DEPARTMENT OF ECONOMICS AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

Fixed-b Inference in the Presence of Time-Varying Volatility

Matei Demetrescu, Christoph Hanck and Robinson Kruse

CREATES Research Paper 2016-1

Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Fixed-b Inference in the Presence of Time-Varying Volatility*

Matei Demetrescu[a], Christoph Hanck[b] and Robinson Kruse[c]

[a]Christian-Albrechts-University of Kiel[†] [b]University of Duisburg-Essen[‡]

[c]Rijksuniversiteit Groningen and CREATES, Aarhus University[§]

January 5, 2016

Abstract

The fixed-b asymptotic framework provides refinements in the use of heteroskedasticity and autocorrelation consistent variance estimators. The resulting limiting distributions of t-statistics are, however, not pivotal when the unconditional variance changes over time. Such time-varying volatility is an important issue for many financial and macroeconomic time series. To regain pivotal fixed-b inference under time-varying volatility, we discuss three alternative approaches. We (i) employ the wild bootstrap (Cavaliere and Taylor, 2008, ET), (ii) resort to time transformations (Cavaliere and Taylor, 2008, JTSA) and (iii) consider to select test statistics and asymptotics according to the outcome of a heteroskedasticity test, since small-b asymptotics deliver standard limiting distributions irrespective of the so-called variance profile of the series. We quantify the degree of size distortions from using the standard fixed-b approach assuming homoskedasticity and compare the effectiveness of the corrections via simulations. It turns out that the wild bootstrap approach is highly recommendable in terms of size and power. An application to testing for equal predictive ability using the Survey of Professional Forecasters illustrates the usefulness of the proposed corrections.

Keywords: Hypothesis testing, HAC estimation, HAR testing, Bandwidth, Robustness

JEL classification: C12 (Hypothesis Testing), C32 (Time-Series Models)

---

*The authors would like to thank seminar and conference participants in Aarhus, Bath, Cologne, Konstanz, Maastricht and Münster, in particular Ulrich Müller for very useful comments and remarks. The first two authors gratefully acknowledge the support of the German Research Foundation (DFG) through the projects DE 1617/4-1 and HA 6766/2-1. Robinson Kruse gratefully acknowledges financial support from CREATES funded by the Danish National Research Foundation (DNRF78).

[†]Institute for Statistics and Econometrics, Christian-Albrechts-University of Kiel, Olshausenstr. 40-60, D-24118 Kiel, Germany, email: `mdeme@stat-econ.uni-kiel.de`.

[‡]**Corresponding author:** Faculty of Economics and Business Administration, University of Duisburg-Essen, Universitätsstraße 12, D-45117 Essen, Germany, e-mail address: `christoph.hanck@vwl.uni-due.de`.

[§]Rijksuniversiteit Groningen, Faculty of Economics and Business, PO Box 800, 9700 AV Groningen, The Netherlands. E-mail address: `y.r.kruse@rug.nl` and CREATES, Aarhus University, School of Economics and Management, Fuglesangs Allé 4, DK-8210 Aarhus V, Denmark, e-mail address: `rkruse@creates.au.dk`.

1 Introduction

Sound statistical inference takes properties of the data such as heteroskedasticity or serial dependence into account. For weakly stationary series, the seminal contributions of Newey and West (1987) and Andrews (1991) provide GMM (Hansen, 1982) hypothesis tests which are robust to the potential presence of such features. Relying on a heteroskedasticity- and autocorrelation consistent [HAC] estimator for the long-run variance, this framework permits to use critical values from standard distributions, like the standard normal or χ^2-distribution. These asymptotic distributions, however, turn out to be rather poor approximations to the actual finite-sample distributions. As a consequence, substantial size distortions may arise in applied work. In particular, test results turn out to be sensitive to the choice of bandwidth B and kernel k employed for estimating the long-run variance. The poor performance of the asymptotic approximation may be explained by the requirement that a vanishing fraction $b := {}^{B}\!/_{T} \to 0$ of the number of observations T be used for estimation. In actual applications, b must of course be positive. In additon, the asymptotic approximation is independent of k.

To tackle these finite-sample issues with HAC long-run variance estimation, a series of contributions, including Kiefer et al. (2000) and Kiefer and Vogelsang (2002a,b, 2005), proposes a new asymptotic framework, labelled fixed-b asymptotics, in which it is not required that $b \to 0$. This leads to nonstandard distributions (reviewed in more detail in Section 2) for the test statistics, then called heteroskedasticity- and autocorrelation robust [HAR]. Conveniently, the new asymptotic distributions reflect the choice of bandwidth and kernel even in the limit. These papers convincingly demonstrate that the new distributions may provide substantially better approximations to actual finite-sample distributions. In fact, the usefulness of such procedures has spawned an active stream of literature; an incomplete list of recent contributions includes Sun et al. (2008), Yang and Vogelsang (2011), Vogelsang and Wagner (2013) or Sun (2014a,b).

The fixed-b framework, however, does not automatically lead to finite-sample improvements in all empirically relevant settings. Importantly, variances varying over time affect limiting distributions in the fixed-b framework and thus lead to a loss of asymptotic pivotality; see Müller (2014). Time-varying volatility is present in many financial (see among others Guidolin and Timmermann, 2006; Amado and Teräsvirta, 2014; Teräsvirta and Zhao, 2011; Amado and Teräsvirta, 2013) and macroeconomic (see e.g. Stock and Watson, 2002; Sensier and van Dijk, 2004; Clark, 2009, 2011; Justiniano and Primiceri, 2008) time series such as excess returns, economic growth or inflation rates. Time-varying volatility includes, but is not limited to, permanent breaks or trends in the variance properties of (the innovations of) the series.[1] Correspondingly, consequences of (and remedies for) time-varying volatility for inference with dependent data have received substantial attention in recent years.[2] Yet, if one lets $b \to 0$ as in the work of Newey

[1]In the macroeconomic literature, a particular such phase of declining volatility at the end of the millennium is known as the "Great Moderation."

[2]For stationary autoregressions see e.g. Phillips and Xu (2006) or Xu (2008); for unit root autoregressions, see Cavaliere and Taylor (2008b) or Cavaliere and Taylor (2009). The effects of time-varying volatility are amplified in panels of (nonstationary) series, making corrections all the more necessary; see e.g. Demetrescu and Hanck (2012) or Westerlund (2014).

and West (1987), time-varying volatilities do not have an asymptotic effect (cf. Cavaliere, 2004). Practitioners thus face a trade-off in the precision of the critical values provided by the fixed-b approach, a trade-off which is determined by the strength of the time variations of the variance in the data generating process [DGP].

Our first contribution is to quantify the extent of distortions due to time-varying volatility in Section 3. They are nontrivial even in the limit, such that fixed-b asymptotics should only be used with additional care. The second contribution of the paper is to discuss methods for valid fixed-b inference even in the presence of time-varying volatility, thus making the trade-off irrelevant. To achieve this, Section 4 builds on techniques inspired by the unit root testing literature, modifying them as required. The first approach builds on the work of Cavaliere and Taylor (2008a) employing a wild bootstrap scheme. Second, we propose to time transform heteroskedastic series following Cavaliere and Taylor (2008b) so as to recover homoskedasticity prior to conducting the test. Since Cavaliere and Taylor (2008b) work with series being integrated of order one, and our setup assumes integration of order zero, the transformation algorithm is modified accordingly. Third, we argue that a pretest for time-varying volatility can also be used for robustification: depending on the outcome of the test, one either uses small-b methods valid under time-varying volatility or fixed-b methods requiring homoskedasticity.

The simulation results presented in Section 5 support the asymptotic discussion. First and as is well-known, the usual HAC-based tests are size-distorted in finite samples when there is serial correlation, a distortion which can be remedied using the fixed-b approach under homoskedasticity. The fixed-b based tests are, however, size distorted under time-varying volatility. Second, the corrections suggested here yield better finite-sample size under time-varying volatility. They also show good size performance under homoskedasticity. The time transformation is somewhat oversized for shorter time series and strong autocorrelation, while the wild bootstrap is only very slightly size-distorted. Third and as one would expect, the pre-test has an intermediate position. Fourth, the wild bootstrap turns out to be substantially more powerful than the time transformation, which may even have zero local power. Summing up, the wild bootstrap approach is highly recommendable for practical purposes.

Section 6 provides an empirical application to testing for equal predictive ability of survey forecasts and simple time-series forecasts, illustrating the potential empirical effect of using testing procedures which are (and which are not) robust to time-varying volatility. Section 7 concludes. The appendices collect proofs, other derivations and additional simulation results.

2 Fixed-b heteroskedasticity- and autocorrelation robust testing

In this paper, we focus on the simple and prototypical case of tests for the finite mean of a series y_t, $\mathrm{E}(y_t) = \mu$. That is, we test H_0: $\mu = \mu_0$. The findings to be presented, however, generalize readily to other testing problems, e.g., to the case of GMM testing of moment restrictions in a regression model, where both shocks or instruments could be subject to time-varying volatility.

Our goal is to provide tests which are robust to the potential presence of both (unconditional) heteroskedasticity and autocorrelation, and the analysis of the univariate case provides good guidance in this respect.

The classical t-test for μ relies on the normalized sample mean,

$$\sqrt{T}\left(\frac{\bar{y}-\mu_0}{\omega}\right),$$

with $\bar{y}=\frac{1}{T}\sum_{t=1}^{T}y_t$ the sample average of y_t and $\omega^2=\lim_{T\to\infty}\mathrm{Var}\big(\sqrt{T}\,(\bar{y}-\mu)\big)$ is the so-called long-run variance of y_t. We thus consider the case of $\sqrt{T}$-consistent sample averages, which are given for independent, both identically and heterogeneously distributed random variables, as well as serially correlated short memory series. The long-run variance—as opposed to the variance of y_t—captures the effect of possible serial correlation or heteroskedasticity on the sample average, hence the acronym HAC for its estimate.

For y_t weakly stationary with absolutely summable autocovariances $\gamma_j=\mathrm{Cov}\,(y_t,y_{t-j})$, it holds that $\omega^2=\sum_{j=-\infty}^{\infty}\gamma_j$. Regularity conditions assumed[3], a central limit theorem applies for $\bar{y}$ and

$$\sqrt{T}\left(\frac{\bar{y}-\mu_0}{\omega}\right)\xrightarrow{d}\mathcal{N}(0,1)$$

under the null.

In practice, the long-run variance ω^2 is unknown and has to be estimated, leading to the feasible t-ratio

$$\mathcal{T}=\sqrt{T}\left(\frac{\bar{y}-\mu_0}{\hat{\omega}}\right)\tag{1}$$

with $\hat{\omega}$ being a suitable estimate of ω.

The most popular HAC estimators rely on suitably weighted sums of sample autocovariances; see Newey and West (1987) and Andrews (1991).[4] Thus,

$$\hat{\omega}^2=\sum_{j=-T+1}^{T-1}k\left(\frac{j}{B}\right)\hat{\gamma}_j$$

where k is the kernel function, B denotes the bandwidth and $\hat{\gamma}_j$ is the jth-order sample autocovariance, i.e. $\hat{\gamma}_j=\frac{1}{T}\sum_{t=j+1}^{T}(y_t-\bar{y})(y_{t-j}-\bar{y})$. Under additional regularity conditions (see e.g. Andrews, 1991), and in particular $b={}^{B}\!/_{T}\to 0$ at suitable rates, consistency follows, $\hat{\omega}\xrightarrow{p}\omega$, and

$$\mathcal{T}\xrightarrow{d}\mathcal{N}(0,1)$$

under H_0. Although this asymptotic result does not depend on the particular choice of a suitable kernel and a bandwidth, the finite-sample behavior of $\mathcal{T}$ hinges on both user inputs. To make

[3]See e.g. Davidson (1994, Chapter 24) for sets of suitable assumptions.

[4]Semiparametric estimates based on AR approximations (e.g. Berk, 1974) or on so-called steep origin kernels (Phillips et al., 2006) are also available in the literature.

this dependence explicit, Kiefer and Vogelsang (2005) let $b \in (0,1]$ for the asymptotic analysis. While the resulting limiting distribution is free of nuisance parameters, it is nonstandard. In particular, it depends directly on the kernel k and indirectly (via b) on the bandwidth B, thus offering second-order refinements to the usual, small-b asymptotics where $b \to 0$; see Sun (2014b). Concretely,

$$\mathcal{T} \xrightarrow{d} \mathcal{B}(k,b),$$

where

$$\mathcal{B}(k,b) = \frac{W(1)}{\sqrt{Q(k,b)}} \tag{2}$$

with $W(s)$ a standard Wiener process,

$$Q(k,b) = -\int_0^1 \int_0^1 \frac{1}{b^2} k''\left(\frac{r-s}{b}\right) (W(r) - rW(1))(W(s) - sW(1))\, \mathrm{d}r\mathrm{d}s$$

for kernels with smooth derivatives, and

$$Q(k,b) = \frac{2}{b}\int_0^1 (W(r) - rW(1))^2 \,\mathrm{d}r - \frac{2}{b}\int_0^{1-b} (W(r+b) - (r+b)W(1))(W(r) - rW(1))\,\mathrm{d}r$$

for the Bartlett kernel.

The corresponding critical values for $\mathcal{T}$ are tabulated as a function of k and b in Kiefer and Vogelsang (2005). For $b \to 0$, $Q(k,b) \xrightarrow{d} 1$, $\mathcal{B}(k,b) \xrightarrow{d} \mathcal{N}(0,1)$ and small-b asymptotics are, in a sense, a special case of the fixed-b approach. Note that the functional $\mathcal{B}(k,b)$ depends on the entire path of the Wiener process and not only on $W(1)$, as is the case with the small-b approach. This has consequences when the volatility of y_t varies in time, as we shall see in the following section.

3 Failure of fixed-b HAR tests under time-varying volatility

In order to analyze fixed-b asymptotics of $\mathcal{T}$ under time-varying volatility, we assume a multiplicative component structure.

Assumption 1. *Let the observed series y_t be generated as*

$$y_t = \mu + h_t v_t, \qquad t = 1, \ldots, T,$$

where the stochastic component v_t is zero-mean stationary as specified below, and time variation in the volatility is induced by the function $h_t = h(t/T)$, also specified below.

This multiplicative structure is common in the literature; see e.g. Cavaliere (2004). This makes y_t a uniformly modulated process (Priestley, 1988, p. 165).[5] To conduct the asymptotic analysis,

[5]Other contributions model v_t explicitly as a linear process with modulated innovations; see e.g. Cavaliere and Taylor (2008a,b). Demetrescu and Sibbertsen (2014) argue that the two DGPs are equivalent for most practical purposes.

we assume the stochastic component to have short memory in the following sense.

Assumption 2. *Let v_t be a zero-mean strictly stationary series with unit long-run variance, $L_{2+\delta}$-bounded for some $\delta > 0$, and strong mixing with coefficients $\alpha(j)$ which satisfy the summability condition $\sum_{j\geq 0} \alpha(j)^{1/p-1/(2+\delta)} < \infty$ for some $2 < p < 2 + \delta$.*

Assuming strong mixing (with coefficients $\alpha(j)$ satisfying a typical summability condition) is a standard way of imposing short memory for v_t (and thus y_t); cf. e.g. Phillips and Durlauf (1986). Restricting the long-run variance to unity is without loss of generality, and allows to interpret h_t^2 as the localized long-run variance of the series y_t. The assumption yields (see e.g. Davidson, 1994, Chapter 29) weak convergence of the partial sums of v_t to a standard Wiener process,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} v_t \Rightarrow W(s),$$

so v_t is integrated of order 0. While y_t, being a modulated version of v_t, is also strong mixing, its partial sums exhibit a limiting behavior depending on the modulating function h_t.

Assumption 3. *Let $h_t = h\left({}^t\!/\!{}_T\right)$ with $h(\cdot)$ a deterministic, piecewise Lipschitz function, positive at all $s \in [0, 1]$.*

This allows for general patterns of smoothly or abruptly changing variances, as long as the abrupt changes are not too frequent.[6]

Under the conditions spelled out by the above assumptions, we have (for details, see Cavaliere, 2004, Lemma 3)

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} (y_t - \mu) \Rightarrow \int_0^s h(v)\, \mathrm{d}W(v) \equiv B_h(s).$$

The process $B_h(s)$ is a Gaussian process, but not a Brownian motion: the covariance kernel of $B_h(s)$ is given by

$$\mathrm{Cov}\left(B_h(s), B_h(r)\right) = \int_0^{\min\{s,r\}} h^2(v)\, \mathrm{d}v$$

which, in general, is not the covariance kernel of the standard Wiener process, $\min\{s, r\}$.

Unsurprisingly, the fact that the normalized partial sums of the centered y_t do not converge weakly to a Brownian motion affects the fixed-b asymptotics of $\mathcal{T}$. The corresponding fixed-b limiting distribution is stated in

Proposition 1. *Under H_0 and Assumptions 1-3, it holds for ${}^B\!/\!{}_T = b \in (0, 1]$ that*

$$\mathcal{T} \xrightarrow{d} \mathcal{B}(h, k, b) \equiv \frac{B_h(1)}{\sqrt{Q_{h,k,b}}}$$

[6] Seasonally varying variances are excluded, for instance. This is not critical, however, since the work of Burridge and Taylor (2001) suggests that seasonally varying variances actually average out and do not affect convergence to Wiener process.

as $T \to \infty$, where

$$Q_{h,k,b} = -\int_0^1 \int_0^1 \frac{1}{b^2} k'' \left(\frac{r-s}{b} \right) (B_h(r) - rB_h(1)) (B_h(s) - sB_h(1)) \, \mathrm{d}r\mathrm{d}s$$

for kernels with smooth derivatives, and

$$Q_{h,k,b} = \frac{2}{b} \int_0^1 (B_h(r) - rB_h(1))^2 \, \mathrm{d}r - \frac{2}{b} \int_0^{1-b} (B_h(r+b) - (r+b)B_h(1)) (B_h(r) - rB_h(1)) \, \mathrm{d}r$$

for the Bartlett kernel.

Proof: *See the Appendix.*

Proposition 1 elaborates upon the statement in Müller (2014, p. 314) that HAR testing via $\mathcal{T}$ is not robust to time-varying volatility for fixed b, giving an exact description of the asymptotic distribution. Although $B_h(1)$ is normal with mean zero and variance $\bar{\omega}^2 = \int_0^1 h^2(s)\mathrm{d}s$, the distribution of $\mathcal{B}(h,k,b)$ is different from that of $\mathcal{B}(k,b)$ whenever h is not constant almost everywhere. This is because $Q_{h,k,b}$ is essentially different from the denominator of (2) under time-varying volatility.

Figure 1 quantifies the lack of pivotality, showing quantile-quantile (QQ) plots for the distributions $\mathcal{B}(h,k,b)$ for $b = \{0.1, 0.5, 0.9\}$ and four different variance patterns h. We take $T = 1000$ and simulate $\mathcal{B}(h,k,b)$ with 50,000 replications. The kernel k is taken to be the Bartlett kernel. Under DGP1, volatility is constant over time. This case is reported as a benchmark, where we compare the quantiles of $\mathcal{B}(k,b)$ with themselves. The first row of graphs in Figure 1 show the results for a small, medium and large value of b. The negligible deviations are due to Monte Carlo variability. An early downward break in volatility is present in DGP2. (For the exact details of the DGPs, see Section 5.) Here, we compare the quantiles of $\mathcal{B}(h,k,b)$ on the y-axis with the corresponding quantiles of $\mathcal{B}(k,b)$ on the x-axis. The results shown in the second row of Figure 1 clearly demonstrate differences between the two distributions. The larger b, the more pronounced is the discrepancy. For DGP 4 (with a double break in volatility), differences are again clearly visible. The results for the linear upward trend in volatility in DGP6 nicely illustrate the difference between the distributions as well as the role of b. The discrepancies are most pronounced in the tails of the distributions, which matter for testing. In our numerical work in Section 5, we err on the safe side so as not to overstate the effects of time-varying volatility and work with a relatively large nominal level of 10%.

For HAC-based tests (or, with some abuse of notation, $b = 0$), robustness to time-varying volatility is recovered. Recall, $B_h(1)$ follows a normal distribution with mean zero and variance $\bar{\omega}^2$, which can be interpreted as the average long-run variance of the series. Moreover, Cavaliere (2004) shows that, under mild conditions on the rate at which b vanishes, the HAC variance estimator is consistent precisely for this variance, $\operatorname{plim} \hat{\omega}^2 = \bar{\omega}^2$. Hence, under H_0,

$$\mathcal{T} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{for} \quad b \to 0.$$

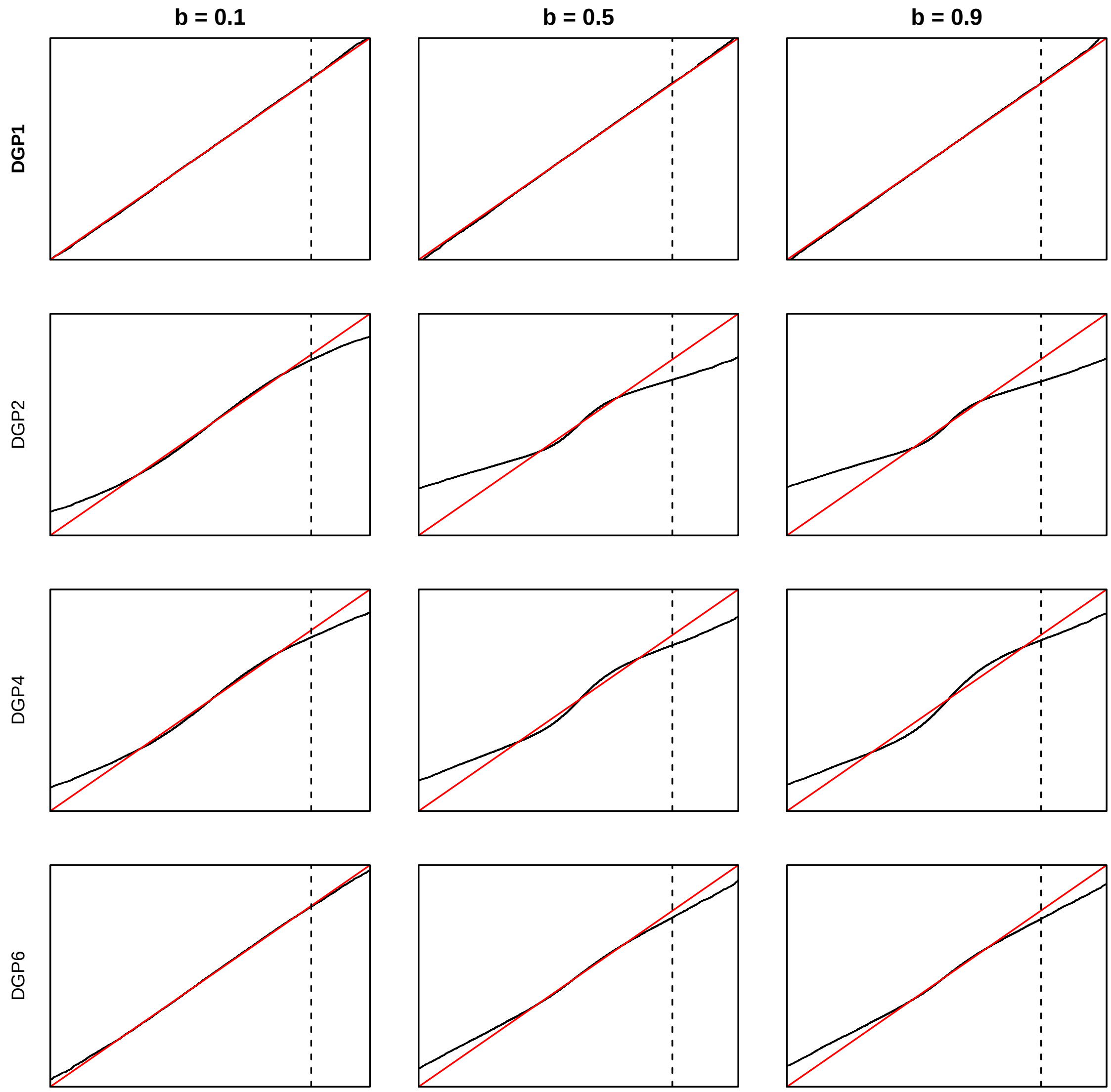

Figure 1: Quantile-quantile plots to compare $\mathcal{B}(k,b)$ (x-axis) to the distributions $\mathcal{B}(h,k,b)$ under various variance profiles h and for different b. DGP1: constant volatility; DGP2: early downward break in volatility; DGP4: double break in volatility; DGP6: linear upward trend in volatility. The Bartlett kernel is employed. The dashed vertical line is the 95% critical value from the $\mathcal{B}(k,b)$ distribution.

In other words, small-b methods asymptotically lead to pivotality under time-varying volatility as it does under weak stationarity. Recall, however, that the finite-sample quality of the asymptotic approximation in the small-b framework is meager, so practitioners essentially have to choose between the devil and the deep blue sea when not knowing the variance properties of the DGP.

Section 5 quantifies the size distortions resulting from ignoring time-varying volatility when using fixed-b asymptotics. The section will also recall that, despite asymptotic robustness, the small-b approach will often exhibit fairly strong size distortions in small samples under both homo- and heteroskedasticity. Hence, corrections for fixed-b inference should be valuable in empirical work.

As a final comment, note that the behavior under local alternatives is affected as well, via the "denominator" of the limiting distribution depending on h.

Corollary 1. *With* $\mathrm{E}(y_t) = {}^{c}/\sqrt{T}$, *we have under the assumptions of Proposition 1 that*

$$\mathcal{T} \xrightarrow{d} \frac{B_h(1) + c}{\sqrt{Q_{h,k,b}}}.$$

Proof: *See the Appendix.*

Remark 1. A partial solution to the problem studied in this paper is provided by Ibragimov and Müller (2010). They construct valid t-statistic based inference when it is possible to partition the data into q groups, such that estimators based on each group are approximately independent, unbiased and Gaussian. Their approach is somewhat restrictive, however, in that it relies on a result of Bakirov and Székely (2005) that guarantees size control only if the nominal level is at most 8.3%.

4 Robust inference under time-varying volatility

The critical issue about the failing asymptotics is that the partial sums of y_t do not converge weakly to Brownian motion, but to a different Gaussian process. In the following, we discuss three different ways to accommodate this.

4.1 The wild bootstrap

Cavaliere and Taylor (2008a) propose the wild bootstrap as a way to deal with time-varying volatility in a unit root testing context. We exploit the wild bootstrap to estimate the actual null distribution of $\mathcal{T}$ under time-varying volatility.[7] The basic algorithm is as follows.

1. Generate T iid standardized random variables r_t^*.

2. Generate the wild bootstrap sample as $y_t^* = r_t^* (y_t - \bar{y})$.

[7] Cavaliere and Taylor (2009) show the wild bootstrap to cope with stochastic volatility as well. The multivariate case has been dealt with in a series of papers starting with Cavaliere et al. (2010).

3. Compute the bootstrap test statistic $\mathcal{T}^*$ based on the resampled series y_t^*.

4. Repeat steps 1-3 to obtain a set of M resampled statistics $\mathcal{T}_m^*$, $m = 1, \ldots, M$.

5. Use the $(1-\alpha)$-quantile of $\{\mathcal{T}_m^*\}_{m=1,\ldots,M}$, say $q_{1-\alpha}^*$, as critical value for the test.[8]

As choice for r_t^* in Step 1 one usually picks the standard normal or the two-point Mammen distribution.

The following proposition shows that the wild bootstrap procedure gives size control in the limit. Given that the replicated series are white noise, whereas the sample is not, size control may be somewhat surprising. The key idea is that the wild bootstrap procedure is intended to obtain correct critical values of an asymptotic distribution that is invariant to short-memory serial correlation, but not to time-varying volatility. Therefore, the bootstrapped residuals need to replicate the variance profile (which they do, see the proof for details), but not necessarily the serial correlation structure. The latter may, however, additionally be incorporated, if desired, cf. Remark 3.

Proposition 2. *Under H_0 and Assumptions 1-3, it holds as $T \to \infty$ and $B/T = b \in (0, 1]$ that*

$$\Pr\left(\mathcal{T} > q_{1-\alpha}^*\right) \to \alpha.$$

Proof: *See the Appendix.*

Remark 2. Alternatively to step 5, one could of course use bootstrap p-values for a test decision; it can be seen from the proof of the proposition that the wild bootstrap p-values converge weakly to a uniform distribution $U[0, 1]$. Moreover, since the bootstrap procedure generates critical values which are invariant to the true mean μ of y_t, the behavior of $\mathcal{T}$ under (local) alternatives remains as implied by Corollary 1.

Remark 3. To reduce the influence of the short-run dynamics in finite samples, one may also use the residuals of a parametric model fit, say ARMA, in Step 2; when doing so here, we resort for simplicity to an AR(1) fit such that $\hat{u}_t = y_t - \hat{a}_0 - \hat{a}_1 y_{t-1}$ are used instead of $y_t - \bar{y}$ when generating the bootstrap sample. It is not necessary to recolor the bootstrap shocks to obtain the correct limiting distribution, although it would (of course) be possible to do so. See the proof of Proposition 2 for further details on this issue.

Remark 4. In fact, employing an autoregressive sieve approximation and recoloring would likely lead to second-order improvements. Moreover, in related work, Gonçalves and Vogelsang (2011) prove that suitable bootstrap procedures may even lead to higher-order accuracy in HAC testing under a location model like the present one. Our focus, however, is not on such refinements but on solving the first-order problem resulting from time-varying volatility. We conjecture that employing block, dependent or autoregressive wild bootstraps may allow to extend our analysis in this direction, too.

[8] This obviously applies, as in this paper, when performing a right-tailed test. The modifications to left-tailed and two-sided tests are, however, entirely standard.

4.2 Time transformations

The wild bootstrap can be computationally demanding with a large number of replications to ensure precision. The second possible correction therefore elaborates on the approach provided by Cavaliere and Taylor (2008b) which modifies the data in such a way that the series are in a sense transformed back to homoskedasticity. Hence, it will be valid to apply fixed-b methods applied to the transformed series. The time transformation approach of Cavaliere and Taylor (2008b) needs to be adapted to our setup, though, since they deal with I(1) processes under the null when testing for a unit root, whereas we deal with I(0) processes. The procedure we suggest is as follows.

1. Subtract the mean of y_t under the null and build the cumulated sums,

$$x_t = \sum_{j=1}^{t} (y_j - \mu_0) .$$

2. Estimate the variance profile of x_t, $\hat{\eta}(s) = \frac{\sum_{t=1}^{[sT]}(y_j-\bar{y})^2}{\sum_{t=1}^{T}(y_j-\bar{y})^2}$, and build its inverse $g(s)$.
3. Time transform x_t via

$$\tilde{x}_t = x_{[Tg(t/T)]}.$$

4. Base the actual test on the differenced series, $\tilde{y}_t = \Delta\tilde{x}_t$, i.e. compute

$$\tilde{\mathcal{T}} = \sqrt{T}\frac{\bar{\tilde{y}}}{\tilde{\omega}}, \tag{3}$$

where $\tilde{\omega}^2$ is an estimator of the long-run variance of $\tilde{y}_t$ using a bandwidth $B = [bT]$.

The following proposition shows that fixed-b asymptotics are recovered.

Proposition 3. *Under H_0 and Assumptions 1-3, it holds as $T \to \infty$ and $B/T = b \in (0, 1]$ that*

$$\tilde{\mathcal{T}} \xrightarrow{d} \mathcal{B}(k, b) .$$

Proof: *See the Appendix.*

Remark 5. In practice one often computes HAC estimators using some form of prewhitening; see Andrews and Monahan (1992). Although serial correlation does not enter the asymptotic distribution, it may still impact the empirical size in small samples. The proposition also holds when the long-run variance estimator is computed on the basis of ARMA residuals and then adjusted for serial correlation; see the proof for details.

Considering local alternatives, it turns out that the time transformation does have an asymptotic effect on the resulting distribution, in contrast to the wild bootstrap (Corollary 1). The precise effect is given in the following

Corollary 2. *With* $\mathrm{E}\left(y_t\right) = {}^{c}\!/\!\sqrt{T}$*, we have under the assumptions of Proposition 3 that*

$$\tilde{\mathcal{T}} \xrightarrow{d} \frac{W(1) + {}^{c}\!/\!\bar{\omega}}{\sqrt{\tilde{Q}_{h,k,b,c}}}$$

where

$$\begin{aligned}\tilde{Q}_{h,k,b,c} \quad = \quad & -\int_0^1 \int_0^1 \frac{1}{b^2} k''\left(\frac{r-s}{b}\right) \left\{ \left(W(r) - rW(1) + {}^{c}\!/\!\bar{\omega}\left(\tilde{g}(r) - r\right)\right) \times \right. \\ & \left. \left(W(s) - sW(1) + {}^{c}\!/\!\bar{\omega}\left(\tilde{g}(s) - s\right)\right) \right\} \mathrm{d}r\mathrm{d}s\end{aligned}$$

for kernels with smooth derivatives, and

$$\begin{aligned}\tilde{Q}_{h,k,b,c} \quad = \quad & \frac{2}{b}\int_0^1 \left(W(r) - rW(1) + {}^{c}\!/\!\bar{\omega}\left(\tilde{g}(r) - r\right)\right)^2 \mathrm{d}r \\ & -\frac{2}{b}\int_0^{1-b} \left\{ \left(W(r+b) - (r+b)W(1) + {}^{c}\!/\!\bar{\omega}\left(\tilde{g}(r+b) - (r+b)\right)\right) \times \right. \\ & \left. \left(W(r) - rW(1) + {}^{c}\!/\!\bar{\omega}\left(\tilde{g}(r) - r\right)\right) \right\} \mathrm{d}r\end{aligned}$$

for the Bartlett kernel, with $\tilde{g}$ *being the inverse of the variance profile* $\eta(s) = \bar{\omega}^{-2}\int_0^s h^2(s)\mathrm{d}s$.

Proof: *See the Appendix.*

The limiting distribution in Corollary 2 exhibits counterintuitive behavior under large deviations from the null. Let e.g. $c \to \pm\infty$ and note that, unless $\tilde{g}(s) = s$ (which is essentially only the case for constant variances h_t^2), we have that

$$G := \mathrm{plim}_{c\to\pm\infty}|\tilde{\mathcal{T}}| = -\int_0^1 \int_0^1 \frac{1}{b^2} k''\left(\frac{r-s}{b}\right)\left(\tilde{g}(r) - r\right)\left(\tilde{g}(s) - s\right)\mathrm{d}r\mathrm{d}s > 0$$

for smooth kernels, and correspondingly

$$G = \frac{2}{b}\int_0^1 \left(\tilde{g}(r) - r\right)^2 \mathrm{d}r - \frac{2}{b}\int_0^{1-b} \left(\tilde{g}(r+b) - (r+b)\right)\left(\tilde{g}(r) - r\right)\mathrm{d}r > 0$$

for the Bartlett kernel. Consequently, c^2 dominates in the denominator and it holds as $T \to \infty$ followed by $c \to \pm\infty$ that

$$\tilde{\mathcal{T}} \xrightarrow{p} \mathrm{sgn}(c)G^{-1/2}. \tag{4}$$

As a consequence, the test based on $\tilde{\mathcal{T}}$ will either always reject (if the constant $\mathrm{sgn}(c)G^{-1/2}$ belongs to the critical region of the fixed-b test) or never (if $\mathrm{sgn}(c)G^{-1/2}$ does not belong to the rejection region). Simulations presented in Section 5 reveal that the test will indeed have no power for several realistic heteroskedastic scenarios.

Some variations of the time transformation approach can be built analogously to the classical weighted least squares method and would consist here of standardizing the observations y_t by some (nonparametric) estimate of their standard deviation, $\hat{h}_t$. See e.g. Xu and Phillips (2008) for an application in time series autoregressions (and note also their additional requirements on the estimator $\hat{h}_t$). The analogy does not go too far, though, given that in the WLS transformed model we would not have a regression with a constant anymore, but rather

$$\frac{y_t}{h_t} = \frac{\mu}{h_t} + v_t.$$

While the error term is strictly stationary, time-varying volatility is re-introduced through the back door, given that the LS estimator for μ in the transformed model is given as

$$\hat{\mu} = \mu + \frac{\sum_{t=1}^{T} \frac{v_t}{h_t}}{\sum_{t=1}^{T} \frac{1}{h_t^2}}$$

and the term relevant for the limiting distribution of the test statistic is $T^{-1/2} \sum_{t=1}^{T} \frac{v_t}{h_t}$.[9] The fact that the true standard deviation was used for the WLS transformation does not affect the argument. One could alternatively standardize the series y_t under the null hypothesis, i.e. work with

$$\tilde{y}_t^w = \frac{y_t - \mu_0}{h_t}$$

and test the equivalent null hypothesis that $\mathrm{E}\left(\tilde{y}_t^w\right) = 0$. Denote the corresponding test statistic $\tilde{\mathcal{T}}^w$. The difference to testing the mean of y_t (either ignoring time-varying volatility or via WLS) is that $\tilde{y}_t^w$ is now strictly stationary under the null hypothesis and fixed-b inference would be applicable. The disadvantages of this approach appear under (local) alternatives, just like those of the time transformation, as illustrated in the following

Corollary 3. *Under the assumptions of Proposition 3 we have that*

$$\tilde{\mathcal{T}}^w \xrightarrow{d} \frac{W(1) + c\bar{h}(1)}{\sqrt{\tilde{Q}^w_{h,k,b,c}}}$$

where $\bar{h}(s) = \int_0^s \frac{1}{h(r)} \mathrm{d}r$ *and*

$$\begin{aligned} \tilde{Q}^w_{h,k,b,c} &= -\int_0^1 \int_0^1 \frac{1}{b^2} k'' \left(\frac{r-s}{b}\right) \left\{ \left(W(r) - rW(1) + c\left(\bar{h}(r) - r\right)\right) \times \right. \\ &\qquad \left. \left(W(s) - sW(1) + c\left(\bar{h}(s) - s\right)\right) \right\} \mathrm{d}r\mathrm{d}s \end{aligned}$$

for kernels with smooth derivatives, and

$$\begin{aligned} \tilde{Q}^w_{h,k,b,c} &= \frac{2}{b} \int_0^1 \left(W(r) - rW(1) + c\left(\bar{h}(r) - r\right)\right)^2 \mathrm{d}r \\ &\quad - \frac{2}{b} \int_0^{1-b} \left\{ \left(W(r+b) - (r+b)W(1) + c\left(\bar{h}(r+b) - (r+b)\right)\right) \times \right. \\ &\qquad \left. \left(W(r) - rW(1) + c\left(\bar{h}(r) - r\right)\right) \right\} \mathrm{d}r \end{aligned}$$

[9] This also illustrates how time-varying volatility may be induced in regression models by nonstationary regressors or instruments.

for the Bartlett kernel.

Proof: *See the Appendix.*

Note that $\bar{h}(s) = s$ only when h is constant, such that, in general, one would obtain a limiting distribution under local alternatives having the exact same properties as $c \to \pm\infty$ as the one in Corollary 2. We therefore do not further pursue rescaling approaches here.

4.3 Pre-testing

The third correction for time-varying volatility we study aims to mimick what practitioners often do: only correct for a problem detected in the data. Essentially, we consider to first test for heteroskedasticity, and then work with either fixed-b or small-b statistics and asymptotics, according to the outcome of the test. The intuition is that if a test for time-varying volatility does not reject, then the departures from constant variances may not be strong enough to seriously distort the fixed-b asymptotics of $\mathcal{T}$. If, on the other hand, the test rejects, then the small bandwidth-choice procedures may be preferable.[10] The success of such a testing strategy obviously depends on the properties of the pre-test.

To this end, we resort to the test proposed by Deng and Perron (2008). Xu (2013) demonstrates that the test has good size and power properties. It is based on the series $z_t = (y_t - \bar{y})^2$. The test statistic is given by

$$Q = \sup_{1 \le t \le T} \frac{1}{\sqrt{T}} \frac{|D_t|}{\hat{\omega}_z}$$

where $D_t = \sum_{j=1}^{t} z_j - \frac{t}{T} \sum_{j=1}^{T} z_j$ and $\hat{\omega}_z$ is a HAC estimator of the long-run variance of z_t. The test rejects for large values of Q.

Like for the wild bootstrap, the third correction just concerns obtaining suitable critical values, such that Corollary 1 applies under local alternatives.

5 Simulation evidence

5.1 Setup

This section studies the finite-sample behavior for the various statistics discussed above in different settings.

We consider one-sided tests of $H_0 : \mu = 0$ against $H_1 : \mu > 0$. The DGP is given by

$$y_t = \mu + \upsilon_t \quad (5)$$

$$(1 - \phi L)\upsilon_t = h_t \varepsilon_t \quad (6)$$

[10] A more complex alternative would be to use either of the two robust versions discussed above when time-varying volatility is detected. We do not pursue this line of research here.

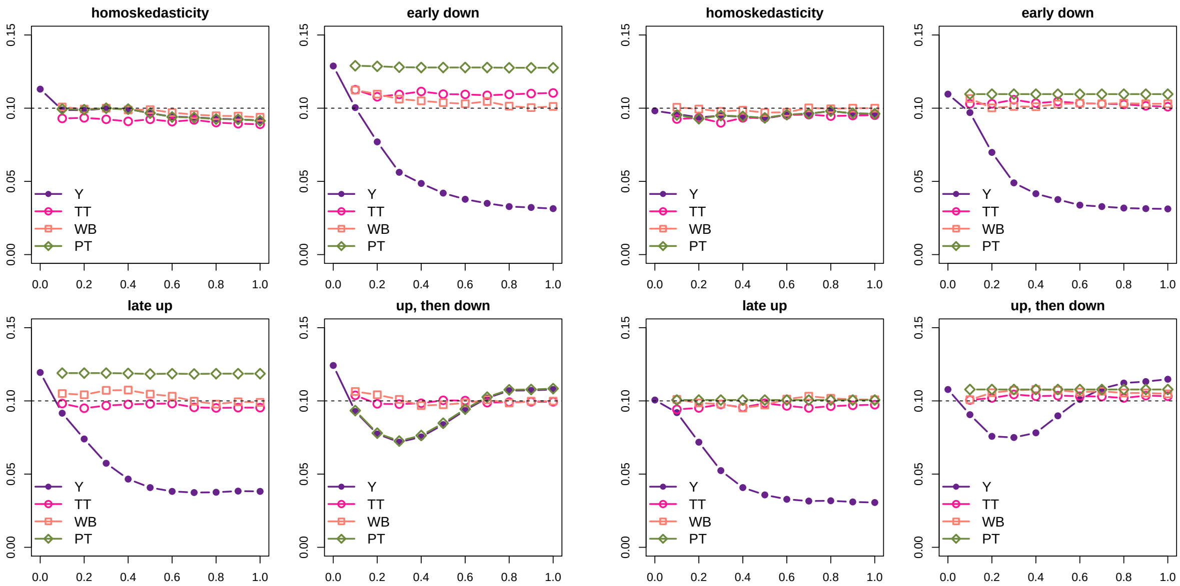

Figure 2: Size, $\phi = 0.1$. Left block: $T = 100$, right block: $T = 500$. Y denotes the standard fixed-b approach (NW for $b = 0$), TT the time transformation statistic (3), WB the wild bootstrap approach (cf. Prop. 2) and PT the pretest (cf. Section 4.3). Rejection frequencies are given on the y-axis, while b-values are given on the x-axis.

with $\varepsilon_t \sim i.i.d.N(0, 1)$ and $\phi = \{0.1, 0.5, 0.85\}$.[11] The following deterministic volatility DGPs for h_t are studied.

1. Constant volatility ($h_t = 1$);
2. Downward break at $t = [0.2T]$ from $\sigma_0 = 5$ to $\sigma_1 = 1$
3. Upward break at $t = [0.8T]$ from $\sigma_0 = 1$ to $\sigma_1 = 5$
4. Double break, upward (as in DGP3) at 0.4 and downward (back to initial level) at 0.6
5. Double break, downward at 0.1 (as in DGP2) and upward (back to initial level) at 0.9
6. Downward trend: $h_t = \sigma_0 + (\sigma_1 - \sigma_0)(t/T)$, $\sigma_0 = 5$, $\sigma_1 = 1$.

The DGPs, and in particular its breakpoints and sizes, are borrowed from Cavaliere and Taylor (2008b). For power results, we take $\mu_T = c(\bar{\omega}^2/T)^{1/2}$ with $\bar{\omega}^2$ being the average variance depending on the particular DGP1-6. Under homoskedasticity (DGP1), we have $\bar{\omega}^2 = \sigma^2$, while $\bar{\omega}^2 = T^{-1}\sum_{t=1}^{T}(\sigma_0^2 + 1(t > [\tau T])\sigma_1^2)$ under DGPs 2 and 3 for instance. The parameter $c = \{2, 4, \ldots, 16\}$ is a localizing constant.

[11] We also studied the ARMA(1,1) case with $(\phi, \theta) = (0.85, -0.45)$. Obviously, the implied AR(∞) structure generally does not improve the performance of the tests. However, it is found not to have an impact on the relative performance of the procedures studied here. In order to focus on the issue of volatility, we therefore work with a relatively simple autocorrelation structure in what follows.

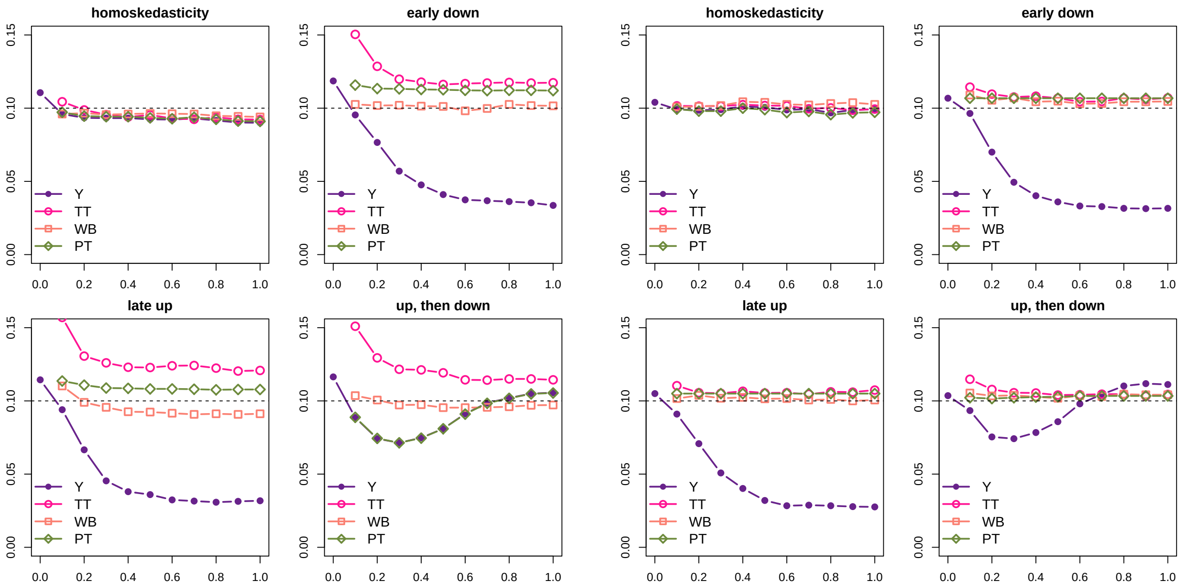

Figure 3: Size, $\phi = 0.5$. Left: $T = 100$, right: $T = 500$. See notes to Figure 2.

Critical values for the fixed-b approach are taken from Kiefer and Vogelsang (2005), Table 1. The nominal significance level is 10%. We use the quadratic spectral [QS] and the Bartlett kernel and values of b ranging from 0.1 to 1 in increments of 0.1. The number of wild bootstrap replications equals $M = 399$, while the number of Monte Carlo replications is 5,000. The sample sizes are $T = \{100, 250, 500\}$. We implement the AR(1) finite-sample correction along the lines of Remarks 3 and 5. For the Newey and West (1987, NW) approach, a data driven bandwidth choice is implemented, see Andrews and Monahan (1992).

5.2 Size

First, we present size results. We focus on the Bartlett kernel for better readability in the size results, as results for the QS kernel were quite similar.[12]

Under homoskedasticity (DGP1), the top-left entries of Figures 2, 3 and 4 reveal that, as is well-known, NW (corresponding to the entry $b = 0$) with automatic bandwidth selection (see Andrews, 1991) faces substantial size distortions for T as large as $T = 100$. Again in line with the literature, these distortions are more pronounced when autocorrelation is stronger. Confirming the results of Kiefer and Vogelsang (2005), the remainder of the "homoskedasticity" panels of Figures 2, 3 and 4 show that fixed-b asymptotics provide a very good approximation to the finite-sample distribution of the t-ratio given in equation (1) for all T, all but eliminating the size distortions of NW. Similarly, the corrections based on time transformations and the wild bootstrap provide accurate tests under homoskedasticity.

[12]Moreover, we waive to report qualitatively similar results for DGPs 5 and 6 here, as well as intermediate results for $T = 250$. **For the referees: Additional simulation results are reported in Appendix B.**

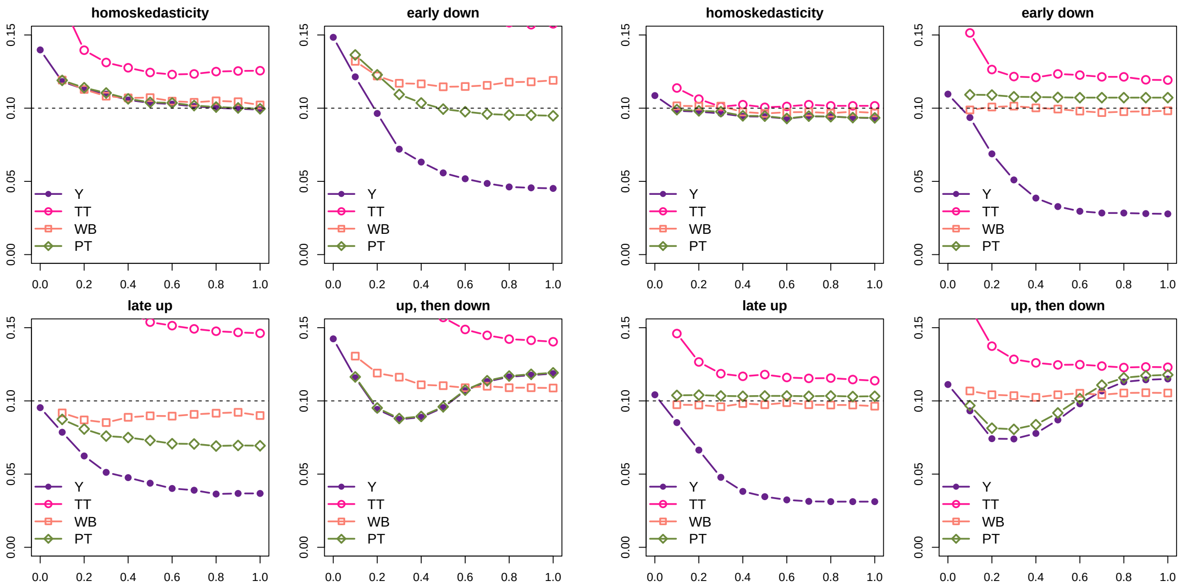

Figure 4: Size, $\phi = 0.85$. Left: $T = 100$, right: $T = 500$. See notes to Figure 2.

DGPs 2 ("early down"), 3 ("late up") and 4 ("up, then down") confirm the analytical prediction from Section 3 that, in general, fixed-b asymptotics are not pivotal under time-varying volatility. In particular, the tests seem to be conservative and increasingly so in b, with the exception of DGP 4 for large b. That fixed-b asymptotics work relatively well for small b is not unexpected, as they then operate very similarly to the standard NW approach which would be valid asymptotically (see above and Cavaliere, 2004). The wild bootstrap continues to be very effective in removing the size distortions, with some mild finite-sample exceptions for large ϕ, small T and small b.

Unsurprisingly, the size of the pretest is intermediate between that of NW and the fixed-b approaches. When the pretest does not reject frequently, e.g., for small T and/or homoskedasticity and the double break scenario ("up, then down"), the size of the pretest tracks that of fixed-b for all choices of b. Looking at the three panels for the different T for the double break scenario confirms that higher power of the pretest results in a behavior more like that of the asymptotically robust NW test.

The time transformation only performs convincingly for large T (as predicted by Proposition 3) and small or moderate autocorrelation. The top left panel of Figure 4 reveals poor empirical size properties of the time transformation for $T = 100$. The reason for this is that the time transformation may, in short series, produce stretches of identical observations that do not properly mimic the dynamics of the underlying series.

The "homoskedasticity" panels of the top-right and bottom entries of Figures 2, 3 and 4 highlight the finite-sample character of the size distortions of NW, which are largely removed for $T = 500$. Similarly, the upward size distortions of the wild bootstrap have all but disappeared for $T = 500$. The top-right and bottom entries of Figures 2, 3 and 4 moreover confirm that the distortions for fixed-b are *not* of a finite-sample nature under time-varying volatility.

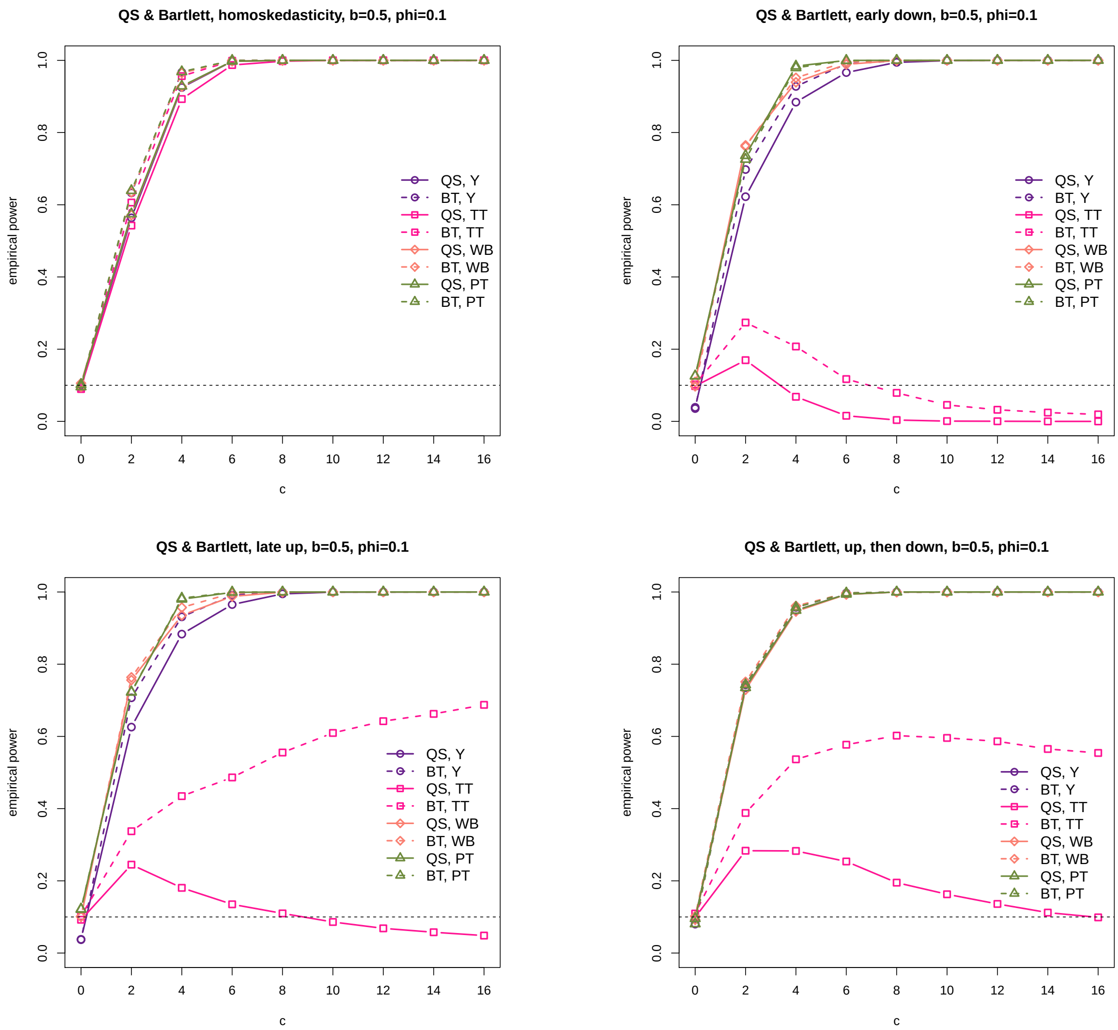

Figure 5: Power vs. c, $\phi = 0.1$, $b = 0.5$, $T = 100$. QS denotes the Quadratic Spectral kernel and BT stands for the Bartlett kernel. Y denotes the standard fixed-b approach (NW for $b = 0$), TT the time transformation statistic (3), WB the wild bootstrap approach (cf. Prop. 2) and PT the pretest (cf. Section 4.3). Rejection frequencies are given on the y-axis, while the localizing coefficient c is given on the x-axis.

5.3 Power

Figures 5 and 6 report power results for $T = 100$. As expected, power generally increases in c. The notable exception here is the time transformation under time-varying volatility, where the probability limit under local alternatives appears to yield zero power as $c \to \infty$ at least for the simulation designs considered here. This is not unexpected, see the discussion following Corollary 2, in particular Eq. (4).

Overall, the wild bootstrap achieves highest power. Essentially, it performs size adjustment for fixed-b, with a power curve that starts near $\alpha = 0.1$ at $c = 0$: it does not suffer from the conservativeness of fixed-b visible at $c = 0$ for “early down” and “late up.” For $T = 100$, the

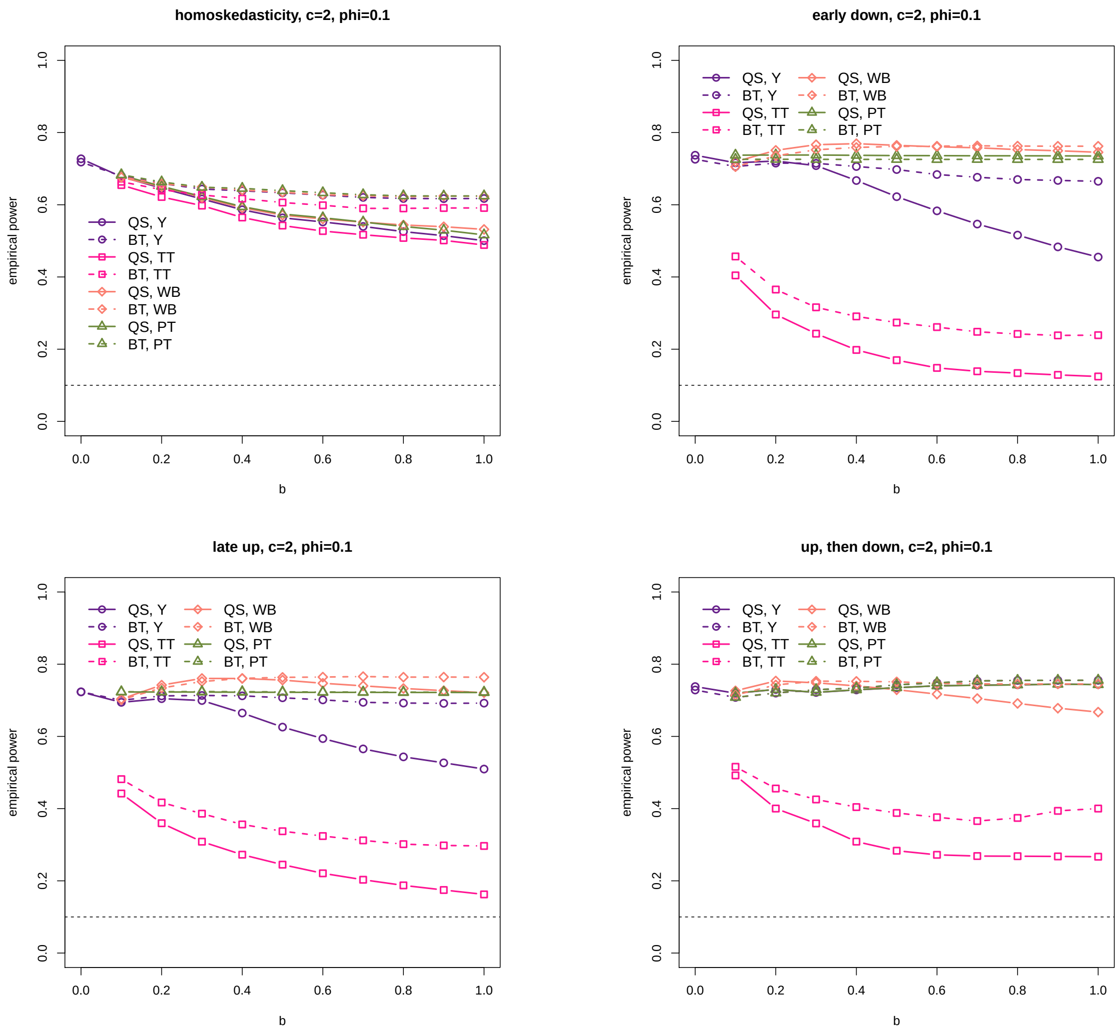

Figure 6: Power vs. b, $\phi = 0.1$, $c = 2$, $T = 100$. QS denotes the Quadratic Spectral kernel and BT stands for the Bartlett kernel. Y denotes the standard fixed-b approach (NW for $b = 0$), TT the time transformation statistic (3), WB the wild bootstrap approach (cf. Prop. 2) and PT the pretest (cf. Section 4.3). Rejection frequencies are given on the y-axis, while the fixed-b values are given on the x-axis.

Deng and Perron (2008) pretest does not have very high power yet, such that it sometimes sides with fixed-b, implying power intermediate between the wild bootstrap and the more conservative fixed-b testing approach.

That the particular type of variance break yields slight variation in power for fixed-b may be explained by its size distortions. (The normalization employed here implies that the power of the robust versions is not affected by the type of time-varying volatility.) Figure 6 reveals that there is no clear pattern of power as a function of b for the different DGPs. Under homoskedasticity, as already noticed by Kiefer and Vogelsang (2005) for fixed-b, the power of all procedures falls in b, albeit not by much. Under the heteroskedastic DGPs, the power of the wild bootstrap and that of the pretest-based procedure is roughly constant in b, with possibly a little peak at $b = 0.4$ in the case of the wild bootstrap. The lack of power of the time transformation is more pronounced for larger b.

The comparison of the dashed and solid lines suggests that the Bartlett kernel may be somewhat more powerful than the QS kernel. However, the difference is only substantial in the case of the time transformation, which is dominated by other approaches anyhow.

The present simulations therefore suggest to use the wild bootstrap with an intermediate value of b (say, $b = 0.4$) in practice. The pretest would be a useful alternative for larger sample sizes.

6 Real-time evaluation of professional output and inflation forecasts

We evaluate forecasts of real output and inflation in the US using the procedures proposed in the previous section. Utilizing the rich Survey of Professional Forecasters (SPF) database from the Federal Reserve Bank of Philadelphia, we shall demonstrate that allowing for changes in volatility has an important implication on the empirical findings. In particular, we make use of the "Forecast Error Statistics for the Survey of Professional Forecasters", see Stark (2010). For exchange rate data, Choi and Kiefer (2010) employ a fixed-b approach to inference on equal predictive accuracy, on the grounds that serial dependence of the loss differentials leads to finite-sample distortions of the usual t-tests. Furthermore, Li and Patton (2015) use fixed-b inference in the context of predictive accuracy testing based on high frequency data, see also Patton (2015) for a discussion of Diebold (2015).

We consider one-quarter and one-year ahead forecasts. Regarding the vintage structure of the real-time data, we study the first and the final data release.[13] The sample ranges from 1969:Q4 to 2015:Q1, yielding $T = 182$ quarterly observations on real output (RGDP) and gross domestic product deflator inflation (PGDP) time series and their respective forecasts. The Great Moderation is thus part of our sample. It is well-documented that the it led to enhanced macroeconomic stability which eased forecasting in general, but also made it more difficult to beat simple time series models, see for instance Stock and Watson (2007). Groen et al. (2013) find along the same lines that structural breaks in the variance play an important role for real-time inflation forecasting. As the sample most likely contains observations with variance breaks, the procedures developed above are important and suitable in this application.

The forecasts in the SPF are either averaged surveys obtained from individual professionals or simple time series forecasts. Among these, we consider models already available in the data set, namely direct and indirect autoregressive models and a no-change random walk model. Thereby, we follow Stark (2010) in our real-time evaluation, who also provides details on the construction of the time series model forecasts. See Groen et al. (2009) for a similar analysis for the real-time evaluation of Bank of England forecasts for output and inflation.

[13]Results for two- and three-quarter ahead forecasts are qualitatively similar and available upon request.

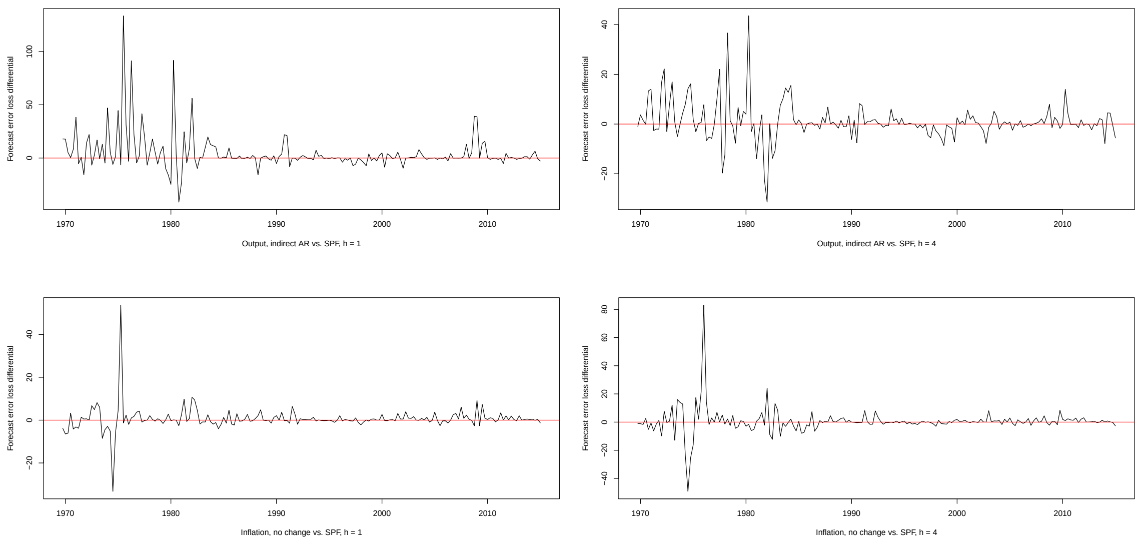

Figure 7: Forecast error loss differentials $e^2_{TSM,t} - e^2_{SPF,t}$ between time series model forecasts and survey forecasts. The left (right) panel shows the series for $h = 1$ ($h = 4$). The upper graphs are for real output, while the lower ones are for inflation. The time series model is in each case the best competitor of the SPF, see Table 1.

Some series contain a few missing values (either only a single one in the mid-nineties (0.55% of the data points) or five ones in the early seventies (2.75% of the data points)). We decide to impute these few missing values via a bootstrap based expectation maximization [EM] algorithm, see Honaker et al. (2011). The algorithm makes use of the standard EM algorithm on multiple bootstrapped samples of the original data set (containing missing values) to obtain imputed values.[14]

Figure 7 shows some exemplary MSE loss differential series. These are computed as $e^2_{TSM,t} - e^2_{SPF,t}$, where $e^2_{TSM,t}$ is the squared forecast error of a time series model (TSM) and $e^2_{SPF,t}$ denotes the corresponding value for the survey forecast. We show the resulting series for SPF real output and inflation forecasts in comparison each to the best competing time series model forecasts. Different horizons and vintages are considered. It can be seen that the volatility of the loss differentials sharply declined during the early eighties for real GDP growth (upper panel) and inflation (lower panel) in conjunction with the Great Moderation. The conjecture of a nonconstant volatility in forecast error loss differentials is strongly confirmed by the Deng and Perron (2008) test. Three out of four rejections are at the nominal significance level of one percent, while the rejection for the case of four-quarters ahead inflation forecasts is significant at the five percent level. The degree of autocorrelation is mild to intermediate, ranging between 0.03 and 0.44 (average 0.17) for output and between 0.14 and 0.5 (average 0.32) for inflation. As before, we employ a first-order autoregressive finite-sample correction.

[14] We use 10,000 bootstrap replications for the EM algorithm and compare the resulting imputed values with the distribution of the original data set to check their plausibility. An appendix (available upon request) provides specific details about the missing and imputed values. **For the referees: Please see Appendix C.**

Table 1: Relative RMSE values $RMSE(\text{TSM})/RMSE(\text{SPF})$ of SPF output and inflation forecasts; 1969:Q4-2015:Q1; evaluated against the first and the final data release.

Variable	Forecast horizon	Vintage	No change	Direct AR	Indirect AR
Real GDP	$h = 1$	First release	1.52	1.36	1.35
		Final release	1.39	1.24	1.24
	$h = 4$	First release	1.41	1.08	1.04
		Final release	1.41	1.08	1.04
Inflation	$h = 1$	First release	1.22	1.24	1.23
		Final release	1.17	1.20	1.20
	$h = 4$	First release	1.17	1.20	1.20
		Final release	1.07	1.22	1.23

Table 1 shows some descriptive statistics for the root mean squared error (RMSE) losses of time series model-based forecasts in relation to the ones for survey forecasts. A ratio greater than unity thus indicates the superiority of SPF forecasts. The results give some interesting insights regarding the performance of time series models in comparison to professional survey forecasts. Based on these descriptive statistics, it seems that survey forecasts outperform simple time series models. In particular for the short horizon ($h = 1$), the figures suggest that SPF forecasts for output and inflation dominate the ones by time series models. In most cases (except for one-year ahead real GDP growth and inflation forecasts) the SPF performs better when being evaluated against the first release rather than the final release. For real GDP growth predictions, the indirect AR is the strongest competitor of SPF forecasts, while the no-change forecast performs best among the time series model forecasts for the inflation series.

Next, we conduct formal tests of equal predictive ability using Diebold and Mariano (1995) statistics. The auxiliary regression is given by

$$e^2_{TSM,t} - e^2_{SPF,t} = \mu + u_t \ .$$

We report results for one-sided tests, i.e. $H_0 : \mu = 0$ versus $H_0 : \mu > 0$ at different conventional significance levels. We thus test the null hypothesis of equal predictive ability against the alternative that time series model-based forecasts have a larger mean squared error. A rejection thus yields evidence in favor of superiority of SPF forecasts relative to simple model-based forecasts. Among the different versions of the Diebold-Mariano statistics are a classic Newey-West type statistic and also the fixed-b versions as suggested by Choi and Kiefer (2010). Moreover, we add the time transformation and wild bootstrap fixed-b versions to our comparison to deal with the time-varying nature of volatility. We use the Bartlett kernel for the fixed-b approach (with $b = \{0.1, \ldots, 1\}$) as it is slightly more powerful than the Quadratic Spectral kernel in our simulations. The number of wild bootstrap replications is 10,000.

Results for the first data release are reported in Figures 8 (real GDP growth) and 9 (inflation).[15] Each figure contains the results for $h = 1$ (left panel consisting of six graphs) and $h = 4$ (right panel). In a given panel, the upper two graphs show the results for the comparison of no-change forecasts to the SPF; the middle ones for the direct autoregressive forecast against the SPF and the two graphs at the bottom show the results for the indirect autoregressive model versus the SPF. The left part of a given panel contains Diebold-Mariano statistics in comparison to standard fixed-b critical values, while the right part presents the outcomes for the wild bootstrap. Both sets of critical values are presented for three conventional nominal significance levels (1%, 5% and 10%). Differences in the asymptotic and bootstrapped critical values give an indirect indication for the presence of time-varying volatility. The null hypothesis of equal predictive ability is rejected if the Diebold-Mariano statistic [DM stat] exceeds the critical value.

First, we find that asymptotic and bootstrap critical values differ remarkably in all considered situations. This finding is in line with the rejections of homoskedasticity by the Deng and Perron (2008) test. Boostrapped critical values are smaller than their asymptotic counterparts, shifting statistical evidence towards the alternative hypothesis. The use of unmodified fixed-b DM statistics most often leads to only weak evidence against the null hypothesis of equal predictive ability and sometimes even to nonrejections. To the contrary, the application of the fixed-b DM statistic with time transformed data [labeled as DM (TT) in the figures] provides strong evidence of superiority of professional forecasts in the case of real GDP forecasts. Similarly, the wild bootstrap versions of fixed-b DM statistics yield strong evidence against the null. For short-term inflation forecasts, we observe that the DM (TT) statistic does not affect the conclusions much for the comparison to the direct AR and indirect AR models, while it does so for the longer horizon of four quarters. Moreover, the wild bootstrap DM statistics clearly provide evidence for superior predictive ability of survey forecasts over simple time series model forecasts for both horizons.

[15] Results for the final release are qualitatively similar and available upon request.

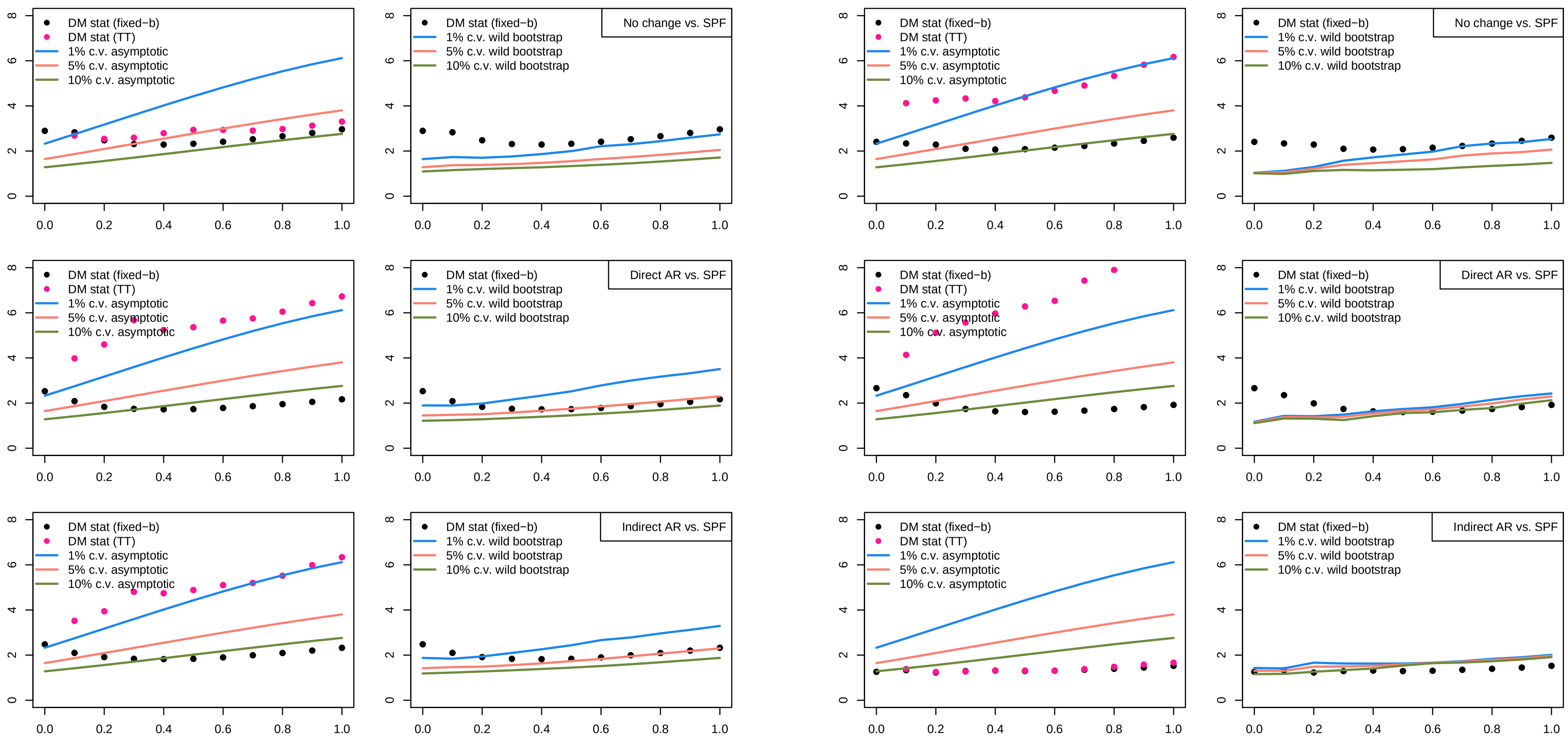

Figure 8: Real GDP growth, first data release, left panel: $h = 1$, right panel: $h = 4$.

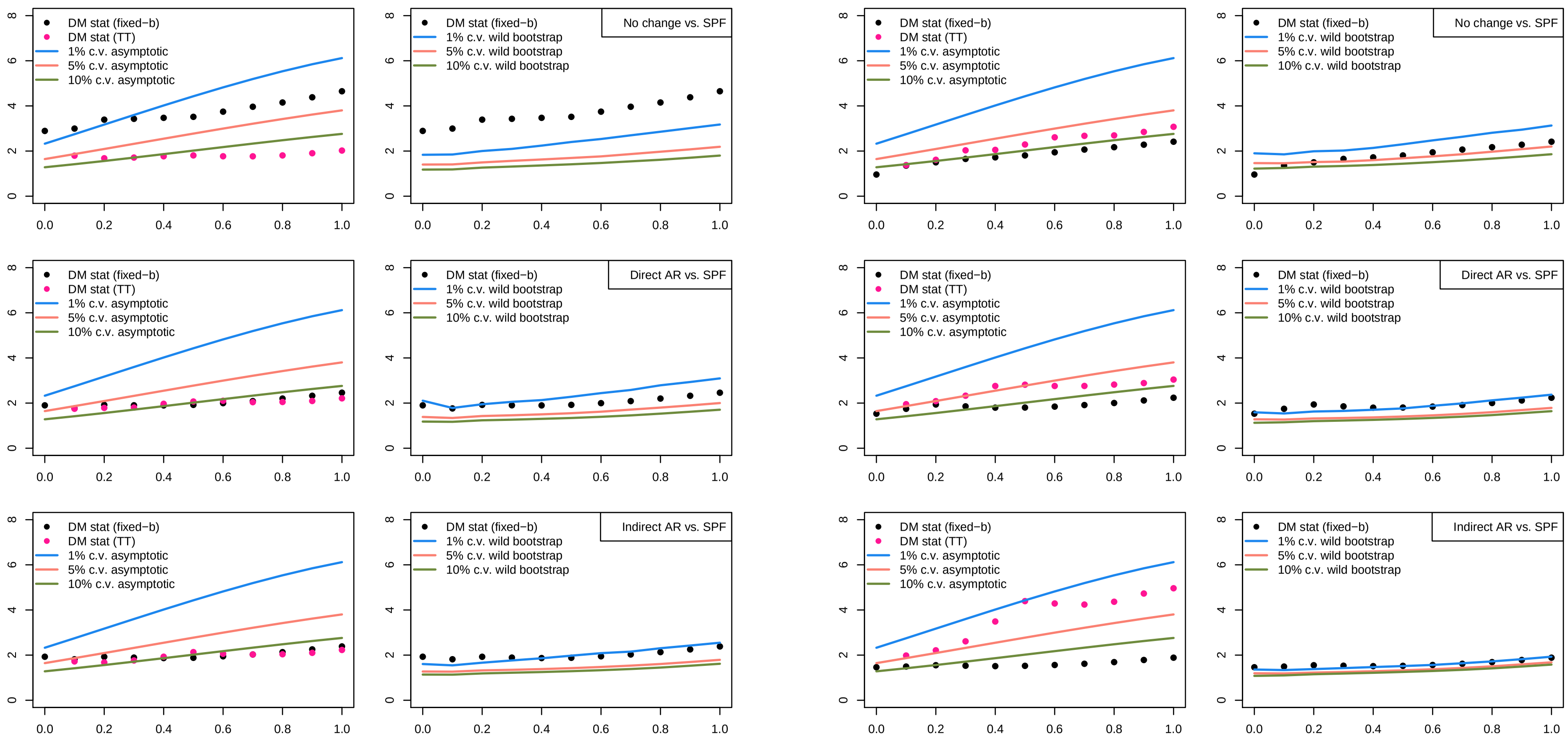

Figure 9: Inflation, first data release, left panel: $h = 1$, right panel: $h = 4$.

7 Concluding remarks

Fixed-b asymptotics are a tremendously useful device to enable more accurate inference when dealing with serially correlated data. Serial correlation is, however, not the only important data feature practitioners need to pay attention to when aiming to conduct reliable hypothesis tests: many important macroeconomic and financial time series are subject to time-varying volatility such as variance breaks. We show that the standard fixed-b approach no longer yields pivotal tests under time-varying volatility and quantify the resulting distortions.

Based on wild bootstrap schemes (Cavaliere and Taylor, 2008a), on time transformations (Cavaliere and Taylor, 2008b) or on a pre-test procedure, we provide corrections that restore size control of fixed-b methods even under time-varying volatility. Simulations illustrate the useful size and power properties of the corrections, in particular of the wild bootstrap approach. The behavior of the pre-test procedure hinges on whether the pre-test has enough power of detecting changes in the variance and is not reliable for very small sample sizes, where the power of the pretest is not close to unity, while the time transformation approach may lead to a nonmonotonic local power function and even zero local power.

The evidence provided here suggests quite plausibly that the multivariate wild bootstrap would provide a robust version of fixed-b tests in a composite-hypothesis situation as well. While the time transformation has inherent difficulties in the multivariate setup which make such extensions difficult, the pre-test may again provide a complement to the wild bootstrap approach, provided that the sample size is large enough.

An application to Survey of Professional Forecasters data suggests that ignoring important time-variation in the variance can seriously affect conclusions regarding predictive ability. In our application, we consider a comparison of the forecasting abilities of professional survey forecasts to time series models. Without accounting for the apparent changes in volatility, Diebold and Mariano (1995) statistics provide only weak evidence (at best) against the null hypothesis of equal predictive ability. The modified versions (based on time transformed data or the wild bootstrap) however indicate that survey forecasts significantly outperform model-based forecasts.

References

Amado, C. and T. Teräsvirta (2013). Modelling volatility by variance decomposition. *Journal of Econometrics 175*(2), 142–153.

Amado, C. and T. Teräsvirta (2014). Modelling changes in the unconditional variance of long stock return series. *Journal of Empirical Finance 25*(1), 15–35.

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

Andrews, D. W. K. and J. C. Monahan (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica 60*(4), 953–966.

Bakirov, N. K. and G. J. Székely (2005). Student's t-test for Gaussian scale mixtures. *Zapiski Nauchnyh Seminarov POMI 328*, 5–19.

Berk, K. N. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics 2*(3), 489–502.

Burridge, P. and A. M. R. Taylor (2001). On regression-based tests for seasonal unit roots in the presence of periodic heteroscedasticity. *Journal of Econometrics 104*(1), 91–117.

Cavaliere, G. (2004). Unit root tests under time-varying variances. *Econometric Reviews 23*(3), 259–292.

Cavaliere, G., A. Rahbek, and A. M. R. Taylor (2010). Testing for co-integration in vector autoregressions with non-stationary volatility. *Journal of Econometrics 158*(1), 7–24.

Cavaliere, G. and A. M. R. Taylor (2008a). Bootstrap unit root tests for time series with nonstationary volatility. *Econometric Theory 24*(1), 43–71.

Cavaliere, G. and A. M. R. Taylor (2008b). Time-transformed unit root tests for models with non-stationary volatility. *Journal of Time Series Analysis 29*(2), 300–330.

Cavaliere, G. and A. M. R. Taylor (2009). Heteroskedastic time series with a unit root. *Econometric Theory 25*(5), 1228–1276.

Choi, H. S. and N. M. Kiefer (2010). Improving robust model selection tests for dynamic models. *The Econometrics Journal 13*(2), 177–204.

Clark, T. E. (2009). Is the Great Moderation over? An empirical analysis. *Economic Review 94*(4), 5–42.

Clark, T. E. (2011). Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics 29*(3), 327–341.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press.

Demetrescu, M. and C. Hanck (2012). Unit root testing in heteroskedastic panels using the Cauchy estimator. *Journal of Business & Economic Statistics 30*(2), 256–264.

Demetrescu, M. and P. Sibbertsen (2014). Inference on the long-memory properties of time series with non-stationary volatility. *unpublished*.

Deng, A. and P. Perron (2008). The limit distribution of the CUSUM of squares test under general mixing conditions. *Econometric Theory 24*(3), 809–822.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. *Journal of Business and Economic Statistics 33*(1), 1–8.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–263.

Gonçalves, S. and T. J. Vogelsang (2011). Block bootstrap HAC robust tests: The sophistication of the naive bootstrap. *Econometric Theory 27*(4), 745–791.

Groen, J. J. J., G. Kapetanios, and S. Price (2009). A real time evaluation of Bank of England forecasts of inflation and growth. *International Journal of Forecasting 25*(1), 74–80.

Groen, J. J. J., R. Paap, and F. Ravazzolo (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics 31*(1), 29–44.

Guidolin, M. and A. Timmermann (2006). An econometric model of nonlinear dynamics in the joint distribution of stock and bond returns. *Journal of Applied Econometrics 21*(1), 1–22.

Hansen, L. P. (1982). Large sample properties of generalized methods of moments estimators. *Econometrica 50*(2), 1029–1054.

Honaker, J., G. King, and M. Blackwell (2011). Amelia II: A program for missing data. *Journal of Statistical Software 45*(7), 1–47.

Ibragimov, R. and U. K. Müller (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics 28*(4), 453–468.

Justiniano, A. and G. Primiceri (2008). The time-varying volatility of macroeconomic fluctuations. *American Economic Review 98*(3), 604–641.

Kiefer, N. M. and T. J. Vogelsang (2002a). Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica 70*(5), 2093–2095.

Kiefer, N. M. and T. J. Vogelsang (2002b). Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size. *Econometric Theory 18*(6), 1350–1366.

Kiefer, N. M. and T. J. Vogelsang (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory 21*(6), 1130–1164.

Kiefer, N. M., T. J. Vogelsang, and H. Bunzel (2000). Simple robust testing of regression hypotheses. *Econometrica 68*(3), 695–714.

Li, J. and A. J. Patton (2015). Asymptotic inference about predictive accuracy using high frequency data. *unpublished*.

Müller, U. K. (2014). HAC corrections for strongly autocorrelated time series. *Journal of Business & Economic Statistics 32*(3), 311–322.

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

Patton, A. J. (2015). Comment on "Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests" by Francis X. Diebold. *Journal of Business and Economic Statistics 33*(1), 22–24.

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

Phillips, P. C. B., Y. Sun, and S. Jin (2006). Spectral density estimation and robust hypothesis testing using steep origin kernels without truncation. *International Economic Review 47*(3), 837–894.

Phillips, P. C. B. and K. L. Xu (2006). Inference in autoregression under heteroskedasticity. *Journal of Time Series Analysis 27*(2), 289–308.

Priestley, M. B. (1988). *Non-linear and non-stationary time series analysis.* Academic Press London.

Sensier, M. and D. van Dijk (2004). Testing for volatility changes in U.S. macroeconomic time series. *The Review of Economics and Statistics 86*(3), 833–839.

Stark, T. (2010). Realistic evaluation of real-time forecasts in the survey of professional forecasters. *Federal Reserve Bank of Philadelphia, Research Department* (Special Report), 726–740.

Stock, J. H. and M. W. Watson (2002). Has the business cycle changed and why? *NBER Macroeconomics Annual 17*(1), 159–218.

Stock, J. H. and M. W. Watson (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking 39*(S1), 3–33.

Sun, Y. (2014a). Fixed-smoothing asymptotics in a two-step generalized method of moments framework. *Econometrica 82*(6), 2327–2370.

Sun, Y. (2014b). Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference. *Journal of Econometrics 178*(3), 659–677.

Sun, Y., P. C. B. Phillips, and S. Jin (2008). Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. *Econometrica 76*(1), 175–194.

Teräsvirta, T. and Z. Zhao (2011). Stylized facts of return series, robust estimates and three popular models of volatility. *Applied Financial Economics 21*(1-2), 67–94.

Vogelsang, T. J. and M. Wagner (2013). A fixed-b perspective on the Phillips-Perron unit root tests. *Econometric Theory 29*(3), 609–628.

Westerlund, J. (2014). Heteroscedasticity robust panel unit root tests. *Journal of Business & Economic Statistics 32*(1), 112–135.

Xu, K.-L. (2008). Bootstrapping autoregression under non-stationary volatility. *Econometrics Journal 11*(1), 1–26.

Xu, K.-L. (2013). Powerful tests for structural changes in volatility. *Journal of Econometrics 173*(1), 126–142.

Xu, K.-L. and P. C. B. Phillips (2008). Adaptive estimation of autoregressive models with time-varying variances. *Journal of Econometrics 142*(1), 265–280.

Yang, J. and T. J. Vogelsang (2011). Fixed-b analysis of LM-type tests for a shift in mean. *The Econometrics Journal 14*(3), 438–456.

Appendix

A Proofs

Proof of Proposition 1

Note that the arguments in the proof of Theorem 2 in Kiefer and Vogelsang (2005) can be used without further modification to conclude that

$$\mathcal{T} = \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(y_t-\mu_0)}{\sqrt{-\frac{1}{T^2}\sum_{i=1}^{T-1}\sum_{j=1}^{T-1}\frac{T^2}{B^2}k''\left(\frac{i-j}{B}\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{i}(y_t-\bar{y})\frac{1}{\sqrt{T}}\sum_{t=1}^{j}(y_t-\bar{y})}} + o_p(1).$$

for kernels with smooth derivatives or

$$\mathcal{T} = \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(y_t-\mu_0)}{\sqrt{\frac{2}{bT}\sum_{i=1}^{T}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{i}(y_t-\bar{y})\right)^2-\frac{2}{bT}\sum_{i=1}^{[(1-b)T]}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{i}(y_t-\bar{y})\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{i+[bT]}(y_t-\bar{y})\right)}} + o_p(1).$$

for the Bartlett kernel. The weak convergence

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}(y_t-\mu) \Rightarrow B_h(s)$$

and the continuous mapping theorem [CMT] then establish the desired limiting null distribution.

Proof of Corollary 1

Under a local alternative we have the weak convergence

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}(y_t-\mu_0) \Rightarrow B_h(s)+cs$$

and the result follows with the same arguments as in the proof of Proposition 1.

Proof of Proposition 2

We begin with the case of Gaussian bootstrap variables r_t^* and no prewhitening. Let $S_T^*(s)$ denote the normalized partial sums of the bootstrapped centered sample,

$$S_T^*(s) = \frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}(y_t-\bar{y})\,r_t^*$$

To guarantee size control in the limit, it suffices to show that the bootstrap partial sums converge weakly in probability to $\sqrt{\mathrm{Var}(\upsilon_t)}B_h(s)$, since $\mathrm{Var}(\upsilon_t)$ would cancel out in the bootstrapped t-ratio. Note that, conditional on the sample y_t, $t = 1, \ldots, T$, $S_T^*(s)$ is a Gaussian process with independent increments. Its covariance kernel is given by

$$\mathrm{Cov}\left(S_T^*(s), S_T^*(r)\right) = \frac{1}{T}\sum_{t=1}^{[\min\{s,r\}T]} (y_t - \bar{y})^2 \,\mathrm{E}\left(\left(r_t^*\right)^2\right) = \frac{1}{T}\sum_{t=1}^{[\min\{s,r\}T]} (y_t - \bar{y})^2 .$$

Then, following the proof of Lemma A.5 in Cavaliere et al. (2010), it suffices to establish the weak convergence

$$\frac{1}{T}\sum_{j=1}^{[sT]} (y_j - \bar{y})^2 \Rightarrow \mathrm{Var}\left(\upsilon_t\right) \int_0^s h^2(r)\mathrm{d}r$$

i.e. that the wild bootstrap correctly replicates the variance profile of the sample y_t in the limit. Assumption 2 guarantees pointwise convergence of $\frac{1}{T}\sum_{j=1}^{[sT]} (y_j - \bar{y})^2$ via a Law of Large Numbers for strong mixing processes (see Davidson, 1994, Section 20.6), and the monotonicity of the quadratic variation function leads to uniformity of the convergence, as required for the result.

We then examine the case of prewhitening. Let a and c be the pseudo-true value of the AR(1) coefficient and of the intercept in the autoregression[16] $y_t = \hat{a}_0 + \hat{a}_1 y_{t-1} + \hat{u}_t$ and let $u_t = h_t\left(\upsilon_t - a_1\upsilon_{t-1}\right)$ such that u_t satisfies indeed the same assumptions as y_t up to the (for this step irrelevant) unity long-run variance requirement. Then,

$$S_T^*(s) = \frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]} u_t r_t^* + \frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]} \left(\hat{u}_t - u_t\right) r_t^*$$

and the result follows along the lines of the case without prewhitening if the second summand on the r.h.s. vanishes uniformly in $s \in [0,1]$. Given that r_t^* are serially independent and independent of y_t, this is implied by

$$\sup_{s\in[0,1]} \frac{1}{T}\sum_{t=1}^{[sT]} \left(\hat{u}_t - u_t\right)^2 \xrightarrow{p} 0$$

which is in turn implied by $\sup_{s\in[0,1]\setminus D} \left|\hat{u}_{[sT]} - u_{[sT]}\right| \xrightarrow{p} 0$ where D denotes the set of discontinuities of $h(\cdot)$, which are negligible given that there is a finite number of jump discontinuities. Now, at all continuity points of h,

$$\begin{aligned} \hat{u}_{[sT]} - u_{[sT]} &= \mu + h_{[sT]}\upsilon_{[sT]} - \hat{a}_0 - \hat{a}_1\left(\mu + h_{[sT]-1}\upsilon_{[sT]-1}\right) - h_{[sT]}\left(\upsilon_{[sT]} - a_1\upsilon_{[sT]-1}\right) \\ &= \mu\left(1 - \hat{a}_1\right) - \hat{a}_0 - \upsilon_{[sT]-1}\left(\left(\hat{a}_1 - a_1\right)h_{[sT]-1} - a_1\left(h_{[sT]} - h_{[sT]-1}\right)\right), \end{aligned}$$

where $\hat{a}_1$ and $\hat{a}_0$ are $\sqrt{T}$-consistent (Phillips and Xu, 2006), $\sup_{t\in\{1,\ldots,T\}} |\upsilon_t| = o_p\left(\sqrt{T}\right)$ thanks to the uniform $L_{2+\delta}$ boundedness of υ_t, and $h_{[sT]} - h_{[sT]-1} = O\left(T^{-1}\right)$ uniformly in s thanks to

[16] See Phillips and Xu (2006) for the precise details.

the piecewise Lipschitz condition on h. Hence $\sup_{s\in[0,1]\setminus D}\left|\hat{u}_{[sT]}-u_{[sT]}\right|\xrightarrow{p}0$ as required for the result.

Finally, should r_t^* follow the Rademacher or Mammen distribution, say, $S_T^*(s)$ is not Gaussian, but weak convergence to a Gaussian process (conditional on the sample) holds and the proof follows along the same lines.

Proof of Proposition 3

Cavaliere and Taylor (2008b, proof of Theorem 1) show that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}\tilde{y}_t \Rightarrow \bar{\omega}W(s).$$

The result then follows like in the proof of Proposition 1 when no prewhitening is used.

To discuss prewhitening for computing the HAC estimator, we focus for simplicity on the AR(1) case; the extension to ARMA is not difficult but tedious.

Use the Phillips-Solo device to write, with $y_0=0$ for convenience,

$$\hat{u}_t = y_t-\hat{a}_1y_{t-1}-\hat{a}_0 = y_t\left(1-\hat{a}_1\right)+\hat{a}_1\Delta y_{t-1}-\hat{a}_0,$$

such that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}\hat{u}_t = \left(1-\hat{a}_1\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}y_t+\hat{a}_1y_{[sT]}-\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}\hat{a}_0.$$

Some OLS algebra indicates that $\hat{a}_0=\left(1-\hat{a}_1\right)\bar{y}+O_p\left(T^{-1}\right)$, and, thanks to the uniform $L_{2+\delta}$-boundedness of y_t, it holds that $\sup_{s\in[0,1]}\left|y_{[sT]}\right|=o_p\left(\sqrt{T}\right)$, hence

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}\hat{u}_t = \left(1-\hat{a}_1\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}\left(y_t-\bar{y}\right)+o_p\left(1\right)$$

with the $o_p\left(1\right)$ term uniform in s. Since the adjustment of the HAC estimator is done precisely by division with $\left(1-\hat{a}_1\right)$, this term cancels out and we have the same behavior of $Q_{b,k}$ as under no prewhitening.

Proof of Corollary 2

Begin by writing

$$x_t=\sum_{j=1}^{t}\left(y_j-\mu_0\right)=\sum_{j=1}^{t}h_jv_j+c\frac{t}{\sqrt{T}}$$

and note that the variance profile estimate is invariant to μ. Now, the transformation is

$$\tilde{x}_t = x_{[Tg(t/T)]}$$

with g the inverse of $\hat{\eta}$, implying that

$$\sum_{j=1}^{t} \tilde{y}_j = \tilde{x}_t = \sum_{j=1}^{[Tg(t/T)]} h_j \upsilon_j + c\sqrt{T} g\left(\frac{t}{T}\right).$$

Since $\hat{\eta}$ converges weakly to the variance profile η, its inverse converges weakly to the inverse $\tilde{g}$ of the variance profile, which is, like η, monotonic and continuous. Hence, using the CMT and Cavaliere and Taylor (2008b, proof of Theorem 1) again, we obtain the required weak convergence

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} \tilde{y}_t \Rightarrow \bar{\omega} W(s) + c\tilde{g}(s).$$

Proof of Corollary 3

Under $\mu = \mu_0 + {}^{c}/\sqrt{T}$, we have that

$$\tilde{y}_t^w = \upsilon_t + \frac{1}{h_t} \frac{c}{\sqrt{T}}.$$

This implies for the partial sums of $\tilde{y}_t^w$ that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} \tilde{y}_t^w = \frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} \upsilon_t + \frac{c}{T} \sum_{t=1}^{[sT]} \frac{1}{h_t}.$$

With $h(\cdot)$ being piecewise Lipschitz, bounded and bounded away from zero, ${}^{1}/h$ is itself piecewise Lipschitz, bounded and bounded away from zero, so we ultimately have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} \tilde{y}_t^w \Rightarrow W(s) + c\int_0^s \frac{1}{h(r)} \mathrm{d}r \equiv W(s) + c\bar{h}(s)$$

as required for the result.

B Additional simulation results—Not for publication

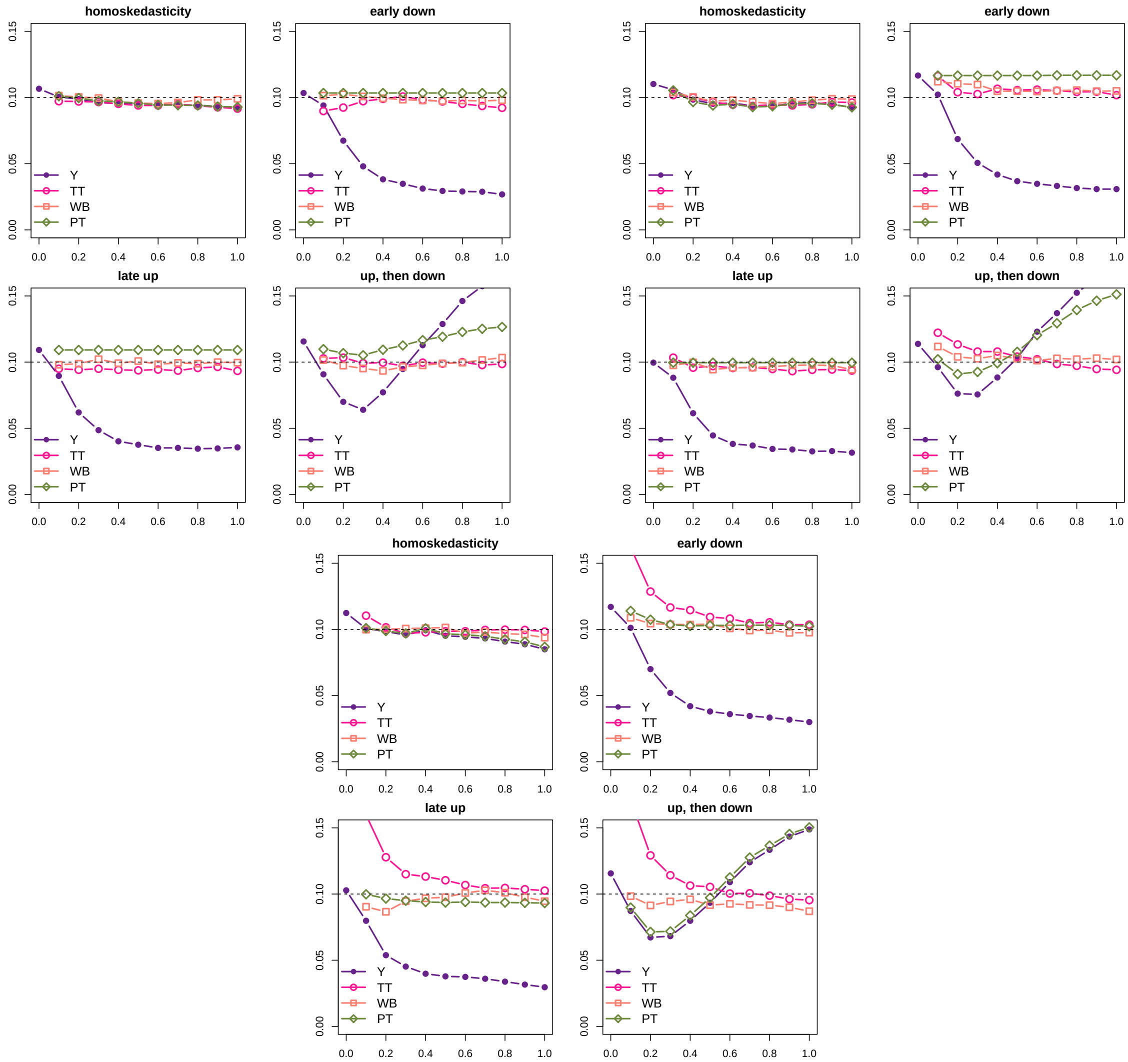

Figure 10: Size, $T = 250$. Top left block: $\phi = 0.1$, top right: $\phi = 0.5$, bottom: $\phi = 0.85$. Y denotes the standard fixed-b approach, TT the time transformation statistic (3), WB the wild bootstrap approach (cf. Prop. 2) and PT the pretest (cf. Section 4.3). Rejection frequencies are given on the y-axis, while b-values are given on the x-axis.

Figure 11 omits results for the time transformation in view of the substantial upward size distortions.

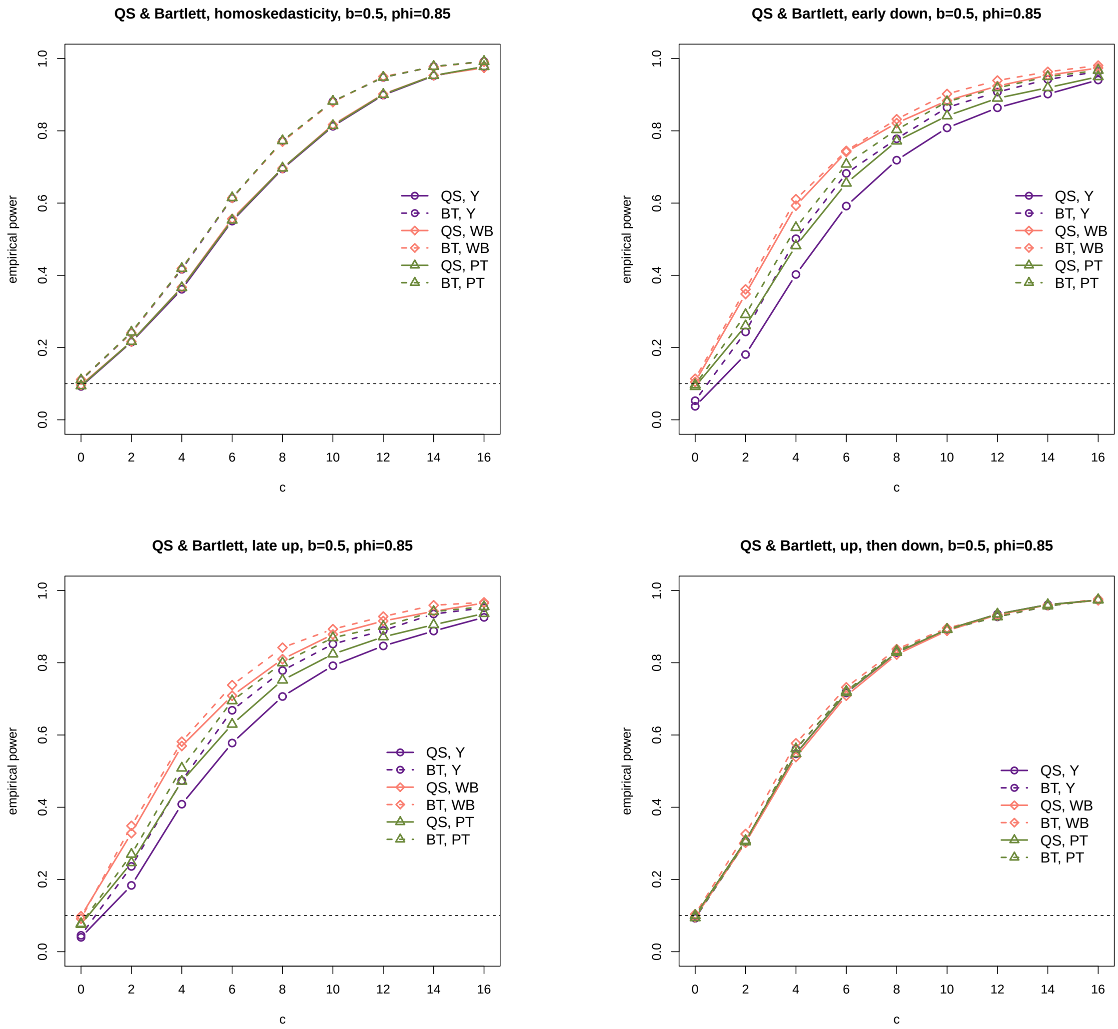

Figure 11: Power vs. c, $\phi = 0.85$, $b = 0.5$, $T = 100$. See notes to Figure 5.

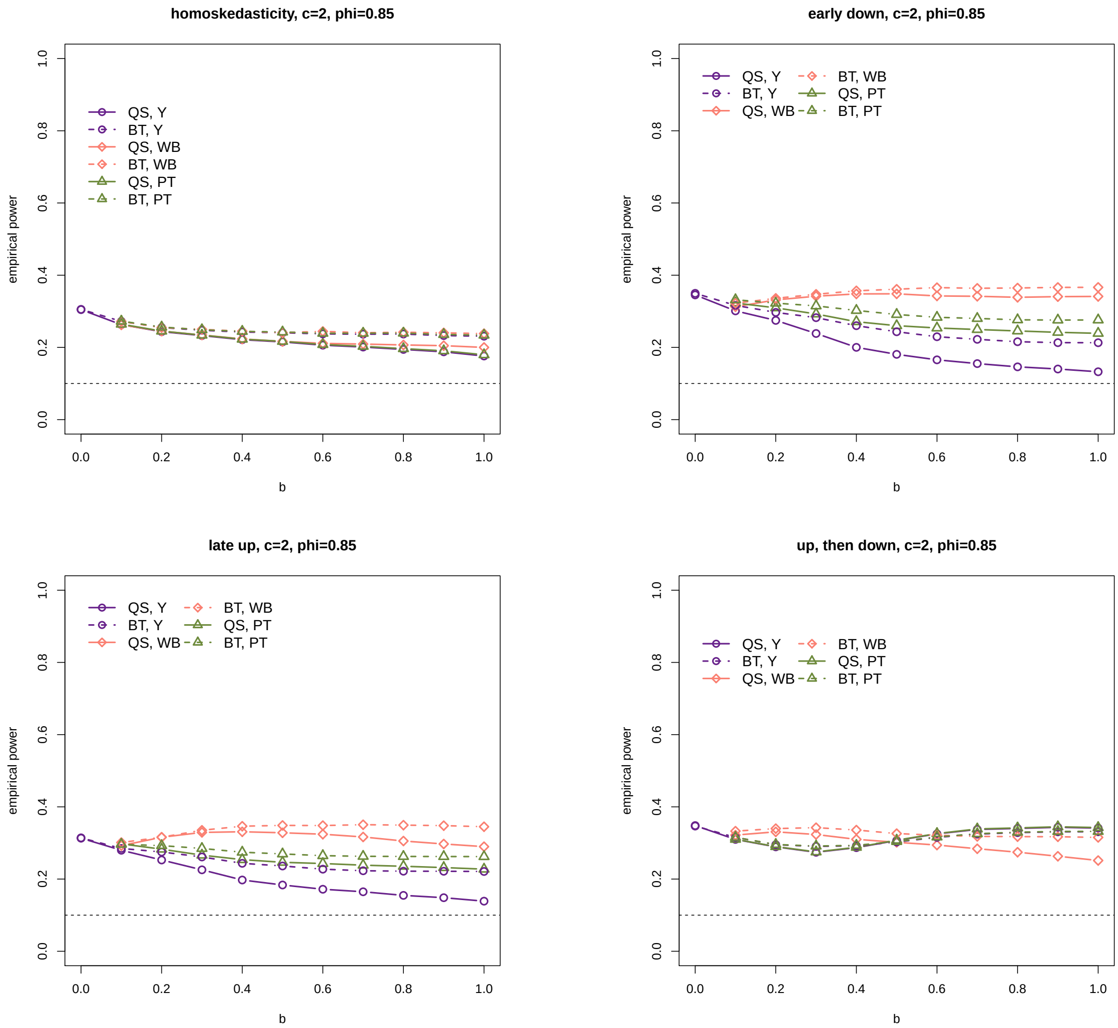

Figure 12: Power vs. b, $\phi = 0.85$, $c = 2$, $T = 100$. See notes to Figure 6.

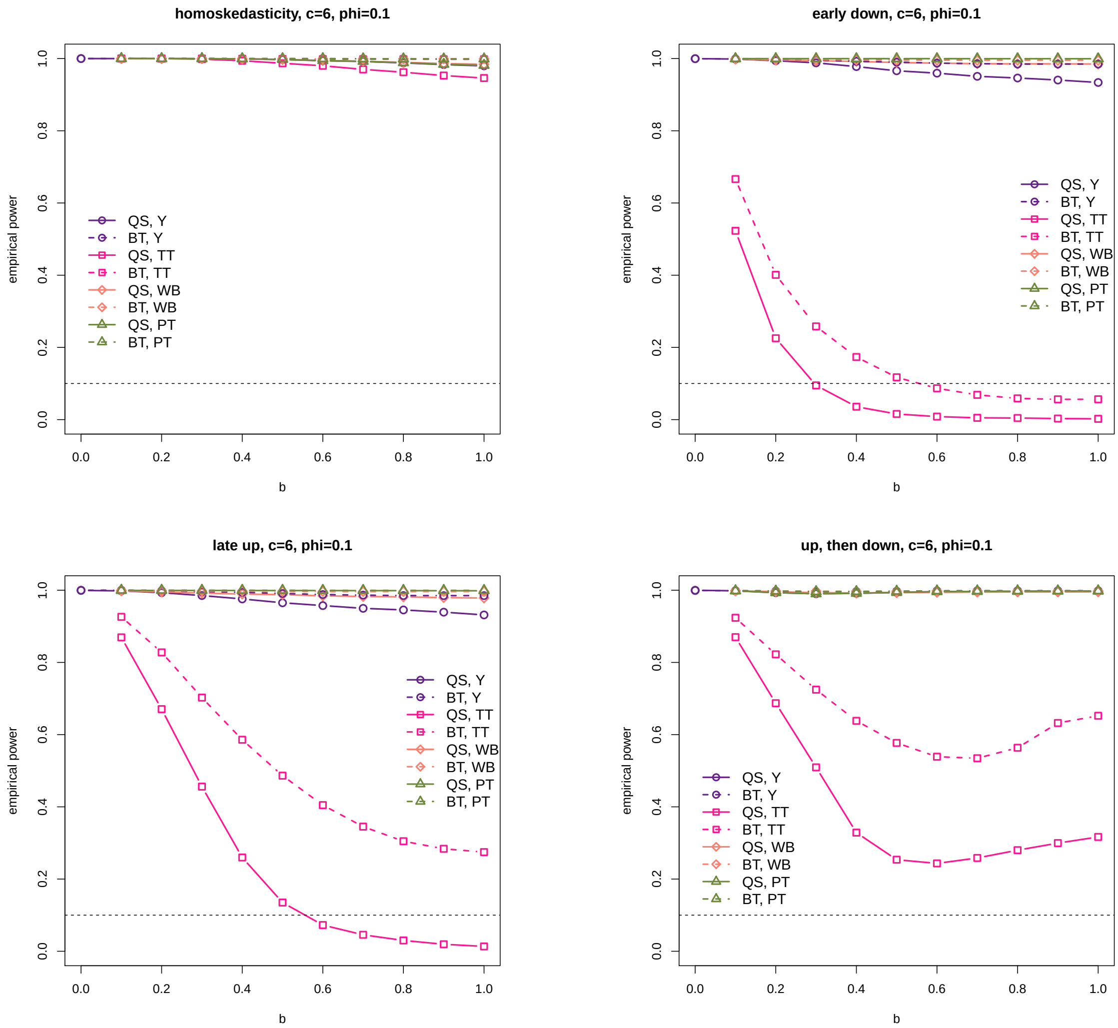

Figure 13: Power vs. b, $\phi = 0.1$, $c = 6$, $T = 100$. See notes to Figure 6.

C Imputation—Not for publication

This appendix contains details on the imputed values for the missing observations in the SPF data set from the ”Forecast Error Statistics for the Survey of Professional Forecasters” obtained from the Federal Reserve Bank of Philadelphia.

Data

In our empirical application, we evaluate forecasts of real output and inflation in the US. The data stems from the Survey of Professional Forecasters (SPF) database of the Federal Reserve Bank of Philadelphia. In particular, we obtain data from the "Forecast Error Statistics for the Survey of Professional Forecasters". This database offers time series based forecasts (direct and indirect autoregression and a no-change random walk forecast) in comparison to averages computed from the distribution of survey forecasts for a wide range of variables and a relatively long time period.

In connection to the previous literature, we focus on real output growth (RGDP) and inflation (PGDP, i.e. the price deflator of the GDP). For both variables, we use one-quarter and four-quarters ahead forecasts. Moreover, we take the first and last vintage of the real-time data for the forecast evaluation. Our sample ranges from 1969:Q4 to 2015:Q1, yielding $T = 182$ quarterly observations on real output and inflation time series and their respective forecasts. Some series contain a few missing values: either only a single one in the mid-nineties (0.55% of the data points) or five ones in the early seventies (2.75% of the data points). Details are reported in Tables 2–5.

These few missing values are imputed via a bootstrap based expectation maximization [EM] algorithm, see Honaker et al. (2011). The algorithm makes use of the standard EM algorithm on multiple bootstrapped samples of the original data set (containing missing values) to obtain imputed values. We use 10,000 bootstrap replications for the EM algorithm. The code is written in `R` (by using the `Amelia` package) and available upon request from the authors. Tables 2–5 contain the imputed values (underlined) in connection to neighboring values. The obtained bootstrap averages serve as imputed values which are reasonable. Figure 14 shows the time series of real output and inflation and the corresponding four-quarters ahead forecasts together with the imputed values (shown as red filled dots). The imputed values seem to plausible. Further details on the time series model forecasts are provided in Tables 4–5.

Table 2: Data entries for realized output and inflation series (first data release). #MV gives the number of missing values in total. Underlined values are imputed values obtained from the bootstrap-based EM algorithm. Neighboring values are reported for comparison.

Output		Inflation	
Date	First release	Date	First release
1995:03	4.20481	1995:03	0.58927
1995:04	$\underline{2.41452}$	1995:04	$\underline{2.26685}$
1996:01	2.80932	1996:01	2.60573
#MV	1	#MV	1

Table 3: Data entries for SPF forecasts of output and inflation series (four-quarters ahead). #MV gives the number of missing values in total. Underlined values are imputed values obtained from the bootstrap-based EM algorithm. Neighboring values are reported for comparison.

Output		Inflation	
Date	SPF ($h = 4$)	Date	SPF ($h = 4$)
1969:04	4.03701	1969:04	3.21260
1970:01	$\underline{3.55115}$	1970:01	$\underline{3.56122}$
1970:02	$\underline{3.90855}$	1970:02	$\underline{3.56355}$
1970:03	$\underline{4.05961}$	1970:03	$\underline{3.90631}$
1970:04	3.10037	1970:04	3.01866
1971:01	$\underline{4.54798}$	1971:01	$\underline{4.15267}$
1971:02	4.26233	1971:02	2.95183
1975:02	5.40498	1975:02	3.50332
1975:03	$\underline{5.33554}$	1975:03	$\underline{6.52413}$
1975:04	5.02638	1975:04	6.57499
#MV	5	#MV	5

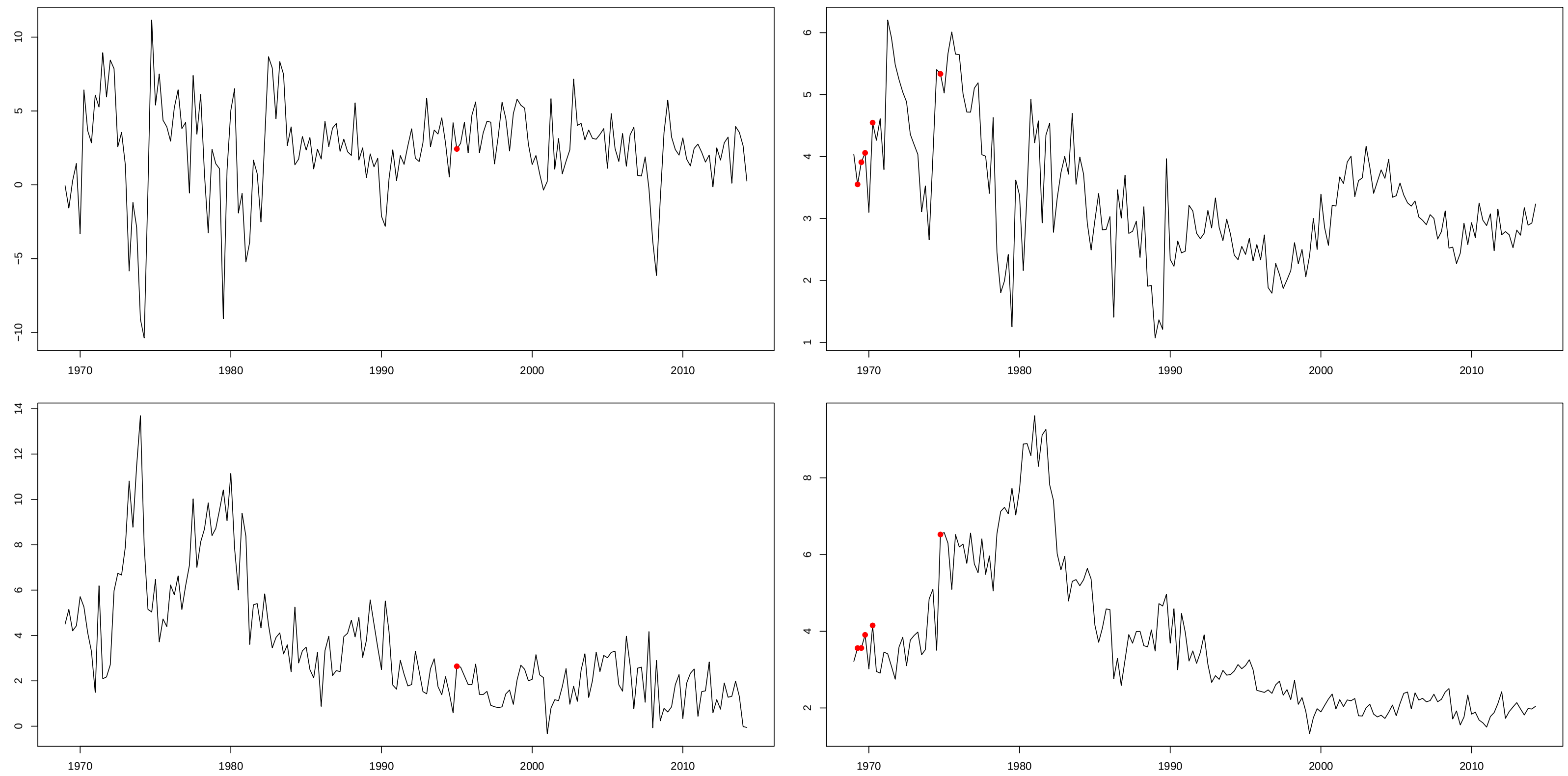

Figure 14: Top: real DGP series (left) and four-quarters ahead SPF forecast (right); bottom: inflation series (left) and four-quarters ahead SPF forecast (right). Imputed values are shown as filled red dots. Details can be obtained from Tables 2 and 3.

Table 4: Data entries for TSM forecasts of output and inflation series (one-quarter ahead). #MV gives the number of missing values in total. Underlined values are imputed values obtained from the bootstrap-based EM algorithm. Neighboring values are reported for comparison.

Output Date	IAR ($h=1$)	NCH ($h=1$)	DAR ($h=1$)	Inflation Date	IAR ($h=1$)	NCH ($h=1$)	DAR ($h=1$)
1996:01	3.03197	4.20481	2.58962	1996:01	1.39023	0.58927	1.54054
1996:02	<u>2.51616</u>	<u>1.97486</u>	<u>2.50171</u>	1996:02	<u>2.38443</u>	<u>2.14000</u>	<u>2.38342</u>
1996:03	2.49993	2.80932	2.45663	1996:03	2.61826	2.60573	2.62880
1996:04	3.01134	4.22119	3.44510	1996:04	2.47629	2.22008	2.45910
1997:01	2.44679	2.17048	2.45901	1997:01	2.47246	1.83734	2.47421
1997:02	3.47569	4.71721	3.40647	1997:02	2.46867	1.82727	2.47512
#MV	1	1	1	#MV	1	1	1

Table 5: Data entries for TSM forecasts of output and inflation series (four-quarters ahead). #MV gives the number of missing values in total. Underlined values are imputed values obtained from the bootstrap-based EM algorithm. Neighboring values are reported for comparison.

Output Date	IAR ($h = 4$)	NCH ($h = 4$)	DAR ($h = 4$)	Inflation Date	IAR ($h = 4$)	NCH ($h = 4$)	DAR ($h = 4$)
1996:01	2.50135	4.53191	2.30798	1996:01	2.04827	1.38913	2.09657
1996:02	2.48018	2.81927	2.07116	1996:02	2.11200	2.18291	2.63003
1996:03	2.72364	0.52754	2.80953	1996:03	1.93867	1.45128	2.16886
1996:04	2.78802	4.20481	2.51189	1996:04	1.95016	0.58927	1.67693
1997:01	<u>2.69266</u>	<u>2.73465</u>	<u>2.53956</u>	1997:01	<u>2.65489</u>	<u>2.35841</u>	<u>2.71337</u>
1997:02	2.45413	2.80932	2.45289	1997:02	2.80127	2.60573	2.78749
#MV	1	1	1	#MV	1	1	1

Research Papers
2016

CREATES
Center for Research in Econometric Analysis of Time Series

2015-44:	Jonas Nygaard Eriksen: Expected Business Conditions and Bond Risk Premia
2015-45:	Kim Christensen, Mark Podolskij, Nopporn Thamrongrat and Bezirgen Veliyev: Inference from high-frequency data: A subsampling approach
2015-46:	Asger Lunde, Anne Floor Brix and Wei Wei: A Generalized Schwartz Model for Energy Spot Prices - Estimation using a Particle MCMC Method
2015-47:	Annastiina Silvennoinen and Timo Teräsvirta: Testing constancy of unconditional variance in volatility models by misspecification and specification tests
2015-48:	Harri Pönkä: The Role of Credit in Predicting US Recessions
2015-49:	Palle Sørensen: Credit policies before and during the financial crisis
2015-50:	Shin Kanaya: Uniform Convergence Rates of Kernel-Based Nonparametric Estimators for Continuous Time Diffusion Processes: A Damping Function Approach
2015-51:	Tommaso Proietti: Exponential Smoothing, Long Memory and Volatility Prediction
2015-52:	Mark Podolskij, Christian Schmidt and Mathias Vetter: On U- and V-statistics for discontinuous Itô semimartingale
2015-53:	Mark Podolskij and Nopporn Thamrongrat: A weak limit theorem for numerical approximation of Brownian semi-stationary processes
2015-54:	Peter Christoffersen, Mathieu Fournier, Kris Jacobs and Mehdi Karoui: Option-Based Estimation of the Price of Co-Skewness and Co-Kurtosis Risk
2015-55	Kadir G. Babaglou, Peter Christoffersen, Steven L. Heston and Kris Jacobs: Option Valuation with Volatility Components, Fat Tails, and Nonlinear Pricing Kernels
2015-56:	Andreas Basse-O'Connor, Raphaël Lachièze-Rey and Mark Podolskij: Limit theorems for stationary increments Lévy driven moving averages
2015-57:	Andreas Basse-O'Connor and Mark Podolskij: On critical cases in limit theory for stationary increments Lévy driven moving averages
2015-58:	Yunus Emre Ergemen, Niels Haldrup and Carlos Vladimir Rodríguez-Caballero: Common long-range dependence in a panel of hourly Nord Pool electricity prices and loads
2015-59:	Niels Haldrup and J. Eduardo Vera-Valdés: Long Memory, Fractional Integration, and Cross-Sectional Aggregation
2015-60:	Mark Podolskij, Bezirgen Veliyev and Nakahiro Yoshida: Edgeworth expansion for the pre-averaging estimator
2016-01:	Matei Demetrescum, Christoph Hanck and Robinson Kruse: Fixed-b Inference in the Presence of Time-Varying Volatility