Kiefer, N. M., Vogelsang, T. J., and Bunzel, H. (2000), "Simple Robust Testing of Regression Hypotheses," *Econometrica*, 68, 695–714. [311,314]
King, M. L. (1980), "Robust Tests for Spherical Symmetry and Their Application to Least Squares Regression," *The Annals of Statistics*, 8, 1265–1271. [316]
——— (1987), "Towards a Theory of Point Optimal Testing," *Econometric Reviews*, 6, 169–218. [315]
Lehmann, E. L., and Romano, J. P. (2005), *Testing Statistical Hypotheses*, New York: Springer. [316]
Müller, U. K. (2004), "A Theory of Robust Long-Run Variance Estimation," Working paper, Princeton University. [311,314]
——— (2007), "A Theory of Robust Long-Run Variance Estimation," *Journal of Econometrics*, 141, 1331–1352. [311,314,318,321]
——— (2011), "Efficient Tests Under a Weak Convergence Assumption," *Econometrica*, 79, 395–435. [315]
Müller, U. K., and Watson, M. W. (2008), "Testing Models of Low-Frequency Variability," *Econometrica*, 76, 979–1016. [314]
——— (2013), "Measuring Uncertainty About Long-Run Forecasts," Working Paper, Princeton University. [315]
Newey, W. K., and West, K. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [311,312]
Phillips, P. (2005), "HAC Estimation by Automated Regression," *Econometric Theory*, 21, 116–142. [311,314]
Phillips, P., Sun, Y., and Jin, S. (2006), "Spectral Density Estimation and Robust Hypothesis Testing Using Steep Origin Kernels Without Truncation," *International Economic Review*, 47, 837–894. [311]
Phillips, P., Sun, Y., and Jin, S. (2007), "Long Run Variance Estimation and Robust Regression Testing Using Sharp Origin Kernels With No Truncation," *Journal of Statistical Planning and Inference*, 137, 985–1023. [311]
Phillips, P. C. B. (1987), "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547. [315]
Robinson, P. M. (2005), "Robust Covariance Matrix Estimation: HAC Estimates With Long Memory/Antipersistence Correction," *Econometric Theory*, 21, 171–180. [315]
Stock, J., and Watson, M. (2011), *Introduction to Econometrics* (3rd ed.), Boston: Addison Wesley. [312]
Sun, Y. (2013), "Heteroscedasticity and Autocorrelation Robust F Test Using Orthonormal Series Variance Estimator," *The Econometrics Journal*, 16, 1–26. [314]
Sun, Y., and Kaplan, D. M. (2012), "Fixed-Smoothing Asymptotics and Accurate F Approximation Using Vector Autoregressive Covariance Matrix Estimator," Working Paper, University of California, San Diego. [311]
Sun, Y., Phillips, P. C. B., and Jin, S. (2008), "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing," *Econometrica*, 76, 175–794. [311]
Whittle, P. (1957), "Curve and Periodogram Smoothing," *Journal of the Royal Statistical Society,* Series B, 19, 38–63. [313]
——— (1962), "Gaussian Estimation in Stationary Time Series," *Bulletin of the International Statistical Institute*, 39, 105–129. [313]

Comment

Nicholas M. Kiefer
Departments of Economics and Statistical Science, Cornell University, Ithaca, NY 14853 and CREATES, University of Aarhus, Aarhus, Denmark (*nicholas.kiefer@cornell.edu*)

1. INTRODUCTION

Müller looks at the problems of interval estimation and hypothesis testing in autocorrelated models from a frequency domain point of view. This leads to good insights as well as proposals for new methods. The new methods may be more robust than existing approaches, though this is more suggested than firmly established. This discussion begins with a speculative overview of the problem and the approach. The theme is that the issue involved is essentially the choice of a conditioning ancillary. Then I turn, perhaps more usefully, to some specific technical comments. Finally, I agree wholeheartedly with the general point that comes through clearly: the more we know or are willing to assume about the underlying process the better we can do.

2. ASYMPTOTICS AND CONDITIONING

The point of asymptotic theory is sometimes lost, especially when new approaches are being considered. The goal is to find a manageable approximation to the sampling distribution of a statistic. The approximation should be as accurate as possible. The assumptions needed to develop the asymptotics are not a model of any actual physical process.

The "trick" is to model the rate of information accumulation leading to the asymptotic approximation, so that the resulting limit distribution can be calculated and as accurately as possible mimics the sampling distribution of interest. There are many ways to do this. These are not "correct" or "incorrect," just different models. What works?

One way to frame the choice of assumptions is as specification of an ancillary statistic. An example will make this specific. Suppose we are estimating μ and the sufficient statistic is S. Suppose S can be partitioned into $(\hat{\mu}, a)$ with a ancillary. With data y sufficiency implies the factorization

$$p(y|\mu) = g(y)p(S|\mu)$$

and ancillarity implies

$$p(S|\mu) = p(\hat{\mu}|a, \mu)p(a).$$

The key is choosing a so its distribution does not depend on μ—or in the local case, does not depend "much." See Christensen and Kiefer (1994). S may have the dimension of the dataset.

It is widely agreed—mostly from examples, not theorems—that inference can (and perhaps should) be based on the conditional distribution. See Barndorff-Nielsen (1984), Berger et al. (1988), and the review by Reid (1995). In the normal mean model, we could set $a = (s^2, a')$ and condition on a, obtaining normal inference, or condition on a' alone obtaining the t. With autocorrelation, $a = (\hat{\rho}, s^2, a') = (\psi, a')$ and conditioning on a

Journal of Business & Economic Statistics
July 2014, Vol. 32, No. 3
DOI: 10.1080/07350015.2014.926816

is "too much" conditioning in that confidence intervals are too small under repeated sampling. There are two sources of error here:

- Specification error: dependence is more general than AR(1).
- Too much conditioning.

3. GLS, HAC, AND KVB

Following Müller, let $\omega^2 = \sum_{j=-\infty}^{j=\infty} \gamma(j)$ with $\gamma(j) = E(y_t - \mu)(y_{t-j} - \mu)$. The generalized least squares approach essentially parameterizes ω by parameterizing the $\gamma(j)$, for example, with ρ in a first-order autocorrelation model. GLS then conditions on the estimate $\hat{\rho}$. There is a clear possibility of specification error if the dependence is different from AR(1). Further, conditioning on $\hat{\rho}$ is probably too much conditioning, as it often leads to understatement of sampling errors in $\hat{\mu}$. Essentially $\hat{\rho}$ is not close enough to being ancillary (Skovgaard 1986). Addressing specification error, the HAC estimators generalize to an NP estimate of the variance, as (4) in Müller (e.g., Newey-West). Thus, the specification error question is addressed. But the conditioning on $\hat{\omega}$ is still too much for our purposes—intervals are too small under repeated sampling. KVB and KV uncondition, in the same way that the "t" distribution generalizes the normal distribution conditional on $\hat{\sigma}$ by unconditioning, thus improving the approximation to the sampling distribution. Still there is conditioning—the distant $\gamma(j)$ are zero and therefore do not depend on parameters.

4. FREQUENCY DOMAIN

Müller represents the data as $(\hat{\mu}, Z^{\sin}, Z^{\cos}) = (\hat{\mu}, a)$. Then a is additionally factored as $a = (p_1, \ldots, p_n, a')$ with

$$p_l = \frac{1}{2}\left((Z_l^{\cos})^2 + (Z_l^{\sin})^2\right)$$

and again to $a = (\hat{\omega}^2, a'')$ with $\hat{\omega}^2 = \frac{1}{n}\sum p_l$. Here, the natural ancillarity assumption is on the interval around zero in which the spectrum is constant. The choice of $n_T < T$ is like a bandwidth choice. We would like an asymptotic theory that can guide the choice of n_T? For strongly autocorrelated series, the procedure apparently does not approximate the sampling distribution well enough. More variation needs to be introduced in the asymptotic approximation.

It is natural to consider possibilities for unconditioning. KVB choose a method like specifying $n_T/T = b$, so that the sampling distribution of $\hat{\omega}$ does not disappear. Müller instead reduces the per-observation information content of the data asymptotically. He considers the case $\rho_T = 1 - c/T$ for fixed c to get a local to unity theory. This theory leads to a limit for the variance that depends on c. Müller addresses this by marginalizing the limit distribution using a distribution for c. Note that c has no operational meaning—like a kernel or a bandwidth choice. Unconditioning is unquestionably a good idea here, and especially good in hindsight, since the resulting statistics appear to perform competitively in simulations. However, I find it difficult to think about a distribution for c. This has mean about 186. What does this mean? Can it be translated into a mean for ρ (a linear function of c)? The result might be a distribution easier to specify plausibly.

5. MISCELLANEOUS COMMENTS

I am intrigued by the good performance of the Ibragimov-Müller approach. Is it the case that the choice of bootstrap method is like the choice of assumptions to use in an asymptotic approximation? That is, the more we know, the better we can do.

In practice, would not a persistent series be prewhitened? This is given short shrift with the comment that "this approach requires the estimation error in the prewhitening stage to be negligible." Why is that? The whole point of the NP part is to mop up dependence, before or after prewhitening.

A minor complaint: It is certainly true that "inconsistent LRV estimators require a strong degree of homogeneity to justify the distributional approximation." Why does it follow that they "are not uniformly more 'robust' than consistent ones?" This would require at least a definition of robust and some demonstration. A poor approximation can be better than a still poorer approximation!

In short, this is a thought-provoking article, covering a lot of ground. A pleasure to read. And, I have something to agree with completely in the conclusion: "There are good reasons to be skeptical of methods that promise to automatically adapt to any given dataset. All inference requires some a priori knowledge of exploitable regularities. The more explicit and interpretable these driving assumptions, the easier it is to make sense of empirical results."

Ulrich will become a Bayesian in time.

REFERENCES

Barndorff-Nielsen, O. E. (1984), "On Conditionality Resolution and the Likelihood Ratio for Curved Exponential Models," *Scandinavian Journal of Statistics*, 11, 157–170. [322]

Berger, J. O., Wolpert, R. L., Bayarri, M. J., DeGroot, M. H., Hill, B. M., Lane, D. A., and LeCam, L. (1988), *The Likelihood Principle* (Lecture Notes-Monograph Series), Vol. 6, Hayward, CA: Institute of Mathematical Statistics. [322]

Christensen, B. J., and Kiefer, N. M. (1994), "Local Cuts and Separate Inference," *Scandinavian Journal of Statistics*, 21, 389–401. [322]

Reid, N. (1995), "The Roles of Conditioning in Inference," *Statistical Science*, 10, 138–157. [322]

Skovgaard, I. (1986), "Successive Improvement of the Order of Ancillarity," *Biometrika*, 73, 516–519. [323]

Comment

Matias D. CATTANEO
Department of Economics, University of Michigan, Ann Arbor, MI 48109 (*cattaneo@umich.edu*)

Richard K. CRUMP
Capital Markets Function, Federal Reserve Bank of New York, New York, NY 10045 (*richard.crump@ny.frb.org*)

1. INTRODUCTION

Conducting valid inference in a time series setting often requires the use of a heteroscedasticity and autocorrelation (HAC) robust variance estimator. Under mild dependence structures, the theory and practice of such estimators is well developed. However, under more severe forms of dependence, the conventional distributional approximation usually employed to describe the finite-sample properties of test statistics based on these estimators tends to be poor. Professor Müller is to be congratulated for this excellent article addressing the important issue of conducting valid inference using HAC estimators in the presence of strong autocorrelation. The class of tests introduced in the article should prove useful both to applied practitioners and as a foundation for future theoretical work.

This comment is comprised of two sections. First, we compare the main contribution of Müller (2014) (hereafter, the "S_q test") to a theoretically valid approach based on the usual t-test. More specifically, to gain further insight into the properties of the S_q test, we compare its finite-sample properties to those of a t-test with limiting distribution obtained under the local-to-unity parameterization and fixed-b asymptotics. We use critical values obtained from a Bonferroni-based procedure to control size in the presence of the nuisance parameter governing the degree of dependence. Second, we employ the new test in an empirical application by revisiting the question of long-horizon predictability in asset returns. We find that the S_q test provides evidence of predictability of equity returns by the dividend yield at shorter horizons when the sample is restricted to end in 1990. The S_q test, when applied to bond returns, produces little evidence of predictability in our application. In both applications the conclusions drawn from the S_q test can be sensitive to the choice of q, suggesting that further work will be necessary to guide the use of the S_q test in empirical applications.

2. FIXED-b ASYMPTOTICS IN A LOCAL-TO-UNITY SETTING

The new testing procedure of Müller (2014) is motivated by the assumption that for a restricted class of frequencies, governed by the user-defined parameter q, the spectral density is well approximated by the spectral density of a nearly integrated autoregressive process. By focusing only on this band of frequencies, the core assumption of the article is of the "semiparametric" variety. The exact form of the S_q test is then derived under the assumption of scale invariance and maximization of a weighted-average power criterion using the results of Elliott, Müller, and Watson (2013).

Suppose that instead of making the semiparametric assumption of Müller (2014), we assume $\{y_t : t = 1, \ldots, T\}$ is generated by

$$y_t = \mu + \varepsilon_t, \qquad \varepsilon_t = \rho\varepsilon_{t-1} + \eta_t \tag{2.1}$$

with initial condition ε_0, where $\{\eta_t\}$ is a weakly dependent process. Furthermore, we impose the local-to-unity parameterization, $\rho_T = 1 - c/T$, where $c \in [0, \infty)$. In words, we assume that the data are generated by a nearly integrated autoregressive process at all frequencies. As in Müller (2014), we are interested in testing $\mathbb{H}_0 : \mu = \mu_0$ versus the alternative that $\mathbb{H}_A : \mu \neq \mu_0$. We use the fixed-$b$ asymptotics of Kiefer and Vogelsang (2005) in this setting. Following, for example, Atchadé and Cattaneo (2014), we can write the long-run variance estimator as

$$\hat{\omega}^2_{k,S_T} = T^{-3}\sum_{\ell=1}^{T}\sum_{j=1}^{T}\left\{k_b\left(\frac{\ell - j}{S_T}\right) - v_T(\ell) - v_T(j) + u_T\right\} \times y_\ell y_j$$

$$v_T(\ell) = T^{-1}\sum_{i=1}^{T} k_b\left(\frac{\ell - i}{S_T}\right), \; u_T = T^{-2}\sum_{i=1}^{T}\sum_{j=1}^{T} k_b\left(\frac{j - i}{S_T}\right),$$

where $k_b(\cdot)$ is a kernel function. In the simulations, we use a Parzen kernel, $b = 0.5$ or $b = 1$, and $S_T = T$. Then, under regularity conditions, and defining $\hat{\mu} = T^{-1}\sum_{t=1}^{T} y_t$, we have

$$\tau_b = \frac{T^{-1/2}(\hat{\mu} - \mu)}{\sqrt{\hat{\omega}^2_{k,S_T}}} \xrightarrow{d} \frac{\int_0^1 B_c(r)dr}{\int_0^1\int_0^1 K_b(s,t)B_c(s)B_c(t)dsdt},$$

$$K_b(s,t) = k_b(s-t) - \int_0^1 k_b(s-w)\,dw - \int_0^1 k_b(t-w)\,dw + \int_0^1\int_0^1 k_b(w_1 - w_2)\,dw_1 dw_2,$$

where $\{B_c(s) : s \in [0,1]\}$ is an Ornstein–Uhlenbeck process. For instance, this result may be obtained from similar steps as in Tanaka (1996, chap. 5). The limiting distribution of the t-test, τ_b, is then solely a function of the user-defined choice of kernel

Journal of Business & Economic Statistics
July 2014, Vol. 32, No. 3
DOI: 10.1080/07350015.2014.928220

Table 1. AR(1), $\varepsilon_0 \sim \mathcal{N}\left(0, \sigma_\eta^2/(1-\rho^2)\right)$

ρ	S_{12}	S_{24}	S_{48}	$\tau_{1/2}$	τ_1	$\tau_{1/2}^\star$	$\tau_1^\star$
			Panel A: size				
0	4.9	4.7	4.7	5.5	5.3	5.5	5.7
0.7	5.0	4.9	4.8	5.2	5.1	5.2	5.5
0.9	5.0	5.0	5.2	4.1	3.9	4.2	4.3
0.95	5.0	5.1	5.2	3.5	3.4	3.5	3.3
0.98	4.9	4.9	5.1	5.0	5.4	3.1	2.9
0.999	4.4	4.4	4.2	48.0	41.6	5.2	4.9
			Panel B: power				
0	34.2	42.1	47.1	36.5	27.8	36.3	29.2
0.7	33.8	40.6	44.3	34.7	26.7	34.9	28.1
0.9	28.0	32.9	35.5	26.1	21.7	26.6	23.0
0.95	20.2	22.5	24.7	19.0	18.8	16.7	14.8
0.98	11.2	11.7	12.0	29.8	27.7	10.1	9.3
0.999	4.9	5.0	4.9	100.0	100.0	8.1	8.0
			Panel C: size-adjusted power				
0	34.6	43.1	47.9	34.0	26.7	34.0	26.7
0.7	33.6	40.9	45.1	34.6	26.2	34.6	26.2
0.9	27.8	33.0	34.3	34.6	27.8	34.6	27.8
0.95	20.1	22.1	23.4	34.3	28.2	34.3	28.2
0.98	11.4	12.2	11.7	36.1	27.9	36.1	27.9
0.999	5.6	5.8	5.8	92.1	80.4	92.1	80.4

and parameter b and the nuisance parameter c, which governs the degree of persistence of the data. To conduct valid inference, it is crucial to control the size of the test with respect to the value of c. See, for example, Andrews and Guggenberger (2009) and Andrews and Guggenberger (2010) for further discussion on the importance of uniformly valid inference in econometrics. We use the Bonferroni-based critical values introduced in McCloskey (2012). We view this approach as possibly the most natural point of comparison to the S_q test.

Table 1 presents the simulation results for the strictly stationary AR(1) model and set of alternatives given in Müller (2014). As in Müller (2014), we set $\sigma_\eta^2 = 1$. The first three columns present the size, power, and size-adjusted power (we show size-adjusted power so the results for the S_q test are comparable to the tables in Müller (2014). Size-adjusted power for the τ_b statistic is (nearly) the same as inference when the value of c is known and so should be interpreted with this in mind) for the S_q test with $q = 12$, $q = 24$, and $q = 48$. The next two columns provide the results for the Bonferroni-based procedure with critical values formed under the assumption that the initial condition is negligible (written as τ_b). As an additional point of comparison, the final two columns report results for the Bonferroni-based procedure with the limiting distributions constructed under a stationary initial condition (since $c > 0$ in this case, we implemented the testing procedure by setting the lower and upper critical values for μ equal to $-\infty$ and ∞, respectively, whenever the confidence interval formed for c included the smallest value in our grid. This approach thus requires a user-defined minimum value of c. In our simulations, we chose $c_{\min} = 0.01$) (written as $\tau_b^\star$). The "S-Bonf-Adj" critical values are formed with $\beta = 0.15$ (i.e., nominal coverage of the parameter c equal to $1 - \beta$). To form confidence intervals for c, we invert an ordinary least square (OLS)-based augmented Dickey–Fuller (ADF) test with lag length chosen by the modified Akaike information criterion (MAIC) of Perron and Qu (2007). Refinements of our procedure could include shifting to the "S-Bonf-Min" critical values of McCloskey (2012), a different choice of confidence interval for the local-to-unity parameter or an alternative test statistic to τ. However, we prefer this formulation for simplicity of interpretation. All results are based on 20,000 simulations.

The results of Table 1 are instructive on how to interpret properties of the S_q test. First, the τ_b test statistic controls size away from values of ρ close to one, but is severely size distorted when $\rho = 0.999$. This reflects the fact that the critical values for τ_b are constructed under the assumption that the initial condition is negligible. Meanwhile, the $\tau_b^\star$ test statistic controls size well across this grid of values of ρ. The power of the $\tau_{1/2}^\star$ test is comparable to that of the S_{12} test. However, the S_{24} and S_{48} tests have higher power when ρ moves away from one. In Table 2, we again consider the AR(1) specification but use the fixed initial condition $\varepsilon_0 = 0$. The pattern of the results is similar to that of Table 1 except that the τ_b test statistic controls size even in cases where the error terms are highly persistent. We can also contrast the results from Tables 1 and 2 to those using the least-favorable critical value. In this setting, because critical values for τ_b grow as $c \downarrow 0$, choosing to use the least-favorable critical value produces severely undersized tests for all areas of the parameter space except when ρ is very close to one.

In Table 3, we present results for the "AR(1) + noise" specification and set of alternatives given in Müller (2014). As in Müller (2014), we set the variance of the additive noise term to 4. Similar to Table 1, the τ_b test is oversized when ρ is near one with excessive size distortion when $\rho = 0.999$. In contrast, the $\tau_b^\star$ test controls size well across the grid of ρ values and the $\tau_{1/2}^\star$

Table 2. AR(1), $\varepsilon_0 = 0$

ρ	S_{12}	S_{24}	S_{48}	$\tau_{1/2}$	τ_1	$\tau^\star_{1/2}$	$\tau^\star_1$
			Panel A: size				
0	4.8	4.8	4.8	5.6	5.4	5.5	5.7
0.7	5.0	4.9	4.8	5.3	5.1	5.3	5.5
0.9	5.3	5.1	5.5	4.2	4.0	4.4	4.5
0.95	5.4	5.4	5.6	3.6	3.5	3.5	3.4
0.98	5.0	4.8	5.0	4.1	4.7	3.3	3.0
0.999	2.5	2.4	2.3	6.0	5.3	3.0	2.5
			Panel B: power				
0	34.5	42.5	47.3	36.7	28.0	36.4	29.3
0.7	34.3	41.3	44.9	35.2	27.2	35.4	28.5
0.9	29.7	35.1	37.7	27.1	22.4	27.6	23.9
0.95	22.9	25.7	28.1	20.1	20.1	17.6	15.9
0.98	13.8	14.9	15.6	34.0	32.5	11.0	10.4
0.999	5.2	5.3	5.1	100.0	100.0	8.3	8.2
			Panel C: size-adjusted power				
0	35.0	43.1	48.1	34.3	27.5	34.3	27.5
0.7	34.4	41.6	45.8	34.9	27.4	34.9	27.4
0.9	28.9	34.5	35.5	35.7	28.7	35.7	28.7
0.95	21.6	23.9	25.4	37.7	29.9	37.7	29.9
0.98	13.8	15.9	15.4	48.7	38.1	48.7	38.1
0.999	10.2	12.6	11.1	100.0	100.0	100.0	100.0

statistic has comparable power properties to that of the S_{12} test. The S_{24} and S_{48} tests suffer from size distortion for larger values of ρ as discussed in Müller (2014). Table 4 reports the companion results under a zero initial condition. As in Table 3, the S_{12} test and the $\tau^\star_b$ control size well across these values of ρ. The S_{24} and S_{48} tests show some size distortion as ρ moves above 0.90 whereas the τ_b test controls size except when $\rho = 0.999$.

There are three main observations from the limited simulation evidence we present. First, it is instructive to see where power is directed by the S_q test. In particular, the S_q test has little power when ρ is near 1 but much higher power elsewhere. From a practitioner's perspective, this is a very appealing property (see next section) as the test controls size well in exactly the region of the parameter space where

Table 3. AR(1) + Noise, $\varepsilon_0 \sim \mathcal{N}\left(0, \sigma_\eta^2/(1-\rho^2)\right)$

ρ	S_{12}	S_{24}	S_{48}	$\tau_{1/2}$	τ_1	$\tau^\star_{1/2}$	$\tau^\star_1$
			Panel A: size				
0	4.9	5.0	5.0	5.5	5.3	5.4	5.5
0.7	4.9	4.9	5.6	5.2	4.9	5.2	5.1
0.9	5.1	5.9	9.2	4.9	4.2	5.0	4.3
0.95	5.2	6.7	12.4	4.7	4.3	4.7	4.0
0.98	5.3	7.0	15.0	5.9	6.0	4.0	3.2
0.999	4.9	7.4	17.0	48.1	41.9	4.8	4.5
			Panel B: power				
0	34.4	42.7	47.2	35.5	27.1	35.1	27.9
0.7	34.1	42.5	49.1	33.9	26.3	33.7	26.8
0.9	28.7	37.9	53.7	27.1	22.8	26.7	21.6
0.95	22.0	29.8	50.0	21.2	20.7	17.6	14.7
0.98	12.4	17.8	37.3	30.9	28.7	9.6	8.6
0.999	5.5	8.4	19.4	100.0	100.0	6.8	6.7
			Panel C: size-adjusted power				
0	34.8	42.7	47.3	34.4	26.2	34.4	26.2
0.7	34.6	42.9	47.0	36.2	26.9	36.2	26.9
0.9	28.3	34.1	38.0	35.7	29.1	35.7	29.1
0.95	21.2	23.0	25.4	35.0	28.2	35.0	28.2
0.98	11.5	11.9	12.0	36.6	28.4	36.6	28.4
0.999	5.7	5.7	5.6	92.4	80.6	92.4	80.6

Table 4. AR(1) + Noise, $\varepsilon_0 = 0$

ρ	S_{12}	S_{24}	S_{48}	$\tau_{1/2}$	τ_1	$\tau^\star_{1/2}$	$\tau^\star_1$
			Panel A: size				
0	4.9	4.9	5.0	5.5	5.3	5.4	5.6
0.7	4.9	4.9	5.5	5.2	4.9	5.1	5.1
0.9	5.3	6.0	9.6	4.9	4.5	5.0	4.6
0.95	5.6	7.1	12.6	4.7	4.4	4.6	4.1
0.98	5.4	6.9	14.7	5.4	5.5	4.5	3.6
0.999	2.9	4.0	9.8	7.1	6.3	3.7	2.8
			Panel B: power				
0	34.5	42.8	47.3	35.5	27.2	35.2	27.9
0.7	34.5	43.0	49.5	34.2	26.6	34.0	27.1
0.9	30.5	40.1	55.4	28.2	24.0	27.6	22.6
0.95	24.7	33.6	54.4	22.5	22.3	18.5	15.8
0.98	15.4	22.4	44.5	35.0	33.6	10.2	9.6
0.999	5.9	8.7	20.3	100.0	100.0	6.9	6.8
			Panel C: size-adjusted power				
0	34.8	43.1	47.2	34.4	26.1	34.4	26.1
0.7	34.9	43.2	47.0	36.7	27.4	36.7	27.4
0.9	29.3	35.9	39.6	36.4	28.4	36.4	28.4
0.95	22.8	25.3	27.4	37.9	30.2	37.9	30.2
0.98	14.1	14.6	15.5	47.7	38.3	47.7	38.3
0.999	10.1	11.6	10.7	100.0	100.0	100.0	100.0

inference is most difficult. Second, the results show clearly that the underlying assumption of how the initial condition is generated plays a key role in the performance of testing procedures for large values of ρ. In particular, procedures that may perform well in terms of size control when the initial condition is negligible can have severe size distortion when the initial condition is drawn from its unconditional distribution. This is not a surprising result, but one that may not be fully appreciated outside of the unit root testing literature. Third, the simulations show that the choice of q can matter and so further guidance for applied practitioners would be an important contribution for future work.

3. APPLICATION: LONG-HORIZON RETURN PREDICTABILITY

Conventional wisdom in applied time series would suggest the presence of model misspecification when products of estimated error terms are highly persistent. The consummate example would be the omission of the lagged left-hand side variable in the canonical spurious regression case. However, an important situation where the prescription to add lags of the dependent variable is unavailable is in long-horizon predictability regressions, such as predicting h-period ahead stock and bond returns. We revisit the question of longer-horizon return predictability for equities and bonds using the S_q test. These two simple empirical exercises demonstrate clearly the applicability of the new testing procedure.

Long-horizon stock return predictability has been studied by a number of authors (see, e.g., Koijen and Nieuwerburgh 2011 or Rapach and Zhou 2013 for general discussions.) Here, we follow the standard approach in the literature and form continuously compounded, cumulative stock returns as

$$rx_{t+h} = \sum_{j=1}^{h} rx_{t+j}, \tag{3.1}$$

where rx_t is the 1 month compounded excess return formed as the Center for Research in Security Prices (CRSP) value-weighted return less the 1 month interest rate (here measured as the Fama–Bliss risk-free rate). We consider two regressors: (1) the log dividend yield formed as the natural logarithm of the sum of monthly dividends over the last 12 months relative to the current price and (2) the Fama–Bliss risk-free rate. We include the latter regressor as Ang and Bekaert (2007) argued that the predictive ability of the dividend yield is enhanced by including this variable. We consider two sample periods: monthly data over the period 1952–2012 and over the period 1952–1990. We include the latter, shorter sample, as it is perceived that the dividend yield was a more reliable predictor of stock returns up to 1990. Finally, we report results for values of $h \in \{6, 12, 24, 36\}$ months. Because the returns in Equation (3.1) are calculated with overlapping periods, they have a considerable degree of persistence, comparable to that of the right-hand side variables. Thus, concerns have arisen about conducting inference in such a setting.

Confidence intervals formed using Newey–West (NW) standard errors with lag truncation parameter h and those using Müller (2014) are reported in Table 5. When the sample is restricted to 1952–1990, the S_q test provides evidence of return predictability at shorter horizons, although the confidence intervals vary considerably depending on the specification and the choice of q. However, there is little evidence of predictability

Table 5. Long-horizon regressions: Equity returns

		1952–1990		
		Dividend yield only		
h	6	12	24	36
$\hat{\beta}_{dy}$	16.93	33.818	52.339	54.685
NW	(7.357, 26.503)	(15.823, 51.814)	(17.213, 87.464)	(14.407, 94.964)
S_{12}	(2.736, 54.803)	(3.619, 118.719)	$(-16.271, 203.980)$	$(-14.234, 221.210)$
S_{24}	(3.652, 33.872)	(5.33, 71.531)	$(-5.834, 194.429)$	$(-22.098, 148.099)$
S_{48}	(4.663, 32.203)	(5.334, 76.591)	$(-29.8, 125.125)$	$(-29.001, 159.285)$
		Dividend yield and risk-free rate		
h	6	12	24	36
$\hat{\beta}_{dy}$	26.435	49.697	69.028	71.753
NW	(17.478, 35.392)	(34.268, 65.125)	(43.526, 94.531)	(44.859, 98.648)
S_{12}	(17.908, 429.574)	(32.07, 874.616)	$(-247.772, 155.535)$	$(-\infty, \infty)$
S_{24}	(15.647, 52.028)	(25.505, 77.308)	$(-115.492, 102.134)$	$(-323.597, 108.84)$
S_{48}	$(-\infty, -548.666)$ $\cup$ (14.642, ∞)	$(-\infty, -375.173)$ $\cup$ (22.021, ∞)	$(-\infty, 97.251)$ $\cup$ (534.008, ∞)	$(-\infty, 118.368)$ $\cup$ (213.433, ∞)
		1952–2012		
		Dividend yield only		
h	6	12	24	36
$\hat{\beta}_{dy}$	5.23	10.564	19.103	23.983
NW	(0.791, 9.668)	(1.268, 19.859)	(1.699, 36.508)	(4.152, 43.815)
S_{12}	$(-\infty, -10.393)$ $\cup$ $(-2.058, \infty)$	$(-\infty, -19.806)$ $\cup$ $(-4.435, \infty)$	$(-\infty, -28.866)$ $\cup$ $(-10.372, \infty)$	$(-\infty, \infty)$
S_{24}	$(-\infty, -22.479)$ $\cup$ $(-3.448, \infty)$	$(-\infty, -47.079)$ $\cup$ $(-6.829, \infty)$	$(-\infty, -63.666)$ $\cup$ $(-16.849, \infty)$	$(-\infty, \infty)$
S_{48}	$(-\infty, -33.192)$ $\cup$ $(-3.911, 14.364)$ $\cup$ $(30.886, \infty)$	$(-\infty, -67.765)$ $\cup$ $(-10.525, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
		Dividend yield and risk-free rate		
h	6	12	24	36
$\hat{\beta}_{dy}$	8.397	15.840	25.917	30.976
NW	(3.924, 12.871)	(6.101, 25.580)	(7.795, 44.038)	(11.943, 50.010)
S_{12}	$(-\infty, -9.893)$ $\cup$ $(2.884, \infty)$	$(-\infty, -14.318)$ $\cup$ $(4.238, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
S_{24}	$(-\infty, -16.464)$ $\cup$ $(-1.34, \infty)$	$(-\infty, -31.862)$ $\cup$ $(-5.298, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
S_{48}	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$

NOTE: This table shows the results for long-horizon predictability regressions of equity market returns on the log dividend yield and the risk-free rate. $\hat{\beta}_{dy}$ is the OLS coefficient corresponding to the log dividend yield. Newey–West (NW) standard errors are constructed with h lags. All confidence intervals have nominal coverage of 95%.

at shorter horizons for the full sample. At longer horizons there is no evidence of predictability for either the restricted or full sample based on the S_q test.

Next, we consider a similar exercise for excess bond returns. Specifically, we revisit the influential work of Cochrane and Piazzesi (2005, 2008), where the authors form a bond-return forecasting factor using linear combinations of forward rates. To proceed, we first must form this return-forecasting factor (hereafter, CP factor). We use excess returns and log forward rates, defined by

$$rx_{t+1}^{(n)} \equiv p_{t+1}^{(n-1)} - p_t^{(n)} - y_t^{(1)}, \qquad f_t^{(n)} \equiv p_t^{(n-1)} - p_t^{(n)},$$

where $p_t^{(n)}$ is the log price of an n year discount bond at time t and $y_t^{(1)}$ is the 1-month GSW rate (GSW refers to zero-coupon bond yields from Gurkaynak, Sack, and Wright (2007), which are available at a daily frequency on the Board of Governors of the Federal Reserve's research data page). We use GSW yields to construct excess returns and Fama–Bliss forward rates as regressors. We follow Cochrane and Piazzesi (2008) and regress 14 excess returns $rx_{t+1} = [rx_{t+1}^{(2)}, rx_{t+1}^{(3)}, \ldots, rx_{t+1}^{(15)}]'$ on a constant $z_t = 1$ and five forward rates $w_t = [y_t^{(1)}, f_t^{(2)}, \ldots, f_t^{(5)}]'$. Cochrane and Piazzesi (2008) formed the CP factor by taking the first principal component of the fitted values from this regression. It can be shown that this is equivalent to the maximum-likelihood estimator (MLE) of a reduced-rank regression (with coefficient matrix of rank one) under the assumption of iid Gaussian errors and a scalar variance matrix. We also consider a weighted version of the CP factor, formed as the MLE under the assumption of a diagonal variance matrix (see Adrian, Crump, and Moench 2014 for further details).

Table 6. Long-horizon regressions: Bond returns

	In-sample	Out-of-sample (5 years)	Out-of-sample (10 years)	Out-of-sample (15 years)
Identity weight matrix				
$\hat{\beta}_{CP}$	0.220	0.052	0.082	0.063
NW	(0.101, 0.339)	(−0.001, 0.105)	(0.047, 0.117)	(−0.009, 0.136)
S_{12}	$(-\infty, 0.328) \cup (3.545, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
S_{24}	(−0.102, 0.324)	(−0.14, 0.197)	$(-\infty, 0.125) \cup (0.225, \infty)$	$(-\infty, -8.129) \cup (-0.031, \infty)$
S_{48}	(−0.106, 0.356)	(−0.165, 0.13)	$(-\infty, -0.277) \cup (-0.036, 0.125) \cup (0.456, \infty)$	(−0.07, 0.741)
Diagonal weight matrix				
$\hat{\beta}_{CP}$	0.231	0.053	0.086	0.066
NW	(0.105, 0.356)	(−0.003, 0.108)	(0.049, 0.122)	(−0.012, 0.144)
S_{12}	$(-\infty, 0.355) \cup (2.257, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
S_{24}	(−0.116, 0.342)	(−0.155, 0.209)	$(-\infty, 0.132) \cup (0.226, \infty)$	$(-\infty, -13.464) \cup (-0.035, \infty)$
S_{48}	(−0.107, 0.376)	(−0.178, 0.134)	$(-\infty, -0.205) \cup (-0.045, 0.13) \cup (0.419, \infty)$	(−0.078, 0.843)

NOTE: This table shows the results for long-horizon predictability regressions of 1 year excess holding period returns on the CP factor. $\hat{\beta}_{CP}$ is the OLS coefficient corresponding to the CP factor. The first column report results for the CP factor constructed on data for 1971–2012. The next three columns report results for the CP factor constructed in real time with a 5, 10, and 15 year burn-in period, respectively. Newey–West (NW) standard errors are constructed with 12 lags. All confidence intervals have nominal coverage of 95%.

We then regress average excess returns on the CP factor,

$$\overline{rx}_{t+1} = \alpha + \beta x_t + \epsilon_t,$$

where $\overline{rx}_{t+1}$ is the average return, across maturities, $\overline{rx}_{t+1} = \frac{1}{14}\sum_{n=2}^{15} rx_{t+1}^{(n)}$.

We use the S_q test statistic to construct confidence intervals (results reported in Table 6). We then repeat the exercise but now we construct the CP factor without using future information after a certain burn-in period. (We also considered specifications that added the term spread as an additional predictor. The results in this case were qualitatively similar to those presented here.) We find that the conclusions drawn from the confidence intervals formed from the S_q test are sensitive to different values of q and different burn-in periods. As in the equity application, we find that confidence intervals can be asymmetric, sometimes disjoint but nonempty (as discussed in Müller 2014). Despite this, we find no evidence that a predictive relationship can be uncovered with these data.

ACKNOWLEDGMENTS

The authors thank Lutz Kilian, Adam McCloskey, Emanuel Moench, and Ulrich Müller for helpful comments and discussions and the editors, Keisuke Hirano and Jonathan Wright, for inviting us to participate in this intellectual exchange. Benjamin Mills provided excellent research assistance. The first author gratefully acknowledges financial support from the National Science Foundation (SES 1122994). The views expressed in this article are those of the authors and do not necessarily represent those of the Federal Reserve Bank of New York or the Federal Reserve System.

REFERENCES

Adrian, T., Crump, R. K., and Moench, E. (2014), "Regression-based Estimation of Dynamic Asset Pricing Models," Staff Report 493, Federal Reserve Bank of New York. [328]

——— (2009), "Hybrid and Size-corrected Subsampling Methods," *Econometrica*, 77, 721–762. [325]

Andrews, D. W. K., and Guggenberger, P. (2010), "Asymptotic Size and a Problem with Subsampling and with the *m* Out of *n* Bootstrap," *Econometric Theory*, 26, 426–468. [325]

Ang, A., and Bekaert, G. (2007), "Stock Return Predictability: Is it There?," *Review of Financial Studies*, 20, 651–707. [327]

Atchadé, Y. F., and Cattaneo, M. D. (2014), "A Martingale Decomposition for Quadratic Forms of Markov Chains (with Applications)," *Stochastic Processes and Their Applications*, 124, 646–677. [324]

Cochrane, J., and Piazzesi, M. (2005), "Bond Risk Premia," *American Economic Review*, 95, 138–160. [328]

——— (2008), "Decomposing the Yield Curve," Working Paper, University of Chicago. [328]

Elliott, G., Müller, U. K., and Watson, M. W. (2013), "Nearly Optimal Tests When a Nuisance Parameter is Present Under the Null Hypothesis," Working Paper, Princeton University. [324]

Gurkaynak, R. S., Sack, B., and Wright, J. H. (2007), "The U.S. Treasury Yield Curve: 1961 to the Present," *Journal of Monetary Economics*, 54, 2291–2304. [328]

Kiefer, N. M., and Vogelsang, T. J. (2005), "A New Asymptotic Theory for Heteroskedasticity-autocorrelation Robust Tests," *Econometric Theory*, 21, 1130–1164. [324]

Koijen, R. S., and Nieuwerburgh, S. V. (2011), "Predictability of Returns and Cash Flows," *Annual Review of Financial Economics*, 3, 467–491. [327]

McCloskey, A. (2012), "Bonferroni-Based Size-Correction for Nonstandard Testing Problems," Working Paper, Brown University. [325]

Müller, U. K. (2014), "HAC Corrections for Strongly Autocorrelated Time Series," *Journal of Business and Economic Statistics*, 32, 311–322. [324,325,327,329]

Perron, P., and Qu, Z. (2007), "A Simple Modification to Improve the Finite Sample Properties of Ng and Perron's Unit Root Tests," *Economics Letters*, 94, 12–19. [325]

Rapach, D. E., and Zhou, G. (2013), "Forecasting Stock Returns," in *Handbook of Economic Forecasting (Vol. II)*, eds. G. Elliott and A. Timmermann, Amsterdam: North-Holland, pp. 329–383. [327]

Tanaka, K. (1996), *Time Series Analysis: Nonstationary and Noninvertible Distribution Theory*, New York: Wiley. [324]

Comment

Yixiao Sun
Department of Economics, University of California, San Diego, La Jolla, CA 92093 (*yisun@ucsd.edu*)

1. ON THE NEARLY OPTIMAL TEST

Müller applies the theory of optimal statistical testing to heteroscedasticity and autocorrelation robust (HAR) inference in the presence of strong autocorrelation. As a starting point, Müller uses Le Cam's idea on the limits of experiments ingeniously and converts a more complicated finite sample testing problem into an asymptotically equivalent and simpler testing problem. The main barrier to optimal testing is that both the null hypothesis and alternative hypothesis are composite, even after the asymptotic reduction based on Le Cam's idea. So the Neyman–Pearson lemma does not directly apply.

To reduce the dimension of the alternative hypothesis space, it is standard practice to employ a weighting function and take a weighted average of the probability distributions under the alternative hypothesis. See, for example, Cox and Hinkley (1974, p. 102). The weighting function should reflect a user's belief about the likelihood of different parameter values and the associated cost of false acceptance under the alternative. Selecting the weighting function is as difficult as selecting one point out of many possible parameter values. A test that is designed to be optimal against a point alternative may not be optimal for other alternatives. The near-optimality of Müller's test should be interpreted with this point in mind.

There are a number of ways to reduce the dimension of the null hypothesis space, including the invariance arguments and the conditioning argument on sufficient statistics. See, for example, Cox and Hinkley (1974, Ch. 5). In fact, Müller uses scale invariance to remove one nuisance parameter. However, as in many other contexts, here the null cannot be reduced to a point by using the standard arguments. Müller follows Wald, Lehmann, Stein, and other pioneers in statistical testing and constructs the so-called least favorable distribution over the null parameter space and uses it to average the probability distributions. This effectively reduces the composite null into a simple one. However, the least favorable distribution has to be found numerically, which can be a formidable task. This is perhaps one of the reasons that the theory of optimal testing has not been widely used in statistics and econometrics. A contribution of Müller's article is to find an approximate least favorable distribution and construct a test that is nearly optimal.

The reason to employ the least favorable distribution is that we want to control the level of the test for each point in the parameter space under the null. While we are content with the average power under the alternative, we are not satisfied with the control of the average level under the null. In fact, the requirement on size control is even stronger: the null rejection probability has to be controlled uniformly over the parameter space under the null. There is some contradiction here, which arises from the classical dogma that puts size control before power maximization. The requirement that the null rejection probability has to be controlled for each possible parameter value under the null, no matter how unlikely a priori a given value is, is overly conservative. The test designed under this principle can suffer from a severe power loss. In fact, in the simulation study, Müller's test often has a lower (size-adjusted) power than some commonly used tests. I am sympathetic with the argument that the power loss is the cost one has to pay to achieve the size accuracy as size-adjustment is not empirically feasible. However, one can always design a test with accurate size but no power. Ultimately, there is a trade-off between size control and power improvement. Using the least favorable distribution does not necessarily strike the optimal trade-off. As a compromise, one may want to control the average level/size of the test over a few empirically relevant regions in the parameter space.

There is some convincing simulation evidence that Müller's test is much more accurate than existing tests for some data-generating processes (DGP). These are the DGP's where the finite sample testing problem can be approximated very well by the asymptotically equivalent testing problem. However, there is not much research on the quality of the approximation. If the approximation error is large, Müller's test, which is nearly optimal for the asymptotically equivalent problem, may not be optimal for the original problem. For example, when the error process in the location model is an AR(1) plus noise, Müller's simulation results show that his test can still over-reject considerably. As another example, if the error process follows the AR(2) model $u_t = 1.90u_{t-1} - 0.95u_{t-2} + e_t$ where $e_t \sim$ iid N(0,1), then Müller's test (and many other tests) suffers from under-rejection. In this example, the modulus of the larger root is about 0.97, which is close to 1. However, the spectral density does not resemble that of an AR(1) process. It does not have a peak at the origin. Instead, there is a peak near the origin. As a result, the quality of the AR(1) approximation is low. If the periodograms used in the variance estimator include the peak, then the variance estimator will be biased upward, leading to a smaller test statistic and under-rejection.

2. NEAR-UNITY FIXED-SMOOTHING ASYMPTOTIC APPROXIMATION

An attractive feature of Müller's test is that the scenarios under which it is optimal or nearly optimal are given explicitly. However, practitioners may find it unattractive because of the computation cost, the unfamiliar form of the test statistic, and its applicability to models beyond the simple Gaussian location model. An alternative approach to deal with strong

Journal of Business & Economic Statistics
July 2014, Vol. 32, No. 3
DOI: 10.1080/07350015.2014.926817
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jbes*.

autocorrelation is to derive a new approximation for the conventional test statistic that captures the strong autocorrelation. This was recently developed by Sun (2014b).

To provide the context for further discussion, I give a brief summary of Sun (2014b) here. Consider a p-dimensional time series y_t of the form

$$y_t = \theta + e_t, t = 1, 2, \ldots, T, \tag{1}$$

where $y_t = (y_{1t}, \ldots, y_{pt})'$, $\theta = (\theta_1, \ldots, \theta_p)'$, and $e_t = (e_{1t}, \ldots, e_{pt})'$ is a zero mean process. We are interested in testing the null $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. The OLS estimator of θ is the average of $\{y_t\}$, that is, $\hat{\theta} = \bar{y} := T^{-1}\sum_{t=1}^{T} y_t$. The F-test version of the Wald statistic based on the OLS estimator is given by

$$F_T = (\hat{\theta} - \theta_0)'\hat{\Omega}^{-1}(\hat{\theta} - \theta_0)/p,$$

where $\hat{\Omega}$ is an estimator of the approximate variance of $(\hat{\theta} - \theta_0)$. When $p = 1$, we can construct the t-statistic $t_T = (\hat{\theta} - \theta_0)/\hat{\Omega}^{1/2}$.

A very general class of variance estimators is the class of quadratic variance estimators, which takes the form

$$\hat{\Omega} = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} Q_h\left(\frac{t}{T}, \frac{s}{T}\right)\hat{e}_t\hat{e}_s', \tag{2}$$

where $\hat{e}_t = e_t - \bar{e}$ for $\bar{e} = T^{-1}\sum_{t=1}^{T} e_t$ and $Q_h(r, s)$ is a weighting function that depends on the smoothing parameter h. When $Q_h(r, s) = k((r - s)b)$ for some kernel function $k(\cdot)$ and smoothing parameter b, $\hat{\Omega}$ is the commonly used kernel variance estimator. When $Q_h(r, s) = K^{-1}\sum_{j=1}^{K}\phi_j(r)\phi_j(s)$ for some basis functions $\{\phi_j(r)\}$ on $\mathbb{L}^2[0, 1]$ satisfying $\int_0^1 \phi_j(r)dr = 0$ and smoothing parameter K, we obtain the so-called series variance estimator. This estimator has a long history. It can be regarded as a multiple-window estimator with window function $\phi_k(t/T)$ (see Thompson 1982). It also belongs to the class of filter-bank estimators and $\hat{\Omega}$ is a simple average of the individual filter-bank estimators. For more discussions along this line see Thompson (1982) and Stoica and Moses (2005, Ch. 5). Recently, there has been some renewed interest in this type of variance estimators, see Phillips (2005), Sun (2006, 2011, 2013), and Müller (2007).

Define

$$Q^*_{T,h}(r, s) = Q_h(r, s) - \frac{1}{T}\sum_{\tau_1=1}^{T} Q_h\left(\frac{\tau_1}{T}, s\right) - \frac{1}{T}\sum_{\tau_2=1}^{T} Q_h\left(r, \frac{\tau_2}{T}\right) + \frac{1}{T}\sum_{\tau_1=1}^{T}\sum_{\tau_2=1}^{T} Q_h\left(\frac{\tau_1}{T}, \frac{\tau_2}{T}\right),$$

then

$$\hat{\Omega} = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} Q^*_{T,h}\left(\frac{t}{T}, \frac{s}{T}\right)e_te_s'. \tag{3}$$

The Wald statistic is then equal to

$$F_T = \left(\sum_{t=1}^{T} e_t\right)'\left[\sum_{t=1}^{T}\sum_{s=1}^{T} Q^*_{Th}\left(\frac{t}{T}, \frac{s}{T}\right)e_te_s'\right]^{-1}\left(\sum_{t=1}^{T} e_t\right)\Big/p.$$

Similarly, the t-statistic becomes

$$t_T = \frac{\sum_{t=1}^{T} e_t}{\left[\sum_{t=1}^{T}\sum_{s=1}^{T} Q^*_{Th}\left(\frac{t}{T}, \frac{s}{T}\right)e_te_s\right]^{1/2}}.$$

The question is how to approximate the sampling distributions of F_T and t_T. If $\{e_t\}$ is stationary and $T^{-1/2}\sum_{t=1}^{[Tr]} e_t$ converges weakly to a Brownian motion process, then under some conditions on Q_h, it can be shown that, for a fixed h:

$$F_T \to^d F_\infty(h) := W_p(1)'\left[\int_0^1\int_0^1 Q^*_h(r, s)\, dW_p(r) \times dW_p'(s)\right]^{-1} W_p(1)/p, \tag{4}$$

$$t_T \to^d t_\infty(h) := \frac{W_p(1)}{\sqrt{\int_0^1\int_0^1 Q^*_h(r, s)\, dW_p(r)\, dW_p'(s)}}, \tag{5}$$

where $W_p(r)$ is a $p \times 1$ vector of standard Wiener processes and

$$Q^*_h(r, s) = Q_h(r, s) - \int_0^1 Q_h(\tau_1, s)\, d\tau_1 - \int_0^1 Q_h(r, \tau_2)\, d\tau_2 + \int_0^1\int_0^1 Q_h(\tau_1, \tau_2)\, d\tau_1 d\tau_2.$$

For easy reference, I refer to the previous approximations as the stationary fixed-smoothing asymptotic approximations. They are more accurate than the chi-square approximation or the normal approximation. As pointed out by Müller's article, these approximations are still not good enough when e_t is highly autocorrelated.

To model the high autocorrelation, we assume that e_t follows an AR(1) process of the form

$$e_t = \rho_T e_{t-1} + u_t \text{ where } e_0 = O_p(1) \text{ and } \rho_T = 1 - \frac{c_m}{T}$$

for some sequence $\{c_m\}$. In spirit, this is similar to Müller's article and many other papers in the literature. See, for example, Phillips, Magdalinous, and Giraitis (2010). Under the assumption that

$$\frac{1}{\sqrt{T}}e_{[Tr]} \to \Lambda J_{c_m}(r)$$

for some matrix Λ, where $J_{c_m}(r)$ is the Ornstein–Uhlenbeck process defined by

$$dJ_{c_m}(r) = -c_m J_{c_m}(r)\, dr + dW_p(r)$$

with $J_{c_m}(0) = 0$, we can obtain the following near-unity fixed-smoothing asymptotics when c_m and h are fixed:

$$F_T \to^d F_\infty(c_m, h) := \left[\int_0^1 J_{c_m}(r)dr\right]'\left[\int_0^1\int_0^1 Q^*_h(r, s)\, J_{c_m}(r) \times J'_{c_m}(s)\, drds\right]^{-1}\left[\int_0^1 J_{c_m}(r)\, dr\right]\Big/p.$$

If we further assume that $Q_h(r, s)$ is positive definite, then for fixed c_m and h:

$$t_T \to^d t_\infty(c_m, h) := \frac{\int_0^1 J_{c_m}(r)\, dr}{\left[\int_0^1\int_0^1 Q^*_h(r, s)\, J_{c_m}(r)\, J'_{c_m}(s)\, drds\right]^{1/2}}.$$

If we let $c_m \to \infty$, then the near-unity fixed-smoothing asymptotic distributions $F_\infty(c_m, h)$ and $t_\infty(c_m, h)$ approach the stationary fixed-smoothing asymptotic distributions given in (4) and (5). On the other hand, if we let $c_m \to 0$, then $F_\infty(c_m, h)$ and $t_\infty(c_m, h)$ approach the unit-root fixed-smoothing asymptotic distributions, which are defined as $F_\infty(c_m, h)$ and $t_\infty(c_m, h)$ but with $J_{c_m}(r)$ replaced by $W_p(r)$. Depending on the value of c_m, the limiting distributions $F_\infty(c_m, h)$ and $t_\infty(c_m, h)$ provide a smooth transition from the usual stationary fixed-smoothing asymptotics to the unit-root fixed-smoothing asymptotics.

In my view, the chi-square/normal approximation, the stationary fixed-smoothing approximation, and the near-unity fixed-smoothing approximation are just different approximations to the same test statistic constructed using the same variance estimator. It is a little misleading to talk about consistent and inconsistent variance estimators. The variance estimator is actually the same but we embed it on different asymptotic paths. When the fixed-smoothing asymptotics are used, we do not necessarily require that we fix the smoothing parameter h in finite samples. In fact, in empirical applications, the sample size T is usually given beforehand and the smoothing parameter h needs to be determined using a priori information and/or information obtained from the data. Very often, the selected smoothing parameter h is larger for a larger sample size but is still small relative to the sample size. So, the empirical situations appear to be more compatible with the conventional increasing-smoothing asymptotics. The beauty of the fixed-smoothing asymptotics is that fixed-smoothing critical values are still correct under the increasing-smoothing asymptotics. In fact, in a sequence of papers (e.g., Sun 2014a), I have shown that the fixed-smoothing critical values are second-order correct under the increasing-smoothing asymptotics. In contrast, increasing-smoothing critical values are not even first-order correct under the fixed-smoothing asymptotics. Given this, the fixed-smoothing approximation can be regarded as more robust than the increasing-smoothing approximation.

The same comment applies to the local-to-unity parameter c_m. When we use the near-unity fixed-smoothing approximation, we do not have to literally fix c_m at a given value in finite samples. Whether we hold c_m fixed or let it increase with the sample size can be viewed as different asymptotic specifications to obtain approximations to the same finite sample distribution. In practice, we can estimate c_m even though a consistent estimator is not available. For a stationary AR(1) process with a fixed autoregressive coefficient, the estimator $\hat{c}_m$ derived from the OLS estimator of $\hat{\rho}_{T,m}$ necessarily converges to infinity in probability. The critical values from the near-unity fixed-smoothing asymptotic distribution are thus close to those from the stationary fixed-smoothing asymptotic distribution. So, the near-unity fixed-smoothing approximation is still asymptotically valid. For this reason, we can say that the near-unity fixed-smoothing approximation is a more robust approximation. Compared to the chi-square or normal approximation, the near-unity fixed-smoothing approximation achieves double robustness—it is asymptotically valid regardless of the limiting behaviors of c_m and h.

3. SOME SIMULATION EVIDENCE

To implement the near-unity fixed-smoothing approximation, we need to pin down the value of c_m, which cannot be consistently estimated. However, a nontrivial and informative confidence interval (CI) can still be constructed. I propose to construct a CI for c_m and use the maximum of the critical values, each of which corresponds to one value of c_m in the CI. An argument based on the Bonferroni bound can be used to determine the confidence level of the CI and the significance level of the critical values. More specifically, for tests with nominal level α, we could employ the critical value defined by

$$\mathrm{CV} = \sup_{c_m \in \mathrm{CI}_{1-\alpha+\delta}} \mathrm{CV}(c_m, 1-\delta),$$

where $\delta \le \alpha$, $\mathrm{CI}_{1-\alpha+\delta}$ is a lower confidence interval for c_m with nominal coverage probability $1-\alpha+\delta$, and $\mathrm{CV}(c_m, 1-\delta)$ is the $100(1-\delta)\%$ quantile from the distribution $F_\infty(c_m, h)$ or $t_\infty(c_m, h)$. This idea of choosing critical values in the presence of unidentified nuisance parameters has been used in various settings in statistics and econometrics. See, for example, McCloskey (2012) and references therein.

One drawback of the approach based on the Bonferroni correction is that the critical value is often too large and the resulting test often under-rejects. There are sophisticated ways to improve on the Bonferroni method. As a convenient empirical strategy, here I employ $\mathrm{CV} = \sup_{c_m \in \mathrm{CI}_{90\%}} \mathrm{CV}(c_m, 95\%)$ for nominal 5% tests. I construct the CI for c_m using the method of Andrews (1993). Other methods such as Stock (1991) and Hansen (1999) can also be used. See Mikusheva (2014) and Phillips (2014) for recent contributions on this matter. Since $\mathrm{CV}(c_m, 95\%)$ is decreasing in c_m, we only need to find the lower limit of $\mathrm{CI}_{90\%}$ to compute CV. The computational cost is very low.

Table 1. Empirical null rejection probability of nominal 5% tests with $T = 200$ under AR(2) errors

	Stationary fixed-smoothing				Near-unity fixed-smoothing				Nearly optimal tests		
(ρ_1, ρ_2)	K_{12}	K_{24}	K_{48}	KVB	K_{12}	K_{24}	K_{48}	KVB	S_{12}	S_{24}	S_{48}
(0, 0)	0.048	0.047	0.048	0.047	0.045	0.042	0.041	0.042	0.045	0.047	0.047
(0.7,0)	0.058	0.083	0.148	0.058	0.040	0.035	0.032	0.040	0.046	0.047	0.047
(0.9,0)	0.133	0.248	0.393	0.084	0.036	0.033	0.033	0.038	0.048	0.047	0.050
(0.95,0)	0.258	0.412	0.553	0.125	0.039	0.037	0.036	0.039	0.048	0.049	0.050
(0.99,0)	0.630	0.738	0.816	0.333	0.079	0.079	0.079	0.071	0.045	0.044	0.045
(1.9,−0.95)	0.005	0.002	0.015	0.015	0.000	0.000	0.000	0.000	0.025	0.001	0.000
(0.8,.1)	0.146	0.272	0.417	0.089	0.050	0.051	0.052	0.045	0.049	0.048	0.052

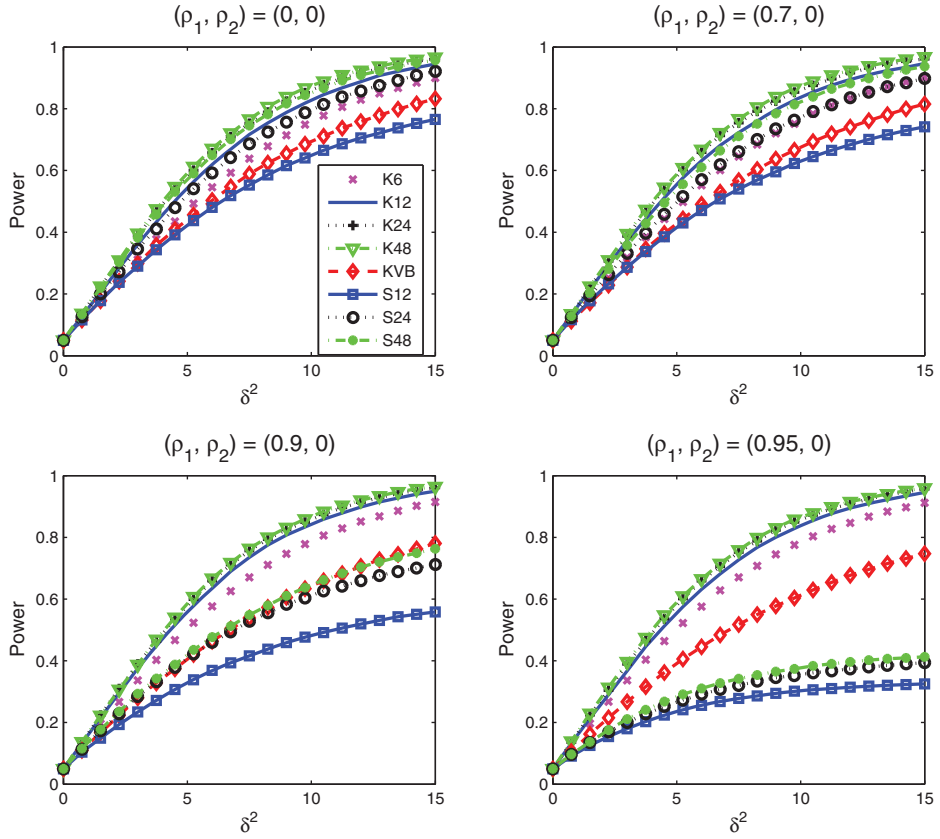

Figure 1. Size-adjusted power of different 5% tests with sample size $T = 200$ ("K6", "K12", "K24", "K48", and "KVB" are the near-unity fixed-smoothing tests while S12, S24, and S48 are Müller's tests).

I consider a univariate Gaussian location model with AR(2) error $e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + u_t$, where $u_t \sim \text{iid}N(0, 1)$. The sample size is 200. The initial value of the error process is set to be standard normal. I generate a time series of length 400 and drop the first 200 observations to minimize the initialization effect. This is similar to generating a time series with 200 observations but with the initial value drawn from its stationary distribution. I consider Müller's test, the KVB test, and the test based on the series variance estimator with the basis functions: $\phi_{2j-1}(x) = \sqrt{2}\cos(2j\pi x)$, $\phi_{2j}(x) = \sqrt{2}\sin(2j\pi x)$, $j = 1, \ldots, K/2$. The values of $K = 12, 24, 48$ correspond to the values of $q = 12, 24, 48$ in Müller's article. The number of simulation replications is 20,000.

Table 1 reports the null rejection probabilities for various two-sided 5% tests. It is clear that the tests based on the near-unity fixed-smoothing approximation are in general more accurate than those based on the usual stationary fixed-smoothing approximation. This is especially true when the process is highly autocorrelated. In term of size accuracy, Müller's test is slightly better than the near-unity fixed-smoothing test. The size accuracy of the latter test is actually quite satisfactory. As I mentioned before, the AR(2) process with $(\rho_1, \rho_2) = (1.9, -0.95)$ posts a challenge to all tests considered.

Figure 1 plots the size-adjusted power against the noncentrality parameter δ^2 in the presence AR(1) errors. The figure is representative of other configurations. It is clear from the figure that the slightly better size control of Müller's tests is achieved at the cost of some power loss.

4 CONCLUSION

Müller's article makes an important contribution to the literature on HAR inference. It has the potential for developing a standard of practice for HAR inference when the process is strongly autocorrelated. The article inspires us to think more about optimality issues in hypothesis testing. Unfortunately, uniformly optimal tests do not exist except in some special cases. This opens the door to a wide range of competing test procedures. In this discussion, I have outlined an alternative test, which is based on the standard test statistic but employs a new asymptotic approximation. The alternative test has satisfactory size but is not as accurate as Müller's test. However, Müller's test is often less powerful. The trade-off between size accuracy and power improvement is unavoidable. A prewhitening testing procedure with good size property may also be crafted. HAR testing is fundamentally a nonparametric problem. A good test requires some prior knowledge about the data generating process. In the present setting, the prior knowledge should include the range of the largest AR root and the neighborhood around origin in which the spectral density remains more or less flat. Equipped with this knowledge, a practitioner can select a testing procedure to minimize their loss function.

REFERENCES

Andrews, D. W. K. (1993), "Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models," *Econometrica*, 61, 139–165. [332]

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, New York: Chapman and Hall. [330]
Hansen, B. E. (1999), "The Grid Bootstrap and the Autoregressive Model," *Review of Economics and Statistics*, 81, 594–607. [332]
McCloskey, A. (2012), "Bonferroni-Based Size-Correction for Nonstandard Testing Problems," Working paper, Department of Economics, Brown University. Available at *http://faculty.wcas.northwestern.edu/~ate721/McCloskey_BBCV.pdf*. [332]
Mikusheva, A. (2014), "Second Order Expansion of the *t*-Statistic in AR(1) Models," *Econometric Theory*, forthcoming. [332]
Müller, U. K. (2007), "A Theory of Robust Long-Run Variance Estimation," *Journal of Econometrics*, 141, 1331–1352. [331]
Phillips, P. C. B. (2005), "HAC Estimation by Automated Regression," *Econometric Theory*, 21, 116–142. [331]
——— (2014), "On Confidence Intervals for Autoregressive Roots and Predictive Regression," *Econometrica*, 82, 1177–1195. [332]
Phillips, P. C. B., Magdalinos, T., and Giraitis, L. (2010), "Smoothing Local-to-Moderate Unit Root Theory," *Journal of Econometrics*, 158, 274–279. [331]
Stock, J. (1991), "Confidence Intervals for the Largest Autoregressive Root in US Macroeconomic Time Series," *Journal of Monetary Economics*, 28, 435–459. [332]
Stoica, P., and Moses, R. (2005), *Spectral Analysis of Signals*, Upper Saddle River, NJ: Pearson Prentice Hall. [331]
Sun, Y. (2006), "Best Quadratic Unbiased Estimators of Integrated Variance in the Presence of Market Microstructure Noise," Working paper, Department of Economics, UC San Diego. Available at *http://econweb.ucsd.edu/~yisun/bqu.pdf*. [331]
——— (2011), "Autocorrelation Robust Trend Inference With Series Variance Estimator and Testing-optimal Smoothing Parameter," *Journal of Econometrics*, 164, 345–366. [331]
——— (2013), "A Heteroscedasticity and Autocorrelation Robust F Test Using Orthonormal Series Variance Estimator," *Econometrics Journal*, 16, 1–26. [331]
——— (2014a), "Let's Fix It: Fixed-b Asymptotics versus Small-b Asymptotics in Heteroscedasticity and Autocorrelation Robust Inference," *Journal of Econometrics*, 178(3), 659–677. [332]
——— (2014b), "Fixed-Smoothing Asymptotics and Asymptotic F and t Tests in the Presence of Strong Autocorrelation," Working paper, Department of Economics, UC San Diego. Available at *http://econweb.ucsd.edu/~yisun/HAR_strong13_revised.pdf*. [331]
Thomson, D. J. (1982), "Spectrum Estimation and Harmonic Analysis," *IEEE Proceedings*, 70, 1055–1096. [331]

Comment

Timothy J. VOGELSANG
Department of Economics, Michigan State University,110 Marshall-Adams Hall, East Lansing, MI 48824
(*tjv@msu.edu*)

1. INTRODUCTION

Inference in time series settings is complicated by the possibility of strong autocorrelation in the data. In general, some aspect of a time series model is assumed to satisfy stationarity and weak dependence assumptions sufficient for laws of large numbers and (functional) central limit theorems to hold. Otherwise, inference is difficult, if not impossible, because information aggregated over time will not be informative. Even if one is willing to allow nonstationarities such as unit root behavior in the data, various transformations of the data (e.g., first differences) are assumed to satisfy stationarity and weak dependence conditions. Time series inference typically performs well when the part of the model that is assumed to satisfy stationarity and weak dependence is far from nonstationary boundaries. However, for a given sample size, when the stationary/weak dependent part of the model approaches a nonstationary boundary, inference is usually distorted in small samples. The distortions can be quite large. Alternatively, if certain parameter values are close to nonstationary boundaries, then very large sample sizes are needed for accurate inference. Strong autocorrelation is the prototypical case where a model becomes close to a nonstationary boundary and accurate inference can be challenging in this case.

For a simple location model and obvious extensions to regressions and models estimated by generalized method of moments, Müller (2014) proposes a class of tests for single parameters that are robust to strong autocorrelation and maximize a weighted power criterion. As is typical in articles by Müller, he takes a systematic and elegant theoretical approach to tackle a difficult econometrics problem. While the test statistic that emerges from his analysis, S_q, has a complicated form and is far from a priori obvious, finite sample simulations reported by Müller indicate that S_q is very robust to strong autocorrelation and retains respectable power. Comparisons to existing autocorrelation robust tests suggest that the S_q test is a useful addition to the time series econometrics toolkit.

In this note I make some additional finite sample comparisons between S_q and widely used t-tests based on nonparametric kernel long run variance estimators. While Müller includes some nonparametric kernel-based tests in his comparison group, bandwidth rules are used that tend to pick bandwidths that are too small to effectively control over-rejection problems caused by strong autocorrelation. When autocorrelation is strong, the use of very large bandwidths in conjunction with the fixed-b critical values of Kiefer and Vogelsang (2005) can lead to t-tests that have similar robustness properties to S_q while retaining substantial power advantages in some cases. Not surprisingly, the relative performance across tests is sensitive to assumptions about initial values further illustrating the inherent difficulty of carrying out robust inference when autocorrelation is strong.

Journal of Business & Economic Statistics
July 2014, Vol. 32, No. 3
DOI: 10.1080/07350015.2014.926818
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jbes*.

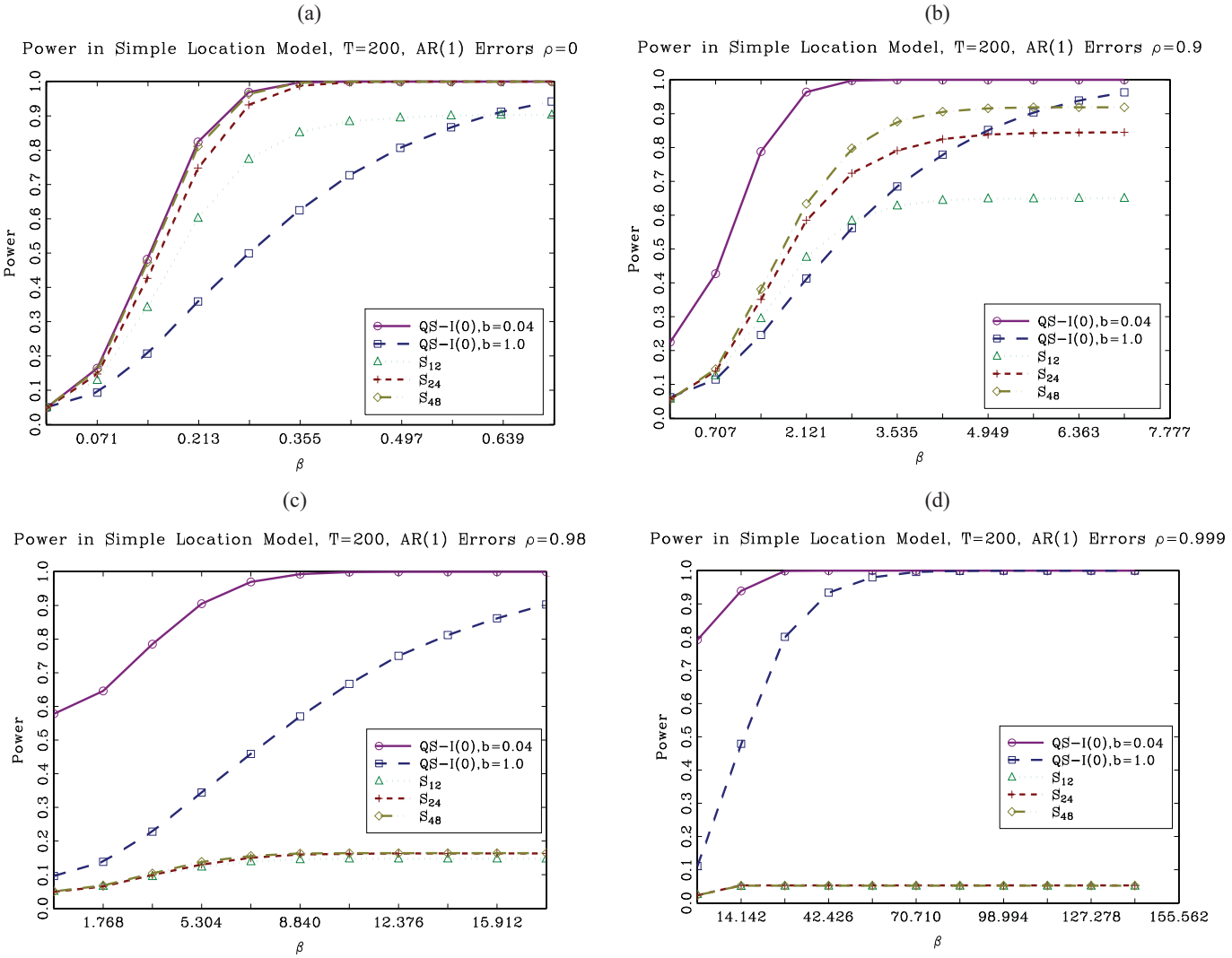

Figure 1. Power of S_q and QS tests. Stationary ($I(0)$) fixed-b critical values used for QS tests with no prewhitening.

2. FINITE SAMPLE PROPERTIES

To conserve space, the same notation used by Müller is adopted here and definitions of the various test statistics can be found there. The data-generating process is given by

$$y_t = \beta + u_t, \quad t = 1, 2, \ldots, T,$$
$$u_t = \rho u_{t-1} + \varepsilon_t, \quad u_0 = 0, \varepsilon_0 = 0,$$
$$\varepsilon_t \sim \text{iid}N(0, 1).$$

Following Müller, consider the following two-sided hypothesis

$$H_0 : \beta = 0, \quad H_1 : \beta \neq 0.$$

Results are reported for $\rho = 0.0, 0.9, 0.98, 0.999$ using 10,000 replications, and $T = 200$ is used in all cases. The nominal level is 5%. Comparisons are made between (i) three configurations of the new tests: S_{12}, S_{24}, and S_{48} and (ii) the ordinary least squares (OLS) t-statistic for β based on the quadratic spectral (QS) nonparametric kernel long run variance estimator (Andrews 1991).

The QS t-statistic was considered by Müller in his simulation study where it was found that the QS test quickly exhibits large over-rejection problems as ρ moves away from 0 and toward 1. Two features of Müller's implementation of the QS statistic preclude the possibility of more robust inference as ρ approaches 1. First, the bandwidth was chosen using the data-dependent rule proposed by Andrews (1991). The Andrew's formula tends to pick relatively small bandwidths for the QS kernel even when autocorrelation is strong. Second, standard normal critical values were used to carry out rejections.

Here, I implement the QS test using two bandwidth choices, one small and one very large. Rejections are calculated using the fixed-b critical values proposed by Kiefer and Vogelsang (2005). Referring to formula (4) of Müller, S_T denotes the bandwidth of the long run variance estimator. Results are reported here for the QS test using $S_T = 8$ (small) and $S_T = 200$ (very large). In the fixed-b framework, these bandwidths map to bandwidth-sample-size ratios, $b = S_T/T$, of $b = 8/200 = 0.04$ and $b = 200/200 = 1$, respectively. The corresponding fixed-b critical values are 2.115 ($b = 0.04$) and 12.241 ($b = 1$). The use of a very large bandwidth with the fixed-b critical value

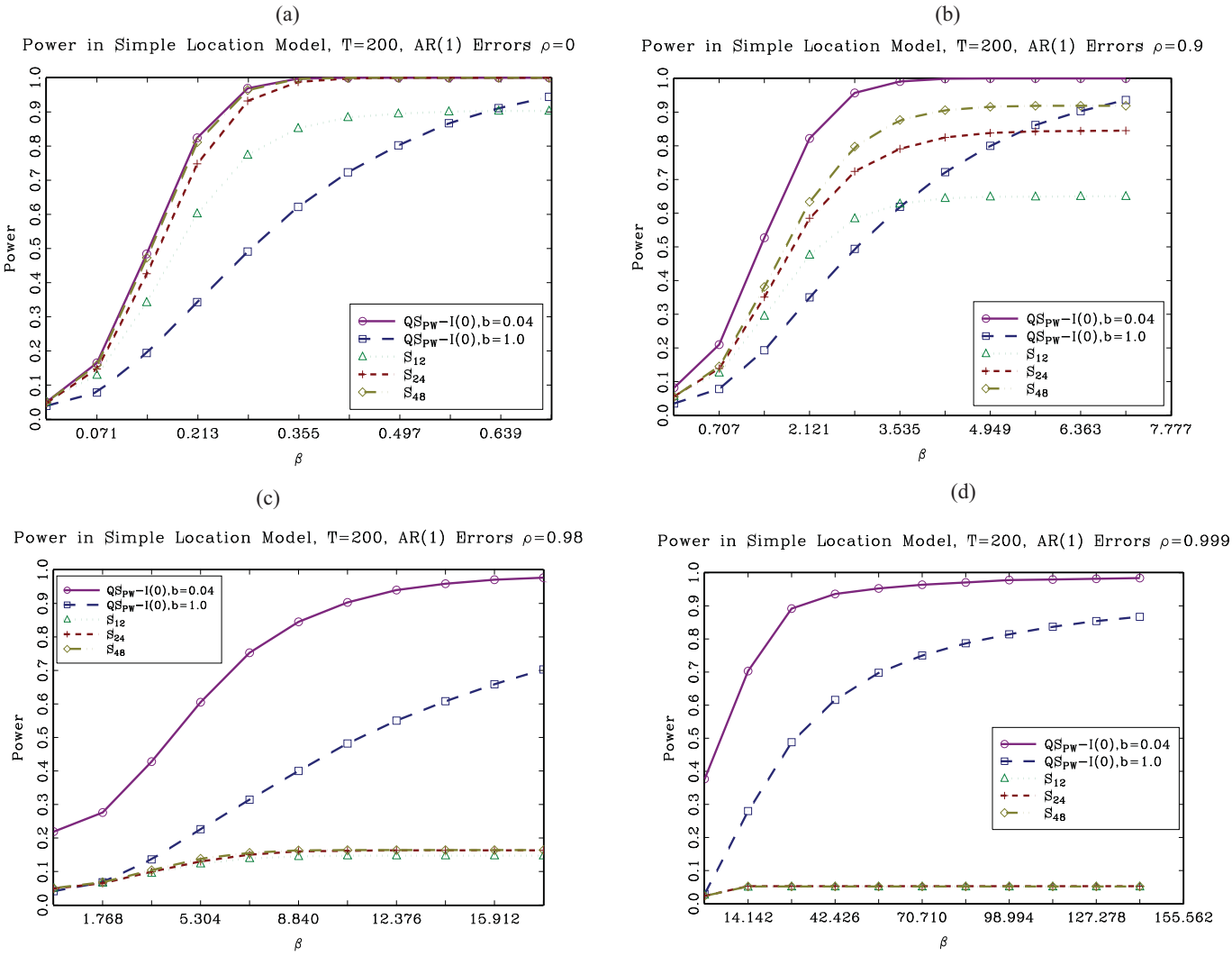

Figure 2. Power of S_q and QS tests. Stationary ($I(0)$) fixed-b critical values used for QS tests with $AR(1)$ prewhitening.

can greatly improve the robustness properties of the QS test when autocorrelation is strong relative to the use of a small bandwidth.

Figure 1(a)–1(d) reports power plots of the S_q and QS statistics. Power is not size-adjusted and therefore rejections for the case of $\beta = 0$ represent null rejection probabilities. Because power is not size-adjusted, we can explicitly see the practical trade-off between robustness to over-rejections under the null and power under the alternative.

When $\rho = 0$ (Figure 1(a)), all tests have empirical null rejections close to 0.05. Power of the S_q tests is increasing in q. Power of the QS tests is high with the small bandwidth ($b = 0.04$) but is substantially lower with the large bandwidth ($b = 1.0$). When ρ is close to 1, we see in Figure 1(b)–1(d) that empirical null rejections of the small bandwidth QS test are above 0.05 and substantially so when ρ is very close to 1. These large over-rejections were also reported by Müller for case where the data-dependent bandwidth was used for QS. In contrast, null rejections for the large bandwidth QS test are much closer to 0.05, and only for $\rho = 0.98, 0.999$ do we begin to see some mild over-rejections. It is important to keep in mind that if the standard normal critical value had been used instead of the fixed-b critical value, rejections would be substantially above 0.05. While the large bandwidth QS test is relatively robust when autocorrelation is strong, the S_q tests are more robust with null rejections very close to 0.05.

Now consider power. When $\rho = 0.9$, power of the S_q tests is higher than the large bandwidth QS test for alternatives close to the null but is lower for alternatives far from the null. For the cases with ρ very close to 1, the large bandwidth QS test has substantially higher power than the S_q tests. This shows that in some cases the remarkable robustness of the S_q statistics to strong autocorrelation comes at a high price with respect to power. *If* an empirical practitioner is willing to accept a small amount of over-rejection under the null, large gains in power are possible with the large bandwidth QS test.

Suppose one implements the QS tests using a prewhitened version of the nonparametric kernel long run variance estimator following Andrews and Monahan (1992). Figure 2(a)–2(d) reports power plots where QS is implemented with AR(1) prewhitening. As before, results are reported for the small and

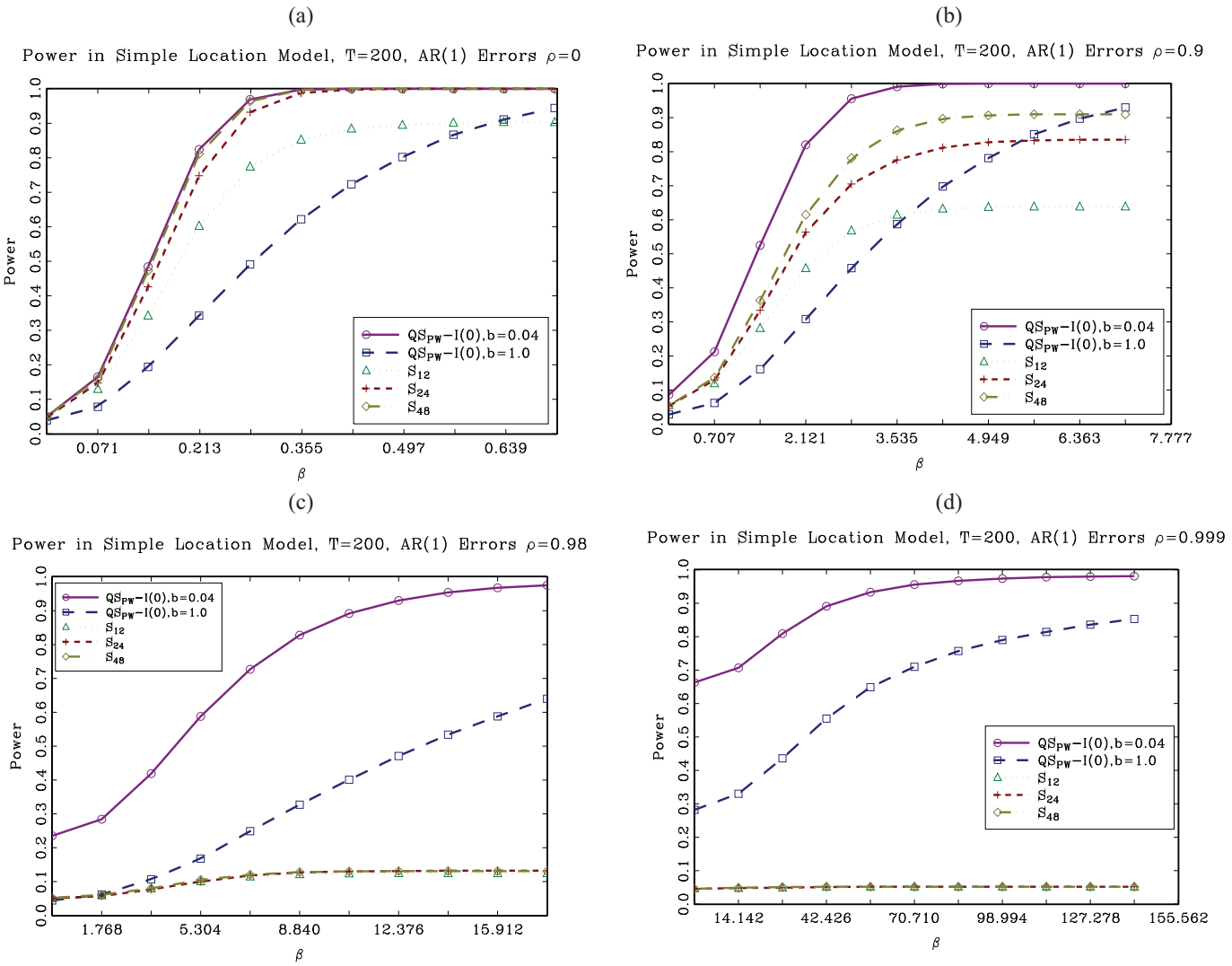

Figure 3. Power of same tests as in Figure 2 but with initial value $u_1 \sim N(0;(1-\rho^2)^{-1})$ in place of $u_1 = 0$.

very large bandwidth QS test with fixed-b critical values used to compute rejections. Prewhitening only partially reduces the over-rejection problems of the small bandwidth QS test. For the large bandwidth QS test, over-rejections are gone even when $\rho = 0.999$ and power is not adversely affected.

These simulation results suggest that judicious choice of bandwidth (including potentially very large ones) for the QS statistic and use of fixed-b critical values can deliver tests that are similarly robust to strong autocorrelation as the S_q statistics while delivering substantially more power when autocorrelation is strong. The challenge is to develop a bandwidth rule that tends to pick small bandwidths when autocorrelation is weak but very large bandwidths when autocorrelation is strong. The data-dependent bandwidth developed by Sun, Phillips, and Jin (2008), which seeks to balance the over-rejection problem with power, does tend to choose larger bandwidths than the Andrews (1991) approach, but the Sun, Phillips, and Jin (2008) bandwidth rule does not tend to pick bandwidths large enough to give robustness when autocorrelation is very strong. It would be interesting to see if the Sun, Phillips, and Jin (2008) bandwidth rule could be modified to choose very large bandwidths when autocorrelation is strong while still choosing small bandwidths when autocorrelation is weak.

Some readers might be wondering why a very large bandwidth is the key to achieving robustness for the QS test when autocorrelation is strong. One way to see this intuitively is to examine the approximate bias of the long run variance estimator given by Equation (4) of Müller. Combining results from the traditional spectral analysis literature and the recent fixed-b literature, one can approximate the bias in $\widehat{\omega}^2_{k,S_T}$ for the QS kernel as

$$\text{bias}\left(\widehat{\omega}^2_{k,S_T}\right) \approx A(b) + B(b), \quad b = S_T/T,$$

where

$$A(b) = -\frac{18}{125}\pi^2\left(\frac{1}{bT}\right)^2 \sum_{j=-\infty}^{\infty} j^2\gamma(j),$$

$$B(b) = -\omega^2\left[\left(\frac{5}{2\pi}\int_0^{\frac{6\pi}{5b}} \frac{\sin(x)}{x}dx\right)b - \frac{25}{6\pi^2} \times \left(1 - 2\cos\left(\frac{6\pi}{5b}\right)\right)b^2 + \left(\frac{125}{72\pi^3}\sin\left(\frac{6\pi}{5b}\right)\right)b^3\right].$$

The first term in the bias, $A(b)$, can be found in Andrews (1991) whereas the second term, $B(b)$, arises because residuals are being used to compute $\widehat{\omega}^2_{k,S_T}$ and was calculated by Hashimzade et al. (2005) using fixed-b theory.

When a small bandwidth is used, $A(b)$ dominates and the bias is negative (downward) for positively autocorrelated data. As the autocorrelation becomes stronger, $A(b)$ becomes larger in magnitude and the downward bias becomes more pronounced leading to the over-rejections seen in the figures. In contrast, when a large bandwidth is used, $A(b)$ becomes small or even negligible and $B(b)$ dominates the bias. The sign of $B(b)$ is always negative and its magnitude is increasing in the bandwidth (increasing in b). Whereas $A(b)$ is not captured by traditional or fixed-b asymptotic theory, $B(b)$ is implicitly captured by the fixed-b limit. Using a large bandwidth minimizes $A(b)$ while maximizing $B(b)$ but fixed-b critical values correct the impact of $B(b)$ on the t-test, and robustness is achieved. If fixed-b critical values are not used to correct the large downward bias induced by $B(b)$, substantial over-rejections would be obtained with large bandwidths.

While the QS approach is promising relative to the S_q in the simple location model, the QS approach does have some drawbacks. Suppose we change the initial condition of $\{u_t\}$ in the data-generating process to

$$u_1 \sim N(0, (1-\rho^2)^{-1}).$$

Figure 3(a)–3(d) shows power plots for this case. For the cases of $\rho = 0.0, 0.9, 0.98$ the change in the initial value has little or no effect on the null rejection probabilities or power of the prewhitened QS statistics or the S_q statistics. However, for $\rho = 0.999$ the large bandwidth QS test now shows nontrivial over-rejections under the null hypothesis. In contrast, the S_q statistics are unaffected by the change in initial condition. This sensitivity of one class of statistics and the relative insensitivity of another class of statistics to the initial value underscores the conclusion in Müller where it is stated that "researchers in the field have to judge which set of regularity conditions makes the most sense for a specific problem." Robust inference when autocorrelation is strong is not easy.

REFERENCES

Andrews, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–854. [335,337]

Andrews, D. W. K., and Monahan, J. C. (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–966. [336]

Hashimzade, N., Kiefer, N. M., and Vogelsang, T. J. (2005), "Moments of HAC Robust Covariance Matrix Estimators Under Fixed-b Asymptotics," Working Paper, Department of Economics, Cornell University. [338]

Kiefer, N. M., and Vogelsang, T. J. (2005), "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 21, 1130–1164. [334,335]

Müller, U. K. (2014), "HAC Corrections for Strongly Autocorrelated Time Series," *Journal of Business and Economic Statistics*, 32, 311–322. [334]

Sun, Y., Phillips, P. C. B., and Jin, S. (2008), "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing," *Econometrica*, 76, 175–194. [337]

Rejoinder

Ulrich K. Müller

Department of Economics, Princeton University, Princeton, NJ, 08544 (*umueller@princeton.edu*)

I would like to start by expressing my sincere gratitude for the time and effort the reviewers spent on their thoughtful and constructive comments. It is a rare opportunity to have one's work publicly examined by leading scholars in the field. I will focus on three issues raised by the reviews: the role of the initial condition, applications to predictive regressions, and the relationship to Bayesian inference.

The role of the initial condition under little mean reversion: Vogelsang, Sun, and Cattaneo and Crump all suggest alternative inference procedures for strongly autocorrelated series. Their simulations show that these procedures are substantially more powerful than the S_q tests, especially for distant alternatives under strong autocorrelation, while coming quite close to controlling size. This is surprising, given that the S_q tests are designed to maximize a weighted average power criterion that puts nonnegligible weight on such alternatives.

Recall that the S_q tests are constructed to control size under any stationary AR(1), including values of ρ arbitrarily close to one. Importantly, the initial condition is drawn from the unconditional distribution. Figure 1 plots four realizations of a mean-zero Gaussian AR(1). For all values of ρ close to one, the series are almost indistinguishable from a random walk, with the initial condition diverging as $\rho \to 1$. Yet all of these series are perfectly plausible realizations for a stationary *mean-zero* AR(1). As $\rho \to 1$, the sample mean is very far from the population mean relative to the in-sample variation. A test that controls size for all values of $\rho < 1$ must not systematically reject the null hypothesis of a zero mean for such series. But the new tests of Sun and Vogelsang, and the τ and $\tau^{1/2}$ tests of Cattaneo and Crump all do so for sufficiently large $\rho < 1$, leading to arbitrarily large size distortions in this model (see the discussion in Section 6 of the article).

In contrast, the S_q tests do not overreject even as $\rho \to 1$. The S_q tests are usefully thought of as joint tests of whether a series is mean reverting, and whether the long-run mean equals the hypothesized value. The power of any test of this joint problem is clearly bounded above by the power of a test that solely

Journal of Business & Economic Statistics
July 2014, Vol. 32, No. 3
DOI: 10.1080/07350015.2014.931769

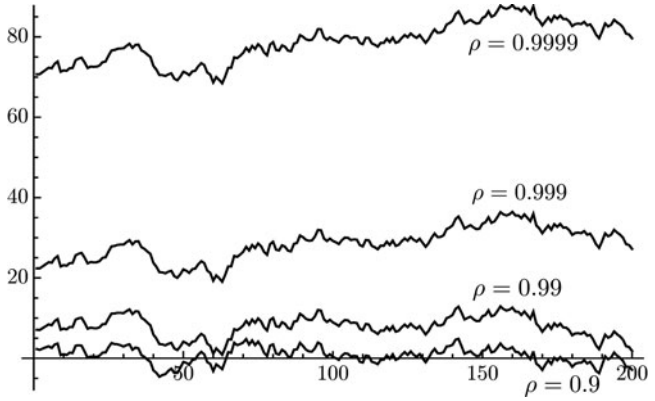

Figure 1. Realizations of a stationary AR(1). Notes: Realizations of a mean-zero stationary AR(1) $y_t = \rho y_{t-1} + \varepsilon_t$, $t = 1, \ldots, 200$ for various ρ and identical draws of $\varepsilon_t \sim$ iid$\mathcal{N}(0, 1)$, and initial condition fixed at 1.0 unconditional standard deviations, $y_1 = 1/\sqrt{1-\rho^2}$.

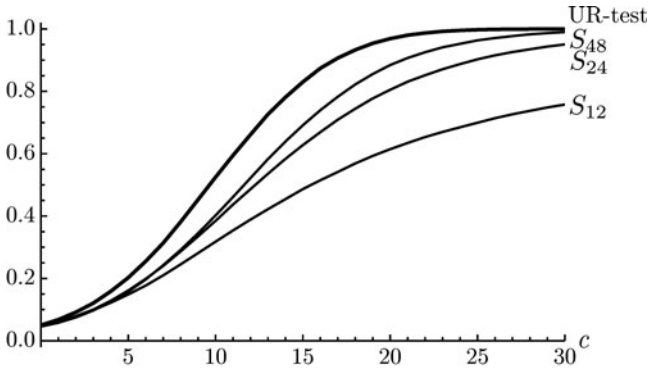

Figure 2. Comparison of power of unit root tests with power of S_q-tests against distant alternatives. Notes: Local asymptotic power in the AR(1) model with $\rho = \rho_T = 1 - c/T$ of 5% level S_q tests against distant alternatives (where $Y_0 = B\sqrt{q^{-1}\sum_{l=1}^{q} Y_l^2}$ from Step 2), and asymptotic power envelope of 5% level translation invariant unit root tests.

focuses on testing the null hypothesis of a unit root. Figure 2 shows that the power of the S_q tests against distant alternatives is not very far from the power envelope of a translation invariant unit root test (see Elliott 1999).[1] The S_q tests thus come close to performing as well as possible in the stationary AR(1) model that allows for values of ρ arbitrarily close to one.

To insist on size control in this comprehensive fashion may be viewed as too restrictive. In their simulations or derivations, Cattaneo and Crump, Sun and Vogelsang all consider a model where the initial condition is not drawn from the unconditional distribution, but rather $y_0 = \mu + \varepsilon_0$. Under this assumption, the realizations in Figure 1 with $\rho > 0.99$ are entirely implausible under $\mu = 0$, and good tests should reject the corresponding hypothesis. Substantively, though, I find it difficult to think of times series that are much closer to the population mean at the beginning of the sample compared to, say, the middle of the sample. If anything, the start of the sampling period is often chosen just after some incisive event, such as a war, which rather suggests a model where the initial condition is even more disperse than the unconditional distribution.

Alternatively, one might assume some lower bound on the amount of mean reversion. With some nonnegligible mean reversion, the realizations in Figure 1 with $\rho > 0.99$ are again safely ruled out as stemming from a mean-zero process. In the local-to-unity parameterization $\rho = \rho_T = 1 - c/T$, a certain degree of mean reversion is imposed by a lower bound on c. Cattaneo and Crump, for instance, set $c \geq 0.1$ in their derivation of the tests $\tau_{1/2}^*$ and τ_1^*. For a given value of the lower bound $\underline{c}$, it is fairly straightforward to adjust the numerical derivation of the tests in Section 4.2 to obtain nearly weighted average power maximizing tests that controls size only under $c \geq \underline{c}$. Using the same weighting function that underlies the derivation of the S_q tests and setting $\underline{c} = 1$, for instance, yields tests with much larger power for moderate and small values of c. Figure 3 depicts these gains. The left panel shows weighted average power of tests that only impose size control for $c \geq 1$.[2] The right panel replicates Figure 3 of the article for comparison and shows weighted average power of tests that remain valid for any $c > 0$. Unreported results show that these gains translate into substantially improved small sample performance in the stationary AR(1) model with $T = 200$ as long as $\rho \leq 0.995$, while inducing size distortions of roughly 23% under $\rho = 0.999$.

Does it make sense to impose a lower bound on the degree of mean reversion, and if so, how should this bound be chosen? The answer surely depends on the specifics of the application. It is an appealing quality of the S_q tests that their validity does not require such a judgment. At the same time, as underlined by the comments and the above calculations, this robustness comes at a substantial cost in terms of power. It is useful to provide practitioners with a menu of tests, with a range of robustness and efficiency properties, as also suggested by Sun and Vogelsang. In the derivation of this menu, however, I think it is important to be as explicit as possible about the imposed conditions.

Predictive regressions: Cattaneo and Crump apply the S_q tests to some standard long-horizon predictive regressions and find very little evidence of significant predictive relationships. In contrast, inference based on a standard Newey-West estimator with lag length equal to the horizon h often indicates significant predictive power.

While entirely standard, this Newey-West correction lacks sound theoretical foundation. Suppose the only source of autocorrelation is the partial overlap of the long-horizon returns, so that the errors follow an MA($h-1$). The long-run variance ω^2 is then given by $\omega^2 = \sum_{j=-(h-1)}^{h-1} \gamma(j)$. The Newey-West estimator with bandwidth equal to h equals $\hat{\omega}_{\text{NW}}^2 = \sum_{j=-(h-1)}^{h-1} \frac{|h-j|}{h} \hat{\gamma}(j)$. By standard arguments, for any fixed j, the sample autocovariances $\hat{\gamma}(j)$ are consistent for the population autocovariances, $\hat{\gamma}(j) \xrightarrow{p} \gamma(j)$, so that $\hat{\omega}_{\text{NW}}^2 \xrightarrow{p} \sum_{j=-(h-1)}^{h-1} \frac{|h-j|}{h} \gamma(j)$. This

[1] Unreported results show these differences to become even smaller relative to the low-frequency unit root tests derived in Müller and Watson (2008), which are based on the same cosine transforms as S_q.

[2] The approximate least favorable distribution under $c \geq 1$ puts a lot of weight on $c = 1$. Under the null hypothesis and $c = 1$, the unconditional variance of y_1 is approximately $\sigma^2 T/(1-\rho_T^2) \approx \sigma^2 T/(2c) = \sigma^2 T/2$. This value corresponds more or less to the variance of $y_1 + \mu$ under the alternative $\mu \sim \mathcal{N}(0, 10T\sigma^2/c^2)$ under $c \approx 5$, explaining the dip in weighted average power.

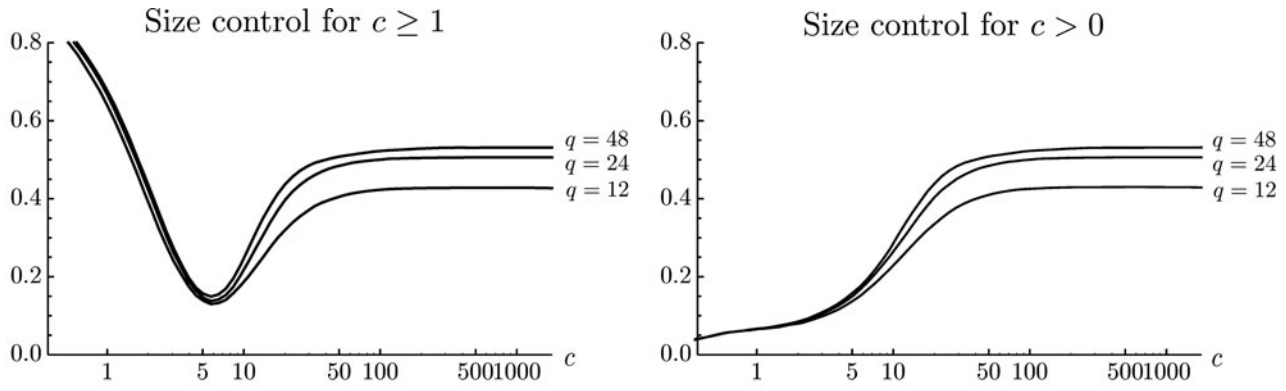

Figure 3. Gains in power from restricting $c \geq 1$. Notes: Asymptotic weighted average power under a $\mathcal{N}(0, 10T/c^2)$ weighting on the difference between population and hypothesized mean in an AR(1) model with unit innovation variance and coefficient $\rho = \rho_T = 1 - c/T$ of weighted average power maximizing 5% level hypothesis tests, based on q low-frequency cosine averages. Tests on left panel control size for $c \geq 1$, and tests on right panel control size for $c > 0$.

probability limit is not generically equal to ω^2. In fact, if $\gamma(j) > 0$ for $|j| < h$, as one would expect for an MA($h-1$) induced by partial overlaps, $\hat{\omega}^2_{\text{NW}}$ is consistent for a value that is strictly smaller than ω^2. Inference based on $\hat{\omega}^2_{\text{NW}}$ thus overstates significance, even in large samples. The asymptotic validity of inference based on the Newey-West estimator rather stems from a promise to increase the bandwidth without bound, so that all sample autocovariances $\hat{\gamma}(j)$ for fixed j eventually receive a weight arbitrarily close to one.

It is hence interesting to consider alternative HAC corrections for such regressions, including those based on inconsistent LRV estimators reviewed in Section 3 of the article, or the S_q tests. One serious issue, however, is that the regressor of interest, such as the dividend yield, is very persistent. Recall from Section 5 that the alternative approximations to test statistics generated by inconsistent long-run variance estimators, and also the S_q tests, require the partial sum of the regressors to be roughly linear ($T^{-1}\sum_{t=1}^{rT} X_t X_t' \approx r\,\Sigma_X$ for $0 \leq r \leq 1$ and some Σ_X). Strong persistence of the regressor renders this a poor approximation.[3] In contrast, the Ibragimov-Müller approach does not depend on this time homogeneity of the design matrix, making it perhaps a more attractive choice for this type of regression problem.

Bayesian inference: Kiefer raises interesting points about appropriate conditioning and a Bayesian perspective. My 2012 paper with Andriy Norets explores similar questions in other nonstandard problems. As others before, we conclude that a description of parameter uncertainty with confidence intervals is generically "unreasonable" conditional on some data draws unless it can be rationalized as a Bayesian posterior set of at least the same level relative to some prior.

In the context of HAC estimation, consider first the Student-t intervals computed from (7). These are rationalizable by a Bayesian, at least approximately: Consider a Gaussian model for y_t with population mean $E[y_t] = \mu$ and spectral density that is flat and equal to $\omega^2/2\pi$ on the interval $[0, \pi q/T]$. Under the (analogue of) the DFT approximation (3), $\hat{\mu} \sim \mathcal{N}(\mu, \omega^2/T)$ and $Y_l \sim \mathcal{N}(0, \omega^2)$, $l = 1, \ldots, q$ are independent and independent of $\{Y_l\}_{l=q+1}^{T}$. Thus, with a prior on the remaining part of the spectral density on the interval $[\pi q/T, \pi]$ that is independent of (ω, μ), the low-frequency information factorizes in both the prior and the likelihood, and the problem reduces to Bayesian inference about μ based on $q+1$ Gaussian observations. With the usual uninformative prior on (μ, ω^2) proportional to $1/\omega^2$, one therefore recovers a Student-t posterior of $\sqrt{T}(\hat{\mu} - \mu)/\hat{\omega}_{Y,q}$ with q degrees of freedom (see, e.g., chap. 3.2 in Gelman et al. 2004), leading to the same description of uncertainty as those derived from frequentist arguments explored in Section 3.

Similar arguments also rationalize the confidence intervals generated by the Ibragimov-Müller approach, at least in a limited information sense with $\hat{\mu}_j \sim \mathcal{N}(\mu, q\omega^2/T)$ the only available sample information.

For S_q, however, the possibility of a Bayesian rationalization is much less clear. The roughly uniform prior on $\log c$ in the weighting function approximates the usual uninformative prior a scale for parameter. This makes sense, since c plays a role akin to a scale parameter, at least for c large. But the denominator of the LR statistic derived in Section 4.2 entails an entirely different, numerically determined approximate least favorable distribution for c. The LR statistic, viewed as a function of μ_0, hence does not map out some posterior density. Müller and Norets (2012) make some concrete suggestions how to ensure that confidence intervals remain reasonable descriptions of uncertainty also conditional on the data, and it would be interesting to incorporate those into the derivations of the S_q tests. I leave this to future research.

[3]Once this approximate linearity fails the implementation details of the S_q tests in a regression context become potentially important; for instance, a variant of (15) such as $y_t = \hat{y}_t + \hat{\beta}_1 - \beta_{1,0}$ may lead to substantially different empirical results.

REFERENCES

Elliott, G. (1999), "Efficient Tests for a Unit Root When the Initial Observation is Drawn From its Unconditional Distribution," *International Economic Review*, 40, 767–783. [339]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC. [340]

Müller, U. K., and Norets, A. (2012), "Credibility of Confidence Sets in Nonstandard Econometric Problems," Working Paper, Economics Department, Princeton University. [340]

Müller, U. K., and Watson, M. W. (2008), "Testing Models of Low-Frequency Variability," *Econometrica*, 76, 979–1016. [339]