

Simulated Score Methods and Indirect Inference for Continuous-time Models

Initial Draft: August 2001
This Draft: December 2007

A. Ronald Gallant^a George Tauchen^b

^a*Duke University*

^b*Duke University*

Abstract

We describe a simulated method of moments estimator that is implemented by choosing the vector valued moment function to be the expectation under the structural model of the score function of an auxiliary model, where the parameters of the auxiliary model are eliminated by replacing them with their quasi-maximum likelihood estimates. This leaves a moment vector depending only the parameters of the structural model. Structural parameter estimates are those parameter values that put the moment vector as closely to zero as possible in a suitable GMM metric. This methodology can also be interpreted as a practical computational strategy for implementing indirect inference. We argue that considerations from statistical science dictate that the auxiliary model should approximate the true data generating process as closely as possible and show that using the SNP model is one means to this end. When the view of close approximation is accepted in implementation, the methodology described here is usually referred to as Efficient Method of Moments (EMM) in the literature because (i) the estimator is asymptotically as efficient as maximum likelihood under correct specification, and (ii) detection of model error is assured under incorrect specification. There are alternative views toward the desirability of close approximation to the data, which we discuss.

Key words: Efficient method of moments, Indirect inference, Simulated method of moments.

1 Introduction and Overview

In both empirical work (Bansal, Gallant, Hussey, and Tauchen, 1993, 1995) and in theoretical work (Gallant and Tauchen, 1996; Gallant and Long, 1997)

we have developed a systematic strategy for choosing the moments for Generalized Method of Moments (GMM) estimation of a structural model. The idea is relatively straightforward: Use the expectation with respect to the structural model of the score function of an auxiliary model as the vector of moment conditions for GMM estimation.

The score function is the derivative of the logarithm of the density of the auxiliary model with respect to the parameters of the auxiliary model. The moment conditions obtained by taking the expectation of the score depend directly upon the parameters of the auxiliary model and indirectly upon the parameters of the structural model through the dependence of expectation operator on the parameters of the structural model. The parameters of the auxiliary model are eliminated from the moment conditions by replacing them with their quasi-maximum likelihood estimates, which are obtained by maximizing the likelihood of the auxiliary model. This leaves a random vector of moment conditions that depends only on the parameters of the structural model; the randomness is due to the random fluctuations of the quasi-maximum likelihood estimates of the parameters of the auxiliary model. When this vector of moment conditions is evaluated at the true values of the structural parameters, it tends to zero as sample size increases, presuming that the structural model is correctly specified. The parameters of the structural model may therefore be estimated by minimizing the magnitude of the vector of moment conditions as measured by the appropriate GMM metric.

This estimation method, which is the main topic of this chapter, is particularly useful in a simulation-based estimation context, where the structural model is readily simulated but the likelihood function of the structural model is intractable. This context applies to many continuous-time estimation problems. In implementation, the expectation of the score with respect to the structural model is computed by simulating the structural model and averaging the score function over the simulations. When the structural model is strictly stationary, one may average the scores over a single very long simulated realization (Case 2 of Gallant and Tauchen, 1996), while in the presence of exogenous covariates, one simulates and averages the scores at each data point conditional on the covariates and then sums across data points (Case 3 of Gallant and Tauchen, 1996).

The estimator is closely related to the indirect inference estimator, due to Smith (1990, 1993) and Gourieroux, Monfort, and Renault (1993). The referee takes a more extreme view and insists that all statistical methods that use an auxiliary model as an adjunct must be called indirect inference methods. However, the way the literature developed we think that if one refers to the score based method as efficient method of moments and those that use a binding function as indirect inference methods, one has less of a chance of being misunderstood. However, ideas matter and names do not. Readers can call

these methods whatever they please. The main difference between score based methods and those that use a binding function is computational. Score based methods are computationally tractable. Methods that use a binding function can be a computational nightmare. The innovative computational methods proposed by Chernozukov and Hong (2003) appear to have further increased this computational advantage. We discuss the Chernozukov and Hong method in Section 7.

We illustrate this point regarding computations in Subsection 2.1.2 using what is sometimes called the Wald variant of the indirect inference estimator. It is a minimum distance estimator that entails minimizing, in a suitable metric, the difference between the parameters of the auxiliary model obtained by quasi-maximum likelihood and those predicted by the structural model. The predicted parameter values are given by the binding function, which in practice is computed by re-estimating the auxiliary model on simulations from the structural model. The binding function computation is trivial for linear auxiliary models, as initially suggested by Smith (1990, 1993), but is very demanding and possibly infeasible for more complicated nonlinear auxiliary models. The score-based approach discussed here circumvents the need to evaluate the binding function, which is why it is more computationally tractable. Nonetheless, for any given auxiliary model, the score-based estimator and indirect inference have the same asymptotic distribution. Thus, as just mentioned, some interpret and view the score-based estimator as a practical way to implement indirect inference in a simulation-based context. Either way, there are strong parallels to the classical simultaneous equations literature, with the auxiliary model playing the role of the reduced form and we recognize that there may be different, but asymptotically equivalent ways to work back from the reduced form parameter estimates to obtain structural parameter estimates.

The practical implication of working from the score function is that the auxiliary model only needs to be estimated once, namely on the observed data. This added flexibility makes it possible to implement the score-based estimator using either very simple, or very complicated and sophisticated auxiliary models as discussed in (Hansen, 2001). Complicated auxiliary models would be appropriate if the observed data exhibit important nonlinearities, and the researcher wants the structural model to confront these nonlinearities. Regardless of the score generator actually used, the estimator is consistent and asymptotically normal, subject only to mild identification conditions. Thus, there is potentially great latitude for choosing the auxiliary model.

We have consistently argued for resolving this choice by making the auxiliary model be a good statistical description of the data. That is, it should be a bona fide reduced form model. As we shall see, by doing so the researcher can ensure that the estimator can achieve the full efficiency of maximum likelihood estimation if the structural model is correct. Furthermore, and more impor-

tant, it assures the researcher of detecting misspecification if the structural model is wrong. In view of these capabilities, we ascribe the term Efficient Method of Moments (EMM) to the estimator.

There are three basic steps to EMM. The first, termed the Projection Step, entails summarizing the data by projecting it onto the reduced form auxiliary model, which we frequently term the score generator. If one knows of a good statistical model for the data, then it should be used in the projection step. That is rarely the case, however, and we have proposed the SNP models of Gallant and Tauchen (1989) as a general purpose score generator. The second step is the Estimation Step, where the parameters are obtained by GMM (minimum chi-squared) using an appropriate weighting matrix. If, in the projection step, care is taken to obtain a good auxiliary model, then the weighting matrix takes a particularly simple form. The estimation step produces an omnibus test of specification along with useful diagnostic t statistics. The third step is termed the Reprojection Step, which entails post-estimation analysis of simulations for the purposes of prediction, filtering, and model assessment.

Section 2, immediately below, discusses and contrasts simulated score methods and indirect inference. Thereafter the discussion is combined with a focus on the EMM estimator. As will be seen, these estimators are so closely related that, after the preliminary discussion of the differences, a unified discussion under the projection, estimation, reprojection paradigm described above is warranted.

Section 3 gives general guidelines for selecting the auxiliary model for the projection step. Section 4 is formal analysis of the efficiency theory and develops the SNP model as a general purpose score generator. Section 5 gives an intuitive overview of reprojection followed by a more formal description of the theory underlying it. Section 6 reviews in detail two selected applications of EMM for estimation of continuous time models. Section 7 discusses software, practical issues, and some interesting capabilities using parallelization.

This chapter is focused on applications to continuous time processes. But one should be aware that indirect inference, efficient method of moments, and simulated method of moments methods have far greater applicability. They apply to cross sectional data, panel data, data with fixed covariates, and spatial data. For details see Gourieroux, Monfort and Renault (1993), Gallant and Tauchen (1996), Pagan (1999), and de Luna and Genton (2002).

2 Estimation and Model Evaluation

The simulated score estimation method was proposed and applied in Bansal, Gallant, Hussey, and Tauchen (1993) where it was used to estimate and evaluate a representative agent specification of a two-country general equilibrium model. The theory was developed in Gallant and Tauchen (1996) and extended to non-Markovian data with latent variables in Gallant and Long (1997).

Indirect inference was proposed and developed by Smith (1990, 1993) and Gourieroux, Monfort, and Renault (1993). These ideas overlap with the simulated method of moments estimators proposed by Ingram and Lee (1991), Duffie and Singleton (1993), McFadden (1989), and Pakes and Pollard (1989).

Here we shall sketch the main ideas of simulated score estimation and indirect inference in a few paragraphs at a modest technical level and then present a more detailed review of the efficient method of moments methodology.

2.1 Overview

2.1.1 Simulated Score Estimation

Suppose that $f(y_t|x_{t-1}, \theta)$ is a reduced form model for the observed data, where x_{t-1} is the state vector of the observable process at time $t - 1$ and y_t is the observable process. An example of such a reduced form model is a GARCH(1,1). If this reduced form model, which we shall call a score generator, is fitted by maximum likelihood to get an estimate $\tilde{\theta}_n$, then the average of the score over the data $\{\tilde{y}_t, \tilde{x}_{t-1}\}_{t=1}^n$ satisfies

$$\frac{1}{n} \sum_{t=1}^n \frac{\partial}{\partial \theta} \log f(\tilde{y}_t | \tilde{x}_{t-1}, \tilde{\theta}_n) = 0 \quad (1)$$

because equations (1) are the first order conditions of the optimization problem. Throughout, as in (1), we shall use a tilde to denote observed values and statistics computed from observed values.

Now suppose we have a structural model that we wish to estimate. We express the structural model as the transition density $p(y_t|x_{t-1}, \rho)$ where ρ is the parameter vector. In relatively simple models, $p(y_t|x_{t-1}, \rho)$ is available in a convenient closed-form expression, and one can estimate ρ directly by classical maximum likelihood. However, for more complicated nonlinear models, $p(y_t|x_{t-1}, \rho)$ is often not available and direct maximum likelihood is infeasible.

But at the same time, it can be relatively easy to simulate the structural model. That is, for each candidate value ρ one can generate a simulated trajectory on $\{\hat{y}_t\}_{t=1}^N$ and the corresponding lagged state vector $\{\hat{x}_{t-1}\}_{t=1}^N$. This situation, of course, is the basic setup of simulated method of moments (Ingram and Lee, 1991; Duffie and Singleton, 1993). It arises naturally in continuous-time models, because the implied discrete time density is rarely available in closed form (Lo, 1988), but continuous time models are often quite easy to simulate. The situation also arise in other areas of economics and finance as well as discussed in Tauchen (1997).

If the structural model is correct and the parameters ρ are set to their true values ρ^o , then there should not be much difference between the data $\{\tilde{y}_t\}_{t=1}^n$ and a simulation $\{\hat{y}_t\}_{t=1}^N$. Therefore, if the first order conditions (1) of the reduced form were computed by averaging over a simulation instead of the sample, viz.,

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

one would expect that

$$m(\rho^o, \tilde{\theta}_n) = 0,$$

at least approximately. This condition will hold exactly in the limit as N and n tend to infinity under the standard regularity conditions of quasi maximum likelihood. One can try to solve $m(\rho, \tilde{\theta}_n) = 0$ to get an estimate $\hat{\rho}_n$ of the parameter vector of the structural model. In most applications this cannot be done because the dimension of θ is larger than the dimension of ρ . To compensate for this, one estimates ρ by $\hat{\rho}_n$ that minimizes the GMM criterion

$$m'(\rho, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\rho, \tilde{\theta}_n)$$

with weighting matrix

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

This choice of weighting matrix presupposes that the score generator fits the data well. If not, then a more complicated weighting matrix, described below, should be considered. The estimator $\hat{\rho}_n$ is asymptotically normal.

If the structural model is correctly specified, then the statistic

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

has the chi-squared distribution on $\dim(\theta) - \dim(\rho)$ degrees freedom. This is the familiar test of overidentifying restrictions in GMM nomenclature and is used to test model adequacy. A chi-squared is asymptotically normally distributed as degrees freedom increase. Therefore, for ease of interpretation, the statistic L_0 is often redundantly reported as a z -statistic, as we do later in our tables.

The vector $m(\hat{\rho}_n, \tilde{\theta}_n)$ can be normalized by its standard error to get a vector of t -statistics. These t -statistics can be interpreted much as normalized regression residuals. They are often very informative but are subject to the same risk as the interpretation of regression residuals; namely, a failure to fit one characteristic of the data can show up not at the score of the parameters that describe that characteristic but elsewhere due to correlation (colinearity). Nonetheless, as with regression residuals, inspecting normalized $m(\hat{\rho}_n, \tilde{\theta}_n)$ is usually the most informative diagnostic available. To protect oneself from misinterpreting these t -statistics, one should confirm all conclusions by means of the test of model adequacy L_0 above.

If the score generator is a poor fit to the data or the chi-squared test of model adequacy L_0 is not passed, then the analysis must be viewed as a calibration exercise rather than classical statistical inference. One might, for instance, deliberately choose a score generator that represents only some characteristics of the data to study the ability of a structural model to represent only those characteristics. We do this below but, as Gallant and McCulloch (2009) illustrate, it should be done with care because it is quite possible to be seriously misled. One might also use a rejected model to price options, arguing that it is the best available even though it was rejected. The use of EMM for calibration is discussed in Gallant, Hsu, and Tauchen (1999).

The score generator can be viewed as a summary of the data. It is accomplished by, in effect, projecting the data onto a reduced form model and is therefore called the projection step of an EMM investigation. Extraction of structural parameters from the summary by minimizing the chi squared criterion is called the estimation step. In a later section, we shall describe a third step, reprojection, that often accompanies an EMM investigation.

2.1.2 Indirect Inference Estimation

There are variants on the indirect inference scheme, some of which we discuss at the end of this subsection. We first describe what is sometimes called the Wald variant. The score based method described in 2.1.1 is sometimes called the Lagrange multiplier variant in this vernacular.

The indirect inference estimator is based on the binding function, which is

defined as

$$b(\rho) = \operatorname{argmax}_{\theta \in \Theta} \iint \log f(y|x, \theta) p(y|x, \rho) dydx,$$

where $f(y|x, \theta)$ and $p(y|x, \rho)$ are the transition densities of the auxiliary model and structural model, respectively, as described above. The binding function can also be defined as the function that satisfies $m[\rho, b(\rho)] = 0$ (where, implicitly, $N = \infty$). According to the asymptotics of quasi maximum likelihood,

$$\sqrt{n}[\tilde{\theta}_n - b(\rho)] \xrightarrow{\mathcal{L}} N(0, \mathcal{J}^{-1} \mathcal{I} \mathcal{J}^{-1}),$$

where $\tilde{\theta}_n$ is as above and

$$\mathcal{I} = \iint \left\{ \frac{1}{\sqrt{n}} \sum_{t=1}^n \frac{\partial}{\partial \theta} \log f[y|x, b(\rho)] \right\}^2 p(y|x, \rho) dydx,$$

$$\mathcal{J} = \iint \left\{ \frac{\partial^2}{\partial \theta \partial \theta'} \log f[y|x, b(\rho)] \right\} p(y|x, \rho) dydx.$$

The matrix \mathcal{I} will likely have to be estimated by a HAC variance estimator as described below because, for reasons mentioned below, the auxiliary model is not apt to be a good approximation to the structural model in most indirect inference applications. A plug-in estimator can be employed to estimate \mathcal{J} ; numerical differentiation or Chernozukov-Hong (2003) method may be required to get the second derivatives. If the auxiliary model is an accurate approximation to the true data generating process, then $\mathcal{I} \doteq \mathcal{J}$ and one can compute whichever is more convenient. Without being specific as to the method employed, let $\tilde{\mathcal{I}}_n$ and $\tilde{\mathcal{J}}_n$ denote estimates of \mathcal{I} and \mathcal{J} .

The indirect inference estimator is

$$\hat{\rho}_n = \operatorname{argmin}_{\rho \in R} [\tilde{\theta}_n - b(\rho)]' (\tilde{\mathcal{J}}^{-1} \tilde{\mathcal{I}} \tilde{\mathcal{J}}^{-1})^{-1} [\tilde{\theta}_n - b(\rho)]$$

where R is the parameter space of the structural model.

Herein lies the computational difficulty with the indirect inference estimator: One must have an expression for $b(\rho)$ in order to compute the estimator. The expression

$$b(\rho) = \operatorname{argmax}_{\theta \in \Theta} \iint \log f(y|x, \theta) p(y|x, \rho) dydx,$$

can be computed numerically, with the integral computed by simulation as discussed above, and $b(\rho)$ computed by numerical optimization for given ρ . This embeds one numerical optimization, that for $b(\rho)$, inside another, that for $\hat{\rho}_n$, which poses two problems: The first is cost, the second is stability. That this computation will be costly is obvious. The stability issue is that a numerical optimizer can only compute the inner optimization, that for $b(\rho)$, to within a tolerance, at best. This will cause jitter which will cause the outer optimization problem to be non-smooth. Non-smooth optimization problems

are very difficult and costly to solve because good curvature information is not available. The Chernozukov-Hong method could lessen some of the problems caused by jitter but would, unfortunately, further increase cost. If the inner problem has local minima, the situation becomes nearly hopeless. For this reason, most practitioners convert a problem formulated as an indirect inference problem to simulated score estimation problem prior to computation so as to eliminate $b(\rho)$ and \mathcal{J} from consideration; see, for instance, Pastorello, Renault, and Touzi (2000). A verification of the equivalence of the indirect inference and simulated score formulations is in Gourieroux, Monfort, and Renault (1993). Of course, if the auxiliary model is sufficiently simple, then analytic expressions for $b(\rho)$ and \mathcal{J} become available and the computation

$$\hat{\rho}_n = \underset{\rho \in R}{\operatorname{argmin}} [\tilde{\theta}_n - b(\rho)]' (\tilde{\mathcal{J}}^{-1} \tilde{\mathcal{I}} \tilde{\mathcal{J}}^{-})^{-1} [\tilde{\theta}_n - b(\rho)]$$

becomes feasible as posed.

As mentioned earlier, there are variants on the scheme outlined above. Other statistical objective functions can be substituted for the likelihood. Another variant is as follows: Once the binding function has been computed from a simulation for given θ , the likelihood of the auxiliary model can be evaluated at the data and the value of the binding function at that θ and used as if it were the likelihood for purposes of inference. This is one way to implement a Bayesian variant of indirect inference as is outlined in Gallant and McCulloch (2009). They develop numerical methods to mitigate against the effects of the jitter in computing the binding function which can be effective in a Bayesian context. Software to implement their method is <http://econ.duke.edu/webfiles/arg/gsm>. Del Negro and Schorfheide (2004) describe another Bayesian approach that makes use of an auxiliary model. In their approach, the structural model is used to build a hierarchical likelihood that contains both parameters from the structural and auxiliary models both of which are estimated simultaneously.

The indirect inference formulation of the estimation problem can be useful device for modifying the estimator to achieve semiparametric or robustness properties. Space does not permit an exploration of those ideas here. For a discussion of seminonparametric properties achieved through indirect inference see Dridi and Renault (1998) and the references therein. For a discussion of robustness properties achieved through indirect inference, see Genton and de Luna (2002).

2.2 Details

We now discuss the ideas above in more detail. We consider nonlinear systems that have the features of the models described in Section 1. Specifically, (i) for a parameter vector ρ in a parameter space R , the random variables determined by the system have a stationary density

$$p(y_{-L}, \dots, y_{-1}, y_0 | \rho), \quad (2)$$

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

$$\mathcal{E}_\rho(g) = \int \cdots \int g(y_{-L}, \dots, y_0) p(y_{-L}, \dots, y_0 | \rho) dy_{-L} \cdots dy_0 \quad (3)$$

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

$$\mathcal{E}_\rho(g) \doteq \frac{1}{N} \sum_{t=1}^N g(\hat{y}_{t-L}, \dots, \hat{y}_{t-1}, \hat{y}_t). \quad (4)$$

As conventions, we 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 (2) is denoted by ρ° .

We presume that the data have been summarized in the projection step, as described in Section 3, and that a score generator of the form

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

and a weighting matrix

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

are available from the projection step. This formula assumes that $f(y | x, \tilde{\theta}_n)$ closely approximates $p(y | x, \rho^\circ)$. If the SNP density $f_K(y | x, \theta)$ is used as the auxiliary model with tuning parameters selected by BIC (Schwarz, 1978), $\tilde{\mathcal{I}}_n$ as above will be adequate (Gallant and Long, 1997; Gallant and Tauchen, 1999; and Coppejans and Gallant, 2000). If the approximation is not adequate, then

a HAC weighting matrix (Andrews, 1991) must be used. A common choice of HAC matrix is

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

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

$$\tilde{\mathcal{I}}_{n\tau} = \begin{cases} \frac{1}{n} \sum_{t=1+\tau}^n \left[\frac{\partial}{\partial \theta} \log f(\tilde{y}_t | \tilde{x}_{t-1}, \tilde{\theta}_n) \right] \left[\frac{\partial}{\partial \theta} \log f(\tilde{y}_{t-\tau} | \tilde{x}_{t-1-\tau}, \tilde{\theta}_n) \right]' & \text{if } \tau \geq 0 \\ \tilde{\mathcal{I}}_{n,-\tau} & \text{if } \tau < 0 \end{cases}$$

(Gallant and White, 1987).

Recall that the moment equations are

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

which can be 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 ρ° denotes the true value of ρ and θ° is an isolated solution of the moment equations $m(\rho^\circ, \theta) = 0$, then under regularity conditions that include holding the parameterization of the structural and auxiliary model fixed (Gallant and Tauchen, 1996)

$$\begin{aligned} \lim_{n \rightarrow \infty} \hat{\rho}_n &= \rho^\circ \quad \text{a.s.} \\ \sqrt{n}(\hat{\rho}_n - \rho^\circ) &\xrightarrow{\mathcal{L}} N\left\{0, [(M^\circ)'(\mathcal{I}^\circ)^{-1}(M^\circ)]^{-1}\right\} \end{aligned} \quad (6)$$

$$\lim_{n \rightarrow \infty} \hat{M}_n = M^o \quad \text{a.s.}$$

$$\lim_{n \rightarrow \infty} \tilde{\mathcal{I}}_n = \mathcal{I}^o \quad \text{a.s.}$$

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

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

if $f(y|x, \theta)$ encompass the data generating process, or

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

if not. Under the null hypothesis that $p(y_{-L}, \dots, y_0 | \rho)$ is the correct model,

$$L_0 = n m'(\hat{\rho}_n, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\hat{\rho}_n, \tilde{\theta}_n) \quad (7)$$

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 ,

$$L_h = n \left[m'(\hat{\rho}_n, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\hat{\rho}_n, \tilde{\theta}_n) - m'(\hat{\rho}_n, \tilde{\theta}_n) (\tilde{\mathcal{I}}_n)^{-1} m(\hat{\rho}_n, \tilde{\theta}_n) \right] \quad (8)$$

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 ρ_i of ρ 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 ρ_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 ρ_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

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

inspection of the t -ratios

$$T_n = S_n^{-1} \sqrt{n} m(\hat{\rho}_n, \tilde{\theta}_n), \quad (9)$$

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.

In practice, one would usually prefer to inspect $\sqrt{n} m(\hat{\rho}_n, \tilde{\theta}_n)$, which are underestimates of the t -ratios, to avoid having to determine the matrix \hat{M}_n numerically and to avoid any potential inaccuracies that numerical differentiation can introduce. Because the statistic L_0 provides an overall test of significance, it is not necessary to have exactly correct values of the t -ratios. That is, one is only relying on the t -ratios for suggestions as to where a structural model fails to fit and one is not relying on them for statistical inference.

3 Projection: General Guidelines on the Score Generator

A sensible question is how to determine the reduced form density $f(y_t|x_{t-1}, \theta)$ that defines the score generator for EMM. Interestingly, there are two natural principles that lead to different strategies. The first principle is data-based: choose $f(y_t|x_{t-1}, \theta)$ to be good approximation to the dynamics of the data, i.e., to $\text{pdf}(y_t|x_{t-1})$, whatever that might be. In other words, $f(y_t|x_{t-1}, \theta)$ should emerge from a carefully-conducted effort to model the data $\{\tilde{y}_t\}_{t=1}^n$ without much regard to the structural model. A flexible parameterization should be used if the dynamics of the data are not well understood *a priori*. The second principle is model-based: choose $f(y_t|x_{t-1}, \theta)$ to be a close approximation to the $p(y_t|x_{t-1}, \rho)$ implied by the structural model, so that the moment function $(\partial/\partial\theta) \log[f(y_t|x_{t-1}, \theta)]$ for EMM should look very much like moment function

$(\partial/\partial\rho)\log[p(y_t|x_{t-1},\rho)]$ of maximum likelihood estimation. Implementing this strategy entails using detailed knowledge of the characteristics of the structural model to build up the score generator.

We initially set forth the arguments for the data-based strategy in Bansal, Gallant, Hussey, and Tauchen (1993, 1995), and we have consistently argued for it over the model-based strategy ever since. The issue is controversial. Dridi and Renault (1998) argue for a more model-based strategy and Hansen (2002) outlines some of the issues. The gist of our argument is that the data-based strategy will be nearly fully efficient if the structural model is correctly specified, and it will reveal the inadequacy of the structural model if it is misspecified. On the other hand, the model-based strategy is fine if the structural model is correct, but it could be potentially very misleading if the structural model is wrong. See Gallant and McCulloch (2009) for an illustration. Rarely do we know for sure that our models are indeed correct.

We now look at some of these considerations in more detail.

3.1 *An Initial Look at Efficiency*

Let \mathcal{V}_f denote the asymptotic covariance matrix of EMM if the score generator is $f(y_t|x_{t-1},\theta)$. In Section 2 we saw that $\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{\mathcal{L}} N(0, \mathcal{V}_f)$, where from (6), \mathcal{V}_f is given by

$$\mathcal{V}_f = [(M^o)'(\mathcal{I}^o)^{-1}(M^o)]^{-1}$$

Let \mathcal{V}_{ML} denote the asymptotic distribution of the maximum likelihood estimator, which is given by

$$\mathcal{V}_{ML} = [\mathcal{E}(s_t s_t')]^{-1} = \mathcal{I}^{-1}$$

where

$$s_t = \frac{\partial}{\partial\rho} \log[p(y_t|x_{t-1},\rho^o)]$$

is the score function of the underlying probability model, presumed correct here. From basic maximum likelihood theory we have that

$$\mathcal{V}_{ML} \leq \mathcal{V}_f$$

Tauchen (1997) considers the iid case, $p(y_t|\rho)$ and shows that \mathcal{V}_f and \mathcal{V}_{ML} are connected as follows. Let

$$\Omega = \text{Var}(s_t - Bs_{ft})$$

where $s_{ft} = (\partial/\partial\theta)\log[f(y_t|\theta^o)]$ is the score of the reduced form model reduced form, and B is the coefficient matrix from a linear projection of $s_t = (\partial/\partial\rho)\log[f(y_t|\rho^o)]$ onto s_{ft} . Then

$$\mathcal{V}_{ML} = (\mathcal{V}_f^{-1} + \Omega)^{-1} \leq \mathcal{V}_f \tag{10}$$

with equality if $\Omega = 0$. Hence, the better the score of f comes to spanning the score of p , then the smaller is Ω and the closer is EMM to full efficiency. Tauchen (1997) also provides some intuition as to how this result would carry over to the dynamic case. Gallant and Long (1997) handle the dynamic case and prove that

$$\lim_{K \rightarrow \infty} \mathcal{V}_{f_K} = \mathcal{V}_{ML}$$

where f_K represents the K^{th} term in the SNP series expansion as described in Section 4 below.

The upshot is that the better the score generator approximates structural model, then the closer is \mathcal{V}_f to \mathcal{V}_{ML} . But since the structural model is presumed to be correct, the data-based approach has to produce a score function that is a close approximation to the true score function. If one knows of a good model for the data, then that model should be used as the auxiliary model. If not, as is often the case, then Gallant and Long's (1997) results provide a systematic strategy based on SNP modeling for getting a close approximation.

3.2 Misspecification

Suppose the structural model is itself misspecified. Will this be detected by the EMM objective function? The issue was first formally considered in Tauchen (1997) and examined in much more detail by Aguire-Torres (2001). The answer is essentially yes if one employs the data-based strategy to determine the score generator but no if one follows the model-based strategy for determining the score generator for EMM.

Define the densities

	conditional	joint
true model:	$\xi(y x)$	$\xi(x, y)$
structural model:	$p(y x, \rho)$	$p(x, y, \rho)$
auxiliary model:	$f(y x, \theta)$	$f(x, y \theta)$

where for simplicity we drop the time subscripts. The pseudo-true values of θ and ρ are

$$\theta^o = \operatorname{argmax}_{\theta} \int \int \log[f(y|x, \theta)] \xi(x, y) dy dx$$

$$\rho^o = \operatorname{argmin}_{\rho} m'(\rho, \theta^o) (\mathcal{I}^o)^{-1} m(\rho, \theta^o)$$

where

$$m(\rho, \theta) = \int \int \frac{\partial}{\partial \theta} \log[f(y|x, \theta)] p(x, y, \rho) dy dx$$

Note that \mathcal{I}^o is the limiting pseudo-information matrix computed under $\xi(x, y)$. Following Geweke (1983) define the approximate slope functional

$$S(f, p, \xi) = m'(\rho^o, \theta^o) (\mathcal{I}^o)^{-1} m(\rho^o, \theta^o)$$

The value of $S(f, p, \xi)$ is the limiting normalized value of the noncentrality parameter of the test of the overidentifying restrictions.

For fixed f and p , it is reasonably easy to come with plausible alternative models ξ such that

$$S(f, p, \xi) = 0$$

In other words, given f , there is no power to detect $p \neq \xi$. It is easy to construct such examples. The danger is fitting a misspecified p -model to the scores of a misspecified f -model, and thinking everything is fine, when in fact $p \neq \xi$.

The problem is with leaving f fixed. If one chooses f nonparametrically, say by SNP, then the preliminary analysis of Tauchen (1997) and detailed calculations of Aguire-Torres (2001) indicate that whenever $p \neq \xi$, then

$$\liminf_{K \rightarrow \infty} S(f_K, p, \xi) > 0$$

so the misspecification is detected with probability one asymptotically. As the editor has pointed out, these computations are not uniform in (n, K) but rather allow n to tend to infinity first and K to tend to infinity second.

The message, again, is to think nonparametrically when choosing the auxiliary model.

3.3 *Non-nested Models*

Another argument in favor of the data-based strategy has emerged after some years of experience with EMM. Frequently, one considers families of non-nested structural models and one faces a model selection problem. EMM using a data-based score generator forces all structural models to confront the same set of moment conditions, and therefore meaningful comparisons of objective values across models are available. For example, Dai and Singleton (2000) use the EMM objective function to guide model selection within and across the non-nested branches of the affine family of term structure models, as do Bansal and Zhou (2002) for regime switching affine term structure models.

3.4 *Dynamic Stability*

The EMM objective function is

$$Q(\rho) = m'(\rho, \tilde{\theta}) \tilde{\mathcal{I}}^{-1} m(\rho, \tilde{\theta})$$

and

$$\hat{\rho} = \underset{\rho \in \mathcal{R}}{\operatorname{argmin}} Q(\rho)$$

Simulated trajectories $\{\hat{y}_t\}_{t=1}^N$ are used to compute the expectation that defines $m(\rho, \theta)$. Since the underlying structural model typically have a nonlinear dynamic autoregressive structure, it is natural to consider potential problems if ρ lies in the explosive region of the parameter space and $|\hat{y}_t| \rightarrow \infty$?

Tauchen (1998) examines the issue of dynamic stability of the structural model (p) and the score generator (f). The upshot is that one really need not worry about imposing dynamic stability on the structural model itself. Dynamic stability is self-enforcing. If the optimizer wanders into the region of the parameter space where the underlying structural model is unstable, then the data simulator generates a wildly explosive simulated realization that induces

a large value of the objective function. The time series properties of this explosive realization are very much unlike the time series properties of the observed data set to which the auxiliary model has been fitted, so the objective function attains an exceedingly high value. The situation is actually a bit more subtle, because automatic stability is ensured only if the auxiliary model itself is dynamically stable. The use of a dynamically unstable auxiliary model can be expected to define a GMM objective function with very poor numerical properties in both the stable and unstable regions of the parameter space.

Dynamic stability is of practical importance. Andersen and Lund (1997) carefully examine a class of generalized GARCH and E-GARCH auxiliary models for the short-term interest rate. They find the former typically unstable, and therefore unusable as auxiliary models, while the latter are stable. Gallant and Tauchen (1998) likewise use model stability as part of the selection criterion. We now incorporate into the SNP code (Gallant and Tauchen, 2001c) nonlinear transformations of the state vector x_{t-1} that attenuate large movements and help enforce stability, but we still recommend checking long simulations to ensure the score generator is a stable model.

As the editor has pointed out, it is possible for explosive nonlinear dynamic processes to linger for extended periods in a strongly dependent state before they become explosive. This sort of behavior on the part of a structural model might not be detected.

4 A General Purpose Score Generator

4.1 Efficiency Comparisons

In Section 2 we defined the EMM estimator as

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

It is essentially a simulated method of moments estimator based on the moment function

$$m(\rho) = \mathcal{E}_\rho \tilde{\psi}$$

where

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

and for now we shall suppress the second argument of $m(\rho, \theta)$.

In Subsection 3.1 we noted that the closer f comes to approximating the condition density implied by the structural model, then the closer will be the asymptotic variance of the EMM estimator to that of maximum likelihood. In fact, a spanning argument can be used to show that the efficiency of EMM can be made asymptotically negligible.

But the same spanning argument applies to estimation using more traditional moments such as means, variances, etc., which we shall call below the Classical Method of Moments. Thus, an open question is whether the moment function of EMM, which entails the extra effort of estimating the score generator, defines a better set of moments, other things equal.

The question is considered and answered affirmatively by Gallant and Tauchen (1999), which we now summarize. They examine the simpler case where the random variables defined by the system (2) 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:

$$\tilde{\psi}_{c,n}(y) = \begin{pmatrix} y - \frac{1}{n} \sum_{i=1}^n \tilde{y}_i \\ y^2 - \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i)^2 \\ \vdots \\ y^K - \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i)^K \end{pmatrix},$$

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

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

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

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

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

ρ is of length p_{ρ} and θ 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 ρ that puts one of these moment functions, denoted generically as $\tilde{\psi}_n$, in the null space of the expectation operator \mathcal{E}_{ρ} 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 ρ 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}_{ρ} for any ρ , 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 ψ^o such that

$$\lim_{n \rightarrow \infty} \tilde{\psi}_n(y) = \psi^o(y) \quad \text{a.s.}$$

$$\lim_{n \rightarrow \infty} \tilde{\mathcal{I}}_n = \mathcal{E}_{\rho^o}(\psi^o)(\psi^o)' \quad \text{a.s.}$$

$$\sqrt{n} m_n(\rho^o) \xrightarrow{\mathcal{L}} N\left[0, \mathcal{E}_{\rho^o}(\psi^o)(\psi^o)'\right]$$

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

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

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

and

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

where

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

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

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

and

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

where

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

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

$$\min_b \mathcal{E}_{\rho^o} \left[a' \psi_p^o - (\psi^o)' b \right]^2 = \mathcal{E}_{\rho^o} \left(a' \psi_p^o \right)^2 - a' C^o a \geq 0. \quad (11)$$

Expression (11) 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 ρ . It is also apparent from (11) 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

$$f_K(y|\theta) = \frac{1}{\sigma} h_K\left(\frac{y - \mu}{\sigma}\right). \quad (12)$$

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 ψ_p is in the span of ψ^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), van der Sluis (1999), and, Zhou (2001). 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 $\{\rho : (\rho - \rho^o)'I_p^o(\rho - \rho^o) \leq \mathcal{X}_d^2/n\}$ has the smallest achievable volume,

$$\text{RE} = \frac{\det(C^o)}{\det(I_p^o)}$$

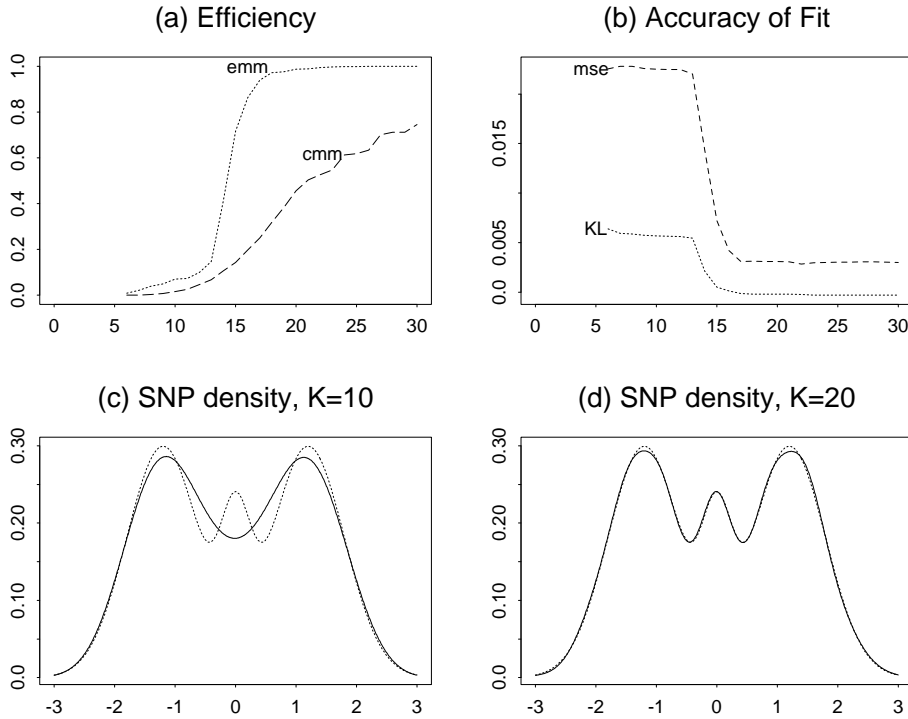


Figure 1. Relative Efficiency for the Trimodal Density. Panel (a) plots the relative efficiency of the EMM and CMM estimators against degree K , for the trimodal density of the Marron-Wand test suite. As seen, the efficiency of the EMM estimator increases rapidly when the degree K of the SNP auxiliary 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.

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 CMM. Moreover, their computations indicate that once $f_K(\cdot|\theta^\circ)$ begins to approximate $p(\cdot|\rho^\circ)$ accurately, the efficiency of the EMM estimator begins to increase rapidly. A representative illustration is provided by Figure 1, which shows the relative efficiency comparison for a trimodal density $p(y|\rho)$ taken from the Marron-Wand suite (Marron and Wand, 1992). As seen in Figure 1, once $f_K(\cdot|\theta^\circ)$ has detected the third mode of the trimodal density, EMM efficiency increases rapidly.

The second question to address is how many moments to include in the moment function ψ_f . As the computations in Gallant and Tauchen (1999) and Figure 1 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 ψ_c in that there seems to be no objective statistical criterion for determining the number of moments to include in ψ_c .

As the editor has pointed out, the Gallant and Long (1997) results that justify extending these ideas to non-Markovian situations rely on several high level assumptions and it is not clear what low level primitives will justify them, especially in continuous time applications. The most serious of these high level assumptions is their Assumption 4 which requires that the asymptotic variance of the maximum likelihood estimator can be accurately approximated by the variance of the maximum likelihood estimator with lags past some large value neglected. It is an important open problem to determine which classes of models will satisfy this assumption.

4.2 SNP: A General Purpose Score Generator

As indicated in Subsection 4.1, the best choice of a moment function ψ 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 (2) of the dynamic system under consideration can be written $p(x, y|\rho)$ and its transition density as

$$p(y|x, \rho) = \frac{p(x, y|\rho)}{\int p(x, y|\rho) dx} \quad (13)$$

If one expands $\sqrt{p(x, y|\rho^o)}$ 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

$$h_K(z|x) = \frac{[\mathcal{P}(z, x)]^2 \phi(z)}{\int [\mathcal{P}(u, x)]^2 \phi(u) du}, \quad (14)$$

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

$$\mathcal{P}(z, x) = \sum_{\alpha=0}^{K_z} \left(\sum_{\beta=0}^{K_x} a_{\beta\alpha} x^\beta \right) z^\alpha, \quad (15)$$

where α and β 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

$$\mu_x = b_0 + Bx_{t-1}, \quad (16)$$

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 $p(y|x, \rho^0)$. With this, the location-scale transformation becomes

$$y = R_x z + \mu_x \quad (17)$$

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 , ρ_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

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

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, called R-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 (17) and innovation density (14) yields the SNP density

$$f_K(y | x, \theta) = \frac{h_K[R_x^{-1}(y - \mu_x) | x]}{\det(R_x)}. \quad (18)$$

Hereafter, we shall distinguish among the lag lengths appearing in the various components of the expansion. The number of lags in μ_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 μ_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 $\{z_t\}$. 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 1.

The empirical work described in this article uses the R-GARCH form of the conditional variance matrix. In 2004 the SNP Fortran code was reimplemented in C++ and a BEKK variance matrix (Engle and Kroner, 1995) modified to add leverage and level effects was substituted for the R-GARCH. It is

$$\Sigma_{x_{t-1}} = R_0 R_0'$$

Table 1. 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.

$$\begin{aligned}
& + \sum_{i=1}^{L_g} Q_i \Sigma_{x_{t-1-i}} Q_i' \\
& + \sum_{i=1}^{L_r} P_i (y_{t-i} - \mu_{x_{t-1-i}}) (y_{t-i} - \mu_{x_{t-1-i}})' P_i' \\
& + \sum_{i=1}^{L_v} \max[0, V_i(y_{t-i} - \mu_{x_{t-1-i}})] \max[0, V_i(y_{t-i} - \mu_{x_{t-1-i}})]' \\
& + \sum_{i=1}^{L_w} W_i x_{(1),t-i} x_{(1),t-i}' W_i'.
\end{aligned}$$

Above, R_0 is an upper triangular matrix. The matrices P_i , Q_i , V_i , and W_i can be scalar, diagonal, or full M by M matrices. The notation $x_{(1),t-i}$ indicates that only the first column of x_{t-i} enters the computation. The $\max(0, x)$ function is applied elementwise. Because $\Sigma_{x_{t-1}}$ must be differentiable with respect to the parameters of $\mu_{x_{t-2-i}}$, the $\max(0, x)$ function actually applied is a twice continuously differentiable cubic spline approximation that agrees with the $\max(0, x)$ function except over the interval $(0, 0.1)$ over which it lies slightly above the $\max(0, x)$ function.

5 Reprojection: Analysis of Post-Estimation Simulations

5.1 Simple Illustration of Volatility Extraction

We start with an illustration that gives the main idea of reprojection. In Gallant and Tauchen (2001a) we estimated via EMM the vector SDE with two stochastic volatility factors:

$$\begin{aligned} dU_{1t} &= \alpha_{10}dt + \exp(\beta_{10} + \beta_{12}U_{2t} + \beta_{13}U_{3t}) dW_{1t} \\ dU_{2t} &= \alpha_{22}U_{2t}dt + dW_{2t} \\ dU_{3t} &= \alpha_{33}U_{3t}dt + dW_{3t} \end{aligned} \tag{19}$$

using daily data on Microsoft (MSFT), 1986-2001. Here U_{1t} is the log price process and

$$y_t = 100 * (U_{1t} - U_{1,t-1})$$

at integer t is the observation equation for the geometric daily return expressed as a percent. U_{2t} and U_{3t} are stochastic volatility factors. We find that 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} .

Thus, from the observed data set $\{\tilde{y}_t\}_1^n$ we generated via EMM the parameter estimate $\hat{\rho}$ for each model under consideration. We now summarize how to 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 $\{\hat{y}_t\}_{t=1}^N$ 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 $\{\tilde{y}_t\}_{t=1}^n$.

In the application we work with the conditional mean functions of the volatility factors. Our targets are

$$\mathcal{E}(U_{it} | \{y_\tau\}_{\tau=1}^t), \quad i = 2, 3 \tag{20}$$

To begin, we generated 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 predictions of U_{2t} and U_{3t} via filtering y_t , 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.

The figures in Gallant and Tauchen (2001a) indicate that \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_{3t} . This result suggest that the volatility increase surround the 87 crash was rather temporary, which appears consistent with raw data plots. Also, the reprojected volatility factor from a model with only one stochastic volatility factor misses much of the crash of 1987, which reflects further on the shortcomings of single-factor stochastic volatility models.

5.2 General Theory of Reprojection

Having the EMM estimate of system parameters $\hat{\rho}_n$ in hand, we should like to elicit the dynamics of the implied conditional density for observables

$$\hat{p}(y_0|x_{-1}) = p(y_0|x_{-1}, \hat{\rho}_n). \quad (21)$$

Recall that x_{-1} represents the lagged state vector, and so in the Markov case (21) is an abbreviated notation for

$$\hat{p}(y_0|y_{-L}, \dots, y_{-1}) = p(y_0|y_{-L}, \dots, y_{-1}, \hat{\rho}_n).$$

Although analytic expressions are not available, an unconditional expectation

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

can be computed by generating a simulation $\{\hat{y}_t\}_{t=-L}^N$ from the system with parameters set to $\hat{\rho}_n$ and using

$$\mathcal{E}_{\hat{\rho}_n}(g) = \frac{1}{N} \sum_{t=0}^N g(\hat{y}_{t-L}, \dots, \hat{y}_t).$$

With respect to unconditional expectation so computed, define

$$\hat{\theta}_K = \operatorname{argmax}_{\theta \in \mathfrak{R}^{p_K}} \mathcal{E}_{\hat{\rho}_n} \log f_K(y_0|x_{-1}, \theta)$$

where $f_K(y_0|x_{-1}, \theta)$ is the SNP density given by (18). Let

$$\hat{f}_K(y_0|x_{-1}) = f_K(y_0|x_{-1}, \hat{\theta}_K). \quad (22)$$

Theorem 1 of Gallant and Long (1997) states that

$$\lim_{K \rightarrow \infty} \hat{f}_K(y_0|x_{-1}) = \hat{p}(y_0|x_{-1}).$$

Convergence is with respect to a weighted Sobolev norm that they describe. Of relevance here is that convergence in their norm implies that \hat{f}_K as well as its partial derivatives in $(y_{-L}, \dots, y_{-1}, y_0)$ converge uniformly over \mathfrak{R}^ℓ , $\ell = M(L+1)$, to those of \hat{p} . We propose to study the dynamics of \hat{p} by using \hat{f}_K as an approximation. This result provides the justification for our approach.

To approximate \hat{p} by \hat{f}_K values of $(L_u, L_r, L_p, K_z, I_z, K_x, I_x)$ must be chosen. It seems natural to reuse the values of the projection that determined $\hat{\rho}_n$ because, among other things, that choice facilitates a comparison of the constrained dynamics determined by the estimated system with the unconstrained dynamics determined by the data. However, if the estimated nonlinear system is to be sampled at a different frequency than was the data, then it will be necessary to redetermine $(L_u, L_r, L_p, K_z, I_z, K_x, I_x)$ by the methods described in Subsection 2.2. We anticipate that the dynamics at a different sampling frequency will not often be of interest and we shall presume in what follows that the sampling frequency is the same as the data. The modifications required when it differs are mentioned as they occur.

Of immediate interest in eliciting the dynamics of observables are the first two one-step-ahead conditional moments

$$\begin{aligned} \mathcal{E}(y_0|x_{-1}) &= \int y_0 f_K(y_0|x_{-1}, \hat{\theta}_K) dy_0 \\ \operatorname{Var}(y_0|x_{-1}) &= \int [y_0 - \mathcal{E}(y_0|x_{-1})][y_0 - \mathcal{E}(y_0|x_{-1})]' f_K(y_0|x_{-1}, \hat{\theta}_K) dy_0 \end{aligned}$$

where $x_{-1} = (y_{-L}, \dots, y_{-1})$. Owing to the form of a Hermite expansion, expressions for these integrals as linear combinations of high order moments of the normal distribution are available (Gallant and Tauchen, 1992). The moments themselves may be obtained from standard recursions for the moments of the normal (Johnson and Kotz, 1970).

Filtered volatility is the one-step-ahead conditional standard deviation evaluated at data values; viz.

$$\sqrt{\text{Var}(y_{k0} | x_{-1})} \Big|_{x_{-1}=(\tilde{y}_{t-L}, \dots, \tilde{y}_{t-1})} \quad t = 0, \dots, n. \quad (23)$$

In (23), \tilde{y}_t denotes data and y_{k0} denotes the k th element of the vector y_0 , $k = 1, \dots, M$. Because filtered volatility is a data dependent concept, the dynamic system must be sampled at the same frequency as the data to determine \hat{f}_K . It has been claimed that filtered volatility could not be recovered from method of moments estimates of a nonlinear dynamic system with partially observed state and that this has been a criticism of such estimates. However, as just seen, filtered volatility is easily computed using the reprojection notion.

We are using the term “filtered volatility” with a purely ARCH-type meaning as in the nonlinear impulse-response literature. Another usage of filtering, perhaps the predominant one, involves estimating an unobserved state variable conditional upon all past and present observables. Filtering according to this notion (for L lags rather than back to the first observation) can be accomplished through reprojection. This may be seen by noting that one can repeat the derivation with y taken to be a contemporaneous unobserved variable and x taken to be contemporaneous and lagged observed variables. Denote y and x thus modified by y^* and x^* , respectively. The result is a density $f_K(y^*|x^*, \theta)$ of the same form as (18) but with altered dimensions. One can simulate $\{y_t^*, x_t^*\}$ from the structural modal and perform the reprojection step to get $\hat{f}_K(y^*|x^*)$ as described above. The proof of Gallant and Long (1997) can be altered to justify these modifications. How one uses $\hat{f}_K(y^*|x^*)$ will be application specific. For instance, one might wish obtain an estimate of

$$y_t^* = \int_t^{t+T} \exp(\beta_{10} + \beta_{12}U_{2t} + \beta_{13}U_{3t}) dt$$

in a system such as (19) for the purpose of pricing an option. In this instance, $x_t^* = (U_{1,t-L}, \dots, U_{1t})$, and $\hat{y}_t^*(x^*) = \int y^* \hat{f}_K(y^*|x^*) dy^*$. To avoid any confusion, we shall refer to (23) as reprojected volatility hereafter. We now return to the main discussion.

One-step-ahead dynamics may be studied by means of plots of (the elements of) $\mathcal{E}(y_0|y_{-L}, \dots, y_{-1} + \Delta)$, $\text{Var}(y_0|y_{-L}, \dots, y_{-1} + \Delta)$, or other conditional moments against δ where Δ is an M -vector with δ in the i th element and zeroes

elsewhere. More general perturbation strategies may be considered such as $\Delta = \delta \tilde{y}_\tau$ where \tilde{y}_τ is a point chosen from the data such that perturbations in the direction $\delta \tilde{y}_\tau$ take contemporaneous correlations among the components of y_t into account. Perturbations to a single element of y_{-1} in a multivariate setting may represent a movement that is improbable according to the dynamics of the system. Some thought must be given to the perturbation scheme in multivariate applications if plots of conditional moments against δ are to be informative. This issue is discussed in Gallant, Rossi, and Tauchen (1993).

Two methods for choosing (y_{-L}, \dots, y_{-1}) for these plots suggest themselves. The first is to put y_{-L}, \dots, y_{-1} to the sample mean, that is, put $(y_{-L}, \dots, y_{-1}) = (\bar{y}, \dots, \bar{y})$ where $\bar{y} = (1/n) \sum_{t=0}^n \tilde{y}_t$, and plot, for instance,

$$\text{Var}(y_0 | \bar{y}, \dots, \bar{y} + \Delta) \quad (24)$$

against δ . The second is to average over the data and plot, for instance,

$$(1/n) \sum_{t=0}^n \text{Var}(y_t | \tilde{y}_{t-L}, \dots, \tilde{y}_{t-1} + \Delta) \quad (25)$$

against δ . If the estimated system is sampled at a different frequency than the data, then one plots the average $(1/N) \sum_{t=0}^n \text{Var}(y_t | \hat{y}_{t-L}, \dots, \hat{y}_{t-1} + \Delta)$ over a simulation $\{\hat{y}_t\}_{t=-L}^N$ at the correct frequency instead.

In an economic system, the graphics just described are interpreted as representing the consequences of a shock to the system that comes as a surprise to the economic agents involved, and similar interpretations hold in other contexts. If one wants to consider the consequences of forcing the system to a different equilibrium, the graphic obtained by plotting $\text{Var}(y_0 | y_{-L} + \Delta, \dots, y_{-1} + \Delta)$ against δ is relevant. They can be quite different.

Multi-step-ahead dynamics may be studied by considering plots of the trajectories

$$\mathcal{E}[g(y_{j-L}, \dots, y_{j-1}) | y_{-L}, \dots, y_{-1} + \Delta], \quad j = 0, 1, \dots, J, \quad (26)$$

where $g(y_{-L}, \dots, y_{-1})$ is a time invariant function whose choice is discussed immediately below. As discussed in Gallant, Rossi, and Tauchen (1993), if one sets the initial condition to $(y_{-L}, \dots, y_{-1} + \Delta) = (\bar{y}, \dots, \bar{y} + \Delta)$ it is helpful to net out transients by plotting either

$$\mathcal{E}[g(y_{j-L}, \dots, y_{j-1}) | \bar{y}, \dots, \bar{y} + \Delta] - \mathcal{E}[g(y_{j-L}, \dots, y_{j-1}) | \bar{y}, \dots, \bar{y}] \quad (27)$$

or

$$\frac{1}{n} \sum_{t=0}^n \mathcal{E} [g(y_{t+j-L}, \dots, y_{t+j-1}) | \tilde{y}_{t-L}, \dots, \tilde{y}_{t-1} + \Delta] \quad (28)$$

against $j = 0, 1, \dots, J$ instead of (26). Although (28) is conceptually superior, because it recognizes the fact that a sequence exactly equal to the stationary mean for L periods can never happen, in the examples considered by Gallant, Rossi, and Tauchen (1993), plots of (27) had nearly the same appearance and are much cheaper to compute.

To compute (26), one exploits the fact that there are efficient algorithms for sampling the density $\hat{f}_K(y_0 | y_{-L}, \dots, y_{-1} + \Delta)$ recursively to obtain R simulated futures

$$\{\hat{y}_{0,i}, \dots, \hat{y}_{J,i}\}, \quad i = 1, \dots, R,$$

each conditional upon $y_{-L}, \dots, y_{-1} + \Delta$ (Gallant and Tauchen, 1992). Prepend $\{y_{-L}, \dots, y_{-1} + \Delta\}$ to each future to obtain the sequences

$$\{\hat{y}_{-L,i}, \dots, \hat{y}_{-1,i}, \hat{y}_{0,i}, \dots, \hat{y}_{J,i}\}, \quad i = 1, \dots, R.$$

$\mathcal{E}[g(y_{j-L}, \dots, y_{j-1}) | y_{-L}, \dots, y_{-1} + \Delta]$ can then be computed as

$$\mathcal{E}[g(y_{j-L}, \dots, y_{j-1}) | y_{-L}, \dots, y_{-1} + \Delta] = \frac{1}{R} \sum_{i=1}^R g(\hat{y}_{j-L,i}, \dots, \hat{y}_{j-1,i}).$$

A general discussion of appropriate choice of $g(y_{-L}, \dots, y_{-1})$ for nonlinear impulse-response analysis, the analysis of turning points, etc. is in Gallant, Rossi, and Tauchen (1993). Of these, the more routinely useful are the conditional mean profiles

$$\begin{aligned} & \mu_j(y_{-L}, \dots, y_{-1} + \Delta) \\ &= \mathcal{E} [\mathcal{E}(y_{k,j} | y_{j-L}, \dots, y_{j-1}) | y_{-L}, \dots, y_{-1} + \Delta], \quad j = -1, \dots, J \end{aligned}$$

for the components $k = 1, \dots, M$ of y , which extend the impulse-response profiles of Sims (1980) to nonlinear systems, and conditional volatility profiles

$$\begin{aligned} & \sigma_j^2(y_{-L}, \dots, y_{-1} + \Delta) \\ &= \mathcal{E} [\text{Var}(y_{k,j} | y_{j-L}, \dots, y_{j-1}) | y_{-L}, \dots, y_{-1} + \Delta], \quad j = 0, \dots, J \end{aligned}$$

which extend the volatility impulse-response profiles of Engle, Ito, and Lin (1990) and Bollerslev and Engle (1993) to nonlinear systems. Plots of the conditional mean profile reveal the future dynamic response of system forecasts to a contemporaneous shock to the system. These will, in general, be nonlinear and can differ markedly when the sign of δ changes. Similarly for volatility.

Persistence can be studied by inspection of profile bundles, which are overplots for $t = 0, \dots, n$ of the profiles

$$\{\mu_j(\tilde{y}_{t-L}, \dots, \tilde{y}_{t-1}), j = -1, \dots, J\} \quad (29)$$

That is, one overplots profiles conditional on each observed datum. If the thickness of the profile bundle tends to collapse to zero rapidly, then the process is mean reverting. If the thickness tends to retain its width, then the process is persistent. Similarly, the profile bundles

$$\{\{\sqrt{\sigma_j^2}(\tilde{y}_{t-L}, \dots, \tilde{y}_{t-1}), j = 0, \dots, J\}, t = 0, \dots, n\} \quad (30)$$

can be used to examine volatility for persistence. These are extensions to nonlinear systems of notions of persistence due to Bollerslev and Engle (1993). Rather than comparing plots, one can instead compare half-lives. A half-life \hat{j} can be obtained by computing the range R_j at each ordinate $j = 0, \dots, J$ of either (29) or (30), regressing $\log R_j$ on $j\beta$, and using $(-\log 2)/\hat{\beta}$ as an estimate of half-life.

Extensive examples of the use of the methods described here for elucidating the joint dynamics of stock prices and volume are in Gallant, Rossi, and Tauchen (1993).

6 Applications

There are now several applications of EMM to substantive problems in continuous time estimation and economics more broadly. For reasons of space, we can only review in detail a few applications. At the end of this section we give a short overview of the other applications of which we are currently aware. Simulation methods for continuous time models are discussed in Kloeden and Platen (1992) in general. The papers that we discuss contain within them the details on the way these methods were adapted to the particular problem.

6.1 Multi-factor stochastic volatility models for stock returns

6.1.1 Jump Diffusions

We start with the application of Andersen, Benzoni, and Lund (2002). They consider the familiar stochastic volatility diffusion for an observed stock price S_t given by

$$\frac{dS_t}{S_t} = (\mu + cV_t)dt + \sqrt{V_t}dW_{1t} \quad (31)$$

where the unobserved volatility process V_t is either log-linear

$$\text{Log linear: } d \log(V_t) = [\alpha - \beta \log(V_t)] + \eta dW_{2t} \quad (32)$$

or square-root (affine)

$$\text{Square root: } dV_t = (\alpha - \beta V_t) + \eta \sqrt{V_t} dW_{2t}. \quad (33)$$

Here, W_{1t} and W_{2t} are standard Brownian motions that are correlated with $\text{corr}(dW_{1t}, dW_{2t}) = \rho$. The notation is self-explanatory taking note that the term cV_t reflects possible GARCH in mean effects. The version with the log-linear volatility dynamics has attracted substantial attention in the econometrics literature, while the version with square-root volatility dynamics has attracted attention in the finance literature because of the availability of closed form solutions for options prices.

Andersen, Benzoni, and Lund use EMM to estimate both versions of the stochastic volatility model with daily S&P 500 Stock Index data, January 2, 1953 – December 31, 1996. Their auxiliary model is an E-GARCH model (Nelson, 1991) with an SNP-like Hermite series representation for the error density. They report that the EMM chi-square test statistic (7) sharply rejects both versions; likewise, the EMM t -ratio diagnostics (9) indicate that these models have difficulty accommodating the tail behavior of the data.

These authors also consider a more general jump diffusion stochastic volatility models

$$\frac{dS_t}{S_t} = (\mu + cV_t - \lambda_t \bar{\kappa})dt + \sqrt{V_t}dW_{1t} + \kappa_t dq_t \quad (34)$$

with jump intensity given by

$$\lambda_t = \lambda_0 + \lambda_1 V_t \quad (35)$$

and jump size κ_t given by

$$\log(1 + \kappa_t) \sim N[\log(1 + \bar{\kappa}) - 0.5\delta^2, \delta^2]$$

The jump diffusion models pass the EMM chi-squared test of fit and the EMM diagnostic t -ratio tests, which suggests an adequate fit. Once jumps are included in the model, the test statistics reveal no substantive difference between the log-linear and square-root specifications for volatility. Also, their estimates suggest little evidence for state-dependent jumps in (35). They go on to compute hypothetical options prices under various assumptions about the risk premiums on volatility and jump risks. They illustrate the role of stochastic volatility and jumps in generating anomalies such as volatility smiles and smirks.

6.1.2 *Alternative Models*

The fact that adding a jump component to a basic stochastic volatility model improves the fit so much reflects two familiar characteristics of financial price movements: thick non-Gaussian tails and persistent time-varying volatility. A model with a single stochastic volatility factor can accommodate either of these characteristics separately, but not both together. The addition of the jump factor accounts for the thick tails. Doing so complicates the estimation, however, because a direct simulation of a jump diffusion entails a discontinuous path and thereby a discontinuous objective function. Andersen, Benzoni, and Lund need to implement a simulation strategy that smooths out the sample path across a jump boundary.

An alternative to adding the jump component is to add another stochastic volatility factor. This step is undertaken via EMM in Gallant, Hsu, and Tauchen (1999), with some encouraging initial results. A more extensive investigation is undertaken in the next paper we review.

6.1.3 *Volatility Index Models*

Chernov, Gallant, Ghysels, and Tauchen (2003) consider a four factor model of the form

$$\begin{aligned} \frac{dP_t}{P_t} &= (\alpha_{10} + \alpha_{12}U_{2t})dt + \sigma(U_{3t}, U_{4t})(dW_{1t} + \psi_{13}dW_{3t} + \psi_{14}dW_{4t}) \\ dU_{2t} &= (\alpha_{20} + \alpha_{22})dt + \beta_{20}dW_{2t} \end{aligned} \quad (36)$$

In the above, P_t represents the financial price series evolving in continuous time; U_{2t} is a stochastic drift factor; U_{3t} and U_{4t} are stochastic volatility factors

that affect price evolution through the volatility index function $\sigma(U_{3t}, U_{4t})$.

These authors consider two broad classes of setups for the volatility index functions and factor dynamics: an affine setup, where the index function and volatility dynamics are

$$\begin{aligned}\sigma(U_{3t}, U_{4t}) &= \sqrt{\beta_{10} + \beta_{13}U_{3t} + \beta_{14}U_{4t}} \\ dU_{it} &= (\alpha_{i0} + \alpha_{ii}U_{it})dt + \sqrt{\beta_{i0} + \beta_{ii}U_{it}}dW_{it} \quad i = 3, 4\end{aligned}\tag{37}$$

and a logarithmic setup where

$$\begin{aligned}\sigma(U_{3t}, U_{4t}) &= \exp(\beta_{10} + \beta_{13}U_{3t} + \beta_{14}U_{4t}) \\ dU_{it} &= (\alpha_{i0} + \alpha_{ii}U_{it})dt + (\beta_{i0} + \beta_{ii}U_{it})dW_{it} \quad i = 3, 4\end{aligned}\tag{38}$$

The simpler stochastic volatility models with only one volatility factor, (31) above, are subsumed in this setup by taking $\beta_{14} = 0$.

Chernov, Gallant, Ghysels, and Tauchen (2003) apply EMM to estimate the above models along with affine jump diffusion models, using daily data on the DOW Index, January 2, 1953, to July 16, 1999. They find that models with two volatility factors, U_{3t} and U_{4t} , do much better on the EMM chi-squared specification test than do models with only a single volatility factor. They also find the logarithmic two-volatility factor models (38) outperform affine jump diffusions and basically provide an acceptable fit to the data. One of the volatility factors is extremely persistent and the other strongly mean reverting. Interestingly, the volatility feedback parameter, β_{ii} , is positive and very important for finding an acceptable fit. This parameter permits the local variability of the volatility factors to be high when the factors themselves are high, a characteristic of volatility that has been noted by others. The strongly mean reverting factor with the volatility feedback acts much like a jump factor in the return process itself.

At this point, it is not clear whether jump diffusions or multiple-factor models with appropriate factor dynamics are the right models for equity prices. The former, with jumps entered directly into the price process, are intuitively appealing models for financial prices. But the jumps generate complications for the simulations and estimation. On the other hand, the multifactor models are far easier to simulate and estimate and might prove more adaptable to derivatives computations, since all sample paths are continuous and standard hedging arguments and the Ito calculus apply.

6.2 Term Structure of Interest Rates

6.2.1 Affine Term Structure Models

Dai and Singleton (2000) apply EMM for estimation of an affine term structure model. In the affine setting, the vector of underlying state variables, Y_t , follows affine dynamics

$$dY_t = \bar{K}[\bar{\theta} - Y_t]dt + \Sigma\sqrt{S_t}d\tilde{W}_t \quad (39)$$

where S_t is a diagonal matrix with entries $S_{ii,t} = \beta_{i0} + \beta'_i Y_t$. The short-rate of interest follows

$$r_t = \delta_0 + \delta'_y Y_t$$

On these assumptions for the risk neutral dynamics, the pure-discount bond prices are given by

$$P_t(\tau) = e^{A(\tau) - B(\tau)'Y_t}$$

where $A(\tau)$ and $B(\tau)$ are given by the solutions to ordinary differential equations.

Dai and Singleton use Euro-dollar swap rates, and the observation equation is a bit more complicated than in other applications due to the nature of swaps. The no-arbitrage swap rate, $r_{s\tau t}$, on a fixed for variable swap at times $t + k\tau_0$, $k = 1, 2, \dots, K$, $\tau = K\tau_0$, is

$$r_{s\tau t} = \frac{1 - P_t(K\tau_0)}{\sum_{k=1}^K P_t(k\tau_0)}$$

They estimate ATSMs using three observed variables $y_t = (y_{1t} \ y_{2t} \ y_{3t})'$:

y_{1t}	$-0.50 \log[P_t(0.50)]$	Six Month LIBOR
y_{2t}	r_{s2t}	Two-Year Swap Rate
y_{3t}	r_{s10t}	Ten-Year Swap Rate

This selection defines the observation function

$$y_t = \phi(Y_t, \rho)$$

where ρ contains all of the parameters of the Affine Term Structure Model (ATSM) to be estimated and tested.

Dai and Singleton focus on two stochastic volatility models for the term structure. One is due to Balduzzi, Das, Foresi, and Sundaram (1996), abbreviated (BDFS), (1996) and the other to Chen (1996). Each lies in a separate branch of the family of ATSMs. Dai and Singleton find that neither model fits the data, in sense that the overall goodness-of-fit chi-squared tests are very large relative to degrees of freedom and the diagnostic t -ratios are well above 2.0 in magnitude. However, if each model is expanded outwards to the maximal identified ATSM within its particular branch, then the chi-squared tests for both models become acceptable at conventional significance. To choose an overall preferred model, Dai and Singleton undertake additional analysis of post estimation simulations, much in the spirit of reprojection analysis described in Section 5 above, to select the extended version of the BDFS model as their preferred model.

6.2.2 Regime Switching Affine Term Structure Models

Bansal and Zhou (2001) examine a class affine models with stochastic regime switching. In their class of models, factor dynamics are constant-parameter affine within each regime, but the economy shifts stochastically between regimes. They deduce appropriate closed-form bond pricing functions that properly account for the regime switching. The use of regime switching models is intuitively appealing in view of potential effects on fixed income markets of various monetary regimes. Bansal and Zhou use monthly data, 1964–1995, on yields of six months and five year maturities for estimation. They use an ARCH-type model with an SNP error density as the auxiliary model. They find that a two-factor regime switching model passes the EMM test of specification while every model in broad class of two- and three-factor constant regime affine models is sharply rejected. They also find that the estimated regime switching model does pricing in pricing the cross section of bond prices beyond the two basis yields use in estimation.

6.2.3 Non-Affine Models

Ahn, Dittmar, and Gallant (2001) use EMM to examine the class of quadratic term structure models (QTSMs) for two monthly data sets, January 1952 – February 1991 and November 1971 – December 1999. They find that the QTSM models generally outperform affine models on the EMM diagnostic test, but no QTSM is capable of explaining the data. Ahn, Dittmar, Gallant, and Gao (2001) use EMM to estimate hybrid models where some underlying factors follow affine dynamics and the others quadratic. They find that hybrid

models do better than either class separately but are still rejected on the EMM chi-squared test of fit.

An interesting and promising line of research would be to combine the findings Bansal and Zhou (2001), who report favorable evidence for regime switching models, with those of Ahn, Dittmar, and Gallant (2001) who find encouraging evidence for quadratic term structure models.

6.3 Exchange Rates

Chung and Tauchen (2001) use EMM to test various target zone models of exchange rates. They consider the basic model where the fundamental k_t evolves as

$$dk_t = \mu dt + \sigma dw_t \quad (40)$$

and more general models with mean reversion

$$dk_t = -\gamma(k_t - k_0) + \sigma dw_t. \quad (41)$$

The central bank is assumed to follow policy actions to keep the fundamental within the band $[\underline{k}, \bar{k}]$. Letting s_t denote the exchange rate, then the target zone model generates the observation equation

$$s_t = G(k_t, \rho), \quad (42)$$

where the functional form of G is determined by the asset pricing equation that connects the dynamics of the exchange rate to the fundamental process k_t and by the boundary and smooth pasting conditions. Above, ρ represents the parameters. See Delgado and Dumas (1991) for details on specification and solution of target zone models. Evidently, it is relatively simple to simulate exchange rate data from a target zone model and thereby implement EMM.

Chung and Tauchen (2001) apply the procedure to weekly French franc-Deutsche mark exchange rates, 1987–1993. Their findings, in brief, are as follows. Consistent with previous empirical work, their specification tests reject all target zone models considered when bounds, \underline{k} and \bar{k} , are determined directly from officially announced bands. However, they find that a very acceptable fit is given by a target zone model with implicit bands, i.e., where \underline{k} and \bar{k} are free parameters, and the fundamental process is with mean reversion (41). Their results indicate that the central banks were operating within an implicit band inside the announced official bands. Interestingly, their results

are consistent with theoretical predictions for a bilateral analysis of exchange rates determined in a multilateral system (Pedroni, 2001). Finally, Chung and Tauchen present rather dramatic graphical evidence on the much better fit to the data provided by the preferred target zone model over a conventional stochastic volatility model for exchange rates.

A recent exchange rate application that uses the C++ MCMC implementation of EMM is Danielsson and Penaranda (2007). They estimate the parameters of a coordination game of market instability which focuses on the endogenous reaction of agents to fundamentals and liquidity. They apply the model to the potential for financial turmoil caused by carry trades using data for various subperiods that bracket the yen-dollar market in 1998. They find that the strategic behavior of agents is required to account for the turmoil in that market rather than just market fundamentals and liquidity.

6.4 *General Equilibrium Models*

Genotte and Marsh (1993) is an early effort to estimate a general equilibrium asset pricing model by simulated method of moments. In Bansal, Gallant, Hussey, and Tauchen (1993, 1995) we employ EMM to estimate small-scale general equilibrium model of international currency markets. More recently, Valderrama (2001) has implemented EMM for estimation of a small-scale real business cycle model and Bansal, Gallant, and Tauchen (2007) contrast the implications of the habit and long run risk models.

Estimation of completely specified equilibrium models, i.e., starting from tastes and technology, faces a computational bottleneck. For candidate values of the parameter the users needs to solve for the equilibrium along the simulated trajectory. This computational requirement is generally more demanding than that required to estimate an SDE, as described in many of the preceding examples. However, recent sharp increases in computational power, in the form of faster processors linked by parallization software, indicate that it will soon be feasible to investigate more extensively via EMM such fully articulated models. In an initial effort, we are exploring the feasibility of confronting the models of and Bansal and Yaron (2000). These models entail complicated state and time nonseparable specifications for the stochastic discount factor and elaborate multi-factor models dynamics for cash flow dynamics, and thereby present serious challenges for estimation.

6.5 *Additional Applications*

Below we give a short summary of additional applications of which we are currently aware. Many of these applications preceded and motivated those described above. We apologize in advance for omissions and would be interested in knowing of applications we might have inadvertently left out; send an e-mail with the citation to either **george.tauchen@duke.edu** or **ron.gallant@duke.edu**.

Discrete time stochastic volatility models are well suited for EMM estimation. Van der Sluis (1997, 1999) implements the method and provides C/C+ code under Ox for discrete time univariate stochastic volatility models. Gallant, Hsieh, and Tauchen (1997) use it to examine an extensive list of discrete time stochastic volatility models and document a set of empirical shortcomings.

Applications to estimation of continuous time stochastic volatility models include Engle and Lee (1996), Gallant and Tauchen (1997), and Gallant and Long (1997). Mixon (1998) generalizes the log-linear Gaussian continuous time model to include a feedback effect in volatility. Gallant, Hsu, and Tauchen (1999) also find this feedback effect to be important as well a second volatility factor in their investigation of daily returns and range data. Chernov, Gallant, Ghysels, and Tauchen (1999) use the technique to explore stochastic volatility and state dependent jump models.

A recent application to options pricing is Chernov and Ghysels (2000), who use the technique for joint estimation of the risk neutral and objective probability distributions using a panel of options data. Pastorello, Renault, and Touzi (2000) use it to deal with the estimation of continuous-time stochastic volatility models of option pricing.

Early applications to interest rate modeling include Pagan, Hall, and Martin (1996), who apply the technique for estimating a variety of factor models of the term structure, and Andersen and Lund (1997), who use the technique to estimate a stochastic volatility model of the short rate. Some evidence from EMM diagnostics on the shortcomings of a one factor model is set forth in Tauchen (1997) and in McManus and Watt (1999). An extensive analysis of multifactor models of short rate dynamics is in Gallant and Tauchen (1998). Other term structure applications include Martin and Pagan (2000) along with Dungey, Martin, and Pagan (2000), who undertake a factor analysis of bond yield spreads.

Some interesting recent applications to microeconomic problems include Nagypal (2007), who uses the method to estimate and compare various models of learning by doing. Her scores are not SNP scores, which, indeed, would be inappropriate in her application. The referee argues that her auxiliary model

may not approximate the true data generating process closely enough for a claim of efficiency. Austin and Katzman (2001) apply the method to estimate and test new models of multi-step auctions using tobacco auction data.

7 Software and Practical Issues

7.1 Code

In this section we first describe methods that are appropriate when a high quality implementation of the BFGS algorithm is used. The BFGS algorithm (Fletcher, 1987) works best when analytical first derivatives of the objective function can be supplied, which is the case with EMM when using an SNP auxiliary model. The BFGS algorithm uses a rank two update scheme to compute a “Hessian” matrix, which is needed to get quadratic convergence. Therefore second derivatives are not required. Next we discuss the MCMC algorithm proposed by Chernozukov and Hong (2003). Its advantages are that it is not as dependent on start values for success, computing sandwich variance matrices becomes feasible, and it can cope with the jitter inherent in estimating jump diffusion models. Its disadvantage is that it can be more computationally intensive.

A Fortran program that implements the BFGS algorithm is available at <http://econ.duke.edu/webfiles/get/emm>. A C++ program implementing the Chernozukov-Hong method is available at <http://econ.duke.edu/webfiles/arg/emm>. A User’s Guide is included with the code as well as the SNP code and an SNP User’s Guide. The C++ code is distributed in both a serial version and a parallel version that runs under MPI (Foster, 1995). The C++ program is actually a general purpose implementation of the Chernozukov-Hong method that can be used with maximum likelihood estimation or other statistical objective functions. It can also be used for Bayesian inference and we remark in passing that the EMM objective function can be used for Bayesian inference in place of a likelihood (Gallant and Hong, 2007).

As supplied the code presumes a CASE 2 structural model in the nomenclature of Gallant and Tauchen (1996). That is the case that we shall describe here. The setup subsumes a wide variety of situations in macroeconomics and finance. The SNP model is the score generator. The code can easily be modified to accommodate other score generators and to accommodate covariates, as in CASE 1 or CASE 3 of Gallant and Tauchen (1996). While we do our work in Unix (Linux or Mac OS), EMM will run under Microsoft Windows. Running under different operating systems is discussed in the Guides.

7.2 Troubleshooting, Numerical Stability, and Convergence Problems

On the whole, the EMM package is useful and practical. An early version was used for estimating asset pricing models (Bansal *et al.*, 1993, 1995). Recent versions of the Fortran package have been used in several applications including, among others, Chernov, Gallant, Ghysels, and Tauchen (2001) and Gallant and Tauchen (2001a) for stochastic volatility, Gallant and Long (1997), Chung and Tauchen (2001) for exchange rate modeling, and Dai and Singleton (2000), Ahn, Dittmar, Gallant, (2001), and Tauchen (1997) for interest rates.

Things can go awry, however. Sometimes, the program may stop prematurely, and there are some key issues of dynamic and numerical stability that the user must be attentive to. These issues affect the speed of the computations and relate to convergence problems in the nonlinear optimization. The following discussion pertains to these issues.

7.2.1 Start Value Problems and Scaling

Sometimes it is hard to get decent start values. We suggest intensive use of randomly perturbed start values. The nonlinear optimizer works best if it sees all parameters as roughly the same order of magnitude. We adopt a scaling so that all parameters as seen by the optimizer lie in the interval (-1,1). Since this scaling may not be the natural scaling for the data generator, one might want to adapt the user supplied portion of the code so that the rescaling is done automatically, as in the logliner example distributed with the Fortran EMM code. We find the proper scaling mitigates many problems and accelerates convergence.

7.2.2 Enforcing Dynamic Stability

As noted in Section 3 above, the score generator should be dynamically stable. The SNP package incorporates a spline and/or logistic transformation feature that directly enforces dynamic stability on the score generator. This feature is discussed at length in the SNP User's Guide (Gallant and Tauchen, 2001c). The transformations only affect the conditioning variable x_{t-1} in the conditional density $f(y_t|x_{t-1}, \theta)$; it has no effect on y_t and it is not a prefiltering of the data. All it does is force a very gentle sort mean reversion so that $(\partial/\partial\theta) \log[f(\hat{y}_\tau|\hat{x}_{\tau-1}, \theta)]$ remains well defined should the optimizer happen to pass back a parameter vector ρ such that the simulation $\{\hat{y}_\tau(\rho), \hat{x}_{\tau-1}(\rho)\}$ is explosive. For series that are very persistent, such as interest rates, we find the spline transformation the best while for series that are nearly *iid*, e.g., stock returns series, we recommend using the logistic transformation instead of the spline transformation. As explained in the SNP User's Manual, the logistic

really serves a different purpose than the spline. The logistic prevents large elements of x_{t-1} from unduly influencing the conditional variance computation.

7.2.3 *Bullet Proofing the DGP*

Recall the basic structure of EMM as outlined in Subsection 2.2 above. The core component of the distributed EMM package is the user-supplied simulator that takes as input a candidate vector ρ and generates a simulated realization. This component computes the mapping $\rho \rightarrow \{\hat{y}_t\}_{t=1}^N$. The EMM package evaluates the objective function

$$s_n(\rho) = m'_n(\rho, \tilde{\theta}_n)(\tilde{\mathcal{I}}_n)^{-1}m_n(\rho, \tilde{\theta}_n)$$

and optimizes it with respect to ρ .

The optimizer should see $s_n(\rho)$ as a smooth surface and care should be taken in writing the DGP code to ensure small perturbations of ρ lead to small perturbations of $s_n(\rho)$. The most common source of a rough surfaces is the failure to control Monte Carlo jitter. One must ensure that when ρ changes the random numbers used to compute $\{\hat{y}_t\}_{t=1}^N$ do not change. Usually taking care that the seeds passed to random number generators do not change when ρ changes is an adequate precaution. However, as mentioned in connection with the discussion of Anderson, Benzoni and Lund (2001), additional precautions may be necessary when adding jumps or other discrete elements to simulated paths. Large values for N also contribute to smoothness.

Our experience is that the optimizer sometimes tries outlandishly extreme values of ρ , especially in the initial phase of the optimization when it's acquiring information on functional form of the objective function. These outlandish values of ρ could entail taking the logs or square roots of negative numbers, dividing by zero, or undertaking other operations that generate numerical exceptions, either within the user's simulator, within SNP (which evaluates to scores), or even within the optimizer itself. Our experience is that things proceed most smoothly when the user-supplied simulator can generate some kind of sensible simulated realization regardless of ρ and be able to compute something for $\rho \rightarrow \{\hat{y}_t\}_{t=1}^N$ given arbitrary ρ . We call this "bullet proofing" the code.

However, sometimes it is extremely difficult to bullet proof completely the simulator (especially for diffusion models) and numerical exceptions occur that generate NaN's. On a Unix workstation, the Fortran compiler usually has produced an executable that can appropriately propagate the NaN's and the EMM objective function evaluates to either Inf or NaN. Typically, the optimizer distributed with the code can recover, as it realizes that the particular value of ρ

that led to the disaster is very unpromising and it tries another. The cost of this is that the program slows down considerably while handling the numerical exceptions along a very long simulated realization.

7.3 The Chernozukov-Hong Method

The computational methods discussed here and implemented by the C++ implementation of the EMM package apply to any discrepancy function $s_n(\rho)$ that produces asymptotically normal estimates; i.e., any discrepancy function for which there exist ρ^o , \mathcal{I} and \mathcal{J} such that

$$\mathcal{J}\sqrt{n}(\hat{\rho}_n - \rho^o) = \sqrt{n}\frac{\partial}{\partial\rho}s_n(\rho) + o_p(1) \text{ and } \sqrt{n}\frac{\partial}{\partial\rho}s_n(\rho) \xrightarrow{\mathcal{L}} N(0, \mathcal{I}) \quad (43)$$

The \mathcal{I} matrix discussed in this subsection pertains to $\hat{\rho}_n$ and is not the $\tilde{\mathcal{I}}_n$ weighting matrix of the EMM auxiliary model.

Quasi maximum likelihood estimation requires the computation of the estimator itself, $\hat{\rho}_n = \underset{\rho}{\operatorname{argmin}} s_n(\rho)$, an estimate of the Hessian

$$\mathcal{J} = \frac{\partial}{\partial\rho\partial\rho'} s^o(\rho^o),$$

where $s^o(\rho) = \lim_{n \rightarrow \infty} s_n(\rho)$, and an estimate of Fisher's information

$$\mathcal{I} = \operatorname{Var} \left[\frac{\partial}{\partial\rho'} \sqrt{n} s_n(\rho^o) \right] = \mathcal{E} \left[\frac{\partial}{\partial\rho'} \sqrt{n} s_n(\rho^o) \right] \left[\frac{\partial}{\partial\rho'} \sqrt{n} s_n(\rho^o) \right]'$$

The variance of $\sqrt{n}(\hat{\rho}_n - \rho^o)$ is then of the sandwich form

$$V_n = \operatorname{Var} [\sqrt{n}(\hat{\rho}_n - \rho^o)] = \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}$$

Put $\ell(\rho) = e^{-n s_n(\rho)}$. Apply Bayesian MCMC methods with $\ell(\rho)$ as the likelihood. From the resulting MCMC chain $\{\rho_i\}_{i=1}^R$ put

$$\hat{\rho}_n = \bar{\rho}_R = \frac{1}{R} \sum_{t=1}^R \rho_i \text{ and } \hat{\mathcal{J}}^{-1} = \left(\frac{n}{R} \right) \sum_{t=1}^R (\rho_i - \bar{\rho}_R) (\rho_i - \bar{\rho}_R)'$$

Alternatively, and definitely for EMM, use the mode of $\ell(\rho)$ as the estimator $\hat{\rho}_n$. The EMM package computes and reports both the mean and the mode.

Actually, the mode is the better choice of an estimator in most applications because the parameter values in the mode actually have generated a simulation. The parameter values in the mean vector may not even satisfy the support conditions of the structural model.

The strategy used to estimate \mathcal{I} is the following. For ρ set to the mode, simulate the model, and generate I approximately independent bootstrap data sets $\{\hat{y}_{t,i}\}_{t=1}^n$, $i = 1, \dots, I$, each of exactly the same sample size n as the original data. Keeping the size to exactly n and using model simulations makes the estimator below a heteroskedastic autocovariance consistent (HAC) estimator. Keeping the size to exactly n does not imply that the simulation size N should be set to n when using the program. The simulation size N should be set much larger than n in most instances. One way to get a bootstrap sample is to split this long simulation into blocks of size n . With this approach, the estimate of \mathcal{I} would be a parametric bootstrap estimate. Alternatively, stationary bootstrap or some other method could be used to construct the blocks. The bootstrap generating mechanism is coded by the user.

Let $\hat{s}_{n,i}(\rho)$ denote the criterion function corresponding to the i th bootstrap data set $\{\hat{y}_{t,i}\}_{t=1}^n$ and let $\hat{\rho}_n$ denote mode of $\ell(\rho)$. Compute $\frac{\partial}{\partial \rho'} \sqrt{n} \hat{s}_{n,i}(\hat{\rho}_n)$ numerically. An estimate of the information matrix is the average

$$\hat{\mathcal{I}} = \frac{1}{I} \sum_{i=1}^I \left[\frac{\partial}{\partial \rho'} \sqrt{n} \hat{s}_{n,i}(\hat{\rho}_n) \right] \left[\frac{\partial}{\partial \rho'} \sqrt{n} \hat{s}_{n,i}(\hat{\rho}_n) \right]' \quad (44)$$

Note that for the EMM estimator one must compute the likelihood of the auxiliary model from the i th bootstrap sample and optimize it in order to get the i th EMM objective function $\hat{s}_{n,i}(\rho)$. This is done using the BFGS method. This is the step that makes computing an accurate numerical derivative accurately both difficult and costly for EMM. The code attempts to detect failure of the optimizer and failure to compute an accurate derivative and discard those instances. An objective function such as maximum likelihood or GMM that does not rely on a preliminary optimization is not as much of a challenge to differentiate numerically. With these one can have more confidence that the code provides the correct answer.

If the SNP model is a good approximation to the true data generating process, the computation of $\hat{\mathcal{I}}$ is not necessary because $\mathcal{I} = \mathcal{J}$. This issue has been discussed extensively above. The same is true for maximum likelihood if the model is correctly specified.

The code provides the option of putting the parameter ρ on a grid. This increases speed by allowing $s_n(\rho)$ and related variables to be obtained by table lookup thus avoiding recomputation for a value of ρ that has already been visited in the MCMC chain. This is a useful feature when the objective function $s_n(\rho)$ is costly to compute.

Note that $S_n(\rho) = \tau s_n(\rho)$ is also a valid criterion according to the theory. This gives one a temperature parameter τ to use for tuning the chain. This feature is implemented in the package.

A random walk, single move, normal proposal is the workhorse of the C++ EMM package. When parameters are put on a grid, a discrete proposal density is used instead that has probabilities assigned to grid points proportionally to this normal. Group moves are also supported. It is easy to substitute an alternative proposal density.

Another advantage of putting parameters on a grid is that it allows the statistical objective function to be computed less accurately because the accept-reject decision is still likely to be correct when parameter values are well separated despite an inaccurately computed objective function, within reason. This helps mitigate against the effects of jitter discussed above. Also, it will allow smaller values of the simulation size N than hill climbers require. For small values of N one should probably multiply variance estimates by the correction factor $(1+n/N)$ discussed in Gourieroux, Monfort, and Renault (1993).

Simulated method of moments is exactly the same as the foregoing but with a GMM criterion replacing $s_n(\rho)$. As with EMM, if the correct weight function is used with the GMM criterion function, then $\mathcal{I} = \mathcal{J}$ so that \mathcal{I} need not be computed and there is no need for any numerical differentiation. But often the effectiveness of the GMM weighting function is doubtful and it can cloud the interpretation of results. One may prefer sandwich standard errors regardless. With GMM there is usually no numerical optimization to compute moments as with EMM so better accuracy can be expected.

The MCMC method described here makes the imposition of support restrictions, inequality restrictions, and informative prior information exceptionally convenient. These restrictions and prior information can be imposed on model parameters or on (nonlinear) functionals of the model that can only be known via simulation. This feature is implemented in the EMM package. As mentioned above, when the EMM criterion function is used in connection with a prior the results can be given a Bayesian interpretation (Gallant and Hong, 2007).

Obviously these ideas are not restricted to simulation estimators. The EMM package is a general purpose implementation of the Chernozhukov-Hong estimator. An illustration of how the code may be used to implement maximum likelihood is included with the package and described in the User's Guide. The application used for this illustration is a translog consumer demand system for electricity by time of day with demand shares distributed as the logistic normal that is taken from Gallant (1987). It shows off the Chernozhukov-Hong estimator to good advantage because a vexing problem with hill climbers is trying to keep model parameters in the region where predicted shares are positive for every observed price/expenditure vector. This is nearly impossible to achieve when using conventional derivative based hill climbing algorithms but is trivially easy to achieve using the the Chernozhukov-Hong estimator.

8 Conclusion

We described a simulated score method of moments estimator based on the following idea: Use the expectation with respect to the structural model of the score function of an auxiliary model as the vector of moment conditions for GMM estimation. Making the procedure operational requires an estimate of the parameters of the auxiliary model and computation of the expectation via simulation. Strategies for doing this were set forth, considerations regarding choice of the auxiliary model were discussed, and the SNP density, which is a sieve, was described as a general purpose auxiliary model. When the auxiliary model is chosen to closely approximate the characteristics of the observed data, the estimation method is termed efficient method of moments (EMM). The SNP density provides a systematic method to achieve a close approximation, though, depending on the nature of the data, other auxiliary models might provide a more convenient way to achieve adequate approximation for EMM.

These ideas were related to indirect inference, which is an asymptotically equivalent methodology, and mention made of the fact that the indirect inference view of the method can be used to facilitate the choice of an auxiliary model that confers seminonparametric or robustness properties on the estimator. Also mentioned was that, as a practical matter, indirect inference will often have to be reformulated as a simulated score method to make it computationally feasible.

There are three steps to EMM. The first, termed the Projection Step, entails summarizing the data by projecting it onto the auxiliary model. The second is the Estimation Step, where the parameters are obtained by GMM. The estimation step produces an omnibus test of specification along with useful diagnostic t statistics. The third step is termed the Reprojection Step, which entails post-estimation analysis of simulations for the purposes of prediction, filtering, and model assessment. It was argued that the last two steps, assessment of model adequacy, and post estimation evaluation, are the real strengths of the methodology in building scientifically valid models.

There have been numerous applications of the EMM methodology in the literature and several of these were discussed in detail. Code is available, its use was broadly discussed with attention given to various pitfalls that need to be avoided. The code is available for both serial and parallel architectures.

References

- [1] Aguirre-Torres, V. (2001) “Local behavior of the efficient method of moments in iid models,” Technical Report DE-C01.3, Division Academica de Actuaría, Estadística Matemáticas, Instituto Tecnológico Autónomo de México (ITAM), Río Hondo #1, México DF 01000, México.
- [2] Aguirre-Torres, V., and A. R. Gallant (2001) “Local behavior of the efficient method of moments in dynamic models,” Technical Report DE-C01.6, Division Academica de Actuaría, Estadística Matemáticas, Instituto Tecnológico Autónomo de México (ITAM), Río Hondo #1, México DF 01000, México.
- [3] Ahn, D-H., R. F. Dittmar and A. R. Gallant (2002) “Quadratic term structure models: theory and evidence,” *The Review of Financial Studies* 15, 243–288.
- [4] Ahn, D-H, R. F. Dittmar, A. R. Gallant and B. Gao (2002) “Purebred or hybrid?: reproducing the volatility in term structure dynamics,” *Journal of Econometrics* 116, 147–180.
- [5] 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.
- [6] 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.
- [7] Andersen, T. G., L. Benzoni and J. Lund (2002) “Towards an empirical foundation for continuous-time equity return models,” *Journal of Finance* 57, 1239–1284.
- [8] Andrews, D. W. K. (1991) “Heteroskedasticity and autocorrelation consistent covariance matrix estimation,” *Econometrica*, 59, 307–346.
- [9] Austin, A. and B. Katzman (2001) “Testing a Model of Multi-Unit Bidder Demands Using Auction Data,” working paper, University of Miami.
- [10] Balduzzi, P., S. R. Das, S. Foresi and R. K. Sundaram (1996) “A simple approach to three factor affine term structure Models,” *Journal of Fixed Income*, 6, 43–53.
- [11] Bansal, R and A. Yaron (2004) “Risks for the long run: a potential resolution of asset pricing puzzles,” *Journal of Finance* 59 1481–1509.
- [12] Bansal, R. and H. Zhou (2002) “Term structure of interest rates with regime shifts,” *Journal of Finance* 57, 1997–2043.
- [13] Bansal, R., A. R. Gallant, R. Hussey and G. Tauchen (1993) “Computational aspects of nonparametric simulation estimation”, in: David A. Belsley, (ed.), *Computational techniques for econometrics and economic analysis* Boston: Kluwer Academic Publishers, 3–22.

- [14] Bansal, R., A. R. Gallant, R. Hussey and G. Tauchen (1995) “Nonparametric estimation of structural models for high-frequency currency market data”, *Journal of Econometrics*, 66, 251–287.
- [15] Bansal, Ravi, A. Ronald Gallant, and George Tauchen (2007), “Rational Pessimism, Rational Exuberance, and Asset Pricing Models,” *Review of Economic Studies*, 74, 1005–1033.
- [16] Bollerslev, T. (1986) “Generalized autoregressive conditional heteroskedasticity,” *Journal of Econometrics*, 31, 307–327.
- [17] Bollerslev, T. and R. F. Engle (1993) “Common persistence in conditional variances,” *Econometrica*, 61, 167–186.
- [18] Chen, L. (1996) “Stochastic mean and stochastic volatility: a three factor model of the term structure of interest rates,” Oxford, U.K.: Blackwell Publishers.
- [19] Chernov, M., A. R. Gallant, E. Ghysels and G. Tauchen (1999) “A new class of stochastic volatility models with jumps: theory and estimation,” Working Paper, University of North Carolina-Chapel Hill.
- [20] Chernov, M., A. R. Gallant, E. Ghysels and G. Tauchen (2001) “Alternative models for stock price dynamics,” *Journal of Econometrics* 116, 225–257.
- [21] Chernov, M. and E. Ghysels (2000) “A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options Valuation,” *Journal Of Financial Economics*, 56, 407-458.
- [22] Chernozhukov, Victor, and Han Hong (2003), “An MCMC approach to classical estimation,” *Journal of Econometrics* 115, 293–346.
- [23] Chumacero, R. (1997) “Finite sample properties of the efficient method of moments,” *Studies in Nonlinear Dynamics and Econometrics* 2, 35–51.
- [24] Chung, C-C., and G. Tauchen (2001) “Testing target zone models using efficient method of moments,” *Journal of Business and Economic Statistics*, 19, 255–269.
- [25] Coppejans, M. and A. R. Gallant (2000) “Cross validated SNP density estimates,” *Journal of Econometrics* 110, 27–65.
- [26] Dai, Q. and K. J. Singleton (2000) “Specification analysis of affine term structure models,” *Journal of Finance* LV 5: 1943–1978.
- [27] Danielsson, J. and F. Penaranda (2007) “On the impact of fundamentals, liquidity, and coordiantaion on market stability,” Working paper, London School of Economics, available at <http://www.RiskResearch.org>.
- [28] Davidian, M. and R. J. Carroll (1987) “Variance function estimation,” *Journal of the American Statistical Association*, 82, 1079–1091.
- [29] Del Negro, Marco, and Frank Schorfheide (2004), “Priors from General Equilibrium Models for VARS,” *International Economic Review* 45, 643–673.

- [30] de Luna, Xavier, and Marc G. Genton (2002), “Simulation-based inference for simultaneous processes on regular lattices,” *Statistics and Computing* 12, 125–134.
- [31] Dridi, R. and E. Renault (1998) “Semiparametric indirect inference,” Manuscript, Toulouse.
- [32] Delgado, F. and B. Dumas, (1991), “Target zones, broad and narrow”, in: P.Krugman and M.Miller, editors, {Exchange Rate Targets and Currency Bands} Cambridge University Press, Cambridge.
- [33] Duffie, D. and K. J. Singleton (1993) “Simulated moments estimation of Markov models of asset prices,” *Econometrica*, 61, 929–952.
- [34] Dungey, M, V. L. Martin, and A. R. Pagan (2000), “A Multivariate Latent Factor Decomposition of International Bond Yield Spreads,” *Journal of Applied Econometrics*, 15, 697–715..
- [35] Engle, R. F. (1982) “Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation,” *Econometrica*, 50, 987–1007.
- [36] Engle, R. F, and K. F. Kroner (1995), “Multivariate Simultaneous Generalized ARCH,” *Econometric Theory* 11, 122–150.
- [37] Engle, R. F. and G. G. J. Lee (1996) “Estimating diffusion models of stochastic volatility,” in Peter E. Rossi, (ed), *Modeling Stock Market Volatility: Bridging the Gap to Continuous Time* (New York: Academic Press, 1996), 333–384.
- [38] Engle, R. F., T. Ito and W-L Lin (1990) “Meteor showers or heat waves? heteroskedastic intra-daily volatility in the foreign exchange market,” *Econometrica*, 58, 525–542.
- [39] Fletcher, R. (1987) *Practical methods of optimization, second edition*, Wiley, New York.
- [40] Foster, D. and S. Viswanathan (1995) “Can speculative trading explain the volume-volatility relation?” *Journal of Business and Economic Statistics*, 13, 379–398.
- [41] Fisher, R. A. (1912) “On an absolute criterion for fitting frequency curves,” *Messages of Mathematics*, 41, 155–157.
- [42] Foster, I. (1995) *Designing parallel programs* Addison-Wesley, Boston. Available online at: <http://www.mcs.anl.gov/dbpp>
- [43] Gallant, A. R. (1980) “Explicit estimators of parametric functions in nonlinear regression,” *Journal of the American Statistical Association*, 75, 182–93.
- [44] Gallant, A. R. (1987) *Nonlinear Statistical Models*. New York, NY: John Wiley and Sons.

- [45] Gallant, A. Ronald, and Han Hong (2007), “A Statistical Inquiry into the Plausibility of Recursive Utility,” *Journal of Financial Econometrics*, 5, 523–590.
- [46] Gallant, A. R., D. A. Hsieh and G. Tauchen (1991) “On fitting a recalcitrant series: the Pound/Dollar exchange rate, 1974–83,” in W. A. Barnett, J. Powell, and 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.
- [47] Gallant, A. R., D. A. Hsieh and G. Tauchen (1997) “Estimation of stochastic volatility models with diagnostics,” *Journal of Econometrics*, 81, 159–192.
- [48] 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.
- [49] Gallant, A. R., and R. E. McCulloch. (2009). “On the Determination of General Statistical Models with Application to Asset Pricing.” *Journal of the American Statistical Association*, forthcoming.
- [50] Gallant, A. R., and J. Long (1997) “Estimating stochastic differential equations efficiently by minimum chi-squared,” *Biometrika*, 84, 125–141.
- [51] Gallant, A. R., P. E. Rossi and G. Tauchen (1993) “Nonlinear dynamic structures,” *Econometrica*, 61, 871–907.
- [52] Gallant, A. Ronald and D. W. Nychka (1987) “Semi-nonparametric maximum likelihood estimation,” *Econometrica*, 55, 363–390.
- [53] Gallant, A. R. and G. Tauchen (1989) “Semi-nonparametric estimation of conditionally constrained heterogeneous processes: asset pricing applications,” *Econometrica*, 57, 1091–1120.
- [54] Gallant, A. R. and G. Tauchen (1992) “A nonparametric approach to nonlinear time series analysis: estimation and simulation” in E. Parzen, D. Brillinger, M. Rosenblatt, M. Taqqu, J. Geweke and P. Caines (eds.), *New dimensions in time series analysis*. New York: Springer-Verlag.
- [55] Gallant, A. R. and G. Tauchen (1996) “Which moments to match?” *Econometric Theory*, 12, 657–681.
- [56] Gallant, A. R. and G. Tauchen (1997) “Estimation of continuous time models for stock returns and interest rates,” *Macroeconomic Dynamics*, 1, 135–168.
- [57] 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.
- [58] Gallant, A. R. and G. Tauchen (1999) “The relative efficiency of method of moments estimators,” *Journal of Econometrics*, 92, 149–172.

- [59] Gallant, A. R. and G. Tauchen (2001a) “Efficient method of moments,” Manuscript, Duke University.
(<http://econ.duke.edu/webfiles/arg/papers/ee.pdf>)
- [60] Gallant, A. R. and G. Tauchen (2001b) “EMM: a Program for efficient method of moments estimation, a user’s guide,” Manuscript, Duke University.
(Available along with code at <http://econ.duke.edu/webfiles/arg/emm>)
- [61] Gallant, A. R. and G. Tauchen (2001c) “SNP: a program for nonparametric time series analysis, a user’s guide,” Manuscript, University of North Carolina.
(Available along with code via <http://econ.duke.edu/webfiles/arg/snp>)
- [62] Gallant, A. R. and H. White (1987) *A unified theory of estimation and inference for nonlinear dynamic Models* Oxford: Basil Blackwell Ltd.
- [63] Gauss, C. F. (1816) “Bestimmung der genauigkeit der beobachtungen,” *Zeitschrift für Astronomie und verwandte Wissenschaften* 1, 185–196.
- [64] Gennotte, G. and Marsh, T. A. (1993) “Variations in economic uncertainty and risk premiums on capital assets,” *European Economic Review*, 37, 1021–41.
- [65] Genton, Marc G. and Xavier de Luna (2000), “Robust simulation-based estimation,” *Statistics & Probability Letter* 48, 253–259.
- [66] Geweke, J. (1983) “The approximate slope of econometric tests,” *Econometrica*, 49, 1427–1442.
- [67] Gourieroux, C., A. Monfort and E. Renault (1993) “Indirect inference,” *Journal of Applied Econometrics*, 8, S85–S118.
- [68] Hansen, L. P. (1982) “Large sample properties of generalized method of moments estimators,” *Econometrica*, 50, 1029–1054.
- [69] Hansen, L. P. (2002) “Generalized method of moments estimation: a time series perspective,” *International Encyclopedia of Social and Behavioral Sciences*. N. J. Smelser and P. B. Bates (editors), Pergamon, Oxford.
- [70] Hansen, L. P. and J. Scheinkman (1995) “Back to the future: generating moment implications for continuous-time markov processes,” *Econometrica*, 63, 767–804.
- [71] Ingram, B. F. and B. S. Lee (1991) “Simulation estimation of time series models,” *Journal of Econometrics*, 47, 197–250.
- [72] Jiang, G.J. and P. J. van der Sluis (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.
- [73] Johnson, N. L., and Kotz, S. (1970) *Distributions in statistics: continuous univariate distributions-1*, New York: Wiley.
- [74] Kloeden, P. E. and E. Platen (1992) *Numerical solution of stochastic differential equations*, New York: Springer-Verlag.

- [75] Lo, A. W. (1988) “Maximum likelihood estimation of generalized Ito process with discretely sampled data,” *Econometric Theory*, 4, 231–247.
- [76] Martin, V. L. and A. R. Pagan (2000) “Simulation based estimation of some factor models in econometrics,” in R. Mariano and T. Schuermann (Eds), *Simulation Based Estimation in Econometrics*, (Cambridge: Cambridge University Press) 235–254.
- [77] McManus D., and D. Watt (1999) “Estimating one factor models of short-term interest rates,” Working Paper 99-18, Bank of Canada, Ottawa, Canada.
- [78] Marron, J. S. and M. P. Wand (1992) “Exact mean integrated squared error,” *The Annals of Statistics*, 20, 712–736.
- [79] Mixon, O. S. (1998) “Essays on financial market volatility,” Ph.D. Thesis, Duke University.
- [80] Nagypal, E. (2007) “Learning-by-doing versus selection: can we tell them apart?,” *Review of Economic Studies* 74, 537–566.
- [81] Nelson, D. (1991) “Conditional heteroskedasticity in asset returns: a new approach,” *Econometrica*, 59, 347–370.
- [82] 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.
- [83] 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.
- [84] Newey, W. K. and K. D. West (1987) “A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix estimator,” *Econometrica*, 55, 703–708.
- [85] Pagan, A. R., A. D. Hall and V. L. Martin (1996) “Modeling the term structure,” in G. S. Maddala and C. R. Rao (Eds), *Statistical Methods in Finance*, Handbook of Statistics, 13, 91–118.
- [86] Pagan, A. (1999) “Some Uses of Simulation in Econometrics,” *Mathematics and Computers in Simulation*, 48, 341–349.
- [87] Pastorello, S, E. Renault and N. Touzi (2000) “Statistical inference for random variance option pricing,” *Journal of Business and Economic Statistics*, 18, 358–367.
- [88] Pakes, Ariel and Pollard, David (1989) “Simulation and the asymptotics of optimization estimators,” *Econometrica*, 57, 1027–1058.
- [89] Pearson, K. (1894) “Contributions to the mathematical theory of evolution,” *Philosophical Transactions of the Royal Society of London, Series A*, 185, 71–78.

- [90] Pedroni, P. (2001) “Comment on Chung and Tauchen,” *Journal of Business and Economic Statistics*, 19, 271–273.
- [91] Schwarz, G. (1978) “Estimating the dimension of a model,” *The Annals of Statistics*, 6, 461–464.
- [92] Sims, C. A. (1980) “Macroeconomics and reality,” *Econometrica*, 48, 1–48.
- [93] Smith, A. A. (1990) “Three essays on the solution and estimation of dynamic macroeconomic models,” Ph.D. Dissertation, Duke University, Durham, NC.
- [94] Smith, A. A. (1993) “Estimating nonlinear time series models using simulated vector autoregressions,” *The Journal of Applied Econometrics* 8, S63–S84.
- [95] Tauchen, G. (1985) “Diagnostic testing and evaluation of maximum likelihood models,” *Journal of Econometrics*, 30, 415–443.
- [96] 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.
- [97] 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.
- [98] Tauchen, G. and R. Hussey (1991) “Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models”, *Econometrica*, 59, 371–396.
- [99] Taylor, J. B. and H. Uhlig (1990) “Solving nonlinear stochastic growth models: a comparison of alternative solution methods,” *Journal of Business and Economic Statistics*, 8, 1–17.
- [100] Valderrama, D. (2001), “Can a standard real business cycle model explain the nonlinearities in U.S. national accounts data?” Ph.D. thesis document, Department of Economics, Duke University.
- [101] 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.
- [102] 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.
- [103] White, H. (1994) *Estimation, inference and specification analysis*. Cambridge University Press.
- [104] Zhou, H. (2001) “Finite sample properties of EMM, GMM, QMLE, and MLE for a square-root interest rate diffusion model” *Journal of Computational Finance* 5, 89–122.