

FINITE-SAMPLE SIZE CONTROL OF IVX-BASED TESTS IN PREDICTIVE REGRESSIONS*

MEHDI HOSSEINKOUCHACK
University of Mannheim

MATEI DEMETRESCU
Christian-Albrechts-University of Kiel

In predictive regressions with variables of unknown persistence, the use of extended IV (IVX) instruments leads to asymptotically valid inference. Under highly persistent regressors, the standard normal or chi-squared limiting distributions for the usual t and Wald statistics may, however, differ markedly from the actual finite-sample distributions which exhibit in particular noncentrality. Convergence to the limiting distributions is shown to occur at a rate depending on the choice of the IVX tuning parameters and can be very slow in practice. A characterization of the leading higher-order terms of the t statistic is provided for the simple regression case, which motivates finite-sample corrections. Monte Carlo simulations confirm the usefulness of the proposed methods.

1. INTRODUCTION

A common inferential task of practical relevance is to decide whether a potential predictor variable does indeed forecast another variable of interest. In the simplest setup, practitioners thus test the null hypothesis of no predictability in the model

$$y_t = \mu + \beta x_{t-1} + u_t, \quad t = 2, \dots, T, \quad (1)$$

where the regressor is usually assumed to have an autoregressive structure,

$$x_t = \rho x_{t-1} + v_t, \quad (2)$$

with initial condition bounded in probability, $x_1 = O_p(1)$. With financial data, predictors such as dividend yields or earnings-price ratios are often quite persistent, even if still mean-reverting (typically captured by a value of ρ close to unity), and its shocks are contemporaneously correlated with the variable to be predicted (see Phillips, 2015, for a recent review). This biases the OLS estimator of the slope

*The authors would like to thank three anonymous referees, the Co-Editor (Giuseppe Cavaliere), as well as Karim Abadir and Uwe Hassler for very useful comments. Demetrescu gratefully acknowledges the support of the German Research Foundation (DFG) through the project DE 1617/4-2. Address correspondence to Mehdi Hosseinkouchack, University of Mannheim, L7, 3-5, 68161 Mannheim, Germany; e-mail: hosseinkouchack@uni-mannheim.de.

parameter and induces heavy non-normality of t statistics (Elliott and Stock, 1994; Stambaugh, 1999) such that tests for predictability are size-distorted.

Near to unity asymptotics, obtained by letting $\rho = 1 - c/T$, offer a better approximation of the actual distribution of the OLS t statistic than the standard normal in this situation; cf. Elliott and Stock (1994). The limiting distribution of the OLS estimator and test is explicitly non-normal under near integration and depends on the mean-reversion parameter c and the correlation between u_t and v_t . Since consistent estimation of c is not possible in such highly persistent cases (Phillips, 1987), the literature suggested several different ways of circumventing the lack of knowledge on ρ . See, among others, Campbell and Yogo (2006), Jansson and Moreira (2006), Maynard and Shimotsu (2009), Camponovo (2015), Phillips (2015), and Breitung and Demetrescu (2015).

Building on the work of Phillips and Magdalinos (2007) and Magdalinos and Phillips (2009), the extended IV (IVX) estimation and testing approach introduced by Phillips and Magdalinos (2009) is gaining momentum for predictive regressions; see, for example, Gonzalo and Pitarakis (2012), Phillips and Lee (2013), Kostakis, Magdalinos, and Stamatogiannis (2015), Demetrescu and Rodrigues (2016), Demetrescu et al. (2020), or Yang et al. (2020). In the IVX framework, x_{t-1} is instrumented by the specifically constructed instrumental variable $z_{t-1} = (1 - \varrho L)_+^{-1} \Delta x_{t-1} = \sum_{j=0}^{t-2} \varrho^j \Delta x_{t-1-j}$ with initial condition $z_1 = 0$ and $\varrho = 1 - a/T^\eta$, where $a > 0$ and $\eta \in (0, 1)$. This “endogenous instrumentation” method has convenient properties: the persistence of z_t is under control, and is below that of the near-integrated x_{t-1} . Regularity conditions assumed, the resulting IV estimator follows a mixed Gaussian distribution in the limit, and the limiting null distribution of the corresponding t ratio is standard normal.

Should the regressor x_t be highly persistent with localization parameter c close to zero, the IVX-based test of no predictability may still be seriously distorted in finite samples, even if less so than the OLS-based test. This is clearly the case when choosing η too close to unity or a too close to zero, and the difference in terms of persistence between the instrument z_t and the regressor x_t becomes small: for example, the rule of thumb proposed by Kostakis et al. (2015), which sets $\varrho = 1 - 1/T^{0.95}$, is actually equivalent to a near unit root with localizing coefficient \tilde{c} between 1 and 2 for sample sizes between $T = 100$ and $T = 10,000$. Kostakis et al. (2015), therefore, recommend the use of a finite-sample correction leading to reliable size control for two-sided tests.¹ We find, however, that the correction is not equally effective for tests against one-sided alternatives. This is relevant in practice, as economic theory often predicts a certain sign of the slope coefficient β .

¹In the more general context of IVX predictive regressions with serially correlated errors, Yang et al. (2020) employ autoregressive prewhitening to eliminate the second-order bias of the IVX estimator induced by correlated regression errors, whereas Phillips and Magdalinos (2009) use nonparametric estimates of one-sided long-run covariance matrices to correct for this bias.

We, therefore, examine in Section 2 the behavior of components of the t statistic that vanish in the limit, but still have an effect in finite samples. We do so in a setup allowing for deterministically varying variances and correlations of the errors u_t and v_t . Since the main source of distortions in finite samples appears to be the fact that the finite-sample distribution is not centered at zero (see also Stambaugh, 1999), we focus on correcting for the noncentrality of the t ratio. One way of doing so is to resort to backward and forward demeaning of the involved variables. In time series analysis, backward (or recursive, or adaptive) demeaning can be traced back to at least the work of So and Shin (1999) where recursive demeaning is shown to reduce bias in estimators of large autoregressive roots. Specifically for (panel) predictive regressions, Westerlund, Karabiyik, and Narayan (2017) resorts to forward and backward demeaning to reduce endogeneity bias. While this is shown to stabilize size, we also find that it has the side effect of reducing power in a nontrivial manner. This is a specific effect of forward demeaning in the context of persistent predictors, and not of IVX. Therefore, we discuss the use of direct approximations of the higher-order terms affecting the finite-sample behavior of the t statistic. Some depend on the localizing coefficient c , which cannot be consistently estimated, so we provide a method of side-stepping this issue. In extensive Monte Carlo experiments (see Section 3), we find it to work reasonably well under various patterns of changing error variances.

The technical details of the proofs can be found in the Appendix and in an Online Supplement, which also contains additional simulation results pertaining to conditional heteroskedasticity.

2. IMPROVED IVX INFERENCE

2.1. Preliminaries

Let us first specify the details of the predictive regression model we work with.

Assumption 1. *The data $\{y_t, x_t\}$, $t = 2, \dots, T$, are generated from (1) and (2) with $x_1 = O_p(1)$.*

To keep a realistic setup, we allow for error heterogeneity in the form of time-varying variances and correlations, as well as short-run dynamics. Specifically, we work under the following assumptions.

Assumption 2. *Let $v_t = \Psi(L)v_t$, where v_t is heteroskedastic white noise as specified below and $\Psi(L)$ is a lag polynomial, $\Psi(L) = \sum_{j \geq 0} b_j L^j$ with $\sum_{j \geq 0} j |b_j| < \infty$ and $\psi = \Psi(1) \neq 0$.*

Assumption 3. *Let $(u_t, v_t)' = \mathbf{H}(\frac{t}{T})(\xi_{1t}, \xi_{2t})'$, with $(\xi_{1t}, \xi_{2t})' \sim iid(0, \mathbf{I}_2)$, L_r -bounded for some $r > 4$, where $\mathbf{H}(\cdot) := \{h_{ij}\}_{i,j=1,2}$ is a matrix of piecewise Lipschitz-continuous bounded functions on $(-\infty, 1]$, which is of full rank at all but a finite number of points.*

This would be a typical structure in predictive regressions for stock returns, where the disturbance u_t is not predictable using the past of v_t . We do not assume a particular distribution for the errors but only require finite fourth-order moments. Although daily returns may exhibit fat tails, standard predictive regression models are used in conjunction with monthly, quarterly, or even annual data, where infinite kurtosis is not an issue. For the same reason, the serial independence assumption we make on the innovations is justifiable. The 1-summability condition placed on the coefficients of the filter Ψ is standard in the literature involving integrated and near-integrated variables. Let

$$\Sigma_t = \mathbf{H}(t/T) \mathbf{H}'(t/T) = \begin{pmatrix} \sigma_u^2(t/T) & \sigma_{uv}(t/T) \\ \sigma_{uv}(t/T) & \sigma_v^2(t/T) \end{pmatrix}$$

and notice that we have time varying variances, covariances, and correlations of the errors, as $\text{Cov}((u_t, v_t)') = \Sigma_t$ is not restricted beyond piecewise smoothness. The off-diagonal elements of Σ_t are not required to be zero, thereby allowing for predictive regression endogeneity. The assumption on $\mathbf{H}(s)$ allows for a wide range of covariance matrices of the innovations, including, for example, single or multiple (co-) variance shifts, smooth transition (co-) variance shifts, or even trending variances.

With \mathbf{W} a vector of two independent standard Wiener processes and “ \Rightarrow ” denoting weak convergence of probability measures on the space of càdlàg real functions on $[0, 1]$ equipped with the Skorokhod topology, we have (see Cavaliere, Rahbek, and Taylor, 2010)

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[sT]} \begin{pmatrix} u_t \\ v_t \end{pmatrix} \Rightarrow \int_0^s \mathbf{H}(r) d\mathbf{W}(r) =: \begin{pmatrix} U_H(s) \\ V_H(s) \end{pmatrix};$$

the normalized levels of x_t converge weakly to a heteroskedastic Ornstein–Uhlenbeck type process,

$$T^{-1/2} x_{[sT]} \Rightarrow \psi \left(V_H(s) - c \int_0^s e^{-c(s-r)} V_H(r) dr \right) := \psi J_{c,H}(s).$$

IVX estimation relies on using the instrument $z_{t-1} = (1 - \varrho L)_+^{-1} \Delta x_{t-1} = \sum_{j=0}^{t-2} \Delta x_{t-1-j}$, where $\varrho = 1 - \frac{a}{T^\eta}$ with $a > 0$ and $\eta \in (0, 1)$. The IVX t statistic for the null $\beta = 0$ is

$$t_{vx} = \frac{\sum_{t=2}^T (z_{t-1} - \bar{z}) y_t}{\sqrt{\sum_{t=2}^T (z_{t-1} - \bar{z})^2 \hat{u}_t^2}}, \tag{3}$$

with Eicker–White standard errors to account for the heteroskedasticity. The residuals \hat{u}_t are computed using the OLS estimator of β , as is common in the predictive IVX regression literature.²

²The IVX estimator has convergence rate $T^{1/2+\eta/2}$ so IVX residuals, although still consistent for $\eta > 0$, are typically less precise; see Kostakis et al. (2015).

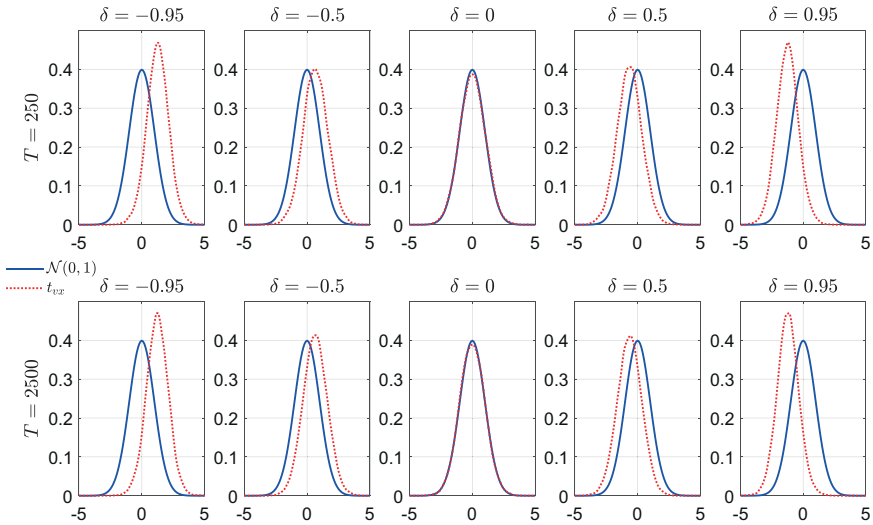


FIGURE 1. Finite sample null distribution of t_{vx} for $a = 1$; $\eta = 0.95$ versus $\mathcal{N}(0, 1)$; DGP: (1) with (2), $\rho = 1$, $(u_t, v_t) \sim iid\mathcal{N}(0, ((1, \delta); (\delta, 1)))$, 25,000 replications, different correlations δ , and sample sizes.

What makes the IVX approach interesting for practitioners is that the terms involving c vanish as $T \rightarrow \infty$ and pivotal inference on β can be obtained asymptotically. See Kostakis et al. (2015) for details on IVX-based predictive regression under strict stationarity of errors, and Demetrescu and Rodrigues (2016) for a case with time-varying variances with some (nontrivial) restrictions on the correlations. In finite samples, however, the actual distribution is not centered at zero because numerator and denominator correlate, and has a variance somewhat smaller than 1, as can be seen in Figure 1. Notice also the slow convergence to the standard normal.

2.2. Higher-Order Terms

We, therefore, study corrections that make IVX-based inference in predictive regressions even more reliable. To this end, we first characterize the leading terms of the IVX t statistic.

Proposition 1. Under Assumptions 1–3 and any $\eta \in (1/2, 1)$, it holds as $T \rightarrow \infty$ that

$$t_{vx} = Z_T + B_T + C_T + o_p(T^{\eta/2-1/2}),$$

where

$$Z_T \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{with} \quad E(Z_T) = 0,$$

$$B_T \rightarrow 0 \quad \text{with} \quad T^{1/2-\eta/2} E(B_T) \rightarrow -\frac{1}{\sqrt{2a}} \frac{\int_0^1 \sigma_{uv}(s) \sigma_u^2(s) \sigma_v^2(s) ds}{\sqrt{\left(\int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds\right)^3}},$$

and

$$T^{1/2-\eta/2} C_T \Rightarrow -\sqrt{\frac{2}{a}} \frac{U_H(1) J_{c,H}(1)}{\sqrt{\int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds}}.$$

Proof: See Appendix B.

Proposition 1 provides an explanation for the finite sample behavior of t_{vx} as observed in Figure 1 for $c = 0$. For instance, the direction and magnitude of the noncentrality depend on the average sign and magnitude of the correlation between the errors u_t and v_t via the two terms B_T and C_T . It can be seen from the discussion below that, under constant correlation δ , the magnitude of the noncentrality is in fact proportional to δ . Moreover, the slow convergence of t_{vx} to the standard normal seen in Figure 1 can also be explained by the behavior of the two terms B_T and C_T : although they do vanish, they do so at rate $T^{\eta/2-1/2}$, which is low whenever η close to unity.

The noncentrality is mainly driven by the two terms B_T and C_T . The first depends on the user-chosen parameters a and η (with the additional restriction $\eta > 1/2$ required for the calculation of $E(B_T)$), as well as on a particular form of average correlation. In fact, under homoskedasticity ($\mathbf{H} = const.$), the expectation of $T^{1/2-\eta/2} B_T$ is asymptotically equivalent to $-\delta/\sqrt{2a}$ with δ the constant correlation of u_t and v_t . If $\sigma_{uv}(s) = 0 \forall s \in [0, 1]$, B_T does not affect the centering of t_{vx} .

The same holds for the second component, C_T : if $\sigma_{uv}(s) = 0 \forall s \in [0, 1]$, then U_H and $B_{c,H}$ are independent; therefore, the expectation of the limit of the normalized C_T is zero as well. Should there, however, be contemporaneous correlation, the behavior of C_T —in particular its expectation—does depend on c . Moreover, the dependence is nonlinear, since it is easily shown that

$$E(U_H(1) J_{c,H}(1)) = \int_0^1 e^{-c(1-s)} \sigma_{uv}(s) ds.$$

As expected, the expectation decreases in magnitude as c increases. This expression simplifies too under homoskedasticity, where $T^{1/2-\eta/2} C_T$ has an asymptotic expectation depending on the (constant) correlation δ , namely $-\delta \sqrt{\frac{2}{a} \frac{1-e^{-c}}{c}}$. For $c = 0$, the case with the largest distortions, we see this expectation to be twice as large as that of the normalized B_T , with the relative importance of C_T diminishing as c increases. This component depends, however, on the localizing coefficient c which cannot be consistently estimated, unlike the expectation of B_T .

Figure 2 plots the contribution of both B_T and C_T to the noncentrality of the t statistic t_{vx} . We note that heteroskedasticity only has a secondary influence

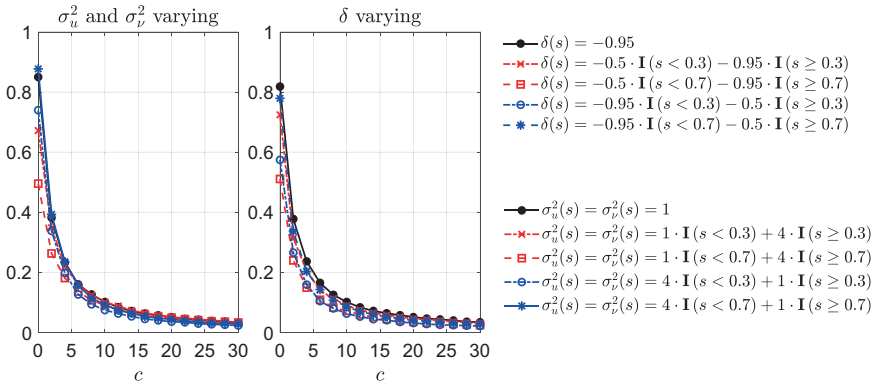


FIGURE 2. Finite sample average of $T^{1/2-n/2} (B_T + C_T)$ under the null of $\beta = 0$ as a function of c . DGP: (1) with (2) where $T = 250$, $a = 1$, $\eta = 0.95$, $\rho = 1 - c/T$, $(u_t, v_t) \sim iidN(0, \Sigma_t)$, 10,000 replications, breaks in variances or in the correlation. Left panel: $\Sigma_t = ((\sigma_u^2(t/T), \delta); (\delta, \sigma_v^2(t/T)))$, $\delta = -0.95$ (see the Section 3 for more details); right panel: $\Sigma_t = ((1, \delta(t/T)); (\delta(t/T), 1))$, $\delta(\cdot)$ switching between -0.5 and -0.95 .

compared to the localizing coefficient c , and that most (but not all) of the finite-sample noncentrality seen in Figure 1 for $c = 0$ is accounted for by the two terms.

2.3. Corrections

The term C_T from Proposition 1 appears because of the full-sample demeaning of the dependent variable (see the proof of Proposition 1 for details). To deal with this, we first discuss recursive demeaning as a possible correction. In particular, we use backward recursive demeaning for the regressor and forward demeaning for the dependent variable, in that we write

$$t_{vx}^{rec} = \frac{\sum_{t=2}^T (z_{t-1} - \bar{z}_{t-1})(y_t - \ddot{y}_t)}{\sqrt{\sum_{t=2}^T (z_{t-1} - \bar{z}_{t-1})^2 \hat{u}_t^2}}, \quad \text{with } \ddot{y}_t = \frac{1}{T-t+1} \sum_{j=t}^T y_j \quad \text{and } \bar{z}_t = \frac{1}{t} \sum_{j=1}^t z_j.$$

The motivation for such demeaning schemes is that the recursively demeaned regressor and the forward demeaned disturbance are now orthogonal irrespective of the correlation between u_t and v_t , which is not the case with usual demeaning. Such orthogonal schemes of mean adjustment have been used before in predictive regressions: for example, Westerlund et al. (2017) uses such a scheme to develop a predictability test in panel predictive regression.³ In fact, in the panel literature, forward and backward demeaning have a much longer history in dealing with the Nickell bias (Nickell, 1981); see Everaert (2013) for a recent contribution.

³Moreover, recursive adjustment schemes are popular in the related unit root testing literature as a means for reducing noncentrality and improving local power; see, for example, Shin and So (2001).

This effect of recursive adjustment is very much in the spirit of the proposal of Kostakis et al. (2015), who point out that not demeaning the instrument z_{t-1} (while still demeaning the dependent variable and the predictor itself to account for a nonzero intercept in the predictive regression) reduces the finite-sample correlation between the numerator and the denominator of the IVX t statistic. We shall closer examine the corrections of Kostakis et al. (2015) after analyzing the effect of the orthogonal mean adjustment scheme in:

Proposition 2. *Under the assumptions of Proposition 1, it holds as $T \rightarrow \infty$ that $t_{vx}^{rec} = Z_T + B_T + o_p(T^{\eta/2-1/2})$ with Z_T and B_T from Proposition 1.*

Proof: See Section I of the Online Supplement.

Proposition 2 shows that dependence on c of the leading higher-order terms may in fact be eliminated. Our Monte Carlo study (see Section 3) shows that t_{vx}^{rec} performs quite well in terms of size in spite of the remaining term B_T , so the first correction we suggest is orthogonal mean adjustment.

The Monte Carlo study also shows that the local power of t_{vx}^{rec} is low. To see why, examine

$$y_t - \ddot{y}_t = \beta (x_{t-1} - \ddot{x}_{t-1}) + u_t - \ddot{u}_t,$$

with $\ddot{\cdot}_t$ denoting forward demeaned quantities. Under the alternative $\beta \neq 0$, power is driven by the cross-product $(z_{t-1} - \bar{z}_{t-1})(x_{t-1} - \ddot{x}_{t-1})$. But the forward demeaned x_{t-1} may be rewritten as

$$x_{t-1} - \frac{1}{T-t+1} \sum_{j=t}^T x_{j-1} = -\frac{1}{T-t+1} \sum_{j=t}^T \sum_{k=0}^{j-t-1} \Delta x_{j+k},$$

so, in the extreme case of x_t being a random walk, $\Delta x_t = v_t$ and $(x_{t-1} - \ddot{x}_{t-1})$ is uncorrelated with $z_{t-1} - \bar{z}_{t-1}$. The effect on t_{vx}^{rec} under the alternative is similar to that of a weak instrument. This phenomenon is caused by the correction itself and not by the IVX instrumentation.

Turning our attention to the corrections proposed by Kostakis et al. (2015), they result in

$$t_{vx}^W = \frac{\sum_{t=2}^T z_{t-1} (y_t - \bar{y})}{\sqrt{\sum_{t=2}^T z_{t-1}^2 \hat{u}_t^2 - T \bar{z}^2 \hat{\omega}_{u|v}^2}},$$

where $\hat{\omega}_{u|v}^2 = \hat{\omega}_u^2 - \hat{\lambda}_{uv} \hat{\omega}_v^{-2}$ with $\hat{\omega}_u^2$ and $\hat{\omega}_v^2$ estimators of the long-run variances of u_t and v_t , and $\hat{\lambda}_{uv}$ an estimator of the long-run covariance of u_t and v_t . The behavior of t_{vx}^W is discussed in

Proposition 3. *Under the assumptions of Proposition 1, it holds as $T \rightarrow \infty$ that $t_{vx}^W = Z_T + B_T + C_T + o_p(T^{\eta/2-1/2})$ with $Z_T, B_T,$ and C_T from Proposition 1.*

Proof: See Appendix B.

Although t_{vx}^W has the same leading terms as t_{vx} , we give in Section II of the Online Supplement some of the higher-order terms of t_{vx} which are of order $O_p(T^{\eta-1})$ and missing from t_{vx}^W . Since they contribute to the noncentrality (cf. Kostakis et al., 2015, p. 1516 and also the differences seen by comparing Figures 1 and 2), their absence likely improves the finite-sample behavior of t_{vx}^W and thus explains our findings in the Monte Carlo section that the two-sided t_{vx}^W statistic performs remarkably well. The impact of the terms B_T and in particular C_T on the one-sided versions of t_{vx}^W is, however, not negligible and we, therefore, move on to propose explicit corrections for B_T and C_T .

The quantities involved in the expectation of B_T may for instance be estimated using smoothed residuals, delivering an estimate of

$$\frac{T^{\eta/2-1/2} \int_0^1 \hat{\sigma}_{uv}(s) \hat{\sigma}_u^2(s) \hat{\sigma}_v^2(s) ds}{\sqrt{2a} \sqrt{\left(\int_0^1 \hat{\sigma}_u^2(s) \hat{\sigma}_v^2(s) ds \right)^3}}$$

In dealing with C_T , it may be tempting to proceed analogously. Yet, with c unknown and no consistent estimator available, this approach seems of limited applicability in general.

Alternatively, one may try to match the functional $\int_0^1 e^{-c(1-s)} \sigma_{uv}(s) ds$ using the expectation of another functional depending on c . We illustrate this idea for the case of homoskedasticity, where we have

$$E(U_H(1) J_c(1)) = \delta \sigma_u \sigma_v (1 - e^{-c}) / c.$$

Recall that $\frac{1}{\psi \sqrt{T}} x_{[sT]} \Rightarrow J_{c,H}(s)$, where it is easily seen that $\text{Var}(J_{c,H}(s)) = \frac{1-e^{-2sc}}{2c}$ under homoskedasticity. Therefore, $2 \text{Var}(J_{c,H}(1/2)) = \frac{1-e^{-c}}{c}$, which suggests employing a quantity with this expectation to accommodate the noncentrality induced by C_T .

While it is likely possible to modify this approach in certain particular cases (say for breaks in variances and covariances at suitable times), the general case seems out of reach. We, therefore, propose the use of the correction for the homoskedastic case (and also point to Figure 2 as additional motivation for this proposal). In fact, our simulations in Section 3 and in Section III of the Online Supplement show this to work reasonably well under heteroskedasticity too.

The resulting correction term for the expectation of B_T is then

$$b_T = -\delta / \sqrt{2T(1-\varrho)},$$

where δ may be estimated as the correlation of v_t and u_t based on \hat{u}_t and \hat{v}_t from an AR(p) approximation of x_t with p selected via an information criterion (we resort to the Akaike IC).

For C_T , the natural choice following from the property of the Ornstein-Uhlenbeck process discussed above is then $1/(\psi^2\sigma_v^2)T^{-1}x_{[T/2]}^2$, leading to

$$c_T = -\frac{\delta\sqrt{2}}{\sqrt{T(1-\varrho)}} \cdot \frac{2x_{[T/2]}^2}{\omega^2 T} = 2b_T \frac{2x_{[T/2]}^2}{\omega^2 T},$$

since $\omega^2 = \psi^2\sigma_v^2$ is simply the (stationary) long-run variance of v_t (which may be estimated either based on Δx_t , or—as we proceed in our simulations—on the residuals of a first-order autoregression of x_t).

It should be noted, however, that this delivers a noisy proxy for the mean of C_T : while it will remove the noncentrality due to C_T (at least under homoskedasticity), it will at the same time marginally inflate the variance of the corrected t statistic. The presence of the estimator $\hat{\omega}^2$ in the denominator further inflates the variance: since we employ a nonparametric estimator, its variability in finite samples is large enough to affect the positive effect of the correction. Concretely, it induces outliers in the distribution of the correction and inflates the variance of the corrected statistic. To deal with these issues, we add finite-sample modifications which do not affect the asymptotics.

First, we use the fact that the term to be corrected is bounded to $[0, 1]$, so we censor $2x_{[T/2]}^2/(\hat{\omega}^2 T)$ on $[0, 1]$. While this reduces the variability of the term, it changes its expectation, so we re-normalize it. The limit of $2x_{[T/2]}^2/(\hat{\omega}^2 T)$ is $\chi^2(1)$ under the worst-case scenario of $c = 0$, which, upon censoring on $[0, 1]$, loses the unity expectation, and has an expectation of

$$\varpi = 1 - \sqrt{2/(\pi e)} \approx 0.5161,$$

and a variance of

$$\varsigma = 4(\Phi(1) - \Phi(0)) - 2/(\pi e) - 2\sqrt{2/(\pi e)} \approx 0.1633.$$

Second, we standardize the t statistic to take the variance of the correction c_T into account.

Finally, should x_t be stationary instead of near-integrated, this bias correction may overcorrect, since, for ρ away from unity, the standard normal asymptotics do relatively well even when ϱ is close to unity; see Kostakis et al. (2015). A practical adjustment of the correction is to restrict ϱ in t_{vx}^* to be smaller than an estimate of ρ . In particular, we suggest $\varrho = \min\{\varrho, \hat{\rho}\}$, where $\hat{\rho}$ is the OLS estimator in a first-order autoregression of x_t . Asymptotically, this restriction makes no difference under near-integration, but prevents the bias correction to “overshoot.”

To sum up, we suggest to use, with $\hat{b}_T = -\hat{\delta}/\sqrt{2T(1 - \min\{\varrho, \hat{\rho}\})}$, the statistic

$$t_{vx}^* = \frac{t_{vx} - \hat{b}_T \left(1 + \frac{2}{\varpi} \min \left\{ 1, \frac{2x_{[T/2]}^2}{\hat{\omega}^2 T} \right\} \right)}{\sqrt{1 + 2\frac{\hat{\delta}}{3} \frac{\sqrt{\varsigma}}{\varpi} \hat{b}_T + \frac{\varsigma}{\varpi^2} \hat{b}_T^2}},$$

where $\hat{\delta}/3$ is intended to capture finite-sample correlation of \hat{c}_T and t_{vx} , and is tuned to homoskedasticity. Our Monte Carlo study in Section 3 and in Section III of the Online Supplement shows that t_{vx}^* works well under heteroskedasticity too.

3. FINITE SAMPLE EVIDENCE

In this section, we provide finite sample evidence on the merits of the remedies proposed in this paper. We use a data generating process (DGP) as outlined under equations (1) and (2) with independent innovation process governed by a bi-variate normal distribution with a correlation coefficient of $\delta = -0.95$ (which is typical for predictive regressions with stock returns; see, e.g., Phillips, 2015), as well as time-varying volatility. Concretely,

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} \sim iid\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_t\right),$$

$$\Sigma_t = \begin{pmatrix} \sigma_u(t/T) & 0 \\ 0 & \sigma_v(t/T) \end{pmatrix} \begin{pmatrix} 1 & \delta \\ \delta & 1 \end{pmatrix} \begin{pmatrix} \sigma_u(t/T) & 0 \\ 0 & \sigma_v(t/T) \end{pmatrix};$$

furthermore, $v_t = \phi v_{t-1} + v_t$ for $\phi = 0.5$. The variance of the innovations is taken to vary in time according to the following scenarios:

1. Homoskedasticity [cst.]: $\sigma_u^2(s) = \sigma_v^2(s) = 1, s \in [0, 1]$,
2. Early upward break [e.u.]: $\sigma_u^2(s) = \sigma_v^2(s) = 1 \cdot \mathbf{I}(s < 0.3) + 4 \cdot \mathbf{I}(s \geq 0.3)$,
3. Late upward break [l.u.]: $\sigma_u^2(s) = \sigma_v^2(s) = 1 \cdot \mathbf{I}(s < 0.7) + 4 \cdot \mathbf{I}(s \geq 0.7)$,
4. Early downward break [e.d.]: $\sigma_u^2(s) = \sigma_v^2(s) = 4 \cdot \mathbf{I}(s < 0.3) + 1 \cdot \mathbf{I}(s \geq 0.3)$,
5. Late downward break [l.d.]: $\sigma_u^2(s) = \sigma_v^2(s) = 4 \cdot \mathbf{I}(s < 0.7) + 1 \cdot \mathbf{I}(s \geq 0.7)$.

The size study results are generated using 10,000 replications and considering $c \in \{0, 1, 5, 10, 30, 50\}$ together with $\beta = 0$ for $T = 250$ and 500. To analyze the behavior of the corrected tests under the alternative, we consider a sequence of local alternatives characterized by $\beta = \frac{b}{T} \sqrt{1 - \delta^2}$, for $b \in \{-26, -24, \dots, -2, 0, 2, \dots, 26\}$. Note under $b = 0$, the size properties of the test will be recovered. Since the sign of β might be known in practice (as is often the case when the choice of the predictor is motivated by economic theory⁴), we consider local alternatives covering both situations, $\beta < 0$ and $\beta > 0$, alongside with cases where two-sided testing is of interest. All through this section, we fix $a = 1$ and $\eta = 0.95$ following the recommendation of Kostakis et al. (2015).

We compare four versions of the IVX statistic testing the null $\beta = 0$: the original IVX t statistic (t_{vx}), the finite-sample adjusted version of Kostakis et al. (2015)

⁴For example, the effect of dividend yields on excess returns can be expected to be positive, while increasing risk-free rates could have a negative effect on excess returns (see, e.g., Ang and Bekaert, 2007).

(t_{vx}^W) ,⁵ the IVX t statistic computed with orthogonal mean adjustment (t_{vx}^{rec}), as well as our bias-corrected proposal (t_{vx}^*).

Table 1 shows the finite-sample rejection frequencies at the 5% nominal level for strong negative contemporaneous correlation $\delta = -0.95$.⁶ The finite-sample noncentrality of the standard IVX t statistic, t_{vx} , leads as expected to huge size distortions that only drop to reasonable levels for $c = 10$ if not $c = 30$. The time variation of the variance influences these distortions, but not by much. Also, they do not drop with increasing T , as predicted by the small rates in Proposition 1. The statistic t_{vx}^W on the other hand shows that the finite-sample corrections introduced in Kostakis et al. (2015) work excellently in the two-sided case. Only for $c = 50$, can one observe a very slight tendency to overreject (with rejection frequencies closer to 6% than to 5% for $T = 250$). However, the t_{vx}^W statistic does not behave too well in each tail taken alone, as it tends to overreject to the right (one sees large rejection frequencies for small c , and even for $c = 50$ we note rejection frequencies above 8%) and to underreject to the left (this is most visible for small c , where the rejection frequencies are below 1%). This also does not significantly improve for larger $T = 500$, and exhibits little variation across the different variance patterns. The statistic with backward and forward recursive demeaning, t_{vx}^{rec} , has very good size control (with some exceptions for $c = 0$, where rejection frequencies of 7% may be observed for the test against right-sided alternatives, and some cases of under-rejections: for left-sided testing under downward breaks and $c = 0, 1$ we observe rejection frequencies of 2 or 3%). Finally, the t_{vx}^* statistic has the best size control of all four tests: while it sometimes underrejects for left-sided testing (in the same situations where the t_{vx}^{rec} statistic was undersized), most rejection frequencies lie between 4% and 6%, with only a handful of cases where the 6% threshold is exceeded, and no rejection frequency above 7%.

Summing up, all three modified statistics may be used in a two-sided testing situation in what concerns size control. For one-sided testing situations, the use of t_{vx}^W is not recommended as it overrejects to the right and severely underrejects to the left, which has a dampening effect on rejection frequencies under the alternative; see below. The simulations will also confirm the power-reducing effect of the orthogonal mean adjustment scheme mentioned after Proposition 2.

We present in Figures 3–5 plots of rejection frequencies of the four statistics compared for $c = 0, 10, 30$ and all variance patterns and test variants (left-, right-, and two-sided).

For left-sided testing, it is t_{vx}^* that has best rejection rates in all cases. Compared to t_{vx} and t_{vx}^W , this is because t_{vx}^* is centered correctly and therefore not undersized. Here, t_{vx} seems to perform a bit better than t_{vx}^W . The test based on backward and forward adjustment has poor rejection properties under the alternative; the gap

⁵Kostakis et al. (2015) consider a Wald statistic W for which $(t_{vx}^W)^2 = W$.

⁶The findings are symmetric in the sign of δ ; moreover, size behavior improves uniformly for decreasing magnitude of δ so we do not include the exact figures to save space.

TABLE 1. Size properties of different tests under short-run dynamics and strong contemporaneous shock correlation

		<i>T</i> = 250												<i>T</i> = 500																																				
<i>c</i>	Var	Two-sided				Left-sided				Right-sided				Two-sided				Left-sided				Right-sided																												
		<i>t</i> _{vx}	<i>t</i> ^{rec} _{vx}	<i>t</i> ^W _{vx}	<i>t</i> [*] _{vx}	<i>t</i> _{vx}	<i>t</i> ^{rec} _{vx}	<i>t</i> ^W _{vx}	<i>t</i> [*] _{vx}	<i>t</i> _{vx}	<i>t</i> ^{rec} _{vx}	<i>t</i> ^W _{vx}	<i>t</i> [*] _{vx}	<i>t</i> _{vx}	<i>t</i> ^{rec} _{vx}	<i>t</i> ^W _{vx}	<i>t</i> [*] _{vx}	<i>t</i> _{vx}	<i>t</i> ^{rec} _{vx}	<i>t</i> ^W _{vx}	<i>t</i> [*] _{vx}	<i>t</i> _{vx}	<i>t</i> ^{rec} _{vx}	<i>t</i> ^W _{vx}	<i>t</i> [*] _{vx}																									
0	cst.	20.90	4.90	4.69	4.48	0.11	2.91	0.08	4.08	33.01	6.83	8.82	5.42	20.65	5.09	4.39	4.42	0.09	2.82	0.05	4.50	33.51	7.27	9.00	5.19	e.u.	16.18	4.97	4.66	4.91	0.20	3.11	0.12	6.28	26.30	6.87	9.38	4.72	14.56	4.80	3.88	4.75	0.16	3.07	0.08	6.16	25.15	6.62	8.35	4.18
	l.u.	16.01	5.73	4.93	5.87	0.76	3.71	0.50	6.38	25.51	7.64	9.07	5.22	15.49	5.93	4.69	5.49	0.75	3.70	0.48	6.54	24.62	7.43	9.06	4.59	e.d.	23.55	4.91	4.62	3.52	0.02	2.70	0.02	1.82	37.00	7.14	9.46	5.54	21.79	4.64	4.37	3.28	0.00	2.33	0.01	2.00	35.27	6.99	9.20	5.18
	l.d.	22.38	3.81	4.94	3.35	0.01	1.94	0.00	2.64	35.38	6.15	9.51	5.44	21.45	4.02	4.13	3.22	0.01	1.94	0.00	2.77	34.86	6.34	8.64	4.99	cst.	15.94	4.48	4.87	6.28	0.24	2.58	0.18	6.42	25.58	6.85	9.22	5.82	15.54	5.01	4.46	6.27	0.23	3.24	0.15	6.59	25.03	7.06	9.22	5.56
	e.u.	12.56	4.73	5.10	5.95	0.33	3.08	0.21	7.04	21.40	6.99	9.55	5.40	12.87	4.94	4.78	6.08	0.32	3.10	0.20	7.44	20.90	7.02	9.19	5.03	e.d.	15.76	4.50	4.84	5.50	0.06	2.76	0.04	5.39	25.28	6.70	10.15	5.65	14.54	4.55	4.53	5.24	0.11	2.52	0.01	6.01	23.80	6.78	9.07	5.07
	l.d.	16.47	4.61	4.95	5.44	0.04	2.16	0.03	5.90	26.98	7.20	10.01	5.98	15.24	4.22	4.72	4.96	0.04	2.10	0.04	5.80	24.72	6.30	9.19	5.44																									
1	cst.	9.96	5.10	5.69	6.03	1.15	3.33	0.89	5.47	16.01	6.91	10.12	5.89	9.39	4.70	5.18	5.67	1.24	3.36	1.02	5.41	15.10	6.46	9.65	5.56	e.u.	8.42	5.29	5.20	6.43	1.04	3.13	0.73	6.47	14.24	6.85	9.54	4.94	8.95	5.16	5.13	5.95	0.91	3.24	0.54	5.67	14.76	7.02	9.95	5.18
	l.u.	9.18	5.22	5.81	4.00	1.35	3.34	0.94	2.31	15.20	7.10	10.51	5.39	9.06	5.39	5.46	3.75	1.35	3.49	0.91	2.36	14.63	7.05	10.45	5.48	e.d.	8.65	4.95	5.60	4.65	1.07	3.02	0.77	3.99	14.18	6.61	10.49	5.48	8.07	4.85	4.86	4.66	1.22	3.12	0.89	4.09	13.62	6.36	9.74	4.92
	l.d.	9.19	4.81	5.52	6.82	0.86	2.31	0.54	6.41	14.97	6.80	10.40	5.57	8.14	4.14	4.62	5.92	0.78	2.41	0.57	6.29	14.14	6.31	9.24	4.86																									

(Continued)

TABLE 1. (Continued)

		$T = 250$												$T = 500$											
c	Var	Two-sided				Left-sided				Right-sided				Two-sided				Left-sided				Right-sided			
		t_{vx}	t_{vx}^{rec}	t_{vx}^W	t_{vx}^*	t_{vx}	t_{vx}^{rec}	t_{vx}^W	t_{vx}^*	t_{vx}	t_{vx}^{rec}	t_{vx}^W	t_{vx}^*	t_{vx}	t_{vx}^{rec}	t_{vx}^W	t_{vx}^*	t_{vx}	t_{vx}^{rec}	t_{vx}^W	t_{vx}^*	t_{vx}	t_{vx}^{rec}	t_{vx}^W	t_{vx}^*
10	cst.	7.70	4.84	5.81	4.34	1.68	3.39	1.47	3.15	12.18	6.12	9.80	5.37	7.71	4.92	5.74	4.15	1.78	2.95	1.53	3.23	11.95	6.43	9.60	5.31
	e.u.	7.69	5.17	5.59	4.87	1.61	3.37	1.24	4.01	12.24	6.56	9.57	5.27	7.29	5.28	5.51	4.83	1.70	3.46	1.33	3.94	11.83	6.66	9.18	5.11
	l.u.	7.88	5.29	6.00	3.65	1.84	3.46	1.24	1.89	12.11	6.82	10.07	5.76	7.44	5.34	5.58	3.26	1.70	3.31	1.14	1.68	11.56	6.96	9.78	5.45
	e.d.	7.50	4.94	5.94	3.47	1.85	3.15	1.53	2.18	11.60	6.70	9.98	5.22	7.04	4.78	5.67	3.29	1.82	3.13	1.56	2.06	11.05	6.62	9.42	5.04
	l.d.	7.85	4.44	6.10	4.86	1.97	2.89	1.68	4.27	11.62	6.11	9.49	5.33	6.31	4.31	4.70	4.20	1.67	2.72	1.33	4.07	10.61	5.87	8.81	4.42
30	cst.	6.72	5.17	6.14	4.07	2.82	4.07	2.74	3.13	9.72	6.57	9.18	5.57	6.07	4.81	5.52	3.78	2.90	3.74	2.82	3.11	8.40	6.03	7.84	4.91
	e.u.	6.95	5.31	6.21	4.44	3.47	4.11	3.18	3.78	9.22	6.40	8.63	5.50	6.02	4.86	5.51	3.83	2.83	3.78	2.65	3.26	9.00	6.36	8.29	5.10
	l.u.	6.68	5.10	5.97	4.29	2.95	3.37	2.52	2.90	9.56	6.61	8.88	5.58	5.99	5.06	5.26	3.64	2.45	3.73	2.06	2.42	8.87	6.41	8.16	5.20
	e.d.	6.28	5.16	5.90	3.65	3.09	4.02	2.86	3.09	8.76	6.38	8.38	5.11	5.77	5.16	5.45	3.43	3.03	3.50	2.85	2.98	8.43	6.30	7.93	4.86
	l.d.	6.00	4.80	5.56	3.97	2.99	3.59	2.76	3.44	8.94	6.16	8.34	5.00	6.03	4.74	5.65	3.84	2.86	3.48	2.73	3.16	8.51	6.22	8.03	4.87
50	cst.	6.14	5.06	5.91	4.22	3.26	4.10	3.09	3.37	8.41	5.99	8.11	5.48	5.44	4.82	5.11	3.59	3.12	3.90	3.04	3.16	7.14	5.63	6.84	4.65
	e.u.	6.50	5.26	6.17	4.58	3.33	4.28	3.21	3.61	8.84	6.23	8.40	5.59	5.97	5.18	5.62	4.22	2.96	4.13	2.84	3.15	7.85	6.28	7.65	5.33
	l.u.	6.60	5.25	6.05	4.70	3.62	4.15	3.31	3.62	8.76	6.13	8.36	5.69	5.43	5.00	5.03	3.70	2.68	3.72	2.37	2.69	7.87	5.93	7.46	5.02
	e.d.	5.97	5.12	5.72	4.17	3.39	3.97	3.28	3.44	7.71	6.58	7.56	5.02	5.78	4.87	5.60	3.88	3.27	3.83	3.18	3.26	8.08	6.21	7.89	5.25
	l.d.	6.35	5.18	6.11	4.71	3.68	4.03	3.63	3.88	7.87	6.30	7.62	5.22	5.65	5.09	5.40	3.75	3.45	4.15	3.35	3.59	7.39	5.87	6.99	4.88

Note: Data generated with (1) and (2) with $v_t = \phi v_{t-1} + v_t$ for $\phi = 0.5$, where $(u_t, v_t) \sim iIN(0, \Sigma_t)$ and Σ_t exhibits constant correlation $\delta = -0.95$ and time-varying variances. We set $\rho = 1 - c/T$ for various c and $\varrho = 1 - 1/T^{0.95}$ and use standard normal critical values. See the text for details.

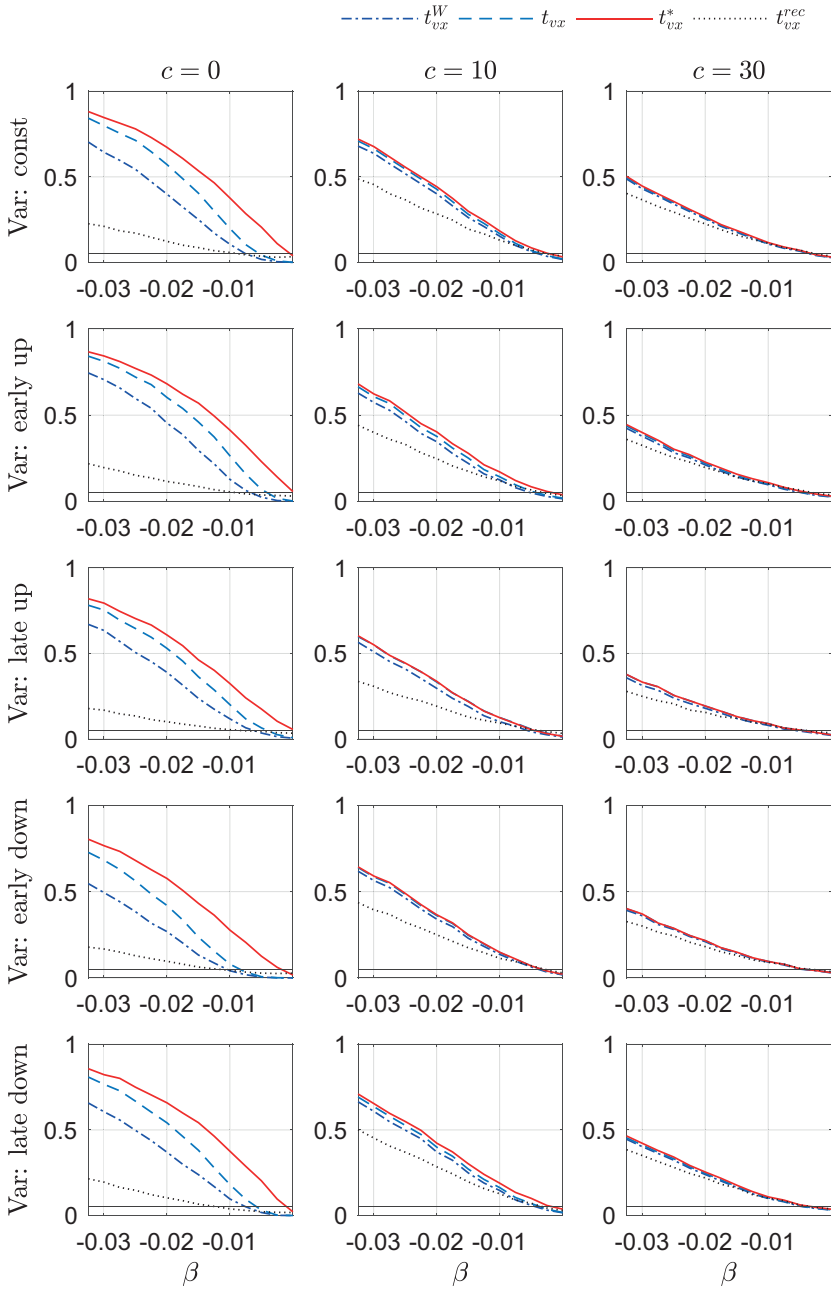


FIGURE 3. Power properties of different test statistics when $T = 250$ for $H_0 : \beta = 0$ versus $H_1 : \beta < 0$ with $\beta = \frac{b}{T} \sqrt{1 - \delta^2}$ for $b \in \{-26, -24, \dots, -2, 0\}$. Data generated with (1) and (2) with $v_t = \phi v_t + v_t$ for $\phi = 0.5$, where $(u_t, v_t) \sim iin(0, \Sigma_t)$ and Σ_t exhibits constant correlation $\delta = -0.95$ and time-varying variances. We set $\rho = 1 - c/T$ for various c and $\varrho = 1 - 1/T^{0.95}$ and use standard normal critical values. See the text for details.

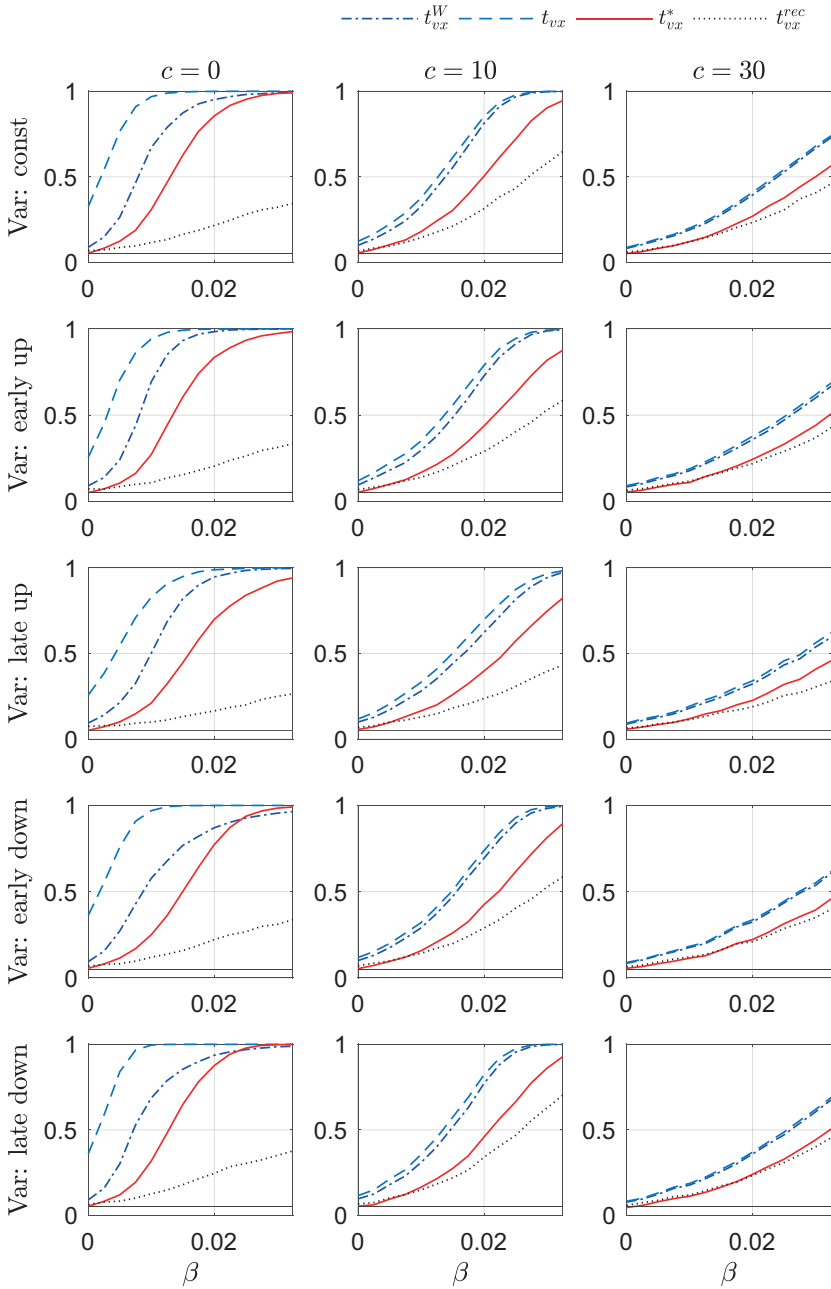


FIGURE 4. Power properties of different test statistics when $T = 250$ for $H_0 : \beta = 0$ versus $H_1 : \beta > 0$ with $\beta = \frac{b}{T}\sqrt{1 - \delta^2}$ for $b \in \{0, 2, 4, \dots, 26\}$. Data generated with (1) and (2) with $v_t = \phi v_{t-1} + v_t$ for $\phi = 0.5$, where $(u_t, v_t) \sim iIN(0, \Sigma_t)$ and Σ_t exhibits constant correlation $\delta = -0.95$ and time-varying variances. We set $\rho = 1 - c/T$ for various c and $\varrho = 1 - 1/T^{0.95}$ and use standard normal critical values. See the text for details.

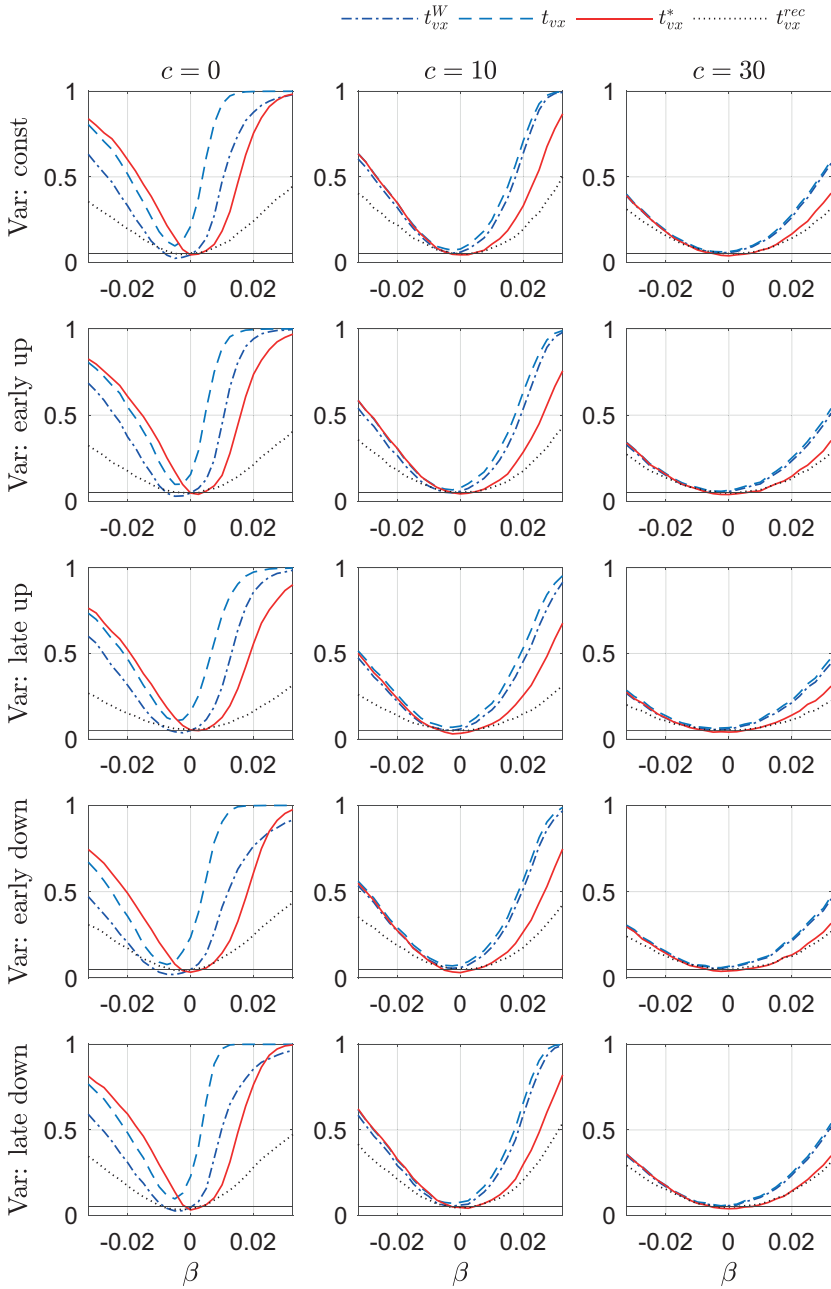


FIGURE 5. Power properties of different test statistics when $T = 250$ for $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ with $\beta = \frac{b}{T} \sqrt{1 - \delta^2}$ for $b \in \{-26, -24, \dots, -2, 0, 2, \dots, 26\}$. Data generated with (1) and (2) with $v_t = \phi v_{t-1} + v_t$ for $\phi = 0.5$, where $(u_t, v_t) \sim iin(0, \Sigma_t)$ and Σ_t exhibits constant correlation $\delta = -0.95$ and time-varying variances. We set $\rho = 1 - c/T$ for various c and $q = 1 - 1/T^{0.95}$ and use standard normal critical values. See the text for details.

to the other tests decreases as c increases, but rejection rates drop anyway with increasing c .

For right-sided testing, t_{vx} rejects very often, but this is of course due to the extreme liberality compared to the other tests. The test based on t_{vx}^{rec} performs, like before, worst (again, with differences decreasing as c increases). To the right, t_{vx}^W typically rejects more often than t_{vx}^* , but keep in mind that it is also quite oversized, even if not as oversized as the uncorrected t_{vx} .

Finally, examining the two-sided tests, we observe as expected a combination of the findings for the left- and right-sided tests, with the difference that the t_{vx}^W test is now correctly sized and the corresponding test decisions are now reliable. The test based on t_{vx}^* is also correctly sized, and the power ranking of the two depends on the sign of β under the alternative. While t_{vx}^W is more powerful against right-sided alternatives, but less powerful against left-sided ones, t_{vx}^* exhibits a more balanced behavior. Again, the larger c , the closer the rejection frequencies of the three corrected tests.

Summing up, when the theory provides clear justifications about using a one-sided test, we could safely recommend the use of t_{vx}^* . For two-sided testing, one has the choice between t_{vx}^W and t_{vx}^* , with the symmetry of the rejection frequencies under the alternative being an argument in favor of t_{vx}^* , and the higher power against right-sided alternatives (or left-sided, should the correlation δ be positive) being an argument in favor of t_{vx}^W . Altogether, as our Monte Carlo results show, we would like to stress that t_{vx}^W has a very good size control for two-sided testing. Allowing for conditional heteroskedasticity does not alter this general recommendation (see Section III of the Online Supplement for further Monte Carlo simulation results supporting this claim).

4. CONCLUDING REMARKS

A convenient approach in the context of predictive regressions where the persistence of the endogenous forecasting variable is unknown, is to turn to IV regressions where a so-called extended instrumental variable with a controlled level of persistence is constructed. The resulting IVX estimator is asymptotically mixed Gaussian and makes for standard asymptotic inference. Finite-sample deviations from the asymptotic limit can, however, be quite serious. Typically manifested in the form of noncentrality, they depend heavily on how the IV estimator is constructed.

In this paper, we provide a structured approach to control the small sample noncentrality of the IVX t statistic for a given instrumental variable, and as a result control the size distortions. First, we develop a higher-order expansion of the corresponding IVX t statistic and as such provide a theoretical understanding of the small sample deviations of the t statistic from its limit. This in turn suggests ways to center the t statistic at the origin under the null. Combining forward and recursive demeaning does account for most leading terms of the bias at the cost of some loss of power. An explicit correction for the noncentrality achieves similar

size control but without the power reduction. These proposals do not assume any parametric restriction on the persistence of the extended instrumental variable, and rather provide, for any given parameterization thereof, a corresponding way of reducing noncentrality.

Our recommendations do not concern Wald tests of the null of no predictability in multiple regressions, the main reason being that the corrections already proposed by Kostakis et al. (2015) are quite effective for the Wald statistic. We leave a full analysis of higher-order terms of the involved quadratic forms to future work.

Our Monte Carlo study shows that all of these proposals provide substantial remedies to small sample size distortions to the IVX t statistic while maintaining relatively good properties under the alternative. Further, when the effect of a forecasting variable is negative, we suggest using a left-sided t statistic with one of the corrections we provided in this paper, since our Monte Carlo study provides evidence that such a strategy is associated with a better statistical power compared to using a two-sided test. For two-sided alternatives, the Wald test of Kostakis et al. (2015) offers the better balance between size and power.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit <http://dx.doi.org/10.1017/S0266466620000298>

REFERENCES

- Ang, A. & G. Bekaert (2007) Return predictability: Is it there? *Review of Financial Studies* 20(3), 651–707.
- Breitung, J. & M. Demetrescu (2015) Instrumental variable and variable addition based inference in predictive regressions. *Journal of Econometrics* 187(1), 358–375.
- Campbell, J.Y. & M. Yogo (2006) Efficient tests of stock return predictability. *Journal of Financial Economics* 81(1), 27–60.
- Camponovo, L. (2015) Differencing transformations and inference in predictive regression models. *Econometric Theory* 31(6), 1331–1358.
- Cavaliere, G., A. Rahbek, & A.M.R. Taylor (2010) Testing for co-integration in vector autoregressions with non-stationary volatility. *Journal of Econometrics* 158(1), 7–24.
- Demetrescu, M., I. Georgiev, P.M.M. Rodrigues, & A.M.R. Taylor (2020) Testing for episodic predictability in stock returns. *Journal of Econometrics*, forthcoming.
- Demetrescu, M. & P.M.M. Rodrigues (2016) Residual-Augmented IVX Predictive Regression. Working paper 5/2016, Banco de Portugal.
- Elliott, G. & J.H. Stock (1994) Inference in time series regression when the order of integration of a regressor is unknown. *Econometric Theory* 10(3–4), 672–700.
- Everaert, G. (2013) Orthogonal to backward mean transformation for dynamic panel data models. *The Econometrics Journal* 16(2), 179–221.
- Gonzalo, J. & J.-Y. Pitarakis (2012) Regime-specific predictability in predictive regressions. *Journal of Business & Economic Statistics* 30(2), 229–241.
- Jansson, M. & M.J. Moreira (2006) Optimal inference in regression models with nearly integrated regressors. *Econometrica* 74(3), 681–714.
- Kostakis, A., T. Magdalinos, & M.P. Stamatogiannis (2015) Robust econometric inference for stock return predictability. *Review of Financial Studies* 28(5), 1506–1553.

- Magdalinos, T. & P.C.B. Phillips (2009) Limit theory for cointegrated systems with moderately integrated and moderately explosive regressors. *Econometric Theory* 25(2), 482–526.
- Maynard, A. & K. Shimotsu (2009) Covariance-based orthogonality tests for regressors with unknown persistence. *Econometric Theory* 25(1), 63–116.
- Nickell, S. (1981) Biases in dynamic models with fixed effects. *Econometrica* 49(6), 1417–1426.
- Phillips, P.C.B. (1987) Towards a unified asymptotic theory for autoregression. *Biometrika* 74(3), 535–547.
- Phillips, P.C.B. (2015) Pitfalls and possibilities in predictive regression. *Journal of Financial Econometrics* 13(3), 521–555.
- Phillips, P.C.B. & J.H. Lee (2013) Predictive regression under various degrees of persistence and robust long-horizon regression. *Journal of Econometrics* 177(2), 250–264.
- Phillips, P.C.B. & T. Magdalinos (2007) Limit theory for moderate deviations from a unit root. *Journal of Econometrics* 136(1), 115–130.
- Phillips, P.C.B. & T. Magdalinos (2009) Econometric Inference in the Vicinity of Unity. CoFie Working Paper 7, Singapore Management University.
- Phillips, P.C.B. & V. Solo (1992) Asymptotics for linear processes. *Annals of Statistics* 20(2), 971–1001.
- Shin, D.W. & B.S. So (2001) Recursive mean adjustment for unit root tests. *Journal of Time Series Analysis* 22(5), 595–612.
- So, B.S. & D.W. Shin (1999) Recursive mean adjustment in time-series inferences. *Statistics & Probability Letters* 43(1), 65–73.
- Stambaugh, R.F. (1999) Predictive regressions. *Journal of Financial Economics* 54(3), 375–421.
- Westerlund, J., H. Karabiyik, & P. Narayan (2017) Testing for predictability in panels with general predictors. *Journal of Applied Econometrics* 32(3), 554–574.
- Yang, B., W. Long, L. Peng, & Z. Cai (2020) Testing the predictability of US housing price index returns based on an IVX-AR model. *Journal of the American Statistical Association*, forthcoming.

APPENDIX

A. PRELIMINARIES

Note that, for large enough T , $\sum_{j=0}^{t-1} \varrho^{kj} = \frac{1-\varrho^{kt}}{1-\varrho^k} = \frac{T^\eta}{a} \left(\frac{1-\varrho^{kt}}{1+\varrho+\dots+\varrho^{k-1}} \right) \leq \frac{1}{ka} T^\eta$ for fixed k and $\varrho = 1 - \frac{a}{T^\eta}$ with $\eta \in (0, 1)$ and $a > 0$. In fact, for $t/T \rightarrow s > 0$, $\varrho^{kt} = \left(1 - \frac{a}{T^\eta}\right)^{kt} = \left(\left(1 - \frac{a}{T^\eta}\right)^{-\frac{T^\eta}{a}} \right)^{-ka \frac{t}{T} T^{1-\eta}} \rightarrow 0$ so $\sum_{j=0}^{t-1} \varrho^{kj} \sim \frac{1}{ka} T^\eta$ in such cases. Let further C denote a generic constant whose value may change from occurrence to occurrence. Before moving on to the proofs, we state three helpful lemmas.

Lemma A.1. *Under the assumptions of Proposition 1, it holds that*

$$\sup_{1 \leq t \leq T} \left\| T^{-1/2} x_t \right\|_4 \leq C \quad \text{and} \quad \sup_{1 \leq t \leq T} \left\| T^{-\eta/2} z_t \right\|_4 \leq C \quad \forall T.$$

Proof. See Section I of the Online Supplement. □

Lemma A.2. *Under the assumptions of Proposition 1, we have under the null $\beta = 0$ that*

$$t_{vx} = \frac{\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T (\tilde{z}_{t-1} - \bar{\tilde{z}})(u_t - \bar{u})}{\sqrt{\frac{1}{T^{1+\eta}} \sum_{t=2}^T (\tilde{z}_{t-1} - \bar{\tilde{z}})^2 u_t^2}} + O_p(T^{-\eta/2}),$$

where $\tilde{z}_{t-1} = (1 - \rho L)_+^{-1} \Delta \tilde{x}_{t-1}$ with $\tilde{x}_t = \sum_{j=0}^{t-2} \rho^j v_{t-1-j}$ and $\bar{\tilde{z}}$ the sample average of \tilde{z}_{t-1} .

Proof. See Appendix B. □

Lemma A.3. *Under the assumptions of Proposition 1, we have*

$$T^{1/2-\eta/2} E(B_T) \rightarrow -\frac{1}{\sqrt{2a}} \frac{\int_0^1 \sigma_{uv}(s) \sigma_u^2(s) \sigma_v^2(s) ds}{\sqrt{\left(\int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds\right)^3}}.$$

Proof. See Section I of the Online Supplement. □

B. PROOFS

Proof of Lemma A.2

With $\tilde{x}_{t-1} = \sum_{j=0}^{t-2} \rho^j v_{t-1-j} = (1 - \rho L)_+^{-1} v_{t-1}$ (i.e., the DGP without short run dynamics), use the Phillips–Solo decomposition (see Phillips and Solo, 1992) of v_t (see the proof of Lemma A.1), to obtain that

$$x_{t-1} = \sum_{j=0}^{t-2} \rho^j v_{t-1-j} = \psi \sum_{j=0}^{t-2} \rho^j v_{t-1-j} + \sum_{j=0}^{t-2} \rho^j \Delta \tilde{v}_{t-1-j}$$

such that $\Delta x_{t-1} = \psi \Delta \tilde{x}_{t-1} + \Delta q_{t-1}$ where

$$q_{t-1} = \sum_{j=0}^{t-2} \rho^j \Delta \tilde{v}_{t-1-j} = \tilde{v}_{t-1} - \rho^{t-3} \tilde{v}_1 + (\rho - 1) \sum_{j=0}^{t-4} \rho^{j-1} \tilde{v}_{t-2-j}.$$

Note that q_t is uniformly L_4 -bounded. Now, Demetrescu et al. (2020, Lemma. A.1) show under slightly stricter assumptions than ours that $\sum_{t=2}^T z_{t-1} = O_p(T^{1/2+\eta})$ such that $\bar{z} = O_p(T^{\eta-1/2})$ and hence

$$\sum_{t=2}^T (\tilde{z}_{t-1} - \bar{z}) u_t = \sum_{t=2}^T z_{t-1} (u_t - \bar{u}) = \sum_{t=2}^T z_{t-1} u_t - \bar{u} \sum_{t=2}^T z_{t-1} = \Theta_p(T^{1/2+\eta/2}).$$

Moreover, $\sqrt{\sum_{t=2}^T (z_{t-1} - \bar{z})^2 \hat{u}_t^2}$ has the same order: using the OLS residuals $\hat{u}_t = u_t - \bar{u} - (x_{t-1} - \bar{x})(\hat{\beta} - \beta)$, we obtain

$$\sum_{t=2}^T (z_{t-1} - \bar{z})^2 \hat{u}_t^2 = \sum_{t=2}^T (z_{t-1} - \bar{z})^2 u_t^2 - 2(\hat{\beta} - \beta) \sum_{t=2}^T (z_{t-1} - \bar{z})^2 (x_{t-1} - \bar{x}) u_t$$

$$+ (\hat{\beta} - \beta)^2 \sum_{t=2}^T (z_{t-1} - \bar{z})^2 (x_{t-1} - \bar{x}) o_p(T^{1+\eta}),$$

and, using Lemma A.1, together with the superconsistency of the OLS estimator $\hat{\beta}$,⁷ it follows that

$$\sum_{t=2}^T (z_{t-1} - \bar{z})^2 \hat{u}_t^2 = \sum_{t=2}^T (z_{t-1} - \bar{z})^2 u_t^2 + O_p(T^{1/2+\eta}).$$

Therefore,

$$\begin{aligned} \frac{1}{T^{1+\eta}} \sum_{t=2}^T (z_{t-1} - \bar{z})^2 u_t^2 &= \frac{1}{T^{1+\eta}} \sum_{t=2}^T z_{t-1}^2 u_t^2 - \frac{2\bar{z}}{T^{1+\eta}} \sum_{t=2}^T z_{t-1} u_t^2 + \frac{\bar{z}^2}{T^{1+\eta}} \sum_{t=2}^T u_t^2 \\ &= \frac{1}{T^{1+\eta}} \sum_{t=2}^T z_{t-1}^2 u_t^2 + o_p(1), \end{aligned}$$

since $\sum_{t=2}^T z_{t-1} u_t^2 = \sum_{t=2}^T z_{t-1} (u_t^2 - \sigma_{u,t}^2) + \sum_{t=2}^T \sigma_{u,t}^2 z_{t-1} = O_p(\max\{T^{1/2+\eta/2}, T^{1/2+\eta}\})$. At the same time, it is easily shown using the exact same arguments as in the proof of Lemma A.4 item 2 of Demetrescu and Rodrigues (2016) that

$$\frac{1}{T^{1+\eta}} \sum_{t=2}^T z_{t-1}^2 u_t^2 \xrightarrow{p} \frac{\psi^2}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds > 0, \tag{B.1}$$

so it follows that

$$\frac{\sum_{t=2}^T (z_{t-1} - \bar{z}) u_t}{\sqrt{\sum_{t=2}^T (z_{t-1} - \bar{z})^2 \hat{u}_t^2}} = \frac{\sum_{t=2}^T (z_{t-1} - \bar{z}) u_t}{\sqrt{\sum_{t=2}^T (z_{t-1} - \bar{z})^2 u_t^2}} + O_p(T^{-1/2}). \tag{B.2}$$

Note now that $z_{t-1} = \psi \bar{z}_{t-1} + \sum_{j=0}^{t-3} \varrho^j \Delta q_{t-1-j} = \psi \bar{z}_{t-1} + r_{t-1}$, where the sum $\sum_{j=0}^{t-3} \varrho^j \Delta q_{t-1-j}$ may be re-arranged to give

$$r_{t-1} = q_{t-1} - \varrho^{t-3} q_1 + (\varrho - 1) \sum_{j=0}^{t-4} \varrho^{j-1} q_{t-1-j};$$

since q_t is uniformly L_4 -bounded, it is easily seen that r_t is itself uniformly L_4 -bounded. We now show that

$$\frac{1}{T^{1+\eta}} \sum_{t=2}^T (\bar{z}_{t-1} - \bar{z})^2 u_t^2 = \frac{\psi^2}{T^{1+\eta}} \sum_{t=2}^T (\bar{z}_{t-1} - \bar{z})^2 u_t^2 + O_p(T^{-\eta/2}), \tag{B.3}$$

and

$$\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T (z_{t-1} - \bar{z}) u_t = \frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T z_{t-1} (u_t - \bar{u})$$

⁷Showing that $T(\hat{\beta} - \beta) \Rightarrow (\psi \int_0^1 J_{c,H}^2(s) ds)^{-1} \int_0^1 J_{c,H}(s) dU_H(s)$ is a standard exercise and we omit the details.

$$= \frac{\psi}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} (u_t - \bar{u}) + O_p \left(T^{-\eta/2} \right). \tag{B.4}$$

For (B.3), the difference between the terms on the l.h.s. and r.h.s. is given by

$$\frac{2\psi}{T^{1+\eta}} \sum_{t=2}^T \tilde{z}_{t-1} r_{t-1} u_t^2 + \frac{1}{T^{1+\eta}} \sum_{t=2}^T r_{t-1}^2 u_t^2 = A_{1T} + A_{2T}.$$

Then,

$$E(|A_{1T}|) \leq \frac{C}{T^{1+\eta}} \sum_{t=2}^T \sqrt{E\left(\left|\tilde{z}_{t-1}^2\right|\right) E\left(\left|r_{t-1}^2\right|\right) E\left(\left|u_t^2\right|\right)} = O\left(T^{-\eta/2}\right)$$

and, using L_r -boundedness again (and again), $A_{2T} = O_p\left(T^{-\eta}\right)$.

For (B.4), the vanishing term is given by

$$\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T r_{t-1} (u_t - \bar{u}) = \frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T r_{t-1} u_t - \bar{u} \frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T r_{t-1} = O_p\left(T^{-\eta/2}\right),$$

since the elements of the first sum have the md property and are uniformly L_4 -bounded, while for the second summand we have $\left\| \sum_{t=2}^T r_{t-1} \right\|_4 = O(T)$ but $\bar{u} = O_p\left(T^{-1/2}\right)$ as required.

Proof of Proposition 1

Begin by applying Lemma A.2; since $T^{-\eta/2} = o(T^{\eta/2-1/2})$, we may focus on

$$\begin{aligned} \frac{\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} (u_t - \bar{u})}{\sqrt{\frac{1}{T^{1+\eta}} \sum_{t=2}^T \left(\tilde{z}_{t-1} - \bar{\tilde{z}}\right)^2 u_t^2}} &= \frac{\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} u_t}{\sqrt{\frac{1}{T^{1+\eta}} \sum_{t=2}^T \left(\tilde{z}_{t-1} - \bar{\tilde{z}}\right)^2 u_t^2}} \\ &\quad - \bar{u} \frac{\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1}}{\sqrt{\frac{1}{T^{1+\eta}} \sum_{t=2}^T \left(\tilde{z}_{t-1} - \bar{\tilde{z}}\right)^2 u_t^2}}. \end{aligned}$$

For the second term, we have (using the arguments in the proof of Lemma A.1 of Demetrescu et al., 2020) that

$$\frac{1}{T^{1/2+\eta}} \sum_{t=2}^T \tilde{z}_{t-1} = \frac{1}{T^{1/2}a} \tilde{x}_{T-1} + o_p(1) \Rightarrow \frac{1}{a} J_{C,H}(1),$$

such that

$$T^{1/2-\eta/2} \frac{\bar{u}}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} \Rightarrow \frac{1}{a} U_H(1) J_{C,H}(1)$$

and the limit of C_T follows since $\frac{1}{T^{1+\eta}} \sum_{t=2}^T (\tilde{z}_{t-1} - \bar{\tilde{z}})^2 u_t^2 = \frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds + o_p(1)$; cf. equation (B.1).

For the first term, use a first-order Taylor series expansion with rest term in differential form for the function $\frac{a}{\sqrt{x}}$ about $x_0 = \frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds$, where $x = \frac{1}{T^{1+\eta}} \sum_{t=2}^T (\tilde{z}_{t-1} - \bar{\tilde{z}})^2 u_t^2$ and $a = \frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} u_t$, such that

$$\frac{\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} u_t}{\sqrt{\frac{1}{T^{1+\eta}} \sum_{t=2}^T (\tilde{z}_{t-1} - \bar{\tilde{z}})^2 u_t^2}} = Z_T + B_T + R_T,$$

where

$$Z_T = \frac{\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} u_t}{\sqrt{\frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds}}, \tag{B.5}$$

$$B_T = -\frac{1}{2} \frac{1}{\sqrt{\left(\frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds\right)^3}} \left(\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} u_t \right) \times \left(\frac{1}{T^{1+\eta}} \sum_{t=2}^T \tilde{z}_{t-1}^2 u_t^2 - \frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds \right) \tag{B.6}$$

and, with ξ_T between $\frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds$ and $\frac{1}{T^{1+\eta}} \sum_{t=2}^T (\tilde{z}_{t-1} - \bar{\tilde{z}})^2 u_t^2$,

$$R_T = \frac{3}{4} \frac{1}{\sqrt{\xi_T^5}} \left(\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} u_t \right) \left(\frac{1}{T^{1+\eta}} \sum_{t=2}^T (\tilde{z}_{t-1} - \bar{\tilde{z}})^2 u_t^2 - \frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds \right)^2 - \frac{1}{2} \frac{1}{\sqrt{\left(\frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds\right)^3}} \left(\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} u_t \right) \left(\frac{-2}{T^{1+\eta}} \sum_{t=2}^T \tilde{z}_{t-1} \bar{\tilde{z}} u_t^2 + \frac{1}{T^{1+\eta}} \sum_{t=2}^T \bar{\tilde{z}}^2 u_t^2 \right).$$

Given that $\xi_T = O_p(1)$, the first term on the r.h.s. is seen to be $o_p(B_T)$, and we have from Lemma A.3 that B_T is $O(T^{\eta/2-1/2})$. Then, $\frac{\bar{\tilde{z}}^2}{T^{1+\eta}} \sum_{t=2}^T u_t^2 = O_p(T^{\eta-1}) = \frac{1}{T^{1+\eta}} \sum_{t=2}^T \tilde{z}_{t-1} \bar{\tilde{z}} u_t^2$, such that the second term on the r.h.s. is also $o_p(T^{\eta/2-1/2})$, and R_T is therefore dominated by B_T and Z_T .

Then, Z_T has the required limiting standard normal (see Lemmata A.3 item (ii) and A.4 item 2 in Demetrescu and Rodrigues, 2016), and has zero mean for all T ; the result follows with Lemma A.3.

Proof of Proposition 3

Write under the null

$$t_{vx}^W = \frac{\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T z_{t-1} (u_t - \bar{u})}{\sqrt{\frac{1}{T^{1+\eta}} \sum_{t=2}^T z_{t-1}^2 \hat{u}_t^2}} + \frac{1}{2} \frac{\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T z_{t-1} (u_t - \bar{u})}{\sqrt{\xi_T}} \left(T^{-\eta} \bar{z}^2 \hat{\omega}_{u|v}^2 \right),$$

where $\zeta_T = \frac{1}{T^{1+\eta}} \left(\sum_{t=2}^T \tilde{z}_{t-1}^2 \hat{u}_t^2 - \gamma_T T \bar{z}^2 \hat{\omega}_{u|v}^2 \right)$ for $\gamma_T \in [0, 1]$ is obviously $O_p(1)$.

The second summand on the r.h.s. is seen to vanish at a sufficient rate: recall that $\bar{z} = O_p(T^{\eta-1/2})$, such that, not surprisingly, $\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} (u_t - \bar{u}) = O_p(1)$; also, $T^{-\eta} \bar{z}^2 \hat{\omega}_{u|v}^2 = O_p(T^{\eta-1})$.

For the first summand, we obtain after some straightforward algebra along the lines of the proof of Lemma A.2 that

$$\begin{aligned} \frac{\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} (u_t - \bar{u})}{\sqrt{\frac{1}{T^{1+\eta}} \sum_{t=2}^T \tilde{z}_{t-1}^2 \hat{u}_t^2}} &= \frac{\sum_{t=2}^T \tilde{z}_{t-1} (u_t - \bar{u})}{\sqrt{\sum_{t=2}^T \tilde{z}_{t-1}^2 u_t^2}} + O_p(T^{-\eta/2}) \\ &= Z_T + B_T + C_T + R_T^W + o_p(T^{\eta/2-1/2}), \end{aligned}$$

where, with $\xi_T^W = O_p(1)$ between $\frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds$ and $\frac{1}{T^{1+\eta}} \sum_{t=2}^T (\tilde{z}_{t-1} - \bar{z})^2 u_t^2$, it follows that

$$R_T^W = \frac{3}{4} (\xi_T^W)^{-5/2} \left(\frac{1}{T^{1/2+\eta/2}} \sum_{t=2}^T \tilde{z}_{t-1} (u_t - \bar{u}) \right) \left(\frac{1}{T^{1+\eta}} \sum_{t=2}^T \tilde{z}_{t-1}^2 u_t^2 - \frac{1}{2a} \int_0^1 \sigma_u^2(s) \sigma_v^2(s) ds \right)^2.$$

Like R_T from the proof of Proposition 1, R_T^W is easily shown to be $o_p(T^{\eta/2-1/2})$ and the result follows.