Contents

1. APPROXIMATELY OPTIMAL INSTRUMENT FOR MULTIPERIOD MOMENT CONDITIONS

1.1 Introduction

Many time series models appear in the form of conditional moment restrictions. They are usually estimated and tested by choosing instruments from the conditioning information set and applying GMM (Hansen 1982). The set of possible instruments is typically infinite, which raises a question of their optimal choice for the purpose of attaining higher efficiency of estimation. When the moment function is a martingale difference with respect to the conditioning information so that the moment restrictions are *single-period* and thus are serially uncorrelated, 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. 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). 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, 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. Anatolyev (in press) gives a more algorithmic description of the optimal instrument. He derives the form of the process followed by the optimal instrument which turns out to be a recursion that generalizes Hansen's formula (Hansen 1985, Lemma 5.7). The process followed by the optimal instrument is parameterized by three auxiliary infinite-dimensional parameters. Estimation of these would constitute the feasible procedure, if there were not the following major difficulty: the laws of motion that govern the dynamics of these parameters are not explicit, but instead solve a system of highly nonlinear functional equations. In rare circumstances, it is possible to solve this system analytically, as in Heaton and Ogaki's (1991) example, but it 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 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 Taylor expansion for the three equations that determine the auxiliary parameters. 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 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.

In constructing the instruments, we act as though the needed parametric forms of conditional expectations are known. Since this knowledge is unlikely to be implied by the model,[1] we conduct simulations on the feasible version of the proposed instrumental variables estimator and its popular competitors, both when the parametric forms of auxiliary conditional expectations are conjectured correctly, and when they are misspecified. It turns out that under both circumstances the feasible approximately optimal instrument has advantageous finite sample properties.

The paper is organized as follows. Section 2 elaborates the case of a single two-period conditional moment restriction (i.e. where the errors are first-order serially correlated). We review the form of the optimal instrument, show how the approximations are taken, and calibrate asymptotic properties in a pilot example. Section 3 presents a generalization to the multiple equation case including careful computations of asymptotic gains and losses in the Heaton–Ogaki example. Section 4 reports the results of simulation experiments. In section 5 we outline what changes when the serial correlation is of higher order than the first, and conclude. The Appendix contains tedious derivations and unwieldy details. We use Euclidean norm $|A| = \sqrt{\varrho\left(A'A\right)}$, where $\varrho\left(\cdot\right)$ is the spectral radius, for vectors and matrices. The $n \times n$ identity matrix is denoted by I_n, the $n \times 1$ vector of zeros – by $\mathbf{0}_n$.

1.2 Theory: single equation case

1.2.1 Optimal instrument

We consider the model

$$f\left(\boldsymbol{\beta}, \mathbf{x}_t\right) = e_t, \tag{1.2.1}$$

where e_t is the error, $\mathbf{x}_t$ is a vector of observable variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters to be estimated, and $f\left(\boldsymbol{\beta}, \mathbf{x}_t\right)$ is a known up to $\boldsymbol{\beta}$ function which is possibly nonlinear in $\boldsymbol{\beta}$. In addition, we are given vector $\mathbf{z}_t$ of observable *basic instruments* (as opposed to just *instruments* that may be generated from the basic ones). Some of $\mathbf{x}_t$'s, along with their lags or functions, may be among $\mathbf{z}_t$. We assume that $(\mathbf{x}_t, \mathbf{z}_t)$ is strictly stationary and ergodic, and each involved variable posseses finite fourth moment. 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 $E_t[\cdot] \equiv E[\cdot|\Im_t]$. The conditional moment restriction

$$E_t\left[e_t\right] = 0 \tag{1.2.2}$$

[1] If one does have precise knowledge on the forms of auxiliary conditional expectations not implied by the model, this knowledge should be exploited on the level of the model's formulation to expand the set of moment restrictions. The trade-off between efficiency and robustness to auxiliary parametrizations is an intrinsic feature of the optimal instrumental variables approach.

implies that all measurable functions of the basic instrument and their lags are valid instruments. Define the $k \times 1$ vector

$$\mathbf{d}_t \equiv E_t \left[\frac{\partial f(\boldsymbol{\beta}, \mathbf{x}_t)}{\partial \boldsymbol{\beta}} \right], \tag{1.2.3}$$

and let

$$\omega_t \equiv E_t \left[e_t^2 \right], \quad \gamma_t \equiv E_t \left[e_t e_{t-1} \right] \tag{1.2.4}$$

be the conditional variance and conditional first-order autocovariance of the errors. We assume the first-order conditional serial correlation structure of the error e_t, that is, $E_t[e_t e_{t-j}] = 0$ for $j > 1$.

Under suitable conditions, the optimal instrumental variables estimator takes the form (Anatolyev in press)

$$\boldsymbol{\zeta}_t = \boldsymbol{\zeta}_{t-1} \phi_t + \rho_t \boldsymbol{\delta}_t, \tag{1.2.5}$$

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

$$\gamma_t + \phi_t \left(\omega_t + E_t \left[\phi_{t+1} \gamma_{t+1} \right] \right) = 0, \tag{1.2.6}$$

$$E_t \left[1 - \rho_t (\omega_t - \rho_{t+1} \gamma_{t+1}^2) \right] = 0, \tag{1.2.7}$$

$$\boldsymbol{\delta}_t = \mathbf{d}_t + E_t \left[\phi_{t+1} \boldsymbol{\delta}_{t+1} \right], \tag{1.2.8}$$

$$E \left[\log |\phi_t| \right] < 0. \tag{1.2.9}$$

The key relation is (1.2.5). 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 the next subsection. Here, in contrast to Hansen (1985), ϕ_t and ρ_t are time varying, and $\boldsymbol{\delta}_t$ is a generalized projection of the discounted sum of future $\mathbf{d}_t$-variables onto the space of instruments. The conditions (1.2.6), (1.2.7) and (1.2.8) determine ϕ_t, ρ_t and $\boldsymbol{\delta}_t$, respectively, while (1.2.9) rules out unstable solutions of the nonlinear equation (1.2.6). The nature of the derivative parameters 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.*

1.2.2 Approximately optimal instrument

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

$$\boldsymbol{\zeta}_t^H = \boldsymbol{\zeta}_{t-1}^H \theta + \frac{1}{\sigma^2} E_t \left[\sum_{i=0}^{\infty} \theta^i \mathbf{d}_{t+i} \right], \tag{1.2.10}$$

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. $e_t = w_{t+1} - \theta w_t$, $\sigma^2 = E[w_t^2]$. In the heteroskedastic environment the instrument $\boldsymbol{\zeta}_t^H$ is no longer optimal since it ignores conditional heteroskedasticity.[2] Note that the construction of $\boldsymbol{\zeta}_t^H$ may be viewed as approximation of ϕ_t, $\boldsymbol{\delta}_t$, ρ_t correspondingly by

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

[2] An instrument of type (1.2.10) was used in empirical work, for example, by West and Wilcox (1996) in homoskedastic environment and by Hansen and Singleton (1996) in both homo- and heteroskedastic environments.

On the other hand, when the error is conditionally serially uncorrelated, the optimal instrument is that of Chamberlain (1987):

$$\boldsymbol{\zeta}_t^C = \frac{\mathbf{d}_t}{\omega_t}. \tag{1.2.11}$$

In the conditionally serially correlated environment instrument $\boldsymbol{\zeta}_t^C$ is no longer optimal since it ignores serial correlation. Note that construction of $\boldsymbol{\zeta}_t^C$ may be viewed as approximation of ϕ_t, $\boldsymbol{\delta}_t$, ρ_t correspondingly by

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

We want to find an approximate but explicit solution to (1.2.6)–(1.2.9) treating the instruments (1.2.10) and (1.2.11) as benchmarks. On the one hand, the dynamic structure of the optimal instrument (1.2.5), i.e. nonzeroness of ϕ_t and complicatedness of $\boldsymbol{\delta}_t$, may be attributed to the presence of serial correlation. On the other hand, nonconstant weighting of $\boldsymbol{\delta}_t$ by ρ_t is due to the presence of conditional heteroskedasticity. Therefore, on the one hand, acknowledging the presence of conditional heteroskedasticity, we find approximate deviations of ϕ_t and $\boldsymbol{\delta}_t$ from ϕ^H and $\boldsymbol{\delta}_t^H$ driven by deviations of the parameters ω_t and γ_t from their homoskedastic counterparts $\omega^H \equiv E[e_t^2]$ and $\gamma^H \equiv E[e_t e_{t-1}]$. On the other hand, acknowledging the presence of serial correlation, we find approximate deviations of ρ_t from ρ_t^C driven by deviations of the parameter γ_t from its no serial correlation counterpart $\gamma^C \equiv 0$.

Recall that the discount process ϕ_t is determined from (1.2.6), which we rewrite as

$$E_t\left[F\left(\phi_t, \phi_{t+1}, \omega_t, \gamma_t, \gamma_{t+1}\right)\right] = 0, \tag{1.2.12}$$

where $F(\phi_t, \phi_{t+1}, \omega_t, \gamma_t, \gamma_{t+1}) \equiv \gamma_t + \phi_t\left(\omega_t + \phi_{t+1}\gamma_{t+1}\right)$. We linearize stochastic equation (1.2.12) 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:

$$E_t\left[\begin{array}{c} \left.\dfrac{\partial F}{\partial \phi_t}\right|_H \Delta\phi_t + \left.\dfrac{\partial F}{\partial \phi_{t+1}}\right|_H \Delta\phi_{t+1} + \left.\dfrac{\partial F}{\partial \omega_t}\right|_H \Delta\omega_t + \left.\dfrac{\partial F}{\partial \gamma_t}\right|_H \Delta\gamma_t \\ + \left.\dfrac{\partial F}{\partial \gamma_{t+1}}\right|_H \Delta\gamma_{t+1} + R_{t+1}^F \end{array}\right] = 0, \tag{1.2.13}$$

where R_{t+1}^F contains higher-order terms. Collecting linear terms together in equation (1.2.13) 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

$$\phi_t^{(1)} = \theta + \frac{1}{\sigma^2}\left(\gamma_t - \sum_{i=0}^{\infty}\theta^{2i} E_t\left[\theta\omega_{t+i} + 2\gamma_{t+i}\right]\right). \tag{1.2.14}$$

Note that by construction $E[\phi_t^{(1)}] = \theta$, i.e. $\phi_t^{(1)}$ fluctuates around ϕ^H. 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:

$$E_t\left[\begin{array}{c} \left.\dfrac{\partial F}{\partial \phi_t}\right|_H \Delta\phi_t + \left.\dfrac{\partial F}{\partial \phi_{t+1}}\right|_H \Delta\phi_{t+1} + \left.\dfrac{\partial F}{\partial \omega_t}\right|_H \Delta\omega_t + \left.\dfrac{\partial F}{\partial \gamma_t}\right|_H \Delta\gamma_t \\ + \left.\dfrac{\partial F}{\partial \gamma_{t+1}}\right|_H \Delta\gamma_{t+1} \left.\dfrac{\partial^2 F}{\partial \phi_t \partial \phi_{t+1}}\right|_H \Delta\phi_t\Delta\phi_{t+1} + \left.\dfrac{\partial^2 F}{\partial \phi_t \partial \omega_t}\right|_H \Delta\phi_t\Delta\omega_t \\ + \left.\dfrac{\partial^2 F}{\partial \phi_t \partial \gamma_{t+1}}\right|_H \Delta\phi_t\Delta\gamma_{t+1} + \left.\dfrac{\partial^2 F}{\partial \phi_{t+1} \partial \gamma_{t+1}}\right|_H \Delta\phi_{t+1}\Delta\gamma_{t+1} + R_{t+1}^F \end{array}\right] = 0, \tag{1.2.15}$$

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 $\phi^{(2)}_t$ for ϕ_t, with a unique stationary solution

$$\phi^{(2)}_t = \phi^{(1)}_t + \frac{1}{\theta}\left(\frac{1}{\sigma^2}\left(\phi^{(1)}_t - \theta\right)\left(\gamma_t + \theta\sigma^2\right) + \sum_{i=0}^{\infty}\theta^{2i}E_t\left[\left(\phi^{(1)}_{t+i} - \theta\right)^2\right]\right). \tag{1.2.16}$$

This kind of expansion may be continued further, if desired.

Now we consider the forcing process $\boldsymbol{\delta}_t$ determined by (1.2.8). We approximate $F(\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t+1}, \phi_{t+1}) \equiv -\boldsymbol{\delta}_t + \mathbf{d}_t + \phi_{t+1}\boldsymbol{\delta}_{t+1}$ around $H = \left(\boldsymbol{\delta}^H_t, \boldsymbol{\delta}^H_{t+1}, \phi^H\right)$ to end up with a linear stochastic difference equation with respect to the first-order approximation $\boldsymbol{\delta}^{(1)}_t$ for $\boldsymbol{\delta}_t$, with a unique stationary solution

$$\boldsymbol{\delta}^{(1)}_t = \boldsymbol{\delta}^H_t + \sum_{i=1}^{\infty}\theta^{i-1}E_t\left[\left(\phi^{(1)}_{t+i} - \theta\right)\boldsymbol{\delta}^H_{t+i}\right]. \tag{1.2.17}$$

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

$$\boldsymbol{\delta}^{(2)}_t = \boldsymbol{\delta}^{(1)}_t + \sum_{i=1}^{\infty}\theta^{i-1}E_t\left[\left(\phi^{(2)}_{t+i} - \phi^{(1)}_{t+1}\right)\boldsymbol{\delta}^H_{t+i} + \phi^{(1)}_{t+i}\left(\boldsymbol{\delta}^{(1)}_{t+i} - \boldsymbol{\delta}^H_{t+i}\right)\right]. \tag{1.2.18}$$

This kind of expansion may be continued further, if desired.

Now we consider the weighting multiplier ρ_t determined by (1.2.7). We approximate $F(\rho_t, \rho_{t+1}, \gamma_{t+1}) \equiv 1 - \rho_t\left(\omega_t - \rho_{t+1}\gamma^2_{t+1}\right)$ around $\mathcal{C} = \left(\rho^C_t, \rho^C_{t+1}, 0\right)$ to find that

$$\rho^{(1)}_t = \frac{1}{\omega_t}, \tag{1.2.19}$$

i.e. first-order approximation $\rho^{(1)}_t$ for ρ_t coincides with ρ^C_t. The second-order approximation for ρ_t is

$$\rho^{(2)}_t = \frac{1}{\omega_t}\left(1 + \frac{1}{\omega_t}E_t\left[\frac{\gamma^2_{t+1}}{\omega_{t+1}}\right]\right). \tag{1.2.20}$$

This kind of expansion may be continued further, if desired.

The *approximately optimal instrument* $\boldsymbol{\zeta}^{(jkl)}_t$ 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, $\boldsymbol{\delta}^H_t$ and ρ^C_t, respectively. The approximately optimal instrument follows

$$\boldsymbol{\zeta}^{(jkl)}_t = \boldsymbol{\zeta}^{(jkl)}_{t-1}\phi^{(j)}_t + \rho^{(l)}_t\boldsymbol{\delta}^{(k)}_t. \tag{1.2.21}$$

Since we will use proxies for ϕ_t, $\boldsymbol{\delta}_t$ and ρ_t instead of the true processes to construct the instrument, we have to ensure proper behavior of $\boldsymbol{\zeta}^{(jkl)}_t$. This means two things. First, we require stationarity: there must exist a unique stationary solution to (1.2.21) with approximated ϕ_t, $\boldsymbol{\delta}_t$ and ρ_t. Second, we need to ensure finiteness of fourth moments of the constructed instrument. The former aim is easily attained: the conditions for existence of a unique stationary solution to $AR(1)$ structures like (1.2.21) are quite weak and verifiable (Brandt 1986). The latter aim is much more challenging, due to time dependence of the "AR coefficient" $\phi^{(j)}_t$, the absolute value of which is not necessarily uniformly bounded by 1. Therefore, we deliberately simplify the task at the expense of possible efficiency losses. The following general Lemma will be of great help now and in the following section.

Lemma 1 *Suppose that $k \times s$ matrix process Ψ_t satisfies the recurrence relation*

$$\Psi_t = \Psi_{t-1} A_t + B_t, \tag{1.2.22}$$

where A_t is $s \times s$ and B_t is $k \times s$ 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 stationary ergodic solution Ψ_t of (1.2.22) such that $E[|\Psi_t|^4] < \infty$. This solution can be represented as

$$\Psi_t = \sum_{i=0}^{\infty} B_{t-i} \prod_{j=0}^{i} A_{t-j} \tag{1.2.23}$$

with the right-hand side converging absolutely almost surely.

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_ϕ, like 10^{-2}, say. Define the trimming operator "$^-$" by

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

That is, "$^-$" trims large $|\phi|$ by setting $\phi > 1 - \epsilon_\phi$ to $1 - \epsilon_\phi$, $\phi < -1 + \epsilon_\phi$ to $-1 + \epsilon_\phi$. Then instead of (1.2.21) we can use the following recursion:

$$\boldsymbol{\zeta}_t^{(jkl)} = \boldsymbol{\zeta}_{t-1}^{(jkl)} (\phi_t^{(j)})^- + \rho_t^{(l)} \boldsymbol{\delta}_t^{(k)}. \tag{1.2.24}$$

starting from, say, $\boldsymbol{\zeta}_0^{(jkl)} = \mathbf{0}$. Then $\operatorname{esssup}(\phi_t^{(j)})^- < 1$ and $E[|\rho_t^{(l)} \delta_t^{(k)}|^4] \le E[|\delta_t^{(k)}|^4] \operatorname{esssup} |\rho_t^{(l)}| < \infty$ if $\delta_t^{(k)}$ has finite fourth moments and $\rho_t^{(l)}$ is bounded below, so the prerequisites of Lemma 1 are satisfied. Of course, the efficiency of the instrument $\boldsymbol{\zeta}_t^{(jkl)}$ depends on the trimming parameter ϵ_ϕ. It is wise to set it to a small number to distort $\phi_t^{(j)}$ the least.

1.2.3 Asymptotic Comparisons

We will use the following data generating mechanism for the demonstration of the technique and calibration of asymptotic gains.

$$\begin{aligned} y_t &= \beta z_t + e_t, \quad e_t = w_{t+1} - \theta w_t, \quad \Im_t = \sigma(z_t, z_{t-1}, \ldots); \\ w_t &= \nu_t \sqrt{1 - \lambda + \lambda (1 - \varphi^2) z_t^2}, \quad \nu_t \sim IID\, \mathcal{N}(0,1); \\ z_t &= \varphi z_{t-1} + \eta_t, \quad \eta_t | \Im_t \sim \mathcal{N}(0,1). \end{aligned}$$

Here $\theta \in (-1,1)$, $\varphi \in (0,1)$, and $\lambda \in [0,1)$. The basic instrument is z_t. The object of estimation is β. The parameters are (all constants $\kappa_{\cdot\cdot}$ may be found in the Appendix):

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

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 (column ”OLS” in the table below) and the West–Wong–Anatolyev instrument z_t^* (West, Wong and Anatolyev 2002) (column ”GMM”, see the Appendix for details of computations). The latter instrument 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. The next column belongs to the approximately optimal instrument $\zeta_t^{(101)}$, the one that uses approximations $\phi_t^{(1)}$, $\delta_t^{(0)}$ and $\rho_t^{(1)}$. This version turns out to be a reasonable compromise between an instrument's complexity and efficiency gains in this pilot example. The next three columns belong to approximately optimal instruments 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 first two estimators the asymptotic variances are computed using closed-form formulae, for the approximately optimal IV estimators – by simulations, with sample sizes of at least 10^8 observations.

θ	OLS	GMM	$\zeta_t^{(101)}$	$\zeta_t^{(201)}$	$\zeta_t^{(111)}$	$\zeta_t^{(102)}$
-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

The table 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 be ±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 since when a more complex approximations tends to yield slight efficiency gains, these gains are not likely to be realized when a feasible estimator is developed.

A quick look at the table reveals significant asymptotic efficiency gains from the use of the approximately optimal instrument $\zeta_t^{(101)}$ relative to asymptotic efficiency 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 restrictedness of the space of instruments that are linear in lags of the basic instrument and a promise of the taken approach. In many unreported experiments the same pattern emerges, and in no case 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. Taking into account the difficulty of its derivation, it seems better 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 and its derivation are 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.

1.3 Theory: multiple equations case

1.3.1 Optimal instrument

In the multiple equation case, the conditional moment restriction is

$$E_t[\mathbf{e}_t] = 0, \tag{1.3.25}$$

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

$$\mathrm{D}_t \equiv E_t\left[\frac{\partial f(\boldsymbol{\beta}, \mathbf{x}_t)}{\partial \boldsymbol{\beta}}\right]. \tag{1.3.26}$$

and $s \times s$ matrices

$$\Omega_t \equiv E_t[\mathbf{e}_t\mathbf{e}_t'], \quad \Gamma_t \equiv E_t[\mathbf{e}_{t-1}\mathbf{e}_t'], \tag{1.3.27}$$

the conditional variance and conditional first-order autocovariance of the error vector. We again assume away higher-order conditional serial correlation in the error $\mathbf{e}_t$, i.e. let $E_t[\mathbf{e}_t\mathbf{e}_{t-j}']$ $= \mathrm{O}$ for $j > 1$.

Under suitable conditions, the optimal instrumental variables estimator takes the form (Anatolyev in press)

$$\Xi_t = \Xi_{t-1}\Phi_t + \Delta_t\mathrm{P}_t, \tag{1.3.28}$$

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:

$$\Gamma_t + \Phi_t(\Omega_t + E_t[\Phi_{t+1}\Gamma_{t+1}']) = 0, \tag{1.3.29}$$

$$E_t[\mathrm{I}_s - \mathrm{P}_t(\Omega_t - \Gamma_{t+1}\mathrm{P}_{t+1}\Gamma'_{t+1}]) = 0, \tag{1.3.30}$$

$$\Delta_t = \mathrm{D}_t + E_t[\Delta_{t+1}\Phi'_{t+1}], \tag{1.3.31}$$

$$\lambda(\Phi) \equiv \lim_{T\to\infty} \frac{1}{T} \log |\Phi_T \Phi_{T-1} \cdots \Phi_2 \Phi_1| < 0. \tag{1.3.32}$$

In the single equation case of the previous section, negativity of the top Lyapounov exponent $\lambda(\Phi)$ is equivalent to negativity of $E[\log|\Phi_t|]$. In the multiple equation case, the condition $E[\log|\Phi_t|] < 0$ 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 2 *(Bougerol and Picard 1992) Let A_t be a stationary ergodic matrix process with finite $E[\max(0, \log|A_t|)]$ such that almost surely*

$$\lim_{T\to\infty} |A_T A_{T-1} \cdots A_2 A_1| = 0.$$

Then $\lambda(A) < 0$.

Thus, when Φ_t has a triangular structure, 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). Alternatively, one may impose existence of matrix process S_t such that $E[\log|S_t \Phi_t S_t^{-1}|] < 0$. For more on these issues, see Pötscher and Prucha (1997, p.70 and footnote 25).

1.3.2 Approximately optimal instrument

The parameter values 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 $\mathbf{e}_t$: $\mathbf{e}_t = \mathbf{w}_{t+1} - \Theta\mathbf{w}_t$ and $\Sigma \equiv E[\mathbf{w}_t\mathbf{w}'_t]$. The zeroth-order approximation $\Phi_t^{(0)} = \Phi^H$ is the one that satisfies the matrix quadratic equation

$$\left(\Phi^H\right)^2 \Theta\Sigma - \Phi^H \left(\Sigma + \Theta\Sigma\Theta'\right) + \Sigma\Theta' = 0. \tag{1.3.33}$$

Note that the unstable solution is trivially Θ^{-1}, but the stable one is *not* Θ. Let representation $\mathbf{e}_t = \mathbf{w}_{t+1} - \Theta\mathbf{w}_t$ be invertible, i.e. all s eigenvalues π_i, $i = 1, \ldots, s$, of Θ lie strictly inside the unit circle of 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:

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

Then the stable solution of (1.3.33) is $\Phi^H = \left(\Psi^{-1}\Pi\Psi\right)'$. This follows, for example, from application of Theorems 3 and 4 of Uhlig (1995).

A linear expansion yields the following difference equation for Φ_t^1 with a stationary solution

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

[3] Some of these eigenvalues may well be complex. However, even then the resulting solution $\Phi^{\mathcal{O}}$ will be real-valued. See Uhlig (1995).

The $k \times s$ matrix Δ_t solves (1.3.31), 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

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

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 $\left[0, (1-\epsilon)^2\right]$ 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 $A^- = P\Delta^- Q'$.

1.3.3 Heaton–Ogaki example

Optimal instrument

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

Let w_t be a Gaussian $q \times 1$ vector white noise which is homoskedastic conditionally on its past history, and u_t be a two-period ahead forecast error with the Wold representation $u_t = \nu_0' w_t + \nu_1' w_{t-1}$, where ν_0 and ν_1 are $q \times 1$ vector constants. Observable at time t is $q \times 1$ vector x_t, so the space of instruments is $\Im_t = \sigma(x_t, x_{t-1}, \ldots)$. Let u_t be linked to x_t via $u_t = \left(1\ \beta\ \mathbf{0}_{q-2}'\right) x_t$, 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 one that is a 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 $\rho \equiv \frac{\nu_0' \nu_1}{\nu_0' \nu_0 + \nu_1' \nu_1} = \frac{1}{4}$. The disturbance vector of the two equation system is

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

The observational equation for x_t is $x_t = Hy_t$, where the law of motion of the $p \times 1$ state vector y_t is

$$y_t = Ay_{t-1} + Cw_t,$$

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 $\left(1\ \beta\ \mathbf{0}_{q-2}'\right) HC = \nu_0'$, $\left(1\ \beta\ \mathbf{0}_{q-2}'\right) HAC = \nu_1'$, $\left(1\ \beta\ \mathbf{0}_{q-2}'\right) HA^iC = 0$, $i \geq 2$. Then

$$\mathrm{D}_t = \begin{pmatrix} -hA^2 y_t & r_d + (hA^2 y_t)(\nu_1' w_t) \end{pmatrix},$$

where $h \equiv \left(0\ 1\ \mathbf{0}_{q-2}'\right) H$, $r_d \equiv -h\left(AC\varrho_1 + C\varrho_0\right)$, $\varrho_0 \equiv 2\rho\nu_0 - \nu_1$, $\varrho_1 \equiv 2\rho\nu_1 - \nu_0$.

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

$$\Phi_t = \frac{1}{\xi_{11}\xi_{12}} \begin{pmatrix} -\xi_{12}\xi_{21} & 0 \\ \left(\xi_{22}\alpha_{11}' - \xi_{12}\alpha_{21}'\right) w_t - \xi_{12}\alpha_{22}' w_{t-1} & -\xi_{11}\xi_{22} \end{pmatrix},$$

$$\Delta_t \mathrm{P}_t = \begin{pmatrix} r_1 y_t + r_2(\nu_1' w_t) & r_2 \end{pmatrix},$$

where constants ξ_{ij} and α_{ij} are defined in equations (4)–(10) of Heaton and Ogaki (1991), and the $1 \times p$ vector r_1 and the scalar r_2 are

$$r_1 = -\tfrac{1}{\xi_{11}^2} hA^2 \left(\mathrm{I}_p + \tfrac{\xi_{21}}{\xi_{11}} A\right)^{-1}, \qquad r_2 = \tfrac{1}{\xi_{12}^2}\left(1 + \tfrac{\xi_{22}}{\xi_{12}}\right)^{-1}\left(r_d - \tfrac{\xi_{21}}{\xi_{11}} r_1 C\varrho_1\right).$$

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 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 section 1.3.1). Finally, $E[|\Xi_t|^4] < \infty$ due to normality of w_t.

Approximately optimal instrument

Now we derive the approximation for the optimal instrument. The heteroskedasticity parameters are:

$$\Omega_t = (\nu_0'\nu_0 + \nu_1'\nu_1) \begin{pmatrix} 1 & -\nu_1' w_t \\ -\nu_1' w_t & \varrho^2 + (\nu_1' w_t)^2 \end{pmatrix},$$

$$\Gamma_t = (\nu_0'\nu_1) \begin{pmatrix} 1 & -\nu_1' w_t \\ \varrho_1' w_t - \nu_1' w_{t-1} & -2\rho\varrho^2 - \nu_1' w_t\left(\varrho_1' w_t - \nu_1' w_{t-1}\right) \end{pmatrix},$$

where $\varrho^2 \equiv \nu_0'\nu_0 - \rho\nu_0'\nu_1$. Consequently,

$$\Omega^H = (\nu_0'\nu_0 + \nu_1'\nu_1) \begin{pmatrix} 1 & 0 \\ 0 & (\nu_0'\nu_0 + \nu_1'\nu_1)(1 - \rho^2) \end{pmatrix}, \quad \Gamma^H = (\nu_0'\nu_1) \begin{pmatrix} 1 & 0 \\ 0 & (\nu_0'\nu_1)(2\rho^2 - 1) \end{pmatrix}.$$

Since both Ω^H and Γ^H are diagonal, Φ^H is too and is equal to $\Theta \equiv \mathrm{diag}(\theta_1, \theta_2)$, 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

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

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

$$\Phi_t^{(1)} = \Theta - \left[\Theta\Omega_t + \Gamma_t + \Theta^2 \nu_0' \nu_1 \begin{pmatrix} 1 & -\nu_1' w_t \\ 0 & (\nu_0'\nu_1)(2\rho^2 - 1) \end{pmatrix} \right] \Sigma^{-1},$$

$$\Delta_t^{(0)} = \begin{pmatrix} \xi_{11}^2 r_1 y_t & \dfrac{r_d + \theta_2 h A^2 C \nu_1}{1 - \theta_2} + (hA^2 y_t)(\nu_1' w_t) \end{pmatrix},$$

$$\mathrm{P}_t^{(1)} = \frac{\rho}{\varrho^2 \nu_0' \nu_1} \begin{pmatrix} \varrho^2 + (\nu_1' w_t)^2 & \nu_1' w_t \\ \nu_1' w_t & 1 \end{pmatrix}.$$

Asymptotic Comparisons

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

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

The following table 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 further an overwhelming part of the efficiency gains provided by the optimal instrument.

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

Thus, the efficiency losses arising from the approximation error turn out to be tiny, and show that the proposed instrument is able to nearly attain the efficiency bound.

1.4 Simulation Evidence

1.4.1 Model and data generating mechanism

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

$$y_t = \alpha + \beta x_t + e_t, \tag{1.4.34}$$

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

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

Apart from the constant, the basic instrument is scalar z_t. We set the auxiliary parameter values as follows. The parameter of the disturbance is θ runs through -0.8, -0.5, -0.3, 0.3, 0.5, 0.8. The value of φ is set to 0.3. The skedastic function is set to

$$\sigma_t^2 = \left(z_t + z_{t-1}\right)^2, \tag{1.4.35}$$

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

$$E\left[x_t|\Im_t\right] = 1 + z_t + z_{t-1}. \tag{1.4.36}$$

The econometrician uses the information that σ_t^2 is quadratic, and $E\left[x_t|\Im_t\right]$ is linear, in z_t and z_{t-1}. However, we also investigate the behavior of the proposed estimator when the true mechanism driving heteroskedasticity or the form of the conditional expectation of the right hand variable given the instrument history is not the one assumed by the econometrician. We explore the following five variations of the DGP.

(1) "Absolute value misspecification": the conditional variance is proportional to the presumed conditional standard deviation: $\sigma_t^2 = 3.12 \cdot |z_t + z_{t-1}|$.

(2) "Exponential misspecification": the conditional variance is proportional to an exponent of the presumed conditional standard deviation: $\sigma_t^2 = 0.22 \cdot \exp(z_t + z_{t-1})$.

(3) "Inverse misspecification": the conditional variance is inversely proportional to the presumed conditional variance: $\sigma_t^2 = 21.5/(1 + (z_t + z_{t-1})^2)$.

(4) "Projection misspecification": the right hand variable includes one more historical value of the basic instrument: $E\left[x_t|\Im_t\right] = 1 + z_t + z_{t-1} + z_{t-2}$.

(5) "Nonlinear misspecification": the right hand variable includes an additional quadratic term: $E\left[x_t|\Im_t\right] = 1 + z_t + z_t^2 + z_{t-1}$.

In (1)–(3), the constants are tuned so that the empirical variance of w_t evaluated from 5 million observations is the same as in (1.4.35).

1.4.2 Estimators and details of their construction

In this subsection, we present simulation evidence on the behavior of the proposed estimator together with that of some frequently used competitors. The latter are:

(1) The simple IV estimator $\widehat{\beta}_{IV}$ with $(1\ z_t)'$ as a just-identifying instrument for $(1\ x_t)'$. 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 $\widehat{\beta}_{2SLS}$ with $(1\ z_t\ z_{t-1})'$ as an overidentifying instrument. This would probably be the most intensively used estimator in this context, in spite of the presence of heteroskedasticity.

(3) The feasible $\widehat{\beta}_{\zeta^H}$ that would be a feasible optimal IV estimator in the absence of heteroskedasticity. This estimator has been previously known and used in the literature (see footnote 2).

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

Other possible competing estimators could be conventional optimal GMM estimators with instruments that contain finite number of 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. Therefore we do not consider such estimators here. For their detailed consideration in conditionally heteroskedastic environments, see Tauchen (1986) and West, Wong and Anatolyev (2002).

Now we give details of construction of $\widehat{\boldsymbol{\zeta}}_t^{(101)}$. Suppose the basic instrument z_t is fitted to an autoregressive model of order 2 (note that this is higher than the true order). Then $\mathbf{u}_t \equiv (z_t\ z_{t-1}\ 1)'$ has the following law of motion:

$$\mathbf{u}_{t+1} = G_{uu}\mathbf{u}_t + \boldsymbol{\eta}_{u,t+1}, \tag{1.4.37}$$

for 3×3 companion matrix G_{uu} and 3×1 noise vector $\boldsymbol{\eta}_{ut}$. The law of motion of the right-hand-side variable x_t is

$$x_t = G_{xu}\mathbf{u}_t + \eta_{xt}, \tag{1.4.38}$$

for 1×3 matrix G_{xu} and scalar noise η_{xt}. Note that due to (1.4.38) $\mathbf{d}_t = G_{du}\mathbf{u}_t$, where G_{du} is 2×3. The zeroth-order parameters are: $\phi^H = \theta$, $\boldsymbol{\delta}_t^{(0)} = \boldsymbol{\delta}_t^H = G_{\delta u}\mathbf{u}_t$, $\rho^H = \sigma^{-2}$, where $G_{\delta u} \equiv G_{du}\left(\mathrm{I}_3 - \theta G_{uu}\right)^{-1}$.

Let $\mathbf{v}_t \equiv \left(z_t^2\ z_t z_{t-1}\ z_{t-1}^2\ 1\right)'$. The quadratic heteroskedasticity is such that $E_t[w_t^2]$ and $E_t[\eta_{zt+1}^2]$ are linear forms in $\mathbf{v}_t$ and $\mathbf{u}_t$. For what follows we need to obtain the law of motion of $\mathbf{v}_t$. Note that $\Im_t$-nonmesurable components of $\mathbf{v}_{t+1}$ are z_{t+1}^2 and $z_{t+1}z_t$. It is not hard to see that $E_t[z_{t+1}^2]$ and $E_t[z_{t+1}z_t]$ are linear forms in $\mathbf{v}_t$ and $\mathbf{u}_t$. As a result, we can represent $\mathbf{v}_t$ as

$$\mathbf{v}_{t+1} = G_{vv}\mathbf{v}_t + G_{vu}\mathbf{u}_t + \boldsymbol{\eta}_{v,t+1}, \tag{1.4.39}$$

where G_{vv} is 4×4, G_{vu} is 4×3, and $\boldsymbol{\eta}_{v,t+1}$ is 4×1 with $E_t\left[\boldsymbol{\eta}_{v,t+1}\right] = 0$. It is also straightforward to see that ω_t and γ_t are linear forms in $\mathbf{v}_t$ and $\mathbf{u}_t$ as well:

$$\omega_t = \mathbf{g}_{\omega v}'\mathbf{v}_t + \mathbf{g}_{\omega u}'\mathbf{u}_t, \quad \gamma_t \equiv \mathbf{g}_{\gamma v}'\mathbf{v}_t + \mathbf{g}_{\gamma u}'\mathbf{u}_t,$$

where $\mathbf{g}_{\omega v}$, $\mathbf{g}_{\omega u}$, $\mathbf{g}_{\gamma v}$ and $\mathbf{g}_{\gamma u}$ are functions of $\mathbf{g}_{wv}$, $\mathbf{g}_{wu}$, $\mathbf{g}_{vv}$, $\mathbf{g}_{vu}$, G_{vv}, G_{vu}, G_{uu} and θ. In constructing the feasible estimator, we presume that the researcher is unaware of the true structure of skedastic function, except that it is quadratic in z_t and z_{t-1}. Therefore, the researcher has to estimate $\mathbf{g}_{\gamma v}$, $\mathbf{g}_{\gamma u}$, $\mathbf{g}_{\omega v}$ and $\mathbf{g}_{\omega u}$ directly from regressions of e_t^2 and $e_t e_{t-1}$ on $\mathbf{v}_t$ and $\mathbf{u}_t$:

$$e_t^2 = \mathbf{g}_{\omega v}'\mathbf{v}_t + \mathbf{g}_{\omega u}'\mathbf{u}_t + \eta_{\omega,t+1}, \quad e_t e_{t-1} = \mathbf{g}_{\gamma v}'\mathbf{v}_t + \mathbf{g}_{\gamma u}'\mathbf{u}_t + \eta_{\gamma,t+1}. \tag{1.4.40}$$

Straightforward computation shows (see the Appendix) that $\phi_t^{(1)} = \mathbf{g}_{\phi v}'\mathbf{v}_t + \mathbf{g}_{\phi u}'\mathbf{u}_t$, where

$$\begin{aligned}
\mathbf{g}_{\phi v}' &\equiv \theta e_4' - \tfrac{1}{\sigma^2}\left\{\left(\theta \mathbf{g}_{\omega v}' + 2\mathbf{g}_{\gamma v}'\right)\left(\mathrm{I}_4 - \theta^2 G_{vv}\right)^{-1} - \mathbf{g}_{\gamma v}'\right\}, \\
\mathbf{g}_{\phi u}' &\equiv -\tfrac{1}{\sigma^2}\left\{\left(\theta^2\left(\theta \mathbf{g}_{\omega v}' + 2\mathbf{g}_{\gamma v}'\right)\left(\mathrm{I}_4 - \theta^2 G_{vv}\right)^{-2} G_{vu} + \theta \mathbf{g}_{\omega u}' + 2\mathbf{g}_{\gamma u}'\right)\left(\mathrm{I}_3 - \theta^2 G_{uu}\right)^{-1} - \mathbf{g}_{\gamma u}'\right\}.
\end{aligned}$$

The two feasible approximately optimal instruments are formed via

$$\widehat{\boldsymbol{\zeta}}_t^H = \widehat{\boldsymbol{\zeta}}_{t-1}^H \widehat{\theta} + \frac{\widehat{G}_{\delta u}\mathbf{u}_t}{\widehat{\sigma}^2} \tag{1.4.41}$$

and

$$\widehat{\boldsymbol{\zeta}}_t^{(101)} = \widehat{\boldsymbol{\zeta}}_{t-1}^{(101)}\left[\widehat{\mathbf{g}}_{\phi v}'\mathbf{v}_t + \widehat{\mathbf{g}}_{\phi u}'\mathbf{u}_t\right]^- + \frac{\widehat{G}_{\delta u}\mathbf{u}_t}{\max(\widehat{\mathbf{g}}_{\omega v}'\mathbf{v}_t + \widehat{\mathbf{g}}_{\omega u}'\mathbf{u}_t, \epsilon_\omega)}. \tag{1.4.42}$$

The trimming parameters are set at: $\epsilon_\phi = 10^{-2}$, $\epsilon_\omega = \frac{1}{5}\widehat{\sigma}_e^2$. In addition, if the sample average of $\widehat{\mathbf{g}}_{\phi v}'\mathbf{v}_t + \widehat{\mathbf{g}}_{\phi u}'\mathbf{u}_t$ exceeds 1 in absolute value, we classify the circumstances as unfavorable and substitute $\widehat{\boldsymbol{\zeta}}_t^H$ in place of $\widehat{\boldsymbol{\zeta}}_t^{(101)}$. To recapitulate, the algorithm consists of the following steps.

(1) Obtain consistent preliminary estimates of $(\alpha\ \beta)'$ by 2SLS. Call the residuals $\widehat{e}_t$ and obtain the sample variance $\widehat{\sigma}_e^2$ and first-order autocovariance $\widehat{\sigma}_{e,1}$. Compute the estimates $\widehat{\sigma}^2$ and $\widehat{\theta}$ of the implied σ^2 and θ.

(2) Estimate the law of motion (1.4.37) of $\mathbf{u}_t$ by OLS and construct $\widehat{G}_{uu}$. Estimate the projection (1.4.38) of regressors $\mathbf{x}_t$ on the space of instruments by OLS, construct $\widehat{G}_{xu}$, $\widehat{G}_{\delta u}$ and fitted $\widehat{\mathbf{d}}_t$. Estimate the law of motion (1.4.39) of $\mathbf{v}_t$ by OLS, obtain estimates of nontrivial entries in G_{vv} and G_{vu} and construct $\widehat{G}_{vv}$ and $\widehat{G}_{vu}$. Estimate the projections (1.4.40) of $\widehat{e}_t\widehat{e}_{t-1}$ and $\widehat{e}_t^2$ on the space of instruments by OLS, obtain $\widehat{\mathbf{g}}_{\gamma v}$, $\widehat{\mathbf{g}}_{\gamma u}$, $\widehat{\mathbf{g}}_{\omega v}$, and $\widehat{\mathbf{g}}_{\omega u}$. Construct $\widehat{\mathbf{g}}_{\phi v}$, $\widehat{\mathbf{g}}_{\phi u}$, $\widehat{G}_{\delta u}$.

(3) Initialize $\widehat{\boldsymbol{\zeta}}_0^H = 0$ or $\widehat{\boldsymbol{\zeta}}_0^{(101)} = 0$ and construct a series for the feasible approximately optimal instrument by (1.4.41) or (1.4.42). Estimate $(\alpha\ \beta)'$ by applying the just-identifying IV estimator with instrument $\widehat{\boldsymbol{\zeta}}_t^H$ or $\widehat{\boldsymbol{\zeta}}_t^{(101)}$.

1.4.3 Results

We set the sample size T to 300 and 900. Table 1 compares sample standard errors of the four estimators (sample means and medians are practically zero for all estimators). It is clear that most of the time the asymptotic gains evaluated earlier are realized in finite samples as well. The last column shows a fraction of cases when the substitution of $\widehat{\boldsymbol{\zeta}}_t^{(101)}$ by $\widehat{\boldsymbol{\zeta}}_t^H$ in adverse circumstances takes place. The fraction abruptly increases when the serial correlation is stronger, and not always larger T eliminates this effect. Considering the efficiency gains provided by the instrument $\widehat{\boldsymbol{\zeta}}_t^H$ relative to traditional IV estimators and further efficiency gains provided by the instrument $\widehat{\boldsymbol{\zeta}}_t^{(101)}$ relative to $\widehat{\boldsymbol{\zeta}}_t^H$, one may say that the former exploits well the serial correlation structure of the problem, while the latter takes further advantage of conditional heteroskedasticity. This "division of labor" is more pronounced for smaller $|\theta|$ when the strength of heteroskedasticity is greater relative to that of serial correlation. For larger $|\theta|$ the instrument $\widehat{\boldsymbol{\zeta}}_t^{(101)}$ captures the overwhelming share of efficiency gains.

Table 2 contains results on the five misspecified DGPs with $\theta = 0.5$ in the format of Table 1. The proposed instrument $\widehat{\boldsymbol{\zeta}}_t^{(101)}$ still delivers efficiency gains comparable to those in Table 1, especially when the misspecified heteroskedasticity is strong. The misspecification in the conditional expectation of the right hand variable may or may not spoil the performance of the approximately optimal instrument depending on the type of that misspecification. Again, as for the correctly specified model, the sample size T does not play a central role.

1.5 Conclusion

For general two-period conditional moment restrictions, both single and multiple equation, characterized by the presence of conditional heteroskedasticity, we show how to construct approximately optimal instruments, the ones that approximately satisfy the system of optimality conditions, evaluate the asymptotic properties of corresponding instrumental variables estimators, and verify their finite sample behavior. We concentrate on the first order serial correlation since 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) or data contaminated by temporal aggregation (Hall 1988).

The applications with conditional moment restrictions that have higher than the first order of serial correlation are less frequent, and in addition the likelihood that a researcher will want to exploit the idea of optimal instruments diminishes. 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 in press). 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 time varying parameters. The analog of the equation (1.3.29) 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 (1.3.30) and (1.3.31) are more involved as well.

There are of course limitations of the approach presented in this paper. First, it is an intrinsic trade-off between efficiency and robustness to auxiliary parametrizations, the general feature of the optimal instrumental variables approach. Second, it is a trade-off between efficiency and a cost of constructing the instrument resulting from its complicatedness. Third, although the approximation error was found to be small in the situations that were considered, it may potentially be larger or smaller, and we are unable to clearly rank the approximately optimal instrument among other instruments in terms of asymptotic efficiency. Last, a theoretical econometrician may be dissatisfied by the undertaken approximation as a sacrifice of achieving the instrumental variables efficiency bound. Another approach that does not invoke approximations lies in an attempt to estimate the auxiliary processes directly from the system that defines them by designing a contractive iterative scheme, and estimating the auxiliary conditional expectations nonparametrically.

1.6 Appendix

1.6.1 Proof of Lemma 1

An extension of Theorem 1 of Brandt (1986) for matrix-valued A_t, B_t and Ψ_t gives existence of a stationary ergodic solution Ψ_t of (1.2.22) and its representation (1.2.23) with the right-hand side converging absolutely almost surely, if $-\infty \le E\left[\log|A_t|\right] < 0$ and $E\left[\max\left(0, \log|\Psi_t|\right)\right] < \infty$. But both conditions are satisfied by the assumptions of Lemma 1: $E\left[\log|A_t|\right] < E[\log 1] = 0$ and $E\left[\max\left(0, \log|B_t|\right)\right] \le E\left[|B_t|\right] \le E[|B_t|^4]^{\frac{1}{4}} < \infty$. Next, there exists $0 \le a < 1$ such that $\operatorname{esssup}|A_t| \le a$. Then from (1.2.22) we get by the triangular inequality applied in the $L^4(\mathrm{Pr})$ space and stationarity, that $E[|\Psi_t|^4]^{\frac{1}{4}} = E[|\Psi_{t-1}A_t + B_t|^4]^{\frac{1}{4}} \le aE[|\Psi_t|^4]^{\frac{1}{4}} + E[|B_t|^4]^{\frac{1}{4}}$, from which it follows that $E[|\Psi_t|^4] \le (1-a)^{-4}E[|B_t|^4] < \infty$. □

1.6.2 Constants

The constants in various approximations of parameters of the optimal instrument in the pilot example are:

$$\begin{aligned}
\kappa_{\omega 1} &= 1+\theta^2-\lambda\left(\varphi^2+\theta^2\right), \quad \kappa_{\omega 2} = \lambda\left(1-\varphi^2\right)\left(\varphi^2+\theta^2\right), \\
\kappa_{\gamma 1} &= -\theta\left(1-\lambda\right), \quad \kappa_{\gamma 2} = -\lambda\theta\left(1-\varphi^2\right), \\
\kappa_{\phi 1} &= \theta\left\{1-\frac{\lambda\left(1-\varphi^2\right)\left(1-\theta^2\right)}{1-\varphi^2\theta^2}\right\}, \quad \kappa_{\phi 2} = \frac{\lambda\theta\left(1-\varphi^2\right)^2\left(1-\theta^2\right)}{1-\varphi^2\theta^2}, \\
\kappa_{\phi 3} &= \lambda\kappa_{\phi 1} + \frac{1}{1-\theta^2}\left\{\frac{\kappa_{\phi 1}^2}{\theta} + \frac{2\theta\kappa_{\phi 1}\kappa_{\phi 2}}{1-\varphi^2\theta^2} + \frac{3\theta\left(1+\varphi^2\theta^2\right)\kappa_{\phi 2}^2}{\left(1-\varphi^2\theta^2\right)\left(1-\varphi^4\theta^2\right)}\right\}, \\
\kappa_{\phi 4} &= \lambda\kappa_{\phi 2} + \frac{1}{\theta}\left\{\kappa_{\phi 1}\kappa_{\gamma 2} + \frac{2\kappa_{\phi 1}\kappa_{\phi 2}}{1-\varphi^2\theta^2} + \frac{6\varphi^2\theta^2\kappa_{\phi 2}^2}{\left(1-\varphi^2\theta^2\right)\left(1-\varphi^4\theta^2\right)}\right\}, \\
\kappa_{\phi 5} &= \frac{\kappa_{\phi 2}}{\theta}\left\{\kappa_{\gamma 2} + \frac{\kappa_{\phi 2}}{1-\varphi^4\theta^2}\right\}, \\
\kappa_{\delta 1} &= \frac{1}{1-\varphi\theta}, \quad \kappa_{\delta 2} = \kappa_{\delta 1} + \varphi\kappa_{\delta 1}^2\left\{\kappa_{\phi 1} + \frac{3\kappa_{\phi 2}}{1-\varphi^3\theta}\right\}, \quad \kappa_{\delta 3} = \frac{\varphi^3\kappa_{\phi 2}\kappa_{\delta 1}}{1-\varphi^3\theta},
\end{aligned}$$

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

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: $w\left(x\right) \equiv e^{-x^2}\left(1 + \frac{2i}{\sqrt{\pi}}\int\limits_0^x e^{-t^2}dt\right) = \sum\limits_{n=0}^{\infty}\frac{(ix)^n}{\Gamma\left(\frac{n}{2}+1\right)}$ (see Gautschi 1974).

1.6.3 Linear IV bound

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

$$
\forall k \geq 0 \qquad E\left[\eta_{t-k}z_t\right] = E\left[\eta_{t-k}z_t^* e_t^2\right] + E\left[\eta_{t-k-1}z_t^* e_t e_{t-1}\right] + E\left[\eta_{t-k}z_{t-1}^* e_t e_{t-1}\right]. \qquad (1.6.43)
$$

The left hand side in (1.6.43) is φ^k. Calculate the three terms on the right hand side using $E_t[e_t^2] = \kappa_{\omega 1} + \kappa_{\omega 2}z_t^2$, $E_t[e_t e_{t-1}] = \kappa_{\gamma 1} + \kappa_{\gamma 2}z_t^2$, $z_t^* = \sum\limits_{i=0}^{\infty} g_i\eta_{t-i}$ and $z_t = \sum\limits_{i=0}^{\infty}\varphi^i\eta_{t-i}$:

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

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

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

Therefore the system (1.6.43) can be written in a matrix form

$$
\Psi = \mathrm{SG},
$$

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

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

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

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

1.6.4 Derivation of auxiliary processes used in simulations

Observe that

$$E_t[\mathbf{u}_{t+i}] = G_{uu}^i \mathbf{u}_t$$

and

$$E_t[\mathbf{v}_{t+i}] = G_{vv}^i \mathbf{v}_t + \sum_{j=0}^{i-1} G_{vv}^{i-j-1} G_{vu} E_t[\mathbf{u}_{t+i}] = G_{vv}^i \mathbf{v}_t + \left[\sum_{j=0}^{i-1} G_{vv}^{i-j-1} G_{vu} G_{uu}^i\right] \mathbf{u}_t.$$

But

$$\sum_{i=1}^{\infty} \theta^{2i} \left[\sum_{j=0}^{i-1} G_{vv}^{i-j-1} G_{vu} G_{uu}^i\right] = \sum_{k=1}^{\infty}\sum_{i=k}^{\infty} \theta^{2i} \sum_{j=0}^{k-1} G_{vv}^{i-j-1} G_{vu} G_{uu}^j = \sum_{k=1}^{\infty}\sum_{j=0}^{k-1}\left[\sum_{i=k}^{\infty} \theta^{2i} G_{vv}^{i-j-1}\right] G_{vu} G_{uu}^j$$
$$= \sum_{k=1}^{\infty}\sum_{j=0}^{k-1}\left(\mathrm{I}_4 - \theta^2 G_{vv}\right)^{-1} \theta^{2k} G_{vv}^{k-j-1} G_{vu} G_{uu}^j = \left(\mathrm{I}_4 - \theta^2 G_{vv}\right)^{-1} \sum_{i=1}^{\infty}\left[\sum_{k=i}^{\infty} \theta^{2k} G_{vv}^{k-i}\right] G_{vu} G_{uu}^{i-1}$$
$$= \left(\mathrm{I}_4 - \theta^2 G_{vv}\right)^{-1} \sum_{i=1}^{\infty} \theta^{2i} \left(\mathrm{I}_4 - \theta^2 G_{vv}\right)^{-1} G_{vu} G_{uu}^{i-1} = \theta^2 \left(\mathrm{I}_4 - \theta^2 G_{vv}\right)^{-2} G_{vu} \left(\mathrm{I}_3 - \theta^2 G_{uu}\right)^{-1}.$$

Therefore,

$$\begin{aligned} \sum_{i=0}^{\infty} \theta^{2i} E_t[\mathbf{v}_{t+i}] &= \sum_{i=0}^{\infty} \theta^{2i} G_{vv}^i \mathbf{v}_t + \sum_{i=1}^{\infty} \theta^{2i} \left[\sum_{j=0}^{i-1} G_{vv}^{i-j-1} G_{vu} G_{uu}^i\right] \mathbf{u}_t \\ &= \left(\mathrm{I}_4 - \theta^2 G_{vv}\right)^{-1} \mathbf{v}_t + \theta^2 \left(\mathrm{I}_4 - \theta^2 G_{vv}\right)^{-2} G_{vu} \left(\mathrm{I}_3 - \theta^2 G_{uu}\right)^{-1} \mathbf{u}_t. \end{aligned}$$

Also,

$$\sum_{i=0}^{\infty} \theta^{2i} E_t[\mathbf{u}_{t+i}] = \sum_{i=0}^{\infty} \theta^{2i} G_{uu}^i \mathbf{u}_t = \left(\mathrm{I}_3 - \theta^2 G_{uu}\right)^{-1} \mathbf{u}_t.$$

Hence, $\phi_t^{(1)} = \mathbf{g}_{\phi v}' \mathbf{v}_t + \mathbf{g}_{\phi u}' \mathbf{u}_t$, as in the text.

Bibliography

[1] Anatolyev, S. (in press) The Form of the Optimal Nonlinear Instrument for Multiperiod Conditional Moment Restrictions. *Econometric Theory*, forthcoming.

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

[3] 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.

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

[5] 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.

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

[7] Gautschi, W. (1974) Error Function and Fresnel Integrals. Chapter 7 in: M. Abramowitz and I.A. Stegun, eds., *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables.* Dover Publications.

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

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

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

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

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

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

Table 1. Sample statistics for $\widehat{\beta}_{IV}$, $\widehat{\beta}_{2SLS}$, $\widehat{\beta}_{\zeta^H}$ and $\widehat{\beta}_{\zeta^{(101)}}$, from simulations.

DGP		Standard deviation, $\times 10^{-2}$				%%
θ	T	$\widehat{\beta}_{IV}$	$\widehat{\beta}_{2SLS}$	$\widehat{\beta}_{\zeta^H}$	$\widehat{\beta}_{\zeta^{(101)}}$	$(101) \to H$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0.8	300	12.36	7.22	5.79	5.45	63%
	900	7.01	4.10	3.22	3.08	79%
0.5	300	11.78	7.16	6.60	5.75	18%
	900	6.71	4.09	3.74	3.09	8%
0.3	300	12.20	8.09	7.87	6.16	$< 1\%$
	900	6.97	4.64	4.52	3.35	$\ll 1\%$
-0.3	300	16.41	13.31	13.10	9.62	$< 1\%$
	900	9.49	7.72	7.60	5.29	$\ll 1\%$
-0.5	300	18.45	15.43	14.84	11.76	14%
	900	10.68	8.96	8.59	6.41	7%
-0.8	300	21.81	18.74	17.25	15.40	57%
	900	12.64	10.88	9.91	9.29	76%

The data generating mechanism is $y_t = \alpha + \beta x_t + e_t$, $\alpha = \beta = 0$, $e_t = w_{t+1} - \theta w_t$, $w_t|\Im_t \sim \mathcal{N}\left(0, \sigma_t^2\right)$, $x_t = E\left[x_t|\Im_t\right] + \eta_{xt}$, where $\Im_t \equiv \sigma(z_t, z_{t-1}, \ldots)$, $z_t = 1 + \varphi(z_{t-1} - 1) + \eta_{zt}$, $(\eta_{xt}, \eta_{zt}) \sim IID\, \mathcal{N}\left(0, \mathrm{I}_2\right)$, $\sigma_t^2 = (z_t + z_{t-1})^2$, $E\left[x_t|\Im_t\right] = 1 + z_t + z_{t-1}$. The number of repetitions is 100,000. For each set of parameters and each sample size columns 3–6 present sample standard errors for the following IV estimators: the just-identifying IV estimator $\widehat{\beta}_{IV}$ that uses $(1\ z_t)'$ as a vector of instruments; the two-stage least squares estimator $\widehat{\beta}_{2SLS}$ that uses $(1\ z_t\ z_{t-1})'$ as a vector of instruments; the feasible estimator $\widehat{\beta}_{\zeta^H}$ that would be a feasible optimal IV estimator in the absence of heteroskedasticity, and the feasible proposed estimator $\widehat{\beta}_{\zeta^{(101)}}$. Column 7 indicates how frequently $\widehat{\beta}_{\zeta^{(101)}}$ was subtituted by $\widehat{\beta}_{\zeta^H}$.

Table 2. Sample statistics for $\widehat{\beta}_{IV}$, $\widehat{\beta}_{2SLS}$, $\widehat{\beta}_{\zeta^H}$ and $\widehat{\beta}_{\zeta^{(101)}}$, from simulations, under misspecification.

Type of		Standard deviation, $\times 10^{-2}$				%%
misspecification	T	$\widehat{\beta}_{IV}$	$\widehat{\beta}_{2SLS}$	$\widehat{\beta}_{\zeta^H}$	$\widehat{\beta}_{\zeta^{(101)}}$	(101) → H
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Absolute value	300	11.35	7.08	6.56	6.55	7%
	900	6.47	4.06	3.72	3.59	$<$ 1%
Exponential	300	13.16	7.26	6.77	5.98	39%
	900	7.59	4.19	3.81	3.38	41%
Inverse	300	11.11	7.07	6.55	5.59	5%
	900	6.31	4.04	3.71	3.02	$<$ 1%
Projection	300	11.13	6.22	4.83	4.25	18%
	900	6.31	3.55	2.74	2.28	8%
Nonlinear	300	4.59	3.78	3.34	3.31	18%
	900	2.64	2.18	1.91	1.81	8%

The data generating mechanism is given in the notes to Table 1, with $\theta = 0.5$, except that now $\sigma_t^2 = 3.12 \cdot |z_t + z_{t-1}|$ in the ”absolute value” case, or $\sigma_t^2 = .22 \cdot \exp(z_t + z_{t-1})$ in the ”exponential” case, or $\sigma_t^2 = 21.5/(1 + (z_t + z_{t-1})^2))$ in the ”inverse” case, or $E[x_t|\Im_t] = 1 + z_t + z_{t-1} + z_{t-2}$ in the ”projection” case, or $E[x_t|\Im_t] = 1 + z_t + z_t^2 + z_{t-1}$ in the ”nonlinear” case. The number of repetitions is 100,000. For the estimators compared and the meaning of numbers see the notes to Table 1.

2. EMPIRICAL LIKELIHOOD, GMM, SERIAL CORRELATION, AND ASYMPTOTIC BIAS

2.1 Introduction

In recent years, one step estimators called "generalized empirical likelihood" (GEL) estimators (Smith 1997, 1998) begin to gain attention as theoretically attractive alternatives to GMM. These estimators are based on information theoretical considerations and include the empirical likelihood (Owen 1991, Qin and Lawless 1994, Imbens 1997) and exponential tilting (Kitamura and Stutzer 1997) estimators, together with an entire class of minimizers of certain divergence criteria (Imbens, Spady and Johnson 1998), continuously updating GMM (Hansen, Heaton and Yaron 1996), and other members. It has been established that the first order asymptotic properties of GEL estimators are identical to those of GMM estimators (Smith 1997, 1998). Moreover, it turns out that GEL estimators have certain advantages related to second order asymptotic properties and thus are expected to have better finite sample behavior. In particular, Newey and Smith (2000, 2001) find that in a cross sectional context the GEL estimators do not have some components of the second order bias that are characteristic of GMM estimators resulting from estimating the optimal linear combination of moment conditions at the preliminary step. The empirical likelihood (EL) estimator is the most distinctive in this respect in that its bias is the smallest, and moreover, its bias corrected version is second order asymptotically efficient. One more striking fact is that in an instrumental variables regression the bias of GEL estimators does not, in contrast to that of GMM estimators, grow with the number of instruments. This property makes the class of (appropriately modified) GEL estimators especially attractive in numerous stationary time series models typically estimated by GMM, with wide possibilities of selecting instruments.

In this paper, we consider stationary time series models where the moment function is serially correlated of known order, directing attention toward the conditional models of multistep prediction, examples of which 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. In such situations certain modifications of the baseline EL estimator are required to attain asymptotic efficiency (Imbens 1997, Smith 1997, 1998). Three possibilities are proposed in the literature. According to one approach (Imbens 1997), the inefficient EL estimator is adjusted by removing the asymptotic covariance between the estimator and the auxiliary parameter. This makes the estimator a multistep one, which nullifies its attractiveness. The other two modifications both based on temporal weighting of the moment function are different in that they produce one step final estimators. The moment function may be temporally added up inside the first order conditions (Back and Brown 1990, Imbens 1997), which gives rise to the modification of the EL estimator we call "CEL" (from "corrected EL"). Alternatively, the moment function may be smoothed with the use of a kernel function from the outset (Kitamura 1997, Smith 1997), which gives rise to the modification we call "SEL" (from "smoothed EL"). We present a more detailed description of these two modifications in section 2 of this paper, concentrating on the empirical likelihood estimator because its second order properties are more promising than those of other GEL estimators in the cross-sectional context.

The CEL version is much simpler in realization in practice in that it does not require kernel weighting, while the SEL version does, with unpleasant concerns about a choice of the bandwidth. However, as our subsequent analysis shows, the CEL estimator exhibits less preferable second order asymptotic properties. In section 3, we make the analysis of the second order bias of the CEL and SEL estimators, together with that of the two-step GMM estimator, for time series models where the moment function is serially correlated of known order. It turns out that although in some special cases the biases of the two EL estimators are equal, in general the CEL estimator has more bias components than the SEL estimator. Moreover, one of the bias components of the SEL estimator may be further removed by a judicious choice of the kernel smoother. In addition, in contrast to the SEL estimator, the bias of the CEL estimator has a component which may grow with the number of instruments, the property shared by the GMM estimator. Thus, despite its greater convenience in use, the CEL estimator is expected to be more biased in finite samples than the SEL. We also obtain a striking side result that smoothing the moment function tends to reduce the asymptotic bias even when the moment function is not serially correlated and smoothing is not necessary.

In section 4 we run a Monte Carlo experiment where we analyze the behavior of various estimators in estimation of the AR coefficient of an ARMA model, and, in particular, confirm our findings related to asymptotic bias. We compare the performance of the traditional GMM with empirical likelihood based CEL and SEL. Consistent with our analytical results, the SEL estimator exhibits smallest bias, and, despite a slightly bigger variance, dominates the other estimators in terms of MSE. In contrast, the GMM and CEL estimators are rather biased, especially for many instruments. We analyze the impact of the choice of kernel smoother and bandwidth for the SEL estimator, as well as observe the influence of the serial correlation order. It turns out that the truncated kernel works more predictably, often overperforming other kernels both in terms of bias and variance, confirming our analytical results, and that the impact of the choice of bandwidth is insignificant. We put special emphasis on selecting an instrumental vector, and its impact on statistical properties of estimators. In practice, when running GMM, the instruments are usually set to a couple or so observable variables dated most recently. We find that because of better asymptotic bias properties of the SEL estimator, the bias–variance trade off weakens so that the SEL estimator endures taking more lagged values into the instrumental vector thus allowing utilization of more information.

Section 5 of the paper suggests directions for future research. The appendix contains proofs and derivations of the second order asymptotic bias for the estimators under investigation.

2.2 Serial correlation consistent GMM and EL estimators

Suppose we have the following system of unconditional moment restrictions:

$$E\left[m\left(w_t,\theta\right)\right]=0, \tag{2.2.1}$$

where w_t is an observable random vector on which data from $t=1$ to $t=T$ are available, θ is $k\times 1$ vector of parameters to be estimated, and $m_t\equiv m\left(w_t,\theta\right)$ is $\ell\times 1$ moment function, $\ell>k$. We make the following assumptions about data generation.

Assumption 1 *The sequence w_t is strictly stationary and strongly mixing with mixing coefficients α_j satisfying $\sum_{j=1}^{\infty} j^2 \alpha_j^{1-1/\nu} < \infty$ for some $\nu > 1$.*

In particular, it follows that $\alpha_j = o\left(j^{-3\nu/(\nu-1)}\right)$. Such tight rate of decay for the mixing coefficients and integrability conditions of assumption 3(c) below are needed in the second order asymptotic analysis. The following assumption imposes serial correlation structure on the moment function.

Assumption 2 *The moment function is serially correlated of known order $q > 1$, i.e. $E\left[m_t m_{t-s}'\right] = 0$ if $|s| > q$ and $E\left[m_t m_{t-q}'\right] \neq 0$.*

Such structure does not imply that the moment function is p-dependent for some $p > q$. Denote $m_{\theta t} = \partial m\left(w_t, \theta\right)/\partial \theta'$, and let hats or bars over functions refer to them evaluated at appropriate estimates. Let $||A||$ denote the norm $\sqrt{\operatorname{tr}\left(A'A\right)}$ for any matrix A. We impose the following regularity conditions:

Assumption 3

(a) $\theta \in \operatorname{int}(\Theta)$, where $\Theta \subseteq \mathbb{R}^k$ is compact;

(b) $m\left(w_t, \theta\right)$ is a Borel measurable function for all $\theta \in \Theta$ and is continuously differentiable in the first argument for all $\theta \in \Theta$ for all w_t in its support;

(c) $E\left[||m_t||^{6\nu}\right] < \infty$ and $E\left[\sup_{\theta\in\Theta} ||m_{\theta t}||^{4\nu}\right]$ are finite for ν of assumption 1.

Let

$$\zeta_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} m_t,$$

and define the matrices

$$Q = E\left[m_{\theta t}\right], \quad V = \sum_{s=-q}^{q} E\left[m_t m_{t-s}'\right],$$

$$\Sigma = \left(Q'V^{-1}Q\right)^{-1}, \quad \Xi = \Sigma Q' V^{-1}, \quad \Omega = V^{-1} - V^{-1} Q \Xi.$$

Generalized method of moments (GMM) estimator Let the first step preliminary (and possibly inefficient) estimator $\bar{\theta}$ is used to form the efficient weight matrix[1]

$$\hat{W} = \left(\frac{1}{T} \sum_{t=1}^{T} \bar{m}_t \bar{m}_t^{\sigma\prime}\right)^{-1}$$

with $p\lim \hat{W} = V^{-1}$. The two-step GMM estimator $\hat{\theta}_{GMM}$ solves the optimization problem

$$\min_{\theta\in\Theta} \left(\frac{1}{T} \sum_{t=1}^{T} m\left(w_t, \theta\right)\right)' \left(\frac{1}{T} \sum_{t=1}^{T} \bar{m}_t \bar{m}_t^{\sigma\prime}\right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} m\left(w_t, \theta\right)\right). \tag{2.2.2}$$

[1] It is well known that the Hansen–Hodrick estimator of the variance is not necessarily positive definite. When it is not, the researcher is likely to discard the estimate and switch to another form of the HAC estimator (for example, Newey–West). This may well affect the second order bias of the GMM estimate since the probability of getting a non-positive definite weight matrix is $O\left(1/\sqrt{T}\right)$ which is the order of the second order bias. We ignore this effect since it makes the asymptotic bias considerations even less favorable for the GMM which we will criticize anyway on this basis in the next section.

Other variants of efficient GMM iterate the weighting matrix one more time or until convergence.

The first order asymptotics of efficient GMM estimators is (Hansen 1982)

$$\sqrt{T}\left(\hat{\theta}_{GMM}-\theta\right) \overset{A}{=} -\Xi\zeta_T. \tag{2.2.3}$$

The second order asymptotic bias will depend on the first order asymptotic variance of $\bar{\theta}$. Let $\hat{W}$ denote the weight matrix used on the preliminary step, and let $W = p\lim \hat{W}$ and $\bar{\Xi} = (Q'WQ)^{-1} Q'W$. For instance, if $\hat{W} = I$, then $\bar{\Xi} = (Q'Q)^{-1} Q'$; if $\bar{\theta}$ is asymptotically efficient, then $\bar{\Xi} = \Xi$.

Empirical likelihood (EL) estimator The baseline empirical likelihood estimator $\hat{\theta}_{EL}$ together with the $\ell \times 1$ vector of additional parameters $\hat{\lambda}_{EL}$ solves the optimization problem

$$\min_{\theta} \sup_{\lambda:\, 1+\lambda' m_t>0} \sum_{t=1}^{T} \log\left(1+\lambda' m_t\right). \tag{2.2.4}$$

The FOCs for $\hat{\theta}_{EL}$ and $\hat{\lambda}_{EL}$ of the optimization problem (2.2.4) are

$$0 = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{m}_t}{1+\hat{\lambda}_{EL}'\hat{m}_t}, \tag{2.2.5}$$

$$0 = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{m}_{\theta t}'\hat{\lambda}_{EL}}{1+\hat{\lambda}_{EL}'\hat{m}_t}. \tag{2.2.6}$$

The solution is generally inefficient when serial correlation in m_t is present. As stated in the introduction, to construct an efficient estimator when the order of the serial correlation is known, three approaches are found in the literature. One leads to a multistep estimator, the other two lead to one-step estimators and are both based on temporal weighting of the moment function.

Adjusted empirical likelihood (AEL) estimator Imbens (1997) proposed to adjust for asymptotic correlation between $\hat{\theta}_{EL}$ and $\hat{\lambda}_{EL}$ to attain asymptotic efficiency:

$$\hat{\theta}_{AEL} = \hat{\theta}_{EL} - \widehat{ACov}(\hat{\theta}_{EL},\hat{\lambda}_{EL})\left[\widehat{AVar}(\hat{\lambda}_{EL})\right]^{-1}\hat{\lambda}_{EL},$$

where $\widehat{AVar}(\hat{\lambda}_{EL})$ and $\widehat{ACov}(\hat{\theta}_{EL},\hat{\lambda}_{EL})$ are consistent estimates of the blocks of the asymptotic variance matrix of $(\hat{\theta}_{EL}',\hat{\lambda}_{EL}')'$. This approach is straightforward in implementation, but it results in a multistep estimator losing attractiveness of the whole EL spproach. Therefore, we will not analyze this estimator any further.

Corrected empirical likelihood (CEL) estimator This approach (Back and Brown 1990, Imbens 1997) suggests summing the moment function over $2q+1$ periods in the denominators of the FOC (2.2.5)–(2.2.6). Define $m_t^{\sigma} = \sum_{s=-q}^{q} m_{t-s}$ and $m_{\theta t}^{\sigma} = \sum_{s=-q}^{q} m_{\theta,t-s}$. The FOCs are modified in the following way[2]:

$$0 = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{m}_t}{1+\hat{\lambda}_{CEL}'\hat{m}_t^{\sigma}} \tag{2.2.7}$$

$$0 = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{m}_{\theta t}'\hat{\lambda}_{CEL}}{1+\hat{\lambda}_{CEL}'\hat{m}_t^{\sigma}} \tag{2.2.8}$$

[2] Our convention is that if an index of some component of a summand is beyond the sample limits, the entire summand is dropped.

We will refer to the solution $\hat{\theta}_{CEL}$ of this system as the CEL estimator. This approach to modifying the moment function is convenient since the order of serial correlation q is known, so there is no problem of selecting a truncation laglength. One serious drawback of this modification is that the system (2.2.7)–(2.2.8) does not correspond to an optimization problem like (2.2.5)–(2.2.6). In sections 3 and in simulations we discover a relatively large second order asymptotic bias of $\hat{\theta}_{CEL}$ compared to the alternative estimator $\hat{\theta}_{SEL}$ (see below).

The first order asymptotics of the CEL estimator is

$$\sqrt{T}\left(\hat{\theta}_{CEL}-\theta\right) \overset{A}{=} -\Xi\zeta_T, \quad \sqrt{T}\hat{\lambda}_{CEL} \overset{A}{=} \Omega\zeta_T, \tag{2.2.9}$$

and $\hat{\theta}$ and $\hat{\lambda}$ are first order asymptotically independent (Imbens 1997).

Smoothed empirical likelihood (SEL) estimator This approach (Kitamura 1997, Smith 1997, 1998, 2000) suggests smoothing the moment function from the outset. Let us choose a kernel with the following properties:

Assumption 4 *The kernel function $k(x)$ satisfies:*

(a) $k(x) : [-b, +b] \to [-1, +1]$ for some finite b;

(b) $k(x) = k(-x)$ for all $x \in [-b, +b]$;

(c) $k(x)$ is continuous at 0 and at all but a finite number of points;

(d) $\int_{-b}^{+b} k(x)dx = 1$.

A variety of popular kernels satisfy assumption 4: truncated, Bartlett, Parzen, Tukey–Hanning. Let us denote

$$\rho_2 = \int_{-b}^{+b} k(x)^2 dx, \quad \rho_3 = \int_{-b}^{+b} k(x)^3 dx.$$

Define the system of weights $\kappa(s) = \delta_T^{-1} k\left(\delta_T^{-1} s\right)$, where δ_T is the bandwidth parameter tending to infinity more slowly than the sample size. To derive the results related to asymptotic bias we require even slower rate:

Assumption 5 *$\delta_T \to \infty$ and $\delta_T = o\left(T^{\frac{1}{5}}\right)$ as $T \to \infty$.*

The moment function is smoothed with the system of weights $\kappa(s)$. Define

$$m_t^{\kappa} = \sum_{s=-r_T}^{r_T} \kappa(s) m_{t-s} \text{ and } m_{\theta t}^{\kappa} = \sum_{s=-r_T}^{r_T} \kappa(s) m_{\theta, t-s},$$

where $r_T = \lfloor \delta_T b \rfloor$. The FOCs are modified in the following way:

$$0 = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{m}_t^{\kappa}}{1+\hat{\lambda}'_{SEL}\hat{m}_t^{\kappa}} \tag{2.2.10}$$

$$0 = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{m}_{\theta t}^{\kappa\prime}\hat{\lambda}_{SEL}}{1+\hat{\lambda}'_{SEL}\hat{m}_t^{\kappa}} \tag{2.2.11}$$

We will refer to the solution $\hat{\theta}_{SEL}$ of this system as the SEL estimator. The following two kernels associated with the original kernel $k(x)$ play an important role in the subsequent asymptotic analysis of $\hat{\theta}_{SEL}$. The first one connected with the first order asymptotics is the *induced kernel* (Smith 1998) proportional to the self-convolution of $k(x)$:

$$k^*(x) = \rho_2^{-1} \int_{-b}^{+b} k\left(x+y\right) k\left(y\right) dy,$$

so that $k^*(0) = 1$. The corresponding system of weights $\kappa^*(s) = \delta_T^{-1} k^*\left(\delta_T^{-1} s\right)$ is such that

$$\frac{\delta_T}{T} \sum_{t=1}^{T} \sum_{s=-2r_T}^{2r_T} \kappa^*(s) m_t m_{t-s}'$$

is a consistent and positive definite estimator of V (Smith 1998). The second associated kernel connected with the second order asymptotics is proportional to the double self-convolution of $k(x)$

$$k^{**}\left(y,z\right) = \rho_3^{-1} \int_{-b}^{+b} k(x)k(x+y)k(x+z)dx.$$

It is called the *bispectral estimating kernel* (Rosenblatt and Van Ness 1965), and is symmetric, continuous in both arguments at $(0,0)$ and normalized so that $k^{**}\left(0,0\right) = 1$. The bispectral estimating kernel plays an important role in estimation of bispectra in the spectral analysis. Here it is needed for consistent estimation of third moments of the moment function.

The approach under discussion to modifying the moment function is less convenient than the previous one since generally there does not exist a kernel $\kappa(s)$ whose corresponding induced kernel $\kappa^*(s)$ is flat over $-q, \cdots, q$. The can be easily seen by noting that $\kappa^*(0) \propto \sum_{s=-r_T}^{r_T} \kappa(s)^2$, $\kappa^*(1) \propto \sum_{s=-r_T}^{r_T-1} \kappa(s+1)\kappa(s)$, and due to the Cauchy–Schwartz inequality,

$$\sum_{s=-r_T}^{r_T-1} \kappa(s+1)\kappa(s) \leq \left(\sum_{s=-r_T}^{r_T-1} \kappa(s)^2\right)^{\frac{1}{2}} \left(\sum_{s=-r_T}^{r_T-1} \kappa(s+1)^2\right)^{\frac{1}{2}} < \sum_{s=-r_T}^{r_T} \kappa(s)^2.$$

Even when $k(x)$ has unbounded support, the equality can occur only with flat $k(x)$ which is impossible. Therefore, one cannot take advantage of the special correlation structure of the problem, and instead is forced to act as if the correlation structure was of unknown form. On the positive side, the system (2.2.10)–(2.2.11) does correspond to an optimization problem

$$\min_{\theta} \sup_{\lambda:\, 1+\lambda' m_t^{\kappa}>0} \sum_{t=1}^{T} \log\left(1+\lambda' m_t^{\kappa}\right). \tag{2.2.12}$$

In addition, in sections 3 and 4 we discover a relatively small second order asymptotic bias of $\hat{\theta}_{SEL}$ compared to that of $\hat{\theta}_{CEL}$. The asymptotic bias considerations also provide an extra guide, in addition to usual ones, to selecting the kernel $k(x)$.

The first-order asymptotics of the SEL estimator is

$$\sqrt{T}\left(\hat{\theta}_{SEL} - \theta\right) \overset{A}{=} -\Xi\zeta_T, \quad \rho_2\sqrt{T}\delta_T^{-1}\hat{\lambda}_{SEL} \overset{A}{=} \Omega\zeta_T, \tag{2.2.13}$$

and $\hat{\theta}$ and $\hat{\lambda}$ are first-order asymptotically independent (e.g., Smith 2000). To be more precise, we have

Lemma 3 *Under assumptions 1–4,* $\sqrt{T}\left(\hat{\theta}_{SEL}-\theta\right) = -\Xi\zeta_T + O_p\left(\sqrt{\frac{\delta_T}{T}}\right)$, $\rho_2\sqrt{T}\delta_T^{-1}\hat{\lambda}_{SEL} = \Omega\zeta_T + O_p\left(\sqrt{\frac{\delta_T}{T}}\right)$, *and* $E\left[\sqrt{T}\left(\hat{\theta}_{SEL}-\theta\right)'\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda}_{SEL}\right] = O_p\left(\sqrt{\frac{\delta_T}{T}}\right)$.

2.3 Second order asymptotic bias of GMM and EL estimators

The analysis of second order bias is important. On the one hand, it allows one to evaluate how big it is, and discriminate between alternative first order asymptotically equivalent estimators. On the other hand, its knowledge permits one to construct analytical bias correction, as a convenient alternative to computationally involved bootstrap (Efron and Tibshirani 1993) and jackknife (Angrist, Imbens and Krueger 1999) bias corrections.

As argued by Rothenberg (1984), the expectation of the higher order term in the expansion of an estimator of interest can be viewed as an approximate value of the bias of the estimate in a finite sample, even when the corresponding order moments do not exist. In the Appendix we derive the second order asymptotic biases of the GMM, CEL and SEL estimators when assumptions 1–5 hold. Denote the j^{th} column of the identity matrix by e_j. The second order bias of $\hat{\theta}_{GMM}$ is (omitting the order factor $1/T$)

$$\begin{aligned} B_{GMM} &= \Xi\sum_{s=-\infty}^{+\infty}E\left[m_t m_t^{\sigma\prime}\Omega m_{t-s}\right] - \Sigma\sum_{s=-\infty}^{+\infty}E\left[m_{\theta t}'\Omega m_{t-s}\right] \\ &\quad +\Xi\sum_{j=1}^{k}E\left[\frac{\partial m_t m_t^{\sigma\prime}}{\partial\theta_j}\Omega V\bar{\Xi}' e_j\right] \\ &\quad +\Xi\left(\sum_{s=-\infty}^{+\infty}E\left[m_{\theta t}\Xi m_{t-s}\right] - E\left[\sum_{j=1}^{k}\frac{\partial m_{\theta t}}{\partial\theta_j}\frac{\Sigma}{2}e_j\right]\right). \end{aligned} \tag{2.3.14}$$

Note that if the first step estimator $\bar{\theta}$ comes from the efficient GMM (for example, if the weighting matrix is iterated one more time or until convergence), then $\bar{\Xi}=\Xi$, $\Omega V\bar{\Xi}'=0$, and the third term vanishes. From now on we will presume that $\bar{\theta}$ is efficient.

The second order bias of $\hat{\theta}_{CEL}$ is (omitting the order factor $1/T$)

$$\begin{aligned} B_{CEL} &= \Xi\sum_{|s|>q}E\left[m_t m_t^{\sigma\prime}\Omega m_{t-s}\right] - \Sigma\sum_{|s|>q}E\left[m_{\theta t}'\Omega m_{t-s}\right] \\ &\quad +\Xi\left(\sum_{s=-\infty}^{+\infty}E\left[m_{\theta t}\Xi m_{t-s}\right] - E\left[\sum_{j=1}^{k}\frac{\partial m_{\theta t}}{\partial\theta_j}\frac{\Sigma}{2}e_j\right]\right). \end{aligned} \tag{2.3.15}$$

The second order bias of $\hat{\theta}_{SEL}$ is (omitting the order factor $1/T$)

$$\begin{aligned} B_{SEL} &= \left(1-\frac{\rho_3}{\rho_2^2}\right)\Xi\sum_{s_1=-\infty}^{+\infty}\sum_{s_2=-\infty}^{+\infty}E\left[m_t m_{t-s_1}'\Omega m_{t-s_2}\right] \\ &\quad +\Xi\left(\sum_{s=-\infty}^{+\infty}E\left[m_{\theta t}\Xi m_{t-s}\right] - E\left[\sum_{j=1}^{k}\frac{\partial m_{\theta t}}{\partial\theta_j}\frac{\Sigma}{2}e_j\right]\right). \end{aligned} \tag{2.3.16}$$

There are two common terms in the three formulas which are present even under exact identification. Apart from those, the biases of $\hat{\theta}_{GMM}$ and $\hat{\theta}_{CEL}$ involve third moments of the moment function and covariances of the score and the moment function. There are "more" (summation for all lags and leads vs. summation for lags and leads separated by more than q periods) such terms of both types in $\hat{\theta}_{GMM}$ than in $\hat{\theta}_{CEL}$. Since usually leading terms (i.e. that correspond to small $|s|$) are larger than the tails of infinite sums, this means that the first two terms in B_{GMM} are likely to much exceed thos in B_{CEL}. The bias of $\hat{\theta}_{SEL}$ has even "more" (double summation for all lags and leads vs. one finite summation and one infinite) terms resulting from third moments. However, the double sum is multiplied by the factor $\left(1-\rho_3\rho_2^{-2}\right)$ which depends only on the choice of the kernel and may be manipulated with. In addition, the bias of $\hat{\theta}_{SEL}$ does not have a component associated with covariances of the score and the moment function. Thus, in situations with a nontrivial dependence the bias of $\hat{\theta}_{GMM}$ is likely to exceed the bias of $\hat{\theta}_{CEL}$, which is likely to exceed the bias of $\hat{\theta}_{SEL}$. This is confirmed in our simulations reported in the next section. Below, we analyze the formulas (2.3.14)–(2.3.16) more thoroughly.

Observation 1: order factors All expressions (2.3.14)–(2.3.16) figure into the bias of associated estimates of θ with the factor $1/T$. This may seem surpising for the SEL estimator, as smoothing may be expected to slow down the convergence rate. Nonetheless, even though the smoothing does affect factors of the higher order expansion for $\hat{\lambda}_{SEL}$, it does not for $\hat{\theta}_{SEL}$.

Observation 2: exact identification When the system of moment conditions is exactly identifying, Ω is zero and thus

$$B_{GMM}=B_{CEL}=B_{SEL}=\Xi\left(\sum_{s=-\infty}^{+\infty}E\left[m_{\theta t}\Xi m_{t-s}\right]-E\left[\sum_{j=1}^{k}\frac{\partial m_{\theta t}}{\partial\theta_j}\frac{\Sigma}{2}e_j\right]\right).$$

Observation 3: kernel smoother The first term in B_{SEL} may be removed by a judicious choice of the kernel function that satisfies $\rho_3\rho_2^{-2}=1$. For the truncated kernel

$$k(x)=\frac{1}{2},\ |x|\le 1, \tag{2.3.17}$$

we have $\rho_2=\frac{1}{2}$, $\rho_3=\frac{1}{4}$, so the condition $\rho_3\rho_2^{-2}=1$ is satisfied. For the Bartlett kernel

$$k(x)=1-|x|,\ |x|\le 1, \tag{2.3.18}$$

we have $\rho_2=\frac{2}{3}$, $\rho_3=\frac{1}{2}$, so the condition $\rho_3\rho_2^{-2}=1$ is violated. For the Parzen kernel

$$k(x)=\begin{cases}1-6x^2+6|x|^3, & |x|\le\frac{1}{2},\\ 2\left(1-|x|\right)^3 & \frac{1}{2}\le|x|\le 1,\end{cases} \tag{2.3.19}$$

we have $\rho_2=\frac{151}{280}$, $\rho_3=\frac{1979}{4480}$, so the condition $\rho_3\rho_2^{-2}=1$ is violated. For the Tukey–Hanning kernel

$$k(x)=\frac{1+\cos(\pi x)}{2},\ |x|\le 1, \tag{2.3.20}$$

we have $\rho_2=\frac{3}{4}$, $\rho_3=\frac{5}{8}$, so the condition $\rho_3\rho_2^{-2}=1$ is violated. If fact, if only positive kernels are considered, the only kernel that satisfies $\rho_3\rho_2^{-2}=1$ is (2.3.17) due to the Cauchy–Schwartz inequality $\left(\int k(x)^2dx\right)^2 \le \int k(x)dx \cdot \int k(x)^3dx$. It is this smoother that was originally proposed in Kitamura (1997) and Kitamura and Stutzer (1997), although the motivation was simplicity.

Not necessarily removing the term under consideration is good. Instead, it can be set to a target value to offset other bias components, although hitting the target does not seem plausible in practice. Typically the difference $|1-\rho_3\rho_2^{-2}|$ is pretty small (it equals $\frac{1}{8}$ for the Bartlett kernel, $\frac{23663}{45602}\approx .52$ for the Parzen kernel, $\frac{1}{9}$ for the Tukey–Hanning kernel).

Observation 4: scale parameter Suppose the moment restrictions identify only an additive scale parameter. Then $m_{\theta t}$ is constant, $\partial m_{\theta t}/\partial\theta_j$ is zero and

$$\begin{aligned}
B_{CEL} &= \Xi \sum_{|s|>q} E\left[m_t m_t^{\sigma\prime}\Omega m_{t-s}\right], \\
B_{SEL} &= \left(1-\frac{\rho_3}{\rho_2^2}\right)\Xi \sum_{s_1=-\infty}^{+\infty}\sum_{s_2=-\infty}^{+\infty} E\left[m_t m_{t-s_1}'\Omega m_{t-s_2}\right].
\end{aligned}$$

In this situation B_{SEL} may be set to zero by choosing a kernel for which $\rho_3\rho_2^{-2}=1$, while B_{CEL} is not necessarily zero.

Observation 5: number of moment restrictions Newey and Smith (2000) establish that in cross sectional models estimated by instrumental variables, the GMM type estimators have a bias component that grows linearly with the number of instruments. This component corresponds to the second term in the formula for B_{CEL} which is absent from the bias for the EL estimator in the IID context and from B_{SEL}. It follows that the bias of the CEL estimator, in contrast to that of the SEL estimator, may grow with the number of moment restrictions.

Observation 6: moment function is q-dependent If w_t is q-dependent (i.e., w_t and w_{t-s} are independent for any $s>q$), we have $E\left[m_t m_t^{\sigma\prime}\Omega m_{t-s}\right]=0$ when $|s|>2q$, $E\left[m_{\theta t}\Xi m_{t-s}\right]=0$ and $E\left[m_{\theta t}'\Omega m_{t-s}\right]=0$ when $|s|>q$, so

$$\begin{aligned}
B_{GMM} &= B+\Xi\sum_{|s|\le 2q} E\left[m_t m_t^{\sigma\prime}\Omega m_{t-s}\right]-\Sigma\sum_{|s|\le q}E\left[m_{\theta t}'\Omega m_{t-s}\right], \\
B_{CEL} &= B+\Xi\sum_{q<|s|\le 2q} E\left[m_t m_t^{\sigma\prime}\Omega m_{t-s}\right], \\
B_{SEL} &= B+\left(1-\frac{\rho_3}{\rho_2^2}\right)\Xi\sum_{|s_1|\le 2q}\sum_{|s_2|\le 2q}E\left[m_t m_{t-s_1}'\Omega m_{t-s_2}\right],
\end{aligned}$$

where

$$B=\Xi\left(E\left[m_{\theta t}\Xi m_t^{\sigma}\right]-E\left[\sum_{j=1}^{k}\frac{\partial m_{\theta t}}{\partial\theta_j}\frac{\Sigma}{2}e_j\right]\right).$$

There are more summands representing the third moments of the modent condition in the formula for B_{SEL} than in those for B_{CEL} and B_{GMM}, but they all can be eliminated by the use of the truncated kernel. The formula for B_{GMM} even in this situation includes covariances between the score and the moment functions, which are now absent from B_{CEL}.

Observation 7: moment function is martingale difference Suppose that w_t has a martingale difference structure relative to own past, but is not IID across time. In this case the CEL estimator is the same as the baseline EL estimator, while the SEL estimator still employs kernel smoothing while it is not necessary. Then, if the truncated kernel is used for the SEL, we have

$$\begin{aligned} B_{CEL} &= B_{SEL} + \Xi \sum_{s \neq 0} E\left[m_t m_t^{\sigma\prime} \Omega m_{t-s}\right] - \Sigma \sum_{s \neq 0} E\left[m'_{\theta t} \Omega m_{t-s}\right], \\ B_{SEL} &= \Xi \left(\sum_{s=-\infty}^{+\infty} E\left[m_{\theta t} \Xi m_{t-s}\right] - E\left[\sum_{j=1}^{k} \frac{\partial m_{\theta t}}{\partial \theta_j} \frac{\Sigma}{2} e_j \right] \right). \end{aligned}$$

The two infinite sums in B_{CEL} are still present, while B_{SEL} does not contain them. Thus we observe a striking fact that *smoothing the moment function when smoothing is not necessary tends to reduce the bias*!

Observation 8: IID sampling If w_t is IID, $E\left[m_t m_t^{\sigma\prime} \Omega m_{t-s}\right] = 0$, $E\left[m_{\theta t} \Xi m_{t-s}\right] = 0$ and $E\left[m'_{\theta t} \Omega m_{t-s}\right] = 0$ when $s \neq 0$, so

$$\begin{aligned} B_{GMM} &= B_{CEL} + \Xi E\left[m_t m'_t \Omega m_t\right] - \Sigma E\left[m'_{\theta t} \Omega m_{t-s}\right], \\ B_{CEL} &= \Xi \left(E\left[m_{\theta t} \Xi m_t\right] - E\left[\sum_{j=1}^{k} \frac{\partial m_{\theta t}}{\partial \theta_j} \frac{\Sigma}{2} e_j \right] \right), \\ B_{SEL} &= B_{CEL} + \left(1 - \rho_3 \rho_2^{-2}\right) \Xi E\left[m_t m'_t \Omega m_t\right]. \end{aligned}$$

The additional component in B_{SEL} is due to smoothing when there is no need to smooth. Everything else coincides with formulae in Newey and Smith (2000).

2.4 Simulation evidence

2.4.1 Model and estimators

Previous simulation studies of EL type estimators occurred in Gospodinov (2002) who used the baseline EL estimator in models with martingale difference errors, and in Gregory, Lamarche and Smith (2002) who used the SEL estimator and found that it possessed higher bias than the GMM estimator. The latter finding is surprising as it is at variance with the theoretical and simulation results of the present paper.

We run simulations for the ARMA(1,1) model with GARCH(1,1) innovations. The primary reason for choosing such a model is convenience. Firstly, the right hand side variable is endogenous and correlated with the disturbance. Secondly, suitable lags of the right hand side variable can be used as instruments, as is often done in practice. Third, it is easy to set ARMA and GARCH parameters so that the time series behavior mimics that of many series found in practice. On the other hand, the empirically non-relevant assumption of an ARMA model, that the Wold innovation is a martingale difference relative to past data, is never exploited by the estimators we consider.

The model is

$$y_t = \mu + \alpha y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1} \tag{2.4.21}$$

$$\varepsilon_t = \sigma_t \eta_t, \quad \eta_t \sim IID\, \mathcal{N}(0,1),$$

$$\sigma_t^2 = (1-\gamma_1-\gamma_2) + \gamma_1 \varepsilon_{t-1}^2 + \gamma_2 \sigma_{t-1}^2, \tag{2.4.22}$$

The object of interest is α, which is estimated with several moment condition based methods. The moment condition is formed after choosing the vector of instruments

$$z_t = (1\ y_{t-2}\ \cdots\ y_{t-1-\ell})'$$

as

$$m\left(y_t\ y_{t-1}\ y_{t-2}\ \cdots\ y_{t-1-\ell}, \mu, \alpha\right) = z_t\left(y_t - \mu - \alpha y_{t-1}\right).$$

In addition, some experiments are done with the following modification that allows higher-order serial correlation structure:

$$y_t = \mu + \alpha y_{t-p} + (1+\theta L)^p \varepsilon_t, \tag{2.4.23}$$

where p is 1 through 6. The vector of instruments is

$$z_t = \left(1\ y_{t-p-1}\ \cdots\ y_{t-p-\ell}\right)',$$

and the right hand side variable in (2.4.23) is put to y_{t-p} in order not to reduce the relevance of the instruments as p increases. To approximately match the time series properties of the variable y_t to real data, we set $\alpha = 0.5$, $\theta = -0.8$, $\gamma_1 = 0.1$, $\gamma_2 = 0.8$, which implies the unconditional kurtosis equal 3.35.

The simulation experiments are performed in Gauss. Each experiment involves $5,000$ Monte Carlo repetitions for sample sizes of 300 and 900. The estimators we compare are: efficient GMM, CEL and SEL. In computing the SEL estimates we use the truncated, Bartlett, Parzen and Tukey–Hanning kernels (2.3.17)–(2.3.20) and investigate the influence of the bandwidth r_T choice. When selecting the instrumental vectors, we set ℓ to 1 through 8 for $T = 300$, and to 1 through 10 for $T = 900$. The performance of the estimators is evaluated through comparison of the means and root mean squared errors of the trimmed (see the following subsection) arrays of simulated estimates.

2.4.2 Details of simulation and estimation

In each repetition we discard $1,000$ "presampling" values that start from $y_{-999} = 1$, $\sigma^2_{-999} = 1$, $\varepsilon_{-999} = \eta_{-999}$. For the GMM estimates we use the following formulae. Let Y be the matrix of regressands y_t, X be the matrix of regressors $x_t = (1\ y_{t-1})'$, and Z be the matrix of instruments z_t, then

$$(\hat{\mu}_{2SLS}\ \widehat{\alpha}_{2SLS})' = \left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1} X'Z\left(Z'Z\right)^{-1}Z'Y$$

and

$$(\hat{\mu}_{GMM}\ \widehat{\alpha}_{GMM})' = \left(X'Z\hat{W}Z'X\right)^{-1} X'Z\hat{W}Z'Y$$

with the weight matrix $\hat{W} = \left(\hat{Q}_0 + \hat{Q}_1 + \hat{Q}_{-1}\right)^{-1}$, where $\hat{Q}_0 = \sum_{t=\ell+2}^{T} z_t z_t' \hat{e}_t^2$, $\hat{Q}_1 = \hat{Q}'_{-1} = \sum_{t=\ell+3}^{T} z_t z'_{t-1} \hat{e}_t \hat{e}_{t-1}$, $\widehat{e}_t = y_t - \hat{\mu}_{2SLS} - \widehat{\alpha}_{2SLS} y_{t-1}$, if it is positive definite, and $\hat{W} = \hat{Q}_0^{-1}$

otherwise. To find CEL and SEL estimates we solve respectively the nonlinear systems (2.2.7)–(2.2.8) where t runs from $\ell + 4$ to $T - 1$ and (2.2.10)–(2.2.11) where t runs from $\ell + 3 + r_T$ to $T - r_T$, using the `EqnSolve` subroutine from the library `EqnSolve.src`. If `EqnSolve` fails to find the solution we set $\widehat{\alpha}_{CEL} = \widehat{\alpha}_{GMM}$ or $\widehat{\alpha}_{SEL} = \widehat{\alpha}_{GMM}$, whichever applicable (the percentage of such failures is very low).

For smaller T a certain proportion of estimates lie outside of $[-1, 1]$, so some trimming scheme is called for. According to one scheme, estimates are corrected by trimming near the boundaries of the permissible region, which leads to pile-ups of estimates that give wrong impression of estimate distributions. Another scheme used in the literature eliminates tails by cutting off certain percentages of extreme observations. We use a slightly different procedure: we exclude the estimators that lie outside of $[-1, 1]$, which is natural to do in the context of a stationary autoregressive model. In any case, the percentage of trimmings is very low (almost zero when $T = 900$).

2.4.3 Results

The following first four comments are based on Figure 1, the other two – on Figures 2 and 3. Figure 1 depict the mean and RMSE of trimmed GMM, CEL, and SEL (with the truncated kernel and lag truncation parameter r_T equal 3) estimates against the number of instruments in the instrumental vector.

Traditional vs EL-based estimates The GMM estimator tends to be less volatile than the CEL and SEL, at least for smaller sample sizes, but the difference in bias between the GMM and SEL is much more pronounced, which is however not true when the GMM and CEL are compared. As a result, the SEL estimator has a much smaller MSE than the GMM or CEL, provided that the number of instruments is large enough.

Central tendencies Consider the bias properties of the estimators represented by plots of the mean (TM) of trimmed estimates. The bias of the GMM estimator is big, and clearly increases rapidly when the number of instruments rises. In contrast, the bias of the SEL estimator does not have this drawback, exhibiting only slight increases. The CEL estimator lies in between, the rise in bias being somewhet slower than that of GMM, but very pronounced. This is a consequence of the presence of the last term in B_{CEL} (see observation 5 in section 3).

Variance and MSE The favorable bias properties directly pass over to the MSE, and a preferable estimator in terms of bias is usually preferable in terms of MSE, even though it may be more volatile. In many cases the SEL estimator has greater variance than the GMM, but thanks to its small bias the SEL estimator is always better in terms of MSE for ℓ big enough. In fact, the TRMSE plots for the SEL tend to lie below those for the other estimators. For the CEL estimator, in contrast, a small advantage in bias over the GMM is not a large enough compensation for the variance gain.

Optimal utilization of instruments The plots of TRMSE against the number of instruments are U-shaped for all estimators reflecting the tradeoff between bias and variance. A noticeable feature of the SEL estimator resulting from the bias behavior is that the TRMSE is practically constant over a range of values of ℓ. For example, equal TRMSEs result from setting

ℓ to 6 through 8 for $T = 300$, and to 7 through 10 for $T = 900$. The other estimators tend to have a unique minimizer of the TRMSE. This minimizer is typically smaller than the optimal range of values for the SEL estimator. This means that with the SEL it is possible to exploit more information in the history of the process, without worrying about the exact location of the optimal ℓ and fearing of overshooting it.

Kernel and bandwidth for SEL Figure 2 depict the mean and RMSE of trimmed SEL estimates against the bandwidth r_T for the truncated, Bartlett, Parzen and Tukey–Hanning kernels (note that the Bartlett/Parzen/Tukey–Hanning kernel and $r_T = 1$ yield the baseline EL estimator), with the minimal number of instruments ℓ yielding superior results according to previous evidence. It is easily seen that in bigger samples the statistical properties of estimates are very close for different kernels, provided that the bandwidth is big enough. The truncated kernel yields uniformly less biased estimates that the other kernels as our theory predicts. With the truncated kernel, for $T = 300$ the bias is less pronounced the smaller the bandwidth, and for $T = 900$ it is quite insensitive to the bandwidth choice. The "optimal" value of r_T with the use of the truncated kernel equals 3 for both sample sizes.

Higher order serial correlation Figure 3 depicts the mean and RMSE of trimmed SEL estimates against the order p of the MA part in (2.4.23) for several bandwidths r_T for the truncated kernel (which showed to advantage before), for the medium sample size $T = 300$ with the number of instruments ℓ equal 6, an "optimal" number according to previous evidence. It is easily seen that when the order of serial correlation exceeds one, the statistical properties of the SEL estimator become even less sensitive to a choice of the bandwidth.

2.5 Conclusion

In this paper we focus attention on precision of estimates and do not consider hypothesis testing. Intuitively, favorable properties of the SEL estimator should carry over to the size properties of the overidentification and other tests. This, together with an analysis of power properties, deserves close attention. Another direction of research is analysis of stationary time series models with serial correlation of infinite or unknown order. The SEL estimator will be the same, but the CEL estimator will have to be adapted to this feature. Intuitively, the SEL estimator will still have more favorable finite sample properties.

An obvious interesting extension is the analysis of even higher order terms in the stochastic expansions of the estimators similar to what Newey and Smith (2001) do in the cross-sectional context. This will require much more tedious computations but in reward will yield the expressions for the higher order asymptotic variance and, as applied to the SEL, allow one to pin down the optimal rate of convergence of the bandwidth and rules of choosing it in practice, in the same manner as in Andrews (1991).

Finally, other members of the GEL class may be also considered. We conjecture that correspondingly modified GEL estimators will inherit the bias properties of the modified EL estimators. At the same time, some new features may arise since the bias of GEL estimators generally involves more components than that of the EL estimator.

2.6 Appendix

Let

$$\Delta_{\partial m} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (m_{\theta t} - Q), \quad \Delta_{mm} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(m_t m_t^{\sigma \prime} - V\right).$$

2.6.1 Preliminary lemmas

Lemma 4 *If x_t is mean zero strictly stationary and strong mixing process with mixing coefficients α_j, and $E\left[|x_t|^{3\nu}\right] < \infty$ for some $\nu > 1$, then for $i, j > 0$*

$$|E[x_t x_{t+j}]| \leq 8\alpha_j^{1-1/\nu} \left(E\left[|x_t|^{2\nu}\right]\right)^{1/\nu}$$

and

$$|E[x_t x_{t+i} x_{t+j}]| \leq 8\alpha_{\max\{i,j\}}^{1-1/\nu} \left(E\left[|x_t|^{3\nu}\right]\right)^{1/\nu}.$$

P roof. The mixing inequality (Hall and Heyde 1980, Corollary A.2) together with Hölder's inequality implies the result. ■

Lemma 5 *Under assumptions 1–3, the following is true:*

(a) $E[\Delta_{\partial m} \Xi \zeta_T] = \sum_{s=-\infty}^{+\infty} E[m_{\theta t} \Xi m_{t-s}] + o(1)$

(b) $E[\Delta'_{\partial m} \Omega \zeta_T] = \sum_{s=-\infty}^{+\infty} E[m'_{\theta t} \Omega m_{t-s}] + o(1)$

(c) $E[\Delta_{mm} \Omega \zeta_T] = \sum_{s=-\infty}^{+\infty} E[m_t m_t^{\sigma \prime} \Omega m_{t-s}] + o(1)$

(d) $E\left[\Xi \zeta_T (\Xi \zeta_T)'\right] = \Sigma + o(1)$

(e) $E\left[\Omega \zeta_T (\Xi \zeta_T)'\right] = o(1)$

(f) $E\left[\Omega \zeta_T \left(\bar{\Xi} \zeta_T\right)'\right] = \Omega V \bar{\Xi}' + o(1)$

(g) $E\left[\Omega \zeta_T (\Omega \zeta_T)'\right] = \Omega + o(1)$

P roof. (a)

$$\begin{aligned} E[\Delta_{\partial m} \Xi \zeta_T] &= T^{-1} E\left[\sum_{t=1}^{T} m_{\theta t} \Xi \sum_{\tau=1}^{T} m_\tau\right] = T^{-1} \sum_{t=1}^{T} \sum_{\tau=1}^{T} E[m_{\theta t} \Xi m_\tau] = \\ &= \sum_{s=-(T-1)}^{T-1} \left(1 - \frac{|s|}{T}\right) E[m_{\theta t} \Xi m_{t-s}] = \sum_{s=-\infty}^{+\infty} E[m_{\theta t} \Xi m_{t-s}] + o(1) \end{aligned}$$

by the Toeplitz lemma. Other results can be obtained in a similar way, noting that $\Xi V \Xi' = \Sigma$, $\Omega V \Xi' = 0$ and $\Omega V \Omega' = \Omega$. ■

Lemma 6 *Under assumptions 1–4,* $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} m_t^{\kappa} = \zeta_T + \Psi_T$*, where* $\Psi_T = O_p\left(\sqrt{\frac{\delta_T}{T}}\right)$ *and* $E\left[\Psi_T\right] = 0$.

P roof. Using the symmetry of $\kappa(s)$,

$$\sum_{t=1}^{T} m_t^{\kappa} = \sum_{t=1+2r_T}^{T-2r_T} m_t + \left(m_{2r_T} + m_{T-2r_T-1}\right)\sum_{s=-r_T}^{r_T-1}\kappa(s) + \cdots + \left(m_1 + m_T\right)\kappa(r_T),$$

so

$$\begin{aligned} \sqrt{T}\Psi_T &= \sum_{t=1}^{T} m_t^{\kappa} - \sum_{t=1}^{T} m_t = \left(m_{2r_T} + m_{T-2r_T-1}\right)\kappa(r_T) \\ &\quad + \left(m_{2r_T} + m_{T-2r_T-1}\right)\left(\kappa(r_T) + \kappa(r_T-1)\right) + \cdots + \left(m_1 + m_T\right)\sum_{s=-r_T+1}^{r_T}\kappa(s), \end{aligned}$$

with $E\left[\sqrt{T}\Psi_T\right] = 0$ and $V\left[\sqrt{T}\Psi_T\right] = O_p\left(\delta_T\right)$ because the number of variances and covariances is at most proportional to r_T. ■

Lemma 7 *Under assumptions 1–5, the following is true:*

(a) $$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(m_{\theta t}^{\kappa} - Q\right) = \Delta_{\partial m} + O_p\left(\frac{\delta_T}{\sqrt{T}}\right)$$

(b) $$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial m_{\theta t}^{\kappa}}{\partial\theta_j} = E\left[\frac{\partial m_{\theta t}}{\partial\theta_j}\right] + O_p\left(\frac{\delta_T}{T} + \frac{1}{\sqrt{T}}\right)$$

(c) $$\frac{1}{T}\sum_{t=1}^{T} m_t^{\kappa} m_t^{\kappa\prime} = \rho_2\delta_T^{-1}V + O_p\left(\frac{\delta_T^2}{T}\right) \text{ and}$$
$$E\left[\frac{\delta_T}{\sqrt{T}}\sum_{t=1}^{T} m_t^{\kappa} m_t^{\kappa\prime}\Omega\zeta_T\right] = \rho_2\sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty} E\left[m_t m_{t-u}'\Omega m_{t-v}\right] + O\left(\frac{\delta_T^2}{T}\right)$$

(d) $$\frac{1}{T}\sum_{t=1}^{T}\left(m_{\theta t}^{\kappa} m_{it}^{\kappa} + m_t^{\kappa} m_{\theta it}^{\kappa}\right) = \rho_2\delta_T^{-1}\sum_{u=-\infty}^{+\infty} E\left[m_{\theta t} m_{i,t-u} + m_t m_{\theta i,t-u}\right] + O_p\left(\frac{\delta_T^2}{T}\right) \text{ and}$$
$$\frac{1}{T}\sum_{t=1}^{T}\left(m_{\theta t}^{\kappa\prime} m_{it}^{\kappa} + m_{\theta t}^{\kappa\prime} e_i m_t^{\kappa\prime}\right) = \rho_2\delta_T^{-1}\sum_{u=-\infty}^{+\infty} E\left[m_{\theta t}' m_{i,t-u} + m_{\theta t}' e_i m_{t-u}'\right] + O_p\left(\frac{\delta_T^2}{T}\right),$$
where $i = 1, \cdots, \ell$

(e) $$\frac{1}{T}\sum_{t=1}^{T} m_{it}^{\kappa} m_t^{\kappa} m_t^{\kappa\prime} = \rho_3\delta_T^{-2}\sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty} E\left[m_{it} m_{t-u} m_{t-v}'\right] + O_p\left(\frac{\delta_T^3}{T}\right)$$

P roof.

(a)

$$\sum_{t=1}^{T} m_{\theta t}^{\kappa} = \sum_{t=1+2r_T}^{T-2r_T} m_{\theta t} + (m_{\theta,2r_T} + m_{\theta,T-2r_T-1}) \sum_{s=-r_T}^{r_T-1} \kappa(s) + \cdots + (m_{\theta,1} + m_{\theta,T})\, \kappa(r_T),$$

so

$$\left\| \sum_{t=1}^{T} m_{\theta t}^{\kappa} - \sum_{t=1}^{T} m_{\theta t} \right\| \leq 2r_T \left\| m_{\theta t} \right\| + (2r_T+1) \cdot 2 \left\| m_{\theta t} \right\| \sum_{s=-r_T}^{r_T} |\kappa(s)| = O_p\left(\delta_T\right).$$

(b) Similarly to (a),

$$\sum_{t=1}^{T} \frac{\partial m_{\theta t}^{\kappa}}{\partial \theta_j} = \sum_{t=1}^{T} \frac{\partial m_{\theta t}}{\partial \theta_j} + O_p\left(\delta_T\right).$$

But by the Law of Large Numbers and Central Limit Theorem for α-mixing sequences,

$$\frac{1}{T} \sum_{t=1}^{T} \frac{\partial m_{\theta t}}{\partial \theta_j} = E\left[\frac{\partial m_{\theta t}}{\partial \theta_j}\right] + O_p\left(\frac{1}{\sqrt{T}}\right).$$

The desired result follows.

(c)

$$\begin{aligned}
\sum_{t=1}^{T} m_t^{\kappa} m_t^{\kappa\prime} &= \sum_{t=1}^{T} \sum_{s=-r_T}^{r_T} \kappa(s) m_{t-s} \sum_{v=-r_T}^{r_T} \kappa(v) m_{t-v}' \\
&= \sum_{t=1}^{T} \left(\sum_{s} \sum_{v} \kappa(s) m_{t-s} \kappa(v) m_{t-v}' \right) + O_p\left(\delta_T^2\right) \\
&= \sum_{t=1}^{T} \left(\sum_{s} \sum_{u} \kappa(s) \kappa(s+u) m_t m_{t-u}' \right) + O_p\left(\delta_T^2\right) \\
&= \rho_2 \sum_{t=1}^{T} \sum_{u} \kappa^*(u) m_t m_{t-u}' + O_p\left(\delta_T^2\right).
\end{aligned}$$

Lemma 4 and assumption 1 imply that $\sum_u |u|^2 \left\| E\left[m_t m_{t-u}'\right] \right\| < \infty$. Let the characteristic exponent (Parzen 1957) of $k^*(x)$ be ϱ, then

$$\frac{\delta_T}{T} \sum_{t=1}^{T} \sum_{u} \kappa^*(u) m_t m_{t-u}' = V + \begin{cases} O_p\left(\delta_T^{-\varrho} + \sqrt{\delta_T/T}\right) & \text{if } \varrho \leq 2 \\ o_p\left(\delta_T^{-2}\right) + O_p\left(\sqrt{\delta_T/T}\right) & \text{if } \varrho > 2 \end{cases}$$

(Parzen 1957, theorem 5), so the first statement holds. Using lemma 4 and assumption 1,

$$E\left[\rho_2^{-1} \sum_{t=1}^{T} m_t^{\kappa} m_t^{\kappa\prime} \sum_{t=1}^{T} m_{it}\right] =$$

$$= E\left[\left(\sum_{t=1}^{T}\sum_{u}\kappa^*(u)m_t m'_{t-u} + O_p\left(\delta_T m_1 m'_1 + \delta_T^2 m_1 m'_{1+r_T}\right)\right)\sum_{t=1}^{T} m_{it}\right]$$

$$= \sum_u E\left[\sum_{t=1}^{T}\kappa^*(u)m_t m'_{t-u}\sum_{t=1}^{T} m_{it}\right] + O\left(\delta_T\sum_{j=1}^{T}\alpha_j^{1-1/\nu} + \delta_T^2\left(r_T\alpha_{r_T}^{1-1/\nu} + \sum_{j=2r_T}^{T}\alpha_j^{1-1/\nu}\right)\right)$$

$$= \sum_u E\left[\sum_{t=1}^{T}\left(\kappa^*(u)m_t m'_{t-u}m_{it} + \sum_s \kappa^*(u)m_t m'_{t-u}m_{i,t-s}\right)\right] + O\left(\delta_T^2\right) + O\left(\delta_T\right).$$

Lemma 4 and assumption 1 imply that $\sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty} E\left[m_t m'_{t-u} m_{i,t-v}\right] < \infty$. Then, since

$$\frac{\delta_T}{T}\sum_u E\left[\sum_{t=1}^{T}\left(\kappa^*(u)m_t m'_{t-u}m_{it} + \sum_s \kappa^*(u)m_t m'_{t-u}m_{i,t-s}\right)\right]$$
$$= \sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty} E\left[m_t m'_{t-u} m_{i,t-v}\right] + o\left(1\right),$$

the desired result follows.

(d) Similarly to (c).

(e) For any combination of indices $i, j, l = 1, \cdots, k$ we have

$$\sum_{t=1}^{T} m_{it}^{\kappa} m_{jt}^{\kappa} m_{lt}^{\kappa} = \sum_{t=1}^{T}\sum_{s=-r_T}^{r_T}\kappa(s)m_{i,t-s}\sum_{s_2=-r_T}^{r_T}\kappa(s_1)m_{j,t-s_1}\sum_{s_2=-r_T}^{r_T}\kappa(s_2)m_{l,t-s_2}$$

$$= \sum_{t=1}^{T}\sum_s\sum_u\sum_w \kappa(s)m_{i,t-s}\kappa(u)m_{j,t-u}\kappa(w)m_{l,t-w} + O_p\left(\delta_T^3\right)$$

$$= \sum_{t=1}^{T}\sum_s\sum_u\sum_v \kappa(s)\kappa(s+u)\kappa(s+v)m_{it}m_{j,t-u}m_{l,t-v} + O_p\left(\delta_T^3\right)$$

$$= \rho_3\sum_{t=1}^{T}\sum_u\sum_v \kappa^{**}(u,v)m_{it}m_{j,t-u}m_{l,t-v} + O_p\left(\delta_T^3\right),$$

where $\kappa^{**}(u,v) = \delta_T^{-2}k^{**}\left(\delta_T^{-1}u, \delta_T^{-1}v\right)$. Lemma 4 and assumption 1 imply that $\sum_u\sum_v\left(|u| + |v|\right)\left\|E\left[m_{it}m_{t-u}m'_{t-v}\right]\right\| < \infty$. Since the order ϱ of $k^{**}(x)$ is at least 1,

$$\frac{\delta_T^2}{T}\sum_{t=1}^{T}\sum_u\sum_v \kappa^{**}(u,v)m_{it}m_{t-u}m'_{t-v} = \sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty} E\left[m_{it}m_{t-u}m'_{t-v}\right] + O_p\left(\delta_T^{-1} + \delta_T/\sqrt{T}\right)$$

(Rosenblatt and Van Ness 1965, theorems 4–5), and the desired result follows. ■

Proof of lemma 3. Using lemmas 6 and 7, we have the expansion of the FOC (2.2.10)–(2.2.11) to the order $O_p\left(1/\sqrt{T}\right)$

$$O_p\left(\frac{1}{\sqrt{T}}\right) = \zeta_T + \Psi_T + Q\sqrt{T}\left(\hat{\theta} - \theta\right) - V\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda},$$
$$O_p\left(\frac{1}{\sqrt{T}}\right) = Q'\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda},$$

where the order terms follow from the higher order expansions in appendix 2.6.4, and the indexes of estimates are dropped. Premultiplying the first equation by $Q'V^{-1}$, adding the second and expressing out $\sqrt{T}\left(\hat{\theta}-\theta\right)$ and $\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda}$, we get the first two results. By taking the expectation of the product and noting that $\Xi V\Omega = 0$, we also get the third result:

$$E\left[\sqrt{T}\left(\hat{\theta}-\theta\right)'\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda}\right] = -E\left[\Xi\zeta_T\zeta_T'\Omega\right] + O_p\left(\sqrt{\frac{\delta_T}{T}}\right) = O_p\left(\sqrt{\frac{\delta_T}{T}}\right).$$

■

2.6.2 Second order asymptotic bias of GMM estimator

The first order asymptotics for the first step (possibly inefficient) estimator $\bar{\theta}$ is

$$\sqrt{T}\left(\bar{\theta}-\theta\right) = -\bar{\Xi}\zeta_T + o_p\left(1\right).$$

The second step GMM $\hat{\theta}$ has FOC

$$\left(\frac{1}{T}\sum \hat{m}_{\theta t}\right)'\left(\frac{1}{T}\sum \bar{m}_t\bar{m}_t^{\sigma\prime}\right)^{-1}\frac{1}{\sqrt{T}}\sum \hat{m}_t = 0.$$

The first order asymptotics for $\hat{\theta}$ is

$$\sqrt{T}\left(\hat{\theta}-\theta\right) = -\Xi\zeta_T + o_p\left(1\right).$$

The FOC has the expansion

$$\begin{aligned} 0 \quad = \quad & \left(\frac{1}{T}\sum m_{\theta t} + \frac{1}{\sqrt{T}}\sum_{j=1}^{k}\frac{1}{T}\sum\frac{\partial m_{\theta t}}{\partial \theta_j}\sqrt{T}\left(\hat{\theta}_j-\theta_j\right) + o_p\left(\frac{1}{\sqrt{T}}\right)\right)' \times \\ & \times\left(\frac{1}{T}\sum m_t m_t^{\sigma\prime} + \frac{1}{\sqrt{T}}\sum_{j=1}^{k}\frac{1}{T}\sum\frac{\partial m_t m_t^{\sigma\prime}}{\partial \theta_j}\sqrt{T}\left(\bar{\theta}_j-\theta_j\right) + o_p\left(\frac{1}{\sqrt{T}}\right)\right)^{-1} \times \\ & \times\left(\frac{1}{\sqrt{T}}\sum m_t + \frac{1}{T}\sum m_{\theta t}\sqrt{T}\left(\hat{\theta}-\theta\right)\right. \\ & \quad \left. + \frac{1}{2\sqrt{T}}\sum_{j=1}^{k}\frac{1}{T}\sum\frac{\partial m_{\theta t}}{\partial \theta_j}\sqrt{T}\left(\hat{\theta}_j-\theta_j\right)\sqrt{T}\left(\hat{\theta}-\theta\right) + o_p\left(\frac{1}{\sqrt{T}}\right)\right) \end{aligned}$$

or

$$\begin{aligned} 0 \quad = \quad & \left(Q + \frac{1}{\sqrt{T}}\left(\Delta_{\partial m} - \sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial \theta_j}\right]e_j'\Xi\zeta_T\right) + o_p\left(\frac{1}{\sqrt{T}}\right)\right)' V^{-1} \\ & \times\left(I - \frac{1}{\sqrt{T}}\left(\Delta_{mm} - \sum_{j=1}^{k}E\left[\frac{\partial m_t m_t^{\sigma\prime}}{\partial \theta_j}\right]e_j'\bar{\Xi}\zeta_T\right)V^{-1} + o_p\left(\frac{1}{\sqrt{T}}\right)\right) \\ & \times\left(\zeta_T + Q\sqrt{T}\left(\hat{\theta}-\theta\right) + \frac{1}{\sqrt{T}}\left(-\Delta_{\partial m}\Xi\zeta_T + \frac{1}{2}\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial \theta_j}\right]\Xi\zeta_T\Xi\zeta_T e_j\right) + o_p\left(\frac{1}{\sqrt{T}}\right)\right) \end{aligned}$$

or

$$\begin{aligned}
o_p\left(\frac{1}{\sqrt{T}}\right) &= Q'V^{-1}\left(\zeta_T + Q\sqrt{T}\left(\hat{\theta}-\theta\right)\right) \\
&\quad -\frac{1}{\sqrt{T}}Q'V^{-1}\Delta_{\partial m}\Xi\zeta_T + \frac{1}{\sqrt{T}}\Delta'_{\partial m}\Omega\zeta_T \\
&\quad -\frac{1}{\sqrt{T}}Q'V^{-1}\Delta_{mm}\Omega\zeta_T \\
&\quad +\frac{1}{2\sqrt{T}}Q'V^{-1}\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial\theta_j}\right]\Xi\zeta_T\Xi\zeta_T e_j \\
&\quad -\frac{1}{\sqrt{T}}\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial\theta_j}\right]\Omega\zeta_T e_j'\Xi\zeta_T \\
&\quad -\frac{1}{\sqrt{T}}Q'V^{-1}\sum_{j=1}^{k}E\left[\frac{\partial m_t m_t^{\sigma\prime}}{\partial\theta_j}\right]\Omega\zeta_T e_j'\bar{\Xi}\zeta_T.
\end{aligned}$$

Premultiplying by $-\Sigma$, expressing out $\sqrt{T}\left(\hat{\theta}-\theta\right)$, we find using lemma 5 that the components of the second order bias are (omitting the order factor of $1/T$) are

$$B_{m\partial m} = \Xi\sum_{s=-\infty}^{+\infty}E\left[m_{\theta t}\Xi m_{t-s}\right] - \Sigma\sum_{s=-\infty}^{+\infty}E\left[m'_{\theta t}\Omega m_{t-s}\right]$$

$$B_{mmm} = \Xi\sum_{s=-\infty}^{+\infty}E\left[m_t m_t^{\sigma\prime}\Omega m_{t-s}\right]$$

$$B_{mmm} = -\Xi\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial\theta_j}\frac{\Sigma}{2}e_j\right] + \Xi\sum_{j=1}^{k}E\left[\frac{\partial m_t m_t^{\sigma\prime}}{\partial\theta_j}\Omega V\bar{\Xi}'e_j\right]$$

Summing up all components of the bias delivers (2.3.14).

2.6.3 Second order asymptotic bias of CEL estimator

The derivatives of the summands in the FOC (2.2.7)–(2.2.8) are:

$$\begin{aligned}
\frac{\partial\dfrac{m}{1+\lambda'm^\sigma}}{\partial\theta'} &= \frac{m_\theta}{1+\lambda'm^\sigma} - \frac{m\cdot\lambda'm_\theta^\sigma}{\left(1+\lambda'm^\sigma\right)^2}, \quad \frac{\partial\dfrac{m}{1+\lambda'm^\sigma}}{\partial\lambda'} = -\frac{mm^{\sigma\prime}}{\left(1+\lambda'm^\sigma\right)^2}, \\
\frac{\partial\dfrac{m_\theta'\lambda}{1+\lambda'm^\sigma}}{\partial\theta_j} &= \frac{\dfrac{\partial m_\theta'}{\partial\theta_j}}{1+\lambda'm^\sigma}\lambda - \frac{m_\theta'\lambda\cdot\lambda'\dfrac{\partial m^\sigma}{\partial\theta_j}}{\left(1+\lambda'm^\sigma\right)^2}, \quad \frac{\partial\dfrac{m_\theta'\lambda}{1+\lambda'm^\sigma}}{\partial\lambda'} = \frac{m_\theta'}{1+\lambda'm^\sigma} - \frac{m_\theta'\lambda\cdot m^{\sigma\prime}}{\left(1+\lambda'm^\sigma\right)^2}.
\end{aligned}$$

Then we have the following expansion of the FOC:

$$o_p\left(\frac{1}{T}\right) = \frac{1}{T}\sum_{t=1}^{T}m_t + \frac{1}{T}\sum_{t=1}^{T}m_{\theta t}\cdot\left(\hat{\theta}-\theta\right) - \frac{1}{T}\sum_{t=1}^{T}m_t m_t^{\sigma\prime}\cdot\hat{\lambda}$$

$$
\begin{aligned}
& +\frac{1}{2}\sum_{j=1}^{k}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial m_{\theta t}}{\partial \theta_j}\right]\left(\hat{\theta}-\theta\right)\left(\hat{\theta}_j-\theta_j\right) \\
& -\sum_{i=1}^{\ell}\left[\frac{1}{T}\sum_{t=1}^{T}\left(m_{\theta t}m_{it}^{\sigma}+m_t m_{\theta it}^{\sigma}\right)\right]\left(\hat{\theta}-\theta\right)\hat{\lambda}_i \\
& +\sum_{i=1}^{\ell}\left[\frac{1}{T}\sum_{t=1}^{T}m_{it}^{\sigma}m_t m_t^{\sigma\prime}\right]\hat{\lambda}\hat{\lambda}_i, \\
o_p\left(\frac{1}{T}\right) \quad = \quad & \frac{1}{T}\sum_{t=1}^{T}m_{\theta t}'\cdot\hat{\lambda}+\sum_{j=1}^{k}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial m_{\theta t}'}{\partial \theta_j}\right]\hat{\lambda}\left(\hat{\theta}_j-\theta_j\right) \\
& -\frac{1}{2}\sum_{i=1}^{\ell}\left[\frac{1}{T}\sum_{t=1}^{T}\left(m_{\theta t}'m_{it}^{\sigma}+m_{\theta t}'e_i m_t^{\sigma\prime}\right)\right]\hat{\lambda}\hat{\lambda}_i,
\end{aligned}
$$

where e_i denotes the i^{th} column of the identity matrix.

Substituting (2.2.9) into the above expansion and using first order asymptotic independence of $\hat{\theta}$ and $\hat{\lambda}$, we have

$$
\begin{aligned}
o_p\left(\frac{1}{\sqrt{T}}\right) \quad = \quad & \zeta_T+Q\sqrt{T}\left(\hat{\theta}-\theta\right)-\frac{1}{\sqrt{T}}\Delta_{\partial mT}\Xi\zeta_T-V\sqrt{T}\hat{\lambda}-\frac{1}{\sqrt{T}}\Delta_{mmT}\Omega\zeta_T \\
& +\frac{1}{2}\frac{1}{\sqrt{T}}\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial \theta_j}\right]\Xi\zeta_T\zeta_T'\Xi' e_j+\frac{1}{\sqrt{T}}\sum_{i=1}^{\ell}E\left[m_{it}^{\sigma}m_t m_t^{\sigma\prime}\right]\Omega\zeta_T\zeta_T'\Omega e_i, \\
o_p\left(\frac{1}{\sqrt{T}}\right) \quad = \quad & Q'\sqrt{T}\hat{\lambda}+\frac{1}{\sqrt{T}}\Delta_{\partial mT}'\Omega\zeta_T-\frac{1}{2}\frac{1}{\sqrt{T}}\sum_{i=1}^{\ell}E\left[m_{\theta t}'m_{it}^{\sigma}+m_{\theta t}'e_i m_t^{\sigma\prime}\right]\Omega\zeta_T\zeta_T'\Omega e_i.
\end{aligned}
$$

Premultiplying the first equation by $Q'V^{-1}$, summing up the equations, and expressing out the $\sqrt{T}\left(\hat{\theta}-\theta\right)$, we get

$$
\begin{aligned}
\sqrt{T}\left(\hat{\theta}-\theta\right) \quad = \quad & -\Xi\zeta_T+\frac{1}{\sqrt{T}}\Xi\Delta_{mmT}\Omega\zeta_T+\frac{1}{\sqrt{T}}\Xi\Delta_{\partial mT}\Xi\zeta_T-\frac{1}{\sqrt{T}}\Sigma\Delta_{\partial mT}'\Omega\zeta_T \\
& -\frac{1}{\sqrt{T}}\frac{\Xi}{2}\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial \theta_j}\right]\Xi\zeta_T\zeta_T'\Xi' e_j \\
& -\frac{1}{\sqrt{T}}\Xi\sum_{i=1}^{\ell}E\left[m_{it}^{\sigma}m_t m_t^{\sigma\prime}\right]\Omega\zeta_T\zeta_T'\Omega e_i \\
& +\frac{1}{\sqrt{T}}\frac{\Sigma}{2}\sum_{i=1}^{\ell}E\left[m_{\theta t}'m_{it}^{\sigma}+m_{\theta t}'e_i m_t^{\sigma\prime}\right]\Omega\zeta_T\zeta_T'\Omega e_i+o_p\left(\frac{1}{\sqrt{T}}\right).
\end{aligned}
$$

Now we will compute various components of the second-order bias (omitting the order factor of $1/T$) using lemma 5:

$$
\begin{aligned}
B_{mmm} \quad = \quad & Bias\left[\Xi\Delta_{mmT}\Omega\zeta_T-\Xi\sum_{i=1}^{\ell}E\left[m_{it}^{\sigma}m_t m_t^{\sigma\prime}\right]\Omega\zeta_T\zeta_T'\Omega e_i\right] \\
= \quad & \Xi\sum_{s=-\infty}^{+\infty}E\left[m_t m_t^{\sigma\prime}\Omega m_{t-s}\right]-\Xi E\left[m_t m_t^{\sigma\prime}\Omega m_t^{\sigma}\right]
\end{aligned}
$$

$$= \Xi \sum_{|s|>q} E\left[m_t m_t^{\sigma\prime} \Omega m_{t-s}\right],$$

since

$$\begin{aligned}
\sum_{i=1}^{\ell} E\left[m_{it}^{\sigma} m_t m_t^{\sigma\prime}\right] E\left[\Omega m_t m_t^{\sigma\prime} \Omega e_i\right] &= \sum_{i=1}^{\ell} E\left[m_{it}^{\sigma} m_t m_t^{\sigma\prime}\right] \Omega V \Omega e_i \\
&= E\left[m_t m_t^{\sigma\prime} \Omega \sum_{i=1}^{\ell} m_{it}^{\sigma} e_i\right] \\
&= E\left[m_t m_t^{\sigma\prime} \Omega m_t^{\sigma}\right].
\end{aligned}$$

Next,

$$\begin{aligned}
B_{\partial^2 m} &= Bias\left[-\frac{\Xi}{2} \sum_{j=1}^{k} E\left[\frac{\partial m_{\theta t}}{\partial \theta_j}\right] \Xi \zeta_T \zeta_T' \Xi' e_j\right] \\
&= -\frac{\Xi}{2} \sum_{j=1}^{k} E\left[\frac{\partial m_{\theta t}}{\partial \theta_j}\right] E\left[\Xi m_t m_t^{\sigma\prime} \Xi' e_j\right] \\
&= -\frac{\Xi}{2} E\left[\sum_{j=1}^{k} \frac{\partial m_{\theta t}}{\partial \theta_j} \Sigma e_j\right].
\end{aligned}$$

Finally,

$$\begin{aligned}
B_{m\partial m} &= Bias\left[\Xi \Delta_{\partial m T} \Xi \zeta_T - \Sigma \Delta_{\partial m T}' \Omega \zeta_T + \frac{\Sigma}{2} \sum_{i=1}^{\ell} E\left[m_{\theta t}' m_{it}^{\sigma} + m_{\theta t}' e_i m_t^{\sigma\prime}\right] \Omega \zeta_T \zeta_T' \Omega e_i\right] \\
&= \Xi \sum_{s=-\infty}^{+\infty} E\left[m_{\theta t} \Xi m_{t-s}\right] - \Sigma \sum_{s=-\infty}^{+\infty} E\left[m_{\theta t}' \Omega m_{t-s}\right] \\
&\quad + \frac{\Sigma}{2} \sum_{i=1}^{\ell} E\left[m_{\theta t}' m_{it}^{\sigma} + m_{\theta t}' e_i m_t^{\sigma\prime}\right] E\left[\Omega m_t m_t^{\sigma\prime} \Omega e_i\right] \\
&= \Xi \sum_{s=-\infty}^{+\infty} E\left[m_{\theta t} \Xi m_{t-s}\right] - \Sigma \sum_{|s|>q} E\left[m_{\theta t}' \Omega m_{t-s}\right],
\end{aligned}$$

since

$$\begin{aligned}
\sum_{i=1}^{\ell} E\left[m_{\theta t}' m_{it}^{\sigma} + m_{\theta t}' e_i m_t^{\sigma\prime}\right] E\left[\Omega m_t m_t^{\sigma\prime} \Omega e_i\right] &= \sum_{i=1}^{\ell} E\left[m_{\theta t}' m_{it}^{\sigma} + m_{\theta t}' e_i m_t^{\sigma\prime}\right] \Omega V \Omega e_i \\
&= E\left[m_{\theta t}' \Omega \sum_{i=1}^{\ell} m_{it}^{\sigma} e_i\right] + E\left[m_{\theta t}' \sum_{i=1}^{\ell} e_i e_i' \Omega m_t^{\sigma}\right] \\
&= 2E\left[m_{\theta t}' \Omega m_t^{\sigma}\right].
\end{aligned}$$

Summing up all components of the bias delivers (2.3.15).

2.6.4 Second order asymptotic bias of SEL estimator

The derivatives of the summands in the FOC (2.2.10)–(2.2.11) are:

$$\begin{aligned}
\frac{\partial \dfrac{m^\kappa}{1+\lambda' m^\kappa}}{\partial\theta'} &= \frac{m^\kappa_\theta}{1+\lambda' m^\kappa} - \frac{m^\kappa\cdot\lambda' m^\kappa_\theta}{\left(1+\lambda' m^\kappa\right)^2}, \quad \frac{\partial \dfrac{m^\kappa}{1+\lambda' m^\kappa}}{\partial\lambda'} = -\frac{m^\kappa m^{\kappa\prime}}{\left(1+\lambda' m^\kappa\right)^2},\\
\frac{\partial \dfrac{m^{\kappa\prime}_\theta\lambda}{1+\lambda' m^\kappa}}{\partial\theta_j} &= \frac{\dfrac{\partial m^{\kappa\prime}_\theta}{\partial\theta_j}}{1+\lambda' m^\kappa}\lambda - \frac{m^{\kappa\prime}_\theta\lambda\cdot\lambda'\dfrac{\partial m^\kappa}{\partial\theta_j}}{\left(1+\lambda' m^\kappa\right)^2}, \quad \frac{\partial \dfrac{m^{\kappa\prime}_\theta\lambda}{1+\lambda' m^\kappa}}{\partial\lambda'} = \frac{m^{\kappa\prime}_\theta}{1+\lambda' m^\kappa} - \frac{m^{\kappa\prime}_\theta\lambda\cdot m^{\kappa\prime}}{\left(1+\lambda' m^\kappa\right)^2}.
\end{aligned}$$

Then we have the expansion

$$\begin{aligned}
o_p\left(\frac{1}{\sqrt{T}}\right) &= \frac{1}{\sqrt{T}}\sum_{t=1}^T m^\kappa_t + \left[Q + \frac{1}{\sqrt{T}}\frac{1}{\sqrt{T}}\sum_{t=1}^T\left(m^\kappa_{\theta t} - Q_{\partial m}\right)\right]\sqrt{T}\left(\hat\theta-\theta\right)\\
&\quad - \left[V + \left(\rho_2^{-1}\frac{\delta_T}{T}\sum_{t=1}^T m^\kappa_t m^{\kappa\prime}_t - V\right)\right]\rho_2\sqrt{T}\delta_T^{-1}\hat\lambda\\
&\quad + \frac{1}{2\sqrt{T}}\sum_{j=1}^k\left[\frac{1}{T}\sum_{t=1}^T\frac{\partial m^\kappa_{\theta t}}{\partial\theta_j}\right]\sqrt{T}\left(\hat\theta-\theta\right)\sqrt{T}\left(\hat\theta_j-\theta_j\right)\\
&\quad - \frac{1}{\sqrt{T}}\sum_{i=1}^\ell\left[\frac{1}{T}\sum_{t=1}^T\left(m^\kappa_{\theta t}m^\kappa_{it} + m^\kappa_t m^\kappa_{\theta it}\right)\right]\sqrt{T}\left(\hat\theta-\theta\right)\sqrt{T}\hat\lambda_i\\
&\quad + \frac{1}{\sqrt{T}}\sum_{i=1}^\ell\left[\frac{1}{T}\sum_{t=1}^T m^\kappa_{it}m^\kappa_t m^{\kappa\prime}_t\right]\sqrt{T}\hat\lambda\sqrt{T}\hat\lambda_i,\\
o_p\left(\frac{\delta_T}{\sqrt{T}}\right) &= \left[Q + \frac{1}{\sqrt{T}}\frac{1}{\sqrt{T}}\sum_{t=1}^T\left(m^\kappa_{\theta t} - Q_{\partial m}\right)\right]'\sqrt{T}\hat\lambda\\
&\quad + \frac{1}{\sqrt{T}}\sum_{j=1}^k\left[\frac{1}{T}\sum_{t=1}^T\frac{\partial m^{\kappa\prime}_{\theta t}}{\partial\theta_j}\right]\sqrt{T}\hat\lambda\sqrt{T}\left(\hat\theta_j-\theta_j\right)\\
&\quad - \frac{1}{2\sqrt{T}}\sum_{i=1}^\ell\left[\frac{1}{T}\sum_{t=1}^T\left(m^{\kappa\prime}_{\theta t}m^\kappa_{it} + m^{\kappa\prime}_{\theta t}e_i m^{\kappa\prime}_t\right)\right]\sqrt{T}\hat\lambda\sqrt{T}\hat\lambda_i,
\end{aligned}$$

or, due to lemmas 6, 7(a)–(c), and 3,

$$\begin{aligned}
o_p\left(\frac{1}{\sqrt{T}}\right) &= \zeta_T + \Psi_T + Q\sqrt{T}\left(\hat\theta-\theta\right) - V\rho_2\delta_T^{-1}\sqrt{T}\hat\lambda\\
&\quad + \frac{1}{\sqrt{T}}\left(\Delta_{\partial m} + O_p\left(\frac{\delta_T}{\sqrt{T}}\right)\right)\left(-\Xi\zeta_T + O_p\left(\sqrt{\frac{\delta_T}{T}}\right)\right)\\
&\quad - \left(\rho_2^{-1}\frac{\delta_T}{T}\sum_{t=1}^T m^\kappa_t m^{\kappa\prime}_t - V\right)\left(\Omega\zeta_T + O_p\left(\sqrt{\frac{\delta_T}{T}}\right)\right)\\
&\quad + \frac{1}{2\sqrt{T}}\sum_{j=1}^k\left(E\left[\frac{\partial m_{\theta t}}{\partial\theta_j}\right] + O_p\left(\frac{\delta_T}{T} + \frac{1}{\sqrt{T}}\right)\right)\left(-\Xi\zeta_T\zeta_T'\Xi' e_j + O_p\left(\sqrt{\frac{\delta_T}{T}}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
& -\frac{1}{\sqrt{T}}\sum_{i=1}^{\ell}\left(\rho_2\delta_T^{-1}\sum_{u=-\infty}^{+\infty}E\left[m_{\theta t}m_{i,t-u}+m_t m_{\theta i,t-u}\right]+O_p\left(\frac{\delta_T^2}{T}\right)\right)\times \\
& \qquad \times\sqrt{T}\left(\hat{\theta}-\theta\right)\sqrt{T}\hat{\lambda}_i \\
& +\frac{1}{\sqrt{T}}\sum_{i=1}^{\ell}\left(\rho_3\delta_T^{-2}\sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty}E\left[m_{it}m_{t-u}m_{t-v}'\right]+O_p\left(\frac{\delta_T^3}{T}\right)\right)\times \\
& \qquad \times\left(\rho_2\delta_T^{-1}\right)^{-2}\left(\Omega\zeta_T\zeta_T'\Omega e_i+O_p\left(\sqrt{\frac{\delta_T}{T}}\right)\right),
\end{aligned}
$$

$$
\begin{aligned}
o_p\left(\frac{1}{\sqrt{T}}\right) \quad = \quad & Q'\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda}+\frac{1}{\sqrt{T}}\left(\Delta_{\partial m}+O_p\left(\frac{\delta_T}{\sqrt{T}}\right)\right)'\left(\Omega\zeta_T+O_p\left(\sqrt{\frac{\delta_T}{T}}\right)\right) \\
& +\frac{1}{\sqrt{T}}\sum_{j=1}^{k}\left(E\left[\frac{\partial m_{\theta t}}{\partial\theta_j}\right]+O_p\left(\frac{\delta_T}{T}\right)\right)'\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda}\sqrt{T}\left(\hat{\theta}_j-\theta_j\right) \\
& -\frac{1}{2\sqrt{T}}\left(\rho_2\delta_T^{-1}\sum_{i=1}^{\ell}\sum_{u=-\infty}^{+\infty}E\left[m_{\theta t}'m_{i,t-u}+m_{\theta t}'e_i m_{t-u}'\right]+O_p\left(\frac{\delta_T^2}{T}\right)\right)\times \\
& \qquad \times\left(\rho_2\delta_T^{-1}\right)^{-1}\left(\Omega\zeta_T\zeta_T'\Omega e_i+O_p\left(\sqrt{\frac{\delta_T}{T}}\right)\right),
\end{aligned}
$$

or after simplifications,

$$
\begin{aligned}
o_p\left(\frac{1}{\sqrt{T}}\right) \quad = \quad & \zeta_T+V\Omega\zeta_T+\Psi_T+Q_{\partial m}\sqrt{T}\left(\hat{\theta}-\theta\right)-V\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda} \\
& -\frac{1}{\sqrt{T}}\Delta_{\partial m}\Xi\zeta_T-\frac{1}{\sqrt{T}}\rho_2^{-1}\frac{\delta_T}{\sqrt{T}}\sum_{t=1}^{T}m_t^{\kappa}m_t^{\kappa\prime}\Omega\zeta_T \\
& -\frac{1}{2\sqrt{T}}\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial\theta_j}\right]\Xi\zeta_T\zeta_T'\Xi' e_j \\
& -\frac{1}{\sqrt{T}}\sum_{i=1}^{\ell}\sum_{u=-\infty}^{+\infty}E\left[m_{\theta t}m_{i,t-u}+m_t m_{\theta i,t-u}\right]\sqrt{T}\left(\hat{\theta}-\theta\right)\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda}_i \\
& +\frac{1}{\sqrt{T}}\frac{\rho_3}{\rho_2^2}\sum_{i=1}^{\ell}\sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty}E\left[m_{it}m_{t-u}m_{t-v}'\right]\Omega\zeta_T\zeta_T'\Omega e_i \\
& +O_p\left(\frac{\delta_T}{T}+\frac{\delta_T^5}{T\sqrt{T}}\right), \\
o_p\left(\frac{1}{\sqrt{T}}\right) \quad = \quad & Q_{\partial m}'\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda}+\frac{1}{\sqrt{T}}\Delta_{\partial m}'\Omega\zeta_T \\
& +\frac{1}{\sqrt{T}}\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}'}{\partial\theta_j}\right]\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda}\sqrt{T}\left(\hat{\theta}_j-\theta_j\right) \\
& -\frac{1}{2\sqrt{T}}\sum_{i=1}^{\ell}\sum_{u=-\infty}^{+\infty}E\left[m_{\theta t}'m_{i,t-u}+m_{\theta t}'e_i m_{t-u}'\right]\Omega\zeta_T\zeta_T'\Omega e_i
\end{aligned}
$$

$$+O_p\left(\frac{\delta_T}{T}+\frac{\delta_T^2}{T\sqrt{T}}\right).$$

Premultiplying the first equation by $Q'_{\partial m}V^{-1}$, adding the second equation, expressing out $\sqrt{T}\left(\hat{\theta}-\theta\right)$, and computing the components of the second-order bias, we get (omitting the order factor of $1/T$) using lemmas 5, 6 and 7(c),

$$\begin{aligned}
B_{mmm} &= Bias\left[\Xi\frac{\delta_T}{\rho_2\sqrt{T}}\sum_{t=1}^{T}m_t^{\kappa}m_t^{\kappa\prime}\Omega\zeta_T-\Xi\frac{\rho_3}{\rho_2^2}\sum_{i=1}^{\ell}\sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty}E\left[m_{it}m_{t-u}m'_{t-v}\right]\Omega\zeta_T\zeta'_T\Omega e_i\right]\\
&= \Xi\sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty}E\left[m_tm'_{t-u}\Omega m_{t-v}\right]-\Xi\rho_3\rho_2^{-2}\sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty}E\left[m_tm'_{t-u}\Omega m_{t-v}\right]\\
&= \left(1-\frac{\rho_3}{\rho_2^2}\right)\Xi\sum_{u=-\infty}^{+\infty}\sum_{v=-\infty}^{+\infty}E\left[m_tm'_{t-u}\Omega m_{t-v}\right],
\end{aligned}$$

$$\begin{aligned}
B_{\partial^2 m} &= Bias\left[-\frac{\Xi}{2}\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial\theta_j}\right]\Xi\zeta_T\zeta'_T\Xi' e_j\right]\\
&= -\frac{\Xi}{2}\sum_{j=1}^{k}E\left[\frac{\partial m_{\theta t}}{\partial\theta_j}\right]\sum_{s=-\infty}^{+\infty}E\left[\Xi m_tm'_{t-s}\Xi' e_j\right]\\
&= -\frac{\Xi}{2}E\left[\sum_{j=1}^{k}\frac{\partial m_{\theta t}}{\partial\theta_j}\Sigma e_j\right],
\end{aligned}$$

$$\begin{aligned}
B_{m\partial m} &= Bias\left[\Xi\Delta_{\partial m}\Xi\zeta_T-\Sigma\Delta'_{\partial m}\Omega\zeta_T\right.\\
&\quad\left.+\frac{\Sigma}{2}\sum_{i=1}^{\ell}\sum_{u=-\infty}^{+\infty}E\left[m'_{\theta t}m_{i,t-u}+m'_{\theta t}e_im'_{t-u}\right]\Omega\zeta_T\zeta'_T\Omega e_i\right]\\
&= \Xi\sum_{u=-\infty}^{+\infty}E\left[m_{\theta t}\Xi m_{t-u}\right]-\Sigma\sum_{u=-\infty}^{+\infty}E\left[m'_{\theta t}\Omega m_{t-u}\right]\\
&\quad+\frac{\Sigma}{2}\sum_{i=1}^{\ell}\sum_{u=-\infty}^{+\infty}E\left[m'_{\theta t}m_{i,t-u}+m'_{\theta t}e_im'_{t-u}\right]\sum_{u=-\infty}^{+\infty}E\left[\Omega m_tm'_{t-u}\Omega e_i\right]\\
&= \Xi\sum_{u=-\infty}^{+\infty}E\left[m_{\theta t}\Xi m_{t-u}\right].
\end{aligned}$$

Expectations of the terms involving the interactions between $\sqrt{T}\left(\hat{\theta}-\theta\right)$ and $\rho_2\delta_T^{-1}\sqrt{T}\hat{\lambda}$ contribute, due to the third result in lemma 3, the bias of order $O_p\left(\sqrt{\delta_T}/T\right)$, i.e. of smaller order. Summing up all components of the bias delivers (2.3.16).

Bibliography

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

[2] Angrist, J.D., Imbens, G.W. and A. Krueger (1999) Jackknife instrumental variables estimation. *Journal of Applied Econometrics* 14, 57–67.

[3] Back, K. and Brown, D. (1990) Estimating distributions from moment restrictions. Working Paper, Indiana University.

[4] Efron, B. and R.J. Tibshirani (1993) An introduction to the bootstrap. New York, Chapman and Hall.

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

[6] Gospodinov, N. (2002) Nonparametric likelihood inference in moment condition models with martingale structure. Working Paper, Concordia University.

[7] Gregory, A.W., J.-F. Lamarche, and G.W. Smith (2002) Information-theoretic estimation of preference parameters: macroeconomic applications and simulation evidence. *Journal of Econometrics* 107, 213–233.

[8] Hall, P. and C.C. Heyde (1980) *Martingale limit theory and its application.* New York: Academic Press.

[9] Hansen, L.-P., Heaton, J. and A. Yaron (1996) Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* 14, 262–280.

[10] Hansen, L.P and R.J. Hodrick (1980) Forward exchange rates as optimal predictors of future spot rates: an econometric analysis. *Journal of Political Economy* 88, 829–853.

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

[12] Hansen, L.P. and Singleton K.J. (1996) Efficient estimation of linear asset pricing models with moving-average errors. *Journal of Business and Economic Statistics* 14, 53–68.

[13] Imbens, G.W. (1997) One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies* 64, 359–383.

[14] Imbens, G.W., Spady, R.H. and P. Johnson (1998) Information theoretic approaches to inference in moment condition models. *Econometrica* 66, 333–357.

[15] Kitamura, Y. (1997) Empirical likelihood methods with weakly dependent processes. *Annals of Statistics* 25, 2084–2102.

[16] Kitamura, Y. and M. Stutzer (1997) An information-theoretic alternative to generalized method of moments estimation. *Econometrica* 65, 861–874.

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

[18] Mittelhammer, R.C., G.G. Judge and R. Schoenberg (2001) Empirical evidence concerning the finite sample performance of EL-type structural equation estimation and inference methods. Working Paper, Washington State University.

[19] Newey, W.K. and R.J. Smith (2000) Asymptotic bias and equivalence of GMM and GEL estimators. Working Paper No. 01/517, University of Bristol.

[20] 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.

[21] Owen, A. (1991) Empirical likelihood for linear models. *Annals of Statistics* 19, 1725–1747.

[22] Parzen, E. (1957) On consistent estimates of the spectrum of a stationary time series. *Annals of Mathematical Statistics* 28, 329–348.

[23] Qin, J. and J. Lawless (1994) Empirical likelihood and general estimating equations. *Annals of Statistics* 22, 300–325.

[24] 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.

[25] Rosenblatt, M. and J.W. Van Ness (1965) Estimation of the Bispectrum. *Annals of Mathematical Statistics* 36, 1120–1136.

[26] Rothenberg, T.J. (1984) Approximating the distributions of econometric estimators and test statistics. In: Griliches, Z. and M.D. Intriligator, eds., *Handbook of Econometrics*, Vol. 2, New York: North-Holland.

[27] Smith, R.J. (1997) Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Economic Journal* 107, 503–519.

[28] Smith, R.J. (1998) General empirical likelihood criteria for generalized method of moments estimation and inference. Working Paper, University of Bristol.

[29] Smith, R.J. (2000) Empirical likelihood estimation and inference. In: Marriott, P. and M. Salmon, eds., *Applications of Differential Geometry to Econometrics.* Cambridge: Cambridge University Press, 119–150.

3. BOOTSTRAP INFERENCE IN MULTIPERIOD PREDICTION PROBLEMS

3.1 Introduction

For linear prediction models a popular method of estimation and inference is OLS with standard errors corrected for serial correlation and sometimes conditional heteroskedasticity.[1] The correction is usually that of Hansen and Hodrick (1980), Newey and West (1987), or a modification of these. It has been noticed more than once that asymptotic approximation works poorly in typically available samples. For example, Mishkin (1990b) finds via simulations that correcting for conditional heteroskedasticity can appreciably distort inference even when the data are conditionally heteroskedastic. Thus, it is important to find more reliable tools of inference specifically for this class of models. A natural candidate for this is the bootstrap (Efron 1979). In this paper, we investigate the performance of various bootstrap resampling schemes in a systematic way with the help of Monte-Carlo simulations, with the forecasting horizon set equal to 2. Some bootstrap algorithms in the context of serially correlated errors have been explored in the literature. For example, for long-run regressions Mark (1995) uses parametric and nonparametric residual bootstrap under the null of no predictability to bias-adjust the OLS estimates and to test the null, and notices severe size distortions in the asymptotic t-tests compared to the bootstrap t-tests. Bose (1990) studies residual bootstrap in moving average models; Gospodinov (2002) uses the grid bootstrap of Hansen (1999a) working with the ML estimator in an MA(1) model with the MA root close to unity. Both Bose and Gospodinov, however, are interested in inference about moving average coefficients. We are instead concerned with the inference about coefficients in the conditional mean while serial correlation in the error is treated as a nuisance feature.

Although a multiperiod prediction model may not necessarily be linear and also may involve exogenous variables, we use a linear model without exogenous variables to isolate distortions caused by the serial correlation structure of the error term. We set the slope parameter equal to zero, which corresponds to the null hypothesis of no forecasting ability. The effects of nonlinearities and persistence on the bootstrap performance are extensively discussed in Kilian (1998, 1999). We tune the parameters of the Data Generating Process (DGP) so that the regression error is conditionally homo- or heteroskedastic, with innovations that are serially independent or only serially uncorrelated but dependent. This allows us to study the impact of prediction error structure on bootstrap performance. As a measure of performance, we use rejection rates for one-sided and symmetric two-sided tests of the null of zero value for the slope parameter from bootstrapping the t-ratio, which is an asymptotically pivotal statistic.

Regarding the covariance matrix estimator, the structure of the model imposes zeroness on the autocovariances beyond the lag corresponding to the prediction horizon, so a natural

[1] See, for example, Hansen and Hodrick (1980), Huizinga and Mishkin (1984), Blake, Beenstock and Brasse (1986), Frankel and Froot (1987), Fama and French (1988), Mishkin (1990a), Lewis (1990), Estrella and Hardouvelis (1991), De Bondt and Bange (1992), Chinn and Frankel (1994), Debelle and Lamont (1997), Lettau and Ludvigson (2001).

estimator is that of Hansen and Hodrick (1980), although many studies employ the Newey and West (1987) estimator for the sake of its positive definiteness. Our simulations indicate that the use of the Hansen–Hodrick estimator with a simple correction in case of its negativity provides a much more precise rejection rates than the use of the Newey–West estimator, to say nothing about a need to choose a truncation parameter in the latter case. Inoue and Shintani (2001) consider performance of the block bootstrap (Carlstein 1986, Künsch 1989) when a HAC variance matrix is used. Thanks to the *a priori* known autocorrelation structure of the moment condition in our problem we can use a simpler estimator that is a function of sample averages, which is known from the bootstrap literature to provide more precise rejection rates. When employing the block bootstrap we can use correction factors described in Hall and Horowitz (1996) and Andrews (2002) extended to the case where the sample size is not a multiple of the block length. The correction factors are meant to correct for independence between blocks in bootstrap samples which is absent in the original sample.

Beside the block bootstrap, we consider the residual bootstrap (Bose 1988) and the wild bootstrap (Wu 1986). The residual bootstrap is ideal when the Wold innovation in the error is a serially independent sequence, but we explore how critical this condition is when the error does not have that ideal property. The wild bootstrap is expected to help in the case of a conditionally on the history heteroskedastic error term. The block bootstrap is robust to the presence or absence of such properties, but it invokes selection of an additional parameter. Our simulation evidence shows that the residual bootstrap performs well even in situations where the non-IID structure of Wold innovations in the error term is expected to contaminate the inference. Small distortions caused by the presence of a strong conditional heteroskedasticity are partly removed by the wild bootstrap. The use of either variation of the block bootstrap, while being able to provide more precise rejection rates, is more problematic due to the need to select a block length, exacerbated by the fact that plug-in rules for the optimal block length suggested in the literature do not work well in practice. Thus, a higher degree of asymptotic refinement delivered by residual based algorithms in comparison with that of block algorithms (see Andrews 2002) may be said to dominate a somewhat faulty applicability of the residual based resampling schemes in the multiperiod prediction context. In addition, the use of the block bootstrap with correction factors is computationally far more intensive than that of alternative schemes.

The rest of the paper is organized as follows. Section 2 characterizes the model and estimators used, section 3 – three data generating mechanisms. Section 4 describes the residual, wild and block bootstrap algorithms in relation to the model used. In section 5 simulation results are reported and discussed. Section 6 concludes.

3.2 Model and estimator

The working model is that of two-step-ahead linear prediction:

$$y_{t+2} = \alpha + \beta y_t + e_{t+2}, \quad E_t\left[e_{t+2}\right] = 0, \tag{3.2.1}$$

where $E_t\left[\cdot\right] \equiv E\left[\cdot|\sigma\left(y_t, y_{t-1}, \ldots\right)\right]$. The true values of α and β are set to zero in all DGPs. The zero value of β allows us not to be distracted by an autoregression bias that is often blamed for unsatisfactory bootstrap performance (Kilian 1998, 1999, Berkowitz and Kilian 2000). We vary properties of the DGP by varying features of the error term e_{t+2}.

The estimator we concentrate on is the OLS estimator of β:

$$\hat{\beta} = \frac{\sum_{t=1}^{T-2} y_{t+2}(y_t - \bar{y})}{\sum_{t=1}^{T-2}(y_t - \bar{y})^2}, \tag{3.2.2}$$

where $\bar{y} = \frac{1}{T-2}\sum_{t=1}^{T-2} y_t$. The OLS estimator of α is $\hat{\alpha} = \frac{1}{T-2}\sum_{t=1}^{T-2} y_{t+2} - \hat{\beta}\bar{y}$. The residuals are computed as $\hat{e}_1 = y_1 - \hat{\alpha} - \hat{\beta}\bar{y}$, $\hat{e}_2 = y_2 - \hat{\alpha} - \hat{\beta}\bar{y}$, $\hat{e}_t = y_t - \hat{\alpha} - \hat{\beta}y_{t-2}$, $t = 3, \cdots, T$. The estimate of the asymptotic variance is set to

$$\begin{aligned} \widehat{V} &= \left(\sum_{t=1}^{T-2} \vec{y}_t\vec{y}_t{}'\right)^{-1} \left(\sum_{t=1}^{T-2} \vec{y}_t\vec{y}_t{}' \left(\widehat{e}_{t+2}\right)^2\right. \\ &\qquad \left. + \sum_{t=1}^{T-3} \left(\vec{y}_{t+1}\vec{y}_t{}' + \vec{y}_t\vec{y}_{t+1}{}'\right) \widehat{e}_{t+2}\widehat{e}_{t+3}\right) \left(\sum_{t=1}^{T-2} \vec{y}_t\vec{y}_t{}'\right)^{-1}, \end{aligned} \tag{3.2.3}$$

where $\vec{y}_t = (1\ y_t)'$ (we omit scalar factors like T since they are immaterial for the bootstrap). Whenever the $(2,2)$ component is needed and it is negative, $\widehat{V}$ is modified by excluding the covariance terms in the middle term of (3.2.3).

3.3 Data-generating processes

We consider the following types of structure of the error e_t in the DGP: a moving average with IID innovations; conditionally homoskedastic with uncorrelated but not independent innovations; conditionally heteroskedastic with an ARCH-type skedastic function. In all DGPs, we make the error unconditionally leptokurtic, with the degree of leptokurtocity falling into a range that is typical for financial data (Stambaugh 1993).

The simplest structure of the error occurs when it is a moving average with independent innovations. We will call this *DGP IID*:

$$e_{t+2} = w_{t+2} - \theta w_{t+1}, \quad w_{t+2} \sim IID\ \sqrt{.6}t_{(5)}. \tag{3.3.4}$$

The distributional specification of the innovation (scaled Student's with 5 degrees of freedom) implies the unconditional kurtosis of e_t equal $\kappa = 9 - 12\theta^2/(1+\theta^2)^2$. In our experiments, κ varies from 6.0 to 8.1.

Another structure of the error implies exactly the same autocovariance function as DGP IID, but the innovations are not independent. We will call this *DGP UC*:

$$\begin{aligned} e_{t+2} &= u_{t+2} - \text{sgn}(\theta)u_{t+1} + v_{t+2}, \\ u_{t+2} &\sim IID\ \sqrt{.6|\theta|}t_{(5)}, \quad v_{t+2} \sim IID\ \sqrt{.6}(1-|\theta|)t_{(5)}, \end{aligned} \tag{3.3.5}$$

and u_{t+2} and v_{t+2} are independent. The variance parameters of the disturbances u_{t+1} and v_{t+1} are set so that the variance and first-order autocovariance of e_t are the same in DGPs IID and UC. The principal difference between error structures in (3.3.4) and (3.3.5) is that the Wold innovation in (3.3.4) is an IID sequence (a structure that is ideal for the residual bootstrap), while in (3.3.5) the innovation is a serially uncorrelated, but not independent, sequence[2] (a structure

[2] Note that if the two disturbances were normal, the Wold innovations would be serially independent.

that may potentially invalidate the residual bootstrap). The distributional specification implies the unconditional kurtosis of e_t equal $\kappa = 9 - 6(4|\theta|(1-|\theta|)^2 - \theta^2)/(1+\theta^2)^2$. In our experiments, κ varies from 6.5 to 10.5.

Finally, we explore a conditionally heteroskedastic structure of the error which we will call this *DGP HS*:

$$e_{t+2} = \mu_{t+1} + \zeta_{t+2}\sqrt{\omega_t},\ \mu_{t+1} = \delta\left(\mu_t - e_{t+1}\right),\ \zeta_{t+2} \sim IID\,\mathcal{N}(0,1), \tag{3.3.6}$$

where $0 < \delta < 1$ and the auxiliary process ω_t may be specified in a variety of ways. It is easy to see that $E_t[\mu_{t+1}] = 0$ and thus $E_t[e_{t+2}] = 0$. The conditional autocovariance structure of e_{t+2} is: $E_t[e_{t+2}^2] = \omega_t + \delta^2\omega_{t-1}$, $E_t[e_{t+2}e_{t+1}] = -\delta\omega_{t-1}$. Since we design this DGP to explore only the impact of conditional heteroskedasticity, we set δ to be time invariant, so that the Wold innovation of e_{t+2} is a martingale difference relative to the history. This, along with the presence of conditional heteroskedasticity, distinguishes DGP HS from DGP UC. We make ω_t time-varying in the ARCH(1) spirit: $\omega_t = 1 - \alpha_\omega(1+\delta^2) + \alpha_\omega e_t^2$, where $0 < \alpha_\omega < 1$. Then $E[e_{t+2}^2] = 1+\delta^2$, $E[e_{t+2}e_{t+1}] = -\delta$, so the implied moving-average coefficient is $\theta = \delta$. The fourth moment of e_t exists if $3\alpha_\omega^2(1+\alpha_\omega\delta^2) < (1-3\alpha_\omega^2\delta^4)(1-\alpha_\omega\delta^2)$, and the implied unconditional kurtosis of e_t is

$$\kappa = 3\frac{(1-\alpha_\omega-\alpha_\omega\delta^2)(1+\alpha_\omega-\alpha_\omega\delta^2)(1+\alpha_\omega\delta^2)}{(1-3\alpha_\omega^2\delta^4)(1-\alpha_\omega\delta^2) - 3\alpha_\omega^2(1+\alpha_\omega\delta^2)}.$$

For a given value of θ (and thus δ), we put α_ω to be such that the unconditional kurtosis is $\kappa = 15$.

3.4 Bootstrap Methods

3.4.1 Residual bootstrap

In the *residual bootstrap* (RB) one resamples the Wold innovation in the error treated as an IID process (Bose 1990, Kreiss and Franke 1992). Denote by ε the innovation, then $e_{t+2} = \varepsilon_{t+2} - \theta\varepsilon_{t+1}$. The innovation is indeed IID under DGP (3.3.4), but it is not under DGP (3.3.5) or (3.3.6). It is important to know if the structure of uncorrelated non-IID Wold innovations, on the one hand, or their conditionally heteroskedastic structure though with a martingale difference property, on the other hand, may have a significant adverse impact on performance of the residual bootstrap.

After the residuals $\hat{e}_t$, $t = 1, \cdots, T$ are computed, we restore estimates of the Wold innovations $\hat{\varepsilon}_t$, $t = 1, \cdots, T$ in the following way. We compute an estimate of θ by the method of moments imposing restrictions on a value of the correlation coefficient: $\hat{\theta} = -2\hat{\rho}/(1+(1-4\hat{\rho}^2)^{1/2})$, where $\hat{\rho} = \min(.499, \max(-.499, (\sum_{t=3}^{T-1}\hat{e}_t\hat{e}_{t+1})/(\sum_{t=3}^{T-1}\hat{e}_t^2)))$. Then we calculate innovations: $\hat{\varepsilon}_t = \sum_{i=0}^{t-1}\hat{\theta}^i\hat{e}_{t-i}$, $t = 1, \cdots, T$.

We then resample $\hat{\varepsilon}_t$ from the original sample randomly, uniformly over t, with replacement. Having obtained a bootstrap sample ε_t^*, $t = 1, \cdots, T$, we generate the e^* and y^* series recursively as $e_1^* = \varepsilon_1^*$, $y_1^* = \hat{\alpha} + \hat{\beta}\bar{y} + e_1^*$, $e_2^* = \varepsilon_2^* - \hat{\theta}\varepsilon_1^*$, $y_2^* = \hat{\alpha} + \hat{\beta}\bar{y} + e_2^*$, $e_t^* = \varepsilon_t^* - \hat{\theta}\varepsilon_{t-1}^*$, $y_t^* = \hat{\alpha} + \hat{\beta}y_{t-2}^* + e_t^*$, $t = 3, \cdots, T$. Using the bootstrap sample we obtain the bootstrap OLS estimator $\hat{\beta}^*$, bootstrap asymptotic variance estimate $\widehat{V}^*$ by (3.2.2) and (3.2.3) evaluated at the bootstrap sample, and bootstrap t-statistic $t_\beta^* = (\hat{\beta}^* - \hat{\beta})/s^*$, where $s^* = \sqrt{\widehat{V}_{2,2}^*}$.

3.4.2 Wild bootstrap

The *wild bootstrap* (WB) proposed by Wu (1986) helps to preserve the pattern of conditional heteroskedasticity in bootstrap samples. In the context of an autoregression it was described in Kreiss (1997) and applied, for instance, in Hafner and Herwartz (2000). We adapt the algorithm to our MA(1) setting. The construction of a bootstrap sample is similar to that of residual bootstrap, but instead of resampling the bootstrap innovations ε_t^+, $t = 1, \cdots, T$ from the set of estimated innovations $\hat{\varepsilon}_t$, $t = 1, \cdots, T$, we obtain them by multiplying the latter by an IID zero mean sequence η_t, $t = 1, \cdots, T$ having properties $E[\eta_t^2] = E[\eta_t^3] = 1$, i.e. $\varepsilon_t^+ = \eta_t \hat{\varepsilon}_t$, $t = 1, \cdots, T$. Then we set $e_1^+ = \varepsilon_1^+$, $e_2^+ = \varepsilon_2^+ - \hat{\theta}\varepsilon_1^+$, $e_t^+ = \varepsilon_t^+ - \hat{\theta}\varepsilon_{t-1}^+$, $t = 3, \cdots, T$. A bootstrap sample is generated recursively: $y_1^+ = \hat{\alpha} + \hat{\beta}\bar{y} + e_1^+$, $y_2^+ = \hat{\alpha} + \hat{\beta}\bar{y} + e_2^+$, $y_t^+ = \hat{\alpha} + \hat{\beta} y_{t-2}^+ + e_t^+$, $t = 3, \cdots, T$. From the bootstrap sample we obtain the bootstrap OLS estimator, bootstrap asymptotic variance estimate and bootstrap t-statistic as in the residual bootstrap.

In our experiment, we use the following probability distribution for η_t: let (η_{1t}, η_{2t}) be standard bivariate normal, then $\eta_t = \eta_{1t}/\sqrt{2} + (\eta_{2t}^2 - 1)/2$ (Mammen 1993).

3.4.3 Block bootstrap

The *block bootstrap* (BB) does not rely on a parametric structure of the error term. Instead, it attempts to capture the true underlying DGP by resampling the original data in blocks. We use both non-overlapping (NOL) (Hall 1985, Carilstein 1986) and overlapping (OL) (Hall 1985, Künsch 1989) versions of the BB. We correct the bootstrap t-statistics with the use of correction factors (Hall and Horowitz 1996, Andrews 2002): $t_\beta^* = \tau_\beta \sqrt{T}(\hat{\beta}^* - \hat{\beta})/s^*$, where τ_β is the correction factor. They are meant to correct for independence between blocks in bootstrap samples which is absent in the original sample. We extend the formulae for the correction factors given in Andrews (2002) to the case where the sample size is not a multiple of the block length. There is some arbitrariness in how to form the last fragmentary block in a bootstrap sample. We make the convention that a slightly lengthier bootstrap sample is drawn from the population of only complete blocks from the original sample, and then it is cut at the end to have the necessary length.

Let the block length be ℓ and denote $b = \lfloor \frac{T-3}{\ell} \rfloor$ and $B = T - 2 - \ell$. We resample blocks of length ℓ of row vectors $\tilde{y}_t = (y_t \; y_{t+1} \; y_{t+2} \; y_{t+3})$, at each bootstrap repetition getting their bootstrapped versions of the type $\tilde{y}_t^* = (y_t^* \; y_{t+1}^* \; y_{t+2}^* \; y_{t+3}^*)$. Let $\varrho = (\varrho_1 \; \varrho_2)$ be the recentering term to be defined shortly. The bootstrap OLS estimator of β is

$$\hat{\beta}^* = \frac{\sum_{t=1}^{T-2}(y_{t+2}^* y_t^* + \varrho_2 - (y_{t+2}^* + \varrho_1)\bar{y}^*)}{\sum_{t=1}^{T-2}(y_t^* - \bar{y}^*)^2},$$

where $\bar{y}^* = \frac{1}{T-2}\sum_{t=1}^{T-2} y_t^*$. The bootstrap OLS estimator of α is $\hat{\alpha}^* = \frac{1}{T-2}\sum_{t=1}^{T-2} y_{t+2}^* + \varrho_1 - \hat{\beta}^* \bar{y}^*$. The bootstrap OLS residuals are obtained as usually. Denote $\vec{y}_t^* = (1 \; \tilde{y}_t^* \imath_1)'$, $\vec{y}_{t+1}^* = (1 \; \tilde{y}_t^* \imath_2)'$, $\widehat{e}_{t+3}^* = \tilde{y}_t^* \imath_4 - \hat{\alpha}^* - \hat{\beta}^* \tilde{y}_t^* \imath_2$, $t = 1, \cdots, T-3$, where $\imath_i$ is i^{th} unit vector. The bootstrap variance estimator is computed as

$$\widehat{V}^* = \left(\sum_{t=1}^{T-2} \vec{y}_t^* \vec{y}_t^{*\prime}\right)^{-1} \left(\sum_{t=1}^{T-2} (\vec{y}_t^{*\prime} \widehat{e}_{t+2}^* + \varrho)'(\vec{y}_t^{*\prime} \widehat{e}_{t+2}^* + \varrho)\right.$$

$$+\sum_{t=1}^{T-3}(\vec{y}_{t+1}^{*\prime}\widehat{e}_{t+3}^{*}+\varrho)'(\vec{y}_{t}^{*\prime}\widehat{e}_{t+2}^{*}+\varrho)+(\vec{y}_{t}^{*\prime}\widehat{e}_{t+2}^{*}+\varrho)'(\vec{y}_{t+1}^{*\prime}\widehat{e}_{t+3}^{*}+\varrho)\Bigg)\left(\sum_{t=1}^{T-2}\vec{y}_t^{*}\vec{y}_t^{*\prime}\right)^{-1}.$$

A number of full blocks in the original sample is b for the NOL or B for the OL. The bootstrap population consists of $T-3$ pairs of observations $(y_1\ y_3)$, $(y_2\ y_3)$, $\cdots$, $(y_{T-3}\ y_{T-1})$. The correction terms are[3]

$$\varrho=\frac{1}{T-2}\frac{1}{b}\sum_{i=0}^{b-1}\sum_{j=1}^{T-2-b\ell}\vec{y}_{i\ell+j}'\widehat{e}_{i\ell+j+2}$$

in the NOL case and

$$\varrho=\frac{1}{T-2}\frac{1}{B}\sum_{t=0}^{B-1}\left[b\sum_{j=1}^{\ell}+\sum_{j=1}^{T-2-b\ell}\right]\vec{y}_{t+j}'\widehat{e}_{t+j+2}$$

in the OL case. The conditional (on the sample) variance of the asymptotic distribution of $\hat{\beta}^{*}-\hat{\beta}$ is

$$\tilde{V}=\left(\sum_{t=1}^{T-2}\vec{y}_t\vec{y}_t'\right)^{-1}\tilde{W}\left(\sum_{t=1}^{T-2}\vec{y}_t\vec{y}_t'\right)^{-1},$$

where

$$\tilde{W}=\sum_{i=0}^{b-1}\left[\sum_{j=1}^{\ell}\sum_{k=1}^{\ell}+\frac{1}{b}\sum_{j=1}^{T-2-b\ell}\sum_{k=1}^{T-2-b\ell}\right](\vec{y}_{i\ell+j}'\widehat{e}_{i\ell+j+2}+\varrho)'(\vec{y}_{i\ell+k}'\widehat{e}_{i\ell+k+2}+\varrho)$$

in the NOL case, and

$$\tilde{W}=\frac{1}{B}\sum_{t=0}^{B-1}\left[b\sum_{j=1}^{\ell}\sum_{k=1}^{\ell}+\sum_{j=1}^{T-2-b\ell}\sum_{k=1}^{T-2-b\ell}\right](\vec{y}_{t+j}'\widehat{e}_{t+j+2}+\varrho)'(\vec{y}_{t+k}'\widehat{e}_{t+k+2}+\varrho)$$

in the OL case. The second terms in these formulae are related to conditional (on the sample) variance for the appendix of a bootstrap sample beyond full b blocks. The correction factor is formed as $\tau_\beta=\sqrt{\widehat{V}_{22}/\tilde{V}_{22}}$.

3.5 Simulation results

We evaluate rejection rates for 5% size tests on the basis of 10,000 simulations. Thus, the estimates will have a standard error of approximately $\sqrt{5\%\cdot(100\%-5\%)/10{,}000}\approx 0.22\%$. In any single simulation loop, 500 repetitions are used to form a bootstrap distribution and read off bootstrap critical values. We generate a time series for y_t of $T+1{,}000$ observations with zero starting values using a DGP of interest and discard the first $1{,}000$ observations.

[3] The recentering for the non-overlapping blocks scheme is usually ignored in theory, and the sample is artificially slightly truncated. We do not do that because we do not want the original sample size (and thus the estimates) to depend on the block length.

Table 1 contains the results for the residual and wild bootstrap schemes, table 2 – for the NOL, table 3 – for the OL block bootstraps. We report actual rejection frequencies for symmetric two-sided (marked as $\beta \neq 0$) and right (marked as $\beta > 0$) and left (marked as $\beta < 0$) one-sides alternatives, i.e. $\Pr\{|t_\beta| > q^*_{5\%}\}$, $\Pr\{t_\beta > q^*_{5\%}\}$ and $\Pr\{t_\beta < q^*_{5\%}\}$, where $q^*_{5\%}$ is an appropriate bootstrap critical value corresponding to nominal size 5%. We run the residual bootstrap for all three DGPs, the wild bootstrap – for the DGP HS, and both variations of the block bootstrap – for the DGPs UC and HS. We choose to consider sample sizes 30, 60, 120, 240, and MA coefficients ∓ 0.3, ∓ 0.8, ∓ 0.95. This allows us to study the impact of sample length and strength of serial correlation. In unreported experiments we tried to use the Newey–West HAC variance estimator, which led to more significant size distortions; therefore we here concentrate on the Hansen–Hodrick estimator correcting for negative definiteness as described earlier.

One can notice immediately that overall actual rejection rates are much closer to nominal ones than those frequently encountered in other bootstrap simulation studies. This is a consequence of linearity of the model and absence of autoregressive persistence. For the largest sample size, most of size distortions fall in the range from 0% to 1.5% for symmetric two-sided alternatives and from 0% to 3% for one-sided ones. The DGP IID has an ideal error structure for the residual bootstrap, hence the corresponding numbers can be considered as lower bounds for size distortions for other DGPs and bootstrap algorithms. However, the results for the DGP UC are nearly identical to those for the DGP IID. For the DGP HS they are but slightly worse, meaning that the actual rejection rates deviate from the nominal sizes by $0 \div 2\%$ when $T = 240$, $1 \div 2.5\%$ when $T = 120$, $1 \div 3\%$ when $T = 60$, and $1 \div 5\%$ when $T = 30$ This implies a perhaps surprising fact: serial uncorrelatedness seems to be a guarantee against big distortions in the residual bootstrap. The serial independence is not necessary, but conditional homoskedasticity is desirable (recall though that the conditional heteroskedasticity is rather strong in the DGP HS). Further, we can analyze what fraction of distortions caused by conditional heteroskedasticity can be corrected by the use of the wild bootstrap. It turns out that this correction is not substantial for smaller sample sizes, but when T exceeds 100 the rejection rates nearly straighten out and are close to nominal sizes.

In tables 2 and 3 we explore the performance of block bootstraps applied to the DGP UC and DGP HS. We employ correction factors as described earlier; without their use rejection rates (not reported) tend to worsen appreciably. The block bootstrap requires selection of the block length, the optimal rate for which is $\ell \sim T^{\frac{1}{3}}$ for both symmetric two-sided and one-sided alternatives (Andrews 2002). In spite of equality of optimal convergence rates, in practice we find that the block length for one-sided alternatives should be maintained consistently at smaller values than for symmetric two-sided alternatives. In tables 2 and 3 we show the results for those block lengths that provide least size distortions and are thus different for various alternatives. By felicitous selection of a block length one can quite successfully reduce the distortion to zero (except for one-sided alternatives and small sample sizes). However, the distortion being zero is essentially a compromise between a usual slight overrejection and an arbitrary underrejection because of a high block length, and *a priori* there is no solid basis for deciding at which ℓ that compromise is realized. It seems easier, however, to control the optimal block length and attain lower size distortions for the non-overlapping BB than for the overlapping BB. This somewhat contradicts the usual econometric practice which prefers the latter to the former, although this practice was evolving when correction factors were not used.

For reference, we here give the optimal block sizes implied for our DGP by the automatic rule derived by Hall and Horowitz (1993) and cited in Maddala and Kim (1998). The formulae are

$$\ell = \left(\frac{(1-\theta)^2}{\theta}\right)^{-\frac{2}{3}} T^{\frac{1}{3}} \quad \text{and} \quad \ell = \left(\frac{(1-\theta)^2}{\theta}\right)^{-\frac{2}{3}} \left(\frac{3}{2}T\right)^{\frac{1}{3}}$$

for the NOL and OL block bootstraps, respectively. The automatic rule yields block sizes that are very sensitive to the serial correlation parameter, sometimes are ridiculously small (so that their use would break serial dependence in bootstrap samples) or unattainably big (so that the entire sample is not sufficient for one block). We find in our simulations that block sizes close to optimal do not vary that much with the strength of serial correlation.

To summarize, the residual bootstrap performs well even in situations where the non-IID structure of the error is expected to contaminate the inference much more. Small distortions caused by the presence of a strong conditional heteroskedasticity are partly removed by the wild bootstrap, while the use of either variation of the block bootstrap is more problematic due to the need to select a block length, exacerbated by the fact that plug-in rules for the optimal block length suggested in the literature do not work well in practice. Thus, higher degree of asymptotic refinement delivered by residual based algorithms in comparison with that of block algorithms (see Andrews 2002) may be said to dominate a somewhat faulty applicability of the residual based resampling schemes in the multiperiod prediction context. In addition, the use of the block bootstrap with correction factors is computationally far more intensive than that of alternative schemes.

3.6 Conclusion

We have studied the performance of the bootstrap inference in small samples in a short horizon linear forecasting model. The residual bootstrap performs well even in situations where the non-IID structure of Wold innovations is expected to contaminate the inference. Small distortions caused by the presence of a strong conditional heteroskedasticity are partly removed by the wild bootstrap, while the use of either variation of the block bootstrap is more problematic because of the need to select a block length, and in addition computationally more intensive.

A promising area of future research is extending the nonparametric procedure of Hansen (1999b) that incorporates the conditional moment restriction with the martingale difference structure into the bootstrap algorithm, to the problems with multiperiod restrictions. This procedure would be most robust to the dependence structure of innovations.

Bibliography

[1] Andrews, D.W.K. (2002) Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators. *Econometrica*, 70, 119–162.

[2] Berkowitz, J. & Kilian, L. (2000) Recent Developments in Bootstrapping Time Series. *Econometric Reviews*, 19, 1–54.

[3] Blake, D., Beenstock, M. & Brasse, V. (1986) The Performance of UK Exchange Rate Forecasters. *Economic Journal*, 96, 986–999.

[4] Bose, A. (1988) Edgeworth Correction by Bootstrap in Autoregressions. *Annals of Statistics*, 16, 1709–1722.

[5] Bose, A. (1990) Bootstrap in moving average models. *Annals of the Institute of Statistical Mathematics*, 42, 753–768.

[6] Carlstein, E. (1986) The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *Annals of Statistics*, 14, 1171–1179.

[7] Chinn, M. & Frankel, J. (1994) Patterns in Exchange Rate Forecasts for Twenty-Five Currencies. *Journal of Money, Credit and Banking*, 26, 759–770.

[8] Debelle, G. & Lamont, O. (1997) Relative Price Variability and Inflation: Evidence from U.S. Cities. *Journal of Political Economy*, 105, 132–152.

[9] De Bondt, W.F.M. & Bange, M.M. (1992) Inflation Forecast Errors and Time Variation in Term Premia. *Journal of Financial and Quantitative Analysis*, 27, 479–496.

[10] Efron, B. (1979) Bootstrap methods: another look at the jacknife. *Annals of Statistics*, 7, 1–26.

[11] Estrella, A. & Hardouvelis, G.A. (1991) The Term Structure as a Predictor of Real Economic Activity. *Journal of Finance*, 46, 555–576.

[12] Fama, E.F. & K.R.French (1988) Permanent and Temporary Components of Stock Prices. *The Journal of Political Economy*, 96, 246–273.

[13] Frankel, J.A. & Froot, K.A. (1987) Using Survey Data to Test Standard Propositions Regarding Exchange Rate Expectations. *American Economic Review*, 77, 133–153.

[14] Gospodinov, N. (2002) Bootstrap-Based Inference in Models with a Nearly Noninvertible Moving Average Component, *Journal of Business and Economic Statistics*, 20, 254–268.

[15] Hafner, C.M. & Herwartz, H. (2000) Testing for linear autoregressive dynamics under heteroskedasticity. *Econometrics Journal*, 3, 177–197.

[16] Hall, P. (1985) Resampling a Coverage Process. *Stochastic Processes and Their Applications*, 19, 259–269.

[17] Hall, P. & Horowitz, J.L. (1993) Corrections and Blocking Rules for the Block Bootstrap with Dependent Data. Working paper 93-11, University of Iowa.

[18] Hall, P. & Horowitz, J.L. (1996) Bootstrap Critical Values for Tests Based on Generalized-Method-of-Moments Estimators. *Econometrica*, 64, 891–916.

[19] Hansen, B.E. (1999a) The Grid Bootstrap and Autoregressive model. *Review of Economics and Statistics*, 81, 594–607.

[20] Hansen, B.E. (1999b) Non-Parametric Dependent Data Bootstrap for Conditional Moment Models. Working paper, University of Wisconsin–Madison.

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

[22] Huizinga, J. & Mishkin, F.S. (1984) Inflation and Real Interest Rates on Assets with Different Risk Characteristics. *Journal of Finance*, 39, 699–714.

[23] Inoue, A. & Shintani, M. (2001) Bootstrapping GMM Estimators for Time Series. Working paper, Vanderbilt University.

[24] Kilian, L. (1998) Finite-Sample Confidence Intervals for Impulse Response Functions. *Review of Economics and Statistics*, 80, 218–230.

[25] Kilian, L. (1999) Finite-Sample Properties of Percentile and Percentile-t Bootstrap Confidence Intervals for Impulse Responses. *Review of Economics and Statistics*, 81, 652–660.

[26] Kreiss, J.-P. (1997) Asymptotic Properties of Residual Bootstrap for Autoregressions. Preprint, Institute for Mathematical Stochastics, Technical University of Braunschweig, Germany.

[27] Kreiss, J.-P. & Franke, J. (1992) Bootstrapping stationary autoregressive moving-average models. *Journal of Time Series Analysis*, 13, 297–317.

[28] Künsch, H.R. (1989) The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17, 1217–1241.

[29] Lettau, M. & Ludvigson, S. (2001) Consumption, Aggregate Wealth, and Expected Stock Returns. *Journal of Finance*, 56, 815–850.

[30] Lewis, K.K. (1990) The Behavior of Eurocurrency Returns Across Different Holding Periods and Monetary Regimes. *Journal of Finance*, 45, 1211–1236.

[31] Maddala, G.S. & Kim, I.-M. (1998) *Unit Roots, Cointegration and Structural Change*. Chapter 10, Small Sample Inference: Bootstrap Methods, 309–336. Cambridge University Press.

[32] Mammen, E. (1993) Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, 21, 255–285.

[33] Mark, N.C. (1995) Exchange Rates and Fundamentals: Evidence on Long-Run Predictability. *American Economic Review*, 85, 201–218.

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

[35] Mishkin, F. (1990b) Does Correcting for Heteroskedasticity Help? *Economics Letters*, 34, 351–356.

[36] Newey, W.K. & West, K.D. (1987) A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55, 703–708.

[37] Stambaugh, R.F. (1993) Estimiting Conditional Expectations when Volatility Fluctuates. NBER Working Paper t0140.

[38] Wu, J. (1986) Jacknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14, 1261–1343.

		Residual bootstrap									Wild bootstrap		
Ha		β≠0	β>0	β<0	β≠0	β>0	β<0	β≠0	β>0	β<0	β≠0	β>0	β<0
T	θ	DGP IID			DGP UC			DGP HS			DGP HS		
30	-0.95	6.8	6.7	8.6	6.7	6.5	8.1	8.5	7.8	9.8	7.5	6.2	9.2
	-0.8	7.0	6.4	8.2	6.6	6.7	8.0	7.8	7.3	9.3	7.6	6.7	9.3
	-0.3	6.1	6.3	7.8	6.0	6.8	7.8	6.3	7.8	7.7	6.7	7.2	7.9
	0.3	5.4	6.3	6.2	5.1	6.3	6.3	6.4	7.3	6.7	6.5	6.5	7.3
	0.8	5.9	6.5	6.5	5.7	6.6	6.3	6.7	6.5	7.9	6.4	5.2	8.1
	0.95	5.4	6.0	6.6	5.1	6.3	5.8	7.1	7.0	7.8	6.6	5.7	7.7
60	-0.95	6.2	6.1	6.4	5.7	6.4	6.3	7.3	6.4	8.1	6.5	5.3	7.3
	-0.8	5.9	6.1	6.6	5.9	6.0	6.6	7.2	7.2	7.5	6.3	5.9	7.2
	-0.3	5.3	6.0	6.7	5.2	6.1	6.4	6.0	6.7	6.7	5.4	5.7	6.2
	0.3	4.9	5.8	6.0	5.1	6.0	5.5	5.8	6.2	6.0	5.8	6.0	6.0
	0.8	5.4	5.7	6.0	5.4	5.6	5.5	6.4	5.7	6.9	5.6	4.8	6.1
	0.95	5.3	5.8	5.4	5.6	6.2	5.3	7.2	6.0	7.8	5.5	5.1	6.4
120	-0.95	5.3	5.9	5.5	5.5	5.7	6.0	6.9	6.8	7.3	5.8	4.9	6.5
	-0.8	5.7	5.5	5.7	5.2	5.5	5.7	6.9	6.6	7.0	5.3	5.2	5.8
	-0.3	5.4	5.8	6.3	5.2	5.5	5.5	5.9	6.7	5.9	5.3	5.6	5.6
	0.3	5.5	5.5	5.1	5.6	5.7	5.2	5.5	5.4	5.8	5.2	5.0	5.3
	0.8	5.2	5.6	5.4	5.5	5.4	5.6	6.6	5.4	6.8	5.3	4.5	5.9
	0.95	5.3	5.3	5.2	5.3	5.3	5.5	6.8	5.9	6.7	5.6	4.6	6.2
240	-0.95	4.8	5.3	5.4	5.4	5.4	5.6	6.5	6.0	6.1	4.8	4.6	5.2
	-0.8	5.2	5.3	5.1	5.6	5.6	5.9	6.2	5.8	6.0	5.6	4.7	6.0
	-0.3	5.3	5.4	5.3	5.3	5.1	5.5	5.6	6.1	5.5	5.0	5.0	5.3
	0.3	5.1	5.0	5.3	5.2	5.4	5.2	5.8	5.6	5.8	5.5	5.5	5.3
	0.8	5.1	5.5	5.0	5.1	5.3	5.2	6.2	5.0	6.5	5.3	4.4	6.1
	0.95	5.7	5.5	5.6	5.4	5.3	5.6	6.5	4.8	6.9	5.2	5.0	5.4

Table 1

		Nonoverlapping block bootstrap											
Ha		ℓ	β≠0	ℓ	β>0	ℓ	β<0	ℓ	β≠0	ℓ	β>0	ℓ	β<0
T	θ	DGP UC						DGP HS					
30	-0.95	4	6.6	3	4.5	3	12.1	4	6.2	3	4.1	3	11.9
		5	6.4	4	9.8	4	13.6	5	6.4	4	10.1	4	13.9
	-0.8	4	6.1	3	4.8	3	11.9	4	6.2	3	3.9	3	12.2
		5	6.3	4	9.7	4	12.7	5	6.4	4	10.4	4	13.2
	-0.3	4	5.6	3	5.0	3	11.2	4	5.3	3	5.2	3	10.1
		5	5.2	4	9.5	4	12.1	5	5.1	4	9.8	4	11.2
	0.3	4	5.0	3	6.5	3	8.6	4	5.1	3	6.0	3	8.7
		5	5.1	4	9.8	4	10.7	5	5.7	4	10.1	4	10.5
	0.8	4	5.5	3	6.4	3	9.4	4	5.8	3	6.0	3	9.7
		5	5.7	4	10.4	4	11.6	5	6.0	4	10.1	4	12.4
	0.95	4	5.5	3	6.8	3	9.0	4	6.1	3	5.6	3	9.6
		5	6.0	4	10.3	4	11.4	5	5.9	4	10.2	4	13.0
60	-0.95	4	5.7	4	5.3	4	8.6	4	5.7	4	5.0	4	9.3
		6	5.5	5	6.1	5	8.4	6	5.4	5	6.2	5	9.4
	-0.8	4	6.1	4	5.3	4	9.0	4	6.1	4	5.2	4	9.2
		6	5.2	5	6.5	5	9.4	6	5.3	5	5.6	5	8.6
	-0.3	4	5.2	4	5.6	4	8.1	4	5.0	4	5.3	4	8.1
		6	4.9	5	6.5	5	8.5	6	4.5	5	6.3	5	8.2
	0.3	4	5.2	4	5.8	4	7.8	4	4.7	4	5.8	4	6.7
		6	5.0	5	6.7	5	7.8	6	4.2	5	6.9	5	7.8
	0.8	4	5.1	4	6.1	4	7.4	4	5.3	4	5.8	4	8.1
		6	5.1	5	6.4	5	8.1	6	5.1	5	6.2	5	8.9
	0.95	4	5.0	4	6.2	4	7.0	4	5.1	4	5.4	4	8.0
		6	5.2	5	7.4	5	8.2	6	5.3	5	7.0	5	8.4
120	-0.95	5	5.4	5	5.2	5	7.3	5	5.7	5	5.3	5	8.3
		7	5.1	6	5.3	6	7.4	7	4.5	6	5.5	6	8.6
	-0.8	5	5.1	5	4.8	5	7.2	5	5.0	5	5.1	5	7.6
		7	5.0	6	5.4	6	7.1	7	4.8	6	5.6	6	7.9
	-0.3	5	5.0	5	5.4	5	6.6	5	5.0	5	5.6	5	7.0
		7	4.6	6	5.6	6	7.0	7	4.5	6	5.5	6	7.0
	0.3	5	4.7	5	5.4	5	6.3	5	4.8	5	5.7	5	6.8
		7	4.9	6	5.9	6	6.8	7	4.6	6	6.0	6	6.1
	0.8	5	5.2	5	5.6	5	6.8	5	5.3	5	5.4	5	7.7
		7	4.9	6	5.8	6	6.6	7	4.2	6	5.6	6	7.5
	0.95	5	5.0	5	5.8	5	6.7	5	5.1	5	5.6	5	7.8
		7	5.0	6	6.0	6	7.2	7	5.1	6	5.9	6	7.8
240	-0.95	6	5.0	5	5.0	5	7.0	6	5.1	5	5.1	5	7.2
		8	5.3	6	5.0	6	6.7	8	4.8	6	5.4	6	7.7
	-0.8	6	5.2	5	4.8	5	6.7	6	5.0	5	4.8	5	7.3
		8	5.1	6	5.0	6	6.7	8	4.7	6	5.2	6	7.2
	-0.3	6	4.8	5	5.0	5	6.4	6	4.5	5	5.2	5	6.4
		8	4.9	6	5.2	6	6.3	8	4.7	6	5.3	6	6.4
	0.3	6	5.0	5	5.1	5	6.0	6	4.6	5	5.2	5	6.2
		8	4.7	6	5.6	6	5.6	8	4.8	6	5.2	6	6.3
	0.8	6	5.1	5	5.2	5	6.4	6	4.6	5	5.2	5	6.5
		8	5.0	6	5.7	6	5.8	8	4.8	6	5.3	6	6.9
	0.95	6	4.9	5	5.4	5	5.8	6	4.8	5	5.1	5	7.1
		8	5.1	6	5.2	6	6.6	8	4.4	6	5.3	6	7.3

Table 2

		Overlapping block bootstrap											
Ha		ℓ	β≠0	ℓ	β>0	ℓ	β<0	ℓ	β≠0	ℓ	β>0	ℓ	β<0
T	θ	DGP UC						DGP HS					
30	-0.95	4	6.4	3	8.0	3	14.6	4	6.0	3	9.4	3	15.0
		5	5.7	4	10.8	4	14.7	5	5.8	4	11.1	4	15.8
	-0.8	4	5.8	3	8.3	3	14.2	4	5.8	3	8.9	3	15.5
		5	5.9	4	10.8	4	13.9	5	6.0	4	11.6	4	15.7
	-0.3	4	5.7	3	9.3	3	14.2	4	5.0	3	8.9	3	13.3
		5	5.7	4	11.7	4	14.2	5	4.8	4	11.5	4	13.8
	0.3	4	5.6	3	9.7	3	11.2	4	5.0	3	10.4	3	10.9
		5	5.1	4	11.7	4	12.6	5	4.6	4	12.2	4	12.8
	0.8	4	5.5	3	10.0	3	11.7	4	5.6	3	10.0	3	12.3
		5	5.4	4	11.5	4	13.3	5	5.4	4	11.4	4	14.0
	0.95	4	5.3	3	9.5	3	11.7	4	5.2	3	9.9	3	13.2
		5	5.3	4	11.8	4	13.3	5	5.2	4	11.4	4	14.2
60	-0.95	4	5.8	3	5.9	3	10.4	4	5.4	3	6.0	3	10.7
		5	5.5	4	6.9	4	10.9	5	5.4	4	7.5	4	11.3
	-0.8	4	5.2	3	5.8	3	10.6	4	5.4	3	6.0	3	10.7
		5	5.0	4	7.0	4	10.5	5	5.3	4	7.3	4	10.6
	-0.3	4	5.1	3	5.9	3	9.7	4	4.3	3	6.0	3	8.9
		5	4.3	4	7.5	4	10.3	5	4.4	4	7.4	4	9.6
	0.3	4	4.2	3	6.9	3	8.1	4	3.9	3	7.0	3	8.1
		5	4.5	4	7.5	4	8.4	5	4.5	4	7.5	4	8.4
	0.8	4	4.7	3	7.2	3	8.5	4	4.6	3	6.9	3	9.1
		5	4.6	4	7.3	4	9.2	5	4.8	4	7.0	4	9.2
	0.95	4	4.8	3	7.3	3	8.9	4	4.9	3	6.7	3	9.9
		5	4.8	4	8.2	4	8.8	5	5.1	4	7.1	4	10.3
120	-0.95	5	5.0	4	5.6	4	7.7	5	5.3	4	5.7	4	9.2
		6	4.9	5	6.1	5	8.4	6	5.2	5	6.0	5	9.5
	-0.8	5	4.7	4	5.5	4	8.3	5	5.2	4	5.6	4	8.7
		6	5.1	5	6.2	5	8.0	6	4.8	5	5.9	5	8.8
	-0.3	5	4.6	4	5.7	4	7.3	5	4.2	4	5.7	4	8.0
		6	4.3	5	6.6	5	7.6	6	4.0	5	6.3	5	7.7
	0.3	5	4.2	4	6.3	4	6.8	5	4.1	4	6.3	4	7.1
		6	4.5	5	6.2	5	7.3	6	4.2	5	6.4	5	7.5
	0.8	5	4.3	4	6.2	4	7.3	5	4.1	4	6.0	4	8.4
		6	4.8	5	6.4	5	7.6	6	4.5	5	6.4	5	7.8
	0.95	5	4.9	4	6.3	4	7.0	5	5.0	4	6.4	4	8.1
		6	4.6	5	7.0	5	7.7	6	4.2	5	6.7	5	8.7
240	-0.95	6	5.0	5	5.4	5	7.5	6	4.7	5	5.3	5	7.8
		7	4.5	6	6.0	6	6.6	7	4.6	6	5.7	6	8.1
	-0.8	6	4.9	5	5.3	5	7.2	6	4.9	5	5.3	5	7.9
		7	5.0	6	5.5	6	7.3	7	5.1	6	5.5	6	7.6
	-0.3	6	4.7	5	5.3	5	6.3	6	4.2	5	5.5	5	6.8
		7	5.1	6	5.9	6	6.4	7	4.5	6	6.0	6	6.3
	0.3	6	4.8	5	5.4	5	6.2	6	4.2	5	6.0	5	6.2
		7	4.5	6	5.8	6	6.3	7	4.5	6	5.7	6	6.2
	0.8	6	4.5	5	5.9	5	6.5	6	4.8	5	5.8	5	7.3
		7	4.6	6	6.0	6	6.1	7	4.5	6	5.8	6	7.5
	0.95	6	4.7	5	5.9	5	6.1	6	4.5	5	5.5	5	7.5
		7	4.7	6	6.1	6	6.7	7	4.5	6	5.8	6	7.9

Table 3

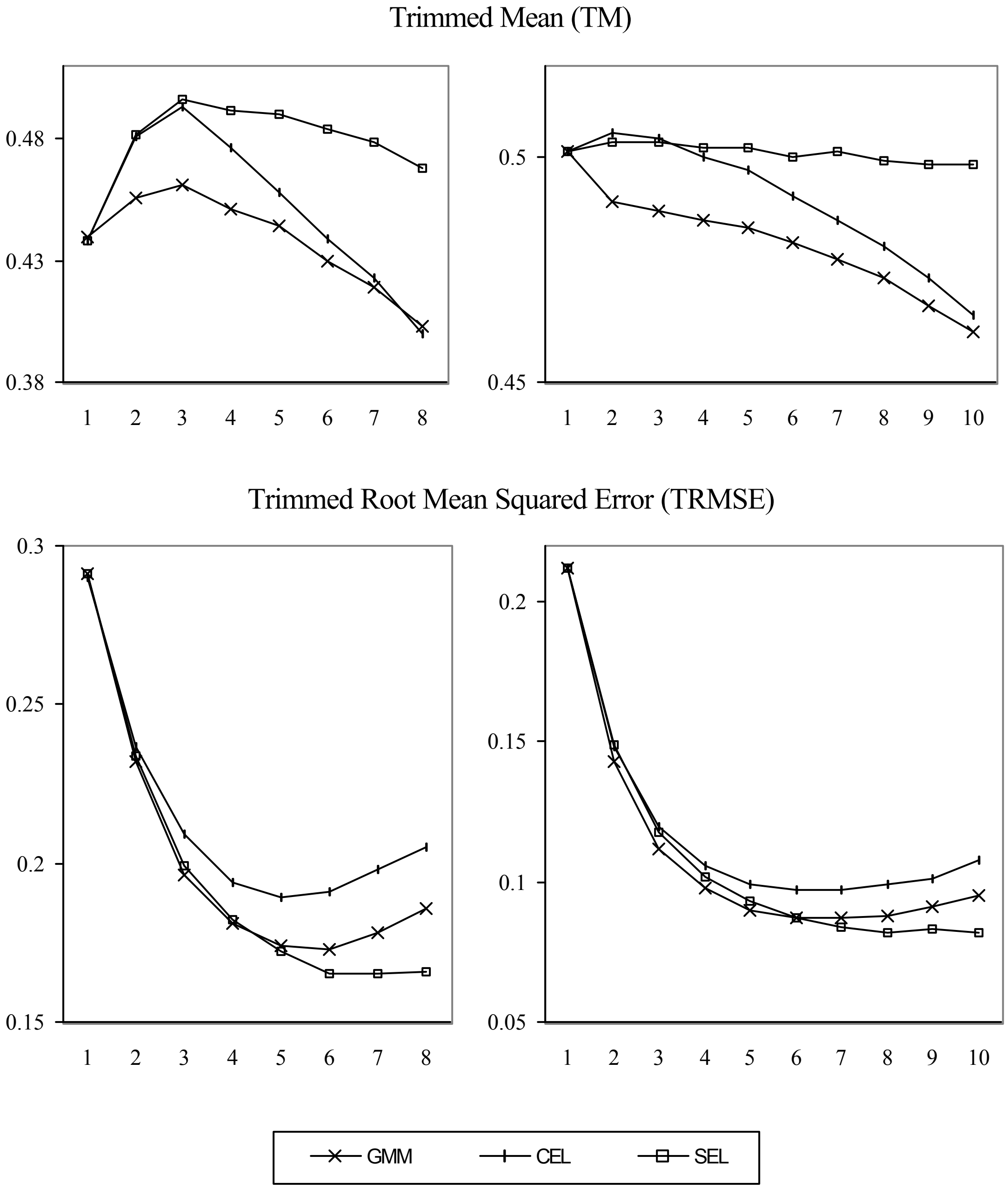


Figure 1. Simulated statistics against the number of instruments ℓ,

for T = 300 (left) and T = 900 (right)

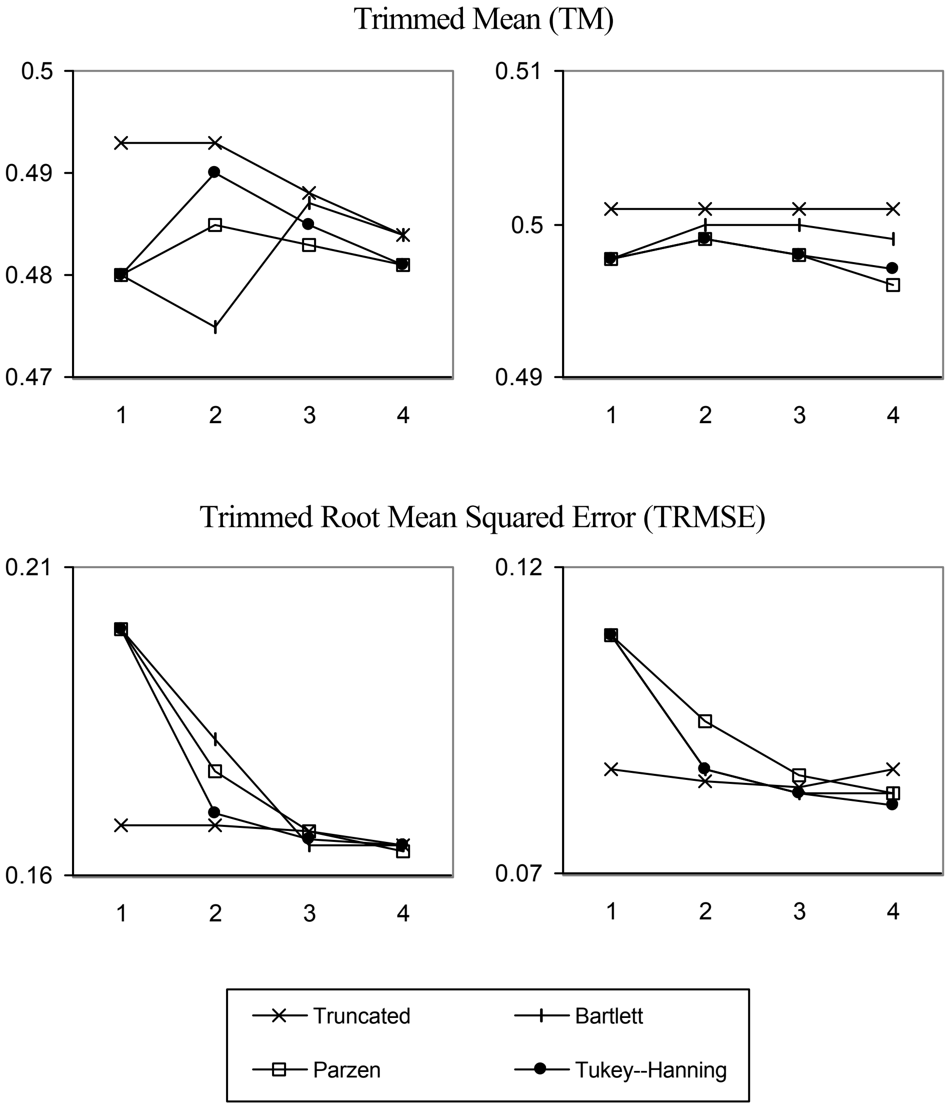
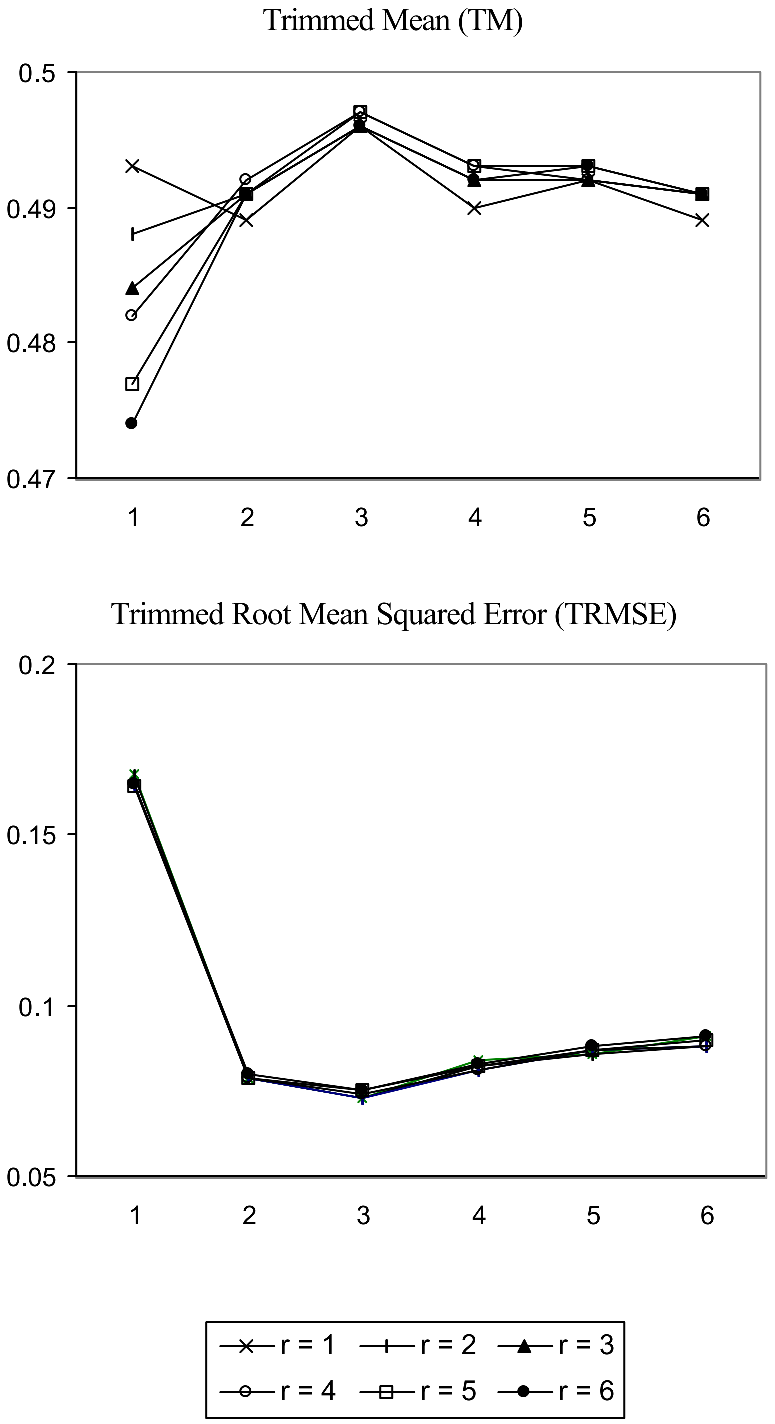


Figure 2. Simulated statistics against the lag truncation parameter r_T,

for SEL with T = 300, ℓ = 6 (left) and T = 900, ℓ = 7 (right)



Trimmed Mean (TM)

Trimmed Root Mean Squared Error (TRMSE)

Figure 3. Simulated statistics against the MA order p ,

for SEL and truncated kernel with T = 300, ℓ = 6