GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study

Torben G. Andersen
Department of Finance, J.L. Kellogg Graduate School of Management, Northwestern University, Evanston, Il 60208

Bent E. Sørensen
Department of Economics, Brown University, Providence, RI 02912

We examine alternative generalized method of moments procedures for estimation of a stochastic autoregressive volatility model by Monte Carlo methods. We document the existence of a tradeoff between the number of moments, or information, included in estimation and the quality, or precision, of the objective function used for estimation. Furthermore, an approximation to the optimal weighting matrix is used to explore the impact of the weighting matrix for estimation, specification testing, and inference procedures. The results provide guidelines that help achieve desirable small-sample properties in settings characterized by strong conditional heteroscedasticity and correlation among the moments.

KEY WORDS: Asymptotic standard errors; Generalized method of moments; Goodness of fit; Simulation techniques; Specification tests; Weighting matrix.

In recent years the interest in estimating stochastic volatility models has been strong. These models are complements as well as alternatives to the autoregressive conditionally heteroscedastic (ARCH) models (Bollerslev 1986; Engle 1982). The distinction between the two models relies on whether volatility is observable or not and may formally be stated in terms of measurability properties of the volatility process (Andersen 1992). Although ARCH models are more tractable, at least in the univariate case, there are several reasons why some researchers have turned their attention to a new class of models. First, multivariate ARCH models induce a proliferation of parameters that must be handled in an, arguably, ad hoc manner. Second, several theoretical models build on the concept of unobservable latent factors generating asset returns—for example, information-flow interpretations of the mixture-of-distributions hypothesis (Andersen 1996; Clark 1973; Epps and Epps 1976; Foster and Viswanathan 1995; Gallant, Hsieh, and Tauchen 1991; Tauchen and Pitts 1983), or low-dimensional factor structures that govern the joint mean and volatility features of returns (Diebold and Nerlove 1989; Engle, Ng, and Rothschild 1990; Ho, Perraudin, and Sørensen 1996; King, Sentana, and Wadhwani 1994; Laux and Ng 1993). Third, allowing for time-varying volatility in diffusions, which are important in modern finance and economics, leads naturally to stochastic volatility specifications. Fourth, the close association between ARCH and diffusion models for high-frequency data (Nelson 1990, 1992; Nelson and Foster 1991) has generated considerable interest in the properties of alternative discrete-time specifications for returns and the interrelations among them.

Discrete-time approximations to diffusion processes have found frequent use in the option pricing literature in which lognormal autoregressive specifications for the volatility process serve as discretized Ornstein–Uhlenbeck processes. Early applications of the model include those of Taylor (1986), Johnson and Shanno (1987), Scott (1987), Hull and White (1987), and Wiggins (1987), and later applications include those of Melino and Turnbull (1990) and Perraudin and Sørensen (1994). In fact, this particular stochastic volatility model has come to dominate the field to the extent that it is referred to as *the* stochastic volatility model although it hinges on particular functional forms and distributional assumptions. It is interchangeably referred to as the lognormal stochastic autoregressive volatility model (Andersen 1994a), the autoregressive random variance model (Taylor 1994), or the stochastic variance model (Harvey, Ruiz, and Shephard 1994). Consequently, most of the accumulated evidence regarding estimation performance in stochastic volatility models applies to this specific model.

The lognormal stochastic volatility model has been estimated by a variety of means, including simple moment matching (MM) (Taylor 1986), generalized method of moments (GMM) (Melino and Turnbull 1990), quasi-maximum likelihood (QML) (Harvey et al. 1994), various simulated method of moment (SMM) procedures (Duffie and Singleton 1989; Gallant and Tauchen in press; Gourieroux, Monfort, and Renault 1993), Bayesian Markov-chain Monte Carlo analysis (MCMC) (Jacquier, Polson, and Rossi 1994, henceforth JPR), and simulation-based maximum likelihood estimation (SML) (Danielsson 1993, 1994; Danielsson and Richard 1993). Apart from MM, GMM, and QML, the approaches are computationally intensive. The Monte Carlo evidence of JPR, however, suggests that GMM and QML have poor finite-sample performance, both in terms of bias and root mean squared error (RMSE) of the estimated parameters when compared to the likelihood-based MCMC. Nonetheless, the relatively simple GMM and QML procedures will undoubtedly be used extensively in the foreseeable future due to the computational demands of the alternative methods. Moreover, different stochastic volatility models are surfacing rapidly, and the more computationally intensive simulation-based strategies (MCMC and SML) either have not been designed to perform estimation outside of the lognormal volatility setting or they remain unproven in these contexts and may turn out to be infeasible. GMM and SMM procedures are, however, likely to apply to a wide set of models (Andersen 1994a).

This article investigates the small-sample properties of GMM estimation of the lognormal stochastic volatility model. The issue was addressed by JPR and by Ruiz (1994). Both concluded that the GMM performs relatively poorly—the former found that GMM weakly dominates QML, but the latter reached the opposite conclusion. Although one can demonstrate that the performance of GMM can be improved relative to the results reported in these articles, and very much so in the latter case (Andersen 1994b; Andersen and Sørensen 1996), it is not surprising that conflicting evidence may be obtained given the number of specific choices that have to be made to implement the procedure. We take a comprehensive look at the relevant issues in a large-scale Monte Carlo study. To retain a benchmark, we rely primarily on the parameter and moment design used by JPR, but we expand on their setup by also exploring larger samples that are more representative of those used in typical studies based on high-frequency returns data.

We first address the choice of the number of moments to include in the estimation procedure. We find that this depends critically on sample size. As the sample expands, one should exploit additional moment restrictions. In small samples, however, the inclusion of an excessive number of moments results in more pronounced biases and larger RMSE. Thus, the use of additional information can be harmful. We conjecture that this occurs due to the need to obtain an estimate of the weighting matrix used in the GMM procedure. When N moments are used, we are implicitly asked to estimate $N(N+1)/2$ separate entries of the weighting matrix along with the sample moments. Clearly, if this dimensionality is large relative to sample size the estimates of the weighting matrix may be poor which, in turn, distorts the metric by which the GMM procedure operates. It suggests a fundamental trade-off for GMM: Inclusion of more information in the form of additional moment restrictions improves estimation performance for a given degree of precision in the estimate of the weighting matrix, but in small samples this must be balanced against the deterioration in the estimate of the weighting matrix as the number of moments expands. In the present model, we are able to provide a fairly transparent characterization of the trade-off. Of course, the optimal trade-off will reflect the particular model under consideration. Nonetheless, the qualitative conclusions are likely to apply to a general class of models, characterized by strong conditional heteroscedasticity and correlation between the sample moments—conditions that are almost universal in high-frequency financial-returns series.

Further evidence on the importance of estimation of the weighting matrix is obtained from Monte Carlo simulations of the GMM procedure in which the weighting matrix is fixed and approximately "true"—that is, estimated separately from a very large sample. This removes the main impact of the estimation error in the weighting matrix and, in addition, ensures that the sample moments are estimated independently of the weighting matrix. Our results confirm the preceding intuition. When the weighting matrix is estimated more precisely and independently of the sample moments, inclusion of additional moments almost uniformly improves estimation performance. Hence, the deterioration of the estimation performance observed in the simulations is, indeed, partly due to the use of poor weighting matrices. Our observations are in line with the motivation behind the study by Altonji and Segal (1993). In a very different setting, they also investigated the bias in GMM procedures induced by the dependence between the estimated moments and the weighting matrix.

In practice the "true" weighting matrix is not available but must be estimated along with the unconditional moments from the given sample. Although it is standard to rely on a nonparametric kernel estimate of the spectral density of the moment vector for this purpose, there is less consensus on the appropriate choice of kernel estimator. The majority of studies apply the White (1984) or the Bartlett kernel procedure with a fixed bandwidth that was advocated by Newey and West (1987). The previous discussion suggests that the choice of weighting-matrix estimator is potentially important. Andrews (1991) and Andrews and Monahan (1992) studied the properties of a general class of heteroscedasticity and autocorrelation consistent (HAC) estimators including the White and Bartlett estimators. Three types of modifications were suggested. First, it is possible to use an automatic (data-dependent) bandwidth. Second, the quadratic spectral (QS) kernel estimator is optimal in terms of truncated mean squared error within the HAC class for autocorrelation and heteroscedasticity of unknown form. Third, note that vector autoregressive prewhitened HAC estimators display superior finite-sample performance in several dimensions. We explore the virtues of these procedures in the present setting. Specifically, we investigate the consequences of (a) using an automatic rather than a fixed bandwidth in the Bartlett procedure, (b) combining the automatic bandwidth with prewhitening, (c) employing the QS kernel estimator of the weighting matrix rather than the Bartlett kernel. Finally, we explore the alternative bandwidth-selection scheme proposed by Newey and West (1994), both with and without prewhitening. In addition, some authors have used diagonal weighting matrices that may be a reasonable choice when estimates of the full weighting matrix are poorly behaved, so we briefly investigate this methodology as well.

An important issue that we do not pursue at length is the selection of which—rather than how many—moments to include in the GMM procedure. As long as we remain within the confines of the traditional GMM approach that requires closed-form solutions for the analytical moments, the choice is both limited and fairly straightforward. Intuitively, estimation efficiency is improved by using moments with low sample variability rather than high sample variability. We confirm this intuition through a few experiments that compare our leading choice of moments to the exclusive use of absolute or squared lagged return moments. In addition we consider some practical guidelines for the choice of alternative moments using the large-sample approximation to the optimal weighting matrix. It is possible, however, that important efficiency gains can be obtained by a more ingenious selection of the moments. For example, Gallant and Tauchen (in press) suggested using an auxiliary model as a moment generator based on the scores of a quasi-likelihood. Analyzing this approach by Monte Carlo methods requires an additional layer of estimation procedures and simulations and thus falls outside the scope of the present study. Nonetheless, thorough analysis of this type of procedures, based on the principles of efficient moment selection or indirect inference, is a logical next step and should be high on the agenda for future research in this area.

In addition, we do not implement the alternative GMM procedure recently advanced by Hansen, Heaton, and Yaron (1996). This involves simultaneous optimization of the GMM criterion function over both the sample and analytical moments of the model and the weighting matrix. The computational demands of this method were deemed impractical for inclusion in our simulation design, but the method provides yet another potential route for improvements of the small-sample properties of the GMM procedure and should be investigated in future research.

Our setup provides an ideal setting for an investigation of the Hansen (1982) χ^2 test of goodness of fit based on the overidentifying restrictions of the model in the context of strong conditional heteroscedasticity in the data. The test is very popular because it may be calculated as a by-product of the estimation procedure. Although the finite-sample properties of the χ^2 test statistic have been explored in several studies, including those of Tauchen (1986), Kocherlakota (1990), and Ferson and Foerster (1994), there is hardly any direct evidence on the finite-sample behavior of the test in a context like this one. We find that the statistic is far from χ^2 distributed in small samples, but, nonetheless, the 5%-level test has approximately the correct size when we adhere to our guidelines regarding the preferred type and number of moments to include in the estimation procedure. Furthermore, the performance of the test deteriorates sharply when those prescriptions are ignored. Indeed, a general pattern emerges: When an excessive number of moments is used we unambiguously find that the test is biased strongly in favor of accepting the model. Alternatively, if a minimal number of moments is used we invariably find that the test overrejects. Studies concerned with GMM estimation in the context of high-frequency return series often include many moments. These studies may suffer from very significant size distortions, and standard hypothesis tests may lack power. On the other hand, macroeconomic applications often rely on the just-identified case. To the extent that our analysis carries over to this environment, we expect poor small-sample behavior of the parameter estimates and a tendency for overrejection by the standard tests. This appears consistent with the findings of recent studies in this area—for example, those of Christiano and den Haan (1996) and Burnside and Eichenbaum (1994). Thus, a large portion of the literature conducts asymptotically motivated inference and specification testing that is tenuous in light of our findings.

The remainder of the article is organized as follows. Section 1 introduces the lognormal stochastic volatility model, discusses the specific choice of parameters we consider, and outlines the GMM estimation procedure. Section 2 describes our general Monte Carlo setup with emphasis on our handling of the simulations that are incompatible with converging estimates within the parameter space. Section 3 reports on the estimation performance of the GMM procedure in terms of bias and RMSE for each of our simulation designs, whereas Section 4 summarizes the evidence on the standard specification test based on overidentifying restrictions. Section 5 considers some issues of inference by studying the small-sample distribution of the studentized parameter estimates, and, finally, Section 6 provides concluding remarks and suggestions for future research.

1. THE STOCHASTIC VOLATILITY MODEL AND THE GMM PROCEDURE

We investigate the following simple version of the lognormal stochastic volatility model:

$$\begin{aligned} y_t &= \sigma_t Z_t \\ \ln \sigma_t^2 &= \omega + \beta \ln \sigma_{t-1}^2 + \sigma_u u_t, \end{aligned}$$

where (Z_t, u_t) is iid $N(0, I_2)$; that is, the error terms are mutually independent standard normals. The parameter vector is $\theta = (\omega, \beta, \sigma_u)$. For $0 < \beta < 1$ and $\sigma_u \geq 0$, the return innovation series, y_t, is strictly stationary and ergodic, and unconditional moments of any order exist. Throughout, we work with parameter values that satisfy these inequalities.

In the model, returns display zero serial correlation but dependency in the higher-order moments is induced through the stochastic volatility term, σ_t, which follows a first-order autoregressive [AR(1)] model in logarithms. The volatility persistence parameter, β, is estimated to be less than, but quite close to, unity in most empirical studies. Finally, the assumption of lognormality of the volatility process is a convenient parameterization that allows for closed-form solutions for the moments and is consistent with the evidence of excess kurtosis or "fat tails" in the unconditional return distribution.

The specification ignores the possibility of a nonzero, potentially time-varying mean return as proposed by, for example, Engle, Lilien, and Robins (1987), and it rules out correlation between the two error terms that would allow for an asymmetric "leverage effect" (e.g., Nelson 1991). This is done to retain the JPR benchmark and to keep the computational demands manageable. In addition, the simplified model remains a good first approximation for a variety of high-frequency financial-return series.

GMM estimation exploits the convergence of selected sample moments to their unconditionally expected values. We denote the vector of sample realizations of the moments at time t by $m_t(\theta) = (m_{1t}(\theta), \ldots, m_{Qt}(\theta))$, where the number of selected moments, Q, exceeds the dimension of θ—that is, the number of parameters to be estimated. The true parameter vector is denoted θ_0, and the sample moments are $M_T(\theta) = (M_{1T}(\theta), \ldots, M_{QT}(\theta))$, where $M_{iT}(\theta) = \sum_{t=j+1}^{T} m_{it}(\theta)/(T-j)$, for $i = 1, \ldots, Q$, and j is the maximum lag between the variables defining the sample moments. Finally, the corresponding vector of analytical moments is denoted $A(\theta)$. The GMM estimator, $\hat{\theta}_T$, minimizes the distance between $A(\theta)$ and $M_T(\theta)$ over the parameter space Θ in the following quadratic form: $\hat{\theta}_T = \arg\min_{\theta\in\Theta}(M_T(\theta) - A(\theta))'\Lambda_T^{-1}(M_T(\theta) - A(\theta))$, where the specific matrix is determined by the choice of the positive definite and possibly random weighting matrix, Λ_T. Under suitable regularity conditions, $\hat{\theta}_T$ is consistent and asymptotically normal (Hansen 1982): $T^{1/2}(\hat{\theta}_T - \theta_0) \sim N(0, \Omega)$. The optimal choice of weighting matrix, Λ^{-1}, in the sense of minimizing the asymptotic covariance matrix, Ω, is given by the inverse of the covariance matrix of the appropriately standardized moment conditions:

$$\Lambda = \lim_{T\to\infty} E\left[\sum_{t,\tau=1}^{T} (m_t - A(\theta_0))(m_\tau - A(\theta_0))'/T\right].$$

This matrix may be estimated by a kernel estimator for the spectral density of the vector of sample moments at frequency 0. The use of an appropriate weighting matrix is important. The return sample moments are likely to be heavily correlated and display strong serial dependence. If these features are ignored, say by using the identity matrix, there is likely to be a serious loss of efficiency. Indeed, when we attempt to estimate the present system with an identity weighting matrix, it becomes extremely ill behaved, and convergence is hardly ever obtained. Some preliminary scaling of the moments through the weighting matrix (e.g., by simple sample moment estimates) is simply a requirement for meaningful inference by GMM in this model.

Thus, to implement the GMM procedure, we face two basic choices, the selection of sample moments, $m_t(\theta)$, to use in estimation and the selection of the estimator of the weighting matrix, $\hat{\Lambda}_T^{-1}$, where $\hat{\Lambda}_T$ is an estimator of Λ based on T observations.

The main guide to moment selection is the erratic finite-sample behavior of higher-order moments, caused by the presence of fat tails in the return series. Asymptotic normality of $\hat{\theta}_T$ requires finite variances of the moment conditions and, for practical purposes, good estimates of these quantities in finite samples. This suggests a focus on the lower-order moments, which is consistent with current practice as well as the approach taken by JPR. Hence, for simplicity, we elect to rely on (subsets of) the 24 moments used by the latter. Letting $\mu = \omega/(1-\beta)$ and $\sigma^2 = \sigma_u^2/(1-\beta^2)$, the analytic expressions are as follows: $E|y_t| = (2/\pi)^{1/2}E(\sigma_t)$, $E(y_t^2) = E(\sigma_t^2), E|y_t^3| = 2\sqrt{2/\pi}E(\sigma_t^3), E(y_t^4) = 3E(\sigma_t^4)$, $E|y_t y_{t-j}| = (2/\pi)E(\sigma_t\sigma_{t-j})(j = 1, \ldots, 10)$, and $E(y_t^2 y_{t-j}^2) = E(\sigma_t^2\sigma_{t-j}^2)(j = 1, \ldots, 10)$, where, for any positive integer j and positive constants $r, s, E(\sigma_t^r) = \exp(r\mu/2 + r^2\sigma^2/8)$ and $E(\sigma_t^r\sigma_{t-j}^s) = E(\sigma_t^r)E(\sigma_t^s)\exp(rs\beta^j\sigma^2/4)$. The finite-sample properties of the GMM estimator for a variety of choices of the weighting matrix are explored in the following sections. Here, we only provide a few general remarks. The class of kernel estimators of the spectral density matrix is of the general form

$$\sum_{j=-T+1}^{T-1} k(j)\hat{\Gamma}_T(j),$$

where $k(j)$ are weights that may become 0 for $|j| > L_T$, a lag truncation parameter that grows toward infinity at a slower rate than T and $\hat{\Gamma}_T(j)$ is a covariance matrix estimator at lag j—that is, for $\hat{\theta}$, a consistent estimator of θ,

$$\hat{\Gamma}_T(j) = \frac{1}{T}\sum_{t=j+1}^{T} (m_t(\hat{\theta}) - A(\hat{\theta}))(m_{t-j}(\hat{\theta}) - A(\hat{\theta}))'.$$

The most obvious difference between kernel estimators is the shape of the weighting scheme $k(j)$, but the length (or bandwidth) of the weighting scheme determined by the parameter L_T, as well as the possibility of prewhitening, is also an important issue. Finally, it is possible that procedures based on weighting matrices outside of the class of kernel estimators possess attractive finite-sample properties. All of these questions are addressed later.

We conclude this section with an account of the parameter values that generate our return samples in the Monte Carlo simulations. We follow JPR and concentrate on an expected value of σ_t^2 of .0009, implying an annual standard deviation in weekly return data of around 22% and a coefficient of variation of σ_t^2 of unity. Then the choice of β determines the remaining parameters, ω and σ_u. They focus on $\beta = .90$ but report results for $\beta = .95$ and $\beta = .98$ as well. Accordingly, we use $\beta = .90$ as our leading case, but we do also experiment with higher values for the persistence parameters due to the plethora of studies reporting very high estimates of persistence in the volatility process. The result is the following three parameter settings:

$$\begin{aligned}(\omega, \beta, \sigma_u) &= (-.736, .90, .363)\\ &= (-.368, .95, .260)\\ &= (-.147, .98, .166).\end{aligned}$$

2. THE MONTE CARLO SETUP

The simulations were performed using GAUSS version 3.1 on RISC/6000 workstations and on 486 PC's. We used the OPTMUM procedure for optimization, predominantly relying on the BFGS algorithm but also sometimes on the NEWTON and other algorithms. We found no discrepancies when we repeated identical jobs with different algorithms or on different platforms. Many of the Monte Carlo experiments were performed using numerical derivatives, but some jobs were later rerun using analytical derivatives. This made absolutely no difference to the results.

We display results for the just-identified model (three moments) and for the number of moments being $M = 5, 9, 14,$

and 24. Our leading choice of moments consists of the selections denoted "Baseline set" in the Appendix. We rely on this set in the vast majority of the study. We consider sample sizes of $T = 500$ (following JPR), $T = 1{,}000, 2{,}000, 4{,}000$ and 10,000. A sample of 1,000–4,000 is not uncommon in studies using daily or weekly data, and the $T = 10{,}000$ simulation is relevant given the increasing availability of transactions data. We perform 1,000 Monte Carlo simulations for each (M, T) combination. For the design $T = 10{,}000$ and $M = 24$ this turns out to be computationally very demanding (several days of central processing unit time on the RISC/6000, model 550).

In each GMM estimation we performed three sets of iterations. In the first step we used a simple estimate of the weighting matrix, derived directly from the sample moments, but in the second and third steps we used the kernel-weighting matrix under examination. We never detected any noticeable difference between the second- and third-step estimations, and it is highly unlikely that a higher number of iterations over the weighting matrix would have made a noticeable difference.

For the lower sample sizes our estimation algorithm was frequently unable to locate a minimum for the criterion function within the parameter space. Inspecting the iterations of the algorithm, we invariably noted a similar pattern in these situations. During the iterations the estimated value for the autoregressive parameter for volatility, $\hat{\beta}$, converged to 1, and as $\hat{\beta}$ became approximately 1, the iterations would crash as the weighting matrix became singular or the criterion function diverged to infinity. To interpret our results for the lower sample sizes, it is critical to identify the source of these nonconvergence problems. The preceding observations and our analysis presented in Section 3 suggest that the main issue is the lack of an interior optimum for the objective function over the open parameter space ($\beta < 1$) rather than a failure of the optimization routines to detect the optimum.

We dealt with the crashes in the following fashion: If the weighting matrix was singular, we trapped the error. If it happened during the third and final estimation step, this simulation was discarded, but if it happened during one of the two preliminary steps, we went on to the next estimation step with $\hat{\beta}$ adjusted to $\min\{\hat{\beta}, \beta_{\max}\}$, where the upper bound, $\beta_{\max}$, equals .999999. If $\hat{\beta}$ went above $\beta_{\max}$ during the iterations, we penalized the criterion function to force the estimate below $\beta_{\max}$. In almost all of these cases the algorithm was unable to obtain convergence. We allowed for a maximum of 50 iterations in the first round, 200 in the second round, and 500 in the third round. An estimation was discarded if it reached the maximum number of iterations in the third round. We are convinced that these maximum numbers of iterations were sufficiently large, so we did not eliminate any (or negligibly few) estimation experiments that eventually would have resulted in convergence. We continued the simulations until we obtained 1,000 sets of third-round iterations that terminated with convergence.

It is not unproblematic to discard simulations that do not result in convergence because we systematically eliminate samples that appear compatible with high values of β, and hence a significant downward bias in our mean β estimate may result when many simulations are discarded. Similar biases will materialize for the remaining mean parameter estimates because they are correlated with the estimates of β. The impact on the RMSE is not predictable, however. If the discarded β estimates were replaced by an estimate near unity (effectively the strategy chosen by JPR), say .99999, then the RMSE of $\hat{\beta}$ is enhanced or reduced depending on the simulation design. Rather than rely on corrective procedures of this nature, we conclude that GMM is poorly equipped to deal with inference problems in cases that correspond to simulation designs for which we find many nonconverging samples.

3. RESULTS

This section reports on our findings for each of the simulation designs.

3.1 Fixed-Bandwidth Bartlett Kernel

Our first set of results relies on weighting matrices estimated by the Bartlett kernel using a fixed lag length of $L_T = 10$, and $M = 3, 5, 9, 14$, and 24. The weighting scheme takes the form $k(j) = 1 - j/L_T$ and $k(j) = 0$ for $j \geq L_T$. This is the kernel estimator advocated by Newey and West (1987), and it is widely used in the literature. The choices of the lag length and, in particular, the number of moments included in the procedure are probably slightly on the low side relative to standard practice, but they are not unreasonable in light of the findings for the automatic bandwidth reported on later. Consequently, it serves as a natural benchmark for the subsequent experiments. Moreover, it generates some interesting qualitative conclusions that hold up across all the designs.

Before turning to the interpretation of the tables, we note that the use of 1,000 replications for each simulation design results in small Monte Carlo errors for the reported statistics. Moreover, a direct and simple upper bound on the Monte Carlo standard error (exact if point estimates are unbiased) is available as $N^{-1/2}$ RMSE. This bound is tight for samples in excess of 500; for example, consider the $M = 14, T = 1{,}000$ entry in Table 1. The upper bound on the standard error is $1{,}000^{-1/2}$ (.657, .088, .143) = (.021, .003, .005), which is small and indistinguishable from the direct estimates of the standard error.

Our first results are given in Table 1. Consider the first rows based on sample size $T = 500$. Some interesting conclusions emerge immediately. The first and somewhat disturbing finding is that approximately a third of the estimations fail to converge. We explicitly assess whether this appears reasonable in a more controlled setting in the following section. A second problem is that both the RMSE and the biases are substantial. In particular, the RMSE of the parameter ω is about as large as its mean estimate. A third, and intriguing, observation is that the preferred choice of the number of moments for this sample size is $M = 9$. Although asymptotic theory may suggest that it is optimal to include as many moments as possible in the estimation

Table 1. Simulated Mean and Root Mean Squared Error: Bartlett Kernel, Fixed Bandwidth (lag length = 10), $(\omega, \beta, \sigma_u) = (-.736, .900, .363)$

# moments	3	5	9	14	24
$T = 500$					
$\hat{\omega}$	−1.951 (1.854)	−1.636 (1.763)	−1.063 (1.063)	−1.076 (1.400)	−1.176 (1.129)
$\hat{\beta}$	.736 (.250)	.786 (.209)	.858 (.137)	.861 (.148)	.844 (.149)
$\hat{\sigma}_u$	.503 (.237)	.413 (.197)	.307 (.167)	.294 (.214)	.286 (.168)
No convergence	519	528	398	402	434
$T = 1{,}000$					
$\hat{\omega}$	−1.475 (1.259)	−1.135 (.913)	−.801 (.571)	−.829 (.657)	−.946 (.732)
$\hat{\beta}$	.800 (.170)	.847 (.123)	.892 (.077)	.888 (.088)	.873 (.097)
$\hat{\sigma}_u$	.458 (.201)	.372 (.142)	.297 (.135)	.287 (.143)	.287 (.137)
No convergence	422	298	129	88	72
$T = 2{,}000$					
$\hat{\omega}$	−1.180 (.889)	−.867 (.564)	−.730 (.382)	−.747 (.388)	−.839 (.429)
$\hat{\beta}$	.840 (.120)	.883 (.075)	.901 (.052)	.899 (.053)	.887 (.058)
$\hat{\sigma}_u$	.422 (.163)	.343 (.126)	.305 (.108)	.302 (.108)	.309 (.101)
No convergence	301	135	20	11	3
$T = 4{,}000$					
$\hat{\omega}$	−.984 (.632)	−.791 (.397)	−.740 (.255)	−.745 (.227)	−.834 (.288)
$\hat{\beta}$	.867 (.086)	.893 (.054)	.900 (.035)	.899 (.031)	.887 (.038)
$\hat{\sigma}_u$	.392 (.135)	.346 (.100)	.331 (.072)	.325 (.068)	.333 (.068)
No convergence	144	38	3	0	0
$T = 10{,}000$					
$\hat{\omega}$	−.796 (.405)	−.763 (.268)	−.744 (.157)	−.740 (.139)	−.795 (.161)
$\hat{\beta}$	.892 (.055)	.896 (.036)	.899 (.021)	.900 (.019)	.892 (.022)
$\hat{\sigma}_u$	.360 (.103)	.354 (.068)	.347 (.043)	.344 (.042)	.348 (.039)
No convergence	65	0	0	0	0

NOTE: The reported statistics are based on 1,000 simulated samples of sample size equal to the indicated T. For each cell, the first number shows the mean and the second the root mean squared error (in parentheses).

procedure to maximize the information extracted from the sample, this is clearly not correct for this sample size. A fourth noteworthy point is that the exactly identified model ($M = 3$) fares extremely poorly. Estimation for this case consists of solving three equations in three unknowns. A "nonconvergence" is reported when the solution falls outside of the parameter space; that is, $\beta \geq 1$. For comparison with the other entries in the table the results provided for $M = 3$ exclude the parameter estimates associated with such nonconvergence. One might conjecture that the just-identified approach is attractive if problems in estimating the weighting matrix are the source of the poor performance of the GMM procedure. Our findings, however, effectively eliminate this procedure from the range of desirable options.

Fortunately, the quality of the inference improves rapidly as the sample size increases. For estimates of ω and β, the RMSE shrinks faster than is to be expected from standard root-T asymptotics, and the RMSE (at least for $T = 1{,}000$ to 10,000) for σ_u declines roughly in line with root T. This reflects the fact that the biases disappear more quickly for the first two parameters. It further indicates that the RMSE for the smaller samples are driven by outliers that tend to disappear at a rapid rate as the sample size increases. Harvey and Shephard (1993) reported similar dramatic reductions in RMSE with increasing sample size when they estimated the model by QML. The extreme number of crashes reported for the smaller samples further reinforces this conclusion. In addition, for the larger samples the parameter bias is all but eliminated except for $\hat{\sigma}_u$ which remains downward biased.

We also find that the results based on a higher number of moments tend to improve relative to the $M = 9$ case as the sample size grows. In fact, for $T = 4{,}000$ the RMSE for the choice of $M = 14$ uniformly dominates the $M = 9$ case, but $M = 24$ generally underperforms relative to both. Moreover, it is evident that for the design $T = 2{,}000$ and $M = 14$ the problems with lack of convergence are no longer of much concern. Interestingly, in the case of a very large sample, $T = 10{,}000$, it often remains preferable to use 14 rather than 24 moments. Nevertheless, even for samples of this size, the exactly identified model still crashes fairly often, and its performance in terms of RMSE is clearly inferior to all other choices of M. It is safe to conclude that our findings soundly refute the usefulness of the just-identified approach in this setting.

Our results are roughly in agreement with the corresponding results of JPR, which are based on 500 simulations with $T = 500$ and $T = 2{,}000$. They deal quite differently with the problems of nonconvergence because they "force" the estimate of β at .99 rather than discarding the results. For $T = 500$, our results based on 24 moments, but excluding the discarded simulations, are slightly better in terms of RMSE than those reported by JPR. For the autoregressive volatility parameter β, this is clearly not due to the elimination of nonconverging estimates. For β, letting the estimate be less than but approximately equal to unity when nonconvergence occurs actually reduces the overall RMSE ($1 - .9 = 1$ is less than the reported RMSE) and

Table 2. Simulated Mean and Root Mean Squared Error: Bartlett Kernel, Fixed Bandwidth ($1.2 * T^{1/3}$), $(\omega, \beta, \sigma_u) = (-.736, .900, .363)$

# moments	9	14	24
$T = 1{,}000$			
$\hat{\omega}$	−.831 (.603)	−.820 (.760)	−.995 (.769)
$\hat{\beta}$	.888 (.081)	.891 (.081)	.867 (.102)
$\hat{\sigma}_u$	.300 (.131)	.283 (.149)	.294 (.134)
No convergence	105	51	48
Fixed lag	12.00	12.00	12.00
$T = 2{,}000$			
$\hat{\omega}$	−.787 (.360)	−.823 (.366)	−.908 (.439)
$\hat{\beta}$	.894 (.049)	.889 (.049)	.878 (.058)
$\hat{\sigma}_u$	.321 (.093)	.320 (.089)	.317 (.090)
No convergence	13	5	0
Fixed lag	15.12	15.12	15.12
$T = 4{,}000$			
$\hat{\omega}$	−.804 (.243)	−.800 (.222)	−.904 (.295)
$\hat{\beta}$	.891 (.033)	.892 (.030)	.878 (.039)
$\hat{\sigma}_u$	.343 (.060)	.335 (.058)	.345 (.055)
No convergence	0	0	0
Fixed lag	19.05	19.05	19.05
$T = 10{,}000$			
$\hat{\omega}$	−.769 (.142)	−.790 (.129)	−.846 (.176)
$\hat{\beta}$	.896 (.019)	.893 (.017)	.886 (.023)
$\hat{\sigma}_u$	.353 (.036)	.352 (.031)	.355 (.032)
No convergence	0	0	0
Fixed lag	25.85	25.85	25.85

NOTE: The reported statistics are based on 1,000 simulated samples of sample size equal to the indicated T. For each cell, the first number shows the mean and the second the root mean squared error (in parentheses).

almost eliminates the bias in the mean estimate. Not surprisingly, similar but even stronger conclusions follow from our $M = 9$ case. For $T = 2{,}000$, our $M = 24$ case produces almost identical results to theirs, both with respect to the mean estimates and to the RMSE. This is particularly encouraging because the number of nonconverging simulations is negligible for this design (less than half a percent), and it confirms basic compatibility between the two studies.

An immediate question concerns the robustness of the findings in Table 1. Thus, we next investigate an alternative set of GMM estimates based on a different, and probably more reasonable, interpretation of the concept of a *fixed-bandwidth* Bartlett kernel. Andrews (1991) pointed out that $\hat{\Lambda}_T$ converges to Λ at the fastest possible rate when the bandwidth grows with $T^{1/3}$. This suggests letting $L_T = \gamma T^{1/3}$ for a given γ. Hence, the bandwidth is fixed for a given sample, but we allow it to grow with the size of the sample. We choose $\gamma = 1.2$ as our leading case, implying a lag length of 12, 15, 19, and 25 for sample sizes $T = 1{,}000, 2{,}000, 4{,}000$, and 10,000, but we also investigate shorter and longer lags by letting $\gamma = .6, .9, 2, 5$, and 10. The lag lengths for $\gamma = .6$ and .9 correspond roughly to the average lag length picked by the automatic bandwidth procedure for the Bartlett kernel with $M = 14$ (Table 5, Sec. 3.5), and straddle $L_T = 10$ (Table 1), but the longer lag length picked by the other choices of γ are in line with those used later in the article. Consequently, comparisons across tables remain meaningful.

Table 2 collects the results from our leading case among the alternative fixed-bandwidth GMM estimates. First, we notice that the relative performance across the designs in Table 2 largely mirrors that of Table 1. Second, we find that the choice of wider bandwidths for the larger samples generally is beneficial. For example, for the $M = 14$ design, which tends to perform well, the RMSE's are strictly lower than before for all sample sizes above $T = 1{,}000$—that is, when the impact of the nonconverging samples is negligible. The identical observation holds true for $M = 9$, but the evidence for the $M = 24$ design is mixed.

Figures 1 and 2 provide evidence for the performance across a wider set of bandwidths. For brevity, only results for the $M = 14$ design are included, and we only consider a small ($T = 1{,}000$), large ($T = 4{,}000$), and very large sample ($T = 10{,}000$). Figure 1 displays the RMSE of β and Figure 2 the RMSE of σ_u as functions of the bandwidth parameter γ. The RMSE for ω is not shown, but it displayed the same pattern as the RMSE for β. According to the asymptotic theory for estimation of the weighting matrix, a fixed value of γ is optimal, but this is not borne out by the RMSE of the estimated parameters. For the parameter β, it is evident that a small bandwidth is optimal for the small sample (although the results for this sample size should be interpreted cautiously due to the elimination of nonconverging samples). For $T = 4{,}000$, a clear U shape emerges, implying that an intermediate choice of bandwidth is preferable. This pattern is also discernible for $T = 10{,}000$, but here the penalty for choosing a very large bandwidth has declined sharply as the right leg of the U shape has flattened. For σ_u the evidence is somewhat different, with a longer bandwidth being optimal in small samples, although the U shape for the larger samples is quite similar to the pattern found for β.

Although the gains obtained from optimizing over the bandwidths thus are nontrivial, the gains realized by including the appropriate number of moments in the estimation procedure appear more substantial. Table 2 reveals that $M = 9$ seems to dominate $M = 14$ and $M = 24$ for small samples (subject to the usual caveat), but, as in Table 1, $M = 14$ dominates $M = 9$ for T larger than 2,000, and in most cases $M = 14$ also dominates $M = 24$, except that the latter occasionally provides the best available estimate of σ_u for the larger samples. This may imply that a choice of M between 14 and 24 might dominate both in some cases.

In summary, we find that the quality of inference is quite sensitive to the number of moments included in the estimation procedure relative to sample size. In addition, there is some evidence that a fairly large number of lags should be incorporated in the kernel estimators for the larger samples.

We conjecture that the eventual deterioration in the performance of GMM, as more moments are incorporated in the procedure (for a given sample size), is linked to problems with the estimated weighting matrix used in the objective function. The sample moments are quite highly correlated, which may result in a badly conditioned weighting matrix. In addition, because this matrix includes, for example, $25 * 24/2 = 300$ elements for $M = 24$, the many implicitly estimated parameters may in part be responsible for the rather disappointing results.

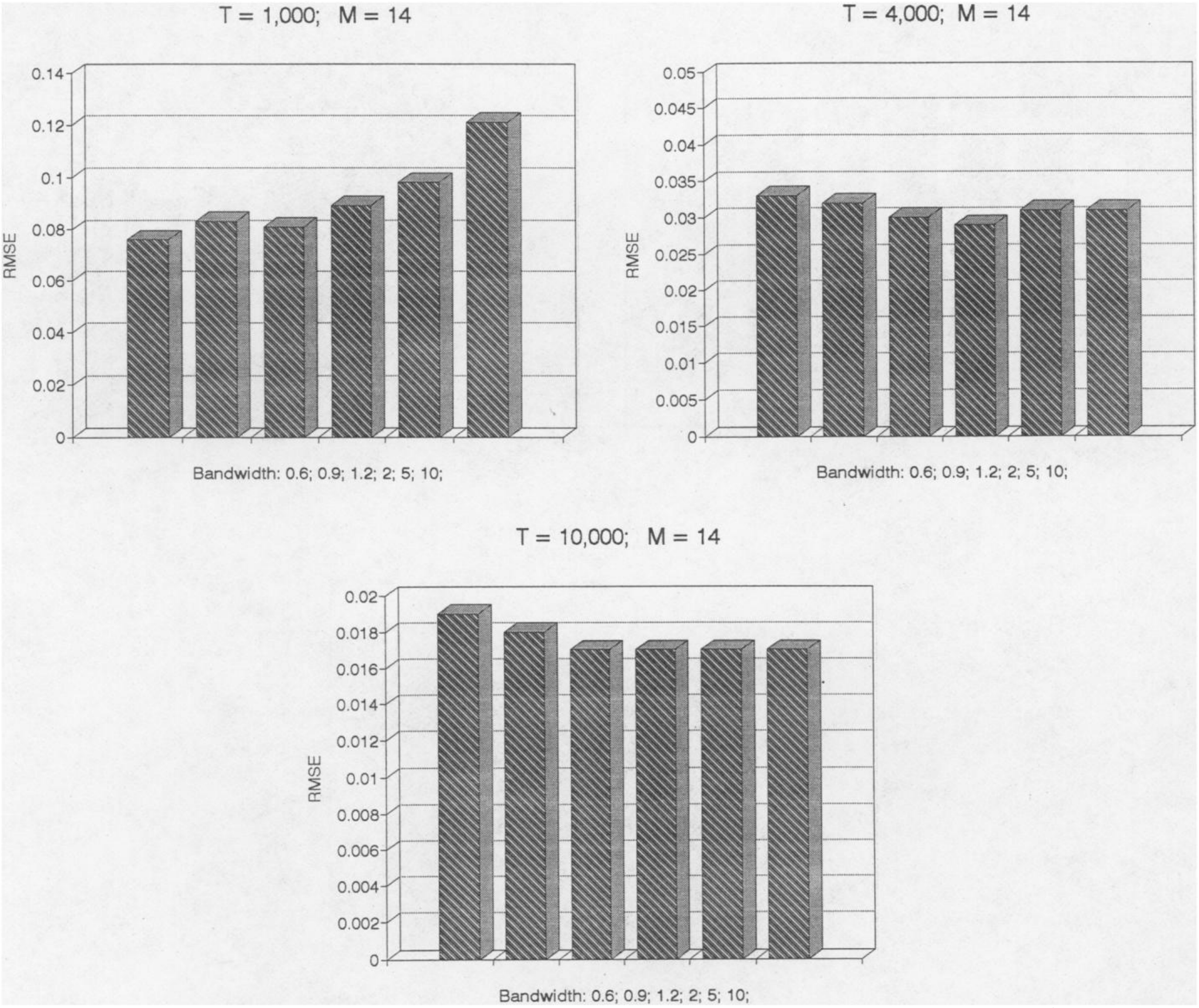


Figure 1. Root Mean Squared Error (RMSE) for GMM Estimates of β. The results are based on 1,000 converging estimates for each combination of sample size and (fixed) bandwidth. The GMM procedure is implemented using an estimated weighting matrix determined by the Bartlett kernel with a lag length given by the bandwidth parameter, γ, as $L_T = \gamma \cdot T^{1/3}$. The figures display findings for sample sizes T = 1,000; 4,000; 10,000. All estimates are based on M = 14 moments.

3.2 GMM Estimation Using the "True" Weighting Matrix

Previously, we conjecture that there is a trade-off between the amount of information used in estimation (the number of sample moments included) and the quality of the objective function (the precision of the estimate of the appropriate weighting matrix). This trade-off changes with sample size because the weighting matrix—for a given number of moments—is more precisely estimated as the sample grows. The empirical results presented in the preceding section provide indirect support for this interpretation, but a more direct exploration of this hypothesis is available in the current setting. Rather than estimate both the sample moments and the weighting matrix from the given simulated sample, we estimate the latter from a separate and very large simulated sample and exploit this as an approximation to the "true" optimal weighting matrix in the subsequent simulations. In this manner the weighting matrix is estimated with higher precision and the estimate is independent of the sample moments. If the estimate of the weighting matrix is critical for the performance of the GMM procedure, this should lead to an appreciable improvement for the larger samples.

Table 3 reports on the results from this simulation experiment. We repeat the estimations from Table 1 (except for the exactly identified model, of course) using a fixed, exogenous approximation to the true weighting matrix. This weighting matrix was constructed from simulations for each choice of M, using 50,000 observations and a lag length of 50. This choice corresponds to $1.36 * T^{1/3}$, which belongs to the suitable range according to our earlier findings.

The findings are revealing. First, notice that the simulations now are much less prone to crash. Second, and even more to the point, there is an almost uniform improvement in the RMSE as more moments are included. This supports our suspicion that a poorly estimated weighting matrix is

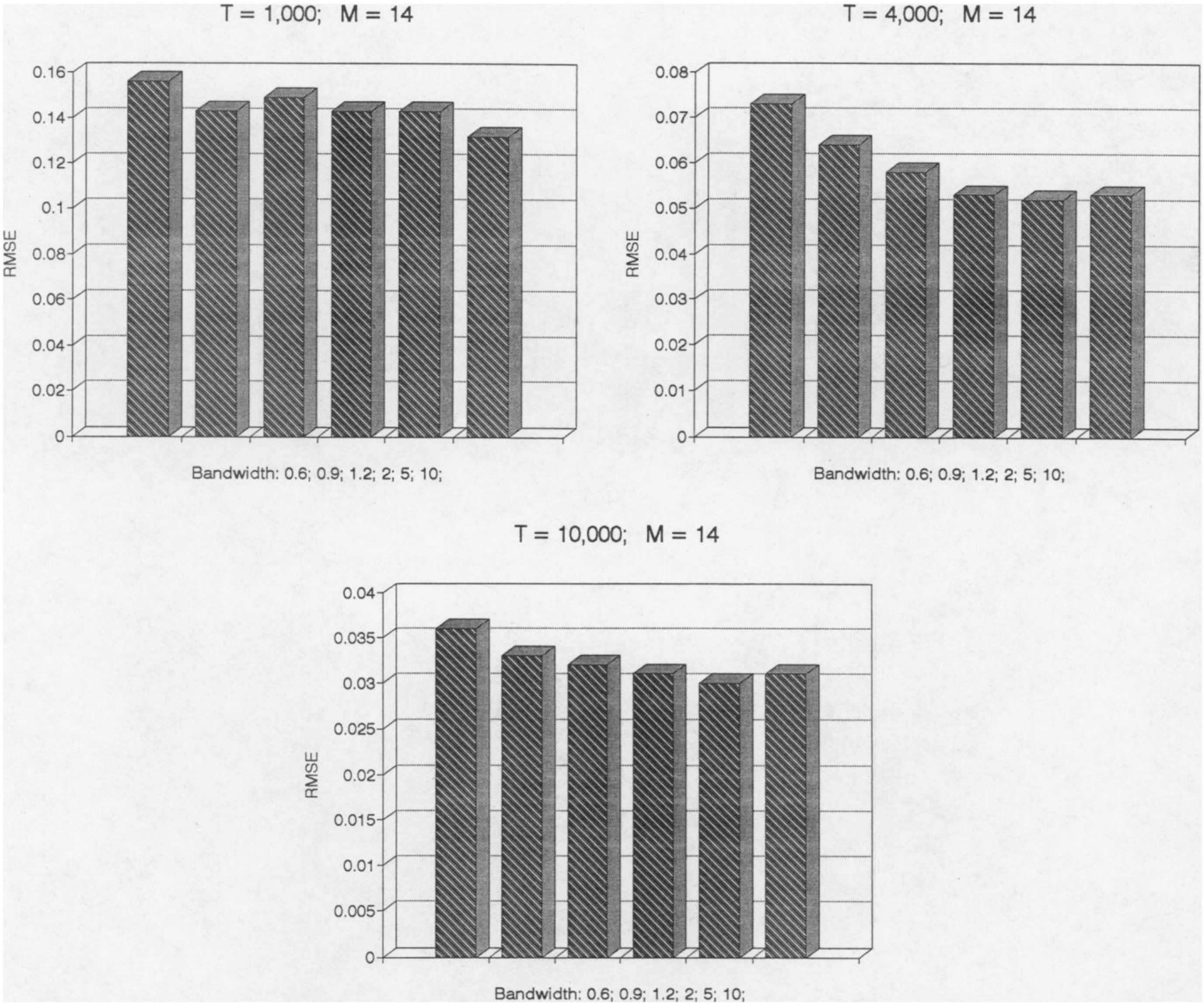


Figure 2. Root Mean Squared Error (RMSE) for GMM Estimates of σ_u. The results are based on 1,000 converging estimates for each combination of sample size and (fixed) bandwidth. The GMM procedure is implemented using an estimated weighting matrix determined by the Bartlett kernel with a lag length given by the bandwidth parameter, γ, as $L_T = \gamma \cdot T^{1/3}$. The figures display findings for sample sizes $T = 1{,}000; 4{,}000; 10{,}000$. All estimates are based on $M = 14$ moments.

the root of many of the problems encountered in Tables 1 and 2. It is, however, clear that, even with this approximation to the true weighting matrix, the inference is poor for $T = 500$, in which the RMSE remains very large. It is also noteworthy that around 10,000 observations are needed before most of the bias in $\hat{\sigma}_u$ is gone.

The procedure underlying the results reported in Table 3 is, of course, not feasible in practice but serves to highlight the potential gains that may be obtained by including the various moments in the estimation procedure. We pursue this issue further in Subsection 3.3. Moreover, the results point to a critical influence from the estimation of the weighting matrix. Thus, we shall examine several different strategies for choosing the weighting matrix.

The results allow an informal comparison to the Bayesian estimator proposed by JPR. The RMSE from the simulations using the "true" weighting matrix provides an approximate lower bound for the RMSE that can be obtained using the same set of moments and an estimated weighting matrix, as we substantiate later in the article. It is interesting to observe that—for the present parameter constellation and moments—GMM cannot be expected to match the efficiency of the Bayes estimator as reported in their table 5. On the other hand, it is feasible to improve the efficiency of their GMM estimator. Specifically, for the three parameters via the Bayes estimator they reported the RMSE (.15, .02, .034) and for the GMM estimator (.42, .06, .10). The corresponding RMSE entries (for $T = 2{,}000$, $M = 24$) in Table 3 are (.275, .037, .070). Whether the relative efficiency gains associated with the use of the Bayesian estimator are similar for alternative designs, including more persistent volatility processes, can be addressed by similar means. The answer is not obvious because higher volatility persistence appears to induce an improvement in the efficiency of the GMM estimator relative to the Bayes estimator.

Table 3. Simulated Mean and Root Mean Squared Error: Exogenous Approximation to the "True" Weighting Matrix: $(\omega, \beta, \sigma_u) = (-.736, .900, .363)$

# moments	5	9	14	24
$T = 500$				
$\hat{\omega}$	−1.372 (1.265)	−1.118 (1.252)	−1.237 (1.446)	−1.132 (1.146)
$\hat{\beta}$	.815 (.166)	.852 (.150)	.839 (.160)	.850 (.136)
$\hat{\sigma}_u$	.427 (.181)	.378 (.148)	.393 (.140)	.388 (.134)
No convergence	356	77	15	23
$T = 1{,}000$				
$\hat{\omega}$	−1.040 (.803)	−.898 (.620)	−.930 (.658)	−.920 (.499)
$\hat{\beta}$	.859 (.108)	.879 (.077)	.874 (.080)	.875 (.067)
$\hat{\sigma}_u$	.392 (.160)	.373 (.109)	.381 (.091)	.381 (.092)
No convergence	200	12	2	3
$T = 2{,}000$				
$\hat{\omega}$	−.874 (.548)	−.800 (.330)	−.824 (.399)	−.835 (.275)
$\hat{\beta}$	.882 (.074)	.891 (.045)	.889 (.041)	.887 (.037)
$\hat{\sigma}_u$	.373 (.132)	.369 (.084)	.374 (.064)	.376 (.070)
No convergence	93	2	1	0
$T = 4{,}000$				
$\hat{\omega}$	−.768 (.404)	−.759 (.217)	−.786 (.175)	−.792 (.186)
$\hat{\beta}$	.896 (.055)	.897 (.029)	.893 (.024)	.892 (.025)
$\hat{\sigma}_u$	.356 (.107)	.366 (.057)	.373 (.050)	.373 (.049)
No convergence	27	1	0	1
$T = 10{,}000$				
$\hat{\omega}$	−.742 (.270)	−.757 (.143)	−.761 (.117)	−.772 (.125)
$\hat{\beta}$	.899 (.037)	.897 (.019)	.897 (.016)	.895 (.017)
$\hat{\sigma}_u$	.359 (.073)	.367 (.040)	.368 (.031)	.371 (.036)
No convergence	2	0	0	0

NOTE: The reported statistics are based on 1,000 simulated samples of sample size equal to the indicated T. For each cell, the first number shows the mean and the second the root mean squared error (in parentheses).

3.3 Asymptotic Efficiency for Alternative Moment Selections

We do not address the general question regarding the optimal choice of moments that has been studied recently by, for example, Gallant and Tauchen (in press). Instead, we explore the implications of choosing different sets of moments among the ones that lead to closed-form, analytic solutions for the moments. This allows us to stay within the classical GMM framework.

The approximation to the true weighting matrix, Λ, allows us to find the asymptotic standard deviations of the parameters estimates for alternative selections of moment conditions. These calculations may be useful for a preliminary selection of moments in the spirit of Ruiz (1994), who also relied on asymptotic standard deviations as a benchmark for finite-sample performance.

From Hansen (1982) we have the following expression for the asymptotic variance–covariance matrix, Ω, of the parameter estimates, $\hat{\theta}_T$:

$$\Omega = a(\theta_0)'\Lambda a(\theta_0), \quad \text{where} \quad a(\theta_0) = \left.\frac{\partial A(\theta)}{\partial \theta}\right|_{\theta=\theta_0}.$$

Because we have an estimate of Λ and we, in addition, have analytic expressions for $A(\theta)$ and thus $a(\theta)$, we may estimate the true Ω by simply plugging in our estimate of Λ and the analytic derivatives evaluated at the true parameter vector. Hence, we obtain a tangible approximation to the asymptotic variance–covariance matrix of the parameter estimates. The implied asymptotic standard errors for the individual parameters should provide a natural lower bound for the RMSE that we can achieve in our finite-sample experiments. The only caveat associated with this interpretation is that the weighting matrix estimated from even this very large sample continues to display a fairly large degree of variability. We investigate this problem further later in the article. The standard errors obtained from a sample of 50,000 should, nonetheless, serve as a gauge for the efficiency that we can hope to attain in our shorter samples in the simulation designs, and this seems to be confirmed by our subsequent results.

Table 4 reports the asymptotic standard errors normalized to correspond to a sample size of 2,000 for alternative selections of moments. We expect the use of more moments to improve inference as additional information is exploited and most of the impact of estimation error in the weighting matrix has been eliminated. This expectation is basically confirmed, but the pattern is nonetheless striking. The decline in the standard errors as we move from $M = 5$ to $M = 9$ is remarkable. Clearly, the extra four moments contribute significant additional information regarding the parameters. The improvement from $M = 9$ to $M = 14$ is also substantial, albeit less dramatic. But the move from $M = 14$ to $M = 24$ is barely noticeable. The inclusion of the final 10 moments apparently adds very little information. This suggests that the use of 24 moments may be excessive in any of our simulation designs. The small gain in information is likely not sufficient to compensate for the loss in efficiency associated with deteriorating estimates of the weighting matrix.

We further study the impact of introducing a new set of lagged moments in the estimation procedure. We chose the following (third order) moments:

$$E|y_t y_{t-j}^2| = (2/\pi)^{(1/2)} E(\sigma_t \sigma_{t-j}^2), \qquad j = 1, \ldots, 10.$$

Including some of these among the 14 moments (and excluding some lagged absolute or squared moments) always results in larger RMSE. This occurs irrespective of whether all lagged moments or only a part of the lagged moments are of this type. Hence, we conjecture that inclusion of such lagged moments is unlikely to improve estimation performance. We experimented with the composition between the lagged squared and absolute moments in the designs with 9 and 14 moments. The changes are in all instances minor, and none provide significant improvements over our leading choice of moments. We conclude that the difference in estimation performance across designs with a different number of included moments is due largely to the increase in the number of moments rather than the specific identity of those moments.

To gauge the empirical relevance of the results, we examined by Monte Carlo simulation whether 14 sample moments of the form $E[y_t^2 y_{t-i}^2]$ (labeled "Quadratic moments" in the Appendix) or the form $E[|y_t y_{t-i}|]$ ("absolute moments") contain more information about the parameters. We found that the results based on the absolute moments have the lower RMSE, but the gains were quite minor. These results were reported by Andersen and Sørensen (1995). When we compared to the results for our baseline set of 14 moments in Table 2, we found even less clear-cut evidence. For $T = 2{,}000$, the RMSE of the absolute-moments-based procedure dominates, but for $T = 4{,}000$, the baseline mix of moments appears better, and again all differences are minor. This may be compared to the asymptotic standard in Table 4 ($m14a, m14c$, and $m14d$). Again, the differences between the three sets are minor, but it is noteworthy that this semianalytic approach ranks the designs in the same way that our simulations do; that is, the uses of mixed and absolute lagged moments are close, but with a minor edge to the mixed moments, whereas relying exclusively on lagged squared moments is inferior to both. This suggests that the semianalytic efficiency bounds may be relevant for econometric practice. The issue appears, however, not to be of first-order importance, and from this evidence it seems that our prior selection of a mix of absolute and quadratic lagged moments performs reasonably well.

Table 4. Asymptotic Std. Deviations Using "True" Kernel—Alternative Models: $(\omega, \beta, \sigma_u) = (-.736, .900, .363)$

Parameter	ω	β	σ_u
5 moments	.5355	.0727	.1316
9 moments			
Baseline set of moments ($m9a$)	.3071	.0417	.0767
Alternative set ($m9b$)	.2934	.0398	.0756
14 moments			
Baseline set of moments ($m14a$)	.2511	.0341	.0651
Alternative set ($m14b$)	.2529	.0344	.0646
"Absolute" moments ($m14c$)	.2526	.0343	.0651
"Quadratic" moments ($m14d$)	.2641	.0359	.0679
Absolute 3rd moments ($m14e$)	.3089	.0420	.0783
Mix of low abs 1st, 2nd, and 3rd moments ($m14f$)	.3361	.0456	.0866
Alternative mix of 1st, 2nd, 3rd moments ($m14g$)	.2670	.0363	.0669
24 moments			
Baseline set	.2414	.0328	.0629
34 moments			
All moments included	.1987	.0270	.0523

NOTE: Standard deviations are normalized to correspond to $T = 2{,}000$. The exact selection of moments for each model is listed in the Appendix.

In conclusion, we note that this analytic procedure may be useful for preliminary assessment of the appropriate estimation design whenever closed-form expressions for the moments can be obtained. Such calculations can potentially eliminate the need for large-scale simulation experiments over various moment designs by providing a reasonable guide to the relative importance of different moments for estimation performance. This insight may be relevant for quite general GMM estimation problems.

3.4 Analysis of Nonconvergence

The present setting is ideal for an assessment of the reported number of "crashes." If the asymptotic normal approximation remains good within the neighborhood of the true parameter vector, then the standard error of $\hat{\beta}$ provides an estimate of the probability with which the $\hat{\beta}$ estimate will exceed unity and thus potentially induce a crash. For example, the reported standard error for $M = 5$ of .0727 in Table 4 implies that for $T = 2{,}000$ $\hat{\beta}$ will exceed unity with probability $1 - \Phi([1 - .9]/.0727) = .084$, where $\Phi(\cdot)$ denotes the cumulate density function of the standard normal distribution. We should thus expect that 8.4% of the estimations crash due to an estimate of β that falls outside of the parameter space. The actual number of crashes for this cell in Table 3 is 93 or 8.5% $[= 93/(1{,}000 + 93)]$. A similar analysis suggests 5, 1, and 1 crashes for $M = 9, 14$, and 24, respectively, whereas the realized number of crashes were 2, 1, and 0. It appears that this analysis is able to rationalize the propensity of the estimations to crash. For the smaller samples, in which the asymptotic standard errors can be obtained by simple transformations of the ones given for $T = 2{,}000$, we expect the corresponding calculations based on the normal approximation to be less precise, which, indeed, is what we find. The orders of magnitude remain correct, however. For $T = 500$ the asymptotic standard errors predict 24.6% , 10.1% , 5.8% , and 5.3% crashes, whereas the actual occurrences numbered 356, 77, 15, and 23, or 26.2% , 7.1% , 1.5% , and 2.2% . When the weighting matrix is estimated from much smaller simulated samples, the parameter estimates become more erratic, and we should expect to find an even higher proportion of crashes, as we do.

The interpretation provided previously suggests that estimates of, for example, β in the right tail of the empirical distribution have been eliminated due to the boundary of the parameter space. This feature is, indeed, quite appar-

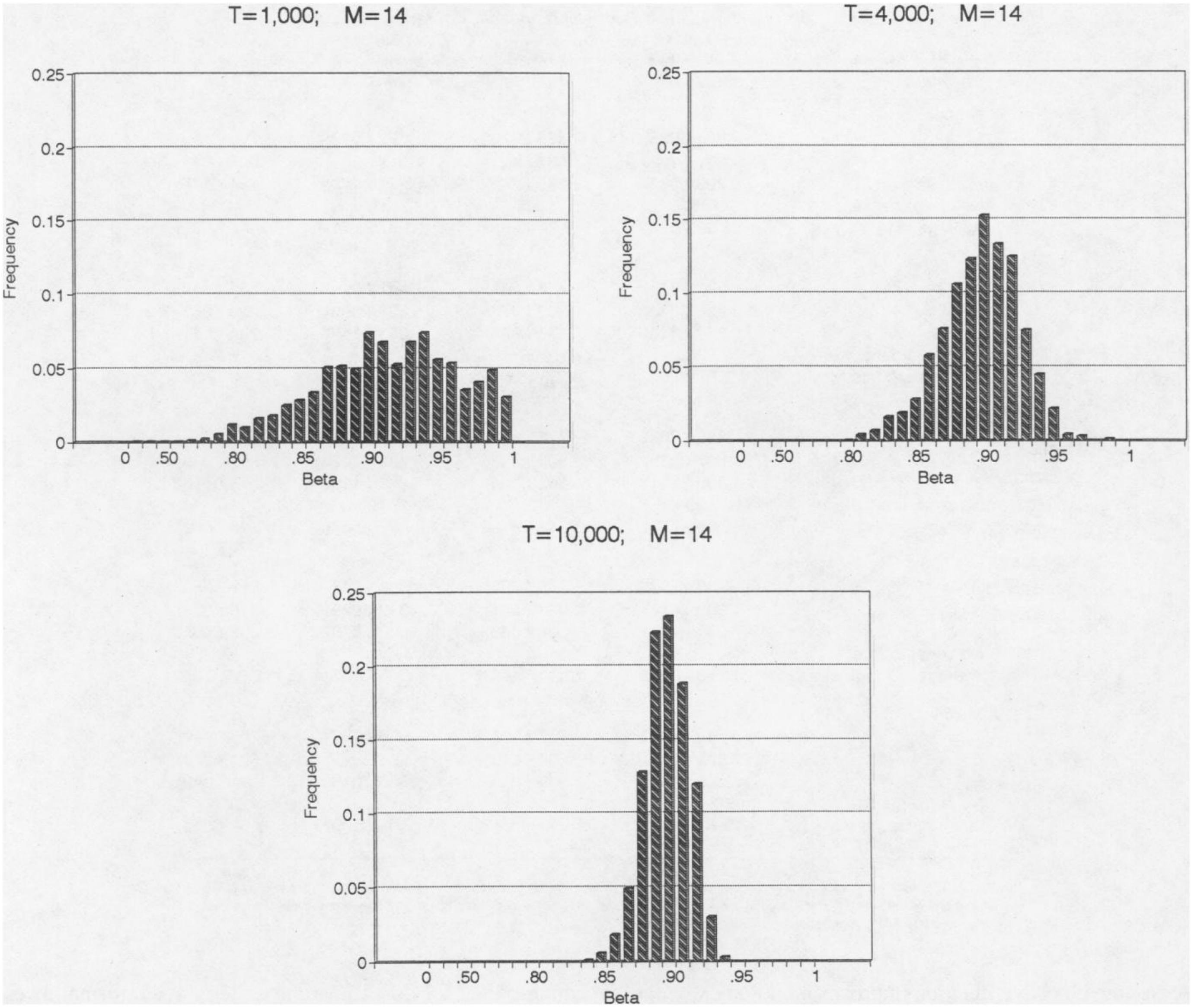


Figure 3. The Distribution of GMM Estimates of β. The figure shows the fraction of estimates that fall within the different 5% fractiles. The results are based on 1,000 converging estimates for each sample size. The GMM procedure is implemented using an estimated weighting matrix determined by the Bartlett kernel using a lag length of 1.2 $\cdot$ $T^{1/3}$. Figures are displayed for sample sizes T = 1,000; 4,000; 10,000. All estimates are obtained with M = 14 moments.

ent in plots of the distribution of the β estimates. Figure 3 shows the distribution for $M = 14$ and $T =$ 1,000, 4,000, and 10,000 when using the simulation design in Table 2. It seems apparent for the small sample that the right tail has been truncated at $\beta = 1$, and furthermore the distribution displays a long left tail. For the larger sample sizes, in which we do not encounter convergence problems, the right tail is bounded away from $\beta = 1$, and the tails become closer to being symmetric as the sample size increases, although there is still some evidence of left skewness in the distribution even for $T =$ 10,000. Andersen and Sørensen (1995) displayed similar results for alternative moment selections.

The preceding is at best indirect evidence for the hypothesis that the crashes are associated with instances in which the objective function does not attain a minimum within the parameter space. We therefore explored the issue further. Specifically, we collected samples that did not converge in the $(M, T) =$ (5,500) design when using the BFGS algorithm. We then subjected these samples to a variety of alternative optimization algorithms. Although the procedures differ in their ability to accommodate estimates very close to the boundary of the parameter space, they all eventually fail for the nonconverging samples as numerical problems terminate the routine. Andersen and Sørensen (1995) reported the value of the objective function as the estimated parameter vector approaches the boundary for a few arbitrarily selected nonconverging samples. The values were obtained by fixing the β grid and optimizing over the other parameters using the NEWTON algorithm in GAUSS, which generally was the best algorithm in terms of accommodating β estimates close to unity (we found that the number of crashes was not sensitive to the choice of algorithm but that the BFGS algorithm sometimes crashed sooner along the increasing β sequence). Nonetheless, in all

Table 5. Simulated Mean and Root Mean Squared Error: Bartlett Kernel, Automatic Bandwidth, $(\omega, \beta, \sigma_u) = (-.736, .900, .363)$

# moments	9 (non-pw)	14 (non-pw)	14 (pw)	24 (pw)
$T = 500$				
$\hat{\omega}$	−.958 (.998)	−.876 (1.200)	−.622 (.588)	−.703 (.650)
$\hat{\beta}$	.871 (.132)	.887 (.121)	.915 (.080)	.905 (.087)
$\hat{\sigma}_u$	.295 (.167)	.265 (.201)	.235 (.170)	.224 (.178)
No convergence	567	365	342	422
Average lag (std. dev.)	6.09 (7.72)	7.73 (7.85)	1.07 (1.70)	1.46 (2.32)
$T = 1{,}000$				
$\hat{\omega}$	−.691 (.515)	−.726 (.541)	−.567 (.414)	−.585 (.452)
$\hat{\beta}$	.906 (.070)	.902 (.073)	.923 (.056)	.920 (.061)
$\hat{\sigma}_u$	.279 (.143)	.273 (.138)	.251 (.148)	.234 (.161)
No convergence	176	66	77	87
Average lag (std. dev.)	7.67 (9.00)	9.35 (8.64)	1.36 (1.78)	1.80 (2.01)
$T = 2{,}000$				
$\hat{\omega}$	−.685 (.368)	−.726 (.352)	−.592 (.311)	−.627 (.315)
$\hat{\beta}$	.907 (.050)	.902 (.048)	.920 (.042)	.915 (.043)
$\hat{\sigma}_u$	.300 (.111)	.301 (.104)	.279 (.115)	.272 (.118)
No convergence	29	3	10	10
Average lag (std. dev.)	9.06 (9.63)	11.06 (9.43)	1.82 (2.37)	2.26 (2.62)
$T = 4{,}000$				
$\hat{\omega}$	−.736 (.253)	−.760 (.227)	−.645 (.217)	−.662 (.225)
$\hat{\beta}$	.900 (.034)	.897 (.031)	.912 (.029)	.910 (.031)
$\hat{\sigma}_u$	.330 (.071)	.328 (.066)	.309 (.078)	.302 (.083)
No convergence	1	0	0	0
Average lag (std. dev.)	10.68 (10.15)	13.37 (10.03)	2.03 (1.95)	2.62 (2.34)
$T = 10{,}000$				
$\hat{\omega}$	−.749 (.154)	−.777 (.131)	−.696 (.132)	−.708 (.140)
$\hat{\beta}$	.898 (.021)	.895 (.018)	.905 (.018)	.904 (.019)
$\hat{\sigma}_u$	.348 (.041)	.349 (.034)	.336 (.045)	.331 (.049)
No convergence	0	0	0	0
Average lag (std. dev.)	12.78 (9.31)	16.54 (11.12)	2.52 (2.00)	3.34 (2.45)

NOTE: The reported statistics are based on 1,000 simulated samples of sample size equal to the indicated T. For each cell, the first number shows the mean and the second the root mean squared error (in parentheses). (pw): The prewhitening technique is applied using univariate AR(1) approximations to each sample moment.

cases the objective function improves monotonically until it eventually explodes. The P values obtained for these nonconverging samples did not correspond to bad fits of the model but seemed evenly distributed over the unit interval.

In addition, we picked 100 nonconverging samples from the preceding design and estimated the parameters in the just-identified case, $M = 3$, using a genuine subset of the five moments $(m1, m2$, and $m6)$ as the identifying moments. In 98 of the 100 cases, the implied estimate of β was above unity and (by construction) that of σ_u was negative. For one of the remaining two samples, the alternative subset of moments $(m1, m2$, and $m15)$ resulted in a similar "crash." The remaining sample was characterized by an exceptionally high fourth moment relative to the second moment, but no further exploration was undertaken.

Our findings support the interpretation that the crashes are associated with the lack of interior optima within the parameter space. This provides a rationale for discarding the nonconverging samples and interpreting the reported results as representative of the subset of GMM results that succeed in achieving convergence. On the other hand, one may suspect that practitioners may experiment with alternative choices of moments for a given sample before abandoning their GMM estimation strategy. An informal investigation of this possibility revealed that such procedures usually will detect a collection of moments that achieves convergence. Such exploratory search over alternative selections of moments will induce a type of bias in reported results that is virtually impossible to quantify within our simulation setting. Consequently, there are numerous reasons to emphasize the difficulty of interpreting the results for the smaller samples. Fortunately, the results for the intermediate and larger samples are basically unaffected by these nonconvergence problems.

3.5 Bartlett Kernel With Endogenous Bandwidth

Table 5 reports on GMM estimation using the Bartlett weighting matrix but with lag length chosen according to the suggestions of Andrews (1991). In this subsection we discuss the results when the weighting matrix is not prewhitened, displayed in the columns labeled "non-pw." We report the results for $M = 9$ and $M = 14$, because these moment selections clearly dominated the results for 5 and 24 moments. Specifically, we chose an AR(1) approximation to the sample moments for the purpose of determining a suitable bandwidth, and we therefore rely on the appropriate part of formula (6.4) of Andrews (1991). The exact formulas are as follows: Let $(\hat{\rho}_m, \hat{\sigma}_m^2), m = 1, \ldots, M$, denote the estimates of the autoregressive and innovation variance parameter for each of the moments, and let $K = \sum_{m=1}^{M} \hat{\sigma}_m^4/(1 - \hat{\rho}_m)^4$. Then the Bartlett lag length is chosen as $L_T = 1.1447(\hat{\alpha}(1)T)^{1/3}$, where $\hat{\alpha}(1) = \sum_{m=1}^{M} (4\hat{\rho}_m^2\hat{\rho}_m^4)/[K(1 - \hat{\rho}_m)^6(1 + \hat{\rho}_m)^2]$.

The impact of the endogenous choice of lag length is apparent. The average number of included lags grows not only with sample size but also, rather significantly, with the number of included moments [this is particularly striking from the full set of simulations reported by Andersen and Sørensen (1995), in which results were also reported for 5 and 24 moments]. The latter is ignored by the fixed-bandwidth procedures. The difficulty of accounting for this factor in an appropriate fashion prior to estimation provides a strong argument in favor of the automatic, or data-dependent, bandwidth choice.

Most of the conclusions from Table 1 still hold up. It remains preferable to use 9 moments for the lower values of T and 14 moments for the higher values of T. For $T = 500$ and $M = 9$, the estimations appear more prone to crash. Moreover, the upward bias in the mean estimates of ω now is less pronounced and in some cases ($M = 9, T = 1{,}000$, or $T = 2{,}000$) it has changed sign. It is also clear from a comparison of Tables 1 and 5 that it is almost always preferable, in terms of RMSE and bias, to use the automatic, or plug-in, bandwidth relative to the rather conservative choice of $L_T = 10$. Note also that the bias in the important β parameter is sharply reduced for the empirically relevant estimation with 14 moments and 1,000–2,000 observations.

Interestingly, comparisons to the fixed bandwidth in Table 2 and Figures 1 and 2 provide a more mixed picture. The automatic bandwidth procedure again performs uniformly well for the small samples ($T = 1{,}000$). For $T = 2{,}000$ the automatic bandwidth also stands up well against the previous procedures, but the evidence is mixed whenever the fixed-bandwidth choices are longer than the average ones chosen by the data-dependent procedure. Finally, for the larger sample sizes ($T = 4{,}000$ and $T = 10{,}000$) the procedures relying on the longest (average) lag length seem to dominate in terms of RMSE. This confirms our earlier findings regarding lag length for the large sample sizes, in which a fixed γ between 1.2 and 2 may be a sensible choice.

Encouraged by the significant improvements form this procedure, we turn to the prewhitening method suggested by Andrews and Monahan (1992).

3.6 Bartlett Kernel With Prewhitening

Prewhitening consists of a preliminary transformation that flattens (prewhitens) the spectral density of the sample moment vector prior to applying the kernel estimator, thus improving the properties of the estimator, and then inverting the transformation to obtain an estimate of the original spectral density at frequency 0. If $V_t(\hat{\theta})$ is the orthogonality condition, $m_t - A(\hat{\theta})$, evaluated at a consistent parameter vector $\hat{\theta}$ obtained from a preliminary estimation step, then the suggestion of Andrews and Monahan (1992) is to (a) fit a first-order vector autoregressive [VAR(1)] (or higher-order VAR) to V_t: $V_t = \hat{B}V_{t-1} + V_t^*$, (b) determine the weighting matrix $\tilde{\Lambda}_T$ using the prewhitened residual V_t^* according to the method of Andrews (1991), and then (c) find an estimate $\hat{\Lambda}_T$ by "recoloring": $\hat{\Lambda}_T = (I - \hat{B})^{-1}\tilde{\Gamma}_T(1 - \hat{B}')^{-1}$.

We performed a few experiments using the VAR(1) prewhitening procedure. Andrews and Monahan suggested that the singular values of $\hat{B}$ be restricted to force the $\hat{B}$ matrix to be stable. We follow them by letting singular values in excess of .97 equal .97. The method did not, however, perform well. This is most likely due to imprecision in the estimates of the $\hat{B}$ matrix. For some designs, the estimations did not converge in many cases. In others, the number of lags selected after prewhitening was often larger than the number selected before prewhitening, and even though the RMSE for σ_u declined slightly, the RMSE of the other parameters deteriorated sharply.

We chose instead to use the simpler expedient of fitting a univariate AR(1) model to each series and then using as our $\hat{B}$ a diagonal matrix with the univariate AR(1) coefficients along the diagonal. This approach remains true to the spirit of Andrews and Monahan because they did not suggest the VAR as the true model but rather as a convenient ad hoc way of flattening the spectrum.

The results from these experiments are reported in Table 5 in the columns labeled "pw." For the prewhitened weighting matrix, $M = 14$ and $M = 24$ were uniformly better than the results for 5 and 9 moments [available from Andersen and Sørensen (1995)]. First, notice the dramatic drop in average lag length relative to that of the preceding section. For the smaller sample sizes the results represent a remarkable improvement in RMSE for ω and β, but the estimate of σ_u is severely biased, and the RMSE's on this parameter generally increase relative to those in Table 2. For the higher sample sizes and 14 or 24 moments, there is generally a trade-off between more precise estimates of ω and β relative to σ_u because the downward bias on the latter remains clearly discernible. Finally, note that the use of 14 moments is almost uniformly the preferred choice for this procedure.

In conclusion, the results for this approach are somewhat mixed. It may appear to improve inference for some parameters in small samples, but this finding should be weighted against the very significant increase in the instances of nonconvergence for these samples. In addition, it improves the RMSE significantly for a subset of the parameters in the designs with large samples and many moments, so the approach may be attractive in certain instances.

3.7 The Quadratic Spectral Kernel

Andrews (1991) showed that the QS kernel dominates the Bartlett kernel according to an asymptotic truncated mean squared error criterion when the system is characterized by heteroscedasticity and autocorrelation of unknown form. The weighting scheme takes the form

Table 6. Simulated Mean and Root Mean Squared Error: Quadratic Spectral, Automatic Bandwidth, $(\omega, \beta, \sigma_u) = (-.736, .900, .363)$

# moments	9 (non-pw)	14 (non-pw)	14 (pw)	24 (pw)
$T = 2{,}000$				
$\hat\omega$	−.716 (.386)	−.711 (.376)	−.585 (.318)	−.600 (.338)
$\hat\beta$	.903 (.052)	.904 (.051)	.920 (.043)	.918 (.046)
$\hat\sigma_u$	.306 (.112)	.295 (.111)	.277 (.117)	.264 (.125)
No convergence	56	0	15	15
Average lag (std. dev.)	5.20 (3.82)	6.02 (3.81)	1.53 (.97)	1.81 (1.08)
$T = 4{,}000$				
$\hat\omega$	−.719 (.266)	−.726 (.251)	−.640 (.219)	−.641 (.237)
$\hat\beta$	.902 (.036)	.901 (.034)	.913 (.030)	.913 (.032)
$\hat\sigma_u$	.324 (.080)	.320 (.076)	.308 (.079)	.296 (.090)
No convergence	3	1	0	0
Average lag (std. dev.)	5.73 (3.28)	6.80 (3.60)	1.69 (.96)	2.05 (1.16)
$T = 10{,}000$				
$\hat\omega$	−.746 (.162)	−.756 (.143)	−.687 (.141)	−.700 (.137)
$\hat\beta$	.899 (.022)	.897 (.019)	.907 (.019)	.905 (.019)
$\hat\sigma_u$	.348 (.043)	.346 (.041)	.334 (.047)	.330 (.049)
No convergence	0	0	0	0
Average lag (std. dev.)	6.72 (3.47)	7.76 (3.83)	2.00 (.93)	2.41 (1.36)

NOTE: The reported statistics are based on 1,000 simulated samples of size equal to the indicated T. For each cell, the first number shows the mean and the second the root mean squared error (in parentheses). (pw): The prewhitening technique is applied using univariate AR(1) approximations to each sample moment.

$$k(j) = \frac{25}{12\pi^2 (j/L_T)^2} \times \left[\frac{\sin(6\pi(j/L_T)/5)}{6\pi(j/L_T)/5} - \cos(6\pi(j/L_T)/5)\right].$$

We examined the performance of the QS kernel in some detail, using both fixed and automatic bandwidths plus prewhitening. From Andrews (1991), the automatic bandwidth takes the form $L_T = 1.3221(\hat\alpha(2)T)^{1/5}$, where $\hat\alpha(2) = \sum_{m=1}^{M} (4\hat\rho_m^2 \hat\sigma_m^4)/[K(1-\hat\rho_m)^8]$ and $K, \hat\sigma_m$, and $\hat\rho_m$ are defined in Section 3.5.

The findings were quite similar, but overall slightly inferior as measured by RMSE, to the results reported previously for the Bartlett kernel. An indication of the findings is provided in Table 6, which reports on a subset of the automatic bandwidth designs. Again, prewhitening is clearly beneficial for the smaller sample sizes, but the same tradeoff between the precision in the estimates of the parameters ω and β versus σ_u that we noted previously shows up for the larger samples. Thus, given the particular nature of the positive second-order moment dependency in our series, it appears that the QS estimator does not improve on the performance of the Bartlett kernel.

3.8 The Newey–West Lag-Selection Scheme

Finally, we implemented the procedure advocated by Newey and West (1994) that is based on the Bartlett kernel but uses a different lag-selection criterion. Specifically, the bandwidth is chosen as follows: If x_t is the $Q \times 1$ residual vector from the AR(1) prewhitened moment series, $n = 4(T/100)^{2/9}$, $w_t = \sum_{q=1}^{Q} x_t$, $\hat\sigma_j = (T-1)^{-1}\sum_{t=j+2}^{T} w_t w_{t-j}$, $j = 0, \ldots, n$, $\hat s^{(1)} = 2\sum_{j=1}^{n} j\hat\sigma_j$, $\hat s^{(0)} = \hat\sigma_0 + 2\sum_{j=1}^{n} \hat\sigma_j$, and $\hat\gamma = 1.1447(\hat s^{(1)}/\hat s^{(0)})^{2/3}$, then the lag-selection parameter is chosen as $L_T = \hat\gamma T^{1/3}$.

We implemented the procedure both with and without prewhitening. The results are provided in Table 7. The most striking aspect of this selection scheme is the long lag length they choose and the fact that the lag lengths barely diminish for the prewhitened series. Given our prior findings, we may expect the long but variable lag length to improve estimation performance for the large sample sizes. This is what happens. In fact, for $T = 4{,}000$, and in particular for $T = 10{,}000$, this method produces close to the best RMSE of any method. It reflects the fact that from $T = 4{,}000$ to $T = 10{,}000$ the RMSE continues to drop at a rate faster than root T, which in part is due to rapidly shrinking biases in the parameter estimates. In some sense the results for $T = 10{,}000$ are about as good as we may hope for because they are only slightly worse than those obtained for the "true" weighting matrix in Table 3, indicating that the imprecision in the estimate of the weighting matrix may no longer be much of a concern for estimation. Finally, notice that the choice of prewhitening appears to be of second-order importance, especially when the sample size is large. Indeed, for $T = 10{,}000$ the procedure without prewhitening provides marginally better inference.

We conclude that, although this procedure is not particularly attractive for the smaller sample sizes, it is our preferred method among the ones investigated when the sample size reaches 4,000. It dominates all prior methods by the RMSE criterion for $T = 10{,}000$, and given the results obtained with the "true" weighting matrix we do not expect any alternative procedure to offer much additional improvement for samples of this size (given the choice of moments).

3.9 Diagonal Weighting Matrix

In some cases researchers find it necessary to use a high number of moments to match different aspects of their

Table 7. Simulated Mean and Root Mean Squared Error: Bartlett Kernel, Newey–West Bandwidth, $(\omega, \beta, \sigma_u) = (-.736, .900, .363)$

# moments	9 (non-pw)	14 (non-pw)	14 (pw)	24 (pw)
$T = 1{,}000$				
$\hat{\omega}$	−.888 (.646)	−.855 (.637)	−.901 (.669)	−.931 (.646)
$\hat{\beta}$	.881 (.087)	.886 (.083)	.879 (.089)	.876 (.085)
$\hat{\sigma}_u$	.309 (.127)	.290 (.131)	.300 (.131)	.291 (.130)
No convergence	70	39	53	41
Average lag (std. dev.)	15.24 (3.28)	16.92 (3.00)	16.10 (3.25)	18.38 (3.57)
$T = 2{,}000$				
$\hat{\omega}$	−.824 (.393)	−.847 (.379)	−.831 (.351)	−.876 (.398)
$\hat{\beta}$	.889 (.053)	.886 (.051)	.888 (.047)	.883 (.052)
$\hat{\sigma}_u$	.327 (.093)	.321 (.085)	.319 (.085)	.313 (.089)
No convergence	3	1	2	0
Average lag (std. dev.)	20.85 (4.07)	23.95 (3.57)	23.17 (3.86)	25.72 (3.00)
$T = 4{,}000$				
$\hat{\omega}$	−.803 (.243)	−.821 (.222)	−.835 (.219)	−.891 (.269)
$\hat{\beta}$	.891 (.033)	.889 (.030)	.887 (.029)	.880 (.035)
$\hat{\sigma}_u$	.343 (.059)	.337 (.053)	.342 (.051)	.342 (.052)
No convergence	0	0	0	0
Average lag (std. dev.)	29.31 (5.49)	35.14 (4.84)	34.75 (4.73)	37.73 (3.92)
$T = 10{,}000$				
$\hat{\omega}$	−.775 (.138)	−.795 (.122)	−.791 (.126)	−.836 (.160)
$\hat{\beta}$	.895 (.019)	.892 (.016)	.893 (.017)	.887 (.021)
$\hat{\sigma}_u$	.353 (.034)	.353 (.028)	.352 (.030)	.352 (.031)
No convergence	0	0	0	0
Average lag (std. dev.)	43.51 (7.51)	51.61 (6.24)	50.76 (6.75)	56.06 (5.39)

NOTE: The reported statistics are based on 1,000 simulated samples of size equal to the indicated T. For each cell, the first number shows the mean and the second the root mean squared error (in parentheses). (pw): The prewhitening technique is applied using univariate AR(1) approximations to each sample moment.

model (e.g., Ho et al. 1996). In these cases it is tempting to avoid the documented estimation problems associated with the asymptotically optimal GMM procedure by restricting the weighting matrix to be diagonal. We examine how the results for our model are affected by this choice. The estimations were all performed with the weighting matrix set equal to the diagonal of the prewhitened Bartlett kernel that seems to perform reasonably well for the model.

The results, presented in Table 8, are interesting. It is clear that using a low number of moments ($M = 9$) and a diagonal weight matrix is inferior to our prior procedures. There seems, however, to be a trade-off between the simpler weighting matrix and the number of moments included. A surprising finding is that for T in excess of 1,000 it seems as good (judged by RMSE) to use the diagonal weighting matrix as to use the standard Bartlett kernel with prewhitening or, for that matter, most other methods we have investigated. The one exception is the Newey–West selection of bandwidth in Table 7, and even here the evidence is not unanimously in favor of the alternative. One key to the improved RMSE is that the bias in σ_u has been all but eliminated for the larger samples. Furthermore, note that for $T = 10{,}000$ it is preferable to exploit all 24 moments rather than just 14. Thus, it seems that it may be useful to exploit additional information as long as some of the estimation problems associated with the weighting matrix are appropriately handled or circumvented. An additional benefit of the approach is that the estimations tend to crash a lot less for low values of T, but this seems to be caused by the downward bias in the estimate of β. The latter observation probably constitutes the largest drawback of the method: The smaller bias in σ_u comes at the expense of a significant downward bias in the important autoregressive volatility parameter β for the smaller samples. Moreover, associated inference and specification test procedures are now less convenient because a consistent estimate of the optimal weighting matrix is not obtained as a by-product of the estimation. Nonetheless, the advantages of this rather simple procedure appear enticing, and this type of approach may provide a fruitful starting point for further progress on the development of well-functioning finite-sample GMM procedures in this context.

3.10 Higher Volatility Persistence

Empirical studies of stochastic volatility models often obtain parameter estimates of β near unity. In Table 9 we examine a few experiments with $\beta = .95$ and $\beta = .98$. The pattern is qualitatively similar to what we found earlier, so we only report a subset of our results, relying exclusively on $M = 14$, which seems reasonable in most cases. Notice that the signal-to-noise ratio for the volatility process has improved, so not unexpectedly we obtain lower RMSE for the larger samples. This is consistent with the observations of JPR and Harvey and Shephard (1993). Moreover, not surprisingly, the problem with nonconverging estimates has grown as we push β closer to the bound of the parameter space, although the use of automatic bandwidth appears to alleviate the problem somewhat. In fact, for $T = 4{,}000$ and $M = 14$, it no longer appears to constitute a practical

Table 8. Simulated Mean and Root Mean Squared Error: Diagonal Bartlett Matrix, Automatic Bandwidth and Prewhitening; (ω, β, σ_u) = (−.736, .900, .363)

# moments	9	14	24
$T = 500$			
$\hat{\omega}$	−1.342 (1.379)	−1.364 (1.274)	−1.270 (1.088)
$\hat{\beta}$	.818 (.185)	.814 (.172)	.827 (.143)
$\hat{\sigma}_u$	.393 (.164)	.387 (.150)	.377 (.134)
No convergence	89	16	12
Average lag (std. dev.)	.85 (1.35)	1.15 (1.85)	1.56 (1.92)
$T = 1{,}000$			
$\hat{\omega}$	−.980 (.746)	−1.014 (.661)	−1.037 (.656)
$\hat{\beta}$	.866 (.102)	.862 (.090)	.859 (.090)
$\hat{\sigma}_u$	.363 (.126)	.373 (.109)	.375 (.106)
No convergene	29	0	1
Average lag (std. dev.)	1.11 (1.39)	1.47 (2.17)	1.82 (2.00)
$T = 2{,}000$			
$\hat{\omega}$	−.855 (.420)	−.886 (.373)	−.872 (.357)
$\hat{\beta}$	.884 (.057)	.879 (.051)	.881 (.049)
$\hat{\sigma}_u$	.362 (.093)	.372 (.075)	.365 (.074)
No convergence	4	0	0
Average lag (std. dev.)	1.37 (1.54)	1.65 (1.88)	2.17 (2.42)
$T = 4{,}000$			
$\hat{\omega}$	−.801 (.273)	−.803 (.219)	−.802 (.208)
$\hat{\beta}$	.891 (.037)	.891 (.030)	.891 (.028)
$\hat{\sigma}_u$	.363 (.067)	.364 (.049)	.362 (.050)
No convergence	0	0	0
Average lag (std, dev.)	1.57 (1.40)	2.01 (2.17)	2.51 (2.35)
$T = 10{,}000$			
$\hat{\omega}$	−.767 (.166)	−.769 (.125)	−.765 (.117)
$\hat{\beta}$	.896 (.023)	.895 (.017)	.896 (.016)
$\hat{\sigma}_u$	.363 (.042)	.364 (.032)	.363 (.030)
No convergence	0	0	0
Average lag (std. dev.)	2.07 (1.64)	2.64 (2.42)	3.25 (2.56)

NOTE: The reported statistics are based on 1,000 simulated samples of size equal to the indicated T. For each cell, the first number shows the mean and the second the root mean squared error (in parentheses). The prewhitening technique is applied using univariate AR(1) approximations to each sample moment.

problem. For $\beta = .98$, we find the trend continuing: The RMSE's are now dramatically reduced, but the problem of crashes is prevalent, even for large samples. In this setting some strategy of forcing estimates at the bounds of the parameter space may be required for practical implementation of GMM estimation.

4. THE SIZE OF THE χ^2 TEST FOR GOODNESS OF FIT

Our simulation setting is ideal for an investigation of the standard χ^2 test for goodness of fit of the overidentifying restrictions. For each of the simulations that produce a convergent set of parameter estimates in an overidentified system ($M > 3$), we calculate the χ^2 test statistic and evaluate the associated P value in the appropriate $\chi^2(q)$ distribution with $q = M - 3$. The findings are qualitatively similar across our alternative procedures so, for the sake of brevity, we focus on the relatively successful method based on the Bartlett kernel and an automatic choice of bandwidth.

Figures 4–6 (pp. 346–348) display the fraction of P values that fall within the indicated 5% fractiles for different sample sizes. Asymptotically, the P values are, of course, uniformly distributed over the fractiles. The question is how well the finite-sample $\chi^2(q)$ statistics conform to their asymptotic distribution. In particular, the 0–5% and 5–10% fractiles shed light on the size of these goodness-of-fit tests at the (asymptotic) 5% and 10% level.

The figures are revealing. There are systematic patterns in the small-sample distribution for the P values both across sample sizes and across the number of moments included in the estimation. For each sample size, increasing the number of moments leads to a very significant rightward shift in the entire distribution. Similarly, for a given choice of moments, an increase in sample size leads to a very significant leftward shift in the entire distribution. Moreover, there is no sense in which the distribution appears to approach its asymptotic counterpart as the sample size grows. Indeed, the leftward shift in the distribution appears to continue, suggesting that the size distortion of the χ^2 test statistic is growing increasingly severe as the sample becomes very large. The same phenomenon is observed for all the other designs. Table 10 (p. 349) provides the extreme 5% and 10% fractiles for a representative set of procedures using 14 moments. In all instances the mass located in the 0–5% fractile increases dramatically with sample size. In the process we move from a scenario in which the test statistics are severely downward biased—the maximum frequency observed for the 0–5% fractile with $T = 1{,}000$ is .022 for the quadratic bandwidth kernels—to one in which they are badly inflated;

Table 9. Simulated Mean and Root Mean Squared Error: Bartlett Kernel, Automatic Bandwidth, 14 Moments, $(\omega, \beta, \sigma_u) = (-.368, .950, .260)$ or $(-.147, .980, .166)$

# moments	$\beta = 95$ (non-pw)	$\beta = .95$ (pw)	$\beta = .98$ (pw)
$T = 2{,}000$			
$\hat{\omega}$	−.363 (.215)	−.286 (.197)	−.140 (.112)
$\hat{\beta}$	.951 (.029)	.961 (.027)	.981 (.015)
$\hat{\sigma}_u$	.208 (.087)	.190 (.099)	.125 (.068)
No convergence	62	131	810
Average lag (std. dev.)	12.25 (10.50)	1.83 (1.97)	1.72 (2.20)
$T = 4{,}000$			
$\hat{\omega}$	−.374 (.148)	−.294 (.152)	−.121 (.089)
$\hat{\beta}$	.949 (.020)	.960 (.021)	.984 (.012)
$\hat{\sigma}_u$	.228 (.059)	.206 (.076)	.126 (.063)
No convergence	10	22	399
Average lag (std. dev.)	14.03 (10.65)	2.45 (2.39)	2.33 (2.00)
$T = 10{,}000$			
$\hat{\omega}$	−.389 (.091)	−.336 (.094)	−.121 (.066)
$\hat{\beta}$	.947 (.012)	.954 (.013)	.983 (.009)
$\hat{\sigma}_u$	.247 (.033)	.234 (.043)	.137 (.048)
No convergence	0	0	49
Average lag (std. dev.)	17.81 (10.99)	3.09 (2.39)	3.42 (2.58)

NOTE: The reported statistics are based on 1,000 simulated samples of size equal to the indicated T. For each cell, the first number shows the mean and the second the root mean squared error (in parentheses). (pw): The prewhitening technique is applied using univariate AR(1) approximations to each sample moment.

that is, the minimum frequency for the 0–5% fractile with $T = 10{,}000$ is .121 for the Bartlett kernel with prewhitening and automatic bandwidth choice. Similarly, the mass located in the right tail decreases almost uniformly as the sample size expands. This confirms the robustness of the systematic leftward shift in the P-value distribution that is captured in the figures.

On the other hand, notice that the guidelines for selection of the number of moments to include in estimation, which were developed in Section 3 on the basis of estimation performance, generally also lead to reasonably sized specification tests. For the lower sample sizes, $M = 9$ clearly produces the most appropriately sized tests, but as the sample size grows, the required number of moments expands as well; for example, $M = 14$ appears appropriate for $T = 2{,}000$ and $M = 24$ seems preferable for $T = 4{,}000$. Two caveats are in order. First, for the smaller samples the results may be somewhat misleading because they fail to account for the discarded simulations that are numerous. Although it may seem appropriate to interpret a nonconverging sample as evidence of a poor fit, recall that we frequently found that the χ^2 statistic was consistent with an acceptable goodness of fit prior to the termination of nonconvergent iterations. Thus, the direction of the potential bias is indeterminate. Second, to obtain the optimal test size, it appears that we should expand the number of moments somewhat more aggressively than our evaluation of estimation performance in Section 3 indicated. Nonetheless, both considerations imply that we should let the number of moments grow quite rapidly with the sample size.

The increased size distortions associated with the larger samples may appear puzzling because they defy predictions based on asymptotic theory. The explanation is again related to the imprecision of the estimated weighting matrices and the extremely high degree of variability and dependency in the sample moments.

First, even for sample sizes as large as $T = 10{,}000$, the bias and dispersion of the weighting matrices are profound. This was demonstrated through an in-depth analysis of the design with $M = 5$. We calculated the average weighting matrix from 1,000 simulations using three different kernel estimation procedures—namely, the fixed bandwidth Bartlett kernel with lag length $\gamma T^{1/3}$ for γ equal to 2 and 10 and for the QS kernel with lag length $2 * T^{1/5}$. They all provide similar results. The estimate of the dominant entry on the diagonal of the weighting matrix, corresponding to the absolute return moment, element (1, 1), varies from 1.93 to 2.02 with standard errors between .42 and .50 (the other diagonal elements of the weighting matrix display variation similar to the first, but we focus on one element for brevity). Thus, the estimates of the weighting matrix fluctuate very substantially across the $T = 10{,}000$ samples, but, perhaps even more significantly, they are strongly biased. The latter was confirmed through the construction of a more precise approximation to the "true" weighting matrix based on 16 samples of 500,000 observations and the Monte Carlo variance-reduction technique of antithetic variables. Thus, the series consist of eight "antithetic" samples that pairwise have negatively correlated volatility processes due to the use of the identical draws for the underlying innovations but with a sign change for the volatility innovation. The negative correlation reduces the sample variability of the estimated sample moments (e.g., see Davidson and MacKinnon 1993). The resulting estimate of element (1, 1) of the weighting matrix is 1.25. Consequently, this element displays a very strong upward small-sample bias (for $T = 10{,}000$). This bias will tend to inflate the test statistics and push the P-value distribution leftward.

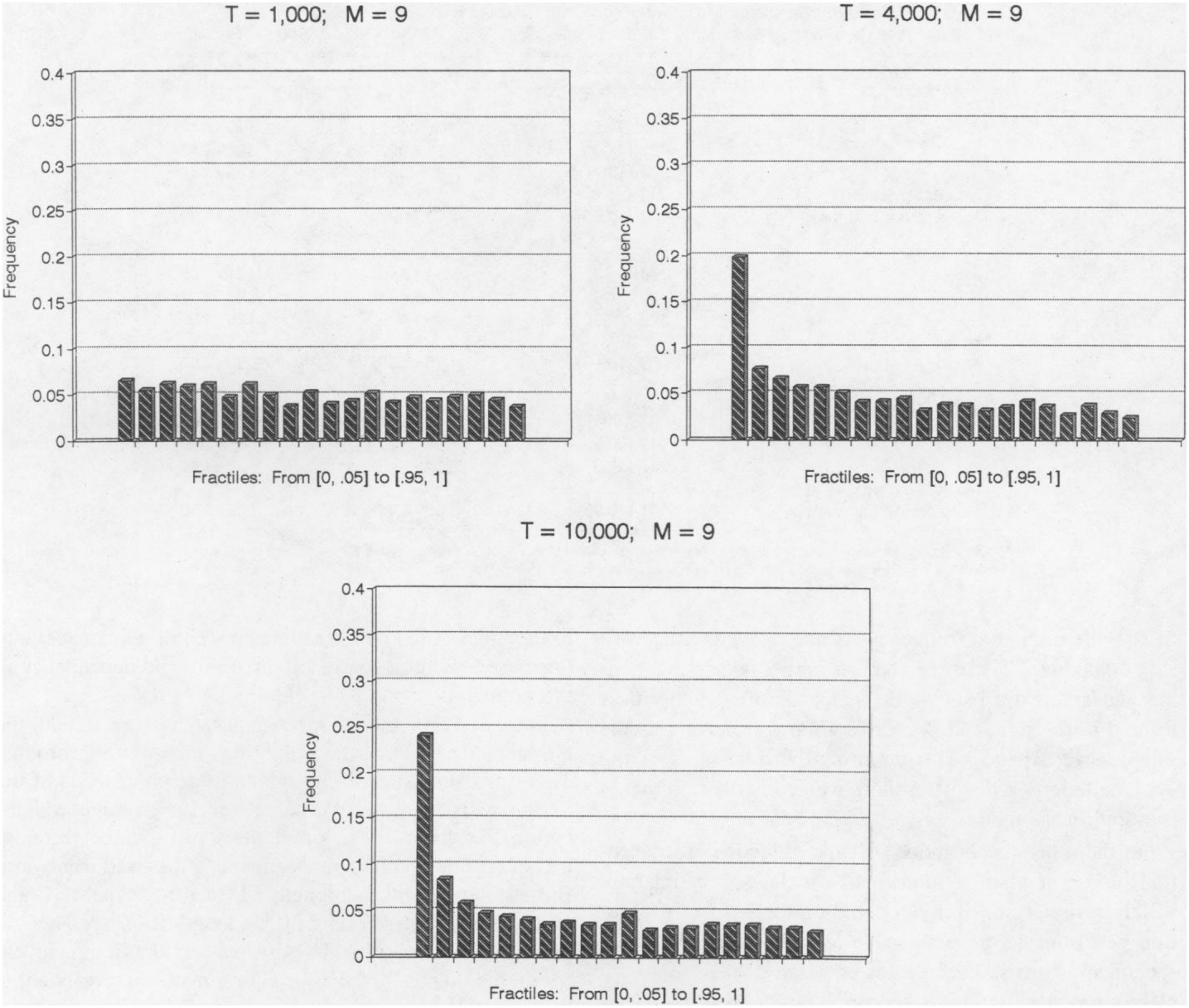


Figure 4. The Distribution of p Values for the Test of Overidentifying Restrictions Based on the GMM Objective Function. The figure shows the fraction of values that fall within the different 5% fractiles. The results are based on 1,000 converging estimates for each sample size. The GMM procedure is implemented using an estimated weighting matrix determined by the Bartlett kernel with an automatic choice of bandwidth. Figures are displayed for sample sizes T = 1,000; 4,000; 10,000. All estimates are obtained with M = 9 moments.

Second, samples of $T = 10{,}000$ remain small in yet another sense. Using the preceding precise estimate of the "true" 5×5 weighting matrix, we find that the P-value distribution is biased to the right. For sample sizes of $T = 50{,}000$, the test statistic finally seems to obey an approximate χ^2 distribution when the "true" weighting matrix is used.

It is important to realize that these rather discouraging findings regarding the finite-sample distribution of the estimated weighting matrices have no direct implications for the performance of the GMM estimation and inference procedures. If the results predominantly reflect a problem in determining the scale of the weighting matrix—which clearly is strongly upward biased, even for very large samples—then the estimation procedure may be relatively immune to this deficiency of the GMM criterion function. In fact, the finite-sample estimation performance reported in Section 3 is quite satisfactory, at least for the larger samples. Section 5 reports on the finite-sample performance for asymptotically motivated inference procedures regarding the model parameters.

In summary, the investigation in this section tends to reinforce our earlier conclusions. The GMM procedure is not well equipped to deal with small samples, and it is essential to increase the included number of moments rather sharply with sample size to avoid serious size distortions for the test of goodness of fit. The extent of the problem is striking. For 500 observations and 24 moments, the P value of the test statistic (given the GMM estimates converge) will exceed 80% seven times out of ten. Without a size correction, the power of the test is therefore likely to be extremely poor. Equally troublesome is the tendency to overreject when an insufficient number of moments is included: For $T = 10{,}000$ and $M = 9$, the test will reject at the 5% level about one

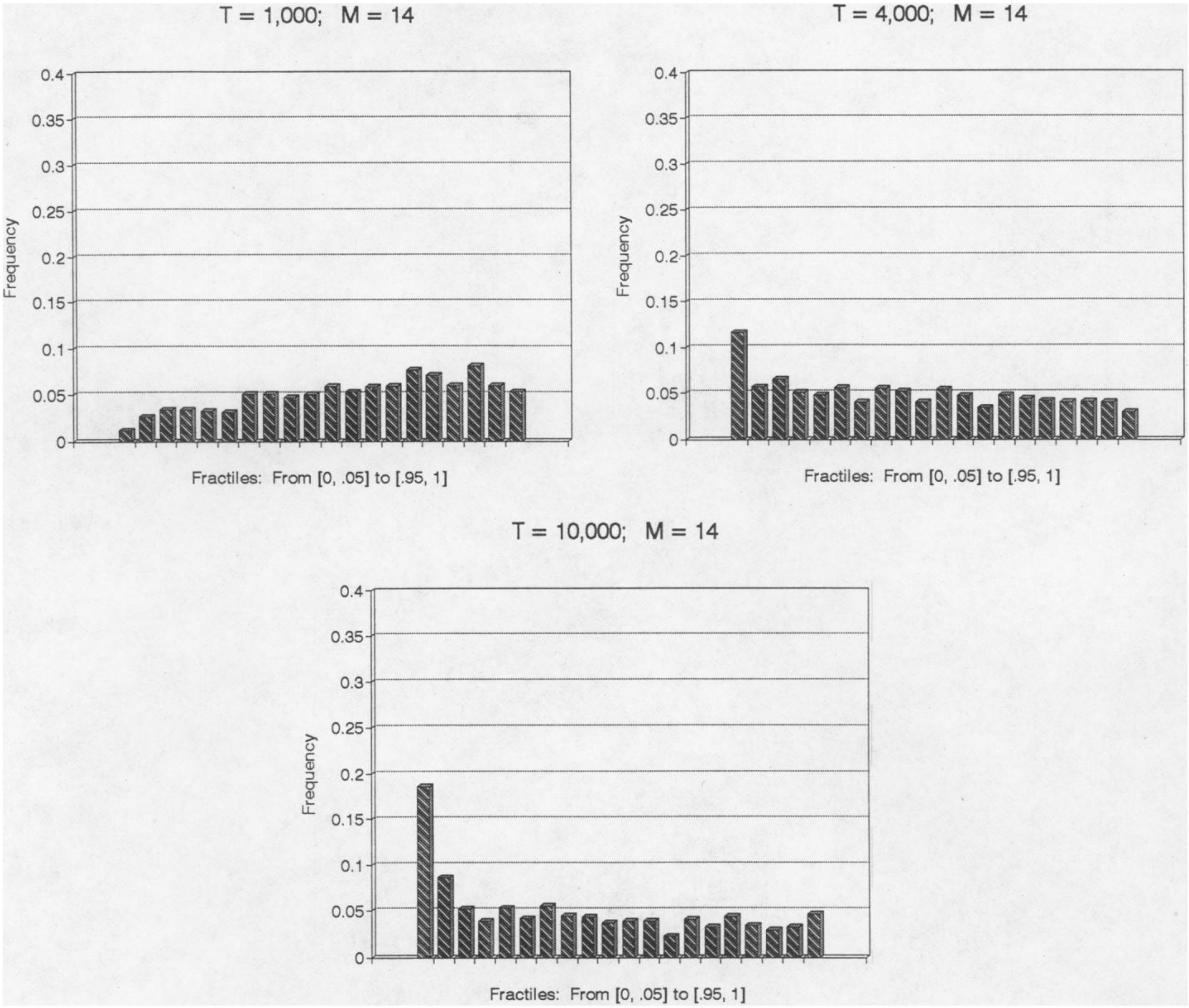


Figure 5. The Distribution of p values for the Test of Overidentifying Restrictions Based on the GMM Objective Function. The figure shows the fraction of values that fall within the different 5% fractiles. The results are based on 1,000 converging estimates for each sample size. The GMM procedure is implemented using an estimated weighting matrix determined by the Bartlett kernel with an automatic choice of bandwidth. Figures are displayed for sample sizes T = 1,000; 4,000; 10,000. All estimates are obtained with M = 14 moments.

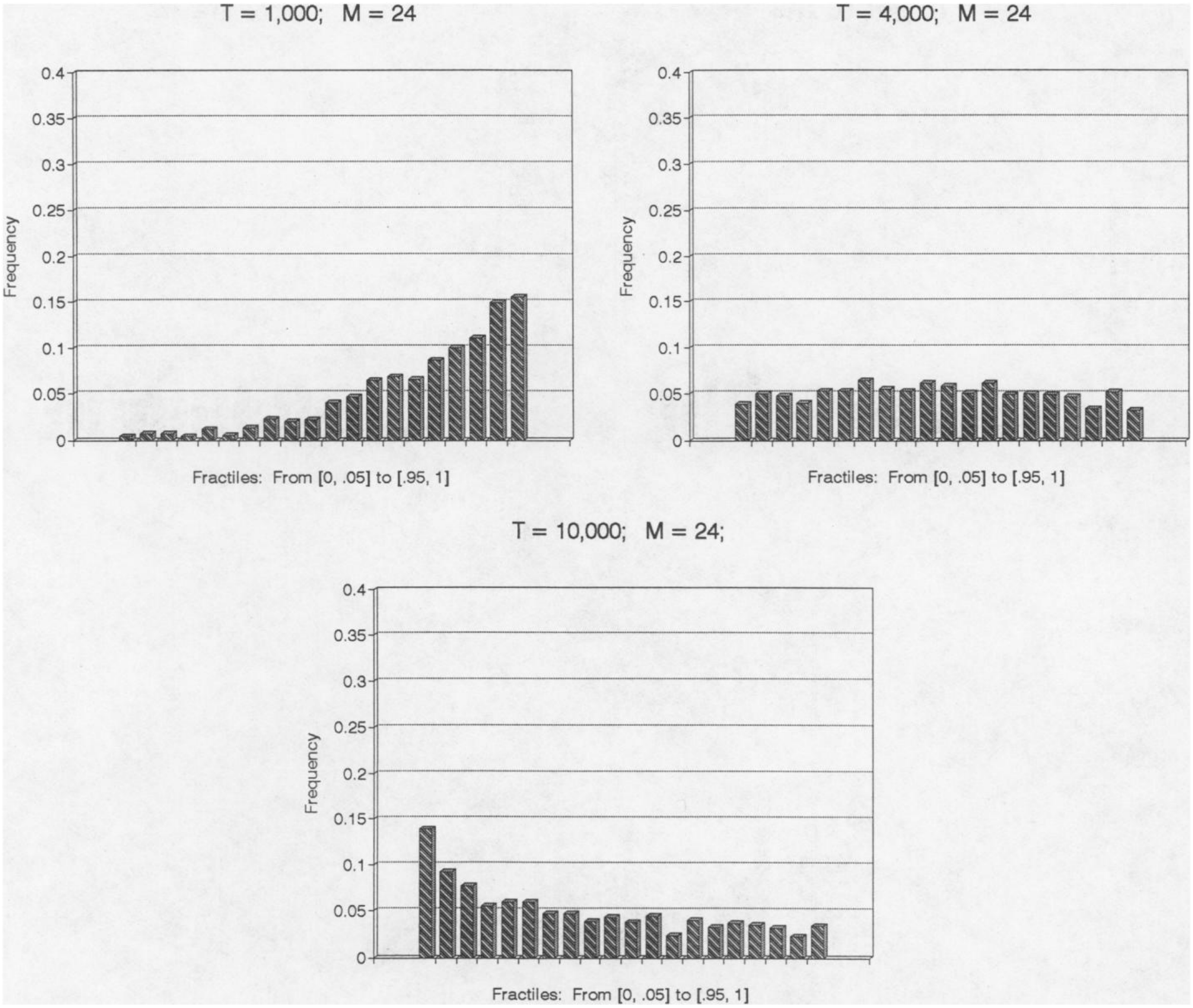
quarter of the time. These findings underscore the importance of paying careful attention to the trade-offs between information and precision involved in the choice of moments for the GMM procedure.

5. HYPOTHESIS TESTS

This section takes a look at some popular inference procedures regarding the parameters of the model. Again, the conclusions are qualitatively similar across the designs, and we present results only for the Bartlett kernel with fixed-bandwidth procedure that corresponds to Table 2. The extreme left and right fractiles for the distribution of the studentized parameters are provided in Table 11 (p. 350). The top panel concerns the difference between the estimated parameters and the true parameters normalized by the estimated standard error. This panel thus reflects both the dispersion and the bias of the normalized parameter estimates. The bottom panel displays the fractiles for the identical studentized parameters, except that the estimated parameters now are centered on the (biased) mean estimate.

The top panel is relevant for assessment of the standard t tests for individual parameters. Asymptotically, the studentized parameters are distributed as standard normals, so the mass located in the tail fractiles approximates the size of one-sided tests for equality of the estimated parameters and their true value.

A few general observations regarding our estimation results are important for the interpretation of Table 11. First, basically all of our GMM estimation procedures, including the one used for Table 2, result in a downward finite-sample bias in all three parameter estimates. Second, we found a large negative correlation between the estimated parameters and the associated standard errors. Finally, the estimates of


T = 1,000; M = 24

T = 4,000; M = 24

T = 10,000; M = 24;

Figure 6. The Distribution of p Values for the Test of Overidentifying Restrictions Based on the GMM Objective Function. The figure shows the fraction of values that fall within the different 5% fractiles. The results are based on 1,000 converging estimates for each sample size. The GMM procedure is implemented using an estimated weighting matrix determined by the Bartlett kernel with an automatic choice of bandwidth. Figures are displayed for sample sizes T = 1,000; 4,000; 10,000. All estimates are obtained with M = 24 moments.

ω and β were extremely highly correlated. This last fact immediately explains the near identical studentized distributions of those two parameters. For the smaller samples, the substantial downward biases are negated by the associated large standard errors. Indeed, for the first two parameters the left tails in the top panel are too thin rather than too thick, but for σ_u the left tail is mostly fat-tailed. This reflects the fact that σ_u has a particularly large bias relative to RMSE compared to the other parameters in Table 2. This also explains the relatively thin right tails in the distribution of this parameter estimate. Notice that for the right tails the downward bias is mitigated by the negative correlation between the estimated parameters and standard errors. This is also clear from the bottom panel where asymmetric distribution of the standard error estimates induces a rightward shift in the studentized distribution. As the sample size grows the bottom panel further shows—quite clearly in the case of $M = 24$—that the (mean corrected) studentized distribution approaches the standard normal.

This suggests that the distribution for the studentized parameters in the top panel will be highly sensitive to biases in the parameter estimates for the larger sample sizes. This is exactly what happens. The distributions acquire heavy left tails, in particular for the designs involving the higher number of moments. This reflects the more significant downward biases for the designs relying on the higher number of moments.

In summary, the evidence on the quality of inference is mixed. For T in excess of 1,000 and a number of moments that is consistent with our prior recommendations, the size distortions are not bad, but there is a clear tendency to underestimate σ_u throughout and a tendency to underestimate the other parameters as well when many moments are used

Table 10. P Values for Selected Models (14 moments): (ω, β, σ_u) = (−.736, .900, .363)

		T			
Model	*Fractile*	*1,000*	*2,000*	*4,000*	*10,000*
Bartlett	0–5%	.005	.042	.105	.156
Bandwidth	5–10%	.023	.059	.067	.084
$1.2{*}10^{1/3}$	90–100%	.109	.093	.076	.049
Bartlett	0–5%	.012	.054	.116	.186
Automatic	5–10%	.027	.050	.057	.086
Bandwidth	90–100%	.113	.093	.070	.079
Bartlett	0–5%	.012	.042	.079	.121
Automatic	5–10%	.010	.037	.028	.021
Bandwidth w. pw	90–100%	.313	.285	.287	.276
Quad spectral	0–5%	.022	.078	.150	.187
Automatic	5–10%	.053	.062	.090	.076
Bandwidth	90–100%	.110	.073	.067	.072
Quad spectral	0–5%	.022	.051	.097	.149
Automatic	5–10%	.015	.040	.048	.049
Bandwidth w. pw	90–100%	.281	.254	.244	.248
Bartlett	0–5%	.006	.024	.079	.144
Newey–West	5–10%	.018	.036	.056	.096
Bandwidth	90–100%	.101	.066	.038	.047
Bartlett	0–5%	.004	.015	.074	.162
Newey–West	5–10%	.006	.037	.063	.083
Bandwidth w. pw	90–100%	.156	.094	.047	.053

NOTE: The reported statistics are based on 1,000 simulated samples of size equal to the indicated *T*. pw: prewhitening applied using univariate AR(1) approximations to each sample moment.

for estimation. The latter is somewhat troublesome because size considerations for the χ^2 test in Section 4 favor the use of many moments for the large samples.

6. CONCLUSION

This article examines the properties of alternative GMM procedures for estimation of the so-called lognormal stochastic autoregressive volatility model. The results are numerous: First, it is generally not optimal to include many moments in the estimation procedure if the sample size is limited. In fact, the preferred number of moments (as measured by RMSE) is typically lower than the standard choice in the literature concerned with estimation on the basis of high-frequency financial data. On the other hand, it is virtually never advisable to rely on the alternative extreme of a just-identified model that underperformed relative to all other models investigated. We document that these results arise because of a fundamental trade-off between the information (number of moments) used in estimation and the quality of the objective function (precision of the estimated weighting matrix) underlying the procedure. Estimation on the basis of a large-sample approximation to the optimal weighting matrix confirms this intuition and provides further insights into the feasible efficiency bounds for this class of GMM estimators. The results suggest that the inclusion of the full 24 moments provides very little additional information regarding the parameters relative to what is contained in the initial 14 moments. Hence, the incorporation of 24 moments is not likely to be beneficial unless the sample is very large.

Second, we find that estimation using a fixed number of lags in the weighting matrix generally is inferior to using the plug-in estimator of lag length suggested by Andrews (1991), although it seems that experimentation with longer lags than indicated by this data-dependent procedure may prove useful.

Third, we find that the prewhitening method for the weighting matrix suggested by Andrews and Monahan (1992) can be helpful in several settings. In particular, the RMSE on the parameters can be substantially reduced via prewhitening when the sample size is relatively small.

Fourth, we find that the QS estimator suggested by Andrews (1991) appears to fare slightly worse than the standard Bartlett kernel estimator for this model.

Fifth, we find that the automatic bandwidth choice proposed by Newey and West (1994) is appropriate for large samples in which the GMM Bartlett-kernel procedure combined with this automatic bandwidth choice provides inference of a quality that other practical methods, arguably, will be hard pressed to improve on.

Sixth, we find indications that a diagonal weighting matrix may be an excellent alternative when many moments are required for estimation.

Seventh, there is some evidence that the choice of less volatile and lower-order moments dominate the choice of more volatile and higher-order moments.

Eight, the popular χ^2 statistic for goodness of fit of the overidentifying restrictions appears fairly well behaved when the general prescriptions regarding choice of mo-

Table 11. Distribution of Studentized Parameter Estimates

	Parameter											
	ω				β				σ_u			
Fractile	0–5	5–10	90–95	95–100	0–5	5–10	90–95	95–100	0–5	5–10	90–95	95–100
$T = 1{,}000$												
$M = 9$	2.1	2.8	1.6	.7	2.0	2.9	1.7	.7	.2	.3	2.1	2.1
$M = 14$	2.2	2.2	4.8	4.5	2.1	2.3	4.9	4.7	2.4	9.2	1.5	2.2
$M = 24$	3.3	3.9	4.6	4.6	3.3	3.8	4.5	5.1	8.3	12.4	1.6	2.5
$T = 2{,}000$												
$M = 9$	2.3	3.2	2.9	1.1	2.2	3.2	2.9	1.4	.0	4.4	1.7	1.3
$M = 14$	2.3	4.9	3.3	2.5	2.3	4.8	3.4	2.7	6.2	1.7	1.7	1.1
$M = 24$	4.0	5.8	3.2	3.5	3.9	5.6	3.2	4.0	11.0	8.9	1.8	.7
$T = 4{,}000$												
$M = 9$	3.6	4.8	2.0	.8	3.5	4.6	1.9	.9	2.1	6.6	2.9	1.3
$M = 14$	3.1	4.2	1.9	1.9	3.1	4.0	1.9	1.9	6.9	11.1	2.1	.8
$M = 24$	9.2	11.1	1.5	1.3	8.6	11.1	1.5	1.3	6.5	7.6	2.2	2.8
$T = 10{,}000$												
$M = 9$	3.8	6.0	2.7	1.6	3.4	6.1	3.1	1.5	4.4	6.5	3.0	1.3
$M = 14$	4.8	6.7	1.2	.6	4.6	6.5	1.1	.7	6.0	6.7	2.3	1.1
$M = 24$	13.8	11.7	.6	.6	12.9	11.7	.6	.8	6.5	6.3	3.0	1.7
					Mean corrected							
$T = 1{,}000$												
$M = 9$	1.6	2.8	2.4	1.2	1.6	2.7	2.3	1.2	.0	.1	3.1	5.4
$M = 14$	1.8	1.8	5.2	7.3	1.8	1.8	4.7	7.1	.1	.5	3.7	5.3
$M = 24$	2.1	2.2	6.8	13.5	2.1	2.3	6.3	13.8	.6	3.5	3.4	6.2
$T = 2{,}000$												
$M = 9$	1.8	2.5	3.8	2.1	1.7	2.7	3.3	2.2	.0	.1	3.0	4.3
$M = 14$	1.7	1.8	5.4	5.0	1.7	2.0	5.5	4.9	1.6	2.3	4.5	4.5
$M = 24$	1.5	2.8	4.8	8.6	1.6	2.6	4.6	8.6	2.4	5.6	4.5	4.1
$T = 4{,}000$												
$M = 9$	1.5	3.6	4.0	2.4	1.4	3.5	3.6	2.4	.3	2.8	4.5	3.9
$M = 14$	2.0	2.7	3.6	3.7	2.1	2.3	3.5	3.7	2.3	3.1	4.4	4.1
$M = 24$	2.0	3.5	4.9	4.7	2.0	1.8	4.9	4.6	3.5	3.2	3.8	4.5
$T = 10{,}000$												
$M = 9$	2.2	3.4	4.8	2.6	2.3	3.3	4.5	2.6	2.8	3.4	4.8	3.6
$M = 14$	1.9	3.0	4.5	2.1	2.0	2.9	4.3	2.0	2.3	4.4	3.4	3.1
$M = 24$	3.4	4.0	6.0	4.3	3.2	4.1	5.8	4.4	4.5	3.8	5.0	3.5

NOTE: Top panel: (parameter—true parameter)/(estimated standard deviation). Bottom panel: (parameter—mean parameter)/(estimated standard deviation). Based on same simulations as Table 2.

ments relative to sample size are obeyed. If too few moments are included, there is a strong tendency for overrejections, and, even more importantly, when too many moments are included, the P values associated with the test statistics are seriously inflated, and the test underrejects. It is, moreover, evident that the satisfactory performance of the test in certain parts of the design matrix is somewhat coincidental. Even for very large samples, the estimates of the elements along the diagonal of the optimal weighting matrix display a very substantial upward bias. In these circumstances, size corrections may generally be necessary to obtain meaningful specification tests and reasonable power properties.

Strictly speaking, the findings are specific to the particular model being studied. The conclusions, nonetheless, are likely to apply to a wide range of economic systems characterized by strongly conditionally heteroscedastic series and highly correlated moment conditions.

Several issues remain of interest in this context. How do we further improve the finite-sample properties of the GMM procedure, especially when sample size is small? Are the conclusions robust across different volatility specifications? What are the efficiency bounds for alternative parameter constellations? What are the power properties of the specification test against some relevant alternatives?

ACKNOWLEDGMENTS

Thanks are due to Tim Bollerslev, Robin Lumsdaine, Angelo Melino, Adrian Pagan, Mark Watson, the editors Larry Christiano and George Tauchen, three anonymous referees, and seminar participants at Boston College, the University of Aarhus, the CIRANO Stochastic Volatility Conference in Montreal, the Midwest Econometrics Group Meeting in Iowa City, and the Econometric Society Meetings, Washington, D.C., for helpful comments, Mike Sherman for research assistance, and colleagues at Brown University for sharing their computers with us.

APPENDIX: CHOICE OF MOMENTS

Our moments are chosen from among the following 34 moments, denoted $m1$–$m24$:

$$\begin{aligned} m1 &= E[|y_t|] \\ m2 &= E[y_t^2] \\ m3 &= E[|y_t|^3] \end{aligned}$$

$$\begin{aligned} m4 &= E[y_t^4] \\ m4+i &= E|y_t y_{t-i}|], \qquad i = 1, \ldots, 10 \\ m14+i &= E[y_t^2 y_{t-i}^2], \qquad i = 1, \ldots, 10 \\ m24+i &= E[|y_t| y_{t-i}^2], \qquad i = 1, \ldots, 10. \end{aligned}$$

3 moments: $m1, m2, m5$

5 moments: $m1, m2, m4, m6, m15$

9 moments:

Baseline set ($m9a$): $m1$–$m4, m5, m7, m9, m16, m18$

Alternative set ($m9b$): $m1$–$m4, m6, m8, m10, m15, m17$

14 moments:

Baseline set: ($m14a$): $m1$–$m4, m6, m8, m10, m12, m14, m15, m17, m19, m21, m23$

Alternative set: ($m14b$): $m1$–$m4, m5, m7, m9, m11, m13, m16, m18, m20, m22, m24$

Absolute moments ($m14c$): $m1$–$m14$

Quadratic moments ($m14d$): $m1$–$m4, m15$–$m24$

Absolute 3rd moments ($m14e$): $m1$–$m4, m25$–$m34$

Mix of low abs 1st, 2nd and 3rd moments ($m14f$): $m1$–$m4, m5$–$m7, m15$–$m17, m25$–$m28$

Alternative mix of 1st, 2nd, 3rd moments ($m14g$): $m1$–$m4, m5, m8, m11, m14, m16, m19, m22, m27, m30, m33$

24 moments: $m1$–$m24$

34 moments: $m1$–$m34$.

[Received February 1994. Revised November 1995.]

REFERENCES

Altonji, J. G., and Segal, L. M. (1993), "Small Sample Bias in GMM Estimation of Covariance Structures," working paper, Northwestern University, Dept. of Economics.

Andersen, T. G. (1992), "Volatility," Working Paper 144, Northwestern University, J. L. Kellogg Graduate School of Management, Dept. of Finance.

——— (1994a), "Stochastic Autoregressive Volatility: A Framework for Volatility Modeling," *Mathematical Finance*, 4, 75–102.

——— (1994b), "Comment," *Journal of Business & Economic Statistics*, 12, 389–392.

——— (1996), "Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility," *Journal of Finance*, 51, 169–204.

Andersen, T. G., and Sørensen, B. E. (1995), "GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study," Working Paper 175, Northwestern University, J. L. Kellogg Graduate School of Management, Dept. of Finance.

——— (1996), "GMM and QML Asymptotic Standard Deviations in Stochastic Volatility Models: A Response to Ruiz (1994)," unpublished manuscript submitted to *Journal of Econometrics*.

Andrews, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Andrews, D. W. K., and Monahan, J. C. (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–967.

Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

Burnside, C., and Eichenbaum, M. (1994), "Small Sample Properties of Generalized Method of Moments Based Wald Tests," Technical Working Paper 155, National Bureau of Economic Research, Cambridge, MA.

Christiano, L. J., and den Haan, W. (1996), "Small-Sample Properties of GMM for Business-Cycle Analysis," *Journal of Business & Economic Statistics*, 14, 309–327.

Clark, P. K. (1973), "A Subordinated Stochastic Process Model With Finite Variance for Speculative Prices," *Econometrica*, 41, 135–155.

Danielsson, J. (1993), "Multivariate Stochastic Volatility," working paper, University of Iceland, Dept. of Economics.

——— (1994), "Stochastic Volatility in Asset Prices: Estimation With Simulated Maximum Likelihood," *Journal of Econometrics*, 64, 375–401.

Danielsson, J., and Richard, J.-F. (1993), "Accelerated Gaussian Importance Sampler With Application to Dynamic Latent Variable Models," *Journal of Applied Econometrics*, 8, S153–S173.

Davidson, R., and MacKinnon, J. G. (1993), *Estimation and Inference in Econometrics*, Oxford, U.K.: Oxford University Press.

Diebold, F. X., and Nerlove, M. (1989), "The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model," *Journal of Applied Econometrics*, 4, 1–21.

Duffie, D., and Singleton, K. J. (1989), "Simulated Moments Estimation of Markov Models of Asset Prices," working paper, Stanford University, Graduate School of Business.

Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation," *Econometrica*, 50, 987–1008.

Engle, R. F., Lilien, D. M., and Robins, R. P. (1987), "Estimating Time-Varying Risk Premia in the Term Structure: The ARCH-M Model," *Econometrica*, 55, 391–407.

Engle, R. F., Ng, V. K., and Rothschild, M. (1990), "Asset Pricing With a Factor-ARCH Covariance Structure," *Journal of Econometrics*, 45, 213–237.

Epps, T. W., and Epps, M. L. (1976), "The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distributions Hypothesis," *Econometrica*, 44, 305–321.

Ferson, W. E., and Foerster, S. R. (1994), "Finite Sample Properties of the Generalized Method of Moments in Tests of Conditional Asset Pricing Models," *Journal of Financial Economics*, 36, 29–55.

Foster, F. D., and Viswanathan, S. (1995), "Can Speculative Trading Explain the Volume–Volatility Relation?" *Journal of Business & Economic Statistics*, 13, 379–396.

Gallant, A. R., Hsieh, D. A., and Tauchen, G. E. (1991), "On Fitting a Recalcitrant Series: The Pound/Dollar Exchange Rate, 1974–83," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics, Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, eds. W. A. Barnett, J. Powell, and G. E. Tauchen, Cambridge, U.K.: Cambridge University Press, pp. 199–240.

Gallant, A. R., and Tauchen, G. E. (in press), "Which Moments to Match?" *Econometric Theory*, 12.

Gourieroux, C., Monfort, A., and Renault, E. (1993), "Indirect Inference," *Journal of Applied Econometrics*, 8, S85–S118.

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

Hansen, L. P., Heaton, J., and Yaron, A. (1996), "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business & Economic Statistics*, 19, 262–280.

Harvey, A. C., Ruiz, E., and Shephard, N. (1994), "Multivariate Stochastic Variance Models," *Review of Economic Studies*, 61, 247–264.

Harvey, A. C., and Shephard, N. (1993), "The Econometrics of Stochastic Volatility," working paper, London School of Economics, Dept. of Statistical and Mathematical Sciences.

Ho, M. S., Perraudin, W. R. M., and Sørensen, B. E. (1996), "A Continuous-Time Arbitrage-Pricing Model With Stochastic Volatility and Jumps," *Journal of Business & Economic Statistics*, 14, 31–43.

Hull, J., and White, A. (1987), "The Pricing of Options on Assets With Stochastic Volatilities," *Journal of Finance*, 42, 281–300.

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994), "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business & Economic Statistics*, 12, 371–389.

Johnson, H., and Shanno, D. (1987), "Option Pricing When the Variance Is Changing," *Journal of Financial and Quantitative Analysis*, 22, 143–152.

King, M., Sentana, E., and Wadhwani, S. (1994), "Volatility and Links Between National Stock Markets," *Econometrica*, 62, 901–933.

Kocherlakota, N. (1990), "On Tests of Representative Consumer Asset Pricing Models," *Journal of Monetary Economics*, 26, 285–304.

Laux, P. A., and Ng, L. K. (1993), "The Sources of GARCH: Empirical Evidence From an Intraday Returns Model Incorporating Systematic and Unique Risks," *Journal of International Money and Finance*, 12, 543–560.

Melino, A., and Turnbull, S. (1990), "Pricing Foreign Currency Options With Stochastic Volatility," *Journal of Econometrics*, 45, 239–265.

Nelson, D. B. (1990), "ARCH Models as Diffusion Approximations," *Journal of Econometrics*, 45, 7–38.

——— (1991), "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 59, 347–370.

——— (1992), "Filtering and Forecasting With Misspecified ARCH Models I: Getting the Right Variance With the Wrong Model," *Journal of Econometrics*, 52, 61–90.

Nelson, D. B., and Foster, D. P. (1991), "Filtering and Forecasting With Misspecified ARCH Models II: Making the Right Forecast With the Wrong Model," *Journal of Econometrics*, 67, 303–335.

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

——— (1994), "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies*, 61, 631–653.

Perraudin, W. R. M., and Sørensen, B. E. (1994), "Modelling Exchange Rates in Continuous Time: Theory, Estimation, and Option Pricing," Working Paper 94-25, Brown University, Dept. of Economics.

Ruiz, E. (1994), "Quasi-Maximum Likelihood Estimation of Stochastic Volatility Models," *Journal of Econometrics*, 63, 289–306.

Scott, L. O. (1987), "Option Pricing When the Variance Changes Randomly: Theory, Estimation and an Application," *Journal of Financial and Quantitative Analysis*, 22, 419–438.

Tauchen, G. E. (1986), "Statistical Properties of Generalized Method-of-Moments Estimators of Structural Parameters Obtained From Financial Market Data," *Journal of Business & Economic Statistics*, 4, 397–425.

Tauchen, G. E., and Pitts, M. (1983), "The Price Variability-Volume Relationship on Speculative Markets," *Econometrica*, 51, 485–505.

Taylor, S. J. (1986), *Modelling Financial Time Series*, Chichester, U.K.: John Wiley.

——— (1994), "Modelling Stochastic Volatility: A Review and Comparative Study," *Mathematical Finance*, 4, 183–204.

White, H. (1984), *Asymptotic Theory for Econometricians*, New York: Academic Press.

Wiggins, J. B. (1987), "Option Values Under Stochastic Volatility: Theory and Empirical Estimates," *Journal of Financial Economics*, 19, 351–372.