Approximately Optimal Instrument for Multiperiod Conditional Moment Restrictions

Stanislav Anatolyev*
New Economic School

Abstract

The form of the optimal instrument for general multiperiod conditional moment restrictions is highly nonlinear and usually cannot be solved analytically. We show how to construct instruments approximately satisfying the optimality conditions, evaluate asymptotic variances of corresponding instrumental variables estimators in specific examples, and verify their behavior in finite samples. The asymptotic properties of approximately optimal instruments are favorable, and the finite sample properties of their feasible versions are advantageous compared to competitors. We also illustrate the proposed method with an application to ultra-high frequency data.

1 Introduction

Many time series models appear in the form of conditional moment restrictions. They are usually estimated and tested by choosing instrumental variables (IV) from the conditioning information set and applying GMM (Hansen 1982). To attain highest efficiency of estimation, instruments should be chosen optimally from an infinite set of possible instruments. When the moment function is a martingale difference with respect to the conditioning information so that the moment restrictions are *single-period*, the optimal instrument is an explicit function of certain conditional expectations, estimation of which constitutes a feasible procedure. However, a variety of intertemporal macroeconomic and financial models give rise to *multiperiod* conditional moment restrictions, the ones that are characterized by the presence of serial correlation of finite order (Hansen and West 2002). The examples are numerous in the asset pricing (e.g., Hansen and Singleton 1982, Ferson and Constantinides 1991, Hansen and Singleton 1996) and forecasting (Hansen and Hodrick 1980, Mishkin 1990, Rich, Raymond and Butler 1992) literatures. Other potential applications include problems with complex decision rules (Eichenbaum, Hansen and Singleton 1988, West and Wilcox 1996) and with temporal aggregation (Grossman, Melino and Shiller 1987, Hall 1988). Recent research on volatility also have led to exploitation of multiperiod restrictions (Meddahi

*Address: Stanislav Anatolyev, New Economic School, Nakhimovsky Prospekt, 47, Moscow, 117418 Russia. E-mail: sanatoly@nes.ru. I thank Kenneth West, Bruce Hansen, Yuichi Kitamura, Russel Davidson, Grigory Kosenok, seminar participants at various places, especially at the University of Wisconsin and CIREQ, and the co-editor Grégoir Stéphane and two anonymous referees whose insightful comments helped improve the presentation.

and Renault 2002, Meddahi, Renault and Werker 2003). The GMM procedure in these circumstances does not change dramatically, but the optimality conditions become significantly more complicated. It turns out, however, that in the special case of conditional homoskedasticity it is still possible to derive an explicit expression for the optimal instrument, which is done in Hansen (1985).

In a general case when both serial correlation and conditional heteroskedasticity are in effect, one approach to handle the problem is to artificially restrict the space of instruments to linear combinations of lags of initially given instruments. This leads to tractable theories that allow one to construct feasible instruments that attain the asymptotic efficiency bound relative to the restricted space of instruments. This is done in Kuersteiner (2002) for conditionally heteroskedastic AR models, in Kuersteiner (2001) for ARMA models with conditionally heteroskedastic innovations, and in West, Wong and Anatolyev (2002) for more general stationary time series models. Generally, employing the subclass of instruments delivers efficiency gains, often substantial, compared to the use of initially given instruments or a finite number of their lags. However, in spite of these efficiency gains, the linear subclass is significantly narrower than the entire class of allowable instruments, and a more appealing idea is to approach the efficiency bound relative to the widest class of instruments.

Hansen (1985) and Hansen, Heaton and Ogaki (1988) presented a characterization of the efficiency bound for GMM estimators that correspond to a given system of conditional moment restrictions and exploit all information in the instruments. Anatolyev (2003) gives a more algorithmic description of the optimal instrument. He finds that the process followed by the optimal instrument is a recursion that generalizes Hansen's formula (Hansen 1985, Lemma 5.7), and is parameterized by three auxiliary processes which can be viewed as infinite-dimensional parameters. Estimation of these would constitute a feasible procedure, if there were not the following major difficulty: the parameter processes solve a system of highly nonlinear functional equations rather than follow explicitly postulated dynamics. In rare circumstances it is possible to solve this system analytically, as in Heaton and Ogaki's (1991) example, but this is not typical.

In order to proceed, we take an approach where the three nonlinear equations are approximated, and the solutions of the approximated versions are used to construct the instrument. By approximation we mean the Taylor expansion around known counterparts that correspond to the two special cases of no conditional heteroskedasticity and of no serial correlation. This procedure results in different versions of the *approximately optimal instrument* according to different orders of the Taylor expansion for the three equations that determine the auxiliary processes. For a simple design with quadratic heteroskedasticity we compute asymptotic variances of approximately optimal instrumental variables estimators, and determine preferable orders in the Taylor expansion. On the other hand, we evaluate the efficiency losses due to the approximation error in the Heaton–Ogaki example where it is possible to explicitly calculate them. These losses turn out to be tiny showing that the proposed instrument is able to nearly attain the efficiency bound.

To illustrate the implementation of the feasible version of the proposed instrument, we run a Monte–Carlo simulation experiment. In constructing the feasible instrument, we estimate auxiliary conditional expectations by a kernel regression on lagged basic instruments, although a researcher may use alternative nonparametric methods. More concretely, we employ the Nadaraya–Watson (NW) nonparametric estimator with a global bandwidth, using the nonparametric corrected asymptotic final prediction error criterion of Tschernig and

Yang (2000) to select significant lags and an optimal value of the bandwidth. It turns out that the feasible approximately optimal instrument has advantageous finite sample properties compared to likely competitors. Finally, we present an empirical application based on a recently developed model by Meddahi, Renault and Werker (2003) using tick-by-tick data for two stocks from the Russian stock market.

The paper is organized as follows. Section 2 presents the setup and reviews the form of the optimal instrument, the starting point of our analysis. Section 3 shows how the approximations are taken, elaborating the case of two-period conditional moment restrictions, and also discusses the approximation method and estimation of auxiliary processes. Section 4 presents computations of asymptotic gains in a simple model with quadratic heteroskedasticity, and losses in the Heaton–Ogaki example. Section 5 reports the results of simulation experiments, while Section 6 contains the empirical application. In Section 7 we outline what changes when the serial correlation is of higher order than the first, and conclude.

2 The optimal instrument

2.1 The single equation case

We consider the equation

$$f(\boldsymbol{\beta}, \mathbf{x}_t) = e_t, \tag{1}$$

where e_t is an error term, $\mathbf{x}_t$ is a vector of observable variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters to be estimated. The function $f(\boldsymbol{\beta}, \mathbf{x}_t)$ is Borel measurable for all $\boldsymbol{\beta} \in \mathbb{B}$ and is continuously differentiable in the first argument for all $\boldsymbol{\beta} \in \mathbb{B}$ for all $\mathbf{x}_t$ in its support. This function is known up to $\boldsymbol{\beta}$, which is the object of estimation, and may be nonlinear in $\boldsymbol{\beta}$. In addition, we are given vector $\mathbf{z}_t$ of observable *basic instruments* (as opposed to simply *instruments* that may be generated from the basic ones). Let us denote by $\Im_t$ the information embedded in $\mathbf{z}_t$ and all its history, i.e. $\Im_t \equiv \sigma(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots)$, and use the shortcut notation $E_t[\cdot] \equiv E[\cdot|\Im_t]$. Some of observable variables in $\mathbf{x}_t$, along with their lags or functions, may be among $\mathbf{z}_t$, but generally they need not be measurable relative to $\Im_t$. For example, in a linear IV regression, $\mathbf{x}_t$ contains both left and right hand variables. In a rational expectation model, $\mathbf{x}_t$ typically contains $\mathbf{z}_t$ together with its several leads, although $\mathbf{z}_t$ may include variables not present in $\mathbf{x}_t$; also, not all information available to decision makers is necessarily contained in $\Im_t$. We assume that the vector $(\mathbf{x}_t', \mathbf{z}_t')'$ is strictly stationary and ergodic.

The model is completed by the conditional moment restriction

$$E_t[e_t] = 0. \tag{2}$$

This restriction implies that all measurable functions of $\mathbf{z}_t$ and their lags are valid instruments. Note that it does not preclude correlatedness of errors with the leads of the basic instrument as the latter typically is not strictly exogenous. Define the $k \times 1$ conditional score vector

$$\mathbf{d}_t \equiv E_t\left[\frac{\partial f(\boldsymbol{\beta}, \mathbf{x}_t)}{\partial \boldsymbol{\beta}}\right], \tag{3}$$

and let

$$\omega_t \equiv E_t\left[e_t^2\right], \quad \gamma_t \equiv E_t\left[e_t e_{t-1}\right] \tag{4}$$

be the conditional variance and conditional first-order autocovariance of the errors. We assume that the error is first-order conditionally serially correlated, i.e. $E_t[e_t e_{t-j}] = 0$ for $j > 1$, and that $\operatorname{ess\,inf} |\gamma_t| > 0$.

We list several examples from applied econometric practice that fit in the described framework.

Example 1 *The Ferson and Constantinides (1991) consumption-based CAPM with habit formation corresponds to* $\theta = (\alpha, \delta, \gamma)'$, $\mathbf{x}_t = (c_{t+2}/c_{t+1}, c_{t+1}/c_t, c_t/c_{t-1}, r_{t+1})'$, *and*

$$f(\boldsymbol{\beta}, \mathbf{x}_t) = \delta(1 + r_{t+1}) \left(\left(\frac{s_{t+1}}{s_t} \right)^{-\gamma} - \alpha\delta \left(\frac{s_{t+2}}{s_t} \right)^{-\gamma} \right) - \left(1 - \alpha\delta \left(\frac{s_{t+1}}{s_t} \right)^{-\gamma} \right),$$

where $s_t = c_t - \alpha c_{t-1}$, c_t *is a consumption, and* r_t *is a market return;* α *is a habit formation/durability parameter,* δ *is a discount factor and* γ *is a coefficient of risk aversion. The basic instrument is* $\mathbf{z}_t = (c_t/c_{t-1}, r_t)'$.

Example 2 *The Meddahi, Renault and Werker (2003) model for ultra-high-frequency returns considered in more detail in Section 6 corresponds to* $\boldsymbol{\beta} = (\theta, \kappa)'$, $\mathbf{x}_t = (r_t, r_{t-1}, d_t, d_{t-1})'$, *and*

$$f(\boldsymbol{\beta}, \mathbf{x}_t) = \left(r_t^2 - \theta\right) - \left(r_{t-1}^2 - \theta\right) \exp(-\kappa d_{t-1}) \frac{c(\kappa d_t)}{c(\kappa d_{t-1})},$$

where d_t *is a duration,* r_t *is a scaled return, and* $c(v) \equiv (1 - \exp(-v))/v$. *The basic instrument is* $\mathbf{z}_t = (r_{t-2}, d_{t-2})'$.

Example 3 *If the AR parameter in an ARMA(1,1) model is estimated by instrumental variables, then* $\boldsymbol{\beta} = \alpha$, $\mathbf{x}_t = (y_t, y_{t-1})'$, $f(\boldsymbol{\beta}, \mathbf{x}_t) = y_t - \alpha y_{t-1}$, *and* $\mathbf{z}_t = y_{t-2}$.

Typically, the parameters in the model (1)–(2) are estimated by GMM (Hansen 1982) after the researcher chooses an instrument vector, which is likely to include the basic instrument, its recent lagged values, and (more rarely) nonlinear functions thereof, using an heteroskedasticity and autocorrelation consistent variance estimator to form a weighting matrix. This procedure yields consistent and asymptotically normal estimates, but intrinsic arbitrariness in forming the instrument vector leaves possibilities of increasing asymptotic efficiency. Hansen (1985) and Hansen, Heaton and Ogaki (1988) present a characterization of the efficiency bound for GMM estimators. Anatolyev (2003) gives a more algorithmic description of the optimal instrument. Under suitable conditions, it has the following form:

$$\boldsymbol{\zeta}_t = \boldsymbol{\zeta}_{t-1}\phi_t + \rho_t \boldsymbol{\delta}_t, \tag{5}$$

where the stationary ergodic $\Im_t$-measurable processes, scalar ϕ_t, $k \times 1$ vector $\boldsymbol{\delta}_t$, and scalar almost surely positive ρ_t, satisfy the following system:

$$\gamma_t + \phi_t \left(\omega_t + E_t\left[\phi_{t+1}\gamma_{t+1}\right]\right) = 0, \tag{6}$$

$$\boldsymbol{\delta}_t = \mathbf{d}_t + E_t\left[\phi_{t+1}\boldsymbol{\delta}_{t+1}\right], \tag{7}$$

$$\rho_t(\omega_t - E_t\left[\rho_{t+1}\gamma_{t+1}^2\right]) = 1, \tag{8}$$

$$E\left[\log|\phi_t|\right] < 0. \tag{9}$$

The key relation is (5). It is a generalization of Hansen's (1985) formula for the process followed by the optimal instrument in a homoskedastic environment, which we will see in Section 3. Here, in contrast to Hansen (1985), ϕ_t and ρ_t are time varying, and $\boldsymbol{\delta}_t$ can be viewed as a generalized projection of the discounted sum of leaded conditional scores onto the space of instruments. The conditions (6), (7) and (8) determine ϕ_t, $\boldsymbol{\delta}_t$ and ρ_t, respectively, while (9) rules out unstable solutions of the nonlinear equation (6). The nature of the parameter processes ϕ_t, $\boldsymbol{\delta}_t$ and ρ_t suggests calling ϕ_t the *discount process*, $\boldsymbol{\delta}_t$ – the *forcing process*, and ρ_t – the *weighting multiplier.*

2.2 The multiple equations case

In the multiple equation case, the system of conditional moment restrictions is

$$E_t\left[\mathbf{e}_t\right] = \mathbf{0}, \tag{10}$$

and $\mathbf{e}_t$ is $s \times 1$, where $s > 1$. Define the $k \times s$ conditional score matrix

$$\mathrm{D}_t \equiv E_t\left[\frac{\partial f(\boldsymbol{\beta}, \mathbf{x}_t)}{\partial \boldsymbol{\beta}}\right], \tag{11}$$

and $s \times s$ matrices

$$\Omega_t \equiv E_t[\mathbf{e}_t\mathbf{e}_t'], \quad \Gamma_t \equiv E_t[\mathbf{e}_{t-1}\mathbf{e}_t'], \tag{12}$$

the conditional variance and conditional first-order autocovariance matrices of the error vector. We again assume away higher-order conditional serial correlation in the error $\mathbf{e}_t$, i.e. $E_t[\mathbf{e}_t\mathbf{e}_{t-j}'] = \mathrm{O}$ for $j > 1$, and use the Euclidean matrix norm $|A| = \sqrt{\varrho\left(A'A\right)}$, where $\varrho\left(\cdot\right)$ is the spectral radius.

Example 4 *The Hansen and Singleton (1996) temporal aggregation model corresponds to* $\boldsymbol{\beta} = (\gamma, \delta, \sigma^2)'$, $\mathbf{x}_t = (c_{t+2}, c_{t+1}, c_t, q_{t+2}, q_{t+1}, q_t)'$, *and*

$$f\left(\boldsymbol{\beta}, \mathbf{x}_t\right) = \begin{pmatrix} u_{t+2} \\ u_{t+2}^2 - 2\sigma^2/3 \\ u_{t+2}^2/4 - u_{t+2}u_{t+1} \end{pmatrix},$$

where $u_{t+1} = \gamma\left(c_{t+1} - c_t\right) + \left(q_{t+1} - q_t\right) - \left(\delta - \sigma^2/2\right)$, c_t *and* q_t *are a consumption and asset price,* γ *and* δ *are preference parameters, and* σ^2 *is a variance measure of underlying Brownian motions. The basic instrument is* $\mathbf{z}_t = \left(c_t, q_t\right)'$.

Under suitable conditions, the optimal instrument takes the form (Anatolyev 2003)

$$\Xi_t = \Xi_{t-1}\Phi_t + \Delta_t\mathrm{P}_t, \tag{13}$$

where the stationary ergodic $\Im_t$-measurable processes, $s \times s$ matrix Φ_t, $s \times s$ symmetric almost surely positive definite matrix P_t, and $k \times s$ matrix Δ_t, satisfy the following system:

$$\Gamma_t + \Phi_t(\Omega_t + E_t[\Phi_{t+1}\Gamma_{t+1}']) = 0, \tag{14}$$

$$\Delta_t = \mathrm{D}_t + E_t[\Delta_{t+1}\Phi_{t+1}'], \tag{15}$$

$$\mathrm{P}_t(\Omega_t - E_t\left[\Gamma_{t+1}\mathrm{P}_{t+1}\Gamma'_{t+1}\right]) = \mathrm{I}_s, \tag{16}$$

$$\lambda(\Phi) \equiv \lim_{T\to\infty} \frac{1}{T} \log |\Phi_T\Phi_{T-1}\cdots\Phi_2\Phi_1| < 0. \tag{17}$$

In the single equation case of the previous subsection, the negativity of the top Lyapounov exponent $\lambda(\Phi)$ is equivalent to the negativity of $E\left[\log|\phi_t|\right]$. In the multiple equation case, the condition $E\left[\log|\Phi_t|\right] < 0$ is too strong, because the inequality in $|\Phi_T\cdots\Phi_1| \leq |\Phi_T|\cdots|\Phi_1|$ may not be tight. For instance, the norm of the companion matrix Φ of a stationary ARMA process is bigger than unity, even though $\lim_{T\to\infty} |\Phi^T| = 0$. The following Lemma may be found useful.

Lemma 1 *(Bougerol and Picard 1992) Let A_t be a stationary ergodic matrix process with finite $E\left[\max\left(0, \log|A_t|\right)\right]$ such that almost surely*

$$\lim_{T\to\infty} |A_TA_{T-1}\cdots A_2A_1| = 0.$$

Then $\lambda(A) < 0$.

Thus, when Φ_t has a triangular structure (as in the Heaton–Ogaki example of Section 4), it is sufficient to verify that $E[\log|\lambda_{\max}(\Phi_t)|] < 0$, where $\lambda_{\max}$ is maximal diagonal element (which is the same as maximal eigenvalue). If this is not a case, one may impose existence of matrix process S_t such that $E[\log|S_t\Phi_tS_t^{-1}|] < 0$. For more on these issues, see Pötscher and Prucha (1997, p.70 and footnote 25).

3 The approximately optimal instrument

3.1 The single equation case

Unfortunately, the system (6)–(9) generally cannot be solved for ϕ_t, $\boldsymbol{\delta}_t$, and ρ_t. Thus, we want to find an approximate but explicit solution to the system treating the known instruments in the special cases of no conditional heteroskedasticity and no serial correlation as benchmarks. We then will employ the approximations to ϕ_t, $\boldsymbol{\delta}_t$, and ρ_t to construct an *approximately optimal instrument* via a recursion similar to (5).

Note first that if there were no conditional heteroskedasticity and serial correlation, the ideal instrument would be the conditional score $\mathbf{d}_t$.

Consider the following instrument which would be optimal if there were no conditional heteroskedasticity (Hansen 1985)[1]:

$$\boldsymbol{\zeta}_t^H = \boldsymbol{\zeta}_{t-1}^H\theta + \frac{1}{\sigma^2}\sum_{i=0}^{\infty}\theta^i E_t\left[\mathbf{d}_{t+i}\right], \tag{18}$$

where σ^2 is a variance of the Wold innovation of e_t, and θ is a negative of its implied moving average coefficient, i.e. $e_t = \varepsilon_{t+1} - \theta\varepsilon_t$, $\sigma^2 = E[\varepsilon_t^2]$. The construction of $\boldsymbol{\zeta}_t^H$ may be viewed

[1] An instrument of type (18) was used in empirical work, for example, by West and Wilcox (1996) in a homoskedastic environment and by Hansen and Singleton (1996) in both homo- and heteroskedastic environments.

as the approximation of ϕ_t, $\boldsymbol{\delta}_t$, and ρ_t correspondingly by

$$\phi^H = \theta, \quad \boldsymbol{\delta}_t^H = \sum_{i=0}^{\infty} \theta^i E_t \left[\mathbf{d}_{t+i}\right], \quad \rho^H = \frac{1}{\sigma^2}.$$

What makes the instrument (18) different from the instrument $\mathbf{d}_t$ is the nonzeroness of ϕ^H and the dynamic structure of $\boldsymbol{\delta}_t^H$. We will therefore treat ϕ^H and $\boldsymbol{\delta}_t^H$ as benchmarks, and look for approximate deviations of ϕ_t and $\boldsymbol{\delta}_t$ from ϕ^H and $\boldsymbol{\delta}_t^H$ when the error is supplemented by conditional heteroskedasticity.

On the other hand, if the error was conditionally serially uncorrelated, the optimal instrument would be[2]

$$\boldsymbol{\zeta}_t^C = \frac{\mathbf{d}_t}{\omega_t}, \tag{19}$$

The construction of $\boldsymbol{\zeta}_t^C$ may be viewed as the approximation of ϕ_t, $\boldsymbol{\delta}_t$, and ρ_t correspondingly by

$$\phi^C = 0, \quad \boldsymbol{\delta}_t^C = \mathbf{d}_t, \quad \rho_t^C = \frac{1}{\omega_t}.$$

What makes the instrument (19) different from the instrument $\mathbf{d}_t$ is the time-varying weighting by ρ_t^C. We will therefore treat ρ_t^C as a benchmark, and look for an approximate deviation of ρ_t from ρ_t^C when the error is supplemented by conditional serial correlation.

Technically, we rewrite the equation (6) that determines the discount process ϕ_t as

$$E_t\left[F_\phi\left(\phi_t, \phi_{t+1}, \omega_t, \gamma_t, \gamma_{t+1}\right)\right] = 0, \tag{20}$$

where $F_\phi(\phi_t, \phi_{t+1}, \omega_t, \gamma_t, \gamma_{t+1}) \equiv \gamma_t + \phi_t\left(\omega_t + \phi_{t+1}\gamma_{t+1}\right)$, and linearize it with respect to all arguments of F around the "homoskedasticity point" $H \equiv \left(\phi^H, \phi^H, \omega^H, \gamma^H, \gamma^H\right)$. For any variable u_t, define $\Delta u_t \equiv u_t - u^H$. Linearization yields the following linear equation for ϕ_t:

$$E_t\left[\begin{array}{c} \left.\frac{\partial F_\phi}{\partial \phi_t}\right|_H \Delta\phi_t + \left.\frac{\partial F_\phi}{\partial \phi_{t+1}}\right|_H \Delta\phi_{t+1} + \left.\frac{\partial F_\phi}{\partial \omega_t}\right|_H \Delta\omega_t + \left.\frac{\partial F_\phi}{\partial \gamma_t}\right|_H \Delta\gamma_t \\ + \left.\frac{\partial F_\phi}{\partial \gamma_{t+1}}\right|_H \Delta\gamma_{t+1} + R_{t+1}^F \end{array}\right] = 0, \tag{21}$$

where R_{t+1}^F contains higher-order terms. Collecting linear terms together in equation (21) and getting rid of higher-order ones, we end up with a linear stochastic difference equation with respect to the first-order approximation $\phi_t^{(1)}$ for ϕ_t, with a unique stationary solution

$$\phi_t^{(1)} = \theta - \frac{1}{\sigma^2}\left(\gamma_t + \theta\omega_t + \sum_{i=1}^{\infty} \theta^{2i} E_t\left[\theta\omega_{t+i} + 2\gamma_{t+i}\right]\right). \tag{22}$$

We can go further and consider the quadratic approximation to get a more refined solution for ϕ_t. Let us expand the ϕ_t-equation in the Taylor series up to quadratic terms:

$$E_t\left[\begin{array}{c} \left.\frac{\partial F_\phi}{\partial \phi_t}\right|_H \Delta\phi_t + \left.\frac{\partial F_\phi}{\partial \phi_{t+1}}\right|_H \Delta\phi_{t+1} + \left.\frac{\partial F_\phi}{\partial \omega_t}\right|_H \Delta\omega_t + \left.\frac{\partial F_\phi}{\partial \gamma_t}\right|_H \Delta\gamma_t \\ + \left.\frac{\partial F_\phi}{\partial \gamma_{t+1}}\right|_H \Delta\gamma_{t+1} \left.\frac{\partial^2 F_\phi}{\partial \phi_t \partial \phi_{t+1}}\right|_H \Delta\phi_t\Delta\phi_{t+1} + \left.\frac{\partial^2 F_\phi}{\partial \phi_t \partial \omega_t}\right|_H \Delta\phi_t\Delta\omega_t \\ + \left.\frac{\partial^2 F_\phi}{\partial \phi_t \partial \gamma_{t+1}}\right|_H \Delta\phi_t\Delta\gamma_{t+1} + \left.\frac{\partial^2 F_\phi}{\partial \phi_{t+1} \partial \gamma_{t+1}}\right|_H \Delta\phi_{t+1}\Delta\gamma_{t+1} + R_{t+1}^F \end{array}\right] = 0, \tag{23}$$

[2]The superscript "C" is an abbreviation of "Chamberlain", with whom such instrument is associated, due to the important paper Chamberlain (1987).

where R^F_{t+1} contain higher-order terms. Collecting the second-order terms together yields a stochastic difference equation with respect to the second-order approximation $\phi_t^{(2)}$ for ϕ_t, with a unique stationary solution

$$\phi_t^{(2)} = \phi_t^{(1)} + \frac{1}{\theta}\left(\frac{1}{\sigma^2}\left(\phi_t^{(1)} - \theta\right)\left(\gamma_t + \theta\sigma^2\right) + \sum_{i=0}^{\infty}\theta^{2i}E_t\left[\left(\phi_{t+i}^{(1)} - \theta\right)^2\right]\right). \tag{24}$$

The expansion may be continued further, but the emerging approximations are too complex to be used in practice.

Now we consider the forcing process $\boldsymbol{\delta}_t$ determined by (7). We approximate $F_\delta(\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t+1}, \phi_{t+1}) \equiv -\boldsymbol{\delta}_t + \mathbf{d}_t + \phi_{t+1}\boldsymbol{\delta}_{t+1}$ around $H = \left(\boldsymbol{\delta}_t^H, \boldsymbol{\delta}_{t+1}^H, \phi^H\right)$ to end up with a linear stochastic difference equation with respect to the first-order approximation $\boldsymbol{\delta}_t^{(1)}$ for $\boldsymbol{\delta}_t$, with a unique stationary solution

$$\boldsymbol{\delta}_t^{(1)} = \boldsymbol{\delta}_t^H + \sum_{i=1}^{\infty}\theta^{i-1}E_t\left[\left(\phi_{t+i}^{(1)} - \theta\right)\boldsymbol{\delta}_{t+i}^H\right]. \tag{25}$$

Similarly, we expand the $\boldsymbol{\delta}_t$-equation up to quadratic terms to get

$$\boldsymbol{\delta}_t^{(2)} = \boldsymbol{\delta}_t^{(1)} + \sum_{i=1}^{\infty}\theta^{i-1}E_t\left[\left(\phi_{t+i}^{(2)} - \phi_{t+1}^{(1)}\right)\boldsymbol{\delta}_{t+i}^H + \phi_{t+i}^{(1)}\left(\boldsymbol{\delta}_{t+i}^{(1)} - \boldsymbol{\delta}_{t+i}^H\right)\right]. \tag{26}$$

Again, the expansion may be continued further, but the emerging approximations are too complex to be used in practice.

Finally, we consider the weighting multiplier ρ_t determined by (8). We approximate $F_\rho(\rho_t, \rho_{t+1}, \gamma_{t+1}) \equiv 1 - \rho_t\left(\omega_t - \rho_{t+1}\gamma_{t+1}^2\right)$ around $C = \left(\rho_t^C, \rho_{t+1}^C, 0\right)$ to find that

$$\rho_t^{(1)} = \frac{1}{\omega_t}, \tag{27}$$

i.e. the first-order approximation $\rho_t^{(1)}$ for ρ_t coincides with ρ_t^C. The second-order approximation for ρ_t is

$$\rho_t^{(2)} = \frac{1}{\omega_t}\left(1 + \frac{1}{\omega_t}E_t\left[\frac{\gamma_{t+1}^2}{\omega_{t+1}}\right]\right). \tag{28}$$

Note that $\operatorname{ess\,inf}|\gamma_t| > 0$ implies $\operatorname{ess\,sup}|\rho_t^{(1)}| > 0$ and $\operatorname{ess\,sup}|\rho_t^{(2)}| > 0$. Once again, the expansion may be continued further, but the emerging approximations are too complex to be used in practice.

The *approximately optimal instrument* $\boldsymbol{\zeta}_t^{(jkl)}$ uses j^{th}-order approximation for ϕ_t, k^{th}-order for δ_t, and l^{th}-order for ρ_t, where by 0^{th} order we mean ϕ^H, $\boldsymbol{\delta}_t^H$ and ρ_t^C, respectively. The approximately optimal instrument follows

$$\boldsymbol{\zeta}_t^{(jkl)} = \boldsymbol{\zeta}_{t-1}^{(jkl)}\phi_t^{(j)} + \rho_t^{(l)}\boldsymbol{\delta}_t^{(k)}. \tag{29}$$

Since we use approximations for ϕ_t, $\boldsymbol{\delta}_t$ and ρ_t in place of the true processes to construct the instrument, we have to ensure proper behavior of $\boldsymbol{\zeta}_t^{(jkl)}$. This means two things. First, we require stationarity: there must exist a unique stationary solution to (29) with approximated ϕ_t, $\boldsymbol{\delta}_t$ and ρ_t. Second, we need to ensure finiteness of fourth moments of the constructed instrument. While the former aim is relatively easily achieved, the latter is much more challenging, due to the time dependence of the "AR coefficient" $\phi_t^{(j)}$, the absolute value of which can take values higher than unity. To kill two birds with one stone, we deliberately simplify the task at the expense of possible efficiency losses. The following general Lemma, which can be easily proved using Brandt (1986) and the triangular inequality, will be helpful.

Lemma 2 *Suppose that the $k \times s$ matrix process Ψ_t satisfies the recurrence*

$$\Psi_t = \Psi_{t-1} A_t + B_t, \tag{30}$$

where A_t is an $s \times s$ and B_t is a $k \times s$ matrix processes such that: (a) A_t and B_t are stationary and ergodic; (b) $\operatorname{esssup} |A_t| < 1$*; (c)* $E[|B_t|^4] < \infty$*. Then there exists a stationary ergodic solution Ψ_t of (30) such that $E[|\Psi_t|^4] < \infty$.*

To force $|\phi_t^{(j)}|$ to be bounded from above by 1, we use the following trimming scheme. Fix a generic small positive number ϵ, like 10^{-2}, say. Define the trimming operator "$^-$" by

$$\phi^- = \min\{1 - \epsilon, \max\{-1 + \epsilon, \phi\}\}.$$

That is, "$^-$" trims large $|\phi|$ by setting $\phi > 1 - \epsilon$ to $1 - \epsilon$, $\phi < -1 + \epsilon$ to $-1 + \epsilon$. Then instead of (29) we can use the following recursion:

$$\boldsymbol{\zeta}_t^{(jkl)} = \boldsymbol{\zeta}_{t-1}^{(jkl)} (\phi_t^{(j)})^- + \rho_t^{(l)} \boldsymbol{\delta}_t^{(k)}. \tag{31}$$

Then $\operatorname{esssup}(\phi_t^{(j)})^- < 1$ and $E[|\rho_t^{(l)} \boldsymbol{\delta}_t^{(k)}|^4] \leq E[|\boldsymbol{\delta}_t^{(k)}|^4] \operatorname{esssup} |\rho_t^{(l)}| < \infty$ if $\delta_t^{(k)}$ has finite fourth moments and $\rho_t^{(l)}$ is bounded from below, so the prerequisites of Lemma 2 are satisfied. Of course, the asymptotic variance of the instrument $\boldsymbol{\zeta}_t^{(jkl)}$ depends on the trimming parameter ϵ, which it is reasonable to set to a small number to distort $\phi_t^{(j)}$ to the least degree.

3.2 The multiple equations case

The covariance and variance parameters under homoskedasticity are given by $\Gamma^H = -\Sigma\Theta'$, $\Omega^H = \Sigma + \Theta\Sigma\Theta'$, where Θ and Σ are determined from the Wold decomposition of $\mathbf{e}_t$: $\mathbf{e}_t = \boldsymbol{\varepsilon}_{t+1} - \Theta\boldsymbol{\varepsilon}_t$ and $\Sigma \equiv E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t']$. The zeroth-order approximation $\Phi_t^{(0)} = \Phi^H$ for Φ_t is one that satisfies the matrix quadratic equation

$$\left(\Phi^H\right)^2 \Theta\Sigma - \Phi^H (\Sigma + \Theta\Sigma\Theta') + \Sigma\Theta' = 0. \tag{32}$$

Note that the unstable solution is trivially Θ^{-1}, but the stable one is *not* Θ. Let representation $\mathbf{e}_t = \boldsymbol{\varepsilon}_{t+1} - \Theta\boldsymbol{\varepsilon}_t$ be invertible, i.e. all s eigenvalues π_i, $i = 1, \ldots, s$, of Θ lie strictly inside the unit circle on the complex plane.[3] Construct $s \times s$ matrices $\Pi \equiv \operatorname{diag}(\pi_1, \ldots, \pi_s)$ and $\Psi \equiv (x_1, \ldots, x_s)$, where each column x_i is a solution of the following system of s equations:

$$\left[\pi_i^2 \Sigma\Theta' - \pi_i (\Sigma + \Theta\Sigma\Theta') + \Theta\Sigma\right] x_i = 0.$$

Then the stable solution of (32) is $\Phi^H = (\Psi^{-1}\Pi\Psi)'$. This follows, for example, from application of Theorems 3 and 4 of Uhlig (1995).

[3] Some of these eigenvalues may be complex. However, even then the resulting solution Φ^H will be real-valued. See Uhlig (1995).

Anticipating the preferred choice of approximation orders $j = 1$, $k = 0$, and $l = 1$ in the single-equation case (see Section 4), we give approximations of same orders here. The linearization of (14) yields a difference equation with the stationary solution

$$\Phi_t^{(1)} = \Phi^H - \sum_{i=0}^{\infty} \left(\Phi^H\right)^i E_t \left[\Phi^H \Omega_{t+i} + \Gamma_{t+i} + \left(\Phi^H\right)^2 \Gamma'_{t+i+1}\right] (\Theta\Sigma)'^{\,-1} \Phi^H \left(\Theta\Sigma\,(\Theta\Sigma)'^{\,-1}\Phi^H\right)^i .$$

The $k \times s$ matrix Δ_t solves (15), with the zeroth-order approximation given by

$$\Delta_t^{(0)} = \Delta_t^H = \sum_{i=0}^{\infty} E_t \left[\mathrm{D}_{t+i} \left(\Phi^{H\prime}\right)^i\right].$$

For P_t the first-order approximation is

$$\mathrm{P}_t^{(1)} = \Omega_t^{-1}.$$

The approximately optimal instrument $\Xi_t^{(101)}$ follows

$$\Xi_t^{(101)} = \Xi_{t-1}^{(101)} (\Phi_t^{(1)})^- + \Delta_t^{(0)} \mathrm{P}_t^{(1)}.$$

For matrices, we generalize the trimming device "$^-$" in the following way. Take an arbitrary nonsingular $s \times s$ matrix A. The *basic structure* of A is decomposition $A = P\Delta Q'$, where P and Q are each orthonormal $s \times s$ matrices, i.e. $P'P = PP' = Q'Q = QQ' = \mathrm{I}_s$, and Δ is a diagonal matrix with strictly positive elements on the diagonal ordered in descending order (Green and Carroll 1976, section 5.7). This decomposition always exists. Observe that the L^2 norm of A is $|A| = \sqrt{\varrho\,(A'A)}$, where $\varrho\,(\cdot)$ is the spectral radius, and $A'A = Q\Delta P'P\Delta Q' = Q\Delta^2 Q'$. But this is the eigenstructure of symmetric positive definite matrix $A'A$, with matrix Δ^2 containing s real positive eigenvalues of $A'A$. If any of these exceed 1, we can deflate them to lie within $[0, (1-\epsilon)^2]$ in the same way we do trimming in the scalar case. By doing so we automatically force $|A|$ to be bounded from above by $1-\epsilon$. Thus, for matrix A the trimming algorithm goes as follows: (1) Compute $A'A$ and find its eigenstructure which yields the matrix of eigenvalues Δ^2 and the matrix Q of eigenvectors. (2) Find the square root Δ of Δ^2. (3) Compute the implied matrix P by $P = AQ\Delta^{-1}$. (4) Trim the diagonal entries of matrix Δ using operator "$^-$", and construct the trimmed A as $A^- = P\Delta^- Q'$.

3.3 Discussion

Similar transformations of nonlinear equations by (log-)linearization and (more rarely) second-order expansion often occur in estimation of Euler equations and are a standard tool in the asset pricing and business cycles literatures. There are several studies analyzing the effects of such transformations on the precision of estimates of parameters in Euler equations. Some of these studies document large biases in those estimates, which, however, largely result from certain econometric problems possibly caused by linearization, rather than from the approximation error *per se* (Attanasio and Low 2002). For example, Carroll (2001) finds out that both OLS and IV estimates of preference parameters are biased, the reason being that the approximation error is endogenous making OLS estimates inconsistent and IV estimates corrupt in the absence of reliable instruments. Ludvigson and Paxson (2001) also obtain large biases, but, as Attanasio and Low (2002) note, there is no identifying information in their model and convincing instruments for IV estimation. Attanasio and Low (2002) find no systematic biases for IV estimates of structural parameters in loglinearized equations, and show that "for all parameter specifications, the performance of the non-linear GMM estimator is considerably worse than that of the estimators based on the log-linearized equations." It can be concluded from this literature that the approximations can indeed be used as a constructive tool, especially when there are no attractive alternatives.

A natural question arises: does a higher-order approximation of parameter processes imply a better approximation of the optimal instrument? Even though this question is rhetorical for practical purposes because approximations beyond the second are prohibitively complex, it is interesting to know if the proposed method is the way to attain the efficiency bound. Unfortunately, the answer is no. It is true that for infinitesimally small conditional heteroskedasticity and serial correlation the parameter processes may be approximated to any desired precision, but it is unlikely to reach perceptible efficiency gains, and it is in fact more likely to incur efficiency losses in feasible implementation with samples of typical size. In contrast, in a situation of strong heteroskedasticity and serial correlation, the Taylor expansion of infinite order does not guarantee yielding true parameter processes, or even convergence of the infinite series. As an illustration, consider a function which is well-behaved in the considered sense, $\exp(x)$, and suppose we want to approximate this function by a Taylor series expansion around $x = 0$. Because the function $x \mapsto \exp(x)$ is analytic, the Taylor series evaluated at $x = 1$, $\sum_{n=0}^{\infty} (n!)^{-1}$, converges to $\exp(1)$. However, even though the function $x \mapsto \exp(x) - 1.9525 \cdot \exp(-1/x^2)$ has the same Taylor expansion around $x = 0$, it has a different value at $x = 1$. At the same time, the first-order approximation $\sum_{n=0}^{1} (n!)^{-1} = 2$ for this function at $x = 1$ happens to be exact (which is purely incidental), while higher order terms introduce discrepancies. We conclude that the approximately optimal instrument should be really viewed as a heuristic *approximation*, and not as an intermediate step towards the truly optimal instrument; whether this approximation delivers efficiency gains in practice can only be evaluated via asymptotic calculations and Monte–Carlo studies. In any case, however, a researcher does not risk consistency of estimates by using the approximation, and most likely wins in terms of efficiency as asymptotic calculations and simulations will show.

Now, to construct a feasible optimal instrument, certain conditional expectations need to be estimated. In the rest of the subsection, we discuss some issues of their estimation, relegating details to the simulation and application sections. As seen from the formulas for components of approximately optimal instruments, one needs estimates of (an infinite number of) conditional expectations of future (from the viewpoint of period t) conditional scores, variances and autocovariances, and possibly other, more complicated, objects. Since the parametric approach requires knowledge not implied by the model and thus is not attractive, we rely on nonparametric estimates of the involved auxiliary functions. More concretely, we employ the Nadaraya–Watson (NW) nonparametric estimator with a global bandwidth, using the nonparametric corrected asymptotic final prediction error (CAFPE) criterion of Tschernig and Yang (2000) to select significant lags and an optimal value of the bandwidth. Alternative possibilities include nearest neighborhood regressions (Robinson 1987) and series expansions (Newey 1990). Among kernel-based methods, we intentionally do not use a more precise local polynomial regression and/or variable bandwidth, for two reasons. First, we do not want to increase computational costs in simulations which are already appreciable. Second, we would like to give conservative evidence on the performance of our estimator so that it may be even more enhanced by using more advanced nonparametric methods.

Whichever approach is adopted, one has to deal with infinite summations present in (18), (22), (24), etc. This is not as hard to do as it may seem. Take, for example, the infinite summation

$$\sum_{i=0}^{\infty} \theta^i E_t \left[\mathbf{d}_{t+i} \right]. \tag{33}$$

Note that the conditional expectation of $\mathbf{d}_{t+i}$ given $\Im_t$ converges to its unconditional counterpart $E\left[\partial f(\boldsymbol{\beta}, \mathbf{x}_t)/\partial \boldsymbol{\beta}\right]$ as $i \to \infty$. Note also that the weight θ^i in front of $E_t\left[\mathbf{d}_{t+i}\right]$ converges to zero as $i \to \infty$. Thus, it is sufficient to estimate but a finite number of summands up to some $i = i_0$, and substitute the tail of the series by a tail composed of corresponding unconditionals. Informally, let us assume that the error from estimation of an expectation is δ_E, and the error from substitution of a conditional by an unconditional estimator is δ_S. Then the maximum estimation error in (33) would be $\delta_E / (1-\theta)$ if all terms could be estimated, while the total error resulting from estimation and substitution is $\left(\delta_E + \theta^{i_0+1}\delta_S\right)/(1-\theta)$. That is, if i_0 is high enough, the difference in the two errors is negligeable. The computations above assume that δ_S is the same for each i, while in fact the substitution error goes down as i increases, which further reduces the difference between the two errors.

We thus estimate nonparametrically, according to the procedure described in detail in Section 5, the summands from $i = 0$ increasing i until the nonparametric procedure selects no significant lags for two i's in a row, say for i_0 and i_0+1, or until $\theta^{i_0}/(1-\theta)$ falls below 0.001, whichever happens earlier. Starting from $i = i_0$ we use the sample mean of $\partial f(\hat{\boldsymbol{\beta}}, \mathbf{x}_t)/\partial \boldsymbol{\beta}$, i.e. a nonparametric estimate of the unconditional expectation $E\left[\partial f(\boldsymbol{\beta}, \mathbf{x}_t)/\partial \boldsymbol{\beta}\right]$, in place of $E_t\left[\mathbf{d}_{t+i}\right]$. Similarly, to estimate

$$\sum_{i=1}^{\infty} \theta^{2i} E_t \left[\theta\omega_{t+i} + 2\gamma_{t+i} \right], \tag{34}$$

we estimate $E_t\left[\theta\omega_{t+i} + 2\gamma_{t+i}\right]$ by nonparametrically regressing $\hat{\theta}\hat{e}_{t+i}^2 + 2\hat{e}_{t+i}\hat{e}_{t+i-1}$ on the basic instruments and their lags, increasing i from $i = 1$ until the procedure selects no significant lags for two i's in a row, or until $\theta^{2i}/\left(1-\theta^2\right)$ falls below 0.001. Here, $\hat{\boldsymbol{\beta}}$, $\hat{\theta}$, and $\hat{e}_t$ are preliminary consistent estimates of $\boldsymbol{\beta}$, θ and e_t, respectively.

4 Asymptotic comparisons

4.1 An example with quadratic heteroskedasticity

We will use the following data generating mechanism for the calibration of asymptotic gains:

$$\begin{aligned}
y_t &= \beta z_t + e_t, \quad e_t = \varepsilon_{t+1} - \theta\varepsilon_t, \quad \Im_t = \sigma\left(z_t, z_{t-1}, \ldots\right); \\
\varepsilon_t &= \nu_t \sqrt{1 - \lambda + \lambda\left(1-\varphi^2\right) z_t^2}, \quad z_t = \varphi z_{t-1} + \eta_t, \quad \left(\eta_t, \nu_t\right)' \sim IID\,\mathcal{N}\left(0, \mathrm{I}_2\right).
\end{aligned}$$

Here $\theta \in (-1,1)$, $\varphi \in (0,1)$, and $\lambda \in [0,1)$. The basic instrument is z_t. The aim is estimation of β from the data on y_t and z_t. We have (all constants $\kappa_{..}$ here and below may be found in the Appendix available from the author on request):

$$\begin{aligned}\omega_t &= \kappa_{\omega 1} + \kappa_{\omega 2} z_t^2, \\ \gamma_t &= \kappa_{\gamma 1} + \kappa_{\gamma 2} z_t^2, \\ d_t &= z_t.\end{aligned}$$

The zeroth-order approximation to the parameters is

$$\begin{aligned}\phi_t^{(0)} &= \theta, \\ \delta_t^{(0)} &= \kappa_{\delta 1} z_t, \\ \rho_t^{(0)} &= \frac{1}{\kappa_{\omega 1} + \kappa_{\omega 2} z_t^2}.\end{aligned}$$

The first-order approximation to the parameters is

$$\begin{aligned}\phi_t^{(1)} &= \kappa_{\phi 1} + \kappa_{\phi 2} z_t^2, \\ \delta_t^{(1)} &= \kappa_{\delta 2} z_t + \kappa_{\delta 3} z_t^3, \\ \rho_t^{(1)} &= \frac{1}{\kappa_{\omega 1} + \kappa_{\omega 2} z_t^2}.\end{aligned}$$

The second-order approximation to the parameters is

$$\begin{aligned}\phi_t^{(2)} &= \kappa_{\phi 3} + \kappa_{\phi 4} z_t^2 + \kappa_{\phi 5} z_t^4, \\ \delta_t^{(2)} &= \kappa_{\delta 3} z_t + \kappa_{\delta 4} z_t^3 + \kappa_{\delta 5} z_t^5, \\ \rho_t^{(2)} &= \frac{\kappa_{\rho 1} + \kappa_{\rho 2} z_t^2 + \kappa_{\rho 3} \left(z_t\right)}{\left(\kappa_{\omega 1} + \kappa_{\omega 2} z_t^2\right)^2}.\end{aligned}$$

One can see that both $\phi_t^{(1)}$ and $\phi_t^{(2)}$ are polynomials in the basic instrument z_t with unbounded support. This points at the importance of using the trimming device "$-$".

The additional instruments that we use in comparisons are the basic instrument z_t implied by the OLS estimator, and the West–Wong–Anatolyev instrument (West, Wong and Anatolyev 2002), which is optimal in the class of linear combinations of the present and past basic instruments, and thus attains the efficiency bound in the class of GMM estimators that use as instruments lags of the basic instrument. An interesting feature of the present example is the asymptotic equivalence of the West–Wong–Anatolyev instrument and the one that would be optimal if there were no conditional heteroskedasticity (the proof of this fact is in the aforementioned Appendix). Approximately optimal instruments include $\zeta_t^{(101)}$, the one that uses approximations $\phi_t^{(1)}$, $\delta_t^{(0)}$ and $\rho_t^{(1)}$, which turns out to be a reasonable compromise between an instrument's complexity and efficiency gains in this example, and three other versions where one of the parameters is higher-order approximated compared to $\zeta_t^{(101)}$, that is, either $\phi_t^{(2)}$ is used in place of $\phi_t^{(1)}$, or $\delta_t^{(1)}$ in place of $\delta_t^{(0)}$, or $\rho_t^{(2)}$ in place of $\rho_t^{(1)}$. For the basic and West–Wong–Anatolyev instruments the asymptotic variances are computed using closed-form formulas, for the approximately optimal instruments – by simulations, with sample sizes of at least 10^8 observations.

Table 1 presents limited but typical evidence on relative asymptotic performance of the considered estimators and corresponds to the case of moderate heteroskedasticity ($\lambda = 0.5$) and moderate persistence in the basic instrument ($\varphi = 0.5$). The degree of serial correlation

θ is set to ± 0.1, ± 0.5, ± 0.9. In judging the applicability potential of various approximately optimal instruments, we are guided by asymptotic efficiency gains, on the one hand, and by complexity of their forms, on the other hand. The latter factor is very important because when a more complex approximation tends to yield slight efficiency gains, these gains are not likely to be realized with a feasible version.

A quick look at the table reveals significant asymptotic efficiency gains from the use of the approximately optimal instrument $\zeta_t^{(101)}$ relative to the asymptotic variance provided by the class of GMM estimators, especially when the serial correlation is strong. Often switching from the linearly optimal instrument to the nonlinear approximately optimal instrument provides more efficiency gains than switching from the basic instrument to the instrument optimal in the entire linear class of instruments. The numbers reveal restrictiveness of the space of instruments that are linear in lags of the basic instrument and a promise of the adopted approach. In many unreported experiments the same pattern emerges, and in no case have we obtained efficiency losses for the instrument $\zeta_t^{(101)}$ (which cannot be excluded in principle).

As far as higher-order approximations are concerned, the use of $\phi_t^{(2)}$ in place of $\phi_t^{(1)}$ tends to decrease efficiency (the entries with dashes indicate awkwardly high variances). Taking into account the difficulty of its derivation, a good advice is to forget about its exploitation. The use of $\delta_t^{(1)}$ in place of $\delta_t^{(0)}$ is able to provide further slight efficiency gains (as well as slight efficiency losses), but its form is far too complex. The use of $\rho_t^{(2)}$ in place of $\rho_t^{(1)}$ has similar effects, and although it never shows efficiency losses, the potential gains seem too small to justify its complexity and computational costs, although one may think of its exploitation in problems with strong serial correlation.

To appraise the significance of the trimming device "$^-$", an additional experiment was run where the trimming of $\phi_t^{(1)}$ in the construction of $\zeta_t^{(101)}$ was abolished. This brought no change in asymptotic variance for small values of $|\theta|$ (and thus for small $|\kappa_{\phi 2}|$) but worsened it for values of θ farther away from zero. Such deterioration indicates the importance of trimming big values of $|\phi_t^{(1)}|$ that the true $|\phi_t|$ is unlikely to take.

4.2 The Heaton–Ogaki example

Heaton and Ogaki (1991) present an econometric example where it is possible to obtain an exact expression for the efficiency bound. Naturally, in this case it is also possible to write out the explicit solution for the parameters of the optimal instrument, which we do below. Unfortunately, in order to accomplish either goal, one has to assume normality of the fundamental process, which nullifies this example's practical significance.

Let $\mathbf{w}_t$ be a serially independent standard normal $q \times 1$ vector, and u_t be a two-period ahead forecast error with the Wold representation $u_t = \boldsymbol{\nu}_0'\mathbf{w}_t + \boldsymbol{\nu}_1'\mathbf{w}_{t-1}$, where $\boldsymbol{\nu}_0$ and $\boldsymbol{\nu}_1$ are $q \times 1$ vectors of constants. Observable at time t is the $q \times 1$ vector $\mathbf{x}_t$, and the space of instruments is $\Im_t = \sigma(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots)$. Let u_t be connected to $\mathbf{x}_t$ via $u_t = \left(1, \beta, \mathbf{0}_{q-2}'\right)\mathbf{x}_t$, where β is a scalar parameter of interest. The rational expectations hypothesis imposes the restriction

$$E_t\left[u_{t+2}\right] = 0.$$

Under the assumptions made, the error in this equation is conditionally homoskedastic. There is conditional heteroskedasticity in another restriction, the conditional analog of Working's (1960) result on temporal aggregation:

$$E_t\left[u_{t+2}\left(\rho u_{t+2}-u_{t+1}\right)\right]=0,$$

where $\rho \equiv \frac{\boldsymbol{\nu}_0'\boldsymbol{\nu}_1}{\boldsymbol{\nu}_0'\boldsymbol{\nu}_0+\boldsymbol{\nu}_1'\boldsymbol{\nu}_1}=\frac{1}{4}$. The disturbance vector of the two equation system is

$$\mathbf{e}_t=\begin{pmatrix} u_{t+2} \\ u_{t+2}\left(\rho u_{t+2}-u_{t+1}\right)\end{pmatrix}.$$

The observational equation for x_t is $\mathbf{x}_t = H\mathbf{y}_t$, where the law of motion of the $p\times 1$ state vector $\mathbf{y}_t$ is

$$\mathbf{y}_t = A\mathbf{y}_{t-1}+C\mathbf{w}_t,$$

where A is a stable $p\times p$ matrix, C is a $p\times q$ matrix, and H is a $q\times p$ matrix. These constants should be consistent with $\left(1,\beta,\mathbf{0}_{q-2}'\right)HC=\boldsymbol{\nu}_0'$, $\left(1,\beta,\mathbf{0}_{q-2}'\right)HAC=\boldsymbol{\nu}_1'$, $\left(1,\beta,\mathbf{0}_{q-2}'\right)HA^iC=\mathbf{0}$, $i\geq 2$. Then

$$\mathrm{D}_t=\begin{pmatrix} -\mathbf{h}A^2\mathbf{y}_t, & r_d+(\mathbf{h}A^2\mathbf{y}_t)(\boldsymbol{\nu}_1'\mathbf{w}_t)\end{pmatrix},$$

where $\mathbf{h}\equiv\left(0,1,\mathbf{0}_{q-2}'\right)H$, $r_d\equiv -\mathbf{h}\left(AC\boldsymbol{\varrho}_1+C\boldsymbol{\varrho}_0\right)$, $\boldsymbol{\varrho}_0\equiv 2\rho\boldsymbol{\nu}_0-\boldsymbol{\nu}_1$, $\boldsymbol{\varrho}_1\equiv 2\rho\boldsymbol{\nu}_1-\boldsymbol{\nu}_0$.

The process Φ_t and the product $\Delta_t\mathrm{P}_t$ are

$$\Phi_t=\frac{1}{\xi_{11}\xi_{12}}\begin{pmatrix} -\xi_{12}\xi_{21} & 0 \\ \left(\xi_{22}\boldsymbol{\alpha}_{11}'-\xi_{12}\boldsymbol{\alpha}_{21}'\right)\mathbf{w}_t-\xi_{12}\boldsymbol{\alpha}_{22}'\mathbf{w}_{t-1} & -\xi_{11}\xi_{22}\end{pmatrix},$$

$$\Delta_t\mathrm{P}_t=\begin{pmatrix} \mathbf{r}_1\mathbf{y}_t+r_2(\boldsymbol{\nu}_1'\mathbf{w}_t) & r_2\end{pmatrix},$$

where constants ξ_{ij} and $\boldsymbol{\alpha}_{ij}$ are defined in equations (4)–(10) of Heaton and Ogaki (1991), and the $1\times p$ vector $\mathbf{r}_1$ and the scalar r_2 are

$$\mathbf{r}_1=-\frac{1}{\xi_{11}^2}\mathbf{h}A^2\left(\mathrm{I}_p+\frac{\xi_{21}}{\xi_{11}}A\right)^{-1},\qquad r_2=\frac{1}{\xi_{12}^2}\left(1+\frac{\xi_{22}}{\xi_{12}}\right)^{-1}\left(r_d-\frac{\xi_{21}}{\xi_{11}}\mathbf{r}_1C\boldsymbol{\varrho}_1\right).$$

Note that $E\left[\max\left(0,\log\left|\Delta_t\mathrm{P}_t\right|\right)\right]<\infty$ and $E\left[\max\left(0,\log\left|\Phi_t\right|\right)\right]<\infty$ are satisfied due to normality of $\mathbf{w}_t$, and $\lambda(\Phi)<0$ because Φ_t has a triangular structure with diagonal elements that are less than unity in absolute value (see remarks in subsection 2.2). Finally, $E[|\Xi_t|^4]<\infty$ due to normality of $\mathbf{w}_t$. Note that unfolding the recursion for the optimal instrument Ξ_t by repeated substitutions gives that the second element of Ξ_t is constant, and the first element is a linear function of the present and all past $\mathbf{w}_t$. That is, the optimal instrument is equivalent to the West–Wong–Anatolyev instrument provided that $\mathbf{x}_t$ is chosen to be the basic instrument.

Now we derive the approximation for the optimal instrument. The variance and covariance processes are:

$$\Omega_t=(\boldsymbol{\nu}_0'\boldsymbol{\nu}_0+\boldsymbol{\nu}_1'\boldsymbol{\nu}_1)\begin{pmatrix} 1 & -\boldsymbol{\nu}_1'\mathbf{w}_t \\ -\boldsymbol{\nu}_1'\mathbf{w}_t & \varrho^2+(\boldsymbol{\nu}_1'\mathbf{w}_t)^2\end{pmatrix},$$

$$\Gamma_t=(\boldsymbol{\nu}_0'\boldsymbol{\nu}_1)\begin{pmatrix} 1 & -\boldsymbol{\nu}_1'\mathbf{w}_t \\ \boldsymbol{\varrho}_1'\mathbf{w}_t-\boldsymbol{\nu}_1'\mathbf{w}_{t-1} & -2\rho\varrho^2-(\boldsymbol{\nu}_1'\mathbf{w}_t)\left(\boldsymbol{\varrho}_1'\mathbf{w}_t-\boldsymbol{\nu}_1'\mathbf{w}_{t-1}\right)\end{pmatrix},$$

where $\varrho^2\equiv\boldsymbol{\nu}_0'\boldsymbol{\nu}_0-\rho\boldsymbol{\nu}_0'\boldsymbol{\nu}_1$. Consequently,

$$\Omega^H=(\boldsymbol{\nu}_0'\boldsymbol{\nu}_0+\boldsymbol{\nu}_1'\boldsymbol{\nu}_1)\begin{pmatrix} 1 & 0 \\ 0 & (\boldsymbol{\nu}_0'\boldsymbol{\nu}_0+\boldsymbol{\nu}_1'\boldsymbol{\nu}_1)(1-\rho^2)\end{pmatrix},\quad \Gamma^H=(\boldsymbol{\nu}_0'\boldsymbol{\nu}_1)\begin{pmatrix} 1 & 0 \\ 0 & (\boldsymbol{\nu}_0'\boldsymbol{\nu}_1)(2\rho^2-1)\end{pmatrix}.$$

Since both Ω^H and Γ^H are diagonal, Φ^H is too and is equal to $\Theta \equiv \operatorname{diag}(\theta_1, \theta_2)$, which is defined together with Σ via $\Theta = -\Gamma^H \Sigma^{-1}$ and $\mathrm{I}_2 + \Theta^2 = \Omega^H \Sigma^{-1}$. Then one may find θ_1 and θ_2 from equations

$$\frac{\theta_1}{1+\theta_1^2} = -\rho, \quad \frac{\theta_2}{1+\theta_2^2} = \rho^2 \frac{1-2\rho^2}{1-\rho^2},$$

subject to $|\theta_1| < 1$ ($\Rightarrow \theta_1 = -\frac{\xi_{21}}{\xi_{11}}$) and $|\theta_2| < 1$. The parameter processes of the approximately optimal instrument have the following forms:

$$\Phi_t^{(1)} = \Theta - \left[\Theta\Omega_t + \Gamma_t + \Theta^2 \boldsymbol{\nu}_0' \boldsymbol{\nu}_1 \begin{pmatrix} 1 & -\boldsymbol{\nu}_1' \mathbf{w}_t \\ 0 & (\boldsymbol{\nu}_0' \boldsymbol{\nu}_1)(2\rho^2 - 1) \end{pmatrix} \right] \Sigma^{-1},$$

$$\Delta_t^{(0)} = \begin{pmatrix} \xi_{11}^2 \mathbf{r}_1 \mathbf{y}_t & \dfrac{r_d + \theta_2 \mathbf{h} A^2 C \boldsymbol{\nu}_1}{1-\theta_2} + (\mathbf{h} A^2 \mathbf{y}_t)(\boldsymbol{\nu}_1' \mathbf{w}_t) \end{pmatrix},$$

$$\mathrm{P}_t^{(1)} = \frac{\rho}{\varrho^2 \boldsymbol{\nu}_0' \boldsymbol{\nu}_1} \begin{pmatrix} \varrho^2 + (\boldsymbol{\nu}_1' \mathbf{w}_t)^2 & \boldsymbol{\nu}_1' \mathbf{w}_t \\ \boldsymbol{\nu}_1' \mathbf{w}_t & 1 \end{pmatrix}.$$

To calibrate asymptotic losses of the approximately optimal IV estimator relative to the optimal IV estimator, let $q = 1$, $\mathbf{x}_t = (z_t \ z_{t-1})'$, $\nu_0 = 1$, $\nu_1 = 2 - \sqrt{3}$, so that $u_t = z_t + \beta z_{t-1} = w_t + \nu_1 w_{t-1}$, and

$$y_t = \begin{pmatrix} z_t \\ z_{t-1} \\ w_t \end{pmatrix}, \quad A = \begin{pmatrix} -\beta & 0 & \nu_1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad H = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad h \equiv (0 \ 1 \ 0).$$

Table 2 presents asymptotic variances of some IV estimators for several values of β. The "truly optimal" IV estimator is most efficient, and significantly beats the optimal IV estimator that ignores the second equation ("first equation optimal"), especially when β is close to ν_1. The "homoskedasticity optimal" instrument that would be optimal if there were no conditional heteroskedasticity captures much of the efficiency gains. However, the proposed "approximately optimal" instrument captures an overwhelming part of the further efficiency gains provided by the optimal instrument. Thus, the efficiency losses arising from the approximation error turn out to be tiny (only $0.8 \div 3.0\%$ for the utilized values of β), and show that the proposed instrument is able to nearly attain the efficiency bound.

5 Simulation Evidence

5.1 The model, data generating mechanism, and estimators

In order to get a feel for finite sample properties of a feasible version of the proposed estimator, we set up the following econometric model:

$$y_t = \alpha + \beta x_t + e_t, \tag{35}$$

where $(\alpha, \beta)'$ is the vector of parameters to be estimated, whose numerical value is set to $(0, 0)'$. The data are generated according to

$$\begin{aligned} e_t &= \varepsilon_{t+1} - \theta\varepsilon_t, \quad \varepsilon_t | \Im_t \sim \mathcal{N}\left(0, \sigma_t^2\right), \\ x_t &= E\left[x_t | \Im_t\right] + \eta_{xt}, \quad \eta_{xt} \sim IID\, \mathcal{N}(0,1), \\ z_t &= 1 + \varphi(z_{t-1} - 1) + \eta_{zt}, \quad \eta_{zt} \sim IID\, \mathcal{N}(0,1). \end{aligned}$$

Apart from the constant, the basic instrument is z_t. We set the parameter values as follows. The parameter of the disturbance is θ takes values ± 0.3, ± 0.6, ± 0.9. The value of φ is set to 0.3. The skedastic function is set to

$$\sigma_t^2 = (z_t + z_{t-1})^2, \tag{36}$$

and the conditional expectation of the right hand variable given the instrument history – to

$$E[x_t|\Im_t] = 1 + z_t + z_{t-1}. \tag{37}$$

We present simulation evidence on the behavior of IV estimators with the following instruments.

(1) The simple IV estimator $\hat{\beta}_{IV}$ with the exactly identifying instrument $(1\ z_t)'$. The use of this "naive" estimator may be attributed to a researcher who wants to avoid complications arising from overidentification.

(2) The two-stage least squares estimator $\hat{\beta}_{2SLS}$ with the overidentifying instrument $(1\ z_t\ z_{t-1})'$. This would probably be the most intensively used estimator in this context, in spite of the presence of conditional heteroskedasticity.

(3) The GMM estimator $\hat{\beta}_{GMM}$ with the overidentifying instrument $(1\ z_t\ z_{t-1})'$. This estimator in contrast to $\hat{\beta}_{2SLS}$ optimally exploits the information in the three instruments in the presence of conditional heteroskedasticity.

(4) The estimator $\hat{\beta}_{\zeta^H}$ that would be a feasible optimal IV estimator in the absence of heteroskedasticity. This estimator has been previously known and used in the literature.

(5) The approximately optimal IV estimator $\hat{\beta}_{\zeta^{(101)}}$, corresponding to the feasible instrument $\hat{\boldsymbol{\zeta}}_t^{(101)}$.

Other possible competing estimators could be standard GMM estimators using instruments containing more lags of z_t. There is sufficient evidence, however, that these suffer a number of small sample deficiencies, mainly due to the need to estimate the efficient weighting matrix, thus we do not consider such estimators here. For their detailed consideration in conditionally heteroskedastic environments, see Tauchen (1986) and West, Wong and Anatolyev (2002).

The feasible instruments $\hat{\boldsymbol{\zeta}}_t^H$ and $\hat{\boldsymbol{\zeta}}_t^{(101)}$ are formed via

$$\hat{\boldsymbol{\zeta}}_t^H = \hat{\boldsymbol{\zeta}}_{t-1}^H \hat{\theta} + \frac{\hat{\boldsymbol{\delta}}_t^{(0)}}{\hat{\sigma}^2} \tag{38}$$

and

$$\hat{\boldsymbol{\zeta}}_t^{(101)} = \hat{\boldsymbol{\zeta}}_{t-1}^{(101)} (\hat{\phi}_t^{(1)})^- + \hat{\rho}_t^{(1)} \hat{\boldsymbol{\delta}}_t^{(0)}, \tag{39}$$

starting from $\hat{\boldsymbol{\zeta}}_0^H = \mathbf{0}$ and $\hat{\boldsymbol{\zeta}}_0^{(101)} = \mathbf{0}$. The 2SLS residuals are used to compute the estimates $\hat{\sigma}^2$ and $\hat{\theta}$ of the implied σ^2 and θ, and to evaluate conditional expectations of future ω_t and γ_t. The trimming parameter ϵ is set at 10^{-2}.

5.2 Details on nonparametric estimation

As mentioned in Section 3, we use the nonparametric corrected asymptotic final prediction error (CAFPE) criterion of Tschernig and Yang (2000) to select significant lags and an optimal value of the bandwidth. The criterion is

$$CAFPE = \left(\hat{A} + 2K(0)^m (T - i_m + 1)^{-1} h_{opt}^{-m} \hat{B}\right) \left(1 + m(T - i_m + 1)^{-4/(m+4)}\right), \tag{40}$$

where $K(u)$ is a kernel function, m is a number of employed lags of the basic instrument, and i_m is its maximal employed lag. Here, $\hat{A}$ and $\hat{B}$ are the following nonparametric estimates of ingredients of the asymptotic final prediction error:

$$\begin{aligned} \hat{A} &= (T - i_m + 1)^{-1} \sum_{t=i_m}^{T} \left(y_t - \hat{m}_y\left(z_{t-i_1}, \cdots, z_{t-i_m}\right)\right)^2 w_t, \\ \hat{B} &= (T - i_m + 1)^{-1} \sum_{t=i_m}^{T} \frac{\left(y_t - \hat{m}_y\left(z_{t-i_1}, \cdots, z_{t-i_m}\right)\right)^2}{\hat{\mu}\left(z_{t-i_1}, \cdots, z_{t-i_m}\right)} w_t, \end{aligned}$$

where $\hat{m}_y\left(z_{t-i_1}, \cdots, z_{t-i_m}\right)$ is a NW estimate of the regression function of y_t on the basic instrument and its lags, $\hat{\mu}\left(z_{t-i_1}, \cdots, z_{t-i_m}\right)$ is a NW estimate of the joint density of the vector of included lags, and w_t equals zero if the associated joint density estimate $\hat{\mu}\left(z_{t-i_1}, \cdots, z_{t-i_m}\right)$ is among the lowest 5% over the values of $\left(z_{t-i_1}, \cdots, z_{t-i_m}\right)$ in the sample, and unity otherwise. Such screening off extreme observations is conventional in nonparametric estimation literature (see also Tjostheim and Auestad 1994). The optimal bandwidth h_{opt} is determined via a grid search procedure. The first term in (40) is a nonparametric estimate of the asymptotic final prediction error, the second term in (40) is a correction aimed at penalizing lag overfitting (i.e. choosing superfluous lags in addition to correct ones). When no lags are used, we set the CAFPE equal to 0.95 times the sample variance of the first 95% of observations on y_t. Here, only 95% are taken in order to compensate for screening off in computing $\hat{A}$ and $\hat{B}$, and the factor 0.95 is used to help reduce lag overfitting and speed up the process.[4]

It is straightforward to verify that the DGP implies that $E_t\left[d_{t+i}\right], E_t\left[\omega_{t+i}\right]$ and $E_t\left[\gamma_{t+i}\right]$ for all $i \geq 0$ are functions of at most z_t and z_{t-1}, so that the basic instrument together with its first lag are sufficient regressors. In nonparametric estimation, we set the possible regressors to be z_t, z_{t-1} and z_{t-2} (i.e. i_m may be no larger than 2 and m may be no larger than 3), thus allowing for lag overfitting in estimating $\phi_t^{(1)}$, $\boldsymbol{\delta}_t^{(0)}$ and $\rho_t^{(1)}$. We use the product kernel with Epanechnikov marginals (thus $K(0) = 0.75$) having the same bandwidth value. We calculate the CAFPE for all combinations of inclusions of lags up to $m = 3$ and for all bandwidth values on a grid from $2h_0$ to $5h_0$ with a step of $\frac{1}{5}h_0$, where $h_0 = \widehat{Var}\left(z_t\right)^{1/2}\left(4/(m+2)\right)^{1/(m+4)}(T - i_m + 1)^{-1/(m+4)}$ (the lower bound is set so because the optimal bandwidth never turns out smaller than $2h_0$). From all estimates, we select the bandwidth that yields a minimum value to the CAFPE.

[4]Without this correction factor, the procedure for nonparametric estimation of infinite sums leads to spurious estimates of terms with negligeable dependence on the instruments and in addition to slower simulations.

5.3 Results

Table 3 reports standard deviations of the five simulated estimators for several combinations of sample sizes and values of the parameter θ. Additional experiments showed that the closer is $|\theta|$ to unity, the greater is the probability of obtaining an outlier in the distribution of approximately optimal IV estimates, such outliers marring the standard deviations. In the table we report the results for such combinations of T and θ that T increases with $|\theta|$ so that the outliers do not tend to appear; the interquantile ranges in the cases when the outliers do appear exhibit a similar pattern. Due to significant computational costs, we set the number of repetitions to 300 when $T = 500$, to 200 when $T = 1,000$, and to 100 when $T = 2,000$.

All estimators are centered around zero, the true parameter value, judging by their sample means and medians (not shown). It is clear that the asymptotic efficiency gains provided by the proposed instrument $\hat{\boldsymbol{\zeta}}_t^{(101)}$ and evaluated earlier are realized in finite samples as well. Among the traditional IV estimators ($\hat{\beta}_{IV}$, $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{GMM}$), the GMM estimator exhibits highest variance, despite asymptotically it is more efficient than the 2SLS. The efficiency gains provided by the instrument $\hat{\boldsymbol{\zeta}}_t^{H}$ relative to the 2SLS instrument are small, in most part due to imprecise non-parametric estimation. The approximately optimal instrument $\hat{\boldsymbol{\zeta}}_t^{(101)}$ relative to the traditional instruments and $\hat{\boldsymbol{\zeta}}_t^{H}$ yields sufficiently big efficiency gains, however.

It may be interesting to know what values various auxiliary parameters take during the nonparametric estimation described in the previous subsection. When conditional scores are estimated, on average 2.2 summands participate in the summation in (33), and the CAFPE criterion selects on average $2.8h_0$ as an optimal bandwidth. When conditional variances and covariances are estimated, on average $1.0 \div 2.1$ summands participate in the summation in (34), with higher value being associated with higher $|\theta|$, and the CAFPE criterion selects on average $4.0h_0$ as an optimal bandwidth.

6 Application to ultra-high frequency returns

Consider tick-by-tick transaction data from a stock or foreign exchange market. Let the white noise ε_t represent the return from the trade at t, and d_t – the duration between this and the previous trade. By Meddahi, Renault and Werker (2003), an exact discretization of a continuous time stochastic volatility process with linear mean reversion implies the conditional moment restriction

$$E\left[\left(r_t^2-\tau\right)-\left(r_{t-1}^2-\tau\right)\exp\left(-\kappa d_{t-1}\right)\frac{c\left(\kappa d_t\right)}{c\left(\kappa d_{t-1}\right)}|r_{t-2},d_{t-2},r_{t-3},d_{t-3},\cdots\right]=0,$$

where $r_t = \varepsilon_t/\sqrt{d_t}$ is a rescaled return, and $c(v) \equiv (1-\exp(-v))/v$. The parameter vector is $(\tau,\kappa)'$, where $\tau > 0$ indexes unconditional volatility, and $\kappa > 0$ indexes mean reversion. The conditional score vector is

$$\begin{pmatrix} f_{\tau,t}(\tau,\kappa) \\ f_{\kappa,t}(\tau,\kappa) \end{pmatrix} = \begin{pmatrix} \exp\left(-\kappa d_{t-1}\right)c\left(\kappa d_{t-1}\right)^{-1}c\left(\kappa d_t\right)-1 \\ \left(r_{t-1}^2-\tau\right)\exp\left(-\kappa d_{t-1}\right)c\left(\kappa d_{t-1}\right)^{-2}\times \\ \times\left(c\left(\kappa d_t\right)\left(c\left(\kappa d_{t-1}\right)+c'\left(\kappa d_{t-1}\right)\right)d_{t-1}-c\left(\kappa d_{t-1}\right)c'\left(\kappa d_t\right)d_t\right) \end{pmatrix}.$$

For conventional IV implementation, we choose the instrument vector

$$z_t = \left(1, r_{t-2}, \cdots, r_{t-1-\ell}, r_{t-2}^2, \cdots, r_{t-1-\ell}^2, \log d_{t-2}, \cdots, \log d_{t-1-\ell}, \log^2 d_{t-2}, \cdots, \log^2 d_{t-1-\ell}\right)'$$

for $\ell = 1, 2, 3$ (we use log durations here and later because their distribution is much less skewed than that of raw durations). We use the identity weight matrix at the initial stage, then iterate seven times the GMM procedure using the estimated efficient weight matrix.

Let $\hat{\tau}$, $\hat{\kappa}$ and $\hat{e}_t$ denote estimates of τ, κ and e_t obtained from conventional IV estimation with $\ell = 2$. From the residuals we compute the estimates $\hat{\theta}$ and $\hat{\sigma}^2$ of the implied θ and σ^2. To implement the approximately optimal IV estimation, we estimate nonparametric regressions of $f_{\tau,t}(\hat{\tau}, \hat{\kappa})$, $f_{\kappa}(\hat{\tau}, \hat{\kappa})$, $\hat{e}_t^2$ and $\hat{e}_t\hat{e}_{t-1}$ on current and past r_{t-2} and $\log d_{t-2}$ standardized by corresponding standard errors. Because many lags turn out to be significant, instead of full search through all possible lag combinations we use a directed search algorithm of Tjostheim and Auestad (1994): a new significant lag is added if it improves upon the CAFPE and if it does it better than other candidates. In every loop, we make the bandwidths for standardized r_t and $\log d_t$ run on a two dimensional grid $[2h_r, 5h_r] \times [2h_{\log d}, 5h_{\log d}]$ with steps of $\frac{1}{3}h_r$ and $\frac{1}{3}h_{\log d}$, where $h_\varsigma = (4/(m+2))^{1/(m+4)}(T - i_m + 1)^{-1/(m+4)}$, where ς is r or $\log d$, to determine optimal bandwidth values $h_{opt,r}$ and $h_{opt,\log d}$. In the standard formula (40) we change h_{opt}^{-m} to $h_{opt,r}^{-m_r} h_{opt,\log d}^{-m_{\log d}}$, where m_ς is a number of employed lags of the basic instrument ς, where ς is r or $\log d$, and $m_r + m_{\log d} = m$. After some experimentation, we set the maximum m_r and $m_{\log d}$ at 1 and 3, respectively, and employ the product kernel with normal marginals (thus $K(0) = 1/\sqrt{2\pi}$).

We use the data on trades in frequently traded common stocks at the Moscow Interbank Currency Exchange (MICEx) of the two oil extraction companies, Yukos and Lukoil.[5] The observation period is the week August 12–16, 2002, i.e. 5 trading days, during which the stocks were being traded at 400–700 transactions a day. Prior to the analysis, we remove intraday and interday deterministic patterns in durations and return volatilities, see Anatolyev and Shakin (2003) for details. To obtain r_t, we take the variable of adjusted raw returns, pass them through an $ARMA(1,1)$ filter (cf. Engle 2000), trim those that exceed 5 in absolute value to get rid of outliers, and demean.

The computations are made in GAUSS 4.0.23. Numerical optimization is performed using the constrained optimization package `co.ext` from the library `co.lib`, with the `bfgs` and `newton` algorithms and `stepbt` line search method. When implementing various estimation techniques, we construct asymptotic variance estimators and GMM efficient weight matrices in the Hansen–Hodrick form, as with the available sample sizes there is little chance to end up with a non-positive definite matrix.

The results are presented in Table 4, with standard errors in parentheses. For Yukos, the serial correlation is weak, and so is the conditional heteroskedasticity (in the conditional variance, no significant lags are selected; in the conditional autocovariance, only r_{t-2} is significant). As a consequence, the point estimates of κ using approximately optimal IV estimators are close to each other and to some of conventional GMM estimates. In contrast, for Lukoil, the serial correlation and conditional heteroskedasticity are much stronger, and as a consequence, the point estimates of κ using approximately optimal IV estimators differ more from each other, and appreciably from conventional GMM estimates. The standard

[5] A detailed description of the companies, their stocks and trades at the MICEx is available in English from `www.micex.com`.

errors for approximately optimal IV estimators rank as expected, with a much higher difference for Lukoil than for Yukos, while those for conventional GMM estimators are implausibly low. Moreover, in unreported experiments with other instrument sets the asymptotic confidence intervals for κ using different instrument sets sometimes do not intersect. This may be due to severe biases of GMM estimates and standard errors due to nonlinearity of the moment function and high dimensionality of the GMM efficient weight matrix in case of many instruments (see, for example, Newey and Smith 2001).

7 Concluding remarks

For general two-period conditional moment restrictions characterized by the presence of conditional heteroskedasticity, we have shown how to construct approximately optimal instruments, the ones that approximately satisfy the system of optimality conditions, evaluated the asymptotic properties of corresponding instrumental variables estimators, and verified their finite sample behavior. We have concentrated on the first order serial correlation because such problems are met most frequently among potential applications. For example, among these are CAPM models with habit formation or durability (Ferson and Constantinides 1991), overlapping data from forecasting surveys (Rich, Raymond and Butler 1992), data contaminated by temporal aggregation (Hall 1988), or ultra-high frequency returns and durations (Meddahi, Renault and Werker 2003).

The applications with conditional moment restrictions that have higher than the first order of serial correlation are less frequent, and in addition it is likely that a researcher will want to exploit the idea of optimal instruments. However, in such cases the approximation technique is absolutely the same as detailed above, with a tendency to become increasingly more complicated as the serial correlation order grows. The reason of this increasing complicatedness lies in a more unwieldy structure of the system defining the process that the optimal instrument follows (Anatolyev 2003). Let p denote the order of serial correlation, then there are $p+1$ processes indexing the conditional heteroskedasticity: $E_t[\mathbf{e}_t\mathbf{e}_t']$, $E_t[\mathbf{e}_{t-1}\mathbf{e}_t']$, $\cdots$, $E_t[\mathbf{e}_{t-p}\mathbf{e}_t']$, and the optimal instrument Ξ_t has the following recursion structure:

$$\Xi_t = \Xi_{t-1}\Phi_{1,t} + \Xi_{t-2}\Phi_{2,t} + \cdots + \Xi_{t-p}\Phi_{p,t} + \Delta_t \mathrm{P}_t,$$

where $\Phi_{1,t}$, $\Phi_{2,t}$, $\cdots$, $\Phi_{p,t}$, Δ_t, P_t are auxiliary processes. The analog of the equation (14) is a polynomial of order $p+2$ with respect to the conditional heteroskedasticity parameters and the p processes $\Phi_{1,t}$, $\Phi_{2,t}$, $\cdots$, $\Phi_{p,t}$ sought for. The analogs of (16) and (15) are more involved as well.

There are of course limitations of the approach presented in this paper. First, in chase of efficiency a researcher faces the need to use nonparametric estimation of various ingredients of the approximately optimal instrument. Second, although the impact of the approximation error was found to be small in the considered situations, it may potentially be larger or smaller, and it is impossible to unambiguously rank the approximately optimal instrument among other instruments in terms of asymptotic efficiency. Third, an econometrician may be dissatisfied by approximations being willing to achieve the efficiency bound. Alterntative approaches that do not invoke approximations are currently under investigation.

References

Anatolyev, S. (2003) The Form of the Optimal Nonlinear Instrument for Multiperiod Conditional Moment Restrictions. *Econometric Theory* 19, 602–609.

Anatolyev, S. and D. Shakin (2003) High Frequency Transactions at an Emerging Stock Market: New Data and New Models. Manuscript, New Economic School.

Attanasio O. and H. Low (2002) Estimating Euler equations. Working paper #02/06, Institute for Fiscal Studies.

Bougerol, P. and N. Picard (1992) Strict Stationarity of Generalized Autoregressive Processes. *Annals of Probability* 20, 1714–1730.

Brandt, A. (1986) The Stochastic Equation $Y_{n+1} = A_n Y_n + B_n$ with Stationary Coefficients. *Advances in Applied Probability* 18, 211–220.

Carroll, C. (2001) Death to the Log-Linearized Euler Equation! (And Very Poor Health to the Second-Order Approximation). *Advances in Macroeconomics* 1, no. 1, article 6.

Chamberlain, G. (1987) Asymptotic Efficiency in Estimation with Conditional Moment Restrictions. *Journal of Econometrics* 34, 305–334.

Eichenbaum, M.S., L.P. Hansen and K.J. Singleton (1988) A Time Series Analysis of Representative Agent Models of Consumption and Leisure Choice Under Uncertainty. *Quarterly Journal of Economics* 103, 51–78.

Engle, R.F. (2000) The econometrics of ultra-high frequency data. *Econometrica* 68, 1–22.

Ferson, W.E. and G.M. Constantinides (1991) Habit persistence and durability in aggregate consumption. *Journal of Financial Economics* 29, 199–240.

Green, P.E. and J.D. Carroll (1976) *Mathematical Tools for Applied Multivariate Analysis.* Academic Press: London.

Grossman, S., A. Melino and R.J. Shiller (1987) Estimating the continuous time consumption-based asset pricing model. *Journal of Business and Economic Statistics* 5, 315–327.

Hall, R.E. (1988) Intertemporal Substitution in Consumption. *Journal of Political Economy* 96, 339–357.

Hansen, B.E. and K.D. West (2002) Generalized Method of Moments and Macroeconomics. *Journal of Business and Economic Statistics* 20, 460–469.

Hansen, L.P. (1982) Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* 50, 1029–1054.

Hansen, L.P. (1985) A Method for Calculating Bounds on the Asymptotic Variance-Covariance Matrices of Generalized Method of Moments Estimators. *Journal of Econometrics* 30, 203–228.

Hansen, L.P., J.C. Heaton and M. Ogaki (1988) Efficiency Bounds Implied by Multiperiod Conditional Moment Restrictions. *Journal of the American Statistical Association* 83, 863–871.

Hansen, L.P and R.J. Hodrick (1980) Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis. *The Journal of Political Economy* 88, 829–853.

Hansen, L.P. and K.J. Singleton (1982) Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–1286.

Hansen, L.P. and K.J. Singleton (1996) Efficient Estimation of Linear Asset Pricing Models with Moving-Average Errors. *Journal of Business and Economic Statistics* 14, 53–68.

Heaton, J.C. and M. Ogaki (1991) Efficiency Bounds Calculations for a Time Series Model, with Conditional Heteroskedasticity. *Economics Letters* 35, 167–171.

Kuersteiner, G.M. (2001) Optimal instrumental variables estimation for ARMA models. *Journal of Econometrics* 104, 359–405.

Kuersteiner, G.M. (2002) Efficient IV estimation for autoregressive models with conditional heteroskedasticity. *Econometric Theory* 18, 547–583.

Ludvigson, S. and C. Paxson (2001) Approximation Bias in Linearized Euler Equations. *Review of Economics and Statistics* 83, 242–256.

Meddahi, N. and E. Renault (2002) Temporal Aggregation of Volatility Models. *Journal of Econometrics*, forthcoming.

Meddahi, N., E. Renault and B. Werker (2003) GARCH and Irregularly Spaced Data. Working Paper, CIREQ and Université de Montréal.

Mishkin, F. (1990) What Does the Term Structure Tell Us About Future Inflation? *Journal of Monetary Economics* 25, 77–95.

Newey, W.K. (1990) Efficient Instrumental Variables Estimation of Nonlinear Models. *Econometrica* 58, 809–837.

Newey, W.K. and R.J. Smith (2001) Higher order properties of GMM and generalized empirical likelihood estimators. Working Paper, MIT and University of Bristol, forthcoming in *Econometrica*.

Pötscher B. M. and I.R. Prucha (1997) *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. Springer-Verlag: Berlin.

Rich, R.W., J.E. Raymond and J.S. Butler (1992) The relationship between forecast dispersion and forecast uncertainty: evidence from a Survey Data – ARCH model. *Journal of Applied Econometrics* 7, 131–148.

Robinson, P. (1987) Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form. *Econometrica* 55, 875–891.

Tauchen, G. (1986) Statistical Properties of Generalized Method-of-Moments Estimators of Structural Parameters Obtained from Financial Market Data. *Journal of Business and Economic Statistics* 4, 397–416.

Tjostheim, D. and B. Auestad (1994) Nonparametric Identification of Nonlinear Time Series: Selecting Significant Lags. *Journal of American Statistical Association* 89, 1410–1419.

Tschernig, R. and L. Yang (2000) Nonparametric Lag Selection for Time Series. *Journal of Time Series Analysis* 21, 457–487.

Uhlig, H. (1995). A Toolkit for Analyzing Nonlinear Dynamic Stochastic Models Easily. CentER for Economic Research Working Paper, Tilburg University.

West, K.D. and D.W. Wilcox (1996) A Comparison of Alternative Instrumental Variables Estimators of a Dynamic Linear Model. *Journal of Business and Economic Statistics* 14, 281–293.

West, K.D., K.-f. Wong and S. Anatolyev (2002) Instrumental Variables Estimation of Heteroskedastic Linear Models Using All Lags of Instruments. Working paper, University of Wisconsin-Madison.

Working, H. (1960) Note on the Correlation of First Differences of Averages in a Random Chain. *Econometrica* 28, 916–918.

Table 1. Asymptotic variances of various IV estimators in the example with quadratic heteroskedasticity.

θ	z_t	ζ_t^H	$\zeta_t^{(101)}$	$\zeta_t^{(201)}$	$\zeta_t^{(111)}$	$\zeta_t^{(102)}$
-0.9	3.503	3.047	2.318	–	2.309	2.307
-0.5	2.063	1.922	1.632	1.795	1.635	1.632
-0.1	1.103	1.097	1.029	1.032	1.029	1.029
$+0.1$	0.803	0.797	0.769	0.772	0.769	0.769
$+0.5$	0.563	0.422	0.392	0.540	0.393	0.392
$+0.9$	0.803	0.347	0.271	–	0.271	0.270

Notes: The model and DGP are $y_t = \beta z_t + e_t$, where $z_t = 0.5z_{t-1} + \eta_t$, $e_t = \varepsilon_{t+1} - \theta\varepsilon_t$, $\varepsilon_t = \nu_t\sqrt{0.5 + 0.375z_t^2}$, $(\eta_t, \nu_t)' \sim IID\,\mathcal{N}(0, \mathrm{I}_2)$. The numbers in the table are asymptotic variances of the IV estimators of β that make use of the following instruments: the basic instrument z_t; the instrument ζ_t^H optimal in the absence of heteroskedasticity; the leading version of the proposed instrument $\zeta_t^{(101)}$ which uses approximations $\phi_t^{(1)}$, $\delta_t^{(0)}$ and $\rho_t^{(1)}$; and three other versions $\zeta_t^{(201)}$, $\zeta_t^{(111)}$, and $\zeta_t^{(102)}$, where $\phi_t^{(2)}$ is used in place of $\phi_t^{(1)}$, or $\delta_t^{(1)}$ in place of $\delta_t^{(0)}$, or $\rho_t^{(2)}$ in place of $\rho_t^{(1)}$, correspondingly.

Table 2. Asymptotic variances of various IV estimators in the Heaton–Ogaki example.

β	-0.8	-0.3	0	$+0.3$	$+0.8$
Truly optimal	0.360	0.910	1.000	0.910	0.360
Approximately optimal	0.363	0.917	1.012	0.923	0.371
Homoskedasticity optimal	0.399	1.070	1.313	1.235	0.430
First equation optimal	0.466	3.293	13.93	749.4	0.786

Notes: The DGP and model are $E_t[(u_{t+2}, u_{t+2}(0.25u_{t+2}-u_{t+1}))'] = \mathbf{0}$, where $u_t = z_t+\beta z_{t-1} = w_t + \left(2-\sqrt{3}\right) w_{t-1}$, and $w_t \sim IID\ \mathcal{N}(0,1)$. The numbers in the table are asymptotic variances of the IV estimators of β that make use of the following instruments: the "truly optimal" instrument that is optimal among nonlinear functions of the history of z_t; the proposed "approximately optimal" instrument which uses approximations $\Phi_t^{(1)}$, $\Delta_t^{(0)}$ and $\mathrm{P}_t^{(1)}$; the "homoskedasticity optimal" instrument that would be optimal if there were no conditional heteroskedasticity; and the "first equation optimal" instrument that is optimal among nonlinear functions of the history of z_t ignoring the second restriction.

Table 3. Standard deviations of various IV estimators, from simulations.

θ	T	$\widehat{\beta}_{IV}$	$\widehat{\beta}_{2SLS}$	$\widehat{\beta}_{GMM}$	$\widehat{\beta}_{\varsigma^H}$	$\widehat{\beta}_{\varsigma^{(101)}}$
-0.3	500	0.125	0.100	0.247	0.099	0.073
$+0.3$	500	0.099	0.065	0.176	0.063	0.048
-0.6	1000	0.102	0.087	0.199	0.081	0.075
$+0.6$	1000	0.060	0.038	0.115	0.033	0.028
-0.9	2000	0.089	0.079	0.187	0.067	0.059
$+0.9$	2000	0.047	0.025	0.077	0.024	0.022

Notes: The DGP is $z_t = 1 + \varphi(z_{t-1} - 1) + \eta_{zt}$, $x_t = 1 + z_t + z_{t-1} + \eta_{xt}$, where $(\eta_{zt}, \eta_{xt})' \sim IID\ \mathcal{N}(0, \mathrm{I}_2)$, and $y_t = \varepsilon_{t+1} - \theta\varepsilon_t$, $\varepsilon_t | z_t, z_{t-1}, \cdots \sim \mathcal{N}(0, (z_t + z_{t-1})^2)$. The table contains standard deviations obtained by simulations of the following IV estimators of β with the true value 0 in the model $y_t = \alpha + \beta x_t + e_t$: the exactly identifying IV estimator $\hat{\beta}_{IV}$ using $(1\ z_t)'$ as a vector of instruments; the two-stage least squares and GMM estimators $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{GMM}$ using $(1\ z_t\ z_{t-1})'$ as a vector of instruments; the feasible estimator $\hat{\beta}_{\varsigma^H}$ that would be an optimal IV estimator in the absence of heteroskedasticity; and the feasible approximately optimal IV estimator $\hat{\beta}_{\varsigma^{(101)}}$. The number of repetitions is 300 when $T = 500$, 200 when $T = 1000$, and 100 when $T = 2000$.

Table 4. The results of application of various IV estimators to the model of ultra-high frequency returns.

Stock	Yukos		Lukoil	
Method	τ	κ	τ	κ
GMM, $\ell = 1$	1.874 (0.179)	6.167 (1.423)	1.835 (0.124)	3.072 (0.158)
GMM, $\ell = 2$	1.913 (0.125)	9.700 (2.048)	1.835 (0.110)	3.134 (0.143)
GMM, $\ell = 3$	1.880 (0.122)	10.407 (2.269)	1.765 (0.109)	2.694 (0.088)
Optimal under homoskedasticity	1.974 (0.122)	9.217 (2.375)	2.010 (0.077)	7.433 (0.912)
Approximately optimal	1.964 (0.122)	9.306 (2.312)	1.976 (0.077)	6.065 (0.430)
$\hat{\theta}$	0.027		0.232	
Number of observations	2308		5572	

Notes: The parameters τ and κ are estimated from the conditional moment restriction $E\left[(r_t^2 - \tau) - \left(r_{t-1}^2 - \tau\right)\exp\left(-\kappa d_{t-1}\right) c\left(\kappa d_t\right)/c\left(\kappa d_{t-1}\right) | r_{t-2}, d_{t-2}, r_{t-3}, d_{t-3}, \cdots\right] = 0$, where $r_t = \varepsilon_t/\sqrt{d_t}$, ε_t is a raw return from the transaction at t, d_t is a duration between this and the previous transactions, and $c(v) \equiv (1 - \exp(-v))/v$. The table contains estimates of τ and κ, with Hansen–Hodrick standard errors in brackets. The estimators are: the iterated GMM estimators using the vector of instruments $(1, r_{t-2}, \cdots, r_{t-1-\ell}, r_{t-2}^2, \cdots, r_{t-1-\ell}^2, \log d_{t-2}, \cdots, \log d_{t-1-\ell}, \log^2 d_{t-2}, \cdots, \log^2 d_{t-1-\ell})'$ for various values of ℓ; the exactly identifying GMM estimator that would be an optimal IV estimator in the absence of heteroskedasticity; and the exactly identifying GMM estimator using the feasible approximately optimal IV instrument $\hat{\boldsymbol{\zeta}}_t^{(101)}$. The line "$\hat{\theta}$" contains estimates of the moving average coefficient in the MA representation of the error.

A Appendix (not intended for publication)

A.1 Constants

The constants in various approximations of parameters of the optimal instrument in the example with quadratic heteroskedasticity are:

$$
\begin{aligned}
\kappa_{\omega 1} &= 1+\theta^2-\lambda\left(\varphi^2+\theta^2\right), \quad \kappa_{\omega 2}=\lambda\left(1-\varphi^2\right)\left(\varphi^2+\theta^2\right),\\
\kappa_{\gamma 1} &= -\theta\left(1-\lambda\right), \quad \kappa_{\gamma 2}=-\lambda\theta\left(1-\varphi^2\right),\\
\kappa_{\phi 1} &= \theta\left\{1-\frac{\lambda\left(1-\varphi^2\right)\left(1-\theta^2\right)}{1-\varphi^2\theta^2}\right\}, \quad \kappa_{\phi 2}=\frac{\lambda\theta\left(1-\varphi^2\right)^2\left(1-\theta^2\right)}{1-\varphi^2\theta^2},\\
\kappa_{\phi 3} &= \lambda\kappa_{\phi 1}+\frac{1}{1-\theta^2}\left\{\frac{\kappa_{\phi 1}^2}{\theta}+\frac{2\theta\kappa_{\phi 1}\kappa_{\phi 2}}{1-\varphi^2\theta^2}+\frac{3\theta\left(1+\varphi^2\theta^2\right)\kappa_{\phi 2}^2}{\left(1-\varphi^2\theta^2\right)\left(1-\varphi^4\theta^2\right)}\right\},\\
\kappa_{\phi 4} &= \lambda\kappa_{\phi 2}+\frac{1}{\theta}\left\{\kappa_{\phi 1}\kappa_{\gamma 2}+\frac{2\kappa_{\phi 1}\kappa_{\phi 2}}{1-\varphi^2\theta^2}+\frac{6\varphi^2\theta^2\kappa_{\phi 2}^2}{\left(1-\varphi^2\theta^2\right)\left(1-\varphi^4\theta^2\right)}\right\},\\
\kappa_{\phi 5} &= \frac{\kappa_{\phi 2}}{\theta}\left\{\kappa_{\gamma 2}+\frac{\kappa_{\phi 2}}{1-\varphi^4\theta^2}\right\},\\
\kappa_{\delta 1} &= \frac{1}{1-\varphi\theta}, \quad \kappa_{\delta 2}=\kappa_{\delta 1}+\varphi\kappa_{\delta 1}^2\left\{\kappa_{\phi 1}+\frac{3\kappa_{\phi 2}}{1-\varphi^3\theta}\right\}, \quad \kappa_{\delta 3}=\frac{\varphi^3\kappa_{\phi 2}\kappa_{\delta 1}}{1-\varphi^3\theta},\\
\kappa_{\delta 4} &= \kappa_{\delta 3}+\frac{\varphi^3}{1-\varphi^3\theta}\left\{\kappa_{\phi 1}\kappa_{\delta 2}+\kappa_{\phi 4}\kappa_{\delta 1}+\frac{10\left(\kappa_{\phi 2}\kappa_{\delta 2}+\kappa_{\delta 1}\kappa_{\phi 5}\right)}{1-\varphi^5\theta}\right\},\\
\kappa_{\delta 5} &= \frac{\varphi^5\left(\kappa_{\phi 2}\kappa_{\delta 2}+\kappa_{\delta 1}\kappa_{\phi 5}\right)}{1-\varphi^5\theta},\\
\kappa_{\rho 1} &= \kappa_{\omega 1}-\frac{\kappa_{\gamma 2}^2}{\kappa_{\omega 2}}\left\{1-\frac{\kappa_{\omega 1}}{\kappa_{\omega 2}}+\frac{2\kappa_{\gamma 1}}{\kappa_{\gamma 2}}\right\}, \quad \kappa_{\rho 2}=\kappa_{\omega 2}-\frac{\varphi^2\kappa_{\gamma 2}^2}{\kappa_{\omega 2}},\\
\kappa_{\rho 3}\left(x\right) &= -\left(\kappa_{\gamma 1}-\frac{\kappa_{\omega 1}\kappa_{\gamma 2}}{\kappa_{\omega 2}}\right)^2\sqrt{\frac{\pi}{2\kappa_{\omega 1}\kappa_{\omega 2}}}\cdot\Re\left(w\left(-\frac{\varphi}{\sqrt{2}}x+i\sqrt{\frac{\kappa_{\omega 1}}{2\kappa_{\omega 2}}}\right)\right),
\end{aligned}
$$

where $\Re\left(\cdot\right)$ is an operator of removing the imaginary part of a complex number, and $w\left(\cdot\right)$ is the error function: $w\left(x\right)\equiv e^{-x^2}\left(1+\frac{2i}{\sqrt{\pi}}\int\limits_0^x e^{-t^2}dt\right)=\sum\limits_{n=0}^{\infty}\frac{(ix)^n}{\Gamma\left(\frac{n}{2}+1\right)}$.

A.2 Linear IV bound for the example with quadratic heteroskedasticity

We find the efficiency bound for linear IV estimators by explicitly deriving the instrument optimal in the linear class (West, Wong and Anatolyev 2002). Let the optimal instrument be $z_t^*=\sum\limits_{i=0}^{\infty}g_i\eta_{t-i}$ and let $\tau\equiv E[\eta_t^4]-1$. The optimality condition is

$$
\forall k\geq 0 \qquad E\left[\eta_{t-k}z_t\right]=E\left[\eta_{t-k}z_t^*e_t^2\right]+E\left[\eta_{t-k-1}z_t^*e_te_{t-1}\right]+E\left[\eta_{t-k}z_{t-1}^*e_te_{t-1}\right]. \qquad (41)
$$

The left hand side in (41) is φ^k. Calculate the three terms on the right hand side using $E_t[e_t^2]=\kappa_{\omega 1}+\kappa_{\omega 2}z_t^2$, $E_t[e_te_{t-1}]=\kappa_{\gamma 1}+\kappa_{\gamma 2}z_t^2$, $z_t^*=\sum\limits_{i=0}^{\infty}g_i\eta_{t-i}$ and $z_t=\sum\limits_{i=0}^{\infty}\varphi^i\eta_{t-i}$:

$$E\left[\eta_{t-k}z_t^*e_t^2\right] = E\left[\eta_{t-k}\left(\sum_{i=0}^{\infty}g_i\eta_{t-i}\right)\left(\kappa_{\omega 1}+\kappa_{\omega 2}\left(\sum_{j=0}^{\infty}\varphi^j\eta_{t-j}\right)^2\right)\right]$$
$$= \left(\kappa_{\omega 1}+\kappa_{\omega 2}\left(\tfrac{1}{1-\varphi^2}+\varphi^{2k}\tau\right)\right)g_k + 2\varphi^k\kappa_{\omega 2}\sum_{i=0,i\neq k}^{\infty}\varphi^i g_i,$$

$$E\left[\eta_{t-k-1}z_t^*e_te_{t-1}\right] = E\left[\eta_{t-k-1}\left(\sum_{i=0}^{\infty}g_i\eta_{t-i}\right)\left(\kappa_{\gamma 1}+\kappa_{\gamma 2}\left(\sum_{j=0}^{\infty}\varphi^j\eta_{t-j}\right)^2\right)\right]$$
$$= \left(\kappa_{\gamma 1}+\kappa_{\gamma 2}\left(\tfrac{1}{1-\varphi^2}+\varphi^{2(k+1)}\tau\right)\right)g_{k+1} + 2\varphi^{k+1}\kappa_{\gamma 2}\sum_{i=0,i\neq k+1}^{\infty}\varphi^i g_i,$$

$$E\left[\eta_{t-k}z_{t-1}^*e_te_{t-1}\right] = E\left[\eta_{t-k}\left(\sum_{i=1}^{\infty}g_{i-1}\eta_{t-i}\right)\left(\kappa_{\gamma 1}+\kappa_{\gamma 2}\left(\sum_{j=0}^{\infty}\varphi^j\eta_{t-j}\right)^2\right)\right]$$
$$= \begin{cases} \left(\kappa_{\gamma 1}+\kappa_{\gamma 2}\left(\tfrac{1}{1-\varphi^2}+\varphi^{2k}\tau\right)\right)g_{k-1} + 2\varphi^{k+1}\kappa_{\gamma 2}\sum_{i=0,i\neq k-1}^{\infty}\varphi^i g_i, & k>0, \\ 2\varphi\kappa_{\gamma 2}\sum_{i=0}^{\infty}\varphi^i g_i, & k=0. \end{cases}$$

Therefore the system (41) can be written in a matrix form

$$\Phi = \mathrm{S}\mathrm{G}, \tag{42}$$

where $\Phi \equiv \begin{bmatrix} 1 \\ \varphi \\ \vdots \\ \varphi^k \\ \vdots \end{bmatrix}$, $\mathrm{G} \equiv \begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_k \\ \vdots \end{bmatrix}$, $\mathrm{S} \equiv \begin{bmatrix} S_{0,0} & S_{0,1} & \cdots & S_{0,k} & \cdots \\ S_{1,0} & S_{1,1} & \cdots & S_{1,k} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \\ S_{k,0} & S_{k,1} & \cdots & S_{k,k} & \cdots \\ \vdots & \vdots & & \vdots & \ddots \end{bmatrix}$, and

$$S_{k,k} = \kappa_{\omega 1}+\kappa_{\omega 2}\left(\tfrac{1}{1-\varphi^2}+\varphi^{2k}\tau\right)+4\varphi^{2k+1}\kappa_{\gamma 2},\ k\geq 0,$$
$$S_{k,k-1} = \kappa_{\gamma 1}+\kappa_{\gamma 2}\left(\tfrac{1}{1-\varphi^2}+\varphi^{2k}\tau\right)+2\varphi^{2k-1}\left(\kappa_{\omega 2}+\varphi\kappa_{\gamma 2}\right),\ k\geq 1,$$
$$S_{k,k+1} = \kappa_{\gamma 1}+\kappa_{\gamma 2}\left(\tfrac{1}{1-\varphi^2}+\varphi^{2(k+1)}\tau\right)+2\varphi^{2k+1}\left(\kappa_{\omega 2}+\varphi\kappa_{\gamma 2}\right),\ k\geq 0,$$
$$S_{k,j} = 2\varphi^{j+k}\left(\kappa_{\omega 2}+2\varphi\kappa_{\gamma 2}\right),\ k\geq 0,\ j<k-1 \text{ or } j>k+1.$$

The optimal instrument then is characterized by the vector of weights $\mathrm{G} = \mathrm{S}^{-1}\Phi$, and the efficiency bound is

$$\mathrm{V}_{z^*} = \left(\Phi'\mathrm{S}^{-1}\Phi\right)^{-1}.$$

Now we prove that instrument z_t^* is identical to ζ_t^H when $\eta_t|\Im_t \sim \mathcal{N}(0,1)$ implying $\tau = 2$. Note that the weighting vector G^H implied by ζ_t^H is a solution of the system

$$\Phi = \mathrm{S}^H\mathrm{G}^H, \tag{43}$$

where Ψ is as above, and S^H is a triagonal matrix with $1+\theta^2$ as diagonal entries and $-\theta$ as off diagonal entries. When $\tau = 2$, the matrix S can be decomposed as $\mathrm{S} = \mathrm{S}^H + \varrho\Phi\Phi'$, where $\varrho \equiv 2\left(\kappa_{\omega 2}+2\varphi\kappa_{\gamma 2}\right)$. Therefore, equation (42) may be written as $\Phi = \left(\mathrm{S}^H+\varrho\Phi\Phi'\right)\mathrm{G}$, or $\Phi\left(1-\varrho\Phi'\mathrm{G}\right) = \mathrm{S}^H\mathrm{G}$, i.e. the same as (43) up to the multiplicative scalar factor $1-\varrho\Phi'\mathrm{G}$. Therefore, the solution G differs from G^H only by the multiplicative factor. This implies that z_t^* and ζ_t^H are essentially the same instrument.