

# Efficient Method of Moments <sup>1</sup>

A. Ronald Gallant  
Department of Economics  
University of North Carolina  
Chapel Hill NC 27599-3305 USA

George Tauchen  
Department of Economics  
Duke University  
Durham NC 27708-0097 USA

Last revised: March 2001

<sup>1</sup>Supported by the National Science Foundation. Corresponding author is A. Ronald Gallant.  
Revisions available at [www.unc.edu/~arg](http://www.unc.edu/~arg).

# Abstract

We describe a computationally intensive methodology for the estimation and analysis of partially observable nonlinear systems. An example from epidemiology is the SEIR model, which is a system of differential equations with random coefficients that describes a population in terms of four state variables: those susceptible to a disease, those exposed to it, those infected by it, and those recovered from it. Only those infected by the disease are known to public health officials. An example from finance is the continuous-time stochastic volatility model, which is a system of stochastic differential equations that describes a security's price and instantaneous variance. Only the security's price can be observed directly.

System parameters are estimated by a variant of simulated method of moments known as efficient method of moments (EMM). The idea is to match moments implied by the system to moments implied by the transition density for observables.

System analysis is accomplished by reprojection. Reprojection is carried out by projecting a long simulation from the estimated system onto an appropriate representation of a relationship of interest. A general purpose representation is a Hermite expansion of the conditional density of state variables given observables. A reprojection density thus obtained embodies all constraints implied by the nonlinear system and is analytically convenient. As an instance, nonlinear filtering can be accomplished by computing the conditional mean of the reprojection density and evaluating it at observed values from the data.

Ideas are illustrated by using the methodology to assess the dynamics of a stochastic volatility model for daily Microsoft closing prices.

Keywords and Phrases: Efficient Method of Moments (EMM), Simulated Method of Moments (SMM), Minimum chi-squared, Estimation of stochastic differential equations, Estimation of diffusions, Partially observed, Latent variables, Semiparametric (SNP), Dynamic models, Finance, Economics, Epidemiology, Pharmacokinetics, Hermite, Stochastic volatility.

# 1 Introduction

Dynamic nonlinear models that have unobserved variables pervade science. Most often they arise from dynamic systems described by a system of deterministic or stochastic differential equations in which the state vector is partially observed. For example, in epidemiology the SEIR model determines those susceptible, exposed, infected, and recovered from a disease whereas usually data are from case reports that report only those infected (Olsen and Schaffer, 1990). Other examples are continuous and discrete time stochastic volatility models of speculative markets from finance (Ghysels, Harvey, and Renault, 1995), general equilibrium models from economics (Genotte and Marsh, 1993), and compartment models from pharmacokinetics (Mallet, Mentré, Steimer, and Lokiec, 1988).

Standard statistical methods, both classical and Bayesian, are usually not applicable in these situations either because it is not practicable to obtain the likelihood for the entire state vector or because the integration required to eliminate unobservables from the likelihood is infeasible. On a case-by-case basis, statistical methods are often available. However, our purpose here is to describe methods that are generally applicable.

Although determining the likelihood of a nonlinear dynamic system that has unobserved variables is often infeasible, simulating the evolution of the state vector is often quite practicable. Our methods rely on this. We describe simulated method of moments methods in general and then focus the discussion on efficient method of moments (EMM).

Briefly, the steps involved in EMM are as follow: Summarize the data by using quasi maximum likelihood to project the observed data onto a a transition density that is a close approximation to the true data generating process. This transition density is called the auxiliary model and its score is called the score generator for EMM. A Hermite series representation of the transition density of the observable process is suggested as a convenient general purpose auxiliary model in this connection. Once a score generator is in hand, given a parameter setting for the system, one may use simulation to evaluate the expected value of the score under the stationary density of the system and compute a chi-squared criterion function. A nonlinear optimizer is used to find the parameter setting that minimizes the criterion.

If the auxiliary model encompasses the true data generating process, then quasi maximum likelihood estimates become sufficient statistics and EMM is fully efficient (Gallant and Tauchen, 1996). If the auxiliary model is a close approximation to the data generating process, then one can expect the efficiency of EMM to be close to that of maximum likelihood (Gallant and Long, 1997; Tauchen, 1997).

Because EMM is a minimum chi-squared estimator, diagnostic tests are available to assess system adequacy as well as are graphics that suggest reasons for failure. Subsequent reprojection of the estimated dynamic system provides a facility for model elucidation that is as convenient as if a likelihood were available. We employ the methodology for estimation and diagnostic assessment of models of stock prices that are partially observed systems of stochastic differential equations.

Due to the fundamental importance of diffusion models for stock price dynamics, there has been significant progress recently using alternative simulation strategies; see Brandt and Santa-Clara (1999), Durham and Gallant (2002), Elerian, Chib and Shephard (2001), and the references therein. Despite this recent progress, the alternatives to the simulation methods discussed here not as general purpose and are limited in their ability to deal with latent variables.

EMM has found several recent applications, and we list a few, mainly from finance. Andersen and Lund (1997), Dai and Singleton (2001), Ahn, Dittmar, and Gallant (2001) use the method for interest rate applications. Liu (2000), Andersen, Benzoni, and Lund (2001), Chernov, Gallant, Ghysels, and Tauchen (2001) use it to estimate stochastic volatility models for stock prices with such complications as long memory and jumps. Chung and Tauchen (2001) use it to estimate and test target zone models of exchange rates. Jiang and van der Sluis (2000) use it to price options. Valderamma (2001) employs it for a macroeconomic analysis and Nagypal (2001) employs it in a labor economics application.

Software for EMM is available from several sources. Our version in Fortran is in the public domain and available at <ftp.econ.duke.edu> in directory `pub/get/emm`. A version for OX is described in the article by van der Sluis (1997). Jefferey Wang, Insightful Corp., is directing an NSF project for the inclusion of EMM in S-PLUS and Donald Erdman, SAS Institute Inc., is directing a project for inclusion in SAS.

## 2 Continuous Time Stochastic Volatility Models

The data used to illustrate ideas are observations on the daily price of a share of Microsoft (MSFT), adjusted for stock splits, from March 13, 1986, through November 16, 2000, yielding 3,712 raw observations. Figure 1 is a plot of the data. The top panel shows the raw price series,  $P_t$ , while the lower panel shows the return series,  $y_t = 100 * [\log(P_t) - \log(P_{t-1})]$ . The series  $\{y_t\}$  is the daily (geometric) return on Microsoft expressed as percentage. Table 1 shows basic statistics on the returns series, both daily and annualized, with 252 used as the annualization factor.

---

Figure 1 about here

---

---

Table 1 about here

---

The finance literature normally treats stock prices as a diffusion, usually expressed as a system of stochastic differential equations. The data shown in Figure 1 would thus be regarded as having resulted from discretely sampling a diffusion.

An example is

$$\begin{aligned}dU_{1t} &= \alpha_{10}dt + \exp(\beta_{10} + \beta_{12}U_{2t} + \beta_{13}U_{3t}) dW_{1t} \mathbf{0} \leq t < \infty \\dU_{2t} &= (\alpha_{20} + \alpha_{22}U_{2t})dt + \beta_{20}dW_{2t}, \\dU_{3t} &= (\alpha_{30} + \alpha_{33}U_{3t})dt + \beta_{30}dW_{3t},\end{aligned}\tag{1}$$

where  $U_{1t}$  is observable at integer  $t$  and represents the continuous-time  $\log(P_t)$  process. The second and third components are not observable and represent factors that affect the volatility of the process. Versions of this model in which there is only one volatility factor, i.e.,  $U_{3t}$  is absent,

$$\begin{aligned}dU_{1t} &= \alpha_{10}dt + \exp(\beta_{10} + \beta_{12}U_{2t})dW_{1t} \mathbf{0} \leq t < \infty \\dU_{2t} &= (\alpha_{20} + \alpha_{22}U_{2t})dt + \beta_{20}dW_{2t},\end{aligned}\tag{2}$$

have a long history in financial economics. More recently, however, finance economists have utilized models with two volatility factors. (See Chernov, Gallant, Ghysels, and Tauchen (2001) for references and an application to the Dow Jones Index.) The need for two volatility factors will become apparent in the empirical example below.

In matrix notation, the system (1) is

$$dU_t = A(U_t) dt + B(U_t) dW_t \quad 0 \leq t < \infty \quad (3)$$

where

$$U_t = \begin{pmatrix} U_{1t} \\ U_{2t} \\ U_{3t} \end{pmatrix}, \quad A(U) = \begin{pmatrix} \alpha_{10} \\ \alpha_{20} + \alpha_{22}U_2 \\ \alpha_{30} + \alpha_{33}U_3 \end{pmatrix},$$

$$W_t = \begin{pmatrix} W_{1t} \\ W_{2t} \\ W_{3t} \end{pmatrix}, \quad B(U) = \begin{pmatrix} \exp(\beta_{10} + \beta_{12}U_2 + \beta_{13}U_3) & 0 & 0 \\ 0 & \beta_{20} & 0 \\ 0 & 0 & \beta_{30} \end{pmatrix}.$$

The process is discretely sampled so that the data available for analysis are

$$y_t = 100 * (U_{1,t} - U_{1,t-1}) \quad t = 1, 2, \dots \quad (4)$$

which corresponds directly to the returns series plotted in the lower panel of Figure 1.

Clearly, not all parameters of the system (1) are separately identified. We impose the restrictions

$$\alpha_{20} = 0, \beta_{20} = 1, \alpha_{30} = 0, \beta_{30} = 1 \quad (5)$$

to achieve identification. These restrictions provide flexibility and numerical stability in the estimation work. The free parameters of the system are thus

$$\rho = (\alpha_{10}, \alpha_{22}, \alpha_{33}, \beta_{10}, \beta_{12}, \beta_{13}) \quad (6)$$

As is well known since Lo (1988), the likelihood of the observed process  $y_t$  given  $\rho$  under the system dynamics

$$dU_t = A(U_t) dt + B(U_t) dW_t \quad (7)$$

is not readily available in closed form. This aspect of the problem motivates method of moments estimation, and, in particular, simulated method of moments as in Ingram and Lee (1991) and Duffie and Singleton (1993), and the essentially equivalent indirect inference method proposed by Gourieroux, Monfort and Renault (1993) and Smith (1993).

The system dynamics (7) suggest that, given a value for  $U_0$ , one could simulate an increment  $U_\Delta - U_0$  from the process  $\{U_t : 0 \leq t < \infty\}$  for a small time value  $\Delta$  as follows: generate three independent normal  $(0, \Delta^2)$  variates  $x_1$ ,  $x_2$  and  $x_3$ ; simulate the Brownian motion increment  $W_\Delta - W_0$  by putting

$$W_\Delta - W_0 = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

and simulate the increment  $U_\Delta - U_0$  by putting

$$U_\Delta - U_0 = A(U_0)\Delta + B(U_0)(W_\Delta - W_0).$$

To simulate a value for  $U_t$ , sum over the increments:

$$\begin{aligned} U_t - U_0 &= \sum_{i=1}^{t/\Delta} (U_{i\Delta} - U_{(i-1)\Delta}) \\ &= \sum_{i=1}^{t/\Delta} A(U_{(i-1)\Delta})\Delta + \sum_{i=1}^{t/\Delta} B(U_{(i-1)\Delta})(W_{i\Delta} - W_{(i-1)\Delta}). \end{aligned} \tag{8}$$

In passing, we note that under regularity conditions (Karatzas and Shreve, 1991), as  $\Delta$  decreases the random variable  $\sum_{i=1}^{t/\Delta} B(U_{(i-1)\Delta})(W_{i\Delta} - W_{(i-1)\Delta})$  converges in mean square to a random variable that has many properties of an integral and is therefore usually denoted as  $\int_0^t B(U_s) dW_s$ . Similarly, the Reimann sum  $\sum_{i=1}^{t/\Delta} A(U_{(i-1)\Delta})\Delta$  converges to  $\int_0^t A(U_s) ds$ . The process  $\{U_t : 0 \leq t < \infty\}$  is interpreted as the solution to the integral equation

$$U_t - U_0 = \int_0^t A(U_s) ds + \int_0^t B(U_s) dW_s,$$

which exists under smoothness and growth conditions on the functions  $A(U)$  and  $B(U)$  (Karatzas and Shreve, 1991).

The important feature of this example, and of all the applications that we consider, is that it can be simulated. That is, given a value for the parameters of the model it is straightforward to generate a simulation  $\{\hat{y}_t\}_{t=1}^N$ , of arbitrary length  $N$ .

The simulation scheme (8) above is known as an Euler scheme. More sophisticated simulations schemes are available in Kloeden and Platen (1992). Those used here (euler.f, strong1.f, and weak2.f) are available by anonymous ftp from host ftp.econ.duke.edu in directory pub/arg/libf.

If the model (1) is stationary at a given value for  $\rho$ , then a time invariant stationary density

$$p(y_{-L}, \dots, y_{-1}, y_0 | \rho) \tag{9}$$

exists for any stretch  $(y_{t-L}, \dots, y_{t-1}, y_t)$  of obseables. If, in addition, (1) is ergodic at  $\rho$ , then the expectation of a time invariant (nonlinear) function  $g(y_{-L}, \dots, y_{-1}, y_0)$  with respect to (9)

$$\mathcal{E}_\rho(g) = \int \cdots \int g(y_{-L}, \dots, y_0) p(y_{-L}, \dots, y_0 | \rho) dy_{-L} \cdots dy_0,$$

can be approximated as accurately as desired by averaging over a simulation, viz.

$$\mathcal{E}_\rho(g) \doteq \frac{1}{N} \sum_{t=1}^N g(\hat{y}_{t-L}, \dots, \hat{y}_{t-1}, \hat{y}_t).$$

Throughout, we presume that the requisite initial lags for this computation are primed via draws from the stationary distribution, which is usually accomplished by letting the system run long enough for transients die out. For examples such as above, we typically use values of  $\Delta$  on the order of 1/7 per week or 1/24 per day, a burn in period of 1,000 to 5,000, and values of  $N$  on the order of 50,000 to 100,000.

The ability to compute  $\mathcal{E}_\rho(g)$  for given  $\rho$  and arbitrary  $g(y_{-L}, \dots, y_{-1}, y_0)$  means that model parameters can be computed by method of moments or minimum chi-squared, as we discuss next.

### 3 Minimum Chi-Squared Estimators

In general, we consider nonlinear systems that have the features of the stochastic volatility model (1) just described. Specifically, (i) for a parameter vector  $\rho$  in a parameter space  $R$ , the random variables determined by the system have a stationary density

$$p(y_{-L}, \dots, y_{-1}, y_0 | \rho), \tag{10}$$

for every stretch  $(y_{t-L}, \dots, y_t)$ ; and (ii) for  $\rho \in R$ , the system is easily simulated so that expectations

$$\mathcal{E}_\rho(g) = \int \cdots \int g(y_{-L}, \dots, y_0) p(y_{-L}, \dots, y_0 | \rho) dy_{-L} \cdots dy_0 \quad (11)$$

can be approximated as accurately as desired by averaging over a long simulation

$$\mathcal{E}_\rho(g) \doteq \frac{1}{N} \sum_{t=1}^N g(\hat{y}_{t-L}, \dots, \hat{y}_{t-1}, \hat{y}_t). \quad (12)$$

Henceforth, we shall use  $\{y_t\}$  to denote the stochastic process determined by the system,  $\{\hat{y}_t\}_{t=1}^N$  to denote a simulation from the system,  $\{\tilde{y}_t\}_{t=1}^n$  to denote data presumed to have been generated by the system, and  $(y_{-L}, \dots, y_{-1}, y_0)$  to denote function arguments and dummy variables of integration. The true value of the parameter vector of the system (10) is denoted by  $\rho^\circ$ .

A method of moments estimator  $\hat{\rho}_n$  of  $\rho^\circ$  is implemented by (i) setting forth a moment function, such as

$$\tilde{\psi}_c(y_{-L}, \dots, y_{-1}, y_0) = \begin{pmatrix} y_0 & - & \tilde{\mu}_1 \\ y_0^2 & - & \tilde{\mu}_2 \\ \vdots & & \\ y_0^k & - & \tilde{\mu}_k \\ y_{-1}y_0 & - & \tilde{\gamma}(1) \\ y_{-2}y_0 & - & \tilde{\gamma}(2) \\ \vdots & & \\ y_{-L}y_0 & - & \tilde{\gamma}(L) \end{pmatrix}$$

where  $\tilde{\mu}_j = \frac{1}{n} \sum_{t=1}^n \tilde{y}_t^j$ ,  $\tilde{\gamma}(h) = \frac{1}{n} \sum_{t=1+h}^n \tilde{y}_t \tilde{y}_{t-h}$ ; (ii) computing the moment equations

$$m_n(\rho) = \mathcal{E}_\rho(\psi_c) = \int \cdots \int \tilde{\psi}_c(y_{-L}, \dots, y_0) p(y_{-L}, \dots, y_0 | \rho) dy_{-L} \cdots dy_0;$$

and (iii) attempting to solve the estimating equations

$$m_n(\rho) = 0$$

for the system parameters  $\rho$ . If a solution  $\hat{\rho}_n$  can be found, then that solution is the method of moments estimate of the system parameters. As indicated earlier, the moment equations

will usually have to be computed by generating a long simulation  $\{\hat{y}_t\}_{t=-L}^N$  from the system at parameter setting  $\rho$  and then averaging over the simulation,

$$m_n(\rho) \doteq \frac{1}{N} \sum_{t=1}^N \tilde{\psi}_c(\hat{y}_{t-L}, \dots, \hat{y}_{t-1}, \hat{y}_t).$$

If there are multiple roots of the estimating equations  $m_n(\rho) = 0$ , a particular solution can be selected as the estimate using methods discussed in Heyde and Morton (1998).

In the event that there is no solution to the estimating equations because, for instance, the dimension of  $\psi_c$  is larger than the dimension of  $\rho$  (so that there are more equations than unknowns), then one must resort to minimum chi-squared estimation (Neyman and Pearson, 1928) as adapted to dynamic models by Hansen (1982). The minimum chi-squared estimator is obtained by using a nonlinear optimizer minimize a quadratic form in the moment equations. Specifically,

$$\hat{\rho}_n = \operatorname{argmin}_{\rho} m'_n(\rho) (\tilde{\mathcal{I}}_n)^{-1} m_n(\rho);$$

the matrix  $\tilde{\mathcal{I}}_n$  appearing in the quadratic form is an estimate of the variance of  $\sqrt{n} m'_n(\rho)$  and may be computed as (Gallant, 1987)

$$\tilde{\mathcal{I}}_n = \sum_{\tau=-\lceil n^{1/5} \rceil}^{\lceil n^{1/5} \rceil} w\left(\frac{\tau}{\lceil n^{1/5} \rceil}\right) \tilde{\mathcal{I}}_{n\tau} \quad (13)$$

where

$$w(u) = \begin{cases} 1 - 6|u|^2 + 6|u|^3 & \text{if } 0 < u < \frac{1}{2} \\ 2(1 - |u|)^3 & \text{if } \frac{1}{2} \leq u < 1, \end{cases}$$

and

$$\tilde{\mathcal{I}}_{n\tau} = \begin{cases} \frac{1}{n} \sum_{t=1+\tau}^n \tilde{\psi}_c(\tilde{y}_{t-L}, \dots, \tilde{y}_t) \tilde{\psi}'_c(\tilde{y}_{t-L-\tau}, \dots, \tilde{y}_{t-\tau}) & \text{if } \tau \geq 0 \\ \tilde{\mathcal{I}}_{n,-\tau} & \text{if } \tau < 0 \end{cases}$$

If  $y_t$  is multivariate, that is,

$$y_t = (y_{1,t}, \dots, y_{M,t})',$$

then, instead of the above, the vector  $\psi_c$  is comprised of the elements

$$\prod_{i=1}^M (y_{i,0})^{\alpha_i} - \tilde{\mu}_{\alpha}$$

$$(y_{i,0})(y_{j,-h}) - \gamma_{ij}(h),$$

where

$$\begin{aligned}\alpha &= (\alpha_1, \dots, \alpha_M)', \quad \alpha_i \geq 0, \quad 0 < \sum_{i=1}^M \alpha_i \leq K \\ \tilde{\mu}_\alpha &= \frac{1}{n} \sum_{t=1}^n \prod_{i=1}^M (\tilde{y}_{i,t})^{\alpha_i} \\ \gamma_{ij}(h) &= \frac{1}{n} \sum_{t=1+h}^n \tilde{y}_{i,t} \tilde{y}_{j,t-h}, \quad 0 \leq i \leq j \leq M, \quad 0 \leq h \leq L.\end{aligned}$$

The use of method of moments together with simulation to estimate the parameters of dynamic models with unobserved variables has been proposed by Ingram and Lee (1991), Duffie and Singleton (1993), Gouriéroux, Monfort, and Renault (1993), Smith (1993), and others. The particular methods that we discuss next are due to Gallant and Tauchen (1996).

## 4 Efficiency Considerations

We have described a minimum chi-squared estimation strategy based on the moment function  $\tilde{\psi}_c$  that can be used to estimate system parameters  $\rho$ . There are two open questions with regard to this estimator: what is the best choice of the moment function  $\tilde{\psi}$ , and, how many moments should be included in  $\tilde{\psi}$ ?

We will consider the questions in the simplest case where the random variables defined by the system (10) generate univariate independently and identically distributed random variables  $\{y_t\}$  with density  $p(y|\rho)$ . The ideas for the general case of a multivariate, non-Markovian, stationary system are the same, but the algebra is far more complicated (Gallant and Long, 1997). Nothing essential is lost by considering the simplest case.

Consider three moment functions  $\tilde{\psi}_{c,n}$ ,  $\tilde{\psi}_{p,n}$ , and  $\tilde{\psi}_{f,n}$  that correspond to Classical Method of Moments, Maximum Likelihood, and Efficient Method of Moments, respectively, defined as follows:

$$\begin{aligned}\tilde{\psi}_{c,n}(y) &= \begin{pmatrix} y - \frac{1}{n} \sum_{i=1}^n \tilde{y}_i \\ y^2 - \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i)^2 \\ \vdots \\ y^K - \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i)^K \end{pmatrix}, \\ \tilde{\psi}_{p,n}(y) &= \frac{\partial}{\partial \rho} \log p(y|\tilde{\rho}_n),\end{aligned}$$

$$\tilde{\psi}_{f,n}(y) = \frac{\partial}{\partial \theta} \log f(y|\tilde{\theta}_n),$$

where the exponent  $K$  that appears in  $\tilde{\psi}_{c,n}(y)$  is the degree of the largest moment used in a method of moments application, the function  $f(y|\theta)$  that appears in  $\tilde{\psi}_{f,n}(y)$  is a density that closely approximates the true data generating process in a sense made precise later, and the statistics  $\tilde{\rho}_n$  and  $\tilde{\theta}_n$  that appear in  $\tilde{\psi}_{p,n}(y)$  and  $\tilde{\psi}_{f,n}(y)$  are

$$\tilde{\rho}_n = \operatorname{argmax}_{\rho} \frac{1}{n} \sum_{i=1}^n \log p(\tilde{y}_i|\rho),$$

$$\tilde{\theta}_n = \operatorname{argmax}_{\theta} \frac{1}{n} \sum_{i=1}^n \log f(\tilde{y}_i|\theta);$$

$\rho$  is of length  $p_{\rho}$  and  $\theta$  of length  $p_{\theta} \geq p_{\rho}$ .

Note that each of the moment functions  $\tilde{\psi}_{p,n}$ ,  $\tilde{\psi}_{c,n}$ , and  $\tilde{\psi}_{f,n}$  is in the null space of the expectation operator corresponding to the empirical distribution of the data, denoted as  $\mathcal{E}_{\tilde{F}_n}$ . That is,  $\mathcal{E}_{\tilde{F}_n} \tilde{\psi}_{p,n} = \mathcal{E}_{\tilde{F}_n} \tilde{\psi}_{c,n} = \mathcal{E}_{\tilde{F}_n} \tilde{\psi}_{f,n} = 0$ . Method of moments is basically an attempt to do the same for the model  $p(y|\rho)$ . That is, method of moments attempts to find a  $\rho$  that puts one of these moment functions, denoted generically as  $\tilde{\psi}_n$ , in the null space of the expectation operator  $\mathcal{E}_{\rho}$  corresponding to  $p(y|\rho)$ .

In addition to computing  $\tilde{\psi}_n$ , one computes

$$\tilde{\mathcal{I}}_n = \mathcal{E}_{\tilde{F}_n}(\tilde{\psi}_n)(\tilde{\psi}_n)'$$

Once  $\tilde{\psi}_n$  and  $\tilde{\mathcal{I}}_n$  have been computed, the data have been summarized, and what we refer to as “the projection step” is finished.

For estimation, define

$$m_n(\rho) = \mathcal{E}_{\rho} \tilde{\psi}_n.$$

If the dimensions of  $\rho$  and  $\tilde{\psi}_n(y)$  are the same, then usually the equations  $m_n(\rho) = 0$  can be solved to obtain an estimator  $\hat{\rho}_n$ . For  $\tilde{\psi}_{p,n}$ , the solution is the maximum likelihood estimator (Gauss, 1816; Fisher, 1912). For  $\tilde{\psi}_{c,n}$  with  $K = p_{\rho}$ , it is the classical method of moments estimator (Pearson, 1894). For  $\tilde{\psi}_{c,n}$  with  $K > p_{\rho}$ , no solution exists and the moment functions  $\tilde{\psi}_{c,n}$  are those of minimum chi-squared or generalized method of moments (Neyman and Pearson, 1928; Hansen, 1982) as customarily implemented.

As just noted, when  $K > p_\rho$ , then  $\tilde{\psi}_n$  cannot be placed in the null space of the operator  $\mathcal{E}_\rho$  for any  $\rho$ , because the equations  $m_n(\rho) = 0$  have no solution. In this case, the minimum chi-squared estimator relies on the fact that, under standard regularity conditions (Gallant and Tauchen, 1996) and choices of  $\tilde{\psi}_n$  similar to the above, there is a function  $\psi^\circ$  such that

$$\begin{aligned}\lim_{n \rightarrow \infty} \tilde{\psi}_n(y) &= \psi^\circ(y) \quad \text{a.s.} \\ \lim_{n \rightarrow \infty} \tilde{\mathcal{I}}_n &= \mathcal{E}_{\rho^\circ}(\psi^\circ)(\psi^\circ)' \quad \text{a.s.} \\ \sqrt{n} m_n(\rho^\circ) &\xrightarrow{\mathcal{L}} N[0, \mathcal{E}_{\rho^\circ}(\psi^\circ)(\psi^\circ)']\end{aligned}$$

where  $\mathcal{E}_{\rho^\circ}$  denotes expectation taken with respect to  $p(y|\rho^\circ)$ . For the three choices  $\tilde{\psi}_{p,n}$ ,  $\tilde{\psi}_{c,n}$ , and  $\tilde{\psi}_{f,n}$  of  $\psi_n(y)$  above, the functions  $\psi_p^\circ$ ,  $\psi_c^\circ$ , and  $\psi_f^\circ$  given by this result are

$$\psi_c^\circ(y) = \begin{pmatrix} y - \mathcal{E}_{\rho^\circ}(y) \\ y^2 - \mathcal{E}_{\rho^\circ}(y^2) \\ \vdots \\ y^K - \mathcal{E}_{\rho^\circ}(y^K) \end{pmatrix}$$

$$\psi_p^\circ(y) = \frac{\partial}{\partial \rho} \log p(y|\rho^\circ)$$

and

$$\psi_f^\circ(y) = \frac{\partial}{\partial \theta} \log f(y|\theta^\circ),$$

where

$$\theta^\circ = \operatorname{argmax}_\theta \mathcal{E}_{\rho^\circ} \log f(\cdot|\theta).$$

With these results in hand,  $\rho$  may be estimated by minimum chi-squared, viz.,

$$\hat{\rho}_n = \operatorname{argmin}_\rho m_n'(\rho) (\tilde{\mathcal{I}}_n)^{-1} m_n(\rho)$$

and

$$\sqrt{n}(\hat{\rho}_n - \rho^\circ) \xrightarrow{\mathcal{L}} N[0, (C^\circ)^{-1}],$$

where

$$C^\circ = [\mathcal{E}_{\rho^\circ}(\psi_p^\circ)(\psi_p^\circ)'] [\mathcal{E}_{\rho^\circ}(\psi^\circ)(\psi^\circ)']^{-1} [\mathcal{E}_{\rho^\circ}(\psi^\circ)(\psi_p^\circ)'].$$

Note that for any nonzero  $a \in \mathcal{R}^{p\rho}$ ,

$$\min_b \mathcal{E}_{\rho^o} [a' \psi_p^o - (\psi^o)' b]^2 = \mathcal{E}_{\rho^o} (a' \psi_p^o)^2 - a' C^o a \geq 0. \quad (14)$$

Expression (14) implies that  $a' C^o a$  cannot exceed  $\mathcal{E}_{\rho^o} (a' \psi_p^o)^2 = a' [\mathcal{E}_{\rho^o} (\psi_p^o) (\psi_p^o)'] a$  and therefore the best achievable asymptotic variance of the estimator  $\hat{\rho}_n$  is  $(\mathcal{I}_p^o)^{-1} = [\mathcal{E}_{\rho^o} (\psi_p^o) (\psi_p^o)']^{-1}$ , which is the variance of the maximum likelihood estimator of  $\rho$ . It is also apparent from (14) that if  $\{\psi_i^o\}_{i=1}^\infty$  spans the  $L_{2,p}$  probability space  $L_{2,p} = \{g : \mathcal{E}_{\rho^o} g^2 < \infty\}$  and  $\psi^o = (\psi_1^o, \dots, \psi_K^o)$ , then  $\hat{\rho}_n$  has good efficiency relative to the maximum likelihood estimator for large  $K$ . The polynomials span  $L_{2,p}$  if  $p(y|\rho)$  has a moment generating function (Gallant, 1980). Therefore, one might expect good asymptotic efficiency from  $\tilde{\psi}_{c,n}$  for large  $K$ .

Rather than just spanning  $L_{2,p}$ , EMM requires, in addition, that the moment functions actually be the score vector  $\psi_{f,n}(y)$  of some density  $f(y|\tilde{\theta}_n)$  that closely approximates  $p(y|\rho^o)$ . Possible choices of  $f(y|\tilde{\theta}_n)$  are discussed in Gallant and Tauchen (1996). Of them, one commonly used in applications is the SNP density, which was proposed by Gallant and Nychka (1987) in a form suited to cross-sectional applications and by Gallant and Tauchen (1989) in a form suited to time-series applications.

The SNP density is obtained by expanding the square root of an innovation density  $h(z)$  in a Hermite expansion

$$\sqrt{h(z)} = \sum_{i=0}^{\infty} \theta_i z^i \sqrt{\phi(z)},$$

where  $\phi(z)$  denotes the standard normal density function. Because the Hermite functions are dense in  $L_2$  (Lebesgue) and  $\sqrt{h(z)}$  is an  $L_2$  function, this expansion must exist. The truncated density is

$$h_K(z) = \frac{\mathcal{P}_K^2(z) \phi(z)}{\int \mathcal{P}_K^2(u) \phi(u) du},$$

where

$$\mathcal{P}_K(z) = \sum_{i=0}^K \theta_i z^i$$

and the renormalization is necessary so that the density  $h_K(z)$  integrates to one. The location-scale transformation  $y = \sigma z + \mu$  completes the definition of the SNP density

$$f_K(y|\theta) = \frac{1}{\sigma} h_K\left(\frac{y - \mu}{\sigma}\right). \quad (15)$$

with  $\theta = (\mu, \sigma, \theta_0, \dots, \theta_K)$ . Gallant and Long (1997) have shown that

$$\psi_f^o(y) = \frac{\partial}{\partial \theta} \log f_K(y|\theta^o),$$

with

$$\theta^o = \underset{\theta}{\operatorname{argmax}} \mathcal{E}_{\rho^o} \log f_K(\cdot|\theta)$$

spans  $L_{2,p}$ .

While a spanning argument can be used to show that high efficiency obtains for large  $K$ , it gives no indication as to what might be the best choice of moment functions with which to span  $L_{2,p}$ . Moreover, if  $\psi_p$  is in the span of  $\psi^o$  for some finite  $K$ , then full efficiency obtains at once (Gallant and Tauchen, 1996). For instance, the score of the normal density is in the span of both  $\tilde{\psi}_{c,n}$  and  $\tilde{\psi}_{f,n}$  for  $K \geq 2$ . These considerations seem to rule out any hope of general results showing that one moment function should be better than another.

With general results unattainable, the best one can do is compare efficiencies over a class of densities designed to stress-test an estimator and over some densities thought to be representative of situations likely to be encountered in practice to see if any conclusions seem to be indicated. Comparisons using Monte Carlo methods are reported by Andersen, Chung, and Sorensen (1999), Chumacero (1997), Ng and Michaelides (2000), and van der Sluis (1999). Overall, their work supports the conjecture that EMM is more efficient than CMM in representative applications at typical sample sizes.

Analytical comparisons are possible for the independently and identically distributed case and are reported in Gallant and Tauchen (1999). Their measure of efficiency is the volume of a confidence region on the parameters of the density  $p(y|\rho)$  computed using the asymptotic distribution of  $\hat{\rho}_n$ . This region has the form  $\{\rho : (\rho - \rho^o)'(C^o)^{-1}(\rho - \rho^o) \leq \mathcal{X}_d^2/n\}$  with volume

$$\frac{2\pi^{d/2}(\mathcal{X}_d^2/n)^d}{d\Gamma(d/2) \det(C^o)},$$

where  $\mathcal{X}_d^2$  denotes a critical value of the chi-squared distribution on  $d$  degrees of freedom. As small volumes are to be preferred, and the region  $\{\rho : (\rho - \rho^o)' \mathcal{I}_p^o(\rho - \rho^o) \leq \mathcal{X}_d^2/n\}$  has the smallest achievable volume,

$$\text{RE} = \frac{\det(C^o)}{\det(\mathcal{I}_p^o)}$$

is a measure of relative efficiency. Over a large collection of densities thought to represent typical applications, their computations support the conclusion that EMM dominates CM-M. Moreover, their computations indicate that once  $f_K(\cdot|\theta^o)$  begins to approximate  $p(\cdot|\rho^o)$  accurately, the efficiency of the EMM estimator begins to increase rapidly. A representative illustration is provided by Figure 2, which shows the relative efficiency comparison for a trimodal density  $p(y|\rho)$  taken from the Marron-Wand test suite (Marron and Wand, 1992). As seen in Figure 2, once  $f_K(\cdot|\theta^o)$  has detected the third mode of the trimodal density, EMM efficiency increases rapidly.

---

Figure 2 about here

---

The second question to address is how many moments to include in the moment function  $\psi_f$ . As the computations in Gallant and Tauchen (1999) and Figure 2 suggest, the answer is as many as is required for  $f$  to well approximate  $p$ . The natural conclusion is that one should use standard statistical model selection criteria to determine  $f$  as we discuss later. This approach has a distinct advantage over the use of  $\psi_c$  in that there seems to be no objective statistical criterion for determining the number of moments to include in  $\psi_c$ .

## 5 A General Purpose Auxiliary Model

As indicated in Section 4, the best choice of a moment function  $\psi$  to implement simulated method of moments is the score of a auxiliary model that closely approximates the density of the data. We have also seen that the SNP density is a useful, general purpose auxiliary model. In this section, we shall extend the SNP density to a general purpose auxiliary model suited to dynamic models. Here,  $y_t$  is multivariate, specifically a column vector of length  $M$ , and we write  $x_{t-1}$  for the lagged state vector, which typically is comprised of lags  $y_{t-j}$ . For simplicity, we often suppress the time subscript and write  $y$  and  $x$  for the contemporaneous value and lagged state vector, respectively. With these conventions, the stationary density (10) of the dynamic system under consideration can be written  $p(x, y|\rho)$  and its transition

density as

$$p(y|x, \rho) = \frac{p(x, y|\rho)}{\int p(x, y|\rho) dx} \quad (16)$$

If one expands  $\sqrt{p(x, y|\rho)}$  in a Hermite series and derives the transition density of the truncated expansion, then one obtains a transition density  $f_K(y_t | x_{t-1})$  that has the form of a location-scale transform

$$y_t = Rz_t + \mu_{x_{t-1}},$$

of an innovation  $z_t$  (Gallant, Hsieh, and Tauchen, 1991). The density function of this innovation is

$$h_K(z|x) = \frac{[\mathcal{P}(z, x)]^2 \phi(z)}{\int [\mathcal{P}(u, x)]^2 \phi(u) du}, \quad (17)$$

where  $\mathcal{P}(z, x)$  is a polynomial in  $(z, x)$  of degree  $K$  and  $\phi(z)$  denotes the multivariate normal density function with dimension  $M$ , mean vector zero, and variance-covariance matrix the identity.

It proves convenient to express the polynomial  $\mathcal{P}(z, x)$  in a rectangular expansion

$$\mathcal{P}(z, x) = \sum_{\alpha=0}^{K_z} \left( \sum_{\beta=0}^{K_x} a_{\beta\alpha} x^\beta \right) z^\alpha, \quad (18)$$

where  $\alpha$  and  $\beta$  are multi-indexes of maximal degrees  $K_z$  and  $K_x$ , respectively, and  $K = K_z + K_x$ . Because  $[\mathcal{P}(z, x)]^2 / \int [\mathcal{P}(u, x)]^2 \phi(u) du$  is a homogeneous function of the coefficients of the polynomial  $\mathcal{P}(z, x)$ ,  $\mathcal{P}(z, x)$  can only be determined to within a scalar multiple. To achieve a unique representation, the constant term  $a_{00}$  of the polynomial  $\mathcal{P}(z, x)$  is put to one. With this normalization,  $h_K(z|x)$  has the interpretation of a series expansion whose leading term is the normal density  $\phi(z)$  and whose higher order terms induce departures from normality.

The location function is linear

$$\mu_x = b_0 + Bx_{t-1}, \quad (19)$$

where  $b_0$  is a vector and  $B$  is a matrix.

It proves advantageous in applications to allow the scale  $R$  of the location-scale transformation  $y = Rz + \mu_x$  to depend on  $x$  because it reduces the degree  $K_x$  required to achieve

an adequate approximation to the transition density  $p(y|x, \rho^o)$ . With this, the location-scale transformation becomes

$$y = R_x z + \mu_x \quad (20)$$

where  $R_x$  is an upper triangular matrix that depends on  $x$ . The two choices of  $R_x$  that have given good results in applications are an ARCH-like moving average specification and a GARCH-like ARMA specification, which we describe next.

For an ARCH specification, let  $R_{x_{t-1}}$  be a linear function of the absolute values of the elements of the vectors  $y_{t-L_r} - \mu_{x_{t-1-L_r}}$  through  $y_{t-1} - \mu_{x_{t-2}}$ , viz.

$$\text{vech}(R_{x_{t-1}}) = \rho_0 + \sum_{i=1}^{L_r} P_{(i)} |y_{t-1-L_r+i} - \mu_{x_{t-2-L_r+i}}|$$

where  $\text{vech}(R)$  denotes a vector of length  $M(M+1)/2$  containing the elements of the upper triangle of  $R$ ,  $\rho_0$  is a vector of length  $M(M+1)/2$ ,  $P_{(1)}$  through  $P_{(L_r)}$  are  $M(M+1)/2$  by  $M$  matrices, and  $|y - \mu|$  denotes a vector containing the absolute values of  $y - \mu$ . The classical ARCH (Engle, 1982) has

$$\Sigma_{x_{t-1}} = R_{x_{t-1}} R'_{x_{t-1}}$$

depending on a linear function of squared lagged residuals. The SNP version of ARCH is more akin to the suggestions of Nelson (1991) and Davidian and Carroll (1987).

Since the absolute value function is not differentiable,  $|u|$  is approximated in the formula for  $R_x$  above by the twice continuously differentiable function

$$a(u) = \begin{cases} (|100u| - \pi/2 + 1) / 100 & |100u| \geq \pi/2 \\ (1 - \cos(100u)) / 100 & |100u| < \pi/2 \end{cases}$$

The scale factor 100 above represents a compromise. Small values, such as 3, improve the stability of the computations but then  $a(\cdot)$  does not approximate  $|\cdot|$  well.

For a GARCH specification, let

$$\text{vech}(R_{x_{t-1}}) = \rho_0 + \sum_{i=1}^{L_r} P_{(i)} |y_{t-1-L_r+i} - \mu_{x_{t-2-L_r+i}}| + \sum_{i=1}^{L_g} \text{diag}(G_{(i)}) \text{vech}(R_{x_{t-2-L_g+i}}) \quad (21)$$

where  $G_{(1)}$  through  $G_{(L_g)}$  are vectors of length  $M(M+1)/2$ .

The classical GARCH (Bollerslev, 1986) has  $\Sigma_{x_{t-1}}$  expressed in terms of squared lagged residuals and lagged values of  $\Sigma_{x_{t-1}}$ . As with the SNP variant of ARCH, the SNP version

of GARCH is expressed in terms of the absolute value of lagged residuals and standard deviations.

Note that when  $L_g > 0$ , the SNP model is not Markovian and that one must know both  $x_{t-1}$  and  $R_{x_{t-2-L_g}}$  through  $R_{x_{t-2}}$  to move forward to the value for  $y_t$ . Thus,  $x_{t-1}$  and  $R_{x_{t-2-L_g}}$  through  $R_{x_{t-2}}$  represent the state of the system at time  $t-1$  and must be retained in order to evaluate the SNP conditional density of  $y_t$  or to iterate the SNP model forward by simulation. If one wants to compute the derivatives of the SNP density with respect to model parameters, one must retain the derivatives of  $R_{x_{t-2-L_g}}$  through  $R_{x_{t-2}}$  with respect to model parameters as well.

The change of variable formula applied to the location-scale transform (20) and innovation density (17) yields the SNP density

$$f_K(y|x, \theta) = \frac{h_K[R_x^{-1}(y - \mu_x)|x]}{\det(R_x)}. \quad (22)$$

Hereafter, we shall distinguish among the lag lengths appearing in the various components of the expansion. The number of lags in  $\mu_x$  is denoted  $L_u$ ; the number of lags in  $R_x$  is  $L_u + L_r$ , and the number of lags in the  $x$  part of the polynomial,  $\mathcal{P}(z, x)$ , is  $L_p$ . We set  $L = \max(L_u, L_u + L_r, L_p)$ .

Large values of  $M$  can generate a large number of interactions (cross product terms) for even modest settings of degree  $K_z$ ; similarly, for  $M \cdot L_p$  and  $K_x$ . Accordingly, we introduce two additional tuning parameters,  $I_z$  and  $I_x$ , to represent filtering out of these high order interactions.  $I_z = 0$  means no interactions are suppressed,  $I_z = 1$  means the highest order interactions are suppressed, namely those of degree  $K_z$ . In general, a positive  $I_z$  means all interactions of order larger than  $K_z - I_z$  are suppressed; similarly for  $K_x - I_x$ .

In summary,  $L_u$ ,  $L_g$ , and  $L_r$  determine the location-scale transformation  $y = R_x z_t + \mu_x$  and hence determine the nature of the leading term of the expansion. The number of lags in the location function  $\mu_x$  is  $L_u$  and the number of lags in the scale function  $R_x$  is  $L_u + L_r$ . The number of lags that go into the  $x$  part of the polynomial  $\mathcal{P}(z, x)$  is  $L_p$ . The parameters  $K_z$ ,  $K_x$ ,  $I_z$  and  $I_x$  determine the degree of  $\mathcal{P}(z, x)$  and hence the nature of the innovation process  $\{z_t\}$ .

Table 2 about here

---

Putting certain of the tuning parameters to zero implies sharp restrictions on the process  $\{y_t\}$ , the more interesting of which are displayed in Table 2.

## 6 The Projection Step

As suggested in Section 4 and verified by Gallant and Tauchen (1996) and Gallant and Long (1997), the best choice of a moment function to implement simulated method of moments for dynamic systems such as (3) and (4) is the score of a auxiliary model  $f(y|x, \theta)$  that closely approximates the transition density implied by the system (16), where the parameter vector  $\theta$  of the auxiliary model is evaluated at its quasi maximum likelihood estimate  $\tilde{\theta}_n$ . That is, the best choice has the form

$$\tilde{\psi}_f(x, y) = \frac{\partial}{\partial \theta} \log f(y|x, \tilde{\theta}_n), \quad (23)$$

where

$$\begin{aligned} \tilde{\theta}_n &= \underset{\theta}{\operatorname{argmin}} s_n(\theta) \\ s_n(\theta) &= -\frac{1}{n} \sum_{t=1}^n \log f(\tilde{y}_t | \tilde{x}_{t-1}, \theta). \end{aligned} \quad (24)$$

A considerable advantage to closely approximating the transition density of the system is that the computational formula (13) for the weighting matrix  $\tilde{\mathcal{I}}_n$  given in Section 3 for the chi-squared estimator simplifies to

$$\tilde{\mathcal{I}}_n = \frac{1}{n} \sum_{t=1}^n \tilde{\psi}_f(\tilde{x}_{t-1}, \tilde{y}_t) \tilde{\psi}'_f(\tilde{x}_{t-1}, \tilde{y}_t) \quad (25)$$

because the covariance terms  $\tilde{\mathcal{I}}_{n\tau}$ ,  $\tau > 0$ , of (13) can be neglected when the auxiliary model closely approximates the transition density of the system in a sense made precise by Gallant and Long (1997).

In Section 5, we have suggested that the SNP density  $f_K(y|x, \theta)$  is a useful general purpose choice of auxiliary model in dynamic applications. The tuning parameters of the SNP density

$$(L_u, L_g, L_r, L_p, K_z, I_z, K_x, I_x)$$

may be determined by a statistical model selection criterion. In our work, we have used the Schwarz BIC criterion. The Schwarz BIC criterion (Schwarz, 1978) is computed as

$$\text{BIC} = s_n(\hat{\theta}) + (1/2)(p_\theta/n) \log(n)$$

with small values of the criterion preferred. The criterion rewards good fits as represented by small  $s_n(\hat{\theta})$  but uses the term  $(1/2)(p_\theta/n) \log(n)$  to penalize good fits gotten by means of excessively rich parameterizations. Our suggestion is to use the Schwarz BIC criterion to move along an upward expansion path until an adequate model is determined.

Coppejans and Gallant (2000) report Monte Carlo evidence suggesting that BIC underestimates  $K_z$  in iid situations. They propose a cross validation method that works well in the examples considered in their simulations. It is possible, even likely, that BIC underestimates the tuning parameters in the time series situation considered here. Coppejans and Gallant propose a cross validation method for time series SNP applications. More work is required to determine if BIC underestimates the tuning parameters in time series applications and to determine if the Coppejans and Gallant proposal determines the tuning parameters effectively.

We shall illustrate with the data from the lower panel of Figure 1, which are 3,711 observations on the daily return on Microsoft. For these data, the preferred fit has

$$(L_u, L_g, L_r, L_p, K_z, I_z, K_x, I_x) = (1, 1, 1, 1, 6, 0, 0, 0) \quad (26)$$

for which the dimension of  $\theta$ , and hence the dimension of  $\tilde{\psi}_f$ , is 11. The specification search follows the now established SNP protocol: The initial model is

$$(L_u, L_g, L_r, L_p, K_z, I_z, K_x, I_x) = (1, 1, 1, 1, 0, 0, 0, 0)$$

which is a Gaussian GARCH model defined in Table 3. From this point,  $K_z$  is expanded from 0 in increments of 2, up to a maximum of 8, until the BIC criterion indicates no further expansion is warranted. The increment of 2 is used because extensive experience suggests that, for financial data, even powers tend to dominate since these powers control tail-thickness, the pervasive feature of financial data. For the Microsoft data, the expansion stops at  $K_z = 6$ . Next,  $L_g$  and  $L_r$  are separately incremented, and for the Microsoft data, no

further expansion in those directions proved warranted. Finally,  $K_x$  is incremented to unity, which permits state dependence of polynomial coefficients. Again, for the Microsoft series, the BIC (and inspection of  $t$ -statistics on the Hermite coefficients) suggests no expansion beyond (26) is warranted.

## 7 The Estimation Step

The objectives are (i) to estimate  $\rho$ , (ii) test the hypothesis that the dynamic system under consideration generated the observed data  $\{\tilde{y}_t\}_{t=1}^n$ , and (iii) provide diagnostics that indicate how a rejected system should be modified to better describe the distribution of the observable process  $\{y_t\}$ .

We presume that the data have been summarized in the projection step, as described in Section 6, and that a moment function of the form

$$\tilde{\psi}_f(x, y) = \frac{\partial}{\partial \theta} \log f(y | x, \tilde{\theta}_n),$$

and a weighting matrix

$$\tilde{I}_n = \frac{1}{n} \sum_{t=1}^n \left[ \frac{\partial}{\partial \theta} \log f(\tilde{y}_t | \tilde{x}_{t-1}, \tilde{\theta}_n) \right] \left[ \frac{\partial}{\partial \theta} \log f(\tilde{y}_t | \tilde{x}_{t-1}, \tilde{\theta}_n) \right]'$$

are available from the projection step. Here we have assumed that  $f(y | x, \tilde{\theta}_n)$  closely approximates  $p(y | x, \rho^o)$ . If not, the weighting matrix given by (13) must be used. If the SNP density  $f_K(y | x, \theta)$  is used as the auxiliary model with tuning parameters selected appropriately,  $\tilde{I}_n$  as computed above will be adequate (Gallant and Long, 1997).

Here, we shall indicate explicitly that the dependence on  $n$  of the moment equations  $m_n(\rho)$  enters through the quasi maximum likelihood estimate  $\tilde{\theta}_n$  by writing  $m(\rho, \tilde{\theta}_n)$  for  $m_n(\rho)$  where

$$m(\rho, \theta) = \mathcal{E}_\rho \frac{\partial}{\partial \theta} \log f(y | x, \theta).$$

Recall that the moment equations of the minimum chi-squared procedure discussed in Section 3 are computed by averaging over a long simulation

$$m(\rho, \tilde{\theta}_n) \doteq \frac{1}{N} \sum_{t=1}^N \frac{\partial}{\partial \theta} \log f(\hat{y}_t | \hat{x}_{t-1}, \tilde{\theta}_n).$$

The EMM estimator is

$$\hat{\rho}_n = \underset{\rho \in R}{\operatorname{argmin}} m'(\rho, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\rho, \tilde{\theta}_n)$$

The asymptotics of the estimator are as follows. If  $\rho^o$  denotes the true value of  $\rho$  and  $\theta^o$  is an isolated solution of the moment equations  $m(\rho^o, \theta) = 0$ , then under regularity conditions (Gallant and Tauchen, 1996; Gallant and Long, 1997),

$$\begin{aligned} \lim_{n \rightarrow \infty} \hat{\rho}_n &= \rho^o \quad \text{a.s.} \\ \sqrt{n}(\hat{\rho}_n - \rho^o) &\xrightarrow{\mathcal{L}} N\left\{0, [(M^o)'(\mathcal{I}^o)^{-1}(M^o)]^{-1}\right\} \\ \lim_{n \rightarrow \infty} \hat{M}_n &= M^o \quad \text{a.s.} \\ \lim_{n \rightarrow \infty} \tilde{\mathcal{I}}_n &= \mathcal{I}^o \quad \text{a.s.} \end{aligned} \tag{27}$$

where  $\hat{M}_n = M(\hat{\rho}_n, \tilde{\theta}_n)$ ,  $M^o = M(\rho^o, \theta^o)$ ,  $M(\rho, \theta) = (\partial/\partial\rho')m(\rho, \theta)$ , and

$$\mathcal{I}^o = \mathcal{E}_{\rho^o} \left[ \frac{\partial}{\partial\theta} \log f(y_0 | x_{-1}, \theta^o) \right] \left[ \frac{\partial}{\partial\theta} \log f(y_0 | x_{-1}, \theta^o) \right]'$$

Under the null hypothesis that  $p(y_{-L}, \dots, y_0 | \rho)$  is the correct model,

$$L_0 = n m'(\hat{\rho}_n, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\hat{\rho}_n, \tilde{\theta}_n) \tag{28}$$

is asymptotically chi-squared on  $p_\theta - p_\rho$  degrees of freedom. Under the null hypothesis that  $h(\rho^o) = 0$ , where  $h$  maps  $R$  into  $\Re^q$ ,

$$L_h = n \left[ m'(\hat{\rho}_n, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\hat{\rho}_n, \tilde{\theta}_n) - m'(\hat{\rho}_n, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\hat{\rho}_n, \tilde{\theta}_n) \right] \tag{29}$$

is asymptotically chi-squared on  $q$  degrees of freedom where

$$\hat{\rho}_n = \underset{h(\rho)=0}{\operatorname{argmin}} m'(\rho, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\rho, \tilde{\theta}_n).$$

A Wald confidence interval on an element  $\rho_i$  of  $\rho$  can be constructed in the usual way from an asymptotic standard error  $\sqrt{\hat{\sigma}_{ii}}$ . A standard error may be obtained by computing the Jacobian  $M_n(\rho, \theta)$  numerically and taking the estimated asymptotic variance  $\hat{\sigma}_{ii}$  to be the  $i$ th diagonal element of  $\hat{\Sigma} = (1/n)[(\hat{M}_n)'(\tilde{\mathcal{I}}_n)^{-1}(\hat{M}_n)]^{-1}$ . These intervals, which are symmetric, are somewhat misleading because they do not reflect the rapid increase in the EMM objective function  $s_n(\rho) = m'(\rho, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\rho, \tilde{\theta}_n)$  when  $\rho_i$  approaches a value for which the system

under consideration is explosive. Confidence intervals obtained by inverting the criterion difference test  $L_h$  do reflect this phenomenon and are therefore more useful. To invert the test one puts in the interval those  $\rho_i^*$  for which  $L_h$  for the hypothesis  $\rho_i^o = \rho_i^*$  is less than the critical point of a chi-squared on one degree of freedom. To avoid re-optimization one may use the approximation

$$\hat{\rho}_n = \hat{\rho}_n + \frac{\rho_i^* - \hat{\rho}_{in}}{\hat{\sigma}_{ii}} \hat{\Sigma}_{(i)}$$

in the formula for  $L_h$  where  $\hat{\Sigma}_{(i)}$  is the  $i$ -th column of  $\hat{\Sigma}$ .

The above remarks should only be taken to imply that confidence intervals obtained by inverting the criterion difference test have more desirable structural characteristics than those obtained by inverting the Wald test and not that they have more accurate coverage probabilities.

When  $L_0$  exceeds the chi-squared critical point, diagnostics that suggest improvements to the system are desirable. Because

$$\sqrt{n} m(\hat{\rho}_n, \tilde{\theta}_n) \xrightarrow{\mathcal{L}} N\{0, \mathcal{I}^o - (M^o)'[(M^o)'(\mathcal{I}^o)^{-1}(M^o)]^{-1}(M^o)'\},$$

inspection of the  $t$ -ratios

$$T_n = S_n^{-1} \sqrt{n} m(\hat{\rho}_n, \tilde{\theta}_n), \quad (30)$$

where  $S_n = \left( \text{diag}\{\tilde{\mathcal{I}}_n - (\hat{M}_n)'[(\hat{M}_n)'(\tilde{\mathcal{I}}_n)^{-1}(\hat{M}_n)]^{-1}(\hat{M}_n)'\} \right)^{1/2}$ , can suggest reasons for failure. Different elements of the score correspond to different characteristics of the data and large  $t$ -ratios reveal those characteristics that are not well approximated.

## 8 Application: Stock Price Diffusion

We illustrate EMM estimation by estimating the stochastic differential equation (1) using the Microsoft data series shown in Figure 1. We use the SNP model determined in Section 5 as the auxiliary model. This auxiliary model has 11 parameters and therefore determines 11 moment conditions.

A special case of (1), with only one stochastic volatility factor, is

$$\begin{aligned} dU_{1t} &= \alpha_{10}dt + \exp(\beta_{10} + \beta_{12}U_{2t}) dW_{1t} & 0 \leq t < \infty \\ dU_{2t} &= \alpha_{22}U_{2t}dt + dW_{2t} \end{aligned} \quad (31)$$

where normalizations have been imposed for identification. We term (31) the SV1 model. The most general case is

$$\begin{aligned} dU_{1t} &= \alpha_{10}dt + \exp(\beta_{10} + \beta_{12}U_{2t} + \beta_{13}U_{3t}) dW_{1t} \\ dU_{2t} &= \alpha_{22}U_{2t}dt + dW_{2t} \\ dU_{3t} &= \alpha_{33}U_{3t}dt + dW_{3t} \end{aligned} \tag{32}$$

We term (32) the SV2 model.

Table 3 shows the EMM objective function values for these two models for various simulation lengths. Evidently, the SV1 model is rejected at the 5 percent level and is borderline at the one percent level. The SV2 model is acceptable at the 5 percent level. The inferences are reasonably robust with respect simulation size, and in what follows all results are report for  $N = 100,000$ .

Table 4 displays parameter estimates, asymptotic standard errors, and 95 percent criterion difference confidence intervals. For SV1, the point estimates reveal moderate mean reversion in the single stochastic volatility. For SV2, on the other hand, the estimates of the stochastic volatility model cleanly separate into two distinct factors: a very persistent factor,  $U_{2t}$ , which displays very little mean reversion, and a very strongly mean-reverting factor,  $U_{3t}$ . Note that a conventional Wald confidence interval  $\hat{\alpha}_{22} \pm \hat{\sigma}_{\alpha}z$  for  $\alpha_{22}$  would include zero, which represents Brownian motion with drift, and extends well into the unstable region  $\alpha_{22} > 0$ . On the other hand, the criterion-difference confidence is very asymmetric and lies almost exclusively in the stable region  $\alpha_{22} < 0$ . The asymmetry of the confidence interval reflects the asymmetry of the objective function due to the explosive behavior of simulations from unstable models (Tauchen, 1998).

Table 5 shows the diagnostic  $t$ -ratios for SV1 and SV2. The SV1 model fails to fit the scores corresponding to the GARCH scale (21) function and the even-numbered parameters of the Hermite polynomial (18). The reason is that the data, and thereby the score, display persistent stochastic volatility and strong conditional non-Gaussian thick tails. A single factor stochastic volatility model cannot simultaneously capture both features of the data. In contrast, the SV2 model adequately fits all of the scores, and thus appears to adequately fit the data.

Figure 3 shows QQ-plots of unconditional quantiles of the data versus those of the SV1 and SV2 models. The model quantiles are determined by simulation. In the top panel, the SV1 model is seen to fail the capture the tail properties of the data. In the bottom panel, the SV2 model captures nearly all aspects of the data.

---

Table 3 about here

---



---

Table 4 about here

---



---

Table 5 about here

---



---

Figure 3 about here

---

## 9 The Reprojection Step

From the observed data set  $\{\tilde{y}_t\}_1^n$  we have generated via EMM the parameter estimate  $\hat{\rho}$  for each model under consideration. We now proceed backwards to infer the unobserved state vector from the observed process as implied by a particular model. The approach follows the reprojection method proposed by Gallant and Tauchen (1998), which is a numerically intensive, simulation-based, nonlinear Kalman filtering technique.

The idea is relatively straightforward. As a by-product of the estimation, we have a long simulated realization of the state vector  $\{\hat{U}_t\}_{t=1}^N$  and the corresponding observables  $\{\hat{y}_t\}_{t=1}^N$  for  $\rho = \hat{\rho}$ . Working within the simulation, we can calibrate the functional form of

the conditional distribution of functions of  $\hat{U}_t$  given  $\{\hat{y}_\tau\}_{\tau=1}^t$ . Given the calibrated functions determined within the simulation, we simply evaluate them on the observed data. More generally, we can determine within the simulation the conditional distribution of functions of  $\hat{U}_t$  given  $\{\hat{y}_\tau\}_{\tau=1}^t$  and then evaluate the result on observed data  $\{\tilde{y}_t\}_{t=1}^n$ .

We illustrate with the conditional mean functions of the volatility factors of the SV1 and SV2 models estimated in the previous section. Our targets are

$$\mathcal{E}(U_{it}|\{y_\tau\}_{\tau=1}^t), \quad i = 2, 3 \quad (33)$$

To begin, we generate simulations  $\{\hat{U}_t\}_{t=1}^N, \{\hat{y}_t\}_{t=1}^N$ , at the estimated  $\tilde{\rho}$  and  $N = 100,000$ . Keep in mind that, in order to generate via filtering  $y_t$  predictions of  $U_{2t}$  and  $U_{3t}$ , we are allowed to use very general functions of  $\{y_\tau\}_{\tau=1}^t$  and that we have a huge data set work with. After some experimentation, we found the following strategy, which seems to work quite well. We estimate an SNP-GARCH model on the  $\hat{y}_t$  because the SNP-GARCH model provides a convenient representation of the one-step ahead conditional variance  $\hat{\sigma}_t^2$  of  $\hat{y}_{t+1}$  given  $\{\hat{y}_\tau\}_{\tau=1}^t$ . We then run regressions of  $\hat{U}_{it}$  on  $\hat{\sigma}_t^2$ ,  $\hat{y}_t$ , and  $|\hat{y}_\tau|$  and lags of these series, with lag lengths generously long. (Keep in mind the huge size of the simulated data set; these regressions are essentially analytic projections.) At this point we have calibrated, inside the simulations, functions that give predicted values of  $U_{2t}$  and  $U_{3t}$  given  $\{y_\tau\}_{\tau=1}^t$ . Lastly, we evaluate these functions on the observed data series  $\{\tilde{y}_\tau\}_{\tau=1}^t$ , which gives reprojected values  $\tilde{U}_{2t}$  and  $\tilde{U}_{3t}$  for the volatility factors at the data points.

Figure 4 shows plots of the raw returns series and the reprojected volatility factors for the SV1 and SV2 models determined as just described. As to be expected, for SV2,  $\tilde{U}_{2t}$  is slowly moving while  $\tilde{U}_{3t}$  is quite choppy. Interestingly, the crash of 1987 is attributed to a large realization of the strongly mean reverting factor,  $U_3$ . This result suggest that the volatility increase surround the 87 crash was rather temporary, which appears consistent with the plot of the raw data in the top panel. Finally, the reprojected volatility factor from the SV1 model lies between those of the SV2 model and seems to miss much of the crash of 1987, which reflects further in the shortcomings of single-factor stochastic volatility models.

Figure 4 about here

---

## 10 Conclusion

In this paper, efficient method of moments (EMM) has been described and illustrated. The main references for the procedure itself are Gallant and Tauchen (1996), Gallant and Long (1997), Tauchen (1997), and Gallant and Tauchen (1998). Efficiency considerations were discussed. Main references for this material are Gallant and Tauchen (1996) and Gallant and Tauchen (1999).

The procedure consists of three steps: Projection, Estimation, and Reprojection. The first two steps must be carried out. The last step, Reprojection, is optional and may need to be modified to suit the application.

The Projection step summarizes the data into two statistics: the quasi maximum likelihood estimator of the parameters of an auxiliary model estimated from the data and an estimate of its variance.

The Estimation step provides estimates of the parameters of the model under consideration. Estimation is accomplished by using a minimum chi-squared criterion defined by the expected score of the auxiliary model determined in the projection step. The chi-squared criterion provides an automatic test of model adequacy. The estimating equations evaluated at the estimated parameter also provide diagnostics.

The Reprojection step is accomplished by generating a long simulation of observables from the system and projecting it onto a nonparametric density estimator. The nonparametric density estimator provides a representation of the transition density of the data subject to the restrictions implied by the model. The SNP density is well suited to this purpose because of the ease with which conditional expectations can be computed. More generally, because state variables together with the corresponding observables can be simulated, reprojections of unobserved state variables onto observables provide filters for extracting unobserved state variables from the observed data. This filtering can be accomplished by projection onto a nonparametric density estimator as in Gallant, Hsu, and Tauchen (1999) or by projection on some reasonable representation of the relationship of interest as in the Microsoft example.

We close with remarks about statistical efficiency and model adequacy. EMM uses the score function of an auxiliary model that describes the data to define the estimating equation for minimum chi-squared estimation. This choice is motivated by efficiency considerations. However, through a series of applications, we and other users have learned that the main benefit of EMM is not so much efficiency *per se*, but rather the diagnostic information on model adequacy that is produced. EMM forces a model to confront a set of comprehensive set of data-determined moment conditions that are valid by construction. This confrontation provides considerable insight into the model's adequacy. For example, in the empirical application in Section 8, the EMM methodology revealed the shortcomings of a single-factor stochastic volatility model as a description of the dynamics of the Microsoft stock price series. These shortcoming might never have been revealed by a direct likelihood-based or Bayesian approach that treated the single-factor stochastic volatility model as correct.

## 11 References

- Ahn, D-H., R. F. Dittmar, and A. R. Gallant (2001), "Quadratic Term Structure Models: Theory and Evidence," *The Review of Financial Studies*, forthcoming.
- Andersen, T. G., and J. Lund (1997), "Estimating Continuous Time Stochastic Volatility Models of the Short Term Interest Rate," *Journal of Econometrics* 77, 343–378.
- Andersen, T. G., L. Benzoni, and J. Lund (2001), "Towards an Empirical Foundation for Continuous-Time Equity Return Models," Manuscript, Kellogg Graduate School of Management, Northwestern University, Evanston, IL.
- Andersen, T. G., H-J. Chung, and B. E. Sorensen (1999), Efficient Method of Moments Estimation of a Stochastic Volatility Model: A Monte Carlo Study, *Journal of Econometrics* 91,61–87.
- Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics* 31, 307–327.
- Brandt, M. W., and P. Santa-Clara (1999), "Simulated Likelihood Estimation of Multivariate Diffusions with an Application to Interest Rates and Exchange Rates with Stochas-

- tic Volatility,” Manuscript, Anderson Graduate School of Management, University of California, Los Angeles.
- Chernov, M., A. R. Gallant, E. Ghysels, G. Tauchen (2001), “Alternative Models for Stock Price Dynamics,” Manuscript, Business School, Columbia University, New York.
- Chumacero, R. (1997), “Finite Sample Properties of the Efficient Method of Moments,” *Studies in Nonlinear Dynamics and Econometrics* 2, 35–51.
- Chung, C-C., and G. Tauchen (2001), “Testing Target Zone Models Using Efficient Method of Moments,” *Journal of Business and Economic Statistics*, forthcoming.
- Coppejans, Mark, and A. Ronald Gallant (2000), “Cross Validated SNP Density Estimates,” Manuscript, Department of Economics, University of North Carolina, Chapel Hill NC.
- Dai, Q., and K. J. Singleton (2001), “Specification Analysis of Affine Term Structure Models,” *Journal of Finance*, forthcoming.
- Davidian, M. and R. J. Carroll (1987), “Variance Function Estimation,” *Journal of the American Statistical Association* 82, 1079–1091.
- Duffie, D., and K. J. Singleton (1993), “Simulated Moments Estimation of Markov Models of Asset Prices,” *Econometrica* 61, 929–952.
- Durham, G. B., and A. R. Gallant (2002), “Numerical Techniques for Maximum Likelihood Estimation of Continuous Time Diffusion Processes,” *Journal of Business and Economic Statistics*, forthcoming.
- Elerian, O., S. Chib, and N. Shephard (2001), “Likelihood Inference for Discretely Observed Nonlinear Diffusions,” *Econometrica*, forthcoming.
- Engle, R. F. (1982), “Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation,” *Econometrica* 50, 987–1007.
- Fisher, R. A. (1912), “On an Absolute Criterion for Fitting Frequency Curves,” *Messages of Mathematics* 41, 155–157.

- Gauss, C. F. (1816), “Bestimmung der Genauigkeit der Beobachtungen,” *Zeitschrift für Astronomie und verwandte Wissenschaften* 1, 185–196.
- Gallant, A. R. (1980), “Explicit Estimators of Parametric Functions in Nonlinear Regression,” *Journal of the American Statistical Association* 75, 182–93.
- Gallant, A. R. (1987), *Nonlinear Statistical Models*, New York: Wiley.
- Gallant, A. R., D. A. Hsieh, and G. E. Tauchen (1991), “On Fitting a Recalcitrant Series: The Pound/Dollar Exchange Rate, 1974–83,” in W. A. Barnett, J. Powell, G. E. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics, Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, Cambridge: Cambridge University Press, Chapter 8, 199–240.
- Gallant, A. R., C-T. Hsu, and G. Tauchen (1999) “Using Daily Range Data to Calibrate Volatility Diffusions and Extract the Forward Integrated Variance” *The Review of Economics and Statistics* 81(4), 617–631.
- Gallant, A. R., and J. Long (1997), “Estimating Stochastic Differential Equations Efficiently by Minimum Chi-squared,” *Biometrika* 84, 125–141.
- Gallant, A. R., and D. W. Nychka (1987), “Seminonparametric Maximum Likelihood Estimation,” *Econometrica* 55, 363–390.
- Gallant, A. R., and G. Tauchen (1989), “Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications,” *Econometrica* 57, 1091–1120.
- Gallant, A. R., and G. Tauchen (1996), “Which Moments to Match?,” *Econometric Theory* 12, 657–681.
- Gallant, A. R., and G. Tauchen (1998), “Reprojecting Partially Observed Systems with Application to Interest Rate Diffusions,” *Journal of the American Statistical Association* 93, 10–24.

- Gallant, A. R., and G. Tauchen (1999), “The Relative Efficiency of Method of Moments Estimators,” *Journal of Econometrics* 92, 149–172.
- Gennotte, G., and T. A. Marsh (1993), “Variations in Economic Uncertainty and Risk Premiums on Capital Assets,” *European Economic Review* 37, 1021–1041.
- Ghysels, E., A. Harvey, and E. Renault (1995), “Stochastic Volatility,” in G. S. Maddala, ed., *Handbook of Statistics, Vol. 14, Statistical Methods in Finance*, Amsterdam: North Holland.
- Gourieroux, C., A. Monfort, and E. Renault (1993), “Indirect Inference,” *Journal of Applied Econometrics* 8, S85–S118.
- Hansen, L. P. (1982), “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica* 50, 1029–1054.
- Heyde, C. C., and R. Morton (1998), “Multiple Roots in General Estimating Equations,” *Biometrika* 85, 949–953.
- Ingram, B. F., and B. S. Lee (1991), “Simulation Estimation of Time Series Models,” *Journal of Econometrics* 47, 197–250.
- Jiang, G.J. and van der Sluis, P.J. (2000), “Option Pricing with the Efficient Method of Moments ”, in Y. S. Abu-Mostafa, B. LeBaron, A. W. Lo, and A. S. Weigend, eds. *Computational Finance*, Cambridge MA: MIT Press.
- Karatzas, I., and S. E. Shreve (1991). *Brownian Motion and Stochastic Calculus*. 2nd Edition. Berlin: Springer-Verlag.
- Kloeden, P. E., and E. Platen (1992), *Numerical Solution of Stochastic Differential Equations*, Berlin: Springer-Verlag.
- Liu, M. (2000), “Modeling Long Memory in Stock Market Volatility,” *Journal of Econometrics* 99, 139–171.
- Lo, A. W. (1988), “Maximum Likelihood Estimation of Generalized Ito Process with Discretely Sampled Data,” *Econometric Theory* 4, 231–247.

- Mallet, A., F. Mentré, J-L. Steimer, and F. Lokiec (1988), “Nonparametric Maximum Likelihood Estimation for Population Pharmacokinetics, with Application to Cyclosporine,” *Journal of Pharmacokinetics and Biopharmaceutics* 16, 311–327.
- Marron, J. S., and M. P. Wand (1992), “Exact Mean Integrated Squared Error,” *The Annals of Statistics* 20, 712–736.
- Nagypal, E. (2001), “Learning-by-Doing versus Selection: Can We Tell Them Apart?,” Manuscript, Department of Economics, Stanford University.
- Nelson, D. (1991), “Conditional Heteroskedasticity in Asset Returns: A New Approach,” *Econometrica* 59, 347–370.
- Neyman, J., and E. S. Pearson (1928), “On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference,” *Biometrika* 20A, 175–240, 263–294.
- Ng, S., and A. Michaelides (2000), “Estimating the Rational Expectations model of Speculative Storage: a Monte Carlo Comparison of Three Simulation Estimators,” *Journal of Econometrics* 96, 231–266.
- Olsen, L. F., and W. M. Schaffer (1990), “Chaos Versus Noisy Periodicity: Alternative Hypotheses for Childhood Epidemics,” *Science* 249, 499–504.
- Pearson, K. (1894), “Contributions to the Mathematical Theory of Evolution,” *Philosophical Transactions of the Royal Society of London, Series A*, 185, 71–78.
- Schwarz, G. (1978), “Estimating the Dimension of a Model,” *Annals of Statistics* 6, 461–464.
- Smith, A. A. (1993), “Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions,” *Journal of Applied Econometrics* 8, S63–S84.
- Tauchen, G. (1997), “New Minimum Chi-square Methods in Empirical Finance,” in D. Kreps, and K. Wallis, eds., *Advances in Econometrics, Seventh World Congress*, Cambridge UK: Cambridge University Press, 279–317.

- Tauchen, G. (1998), “The Objective Function of Simulation Estimators Near the Boundary of the Unstable Region of the Parameter Space,” *Review of Economics and Statistics* 80, 389–398.
- Valderrama, Diego (2001), “Can a Standard Real Business Cycle Model Explain the Non-linearities in U.S. National Accounts Data?” Ph.D. thesis document, Department of Economics, Duke University.
- van der Sluis, P. J. (1997) “EmmPack 1.01: C/C++ Code for Use with Ox for Estimation of Univariate Stochastic Volatility Models with the Efficient Method of Moments,” *Nonlinear Dynamics and Econometrics* 2, 77–94
- van der Sluis, P. J. (1999) “Estimation and Inference with the Efficient Method of Moments: with Application to Stochastic Volatility Models and Option Pricing, Manuscript, Tinbergen Institute, Amsterdam, Research Series Paper No. 204.

## Tables and Figures

Table 1. Sample Statistics: Percent Return on Microsoft,  
March 14, 1986 – November 16, 2000

	Daily	Annualized
Mean:	0.158	39.868
Std Dev:	2.502	39.720
Variance	6.260	1577.641
n:	3711	
Min:	-35.828	-9028.656
Q01:	-6.243	-1573.337
Q05:	-3.407	-858.504
Q10:	-2.598	-654.583
Q50:	0.000	0.000
Q90:	3.010	758.507
Q95:	3.961	998.150
Q99:	6.255	1576.271
Max:	17.869	4503.033

Table 2. Restrictions Implied by Settings of the Tuning Parameters.

Parameter setting	Characterization of $\{y_t\}$
$L_u = 0, L_g = 0, L_r = 0, L_p \geq 0, K_z = 0, K_x = 0$	iid Gaussian
$L_u > 0, L_g = 0, L_r = 0, L_p \geq 0, K_z = 0, K_x = 0$	Gaussian VAR
$L_u > 0, L_g = 0, L_r = 0, L_p \geq 0, K_z > 0, K_x = 0$	semiparametric VAR
$L_u \geq 0, L_g = 0, L_r > 0, L_p \geq 0, K_z = 0, K_x = 0$	Gaussian ARCH
$L_u \geq 0, L_g = 0, L_r > 0, L_p \geq 0, K_z > 0, K_x = 0$	semiparametric ARCH
$L_u \geq 0, L_g > 0, L_r > 0, L_p \geq 0, K_z = 0, K_x = 0$	Gaussian GARCH
$L_u \geq 0, L_g > 0, L_r > 0, L_p \geq 0, K_z > 0, K_x = 0$	semiparametric GARCH
$L_u \geq 0, L_g \geq 0, L_r \geq 0, L_p > 0, K_z > 0, K_x > 0$	nonlinear nonparametric

Notes:  $L_u$  is the lag length of the location function.  $L_g$  is the lag length of the GARCH (autoregressive) part of the scale function.  $L_r$  is the lag length of the ARCH (moving average) part of the scale function.  $L_p$  is the lag length of the polynomials in  $x$  that determine the coefficients of the Hermite expansion of the innovation density.  $K_z$  is the degree of the Hermite expansion of the innovation density.  $K_x$  is the degree of polynomials in  $x$  that determine the coefficients of the Hermite expansion of the innovation density.

Table 3. Model Definitions and Minimized Chi-Squared Criterion.

Specification	$\alpha_{10}$	$\alpha_{22}$	$\alpha_{33}$	$\beta_{10}$	$\beta_{12}$	$\beta_{13}$	$N$	$\chi^2(\hat{\rho})$	df	p-value
SV1	*	*		*	*		50k	18.348	7	0.010
SV1	*	*		*	*		75k	17.888	7	0.012
SV1	*	*		*	*		100k	18.189	7	0.011
SV2	*	*	*	*	*	*	50k	9.978	5	0.076
SV2	*	*	*	*	*	*	75k	9.254	5	0.099
SV2	*	*	*	*	*	*	100k	8.642	5	0.124

Notes: \* denotes a free parameter. 100k denotes a simulation of length  $N = 100,000$  simulated simulated at step size  $\Delta = 1/6048$ , corresponding to 24 steps per day and 252 trading days per year.

Table 4. Parameter Estimates, Standard Errors, and Criterion-Difference Confidence Intervals

Specification	$\alpha_{10}$	$\alpha_{22}$	$\alpha_{33}$	$\beta_{10}$	$\beta_{12}$	$\beta_{13}$
SV1						
Estimate	0.4215	-12.3711		-1.1441	1.4656	
Std. Err.	0.0638	2.2803		0.0384	0.0566	
95% lower	0.2956	-16.5289		-1.2171	1.3592	
95% upper	0.5498	-8.3963		-1.0718	1.5726	
SV2						
Estimate	0.4247	-0.000861	-102.9206	-0.6759	0.0371	5.0979
Std. Err.	0.0874	0.001633	21.2190	0.1489	0.0027	0.2444
95% lower	0.3517	-0.0015322	-114.6309	-0.7402	0.0346	4.9954
95% upper	0.5060	0.0000190	-94.2487	-0.5868	0.0396	5.2315

Table 5. Diagnostic  $t$ -Statistics.

Coefficient		SV1	SV2
Location Function: $b_0$	psi( 1)	0.424	-1.120
$b_1$	psi( 2)	1.244	1.420
Scale Function:			
$\tau_0$	tau( 1)	2.564	0.503
$\tau_{gx}$	tau( 2)	1.956	1.024
$\tau_{gx}$	tau( 3)	2.318	0.722
Hermite Polynomial:			
$a_{0,1}$	A( 2)	0.381	0.563
$a_{0,2}$	A( 3)	3.390	-1.013
$a_{0,3}$	A( 4)	0.456	0.660
$a_{0,4}$	A( 5)	3.570	-0.734
$a_{0,5}$	A( 6)	0.203	0.639
$a_{0,6}$	A( 7)	3.099	-0.830

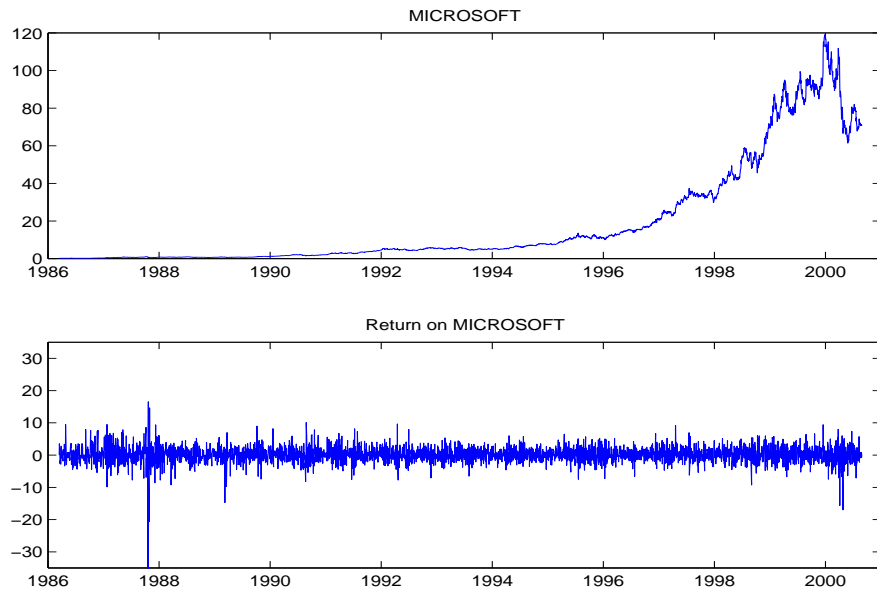
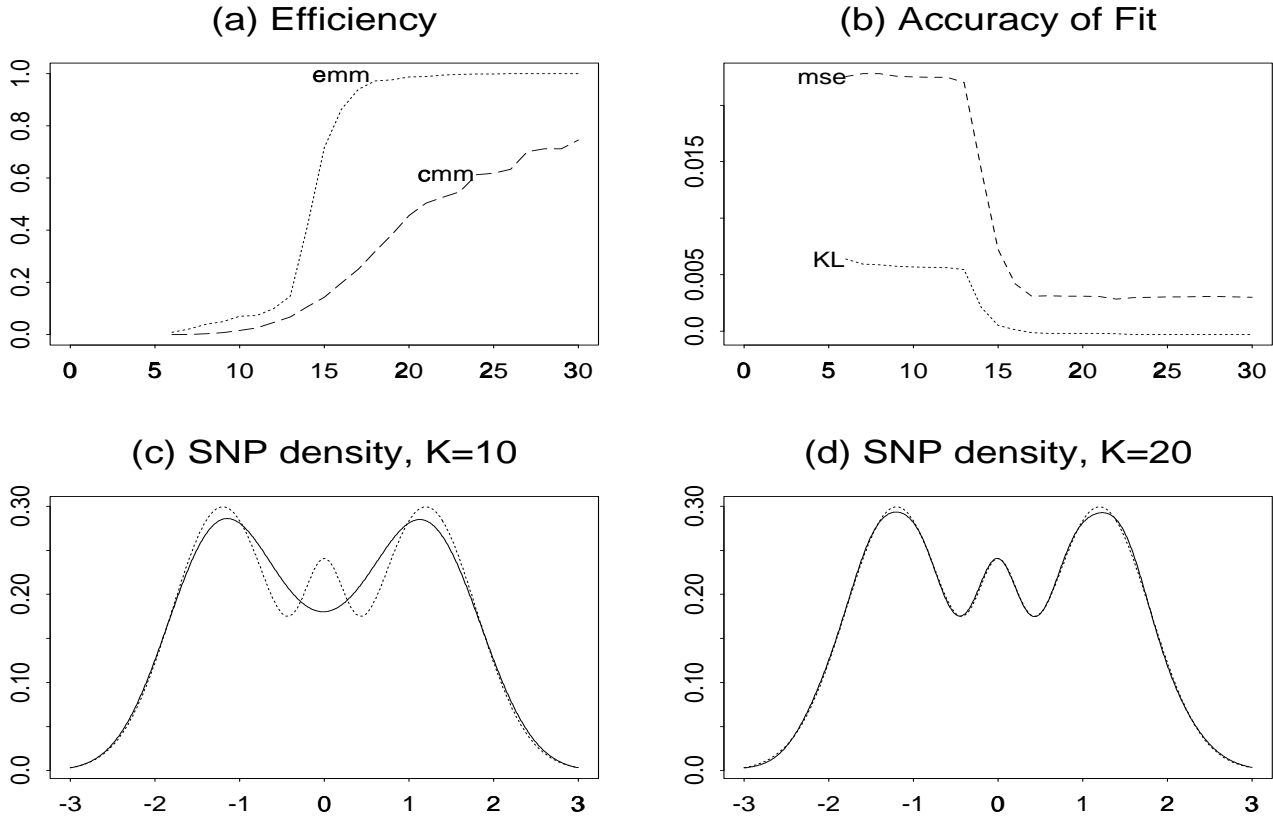


Figure 1. Microsoft, March 13, 1986 – November 16, 2000. The top panel shows the raw (adjusted for splits) price series  $P_t$ . The bottom panel shows the geometric return  $100 * [\log(P_t) - \log(P_{t-1})]$  expressed in daily percent form.



**Figure 2. Relative Efficiency for the Trimodal Density.** Panel (a) plots the relative efficiency of the EMM and CMM estimators against degree  $K$ , for the trimodal density of the Marron-Wand test suite. As seen, the efficiency of the EMM estimator increases rapidly when the degree  $K$  of the SNP auxiliary model is between 10 and 20. Panel (b) plots the root mean squared error and Kullback-Leibler divergence of the SNP approximation to the trimodal density against  $K$ , labeled mse and KL, respectively. As seen, the region  $10 \leq K \leq 20$  is the region where the error in the SNP approximation to the trimodal density decreases rapidly. Panel (c) plots the SNP approximation at  $K = 10$ , shown as a solid line, to the trimodal density, shown as a dotted line. As seen, at  $K = 10$  the SNP density approximates a trimodal density by a bimodal density. Panel (d) is the same at  $K = 20$ . As seen, at  $K = 20$  the SNP density has correctly determined the number of modes.

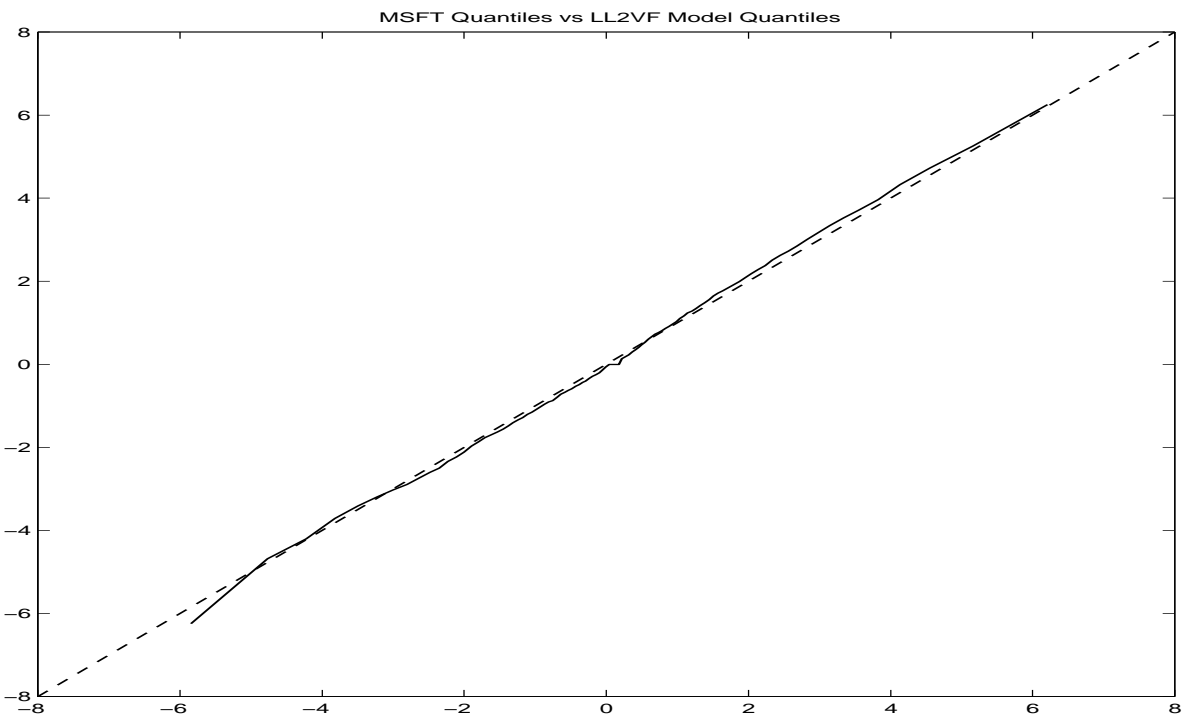
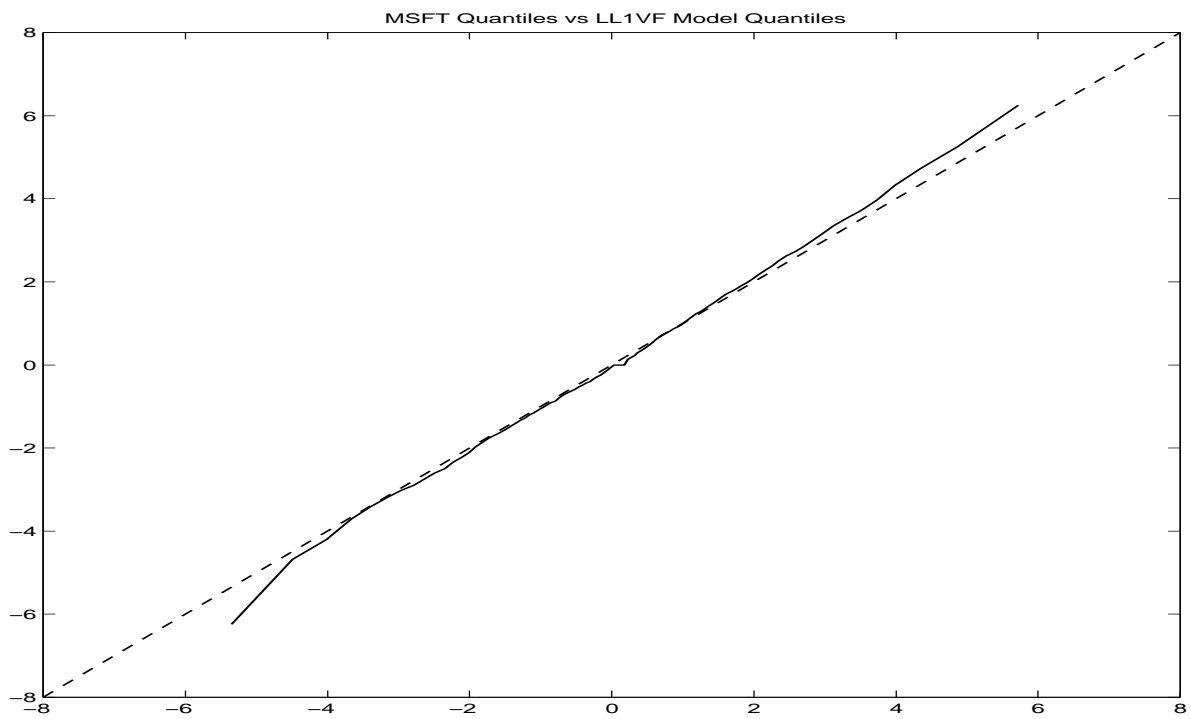


Figure 3. QQ Plots: Top panel is a QQ plot of the quantiles of the data versus those of the SV1 model; bottom panel is a plot of quantiles of the data versus those of the SV2 model.

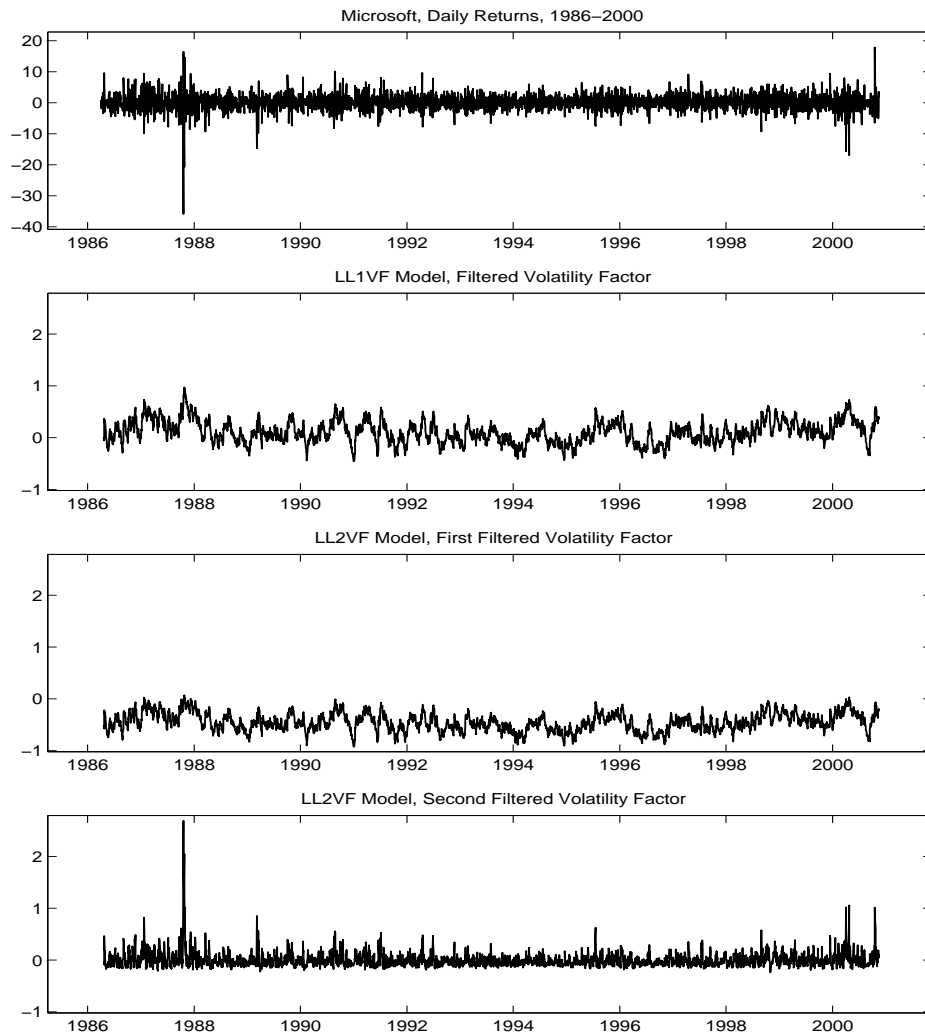


Figure 4. Reprojected (extracted) Volatility Factors, 1986–2000. The top panel is the raw return series; the second panel is the reprojected volatility factor from the SV1 model; the third and fourth panels are reprojected volatility factors from the SV2 model.