Pre and Post Break Parameter Inference[*]

Graham Elliott

University of California, San Diego

9500 Gilman Drive

La Jolla, CA, 92093-0508

Ulrich K. Müller

Princeton University

Princeton, NJ, 08544-1021

August 2012

Abstract

This paper provides a method for conducting inference about the pre and post break value of a scalar parameter in GMM time series models with a single break at an unknown date. We show that treating the break date estimated by least squares as the true break date leads to substantially oversized tests and confidence intervals unless the break is large. We develop an alternative test that controls size uniformly and that is approximately efficient in a well defined sense.

JEL Classification: C32, C12

Keywords: Nuisance Parameters; Least Favorable Distribution; Size of Test; Monte Carlo; Uniformity

1 Introduction

This paper is concerned with testing hypotheses about the pre and post break value of a parameter in a time series model with a single break. By reversing the time ordering, inference about the post break value becomes inference about the pre break value, so that for simplicity, we will exclusively discuss the post break case. If the break date is known, inference is straightforward, as standard results apply after restricting attention to the stable post break data. The effect of an incorrectly chosen break date crucially depends on whether the chosen date is earlier or later than the true break date. On the one hand, if the chosen date is later, then a restriction to the presumed post break data still yields a stable model, and inference remains valid. There is a cost of efficiency, though, since more post break data could have been used. On the other hand, if the chosen date is earlier than the true break date, then the presumed post break data stems from an unstable model. Parameter estimators in this unstable model tend to estimate the average parameter value, which is different from the true post break value, so standard inference yields tests and confidence intervals with distorted size.

These distortionary effects are small if the chosen break date is sufficiently close to the true break date. With the break date unknown, this requires a precise break date estimator. As formally shown by Bai (1994, 1997) and Bai and Perron (1998) for linear regressions, and Hall, Han, and Boldea (2010) and Perron and Yamamoto (2008) for two stage least squares, the least squares break date estimator is indeed sufficiently precise for these distortionary effects to become asymptotically negligible *if* the parameter shift is sufficiently pronounced. Formally, these papers study asymptotics in which the break magnitude, while possibly shrinking, is outside the local $T^{-1/2}$ neighborhood of zero. Relative to the sampling uncertainty of the parameter estimator in a stable model, the parameter shift thus becomes infinitely large. We consider the behavior of the 5% nominal level two-sided test based on the least squares break date estimator under local asymptotics in Section 2.1.2 below, where the break magnitude is measured in multiples of standard deviations of the full sample parameter estimator in a stable model. With the break date restricted to the middle 70% of the sample, the largest null rejection probability is almost 30%, a break magnitude of less than 5 standard deviations leads to effective size of more than 10% for all break dates, and a break of 11 standard deviations still yields size greater than 10% for some break dates. These distortions are further exacerbated by attempts to pre-test for parameter stability. In other words, for any sample size, there exists a break magnitude for which these standard methods yield incorrect inference, so they are not uniformly valid.

It is instructive to consider these issues in the context of an example. Productivity growth in most developed nations appears to have suffered shifts in its mean a number of times. Most researchers agree on a downward shift in productivity following the oil crisis of 1973, typically dated at the third quarter. In post 1973 data, there is a suggestion of a mid 1990's upward shift in average productivity growth and speculation about its cause. A recent summary appears in Jorgenson, Ho, and Stiroh (2008), who informally suggest a break in the fourth quarter of 1995. Gordon (2003) also dates the increase to 1995. Anderson and Klieson (2006) refer to the increase in productivity as the 'defining economic event of the
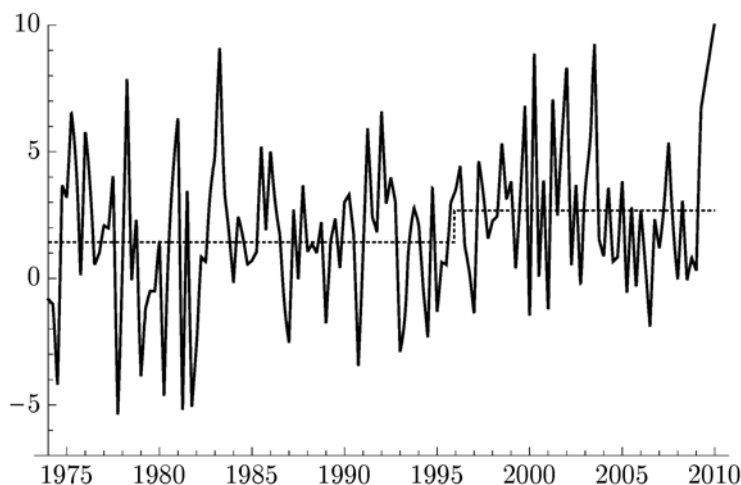
Figure 1: Quarterly growth in the non farm business sector and average growth pre and post 1995Q4 (Fred Database series OPHNFB)

past decade'.

Figure 1 plots quarterly observations of US non farm business productivity growth, along with the pre and post break means using Jorgenson, Ho, and Stiroh (2008) break date of 1995 Q4. With this break date, quarterly productivity growth is estimated to have increased by 1.1% in 1995 Q4. The least-squares estimate of the break date is slightly later at 1997 Q1, although the parameter stability test of Elliott and Müller (2006) fails to reject the null hypothesis of stable mean growth at the 5% level. With a (long-run) standard deviation of productivity growth of approximately 2.6% and a sample of $T = 143$ observations, 5-11 standard deviations of the full sample estimator correspond to a 1.0-2.3% shift in productivity growth. In this example, one would thus not want to assume that the true break magnitude is sufficiently large for least-squares break date based inference to be reliable. These concerns are corroborated in a small sample Monte Carlo study modelled around the productivity example in Section 4.3. This illustrates that lack of uniform asymptotic validity directly translates into poor small sample properties with practically relevant parameter values.

More generally, it seems hard to argue that break magnitudes of less than 11 standard deviations of the full sample estimator are not part of the empirically relevant parameter space in most applications. One way to see this is that shifts of, say, 8 standard deviations tend to induce highly significant rejections of the null hypothesis of parameter stability with high probability—see Elliott and Müller (2007). But in practice, rejections of the stable model are rarely overwhelming. For instance, there is continued debate about the stability of monetary policy with Orphanides (2004) arguing for rather stable relationships, while Cogley and Sargent (2005) find instabilities that they deem 'economically important', but of a magnitude that would be detected by a formal parameter stability test less than 25% of the time.

Simple adjustments to the standard procedure that ensure size control over all break magnitudes

2

do not deliver reasonable tests. For instance, in the set-up mentioned above and described in detail in Section 2.1.2, the critical value for the 5% two-sided t-statistic would need to be increased from 1.96 to approximately 4.9 to ensure uniform size control, with obvious adverse effects on power.[1] A Bonferroni procedure based on uniformly valid confidence sets for the break date developed by Elliott and Müller (2007) performs well for large breaks, but has poor power for breaks of moderate magnitude. Also, given the results of Andrews and Guggenberger (2010a), it seems clear that standard bootstrap or subsampling techniques would not yield uniformly valid inference in this problem.

The main contribution of this paper is the construction of a powerful test about the post break parameter value in a general GMM time series model with unknown break date in the middle 70% of the sample that controls size uniformly over the break magnitude. This test follows a switching rule: if there is strong evidence for a large break, then the test essentially reduces to standard inference using post break data determined by a break date estimator, with a slight increase of the 5% critical value from 1.96 to 2.01. In absence of strong evidence for a large break, the test switches to a likelihood ratio test. This likelihood ratio test is carefully constructed to ensure both overall size control and approximate efficiency in the sense that for a particular weighting function, (local asymptotic) weighted average power of the suggested test is demonstrably at most 1.0% smaller than the weighted average power of any test that controls size. Section 4.1 gives a fully self-contained explanation of how to construct the test statistic and undertake the test with data, and is the relevant section for readers wishing to apply the method to data.

From a statistical point of view, the problem is one of constructing powerful tests in the presence of a nuisance parameter under the null hypothesis—the nuisance parameter being the break data and break magnitude, neither of which can be consistently estimated under local asymptotics. We apply the approach of Elliott, Müller, and Watson (2012), which consists of identifying an approximate least favorable distribution for this two dimensional nuisance parameter. This approximate least favorable distribution corresponds to the null density for the likelihood ratio part of the test statistic, and it allows the construction of an upper bound on the power of all tests that control size. Beyond overcoming the 'engineering' challenges in this example with a two dimensional nuisance parameter, exacerbated by the impossibility of accurately representing continuous time Gaussian processes on a discrete computer, a further contribution of this paper to this methodology is a careful consideration of size control. A test controls size if the rejection probability, viewed as a function on the null parameter space, takes on values below or at the nominal level. The standard approach is to evaluate the rejection probability for a grid of parameter points by Monte Carlo, and to conclude size is controlled if none of the estimates is larger than the nominal level. We develop an alternative algorithm that, while still based on Monte Carlo estimation, instead estimates (an upper bound on) the whole rejection probability function, ensuring that no additional implicit smoothness assumptions are necessary. It is important to note, however, that

---

[1] This adjustment to the critical value is also the end result of the size corrected hybrid subsampling method advanced by Andrews and Guggenberger (2009, 2010b) when applied to this problem.

all of these 'engineering' challenges only concern the *derivation* of the suggested test; its *application* for a given sample is essentially no more difficult than the estimation of the GMM model over 142 subsamples, for any sample size.

In many ways, the analysis in this paper mirrors the developments in the weak instrument literature over the last decade: Similar to Staiger and Stock (1997), we consider an alternative asymptotic embedding that formalizes the lack of uniform validity of the standard method. Akin to their rule of thumb that standard asymptotics are approximately reliable when the first stage F-statistic is sufficiently large, we suggest a formal switching rule to inference based on a break date estimator when a parameter stability test is highly significant. Finally, as Andrews, Moreira and Stock (2006, 2008), we consider tests that are efficient relative to a weighted average power criterion, and identify an approximately optimal test.

Returning to the US productivity example, Table 1 contains nominal 95% confidence intervals for current US productivity growth, constructed using different methods. The first method assumes the Jorgenson, Ho, and Stiroh (2008) break date to be correct and reports the post 1995 Q4 sample mean ±1.96 times the standard error. The second and third method are the same, with the break date conservatively chosen as 2004Q1, and with a break date determined by minimizing the squared residuals. The fourth method also relies on the least-squares break date, but uses a critical value of 4.9 instead of 1.96 to ensure uniform asymptotic coverage control, as discussed in detail in Section 2.1.2 below. The final row reports the 95% confidence interval constructed by inverting the 5% level test developed in this paper.

As discussed above, in this example it is not reasonable to argue that the true break magnitude is large enough for the (uncorrected) least-squares based method to yield reliable inference. In contrast, the last three methods are asymptotically uniformly valid. Among those valid methods, the inference developed here is by far the most informative. The resulting confidence interval is nevertheless wider than those with the break date presumed known, and its center is shifted towards smaller values. This makes intuitive sense: With unknown break date, part of the low productivity growth quarters of, say, the early 90s are potentially realizations from the post break period, which suggests a relatively lower post break population mean.

The remainder of the paper is organized as follows. The next section defines the asymptotic analogues to the small sample post break parameter inference problem, quantifies the lack of size control of inference based on the least squares break date estimator, summarizes our solution and discusses the relationship of the small sample problem in parametric and GMM models to the limiting problem. In Section 3, we discuss the construction of the suggested test, the determination of its size and the bound on power. The test statistic is defined in terms of partial sample GMM statistics (cf. Andrews (1993)) in Section 4, where we also consider its asymptotic and small sample properties. Proofs not given in the main text are collected in an appendix.

Table 1: Inference about current US productivity growth

Method	95% Confidence Interval
Break date chosen at 1995Q4	[2.00,3.36]
Least squares break date estimator	[2.02,3.49]
Corrected LS break date estimator	[0.92,4.59]
Using last 15% of data	[0.89,2.99]
Test of this paper	[1.74,3.43]

Note: Sample size is $T = 143$. All confidence intervals use Newey and West (1987) standard errors with 4 lags.

2 Limiting Problem and Underlying Models

The natural large sample analogue to small sample post break parameter inference is an inference problem about the drift component of a Gaussian process on the unit interval. The scalar version of this limiting problem is discussed first. We investigate the corresponding asymptotic properties of least squares break date based inference, and provide an overview of our line of attack, along with the key properties of the suggested test. The second subsection considers the multivariate limiting problem, and shows in which sense the approximate efficiency of our solution in the scalar case carries over to the multivariate case. Finally, Section 2.3 provides a formal link between the underlying small sample problem to the multivariate limiting problem for both parametric and GMM time series models.

2.1 Scalar Limiting Problem

2.1.1 Statement of the problem

The main focus of this paper is the following hypothesis testing problem: Suppose we observe the scalar Gaussian process $G(\cdot)$ on the unit interval (a random element with values in $D_{[0,1]}$, the space of cadlag functions on the unit interval)

$$G(s) = W(s) + \beta s + \delta \min(\rho, s), \qquad s \in [0, 1] \tag{1}$$

where $W(\cdot)$ is standard Wiener process and $\beta, \delta \in \mathbb{R}$, and $\rho \in [0.15, 0.85]$. The process $G(s)$ is a Wiener process with drift $\beta + \delta$ for $s < \rho$, and drift β for $s \geq \rho$. The testing problem is $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$, so that geometrically, the question is whether the slope of the deterministic component $\beta s + \delta \min(\rho, s)$ is equal to zero after the (potential) kink at $s = \rho$. The location ρ and magnitude δ of the kink are nuisance parameters. This composite nature of the hypotheses can be made more explicit by introducing the notation $\theta = (\beta, \delta, \rho)$, $\Theta_0 = \{\theta = (0, \delta, \rho) : \delta \in \mathbb{R} \text{ and } \rho \in [0.15, 0.85]\}$ and $\Theta_1 = \{\theta = (\beta, \delta, \rho) : \beta \neq 0, \delta \in \mathbb{R} \text{ and } \rho \in [0.15, 0.85]\}$, so that the hypothesis testing problem becomes

$$H_0 : \theta \in \Theta_0 \qquad \text{against} \qquad H_1 : \theta \in \Theta_1. \tag{2}$$

5

Let $\varphi : D_{[0,1]} \mapsto [0,1]$ be a possibly randomized test of (2), where $\varphi(g)$ indicates the probability of rejecting the null hypothesis conditional on observing $G = g$. If the range of φ only consists of the two values $\{0,1\}$, then φ is not randomized. The unconditional rejection probability is then simply the expectation of $\varphi(G)$, which we write as $E_\theta[\varphi(G)]$ to indicate that this expectation depends on the value of θ. With these definitions, a test φ is of level 5% if

$$E_\theta[\varphi(G)] \leq 5\% \text{ for all } \theta \in \Theta_0. \tag{3}$$

To illustrate the link to the post break parameter inference problem, consider observations $\{y_t\}_{t=1}^T$ from a scalar Gaussian model with a break in the mean at time τ

$$y_t = \mu + \Delta \mathbf{1}[t \leq \tau] + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1). \tag{4}$$

We then have the following equality in distribution

$$T^{-1/2} \sum_{t=1}^{sT} y_t \sim G(s) \quad \text{for any } s \in \{\tfrac{t}{T}\}_{t=1}^T$$

where $\rho = \tau/T$, $\beta = T^{1/2}\mu$ and $\delta = T^{1/2}\Delta$. Except for the scaling by $T^{-1/2}$ and the discretization $s \in \{\tfrac{t}{T}\}_{t=1}^T$, the testing problem (2) involving the observation G is therefore identical to inference about the post-break mean in Gaussian location model (4), with the break date restricted to be in the middle 70% of all observations. We argue below how the testing problem (2) based on the observation $G(\cdot)$ naturally arises as the asymptotically relevant problem in more complicated settings, but first discuss this limiting problem in more detail.

2.1.2 Tests Based on Least-Squares Break Date Estimator

As discussed in the introduction, Bai (1994, 1997) and Bai and Perron (1998) suggest and analyze the following procedure for conducting inference about the post break value of a coefficient in a linear regression that is subject to a break: Estimate the break date by least-squares, construct a post break dummy using this estimate, and perform the usual t-test on that dummy. These papers show that this procedure results in asymptotically correct inference for break magnitudes that diverge when multiplied by the square root of the sample size. This corresponds to $|\delta| \to \infty$ in the notation developed here. Another procedure, arguably most prevalent in applied work, first performs a pretest for a break on some conventional significance level, followed by standard full sample inference about the post break value if no significant break is found, and performs least-squares based post break inference as above if the pretest rejects.

In the limiting problem (1), the nominally 5% level least squares and pretest procedure (based on the supF statistic) correspond to the test

$$\mathbf{1}[\text{supF} > \text{cv}_{pre}(\alpha_{pre})]\mathbf{1}[|\hat{t}| > \Phi(0.975)] + \mathbf{1}[\text{supF} < \text{cv}_{pre}(\alpha_{pre})]\mathbf{1}[|G(1)| > \Phi(0.975)] \tag{5}$$

where $\mathrm{cv}_{pre}(\alpha_{pre})$ is the critical value of the supF statistic of level α_{pre} as tabulated in Andrews (1993), (and $\mathrm{cv}_{pre}(1) = 0$ corresponds to always using the least squares break date estimator, with no pretest), $\Phi(0.975) \simeq 1.96$ is the 97.5th percentile of a standard normal and

$$\hat{\rho} = \mathrm{argmax}_{r \in [0.15; 0.85]} \frac{(G(r) - rG(1))^2}{r(1-r)}, \quad \mathrm{supF} = \frac{(G(\hat{\rho}) - \hat{\rho}G(1))^2}{\hat{\rho}(1-\hat{\rho})}, \quad \hat{t} = \frac{G(1) - G(\hat{\rho})}{\sqrt{1-\hat{\rho}}}. \quad (6)$$

Under standard assumptions, the large sample properties of the small sample pretest and least square based tests in a linear regression with a single coefficient converge in distribution to those in (5) for breaks of magnitude $T^{-1/2}$—this follows, for instance, from Proposition 1 of Elliott and Müller (2007).

Figure 1 displays the null rejection probability of (5) for $\alpha_{pre} = 1$ (no pretest), $\alpha_{pre} = 0.10$, $\alpha_{pre} = 0.05$ and $\alpha_{pre} = 0.01$ as a function of δ for selected values of ρ, based on 50,000 Monte Carlo draws and 1,000 step random walk approximations to Wiener processes. None of these tests comes close to controlling size uniformly over δ. The approximately largest null rejection probability of the pure least-squares break date based test is approximately 29% at $(\rho, \delta) = (0.85, 2.6)$. Pre-testing for a break seems to both substantially exacerbate and shift the size control problem to larger values of δ.[2]

In fact, to obtain a 5% level test in the no pretest case, one must employ a critical value of approximately 4.9 instead of $\Phi(0.975) \simeq 1.96$ for the t-statistic, with size still approximately equal to the nominal 5% level at the point $(\rho, \delta) = (0.85, 6.8)$. Denote this size corrected test φ_{LS}. Because the worst size distortion occurs for a strictly positive break magnitude, the size corrected hybrid subsampling method of Andrews and Guggenberger (2007a, 2009) reduces to the size corrected fixed critical value test φ_{LS}. Alternatively, one could always use the test $\varphi_{0.85}(G) = \mathbf{1}[|G(1) - G(0.85)|/\sqrt{0.15} > 1.96]$ which corresponds to a usual 5% level t-test based on the last 15% of the data. As one would expect, though, the power of both φ_{LS} and $\varphi_{0.85}$ is quite poor (we provide power computations in Section 4 below), motivating the construction of an alternative test. The construction of this test in the following subsection closely follows Elliott, Müller, and Watson (2012).

2.1.3 Approximately Weighted Average Power Maximizing Test

No uniformly most powerful test exists for inference about β in absence of knowledge of ρ and δ. Some tests will have good power for certain ranges of value of ρ and δ, while performing poorly for other values. To obtain a reasonable assessment of the overall quality of tests, we therefore adopt a weighted average power criterion, as, for instance, in Wald (1943) and Andrews and Ploberger (1994). Specifically, we seek to determine a 5% level test φ that comes close to maximizing

$$\mathrm{WAP}(\varphi) = \int E_\theta[\varphi(G)]dF(\theta) \quad (7)$$

where the weighting function F is such that ρ is distributed uniform over $[0.15, 0.85]$ and (β, δ) is bivariate normal independent of ρ, with $\beta \sim \mathcal{N}(0, \sigma_\beta^2)$, $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$ and β independent of $\beta + \delta$, where $\sigma_\beta^2 = 22$ and $\sigma_\delta^2 = 400$.

---

[2]Unreported results show that these asymptotic results provide very good approximations for the small sample Gaussian location problem (4) with $T = 100$.
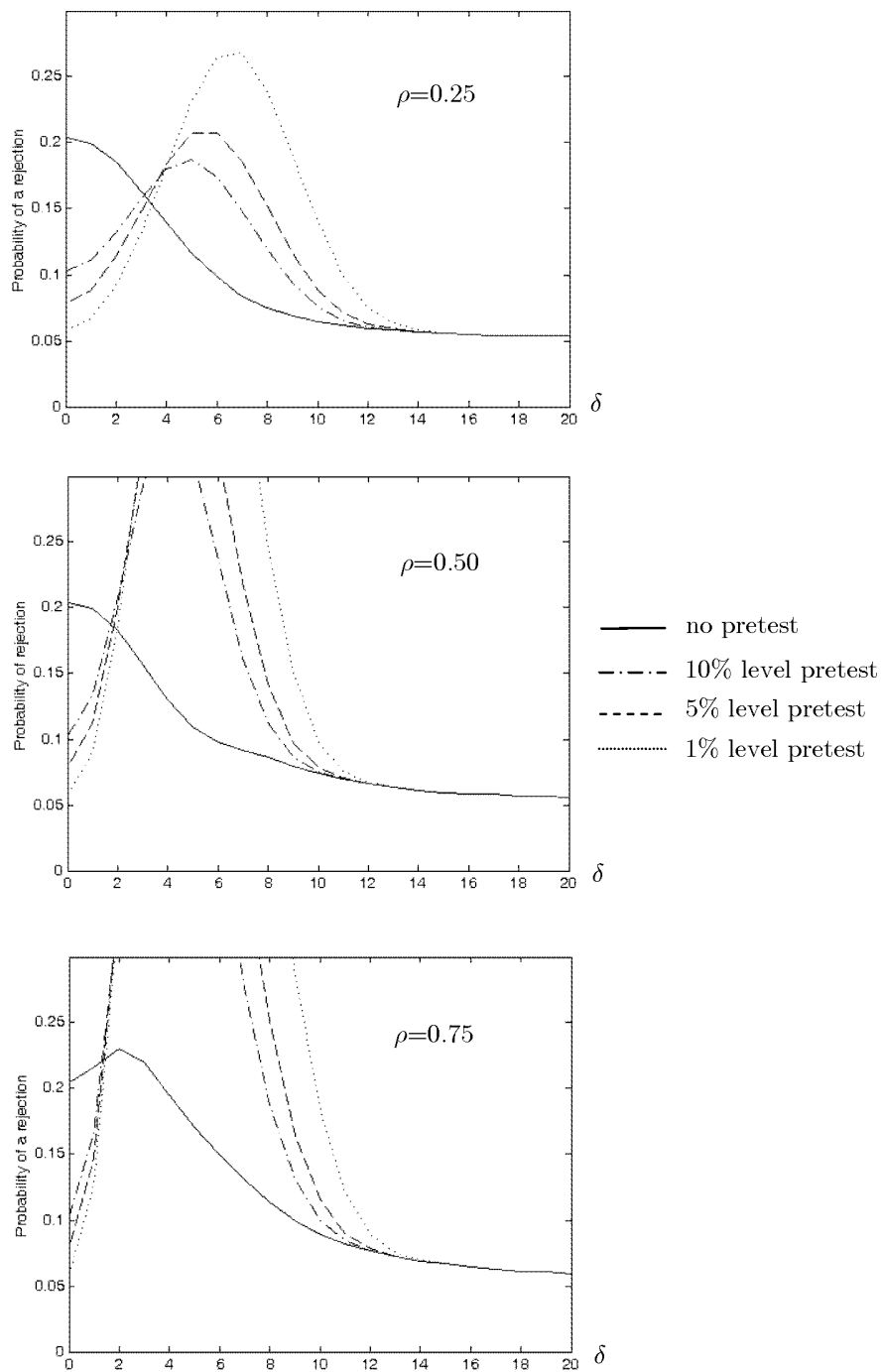
Figure 2: Asymptotic null rejection probabilities of nominal 5% level tests for post break parameter value based on least squares break date estimator.

The Gaussianity of the marginal distribution of β and δ in F is chosen for analytical convenience, and the independence of the pre and post break parameter values β and $\beta + \delta$ mirrors the independence of the pre and post break sample information $G(\rho)$ and $G(1) - G(\rho)$. The value of $\sigma_\beta^2 = 22$ is motivated by King's (1988) discussion of overall reasonable point-optimal tests, since it turns out that for $\sigma_\beta^2 = 22$, the best 5% level test has approximately 50% weighted average power. The uniform weighting over ρ accords to the choice in Andrews and Ploberger (1994) and is intended to generate reasonable power for all break dates $\rho \in [0.15, 0.85]$. Finally, the value $\sigma_\delta^2 = 400$ is motivated as follows: For large δ, (say, $|\delta| > 12$ or so) there is a lot of information about the break date ρ, and good tests can come close to performing as well as if ρ was known. By putting enough weight on rather large values of δ, the choice $\sigma_\delta^2 = 400$ ensures that tests that perform well according to WAP share this desirable feature. At the same time, the distribution $\mathcal{N}(0, 400)$ concentrates about half of its mass on $|\delta| < 12$, so tests with high WAP must also perform reasonably in the arguably empirically relevant region for δ where the lack of knowledge of the break date severely complicates good inference about β.

By Girsanov's Theorem, the Radon-Nikodym derivative of the measure of G in (1) with parameter $\theta \in \Theta_0 \cup \Theta_1$ relative to the measure ν of the standard Wiener process W, evaluated at G, is given by

$$f_\theta(G) = \exp[\delta G(\rho) + \beta G(1) - \tfrac{1}{2}(\beta^2 + 2\beta\delta\rho + \delta^2\rho)]. \tag{8}$$

Expressed in terms of the density f_θ, the hypothesis test (2) becomes

$$\begin{aligned}
&H_0 : \text{The density of } G \text{ is } f_\theta,\ \theta \in \Theta_0 \\
&H_1 : \text{The density of } G \text{ is } f_\theta,\ \theta \in \Theta_1
\end{aligned} \tag{9}$$

and, as noted above, both the null and alternative hypothesis involve the two nuisance parameters δ and ρ. Note, however, that weighted average power can be rewritten as

$$\text{WAP}(\varphi) = \int E_\theta[\varphi(G)] dF(\theta) = \int \varphi(\int f_\theta dF(\theta)) d\nu$$

as Fubini's Theorem justifies the change of order of integration. Thus, maximizing weighted average power is equivalent to maximizing power against the single alternative

$$H_F : \text{The density of } G \text{ is } h = \int f_\theta dF(\theta) \tag{10}$$

and the remaining challenge is to deal with the composite nature of the null hypothesis. The key insight that allows us to make further progress is Lemma 1 of Elliott, Müller, and Watson (2012),[3] which we reproduce here for convenience.

Lemma 1 *Let Λ be a probability distribution on Θ_0, and let φ_Λ be the best level α test of the null hypothesis H_Λ : "The density of G is $\int f_\theta d\Lambda(\theta)$" against H_F. Then for any level α test φ of H_0 against H_F, $\int \varphi_\Lambda h d\nu \geq \int \varphi h d\nu$.*

---

[3]The same basic insight was already applied by Andrews, Moreira, and Stock (2008) in the weak instrument problem.

Proof. Since φ is a level α test of H_0, $\int \varphi f_\theta d\nu \leq \alpha$ for all $\theta \in \Theta_0$. Therefore, $\int \int \varphi f_\theta d\nu d\Lambda(\theta) = \int \int \varphi f_\theta d\Lambda(\theta) d\nu \leq \alpha$ (where the change in the order of integration is allowed by Fubini's Theorem), so that φ is also a level α test of H_Λ against H_1. The result follows by the definition of a best test. $\blacksquare$

Lemma 1 formalizes the perfectly intuitive result that replacing the composite null hypothesis H_0 with the single mixture null hypothesis H_Λ can only simplify the testing problem in the sense of allowing for more powerful tests. Its appeal lies in the fact that the best test of H_Λ against H_F is easily constructed: by the Neyman-Pearson Lemma, the best test rejects for large values of the likelihood ratio statistic $h/ \int f_\theta d\Lambda(\theta)$. Thus, Lemma 1 provides a set of explicit power bounds on the original problem, indexed by the distribution Λ.

Lemma 1 can usefully be thought of as generalizing a standard result concerning tests with a composite null hypothesis; see, for instance, Theorem 3.8.1 of Lehmann and Romano (2005): A distribution Λ^{**} is *least favorable* if the best level α test of $H_{\Lambda^{**}}$ against the single alternative H_F is also of level α in the testing problem with the composite null hypothesis H_0 against H_F, so that—using the same reasoning as in the proof of Lemma 1—this test is also the best test of H_0 against H_F. In contrast to this standard result, Lemma 1 is formulated without any restriction on the probability distribution Λ. This is useful because in many contexts, it is difficult to identify the least favorable distribution Λ^{**} (and it may not even exist).

The strategy suggested by Lemma 1 is to work instead with a numerically determined *approximately least favorable* distribution Λ^*: Suppose we knew of a Λ^* such that (i) the best 5% level test of H_{Λ^*} against H_F, φ_{Λ^*}, has power 50.0% and (ii) the best 4.7% level test $\tilde{\varphi}_{\Lambda^*}$ of H_{Λ^*} against H_F has power 49.0% and $\tilde{\varphi}_{\Lambda^*}$ is a 5% level test of H_0. Using Lemma 1, this would imply that no 5% level test of H_0 against H_1 can exist with weighted average power larger than 50.0%, and at the same time, we would have identified the 5% level test $\tilde{\varphi}_{\Lambda^*}$ of H_0 against H_1 whose weighted average power is only 1.0% lower than this bound.

This is the basic idea underlying the construction of our suggested test φ^* detailed in Section 3. For reasons discussed there, the test φ^* is only roughly equivalent to a Neyman-Pearson test $\tilde{\varphi}_{\Lambda^*}$ relative to an approximately least favorable distribution. This is irrelevant for the logic outlined above, only the validity of the power bound via Lemma 1 depends on its derivation from a best (i.e. Neyman-Pearson) test.

We postpone further details of the construction of the power bound and the test φ^* to Section 3 below. The following proposition summarizes the main finding.

Proposition 1 *In model (1),*

(i) any 5% level test φ of (2) has $WAP(\varphi) \leq \bar{\pi} \simeq 50.0\%$;

(ii) the (nonrandomized) test φ^* defined in the appendix is of level $\alpha^* \simeq 5\%$, and has $WAP(\varphi^*) = \pi^* \simeq 49.0\%$.

The numbers for $\bar{\pi}$, α^* and π^* in Proposition 1 are estimated based on 50,000 independent Monte Carlo draws. As discussed in detail in Section 3, they are *not* subject to any other qualifications, such

as a replacement of the level requirement (3) by a finite set of values for (δ, ρ) or a finite step Random Walk approximation for draws of G.

Figure 3 plots, for selected values of ρ and δ, the power of the suggested test φ^*, of the size corrected least squares break date based test φ_{LS}, of the test using the last 15% of the sample $\varphi_{0.85}$ only, and the infeasible test using actual post break data φ_ρ introduced in Section 2.1.2 above. All tests are invariant under the transformation $G \to -G$, so there is not need to plot the the power for negative δ. As one might expect given the large critical value of 4.9, φ_{LS} has very low power. In contrast, φ^* is almost always more powerful than $\varphi_{0.85}$, often substantially so, and comes close to the power of the infeasible benchmark φ_ρ for large $|\delta|$.

We also investigated size control of φ^* in models that correspond to a local break where the transition to the post break parameter value is smooth rather than a sudden shift at ρ. In particular, we considered parameter paths of the form $f(s) = \delta(\Phi_{0.05}(s - \rho) - 1)$, so that the deterministic part of $G(s)$ equals $\beta s + \int_0^s f(\lambda) d\lambda$, where $\Phi_{0.05}$ is the cdf of a mean zero normal with standard deviation 0.05. For large $|\delta|$, φ^* turns out to be substantially oversized, but for $|\delta| < 15$, the rejection probability of φ^* is always below 7%, at least for $\rho \in \{0.25, 0.5, 0.75\}$. These results suggest that small sample analogues to φ^* discussed below continue to yield approximately valid inference about the post break parameter value even if the transition to the post break value takes place over a nonnegligible fraction of the sample.

2.2 Multivariate Limiting Problem

Now assume that in the hypothesis testing problem (2) we observe a multivariate version of (1), with

$$\begin{pmatrix} G(s) \\ \tilde{G}(s) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} W(s) \\ \tilde{W}_k(s) \end{pmatrix} + \int_0^s \begin{pmatrix} \beta + \delta \mathbf{1}[\lambda \le \rho] \\ \tilde{f}(\lambda) \end{pmatrix} d\lambda \tag{11}$$

where $\tilde{G}$ is $k \times 1$, the $k \times 1$ vector A_{21} and the $k \times k$ full rank matrix A_{22} are known, $\tilde{W}_k$ is a k vector standard Wiener process independent of W, and $\tilde{f}(\lambda)$ is a $D_{[0,1]} \mapsto \mathbb{R}^k$ cadlag function, so that $(G, \tilde{G}')' \in D_{[0,1]}^{k+1}$. Except for $\mathrm{Var}[G(1)] = 1$, the transformation by A_{21} and A_{22} allows for any full rank covariance matrix of $(G(1), \tilde{G}(1)')')'$. With $(G, \tilde{G}')'$ as the observation, tests φ are now mappings from $D_{[0,1]}^{k+1} \mapsto [0,1]$, and the rejection probability is given by $E_{\theta, \tilde{f}}[\varphi(G, \tilde{G})]$ (the rejection probability also depends on A_{21} and A_{22}, but we omit this dependence for ease of notation).

For the link to the post break parameter inference problem, think again of a Gaussian location problem

$$\begin{pmatrix} y_t \\ \tilde{y}_t \end{pmatrix} = \begin{pmatrix} \mu + \Delta \mathbf{1}[t \le \tau] \\ \tilde{\mu}_t \end{pmatrix} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma) \quad \text{with } \Sigma = \begin{pmatrix} 1 & A'_{21} \\ A_{21} & A_{21} A'_{21} + A_{22} A'_{22} \end{pmatrix} \tag{12}$$

where $\tilde{y}_t, \tilde{\mu}_t \in \mathbb{R}^k$, so that partial sums $T^{-1/2} \sum_{t=1}^s (y_t, \tilde{y}_t')'$ have the same distribution as $(G(s), \tilde{G}(s)')')'$ for $s \in \{t/T\}_{t=1}^T$, where $\rho = \tau/T$, $\beta = T^{1/2}\mu$, $\delta = T^{1/2}\Delta$ and $\tilde{f}(t/T) = T^{1/2}\tilde{\mu}_t$.

The first element in (11), G, is distributed exactly as in the scalar problem (1) above. Thus, if we define φ^* to be the same test as above, that is, φ^* is a function of G only, then φ^* continues to have the

11

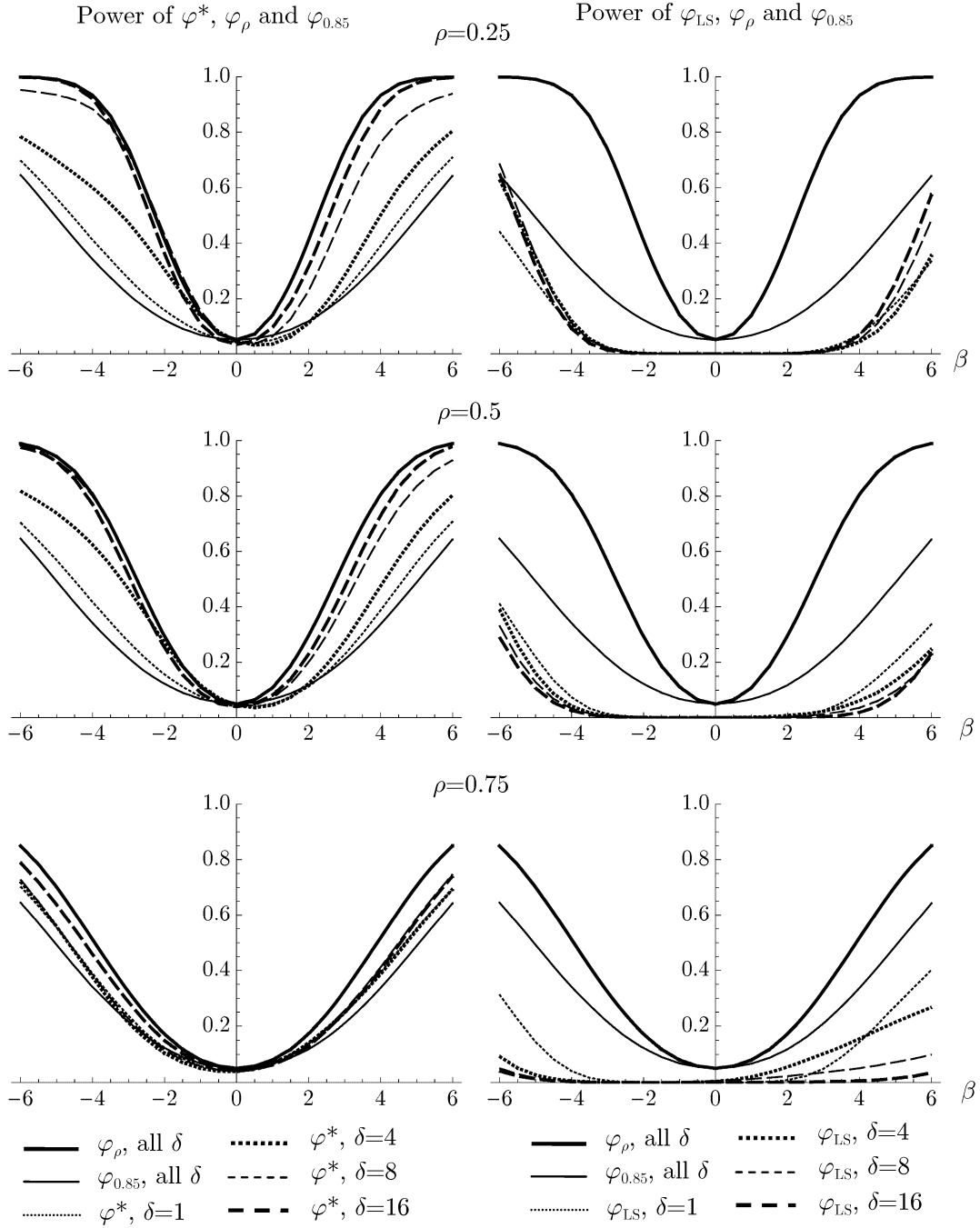Figure 3: Asymptotic power of the approximately optimal test φ^* and of the size corrected least-squares date estimator based test φ_{LS}, compared to the infeasible test φ_ρ with break date known, and the test using only the last 15% of the observations $\varphi_{0.85}$

properties indicated in Proposition 1. The rejection probability of φ^* obviously does not depend on the distribution of $\tilde{G}$, so that in particular

$$E_{\theta,\tilde{f}}[\varphi^*(G)] \quad \leq \quad \alpha^* \text{ for all } \tilde{f} \text{ and } \theta \in \Theta_0 \tag{13}$$

$$\inf_{\tilde{f}(\cdot)\in D_{[0,1]}^k} \text{WAP}_{\tilde{f}}(\varphi^*) \quad = \quad \sup_{\tilde{f}(\cdot)\in D_{[0,1]}^k} \text{WAP}_{\tilde{f}}(\varphi^*) = \pi^* \tag{14}$$

where for a generic test $\varphi : D_{[0,1]}^{k+1} \mapsto [0,1]$, $\text{WAP}_{\tilde{f}}(\varphi) = \int E_{\theta,\tilde{f}}[\varphi(G,\tilde{G})]dF(\theta)$. In other words, φ^* is a robust test of $H_0 : \theta \in \Theta_0$ from observing (11) in the sense that its size and weighted average power does not depend on the nuisance function $\tilde{f}$.

In the following, we restrict attention to the case where $\tilde{f}$ is known to be of the form $\tilde{f}(\lambda) = \tilde{\beta} + \tilde{\delta}\mathbf{1}[\lambda \leq \rho]$ for $\tilde{\beta}, \tilde{\delta} \in \mathbb{R}^k$. This corresponds to a Gaussian location problem (12) where $\tilde{y}_t$ undergoes a single mean shift at the same time as y_t. Write $E_{\theta,\tilde{\beta},\tilde{\delta}}$ for the expectation with respect to the distribution of $(G, \tilde{G}')'$ with $\tilde{f}$ of this form, so that a 5% level test $\varphi : D_{[0,1]}^{k+1} \mapsto [0,1]$ now satisfies $E_{\theta,\tilde{\beta},\tilde{\delta}}[\varphi(G,\tilde{G})] \leq 0.05$ for all $\theta \in \Theta_0$, $\tilde{\beta}, \tilde{\delta} \in \mathbb{R}^k$.[4]

Proposition 2 *In model (11),*

(i) for any unbiased 5% level test φ,

$$E_{\theta,\tilde{\beta},\tilde{\delta}}[\varphi_\rho(G)] - E_{\theta,\tilde{\beta},\tilde{\delta}}[\varphi(G,\tilde{G})] \geq 0 \text{ for all } \theta \in \Theta_1, \tilde{\beta}, \tilde{\delta} \in \mathbb{R}^k$$

where $\varphi_\rho(G) = \mathbf{1}[|\frac{G(1)-G(\rho)}{\sqrt{1-\rho}}| > \Phi(0.975)]$;

(ii) let $\tilde{F}$ be the probability distribution for $(\theta, \tilde{\beta}, \tilde{\delta})$ with the marginal distribution of θ equal to F, and $(\tilde{\beta}, \tilde{\delta}) = (\beta A_{21}, \delta A_{21})$ conditional on θ. For any 5% level test

$$\int E_{\theta,\tilde{\beta},\tilde{\delta}}[\varphi(G,\tilde{G})]d\tilde{F}(\theta,\tilde{\beta},\tilde{\delta}) \leq \bar{\pi} \simeq 50.0\%$$

(iii) suppose in addition that $\tilde{\delta} = 0$ known, so that tests $\varphi : D_{[0,1]}^{k+1} \mapsto [0,1]$ are of 5% level if $E_{\theta,\tilde{\beta},0}[\varphi(G,\tilde{G})] \leq 0.05$ for all $\theta \in \Theta_0$ and $\tilde{\beta} \in \mathbb{R}^k$. Consider the group of transformations

$$\begin{pmatrix} G(s) \\ \tilde{G}(s) \end{pmatrix} \rightarrow \begin{pmatrix} G(s) \\ \tilde{G}(s) + \tilde{b}s \end{pmatrix}, \tilde{b} \in \mathbb{R}^k. \tag{15}$$

For any 5% level test φ invariant to (15), the rejection probability $E_{\theta,\tilde{\beta},0}[\varphi(G,\tilde{G})]$ does not depend on $\tilde{\beta}$, and

$$WAP_{inv}(\varphi) = \int E_{\theta,\tilde{\beta},0}[\varphi(G,\tilde{G})]dF(\theta) \leq \bar{\pi}_{inv}(\bar{R}^2)$$

where $\bar{R}^2 \geq R^2 = A_{21}'(A_{21}A_{21}' + A_{22}A_{22}')^{-1}A_{21}$, and $\bar{\pi}_{inv}(0) = \bar{\pi} \simeq 50.0\%$, $\bar{\pi}_{inv}(0.2) \simeq 53.3\%$, $\bar{\pi}_{inv}(0.4) \simeq 56.4\%$, $\bar{\pi}_{inv}(0.6) \simeq 60.0\%$, and $\bar{\pi}_{inv}(0.8) \simeq 64.6\%$.

---

[4]Propositions 2, 3 and 4 also hold for the scalar case ($k = 0$) with the natural interpretation of notation.

Part (i) of Proposition 2 establishes that the usual two-sided t-test using post break data $G(1) - G(\rho)$, φ_ρ, is the uniformly most powerful unbiased test in the multiparameter problem with (11) observed. With ρ unknown this test is infeasible. It nevertheless provides an arguably relevant benchmark to assess the performance of other tests. In particular, note that φ_ρ, just like φ^*, does not depend on $\tilde{G}$, but is a function of G only. We compare the power of φ^* to that of φ_ρ in detail in Section 4 below, and find that for large $|\delta|$, the power becomes quite close. Thus, for large $|\delta|$, φ^* has similar rejection properties as the uniformly most powerful unbiased test, so that there is at most a small cost in terms of power for ignoring the additional information in $\tilde{G}$. (Although φ^* is not unbiased, as it is not similar—see Section 3 below.)

Part (ii) of Proposition 2 shows that for a particular weighting function $\tilde{F}$ with the same marginal distribution as F on θ, no 5% level test can have weighted average power larger than $\bar{\pi}$. At the same time, (14) implies that the weighted average power of φ^* with respect to $\tilde{F}$ is equal to π^*, and, by Proposition 1, π^* is only slightly smaller than $\bar{\pi}$. In this sense, φ^* is close to being admissible also in the hypothesis testing problem with (11) observed. Thus, any test with "substantially" larger power than φ^* for some $(\theta, \tilde{\beta}, \tilde{\delta})$ is less powerful for some other $(\theta, \tilde{\beta}, \tilde{\delta})$. For instance, it is of course true that if $||\tilde{\delta}||$ is known to be large, observing $\tilde{G}$ adds a lot of information about the common break fraction ρ. But methods that attempt to exploit this will do worse than φ^* if $||\tilde{\delta}||$ is small. A perhaps appealing feature of the weighting function $\tilde{F}$ is that it yields a test whose weighted average power over θ does not depend on the nuisance parameters $(\tilde{\beta}, \tilde{\delta})$.

Finally, part (iii) of Proposition 2 provides bounds on the gains in weighted average power that are possible if it is known that the other parameters do not undergo a break, $\tilde{\delta} = 0$ (or, more generally, for known $\tilde{\delta}$) under an invariance requirement that corresponds to changes of the mean of the additional k observations. A maximal invariant to this group is given by $(G(s), \tilde{G}(s) - s\tilde{G}(1))$. Intuitively, these power gains are possible because if $\tilde{G}$ is correlated with G ($R^2 > 0$), observing $\tilde{G}(s) - s\tilde{G}(1)$ provides information about $W(s) - sW(1)$, which is useful for for learning about ρ and δ. In the extreme case of perfect correlation ($R^2 = 1$), $W(s) - sW(1)$ is effectively observed, and by comparing $G(s) - sG(1)$ to $W(s) - sW(1)$, one can back out ρ and δ and obtain as powerful inference about β as if there was no break in the model. The numbers for $\bar{\pi}_{inv}$ were estimated using 10,000 independent Monte Carlo draws, analogous to the power computations described in detail in Section 3.4 below. They show that rather large values of R^2 are necessary before knowledge of $\tilde{\delta} = 0$ could potentially be used to generate tests with substantially larger weighted average power than φ^*.

In summary, the test φ^* that ignores the additional observation $\tilde{G}$ in (11) is (i) robust in the sense of providing reliable inference about β regardless of the nuisance function $\tilde{f}$; (ii) for large $|\delta|$ has a rejection profile that is close to that of the uniformly most powerful unbiased test; (iii) is close to admissible in a model where other parameters undergo a break at the same time; (iv) as long as R^2 is small, approximately maximizes weighted average power even when it is known that only the parameter of interest is subject to a break. For these reasons, the remainder of this paper focusses on φ^* as a

reasonable test also in the multiparameter context of (11).

2.3 Asymptotic Efficiency Implications for Underlying Models

In this subsection, we discuss how the scalar and multiparameter limiting problems (1) and (11) arise as the relevant asymptotic problems in standard small sample post break parameter inference. We consider two distinct approaches to asymptotic efficiency of tests: On the one hand, we rely on LeCam's Limits of Experiments theory to derive upper bounds on power in correctly specified parametric models. On the other hand, we exploit the recent results of Müller (2011) and provide upper bounds on post break parameter tests in time series GMM models under an asymptotic robustness constraint.

2.3.1 Parametric Models

Let $X_T = (x_{T,1}, \cdots, x_{T,T}) \in \mathbb{R}^{qT}$ be the available data in a sample of size T, so that tests φ_T are sequences of (measurable) functions $\varphi_T : \mathbb{R}^{qT} \mapsto [0,1]$. The density of X_T, relative to some σ-finite measure μ_T, is given by the product $\prod_{t=1}^{T} f_{T,t}(\Gamma_t)$, when the parameter $\Gamma \in \mathbb{R}^{k+1}$ takes on the value Γ_t at time t. This form of the likelihood arises naturally when $f_{T,t}(\Gamma_t)$ is the density of $x_{T,t}$ conditional on $\mathfrak{F}_{T,t-1}$, the σ-field generated by $\{x_{T,s}\}_{s=1}^{t-1}$. We will refer to the model with density $\prod_{t=1}^{T} f_{T,t}(\Gamma_0)$ as the 'stable' model. Define $l_{T,t}(\Gamma) = \ln f_{T,t}(\Gamma)$, $s_{T,t}(\Gamma) = \partial l_{T,t}(\Gamma)/\partial \Gamma$ and $h_{T,t} = \partial s_{T,t}(\Gamma)/\partial \Gamma'$, and suppose $T^{-1} \sum_{t=1}^{T} h_{T,t} \xrightarrow{p} -H$ in the stable model for some positive definite $(k+1) \times (k+1)$ matrix H, so that H is the Fisher information. In the unstable model, the parameter vector Γ evolves as

$$\Gamma_{T,t} = \Gamma_0 + T^{-1/2}\omega \begin{pmatrix} \beta \\ \tilde{\beta} \end{pmatrix} + T^{-1/2}\omega \begin{pmatrix} \delta \\ \tilde{\delta} \end{pmatrix} \mathbf{1}[t \le \rho T] \tag{16}$$

where ω^2 is the 1,1 element of H^{-1}, $\beta, \delta, \rho \in \mathbb{R}$ and $\tilde{\beta}, \tilde{\delta} \in \mathbb{R}^k$, and $\rho \in [0.15; 0.85]$. We assume Γ_0 and H to be known; while this unrealistic, such knowledge can only increase the upper bounds on power derived in Proposition 3 below. The hypothesis of interest is whether the post break value of the first element of Γ equals the first element of Γ_0, so under (16), this corresponds to

$$H_0 : \beta = 0 \qquad \text{against} \qquad H_1 : \beta \ne 0. \tag{17}$$

The factor $\omega > 0$ in (16) ensures that asymptotically, the small sample problem (17) maps to the limiting problem (11) where the disturbance in G is a *standard* Wiener process. Tests $\varphi_T : \mathbb{R}^{qT} \mapsto [0,1]$ of (17) have unconditional rejection probability $E_{\theta,\tilde{\beta},\tilde{\delta}}[\varphi_T(X_T)]$ in a sample of size T, and φ_T is defined to be of asymptotic level 5% if $\limsup_{T\to\infty} E_{\theta,\tilde{\beta},\tilde{\delta}}[\varphi_T(X_T)] \le 5\%$ for all $\theta \in \Theta_0$, $\tilde{\beta}, \tilde{\delta} \in \mathbb{R}^k$.

Under suitable regularity conditions on the parametric model, one can show that the likelihood ratio statistic LR_T between the model with parameter evolution described by (16) and the stable model

satisfies

$$\ln \mathrm{LR}_T = \sum_{t=1}^{T}(\Gamma_{T,t} - \Gamma_0)'s_{T,t}(\Gamma_0) + \tfrac{1}{2}\sum_{t=1}^{T}(\Gamma_{T,t} - \Gamma_0)'h_{T,t}(\Gamma_0)(\Gamma_{T,t} - \Gamma_0) + o_p(1)$$

$$\Rightarrow \int_0^1 \begin{pmatrix} \beta + \delta \mathbf{1}[\lambda \le \rho] \\ \tilde\beta + \tilde\delta \mathbf{1}[\lambda \le \rho] \end{pmatrix}' \Sigma^{-1} d \begin{pmatrix} G(\lambda) \\ \tilde G(\lambda) \end{pmatrix} - \tfrac{1}{2}\int_0^1 \begin{pmatrix} \beta + \delta \mathbf{1}[\lambda \le \rho] \\ \tilde\beta + \tilde\delta \mathbf{1}[\lambda \le \rho] \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} \beta + \delta \mathbf{1}[\lambda \le \rho] \\ \tilde\beta + \tilde\delta \mathbf{1}[\lambda \le \rho] \end{pmatrix} d\lambda$$

(18)

where $\Sigma = \omega^{-2} H^{-1}$. The r.h.s. of (18) may be recognized as log of the Radon-Nikodym derivative of the distribution of $(G, \tilde G')'$ with respect to the distribution of the $(k+1) \times 1$ Wiener process $(W, (A_{21}W + A_{22}\tilde W_k)')'$. This suggests that the information regarding $(\theta, \tilde\beta, \tilde\delta)$ from observing X_T converges in large samples to that contained in $(G, \tilde G')'$. This intuition is made formally precise in the Limit of Experiments theory pioneered by LeCam; see, for instance, van der Vaart (1998) for an introduction.

Condition 1 in the appendix states sufficient regularity assumptions that imply convergence of experiments in this sense. The Asymptotic Representation Theorem (see, for instance, Theorem 9.3 of van der Vaart (1998)) and Prohorov's Theorem then ensure that for any test $\varphi_T : \mathbb{R}^{qT} \mapsto [0,1]$ and any subsequence T' of T, there exists a further subsequence T' of T'' and a test $\varphi : D_{[0,1]}^{k+1} \mapsto [0,1]$ such that

$$E_{\theta,\tilde\beta,\tilde\delta}[\varphi_{T''}(X_{T''})] \to E_{\theta,\tilde\beta,\tilde\delta}[\varphi(G, \tilde G)] \text{ for all } \theta \in \Theta_0 \cup \Theta_1, \ \tilde\beta, \tilde\delta \in \mathbb{R}^k.$$

(19)

Thus, along any such subsequence T'', the asymptotic properties of φ_T are exactly equivalent to a particular test φ in the limiting problem discussed in Section 2.2 above. Define a test φ_T as *asymptotically unbiased* if any test φ satisfying (19) is unbiased. Similarly, define a test φ_T as *asymptotically invariant* if any test φ satisfying (19) is invariant to (15).

Proposition 3 *In a parametric model satisfying (18),*

(i) for any asymptotically unbiased test φ_T of (17) of asymptotic level 5%,

$$\limsup_{T \to \infty} E_{\theta,\tilde\beta,\tilde\delta}[\varphi_T(X_T)] \le E_{\theta,\tilde\beta,\tilde\delta}[\varphi_\rho(G)] \text{ for all } \theta \in \Theta_1, \tilde\beta, \tilde\delta \in \mathbb{R}^k;$$

(ii) for any test φ_T of (17) of asymptotic level 5%,

$$\limsup_{T \to \infty} \int E_{\theta,\tilde\beta,\tilde\delta}[\varphi_T(X_T)] d\tilde F(\theta) \le \bar\pi;$$

(iii) for any asymptotically invariant test φ_T of (17) of asymptotic level 5%,

$$\limsup_{T \to \infty} \int E_{\theta,\tilde\beta,0}[\varphi_T(X_T)] dF(\theta) \le \pi_{inv}(\bar R^2) \quad \text{for all } \tilde\beta \in \mathbb{R}^k.$$

Proposition 3 mirrors the results of Proposition 2 above. We argue in Section 4 below that the feasible test $\hat\varphi^* : \mathbb{R}^{qT} \mapsto [0,1]$ has the same asymptotic rejection properties as the test φ^* under (16). Thus, to the extent that Proposition 2 implies φ^* to be an attractive test in the context of the multivariate liming problem (11), $\hat\varphi^*$ is correspondingly an attractive test in the context of post break inference in a correctly specified parametric model that is subject to a local break.

2.3.2 GMM Models

We now turn to an application of the asymptotic efficiency concept introduced by Müller (2011), which we use here to state asymptotic power bounds on post break parameter inference in a class of GMM models. Let the moment condition be the $\mathbb{R}^m$-valued function $g(\cdot, \cdot)$, so that $E[g(x_{T,t}, \Gamma_0)] = 0$ when the true parameter at date t is given by $\Gamma = \Gamma_0$. Let the $\mathbb{R}^{(k+1) \times m}$ valued function $\Upsilon(\cdot, \Gamma)$ be the partial derivative of g with respect to Γ. Under the parameter evolution (16), conditional on τ, the natural estimators for the pre and post break value of Γ are given by the GMM estimator using pre and post break data,

$$\hat{\Gamma}_{pre}(\tau) = \arg \min_{\Gamma} \left(\sum_{t=1}^{\tau} g(x_{T,t}, \Gamma) \right)' \hat{V}_{pre} \left(\sum_{t=1}^{\tau} g(x_{T,t}, \Gamma) \right)$$

$$\hat{\Gamma}_{post}(\tau) = \arg \min_{\Gamma} \left(\sum_{t=\tau+1}^{T} g(x_{T,t}, \Gamma) \right)' \hat{V}_{post} \left(\sum_{t=\tau+1}^{T} g(x_{T,t}, \Gamma) \right)$$

with $\hat{V}_{pre}$ and $\hat{V}_{post}$ possibly data dependent $m \times m$ positive definite weighting matrices, and associated covariance matrix estimators $\hat{\Omega}_{pre}(\tau)$ and $\hat{\Omega}_{post}(\tau)$ of $\hat{\Gamma}_{pre}(\tau)$ and $\hat{\Gamma}_{post}(\tau)$, respectively. A natural starting point for inference about the post break parameter value in absence of knowledge of $\rho \in [0.15; 0.85]$ are the processes $\{\hat{\Gamma}_{pre}(\lfloor \cdot T \rfloor), \hat{\Gamma}_{post}(\lfloor \cdot T \rfloor), \hat{\Omega}_{pre}(\lfloor \cdot T \rfloor), \hat{\Omega}_{post}(\lfloor \cdot T \rfloor)\} \in D_{[0.15,0.85]}^{4(k+1)}$, where $\lfloor x \rfloor$ denotes the largest integer smaller or equal to $x \in \mathbb{R}$. Assuming that $T^{-1/2} \sum_{t=1}^{\lfloor \cdot T \rfloor} g(x_{T,t}, \Gamma_0) \Rightarrow \Sigma_g^{1/2} W_{k+1}(\cdot)$ for some positive definite matrix Σ_g and $T^{-1} \sum_{t=1}^{\lfloor \cdot T \rfloor} \Upsilon(x_{T,t}, \Gamma_0) \xrightarrow{p} \cdot \bar{\Upsilon}$ for some full column rank matrix $\bar{\Upsilon}$, the usual Taylor expansion arguments yield, for the case of efficient GMM estimation,

$$\{T^{1/2}(\hat{\Gamma}_{pre}(\lfloor sT \rfloor) - \Gamma_0), T^{1/2}(\hat{\Gamma}_{post}(\lfloor sT \rfloor) - \Gamma_0), T\hat{\Omega}_{pre}(\lfloor sT \rfloor), T\hat{\Omega}_{post}(\lfloor sT \rfloor)\}$$

$$\Rightarrow \left\{ \omega \begin{pmatrix} G(s) \\ \tilde{G}(s) \end{pmatrix}, \omega \begin{pmatrix} G(1) - G(s) \\ \tilde{G}(1) - \tilde{G}(s) \end{pmatrix}, s\omega^2 \Sigma, (1-s)\omega^2 \Sigma \right\} \tag{20}$$

in $D_{[0.15,0.85]}^{(k+1) \times 4}$, where ω^2 now is the 1,1 element of $(\bar{\Upsilon}' \Sigma_g^{-1} \bar{\Upsilon})^{-1}$ and $\Sigma = \omega^{-2} (\bar{\Upsilon}' \Sigma_g^{-1} \bar{\Upsilon})^{-1}$ (cf. Andrews (1993) and the discussion in Section 4.2 below).

Now consider the set of data generating processes for X_T that satisfy (20). One might want to choose the tests φ_T of (17) in a way that whenever (20) holds with $\beta = 0$, the test does not overreject asymptotically, that is $\limsup_{T \to \infty} E_{\theta, \tilde{\beta}, \tilde{\delta}}[\varphi_T(X_T)] \leq 0.05$ for all $\theta \in \Theta_0, \tilde{\beta}, \tilde{\delta} \in \mathbb{R}^k$ for *any* sequence of distributions of X_T that satisfies (20). Müller (2011) shows that under this robustness constraint, the best small sample test becomes the best test in the limiting problem (that is, with the r.h.s. of (20) assumed observed), evaluated at sample analogues (that is, at $\{T^{1/2}\hat{\Gamma}_{pre}(\lfloor \cdot T \rfloor), T^{1/2}\hat{\Gamma}_{post}(\lfloor \cdot T \rfloor), T\hat{\Omega}_{pre}(\lfloor \cdot T \rfloor), T\hat{\Omega}_{post}(\lfloor \cdot T \rfloor)\}$). Proposition 2 above shows in which sense φ^* may be considered approximately best in the equivalent problem of observing directly $(G(\cdot), \tilde{G}(\cdot)')' \in D_{[0,1]}^{k+1}$ with A_{12} and A_{22} known. Thus, part (ii) of Proposition 2 implies, for instance, that the weighted average asymptotic power (defined as in part (ii) of Proposition 3) of any 5% level test is no larger than $\bar{\pi}$. We discuss in detail in Section 4 below the construction of the small sample analogue $\hat{\varphi}^*$, and show in Proposition 4 that its asymptotic rejection

profile is equal to the rejection profile of φ^* whenever (20) holds, so its asymptotic weighted average power is equal to π^*. Thus, $\hat{\varphi}^*$ is the approximately local asymptotic power maximizing test among all robust tests in the sense of Müller (2011). We omit a formal statement analogous Proposition 3 to conserve space.

3   Determination of an Approximately Efficient Test with Uniform Size Control

As outlined in Section 2.1.3 above, the key challenge is to identify an approximately least favorable distribution over the two dimensional nuisance parameter of the break date and break magnitude. We employ the algorithm of Elliott, Müller, and Watson (2012) for this numerical search; Section 3.1 provides some details and the specific features of the problem at hand. In the second subsection, we introduce a new approach to the numerical study of size control of tests of a composite null hypothesis, which underlies our claim of uniform size control of the test φ^*. Finally, the third subsection contains details on the Monte Carlo determination of the power bound, which is complicated by the impossibility of generating and representing (pseudo) random Wiener process draws on a discrete computer.

3.1   Approximately Least Favorable Distribution

The approach of Elliott, Müller, and Watson (2012) partitions the sample space into two regions: In a "simple" region, the problem is approximately standard, so that a standard test may be applied. In the "difficult" region, the problem is substantively nonstandard, and the approximately weighted average power maximizing test is of the Neyman-Pearson form, with the approximately least favorable distribution as the null density. A data dependent switching rule determines which of the two tests are performed. For the problem at hand, we switch to the "standard" test if a supF statistic takes on very large values, and the standard test is (close to) the usual two-sided t-test on least-squares estimated post break data, $\hat{t}$ in (6), but with critical value of 2.01 rather than 1.96 on the 5% level. The adjustment of the critical value is to ensure overall size control.

Recall that the approximate least favorable distribution Λ^* is a probability distribution on the null parameter space Θ_0. The algorithm of Elliott, Müller, and Watson (2012) determines Λ^* as a mixture of a set of baseline distributions on Θ_0, which here consists of all possible values of the nuisance parameter (δ, ρ). We choose these baseline distributions to be of the form "ρ is uniform on $[\max(0.15, a/100 - \frac{1}{2}), \min(0.85, b/100 + \frac{1}{2})]$ and δ is an equal probability mixture of $\delta \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2)$ and $\delta \sim \mathcal{N}(-\mu_\delta, \sigma_\delta^2)$" for some integers $a, b \in \{15, \cdots, 85\}$, $b \geq a$, and $\mu_\delta \in \mathbb{R}, \sigma_\delta^2 > 0$. Values for the parameters $(a, b, \mu_\delta, \sigma_\delta^2)$ are given in Table 5 of the appendix. The algorithm now proceeds as follows: Set the candidate Λ^* equal to some baseline distribution, and construct the optimal test against the weighted average alternative distribution h (under the constraint of the switching to the standard test if supF > 90). Slightly adjust

the critical value of this test upwards so that it has one percentage point less power. Check whether the adjusted test controls size for data drawn from the other baseline null distributions. If it doesn't, determine a new Λ^* that is a mixture of the previous Λ^* and the violating baseline mixture. Iterate until the adjusted test controls size uniformly over all baseline distributions.

The result of this process is a test that has power within one percentage of the power bound, and that controls size under all baseline null mixtures. It may not, however, control size for all possible values of the nuisance parameters (δ, ρ). So the final step is to check such size control (details of which are in section 3.2). If size control is not obtained, then additional null basline distributions with smaller σ_δ^2 and/or smaller values of $b - a$ are added to the set of baseline distributions, and the iterative process goes through another round.

The test statistic (in the "difficult" region) is of the Neyman-Pearson form, with null and alternative densities given by f_θ in (8) integrated over the distribution of parameters under the null and alternative. Under a baseline null distribution as described above, we obtain $\frac{1}{\min(0.85, b/100+1/2)-\max(0.15, a/100-1/2)} \int_{\max(0.15, a/100-1/2)}^{\min(0.85, b/100+1/2)} \tilde{f}(r, \mu_\delta, \sigma_\delta^2; G) dr$, where

$$\tilde{f}(r, \mu_\delta, \sigma_\delta^2; G) = (r\sigma_\delta^2 + 1)^{-1/2} \exp \tfrac{1}{2} \left[\frac{\sigma_\delta^2 G(r)^2 - r\mu_\delta^2}{r\sigma_\delta^2 + 1} \right] \cosh \left[\frac{G(r)^2 \mu_\delta}{r\sigma_\delta^2 + 1} \right]$$

and the density of $G(\cdot)$ with $(\beta, \delta, \rho) \sim F$ is given by $\frac{1}{0.70} \int_{0.15}^{0.85} \tilde{h}(r; G) dr$, where

$$\tilde{h}(r; G) = (r\sigma_\delta^2 + 1)^{-1/2}((1-r)\sigma_\beta^2 + 1)^{-1/2} \exp \left[\tfrac{1}{2} \frac{\sigma_\delta^2 G(r)^2}{r\sigma_\delta^2 + 1} + \tfrac{1}{2} \frac{\sigma_\beta^2 (G(1) - G(r))^2}{(1-r)\sigma_\beta^2 + 1} \right].$$

In the application of this paper, we slightly modify the algorithm of Elliott, Müller, and Watson (2012). The observation G is a transformation of a Wiener process, so that computation of the Neyman-Pearson tests and their rejection profile requires the simulation of (pseudo) random continuous time processes. Discrete time Gaussian random walk approximation based on, say, 1000 steps are computationally expensive and of uncertain accuracy. For this reason, we approximate all integrals over ρ by sums with $\rho \in \{1/100, 2/100, \cdots, 1\}$. The rejection probability of the (approximate) Neyman-Pearson test thus becomes a function of

$$\left\{ G \left(\tfrac{l}{100} \right) \right\}_{l=1}^{100}. \tag{21}$$

The 100 random variables (21) can be simulated without any approximation error, as they are exactly jointly Gaussian. What is more, the fact that φ^* only depends on G through (21) significantly simplifies the determination of the size of φ^* discussed in Section 3.3 below. As a result we approximate integrals with sums over the 100 point grid $r \in \{\frac{1}{100}, \frac{2}{100}, \cdots, 1\}$. Thus, we replace $\frac{1}{0.70} \int_{0.15}^{0.85} \tilde{h}(r; G) dr$ by $h(G) = \frac{1}{71} \sum_{l=15}^{85} h(l/100, G)$, and for the baseline null distributions, we approximate the resulting density by $\frac{1}{b-a-1} \sum_{l=a}^{b} \tilde{f}(l/100, \mu_\delta, \sigma_\delta^2; G)$.

A power bound on tests that are functions of (21) is not, however, in general a valid power bound on the original problem involving the observation G. The correct implementation of the power bound via Lemma 1 requires evaluation of the actual density (8) of G, a complication we address in Section
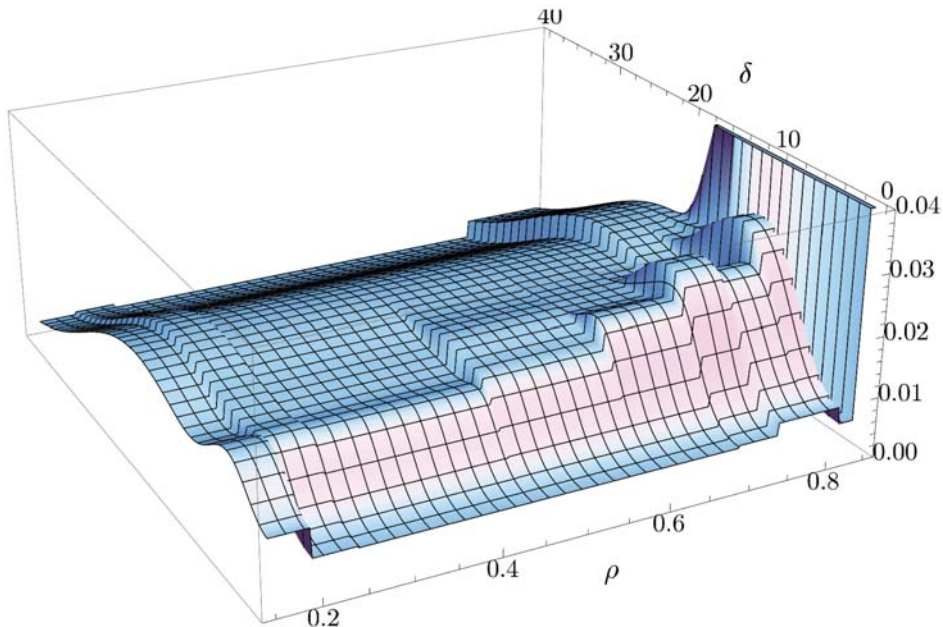
Figure 4: Approximate Least Favorable Distribution Λ^*

3.4 below. In the end, the comparison of this power bound with the power of the feasible test φ^* reveals whether working with (21) instead of the whole process G might have led to substantial inaccuracies in the resulting approximately least favorable distribution, and Proposition 1 shows that these are small.

Figure 4 plots the approximate least favorable distribution obtained by the algorithm. It is a mixture of 18 baseline distributions; see Table 2 below for the corresponding weights and values of $(a, b, \mu_\delta, \sigma_\delta^2)$.

3.2   Size Control

With a candidate test φ^* determined as outlined in Section 3.1 above, the question is whether it actually controls size, i.e. whether

$$\sup_{\theta \in \Theta_0} E_\theta[\varphi^*(G)] \leq 0.05. \tag{22}$$

By construction of φ^*, this is "approximately" true, but we now discuss how to establish (22) more rigorously, that is part (ii) of Proposition 1.

A standard approach is to resort to a Monte Carlo grid search: Choose a finite set $\Theta_{grid} \subset \Theta_0$, estimate the rejection probability by Monte Carlo for each $\theta \in \Theta_{grid}$, and conclude that the test controls size if the largest of these rejection probabilities is smaller or equal to the nominal level. Recent examples for this approach include Bunzel and Vogelsang (2005), Sriananthakumar and King (2006), Andrews and Guggenberger (2009, 2010a, 2010b), Guggenberger (2010), and Elliott, Müller, and Watson (2012), among others.[5] For the specific problem here, this approach would amount to computing the Monte

[5]Dufour's (2006) Maximum Monte Carlo approach also requires the computation of an expression just like

Carlo estimate of the null rejection probability

$$\hat{p}(\theta) = N^{-1} \sum_{i=1}^{N} \varphi^*(G_i) \quad \text{with} \ \ G_i(s) = W_i(s) + \delta \min(\rho, s) \tag{23}$$

for all $\theta = (0, \delta, \rho) \in \Theta_{grid}$, where W_i are N (pseudo) random draws of a standard Wiener process[6] (which may or may not be the same across different θ), and to conclude that φ^* is of level 5% if $\sup_{\theta \in \Theta_{grid}} \hat{p}(\theta) < 0.05$.

Clearly, though, this approach is not fully satisfying. Without some knowledge about the smoothness of $\hat{p}(\theta)$, even a fine grid search of this form does not provide any upper bound on $\sup_{\theta \in \Theta_0} \hat{p}(\theta)$, simply because the rejection probability $\hat{p}(\theta)$ could be very different between the grid points. What is more, Θ_0 is typically unbounded, so that it is impossible from the start to finely cover Θ_0. We now describe an approach that, while still based on Monte Carlo estimation,[7] can handle both these difficulties by bounding the function $\hat{p}(\theta)$ on the whole set Θ_0.

The key idea is the following change of perspective: Rather than to only regard the average $\hat{p}(\theta)$ in (23) as a function of θ, consider instead $\varphi^*(G_i)$ as a function of $\theta = (0, \delta, \rho) \in \Theta_0$, conditional on W_i. The actual data G_i is a function of the (pseudo) randomness W_i and the nuisance parameter θ, so that for each realization of W_i, the test statistic $\varphi^*(G_i)$ may be regarded as a function of θ, $\varphi^*(G_i) = S_i(\theta)$ with $S_i : \Theta_0 \mapsto \{0, 1\}$ (as long as φ^* is non-randomized). The function $S_i(\theta)$ is, of course, known, and one can analytically study its properties. So suppose initially that it was possible to exactly determine and store the "critical regions" $C_i = \{\theta : S_i(\theta) = 1\}$, $i = 1, \cdots, N$, that is the range of values for θ for which the test $\varphi^*(G_i)$ rejects.[8] For instance, if the nuisance parameter space was one dimensional, C_i would typically be a finite union of intervals, and it would suffice to determine their endpoints (possibly including $\pm\infty$). Clearly, the overall Monte Carlo estimate of the null rejection probability function $\hat{p} : \Theta_0 \mapsto [0, 1]$ is simply the average of these critical regions

$$\hat{p}(\theta) = N^{-1} \sum_{i=1}^{N} \mathbf{1}[\theta \in C_i]$$

and size control amounts to $\sup_{\theta \in \Theta_0} \hat{p}(\theta) \leq 0.05$.

Now for our problem, the nuisance parameter space is two dimensional, and the form of φ^* does not make it possible to usefully describe the sets C_i by a small set of numbers. So consider instead a finite partition $\mathcal{Q} = \{Q_1, Q_2, \cdots, Q_K\}$ of Θ_0, such as the *rectangles* of a grid, and suppose study of S_i allows

---

$\sup_{\theta \in \Theta_0} \hat{p}(\theta)$—cf. his equation (4.22). He suggests obtaining the supremum by simulated annealing techniques, which may also fail to yield the global maximum.

[6] As noted above, suffices to generate $\{W(l/100)\}_{l=1}^{100}$, which is actually feasible.

[7] In principle, one could rely on the insight of Dufour (2006) to obtain an overall randomized test of *exact* level 5% by rejecting if and only if the maximized (over Θ_0) p-value, computed from a new set of Monte Carlo simulations, is below 5%.

[8] This terminology is meant to be suggestive only; the actual critical region, of course, is the subset of the sample space $D_{[0,1]}$ for which φ^* rejects.

21

one to conclude that the sets $\{Q_j : j \in J_i\} \subset \mathcal{Q}$ with $J_i \subset \{1, 2, \cdots, K\}$ cover C_i, i.e. $C_i \subset \bigcup_{j \in J_i} Q_j$. It then suffices to keep track of the index sets J_i, $i = 1, \cdots, N$ to establish the bound

$$\sup_{\theta \in \Theta_0} \hat{p}(\theta) \leq \overline{\hat{p}} = \sup_{\theta \in \Theta_0} N^{-1} \sum_{i=1}^{N} \mathbf{1}[\theta \in \bigcup_{j \in J_i} Q_j] = \max_j N^{-1} \sum_{i=1}^{N} \mathbf{1}[j \in J_i] \qquad (24)$$

and one can conclude that φ^* is a 5% level test if $\overline{\hat{p}} \leq 0.05$. Note that this bound remains valid (although it becomes more conservative) when the sets J_i are larger than necessary, that is if $j \in J_i$ despite $C_i \cap Q_j = \varnothing$.

This approach benefits from a finer partition $\mathcal{Q}$ of Θ_0 with K large in two ways. On the one hand, a fine $\mathcal{Q}$ allows for a relatively accurate description of the actual critical region C_i, thus making the bound (24) less conservative. On the other hand, it will typically become easier to decide whether or not $C_i \cap Q_j = \varnothing$, for each $j = 1, \cdots, K$. But the computational cost of making K large is also substantial, as the brute force implementation requires a total of NK such evaluations. For our problem, it turns out that such a brute force implementation is not practical on today's PCs, as K needs to chosen of the order of magnitude of 10^8 for the bound (24) to become sufficiently sharp.

Thus, instead of trying to decide individually whether $C_i \cap Q_j = \varnothing$ for each $Q_j \in \mathcal{Q}$ and $i = 1, \cdots, N$, consider the following divide and conquer algorithm: For each i, initially try to decide for the whole set Θ_0 whether or not $C_i \cap \Theta_0 = \varnothing$. If such a determination can be made, set $J_i = \varnothing$ or $J_i = \{1, 2, \cdots, K\}$, respectively. Otherwise, divide Θ_0 into a coarse partition of 4 subsets $\Theta_0 = \{Q_1^c, Q_2^c, Q_3^c, Q_4^c\}$, where each Q_j^c, $j = 1, \cdots, 4$ can be covered exactly by a subset of $\mathcal{Q}$. Try to decide for each Q_j^c whether or not $C_i \cap Q_j^c = \varnothing$. If for any Q_j^c such a determination can be made, include (or not) the appropriate indices in J_i, and do not consider the subspace Q_j^c further. Otherwise, partition Q_j^c again into 4 subsets that can be covered exactly by a subset of $\mathcal{Q}$, and iterate until either the determination could be made, or the partition is as fine as $\mathcal{Q}$. The advantage of this algorithm is that it "zooms in" only when necessary. In our implementation, it cuts the number of required evaluations from 10^8 to around 10^4 to 10^7, depending on the draw of W_i.

We again relegate the details to the appendix, where we also discuss the treatment of unbounded subsets $Q_j \in \mathcal{Q}$.

The bound $\overline{\hat{p}}$ in (24), of course, is still subject to Monte Carlo error, and one might wonder about the effect of the supremum over Θ_0. Let P^* be the distribution of the N Monte Carlo draws of W_i, $i = 1, \cdots, N$.

Lemma 2 *For any $x \in \mathbb{R}$, $P^*(\sup_{\theta \in \Theta_0} \hat{p}(\theta) > x) \geq \sup_{\theta \in \Theta_0} P^*(\hat{p}(\theta) > x)$.*

Proof. Write E^* for the expectation with respect to P^*. Then, applying Jensen's inequality, $P^*(\sup_{\theta \in \Theta_0} \hat{p}(\theta) > x) = E^*[\mathbf{1}[\sup_{\theta \in \Theta_0} \hat{p}(\theta) > x]] = E^*[\sup_{\theta \in \Theta_0} \mathbf{1}[\hat{p}(\theta) > x]] \geq \sup_{\theta \in \Theta_0} E^*[\mathbf{1}[\hat{p}(\theta) > x]] = \sup_{\theta \in \Theta_0} P^*(\hat{p}(\theta) > x)$. $\blacksquare$

Lemma 2 shows that the P^* distribution of $\sup_{\theta \in \Theta_0} \hat{p}(\theta)$ (weakly) stochastically dominates that of $\hat{p}(\theta)$ for any fixed $\theta \in \Theta_0$. Thus, the P^* probability of observing $\overline{\hat{p}} \leq 0.05$ is at most as large as the

probability of observing $\hat{p}(\theta) \leq 0.05$, even if $\theta \in \Theta_0$ is chosen to maximize the actual null rejection probability $E_\theta[\varphi^*(G)]$. Both $\sup_{\theta \in \Theta_0} \hat{p}(\theta)$ and $\overline{\overline{p}}$ are, in general, upward biased estimates of the size of the test φ^*. A finding of $\overline{\overline{p}} \leq 0.05$ is thus *stronger* evidence that φ^* is of level 5% compared to any other 5% nominal level test of a single null hypothesis whose critical value was determined by N Monte Carlo draws.

3.3 Power Bound

We now turn to numerical issues that arise in the computation of the power bound via Lemma 1. Without the approximation of integrals by sums as in (21), the Neyman-Pearson test rejects for large values of a statistic of the form

$$\text{LR} = \frac{\int f_\theta dF(\theta)}{\int f_\theta d\Lambda^*(\theta)} = \frac{\int_{0.15}^{0.85} d_0(r) \exp\left[A_0(r)G(r)^2 + B_0(r)G(r)G(1) + C_0(r)G(1)^2\right] dr}{\sum_{i=1}^{L} \int_{a_i}^{b_i} d_i(r) \exp\left[A_i(r)G(r)^2 + B_i(r)G(r)\right] dr} \tag{25}$$

where d_i, A_i, B_i and C_i are functions $[0.15, 0.85] \mapsto \mathbb{R}$, and $a_i, b_i \in [0.15, 0.85]$, $i = 1, \cdots, L$, and L, a_i, b_i, d_i, A_i, B_i are determined by the algorithm described in Section 3.1 above. The right hand side in (25) arises from analytically integrating over the conditionally Gaussian distribution of (β, δ) in F in the numerator, and over a mixture of conditionally Gaussian distributions for δ in the denominator (conditional on ρ). As mentioned above, proper application of Lemma 1 requires the determination of the power of a 5% level test of H_{Λ^*} : "The density of G is $\int f_\theta d\Lambda^*(\theta)$" against the alternative H_F : "The density of G is $\int f_\theta dF(\theta)$", based on (25).

The usual approach would be to run a Monte Carlo experiment to estimate the critical value and power, using step function approximations to G and Riemann sums instead of integrals in (25). With enough steps (and Monte Carlo draws), this approach delivers any degree of accuracy. However, the test statistic (25) is poorly behaved for this approach: For large breaks (large $|\delta|$), the exponents in (25) take on large values for some r (of the order of magnitude of $\frac{1}{2}\rho\delta^2$), which renders the integrand extremely peaked. Intuitively, for large $|\delta|$, there is ample information about ρ, reflected in a very peaked likelihood. This leads to potentially very poor properties of Riemann sum approximations, casting doubt on the validity of a power bound computed in this way.

For this reason, we develop a lower and upper bound for LR that can be simulated exactly, and use Monte Carlo draws of the lower bound for the computation of the critical value, and Monte Carlo draws from the upper bound for the power computation. This ensures that the resulting upper bound on the power of the test statistic LR is only subject to the usual Monte Carlo error.

The basic element of the bounds is the following inequality: Let $\eta : [a, b] \mapsto \mathbb{R}$ for some $a < b$ with $\sup_{a \leq x \leq b} |\eta(r)| < M_\eta$, and let $\bar{\eta} = \frac{1}{b-a} \int_a^b \eta(r) dr$. Then Jensen's inequality and $e^x \leq x \cosh[M_\eta]$ for all $|x| < M_\eta$ implies

$$\exp[\bar{\eta}] \leq \frac{1}{b-a} \int_a^b \exp[\eta(r)] dr \leq \exp[\bar{\eta}] \cosh[M_\eta]. \tag{26}$$

Write any of the integrals in (25) as a sum of integrals over a fine partition of $[0.15; 0.85]$ into n intervals $[c_{j-1}, c_j]$, $j = 1, \cdots, n$ with $c_{j-1} < c_j$ for all j and $c_0 = 0.15$ and $c_n = 0.85$. In terms of $W(s) = G(s) - \beta s - \delta \min(\rho, s)$, the representative short integral over $r \in [c_{j-1}, c_j]$ in the denominator has the form

$$\int_{c_{j-1}}^{c_j} d(r) \exp\left[A(r)W(r)^2 + B(r)W(r)\right] dr$$

where now $d(r)$, $A(r)$ and $B(r)$ are also functions of $\theta = (\beta, \delta, \rho)$. Provided $W(r) > 0$ for $c_{j-1} \leq r \leq c_j$, using (26), we have

$$
\begin{aligned}
\underline{d} \exp\left[\underline{A}\overline{W_j^2} + \underline{B}\bar{W}_j\right] &\leq \tfrac{1}{c_j - c_{j-1}} \int_{c_{j-1}}^{c_j} d(r) \exp\left[A(r)W(r)^2 + B(r)W(r)\right] dr \\
&\leq \bar{d} \exp[\bar{A}\overline{W_j^2} + \bar{B}\bar{W}_j] \cosh[M_j]
\end{aligned}
\tag{27}
$$

where $\overline{W_j^2} = \tfrac{1}{c_j - c_{j-1}} \int_{c_{j-1}}^{c_j} W(r)^2 dr$, $\bar{W}_j = \tfrac{1}{c_j - c_{j-1}} \int_{c_{j-1}}^{c_j} W(r) dr$, $S_{1,j} = \sup_{c_{j-1} \leq r \leq c_j} |W(r) - \bar{W}_j|$, $S_{2,j} = \sup_{c_{j-1} \leq r \leq c_j} |W(r)^2 - \overline{W_j^2}|$, $M_j = \bar{A}S_{2,j} + \bar{B}S_{1,j}$, and lower and upper bars on d, A and B indicate maximum and minimum values of $d(r)$, $A(r)$ and $B(r)$ on $c_{j-1} \leq r \leq c_j$, respectively. Similar inequalities hold when $W(r) < 0$, when $W(r)$ changes sign on $c_{j-1} \leq r \leq c_j$, and also for the numerator of (25). A lower bound on LR is then given by replacing integrals in the denominator of (25) with sums over the upper bound in (27) and to replace the integrals in the numerator with sums over corresponding lower bounds, and vice versa for an upper bound on LR.

A Monte Carlo draw from these bounds for LR requires generation of the $4n$ random variables $\{\overline{W_j^2}, \bar{W}_j, S_{1,j}, S_{2,j}\}_{j=1}^n$. Although $\{\bar{W}_j\}_{j=1}^n$ is jointly Gaussian, this is non-trivial, as even the marginal distributions of $\overline{W_j^2}$, $S_{1,j}$ and $S_{2,j}$ are nonstandard. Fortunately, as detailed in the appendix, it is possible to rely on existing analytical results on the distribution of a related statistic of a Wiener process to generate draws $\{\overline{W_j^2}^l, \overline{W_j^2}^u, \bar{W}_j, S_{1,j}^u, S_{2,j}^u\}_{j=1}^n$ where with very high probability, $\overline{W_j^2}^l \leq \overline{W_j^2} \leq \overline{W_j^2}^u$, $S_{1,j} \leq S_{1,j}^u$ and $S_{2,j} \leq S_{2,j}^u$ for all $j = 1, \cdots, n$. By amending the above bounds accordingly, this is sufficient for the computation of a Monte Carlo estimate of an upper bound on the power of LR, entirely avoiding any qualifications that arise through the impossibility of generating exact draws of LR under $\int f_\theta d\Lambda^*(\theta)$ and $\int f_\theta dF(\theta)$.

4 Test Statistic and Properties

This section gives the precise form of the statistic and provides the asymptotic justification for the test as well as some evidence on its small sample performance. Section 4.1 details the construction of the statistic, and is the relevant subsection for those wishing to apply the method to data. The suggested statistic applies to inference on a single element of a $k \times 1$ vector in a GMM problem. Hence the method applies to linear and nonlinear regression, maximum likelihood models with differentiable likelihoods, structural and reduced form models. The method takes as its input the output of a standard GMM estimation over various subsamples, as detailed below. In section 4.2 we examine the large sample properties of the statistic, establishing limit results and showing that the statistics are consistent both

24

when the break is local to zero and nonlocal. The conditions on the data and models are similar to the set of assumptions in the breaks literature (for example Andrews (1993)), allowing for time dependent data and a wide range of models but ruling out unit (or near unit) roots and time trends. We examine small sample size and power properties in Monte Carlo experiments, both for inference on a sample mean and also for a parameter estimated by MLE.

4.1 Definition of Test Statistic

We define the statistic for a general method of moments framework with the $k+1$ dimensional parameter $\Gamma = (\gamma, \tilde{\gamma}')'$, with $\gamma \in \mathbb{R}$ and $\tilde{\gamma} \in \mathbb{R}^k$. The data in a sample of size T is given by $X_T = (x_1, \cdots, x_T)$, and the population moment condition is $E[g(x_t, \Gamma_0)] = 0$ when the true parameter at date t is given by $\Gamma = \Gamma_0$ for some known, $\mathbb{R}^m$ valued function $g(\cdot, \cdot)$. The moment condition can arise as a first order condition from maximum likelihood, a linear or nonlinear regression problem (structural or reduced form). For example in a linear IV/2SLS problem we might have $g(x_t, \Gamma_0) = (x_{1t} - \Gamma_0' x_{2t}) x_{3t}$ where the regression is of the univariate random variable x_{1t} on a vector of regressors x_{2t} with instruments x_{3t}. The weight matrix may be chosen in the usual way.

The parameter Γ changes its value from $\Gamma_{pre} = (\gamma_{pre}, \tilde{\gamma}'_{pre})'$ to $\Gamma_{post} = (\gamma_{post}, \tilde{\gamma}'_{post})'$ at some unknown date $\tau \in \mathbb{N}$ in the middle 70% of the sample, $\tau \in [0.15T, 0.85T]$ (the methods apply for any size of the change so are applicable where subsets or even all of the parameters in Γ do not change). The hypotheses of interest concerns the post break value of the first element γ of Γ,

$$H_0 : \gamma_{post} = \gamma_{post,0} \qquad \text{against} \qquad H_1 : \gamma_{post} \neq \gamma_{post,0}. \tag{28}$$

Denote by $\hat{\Gamma}_{pre}(t)$ and $\hat{\Gamma}_{post}(t)$ standard GMM estimators of Γ using data $\{x_s\}_{s=1}^t$ and $\{x_s\}_{s=t+1}^T$, and denote by $\hat{\Omega}_{pre}(t)$ and $\hat{\Omega}_{post}(t)$ the estimators of the covariance matrix of $\hat{\Gamma}_{pre}(t)$ and $\hat{\Gamma}_{post}(t)$, respectively. These covariance matrix estimators, as well as the weighting matrix for efficient GMM estimation in the overidentified case, are based on data $\{x_s\}_{s=1}^t$ and $\{x_s\}_{s=t+1}^T$ for the pre and post break estimators. If necessary, the estimators $\hat{\Omega}_{pre}(t)$ and $\hat{\Omega}_{post}(t)$ account for serial correlation by employing a correction as in Newey and West (1987) or Andrews (1991), for instance, so that in an overall stable model with parameter Γ_0, approximately, $\hat{\Omega}_{pre}(t)^{-1/2}\hat{\Gamma}_{pre}(t) \sim \mathcal{N}(\Gamma_0, I_{k+1})$ and $\hat{\Omega}_{post}(t)^{-1/2}\hat{\Gamma}_{post}(t) \sim \mathcal{N}(\Gamma_0, I_{k+1})$.

Our test statistic of (28) only requires evaluation of $\{\hat{\Gamma}_{pre}(t), \hat{\Gamma}_{post}(t), \hat{\Omega}_{pre}(t), \hat{\Omega}_{post}(t)\}$ at the 71 numbers $t = \lfloor lT/100 \rfloor$ for $l = 15, 16, \cdots, 85$ where $\lfloor x \rfloor$ denotes the largest integer smaller or equal to $x \in \mathbb{R}$.. Let $\hat{\gamma}_{pre}(l)$ and $\hat{\gamma}_{post}(l)$ be the first element of $\hat{\Gamma}_{pre}(\lfloor lT/100 \rfloor)$ and $\hat{\Gamma}_{post}(\lfloor lT/100 \rfloor)$, and denote by $\hat{\omega}^2_{pre}(l)$ and $\hat{\omega}^2_{post}(l)$ the 1,1 element of $\hat{\Omega}_{pre}(\lfloor lT/100 \rfloor)$ and $\hat{\Omega}_{post}(\lfloor lT/100 \rfloor)$, respectively. Define[9]

$$\widehat{\text{supF}} = \max_{16 \leq l \leq 85} \frac{\left(\hat{\gamma}_{post}(l) - \hat{\gamma}_{pre}(l-1)\right)^2}{\hat{\omega}^2_{post}(l) + \hat{\omega}^2_{pre}(l-1)}$$

---

[9]One obtains the test φ^* of Proposition 1 and defined in the appendix by setting $\hat{\gamma}_{pre}(l) = 100G(l/100)/l$, $\hat{\gamma}_{post}(l) = (G(1) - G(l/100))/(1 - l/100)$, $\hat{\omega}^2_{pre}(l) = 100/l$ and $\hat{\omega}^2_{post}(l) = 100/(100 - l)$.

Table 2: Weights, Intervals, Means and Variances of the Approximate Least Favorable Distribution

j	1	2	3	4	5	6	7	8	9
p_j	0.588	0.123	0.067	0.057	0.038	0.032	0.026	0.02	0.009
a_j	15	85	85	20	75	20	20	75	45
b_j	85	85	85	74	85	74	74	82	59
$\sigma_{\delta,j}^2$	100	10	4	300	200	10	3	10	10
μ_j	20	5	3	16	28	9	6	7	11
j	10	11	12	13	14	15	16	17	18
p_j	0.009	0.008	0.006	0.005	0.004	0.004	0.002	0.001	0.001
a_j	70	15	15	60	80	60	83	85	75
b_j	74	19	24	69	82	69	84	85	82
$\sigma_{\delta,j}^2$	10	10	200	10	10	3	10	3	3
μ_j	9	5	28	12	11	8	13	15.5	13

$$
\begin{aligned}
\hat{\Delta}_{pre}(l) &= l\hat{\gamma}_{pre}(l) - (l-1)\hat{\gamma}_{pre}(l-1), \ \hat{\Delta}_{post}(l) = (101-l)\hat{\gamma}_{post}(l-1) - (100-l)\hat{\gamma}_{post}(l) \\
\hat{l} &= \arg\min_{16 \le l \le 85} \left(\sum_{j=16}^{l-1} \hat{\Delta}_{pre}(j)^2 - (l-1)\hat{\gamma}_{pre}(l-1)^2 + \sum_{j=l+1}^{85} \hat{\Delta}_{post}(j)^2 - (100-l)\hat{\gamma}_{post}(l)^2 \right) \\
\hat{t}_{post} &= \frac{\hat{\gamma}_{post}(\min(\hat{l}+1,85)) - \gamma_{post,0}}{\sqrt{\hat{\omega}_{post}^2(\min(\hat{l}+1,85))}} \\
\hat{\omega}^2 &= \frac{(\hat{l}-1)^2}{9900}\hat{\omega}_{pre}^2(\hat{l}-1) + \frac{(100-\hat{l})^2}{9900}\hat{\omega}_{post}^2(\hat{l}) \\
\widehat{LR} &= \frac{\sum_{l=15}^{85} \dfrac{\exp\left[\frac{1}{2}\frac{\sigma_{pre}^2(\hat{\gamma}_{pre}(l)-\gamma_{post,0})^2 l^2}{100^2\hat{\omega}^2 v(l,\sigma_{pre}^2)} + \frac{1}{2}\frac{\sigma_\beta^2(\hat{\gamma}_{post}(l)-\gamma_{post,0})^2(100-l)^2}{100^2\hat{\omega}^2 v(100-l,\sigma_\beta^2)}\right]}{71\sqrt{v(l,\sigma_{pre}^2)v(100-l,\sigma_\beta^2)}}}{\sum_{j=1}^{18}\sum_{l=a_j}^{b_j} \dfrac{p_j \exp\left[-\frac{1}{2}\frac{\mu_{\delta,j}^2 l}{100 v(l,\sigma_{\delta,j}^2)} + \frac{1}{2}\frac{\sigma_{\delta,j}^2(\hat{\gamma}_{pre}(l)-\gamma_{post,0})^2 l^2}{100^2\hat{\omega}^2 v(l,\sigma_{\delta,j}^2)}\right]\cosh\left[\frac{(\hat{\gamma}_{pre}(l)-\gamma_{post,0})\mu_j l}{100 v(l,\sigma_{\delta,j}^2)\hat{\omega}}\right]}{(b_j-a_j+1)\sqrt{v(l,\sigma_{\delta,j}^2)}}}
\end{aligned}
\tag{29}
$$

where $v(l,\sigma^2) = 1 + \sigma^2 l/100$, $\sigma_{pre}^2 = 378$, $\sigma_\beta^2 = 22$, and a_j, b_j, p_j, μ_j and $\sigma_{\delta,j}^2$ are defined in Table 2. The 5% level test $\hat{\varphi}^*$ of (28) rejects H_0 if and only if either (i) $\widehat{supF} > 90$ and $|\hat{t}_{post}| > 2.01$, or (ii) $\widehat{supF} \le 90$ and $\widehat{LR} > 2.41$.[10]

The intuition for these computations is roughly as follows. The statistic $\widehat{supF}$ is the largest F-statistic of the null hypothesis that the value of γ in the first $(l-1)\%$ is equal to the value of γ in the last $(100-l)\%$, maximized over $16 \le l \le 85$. By leaving out the middle 1% of observations, it is ensured that for any true break fraction within the middle 70% of the sample, one of these F-statistics (often, the largest) only involves estimators $(\hat{\gamma}_{pre}(l), \hat{\omega}_{pre}^2(l))$ and $(\hat{\gamma}_{post}(l), \hat{\omega}_{post}^2(l))$ from stable periods. When

---

[10]Replace $(2.01, 2.41)$ by $(2.36, 10.6)$ for a 1% level test.

$\widehat{\text{supF}} > 90$, that is, when there is very strong evidence for the occurrence of a break, the test $\hat{\varphi}^*$ rejects if the usual t-statistic $\hat{t}_{post}$ is larger than 2.01 (rather than the usual 1.96) in absolute value, where $\hat{t}_{post}$ uses $\hat{l}$ to determine the appropriate post-break data.

In (approximately) linear and stationary models, $\hat{\gamma}_{pre}(l)$ is centered at the average parameter value of the first $l\%$ of the sample, $\lfloor lT/100 \rfloor^{-1} \sum_{t=1}^{\lfloor lT/100 \rfloor} (\gamma_{pre} + (\gamma_{post} - \gamma_{pre})\mathbf{1}[t > \tau])$, and similarly, $\hat{\gamma}_{post}(l)$ is centered at the average parameter value of the last $(100 - l)\%$ of the sample. The statistics $\hat{\Delta}_{pre}(l)$ and $\hat{\Delta}_{post}(l)$ thus approximately estimate the value of the parameter in the lth percent of the sample. If the true break τ is in the l_0th percent of the sample, then $\hat{\Delta}_{pre}(l)$ should be approximately equal to $\hat{\gamma}_{pre}(l_0)$ for all $l < l_0$, and similarly for $\hat{\Delta}_{post}(l)$. The break fraction estimator $\hat{l}$ is determined by minimizing a corresponding least squares criterion, again dropping the middle one percent. By adding one to $\hat{l}$, it is ensured that with high probability, $\hat{l} + 1$ is at least as large as the true break fraction in percentage points, so that $\hat{t}_{post}$ is based on estimates from a stable post-break period. The advantage of the least squares criterion (29) based on $(\hat{\Delta}_{pre}(l), \hat{\Delta}_{post}(l))$ over, say, an analogue to the F-statistics that underlie $\widehat{\text{supF}}$, is that $\hat{l}$ in (29) has appealing properties under relatively weak conditions also when the break magnitude is not small—cf. Proposition 4 below.

In absence of strong evidence for a break, that is $\widehat{\text{supF}} \leq 90$, the test $\hat{\varphi}^*$ switches to deciding the null hypothesis based on a likelihood ratio statistic. The numerator of this statistic is the result of the weighting over alternative values for γ_{post}, break dates and break magnitudes with the weighting function F defined in Section 2.1.3. The denominator is the result of a weighting over break dates and break magnitudes that make detection of this alternative as difficult as possible, the numerically determined approximate least favorable distribution discussed in Section 3, ensuring size control of $\hat{\varphi}^*$ even when information about the true break date is scarce. In this likelihood ratio statistic, the scale of $\hat{\gamma}_{pre}(l)$ and $\hat{\gamma}_{post}(l)$ is normalized by $\hat{\omega}$. If the break data τ satisfies $(\hat{l} - 1)/100 \leq \tau/T \leq \hat{l}/100$, then the estimator $\hat{\omega}^2$ is based on variance estimators of stable models, which improves the small sample properties of $\widehat{\text{LR}}$.

4.2   Asymptotic Properties

The following proposition establishes formally that under suitable conditions, the suggested test $\hat{\varphi}^*$ is consistent for non-local alternatives for both local and non-local break magnitudes; it has local power close to that of the infeasible test that uses only actual post break data when the break magnitude is large; it has local asymptotic power equal to the power of φ^* under a local break and local alternative; and it controls size uniformly. In particular, this implies that the weighted average power in Proposition 1, $\pi^* \simeq 49.0\%$, and the power in Figure 2 of Section 2.1.3 above is also the (weighted average) asymptotic local power of $\hat{\varphi}^*$ under local breaks.

Proposition 4 *Define* $\delta_T = T^{1/2}(\gamma_{pre} - \gamma_{post})/\omega$, $\beta_T = T^{1/2}(\gamma_{post} - \gamma_{post,0})/\omega$, *and* $\rho_T = \tau/T$ *for some* $\omega > 0$.

(i) Suppose $\delta_T = \delta \in \mathbb{R}$, $\rho_T = \rho \in [0.15, 0.85]$ and

$$\{T^{1/2}(\hat{\gamma}_{pre}(l) - \gamma_{post}), T^{1/2}(\hat{\gamma}_{post}(l) - \gamma_{post}), T\hat{\omega}_{pre}^2(l), T\hat{\omega}_{post}^2(l)\}_{l=15}^{85}$$
$$\Rightarrow \{\omega \frac{G_0(l/100)}{l/100}, \omega \frac{G_0(1) - G_0(l/100)}{1 - l/100}, \frac{\omega^2}{l/100}, \frac{\omega^2}{1 - l/100}\}_{l=15}^{85} \tag{30}$$

where $G_0(s) = W(s) + \delta \min(\rho, s)$. (i.a) If $\beta_T = \beta \in \mathbb{R}$, then $\hat{\varphi}^*(X_T) \Rightarrow \varphi^*(G)$. (i.b) If $\beta_T \to \pm\infty$, then $\hat{\varphi}^*(X_T) \xrightarrow{p} 1$.

(ii) Suppose $\delta_T \to \pm\infty$, $\rho_T = \rho \in [0.15, 0.85]$, and, for some $\omega_{pre} > 0$,

$$\{T^{1/2}(\hat{\gamma}_{pre}(l) - \gamma_{pre}), T\hat{\omega}_{pre}^2(l)\}_{l=15}^{\lfloor 100\rho \rfloor} \Rightarrow \{\omega_{pre} \frac{W(l/100)}{l/100}, \frac{\omega_{pre}^2}{l/100}\}_{l=15}^{\lfloor 100\rho \rfloor}$$
$$\{T^{1/2}(\hat{\gamma}_{post}(l) - \gamma_{post}), T\hat{\omega}_{post}^2(l)\}_{l=\lceil 100\rho \rceil}^{85} \Rightarrow \{\omega \frac{W(1) - W(l/100)}{1 - l/100}, \frac{\omega^2}{1 - l/100}\}_{l=\lceil 100\rho \rceil}^{85} \tag{31}$$
$$P(T^{1/2}|\hat{\gamma}_{pre}\lfloor 100\rho \rfloor + 1) - \gamma_{pre}| > M) \to 1, \ P(T^{1/2}|\hat{\gamma}_{post}(\lceil 100\rho \rceil - 1) - \gamma_{post}| > M) \to 1$$

for any $M \in \mathbb{R}$, where $\lceil x \rceil = -\lfloor -x \rfloor$. (ii.a) If $\beta_T = 0$, then $\limsup_{T\to\infty} E[\hat{\varphi}^*(X_T)] \leq 5\%$. (ii.b) If $\beta_T = \beta \in \mathbb{R}$ and $\rho \notin R_{grid} = \{0.15, 0.16, \cdots, 0.84\}$, then $\hat{\varphi}^*(X_T) \Rightarrow \mathbf{1}[|Z + \beta\sqrt{0.98 - \lfloor 100\rho \rfloor/100}| > 2.01]$, where $Z \sim \mathcal{N}(0,1)$. (ii.c) If $\beta_T \to \pm\infty$, then $\hat{\varphi}^*(X_T) \xrightarrow{p} 1$.

(iii) Suppose $\beta_T = 0$; (30) holds for all sequences $(\rho_T, \delta_T) \to (\rho, \delta) \in [0.15, 0.85] \times \mathbb{R}$; (31) holds for all sequences $(\rho_T, \delta_T) \to (\rho, \delta) \in ([0.15, 0.85] \backslash R_{grid}) \times \{-\infty, +\infty\}$; and for all for all sequences $(\rho_T, \delta_T) \to (\rho, \delta) \in R_{grid} \times \{-\infty, +\infty\}$, there exists a subsequence $(\rho_{T'}, \delta_{T'}) \to (\rho, \delta)$ such that along this subsequence, (31) holds except for

$$(T^{1/2}(\hat{\gamma}_{pre}(l_0) - \gamma_{pre}), T\hat{\omega}_{pre}^2(l_0)) \Rightarrow (\omega_{pre} \frac{W(l_0/100)}{l_0/100} + \nu_{pre}(l_0), \frac{\omega_{pre}^2}{l_0/100})$$
$$(T^{1/2}(\hat{\gamma}_{post}(l_0) - \gamma_{post}), T\hat{\omega}_{post}^2(l_0)) \Rightarrow (\omega \frac{W(1) - W(l_0/100)}{1 - l_0/100} + \nu_{post}(l_0), \frac{\omega^2}{1 - l_0/100}) \tag{32}$$

where $l_0 = 100\rho$ and $\nu_{pre}(l_0), \nu_{post}(l_0) \in \mathbb{R} \cup \{-\infty, +\infty\}$ are nonrandom constants, of which at most one is nonzero. Then $\limsup_{T\to\infty} \sup_{\theta \in \Theta_0} E_\theta[\hat{\varphi}^*(X_T)] \leq \alpha^* \simeq 5\%$.

Part (i) of Proposition 4 considers the case where the pre and post break value of γ only differ of the order $T^{-1/2}$, i.e. when the break size is small in a statistical sense. In that neighborhood, the GMM estimators do not contain enough information to pin down the true break fraction exactly, even asymptotically. This asymptotic embedding mirrors the small sample problem with substantial uncertainty about the true break date. The proposition establishes that for local alternatives, where correspondingly, the true value γ_{post} differs by the order $T^{-1/2}$ from the hypothesized value $\gamma_{post,0}$, the asymptotic properties of $\hat{\varphi}^*$ are just like those of φ^* in the limiting problem discussed in Section 2.1. So in particular, by Proposition 1, the test $\hat{\varphi}^*$ has asymptotic level $\alpha^* \simeq 5\%$, and it has asymptotic weighted average power of $\pi^* \simeq 49\%$. Furthermore, $\hat{\varphi}^*$ is consistent in this scenario against any alternative where γ_{post} is outside the $T^{-1/2}$ neighborhood of the hypothesized value.

These results rely on the high level condition (30). The sequence of statistics $\{\hat{\gamma}_{pre}(l), \hat{\gamma}_{post}(l), \hat{\omega}_{pre}(l), \hat{\omega}(l)\}$ are a special case of partial sample GMM estimators analyzed by Andrews (1993). In particular, his primitive Assumption 1 with an appropriate modification to account for the local break in the parameter value imply (30)—see the appendix for details. Alternatively, the approach

of Li and Müller (2009) could be amended to yield (30) under a different set of assumptions. Also, for the special case of maximum likelihood estimation, Condition 1 in the appendix can be shown to imply (30) for both the average Hessian and outer product of scores covariance estimators $\hat{\Omega}_{pre}$ and $\hat{\Omega}_{post}$. The conclusions of part (i) thus hold for a wide range of models.[11]

Part (ii) of Proposition 4 establishes the asymptotic properties of $\hat{\varphi}^*$ when the break is (very) large relative to the sample information about γ. This is the situation that the majority of the literature on breaks has considered, where breaks are large enough that they are essentially 'known' for purposes of the asymptotic theory. The test again controls size and is consistent against any non-local alternative. Under local alternatives $\beta_T = \beta \neq 0$, it asymptotically corresponds to a two-sided t-test about the mean of a unit variance Gaussian variate with non-centrality parameter $\beta\sqrt{0.98 - \lfloor 100\rho \rfloor/100}$ and critical value 2.01, at least as long $\rho \neq R_{grid}$.[12] In comparison, the two-sided 5% level t-test based on post break data (assuming the break date was known) converges to $\varphi_\rho(G) = \mathbf{1}[|\frac{G(1)-G(\rho)}{\sqrt{1-\rho}}| > \Phi(0.975)]$, and thus has non-centrality parameter $\beta\sqrt{1-\rho}$. With $\beta \sim \mathcal{N}(0, 22)$ as in the weighting function F of Section 2.1.3, the average asymptotic power loss of $\hat{\varphi}^*$ relative to φ_ρ over $\rho \in [0.15, 0.85]$ is approximately 1.7 percentage points, with a largest difference of 3.6 percentage points occurring at $\rho \to 0.83$.

The results in part (ii) of Proposition 4 require two sets of assumptions. On the one hand, the statistics $\{\hat{\gamma}_{pre}(l), \hat{\omega}^2_{pre}(l)\}$ and $\{\hat{\gamma}_{post}(l), \hat{\omega}^2_{post}(l)\}$ have to behave in the usual way *over the stable pre and post break periods*, respectively, and be asymptotically independent. Note that the limiting variance may change at the parameter break date; this accommodates, say, changes in the variance of the AR(1) coefficient estimator that are induced by a non-local break in the AR(1) coefficient. One can again invoke the primitive conditions of Andrews (1993) to justify these convergences. On the other hand, the estimators $\hat{\gamma}_{pre}(l)$ and $\hat{\gamma}_{post}(l)$ must diverge from the $T^{-1/2}$ neighborhood of the pre and post parameter values γ_{pre} and γ_{post} for two values of l that involve a positive fraction of post and pre break data, respectively.[13] For a non-local but shrinking break magnitude $|\gamma_{pre} - \gamma_{post}| \to 0$ and $T^{1/2}(\gamma_{pre} - \gamma_{post}) \to \pm\infty$, $\hat{\gamma}_{pre}(l)$ typically estimates the average parameter value in the first $l\%$ of the sample, that is

$$\frac{\hat{\gamma}_{pre}(l) - \gamma_{pre}}{\gamma_{post} - \gamma_{pre}} \xrightarrow{p} \mathbf{1}[l > \lfloor 100\rho \rfloor]\frac{l - \lfloor 100\rho \rfloor}{l} \neq 0, \tag{33}$$

and similarly for $\hat{\gamma}_{post}(l)$, which is clearly sufficient for the purposes of Proposition 4 (ii). The convergence (33) can be shown to hold, for instance, under the high-level Condition 1 of Li and Müller (2009) by

---

[11]As pointed out by Hansen (2000), most tests for parameter stability do not control asymptotic size under 'global' heterogeneity of the average variance across different parts of the sample. This also applies to $\hat{\varphi}^*$, at least for local breaks. However, by picking the 71 grid points in the definition of $\hat{\varphi}^*$ not relative to the sample proportion, but relative to a suitably chosen cumulative information (that typically would be need to be estimated with sufficient precision), one could induce the convergence (30) even for models with 'global' variance heterogeneity.

[12]This qualification is necessary because if $\rho \in R_{grid}$, $\hat{l}$ potentially takes on the two values $\lfloor 100\rho \rfloor + 1$ and $\lfloor 100\rho \rfloor + 2$ with positive probability, even asymptotically, despite the diverging break magnitude $T^{1/2}(\gamma_{pre} - \gamma_{post})$.

[13]Note, however, that nothing is assumed about the behavior of the variance estimators $\{\hat{\omega}_{pre}(l), \hat{\omega}_{post}(l)\}$ that involve subsets of unstable periods.

proceeding as in their Theorem 1. For a fixed break magnitude $\gamma_{post} - \gamma_{pre}$, the distortionary effects of the break becomes even stronger, of course, and barring pathological cancellations, one would expect the condition in Proposition 4 (ii) to be satisfied. We refrain from a detailed discussion of sufficient primitive conditions for the sake of brevity.

Finally, part (iii) of Proposition 4 shows that the test $\hat{\varphi}^*$ controls size uniformly in large samples—for T large enough, no choice of $\theta \in \Theta_0$ leads to a rejection probability substantially above 5%. This result is established under the assumption that for local breaks $\delta_T = O(1)$, the convergences in part (i) hold for all converging sequences $(\rho_T, \delta_T) \to (\rho, \delta)$. We show in the appendix that the primitive assumptions of Andrews (1993) are again sufficient. For non-local breaks, care must be taken for the case $\rho_T \to \rho \in R_{grid}$. If $\rho_T \to \rho = l_0/100 \in R_{grid}$ with $\rho_T < \rho$, then $\hat{\gamma}_{pre}(l_0)$ is an estimator from an unstable model, with a shrinking fraction of the data stemming from the post break model. Depending on the rate at which $\delta_T \to \pm\infty$, this contamination shifts the center of the asymptotic distribution of $\hat{\gamma}_{pre}(l_0)$ by $\nu_{pre}(l_0)$—typically, given the arguments above (33), one would expect $\nu_{pre}(l_0) \in \mathbb{R}$ if $\delta_T(\rho_T - \rho) \to \xi \in \mathbb{R}$, and it is sufficient for the last assumption in Proposition 4 (iii) to assume that when $(\rho_T, \delta_T) \to (\rho, \delta) \in R_{grid} \times \{-\infty + \infty\}$, (31) holds except for (32) for all sequences $\delta_T(\rho_T - \rho) \to \xi \in \mathbb{R} \cup \{-\infty, +\infty\}$.[14]

4.3   Small Sample Properties

We now turn to the small sample properties of the test $\hat{\varphi}^*$ suggested here, and compare it to the infeasible test $\hat{\varphi}_\rho$ that corresponds to standard inference using actual post break data only. We consider two Monte Carlo designs: Inference about the post break mean of an otherwise stationary scalar time series, and inference about the post break value of the moving average parameter of a scalar time series.

For inference about the mean, we design the data generating process to mimic the productivity growth example of the introduction. The data does not exhibit significant autocorrelation, and the sample kurtosis is 3.24. Accordingly, we set $T = 143$, and draw the data $\{x_t\}_{t=1}^T$ from

$$x_t = \beta + \delta T^{-1/2} \mathbf{1}[t \leq \rho T] + u_t$$

where u_t is i.i.d. student-t with 29 degrees of freedom, and scaled to be of unit variance. Table 3 reports size and power of various 5% nominal tests of the null hypothesis that $\beta = 0$ (this choice is without loss of generality, as the testing problem is translation and scale invariant). The results correspond closely to the asymptotic properties of the corresponding tests discussed in Section 2.1, and similar comments apply. The test developed in this paper, $\hat{\varphi}^*$, has close to nominal rejection probability for all value of

---

[14]Inspection of the proof of Proposition 4 shows that these assumptions are only needed to ensure that $(\{\hat{\gamma}_{pre}(l)\}_{l=1}^{l_0}, \hat{\Delta}_{post}(l_0 + 1))$ are asymptotically independent of $\{\hat{\gamma}_{post}(l), \hat{\omega}_{post}^2(l)\}_{l=l_0+1}^{l_0+2}$, as this independence is enough to establish that the mixture of the two $\hat{t}_{post}$ statistics with $\hat{l} = l_0 + 1$ and $\hat{l} = l_0 + 2$ controls size. The condition could dispensed with entirely by increasing the critical value of $\hat{t}_{post}$ to 2.07, since $P(|Z_1| < 2.07 \text{ and } |Z_2| < 2.07) > 0.95$ for all bivariate normal Z_1, Z_2 with $Z_1 \sim Z_2 \sim \mathcal{N}(0,1)$ and correlation of at least $\sqrt{15/16}$, so that no mixture of the two $\hat{t}_{post}$-statistics can induce overrejections.

Table 3: Small Sample Size and Power in Location Model

ρ	Size ($\beta = 0$)				Power ($\beta = 4$)				Power ($\beta = -4$)			
	$\delta =$				$\delta =$				$\delta =$			
	1	4	8	16	1	4	8	16	1	4	8	16
$\hat{\varphi}_\rho$, t-test with break date known												
0.25	0.06	0.06	0.06	0.06	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
0.50	0.07	0.07	0.07	0.07	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
0.75	0.09	0.09	0.09	0.09	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
t-test with least-squares break date, standard critical value												
0.25	0.22	0.19	0.14	0.10	0.66	0.73	0.78	0.78	0.62	0.58	0.56	0.57
0.50	0.22	0.17	0.10	0.07	0.63	0.66	0.74	0.80	0.66	0.77	0.85	0.85
0.75	0.22	0.23	0.16	0.11	0.63	0.56	0.54	0.56	0.65	0.67	0.65	0.61
t-test with least-squares break date, asymptotic uniform critical value												
0.25	0.00	0.00	0.00	0.00	0.10	0.09	0.08	0.08	0.11	0.12	0.11	0.08
0.50	0.00	0.00	0.00	0.00	0.10	0.07	0.06	0.06	0.11	0.11	0.09	0.07
0.75	0.00	0.00	0.01	0.00	0.13	0.15	0.08	0.04	0.08	0.05	0.04	0.03
$\hat{\varphi}_{0.85}$, t-test with last 15% of data												
0.25	0.11	0.11	0.11	0.11	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
0.50	0.11	0.11	0.11	0.11	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
0.75	0.11	0.11	0.11	0.11	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
$\hat{\varphi}^*$, test of this paper												
0.25	0.06	0.06	0.05	0.04	0.46	0.57	0.78	0.90	0.48	0.64	0.90	0.93
0.50	0.06	0.06	0.06	0.05	0.45	0.53	0.69	0.76	0.48	0.67	0.82	0.79
0.75	0.06	0.05	0.06	0.07	0.45	0.44	0.44	0.45	0.46	0.43	0.41	0.50

Notes: The variance of sample means is estimated with a Newey and West (1987) estimator with four lags. The tests are the same as those in Table 1, except that the oracle test $\hat{\varphi}_\rho$ is based on the actual post break data. Power is not size-corrected. Based on 10,000 replications.

Table 4: Small Sample Size and Power of Suggested Test in MA(1) Model

ρ	Size ($\psi_{post} = 0$)					Power ($\psi_{post} = 0.2$)				
	$\hat{\varphi}^*, \psi_{pre} =$				$\hat{\varphi}_\rho$	$\hat{\varphi}^*, \psi_{pre} =$				$\hat{\varphi}_\rho$
	0.1	0.3	0.5	0.7		-0.1	0.1	0.3	0.5	
0.25	5.7	5.6	5.0	5.2	5.2	60.3	85.5	96.9	96.9	97.5
0.50	6.0	7.1	8.4	8.3	5.9	61.3	85.4	88.6	89.1	89.7
0.75	5.7	8.9	12.0	13.3	7.0	53.6	51.0	56.7	66.6	64.4

Notes: All entries are based on asymptotic critical values. The variance of the unconditional maximum likelihood estimators $\hat{\psi}$ is estimated by $(1 - \hat{\psi}^2)/\lfloor rT \rfloor$ in a sample of size $\lfloor rT \rfloor$. Based on 10,000 replications.

the break magnitude, in contrast to the (usual) t-test on post break data with a break date determined by minimizing the sum of squared residuals. The mild overrejections of $\hat{\varphi}^*$ are caused by the variance estimation, and are smaller than those of the oracle test $\hat{\varphi}_\rho$ with known break date.

In the MA(1) design, we generate data from the model

$$x_t = \mu + \varepsilon_t + \mathbf{1}[t \leq \rho T]\psi_{pre}\varepsilon_{t-1} + \mathbf{1}[t > \rho T]\psi_{post}\varepsilon_{t-1}$$

with $\varepsilon_t \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$ and $T = 480$ (think of 40 years of monthly data). We test the hypothesis $H_0 : \psi_{post} = 0$, which corresponds to the null hypothesis of no current predictability of x_t. This design is outside the linear regression framework for which least-squares methods have been developed, so with unknown break date, $\hat{\varphi}^*$ seems to be the only known way of performing this inference. Table 4 reports the null and alternative rejection probabilities of $\hat{\varphi}^*$ and the oracle test $\hat{\varphi}_\rho$, where the parameters are estimated by maximum likelihood. It can be shown that Condition 1 in the appendix holds for this model. By Propositions 3 and 4, $\hat{\varphi}^*$ is therefore the approximately weighted average local asymptotic power maximizing test of current predictability of x_t. Small sample size distortions are mostly larger for $\hat{\varphi}^*$ compared to $\hat{\varphi}_\rho$, but remain modest for most considered parameter values. We experimented with smaller sample sizes and found worse size control, sometimes substantially so. Good properties of $\hat{\varphi}^*$ rely on reasonable accuracy of the usual large sample approximations over all partial sample estimators. But with small sample sizes T, the estimators over, say, the first and last 15% are based on very few observations, which leads to well known problems with the MA(1) maximum likelihood estimator.

5   Conclusion

Models with discrete breaks in the parameters at an unknown or uncertain date have become popular in empirical work. This paper shows that inference about pre and post break parameters using the standard approach of using the estimated break date as the true break date leads to substantially oversized tests and confidence intervals as long as the break magnitude is not very large relative to

the sampling uncertainty about the parameters. For the important special case of a single break at an unknown date and a single parameter of interest, we derive an alternative test with uniform asymptotic size control that demonstrably comes close to maximizing a weighted average power criterion.

The method applies to a very wide range of econometric models—those that can be written as a GMM problem. This includes breaks in means and variances, in coefficients in linear and nonlinear regressions, and maximum likelihood models with differentiable likelihood (where the first order condition becomes the GMM problem). The required regularity conditions are similar to those in Andrews (1993). And while the derivation of the test statistic is somewhat complicated, its application is fairly straightforward: it only involves the estimation of four standard GMM objects (parameter estimates and their covariance matrices) over 142 specific subsamples, irrespective of the sample size.

6 Appendix

Definition of φ^*:

$\varphi^*(G) = \mathbf{1}[\sup\mathrm{F}_0 > 90]\mathbf{1}[|\hat{t}_0| > 2.01] + \mathbf{1}[\sup\mathrm{F}_0 \le 90]\mathbf{1}[\mathrm{LR}_0 > 2.41]$, where

$$\sup\mathrm{F}_0 = \max_{16 \le l \le 85} \frac{100}{99} \frac{((100-l)G(\frac{l-1}{100}) - (l-1)(G(1) - G(\frac{l}{100})))^2}{(100-l)(l-1)}$$

$$S_0(l) = \frac{100(100G(\frac{l}{100}) - lG(1))^2}{l(100-l)} + 100^2 \frac{(lG(\frac{l-1}{100}) - (l-1)G(\frac{l}{100}))^2}{l(l-1)}$$

$$\hat{l}_0 = \arg \max_{16 \le l \le 85} S_0(l)$$

$$\hat{t}_0 = \frac{G(1) - G(\min(\hat{l}_0 + 1, 85)/100)}{\sqrt{1 - \min(\hat{l}_0 + 1, 85)/100}}$$

$$\mathrm{LR}_0 = \frac{\sum_{l=15}^{85} v(l, \sigma_{pre}^2)^{-1/2} v(100-l, \sigma_{\beta}^2)^{-1/2} \exp[\frac{1}{2}\frac{\sigma_{pre}^2 G(l/100)^2}{v(l,\sigma_{pre}^2)} + \frac{1}{2}\frac{\sigma_{\beta}^2 (G(1)-G(l/100))^2}{v(100-l,\sigma_{\beta}^2)}]}{\sum_{j=1}^{18} \sum_{l=a_j}^{b_j} \frac{p_j}{b_j-a_j+1} v(l, \sigma_{\delta,j}^2)^{-1/2} \exp[-\frac{1}{2}\frac{\mu_{\delta,j}^2 l}{100 v(l,\sigma_{\delta,j}^2)} + \frac{1}{2}\frac{\sigma_{\delta,j}^2 G(l/100)^2}{v(l,\sigma_{\delta,j}^2)}] \cosh[\frac{\mu_j G(l/100)}{v(l,\sigma_{\delta,j}^2)}]}.$$

and $v(l, \sigma^2) = 1 + \sigma^2 l/100$, $\sigma_{pre}^2 = 378$, $\sigma_{\beta}^2 = 22$, and $p_j, a_j, b_j, \sigma_{\delta,j}^2$ and μ_j are defined in Table 2.

Proof of Proposition 2:

By Girsanov's Theorem, the log of the Radon-Nikodym (RN) derivative of the distribution of $(G, \tilde{G}')'$ with respect to the distribution of $(W, (A_{21}W + A_{22}\tilde{W}_k)')'$ is the r.h.s. of (18).

(i) Fix arbitrary values $\rho_0 \in [0.15, 0.85]$, $\delta_0 \in \mathbb{R}$, $\tilde{\beta}_0, \tilde{\delta}_0 \in \mathbb{R}^k$, and consider the construction of the most powerful unbiased test against the specific alternative $H_1^s(b) : \beta = b, \delta = \delta_0 - b, \tilde{\beta} = \tilde{\beta}_0 + bA_{21}, \tilde{\delta} = \tilde{\delta}_0 - bA_{21}$ $\rho = \rho_0$, which is indexed by $b \in \mathbb{R}$. Any 5% level test for the unrestricted null hypothesis is also of level 5% under the specific null hypothesis $H_0^s : \beta = 0, \delta = \delta_0, \tilde{\beta} = \tilde{\beta}_0, \tilde{\delta} = \tilde{\delta}_0, \rho = \rho_0$. The log of the RN derivative of the distribution of $(G, \tilde{G}')'$ under $H_1^s(b)$ with respect to the distribution under H_0^s then becomes $b(G(1) - G(\rho_0)) - \frac{1}{2}(1 - \rho_0)b^2$. The experiment of testing H_0^s against $H_1^s(b)$ based on the observation $(G, \tilde{G}')'$ is therefore *equivalent* (in the sense of equivalence of statistical experiments) to inference about b in the Gaussian shift experiment of observing $X \sim \mathcal{N}(b(1 - \rho_0), 1 - \rho_0)$. In particular, for any test of H_0^s against $H_1^s(b)$, there exists a test $\varphi_X : \mathbb{R} \mapsto [0, 1]$ that has the same distribution for all $b \in \mathbb{R}$ (this follows, for instance, from

33

Theorem 9.3 of van der Vaart (1998)). It is well known that the best 5% level unbiased test in the Gaussian shift experiment is $\varphi_X^* = \mathbf{1}[|X/\sqrt{1-\rho_0}| > \Phi(0.975)]$, so that the best unbiased 5% level test of H_0^s against $H_1^s(b)$ is $\varphi_{\rho_0}(G)$. Since $\varphi_{\rho_0}(G)$ does not depend on $\delta_0, \tilde{\beta}_0, \tilde{\delta}_0$, and $\rho_0, \delta_0, \tilde{\beta}_0, \tilde{\delta}_0$ were arbitrary, this proves the claim.

(ii) The bound $\bar{\pi}$ in Proposition 1 is constructed via Lemma 1, i.e., $\bar{\pi}$ is the power of the Neyman-Pearson test φ_{Λ^*} of H_{Λ^*} against H_F, treating G as the observation, with Λ^* the approximately least favorable distribution Λ^* determined as described in Section 3.1.

Now let $\tilde{\Lambda}^*$ be the probability distribution for $(\theta, \tilde{\beta}, \tilde{\delta})$ with the same marginal distribution of θ as Λ^*, and $(\tilde{\beta}, \tilde{\delta}) = (\beta A_{21}, \delta A_{21})$ conditional on θ. The Neyman-Pearson test of $H_{\tilde{\Lambda}^*}$ against $H_{\tilde{F}}$, based on the observation $(G, \tilde{G}')'$, is then seen to be identical to φ_{Λ^*}. The result thus follows from Lemma 1.

(iii) A maximal invariant is $(\Psi, G(1))$, where $\Psi(s) = (G(s) - sG(1), \tilde{G}(s)' - s\tilde{G}(1)')'$. Note that the distribution of $(\Psi, G(1))$ does not depend on $\tilde{\beta}$, and that $G(1)$ is independent of Ψ. The log of the RN derivative of $(\Psi, G(1))$ with respect to the distribution of (B_{k+1}, Z), where B_{k+1} is a $k+1$ Brownian Bridge with covariance matrix Σ and $Z \sim \mathcal{N}(0,1)$ independent of B_{k+1}, is equal to

$$\delta e_1' \Sigma^{-1} \Psi(\rho) + (\beta + \delta\rho)G(1) - \tfrac{1}{2}\rho(1-\rho)e_1' \Sigma^{-1} e_1 \delta^2 - \tfrac{1}{2}(\beta + \delta\rho)^2$$

where e_1 is the first column of I_{k+1}. By the same arguments as employed in part (i), this experiment is seen to be equivalent to the observation of $(\bar{\Psi}, G(1))$, where $\bar{\Psi}(s) = W(s) - sW(1) + (\rho(1-s) + \mathbf{1}[s \leq \rho](s-\rho))\delta/\sqrt{1-R^2}$. A calculation yields that the RN derivative of the distribution of $(\bar{\Psi}, G(1))$ under $\theta \sim F$ with respect to the distribution of $(B_1, W(1))$ (where $B_1(s) = W(s) - sW(1)$) is

$$\frac{1}{0.70} \int_{0.15}^{0.85} \left(\frac{(1 + (1-r)\sigma_\beta^2)(1 + r(\sigma_\delta^2 - \sigma_\beta^2)) - R^2(1 + \sigma_\beta^2(1-2r) + r^2\sigma_\delta^2)}{\sigma_\beta^2(\sigma_\delta^2 - \sigma_\beta^2)(1 - R^2)} \right)^{-1/2}$$
$$\times \exp\left[\tfrac{1}{2}v(r)' \begin{pmatrix} 1 + \frac{1}{\sigma_\beta^2} - \frac{r-(2-r)rR^2}{1-R^2} & -\frac{(1-r)rR^2}{1-R^2} \\ -\frac{(1-r)rR^2}{1-R^2} & \frac{1}{\sigma_\delta^2 - \sigma_\beta^2} + \frac{r(1-rR^2)}{1-R^2} \end{pmatrix}^{-1} v(r) \right] dr$$

where $v(r) = ((1-r)G(1) - \bar{\Psi}(r)/\sqrt{1-R^2}, rG(1) + \bar{\Psi}(r)/\sqrt{1-R^2})'$, and under $\delta \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\rho = r$, it is

$$\left(1 + \frac{r(1 - R^2 r)\sigma_1^2}{1 - R^2}\right)^{-1/2} \exp\left[-\tfrac{1}{2}\frac{\mu_1^2}{\sigma_1^2} + \tfrac{1}{2}\frac{\sigma_1^2(\frac{\bar{\Psi}(\rho)}{\sqrt{1-R^2}} + rG(1) + \mu_1/\sigma_1^2)^2}{1 + \sigma_1^2 \frac{r(1-R^2 r)}{1-R^2}} \right].$$

The power bounds $\bar{\pi}_{inv}$ were computed via Lemma 1 using these densities as described in Section 3.4. The inequality $\bar{\pi}_{inv}(R^2) \leq \bar{\pi}_{inv}(\bar{R}^2)$ for $R^2 \leq \bar{R}^2$ holds because one can always reduce the information of the experiment with $\bar{R}^2$ to that with R^2 by adding a data independent Brownian Bridge of appropriate variance to $\bar{\Psi}$.

Condition 1 *In the stable model with parameter Γ_0*

(i) in some neighborhood $\mathcal{B}_0$ of Γ_0, $l_{T,t}(\Gamma)$ is twice differentiable a.s. with respect to Γ for $t = 1, \cdots, T$, $T \geq 1$;

(ii) for all $s \in [0.15, 0.85]$ and $\epsilon > 0$ there exists $K(s, \epsilon) > 0$ such that $P(\sup_{||\Gamma - \Gamma_0|| \geq \epsilon} T^{-1} \sum_{t=1}^{\lfloor sT \rfloor} (l_{T,t}(\Gamma) - l_{T,t}(\Gamma_0)) < -K(s, \epsilon)) \to 1$ and $P(\sup_{||\Gamma - \Gamma_0|| \geq \epsilon} T^{-1} \sum_{t=\lfloor sT \rfloor + 1}^{T} (l_{T,t}(\Gamma) - l_{T,t}(\Gamma_0)) < -K(s, \epsilon)) \to 1;$

34

Table 5: Baseline Distributions for Approximate Least Favorable Distribution

i	$a, b+1$	μ_δ	σ_δ^2
1	$15, 20, 75, 83, 85, 86$	$0, 2, \cdots, 20$	300
2	$15, 20, 75, 83, 85, 86$	$0, 2, \cdots, 16$	50
3	$15, 20, 75, 83, 85, 86$	$0, 2, \cdots, 22$	10
4	$15, 20, 75, 83, 85, 86$	$0, 2, \cdots, 22$	3
5	$15, 20, 30, 45, 60, 70, 75, 80, 83, 85, 86$	$0, 2, \cdots, 22$	10
6	$83, 84, 85, 86$	$0, \frac{1}{2}, \cdots, 18$	1
7	$15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 83, 85, 86$	$0, 1, \cdots, 22$	10

Note: The algorithm described in Section 3.1 is initialized with $(a, b, \mu_\delta, \sigma_\delta^2) \in \{(15, 24, 28, 200), (75, 85, 28, 200), (15, 85, 20, 4), (85, 85, 3, 4), (15, 85, 0, 300)\}$ and goes through 7 rounds in total, indexed by i. Each round increases the set of baseline distributions by the indicated values for $(a, b, \mu_\delta, \sigma_\delta^2)$, with $(a, b+1)$ equal to all consecutive pairs of values in the second column.

(iii) $\{s_{T,t}(\Gamma_0), \mathfrak{F}_{T,t}\}$ is a square-integrable martingale difference array with $T^{-1} \sum_{t=1}^{\lfloor sT \rfloor} E[s_{T,t}(\Gamma_0) s_{T,t}(\Gamma_0)' | \mathfrak{F}_{T,t-1}] \xrightarrow{p} sH$ for all $0 \leq s \leq 1$ and some $(k+1) \times (k+1)$ dimensional matrix H, $T^{-1} \sup_{t \leq T} ||E[s_{T,t}(\Gamma_0) s_{T,t}(\Gamma_0)' | \mathfrak{F}_{T,t-1}]|| \xrightarrow{p} 0$ and there exists $\nu > 0$ such that $T^{-1} \sum_{t=1}^{T} E[||s_{T,t}(\Gamma_0)||^{2+\nu} | \mathfrak{F}_{T,t-1}] = O_p(1)$;

(iv) $T^{-1} \sum_{t=1}^{T} ||h_{T,t}(\Gamma_0)|| = O_p(1)$, $T^{-1} \sup_{t \leq T} ||h_{T,t}(\Gamma_0)|| \xrightarrow{p} 0$ and for any decreasing neighborhoods $\mathcal{B}_T \subset \mathcal{B}_0$ of Γ_0 that converge to $\{\Gamma_0\}$, $T^{-1} \sum_{t=1}^{T} \sup_{\Gamma \in \mathcal{B}_T} ||h_{T,t}(\Gamma) - h_{T,t}(\Gamma_0)|| \xrightarrow{p} 0$;

(v) For all $0 \leq s \leq 1$, $T^{-1} \sum_{t=1}^{\lfloor sT \rfloor} h_{T,t}(\Gamma_0) \xrightarrow{p} -sH$.

Part (ii) of Condition 1 is an identification condition that ensures consistency of the pre and post potential break maximum likelihood estimator—cf., for instance, Condition 6 on page 436 of Schervish (1995). Parts (ii)-(v) are a special case of Condition 2 in Li and Müller (2009) and Condition 1 in Müller and Petalas (2010), who provide further discussion and references.

Details on Size Control Computations:

The test φ^* as defined above only depends on $\{G(l/100)\}_{l=1}^{100}$, and conditional on the Monte Carlo draw of the random walk $\{W_i(l)\}_{l=1}^{100}$, $G_i(l/100) = W_i(l/100) + \delta \min(\rho, l/100)$ is a function of δ and ρ. Write $S_i(\delta, \rho)$ for the induced function $\varphi^*(G_i)$. The approach outlined in Section 3.3 requires that for any given values $0 \leq \delta_1 < \delta_2$ and $0.15 \leq \rho_1 < \rho_2 \leq 0.85$, an algorithm must (attempt to) determine $\min_{\delta_1 \leq \delta < \delta_2, \rho_1 \leq \rho < \rho_2} S_i(\delta, \rho)$ and $\max_{\delta_1 \leq \delta < \delta_2, \rho_1 \leq \rho < \rho_2} S_i(\delta, \rho)$ (it suffices to consider non-negative δ, because $\varphi^*(G) = \varphi^*(-G)$). For this determination, we now consider how $\sup F_0$, $\hat{l}_0$, $\hat{t}_0$ and LR_0 in the definition of φ^* at the beginning of the appendix behave as functions of (δ, ρ) on the rectangles $\delta_1 \leq \delta < \delta_2, \rho_1 \leq \rho < \rho_2$ where ρ_1 and ρ_2 are such that $l_0 - 1 \leq 100\rho_1$ and $100\rho_2 < l_0$ for some $l_0 \in \{16, \cdots, 85\}$, conditional on W_i.

For $\sup F_0$, note that $((100-l)G(\frac{l-1}{100}) - (l-1)(G(1) - G(\frac{l}{100})))/\sqrt{(100-l)(l-1)} = a^{\sup F}(l, W_i) + \delta \cdot b^{\sup F}(l, \rho)$, where $a^{\sup F}(l, W_i)$ does not depend on (ρ, δ), and $b^{\sup F}(l, \rho) = b_{pre}^{\sup F}(l, \rho) = \sqrt{(l-1)/(100-l)}(1-\rho)$ for $l < l_0$,

$b^{\text{supF}}(l,\rho) = b^{\text{supF}}_{mid}(l) = \sqrt{(100-l)(l-1)}/100$ for $l = l_0$ and $b^{\text{supF}}(l,\rho) = b^{\text{supF}}_{post}(l,\rho) = \sqrt{(100-l)/(l-1)}\rho$ for $l > l_0$. Thus,

$$\begin{aligned}
\delta_1 b^{\text{supF}}_{pre}(l,\rho_2) &\leq \delta b^{\text{supF}}(l,\rho) \leq \delta_2 b^{\text{supF}}_{pre}(l,\rho_1) && \text{for } 16 \leq l < l_0 \\
\delta_1 b^{\text{supF}}_{mid}(l) &\leq \delta b^{\text{supF}}(l,\rho) \leq \delta_2 b^{\text{supF}}_{mid}(l) && \text{for } l = l_0 \\
\delta_2 b^{\text{supF}}_{post}(l,\rho_1) &\leq \delta b^{\text{supF}}(l,\rho) \leq \delta_2 b^{\text{supF}}_{pre}(l,\rho_2) && \text{for } l_0 < l \leq 85
\end{aligned}$$

which in turn allows the construction of straightforward bounds for $\sqrt{\text{supF}_0}$.

For $\hat{l}_0$, note that with $\xi = \rho\delta$, $S_0(l)$ is of the form $S_0(l) = S_0(l,\delta,\xi) = \sum_{j=1}^{2}(a_j^S(l,W_i) + b_{i,j}^S(l)\delta + c_{i,j}^S(l)\xi)^2$ for $i \in \{pre, mid, post\}$ in self-explanatory notation. In particular, $b_{pre,1}^S(l) = l$, $b_{post,1}^S(l) = b_{mid,1}^S(l) = 0$, $b_{pre,2}^S(l) = 0$, $b_{mid,2}^S(l) = (l-1)l/100$, $b_{post,2}^S(l) = 0$, $c_{pre,1}^S(l) = -l$, $c_{post,1}^S(l) = c_{mid,1}^S(l) = 100 - l$, $c_{pre,2}^S(l) = 0$, $c_{mid,2}^S(l) = -(l-1)$, and $c_{post,2}^S(l) = 1$. Let $\bar{\rho} = \frac{1}{2}(\rho_1 + \rho_2)$, $\bar{\delta} = \frac{1}{2}(\delta_1 + \delta_2)$, and $\bar{\xi} = \bar{\rho}\bar{\delta}$, and define $\widehat{\bar{l}} = \arg\max_{16 \leq l \leq 85} S_0(l,\bar{\delta},\bar{\xi})$. With $S_\Delta(l,\delta,\xi) = S_0(l,\delta,\xi) - S_0(\widehat{\bar{l}},\delta,\xi)$ and $\mathcal{T} = [\delta_1,\delta_2] \times [\xi_1,\xi_2)$, $\xi_1 = \delta_1\rho_1$, $\xi_2 = \delta_2\rho_2$, the set of possible values for $\hat{l}_0$ is a (possibly proper) subset of $\hat{L}_0 = \{16 \leq l \leq 85 : \sup_{(\delta,\xi)\in\mathcal{T}} S_\Delta(l,\delta,\xi) \geq 0\}$, because the range of $S_\Delta(l,\delta,\xi)$ on $\mathcal{T}$ is at least as large as the range of $S_\Delta(l,\delta,\rho\delta)$ on $[\delta_1,\delta_2] \times [\rho_1,\rho_2]$, for each l. Since $S_\Delta(l,\delta,\xi)$, viewed as a function of δ and ξ, is a quadratic polynomial, the minimum and maximum occur either in a corner $(\delta,\xi) \in \mathcal{C} = \{(\delta_1,\xi_1),(\delta_1,\xi_2),(\delta_1,\xi_2),(\delta_2,\xi_2)\}$, or, possibly, at a local maximum along one of the four sides, or, possibly, at an interior extremum. The five potential interior local extrema are easily computed by solving the appropriate linear first order conditions.

Given $\hat{L}_0$, it is straightforward to construct upper and lower bounds on $\hat{t}_0$.

Finally, for LR_0, note that the exponents in the numerator is a positive definite quadratic polynomial in δ and $\xi = \rho\delta$. Expanding $\cosh(x) = \frac{1}{2}\exp(x) + \frac{1}{2}\exp(-x)$, also the denominator can be written as a sum of exponentials where each exponent is a positive definite quadratic polynomial in (δ,ξ). Since positive definite quadratic polynomials are convex, and the sum of log-convex functions is log convex, both the numerator and denominator are log-convex functions of (δ,ξ), say, $N_{\text{LR}}(\delta,\xi)$ and $D_{\text{LR}}(\delta,\xi)$. Let $\bar{v}_D$ be the 2×1 vector of partial derivatives of $\ln D_{\text{LR}}(\delta,\xi)$ at $(\delta,\xi) = (\bar{\delta},\bar{\xi})$. Since $\ln D_{\text{LR}}(\delta,\xi) - (\delta - \bar{\delta}, \xi - \bar{\xi})\bar{v}_D$ is convex with zero derivative at $(\delta,\xi) = (\bar{\delta},\bar{\xi})$, it takes on its minimum at $(\bar{\delta},\bar{\xi})$, and since $\ln N_{\text{LR}}(\delta,\xi) - (\delta - \bar{\delta}, \xi - \bar{\xi})\bar{v}_D$ is convex, it takes on its maximum in one of the corners. Thus,

$$\max_{\delta_1 \leq \delta < \delta_2, \rho_1 \leq \rho < \rho_2} \text{LR}_0 \leq \frac{\max\limits_{(\delta,\xi)\in\mathcal{C}} N_{\text{LR}}(\delta,\xi)\exp[-(\delta-\bar{\delta},\xi-\bar{\xi})\bar{v}_D]}{D_{\text{LR}}(\bar{\delta},\bar{\xi})}$$

and, reversing the role of the numerator and denominator, we can construct an analogous lower bound.

These arguments were applied in the context of the divide and conquer algorithm on $(\rho,\delta) \in [0.15, 0.85] \times [0,110]$, with rectangles of dimension $1/10000 \times 1/100$. For $\delta \in (110,\infty)$, consider a rectangle $[\rho_1,\rho_2] \times (110,\infty)$ with $l_0 - 1 \leq 100\rho_1$ and $100\rho_2 < l_0$ for some $l_0 \in \{16,\cdots,85\}$. If $\hat{l}_0 \in \{l_0 - 1, l_0\}$ occurred with certainty, then $\hat{t}_0$ is based on a stable model, and the same arguments as employed in the proof of Proposition 4 (ii) below show that the rejection probability is slightly smaller than 5% for all l_0. It thus suffices to employ the (appropriately modified) algorithm for the determination of $\hat{L}_0$ above to check that $\hat{l}_0 \notin \{l_0 - 1, l_0\}$ rarely enough.

Details on Power Bound Computations:

A standard Wiener process on the unit interval W can be decomposed as $W(s) = D_W^\tau(s) + \alpha_W + (s - 1/2)\beta_W$, where α_W and β_W are the coefficients of a continuous time least squares regression of $W(s)$ on $(1, s - 1/2)$, and $D_W^\tau(s)$ are the residuals. Let $D_W^{\mathrm{sq}} = \int_0^1 D_W^\tau(s)^2 ds$. With these definitions, a straightforward calculation yields $\int_0^1 W(s)^2 ds = D_W^{\mathrm{sq}} + \alpha_W^2 + \frac{1}{12}\beta_W^2$. Further, define $D_W(s)$ as the demeaned Brownian Bridge $W(s) - sW(1)$, $D_W(s) = W(s) - sW(1) - \alpha_W + \frac{1}{2}W(1)$, and note that $\Delta_W = \sup_{0 \le s \le 1} |W(s) - \alpha_W| \le \frac{1}{2}|W(1)| + \sup_{0 \le s \le 1} |D_W(s)|$ and $\Delta_W^{\mathrm{sq}} = \sup_{0 \le s \le 1} |W(s)^2 - \int_0^1 W(s)^2 ds| \le 2\Delta_W |\alpha_W| + \Delta_W^2 + |\int_0^1 W(s)^2 ds - \alpha_W^2|$. These expressions are useful because (i) $(W(1), \alpha_W, \beta_W)$ are jointly Gaussian with known covariance matrix and independent of $D_W^\tau(\cdot)$ (and thus D_W^{sq}); (ii) the distribution of D_W^{sq} can be computed exactly by inverting the characteristic function derived in Theorem 9.12 of Tanaka (1996); (iii) Darling (1983) provides a formula for the c.d.f. of $\sup_{0 \le s \le 1} D_W(s)$.

Evaluation of the expression in Darling (1983) yields $P(\sup_{0 \le s \le 1} D_W(s) > 2.2) < \varepsilon_D = 5 \cdot 10^{-12}$, so that $P(\sup_{0 \le s \le 1} |D_W(s)| > 2.2) < 2\varepsilon_D$. Also, evaluation of the c.d.f. of D_W^{sq} at the 15 points $0.1, 0.2, \cdots, 1.5$ by inverting the characteristic function, as discussed in Chapter 6 of Tanaka (1996), permits the straightforward construction of the two random variables $D_W^{\mathrm{sq}\,l}$ and $D_W^{\mathrm{sq}\,u}$ such that $P(D_W^{\mathrm{sq}\,l} \le D_W^{\mathrm{sq}} \le D_W^{\mathrm{sq}\,u}) = 1$ by suitably transforming uniform random variable on the unit interval (with $D_W^{\mathrm{sq}\,u}$ taking on the value $+\infty$ with probability $P(D_W^{\mathrm{sq}} > 1.5) \simeq 4 \cdot 10^{-14}$). Let $(Z_j, \alpha_{Wj}, \beta_{Wj})$ be n i.i.d. Monte Carlo draws from the distribution of $(W(1), \alpha_W, \beta_W)$, and let $(D_{Wj}^{\mathrm{sq}\,l}, D_{Wj}^{\mathrm{sq}\,u})$ be n independent i.i.d. Monte Carlo draws from the distribution of $(D_W^{\mathrm{sq}\,l}, D_W^{\mathrm{sq}\,u})$, $j = 1, \cdots, n$. Then define $S_{1,j}^u = 2.2 + 0.5|Z_j|$ and $S_{2,j}^u = 2|\alpha_{Wj}|S_{1,j}^u + (S_{1,j}^u)^2 + |D_{Wj}^{\mathrm{sq}\,u} + \frac{1}{12}\beta_{Wj}^2|$, $W_{b,j+1} = W_{b,j} + Z_j/\sqrt{n}$, $\bar{W}_j = W_{bj} + \alpha_{Wj}/\sqrt{n}$, $\overline{W_j^2}^l = (D_{Wj}^{\mathrm{sq}\,l} + \alpha_{Wj}^2 + \frac{1}{12}\beta_{Wj}^2)/n + 2\alpha_{Wj}W_{bj}/\sqrt{n} + W_{bj}^2$ and $\overline{W_j^2}^u = \overline{W_j^2}^l + (D_{Wj}^{\mathrm{sq}\,u} - D_{Wj}^{\mathrm{sq}\,l})/n$ for $j = 1, \cdots, n$, where $W_{b,0} = 0$. With the implemented number of $n = 800,000$ intervals, a draw $\{\overline{W_j^2}^l, \overline{W_j^2}^u, \bar{W}_j, S_{1,j}^u, S_{2,j}^u\}_{j=1}^n$ satisfies $\overline{W_j^2}^l \le \overline{W_j^2} \le \overline{W_j^2}^u$, $S_{1,j} \le S_{1,j}^u$ and $S_{2,j} \le S_{2,j}^u$ for all $j = 1, \cdots, n$ with probability of at least $(1 - 2\varepsilon_D)^n > 1 - 7 \cdot 10^{-5}$ by construction. Thus, very slight corrections to the estimated critical value and Monte Carlo rejection probability yields an unconditionally conservative Monte Carlo estimator of the power bounds.

Proof of Proposition 4:

(i) The CMT, and, in the case of $\hat{l}$, some tedious algebra, imply that $\hat{l} \Rightarrow \hat{l}_0$ and $\widehat{\mathrm{supF}} \Rightarrow \mathrm{supF}_0$ with $\hat{l}_0$ and supF_0 as defined the definition of φ^* above (note that replacing G by G_0 leaves $\hat{l}_0$ and supF_0 unchanged).

(i.a) Since $G(s) = G_0(s) + \beta s$, the result follows from the CMT.

(i.b) We have $\hat{t}_{post}/\beta_T \Rightarrow 1/\sqrt{1 - \min(\hat{l}_0 + 1, 85)/100}$ and a calculation yields $\beta_T^{-2} \ln \widehat{\mathrm{LR}} \xrightarrow{p} c > 0$, so that $P(|\hat{t}_{post}| > 2.01) \to 1$ and $P(\widehat{\mathrm{LR}} > 2.41) \to 1$.

(ii) Define $l_0 = \lceil 100\rho \rceil$. Note that $\delta_T^{-2}T \left(\hat{\gamma}_{post}(l_0) - \hat{\gamma}_{pre}(l_0 - 1)\right)^2 \xrightarrow{p} \omega^2 > 0$, $T\hat{\omega}_{post}^2(l_0) \xrightarrow{p} \omega^2/(1 - l_0/100) > 0$, and $T\hat{\omega}_{pre}^2(l_0 - 1) \xrightarrow{p} 100\omega_{pre}^2/(l_0 - 1) > 0$ so that $P(\widehat{\mathrm{supF}} > 90) \to 1$. Let

$$\hat{S}(l) = \sum_{j=1}^{l-1}(\hat{\Delta}_{pre}(j) - \hat{\gamma}_{pre}(l - 1))^2 + \sum_{j=l+1}^{100}(\hat{\Delta}_{post}(j) - \hat{\gamma}_{post}(l))^2$$

where $\hat{\Delta}_{pre}(j) = \hat{\gamma}_{pre}(15)$ for $j \le 15$ and $\hat{\Delta}_{post}(j) = \hat{\gamma}_{post}(85)$ for $j > 85$. Note that $\hat{l} = \arg\min_{16 \le l \le 85} \hat{S}(l)$, because $\sum_{j=1}^l \hat{\Delta}_{pre}(j) = l\hat{\gamma}_{pre}(l)$ and $\sum_{j=l+1}^{100} \hat{\Delta}_{post}(j) = (100 - l)\hat{\gamma}_{post}(l)$. If $\rho \notin R_{grid}$, then $\hat{S}(l_0) = O_p(T^{-1})$, and for all $l \ne l_0$, $P(T\hat{S}(l) > M) \to 1$ for any $M \in \mathbb{R}$, so that $\hat{l} \xrightarrow{p} l_0$. Part (ii.b), and part (ii.a) for $\rho \notin R_{grid}$, now follow from the CMT.

37

Consider thus the proof of part (ii.a) if $\rho \in R_{grid}$. By the same argument, $P(\hat{l} \notin \{l_0, l_0 + 1\}) \to 0$. Define $t_{post}(l_0 + 1) = (W(1) - W(\frac{l_0+1}{100}))/\sqrt{1 - \frac{l_0+1}{100}} \sim \mathcal{N}(0,1)$ and $t_{post}(l_0 + 2) = (W(1) - W(\frac{l_0+2}{100}))/\sqrt{1 - \frac{l_0+2}{100}} \sim \mathcal{N}(0,1)$, and consider

$$\hat{S}(l_0) - \hat{S}(l_0 + 1) = A_{pre} + \hat{\Delta}_{post}(l_0 + 1)^2 - (100 - l_0)\hat{\gamma}_{post}(l_0)^2 + (99 - l_0)\hat{\gamma}_{post}(l_0 + 1)^2$$
$$\Rightarrow \Delta S(l_0) = a \cdot t_{post}(l_0 + 1)^2 + t_{post}(l_0 + 1)B(l_0) + C(l_0)$$

where A_{pre} is a function of $\{\hat{\gamma}_{pre}(l)\}_{l=15}^{l_0}$, the convergence follows from the CMT, a is a positive constant, and $B(l_0)$ and $C(l_0)$ are random variables that are independent of $t_{post}(l_0 + 1)$ and $t_{post}(l_0 + 2)$. Thus, by a further application of the CMT, we obtain

$$\hat{t}_{post} \Rightarrow \mathbf{1}[\Delta S(l_0) \geq 0]\mathbf{1}[|t_{post}(l_0 + 2)| > 2.01] + \mathbf{1}[\Delta S(l_0) < 0]\mathbf{1}[|t_{post}(l_0 + 1)| > 2.01]. \tag{34}$$

Consider now the rejection probability of the r.h.s. of (34) conditional on $B(l_0) = b$ and $C(l_0) = c$. Noting that $\Delta S(l_0)$ is a quadratic function of $t_{post}(l_0 + 1)$ with positive coefficient on the square term, one obtains that the rejecting probability is bounded from above by the replacement of $\mathbf{1}[\Delta S(l_0) \geq 0]$ by $\mathbf{1}[t_{post}(l_0 + 1) < -2.01]$, for any value of b, c. A numerical calculation now shows that the conditional rejection probability remains below 5% even in that case for all $l_0 = 15, \cdots, 83$, and the result follows.

(ii.b) Immediate from $P(\hat{l} \notin \{l_0, l_0 + 1\}) \to 0$ and $\beta_T^{-1}(\hat{\gamma}_{post}(l) - \gamma_{post,0})/\hat{\omega}_{post}(l) \xrightarrow{p} \sqrt{1 - l/100}/\omega > 0$ for all $l \geq l_0$.

(iii) Let θ_T be a parameter sequence such that $\limsup_{T \to \infty} E_{\theta_T}[\hat{\varphi}^*(X_T)] = \limsup_{T \to \infty} \sup_{\theta \in \Theta_0} E_\theta[\hat{\varphi}^*(X_T)]$. Pick a subsequence T' of T such that $\lim_{T' \to \infty} E_{\theta_{T'}}[\hat{\varphi}^*(X_{T'})] = \limsup_{T \to \infty} E_{\theta_T}[\hat{\varphi}^*(X_T)]$. Since $\bar{\Theta}_0 = \Theta_0 \cup \{(0, \rho, \delta) : \rho \in [0.15, 0.85], \delta \in \{+\infty, -\infty\}\}$ is compact under an appropriate metric, there exists a further subsequence $\theta_{T''}$ of $\theta_{T'}$ such that $\theta_{T''} \to \tilde{\theta} = (0, \tilde{\rho}, \tilde{\delta}) \in \bar{\Theta}_0$. If $\tilde{\rho} \notin R_{grid}$, then the result follows as in parts (i) and (ii). If $\tilde{\rho} \in R_{grid}$, then by assumption, there is a further subsequence such that (31) holds, except for (32). Proceeding as in the proof of part (ii.a) now again yields the result (and for the case $\nu_{pre}(l_0) \in \{+\infty, -\infty\}$ or $\nu_{post}(l_0) \in \{+\infty, -\infty\}$, $P(|\hat{S}(l_0) - \hat{S}(l_0 + 1)| < M) \to 0$ for all M, so that $\mathbf{1}[\Delta S(l_0) \geq 0]$ in (34) is replaced by either zero or one).

Justification of (30) via Andrews (1993):

The true parameters are given by $\Gamma_{T,t} = \Gamma_{post} + T^{-1/2}\mathbf{1}[\tau \leq 100\rho_T]\Delta_T$, where $\Delta_T = -(\omega\delta_T, \omega\tilde{\delta}'_T)' \to \Delta_0 \in \mathbb{R}^{k+1}$ and $\rho_T \to \rho_0 \in [0.15, 0.85]$, so that $E[g(X_{T,t}, \Gamma_{T,t})] = 0$. Write $g_t(\Gamma)$ for $g(X_{T,t}, \Gamma)$, and define $\Upsilon_t(\Gamma) = \partial g_t(\Gamma)/\partial\Gamma'$. Our set-up corresponds to what Andrews (1993) refers to as 'pure' structural change. We now show that under Assumption 1 of Andrews (1993), with the assumption in part (b) $E[g_t(\Gamma_0)] = 0$ replaced by $E[g_t(\Gamma_{T,t})] = 0$, (30) holds. These computations closely correspond to what is derived in Section 5.4 of Andrews (1993), with two differences: (i) we do not assume that $\max_{t \leq T} \sup_{\|\Gamma - \Gamma_{post}\| \leq T^{-1/2}K} \|E[\Upsilon_t(\Gamma)] - \bar{\Upsilon}\| \to 0$ for some $K > 0$ (cf. first line on page 832), where in his Assumption 1(f), $\bar{\Upsilon}$ is defined as the unique limit $T^{-1}\sum_{t=1}^{\lfloor \cdot T\rfloor} E\Upsilon_t(\Gamma_{post}) \xrightarrow{p} \cdot\bar{\Upsilon}$; (ii) we consider sequences for Δ_T and ρ_T, so that $(\Delta_T, \rho_T) \to (\Delta_0, \rho_0)$. (We note that there is a typo in Assumption 1-LP of Andrews (1993); it should read $\sup_{\pi\in\Pi} \|\sqrt{T}E\bar{m}_T(\theta_0, \pi) - \mu(\pi)\| \to 0$ in his notation.)

Define $\eta_T : [0,1] \mapsto \mathbb{R}^{k+1}$ as $\eta_T(x) = \mathbf{1}[x \leq \rho_T]\Delta_T$ for $T = 0, 1, \ldots$. By the argument in Andrews' proof of his Theorem 4 (a), it suffices to show that $\sup_{0\leq s\leq 1} \|T^{-1/2}E[\sum_{t=1}^{\lfloor sT\rfloor} g_t(\Gamma_{post})] - \bar{\Upsilon}\int_0^s \eta_0(l)dl\| \to 0$. Now clearly $\sup_{0\leq s\leq 1} \|\lim_{T\to\infty} T^{-1}\sum_{t=1}^{\lfloor sT\rfloor} \eta_T(t/T) - \int_0^s \eta_0(l)dl\| \to 0$. Also,

$$T^{-1/2}E[\sum_{t=1}^{\lfloor sT\rfloor} g_t(\Gamma_{post}) - T^{-1/2}\bar{\Upsilon}\eta_T(t/T)] = T^{-1/2}E[\sum_{t=1}^{\lfloor \min(s,\rho_T)T\rfloor} g_t(\Gamma_{post}) - T^{-1/2}\bar{\Upsilon}\eta_T(t/T)]$$
$$= (E[\tilde{\Upsilon}_T(\min(s, \rho_T))] - \bar{\Upsilon})\Delta_T$$

where $\tilde{\Upsilon}_T(r)$ is equal to $T^{-1}\sum_{t=1}^{\lfloor rT\rfloor}\int_0^1 \Upsilon_t(\Gamma_{post} + \lambda T^{-1/2}\Delta_T)d\lambda$. Furthermore,

$$\sup_{0\leq s\leq 1} \|E[\tilde{\Upsilon}_T(\min(s, \rho_T) - \tilde{\Upsilon}_T(\min(s, \rho_0)]\| \leq T^{-1}E \sum_{t=\lfloor\min(\rho_0,\rho_T)T\rfloor}^{\lfloor\max(\rho_0,\rho_T)T\rfloor} \|\int_0^1 \Upsilon_t(\Gamma_{post} + \lambda T^{-1/2}\Delta_T)d\lambda\|$$
$$\leq (\lfloor\max(\rho_0,\rho_T)T\rfloor - \lfloor\min(\rho_0,\rho_T)T\rfloor)\sup_{t\leq T}E[\sup_{\Gamma\in\mathcal{G}_0} \|\Upsilon_t(\Gamma)\|] \to 0$$

where the convergence follows, since by Andrews' Assumption 1(f), $\sup_{t\leq T} E[\sup_{\Gamma\in\mathcal{G}_0} \|\Upsilon_t(\Gamma)\|] = O(1)$ for some neighborhood $\mathcal{G}_0$ of Γ_{post}. Finally, $\sup_{0\leq s\leq 1} \|E[\tilde{\Upsilon}_T(\min(s, \rho_0))] - \bar{\Upsilon}\| \to 0$ follows from the same reasoning as below Andrews' equation (A.11).

References

ANDERSON, R. J., AND K. L. KLIESON (2006): "The 1990's Acceleration in Labor Productivity: Causes and Measurement," *Federal Rserve Bank of St. Louis Review*, 88, 181–202.

ANDREWS, D., AND G. GUGGENBERGER (2010a): "Asymptotic Size and a Problem with Subsampling and with the M Out of N Bootstrap," *Econometric Theory*, 26, 426–468.

ANDREWS, D., AND P. GUGGENBERGER (2009): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77, 721–762.

——— (2010b): "Applications of subsampling, hybrid, and size-correction methods," *Journal of Econometrics*, 158, 285–305.

ANDREWS, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

——— (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.

ANDREWS, D. W. K., M. J. MOREIRA, AND J. H. STOCK (2006): "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74, 715–752.

——— (2008): "Efficient Two-Sided Nonsimilar Invariant Tests in IV Regression with Weak Instruments," *Journal of Econometrics*, 146, 241–254.

ANDREWS, D. W. K., AND W. PLOBERGER (1994): "Optimal Tests When a Nuisance Parameter Is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.

BAI, J. (1994): "Least Squares Estimation of a Shift in Linear Processes," *Journal of Time Series Analysis*, 15, 453–470.

——— (1997): "Estimation of a Change Point in Multiple Regressions," *Review of Economics and Statistics*, 79, 551–563.

BAI, J., AND P. PERRON (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47–78.

BUNZEL, H., AND T. J. VOGELSANG (2005): "Powerful Trend Function Tests That are Robust to Strong Serial Correlation with an Application to the Prebisch-Singer Hypothesis," *Journal of Business and Economic Statistics*, 23, 381–394.

COGLEY, T., AND T. J. SARGENT (2005): "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII US," *Review of Economic Dynamics*, 8, 262–302.

DARLING, D. A. (1983): "On the Supremum of a Certain Gaussian Process," *The Annals of Probability*, 11, 803–806.

DUFOUR, J.-M. (2006): "Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics," *Journal of Econometrics*, 133, 443–477.

ELLIOTT, G., AND U. K. MÜLLER (2006): "Efficient Tests for General Persistent Time Variation in Regression Coefficients," *Review of Economic Studies*, 73, 907–940.

——— (2007): "Confidence Sets for the Date of a Single Break in Linear Time Series Regressions," *Journal of Econometrics*, 141, 1196–1218.

ELLIOTT, G., U. K. MÜLLER, AND M. W. WATSON (2012): "Nearly Optimal Tests When a Nuisance Parameter is Present Under the Null Hypothesis," *Working Paper, Princeton University*.

GORDON, R. J. (2003): "Exploding Productivity Growth: Context, Causes, and Implications," *Brookings Papers on Economic Activity*, 2, 207–279.

GUGGENBERGER, P. (2010): "The impact of a Hausman pretest on the asymptotic size of a hypothesis test," *Econometric Theory*, 26, 369–382.

HALL, A., S. HAN, AND O. BOLDEA (2010): "Inference regarding multiple structural changes in linear models with endogenous regressors," *Journal of Econometrics*, 170, 281–302.

HANSEN, B. E. (2000): "Testing for Structural Change in Conditional Models," *Journal of Econometrics*, 97, 93–115.

JORGENSON, D., M. S. HO, AND K. J. STIROH (2008): "A Retrospective Look at the US Productivity Growth Resurgence," *Journal of Economic Perspectives*, 22, 3–24.

KING, M. L. (1988): "Towards a Theory of Point Optimal Testing," *Econometric Reviews*, 6, 169–218.

LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*. Springer, New York.

LI, H., AND U. K. MÜLLER (2009): "Valid Inference in Partially Unstable General Method of Moment Models," *Review of Economic Studies*, 76, 343–365.

MÜLLER, U. K. (2011): "Efficient Tests under a Weak Convergence Assumption," *Econometrica*, 79, 395–435.

MÜLLER, U. K., AND P. PETALAS (2010): "Efficient Estimation of the Parameter Path in Unstable Time Series Models," *Review of Economic Studies*, 77, 1508–1539.

NEWEY, W. K., AND K. WEST (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

ORPHANIDES, A. (2004): "Monetary Policy Rules, Macroeconomic Stability, and Inflation: A view from the Trenches," *Journal of Money, Credit, and Banking*, 36, 151–175.

PERRON, P., AND Y. YAMAMOTO (2008): "Estimating and Testing Multiple Structural Changes in Models with Endogenous Regressors," *Working paper, Boston University*.

SCHERVISH, M. J. (1995): *Theory of Statistics*. Springer, New York.

SRIANANTHAKUMAR, S., AND M. L. KING (2006): "A New Approximate Point Optimal Test of a Composite Null Hypothesis," *Journal of Econometrics*, 130, 101–122.

STAIGER, D., AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

TANAKA, K. (1996): *Time Series Analysis — Nonstationary and Noninvertible Distribution Theory*. Wiley, New York.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

WALD, A. (1943): "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Transactions of the American Mathematical Society*, 54, 426–482.