

INSTRUMENTAL VARIABLES ESTIMATION OF HETEROSKEDASTIC LINEAR  
MODELS USING ALL LAGS OF INSTRUMENTS

Kenneth D. West  
University of Wisconsin

Ka-fu Wong  
Chinese University of Hong Kong

Stanislav Anatolyev  
New Economic School, Moscow

October 1997  
Last Revised August 2001

ABSTRACT

We propose and evaluate a technique for instrumental variables estimation of linear models with conditional heteroskedasticity. The technique uses approximating parametric models for the projection of right hand side variables onto the instrument space, and for conditional heteroskedasticity and serial correlation of the disturbance. Use of parametric models allows one to exploit information in all lags of instruments, unconstrained by degrees of freedom limitations. Analytical calculations and simulations indicate that there sometimes are large asymptotic and finite sample efficiency gains relative to conventional estimators (Hansen (1982)). These efficiency gains are robust to minor misspecification of the parametric models.

The authors are listed in the order that they became involved in this project. We thank seminar audiences at the 1997 NBER-NSF Time Series Conference, the University of Michigan, Michigan State University and the 2000 Econometric Society World Congress for helpful comments, and the National Science Foundation and a Direct Grant from the Chinese University of Hong Kong for financial support. Correspondence: Kenneth D. West, Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706. Email: [kdwest@facstaff.wisc.edu](mailto:kdwest@facstaff.wisc.edu).

This paper proposes and evaluates a technique for instrumental variables estimation of linear time series models with conditionally heteroskedastic disturbances that may also be serially correlated. Our aim is to provide a set of tools that will yield improved estimation and inference.

Equations such as the ones we consider arise often in macroeconomics and finance. One class of applications evaluates the ability of one variable or set of variables to predict another, perhaps over a multiperiod horizon. Examples include forward exchange rates as predictors of spot rates (e.g., Hodrick (1987)), nominal interest rates as predictors of inflation (e.g., Mishkin (1992)), dividend yields and interest rate spreads as predictors of stock returns (e.g., Fama and French (1988)), and survey responses as predictors of economic data (e.g., Ball and Croushore (1995)). A second class evaluates a first order condition or decision rule from an economic model. Examples here include the log-linearized consumption based intertemporal asset pricing model (e.g., Kaminsky and Peruga (1990)), and, more generally, models with costs of adjustment (e.g., Kennan (1979)), moving average shocks (e.g., Kollintzas (1993)), time aggregation (e.g., Campbell and Mankiw (1990)), or combinations of these (e.g., Oliner et al. (1996)).

Two techniques are commonly used in these and related applications. The first is maximum likelihood (Bollerslev and Wooldridge (1992)). In models with many variables and moving average disturbances, however, maximum likelihood can be computationally cumbersome. Such applications are therefore often estimated with a second technique, instrumental variables. Typically, investigators use an instrument list of fixed, small dimension, applying Hansen (1982). We call this technique “conventional GMM” or “conventional instrumental variables.” A recent literature has, however, documented that in some environments conventional GMM suffers from a number of finite sample deficiencies. See, for example, the January 1996 issue of the Journal of Business and Economic Statistics.

These deficiencies motivate our attempt to develop an alternative estimator with better asymptotic and therefore (one hopes) better finite sample properties. Our starting point is the observation that in many time series models, the number of potential instruments is arbitrarily large for an arbitrarily large sample: usually, if a given variable is a legitimate instrument, so,

too, are lags of that variable. Moreover, when the regression disturbance displays conditional heteroskedasticity or serial correlation, use of additional instruments typically delivers increased asymptotic efficiency. In conditionally homoskedastic environments, instrumental variables estimators that efficiently use all available lags are developed in Hayashi and Sims (1983) and Hansen (1985, 1986), applied in Hall (1988) and simulated in West and Wilcox (1996). This work has shown that the asymptotic benefits of using all available lags as instruments sometimes are large, and that the asymptotic benefits may be realized in samples of size available.

A less well developed literature has studied similar environments in which conditional heteroskedasticity is present. Asymptotic calculations in Broze et al. (2001) indicate that using lags can greatly increase efficiency, even in a simple autoregressive model. A basic theoretical reference is Hansen (1985), applied by Tauchen (1986) in a model with a serially uncorrelated disturbance, and exposted in West (2000). Bates and White (1993) provide extensions to environments with heterogenous data. Hansen, Heaton and Ogaki (1988) build on Hansen (1985) to present an elegant and general characterization of an efficiency bound; they do not, however, indicate how to construct a feasible estimator that achieves the bound. Heaton and Ogaki (1991) show how to achieve the bound in a specific example, but their results do not appear to immediately generalize. Finally, Kuersteiner (1996) characterizes a bound for a univariate autoregressive model with a serially uncorrelated disturbance, and Breusch et al. (1999) give conditions under which a finite set of additional lags will increase efficiency.

In this paper, we propose and evaluate a technique for instrumental variables estimation of linear models in which the disturbances display conditional heteroskedasticity and, possibly serial correlation. The set of instruments that we allow consists of time-invariant distributed lags on a pre-specified set of variables that we call the “basic instruments.” The disturbances may be correlated with right hand side variables. As well, the Wold innovations in the disturbances may be correlated with the instruments (though the disturbances themselves are of course uncorrelated with the instruments). As Hayashi and Sims (1983) emphasize, non-zero correlations between

instruments and disturbance innovations arises in a wide class of economic models, and precludes use of filtering such as that of generalized least squares.

Our estimator posits parametric forms for conditional heteroskedasticity and for the process driving the instruments and regressors. The procedure does not require correct parameterization; we allow for the possibility that (say) the investigator models conditional heteroskedasticity as an ARCH(1) process (Engle (1982)) when in fact the correct model is GARCH(1,1) (Bollerslev (1986)). An Additional Appendix available on request shows that under commonly assumed technical conditions, the estimator converges at the usual  $\sqrt{T}$  rate, with a variance-covariance matrix that can be consistently estimated in the usual way. If, as well, the assumed parametric forms are correct, the estimator achieves an asymptotic efficiency bound.

We use asymptotic theory and simulations to compare our estimator to one that uses a small and fixed number of instruments, in a simple scalar model, with conditional heteroskedasticity. As in the conditionally homoskedastic environments of Hansen and Singleton (1991) and West and Wilcox (1996), we find that our estimator has decided asymptotic advantages when the regression disturbance has a moving average root near unity. Whatever the characteristics of the moving average roots, the estimator's asymptotic advantages are larger the greater the persistence in the conditional heteroskedasticity of the disturbance, a result consistent with the calculations in Stambaugh (1994). Simulations indicate that the asymptotic approximation can work well. The finite sample behavior of our estimator generally dominates that of the conventional estimator, even when we misspecify, albeit in a minor way, the parametric form of the data generating process. As in Tauchen (1986) and West and Wilcox (1996), we find that the conventional estimator works poorly when the dimension of the instrument vector is large.

Section 2 describes our setup and estimator. For some simple, stylized data generating processes, Section 3 provides asymptotic comparisons of the optimal and conventional GMM estimators, Section 4 simulation evidence. Section 5 concludes. Throughout, our presentation is

relatively non-technical. A lengthy appendix that is available on request has formal assumptions and proofs, as well as simulation results omitted from the paper to save space.

## 2. THE ENVIRONMENT AND OUR ESTIMATOR

The linear regression equation and vector of what we call “basic” instruments are:

$$(2.1) \quad y_t = X_t' \beta + u_t, \quad u_t \sim \text{MA}(q), \quad z_t \text{ the “basic” instruments, with Wold innovation } e_t.$$

$(1 \times 1) \quad (1 \times k)(k \times 1) \quad (1 \times 1) \quad (r \times 1) \quad (r \times 1)$

In (2.1), the scalar  $y_t$  and the vectors  $X_t$  and  $z_t$  are observed, and  $\beta$  is the unknown parameter vector to be estimated. For simplicity, and in accordance with the leading class of applications (see the references in the previous section), the unobservable disturbance  $u_t$  is assumed to follow a finite order MA process of known order  $q$  ( $q=0 \Rightarrow u_t$  is serially uncorrelated). In addition to a constant term, there is a  $(r \times 1)$  vector of “basic” instruments  $z_t$  that can be used in estimation, with  $(r \times 1)$  Wold innovation  $e_t$ . The adjective “basic” distinguishes  $z_t$  from its lags  $z_{t-j}$ , which also can be used as instruments. The dimension of the basic instrument vector ( $r$ ) may be larger or smaller than that of the coefficient vector ( $k$ ).

We assume that there is a single equation rather than a set of equations, and that the only non-stochastic instrument is a constant term, for algebraic clarity and simplicity. The results directly extend to multiple equation systems (see the Appendix). They do so as well if one (say) uses four seasonal dummies instead of a constant or if one omits non-stochastic terms altogether from the instrument list (see the discussion below).

Let  $T$  be the sample size. It is notationally convenient for us to express GMM estimators as instrumental variables estimators. We consider estimators that can be written

$$(2.2) \quad \hat{\beta} = (\Sigma_{t=1}^T \hat{Z}_t X_t')^{-1} (\Sigma_{t=1}^T \hat{Z}_t y_t)$$

for a  $(k \times 1)$  vector  $\hat{Z}_t$  that depends on  $z_t, z_{t-1}, \dots, z_1$  in a (possibly) sample dependent way.

Let us map conventional GMM in this framework, using an illustrative but arbitrarily

chosen set of lags of  $z_t$ . Define the  $(2r+1) \times 1$  vector  $W_t = (1 \ z_t' \ z_{t-1}')$ . Suppose that we optimally exploit the moment condition  $EW_t u_t = 0$ . Define the  $(2r+1) \times (2r+1)$  matrix  $B = \Sigma_{i=-q}^q E(W_{t-i} u_{t-i} u_t' W_t')$ , assumed to be of full rank. Let  $\hat{B}$  be a feasible counterpart that converges in probability to  $B$ . The GMM estimator chooses  $\hat{\beta}$  to minimize  $(T^{1/2} \sum_{s=1}^T W_s u_s)' \hat{B}^{-1} (T^{1/2} \sum_{s=1}^T W_s u_s)$ . Then of course  $\hat{\beta} = (\Sigma_{t=1}^T \hat{Z}_t X_t')^{-1} (\Sigma_{t=1}^T \hat{Z}_t y_t)$  with  $\hat{Z}_t = (T^{-1} \sum_{s=1}^T X_s W_s') \hat{B}^{-1} W_t$ . Evidently,  $\hat{Z}_t$  may also be written as  $\hat{Z}_t = \hat{A} + \hat{A}_0 z_t + \hat{A}_1 z_{t-1}$  for  $(k \times 1)$   $\hat{A}$  and  $(k \times r)$   $\hat{A}_0$  and  $\hat{A}_1$  that satisfy  $(\hat{A} \ \hat{A}_0 \ \hat{A}_1) = (T^{-1} \sum_{s=1}^T X_s W_s') \hat{B}^{-1}$ .

Our aim is to efficiently exploit the information in not just two lags of  $z_t$  but in all lags. One way to do so is to use conventional GMM estimation, with the number of lags of  $z_t$  used increasing suitably with sample size. Koenker and Machado (1997) establish a suitable rate of increase for a linear model with disturbances that are independent over time. Related theoretical work includes Newey (1988) and Kuersteiner (1996), while Tauchen (1986) presents simulation evidence. Unfortunately, much simulation evidence, including the evidence presented below, has shown that in samples of size typically available, estimators that use many lags have poor finite sample performance. Accordingly, we try another approach.

In our approach, we work with  $z_t$ 's Wold innovation  $e_t$  rather than with  $z_t$  because popular models for conditional heteroskedasticity, such as GARCH, typically are written in terms of innovations. Thus, we shall describe how we propose to fully exploit information available in linear combinations of lags of  $e_t$ , with obvious mapping back to  $z_t$ . To describe our procedure, we begin with a non-feasible estimator. Let  $T$  be the sample size. Define

$$(2.3) \quad \underset{(1+Tr) \times 1}{e(t)} = (1, e_t', \dots, e_{t-T+1}')', \quad \Psi = \underset{(1+Tr) \times k}{Ee(t)X_t'}, \quad S = \underset{(1+Tr) \times (1+Tr)}{\Sigma_{i=-q}^q E[e(t-i)u_{t-i}u_t e(t)']}, \quad \underset{(k \times 1)}{Z_t} = \Psi' S^{-1} e(t).$$

We omit a  $T$  subscript on each of these quantities for notational simplicity.

Consider the nonfeasible estimator of  $\beta$  that uses  $Z_t$  as an instrument:

$(\Sigma_{t=1}^T Z_t X_t')^{-1} (\Sigma_{t=1}^T Z_t y_t)$ . (This is not feasible since the moments required to compute  $\Psi$  and  $S$  are



estimate of the long-run variance of  $Z_t^* u_t$ . ( $Z_t^*$  is the large sample  $[T \rightarrow \infty]$  counterpart to  $Z_t$  defined in (2.3).)  $\hat{\Omega}$  may be computed with techniques such as Andrews (1991), Newey and West (1994), or den Haan and Levin (1996), using data on  $\hat{Z}_t^*$  and  $\hat{u}_t$ , where  $\hat{u}_t$  is a residual obtained with an initial consistent estimate of  $\beta$ .  $\hat{V}$  provides a consistent estimator of the asymptotic variance-covariance matrix of  $\hat{\beta}$  even if the parametric specification is not correct. The second method is to set  $\hat{V} = (\hat{\Psi}' \hat{S}^{-1} \hat{\Psi})^{-1}$ . This method has the advantage of computational simplicity, since one will compute  $\hat{\Psi}$  and  $\hat{S}^{-1}$  in any case. It has the disadvantage that it is consistent only if the parametric specification is correct. We show in our asymptotic calculations and simulations, however, that if the parameter specification is incorrect in minor ways, this second method still works tolerably well.

A simple example may clarify. Suppose  $y_t = \beta_0 + \beta_1 z_t + u_t$ , where  $u_t \sim \text{MA}(1)$  ( $q=1$ ), the scalar  $z_t$  is the sole element of the basic instrument vector ( $r=1$ ), and  $X_t' = (1 \ z_t)$  ( $k=2$ ). (So in this simple example least squares is consistent.) Suppose as well that  $u_t$  and  $e_t$  satisfy  $0 = Eu_t^2 e_{t-j} e_{t-m} = Eu_t^2 e_{t-j} = Eu_t u_{t+1} e_{t-j} e_{t-m} = Eu_t u_{t+1} e_{t-j}$ ,  $j \neq m$ ,  $j, m \geq 0$ , a condition that holds for GARCH models with conditionally symmetric disturbances. Then

(2.5)

$$\begin{aligned}
 S = & \begin{pmatrix} (Eu_t^2 + 2Eu_t u_{t+1}) & 0 & 0 & 0 & \dots & 0 & 0 & 0 & ) \\
 (0 & Ee_t^2 u_t^2 Ee_t^2 u_{t+1} u_t & 0 & 0 & \dots & 0 & 0 & 0 & ) \\
 (0 & Ee_t^2 u_{t+1} u_t & Ee_{t-1}^2 u_t^2 & Ee_{t-1}^2 u_{t+1} u_t & 0 & \dots & 0 & 0 & ) \\
 (0 & 0 & Ee_{t-1}^2 u_{t+1} u_t & Ee_{t-2}^2 u_t^2 Ee_{t-2}^2 u_{t+1} u_t & \dots & 0 & 0 & 0 & ), \\
 (0 & 0 & 0 & Ee_{t-2}^2 u_{t+1} u_t & Ee_{t-3}^2 u_t^2 & \dots & 0 & 0 & 0 & ) \\
 (\dots & & & & & & & & & ) \\
 (0 & 0 & 0 & 0 & 0 & \dots & Ee_{t-T+3}^2 u_t^2 & Ee_{t-T+3}^2 u_{t+1} u_t & 0 & ) \\
 (0 & 0 & 0 & 0 & 0 & \dots & Ee_{t-T+3}^2 u_{t+1} u_t & Ee_{t-T+2}^2 u_t^2 & Ee_{t-T+2}^2 u_{t+1} u_t & ) \\
 (0 & 0 & 0 & 0 & 0 & \dots & 0 & Ee_{t-T+2}^2 u_{t+1} u_t & Ee_{t-T+1}^2 u_t^2 & ) \end{pmatrix} \\
 \Psi = & \begin{pmatrix} (1 & Ez_t & ) \\
 (0 & Ee_t z_t & ) \\
 (0 & Ee_{t-1} z_t & ) \\
 (\dots & & ) \\
 (0 & Ee_{t-T+1} z_t & ) \end{pmatrix} .
 \end{aligned}$$

$\hat{\Psi}$  and  $\hat{S}$  are obtained by replacing the elements of  $\Psi$  and  $S$  with estimates. Suppose, for example, that  $z_t$  is modeled as an AR(1),  $z_t = \phi_0 + \phi z_{t-1} + e_t$ ,  $|\phi| < 1$ ,  $Ee_t^2 = \sigma_e^2$ , with corresponding estimates  $\hat{\phi}_0$ ,  $\hat{\phi}$  and  $\hat{\sigma}_e^2$ . (Here,  $\phi_0$ ,  $\phi$  and  $\sigma_e^2$  are elements of the parameter vector  $b$ .) Then in  $\hat{\Psi}$ , the estimate of  $Ee_{t-j}z_t$  is  $\hat{\phi}^j \hat{\sigma}_e^2$ .<sup>2</sup>  $\hat{S}(1,1)$  may be set to  $\hat{\sigma}_u^2 + 2\hat{\sigma}_{u,1}$ , where  $\hat{\sigma}_u^2$  and  $\hat{\sigma}_{u,1}$  are estimates of  $\sigma_u^2 \equiv Eu_t^2$  and  $\sigma_{u,1} \equiv Eu_t u_{t+1}$ ;  $\sigma_u^2$  and  $\sigma_{u,1}$  are also elements of  $b$ . One obtains  $\hat{\sigma}_u^2$  and  $\hat{\sigma}_{u,1}$  from the residuals from an initial consistent estimator of  $\beta$  (e.g., least squares, in the present example), either by directly computing moments or estimating an MA model. The other diagonal elements of  $\hat{S}$  may be constructed from a GARCH or other model applied to  $\hat{e}_t$  and  $\hat{u}_t$ , as illustrated in the simulations below.

Let us now return to our general discussion to make several remarks. First, one could write the instrument as a distributed lag on  $z_t$  rather than  $\hat{e}_t$ ; in the scalar AR(1) illustration of the previous paragraphs, for example, one could substitute out for  $\hat{e}_t$  using  $\hat{e}_t = z_t - \hat{\phi}_0 - \hat{\phi}z_{t-1}$ . We formulate our estimator in terms of the  $\hat{e}_t$ 's because in most applications it will be a little simpler and more convenient: popular models for conditional heteroskedasticity such as GARCH and stochastic volatility models are written in terms of innovations.

Second, to use an alternative set of nonstochastic instruments, simply replace the “1” that appears in the equation (2.3) definition of  $e(t)$  with the relevant set of nonstochastic terms. The equation (2.3) definitions of  $S$ ,  $\Psi$ , and the mechanics described below equation (2.3), remain unchanged. For example, if one is using zero mean data, and thus has no need of a constant term as an instrument,  $e(t)$  is redefined to omit the constant term;  $e(t)$  will then have dimension  $Tr \times 1$ ,  $\hat{S}$  will have dimension  $Tr \times Tr$ , etc.

Third, our feasible estimator has attractive asymptotic properties. Let  $b$  denote the  $(m \times 1)$  probability limit of  $\hat{b}$ :  $\hat{b} \xrightarrow{p} b$ . Suppose first that our parametric models for  $\Psi$  and  $S$  are correct. That is, suppose that  $S = S(b)$ ,  $\Psi = \Psi(b)$ . ( $S$  and  $\Psi$  are defined in terms of moments of the data in (2.3);  $S(b)$  and  $\Psi(b)$  are the quantities that result when the parametric models are used, evaluated at the population parameter vector  $b$ .) Then under standard conditions, our estimator attains an

asymptotic efficiency bound, and uses information in all lags of  $z_t$ . Suppose, instead, that the difference between  $S$  and  $S(b)$ , or between  $\Psi$  and  $\Psi(b)$ , is not zero. Then our estimator is still asymptotically normal with a variance-covariance matrix that can be estimated in familiar ways. Again, see the Additional Appendix for a formal statement.

Fourth, we have emphasized that our procedure allows one to use all available lags of  $e_t$ . There are, of course, diminishing returns to such usage; as a formal matter, one can capture an arbitrarily large amount of the efficiency gains of all available lags by using an arbitrarily large but finite number of such lags. That is, one can use (2.3) and (2.4) but with  $e(t) \equiv (1, e_t', \dots, e_{t-J+1}')$  for some  $J < T$ . In practice, one can see how rapidly the  $\hat{g}_j$ 's die down for a couple of trial  $J$ 's. In our data generating processes, which included some highly persistent specifications,  $J=50$  was pretty much sufficient to yield an estimator whose asymptotic variance was indistinguishable from that of the optimal estimator (though because we are compulsive we set  $J=100$ ).

Fifth, as noted above, an alternative way to fully exploit information in linear combinations of past  $z_t$ 's would be to estimate with conventional GMM, letting the number of lags of  $z_t$  used in estimation increase with sample size. We view our parametric approach as complementary rather than competing. Our procedure has the disadvantage that if the parametric specification is incorrect, we will not obtain the efficiency bound: under such misspecification, our estimator may be more efficient or less efficient asymptotically than conventional GMM with a given number of instruments. On the other hand, our procedure appears to have some finite sample advantages relative to conventional GMM, at least if the parametric specification is approximately correct. (See the simulations presented below.) The improved performance may well be tied to the smaller number of parameters required by our estimator to construct the linear combination of instruments. We require estimation of a vector  $b$  that includes the parameters of the time series processes for  $(X_t' z_t)'$  and  $(e_t' u_t)'$ . The reader familiar with the forecasting and conditional volatility literature will recognize that with much data a handful of parameters will likely be adequate. That may not be the case if one is attempting to nonparametrically pick up the

information in many lags of  $z_t$ .

Sixth, in certain cases our estimator specializes to ones discussed in earlier work. If conditional heteroskedasticity is absent, i.e., if  $E(u_t u_{t-j} | e_t, e_{t-1}, \dots) = E u_t u_{t-j}$ , our instrument is asymptotically that given in Hansen (1985, section 5.2). If conditional heteroskedasticity is present but there is no serial correlation in  $u_t$ , and, further, the model is a univariate autoregression ( $X_t$  consists of a constant and lags of  $y_t$ ,  $u_t = e_{t+1}$ ,  $z_t = y_{t-1}$ ), our instrument is asymptotically that of Kuersteiner (1996). Our estimator allows for both serial correlation and conditional heteroskedasticity.

Seventh and finally, while the class of estimators we consider includes the ones we have called conventional GMM, it does not include maximum likelihood (except in certain cases that are special and from our point of view uninteresting) or instrumental variables estimators in which the instruments asymptotically depend on stochastic combinations of lagged  $z_t$ 's or  $e_t$ 's. An example of the latter is weighted least squares. This broader class of instruments brings no asymptotic efficiency gains when the regression disturbance  $u_t$  is homoskedastic conditional on the  $z_t$ 's, but it does improve efficiency in the presence of conditional heteroskedasticity (Hansen (1985)). For instrumental variables estimators with asymptotically stochastic combinations, see Hansen, Heaton and Ogaki (1988) for a theoretical efficiency bound and Anatolyev (1998) for theory and simulation evidence. Once again, we consider our research complementary to parallel research on feasible procedures to exploit asymptotically stochastic combinations. Under misspecification of the parametric model needed to construct the optimal instruments, there is no theoretical presumption that a procedure to exploit stochastic combinations is asymptotically more efficient than ours. And of course there is no presumption that one class of estimators will perform better than the other in finite samples. In short, for empirical work, it will be important to have asymptotic and finite sample evidence on the behavior of both classes of estimators.

### 3. ASYMPTOTIC COMPARISONS

This section uses a very simple model to compare the asymptotic variances of conventional and optimal GMM estimators of a scalar regression parameter. The aim is to see what data characteristics imply large efficiency gains when moving from the conventional to the optimal estimator. A secondary aim is to see whether minor misspecification of the parametric form of the data generating process (DGP) substantially lessens efficiency gains that result under correct specification.

The model we use has an MA(1) disturbance driven in whole or in part by GARCH(1,1) innovations: For all but the DGPs involving misspecification (detailed below), the DGP was:

$$(3.1a) \quad y_t = \beta_0 + z_t\beta_1 + u_t,$$

$$(3.1b) \quad u_t = e_{t+2} - \theta e_{t+1} + v_{t+2} - \delta v_{t+1}$$

$$(3.1c) \quad z_t = \phi z_{t-1} + e_t,$$

$$(3.1d) \quad e_t = \sigma_t \eta_t, \quad \eta_t \sim \text{i.i.d.}(0,1), \quad \sigma_t^2 = \omega + \gamma_1 e_{t-1}^2 + \gamma_2 \sigma_{t-1}^2, \quad \gamma \equiv \gamma_1 + \gamma_2, \quad E\eta_t^4 = 3 + \kappa_\eta,$$

$$(3.1e) \quad v_t \sim \text{i.i.d.}(0, \sigma_v^2), \quad v_t \text{ independent of } e_t.$$

All variables are scalars, and, as detailed below, parameters are restricted to insure stationarity (e.g., in (3.1c)  $|\phi| < 1$ ). The parameter of interest is  $\beta_1$  in (3.1a). A constant and lags of  $z_t$  (equivalently,  $e_t$ ) may be used as instruments. For simplicity, most of the simulations set  $\sigma_v^2 = 0$  ( $\Rightarrow u_t = e_{t+2} - \theta e_{t+1}$ ) and  $\kappa_\eta = 0$  ( $\eta_t \sim N(0,1) \Rightarrow \kappa_\eta = 0$ ); we discuss below our motivation for allowing  $\sigma_v^2 \neq 0$  and  $\kappa_\eta \neq 0$ .

We let GMM $n$  denote conventional GMM with an instrument vector that includes a constant and lags 0 through  $n-1$  of  $z_t$  ( $n=1 \Rightarrow$  OLS). The familiar formulas for this estimator are presented in the next section.

When  $\sigma_v^2 = 0$  and  $\kappa_\eta = 0$ , we used 2 values of the autoregressive parameter  $\phi$ , 7 values of the moving average parameter  $\theta$ , 5 values of the GARCH parameter  $\gamma$ , 2 values of the GARCH parameter  $\gamma_1$ , 140 ( $= 2 \times 7 \times 5 \times 2$ ) combinations of parameters altogether. We then used a representative subset of these in conjunction first with some non-zero values of  $\kappa_\eta$ , and then with

some nonzero values  $\sigma_v^2$  and  $d$ . For each combination, we computed the asymptotic variances of GMM1 (=least squares), GMM4, GMM12 and optimal GMM. We will report typical results, in the form of the ratio of the variance of conventional to optimal GMM, commenting briefly on patterns reflected in the many unreported results.

The ratio of the asymptotic variance of conventional to optimal GMM is strictly greater than one, with the ratio declining towards one as the number of lags increases. We aim to see what characteristics of the data lead to large ratios, and how rapid is the approach towards one. In connection with data characteristics, we observe that if  $u_t$  were a textbook disturbance (serially uncorrelated and homoskedastic, conditional on  $z_t$ ), the ratio would be one for GMM with any set of lags (that is, ordinary least squares is efficient). Intuition thus suggests that there will be relatively big gains when serial correlation or conditional heteroskedasticity in  $u_t$  is particularly marked.

Specifically, when  $\sigma_v^2=0$  and  $\kappa_\eta=0$ , the parameter values were:

$$(3.2) \quad \phi = 0.5, 0.9; \theta = -0.9, -0.5, 0, 0.5, 0.7, 0.9, 0.95; \gamma = 0.5, 0.6, 0.7, 0.8, 0.9; \\ \gamma_1 = 0.1, 0.3.$$

The positive values of  $\phi$  were chosen to reflect the positive autocorrelation typically present in time series data, with the larger value of  $\phi$  capturing near unit root behavior. The wide range of values of  $\theta$  reflect the wide range found in empirical work. For example, negative first order autocorrelation of regression residuals (implying positive  $\theta$ ) has been found in inventory work (West and Wilcox (1996)); positive first order autocorrelation (implying negative  $\theta$ ) will result from time aggregation. High persistence in conditional variance is captured by the relatively high values of  $\gamma$ , with exceptionally fat tails in  $e_t$  resulting when  $\gamma_1=0.3$ .

Table 1 reports a few representative results when  $\sigma_v^2=0$  and  $\kappa_\eta=0$ . Lines 1 and 2 present results when  $\theta=0$ , so that  $u_t$  is serially uncorrelated. To read the table, consider line 2. The "1.23" in the "GMM1" column says that when  $\phi=0.9$ ,  $\gamma=.9$ ,  $\gamma_1=.3$ , the asymptotic variance of

the least squares estimator is 23 percent higher than that of the optimal estimator. The “1.00” in all three “GMM” columns in line 1 indicates that when  $\gamma_1 = .1$ , efficiency gains show up only in the third decimal point or later: the asymptotic variance of GMM $n$  is at most .5 percent higher than that of optimal GMM. Recall that  $\gamma_1 = .3$  implies fatter tails (more kurtosis) than  $\gamma_1 = .1$ . As noted in Stambaugh (1994), fatter tails implies greater efficiency to use of additional lags.

Lines 3 through 6 hold fixed the GARCH parameters, at values that involve persistence in conditional variances ( $\gamma = .9$ ) but not particularly fat tails ( $\gamma_1 = .1$ ), and mild persistence of the regressor ( $\phi = .5$ ). These lines differ only in the value of the moving average coefficient  $\theta$ , which increases from  $\theta = -.5$  (implying a positive autocorrelation to  $u_t$ ) to  $\theta = .95$  (implying a negative autocorrelation). Values of  $\theta$  near 1 yield sharp efficiency gains relative to least squares: for  $\theta = .9$  and  $\theta = .95$ , the least squares variance is over three times that of the optimal estimator. By the time  $n = 12$  lags are used, sharply diminishing returns have set in; the largest efficiency loss is when  $\theta = .95$ , and even here the conventional estimator with 12 lags has an asymptotic variance only 11 percent larger than the optimal. Negative autocorrelation in the disturbance ( $\theta > 0$ ) leads to larger efficiency gains than positive autocorrelation ( $\theta < 0$ ), a result also found in conditionally homoskedastic environments (Hansen and Singleton (1996), West and Wilcox (1996)).

Lines 7 through 12 increase the autoregressive parameter  $\phi$  to .9, with a variety of values for the other parameters. Upon comparing lines 7 and 5, or lines 10 and 6, we see the larger value of  $\phi$  increases the relative efficiency of optimal GMM. (This result, however, is not uniform; for  $\phi = .9$ ,  $\theta = .5$ ,  $\gamma = .9$ ,  $\gamma_1 = .1$ , the ratio for GMM1 is 1.21 [not reported in the table], which is slightly lower than the 1.36 reported in line 3.) Line 8 suppresses conditional heteroskedasticity in  $u_t$ . Upon comparing the entries in line 8 with those in lines 9 and 10, and similarly comparing lines 10 and 1, we see that the relative efficiency of the optimal estimator is larger when there is both conditional heteroskedasticity and serial correlation than just heteroskedasticity or correlation. Dramatic gains in efficiency, however, are attributable to correlation rather than heteroskedasticity. This last result may, however, be sensitive to the

assumed form of heteroskedasticity (Broze et al. (2001)).

Finally, line 12 allows  $|\theta| > 1$ . This specification is included largely to remind the reader that in the relevant class of applications, the Wold innovation in the disturbances may be correlated with the instruments (Hayashi and Sims (1983), Hansen and Sargent (1980)). (Recall that if  $u_t = e_{t+2} - \theta e_{t+1}$  with  $|\theta| > 1$ , the Wold representation of  $u_t$  is  $u_t = \epsilon_t - (1/\theta)\epsilon_{t-1}$  with  $\epsilon_t$  a distributed lag on current and past  $e_{t+2}$ 's.) In a conditionally homoskedastic environment, the efficiency gains would be the same for  $u_t = e_{t+2} - \theta e_{t+1}$  and  $e_{t+2} - (1/\theta)e_{t+1}$ ; the presence of conditional heteroskedasticity, however, changes instruments, and, accordingly, the numbers in line 12 are different, though not by much, from those in line 10.

Table 2 presents selected results when the sequence driving  $e_t/\sigma_t$  has fatter tails than does a normally distributed variable (panel A) and when two noises rather than one drive the regression disturbance  $u_t$  (panel B). The baseline specification is in line 10 of Table 1. For convenience, the results from this specification are repeated in line 1 of panel A.

In line 2 of panel A, conditional excess kurtosis is 1, about what was estimated for daily exchange rate data by Bollerslev (1987); in line 3 the value is 3.9, which implies unconditional excess kurtosis of about 7, a figure reported for monthly stock market data by Stambaugh (1994). The increase in the values from line 1 to line 2 and from line 2 to line 3 indicates that fatter tails (increased excess kurtosis) lead to greater efficiency gains for the optimal estimator. This result was already noted in Table 1's discussion of results with  $\gamma_1 = .3$ , and is also reflected in unreported experiments with still other values for conditional and unconditional excess kurtosis. But the gains are the same order of magnitude as for conditionally normal disturbances.

Panel B of Table 2 allows two noises rather than one to drive the regression disturbance  $u_t$ ,  $u_t = e_{t+2} - \theta e_{t+1} + v_{t+2} - \theta v_{t+1}$ ,  $v_t \sim \text{i.i.d. } N(0, \sigma_v^2)$ . In some of the relevant applications, the regression disturbance is a moving average of an unknown number of white noise signals. (In panel B, the number is two, with  $e_t$  and  $v_t$  being the signals.) This multiplicity of signals can lead to the innovation in the Wold decomposition of  $u_t$  being correlated with the instruments, thus

precluding conventional GLS filtering.<sup>3</sup>

Panel B presents results in which parameters were chosen so that half the variance of  $u_t$  was due to each signal. The first row of the panel sets  $d=0.95$ , implying that the MA coefficient in the univariate Wold representation of  $u_t$  (presented in the column “implied MA parameter for  $u_t$ ”) also is 0.95. The disturbance is then a mixture of the heteroskedastic one in line 1 of panel A of the table and the homoskedastic one in line 8 of Table 1, and, unsurprisingly, the figures in line 1 of panel B lie between those in line 1 of panel A and line 8 of Table 1. If one increases (decreases) the fraction of the variance of  $u_t$  due to  $e_{t+2}-\theta e_{t+1}$ , holding  $d$  and  $\theta$  fixed at 0.95, the figures move closer to (farther from) those in line 1 of panel A (not reported in the tables). Rows 2 and 3 experiment with values of  $d$  not equal to 0.95. As the implied MA parameter falls, efficiency gains fall as well, a result also confirmed by experiments with values of  $d$  not reported in the tables. The perhaps unsurprising implication of this panel, then, is that the serial correlation and conditional heteroskedasticity of the regression disturbance, rather than the serial correlation of one or more its underlying components, determines the asymptotic benefits of using the optimal estimator.

Our final asymptotic calculations involve the DGPs and procedures used in the simulations presented in the next section. These procedures misspecify the parametric process driving the data, because in practice there will be some ambiguity about parametric specification. In the present section we use misspecified processes to see whether our estimator's asymptotic efficiency gains hinge on nailing the parametric specification exactly; in the next section we use them to see whether any such gains have a reasonable chance of being realized in practice.

We impose misspecification of both the  $z_t$  process and the conditional variance process for  $e_t$ . For  $z_t$ , we use DGPs in which  $z_t \sim \text{ARMA}(1,1)$ , while  $z_t$  is wrongly modeled as an AR(4). We believe that this captures a common element of econometric practice, in which the investigator uses an unrestricted autoregression involving more parameters than would be required by Box-Jenkins techniques, choosing a lag length sufficiently long that the residual seems to be white

noise. Let  $e_t^\dagger$  denote the residual to this autoregression,

$$(3.3) \quad e_t^\dagger = z_t - E(z_t | z_{t-1}, z_{t-2}, z_{t-3}, z_{t-4}).$$

This residual is a distributed lag on  $e_t$  that in our processes is almost but not quite white noise.

For example, when  $z_t = .9z_{t-1} + e_t - .5e_{t-1}$  (one of our ARMA processes), the absolute value of all the autocorrelations of  $e_t^\dagger$  are below .03 and all past the fifth are less than .01.

For the conditional variance process, we continue to use a GARCH(1,1) as the DGP, while  $e_t^\dagger$ 's conditional moments are computed as described in the next section from an autoregressive forecast of  $|e_{t+j}^\dagger|$ . This technique, which is based on an alternative to GARCH models proposed by Schwert (1989), can be interpreted as trading parsimony for computational ease. With our DGPs it seems to fit the data sufficiently well that we find it plausible that a reasonable person would adopt the technique when faced with data such as ours. Consider, for example, this technique applied to  $|e_t|$  (rather than  $|e_t^\dagger|$ ), with GARCH parameters as in the table ( $\omega = .1$ ,  $\gamma_1 = .1$ ,  $\gamma_2 = .8$ ). Then  $Ee_t^2 e_{t+1}^2 = 1.33$ ,  $Ee_t^2 e_{t+2}^2 = 1.29$ ; the comparable values from the misspecified technique are 1.28 and 1.23.

We simulate with three DGPs, called DGPs A, B, and C. DGP A is one in which our estimator has very substantial asymptotic advantages relative to conventional GMM, even under misspecification. In DGP B, the advantages are modest, and in DGP C the advantages nonexistent for all practical purposes even in the absence of misspecification. We hope that these three stylized DGPs capture a salient feature from a wide range of possible datasets.

Table 3 lists the parameters and asymptotic variances of each of the DGPs. Line A of Table 3 presents asymptotic results for DGP A, in which  $z_t$ 's ARMA parameters are  $\phi = .9$ ,  $\zeta = .5$  and  $u_t$ 's moving average parameter  $\theta = -.95$ . We note first of all that inclusion of the moving average component in  $z_t$  raises considerably the relative efficiency of the optimal estimator, indicating that the figures in Table 1 by no means yield maximum figures. The "23.63" in column 1, line 1, is larger than any of the Table 1 or 2 figures for GMM1 with comparable GARCH parameters. More to the point, we see in the "Proposed Estimator" column that these

forms of misspecification little affect asymptotic efficiency, causing only a .4 percent increase in asymptotic variance. In the other DGPs, with parameters as indicated in the table, asymptotic efficiency is also little affected by our misspecification. Consistent with Table 1, DGPs B and C, the processes with less persistence in  $z_t$  and smaller moving average coefficients, yield smaller asymptotic efficiency gains for our estimator.

#### 4. SIMULATION EVIDENCE

In this section, we present some simulation evidence on the behavior of our estimator. Our intention is not to provide an exhaustive characterization of finite sample behavior, but to get a feel for whether the estimator can work well in samples of size typically available, and in the presence of the minor forms of misspecification described in the previous section. We present results for the processes in Table 3, for sample sizes of  $T=250, 500, 1000$  and  $10,000$ . The last sample size is one not often seen in practice. We include it not only because it is relevant for some data sets, particularly those with asset pricing data, but to gauge how large a sample size is required for the asymptotic approximation to be tight. To conserve space details of data generation and mechanics of estimation are relegated to the Additional Appendix.

##### 4.1 Data Generating Process and Estimators

We use the MA(1) model and estimation techniques underlying the results presented in Table 3:

$$(4.1a) \quad y_t = \beta_0 + z_t \beta_1 + u_t \equiv X_t' \beta + u_t, \quad X_t \equiv (1, z_t)',$$

$$(4.1b) \quad u_t = c_2 e_{t+2} + c_1 e_{t+1} = e_{t+2} - \theta e_{t+1},$$

$$(4.1c) \quad z_t = \phi z_{t-1} + e_t - \zeta e_{t-1},$$

$$(4.1d) \quad e_t \sim N(0, \sigma_t^2), \quad \sigma_t^2 = \omega + \gamma_1 e_{t-1}^2 + \gamma_2 \sigma_{t-1}^2, \quad \gamma \equiv \gamma_1 + \gamma_2.$$

Table 3 has the parameter values, apart from  $\beta_0$  and  $\beta_1$ , which were set to zero for simplicity. Consistent with Table 3, we constructed estimates of  $S$  and  $\Psi$  assuming (incorrectly) that  $z_t \sim$

AR(4) and that conditional variances of the residual to the AR(4), call it  $e_t^\dagger$ , depend only on the autoregressive forecasts of the absolute value of this residual. We write

$$(4.2a) \quad z_t = \phi_0 + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \phi_3 z_{t-3} + \phi_4 z_{t-4} + e_t^\dagger, \quad e_t^\dagger \equiv z_t\text{-projection}(z_t | 1, z_{t-1}, z_{t-2}, z_{t-3}, z_{t-4});$$

$$(4.2b) \quad |e_t^\dagger| = \alpha_0 + \alpha_1 |e_{t-1}^\dagger| + \alpha_2 |e_{t-2}^\dagger| + \alpha_3 |e_{t-3}^\dagger| + \alpha_4 |e_{t-4}^\dagger| + v_t$$

$$v_t \equiv |e_t^\dagger| - \text{projection}(|e_t^\dagger| | 1, |e_{t-1}^\dagger|, |e_{t-2}^\dagger|, |e_{t-3}^\dagger|, |e_{t-4}^\dagger|);$$

$$(4.2c) \quad b = (\beta_0, \beta_1, \sigma_u^2, \sigma_{u,1}; \phi_0, \phi_1, \phi_2, \phi_3, \phi_4, \sigma_{e^\dagger}^2; \alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \sigma_v^2; c_2, c_1);$$

$$(4.2d) \quad b^\dagger \approx (0, 0, 1.9025, .95; 0, .40, .21, .11, .07, 1.003; 0.55, 0.09, 0.08, 0.07, 0.06, 3.37; 1, -0.95)', \quad m = 18.$$

In (4.2a,b), “projection” means linear least squares forecast; in (4.2d),  $m$  is the dimension of  $b$  and of  $b^\dagger$ , while  $b^\dagger$  is the numerical value of  $b$ , which in turn is the probability limit of  $\hat{b}$ . In  $b^\dagger$ , the figures for  $\{\phi_i\}$  and  $\sigma_{e^\dagger}^2$  were computed analytically, those for  $\{\alpha_i\}$  and  $\sigma_v^2 \equiv E v_t^2$  from a simulation.

As in previous sections, we focus on estimation of  $\beta_1$ . We simulated the behavior of four estimators. The first was a feasible version of our proposed estimator,  $\hat{\beta}^* \equiv (\sum_{t=1}^T \hat{Z}_t^* X_t')^{-1} (\sum_{t=1}^T \hat{Z}_t^* y_t)$ . An overview of our implementation (see the Additional Appendix for details): We began with least squares estimation of  $\beta$  in (4.1a) and  $\phi_0, \dots, \phi_4$  in (4.2a) to obtain residuals  $\hat{u}_t$  and  $\hat{e}_t^\dagger$ . These residuals were then used in least squares estimation of  $c_2$  and  $c_1$  in (4.1b) and  $\alpha_0, \dots, \alpha_4$  in (4.2b).  $\hat{\Psi}$  was constructed from the first 100 moving average weights implied by the estimates of (4.2a).  $\hat{S}$  was constructed in accordance with part 3 of the Additional Appendix, again relying on only the first 100 rather than all  $T-1$  lags of  $\hat{e}_t^\dagger$ .  $\hat{Z}_t^*$  was then computed according to equation (2.4), with the upper bound on the summation set to the smaller of  $\{t-1, 100\}$ . Finally, for inference, the estimate of the asymptotic variance-covariance matrix of  $\hat{\beta}^*$  was computed as  $(\hat{\Psi}' \hat{S}^{-1} \hat{\Psi})^{-1}$ .

The second through fourth estimators were conventional GMM with an instrument vector  $W_t$  that includes a constant and lags 0 through  $n-1$  of  $z_t$ ,  $W_t = (1, z_t, z_{t-1}, \dots, z_{t-n+1})'$  ( $n=1 \Rightarrow$  OLS). These estimators proceed in a familiar fashion:

$$(4.3) \quad \hat{\beta} = (\Sigma_{t=1}^T \hat{Z}_t X_t')^{-1} (\Sigma_{t=1}^T \hat{Z}_t y_t), \quad \sqrt{T}(\hat{\beta} - \beta) \sim_A N(0, V),$$

$$\hat{Z}_t = (T^{-1} \Sigma_{t=n}^T X_t W_t') \hat{\Omega}^{-1} W_t, \quad \hat{\Omega} \rightarrow_p \Omega \equiv \Sigma_{i=-1}^1 E(W_{t,i} u_{t-i} u_t W_t'),$$

$$V = [(EX_t W_t') \Omega^{-1} E(W_t X_t')]^{-1} = \text{plim } \hat{V} \equiv \text{plim} [(T^{-1} \Sigma_{t=n}^T X_t W_t') \hat{\Omega}^{-1} (T^{-1} \Sigma_{t=n}^T W_t X_t')]^{-1}.$$

In (4.3), the  $(n+1) \times (n+1)$  matrix  $\hat{\Omega}$  was computed with a Bartlett kernel with VAR(1) prewhitening and a bandwidth set to the integer part of  $[4(T/100)^{1/3}]$  (see Newey and West (1994)).

## 4.2 Results

Information on the distribution of the estimates of  $\beta_1$  is presented in Table 4 (DGP A) and Table 5 (DGPs B and C). Let us begin with Table 4. Panel A of Table 4 normalizes all estimates by dividing by the asymptotic standard error (square root of asymptotic variance) of the optimal estimator. The resulting quantities will asymptotically be distributed as  $N(0, v/v^*)$ , where  $v^* = 1.004$  is the proposed estimator's entry in line A of Table 3 and  $v$  is the corresponding entry for GMM $n$  (e.g., 23.63 for GMM1). We can see from the “variance” and “RMSE” (root mean squared error) columns that even with misspecification, the estimator we propose is, indeed, distinctly more concentrated around  $\beta_1 (=0)$  than are the other estimators. Its variance of 2.05, for example, is less than a tenth of that of the OLS variance of 24.93, and less than a third of that of GMM4 and GMM12. The interquartile range indicates the same: the 1.83 value for our estimator ( $1.83 = 0.61 - (-1.22)$ ) is about a third of that of GMM1, a little over half that of GMM4 and GMM12.

Consistent with the asymptotic theory, then, the optimal estimator is more concentrated around the true value than are the conventional GMM estimators. Evidently, however, the asymptotic approximation is not a particularly accurate one. The estimators are mean- and median-biased downwards (this is the bias familiar from the unit root literature), and the variances depart significantly from the asymptotic (theoretical ratio of  $v/v^*$  for GMM1 = 23.63, actual =  $24.93/2.05 \approx 12$ ).

To better characterize the quality of the asymptotic approximation for the conventional GMM estimators, panel B divides each set of estimates by own asymptotic variance, repeating in its first line the first line of panel A. If the asymptotic approximation worked perfectly, the entries would be those of a  $N(0,1)$  random variable (values for which are presented in the “reference” line below panel C). One can see that in our sample size of 1000 the least squares estimator (GMM1) is well-characterized by the asymptotic approximation; for example, the .05 and .95 quantiles are -1.58 and 1.77, matching tolerably well the asymptotic values of -1.65 and 1.65 given in the “reference” line. The quality of the approximation is not as good for the other estimators, including the one that we propose. All are too variable, and downward biased. The approximation is especially poor for GMM12, a result consistent with Tauchen (1986), who also found especially poor performance of conventional GMM estimators when the instrument vector contained many lags.

The quality of the asymptotic approximation is, of course, even worse for smaller  $T$ , but better for larger  $T$ . This is illustrated for our estimator in panel C of Table 4. Upon scanning down the rows of the table, one sees that the figures move closer and closer to those of a  $N(0,1)$  variable. With  $T=10,000$ , the asymptotic approximation is, perhaps, roughly accurate, though our estimator still is downward biased and a bit too variable.

The results in Tables 4A and 4C are presented graphically in Figures 1 and 2. These figures plot smoothed density estimates, computed using a Gaussian kernel. Figure 1 depicts information summarized in Table 4A. Each of the three panels includes a  $N(0,1)$  density (dotted line), the density of our estimator (solid line, identical in all three panels), and the density of one of the conventional estimators (dashed line). That our estimator is far more concentrated around 0 is evident from the pictures, as is the fact that our estimator is not particularly well characterized by the asymptotic approximation. Figure 2 depicts information summarized in Table 4C. The Figure makes clear that the asymptotic approximation improves with  $T$ , and that there are notable departures from normality for all four values of  $T$ .

Table 5 presents information analogous to that in the top and bottom panels of Table 4, but for DGPs B and C.<sup>4</sup> Table 5A indicates that for DGP B, even though GMM12 and our estimator are roughly equally efficient asymptotically, our estimator is notably less variable for  $T=1000$ : the ratio of their variances is about 1.5 ( $\approx 1.84/1.21$ ). Similar behavior occurs for  $T=250$  and  $T=500$ , though by  $T=10,000$  the relative variability is reasonably close to the asymptotic value of 1.06 given in line B of Table 3 (not shown in the Table). Table 5B indicates that for DGP C, the asymptotic approximation works reasonably well for all four estimators for  $T=1000$ : mean and median bias is small, and variances are within a few percent of their asymptotic values. Indeed, the quality of the asymptotic approximation for all values of  $T$ , even for  $T=250$ , is comparable to that displayed for  $T=1000$  in Table 5B. The bottom panel in Table 5 presents information on the behavior of our estimator for various sample sizes. For DGP B, the asymptotic approximation looks good for  $T=10,000$  and perhaps for  $T=1000$  as well, while for DGP C we find little to complain about even for  $T=250$ .

We turn now from parameter estimation to hypothesis tests. Table 6A presents the actual size of two sided t-tests of  $H_0: \beta_1=0$  for nominal sizes of .10 and .05, for a sample size  $T=1000$ , for all three DGPs. Figure 3 does the same for DGP A for nominal sizes running from 0 to .25. The solid line in each box in Figure 3 is a 45 degree line. The dashed line maps nominal into actual size. In all four boxes, the proposed estimator's dashed line is above the solid line. This means that they tend to reject too much, a result consistent with the Table 4 information that the estimators are too spread out. For example, we see in Table 6A that for our proposed estimator, the absolute value of the t-statistic was above 1.96 in 98 data sets (the ideal is 50). We also see in Table 6A that for our estimator, tests are worst sized in DGP A, best sized in DGP C; for the conventional estimators, the performance is roughly comparable across DGPs. Table 6B indicates that for our estimator the asymptotic approximation is better with larger sample sizes, with size distortions quite moderate when  $T=10,000$ , results that also apply to the other estimators (not presented in the table).

We see from Tables 4-6 that for all estimators, for a given sample size  $T$ , the asymptotic approximation works worst for DGP A, which is the process with the strongest serial correlation, best for DGP C, which is the process with the least serial correlation.

We also conclude from these tables that our estimator's parameter estimates are preferable to those of any of the conventional estimators if the measure of performance is variability as measured by interquartile range or variance. If the measure of performance is mean or median bias, ordinary least squares (GMM1) is preferable to our estimator. If one combines mean bias and variance into RMSE, our estimator is preferable. If the measure of performance is accuracy of size of hypothesis tests of the usual nominal size (equivalently, accuracy of confidence interval coverage), least squares (GMM1) and GMM4 probably perform best, GMM12 the worst, with our estimator falling in the middle.

## 5. CONCLUSIONS

We have proposed and evaluated an instrumental variables estimator for linear models with conditionally heteroskedastic disturbances. The estimator is efficient in a class of estimators that are linear in a possibly infinite set of lags of a finite number of basic instruments. Implementation of the estimator requires specification of a parametric model. Simulations indicate that the estimator often works well relative to a conventional estimator (Hansen (1982)) in common use, even when the parametric model is misspecified. Priorities for future research include development and evaluation of efficient estimators that are nonlinear in lags of basic instruments, and alternative asymptotic approximations to better characterize the small sample distortions evident in many of the simulations.

## FOOTNOTES

1. Note that use of a vector AR or other model does not require knowledge of the entire set of structural equations relating the variables. This model is merely a device for computing  $E(X_t | \text{current and lagged } z_t \text{'s})$ . See West and Wilcox (1996) for an illustration of this point in a conditionally homoskedastic environment.

2. One may of course omit the factor of  $\hat{\sigma}_e^2$  without changing the estimate of  $\beta$ . Note, however, one cannot then use  $(\Psi' S^{-1} \Psi)^{-1}$  to estimate the asymptotic variance-covariance matrix.

3. To prevent confusion, we note that while the regression disturbance in panel B may be split into a component that is correlated with future instruments ( $e_{t+2} - \theta e_{t+1}$ ) and one that is not ( $v_{t+2} - d v_{t+1}$ ), such a split need not always obtain. After projecting  $u_t$  onto appropriately dated  $e_t$ 's, the residual that remains might be correlated with future  $e_t$ 's, and might be conditionally heteroskedastic.

4. We omit information such that in Table 4B because behavior when normalized by own asymptotic standard error is usually indistinguishable from behavior when normalized by the asymptotic standard error of the proposed estimator.

## APPENDIX

Estimation of multiple equation systems proceeds as follows. Consider an  $\ell$  equation system,  $y_t = X_t' \beta + u_t$ , where  $y_t$  and  $u_t$  are  $(\ell \times 1)$ ,  $X_t$  is  $(k \times \ell)$  and  $\beta$  is  $(k \times 1)$ . The  $(r \times 1)$  vector of basic instruments  $z_t$  has an  $(r \times 1)$  vector of innovations  $e_t$  that satisfies  $E(u_t \otimes e_{t-j}) = 0$  for all  $j \geq 0$ .

Define

$$\begin{aligned} e(t) &= (1, e_t', \dots, e_{t-T+1}')', & \tilde{e}(t) &= I_\ell \otimes e(t), & S &= \sum_{i=-q}^q E[u_{t-i} \otimes e(t-i)][u_t' \otimes e(t)'], \\ (1+Tr) \times 1 & & (1+Tr)\ell \times \ell & & (1+Tr)\ell \times (1+Tr)\ell \end{aligned}$$

$$\begin{aligned} \Psi &= EX_t' \otimes e(t), & G &= S^{-1} \Psi. \\ (1+Tr)\ell \times k & & (1+Tr)\ell \times k \end{aligned}$$

Define  $\hat{e}_{t-j} = 0$  for  $t-j < 0$ , and otherwise use a “ $\hat{\phantom{x}}$ ” to denote a sample counterpart constructed by evaluating the indicated random variable or matrix of parameters at  $\hat{b}$ . Then the optimal  $(k \times \ell)$  instrument is  $\hat{Z}_t = \hat{G}' \hat{\tilde{e}}(t)$ ,  $\hat{G} = \hat{S}^{-1} \hat{\Psi}$ , with corresponding estimate  $\hat{\beta} = (\sum_{t=1}^T \hat{Z}_t X_t')^{-1} (\sum_{t=1}^T \hat{Z}_t y_t)$ .

## References

- Anatolyev, Stanislav, 1998, "Optimal Instrumental Variables Estimation in Time Series Models with Serial Dependence," manuscript, University of Wisconsin.
- Andrews, Donald W. K., 1991, "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," Econometrica 59, 817-858.
- Ball, Laurence and Dean Croushore, 1995, "Expectations and the Effects of Monetary Policy," National Bureau of Economic Research Working Paper 5344.
- Bates, Charles E. and Halbert White, 1993, "Determination of Estimators with Minimum Asymptotic Covariance Matrices," Econometric Theory 9, 633-48.
- Bollerslev, Tim, 1986, "Generalized Autoregressive Conditional Heteroskedasticity," Journal of Econometrics 31 302-327.
- Bollerslev, Tim, 1987, "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return," Review of Economics and Statistics 69, 542-47.
- Bollerslev, Tim, and Jeffrey M. Wooldridge, 1992, "Quasi Maximum Likelihood Estimation and Inference in Dynamic Models With Time Varying Covariances," Econometric Reviews 11, 143-172.
- Broze, Laurence, Francq, Christian and Jean-Michel Zakoïan, 2001, "Non-redundancy of High Order Moment Conditions for Efficient GMM Estimation of Weak AR Processes," Economics Letters 71, 317-322.
- Breusch, Trevor, Qian, Hialong, Schmidt, Peter, and Donald Wyhowski, 1999, "Redundancy of Moment Conditions," Journal of Econometrics 91, 89-111.
- Campbell, John Y. and N. Gregory Mankiw, 1990, "Permanent Income, Current Income, and Consumption," Journal of Business and Economic Statistics 8, 265-79.
- den Haan, Wouter and Andrew Levin, 1996, "Inferences from Parametric and Non-Parametric Covariance Matrix Estimators," National Bureau of Economic Research Technical Working Paper No. 195.
- Engle, Robert F., 1982, "Autoregressive Conditional Heteroskedasticity, With Estimates of the Variance of United Kingdom Inflation," Econometrica 50, 987-1007.
- Fama, Eugene F. and Kenneth R. French, 1988, "Permanent and Temporary Components of Stock Prices," Journal of Political Economy 96, 246-73.
- Hall, Robert E., 1988, "Intertemporal Substitution in Consumption," Journal of Political Economy 96, 339-357.
- Hansen, Lars Peter, 1982, "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica 50, 1029-54.
- Hansen, Lars Peter, 1985, "A Method for Calculating Bounds on the Asymptotic Variance-Covariance Matrices of Generalized Method of Moments Estimators," Journal of Econometrics 30, 203-228.

- Hansen, Lars Peter, 1986, "Asymptotic Covariance Matrix Bounds for Instrumental Variables Estimators of Linear Time Series Models," manuscript, University of Chicago.
- Hansen, Lars Peter, Heaton, John C. and Masao Ogaki, 1988, "Efficiency Bounds Implied by Multiperiod Conditional Moment Restrictions," Journal of the American Statistical Association **83**, 863-871.
- Hansen, Lars Peter, and Thomas J. Sargent, 1980, "Formulating and Estimating Dynamic Linear Rational Expectations Models," Journal of Economic Dynamics and Control **2**, 7-46.
- Hansen, Lars Peter and Kenneth J. Singleton, 1991, "Computing Semi-Parametric Efficiency Bounds for Linear Time Series Models," 387-412 in Barnett, W., Powell, J. and G. Tauchen (eds), Nonparametric and Semiparametric Methods in Econometrics and Statistics, Cambridge: Cambridge University Press.
- Hansen, Lars Peter and Kenneth J. Singleton, 1996, "Efficient Estimation of Linear Asset Pricing Models with Moving-Average Errors," Journal of Business and Economic Statistics **14**, 53-68.
- Hayashi, Fumio and Christopher A. Sims, 1983, "Nearly Efficient Estimation of Time Series Models with Predetermined, but not Exogenous, Instruments", Econometrica **51**, 783-798,
- Heaton, John C. and Masao Ogaki, 1991, "Efficiency Bounds for a Time Series Model With Conditional Heteroskedasticity," Economics Letters **35**, 167-171.
- Hodrick, Robert J., 1987, The Empirical Evidence on the Efficiency of Forward and Futures Foreign Exchange Markets, Harwood Academic Publishers: New York.
- Kaminsky, Graciela and Rodrigo Peruga, 1990, "Can a Time-Varying Risk Premium Explain Excess Returns in the Forward Market for Foreign Exchange?" Journal of International Economics, **28**, February, 47-70.
- Kennan, John, 1979, "The Estimation of Partial Adjustment Models with Rational Expectations," Econometrica **47**, 1441-1456.
- Koenker, Roger and José A. F. Machado, 1997, "GMM Inference When the Number of Moment Conditions is Large," manuscript, University of Illinois.
- Kollintzas, Tryphon, 1993, "A Generalized Variance Bounds Test, With an Application to the Holt et al. Inventory Model," manuscript, Athens University of Economics and Business.
- Kuersteiner, Guido M., 1996, "Efficient IV Estimation for Autoregressive Models with Conditional Heterogeneity," manuscript, Yale University.
- Mishkin, Frederic S., 1992, "Is the Fisher Effect for Real? A Reexamination of the Relationship between Inflation and Interest Rates," Journal of Monetary Economics, **30**(2), 195-215.
- Newey, Whitney K., 1988, "Adaptive Estimation of Regression Models via Moment Restrictions," Journal of Econometrics **38**, 301-339.
- Newey, Whitney K. and Kenneth D. West, 1994, "Automatic Lag Selection in Covariance Matrix Estimation," Review of Economic Studies **61** (1994), 631-654.
- Oliner, Stephen D., Rudebusch, Glenn D. and Daniel Sichel, 1996, "The Lucas Critique Revisited: Assessing the Stability of Empirical Euler Equations for Investment," Journal of

Econometrics 70, 291-316.

Schwert, G. William, 1989, "Why Does Stock Market Volatility Change over Time?," Journal of Finance 44, 1115-53.

Stambaugh, Robert F., 1994, "Estimating Conditional Expectations When Volatility Fluctuates," NBER Technical Working Paper No. 140.

Tauchen, George, 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.

West, Kenneth D., 2000, "On Optimal Instrumental Variables Estimation of Stationary Time Series Models," National Bureau of Economic Research Technical Working Paper No. 249; forthcoming, International Economic Review.

West, Kenneth D. and David W. Wilcox, 1996, "A Comparison of Alternative Instrumental Variables Estimators of a Dynamic Linear Model," Journal of Business and Economic Statistics 14, 281-293.

Table 1

Asymptotic Variances Relative to Optimal GMM,  $u_t = e_{t+2} - \theta e_{t+1}$ 

	$\phi$	$\theta$	$\gamma$	$\gamma_1$	GMM1	GMM4	GMM12
1.	.9	0.	.9	.1	1.00	1.00	1.00
2.	.9	0.	.9	.3	1.23	1.14	1.04
3.	.5	-.5	.9	.1	1.11	1.00	1.00
4.	.5	.5	.9	.1	1.36	1.00	1.00
5.	.5	.9	.9	.1	3.13	1.38	1.04
6.	.5	.95	.9	.1	3.57	1.54	1.11
7.	.9	.9	.9	.1	6.13	1.92	1.11
8.	.9	.95	.0	.0	9.16	2.73	1.36
9.	.9	.95	.5	.1	10.45	2.85	1.37
10.	.9	.95	.9	.1	10.65	3.02	1.41
11.	.9	.95	.9	.3	36.36	8.57	2.62
12.	.9	$\frac{1}{.95}$	.9	.1	9.92	2.88	1.38

Notes:

1. The model is  $y_t = \beta_0 + z_t \beta_1 + u_t$ ,  $u_t = e_{t+2} - \theta e_{t+1}$ ,  $z_t = \phi z_{t-1} + e_t$ ,  $e_t \sim \text{GARCH}(1,1)$ ,  $e_t = \sigma_t \eta_t$ ,  $\eta_t \sim \text{iid}(0,1)$ ,  $E\eta_t^4 = 3$ ,  $\sigma_t^2 = \omega + \gamma_1 e_{t-1}^2 + \gamma_2 \sigma_{t-1}^2$ ,  $\gamma = \gamma_1 + \gamma_2$ . The figures are invariant to choice of  $\omega$  (set to 0.1) and  $\beta_0$  and  $\beta_1$  (both set to zero).

2. GMM $n$  is the conventional GMM estimator (Hansen (1982)) with a constant and lags 0 through  $n-1$  of  $z_t$  used as instruments (GMM1 = ordinary least squares). The optimal GMM estimator asymptotically uses all lags of  $z_t$  as instruments. The table presents the ratio of asymptotic variances of estimators of  $\beta_1$  to that of the optimal estimator.

Table 2

## Asymptotic Variances Relative to Optimal GMM, Alternative Specifications

$$A. u_t = e_{t+2} - \theta e_{t+1}, \theta = .95, e_t = \sigma_t \eta_t, \eta_t \sim \text{iid}(0,1), E\eta_t^4 = 3 + \kappa_\eta$$

	$\kappa_\eta$	GMM1	GMM4	GMM12
1.	0.0	10.65	3.02	1.41
2.	1.0	11.43	3.18	1.43
3.	3.9	13.90	3.67	1.52

$$B. u_t = e_{t+2} - .95e_{t+1} + v_{t+2} - dv_{t+1}, e_t = \sigma_t \eta_t, \eta_t \sim \text{iid}(0,1), E\eta_t^4 = 3, v_t \sim \text{iid } N(0, \sigma_v^2)$$

	$d$	implied MA parameter for $u_t$	GMM1	GMM4	GMM12
1.	0.95	0.95	9.93	2.88	1.39
2.	2.0	0.63	1.47	1.02	1.00
3.	-0.9	0.001	1.00	1.00	1.00

Notes:

1. See notes to Table 1 for the model and definition of “GMM1”, “GMM4” and “GMM12”. In all specifications,  $\phi = .9$ ,  $\omega = .1$ ,  $\gamma_1 = .1$ ,  $\gamma = .9$ :  $z_t = .9z_{t-1} + e_t$ ,  $\sigma_t^2 = 0.1 + 0.1e_{t-1}^2 + 0.8\sigma_{t-1}^2$ . The results in line 1 of panel A repeat those in line 10 of Table 1.

2. In panel B,  $\sigma_v^2$  was chosen so that half the variance of  $u_t$  was due to  $v_t$ , i.e.,  $(1 + .95^2)\sigma_v^2 = (1 + d^2)\sigma_\varepsilon^2$ . The “implied MA parameter” is the value  $\xi$  such that  $|\xi| < 1$  and  $\sigma_\varepsilon^2 > 0$  satisfy  $(1 + \xi^2)\sigma_\varepsilon^2 = Eu_t^2$ ,  $-\xi\sigma_\varepsilon^2 = Eu_t u_{t-1}$ .

Table 3

## Asymptotic Variances Relative to Optimal GMM, Processes Used In Simulations

	GMM1	GMM4	GMM12	Proposed estimator
A. $z_t = .9z_{t-1} + e_t - .5e_{t-1}$ , $\theta = .95$ 23.63	4.28	1.52		1.004
B. $z_t = .7z_{t-1} + e_t - .5e_{t-1}$ , $\theta = .9$ 3.73	1.56	1.06		1.002
C. $z_t = .5z_{t-1} + e_t + .5e_{t-1}$ , $\theta = .5$	1.20	1.00	1.00	1.00

## Notes:

1. See notes to Table 1 for the model and definition of “GMM1”, “GMM4” and “GMM12”. In all three DGPs,  $e_t \sim \text{GARCH}(1,1)$ ,  $e_t = \sigma_t \eta_t$ ,  $\eta_t \sim \text{iid}(0,1)$ ,  $E\eta_t^4 = 3$ ,  $\sigma_t^2 = 0.1 + 0.1e_{t-1}^2 + 0.8\sigma_{t-1}^2$ .

2. The column labeled “proposed estimator” presents the ratio of the asymptotic variance of the estimator we propose to that of the optimal estimator. In contrast to Tables 1 and 2, we now assume that the proposed estimator uses a misspecified parametric model and thus is not optimal asymptotically. It is misspecified in two ways. First, the investigator wrongly models  $z_t$  as an AR(4) when in fact  $z_t$  follows the indicated ARMA(1,1) processes. Second, the investigator computes  $Ee_t^2 e_{t+j}^2$  from an AR(4) in  $|e_t|$  when in fact  $e_t$  follows the GARCH process given in note 1. See text and notes to Table 1 for additional details.

Table 4

## Distributions of Parameter Estimates, From Simulations, DGP A

## A. Standardized by Asymptotic Standard Error of Proposed Estimator, T=1000

	Quantiles					Moments		
	.05	.25	.50	.75	.95	mean	variance	RMSE
Proposed Estimator	-2.78	-1.22	-0.23	0.61	1.88	-0.33	2.05	1.47
GMM1	-7.66	-3.37	-0.22	3.01	8.61	-0.00	24.93	4.99
GMM4	-4.32	-1.78	-0.16	1.29	3.77	-0.19	6.16	2.49
GMM12	-4.69	-1.83	-0.29	1.31	4.36	-0.19	7.65	2.77

## B. Standardized by Own Asymptotic Standard Error, T=1000

	Quantiles					Moments		
	.05	.25	.50	.75	.95	mean	variance	RMSE
Proposed Estimator	-2.78	-1.22	-0.23	0.61	1.88	-0.33	2.05	1.47
GMM1	-1.58	-0.69	-0.05	0.62	1.77	-0.00	1.06	1.03
GMM4	-2.09	-0.86	-0.08	0.62	1.83	-0.09	1.44	1.20
GMM12	-3.82	-1.49	-0.24	1.06	3.55	-0.15	5.07	2.26

## C. Proposed Estimator, Standardized by Own Asymptotic Standard Error, Various T

	Quantiles					Moments		
	.05	.25	.50	.75	.95	mean	variance	RMSE
T=250	-4.69	-1.94	-0.59	0.53	2.71	-0.72	5.86	2.52
T=500	-3.25	-1.50	-0.36	0.57	1.99	-0.48	2.83	1.75
T=1000	-2.78	-1.22	-0.23	0.61	1.88	-0.33	2.05	1.47
T=10,000	-2.10	-0.90	-0.10	0.71	1.78	-0.12	1.36	1.17
Reference: asymptotic	-1.65	-0.68	0.	0.68	1.65	0	1.00	1.00

Notes:

1. The data generating process is in line 1 of Table 3:  $y_t = \beta_0 + z_t \beta_1 + u_t$ ,  $u_t = c_2 e_{t+2} + c_1 e_{t+1}$ ,  $z_t = \phi z_{t-1} + e_t - \zeta e_{t-1}$ ,  $e_t \sim \text{GARCH}(1,1)$ ,  $e_t / \sigma_t \sim N(0,1)$ ,  $\sigma_t^2 = \omega + \gamma_1 e_{t-1}^2 + \gamma_2 \sigma_{t-1}^2$ ,  $\beta_0 = \beta_1 = 0$ ,  $c_2 = 1$ ,  $c_1 = -0.95$ ,  $\phi = 0.9$ ,  $\zeta = .5$ ,  $\omega = 0.1$ ,  $\gamma_1 = 0.1$ ,  $\gamma_2 = 0.8$ . All variables are scalars. The number of repetitions was 1000. Additional details on the data generation may be found in the text.

2. The proposed estimator is misspecified as described in the text and Table 3 and so is not optimal. GMM $n$  is conventional GMM (Hansen (1982)) with a constant lags 0 through  $n-1$  of  $z_t$  as instruments. Implementation of the estimators is described in the text.

3. Each panel presents statistics that are asymptotically normal (standard normal in panels B and C). In the four lines of panel A, the variance of the asymptotic distribution may be computed from line A of Table 3 as: Proposed: 1; GMM1:  $23.63/1.004 \approx 23.63$ ; GMM4:  $4.38/1.004 \approx 4.28$ ; GMM12:  $1.52/1.004 \approx 1.52$ .

4. To compute the quantiles, the 1000 sets of estimates were sorted, multiplied by  $\sqrt{T}$  and divided by the indicated standard error. The .05 entry presents the 50<sup>th</sup> smallest entry, the .50 entry the median, etc..

Table 5

Distributions of Parameter Estimates, From Simulations, DGPs B and C

A. Standardized by Asymptotic Standard Error of Proposed Estimator, T=1000, DGP B

	Quantiles					Moments		
	.05	.25	.50	.75	.95	mean	variance	RMSE
Proposed Estimator	-2.07	-0.96	-0.18	0.55	1.51	-0.22	1.21	1.12
GMM1	-2.98	-1.37	-0.06	1.18	3.47	-0.02	3.74	1.93
GMM4	-2.27	-0.95	-0.11	0.78	2.11	-0.07	1.78	1.34
GMM12	-2.34	-0.99	-0.11	0.72	2.24	-0.09	1.84	1.36

B. Standardized by Asymptotic Standard Error of Proposed Estimator, T=1000, DGP C

	Quantiles					Moments		
	.05	.25	.50	.75	.95	mean	variance	RMSE
Proposed Estimator	-1.70	-0.70	-0.01	0.58	1.42	-0.07	0.93	0.97
GMM1	-1.78	-0.75	0.01	0.71	1.76	-0.01	1.16	1.08
GMM4	-1.72	-0.72	0.03	0.65	1.55	-0.03	1.00	1.00
GMM12	-1.69	-0.67	0.03	0.70	1.62	-0.00	1.03	1.01

C. Proposed Estimator, Standardized by Own Asymptotic Standard Error, Various T

	DGP B				DGP C			
	Quantiles			variance	Quantiles			variance
	.25	.50	.75		.25	.50	.75	
T=250	-1.23	-0.45	0.38	1.59	-0.80	-0.17	0.49	0.90
T=500	-1.06	-0.25	0.46	1.38	-0.77	-0.07	0.60	0.95
T=1000	-0.96	-0.18	0.55	1.21	-0.70	-0.01	0.58	0.93
T=10,000	-0.78	-0.06	0.59	1.02	-0.75	-0.05	0.62	0.99
Reference: asymptotic	-0.68	0.	0.68	1.00	-0.68	0.	0.68	1.00

Notes:

1. The data generating process is as given in lines 2 and 3 of Table 3. The full set of parameters are given in note 1 to Table 4, with the following changes: DGP B:  $c_1 = -0.90$ ,  $\phi = 0.7$ ,  $\zeta = .5$ ; DGP C:  $c_1 = -0.50$ ,  $\phi = 0.5$ ,  $\zeta = -.5$ .

2. Each panel presents statistics that are asymptotically normal (standard normal in panel C). In the four lines of panel A, the variance of the asymptotic distribution may be computed from line B of Table 3 as: Proposed: 1; GMM1:  $3.73/1.002 \approx 3.73$ ; GMM4:  $1.56/1.002 \approx 1.56$ ; GMM12:  $1.06/1.002 \approx 1.06$ . In the four lines of panel B, the variance of the asymptotic distribution may be computed from line C of Table 3 as: Proposed: 1; GMM1: 1.20; GMM4: 1.00; GMM12: 1.00.

3. See notes to Tables 3 and 4 for additional information.

Table 6

## Size of Nominal .10 and .05 Tests, From Simulations

## A. Various Estimators, T=1000

	DGP A		DGP B		DGP C	
	.10	.05	.10	.05	.10	.05
Nominal Size						
Actual Size						
Proposed Estimator	.150	.098	.135	.077	.093	.054
GMM1	.128	.075	.138	.079	.116	.066
GMM4	.110	.058	.145	.093	.139	.080
GMM12	.233	.164	.245	.182	.221	.144

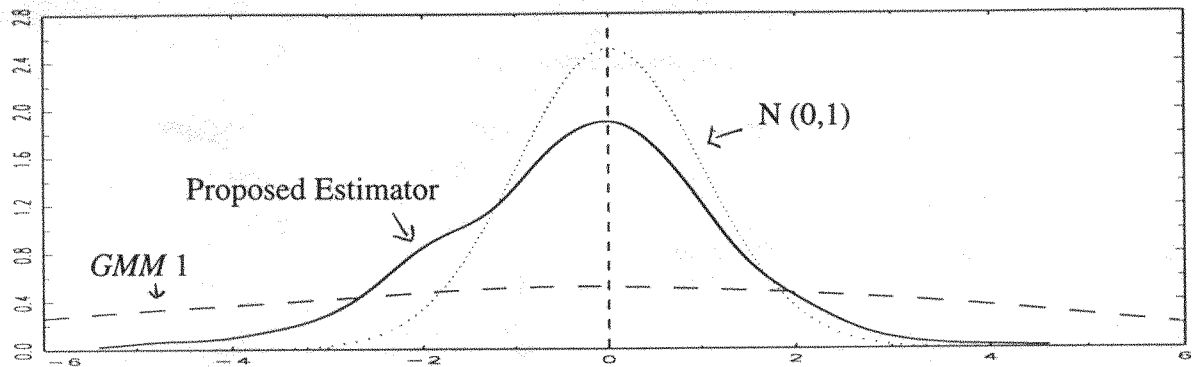
## B. Proposed Estimator, Various T

	DGP A		DGP B		DGP C	
	.10	.05	.10	.05	.10	.05
Nominal Size						
Actual Size						
T=250	.230	.191	.197	.136	.125	.080
T=500	.184	.144	.156	.103	.103	.062
T=1000	.150	.098	.135	.077	.093	.054
T=10,000	.116	.068	.108	.047	.101	.051

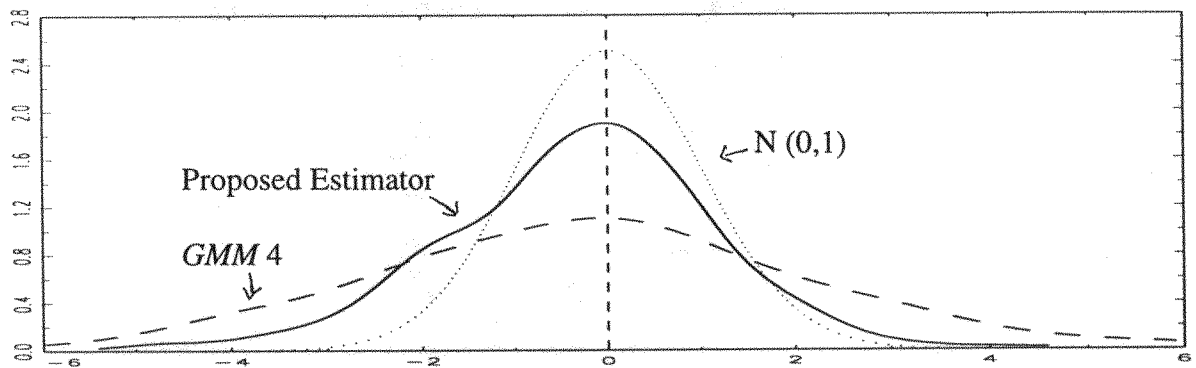
## Notes:

1. See notes 1 and 2 to Tables 3 and 4 for description of data and definition of estimators.
2. The entries are based on 1000 sets of two-sided t-tests for  $H_0: \beta_1=0$ . The “.150” in the first column in the “Proposed Estimator” row in panel A, for example, indicates that in 150 of the 1000 simulated data sets, this t-statistic was greater than 1.65 in absolute value, the “0.098” that in 98 the t-statistic was greater than 1.96 in absolute value.

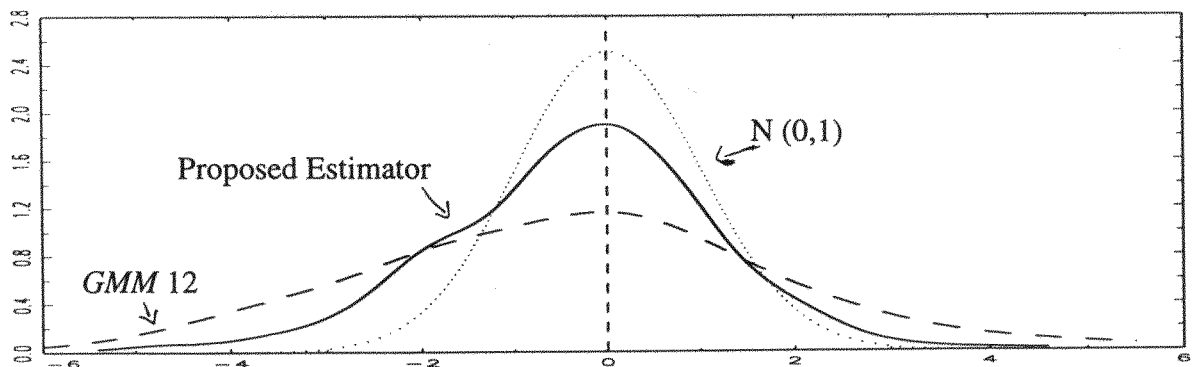
Figure 1: Density of Parameter Estimates,  $GMM\ n$  vs. Proposed Estimator,  $T = 1000$ , DGP A



(a)  $GMM\ 1$



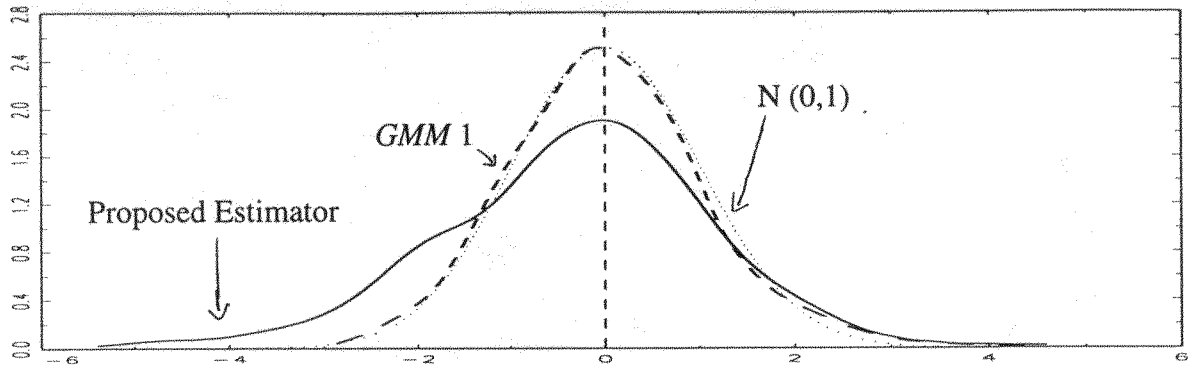
(b)  $GMM\ 4$



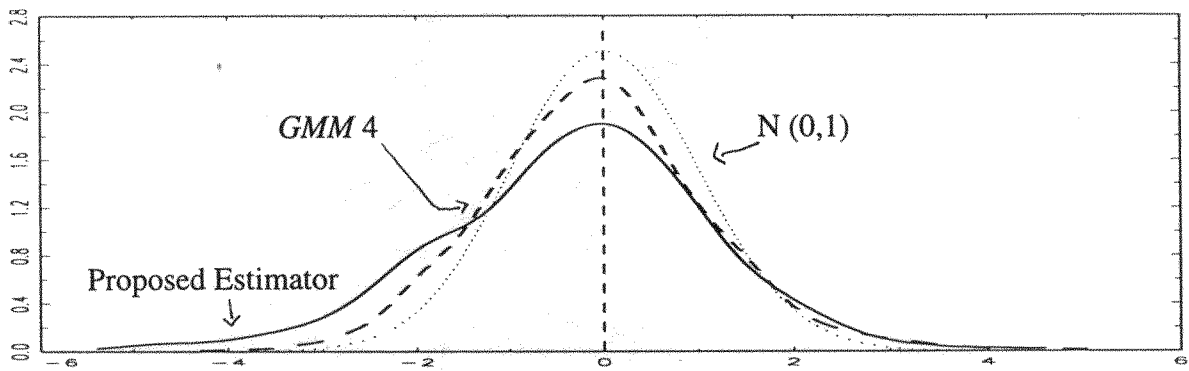
(c)  $GMM\ 12$

Notes: The densities are asymptotically  $N(0, v/v^*)$ , where  $v$  is the asymptotic variance of conventional GMM,  $v^*$  the asymptotic variance of our proposed estimator.

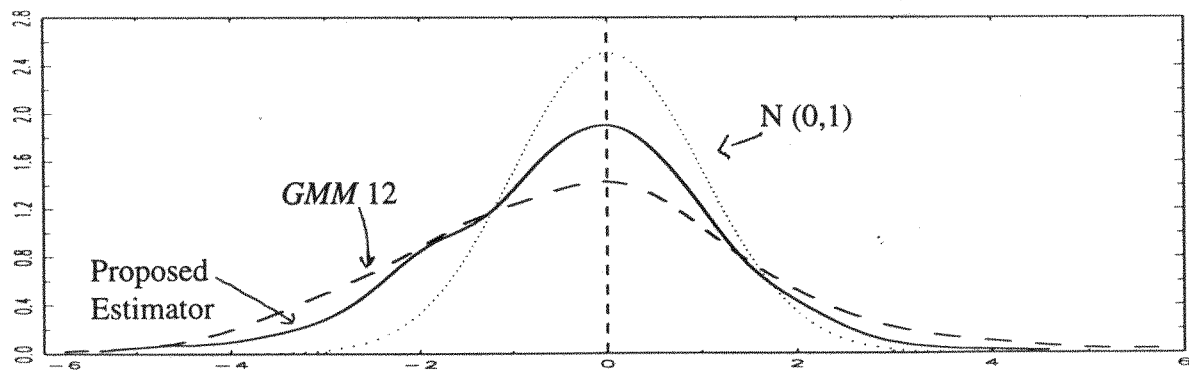
Figure 2: Density of Parameter Estimates,  $GMM_n$  vs. Proposed Estimator,  $T = 1000$ , DGP A



(a)  $GMM_1$



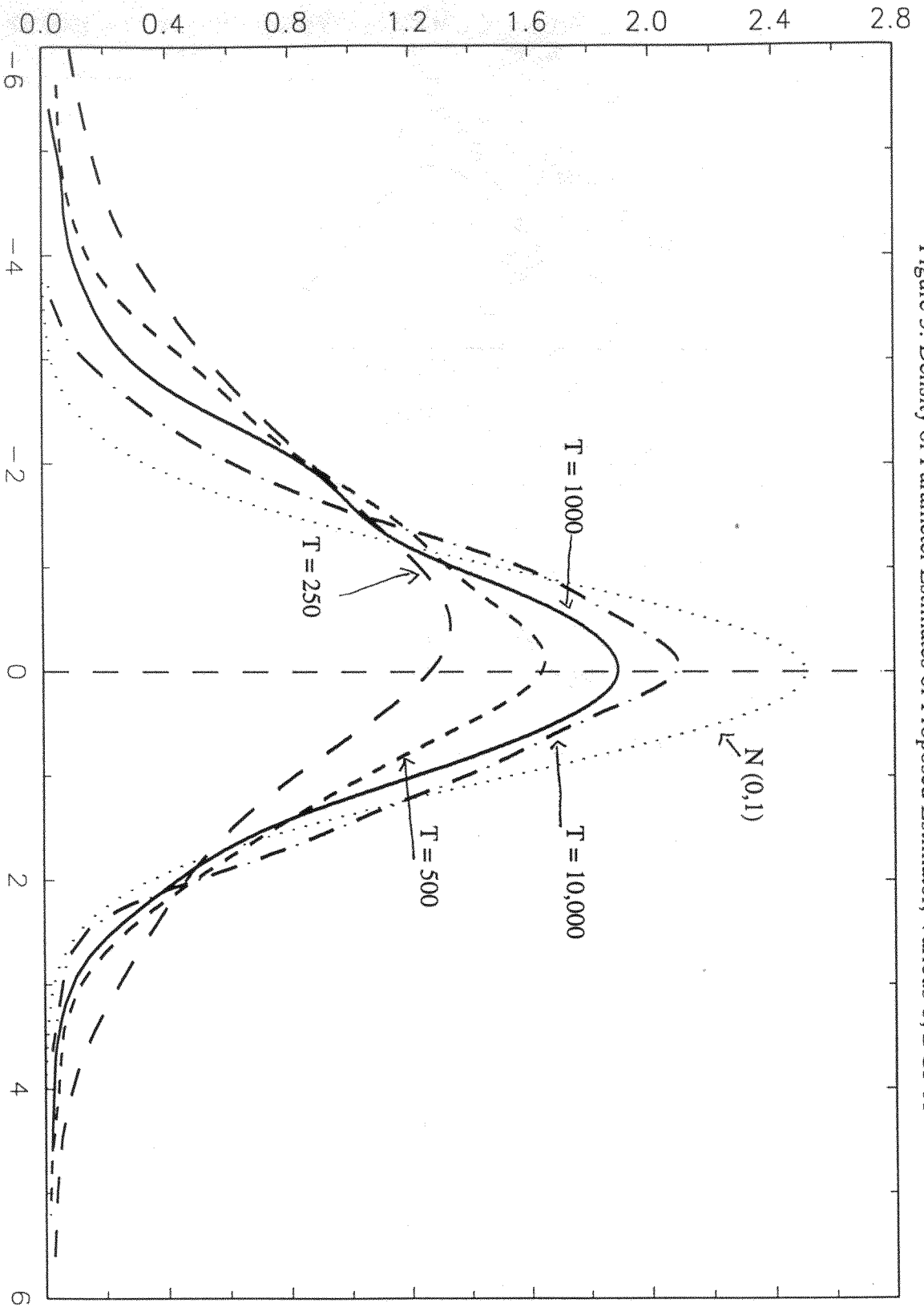
(b)  $GMM_4$



(c)  $GMM_{12}$

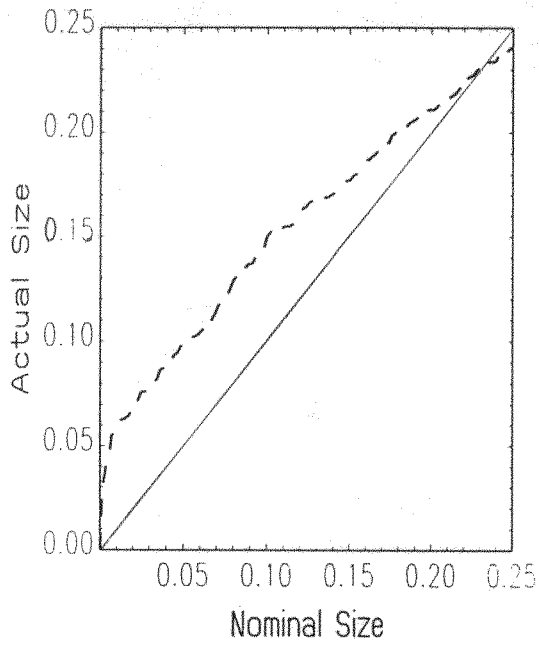
Notes: The densities are asymptotically  $N(0,1)$ .

Figure 3: Density of Parameter Estimates of Proposed Estimator, Various T, DGP A

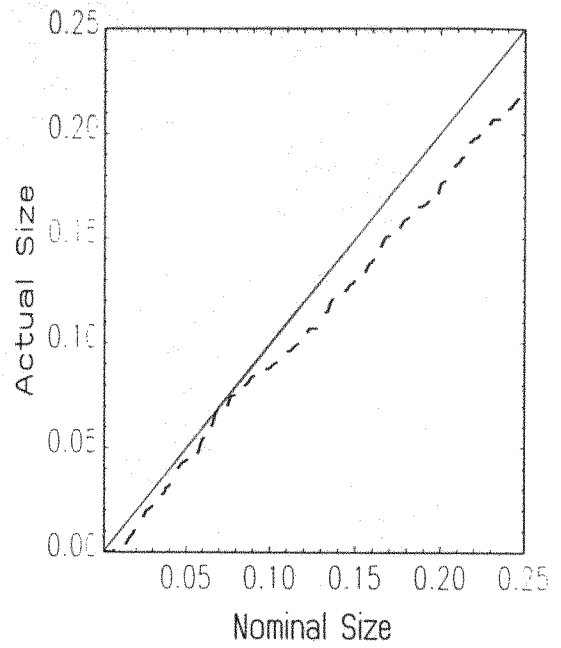


Notes: The densities are asymptotically  $N(0,1)$ .

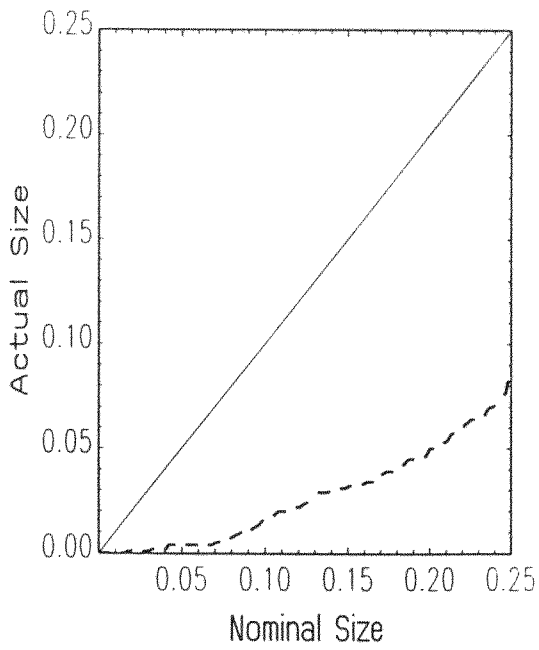
Figure 4: Actual and Nominal Size,  $T = 1000$ , DGP A



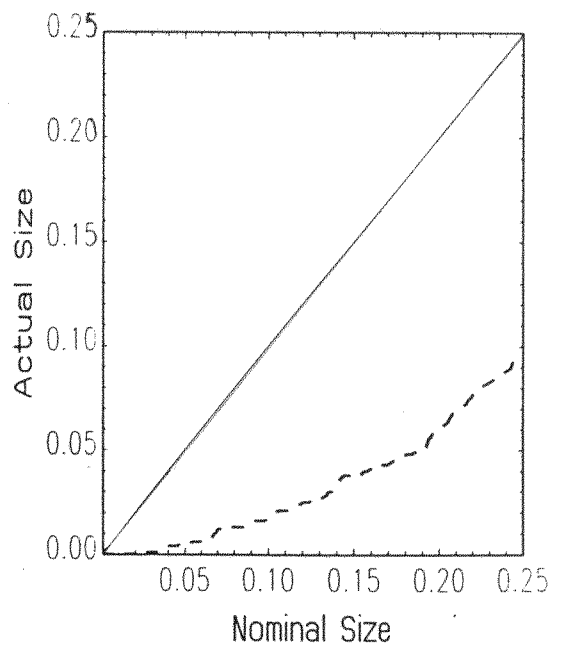
(a) *GMM\**



(b) *GMM1*



(c) *GMM4*



(d) *GMM12*