-element mean value expansion of $f(z_{t+h}, \beta, \delta)$ around δ_0 and evaluated at $\hat{\delta}_t$ for $t = R, \ldots, T$ yields.

$$f(z_{t+h}, \beta_0, \hat{\delta}_t) = f(z_{t+h}, \beta_0, \delta_0) + \nabla_\delta f(z_{t+h}, \beta_0, \delta_0)(\hat{\delta}_t - \delta_0) + w_{t+h} \tag{C.16}$$

where the i-th element of w_{t+h} is

$$w_{t+h,i} \equiv \frac{1}{2}(\hat{\delta}_t - \delta_0)' \, \frac{\partial^2 f_i(z_{t+h}, \beta_0, \tilde{\delta}_{t,i})}{\partial \delta \, \partial \delta'} \, (\hat{\delta}_t - \delta_0) \tag{C.17}$$

and $\tilde{\delta}_{t,i}$ lies between $\hat{\delta}_t$ and δ_0.

Summing between $t = [rT]+1, \ldots, [sT]$ and pre-multiplying by $P^{-1/2}$ gives

$$\begin{aligned} P^{-1/2} \sum_{t=[rT]+1}^{[sT]} f(z_{t+h}, \beta_0, \hat{\delta}_t) = P^{-1/2} & \sum_{t=[rT]+1}^{[sT]} f(z_{t+h}, \beta_0, \delta_0) \\ & + P^{-1/2} \sum_{t=[rT]+1}^{[sT]} \nabla_\delta f(z_{t+h}, \beta_0, \delta_0)(\hat{\delta}_t - \delta_0) \\ & + P^{-1/2} \sum_{t=[rT]+1}^{[sT]} w_{t+h} \end{aligned} \tag{C.18}$$

The second term in (C.18) can be written

$$\begin{aligned} P^{-1/2} \sum_{t=[rT]+1}^{[sT]} \nabla_\delta f(z_{t+h}, \beta_0, \delta_0)(\hat{\delta}_t - \delta_0) &= P^{-1/2} \sum_{t=[rT]+1}^{[sT]} \nabla_\delta f(z_{t+h}, \beta_0, \delta_0) B_t H_t(\delta_0) \\ &= P^{-1/2} F B \sum_{t=[rT]+1}^{[sT]} H_t(\delta_0) \\ &\quad + P^{-1/2} \sum_{t=[rT]+1}^{[sT]} \big(\nabla_\delta f(z_{t+h}, \beta_0, \delta_0) - F\big) B H_t(\delta_0) \\ &\quad + P^{-1/2} \sum_{t=[rT]+1}^{[sT]} F\big(B_t - B\big) H_t(\delta_0) \\ &\quad + P^{-1/2} \sum_{t=[rT]+1}^{[sT]} \big(\nabla_\delta f(z_{t+h}, \beta_0, \delta_0) - F\big)\big(B_t - B\big) H_t(\delta_0) \end{aligned} \tag{C.19}$$

where the first step follows from Assumption 4.1.(ii) and the second step by adding and subtracting the relevant terms involving F and B. The second term after the last equality is $o_{p,rs}(1)$ by Assumption 4.1.(xi), the third by part 4.1.(xii) and the fourth by 4.1.(xiii). Further, it can be shown that the remainder term $P^{-1/2} \sum_{t=[rT]+1}^{[sT]} w_{t+h}(r,s) = o_{p,rs}(1)$ from an argument similar to the one in the proof of equation (4.1) in West (1996).

Substituting into (C.18) gives

$$P^{-1/2} \sum_{t=[rT]+1}^{[sT]} f(z_{t+h}, \beta_0, \hat{\delta}_t) = P^{-1/2} \sum_{t=[rT]+1}^{[sT]} f(z_{t+h}, \beta_0, \delta_0) + P^{-1/2} F B \sum_{t=[rT]+1}^{[sT]} H_t(\delta_0) + o_{p,rs}(1) \tag{C.20}$$

Finally, multiplying and dividing by $\sqrt{T}$, splitting the sums and re-arraging terms gives

$$P^{-1/2}\sum_{t=[rT]+1}^{[sT]} f(z_{t+h},\beta_0,\hat{\delta}_t) = (T/P)^{1/2}\left\{\frac{1}{\sqrt{T}}\sum_{t=R}^{[sT]} f(z_{t+h},\beta_0,\delta_0) - \frac{1}{\sqrt{T}}\sum_{t=R}^{[rT]} f(z_{t+h},\beta_0,\delta_0)\right\}$$
$$+ (T/P)^{1/2}\, FB\left\{\frac{1}{\sqrt{T}}\sum_{t=R}^{[sT]} H_t(\delta_0) - \frac{1}{\sqrt{T}}\sum_{t=R}^{[rT]} H_t(\delta_0)\right\} + o_{p,rs}(1)$$

which proves the Lemma. □

THEOREM 4.1 (OOS Inference): *Assume that the regularity conditions in Assumption 4.1 hold. Under the null hypothesis defined in (3), it holds that*

$$\sup \Phi_T(K) \;\Rightarrow\; \sup_{\lambda^K \in \Lambda_{\epsilon,\rho}} \sum_{j=1}^{K+1} \Phi_j(\lambda_{j-1},\lambda_j)$$
$$D\sup \Phi_T(\bar{K}) \;\Rightarrow\; \max_{1\le k\le \bar{K}} (1/k) \sup_{\lambda^K \in \Lambda_{\epsilon,\rho}} \sum_{j=1}^{K+1} \Phi_j(\lambda_{j-1},\lambda_j)$$
$$\Lambda_{\epsilon,\rho} = \Big\{\lambda_j,\ j=1,\ldots,K:\ \lambda_j \in (\rho+\epsilon, 1-\epsilon),\ \lambda_j > \lambda_{j-1}+\epsilon\Big\}, \quad \lambda_0 \equiv \rho,\ \lambda_{K+1} \equiv 1$$

where

$$\Phi_j(\lambda_{j-1},\lambda_j) \equiv \Big[\,\mathcal{B}_m\left(\int_0^{\lambda_j} \omega(u,\lambda_{j-1},\lambda_j)\omega(u,\lambda_{j-1},\lambda_j)'\,du\right)\Big]' \times$$
$$\left\{\int_0^{\lambda_j} \omega(u,\lambda_{j-1},\lambda_j)\,\omega(u,\lambda_{j-1},\lambda_j)'\,du\right\}^{-1}$$
$$\times\Big[\,\mathcal{B}_m\left(\int_0^{\lambda_j} \omega(u,\lambda_{j-1},\lambda_j)\omega(u,\lambda_{j-1},\lambda_j)'\,du\right)\Big]$$

with

$$\omega(u,r,s) \equiv M'\,\Sigma_{ff}^{-1}\,(1-\rho)^{-1/2}\Big[I_m \quad FB\Big] \times \Big\{\left[\Omega(u,s)^{1/2} - \Omega(u,r)^{1/2}\right]\mathbb{1}(u\le r)$$
$$+\,\Omega(u,s)^{1/2}\,\mathbb{1}(r<u\le s)\Big\}\,\Sigma^{1/2}$$

and where $\Omega(s,\tau)$ is as defined as

$$\Omega(s,\tau)^{1/2} \equiv \begin{pmatrix} \mathbb{1}(s\le\rho)\cdot I_m & 0_{m\times d} \\ 0_{d\times m} & \left\{[\ln\tau - \ln\rho]\,\mathbb{1}(s\le\rho) + [\ln(\tau) - \ln(s)]\,\mathbb{1}(\rho<s\le\tau)\right\}\cdot I_d \end{pmatrix}$$

Proof. Note as in the in-sample case above that under the null hypothesis in (3), it holds that $\theta_t = \theta_0 = (0_{p\times 1},\delta)\ \forall\ t$. Therefore, under the null hypothesis the function defining the

moment condition $f(z_t, \theta_t) = f(z_t, \beta_t, \delta)$ can be written as a function of a constant parameter $f(z_t, \theta)$. This notation will be used in the subsequent derivations.

To prove the weak convergence results stated in the theorem, I start by showing that the partial sample moments satisfy the following invariance principle under the null hypothesis.

$$\frac{1}{\sqrt{T}} \sum_{t=R}^{[sT]} \begin{pmatrix} f(z_{t+h}, \beta_0, \delta_0) \\ H_t(\delta_0) \end{pmatrix} \Rightarrow \int_0^s \Omega(u,s)^{1/2} \, d\xi(u) \tag{C.21}$$

where $\Omega(u,\tau)^{1/2}$ is as defined in the theorem and $\xi(u) \equiv \Sigma^{1/2} \mathcal{B}_{m+d}(u)$ where $\mathcal{B}_{m+d}(u)$ is an $(m+d) \times 1$ vector of independent standard Brownian motions.

This result can be proven by following the same reasoning as in the proof of Proposition 1 of Rossi and Sekhposyan (2016). Start by defining $b_{R,t,j} \equiv \mathbb{1}(t \geq R)$. Then, by direct calculations, for any $j \geq R$, it holds that

$$\sum_{t=R}^{j} f(z_{t+h}, \beta_0, \delta_0) = \sum_{t=1}^{j} b_{R,t,j} \; f(z_{t+h}, \beta_0, \delta_0) \tag{C.22}$$

Under the recursive estimation scheme in Assumption 4.1.(ii) and defining

$$a_{R,t,j} \equiv \left(R^{-1} + \ldots + j^{-1}\right) \cdot \mathbb{1}(t \leq R) \; + \; \left(t^{-1} + \ldots + j^{-1}\right) \cdot \mathbb{1}(R < t \leq j) \tag{C.23}$$

for any $j \geq R$ it holds by direct calculation that

$$\sum_{t=R}^{j} H_t(\delta_0) = \sum_{t=R}^{j} t^{-1} \left(\sum_{r=1}^{t} h(z_r, \delta_0) \right) = \sum_{t=1}^{j} a_{R,t,j} \; h(z_t, \delta_0) \tag{C.24}$$

Using (C.22) and (C.24) it holds that

$$\frac{1}{\sqrt{T}} \sum_{t=R}^{[sT]} \begin{pmatrix} f(z_{t+h}, \beta_0, \delta_0) \\ H_t(\delta_0) \end{pmatrix} = \frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} \begin{pmatrix} b_{R,t,[sT]} \cdot I_m & 0_{m \times d} \\ 0_{d \times m} & a_{R,t,[sT]} \cdot I_d \end{pmatrix} \begin{pmatrix} f(z_{t+h}, \beta_0, \delta_0) \\ h_t(\delta_0) \end{pmatrix} \tag{C.25}$$

To derive the limiting distribution, as in Rossi and Sekhposyan (2016), I consider an asymptotic approximation for the weights $a_{R,t,j}$ and $b_{R,t,j}$. From Assumption 4.1.(i), we have $\rho := \lim_{T \to \infty} R/T$ and thus

$$b_{R,t,j} \equiv \mathbb{1}(t \geq R) \approxeq \mathbb{1}(s \geq \rho) \qquad s \equiv \lim_{T \to \infty} t/T \tag{C.26}$$

Following West (1996) and Rossi and Sekhposyan (2016) it can further be shown that

$$\begin{aligned} a_{R,t,j} &\approxeq \left(\int_R^j \frac{1}{k}\, dk\right) \mathbb{1}(t \leq R) + \left(\int_t^j \frac{1}{k}\, dk\right) \mathbb{1}(R < t \leq j) \\ &\approxeq [\ln(\tau) - \ln(\rho)]\, \mathbb{1}(s \leq \rho) + [\ln(\tau) - \ln(s)]\, \mathbb{1}(\rho < s \leq \tau) \end{aligned} \tag{C.27}$$

To prove the weak convergence in (C.21), I employ the result for weak convergence of stochastic integrals based on mixing sequences of Hansen (1992). In particular, define $\{\xi_{j,T}\}$ to be the following normalized stochastic sum process

$$\xi_j \equiv \frac{1}{\sqrt{T}} \sum_{t=1}^{j} \xi(z_{t+h}, \beta_0, \delta_0) \equiv \frac{1}{\sqrt{T}} \sum_{t=1}^{j} \begin{pmatrix} f(z_{t+h}, \beta_0, \delta_0) \\ h_t(\delta_0) \end{pmatrix} \tag{C.28}$$

where $\xi(z_{t+h}, \theta)$ is defined in the main text of Section 4. Further, define the stochastic integral of interest as

$$\int_0^\tau \begin{pmatrix} \sigma_f(s) \cdot I_m & 0_{m\times d} \\ 0_{d\times m} & \sigma_h(s,\tau) \cdot I_d \end{pmatrix} d\, \xi_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{j} \begin{pmatrix} b_{R,t,j} \cdot I_m & 0 \\ 0 & a_{R,t,j} \cdot I_d \end{pmatrix} \begin{pmatrix} f(z_{t+h}, \beta_0, \delta_0) \\ h_t(\delta_0) \end{pmatrix} \tag{C.29}$$

To apply Theorem 3.1 of Hansen (1992) we need to verify its conditions. The first requirement is Assumption 1 of Hansen (1992) which is satisfied by the mixing condition in Assumption 4.1.(iii) and Assumption 4.1.(v). To satisfy the second requirement, we need to show that $T^{1/2}\ \xi_T \Rightarrow \Sigma^{-1/2}\xi$ where $\xi(s) \equiv \Sigma^{1/2}\, \mathcal{B}_{m+d}(s)$. This follows from applying Corollary 2.2 of Phillips and Durlauf (1986) under Assumptions 4.1.(iii), 4.1.(iv), 4.1.(v) and 4.1.(vi). Applying Theorem 3.1 of Hansen (1992), we get

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{j} \begin{pmatrix} b_{R,t,j} \cdot I_m & 0 \\ 0 & a_{R,t,j} \cdot I_d \end{pmatrix} \begin{pmatrix} f(z_{t+h}, \beta_0, \delta_0) \\ h_t(\delta_0) \end{pmatrix} - C_T^*(\tau) \ \Rightarrow\ \int_0^\tau \Omega(s,\tau)^{1/2}\, d\xi(s) \tag{C.30}$$

where

$$\begin{aligned} C_T^*(\tau) = \Bigg\{ T^{-1/2} \sum_{t=1}^{[\tau T]} \Bigg[&\begin{pmatrix} b_{R,t,j} \cdot I_m & 0 \\ 0 & a_{R,t,j} \cdot I_d \end{pmatrix} - \begin{pmatrix} b_{R,t-1,j} \cdot I_m & 0 \\ 0 & a_{R,t-1,j} \cdot I_d \end{pmatrix} \Bigg] \zeta_t \\ - T^{-1/2} &\begin{pmatrix} b_{R,t-1,j} \cdot I_m & 0 \\ 0 & a_{R,t-1,j} \cdot I_d \end{pmatrix} \zeta_{j+1} \Bigg\} \end{aligned}$$

with $j := [\tau T]$ and $\zeta_t = \sum_{k=1}^{\infty} \mathbb{E}_t \left(\left[f(z_{t+h+k}, \theta_0)', h(z_t, \delta_0)' \right]' \right)$. Using the same reasoning as in Rossi and Sekhposyan (2016), based on the steps in the proof of Cavaliere (2005), Theorem

4 and the fact that the variances $\sigma_f(s), \sigma_h(s,\tau)$ are square integrable and bounded, we get

$$\frac{1}{\sqrt{T}}\sum_{t=R}^{[sT]}\begin{pmatrix} f(z_{t+h},\beta_0,\delta_0) \\ H_t(\delta_0)\end{pmatrix} \Rightarrow \int_0^{\tau}\Omega(s,\tau)^{1/2}\,d\xi(s) \tag{C.31}$$

Lagrange-Multiplier Form, Φ_T^{LM}

To derive the limiting distribution of the $\sup\Phi_T(K)$ test statistic, we follow the same strategy as in the proof of Theorem 3.1. The test statistic is formed by applying a continuous functional to the stochastic process derived above. We then derive the associated limiting stochastic process under the null hypothesis and apply the same functional to characterize the limting distribution of the test statistic. Given K define the continous mapping

$$\sup\Phi_T^{LM}(K) := \sup_{\lambda^K\in\Lambda_{\epsilon,\rho}}\sum_{j=1}^{K+1}A_{j,T}(\lambda)'\left\{V_{j,T}(\lambda)\right\}^{-1}A_{j,T}(\lambda) \tag{C.32}$$

where

$$A_{j,T}(\lambda) := M'\Sigma_{ff}^{-1}\,P^{-1/2}\sum_{t=[\lambda_{j-1}T]+1}^{[\lambda_jT]}f(z_{t+h},\beta_0,\hat{\delta}_t) \tag{C.33}$$

To characterize the limiting stochastic process of $A_{j,T}(\lambda)$, note that under the regularity conditions in Assumptions 4.1, we can apply Lemma 4.1 to get

$$\begin{aligned}P^{-1/2}\sum_{t=[\lambda_{j-1}T]+1}^{[\lambda_jT]}f(z_{t+h},\beta_0,\hat{\delta}_t) = (T/P)^{1/2}&\left\{\frac{1}{\sqrt{T}}\sum_{t=R}^{[\lambda_jT]}f(z_{t+h},\beta_0,\delta_0)-\frac{1}{\sqrt{T}}\sum_{t=R}^{[\lambda_{j-1}T]}f(z_{t+h},\beta_0,\delta_0)\right\}\\ &+(T/P)^{1/2}\,FB\left\{\frac{1}{\sqrt{T}}\sum_{t=R}^{[\lambda_jT]}H_t(\delta_0)-\frac{1}{\sqrt{T}}\sum_{t=R}^{[\lambda_{j-1}T]}H_t(\delta_0)\right\}+o_{p,rs}(1)\end{aligned}$$

Plugging into (C.33) and grouping terms and omitting the o_p term, we get

$$A_{j,T}(\lambda) = M'\Sigma_{ff}^{-1}(T/P)^{1/2}\begin{bmatrix}I_m & FB\end{bmatrix}\left\{\frac{1}{\sqrt{T}}\sum_{t=R}^{[\lambda_jT]}\begin{pmatrix}f(z_{t+h},\beta_0,\delta_0)\\ h_t(\delta_0)\end{pmatrix}-\frac{1}{\sqrt{T}}\sum_{t=R}^{[\lambda_{j-1}T]}\begin{pmatrix}f(z_{t+h},\beta_0,\delta_0)\\ h_t(\delta_0)\end{pmatrix}\right\} \tag{C.34}$$

Using the weak convergence result we derived in (C.31) above, we get

$$\frac{1}{\sqrt{T}}\sum_{t=R}^{[\lambda_jT]}\begin{pmatrix} f(z_{t+h},\beta_0,\delta_0) \\ H_t(\delta_0)\end{pmatrix} \Rightarrow \int_0^{\lambda_j}\Omega(s,\lambda_j)^{1/2}S^{1/2}\,d\,\mathcal{B}_{m+d}(s) \tag{C.35}$$

and applying the continous mapping theorem,

$$
\begin{aligned}
A_{j,T} \;\Rightarrow\; & M'\Sigma_{ff}^{-1}(1-\rho)^{-1/2}\begin{bmatrix} I_m & FB \end{bmatrix}\left\{ \int_0^{\lambda_j} \Omega(s,\lambda_j)^{1/2} S^{1/2}\, d\mathcal{B}_{m+d}(s) \right. \\
& \left. - \int_0^{\lambda_{j-1}} \Omega(s,\lambda_{j-1})^{1/2} S^{1/2}\, d\mathcal{B}_{m+d}(s) \right\} \\
& = M'\Sigma_{ff}^{-1}(1-\rho)^{-1/2}\begin{bmatrix} I_m & FB \end{bmatrix}\left\{ \int_0^{\lambda_{j-1}} \left[\Omega(s,\lambda_j)^{1/2} - \Omega(s,\lambda_{j-1})^{1/2}\right] S^{1/2}\, d\mathcal{B}_{m+d}(s) \right. \\
& \left. + \int_{\lambda_{j-1}}^{\lambda_j} \Omega(s,\lambda_j)^{1/2} S^{1/2}\, d\mathcal{B}_{m+d}(s) \right\} \\
& = M'\Sigma_{ff}^{-1}(1-\rho)^{-1/2}\begin{bmatrix} I_m & FB \end{bmatrix} \times \\
& \int_0^{\lambda_j} \left\{ \left[\Omega(s,\lambda_j)^{1/2} - \Omega(s,\lambda_{j-1})^{1/2}\right] \mathbb{1}(s \le \lambda_{j-1}) + \Omega(s,\lambda_j)^{1/2}\, \mathbb{1}(\lambda_{j-1} < s \le \lambda_j) \right\} S^{1/2}\, d\mathcal{B}_{m+d}(s) \\
& = \int_0^{\lambda_j} \omega(s,\lambda_j,\lambda_{j-1})\, d\mathcal{B}_{m+d}(s) \\
& = \mathcal{B}_m\left(\int_0^{\lambda_j} \omega(s,\lambda_j,\lambda_{j-1})\omega(s,\lambda_j,\lambda_{j-1})'\, ds \right)
\end{aligned}
\tag{C.36}
$$

where

$$
\begin{aligned}
\omega(s,\lambda_{j-1},\lambda_j) \equiv M'\Sigma_{ff}^{-1}(1-\rho)^{-1/2}\begin{bmatrix} I_m & FB \end{bmatrix}\left\{ \left[\Omega(s,\lambda_j)^{1/2} - \Omega(s,\lambda_{j-1})^{1/2}\right] \mathbb{1}(s \le \lambda_{j-1}) \right. \\
\left. + \Omega(s,\lambda_j)^{1/2}\, \mathbb{1}(\lambda_{j-1} < s \le \lambda_j) \right\} S^{1/2}
\end{aligned}
\tag{C.37}
$$

The result of the theorem then follows from applying the continous mapping theorem to get the distribution of $\Phi_T(\lambda)$ and the distributions of $\sup \Phi_T(K)$ and $D \sup \Phi_T(\bar{K})$ as defined in the theorem and analog to the proof of Theorem 3.1.

Wald Form

The proof of the Wald form is available on request.

□

Corollary 4.1 (OOS Inference in Special Cases): *If (a) $F = 0$, that is parameter estimation error is irrelevant, or (b) the following condition holds*

$$
\Sigma_{ff} = -\frac{1}{2}(FB\Sigma_{hf} + \Sigma_{fh}B'F') = FB\Sigma_{hh}B'F'
$$

then, the result of Theorem 4.1 simplifies to

$$\sup \Phi_T(K) \;\Rightarrow\; \sup_{\lambda \,\in\, \Lambda_{\epsilon,\rho}} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j - \rho) - \mathcal{B}_p(\lambda_{j-1} - \rho)\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

$$D \sup \Phi_T(\bar{K}) \;\Rightarrow\; \max_{1 \le k \le \bar{K}} (1/k) \sup_{\lambda \,\in\, \Lambda_{\epsilon,\rho}} \sum_{j=1}^{K+1} \left\{ \frac{\|\mathcal{B}_p(\lambda_j - \rho) - \mathcal{B}_p(\lambda_{j-1} - \rho)\|^2}{\lambda_j - \lambda_{j-1}} \right\}$$

Proof. The proof follows from directly calculating $\int_0^{\lambda_j} \omega(s, \lambda_j, \lambda_{j-1})\omega(s, \lambda_j, \lambda_{j-1})' \, ds$, imposing the condition given in the corollary. The proof follows similar steps to the proofs of Proposition 3,4, & 7 of Rossi and Sekhposyan (2016). □