Econometric Theory, **28**, 2012, 1003–1036.
doi:10.1017/S0266466612000035

ESTIMATORS FOR PERSISTENT AND POSSIBLY NONSTATIONARY DATA WITH CLASSICAL PROPERTIES

YURIY GORODNICHENKO
University of California at Berkeley

ANNA MIKUSHEVA
MIT

SERENA NG
Columbia University

This paper considers a moments-based nonlinear estimator that is $\sqrt{T}$-consistent and uniformly asymptotically normal irrespective of the degree of persistence of the forcing process. These properties hold for linear autoregressive models, linear predictive regressions, and certain nonlinear dynamic models. Asymptotic normality is obtained because the moments are chosen so that the objective function is uniformly bounded in probability and so that a central limit theorem can be applied. Critical values from the normal distribution can be used irrespective of the treatment of the deterministic terms. Simulations show that the estimates are precise and the t-test has good size in the parameter region where the least squares estimates usually yield distorted inference.

1. INTRODUCTION

This paper considers a quasi-differencing (QD) framework that can yield $\sqrt{T}$-consistent and uniformly asymptotically normal estimators of autoregressions and multiple regressions when the predictors are persistent and possibly nonstationary. The approach can also be used to estimate dynamic stochastic general equilibrium (DSGE) models. The critical values are invariant to the presence of deterministic trends.

Let θ be an unknown parameter vector and let θ^0 be its true value. We propose *nonlinear* QD estimators that can generically be defined as

$$\widehat{\theta}_K = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \overline{g}(\theta)' W_T \overline{g}(\theta), \tag{1}$$

where $\overline{g}(\theta)$ is a $K \times 1$ vector of moments, W_T is a $K \times K$ positive-definite matrix, and Θ is a bounded set containing values of θ.[1] The basic premise of QD

The authors thank the editor and three referees for many helpful comments and suggestions. Mikusheva acknowledges financial support from the Castle-Krob Career Development Chair. Ng acknowledges financial support from the National Science Foundation (SES-0962431). Address correspondence to Serena Ng, Department of Economics, Columbia University, 420 W. 118 St., New York, NY 10027, USA; e-mail: Serena.Ng@columbia.edu.

estimation is that for $\widehat{\theta}_K$ to have classical properties, $\overline{g}(\theta)$ needs to be uniformly bounded in probability and that a central limit theorem can be applied. To this end, $\overline{g}(\theta)$ is defined as the difference between the normalized autocovariances of the variables in the model and the data, all quasi-differenced at a persistence parameter that is to be estimated jointly with other parameters of the model. The normalization and QD together provide a nonlinear transformation of the autocovariances to result in estimators that are robust to possible nonstationarity in the data.

Achieving asymptotic normality without knowing when the exogenous process has an autoregressive unit root can be very useful in applied work because the answers to many macroeconomic questions are sensitive to assumptions about the nature of the trend and to whether the corresponding regressions are estimated in levels or in first-differences. The price to pay for practical simplicity and robustness is that the proposed estimators are $\sqrt{T}$-consistent rather than superconsistent when the regressors are truly nonstationary. Although other asymptotically normal estimators robust to nonstationary regressors are available, they apply only to specific linear models. The QD estimation framework is general and can be used whenever the variables can be quasi-differenced in the way discussed subsequently.

We establish uniform asymptotic normality of QD-based estimators in many different settings. Throughout, we use the notion of uniformity given by Mikusheva (2007a), who studied uniform coverage properties of various inference procedures for the AR(1) model.

DEFINITION 1. *A family of distributions* $F^{(1)}_{\theta,T}(x) = P_{\theta,T}\{\xi_1 < x\}$ *is asymptotically approximated by (converges to) a family of distributions* $F^{(2)}_{\theta}(x) = P_{\theta}\{\xi_2 < x\}$ uniformly *over* $\theta \in \Theta$ *if*

$$\lim_{T\to\infty} \sup_{\theta\in\Theta} \sup_{x} \left| F^{(1)}_{\theta,T}(x) - F^{(2)}_{\theta}(x) \right| = 0.$$

In our analysis, $F^{(2)}_{\theta}(x)$ is in the family of Gaussian distributions.

The paper is structured as follows. Section 2 provides a rigorous analysis of the AR(1) model. Section 3 extends the analysis to AR(p) models, and Section 4 studies predictive regressions. Section 5 considers nonlinear estimation of structural parameters. Simulations are presented in Section 6. The relation of QD estimation to other $\sqrt{T}$-consistent estimators is discussed in Section 7. All proofs are in the Appendix. As a matter of notation, the indicator function $\mathbb{I}(a)$ is one if a is true and zero otherwise. We let $W(\cdot)$ be the standard Brownian motion and use $\Rightarrow$ to denote weak convergence.

2. THE AR(1) MODEL

To systematically motivate the idea behind QD estimation, we begin with the simple AR(1) model with parameter α and whose true value is α^0. For $t = 1,\ldots,T$, the data are generated by

$$y_t = \alpha^0 y_{t-1} + \varepsilon_t, \qquad y_0 = 0. \tag{2}$$

Hereafter, we let ε_t be the deviation between the dependent variable and the conditional mean evaluated at the true parameter value, whereas e_t is the deviation evaluated at an arbitrary value of the parameter vector. The error ε_t does not need to be independent and identically distributed or Gaussian, but it cannot be conditional heteroskedastic or heteroskedastic.

Assumption A. $(\varepsilon_t, \mathcal{F}_t)$ is a stationary ergodic martingale-difference sequence with conditional variance $\mathbb{E}(\varepsilon_t^2|\mathcal{F}_{t-1}) = \sigma^2 = \gamma_0$ and $\mathbb{E}((\varepsilon_t^2 - \sigma^2)^2|\mathcal{F}_{t-1}) = \mu_4$.

The least squares estimator $\widehat{\alpha}_{OLS}$ is defined as the solution to $\overline{g}(\alpha) = 0$, where $\overline{g}(\alpha) = 1/T\sum_{t=1}^T e_t y_{t-1}$ is the sample analogue of the moment condition

$$\mathbb{E}g_t(\alpha^0) = \mathbb{E}[\varepsilon_t y_{t-1}] = 0,$$

with $e_t = (y_t - \alpha y_{t-1})$. When $\alpha^0 < 1$, $\widehat{\alpha}_{OLS}$ is $\sqrt{T}$-consistent and asymptotically normal. Although $\widehat{\alpha}_{OLS}$ is superconsistent at $\alpha^0 = 1$, its distribution is nonstandard, which makes inference difficult. In particular, the t-statistic for testing $\alpha^0 = 1$ is nonnormal in finite samples and converges to the so-called Dickey–Fuller distribution. The issue of nonstandard inference arises because of two problems. First, when $\alpha^0 = 1$, the sample moment evaluated at a value $\alpha \neq \alpha^0 = 1$ explodes, and second, the normalized sample moment evaluated at the true value does not obey a central limit theorem. In other words, $\overline{g}(\alpha) = 1/T\sum_{t=1}^T e_t y_{t-1}$ is stochastically unbounded, and $1/T\sum_{t=1}^T \varepsilon_t y_{t-1} \Rightarrow \sigma^2\int_0^1 W(r)dW(r)$, where $W(r)$ is the standard Brownian motion.

Our starting point is to resolve the second issue by exploiting the autocovariance structure of the errors. Specifically, for $j \geq 1$, it holds that

$$\mathbb{E}(\varepsilon_t\varepsilon_{t-j}) = 0. \tag{3}$$

Furthermore, for all $|\alpha^0| \leq 1$ and $\varepsilon_t = y_t - \alpha^0 y_{t-1}$, the population moment condition has a sample analogue that obeys a central limit theorem:

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=j+1}^{T}\varepsilon_t\varepsilon_{t-j} - \mathbb{E}(\varepsilon_t\varepsilon_{t-j})\right) \Rightarrow \xi_j \sim N(0, \sigma^4). \tag{4}$$

Obviously, α^0 is unknown, and ε_t is not observed. However, we can quasi-difference y_t at some α and then optimize over all possible values of α by matching the sample autocovariances of the quasi-differenced data[2]

$$\widehat{\gamma}_j(\alpha) = \frac{1}{T}\sum_{t=j+1}^{T} e_t e_{t-j}$$

with those of the model evaluated under the assumption that α is the true value. Precisely, the model implied moments are

$$\gamma_j(\alpha) = \mathbb{E}_\alpha e_t e_{t-j} = \mathbb{I}(j=0)\sigma^2,$$

where $\mathbb{E}_\alpha$ is the expectation taken under the (not necessarily correct) assumption that α is the true value. In cases when $\gamma_j(\alpha)$ is constant, the dependence of γ_j on α can be suppressed. For example, in the AR(1) under consideration, $\gamma_j(\alpha) = 0$ for all $j \geq 1$, and we may write γ_j instead of $\gamma_j(\alpha)$. This is also true of the AR(p) and predictive regressions considered in Sections 3 and 4. However, in more complex models such as the one considered in Section 5, γ_j often depends on α in a complicated and analytically intractable way. For clarity, we keep the explicit dependence of γ_j on α throughout.

Let $\overline{g}_{NQD}(\alpha) = (\overline{g}_{0,NQD}(\alpha), \ldots, \overline{g}_{K,NQD}(\alpha))'$, where

$$\overline{g}_{j,NQD}(\alpha) = \widehat{\gamma}_j(\alpha) - \gamma_j(\alpha)$$

is the difference between the model-implied and the sample autocovariance of e_t. The NQD (nonstandard quasi-difference) estimator is defined as

$$\widehat{\alpha}_{K,NQD} = \underset{\alpha}{\operatorname{argmin}}\, \overline{g}_{NQD}(\alpha)' W_T \overline{g}_{NQD}(\alpha).$$

Although $\widehat{\alpha}_{K,NQD}$ is a standard generalized method of moments (GMM) estimator, viewing it from the perspective of a covariance structure estimator helps understand the analysis to follow. In standard covariance structure estimation where a typical element of $\overline{g}(\alpha)$ is $\overline{g}_j(\alpha) = 1/T\sum_{t=1}^T y_t y_{t-1} - \mathbb{E}_\alpha y_t y_{t-1}$, each sample autocovariance is a function of the data y_t and hence depends on α^0, but it does not depend on α. In our $\overline{g}_{j,NQD}(\alpha)$, both the sample and model moments depend on α. It differs from the standard formulation of covariance structure estimation, but it is necessary because $\sqrt{T}\left(1/T\sum_{t=1}^T y_t y_{t-1} - \mathbb{E}_\alpha y_t y_{t-1}\right)$ does not obey a central limit theorem at $\alpha = \alpha^0 = 1$.

The NQD solves one of the two problems inherent in least squares estimation by making $\sqrt{T}\overline{g}_{j,NQD}\left(\alpha^0\right)$ asymptotically normal for all $|\alpha^0| \leq 1$. However, the estimator still has nonstandard properties because $\widehat{\gamma}_j(\alpha)$ is stochastically unbounded when $\alpha^0 = 1$ and $\alpha \neq 1$. Thus the moment $\overline{g}_{j,NQD}(\alpha)$ explodes at $\alpha \neq \alpha^0$ when α^0 is unity or in the neighborhood of one. To resolve this problem, suppose γ_0 is known. Define the fixed quasi-differencing (FQD) estimator as

$$\begin{aligned}\overline{g}_{j,FQD}(\alpha) &= \overline{g}_{j,NQD}(\alpha) - \overline{g}_{0,NQD}(\alpha)\\ &= \left(\widehat{\gamma}_j(\alpha) - \widehat{\gamma}_0(\alpha)\right) - \left(\gamma_j(\alpha) - \gamma_0\right), \qquad \textbf{(5)}\\ \widehat{\alpha}_{K,FQD} &= \underset{\alpha}{\operatorname{argmin}}\, \overline{g}_{FQD}(\alpha)' W_T \overline{g}_{FQD}(\alpha).\end{aligned}$$

Obviously, $\overline{g}_{j,FQD}(\alpha^0)$ obeys a central limit theorem. More important is that normalizing $\widehat{\gamma}_j(\alpha)$ by $\widehat{\gamma}_0(\alpha)$ and $\gamma_j(\alpha)$ by γ_0 yields

$$\overline{g}_{j,FQD}(\alpha) = \frac{1}{T}\sum_{t=j+1}^{T} e_t\left(e_{t-j} - e_t\right) + \gamma_0.$$

As shown in Lemma A.2 in the Appendix, $\widehat{\gamma}_j(\alpha)-\widehat{\gamma}_0(\alpha)$ is bounded in probability in the limit even when $\alpha^0=1$ and $\alpha\neq\alpha^0$. Because $\overline{g}_{j,FQD}(\alpha)$ is uniformly bounded in probability for *all* values over which α is optimized and $\alpha^0\in(1-\delta,1]$, the FQD has very different properties from the NQD in the local-to-unity framework.

To gain insight about the importance of normalization, we let W_T be an identity matrix as it does not affect uniformity arguments and this simplifying assumption makes it possible to obtain useful closed form expressions.

PROPOSITION 1. *Let y_t be generated as in equation (2) with error terms satisfying Assumption A. Assume that W_T is a $K\times K$ identity matrix.*

(i) Let $n\in\{1,2\}$ be the number of local minima in the optimization problem $\min_\alpha \sum_{k=1}^K \overline{g}_{k,NQD}^2(\alpha)$. In the local-to-unity framework in which $\alpha^0=1+c/T$ with $c\leq 0$, $\widehat{\alpha}_{K,NQD}$ is superconsistent and

$$\begin{pmatrix} \mathbb{I}\{n=2\}\cdot T^{3/2}\left(\widehat{\alpha}_{K,NQD}-\alpha^0\right)^2 \\ \mathbb{I}\{n=1\}\cdot T\left(\widehat{\alpha}_{K,NQD}-\alpha^0\right) \end{pmatrix} \Rightarrow \begin{pmatrix} \dfrac{-\xi}{\int_0^1 J_c^2(s)ds}\mathbb{I}\{\xi<0\} \\ \dfrac{1+2\int_0^1 J_c(s)dW(s)}{2\int_0^1 J_c^2(s)ds}\mathbb{I}\{\xi>0\} \end{pmatrix}, \tag{6}$$

where J_c is an Ornstein–Uhlenbeck process generated by the Brownian motion W that is independent of $\xi\sim N(0,1/K)$.

(ii) Let γ_0 be the true value of σ^2. For any fixed $K>1$, the estimator $\widehat{\alpha}_{K,FQD}$ is consistent. Furthermore, uniformly over $-1+\delta<\alpha^0\leq 1$:

$$\sqrt{T}\left(\widehat{\alpha}_{K,FQD}-\alpha^0\right)\Rightarrow N\left(0,\sigma_{K,FQD}^2\right), \quad \textit{where}$$

$$\sigma_{K,FQD}^2=\frac{\left(\sum_{k=1}^K\left(\alpha^0\right)^{(k-1)}\right)^2\frac{\mu_4}{\sigma^4}+\sum_{k=1}^K\left(\alpha^0\right)^{2(k-1)}}{\left(\sum_{k=1}^K\left(\alpha^0\right)^{2(k-1)}\right)^2}.$$

The NQD estimator is the basis of the estimators we subsequently investigate. It is consistent and has a data dependent convergence rate. Because the objective function is a polynomial of the fourth order, there are multiple solutions. If the realization of data is such that there is a unique minimum to the optimization problem, the convergence rate is T. If there are two minima, a slower convergence rate of $T^{3/4}$ is obtained. In either case, the distribution of $\widehat{\alpha}_{K,NQD}$ is not asymptotically normal because $\overline{g}_{NQD}(\alpha)$ is not well behaved for all values of α^0. However, the problematic term that frustrates a quadratic expansion of $\widehat{\gamma}_j(\alpha)$ around α^0 is asymptotically collinear with the corresponding term in $\widehat{\gamma}_0(\alpha)$ in the local-to-unity framework. Normalizing each $\widehat{\gamma}_j(\alpha)$ by $\widehat{\gamma}_0(\alpha)$ and $\gamma_j(\alpha)$ by γ_0 results in an FQD estimator whose asymptotic distribution is normal uniformly over $\alpha^0 \in (-1+\delta, 1]$. When $K > 1$, the FQD objective function has only one minimum asymptotically.[3]

The properties for $\widehat{\alpha}_{K,FQD}$ are stated assuming the true value of σ^2 is known. The reason why $\gamma_0 = \sigma^2$ is not freely estimated along with α is that doing so would yield multiple solutions. The objective function is zero not only at the true solution $\alpha = \alpha^0, \sigma^2 = \gamma_0$, but also at $\alpha = 1/\alpha^0, \sigma^2 = \gamma_0/\left(\alpha^0\right)^2$. Without additional information, the FQD cannot uniquely identify α and σ^2.

In practice, the true value of σ^2 is not known, and the FQD estimator is infeasible. This can be overcome by finding another moment that can identify σ^2. Let $\theta = \left(\alpha, \sigma^2\right)$ and consider

$$\overline{g}_{QD}(\theta) = \begin{pmatrix} s^2 - \sigma^2 \\ \widehat{\gamma}_1(\alpha) - \widehat{\gamma}_0(\alpha) - (\gamma_1(\alpha) - s^2) \\ \vdots \\ \widehat{\gamma}_K(\alpha) - \widehat{\gamma}_0(\alpha) - (\gamma_K(\alpha) - s^2) \end{pmatrix}, \tag{7}$$

where $s^2 = (1/T)y'My$ and $M = I_T - z(z'z)^{-1}z'$ is the matrix that projects onto the space orthogonal to z with $z_t = y_{t-1}$. Observe that the first component of $\overline{g}_{QD}(\theta)$ is $s^2 - \sigma^2$ or equivalently $s^2 - \gamma_0(\alpha)$, but not $s^2 - \widehat{\gamma}_0(\alpha)$. Thus, $\overline{g}_{QD}(\theta)$ is not a linear transformation of $\overline{g}_{NQD}(\theta)$. Using $s^2 - \widehat{\gamma}_0(\alpha)$ in the first entry would result in an estimator with the same nonstandard properties as $\widehat{\alpha}_{NQD}$.

Let

$$\widehat{\theta}_{K,QD} = \left(\widehat{\alpha}_{K,QD}, \widehat{\sigma}^2_{K,QD}\right) = \underset{\theta}{\operatorname{argmin}}\, \overline{g}_{QD}(\theta)' W_T \overline{g}_{QD}(\theta).$$

PROPOSITION 2. *Let W_T be a $(K+1) \times (K+1)$ identity matrix. For any fixed $K > 1$, the estimator $\widehat{\alpha}_{K,QD}$ is consistent, and the following convergence holds uniformly over $-1+\delta < \alpha^0 \leq 1$:*

$$\sqrt{T}(\widehat{\alpha}_{K,QD} - \alpha^0) \Rightarrow N(0, \sigma^2_{K,QD}), \quad \text{where } \sigma^2_{K,QD} = \frac{\sum_{k=1}^K \left(\alpha^0\right)^{2(k-1)}}{\left(\sum_{k=1}^K \left(\alpha^0\right)^{2(k-1)}\right)^2}.$$

Estimator $\widehat{\alpha}_{K,QD}$ can also be implemented as a two-step estimator in which s^2 is first obtained and its value would then be used as γ_0 in the moment function $\overline{g}_{FQD}(\alpha) = \left(\overline{g}_{1,FQD}(\alpha), \ldots, \overline{g}_{K,FQD}(\alpha)\right)'$.

The asymptotic variance of $\widehat{\alpha}_{K,QD}$ takes into account the sampling uncertainty of s^2. The surprising aspect of Proposition 2 is that $\widehat{\alpha}_{K,QD}$ does not have an inflated variance as is typical of two-step estimators. Instead, the estimator is more efficient than $\widehat{\alpha}_{K,FQD}$ that has a known σ^2. Pierce (1982) showed in a framework for stationary data that using estimated values of nuisance parameters can yield statistics with smaller variance than if the nuisance parameters were known. This somewhat paradoxical result was also reported by Prokhorov and Schmidt (2009) and Han and Kim (2011) for GMM estimators. Our results suggest that this feature may also arise in the local-to-unity framework.

The closed form expression for the asymptotic variance of $\widehat{\alpha}_{K,QD}$ in Proposition 2 was obtained under the assumption that W_T is an identity matrix. For an arbitrary positive-definite weighting matrix, the asymptotic variance of $\widehat{\alpha}_{K,QD}$ is the (1,1)th element of asymptotic variance matrix

$$A\operatorname{var}(\widehat{\theta}_{K,QD}) = \left(G^{0\prime}WG^0\right)^{-1}G^{0\prime}WS^0WG^0\left(G^{0\prime}WG^0\right)^{-1}, \tag{8}$$

where W, G^0, and S^0 are the probability limits of W_T, the derivative of $\overline{g}_{QD}(\theta)$ with respect to θ evaluated at θ^0, and the asymptotic variance of $\overline{g}_{QD}(\theta^0)$, respectively. The asymptotic variance can thus be estimated as though the GMM estimator were developed in the stationary framework under regularity conditions such as those given in Newey and McFadden (1994). In theory, more efficient estimates can be obtained if W_T is an optimal weighting matrix. However, it has been documented in Abowd and Card (1989) and Altonji and Segal (1996) that an optimal weighting matrix may not be desirable for covariance structure estimation for empirically relevant sample sizes.

The key to the classical properties of $\widehat{\alpha}_{K,QD}$ is the ability to exploit the autocovariance properties of the quasi-differenced data in an appropriate way. QD has a long tradition in econometrics and underlies generalized least squares estimation; see Phillips and Xiao (1998). Canjels and Watson (1997) and Phillips and Lee (1996) found that QD gives more precise estimates of trend parameters when the errors are highly persistent. Pesavento and Rossi (2006) suggest that for such data, QD can improve the coverage of impulse response functions. In both studies, the data are quasi-differenced at $\alpha = \overline{\alpha}$, which is fixed at the value suggested by the local-to-unity framework. In contrast, the FQD and QD simultaneously estimate this parameter and use the normalized autocovariances of the quasi-differenced data for estimation. Notably, both the FQD and the QD have classical properties that hold even in the presence of deterministic terms. Consider data generated as

$$y_t = d_t + x_t, \tag{9a}$$

$$x_t = \alpha^0 x_{t-1} + \varepsilon_t. \tag{9b}$$

The deterministic terms are captured by $d_t = \sum_{j=0}^{r} \psi_j t^j$ where r is the order of the deterministic trend function. In the intercept-only case, $d_t = \psi_0$, and in the linear trend case, $d_t = \psi_0 + \psi_1\, t$. Once the parameters of the trend function are consistently estimated, QD estimation proceeds by replacing y_t with demeaned or detrended data, $\widehat{x}_t = y_t - \widehat{d}_t$. Let $\widehat{e}_t = \widehat{x}_t - \alpha\widehat{x}_{t-k}$. The sample autocovariances can be constructed as

$$\widehat{\gamma}_k(\alpha) = \frac{1}{T}\sum_{t=k+1}^{T} \widehat{e}_t\widehat{e}_{t-k}.$$

Demeaning and detrending do not affect the asymptotic distribution of the QD.[4]

The practical appeal of QD estimation is that asymptotic normality permits standard inference. The usual critical values of ±1.96 and ±1.64 can be used for two-tailed tests at the 5 and 10% significance levels, respectively. We will see in simulations that the size of tests and the coverage of the confidence sets based on the asymptotic normality of $\widehat{\alpha}_{K,QD}$ are stable over the parameter set $\alpha^0 \in (-1+\delta, 1]$. The cost of imposing the stronger assumption of conditional homoskedasticity seems well justified.

To recapitulate, the proposed QD estimation of the AR(1) model is based on two simple premises: first, that for all $j \geq 1$, $\mathbb{E}(\varepsilon_t \varepsilon_{t-j}) = 0$ and its sample analogue obeys a central limit theorem, and, second, that the objective function is uniformly bounded in probability for all values of α and α^0. The idea can be used whenever the variables can be quasi-differenced to form suitably normalized moment conditions that satisfy these two properties. The next two sections consider the AR(p) model and predictive regressions, respectively. We then show that the quasi-differenced variables can be serially correlated and that the QD framework can be used in nonlinear estimations.

3. AR(*p*) MODELS

Consider the data generating process

$$y_t = \alpha^0 y_{t-1} + \sum_{j=1}^{p-1} b_j^0 \Delta y_{t-j} + \varepsilon_t. \tag{10}$$

Let $\beta = (\alpha, b_1, \ldots, b_{p-1})$ be a $p \times 1$ parameter vector of interest. The true parameter vector is denoted β^0, and the correct lag length is denoted p. Let $|\lambda_1| \leq |\lambda_2| \ldots \leq |\lambda_p|$ be defined implicitly by the identity

$$1 - \alpha L - \sum_{j=1}^{p-1} b_j L^j (1-L) = (1-\lambda_1 L)\ldots(1-\lambda_p L).$$

We restrict the parameter set in such a way that the $p-1$ smallest roots do not exceed δ in absolute value for some fixed $0 < \delta < 1$. If the largest root exceeds δ in absolute value, then it is positive and not larger than one.

DEFINITION 2. *The parameter set $\Re_\delta$ consists of all β such that the corresponding roots satisfy the following two conditions:*

(i) $|\lambda_{p-1}| < \delta$;
(ii) if $\lambda_p \in \mathbb{R}$, *then* $-\delta \leq \lambda_p \leq 1$.[5]

Define the quasi-differenced series e_t by

$$e_t = y_t - \alpha y_{t-1} - \sum_{j=1}^{p-1} b_j \Delta y_{t-j}.$$

Obviously, $e_t = \varepsilon_t$ is white noise when $\beta = \beta^0$, but e_t is in general serially correlated. Thus, as in the AR(1) model, the model-implied autocovariances satisfy

$$\gamma_j(\beta) = \mathbb{E}_\beta\left(e_t e_{t-j}\right) = 0, \qquad j \geq 1 \quad \forall \beta \in \mathfrak{R}_\delta$$

with $\gamma_0(\beta) = \sigma^2$. The sample autocovariances of e_t are

$$\widehat{\gamma}_j(\beta) = \frac{1}{T} \sum_{t=j+p+1}^{T} e_t e_{t-j}.$$

Let $s^2 = (1/T)y'My$ where M projects onto the space orthogonal to the one spanned by $X_t = (y_{t-1}, \Delta y_{t-1}, \ldots, \Delta y_{t-p+1})'$, $t = 1, \ldots, T$. Let γ_0 be the true value of σ^2. Define

$$\widehat{\beta}_{K,FQD} = \underset{\beta}{\operatorname{argmin}}\, \overline{g}_{FQD}(\beta)' W_T \overline{g}_{FQD}(\beta),$$

where

$$\overline{g}_{FQD}(\beta) = \begin{pmatrix} \overline{g}_{1,FQD}(\beta) \\ \vdots \\ \overline{g}_{K,FQD}(\beta) \end{pmatrix} = \begin{pmatrix} \widehat{\gamma}_1(\beta) - \widehat{\gamma}_0(\beta) - (\gamma_1(\beta) - \gamma_0) \\ \vdots \\ \widehat{\gamma}_K(\beta) - \widehat{\gamma}_0(\beta) - (\gamma_K(\beta) - \gamma_0) \end{pmatrix}.$$

Define

$$\left(\widehat{\beta}_{K,QD}, \widehat{\sigma}^2_{K,QD}\right) = \underset{\beta,\sigma^2}{\operatorname{argmin}}\, \overline{g}_{QD}\left(\beta,\sigma^2\right)' W_T \overline{g}_{QD}\left(\beta,\sigma^2\right),$$

where

$$\overline{g}_{QD}\left(\beta,\sigma^2\right) = \begin{pmatrix} \overline{g}_{0,QD}(\beta) \\ \overline{g}_{1,QD}(\beta) \\ \vdots \\ \overline{g}_{K,QD}(\beta) \end{pmatrix} = \begin{pmatrix} s^2 - \sigma^2 \\ \widehat{\gamma}_1(\beta) - \widehat{\gamma}_0(\beta) - (\gamma_1(\beta) - s^2) \\ \vdots \\ \widehat{\gamma}_K(\beta) - \widehat{\gamma}_0(\beta) - (\gamma_K(\beta) - s^2) \end{pmatrix}.$$

PROPOSITION 3. *Let y_t be generated as in equation (10) with error terms satisfying Assumption A. Let $a_k = \mathbb{E}[(X_{t+k} + X_{t-k} - 2X_t)\varepsilon_t]$ and $G = \left(\sum_{k=1}^K a_k a_k'\right)^{-1}$. For any fixed $K > p > 1$, the estimators $\widehat{\beta}_{K,QD}$ and $\widehat{\beta}_{K,FQD}$ are consistent. Furthermore, when W_T is an identity weighting matrix, the following results hold uniformly over $\beta^0 \in \mathfrak{R}_\delta$:*

(i) $\sqrt{T}(\widehat{\beta}_{K,FQD} - \beta^0) \Rightarrow N(0, \Sigma_{K,FQD})$, where $\Sigma_{K,FQD} = \sigma^4 G + \mu_4 G \left(\sum_{k=1}^K a_k\right)\left(\sum_{k=1}^K a_k\right)' G$;

(ii) $\sqrt{T}(\widehat{\beta}_{K,QD} - \beta^0) \Rightarrow N(0, \Sigma_{K,QD})$, where $\Sigma_{K,QD} = \sigma^4 G$.

The proof is a generalization of Propositions 1 and 2. A sketch of the arguments is as follows. From the definition that $e_t(\beta) = \varepsilon_t + (\beta^0 - \beta)' X_t$, we have

$$\widehat{\gamma}_j(\beta) = \frac{1}{T}\sum_{t=j+p+1}^{T} \varepsilon_t \varepsilon_{t-j} + \left(\beta^0 - \beta\right)' \frac{1}{T}\sum_{t=j+p+1}^{T} \left(X_t \varepsilon_{t-j} + X_{t-j}\varepsilon_t\right)$$
$$+ \left(\beta^0 - \beta\right)' \frac{1}{T}\sum_{t=j+p+1}^{T} X_t X'_{t-j}\left(\beta^0 - \beta\right).$$

The moment function can be rewritten as

$$\overline{g}_{j,FQD}(\beta) = A_{j,FQD} + \left(\beta^0 - \beta\right)' B_{j,FQD} + \left(\beta^0 - \beta\right)' C_{j,FQD}\left(\beta^0 - \beta\right).$$

The thrust of the proof is to show that for each $1 \leq j \leq K$ uniformly over $\mathfrak{R}_\delta$,

$$A_{j,FQD} = \left(\frac{1}{T}\sum_{t=j+p+1}^{T} \varepsilon_t \varepsilon_{t-j} - \gamma_j(\beta)\right) - \left(\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_t^2 - \gamma_0\left(\beta^0\right)\right) = O_p\left(T^{-1/2}\right), \tag{11}$$

$$B_{j,FQD} = \frac{1}{T}\sum_{t=j+p+1}^{T} \left(X_t \varepsilon_{t-j} + X_{t-j}\varepsilon_t - 2X_t\varepsilon_t\right) \rightarrow^p a_j, \tag{12}$$

$$C_{j,FQD} = \frac{1}{2T}\sum_{t=j+p+1}^{T} \left(X_t X'_{t-j} + X_{t-j}X'_t - 2X_t X'_t\right) = O_p(1). \tag{13}$$

Equations (11)–(13) imply that the function $\overline{g}_{j,FQD}(\beta)$ is bounded in probability uniformly for all β in the optimization set and $\beta_0 \in \mathfrak{R}_\delta$. It also follows from equations (11)–(13) that

$$\frac{\partial \overline{g}_{j,FQD}}{\partial \beta}\left(\widehat{\beta}_{K,FQD}\right) \rightarrow^p a_j$$

and

$$\sqrt{T}\,\overline{g}_{k,FQD}\left(\widehat{\beta}_{K,FQD}\right) = \sqrt{T} A_{k,FQD} + a'_k \sqrt{T}\left(\widehat{\beta}_{K,FQD} - \beta^0\right) + o_p(1).$$

The first-order condition for the optimization problem implies

$$\sqrt{T}\left(\widehat{\beta}_{K,FQD} - \beta^0\right) = G^{-1}\sum_{j=1}^{K} \sqrt{T} A_{j,FQD} a_j + o_p(1).$$

In view of equation (4), $\sqrt{T} A_{j,FQD} = \sqrt{T}\,\overline{g}_{j,FQD}\left(\beta^0\right) \Rightarrow \xi_j - \xi_0$ uniformly over $\mathfrak{R}_\delta$; part (i) of the proposition follows. Part (ii) uses a similar argument with one exception: $A_{j,QD} = A_{j,FQD} + s^2 - \sigma^2 = \frac{1}{T}\sum_{t=j+1}^{T} \varepsilon_t \varepsilon_{t-j}$. As in the AR(1) case, $\widehat{\beta}_{K,QD}$ has a smaller variance than $\widehat{\beta}_{K,FQD}$. Furthermore, one can use other weighting matrices in the estimation. The asymptotic variance of $\widehat{\beta}_{K,QD}$ can be computed from the expression given in equation (8).

4. PREDICTIVE REGRESSIONS

Consider the predictive regression with scalar predictor x_{t-1}:

$$y_t = \beta^0 x_{t-1} + \varepsilon_{yt}, \tag{14a}$$

$$x_t = \alpha^0 x_{t-1} + \varepsilon_{xt}. \tag{14b}$$

If $\alpha^0 = 1$, then $\left(1; -\beta^0\right)$ is a cointegrating vector and ordinary least squares (OLS) provide superconsistent estimates but inference is nonstandard. Unfortunately, the finite-sample distribution of $\widehat{\beta}_{OLS}$ is not well approximated by the normal distribution if x_t is highly persistent. The challenge is how to conduct inference robust to the dynamic properties of x_t. Let $\varepsilon_t = \left(\varepsilon_{yt}, \varepsilon_{xt}\right)'$ be a martingale-difference sequence with $\mathbb{E}(\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-1}) = \Omega^0 = \begin{pmatrix} \sigma_{yy} & \sigma_{yx} \\ \sigma_{yx} & \sigma_{xx} \end{pmatrix}$. Consider quasi-differencing the data at $\theta = (\beta, \alpha)$ to obtain

$$e_{yt} = y_t - \beta x_{t-1},$$
$$e_{xt} = x_t - \alpha x_{t-1}.$$

Now $Y_t = \theta^0 x_{t-1} + \varepsilon_t$ where $Y_t = \begin{pmatrix} y_t \\ x_t \end{pmatrix}$ and $\theta^0 = (\beta^0, \alpha^0)'$. Let $e_t = \begin{pmatrix} e_{yt} \\ e_{xt} \end{pmatrix}$. Then

$$e_t = (\theta^0 - \theta) x_{t-1} + \varepsilon_t.$$

Let $\Gamma_j(\theta) = \mathbb{E}_\theta(e_t e_{t-j}')$ where $\mathbb{E}_\theta$ is the expectation taken under the assumption that θ is the true value. The model implies

$$\Gamma_j(\theta) = 0, \qquad j \neq 0,$$
$$\Gamma_0(\theta) = \Omega.$$

The sample autocovariance at lag j is

$$\widehat{\Gamma}_j(\theta) = \frac{1}{T} \sum_{t=j+1}^{T} e_t e_{t-j}'.$$

A central limit theorem applies to $\sqrt{T}\, \widehat{\Gamma}_j\left(\theta^0\right)$. Evaluating Γ_0 at the true value of Ω and letting $S = (1/T) Y' M Y$, $M = I_T - z(z'z)^{-1} z'$, $z_t = x_{t-1}$, we can define, for $j = 1, \ldots, K$:

$$\overline{g}_{j,FQD}(\theta) = \operatorname{vec}\left(\widehat{\Gamma}_j(\theta) - \widehat{\Gamma}_0(\theta) - (\Gamma_j(\theta) - \Gamma_0)\right).$$

Let $\overline{g}_{FQD}(\theta) = \left(\overline{g}_{1,FQD}(\theta)', \ldots, \overline{g}_{K,FQD}(\theta)'\right)'$. The FQD estimator is

$$\widehat{\theta}_{K,FQD} = \arg\min_{\theta} \overline{g}_{FQD}(\theta)' W_T \overline{g}_{FQD}(\theta).$$

Analogously, let $\overline{g}_{QD}(\theta,\Omega)=(\overline{g}_{0,QD}(\theta,\Omega)',\dots,\overline{g}_{K,QD}(\theta,\Omega)')'$ where

$$\overline{g}_{j,QD}(\theta,\Omega)=\operatorname{vec}\left(\widehat{\Gamma}_j(\theta)-\widehat{\Gamma}_0(\theta)-(\Gamma_j(\theta)-S)\right),\qquad j\geq 1,$$
$$\overline{g}_{0,QD}(\theta,\Omega)=\operatorname{vech}(S-\Omega).$$

Define the QD estimator as

$$\left(\widehat{\theta}_{K,QD},\widehat{\Omega}_{K,QD}\right)=\arg\min_{\theta,\Omega}\overline{g}_{QD}(\theta,\Omega)'W_T\overline{g}_{QD}(\theta,\Omega).$$

PROPOSITION 4. *Suppose that the data are generated according to formulas (14a) and (14b). Suppose also that the error terms are stationary martingale-difference sequence with $\mathbb{E}(\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1})=\Omega^0$ and finite four moments. Define $a_j=\mathbb{E}[x_{t-1}(\varepsilon_{t-j}-\varepsilon_t)]$. Then for any fixed $K>1$, the estimators $\widehat{\theta}_{K,FQD}$ and $\widehat{\theta}_{K,QD}$ are consistent. Furthermore, when W_T is an identity matrix, the following asymptotic results hold uniformly over all possible values of β and uniformly over all possible values of $\alpha\in(-1+\delta,1]$:*

(i) Let $\Gamma_0=\Omega^0$. Then $\sqrt{T}(\widehat{\theta}_{K,FQD}-\theta^0)\Rightarrow N(0,\Sigma_{K,FQD})$, where

$$\Sigma_{K,FQD}=\left(\frac{1}{\sum_{k=1}^K a_k'a_k}\right)^2\sum_{k=1}^K(a_k'\Omega^0a_k)\Omega^0$$
$$+\left(\frac{1}{\sum_{k=1}^K a_k'a_k}\right)^2\mathbb{E}\left[(\varepsilon_t'\sum_{k=1}^K a_j)^2\varepsilon_t\varepsilon_t'\right];$$

(ii) $\sqrt{T}(\widehat{\theta}_{K,QD}-\theta^0)\Rightarrow N(0,\Sigma_{K,QD})$ where

$$\Sigma_{K,QD}=\left(\frac{1}{\sum_{k=1}^K a_k'a_k}\right)^2\sum_{k=1}^K(a_k'\Omega^0a_k)\Omega^0.$$

As in the case of autoregressions, the FQD moments alone cannot globally identify both θ and Ω. Thus, the properties of $\widehat{\theta}_{K,FQD}$ are stated by evaluating Ω at the true value of Ω^0. Proposition 4 shows that $\widehat{\theta}_{K,QD}$ has classical properties in both the stationary and the local-to-unity frameworks and is more efficient than the estimator $\widehat{\theta}_{K,FQD}$ that uses the known Ω. The QD can be implemented as a sequential estimator in which the covariance matrix is computed for shocks obtained from two least squares regressions: one by regressing y_t on x_t to get $\widehat{e}_{yt}$, and another autoregression in x_t to obtain $\widehat{e}_{xt}$.

Proposition 4 has useful implications for applied work because there does not exist an estimator that is robust to the persistent properties of the predictors. The approach of Jansson and Moreira (2006) relies on model-specific conditional critical values, and, in any event, their inference procedure does not yield an estimator per se. In contrast, the QD estimator is simple and robust.

The predictive regression can be generalized to accommodate stationary and predetermined regressors, z_t. Suppose the data generating process is

$$y_t = x_{t-1}\beta^0 + z_t\gamma^0 + \varepsilon_{yt},$$

$$x_t = \alpha^0 x_{t-1} + \varepsilon_{xt}, \qquad (\varepsilon_{yt}, \varepsilon_{xt})' \sim (0, \Omega), \qquad \Omega = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{xy} & \sigma_x^2. \end{pmatrix}.$$

Let $\theta = (\beta, \alpha)$ and as before, $\Gamma_j = \mathbb{E}_\theta(\varepsilon_t\varepsilon'_{t-j}) = 0$ for all $j \neq 0$ with $\Gamma_0 = \Omega^0$. Let $\widehat{\gamma}_{OLS}$ be obtained from least squares regression of y_t on x_{t-1} and z_t and let $\widehat{\Omega}_{OLS}$ be the estimated covariance matrix of the errors. Because z_t is stationary, the estimator $\widehat{\gamma}_{OLS}$ is $\sqrt{T}$-consistent and asymptotically normal uniformly over $\alpha \in (-1+\delta, 1]$. Define the quasi-differenced sequence

$$e_{yt} = y_t - x_{t-1}\beta - z_t\widehat{\gamma}_{OLS}, \qquad e_{xt} = x_t - \alpha x_{t-1}, \qquad e_t = \begin{pmatrix} e_{yt} \\ e_{xt} \end{pmatrix}.$$

Let $\widehat{\Gamma}_j(\theta) = \frac{1}{T}\sum_{t=j+1}^T e_t e'_{t-j}$ and define $\overline{g}_{j,QD}(\theta, \Omega)$ as in the absence of z_t. Using arguments analogous to Proposition 4, it can be shown that $\widehat{\theta}_{K,QD}$ is still $\sqrt{T}$-consistent and asymptotically normal. Proposition 4 assumes that the regression error ε_{yt} is white noise. This is not restrictive as lags of Δy_t and Δx_t can be added to z_t to control for residual serial correlation.

5. NONLINEAR MODELS AND MINIMUM DISTANCE ESTIMATION

So far, the QD framework has been used to estimate linear models, where the model autocovariances are such that $\gamma_j(\theta) = 0$ for all θ assumed to be the true value and $j \geq 1$. The analysis also holds if $\gamma_j(\theta)$ equals a constant vector other than zero provided that the constant vector is known or can be computed numerically. For example, if x_t is an ARMA(1,1) instead of an AR(1), $\gamma_j(\theta)$ will depend on the parameters of the model. Another example is DSGE models, which we now consider.

To fix ideas, consider the simple one-sector stochastic growth model presented in Uhlig (1999). Let Q_t, C_t, K_t, I_t be output, consumption, capital stock, and investment, respectively. The problem facing the central planner is to maximize expected utility $\mathbb{E}_{t-1}\sum_{t=0}^\infty (1+\rho)^{-t}\log C_t$ subject to the constraints $Q_t = K_{t-1}^\psi Z_t^{1-\psi} = C_t + I_t$ and $K_t = (1-\delta)K_{t-1} + I_t$ where $u_t = \log Z_t$ evolves as

$$u_t = \alpha u_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \left(0, \sigma^2\right).$$

Denote the deviation of a variable from its mean by lowercase letters. Let $Y_t = (y_{1t}, \ldots, y_{Nt})'$ be the collection of endogenous variables in the model (such as consumption, output, etc.). As shown in Uhlig (1999), this simple model has an analytic solution:

$$k_t = v_{kk}k_{t-1} + v_{kz}u_t,$$

where $\upsilon_{kk} < 1$ does not depend on α but υ_{kz} depends on α. For each $y_{nt} \in Y_t$,

$$(1 - \upsilon_{kk} L)(1 - \alpha L) y_{nt} = u_t + \vartheta_n u_{t-1}$$

is an ARMA(2,1) with a moving average (MA) parameter ϑ_n that is a function of the structural parameters. Note that all series in Y_t have the same autoregressive dynamics as k_t. The parameters of the linearized solution are $\beta = (\upsilon_{kk}, \upsilon_{kz}, \vartheta_1, \ldots, \vartheta_N)$. The parameters of the model are $\theta = (\psi, \alpha, \sigma^2, \rho, \delta)$. Let Θ be a compact set containing possible values of θ.

In analysis of DSGE models, whether the shocks have permanent or transitory effects matters for how a model is to be linearized. For this reason, researchers typically need to decide whether to difference the data ahead of estimation even though it is understood that the assumption affects the estimates and policy analysis. To date, there does not exist an estimator of DSGE models that has classical properties for all values of α within the likelihood framework because the likelihood function is not well defined when the data are nonstationary.[6]

We propose to estimate the parameters of the model without making a priori assumptions about the degree of persistence of the shocks. We use the fact that the features of covariance stationary processes are completely summarized by their second moments. Conveniently, the software DYNARE automatically calculates the covariance structure of the data. Even though the analysis is not likelihood based, priors can still be incorporated using the approach of Chernozhukov and Hong (2003).[7] The key is to construct the moments $\overline{g}(\theta)$ appropriately.

Two variations of the QD framework are considered. The first method proceeds as follows. For given θ, let $e_t = Y_t - \alpha Y_{t-1}$ with $\Gamma_j(\theta) = \mathbb{E}(e_t e_t')$. Define the moment $\omega_j(\theta) = (\Gamma_j(\theta) - \Gamma_0(\theta))$ whose sample analogue is $\widehat{\omega}_j(\theta) = (\widehat{\Gamma}_j(\theta) - \widehat{\Gamma}_0(\theta))$. Note that because $e_t(\theta)$ can be serially correlated, $\Gamma_j(\theta)$ need not be a null matrix as in the applications considered thus far. Let $\overline{g}_{QD}(\theta) = (\overline{g}_{1,QD}(\theta)', \ldots, \overline{g}_{K,QD}(\theta)')'$ where $\overline{g}_{j,QD}(\theta) = \text{vec}(\widehat{\omega}_j(\theta) - \omega_j(\theta))$. The QD estimator considered in Gorodnichenko and Ng (2010) is defined as[8]

$$\widehat{\theta}_{K,QD} = \underset{\theta \in \Theta}{\text{argmin}}\, \overline{g}_{QD}(\theta)' \overline{g}_{QD}(\theta). \qquad \textbf{(15)}$$

As written, $\widehat{\theta}_{K,QD}$ is an equally weighted estimator. An optimal weighting matrix can be used subject to constraints imposed by stochastic singularity. In the one-shock stochastic growth model considered, the autocovariance at lag one of both output and consumption can be used to construct an efficient $\widehat{\theta}_{K,QD}$, but additional autocovariances will not add independent information. In contrast, the use of data for both output and consumption in likelihood estimation would not even be possible.

Another QD-based estimator can be obtained if we entertain the possibility of a reduced form model. Consider a finite-order AR(p) model:

$$y_t = a^0 y_{t-1} + \sum_{j=1}^{p-1} b_j^0 \Delta y_{t-j} + \varepsilon_{tp}, \qquad \textbf{(16)}$$

where $\beta^0 = \left(a^0, b_1^0, \ldots, b_{p-1}^0\right) = \beta\left(\theta^0\right)$ are the true "reduced form" parameters that can be computed analytically or numerically. We also need the following assumption.

Assumption B (Identification)**.**

(a) There is a unique θ^0 such that $\beta\left(\theta^0\right) = \beta^0$;
(b) the function $\beta(\theta)$ is twice continuously differentiable;
(c) $B\left(\theta^0\right) = \nabla\beta\left(\theta^0\right)$ has full rank $k = \dim(\theta) \leq p$.

The method proceeds as follows. For $y_{nt} \in Y_t$, define $e_{nt}(\beta) = y_{nt} - a y_{nt-1} - \sum_{j=1}^{p-1} b_j y_{nt-j}$ where $\beta = (a, b_1, \ldots, b_{p-1})$. Note that Y_t is now quasi-differenced using the "reduced form" parameter β instead of the structural parameter α as in method 1. Once the data are quasi-differenced, estimation proceeds by defining $\omega_j(\theta; p) = (\Gamma_j(\theta; p) - \Gamma_0(\theta; p))$ with $\overline{g}_{j,QD}(\theta; p) = \widehat{\omega}_j(\theta; p) - \omega_j(\theta; p)$ and sample analogue as in (15). Let $\overline{g}_{QD}(\theta; p) = (\overline{g}_{1,QD}(\theta; p)', \ldots, \overline{g}_{K,QD}(\theta; p)')'$. The estimator is

$$\widehat{\theta}_{K,QD} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \overline{g}_{QD}(\theta; p)'\overline{g}_{QD}(\theta; p). \tag{17}$$

Because β is p-dimensional, this second estimator also depends on the choice of p. Because e_{nt} is not necessarily exactly white noise, $\Gamma_j(\theta)$ will not be zero. However, its autocovariances can be computed for any given θ.

We have presented two uses of the QD framework that can yield estimators that are robust to nonstationary exogenous variables in DSGE models. The data-dependent transformations allow us to construct moments that are uniformly bounded. Applying the central limit theorem to the sample moments yields estimators with classical properties.

6. SIMULATIONS

We consider the finite-sample properties of OLS, FQD with γ_0 fixed at the true σ_0^2, and QD. Even though the FQD estimator is infeasible in practice, it is a useful benchmark. The simulations are based on 2,000 replications. We use the standard Newey–West plug-in estimator for the variance of the moments. As starting values, we use 0.9 times the true values of the parameters. The QD estimator requires evaluation of the model implied autocovariances $\Gamma_j(\theta)$. This is straightforward once a model is cast in a state space. For example, the system

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} 0 & \beta \\ 0 & \alpha \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{xt} \end{bmatrix}$$

in quasi-differenced form is

$$\begin{bmatrix} e_{yt}(\theta) \\ e_{xt}(\theta) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} e_{yt-1}(\theta) \\ e_{xt-1}(\theta) \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{xt} \end{bmatrix}.$$

More generally, every autoregressive moving average model has a state-space representation from which a state-space model for the quasi-differenced data can be expressed as

$$w_t = D_0 w_{t-1} + D_1 \varepsilon_t,$$

where w_t includes $e_t(\theta)$ (and possibly its lags) and ε_t is the set of exogenous white noise shocks with variance $\boldsymbol{\Omega}_\varepsilon$. The variance matrix $\boldsymbol{\Omega}_w(0) = \mathbb{E}(w_t w_t')$ can be computed by iterating the equation

$$\boldsymbol{\Omega}_w^{(i)}(0) = D_0 \boldsymbol{\Omega}_w^{(i-1)}(0) D_0' + D_1 \boldsymbol{\Omega}_\varepsilon D_1' \tag{18}$$

until convergence. The autocovariance matrices can then be computed as $\boldsymbol{\Omega}_w(j) = D_0^j \boldsymbol{\Omega}_w(0)$. Now $\Gamma_j(\theta)$ are submatrices of $\boldsymbol{\Omega}_w(j)$. If we are only interested in computing the moments $w_t^d \subset w_t$, we iterate equation (18) until the block that corresponds to w_t^d converges; that is, $\|\boldsymbol{\Omega}_{w^d}^{(i)}(0) - \boldsymbol{\Omega}_{w^d}^{(i-1)}(0)\| < c$.

Data are generated from the AR(2) model:

$$\left(1 - \lambda_1^0 L\right)\left(1 - \lambda_2^0 L\right) y_t = \varepsilon_t, \qquad \varepsilon_t \sim N(0,1).$$

The process can be written as

$$y_t = \alpha^0 y_{t-1} + b^0 \Delta y_{t-1} + \varepsilon_t.$$

The parameter of interest is $\alpha^0 = \lambda_1^0 + \lambda_2^0 - \lambda_1^0 \lambda_2^0$ with $b^0 = \lambda_1^0 \lambda_2^0$. The OLS estimate of α has a nonstandard distribution when the roots are unity, in which case $\alpha^0 = 1$.

We estimate an AR(2) model when $\lambda_2^0 = 0$ (Table 1). Demeaned data are used to compute the sample autocovariances in the intercept case, and linearly detrended data are used in the linear trend case. We report the mean of the QD, FQD, and OLS estimates when $T = 200$ and 500, the J test for overidentifying restrictions, and the finite-sample power for one-sided t-tests evaluated at $\alpha = \alpha^0 - 0.05$.

Table 1 shows that all three estimators are precise when $\alpha^0 \leq 0.8$. The t-statistic for the null hypothesis that $\alpha = \alpha^0$ for all three estimators has rejection rates close to the nominal size of 0.05 when $\alpha^0 \leq 0.8$. The picture is, however, very different at larger values of α^0. The FQD has slightly smaller bias but is much less efficient. Although OLS has the largest bias, its root-mean-squared error (RMSE) is much smaller than the RMSE of FQD. The QD is neither the most accurate nor the most efficient but has RMSE closer to OLS and much smaller than the RMSE of FQD, in support of Proposition 2.

Efficiency of OLS comes at the cost of size distortion, however. At $T = 200$, the OLS-based t-statistic has a rejection rate of 0.473 when $\alpha^0 = 1$ and 0.15 when $\alpha^0 = 0.95$, much larger than the nominal size of 0.05. Even at $T = 500$, the rejection rates are 0.462 and 0.127, well above the nominal rate of 5%. The rejection rates for the FQD and QD are 0.107 and 0.033 when $T = 200$ and are 0.066 and 0.023 when $T = 500$, much closer to the nominal size of 0.05. The QD has accurate rejection rates that are always around 0.05 for all values of α^0, but

TABLE 1. AR(1) model

		QD					FQD					OLS			
				t-test		J-test			t-test		J-test			t-test	
T	α	Mean	RMSE	Size	Power	Size	Mean	RMSE	Size	Power	Size	Mean	RMSE	Size	Power
(a) Intercept model															
200	1.00	0.975	0.059	0.107	0.533	0.085	0.991	0.122	0.033	0.098	0.078	0.973	0.035	0.473	1.000
200	0.98	0.960	0.052	0.102	0.484	0.083	0.971	0.113	0.032	0.085	0.069	0.956	0.035	0.219	1.000
200	0.95	0.934	0.050	0.080	0.428	0.101	0.942	0.114	0.028	0.066	0.095	0.929	0.036	0.150	0.969
200	0.90	0.883	0.054	0.098	0.417	0.088	0.888	0.117	0.030	0.084	0.083	0.880	0.040	0.111	0.687
200	0.80	0.782	0.058	0.091	0.359	0.075	0.794	0.126	0.035	0.081	0.077	0.782	0.050	0.097	0.413
500	1.00	0.990	0.030	0.066	0.748	0.069	0.997	0.072	0.023	0.136	0.061	0.989	0.014	0.462	1.000
500	0.98	0.974	0.028	0.064	0.681	0.056	0.977	0.073	0.035	0.150	0.054	0.971	0.015	0.138	1.000
500	0.95	0.943	0.030	0.076	0.676	0.077	0.946	0.072	0.037	0.154	0.071	0.941	0.019	0.127	1.000
500	0.90	0.894	0.031	0.073	0.598	0.068	0.896	0.074	0.029	0.144	0.067	0.893	0.022	0.082	0.930
500	0.80	0.793	0.035	0.085	0.545	0.047	0.792	0.080	0.043	0.146	0.052	0.792	0.030	0.082	0.687
(b) Linear trend model															
200	1.00	0.964	0.063	0.224	0.641	0.108	0.991	0.112	0.025	0.091	0.101	0.950	0.057	0.775	1.000
200	0.98	0.954	0.058	0.164	0.547	0.099	0.967	0.114	0.028	0.081	0.098	0.940	0.050	0.454	1.000
200	0.95	0.926	0.056	0.146	0.511	0.090	0.945	0.111	0.022	0.078	0.077	0.916	0.047	0.278	0.986
200	0.90	0.876	0.059	0.143	0.481	0.108	0.892	0.119	0.032	0.071	0.105	0.869	0.049	0.180	0.775
200	0.80	0.775	0.060	0.121	0.432	0.075	0.788	0.127	0.043	0.081	0.078	0.770	0.054	0.133	0.512
500	1.00	0.985	0.033	0.137	0.775	0.072	0.999	0.069	0.020	0.160	0.063	0.980	0.023	0.777	1.000
500	0.98	0.972	0.029	0.080	0.717	0.082	0.976	0.071	0.041	0.153	0.081	0.967	0.018	0.245	1.000
500	0.95	0.942	0.031	0.086	0.678	0.073	0.950	0.074	0.034	0.161	0.068	0.938	0.021	0.185	1.000
500	0.90	0.891	0.032	0.090	0.654	0.061	0.898	0.076	0.041	0.151	0.068	0.889	0.025	0.123	0.957
500	0.80	0.792	0.034	0.089	0.567	0.043	0.797	0.078	0.025	0.122	0.054	0.789	0.030	0.090	0.729

Note: The data generating process is an AR(1) model, whereas the AR(2) model is fitted. Three autocovariances are used in QD estimation. The t-test and J-test sizes are for the 5% level. Power of the t-test is computed for the null of $H_0 : \alpha = \alpha_0 - 0.05$. Additional results are available in the online Appendix (www.columbia.edu/∼sn2294/pub.html).

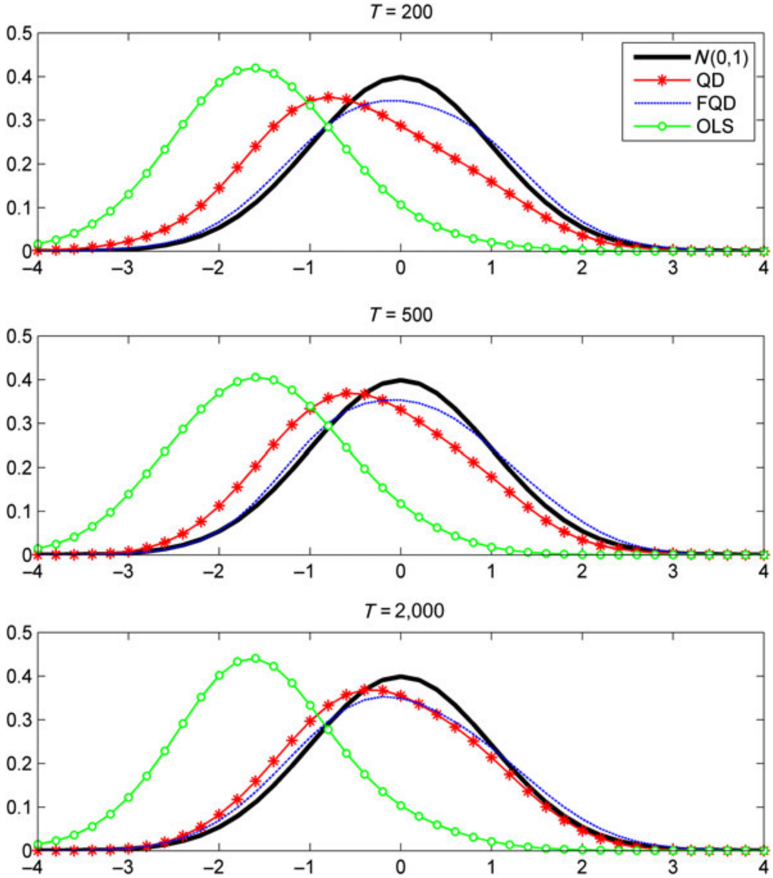

FIGURE 1. Distribution of the t-statistic for the largest autoregressive root in the intercept-only model with $\alpha^0 = 1$. See Table 1 and the text for more details.

it has less power than OLS. Figure 1 plots the distribution of t-statistics for QD at $T = 200$, $T = 500$, and 2000. The normal approximation to the finite-sample distribution is good.

We also consider overparameterized AR(p) models and predictive regressions. The general result is that the QD estimates are precise and the t-statistic is well approximated by the normal distribution for all values of the persistent parameter. To conserve space, these results are available on request.

Next, data are generated from the stochastic growth model presented in Section 5. We fix the true value of capital intensity ψ to 0.25 and σ^2 to 1 and consider five values of α: 1, 0.98, 0.95, 0.9, 0.8. We use data on consumption to estimate ψ, α, and σ^2 using the second method discussed in Section 5. This consists of solving for ψ, α, and σ from the six autocovariances of $e_t(\beta)$, where β are the parameters of an AR(3) model. For the sake of comparison, we also use the Kalman filter

TABLE 2. DSGE model, capital intensity ψ

		QD					MLE			
				t-test		J-test			t-test	
T	α	Mean	RMSE	Size	Power	Size	Mean	RMSE	Size	Power
(a) Intercept model										
200	1.00	0.255	0.122	0.077	0.245	0.088	0.181	0.117	0.309	0.577
200	0.98	0.255	0.120	0.098	0.259	0.085	0.220	0.094	0.142	0.409
200	0.95	0.268	0.133	0.076	0.238	0.095	0.246	0.094	0.076	0.294
200	0.90	0.284	0.147	0.066	0.201	0.094	0.266	0.114	0.065	0.235
200	0.80	0.289	0.144	0.046	0.148	0.079	0.285	0.131	0.065	0.184
500	1.00	0.260	0.072	0.050	0.304	0.049	0.173	0.110	0.517	0.830
500	0.98	0.261	0.071	0.059	0.322	0.053	0.231	0.058	0.151	0.626
500	0.95	0.262	0.073	0.066	0.331	0.049	0.245	0.054	0.076	0.522
500	0.90	0.266	0.082	0.052	0.306	0.043	0.253	0.062	0.053	0.428
500	0.80	0.281	0.110	0.035	0.197	0.035	0.270	0.096	0.062	0.315
(b) Linear trend model										
200	1.00	0.255	0.123	0.088	0.264	0.093	0.223	0.094	0.153	0.399
200	0.98	0.267	0.128	0.075	0.222	0.090	0.243	0.091	0.102	0.315
200	0.95	0.272	0.138	0.085	0.233	0.090	0.257	0.097	0.065	0.249
200	0.90	0.282	0.145	0.066	0.204	0.083	0.273	0.113	0.057	0.205
200	0.80	0.298	0.146	0.053	0.130	0.081	0.296	0.135	0.057	0.150
500	1.00	0.259	0.071	0.050	0.308	0.041	0.209	0.079	0.324	0.730
500	0.98	0.262	0.072	0.057	0.309	0.040	0.237	0.059	0.139	0.593
500	0.95	0.263	0.073	0.058	0.315	0.039	0.250	0.056	0.069	0.477
500	0.90	0.268	0.082	0.042	0.280	0.034	0.259	0.064	0.042	0.387
500	0.80	0.284	0.108	0.024	0.162	0.028	0.277	0.097	0.053	0.295

Note: The true value of capital intensity is $\psi = 0.25$. The observed series is consumption. *QD* uses an OLS estimate of the standard deviation of innovations consumption for QD estimation. MLE corresponds to the maximum likelihood estimation (Kalman filter) of the structural parameters. Three autocorrelation coefficients (i.e, the fitted model is AR(3)) and six autocovariances are used in QD estimation. The t-test and J-test sizes are for the 5% level. Power of the t-test is computed for the null of $H_0 : \psi = \psi_0 - 0.1$.

to obtain the maximum likelihood estimates. The results are reported in Table 2. Evidently, the QD estimates of ψ are close to the true value of 0.25, and the size of the t-test is close to the nominal value for all values of α. In contrast, the maximum likelihood estimates are biased when α is close to one, and the t-test for the null hypothesis that $\psi = 0.25$ is severely distorted.

7. RELATION TO OTHER $\sqrt{T}$-CONSISTENT LINEAR ESTIMATORS

As discussed in Section 2, one of the problems with the OLS estimator when a unit root is present is that the moment condition at the true value $\sqrt{T}\overline{g}(\alpha^0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_{t-1}\varepsilon_t$ does not satisfy a central limit theorem. Although y_{t-1} is orthogonal to ε_t, the persistence of y_{t-1} requires a stronger normalization, and standard distribution theory cannot be used. The thrust of the QD estimator is to use moment conditions that satisfy a central limit theorem uniformly over values of α^0. The approach can be used to estimate a broad range of models. However, for the simple AR(1) model, the ideas underlying the QD estimation can be used to construct a two-step *linear* estimator. We now show how this can be done and then relate this approximate QD estimator to other known linear estimators of the AR(1) model with classical properties in the local-to-unity framework.

For the AR(1) model, the QD moment condition (3) replaces y_{t-1} with ε_{t-j}. As seen from equation (4), the central limit theorem holds whether or not there is a unit root present. But the moment condition can be understood in an instrumental variable setup because ε_{t-j} is uncorrelated with ε_t and is hence a valid instrument. The only problem is that ε_{t-j} is not observed. But $\widehat{\alpha}_{OLS}$ is consistent for all $|\alpha^0| \leq 1$. Thus, let $\widetilde{e}_{t-1} = y_t - \widehat{\alpha}_{OLS} y_{t-1}$, noting that generated instruments do not require a correction for the standard errors as generated regressors do. We can now define a (hybrid differencing) HD estimator using the following moment condition:[9]

$$\bar{g}_{HD}(\widehat{\alpha}_{HD}) = \frac{1}{T}\sum_{t=k+1}^{T} \widetilde{e}_{t-k}(y_t - \widehat{\alpha}_{HD} y_{t-1}) = 0.$$

This leads to the estimator

$$\widehat{\alpha}_{HD} = \frac{\sum_{t=k}^{T} y_t \widetilde{e}_{t-k}}{\sum_{t=k}^{T} y_{t-1} \widetilde{e}_{t-k}} = \alpha^0 + \frac{\sum_{t=k}^{T} e_t \widetilde{e}_{t-k}}{\sum_{t=k}^{T} y_{t-1} \widetilde{e}_{t-k}}.$$

We refer to $\widehat{\alpha}^{HD}$ as a hybrid estimator because it is based on the covariance between the quasi-difference of y_t and a stationary random variable. Notice that the HD and QD use the same moment condition. What distinguishes the HD from the QD is that the objective function of the HD is now linear in α. Consistency of $\widehat{\alpha}_{HD}$ follows from the fact that $1/T \sum_{t=1}^{T} e_t(\alpha^0)\widetilde{e}_{t-k} \xrightarrow{p} \mathbb{E}\varepsilon_t\varepsilon_{t-k} = 0$. It is straightforward to show that in the local-to-unity framework,

$$\sqrt{T}(\widehat{\alpha}_{HD} - \alpha^0) \Rightarrow 2\big(1 + J_c(1)^2\big)^{-1} N(0,1).$$

Once the HD is understood as an instrumental estimator, other possibilities arise. Instead of $\widetilde{e}_{t-1}$, we can use any stationary series uncorrelated with the error term.[10] For example, using Δy_{t-1} would bypass the need for a preliminary least squares estimation. The first-differencing (FD) estimator is

$$\widehat{\alpha}_{FD} = \frac{\sum_{t=2}^{T} \Delta y_{t-1} y_t}{\sum_{t=2}^{T} \Delta y_{t-1} y_{t-1}}.$$

The FD is a special case of estimators analyzed in So and Shin (1999). These authors used the sign of y_{t-1} as instrument x_t to construct

$$\widehat{\alpha}_{SS} = \frac{\sum_{t=2}^{T} x_t y_t}{\sum_{t=1}^{T} x_t y_{t-1}}.$$

Another estimator with classical properties in the local-to-unity framework is that of Phillips and Han (2008). The Phillips and Han (PH) estimator, defined as

$$\widehat{\alpha}_{PH} = \frac{\sum_{t=2}^{T} \Delta y_{t-1}(2\Delta y_t + \Delta y_{t-1})}{\sum_{t=2}^{T}(\Delta y_{t-1})^2},$$

has the property that $\sqrt{T}(\widehat{\alpha}_{PH} - \alpha^0) \Rightarrow N(0, 2(1+\alpha^0))$ for all $\alpha^0 \in (-1, 1]$. As shown in the Appendix, the FD estimator is asymptotically equivalent to the PH estimator in the stationary case when α^0 is far from unit circle, that is, for the AR(1) model, $\widehat{\alpha}_{PH} = \widehat{\alpha}_{FD} + O_p(1/T)$, under stationary classical asymptotics. However, these two estimators differ in the local-to-unity setting. Whereas $\sqrt{T}(\widehat{\alpha}_{FD} - \alpha^0) \Rightarrow 2(1 + W(1)^2)^{-1}N(0,1)$ when $\alpha^0 = 1$, $\sqrt{T}(\widehat{\alpha}_{PH} - \alpha^0) \Rightarrow N(0, 4)$. The FD is thus more efficient at $\alpha^0 = 1$. Simulations presented in the Web Appendix (www.columbia.edu/∼sn2294/pub.html) show that the QD dominates the FD and PH but is comparable to the HD. The simulations also support the theoretical predictions that the FD, HD, and QD are all asymptotically normal and $\sqrt{T}$-consistent.

No estimator is perfect, and QD estimation has its drawbacks. As mentioned in the Introduction, the price we pay for asymptotic normality is that $\widehat{\alpha}_{QD}$ converges at a rate of $\sqrt{T}$ instead of T when there is a unit root. Han, Phillips, and Sul (2011) aggregated L stationary moment conditions and showed that by suitable choice of L, uniform asymptotic normality can be achieved at a rate faster than $\sqrt{T}$. Extension of their result outside of the AR(p) model is, however, not straightforward. In contrast, the QD framework is broadly applicable.

8. CONCLUDING COMMENTS

In this paper, we use a quasi-differencing framework to obtain estimators with classical properties even when the underlying data are highly persistent. QD can render nonstationary processes stationary so that classical limit theorems can be applied. However, the QD estimator is $\sqrt{T}$-consistent rather than superconsistent in the local-to-unity framework. In exchange for this slower convergence is generality, as QD estimation can be used in a broad range of linear and nonlinear models. However, there are several issues that remain to be solved. The first is allowing J to be data dependent and increase with the sample size. The second is to allow for conditional heteroskedasticity. Third, simulations suggest that the QD works well even when the forcing process is mildly explosive. Relaxing the assumption that the largest autoregressive root is inside the unit disk may well be useful for practitioners. These issues are left for future investigation.

NOTES

1. The optimization is performed over an expanded neighborhood of the set of admissible values of θ so that the parameter of interest is not on the boundary of the support. For the AR(1) model, the

admissible values are $(-1+\delta, 1]$ where $\delta > 0$. We optimize over $\Theta = [C_1, C_2]$, where $C_1 < -1+\delta < 1 < C_2$. When the context is clear, Θ will not be explicitly specified.

2. To be more precise, we should write $\widehat{\gamma}_j(\alpha, \alpha^0, \sigma^2)$ because the data are generated under α^0 and σ^2 and we quasi-difference the data at α. For notational simplicity, the dependence of $\widehat{\gamma}_j$ on α^0 and σ^2 is suppressed.

3. When the NQD has two local minima, it does not matter which one is chosen as they are asymptotically symmetric around the true value. This was why we state our result as $T^{3/2}(\widehat{\alpha}_{K,NQD} - \alpha^0)^2$ rather than $T^{3/4}(\widehat{\alpha}_{K,NQD} - \alpha^0)$. If $K = 1$, the FQD objective function has two minima, only one of which is consistent for α^0.

4. Detrending does not affect the asymptotic distribution of FQD, but the J_c in the distribution of the NQD estimator will depend on d_t. In the intercept-only case, one should use the demeaned Ornstein–Uhlenbeck process $\overline{J}_c(r) = J_c(r) - \int_0^1 J_c(s)ds$. In the linear trend case, the detrended process is $\widetilde{J}_c(r) = J_c(r) - \int_0^1 (4-6s)J_c(s)ds - r\int_0^1 (12-6s)J(s)ds$.

5. The optimization in equation (1) is done over a bounded set that includes a neighborhood of $\mathfrak{R}_\delta$ to avoid the boundary problem.

6. Likelihood estimation is also problematic when there are more variables than shocks, a problem known as stochastic singularity.

7. For an example of this implementation, see Coibion and Gorodnichenko (2011).

8. Identification requires rank $(\partial \overline{g}_{j,QD}(\theta))/\partial\theta = \dim\theta$. Now $\overline{g}_j(\theta)$ depends on θ through the parameters in the solution to expectation equations. Formal identification conditions are given in Komunjer and Ng (2011).

9. Laroque and Salanie (1997) used two OLS regressions in stationary variables to obtain a $\sqrt{T}$-consistent estimate of the cointegrating vector.

10. As suggested by a referee, $\mathbb{E}(e_t(e_{t-j} - e_{t-k})) = 0$, $1 \le j < k$ is also a valid moment condition.

REFERENCES

Abowd, J.M. & D. Card (1989) On the covariance structure of earnings and hours changes. *Econometrica* 57, 411–445.

Altonji, J.G. & L.M. Segal (1996) Small-sample bias in GMM estimation of covariance structures. *Journal of Business & Economic Statistics* 14, 353–366.

Canjels, E. & M.W. Watson (1997) Estimating deterministic trends in the presence of serially correlated errors. *Review of Economics and Statistics* 79, 184–200.

Chernozhukov, V. & H. Hong (2003) An MCMC approach to classical estimation. *Journal of Econometrics* 115, 293–346.

Coibion, O. & Y. Gorodnichenko (2011) Strategic interaction among heterogeneous price setters in estimated DSGE model. *Review of Economics and Statistics* 93, 920–940.

Gorodnichenko, Y. & S. Ng (2010) Estimation of DSGE models when the data are persistent. *Journal of Monetary Economics* 57, 325–340.

Han, C. & B. Kim (2011) A GMM interpretation of the paradox in the inverse probability weighting estimation of the average treatment effect on the treated. *Economics Letters* 110, 163–165.

Han, C., P.C.B. Phillips, & D. Sul (2011) Uniform asymptotic normality in stationary and unit root autoregression. *Econometric Theory* 27, 1117–1151.

Jansson, M. & M.J. Moreira (2006) Optimal inference in regession models with nearly integrated regressors. *Econometrica* 74, 681–714.

Komunjer, I. & S. Ng (2011) Dynamic identification of dynamic stochastic general equilibrium models. *Econometrica* 79, 1995–2032.

Laroque, G. & B. Salanie (1997) Normal estimators for cointegrating relationships. *Economics Letters* 55, 185–189.

Mikusheva, A. (2007a) Uniform inference in autoregressive models. *Econometrica* 75, 1411–1452.

Mikusheva, A. (2007b) Uniform inference in autoregressive models. *Econometrica* 75, online supplementary material.

Mikusheva, A. (2012) One dimensional inference in autogressive models with potential presence of a unit root. *Econometrica* 80, 163–212.

Newey, W.K. & D. McFadden (1994) Large sample estimation and hypothesis testing. In R.F. Engle & D. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2111–2245. North-Holland.

Pesavento, E. & B. Rossi (2006) Small-sample confidence interevals for multivariate impulse response functions at long horizons. *Journal of Applied Econometrics* 21, 1135–1155.

Phillips, P.C.B. & C. Han (2008) Gaussian inference in AR(1) time series with or without a unit root. *Econometric Theory* 24, 631–650.

Phillips, P.C.B. & C.C. Lee (1996) Efficiency gains from quasi-differencing under nonstationarity. In P. Robinson & M. Rosenblatt (eds.), *Athens Conference on Applied Probability and Time Series*, vol. II, *Time Series Analysis in Honor of E.J. Hannan*, pp. 300–314. Springer.

Phillips, P.C.B. & V. Solo (1992) Asymptotics for linear processes. *Annals of Statistics* 20, 971–1001.

Phillips, P.C.B. & Z. Xiao (1998) A primer on unit root testing. *Journal of Economic Surveys* 12, 423–469.

Pierce, D.A. (1982) The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Annals of Statistics* 10, 475–478.

Prokhorov, A. & P. Schmidt (2009) GMM redundancy results for general missing data problem. *Journal of Econometrics* 151, 47–55.

So, B.S. & D.W. Shin (1999) Cauchy estimators for autoregressive processes with applications to unit root tests and confidence intervals. *Econometric Theory* 15, 165–176.

Uhlig, H. (1999) A Toolkit for analyzing nonlinear dynamic stochastic models easily. In R. Marimon & A. Scott (eds.), *Computational Methods for the Study of Dynamic Economies*, pp. 30–61. Oxford University Press.

APPENDIX A: Proofs

The proofs proceed with the assumption that the weighting matrix W_T is an identity matrix.

Proof of Proposition 1(i). First, consider the problem of matching the jth autocovariance. That is, $Q_j(\alpha) = (\widehat{\gamma}_j(\alpha) - \gamma_j(\alpha))^2$, and $\widehat{\alpha}_j = \arg\min_\alpha Q_j(\alpha)$. Under the assumption that α is the true value, $\gamma_0(\alpha) = \sigma^2$, and $\gamma_j(\alpha) = 0$ for all $j > 0$. Note that

$$\widehat{\gamma}_j(\alpha) - \gamma_j(\alpha) = \frac{1}{T}\sum_{t=j+1}^{T} \varepsilon_t\varepsilon_{t-j} - \gamma_j(\alpha) - \frac{\alpha-\alpha^0}{T}\sum_{t=j+1}^{T}\left[\varepsilon_t y_{t-j-1} + \varepsilon_{t-j} y_{t-1}\right]$$

$$+ \frac{\left(\alpha-\alpha^0\right)^2}{T}\sum_{t=j+1}^{T} y_{t-1}y_{t-j-1}.$$

As a result, the NQD objective function is the fourth-order polynomial:

$$Q_j(\alpha) = Q_j^{(0)} + \left(\alpha-\alpha^0\right)Q_j^{(1)} + \left(\alpha-\alpha^0\right)^2 Q_j^{(2)} + \left(\alpha-\alpha^0\right)^3 Q_j^{(3)} + \left(\alpha-\alpha^0\right)^4 Q_j^{(4)}. \tag{A.1}$$

In the local-to-unity framework with $\alpha_0 = 1 + c/T$, the following results hold as $T \to \infty$:

$$\frac{1}{T}\sum \varepsilon_{t-j} y_{t-1} \Rightarrow \sigma^2 + \sigma^2\int_0^1 J_c(s)dW(s), \tag{A.2}$$

$$\frac{1}{T}\sum \varepsilon_t y_{t-1-j} \Rightarrow \sigma^2 \int_0^1 J_c(s)dW(s), \tag{A.3}$$

$$\frac{1}{T^2}\sum y_{t-1}y_{t-j-1} \Rightarrow \sigma^2 \int_0^1 J_c^2(s)ds. \tag{A.4}$$

It follows from equations (4) and (A.2)–(A.4) that

$$\begin{aligned} T^{1/2}Q_j^{(1)} &= -2\sqrt{T}\left(\frac{1}{T}\sum_{t=j+1}^{T}\varepsilon_t\varepsilon_{t-j}-\gamma_j(\alpha)\right)\frac{1}{T}\sum_{t=j+1}^{T}\left[\varepsilon_t y_{t-j-1}+\varepsilon_{t-j}y_{t-1}\right] \\ &\Rightarrow -2\xi_j\left(\sigma^2+2\sigma^2\int_0^1 J_c(s)dW(s)\right); \end{aligned}$$

$$\begin{aligned} T^{-1/2}Q_j^{(2)} = 2\sqrt{T}\left(\frac{1}{T}\sum_{t=j+1}^{T}\varepsilon_t\varepsilon_{t-j}-\gamma_j(\alpha)\right)\left(\frac{1}{T^2}\sum_{t=j+1}^{T}y_{t-1}y_{t-j-1}\right) \\ +\frac{1}{\sqrt{T}}\left(\frac{1}{T}\sum_{t=j+1}^{T}\left[\varepsilon_t y_{t-j-1}+\varepsilon_{t-j}y_{t-1}\right]\right)^2 \Rightarrow 2\xi_j\sigma^2\int_0^1 J_c^2(s)ds; \end{aligned}$$

$$\begin{aligned} T^{-1}Q_j^{(3)} &= -2\left(\frac{1}{T^2}\sum_{t=j+1}^{T}y_{t-1}y_{t-j-1}\right)\left(\frac{1}{T}\sum_{t=j+1}^{T}\left[\varepsilon_t y_{t-j-1}+\varepsilon_{t-j}y_{t-1}\right]\right) \\ &\Rightarrow -2\sigma^2\int_0^1 J_c^2(s)ds\left(\sigma^2+2\sigma^2\int_0^1 J_c(s)dW(s)\right); \end{aligned}$$

$$T^{-2}Q_j^{(4)} = \left(\frac{1}{T^2}\sum_1^T y_{t-1}y_{t-2}\right)^2 \Rightarrow \left(\sigma^2\int_0^1 J_c^2(s)ds\right)^2 > 0,$$

where $\sqrt{T}(\frac{1}{T}\sum\varepsilon_t\varepsilon_{t-j}-\gamma_j) \Rightarrow \xi_j = N(0,\sigma^4)$. To summarize:

$$Q_j^{(1)} = O_p\left(T^{-1/2}\right), \qquad Q_j^{(2)} = O_p\left(T^{1/2}\right), \qquad Q_j^{(3)} = O_p\left(T^1\right), \qquad Q_j^{(4)} = O_p\left(T^2\right). \tag{A.5}$$

It follows that

$$\frac{Q_j(\alpha)-Q_j^{(0)}}{T^2} \Rightarrow \left(\sigma^2\int_0^1 J_c^2(s)ds\right)^2(\alpha-\alpha^0)^4$$

uniformly over a bounded parameter space for α. As a result, $\widehat{\alpha}_j$ is a consistent estimate of α^0.

To study the large-sample properties of $\widehat{\alpha}_j$, consider the first-order condition

$$Q_j^{(1)}+2\left(\widehat{\alpha}_j-\alpha^0\right)Q_j^{(2)}+3\left(\widehat{\alpha}_j-\alpha^0\right)^2Q_j^{(3)}+4\left(\widehat{\alpha}_j-\alpha^0\right)^3Q_j^{(4)}=0. \tag{A.6}$$

This is a cubic equation of the form $ax^3+bx^2+cx+d=0$, where x stands for $(\widehat{\alpha}_j-\alpha_0)$ with the obvious correspondence between the coefficients. The cubic equation may have one or three real roots depending on the sign of the determinant:

$$\Delta = 18abcd-4b^3d+b^2c^2-4ac^3-27a^2d^2.$$

Given the orders established in (A.5), it can be shown that

$$T^{-7/2}\Delta \Rightarrow -32\sigma^{10}\left(\int_0^1 J_c^2(s)ds\right)^5 \xi_j^3.$$

The sign of the determinant Δ is asymptotically defined by the sign of ξ_j. When the sign of Δ is negative, there is a unique real root to equation (A.6); otherwise, there are three real roots. However, in the case of three real roots, the middle one corresponds to the local maxima, whereas the other two roots are the local minima of (A.1).

The next step is to work out the formulas for the roots and to check their rates of convergence toward zero. For example, when there is only one real root, the formula is

$$x_1 = -\frac{1}{3a}\left(b + \sqrt[3]{b^3 - \frac{9}{2}abc + \frac{27}{2}a^2d + \frac{1}{2}\sqrt{-27a^2\Delta}}\right.$$
$$\left. + \sqrt[3]{b^3 - \frac{9}{2}abc + \frac{27}{2}a^2d - \frac{1}{2}\sqrt{-27a^2\Delta}}\right).$$

Using the asymptotic orders of the terms in (A.5) and after tedious algebra, we can deduce that $Tx_1 = O_p(1)$ (i.e., $T(\widehat{\alpha}_j - \alpha^0) = O_p(1)$). Similarly, using the explicit formula for cubic roots and denoting the two noncentral roots by x_2 and x_3, we can deduce that when $\Delta > 0$, $T^{3/4}x_2 = O_p(1)$ and $T^{3/4}x_3 = O_p(1)$ (i.e., $T^{3/4}(\widehat{\alpha}_j - \alpha^0) = O_p(1)$ in this case).

To find the asymptotic distribution of these roots, we start with the case of one root. Because $T(\widehat{\alpha}_j - \alpha^0) = O_p(1)$, some terms in equation (A.6) are asymptotically negligible. Thus, asymptotically we have $Q_j^{(1)} + 2(\widehat{\alpha}_j - \alpha^0)Q_j^{(2)} = 0$. Equivalently,

$$T\left(\widehat{\alpha}_j - \alpha^0\right) = -T\frac{Q_j^{(1)}}{2Q_j^{(2)}} + o_p(1) \Rightarrow \frac{\sigma^2 + 2\sigma^2\int_0^1 J_c(s)dw(s)}{2\sigma^2\int_0^1 J_c^2(s)ds}.$$

Similarly, for the case of two local maxima $T^{3/4}\left(\widehat{\alpha}_j - \alpha^0\right) = O_p(1)$, and asymptotically $2\left(\widehat{\alpha}_j - \alpha^0\right)Q_j^{(2)} + 4\left(\widehat{\alpha}_j - \alpha^0\right)^3 Q_j^{(4)} = 0$. Equivalently,

$$T^{3/2}\left(\widehat{\alpha}_j - \alpha^0\right)^2 = -T^{3/2}\frac{Q_j^{(2)}}{2Q_j^{(4)}} \Rightarrow \frac{-\xi_j}{\sigma^2\int_0^1 J_c^2(s)ds}.$$

The equation has a solution only when $\xi_j < 0$. As shown previously, this is the condition for the cubic equation to have three roots. Thus, we proved that equation (6) holds for $\widehat{\alpha}_j$ with $K = 1$. Analogous arguments hold in the general case $\widehat{\alpha}_{K,NQD} = \arg\min_\alpha \sum_{j=1}^K Q_j(\alpha)$ with ξ_j replaced by $\sum_{j=1}^K \xi_j/K \sim N(0, \sigma^4/K)$. This also shows that in the AR(1) case, matching more than one auto-covariance leads to increase in efficiency. ■

Proposition 1(ii) and Proposition 2 are special cases of Proposition 3. Observe that

$$\overline{g}_{j,FQD}(\beta) = A_{j,FQD} + \left(\beta^0 - \beta\right)' B_{j,FQD} + \left(\beta^0 - \beta\right)' C_{j,FQD}\left(\beta^0 - \beta\right),$$

where $A_{j,FQD}, B_{j,FQD}$, and $C_{j,FQD}$ are defined in equations (11)–(13). The following three lemmas will be used to prove Proposition 3.

LEMMA A.1 (Uniform law of large numbers). *Let $\varepsilon_t = (\varepsilon_{t,1}, \varepsilon_{t,2})'$ be a martingale-difference sequence with $\Omega = \mathbb{E}(\varepsilon_t \varepsilon_t' | \mathcal{F}_{t-1})$ and finite fourth moments, $\Omega = (\sigma_{i,j})$ and $\eta_{t,i} = \sum_{j=0}^{\infty} c_j^i \varepsilon_{t-j,i}$ for $i = 1, 2$. Uniformly over the set of all sequences c_j^i satisfying $\sum_{j=0}^{\infty} |c_j^i| < C$ for some constant C,*

$$\frac{1}{T}\sum_{t=1}^{T} \eta_{t,1}\eta_{t,2} \to^p \mathbb{E}[\eta_{t,1}\eta_{t,2}].$$

Proof of Lemma A.1. Notice first that

$$\gamma_j^{i_1,i_2} = \text{cov}(\eta_{t,i_1}, \eta_{t+j,i_2}) = \sigma_{i_1,i_2} \sum_{n=0}^{\infty} c_{n+j}^{i_1} c_n^{i_2}$$

and for any i_1 and i_2, $\sum_{j=0}^{\infty} |\gamma_j^{i_1,i_2}| < \|\Omega\| C^2$. Furthermore,

$$\mathbb{E}[\eta_{t,1}\eta_{t,2}\eta_{t+j,1}\eta_{t+j,2}] = \left(\gamma_0^{1,2}\right)^2 + \gamma_j^{1,2}\gamma_j^{2,1} + \gamma_j^{1,1}\gamma_j^{2,2} + \mathbb{E}\left(\varepsilon_{1,t}^2 \varepsilon_{2,t}^2\right) \sum_{n=0}^{\infty} c_n^1 c_{n+j}^1 c_n^2 c_{n+j}^2.$$

As a result,

$$\text{cov}(\eta_{t,1}\eta_{t,2}, \eta_{t+j,1}\eta_{t+j,2}) = \gamma_j^{1,2}\gamma_j^{2,1} + \gamma_j^{1,1}\gamma_j^{2,2} + \mathbb{E}\left(\varepsilon_{1,t}^2 \varepsilon_{2,t}^2\right) \sum_{n=0}^{\infty} c_n^1 c_{n+j}^1 c_n^2 c_{n+j}^2$$

and

$$\sum_{j=1}^{\infty} \text{cov}((\eta_{t,1}\eta_{t,2}), (\eta_{t+j,1}\eta_{t+j,2})) \leq 2\|\Omega\|^2 C^4 + \mathbb{E}\left(\varepsilon_{t,1}^2 \varepsilon_{t,2}^2\right) \left(\sum_{n=0}^{\infty} |c_n^1 c_n^2|\right)^2$$
$$\leq \left(2\|\Omega\|^2 + \mathbb{E}(\varepsilon_{t,1}^2 \varepsilon_{t,2}^2)\right) C^4.$$

Chebyshev's inequality implies the statement of the lemma. ■

LEMMA A.2. *The following three statements hold asymptotically uniformly over $\mathfrak{R}_\delta$ and uniformly over $1 \leq j \leq K$:*

(i) $\sqrt{T}(A_{1,FQD}, \ldots, A_{K,FQD})' \Rightarrow (\xi_1 - \xi_0, \ldots, \xi_K - \xi_0)$, where $(\xi_0, \xi_1, \ldots, \xi_K)'$ is a normally distributed random vector with mean zero and diagonal covariance matrix, $\mathbb{E}\xi_0^2 = \mu_4$, $\mathbb{E}\xi_j^2 = \sigma^4$ for all $1 \leq j \leq K$;
(ii) $B_{j,FQD} \to^p a_j = \mathbb{E}\left[(X_{t+j} + X_{t-j} - 2X_t)\varepsilon_t\right]$;
(iii) $C_{j,FQD} = O_p(1)$.

Proof of Lemma A.2. Part (i) follows from applying the central limit theorem (4) to the sums of $\varepsilon_t^2 - \sigma^2$ and $\varepsilon_t \varepsilon_{t-j}$ for $1 \leq j \leq K$. For (ii) we need to show that the uniform law of large numbers holds for

$$B_{j,FQD} = \frac{1}{T}\sum_{t=p+j+1}^{T}\left(X_t\varepsilon_{t-j} + X_{t-j}\varepsilon_t - 2X_t\varepsilon_t\right)$$

$$= \frac{1}{T}\sum_{t=p+j+1}^{T-j}\left(X_{t+j} + X_{t-j} - 2X_t\right)\varepsilon_t + O_p\left(1/\sqrt{T}\right),$$

where the $O_p(1/\sqrt{T})$ term appears as a result of the change of limits of summation by a finite number of summands. To apply Lemma A.1, we need to show that the process $X_{t+j} + X_{t-j} - 2X_t$ has absolutely summable MA coefficients. From Lemma S8 in the Web Appendix of Mikusheva (2007b), the process $Z_t = X_t - X_{t-1}$ has absolutely summable MA coefficients uniformly over $\mathfrak{R}_\delta$. Now

$$X_{t+j} + X_{t-j} - 2X_t = \sum_{k=1}^{j} Z_{t+k} - \sum_{k=0}^{j-1} Z_{t-k}.$$

Our process of interest is the sum of a finite number of processes each with summable MA coefficients, and thus its MA coefficients are absolutely summable. Lemma A.1 implies that uniformly over $\mathfrak{R}_\delta$

$$B_{j,FQD} \to^p \mathbb{E}\Big[\left(X_{t+j} + X_{t-j} - 2X_t\right)\varepsilon_t\Big] = a_j.$$

Turning to (iii), the object of interest is the $p \times p$ matrix:

$$C_{j,FQD} = \frac{1}{2T}\sum_{t=j+p+1}^{T}\left(X_t X'_{t-j} + X_{t-j}X'_t - 2X_t X'_t\right), \tag{A.7}$$

where all elements except possibly the top-left element satisfy the uniform law of large numbers and thus are of order $O_p(1)$. Now the last $p-1$ elements of X_t are $\Delta y_{t-1}, \ldots, \Delta y_{t-p+1}$. From Lemma S8 in Mikusheva (2007b), they have absolutely summable MA coefficients uniformly over $\mathfrak{R}_\delta$. Thus, the elements of the right-bottom $(p-1)\times(p-1)$ submatrix satisfy the conditions of Lemma A.1. The elements in the first row and the first column (except the top-left element) are of the form

$$\frac{1}{2T}\sum_{t=j+p+1}^{T}\left(y_{t-1}z_{t-j} + y_{t-1-j}z_t - 2y_{t-1}z_t\right)$$

$$= \frac{1}{2T}\sum_{t=j+1}^{T}\left(y_{t-1+j} + y_{t-1-j} - 2y_{t-1}\right)z_t + O_p\left(1/\sqrt{T}\right),$$

where z_t is one of $\Delta y_{t-1}, \ldots, \Delta y_{t-p+1}$. From the proof of (ii), the series $y_{t-1+j} + y_{t-1-j} - 2y_{t-1}$ also has absolutely summable MA coefficients. Thus, the conditions of Lemma A.1 are satisfied.

It remains to consider the top-left element of the matrix $C_{j,FQD}$, which is given by

$$\left(C_{j,FQD}\right)_{11} = \frac{1}{T}\sum_{t=j+p+1}^{T}\left[y_{t-1}y_{t-j-1} - y_{t-1}^2\right].$$

If the largest (in absolute value) root λ_p is not real, then by definition of $\mathfrak{R}_\delta$, it is less than $\delta < 1$ in absolute value, and the process y_t is uniformly stationary. Thus $\left(C_{j,FQD}\right)_{11}$ satisfies the conditions of Lemma A.1. Assume now that the largest root λ_p is a real number. We have $1-\alpha L-\sum_{j=1}^{p-1} b_j L^j(1-L)=(1-\lambda_p)B(L)$, where all inverse roots of $B(L)$ are strictly inside the circle of radius δ.

Let $u_t = y_t - \lambda_p y_{t-1}$ and thus $B(L)u_t = \varepsilon_t$. Now u_t has absolutely summable MA coefficients uniformly over $\mathfrak{R}_\delta$.

$$\frac{1}{T}\sum_{t=j+1}^{T} y_{t-1}y_{t-j-1} = \frac{1}{T}\sum_{t=j+1}^{T} y_{t-j-1}\left(\lambda_p^j y_{t-j-1} + \sum_{k=0}^{j-1}\lambda_p^k u_{t-k-1}\right)$$

$$= \lambda_p^j \frac{1}{T}\sum_{t=1}^{T-j} y_{t-1}^2 + \sum_{k=0}^{j-1}\lambda_p^k \frac{1}{T}\sum_{t=j+1}^{T} y_{t-j-1}u_{t-k-1}.$$

As a result,

$$\left(C_{j,FQD}\right)_{11} = -\left(1-\lambda_p^j\right)\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2 + \sum_{k=1}^{j-1}\lambda_p^{k-1}\frac{1}{T}\sum_{t=j+1}^{T} y_{t-j-1}u_{t-k-1} + O_p\left(T^{-1/2}\right);$$

again the O_p term appears because of change of summation bounds. First, observe that

$$\operatorname{Var}\left(\frac{1}{T}\sum_{t=1}^{T-j} y_t u_{t+k}\right) = \operatorname{Var}\left(\frac{1}{T}\sum_{t=1}^{T-j}\sum_{s=0}^{t}\lambda_p^s u_{t-s}u_{t+k}\right)$$

$$= \frac{1}{T^2}\sum_{t=1}^{T-j}\sum_{s=0}^{t}\lambda_p^s \operatorname{cov}(u_{t+k}, u_{t-s}) < \operatorname{Var}(u_t) < \operatorname{const}(\delta).$$

The variance of u_t is uniformly bounded because all roots of this process are uniformly separated from the unit circle. That is, $\frac{1}{T}\sum_{t=1}^{T-j} y_t u_{t+k} = O_p(1)$ uniformly over $\beta^0 \in \mathfrak{R}_\delta$ and for all $1 \le k \le j \le K$.

Next, consider the term $(1-\lambda_p^j)\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2$. From Theorem 1 in Mikusheva (2012), $(1-\lambda_p)\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2$ is uniformly approximated by $\sigma^2/(g(c))\int_0^1 J_c^2(t)dt$, where $g(c) = \mathbb{E}\int_0^1 J_c^2(t)dt$, $J_c(t)$ is the Ornstein–Uhlenbeck process, and $c = T\log(|\lambda_p|)$. It follows from Lemma 4(h) and Lemma 10 in Mikusheva (2007a) that $1/(g(c))\int_0^1 J_c^2(t)dt$ is uniformly bounded in probability over c. Summing up, $C_{j,FQD}$ is asymptotically uniformly $O_p(1)$ over $\mathfrak{R}_\delta$, and the proof of Lemma A.2 is complete. ■

LEMMA A.3. *Under the assumptions of Proposition 3 the estimator $\widehat{\beta}_{K,FQD}$ is consistent for any $K > p$.*

Proof of Lemma A.3. Let $f(x) = (f_1(x), \ldots, f_{p+1}(x))$, where $f_j(x) = x'B_{j,FQD} + x'C_{j,FQD}x$ and $Q(x) = \sum_{j=1}^{K} f_j^2(x)$. Any $K \ge p+1$ suffices for consistency of $\widehat{\beta}_{K,FQD}$, though additional moments may improve efficiency. For any bounded set C in the parameter space, and by Lemma A.2, it holds that

$$\sup_{\beta\in C}\left|Q\left(\beta^0-\beta\right) - \sum_{j=1}^{K}\left(\overline{g}_{j,FQD}(\beta)\right)^2\right| = o_p(1).$$

Because $Q(0)=0$, for consistency of $\widehat{\beta}_{K,FQD}$, it is enough to show that for any $\varsigma>0$, there is $\varepsilon>0$ such that

$$\lim_{T\to\infty} P\left(\inf_{|x|>\varsigma} Q(x)>\varepsilon\right)=1, \tag{A.8}$$

where $x=\beta^0-\beta$. Because $\beta^0\in\Re_\delta$ and β belongs to a bounded neighborhood of $\Re_\delta$, x is bounded. There are two cases to consider: $|\lambda_p|<\delta_1<1$, and $\lambda_p\geq\delta_1$.

Case (i) $|\lambda_p|<\delta_1$**.** We will show that for *any* fixed $0<\delta_1<1$ statement (A.8) holds uniformly over $\beta^0\in\Re_\delta\cap\{|\lambda_p|<\delta_1\}$.

Because $K>p$, $Q(x)\geq f(x)'f(x)$. For any orthonormal transformation A, $Q(x)\geq(Af(x))'(Af(x))\geq(Af(x))_1^2$, where $(Af)_1$ is the first component of vector $Af(x)$. Consider a linear transformation, the first component of which is

$$(Af(x))_1=\left(f_{p+1}-\alpha^0 f_p-\sum_{j=1}^{p-1} b_j^0(f_{p+1-j}-f_{p-j})\right)\frac{1}{a},$$

where $a=\sqrt{1+(\alpha^0+b_1^0)^2+(b_2^0-b_1^0)^2+\cdots}$ is a (nonzero) multiplier that normalizes the linear transformation. Let $\mathcal{A}(L)=1-\alpha^0L-\sum_{j=1}^{p-1}b_j^0L^j(1-L)$ be a lag operator. Given the definition of f_j and linearity of the transformation,

$$(Af(x))_1=\frac{1}{a}\mathcal{A}(L)f_{p+1}=\frac{1}{a}\left(x'(\mathcal{A}(L)B_{p+1,FQD})+x'(\mathcal{A}(L)C_{p+1,FQD})x\right). \tag{A.9}$$

From (ii) of Lemma A.2, $B_{j,FQD}\to^p a_j=\mathbb{E}[X_{t+j}\varepsilon_t]$. From (10) and the definition of X_t,

$$\mathcal{A}(L)X_{t+p+1}=[\varepsilon_{t+p+1},\Delta\varepsilon_{t+p+1},\ldots,\Delta\varepsilon_{t+2}]'=\widetilde{e}_{t+p+1}. \tag{A.10}$$

Because $\mathbb{E}\widetilde{e}_{t+p+1}\varepsilon_s=0$ for any $s\leq t$,

$$\mathcal{A}(L)B_{p+1,FQD}\to^p\mathcal{A}(L)\mathbb{E}[X_{t+j}\varepsilon_t]=\mathbb{E}[\mathcal{A}(L)X_{t+j}\varepsilon_t]=\mathbb{E}[\widetilde{e}_{t+p+1}\varepsilon_t]=0.$$

Thus, uniformly over all x in a bounded set,

$$(Af(x))_1=\frac{1}{a}\left(x'(\mathcal{A}(L)C_{p+1,FQD})x\right)+o_p(1). \tag{A.11}$$

It follows from (13) and (A.10) that

$$\mathcal{A}(L)C_{p+1,FQD}=\frac{1}{2T}\sum_{t=p+1}^{T}\left(\widetilde{e}_{t+p+1}X_t'+X_t\widetilde{e}_{t+p+1}'\right)-\mathcal{A}(1)\frac{1}{T}\sum_{t=p+1}^{T}X_tX_t'+o_p(1), \tag{A.12}$$

where $o_p(1)$ appears from change in the bounds of summation. Because $|\lambda_p|<\delta_1$ by assumption, the process is stationary. Thus $\frac{1}{2T}\sum_{t=p+1}^{T}\widetilde{e}_{t+p+1}X_t'\to^p 0$ uniformly over $|\lambda_p|<\delta_1$ and $\mathcal{A}(1)\frac{1}{T}\sum_{t=p+1}^{T}X_tX_t'$ is uniformly positive definite. This gives us the needed bound in (A.8) for processes with $|\lambda_p|<\delta_1$.

Case (ii) $|\lambda_p| \geq \delta_1$**.** To show that for $\delta_1 < 1$ close enough to the unity, (A.8) holds uniformly over $\beta^0 \in \mathfrak{R}_\delta \cap \{|\lambda_p| \geq \delta_1\}$, we divide the area $|x| > \varsigma$ from (A.8) into two regions: $I_1 = \{x : |x| > \varsigma, |x_1| > \varsigma_1\}$ and $I_2 = \{x : |x| > \varsigma, |x_1| \leq \varsigma_1\}$, where $0 < \varsigma_1 < \varsigma$.

Consider $x \in I_1$. We will prove that for *any* fixed $\varsigma_1 > 0$, one can choose δ_1 close enough to the unity such that uniformly over $\beta^0 \in \mathfrak{R}_\delta \cap \{|\lambda_p| > \delta_1\}$, an analogue of (A.8) holds where the infinimum is taken over $x \in I_1$.

Applying the arguments and transformation as in (A.9), it can be shown that equations (A.11) and (A.12) hold. Because $\mathcal{A}(1)$ converges to zero as δ_1 converges to 1, one can choose δ_1 close enough to one to make all terms except the (1,1)th element of $\mathcal{A}(1)\frac{1}{T}\sum_{t=p+1}^{T} X_t X_t'$ sufficiently small, and all but the (1,1)-th element of $\frac{1}{T}\sum_{t=p+1}^{T} \widetilde{e}_{t+p+1} X_t'$ converge in probability to its expected value of zero. In consequence, the following result holds uniformly over $\beta^0 \in \mathfrak{R}_\delta \cap \{|\lambda_p| > \delta_1\}$ and $x \in I_1$:

$$x'\mathcal{A}(L)C_{p+1,FQD}x = x_1^2\left(\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_{t+p+1}y_{t-1} - \mathcal{A}(1)\frac{1}{T}\sum_{t=p+1}^{T} y_{t-1}^2\right) + o_p(1) + o_p(1-\delta_1).$$

It remains to show that $\frac{1}{T}\sum_{t=p+1}^{T}\left(\varepsilon_{t+p+1} - \mathcal{A}(1)y_{t-1}\right)y_{t-1}$ satisfies the uniform law of large numbers and thus converges uniformly to a nonzero constant. To do so, we use the decomposition as in Phillips and Solo (1992) that $\varepsilon_{t+p+1} - \mathcal{A}(1)y_{t-1} = u_t - u_{t-1}$, where u_t is a series with absolutely summable MA coefficients. Because $\frac{1}{T}(u_t - u_{t-1})y_{t-1} = -(1/T)(y_t - y_{t-1})u_t + O_p(1/\sqrt{T})$, Lemma A.1 applies, and $\frac{1}{T}\sum_{t=p+1}^{T}\left(\varepsilon_{t+p+1} - \mathcal{A}(1)y_{t-1}\right)y_{t-1}$ converges in probability to its expectation. Because $(\mathcal{A}(1))/T\mathbb{E}\sum_{t=p+1}^{T} y_{t-1}^2$ is uniformly different from zero, this implies that for any fixed $\varsigma_1 > 0$ there exists $\delta_1 < 1$ such that uniformly over $|\lambda_p| > \delta_1$ an analogue of (A.8) holds where the infinimum is taken over $x \in I_1$.

Consider now $x \in I_2$. One can choose ς_1 small enough and δ_1 close enough to the unity such that uniformly over $\beta^0 \in \mathfrak{R}_\delta \cap \{|\lambda_p| > \delta_1\}$, an analogue of (A.8) holds, where the infinimum is taken over $x \in I_2$. Given that $B_{j,FQD}$ and $C_{j,FQD}$ are uniformly bounded,

$$f_j(x) = x_{-1}'B_{j,-1} + x_{-1}'C_{j,-1}x_{-1} + o_p(\varsigma_1),$$

where $x_{-1} = (x_2, \ldots, x_p)$ is the $(p-1) \times 1$ subvector of x and $B_{j,-1}$ and $C_{j,-1}$ are the $(p-1) \times 1$ and $(p-1) \times (p-1)$ submatrices of $B_{j,FQD}$ and $C_{j,FQD}$ corresponding to the last $p-1$ components of β.

Let $Z_t = (\Delta y_{t-1}, \ldots, \Delta y_{t-p+1})$ and $\widetilde{Z}_t = (y_{t-1} - \lambda_p y_{t-2}, \ldots, y_{t-p+1} - \lambda_p y_{t-p})'$ be two $(p-1) \times 1$ uniformly stationary vector processes. Note that the matrices $B_{j,-1}$ and $C_{j,-1}$ satisfy equations analogous to (12) and (13) with Z_t in place of X_t. Similarly, $\widetilde{B}_j$ and $\widetilde{C}_j$ are defined as in (12) and (13) with $\widetilde{Z}_t$ in place of X_t. Observe that $Z_t' = \widetilde{Z}_t' - (1-\lambda_p)(y_{t-2}, \ldots, y_{t-p})$. It is easy to see that

$$f_j(x) = x_{-1}'\widetilde{B}_j + x_{-1}'\widetilde{C}_j x_{-1} + o_p(1-\delta_1) + o_p(\varsigma_1).$$

The function $\widetilde{f}_j = x_{-1}'\widetilde{B}_j + x_{-1}'\widetilde{C}_j x_{-1}$ corresponds to that of the uniformly stationary process $y_t - \lambda_p y_{t-1}$ with all roots smaller than δ in absolute value. The rest of the proof follows arguments as in case (i). ■

Proof of Proposition 3(i). To establish the asymptotic distribution of $\widehat{\beta}_{K,FQD}$, consider the first-order condition

$$\sum_{j=1}^{K} \overline{g}_{j,FQD}\left(\widehat{\beta}_{K,FQD}\right) \frac{\partial \overline{g}_{j,FQD}}{\partial \beta}\left(\widehat{\beta}_{K,FQD}\right) = 0.$$

From Lemma A.2 and consistency of $\widehat{\beta}_{K,FQD}$,

$$\frac{\partial \overline{g}_{j,FQD}}{\partial \beta}\left(\widehat{\beta}_{K,FQD}\right) = -B_{j,FQD} + o_p(1) \to^p a_j,$$

and uniformly over $\mathfrak{R}_\delta$:

$$\sqrt{T}\,\overline{g}_{j,FQD}\left(\widehat{\beta}_{K,FQD}\right) = \sqrt{T} A_{j,FQD} + a_j' \sqrt{T}\left(\widehat{\beta}_{K,FQD} - \beta^0\right) + o_p(1).$$

As a consequence, the following result holds uniformly:

$$\sqrt{T}\left(\widehat{\beta}_{K,FQD} - \beta^0\right) \Rightarrow \left(\sum_{j=1}^{K} a_j a_j'\right)^{-1} \left(\sum_{j=1}^{K} a_j (\xi_j - \xi_0)\right) = N(0, \Sigma_{K,FQD}),$$

where $G = \left(\sum_{j=1}^{K} a_j a_j'\right)^{-1}$ and $\Sigma_{K,FQD} = \sigma^4 G + \mu_4 G \left(\sum_{j=1}^{K} a_j\right)\left(\sum_{j=1}^{K} a_j\right)' G.$ ■

Proof of Proposition 3(ii). The proof proceeds by treating QD as a two-step estimator. First, note that

$$s^2 - \frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_t^2 = -\frac{1}{T}\left(\sum_t X_t \varepsilon_t\right)' \left(\sum_t X_t X_t'\right)^{-1} \left(\sum_t X_t \varepsilon_t\right).$$

Theorem 1 in Mikusheva (2012) shows that the statistic $(\sum_t X_t \varepsilon_t)'(\sum_t X_t X_t')^{-1}(\sum_t X_t \varepsilon_t)$ is uniformly approximated by the distribution $(t^c + N(0, p-1))^2$, where $t^c = \left(\int J_c(t) dw(t)\right) \Big/ \sqrt{\int J_c^2(t) dt}$ is a local-to-unity limit of a t-statistic and $c = T\log(|\lambda_p|)$. Given that t^c is uniformly bounded in probability over all possible values of $c \le 0$, the following result holds uniformly over $\mathfrak{R}_\delta$:

$$s^2 = \frac{1}{T}\sum_t \varepsilon_t^2 + O_p(1/T). \tag{A.13}$$

Because $\overline{g}_{j,QD}(\beta) = \overline{g}_{j,FQD}(\beta) - \gamma_0 + s^2$,

$$\overline{g}_{j,QD}(\beta) = A_{j,QD} + \left(\beta^0 - \beta\right)' B_{j,QD} + \left(\beta^0 - \beta\right)' C_{j,QD}\left(\beta^0 - \beta\right),$$

where $A_{j,QD} = A_{j,FQD} + s^2 - \sigma^2$, $B_{j,QD} = B_{j,FQD}$, and $C_{j,QD} = C_{j,FQD}$. A result analogous to Lemma A.2 holds for $A_{j,QD}$, $B_{j,QD}$, and $C_{j,QD}$ with one correction: $\sqrt{T}(A_{1,QD}, \ldots, A_{K,QD}) \Rightarrow (\xi_1, \ldots, \xi_K)$. This gives us consistency and asymptotic normality of $\widehat{\beta}_{K,QD}$ with asymptotic covariance matrix $\Sigma_{K,QD} = \sigma^4 G$. ■

The following lemma will be used to prove Proposition 4.

LEMMA A.4. *Uniformly over all possible values of* θ*,*

(i) $\sqrt{T}(A_j + \Omega^0) \Rightarrow \xi_j - \xi_0$*;*
(ii) $\sqrt{T}(A_j + S) \Rightarrow \xi_j$*;*
(iii) $B_{j,1} \to^p \mathbb{E}[\varepsilon_t(x_{t-j-1} - x_{t-1})] = 0$*;*
(iv) $B_{j,2} \to^p \mathbb{E}[x_{t-1}(\varepsilon_{t-j} - \varepsilon_t)'] = a_j$*;*
(v) $C_j = O_p(1)$*,*

where $\frac{1}{\sqrt{T}}\sum_{t=j+1}^{T} \varepsilon_t \varepsilon_{t-j}' \Rightarrow \xi_j$ *and* ξ_j *is a* 2×2 *matrix with normally distributed components such that for any nonrandom vector* $\mathbf{a}$ *the vector* $\xi_j \mathbf{a}$ *is normally distributed with variance-covariance matrix* $\Omega^0\mathbf{a}'\Omega^0\mathbf{a}$*. We also have* $\frac{1}{\sqrt{T}}\sum_{t=j+1}^{T} \varepsilon_t \varepsilon_t' \Rightarrow \xi_0$ *where* ξ_0 *is a* 2×2 *matrix with normally distributed components such that for any nonrandom vector* $\mathbf{a}$*, the vector* $\xi_0\mathbf{a}$ *is normally distributed with variance-covariance matrix* $\mathbb{E}\left[(\varepsilon_t'\mathbf{a})^2 \varepsilon_t \varepsilon_t'\right]$*. The variables* ξ_j *are independent for any* $j \geq 0$*.*

Proof of Lemma A.4. Result (i) follows from the central limit theorem. To prove (ii), note that

$$S = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t \varepsilon_t' - \frac{1}{T}\frac{1}{\sum_s x_{s-1}^2}\begin{pmatrix} (\sum_s \varepsilon_{xs} x_{s-1})^2 & (\sum_s \varepsilon_{xs} x_{s-1})(\sum_s \varepsilon_{ys} x_{s-1}) \\ (\sum_s \varepsilon_{ys} x_{s-1})(\sum_s \varepsilon_{xs} x_{s-1}) & (\sum_s \varepsilon_{ys} x_{s-1})^2 \end{pmatrix}.$$

Now $(\sum_{s=1}^T \varepsilon_{xs} x_{s-1}) \Big/ \sqrt{\sum_{s=1}^T x_{s-1}^2} \Rightarrow t^c$ uniformly over $\alpha^0 \in (-1+\delta, 1]$, and the family t^c is uniformly bounded. The sum $(\sum_{s=1}^T \varepsilon_{ys} x_{s-1}) \Big/ \sqrt{\sum_{s=1}^T \mathbb{E}x_{s-1}^2}$ has a bounded second moment because $\varepsilon_{ys} x_{s-1}$ is a martingale-difference sequence and thus it is uniformly bounded by Chebyshev's inequality. Last, $\left(\sum_{s=1}^T x_{s-1}^2\right) / \left(\sum_{s=1}^T \mathbb{E}x_{s-1}^2\right)$ is uniformly separated from zero, a result that follows from Lemma 4(h) and Lemma 10 in Mikusheva (2007a). Summing up, we have $S = \frac{1}{T}\sum_{t=1}^T \varepsilon_t \varepsilon_t' + O_p(\frac{1}{T})$. As a result, $\sqrt{T}(A_j + S) = \frac{1}{\sqrt{T}}\sum_{t=j+1}^T \varepsilon_t \varepsilon_{t-j}' \Rightarrow \xi_j$.

The proof of part (iii) follows from Lemma A.1 because we show in the proof of Lemma A.2 that $x_{t-j-1} - x_{t-1}$ has absolutely summable MA coefficients uniformly over α.

To prove (iv), rewrite

$$B_{j,2} = \frac{1}{T}\sum_{t=j+1}^{T} x_{t-1}(\varepsilon_{t-j} - \varepsilon_t)' = \frac{1}{T}\sum_{t=j+1}^{T} (x_{t+j-1} - x_{t-1})\varepsilon_t'$$
$$+ \frac{1}{T}\sum_{t=j+1}^{2j} x_{t-1}\varepsilon_{t-j} - \frac{1}{T}\sum_{t=T+1}^{T+j} x_{t-1}\varepsilon_{t-j}.$$

The terms $\frac{1}{T}\sum_{t=j+1}^{2j} x_{t-1}\varepsilon_{t-j}$ and $\frac{1}{T}\sum_{t=T+1}^{T+j} x_{t-1}\varepsilon_{t-j}$ both have j summands each of which is of order $O_p(\sqrt{T})$. This means that for any $j \leq K$ where K is fixed, the following result holds uniformly:

$$B_{j,2} = \frac{1}{T}\sum_{t=j+1}^{T}(x_{t+j-1} - x_{t-1})\varepsilon_t' + O_p\left(\frac{1}{\sqrt{T}}\right).$$

The rest of the proof is the same as for part (iii). Part (v) follows from Lemma A.2(iii). ■

Proof of Proposition 4. Note first that

$$\begin{aligned}\widehat{\Gamma}_j(\theta) - \widehat{\Gamma}_0(\theta) &= \frac{1}{T}\sum_{t=j+1}^{T} e_t(\theta)(e_{t-j}(\theta) - e_t(\theta))' \\ &= \frac{1}{T}\sum_{t=j+1}^{T}((\theta^0 - \theta)x_{t-1} + \varepsilon_t)((\theta^0 - \theta)(x_{t-j-1} - x_{t-1}) + \varepsilon_{t-j} - \varepsilon_t)' \\ &= A_j + B_{j,1}(\theta^0 - \theta)' + (\theta^0 - \theta)B_{j,2} + (\theta^0 - \theta)C_j(\theta^0 - \theta)',\end{aligned}$$

where

$$A_j = \frac{1}{T}\sum_{t=j+1}^{T}\varepsilon_t(\varepsilon_{t-j} - \varepsilon_t)'; \qquad B_{j,1} = \frac{1}{T}\sum_{t=j+1}^{T}\varepsilon_t(x_{t-j-1} - x_{t-1});$$

$$B_{j,2} = \frac{1}{T}\sum_{t=j+1}^{T}x_{t-1}(\varepsilon_{t-j} - \varepsilon_t)'; \qquad C_j = \frac{1}{T}\sum_{t=j+1}^{T}x_{t-1}(x_{t-j-1} - x_{t-1}).$$

Lemma A.4 showed that uniformly over α

$$\|\widehat{\Gamma}_j(\theta) - \widehat{\Gamma}_0(\theta) + \Omega^0\|_2^2 = \|(\theta^0 - \theta)B_{j,2} + C_j(\theta^0 - \theta)(\theta^0 - \theta)'\|_2^2 + o_p(1)$$

and

$$\|\widehat{\Gamma}_j(\theta) - \widehat{\Gamma}_0(\theta) + S\|_2^2 = \|(\theta^0 - \theta)B_{j,2} + C_j(\theta^0 - \theta)(\theta^0 - \theta)'\|_2^2 + o_p(1).$$

We minimize the sum of such functions for $j = 1, \dots, K$. Obviously, the minimized function is nonnegative, and one of its minimal value of zero is achieved at $\theta = \theta^0$. The question is whether there are any other minima. For this, there should exist θ such that $\|(\theta^0 - \theta)B_{j,2} + C_j(\theta^0 - \theta)(\theta^0 - \theta)'\|_2$ is zero for all j. For a given j, the only nontrivial null of function $\|(\theta^0 - \theta)B_{j,2} + C_j(\theta^0 - \theta)(\theta^0 - \theta)'\|_2$ implies $\theta_j = \theta^0 + (1/C_j)B_{j,2}$, which is asymptotically different for different j. This implies that for $K \geq 2$ no other asymptotic null of the objective function other than $\theta = \theta^0$ exists, and thus $\widehat{\theta}_{K,FQD}$ and $\widehat{\theta}_{K,QD}$ are consistent.

To derive the limit distribution of $\widehat{\theta}_{K,FQD}$, we use the fact that the first-order condition must be satisfied at $\theta = \widehat{\theta}_{K,FQD}$. Now the first-order condition is

$$\begin{aligned}&\nabla_\theta\|\widehat{\Gamma}_j(\theta) - \widehat{\Gamma}_0(\theta) + \Omega^0\|_2^2 \\ &\quad = -2\left(A_j + \Sigma^0 + B_{j,1}(\theta^0 - \theta)' + (\theta^0 - \theta)B_{j,2} + C_j(\theta^0 - \theta)(\theta^0 - \theta)'\right) \\ &\quad\quad \times \left(B_{j,1}' + B_{j,2} + 2C_j(\theta^0 - \theta)'\right)'.\end{aligned}$$

Because we proved that $\widehat{\theta}_{K,FQD}$ is uniformly consistent, and given statements (iv) and (v) of Lemma A.4,

$$B_{j,2}+2C_j(\theta^0-\widehat{\theta}_{K,FQD})' \to^p a_j.$$

Furthermore,

$$\begin{aligned}\sqrt{T}\Big(&A_j+\Omega^0+B_{j,1}(\theta^0-\widehat{\theta}_{K,FQD})+(\theta^0-\widehat{\theta}_{K,FQD})B_{j,2}\\ &+C_j(\theta^0-\widehat{\theta}_{K,FQD})(\theta^0-\widehat{\theta}_{K,FQD})'\Big)\\ &=\sqrt{T}(A_j+\Omega^0)+\sqrt{T}(\theta^0-\widehat{\theta}_{K,FQD})a_j+o_p(1).\end{aligned}$$

As a result,

$$\sqrt{T}(\widehat{\theta}_{K,FQD}-\theta^0) \Rightarrow \frac{1}{\sum_{j=1}^K a_j a_j'}\sum_{j=1}^K(\xi_j-\xi_0)a_j$$

uniformly over α. Similarly,

$$\sqrt{T}(\widehat{\theta}_{K,QD}-\theta^0) \Rightarrow \frac{1}{\sum_{j=1}^K a_j a_j'}\sum_{j=1}^K \xi_j a_j.$$

The last two formulas lead to the conclusion of Proposition 4. ■

Relation between PH and FD. Observe that

$$\begin{aligned}\sum_{t=2}^T(\Delta y_{t-1})^2 &= \sum_{t=2}^T \Delta y_{t-1}y_{t-1}-\sum_{t=2}^T(y_{t-1}-y_{t-2})y_{t-2}\\ &=\sum_{t=2}^T \Delta y_{t-1}y_{t-1}+\sum_{t=2}^T y_{t-1}(y_{t-1}-y_{t-2})-y_{T-1}^2+y_0^2\\ &=2\sum_{t=2}^T \Delta y_{t-1}y_{t-1}-y_{T-1}^2+y_0^2.\end{aligned}$$

Thus if $|\alpha^0|<1$ is fixed and $T\to\infty$,

$$\frac{1}{T}\sum_{t=2}^T(\Delta y_{t-1})^2=2\frac{1}{T}\sum_{t=2}^T\Delta y_{t-1}y_{t-1}+O_p(1/T).$$

Similarly,

$$\begin{aligned}\sum_{t=2}^T\Delta y_{t-1}(2\Delta y_t+\Delta y_{t-1}) &= 2\sum_{t=2}^T\Delta y_{t-1}y_t-2\sum_{t=2}^T\Delta y_{t-1}y_{t-1}+\sum_{t=2}^T(\Delta y_{t-1})^2\\ &=2\sum_{t=2}^T\Delta y_{t-1}y_t-y_{T-1}^2+y_0^2\end{aligned}$$

and

$$\frac{1}{T}\sum_{t=2}^T\Delta y_{t-1}(2\Delta y_t+\Delta y_{t-1})=2\frac{1}{T}\sum_{t=2}^T\Delta y_{t-1}y_t+O_p(1/T).$$

This leads us to the result that $\widehat{\alpha}_{PH}=\widehat{\alpha}+O_p(T^{-1})$ under stationary asymptotics.