

The Continuum-GMM Estimation: Theory and Application

Rachidi Kotchoni*

Marine Carrasco†

Université Paris Nanterre

Université de Montréal

November 4, 2017

Abstract

By avoiding discretization, the Generalized Method of Moment based on a Continuum of moment conditions (CGMM) permits to efficiently use the information content of a continuum moment restrictions. When the moment restrictions are deduced from a characteristic function, the CGMM has the potential to achieve the maximum likelihood efficiency. This chapter reviews the theory underlying the CGMM procedure, discusses the properties of the CGMM estimator and presents numerical algorithms for its implementation. An empirical application is proposed where a Variance Gamma model is fitted to the monthly increments of the USD/GBP exchange rates. We find that the variance forecasts inferred from the Variance Gamma model are of poor quality. A model that specifies the variance as a dependent process should deliver better forecasts.

Keywords: Characteristic Function, Continuum of Moment Conditions, Generalized Method of Moments, Quadrature Methods, Regularization, Variance Gamma Model. JEL Classification: C00, C13, C15

*Corresponding Author: Associate Professor (Maître de Conférence) of Economics. Université Paris Nanterre, CNRS-EconomiX. Email: rachidi.kotchoni@parisnanterre.fr

†Professor of Economics. Université de Montréal, CIRANO and CIREQ. Email: marine.carrasco@umontreal.ca

1 Introduction

Many econometrics models are naturally specified in terms of their characteristic function (CF) as their densities are unknown in closed form. This is the case for most stable distributions¹ and discretely sampled continuous-time processes². For some models, the density takes the form of an infinite mixture (e.g., autoregressive Gamma processes; see [Gourieroux and Jasiak, 2005](#)) or an integral (e.g., variance Gamma processes; see [Madan and Seneta, 1990](#)). For other models, the density has a rather simple analytical expression but the likelihood function is numerically ill-behaved in finite sample (e.g., finite mixtures). In these models, parameter estimation may be done via the continuum of moment conditions that is obtained by taking the difference between the theoretical and empirical CF. As there is a one-to-one relationship between the CF and the probability distribution function of a random variable, an inference procedure based on the CF has the potential to be as efficient as one that is based on the likelihood function.

Indeed, many authors have relied on the CF to conduct parametric inference in various contexts. For instance, [Paulson et al. \(1975\)](#) used a weighted modulus of the difference between the theoretical CF and its empirical counterpart to estimate the parameters of a stable distribution. [Feuerverger and Mureika \(1977\)](#) studied the convergence properties of the empirical CF and suggested that it may be useful for tackling numerous statistical problems. Since then, various GMM-like estimators that are based on the CF have been advocated by several authors, including [Feuerverger and McDunnough \(1981a, 1981b, 1981c\)](#), [Carrasco and Florens \(2000\)](#), [Singleton \(2001\)](#), [Jiang and Knight \(2002\)](#), [Knight and Yu \(2002\)](#), [Chacko and Viceira \(2003\)](#), [Yu \(2004\)](#) and [Carrasco, Chernov, Florens, and Ghysels \(2007\)](#).

The Continuum-GMM (henceforth, CGMM) procedure of [Carrasco and Florens \(2000\)](#) closely mimics the efficient two-steps GMM of [Hansen \(1982\)](#). It is obtained by minimizing a quadratic form in a Hilbert space that spans the continuum of moment condition of interest. As the standard GMM estimator, it is implemented in two steps, the first step leading to a consistent estimator and the second step delivering an efficient estimator. The objective function of the efficient CGMM is a quadratic form

¹The density of α -stable distributions is known in closed form only at isolated points of the parameter space. [Nolan \(2016\)](#) proposes numerical methods to approximate the likelihood of α -stable models.

²The density of a discretely sampled continuous time process is known for the special case of a Square-Root diffusion (see [Zhou, 1996](#); [Singleton, 2001](#)). [Ait-Sahalia and Kimmel \(2007\)](#) propose closed form approximations to the log-likelihood function of certain continuous-time stochastic volatility models.

with metrics K^{-1} , where K is the asymptotic covariance operator associated with the continuum of moment conditions of interest. To obtain a feasible second-step CGMM estimator, one has to replace K by a sample analogue K_T , where T is the sample size. However, K_T is not invertible and its limit as T goes to infinity, K , is not continuous. To circumvent this difficulty, Carrasco and Florens (2000) use a regularized inverse of the form $K_{\lambda T}^{-1} = (K_T^2 + \lambda I)^{-1} K_T$, where I is the identity operator and λ is a regularization parameter. The resulting estimator is root- T consistent and asymptotically normal for any fixed λ . Asymptotic efficiency is obtained by letting T go to infinity and λ go to zero at a specific rate.

This chapter reviews the theory underlying the CGMM procedure, discusses the properties of the CGMM estimator and presents numerical algorithms for its implementation. An empirical application is proposed where a Variance Gamma model is fitted to the monthly increments of the USD/GBP exchange rates. This model is appealing for exchange rates because it features both mean reversion and heteroscedasticity. However, it has limited scope as it specifies the latent variance of the exchange rate increments as an independent and identically distributed (IID) process. Consequently, time varying predictions of the variance can only be obtained as its expectation conditional on the ex-post realizations of exchange rate increments. We find that these posterior variances are poor predictions of the monthly realized variances inferred from daily data. A model that specifies the variance as a dependent process should deliver better forecasts than the Variance Gamma model considered here.

The remainder of the presentation is organized as follows. In Section 2, we show how to construct a continuum of moment conditions in five different contexts. Section 3 presents a review of estimation methods that are based on the CF. In Section 4, we review the construction of the objective function of the CGMM procedure and discuss the asymptotic properties of the CGMM estimator. In Section 5, we discuss the selection of the regularization parameter, λ , in practice. Section 6 presents other implementation issues and proposes algorithmic solutions. Section 7 presents the empirical application and Section 8 concludes. A short appendix presents the derivation of the prediction formula inferred from the variance Gamma process for the latent variance process.

2 Moment Conditions Based on the Characteristic Function

This section presents five empirically relevant situations that motivate the use of a continuum of moment condition for parametric estimation. Although most cases involve the use of a CF, an important example concerns the conversion of a conditional moment restriction into a continuum of unconditional moment restrictions.

2.1 Finite Mixtures

Consider a random variable y_t that may have been generated by a distribution $f_j(y_t, \theta_j)$ with probability $\omega_j, j = 1, \dots, J$. These distributions can be viewed as linked with a discrete state variable $s_t \in \{1, \dots, J\}$ so that:

$$\Pr(s_t = j) = \omega_j, \quad j = 1, \dots, J, \quad \text{and} \quad f(y_t | s_t = j) = f_j(y_t, \theta_j).$$

Under this convention, the marginal density of y is given by:

$$f(y_t) = \sum_{j=1}^J \omega_j f_j(y_t, \theta_j).$$

Assuming independence over time, the sample log-likelihood of y_1, \dots, y_T is:

$$\mathcal{L}(\theta) = \sum_{t=1}^T \log \left(\sum_{j=1}^J \omega_j f_j(y_t, \theta_j) \right). \quad (1)$$

where θ collects all the parameters.

The log-likelihood function above may not be globally concave because of the summation across states inside the logarithm. Moreover, when $f_j(y_t, \theta_j)$ is Gaussian, $\mathcal{L}(\theta)$ can be unbounded in finite sample. This occurs when the variance of y_t in one of the regimes eventually take a very small value during the numerical iterations of the likelihood maximization algorithm. However, the CF of y_t is always bounded in modulus. We have:

$$E(\exp(i\tau y_t)) = \sum_{j=1}^J \omega_j \varphi_j(\tau, \theta_j) \equiv \varphi(\tau, \theta), \quad \tau \in \mathbb{R}.$$

where $\varphi_j(\tau, \theta_j) = E(\exp(i\tau y_t) | s_t = j)$ is a regime specific CF and i is the imaginary

number such that $i^2 = -1$.

The following moment function is easily deduced from this CF:

$$h_t(\tau, \theta) = \exp(i\tau y_t) - \varphi(\tau, \theta), \quad \tau \in \mathbb{R}. \quad (2)$$

Indeed, $E(h_t(\tau, \theta)) = 0$ for all $\tau \in \mathbb{R}$. Inference procedures that are based on the CF can take advantage of the fact that $h_t(\tau, \theta)$ is bounded. Indeed:

$$\begin{aligned} |h_t(\tau, \theta)|^2 &= |\exp(i\tau y_t) - \varphi(\tau, \theta)|^2 \\ &\leq |\exp(i\tau y_t)|^2 + |\varphi(\tau, \theta)|^2 \leq 2 \end{aligned}$$

for all τ and θ .

2.2 Stable Distributions

The stable distribution has been introduced in finance in an effort to fit the asymmetry and fat tail observed empirically in the distributions of assets returns (Mandelbrot, 1963; McCulloch, 1986). A random variable Z_t is said to follow the standard α -stable distribution if and only if its CF is given by:

$$E[\exp(i\tau Z_t)] = \exp\{-|\tau|^\alpha [1 + \text{sign}(\tau) i\beta g(\tau, \alpha)]\}, \quad (3)$$

where $\alpha \in (0, 2]$ is a stability parameter that controls the tail behavior, $\beta \in [-1, 1]$ is a skewness parameter, $g(\tau, \alpha) = -\tan \frac{\alpha\pi}{2}$ if $\alpha \neq 1$ and $g(\tau, \alpha) = \frac{2}{\pi} \ln |\tau|$ if $\alpha = 1$.

A random variable y_t follows a general α -stable distribution if and only if it is linked to a standard α -stable random variable Z_t through:

$$y_t = \begin{cases} \sigma Z_t + \mu, & \alpha \neq 1 \\ \sigma Z_t + \mu + \frac{2}{\pi} \beta \sigma \ln \sigma, & \alpha = 1 \end{cases}.$$

The moments of order larger than α do not exist for the stable distribution when $\alpha < 2$. When $\alpha = 2$, all moments exist but the asymmetry parameter β is not identified.

Closed form expressions for stable densities are available only in a few cases. For example, $\alpha = 2$ leads to the normal distribution $N(\alpha, 2\sigma^2)$ whereas $\alpha = 1$ and $\beta = 0$ leads to Cauchy distributions. Levy distributions are obtained by letting $\alpha = 1/2$ and $\beta = 1$. The density for the case $\alpha = 1/2$ and $\beta = -1$ can be deduced from the previous one via an identity (see Zorotalev, 1986; Weron, 1996). However, the

knowledge of the likelihood function at isolated values of the parameter space is not helpful when one is trying to fit the model to real data. This difficulty has often led some empirical researchers to rely on numerical approximations of the likelihood of the stable distribution. For example, McCulloch (1998) discusses an approximate maximum likelihood procedure for symmetric stable distributions while Nolan (1997) proposes a numerical procedure for $\alpha > 0.1$. Mittnik et al. (1999) and Paoletta (2007) use Fast Fourier Transform to approximate the likelihood function while quantile methods are advocated in McCulloch (1986).

As argued in Kotchoni (2012), the parameters of a stable distribution can be estimated using a CF-based continuum of moment condition of the form $E(h_t(\tau, \theta)) = 0$ for all $\tau \in \mathbb{R}$, where:

$$h_t(\tau, \theta) = \exp(i\tau y_t) - \varphi(\tau, \theta), \quad \tau \in \mathbb{R}.$$

where $\varphi(\tau, \theta) = E[\exp(i\tau y_t)]$ and $\theta = (\alpha, \beta, \sigma, \mu)$. The CF of y_t is given by:

$$E[\exp(i\tau y_t)] = \exp\{i\mu\tau - |\sigma\tau|^\alpha [1 + i\beta \text{sign}(\tau) g(\tau, \alpha)]\}, \quad (4)$$

where μ is a location parameter and σ is the scale parameter. Kotchoni (2012) performed a simulation study based on an AR(1) model with stable innovations and found that the CGMM estimator significantly outperforms a GMM estimator based on a discrete subset of moment conditions.

2.3 Discretely Sampled Continuous-Time Processes

Similarly to stable distributions, discretely sampled continuous-time processes are more naturally characterized by their CFs rather than their densities. For example, consider the following stochastic differential equation specified for the short-term nominal interest rate:

$$dx_t = \mu(y_t) dt + \sigma(y_t) dW_t,$$

where y_t denotes an interest rate process, $\mu(y_t)$ is a drift function, $\sigma(y_t)$ is an instantaneous volatility function and W_t is a standard Brownian motion.

The transition density of discretely sampled observations from this model is known only in the special case of a Cox, Ingersoll and Ross (1985) model, that is, when $\mu(y_t) = \kappa(\rho - y_t)$ and $\sigma(y_t) = \sigma\sqrt{y_t}$. In this case, the conditional density of y_t given

$y_{t-\delta}$ ($\delta > 0$) can be represented as an infinite mixture of Gamma densities with Poisson weights:

$$f(y_t|y_{t-\delta}) = \sum_{j=0}^{\infty} p_j \frac{y_t^{j+q-1} e^{j+q}}{\Gamma(j+q)} \exp(-cy_t), \quad (5)$$

where $c = \frac{2\kappa}{\sigma^2(1-e^{-\delta\kappa})}$, $q = \frac{2\kappa\rho}{\sigma^2}$ and $p_j = \frac{(ce^{-\delta\kappa}y_{t-\delta})^j \exp(-ce^{-\delta\kappa}y_{t-\delta})}{j!}$ (see Devroye, 1986; Singleton, 2001). The parameter κ captures the strength of the mean reversion in the interest rate process, ρ is the long run value around which the interest rate process oscillates along the business cycles while σ captures the instantaneous volatility of the interest rate process.

The infinite mixture of densities given above will have to be truncated for the purpose of performing maximum likelihood. However, imprecise estimations of these parameters may have dramatic implications for the valuation of interest rate sensitive assets, the prediction of the term structure of the market risk premium and the assessment of interest rate risk. Contrasting with the infinite mixture (5), the conditional CF of the CIR model has a simple closed form expression given by:

$$\varphi(\tau, \theta; y_{t-\delta}) \equiv E(e^{i\tau y_t} | y_{t-\delta}) = (1 - i\tau/c)^{-q} \exp\left(\frac{i\tau e^{-\delta\kappa} y_{t-\delta}}{1 - i\tau/c}\right), \quad (6)$$

with $\theta = (\kappa, \rho, \sigma)'$. Moreover, the CF is available in closed form for more sophisticated affine-jump diffusions (Jiang and Knight, 2002) and stochastic volatility models (Yu, 2004).

Carrasco and Kotchoni (2017) study the performance of the CGMM by simulation based on the CIR model. The moment function that they consider is:

$$h_t(\tau, \theta) = (\exp(ir_1 y_t) - \varphi(r_1, \theta; y_{t-1})) \exp(ir_2 y_{t-1}), \quad \tau = (r_1, r_2)' \in \mathbb{R}^2. \quad (7)$$

where $\varphi(r_1, \theta; y_{t-1}) = E[\exp(ir_1 y_t) | y_{t-1}]$ and $\theta = (\kappa, \rho, \sigma)$. Here, $\exp(ir_2 y_{t-1})$, $r_2 \in \mathbb{R}$ is being used as a continuum of instruments. Carrasco, Florens, Chernov and Ghysels (2007) showed that this set of instruments is optimal given that the process y_t is Markov.

2.4 Weakly Dependent Processes

The first two examples of CF-based continuum of moment conditions given above deal with IID models while the third example deals with Markov data. In all three examples,

the moment function $h_t(\tau, \theta)$ is uncorrelated over time. Indeed, $h_t(\tau, \theta)$ is IID in the first two cases and a Martingale Difference Sequence in the third case. These cases can easily be extended to account for a type of heteroskedasticity that leaves the observations uncorrelated over time.

When y_t is non-Markov but weakly dependent, its distribution conditional on the past depends on its whole past history. In this case, one has to truncate the history of y_t to J lags by relying on the joint CF of $(y_t, y_{t-1}, \dots, y_{t-J+1})$. This leads to the moment function:

$$h_t(\tau, \theta) = e^{i\tau'Y_t} - E^\theta(e^{i\tau'Y_t}), \quad \tau \in \mathbb{R}^J, \quad (8)$$

where $Y_t = (y_t, y_{t-1}, \dots, y_{t-J+1})'$. The number of lags J must be large enough to guarantee that the parameters of the model are all identified. Moment functions of type (8) are correlated over time.

A CGMM procedure based on the continuum of moment condition (8) is roughly equivalent to the estimator that maximizes the sample conditional likelihood of y_t given J lags. As y_t is non-Markov, the larger J the more efficient the CGMM estimator. In practice, the quest for efficiency must be balanced with the computing cost. For more discussion on this point, see Jiang and Knight (2002), Yu (2004) and Carrasco, Florens, Chernov and Ghysels (2007).

2.5 Conditional Moment Restrictions

Many prominent economic theories lead to testable implications that can be summarized into a set of conditional moment restrictions. Examples of such theories include the consumption capital asset pricing model and dynamic stochastic general equilibrium model. It is customary to convert conditional moment restrictions into a few number of unconditional moment restrictions for the purpose of parameter estimation using selected instruments. However, this approach entails a loss of information given that a conditional moment restriction is equivalent to an infinite number of unconditional moment restrictions. Interestingly, a conditional moment restriction can always be converted into a continuum of moment conditions with same information content (Bierens, 1982).

For example, consider the following conditional moment restriction:

$$E[f(y, \theta) | x] = 0, \quad (9)$$

where y and x are random variables. Dynamic Stochastic General Equilibrium models often lead to first order optimality conditions that are of this form (see Cochrane, 2009). This conditional moment restriction is equivalent to an infinite number of moment conditions since:

$$E [f (y, \theta) g (x)] = 0, \text{ for all possible function } g. \quad (10)$$

For specific classes of functions g , the set of unconditional moment restrictions (10) is equivalent to the conditional moment restriction (9).

For instance, Bierens (1982) and Lavergne and Patilea (2008) used the following type of continuum of moment condition:

$$E [f (y, \theta) \exp (i\tau'x)] = 0, \text{ for all } \tau \in \mathbb{R}^{\dim(x)}, \quad (11)$$

whereas Dominguez and Lobato (2004) propose:

$$E (f (y, \theta) 1(x < \tau)) = 0, \tau \in \mathbb{R}^q \quad (12)$$

The type of continuum of moment condition used in this paper is closer to (11), being complex-valued. If the information content of the conditional moment restriction (9) is equivalent to that of the likelihood of y given x , an estimator based on (11) has the potential to be as efficient as the estimator that maximized this conditional likelihood.

In the particular case of asset pricing models (e.g., Consumption Capital Asset Pricing Model), the first order optimality conditions take the form:

$$E [m_{t+1} (\theta) (r_{i,t+1} - r_{f,t+1}) | I_t] = 0, \text{ for all } i = 1, \dots, N. \quad (13)$$

where $m_{t+1} (\theta)$ is a stochastic discount factor, $r_{i,t+1}$ and $r_{f,t+1}$ are respectively the return on a risky asset and the risk free rate for period $[t, t + 1]$ and I_t is the information set of the investor (see Cochrane, 2009). Following the same intuition as previously, one may consider using the following continuum of moment condition for the estimation of the asset pricing model:

$$E [m_{t+1} (\theta) (r_{i,t+1} - r_{f,t+1}) \exp (i\tau'X_t)] = 0, \text{ for all } \tau \in \mathbb{R}^{\dim(X_t)}. \quad (14)$$

for $i = 1, \dots, N$, where X_t is a vector summarizing the investor's information set at time

t . A suggestion is to restrict (14) to one continuum of moment condition by focusing on the market portfolio only.

3 Review of Estimation Methods Based on the CF

Different approaches have been employed in the literature to estimate a vector of parameters θ_0 from a CF-based continuum of moment conditions (or moment function) $h_t(\tau, \theta), \tau \in \mathbb{R}^d$. The simplest of these approaches consists of applying the GMM procedure to a discrete subset of moment conditions obtained by evaluating $h_t(\tau, \theta)$ at several points $\{\tau_1, \tau_2, \dots, \tau_q\}$ in \mathbb{R}^d (see Feuerverger and McDunnough, 1981b; Singleton, 2001; Chacko and Viceira, 2003). This approach yields a vector of moment conditions of the following type:

$$g_t(\theta) = (\text{Re } h_t(\tau_1, \theta), \dots, \text{Re } h_t(\tau_q, \theta), \text{Im } h_t(\tau_1, \theta), \dots, \text{Im } h_t(\tau_q, \theta))'. \quad (15)$$

Typically, GMM estimators are obtained by minimizing a quadratic form of $\widehat{g}(\theta) = \frac{1}{T} \sum_{t=1}^T g_t(\theta)$ with respect to θ .

The efficient GMM estimator of Hansen (1982) is given by:

$$\widehat{\theta}_{\text{GMM}}^* = \arg \min_{\theta} \widehat{g}(\theta)' S^{-1} \widehat{g}(\theta), \quad (16)$$

where $\widehat{g}(\theta) = \frac{1}{T} \sum_{t=1}^T g_t(\theta)$ and S is the asymptotic covariance matrix of $g_t(\theta_0)$, that is:

$$S = \lim_{T \rightarrow \infty} \text{Var} \left[\sqrt{T} \widehat{g}(\theta_0) \right] = \Gamma_0 + \sum_{j=1}^{\infty} (\Gamma_j + \Gamma_j'), \quad (17)$$

with $\Gamma_j = E [g_t(\theta_0) g_{t-j}(\theta_0)']$. However, $\widehat{\theta}_{\text{GMM}}^*$ is infeasible as S is generally unknown.

If $g_t(\theta)$ is uncorrelated over time, the asymptotic covariance matrix of $g_t(\theta_0)$ reduces to $S = \Gamma_0$ and is therefore estimated by:

$$\widehat{S} = \frac{1}{T} \sum_{t=1}^T g_t(\widehat{\theta}^1) g_t(\widehat{\theta}^1)', \quad (18)$$

where $\widehat{\theta}^1$ is a consistent first step estimator of θ_0 . Such an estimator is given by:

$$\widehat{\theta}^1 = \arg \min_{\theta} \|\widehat{g}(\theta)\|^2. \quad (19)$$

When $g_t(\theta)$ is autocorrelated, the suitable estimator of S takes the form:

$$\widehat{S} = \widehat{\Gamma}_0 + \sum_{j=1}^J w\left(\frac{j}{J}\right) (\widehat{\Gamma}_j + \widehat{\Gamma}'_j), \quad (20)$$

where $\widehat{\Gamma}_j = \frac{1}{T} \sum_{t=1}^T g_t(\widehat{\theta}^1) g_{t-j}(\widehat{\theta}^1)'$, $w(x)$ is a positive and decreasing kernel function on $[0, 1]$ satisfying $w(0) = 1$ and $w(1) = 0$ and J is a bandwidth that must be selected by the econometrician. Intuitively, the optimal bandwidth is increasing in the degree of persistence of $g_t(\theta)$ as well as in the sample size T . See Newey and West (1987) and Andrews and Monahan (1992) for more details on the properties of \widehat{S} .

The feasible version of $\widehat{\theta}_{\text{GMM}}^*$, commonly referred to as the "two-step efficient GMM estimator," is given by:

$$\widehat{\theta}_{\text{GMM}} = \arg \min_{\theta} \widehat{g}(\theta)' \widehat{S}^{-1} \widehat{g}(\theta). \quad (21)$$

Hansen (1982) showed that $\widehat{\theta}_{\text{GMM}}$ and $\widehat{\theta}_{\text{GMM}}^*$ are asymptotically equivalent, meaning that $\widehat{\theta}_{\text{GMM}}$ is asymptotically efficient within the family of GMM estimators that are based on $g_t(\theta)$. However, $\widehat{\theta}_{\text{GMM}}$ remains inefficient in the sense that it only utilizes a portion of the information content of the continuum of moment conditions at hand.

Feuerverger and McDunnough (1981b) claim that the asymptotic variance of $\widehat{\theta}_{\text{GMM}}$ can be made arbitrarily close to the Cramer-Rao bound by selecting the grid (τ_1, \dots, τ_q) sufficiently refined and extended in \mathbb{R}^d as T goes to infinity. However, the asymptotic covariance matrix of the moment conditions, S , becomes singular as one refines and extends the grid. Indeed, the discrete set of moment conditions converges to the continuous moment function $h_t(\tau, \theta)$, $\tau \in \mathbb{R}^d$ while the matrix S converges to the covariance operator associated with that moment function. This suggests that the GMM based on discretization is better handled in a Hilbertian functional space as $q \rightarrow \infty$.

GMM procedures that continuously match the empirical CF to its theoretical counterpart are studied in Press (1972), Paulson et al. (1975), Thornton and Paulson (1977) and more recently in Yu (2004). The corresponding estimators are generally of the form:

$$\arg \min_{\theta} \int |h_t(\tau, \theta)|^2 \pi(\tau) d\tau, \quad (22)$$

where $\pi(\tau)$ is continuous weighting function. As shown by Yu (2004, p. 98), these procedures are less efficient than the maximum likelihood.

When the moment function is deduced from a conditional CF, it is necessary to *instrument* it in order to guaranty the identification of θ_0 . The moment function therefore take the form:

$$h_t(\tau, \theta) = \left(e^{i\tau'y_{t+1}} - E \left[e^{i\tau'y_{t+1}} | y_t \right] \right) w(\tau, y_t), \tau \in \mathbb{R}^d, \quad (23)$$

where $w(\tau, y_t), \tau \in \mathbb{R}^d$ is a continuum of instruments. Singleton (2001) shows that an estimator (22) based on such a moment function achieves the maximum likelihood efficiency when the continuum of instruments is of the following form:

$$w(\tau, y_t) = \frac{1}{2\pi} \int e^{i\tau'y_{t+1}} \frac{\partial \mathcal{L}(\theta, y_{t+1} | y_t)}{\partial \theta} dy_{t+1}, \tau \in \mathbb{R}^d, \quad (24)$$

where $\mathcal{L}(\theta, y_{t+1} | y_t)$ is the log-likelihood of y_{t+1} conditional on y_t . In most cases of practical interest, these optimal instruments cannot be computed as the conditional likelihood is unknown.

Carrasco and Florens (2000) proposed a two-step GMM procedure that avoids the discretization of the CF-based continuum of moment conditions. Interestingly, their procedure achieves the maximum likelihood efficiency and it can be implemented without knowing the conditional likelihood function of the data. This procedure is presented in details in the next section.

4 Derivation and Properties of the CGMM Estimator

This section presents the CGMM estimator and its asymptotic properties. The first subsection presents the objective function that the CGMM estimator minimizes. The second subsection presents the consistency and asymptotic normality of the CGMM estimator. In the third subsection, we discuss the higher order bias and variance of the CGMM estimator.

4.1 The CGMM Estimator

For the sake of generality, we assume that y_t is a d -dimensional random vector. We consider the following CFs:

$$\begin{aligned}\varphi(\tau, \theta) &= E(\exp(i\tau'y_t)), \tau \in \mathbb{R}^d \text{ if } y_t \text{ is IID,} \\ \varphi(r, \theta; y_{t-1}) &= E(\exp(ir'y_t) | y_{t-1}), r \in \mathbb{R}^d \text{ if } y_t \text{ is Markov and} \\ \varphi(\tau, \theta) &= E^\theta(e^{i\tau'Y_t}), \tau \in \mathbb{R}^{dJ} \text{ if } y_t \text{ is non-Markov.}\end{aligned}$$

where $Y_t = (y'_t, y'_{t-1}, \dots, y'_{t-J+1})'$. In the IID and non-Markov cases, the moment function is written as:

$$h_t(\tau, \theta) = e^{i\tau'Y_t} - \varphi(\tau, \theta) \quad (25)$$

whereas in the Markov case we have:

$$h_t(\tau, \theta) = (\exp(ir_1 y_t) - \varphi(r_1, \theta; y_{t-1})) \exp(ir_2 y_{t-1}), \tau = (r_1, r_2). \quad (26)$$

If the starting point is a conditional moment restriction of the form $E[f(y_t, \theta) | x_t] = 0$, then we let:

$$h_t(\tau, \theta) = f(y_t, \theta) \exp(i\tau'x_t), \text{ for all } \tau \in \mathbb{R}^q \quad (27)$$

where $q = \dim(x_t)$. Note that $h_t(\tau, \theta)$ is correlated over time in the non-Markov case while it is uncorrelated in the other cases.

The objective function of the CGMM is a quadratic form that is defined on the Hilbert space of square integrable functions with respect to a well-chosen measure π on \mathbb{R}^d , denoted $L^2(\pi)$. We have:

$$L^2(\pi) = \{f : \mathbb{R}^d \rightarrow \mathbb{C} \mid \int f(\tau) \overline{f(\tau)} \pi(\tau) d\tau < \infty\}, \quad (28)$$

where $\overline{f(\tau)}$ denotes the complex conjugate of $f(\tau)$. For simplicity, π may be taken as a probability density function on \mathbb{R}^d . As $|h_t(\cdot, \theta)|^2 \leq 2$ for all $\theta \in \Theta$, the function $h_t(\cdot, \theta)$ belongs to $L^2(\pi)$ for all $\theta \in \Theta$ and for any finite measure π .

We consider the following scalar product on $L^2(\pi) \times L^2(\pi)$:

$$\langle f, g \rangle = \int f(\tau) \overline{g(\tau)} \pi(\tau) d\tau.$$

Based on this notation, the efficient CGMM estimator of Carrasco and Florens (2000)

is given by:

$$\hat{\theta} = \arg \min_{\theta} \left\langle K^{-1} \hat{h}_T(\cdot, \theta), \hat{h}_T(\cdot, \theta) \right\rangle. \quad (29)$$

where $\hat{h}_T(\tau, \theta) = \frac{1}{T} \sum_{t=1}^T h_t(\tau, \theta)$ and K is the asymptotic covariance operator associated with the moment function. Namely, K is an integral operator that satisfies:

$$Kf(\tau_2) = \int_{-\infty}^{\infty} k(\tau_1, \tau_2) f(\tau_1) \pi(\tau_1) d\tau_1, \text{ for any } f \in L^2(\pi),$$

where $k(\tau_1, \tau_2)$ is the kernel of K . The expression of $k(\tau_1, \tau_2)$ is given by:

$$\begin{aligned} k(\tau_1, \tau_2) &= E \left[h_t(\tau_1, \theta) \overline{h_t(\tau_2, \theta)} \right] \\ &+ \sum_{j=1}^{\infty} E \left[h_t(\tau_1, \theta) \left(\overline{h_{t-j}(\tau_2, \theta)} + \overline{h_{t+j}(\tau_2, \theta)} \right) \right]. \end{aligned} \quad (30)$$

When $h_t(\tau_1, \theta)$ is uncorrelated, $k(\tau_1, \tau_2)$ reduces to:

$$k(\tau_1, \tau_2) = E \left(h_t(\tau_1, \theta) \overline{h_t(\tau_2, \theta)} \right). \quad (31)$$

Like Hansen (1982)'s GMM estimator, the CGMM estimator must be implemented in two steps. In the first step, one estimates K by a sample analogue K_T based on a consistent first step estimator of θ . Such an estimator is given by:

$$\hat{\theta}_T^{(1)} = \arg \min_{\theta} \left\langle \hat{h}_T(\cdot, \theta), \hat{h}_T(\cdot, \theta) \right\rangle.$$

The feasible second step CGMM estimator is then obtained as:

$$\hat{\theta}_T(\lambda) = \arg \min_{\theta} \left\langle K_{\lambda T}^{-1} \hat{h}_T(\cdot, \theta), \hat{h}_T(\cdot, \theta) \right\rangle, \quad (32)$$

where $K_{\lambda T}^{-1}$ is a regularized inverse of K_T . In this paper, we consider:

$$K_{\lambda T}^{-1} = (K_T^2 + \lambda I)^{-1} K_T.$$

When the data generating process is IID or Markov, it is straightforward to estimate $k(\tau_1, \tau_2)$ by:

$$k_T(\tau_1, \tau_2, \hat{\theta}_T^1) = \frac{1}{T} \sum_{t=1}^T h_t(\tau_1, \hat{\theta}_T^1) \overline{h_t(\tau_2, \hat{\theta}_T^1)}. \quad (33)$$

In the specific case of IID data, the first step estimator of θ can be bypassed by considering:

$$k_T(\tau_1, \tau_2) = \frac{1}{T} \sum_{t=1}^T \left(e^{i\tau_1' x_t} - \widehat{\varphi}_T(\tau_1) \right) \overline{\left(e^{i\tau_2' x_t} - \widehat{\varphi}_T(\tau_2) \right)}. \quad (34)$$

where $\widehat{\varphi}_T(\tau_1) = \frac{1}{T} \sum_{t=1}^T e^{i\tau_1' x_t}$.

When the data are non-Markov, $k(\tau_1, \tau_2)$ may be estimated as in Newey and West (1987) or Andrews and Monahan (1992) using:

$$\begin{aligned} \widehat{k}_T(\tau_1, \tau_2, \widehat{\theta}_T^1) &= \frac{1}{T} \sum_{t=1}^T h_t(\tau_1, \widehat{\theta}_T^1) \overline{h_t(\tau_2, \widehat{\theta}_T^1)} \\ &+ \sum_{j=1}^J \omega \left(\frac{j}{J} \right) \sum_{t=1}^T h_t(\tau_1, \widehat{\theta}_T^1) \left(\overline{h_{t-j}(\tau_2, \widehat{\theta}_T^1)} + \overline{h_{t+j}(\tau_2, \widehat{\theta}_T^1)} \right), \end{aligned} \quad (35)$$

where J is a bandwidth and $\omega : [0, 1] \rightarrow [0, 1]$ is a positive and decreasing kernel function from such that $\omega(0) = 1$ and $\omega(1) = 0$. An example of kernel function (the Bartlett kernel) is given by $\omega(x) = 1 - x$.

A regularized inverse is needed in (32) because the empirical operator K_T is degenerate and non invertible. The problem is worsened by the fact that the inverse of K (i.e., the theoretical counterpart of K_T) exists only on a dense subset of $L^2(\pi)$. Moreover, the integral equation $Kf = g$ is ill-posed in the sense that when $f \equiv K^{-1}g$ exists, a small perturbation in g may lead to an arbitrarily large variation in f (see Carrasco, Florens and Renault, 2007). Regularization permits to elude this problem while allowing $K_{\lambda T}^{-1}f$ to exist for all f in $L^2(\pi)$.

The remainder of the current section is rather theoretical as it is devoted to the asymptotic properties of the CGMM estimator. Practical issues regarding the implementation of the CGMM estimator are discussed in subsequent sections.³

4.2 Consistency and Asymptotic Normality

Carrasco and Kotchoni (2017) study the properties of CGMM estimators when the continuum of moment conditions are based on the CF of an IID or a Markov model and under the following assumptions:

Assumption 1: The measure π is strictly positive on \mathbb{R}^d and admits all its moments.

³Readers who are more interested in the practical aspects than in the theory may jump to the next section.

Assumption 2: The equation

$$E^{\theta_0} (h_t(\tau, \theta)) = 0 \text{ for all } \tau \in \mathbb{R}^d, \pi - \text{almost everywhere,}$$

has a unique solution θ_0 which is an interior point of a compact set Θ , where E^{θ_0} is the expectation operator under the true model and $h_t(\tau, \theta)$ is given by (25) in the IID case or (26) in the Markov case.

Assumption 3: (i) $h_t(\tau, \theta)$ is three times continuously differentiable with respect to θ . Furthermore, (ii) the first two derivatives of $h_t(\tau, \theta)$ with respect to θ satisfy:

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \frac{\partial h_t(\tau, \theta)}{\partial \theta_j} - E \left(\frac{\partial h_t(\tau, \theta)}{\partial \theta_j} \right) &= O_p(T^{-1/2}) \text{ and} \\ \frac{1}{T} \sum_{t=1}^T \frac{\partial^2 h_t(\tau, \theta)}{\partial \theta_j \partial \theta_k} - E \left(\frac{\partial^2 h_t(\tau, \theta)}{\partial \theta_j \partial \theta_k} \right) &= O_p(T^{-1/2}), \end{aligned}$$

for all j and k .

Assumption 4: (i) $E^{\theta_0} (h_T(\tau, \theta)) \in \Phi_\beta$ for all $\theta \in \Theta$ and for some $\beta \geq 1$, where

$$\Phi_\beta = \{f \in L^2(\pi) \text{ such that } \|K^{-\beta} f\| < \infty\}. \quad (36)$$

Furthermore, (ii) the first two derivatives of $E^{\theta_0} (h_T(\tau, \theta))$ with respect to θ belong to Φ_β for all θ in a neighborhood of θ_0 and for the same β as above.

Assumption 5: (i) The random variable y_t is stationary Markov and satisfies $y_t = D(y_{t-1}, \theta_0, \varepsilon_t)$ where $D(y_{t-1}, \theta^0, \varepsilon_t)$ is three times continuously differentiable with respect to θ^0 and ε_t is a IID white noise whose distribution is known and does not depend on θ^0 . Furthermore, (ii) the gradient $G(\tau, \theta; \theta^0) = E \left(\frac{\partial h_t(\tau, \theta; \theta^0)}{\partial \theta} \right)$ and covariance operator K are continuously differentiable with respect to θ^0 , where $h_t(\tau, \theta; \theta^0)$ is a moment function evaluated at parameter value θ but using data that are generated by θ^0 .

Assumption 1 implies that the scalar product $\langle \cdot, \cdot \rangle$ underlying the objective function of the CGMM is well-defined and bounded. Assumption 2 imposes the global identification of θ . Assumption 3(i) ensures some smoothness properties for $\widehat{\theta}_T(\lambda)$ while Assumption 3(ii) requests that the Gradient and Hessian of the moment function converge in probability to their theoretical counterparts.

Assumption 4 guarantees that $E^{\theta_0} (h_T(\tau, \theta))$ lies the range of K : there exists g in

$L^2(\pi)$ such that $E^{\theta^0}(h_T(\tau, \theta)) = Kg$. The largest real number β such that $f \in \Phi_\beta$ in Assumption 4 is the level of regularity of f with respect to K . The larger β is, the better f is approximated by a linear combination of the eigenfunctions of K associated with the largest eigenvalues. Because $Kf(\cdot)$ involve a d -dimensional integration, β may be affected by both the dimensionality of the index τ and the smoothness of f .

Assumption 5-(i) implies that the data can be simulated upon knowing how to draw from the distribution of ε_t . It is satisfied for all random variables that can be written as a location parameter plus a scale parameter times a standardized representative of the family of distribution. Examples include the exponential family and the stable distribution. An IID model is a special case of Assumption 5 where $D(y_{t-1}, \theta_0, \varepsilon_t)$ takes the simpler form $D(\theta_0, \varepsilon_t)$. Note that the function $D(y_{t-1}, \theta_0, \varepsilon_t)$ may not be explicitly available in analytical form. For instance, the relation $y_t = D(y_{t-1}, \theta_0, \varepsilon_t)$ can be the numerical solution of a general equilibrium asset pricing model (e.g., as in Duffie and Singleton, 1993).

By Assumptions 3 and 5-(i), $\frac{\partial \widehat{h}_T(\tau, \theta; \theta^0)}{\partial \theta}$ is twice continuously differentiable with respect to θ^0 while the kernel $k_T(\tau_1, \tau_2, \widehat{\theta}_1)$ is three times continuously differentiable with respect to θ^0 . Therefore, the differentiability requirement of Assumption 5-(ii) holds for the empirical gradient $G_T(\tau, \theta; \theta^0) = \frac{\partial \widehat{h}_T(\tau, \theta; \theta^0)}{\partial \theta}$ and the empirical covariance operator K_T . Assumption 5-(ii) extends this differentiability to the probability limits $G(\tau, \theta; \theta^0)$ and K .

Carrasco and Kotchoni (2017) establish the following result under the previous assumptions.

Theorem 1 *Under Assumptions 1 to 5, the CGMM estimator is consistent and satisfies:*

$$T^{1/2} \left(\widehat{\theta}_T(\lambda) - \theta_0 \right) \xrightarrow{L} N(0, I_{\theta_0}^{-1}).$$

as T and $\lambda T^{1/2}$ go to infinity and λ goes to zero, where $I_{\theta_0}^{-1}$ denotes the inverse of the Fisher Information Matrix.

Theorem 1 stipulates that the CGMM estimator is consistent, asymptotically normal and efficient. Moreover, the asymptotic distribution does not depend on the regularization parameter λ nor on the measure π used to build the objective function. A more general statement of this result is provided in Proposition 3.2 of Carrasco, Chernov, Florens, and Ghysels (2007). Due to Assumption 5-(i), non-Markov processes do not share the result of Theorem 1. This assumption can be modified to include non-Markov

processes by letting $y_t = D(\chi_{t-1}, \theta_0, \varepsilon_t)$, where χ_{t-1} is a finite dimensional state variable summarizing the past history of y_t . Under the modified assumption, consistency and asymptotic normality stills hold. However, asymptotic efficiency is no longer guaranteed because the moment function used in the non-Markov case is deduced from a truncated information set and not on the density of y_t conditional on its entire history. See Proposition 4.3 of Carrasco, Chernov, Florens, and Ghysels (2007).

4.3 Stochastic Expansion and Approximate MSE

According to Theorem 1, any sequence of regularization parameters of type $\lambda_T = cT^{-a}$ with $c > 0$ and $0 < a < 1/2$ leads to an efficient CGMM estimator. Among the admissible convergence rates, we would like to find the one that minimizes the trace of the Mean Square Error matrix of the CGMM estimator for a given sample size T :

$$MSE(\lambda, \theta_0) = E \left[T \left(\widehat{\theta}_T(\lambda) - \theta_0 \right) \left(\widehat{\theta}_T(\lambda) - \theta_0 \right)' \right], \quad (37)$$

Unfortunately, there is no theoretical basis for claiming that $MSE(\lambda, \theta_0)$ is finite for any data generating process and any sample size. Indeed, the large sample properties of GMM-type estimators like $\widehat{\theta}_T(\lambda)$ are well-known but their finite sample properties can be established only in special cases. In particular, the MSE of $\widehat{\theta}_T(\lambda)$ can be infinite in finite samples even though $\widehat{\theta}_T(\lambda)$ is consistent for θ_0 .

To hedge against situations where $MSE(\lambda, \theta_0)$ is infinite in finite sample, we consider approximating the MSE of $\widehat{\theta}_T(\lambda)$ by that of the leading terms of its stochastic expansion. Our goal is to examine the nature of the dependence of the higher order variance and bias of the CGMM estimator on the regularization parameter λ . The higher order properties of GMM-type estimators have been studied by Rothenberg (1983, 1984), Koenker et al. (1994), Rilstone et al. (1996), and Newey and Smith (2004). For estimators derived in the linear simultaneous equation framework, examples include Nagar (1959), Buse (1992) and Donald and Newey (2001). The approach followed by Carrasco and Kotchoni (2017) is similar to Nagar (1959) and in particular, Donald and Newey (2001), who select the number of instruments to include in a linear instrumental variable model by minimization of an AMSE criterion.

Carrasco and Kotchoni (2017) derived an expansion of $\widehat{\theta}_T(\lambda) - \theta_0$ that is of the

following form for both the IID and Markov cases:

$$\widehat{\theta}_T(\lambda) - \theta_0 = \Delta_1 + \Delta_2 + \Delta_3 + o_p(\lambda^{-1}T^{-1}) + o_p\left(\lambda^{\min(1, \frac{2\beta-1}{2})}T^{-1/2}\right), \quad (38)$$

where

$$\begin{aligned} \Delta_1 &= O_p(T^{-1/2}), \\ \Delta_2 &= O_p\left(\lambda^{\min(1, \frac{2\beta-1}{2})}T^{-1/2}\right) \text{ and} \\ \Delta_3 &= O_p(\lambda^{-1}T^{-1}) \end{aligned}$$

See Appendix B of Carrasco and Kotchoni (2017) for details.⁴ Carrasco and Kotchoni (2017) used this expansion to calculate an Approximate MSE (AMSE) for $\widehat{\theta}_T(\lambda)$ and establish results on the optimal regularization parameter.

The following result is established in Carrasco and Kotchoni (2017) for the IID and Markov cases when the moment function is deduced from a CF.

Theorem 2 *Assume that Assumptions 1 to 5 hold. Then we have:*

(i) *The AMSE matrix of $\widehat{\theta}_T(\lambda)$ up to order $O(\lambda^{-1}T^{-1/2})$ is decomposed as the sum of the squared bias and variance:*

$$AMSE(\lambda, \theta_0) = TBias * Bias' + TVar$$

where

$$\begin{aligned} TBias * Bias' &= O(\lambda^{-2}T^{-1}), \\ TVar &= I_{\theta_0}^{-1} + O\left(\lambda^{\min(2, \frac{2\beta-1}{2})}\right) + O(\lambda^{-1}T^{-1/2}). \end{aligned}$$

as $T \rightarrow \infty$, $\lambda^2 T \rightarrow \infty$ and $\lambda \rightarrow 0$.

(ii) *The λ that minimizes the trace of $AMSE(\lambda, \theta_0)$, denoted $\lambda_T \equiv \lambda_T(\theta_0)$, satisfies:*

$$\lambda_T = O\left(T^{-\max(\frac{1}{6}, \frac{1}{2\beta+1})}\right).$$

First, note that the expansion is consistent with the condition of Theorem 1 since the optimal regularization parameter satisfies: $\lambda_T \rightarrow 0$ and $\lambda_T^2 T \rightarrow \infty$. Second, we

⁴The expressions of Δ_1 , Δ_2 and Δ_3 are not reported here as we want to focus on their convergence rates.

have a trade-off between a term that is decreasing in λ and another that is increasing in λ . The squared bias term is dominated by two higher order variance terms whose rates of convergence are equated to obtain the optimal rate for the regularization parameter. The same situation happens for the Limited Information Maximum Likelihood estimator for which the bias is also dominated by variance terms (see Donald and Newey, 2001). Third, note that the rate of convergence for the $O\left(\lambda^{\min(2, \frac{2\beta-1}{2})}\right)$ variance term does not improve for $\beta > 2.5$. This is due to a property of Tikhonov regularization that is well documented in the literature on inverse problems, see e.g. Carrasco, Florens and Renault (2007).

It follows from Theorem 2 that the optimal regularization parameter λ_T is necessarily of the form:

$$\lambda_T = cT^{-g(\beta)}, \quad (39)$$

for some positive function c that does not depend on T and a positive function $g(\beta)$ that satisfies $\max\left(\frac{1}{6}, \frac{1}{2\beta+1}\right) \leq g(\beta) < 1/2$. As the derivations of Theorem 2 all rely on big $O(\cdot)$ approximation, the optimal λ_T can go to zero slightly faster than $T^{-\max(\frac{1}{6}, \frac{1}{2\beta+1})}$ but slower than $T^{-1/2}$.

From Appendix B of Carrasco and Kotchoni (2007), we have:

$$\widehat{\theta}_T(\lambda) - \theta_0 \simeq \Delta_T(\lambda, \theta_0),$$

with

$$\begin{aligned} \Delta_T(\lambda, \theta_0) = & -W_0^{-1}(\theta_0) \left\langle K_{\lambda T}^{-1}G(\cdot, \theta_0), \widehat{h}_T(\cdot, \theta_0) \right\rangle \\ & + W_0^{-1}(\theta_0) \left[\left\langle K_{\lambda T}^{-1}G(\cdot, \theta_0), G(\cdot, \theta_0) \right\rangle - W_0(\theta_0) \right] W_0^{-1}\Psi_{T,0}(\theta_0), \end{aligned} \quad (40)$$

where

$$\begin{aligned} G(\tau, \theta_0) &= P \lim \frac{1}{T} \sum_{t=1}^T \frac{\partial h_t(\tau, \theta_0)}{\partial \theta}, \\ \Psi_{T,0}(\theta_0) &= \text{Re} \left\langle K^{-1}G(\cdot, \theta_0), \widehat{h}_T(\cdot, \theta_0) \right\rangle \text{ and} \\ W_0(\theta_0) &= \left\langle K^{-1}G(\cdot, \theta_0), G(\cdot, \theta_0) \right\rangle. \end{aligned}$$

The AMSE of $\widehat{\theta}_T(\lambda)$ coincides with the exact MSE of Δ_T . The trace of the MSE

matrix of Δ_T is given by:

$$\Sigma_T(\lambda, \theta_0) = E [T \|\Delta_T(\lambda, \theta_0)\|^2]. \quad (41)$$

This quantity is always finite and its limit as $T \rightarrow \infty$ coincides with the trace of the approximate MSE matrix of $\widehat{\theta}_T(\lambda)$. Hence, the optimal λ can be selected by minimizing an estimate of the MSE of $\Delta_T(\lambda, \theta_0)$ as explained in the next section.

5 Selection of the regularization parameter

The CGMM estimator is consistent for any reasonable choice of the regularization parameter λ_T . In most applications, an arbitrary choice of λ_T between 10^{-6} and 10^{-2} would work quite well. However, an arbitrary choice is not advised if the spectrum of the empirical covariance operator is severely discontinuous. In order to approximate the efficient CGMM as accurately as possible, two approaches may be used. The first method is a naive parametric bootstrap that is quite easy to implement while the second approach combines a parametric bootstrap and the stochastic expansion presented earlier. We briefly review these two methods below.

5.1 Naive Bootstrap

The naive bootstrap approach assumes without formal proof that the MSE of $\widehat{\theta}_T(\lambda)$ is finite. If θ_0 were known, the optimal regularization parameter would be:

$$\lambda_T(\theta_0) = \arg \min_{\lambda \in [0,1]} E \left[T \left\| \widehat{\theta}_T(\lambda) - \theta_0 \right\|^2 \right].$$

where $\widehat{\theta}_T(\lambda)$ is computed using a sample of size T generated by θ_0 . The multiplication by T inside the expectation operator ensures that this objective function is $O(1)$ as $T \rightarrow \infty$. In the current context, it is useful to make the dependence of $\widehat{\theta}_T(\lambda)$ on θ_0 explicit by writing:

$$\lambda_T(\theta_0) = \arg \min_{\lambda \in [0,1]} E \left[T \left\| \widehat{\theta}_T(\lambda; \theta_0) - \theta_0 \right\|^2 \right], \quad (42)$$

This objective function is nothing but the trace of the MSE matrix of $\widehat{\theta}_T(\lambda; \theta_0)$.

To approximate this MSE, assume that we can draw samples of size T from the true data generating process of $\{y_t\}$, and let $\hat{\theta}_T^j(\lambda; \theta_0)$ denote the CGMM estimator of θ_0 computed using the j^{th} independently simulated sample. A good estimator of the MSE is given by:

$$\frac{T}{M} \sum_{j=1}^M \left\| \hat{\theta}_T^j(\lambda; \theta_0) - \theta_0 \right\|^2.$$

The Law of Large Numbers ensures that this criterion converges to its expected value as $M \rightarrow \infty$. Note that this simulation can be replaced by an appropriate resampling method, in which case $\hat{\theta}_T^j(\lambda; \theta_0)$ becomes the CGMM estimator inferred from the j^{th} bootstrap sample. In the current context, bootstrap samples are obtained by resampling the set of moment conditions $\{h_t(\tau, \theta), \tau \in \mathbb{R}^d\}_{t=1}^T$ in the time domain.

If θ_0 were known, (42) would suggest an estimator of $\lambda_T(\theta_0)$ of the form:

$$\hat{\lambda}_{TM}(\theta_0) = \arg \min_{\lambda \in [0,1]} \frac{T}{M} \sum_{j=1}^M \left\| \hat{\theta}_T^j(\lambda; \theta_0) - \theta_0 \right\|^2.$$

As θ_0 is unknown, a feasible approach consists of replacing θ_0 with a consistent first step estimator $\hat{\theta}^1$. This leads to estimate the optimal regularization parameter as:

$$\hat{\lambda}_{TM}(\hat{\theta}^1) = \arg \min_{\lambda \in [0,1]} \frac{T}{M} \sum_{j=1}^M \left\| \hat{\theta}_T^j(\lambda; \hat{\theta}^1) - \hat{\theta}^1 \right\|^2. \quad (43)$$

Hence, the optimal feasible CGMM estimator is $\hat{\theta}_T(\hat{\lambda}_{TM}(\hat{\theta}^1); \theta_0)$, *i.e.* the second step estimator of θ_0 computed with the actual data by using the point estimate of the optimal regularization parameter $\hat{\lambda}_{TM}(\hat{\theta}^1)$. The MSE must be simulated using common random numbers in order to eliminate the impact of the integration error when comparing the MSE across the different values of λ (see Kotchoni, 2012).

The naive parametric bootstrap is quite easy to implement. Unfortunately, the theoretical properties of $\hat{\lambda}_{TM}(\hat{\theta}^1)$ cannot be established as we do not know whether the theoretical counterpart of the feasible objective function is finite. If one is able to prove that the MSE of $\hat{\theta}_T(\lambda)$ is finite, then the theoretical properties of $\hat{\lambda}_{TM}(\hat{\theta}^1)$ given at (43) would be the same as when the alternative approach described below is used.

5.2 Stochastic Expansion plus Bootstrap

The current approach is aimed at avoiding the drawback of the naive bootstrap by basing the selection of the regularization parameter on the AMSE deduced from the stochastic expansion. Here, the optimal regularization parameter solves:

$$\lambda_T(\theta_0) = \arg \min_{\lambda \in [0,1]} \Sigma_T(\lambda, \theta_0), \quad (44)$$

where $\Sigma_T(\lambda, \theta_0)$ is the exact MSE of $\Delta_T(\lambda, \theta_0)$, the leading terms of the expansion of $\widehat{\theta}_T(\lambda) - \theta_0$ (see Equation 41).

The expressions of $\Delta_T(\lambda, \theta_0)$ depends on both deterministic and random quantities. The deterministic quantities are the true parameter θ_0 , the covariance operator K , the probability limit of the gradient of the moment function $G(\tau, \theta_0)$ and the regularization parameter λ . The random quantities are the moment function $\widehat{h}_T(\tau, \theta_0)$ and the empirical covariance operator K_T . We therefore have:

$$\Delta_T(\lambda, \theta_0) = \Delta \left(\lambda, K, G(\cdot, \theta_0), K_T(\theta_0), \widehat{h}_T(\cdot, \theta_0) \right),$$

In the IID case, $G(\tau, \theta_0)$ and $K(\theta_0)$ are known in closed form since $G(\tau, \theta_0) = \frac{\partial \varphi(\tau, \theta_0)}{\partial \theta}$ and

$$k(\tau_1, \tau_2) = \varphi(\tau_1 - \tau_2, \theta_0) - \varphi(\tau_1, \theta_0) \overline{\varphi(\tau_2, \theta_0)}.$$

In the Markov and weakly dependent cases, these quantities can be consistently estimated.

Let $\widehat{\theta}_T^1$ be a consistent but inefficient estimator of θ . The steps to estimate the AMSE of $\widehat{\theta}_T(\lambda)$ and the optimal regularization parameter $\lambda_T(\theta_0)$ are as follows:

Step 1. Obtain an estimate $\widetilde{K}(\widehat{\theta}_T^1)$ of $K(\theta_0)$ and an estimate $\widetilde{G}(\cdot, \widehat{\theta}_T^1)$ of $G(\cdot, \theta_0)$ from a very large sample that is simulated using $\widehat{\theta}_T^1$. The simulated sample must be large enough to ensure that the approximation errors of \widetilde{K} and \widetilde{G} are negligible. Note that this step is not necessary in the IID case as closed form expressions are available for $K(\theta)$ and $\widetilde{G}(\cdot, \theta)$.

Step 2. For $j = 1, 2, \dots, M$:

Step 2.1. Draw independent samples $X_T^{(j)}(\widehat{\theta}_T^1)$ of size T from the data generating process using $\widehat{\theta}_T^1$.

Step 2.2. Use the sample $X_T^{(j)}(\widehat{\theta}_T^1)$ to compute the moment function $\widehat{h}_T^{(j)}(\tau, \widehat{\theta}_T^1)$

and the empirical covariance operator $K_T^{(j)}(\hat{\theta}_T^1)$.

Step 2.3. Compute $\Delta_T^{(j)}(\lambda, \hat{\theta}_T^1) = \Delta(\lambda, \tilde{K}(\hat{\theta}_T^1), \tilde{G}(\cdot, \hat{\theta}_T^1), K_T^{(j)}(\hat{\theta}_T^1), \hat{h}_T^{(j)}(\cdot, \hat{\theta}_T^1))$.

Step 3. Estimate the AMSE of $\hat{\theta}_T(\lambda)$ as:

$$\hat{\Sigma}_{TM}(\lambda, \hat{\theta}_T^1) = \frac{T}{M} \sum_{j=1}^M \left\| \Delta_T^{(j)}(\lambda, \hat{\theta}_T^1) \right\|^2,$$

where T and M denote the sample size and the number of Monte Carlo replications.

Step 4: Estimate the optimal regularization parameter as:

$$\hat{\lambda}_{TM}(\hat{\theta}_T^1) = \arg \min_{\lambda \in [0,1]} \hat{\Sigma}_{TM}(\lambda, \hat{\theta}_T^1). \quad (45)$$

In practice, one will have to loop over λ on a finite grid.

The current approach to estimate $\lambda_T(\theta_0)$ is rather fast as it does not require a numerical optimization at each Monte Carlo replication. However, this procedure rests on the presumptions that the expression of the moment function allows for a doable stochastic expansion and that \tilde{K} is a highly accurate approximation of K . The latter presumption is reasonable given that the simulated sample used to estimate K at Step 1 can be made arbitrarily large.

Carrasco and Kotchoni (2017) establish the consistency of $\hat{\lambda}_{TM}(\hat{\theta}^1)$ for $\lambda_T(\theta_0)$ by neglecting the estimation errors of \tilde{K} and \tilde{G} and by making the following additional assumption:

Assumption 6: The regularization parameter λ that minimizes $\Sigma_T(\lambda, \theta_0)$ is of the form $\lambda_T(\theta_0) = c(\theta_0) T^{-g(\beta)}$, for some continuous positive function $c(\theta_0)$ that does not depend on T and a positive function $g(\beta)$ that satisfies $\max\left(\frac{1}{6}, \frac{1}{2\beta+1}\right) \leq g(\beta) < 1/2$.

This assumption is reasonable given the findings of Theorem 2. The consistency result for $\hat{\lambda}_{TM}(\hat{\theta}^1)$ is stated in the next theorem.

Theorem 3 *Let $\hat{\theta}^1$ be a \sqrt{T} -consistent estimator of θ_0 . Then under assumptions 1 to 6,*

$$\frac{\hat{\Sigma}_{TM}(\hat{\lambda}_{TM}(\hat{\theta}^1), \hat{\theta}^1)}{\Sigma_T(\lambda_T(\theta_0), \theta_0)} - 1 = o_p(1)$$

as M goes to infinity first and T goes to infinity second.

Finally, Carrasco and Kotchoni (2017) show that $\sqrt{T}(\hat{\theta}(\hat{\lambda}_{TM}) - \theta_0)$ and $\sqrt{T}(\hat{\theta}(\lambda_T) - \theta_0)$

have the same asymptotic distribution as soon as M goes to infinity first and T goes to infinity second.

Theorem 4 *Let $\widehat{\lambda}_{TM} \equiv \widehat{\lambda}_{TM}(\widehat{\theta}^1)$ and $\widehat{\theta}_T(\lambda, \theta_0) \equiv \widehat{\theta}_T(\lambda)$. Then under assumptions 1 to 6,*

$$\sqrt{T} \left(\widehat{\theta}_T \left(\widehat{\lambda}_{TM} \right) - \widehat{\theta}_T \left(\lambda_T \left(\theta_0 \right) \right) \right) = O_p \left(T^{1/2} M^{-1/2} \right),$$

as M goes to infinity first and T goes to infinity second.

Basically, theorem 4 implies that replacing λ_T by $\widehat{\lambda}_{TM}$ does not affect the consistency, asymptotic normality and efficiency of the final CGMM estimator $\widehat{\theta} \left(\widehat{\lambda}_{TM} \right)$.

6 Other Implementation issues

This section tackles algorithmic issues related to the implementations of the CGMM estimators. We present methods to evaluate the objective function of the CGMM and explain how to compute the covariance matrix of the CGMM estimator.

6.1 Computing the Objective Function

The objective function of the CGMM involves a d -dimensional integral against the measure π , where $d = \dim(\tau)$. We have:

$$\widehat{Q}_{T,\lambda} = \int_{\mathbb{R}^d} K_{T,\lambda}^{-1/2} \widehat{h}_T(\tau, \theta) \overline{K_{T,\lambda}^{-1/2} \widehat{h}_T(\tau, \theta)} \pi(\tau) d\tau.$$

When $d \geq 2$, the fast and accurate numerical evaluation of these integrals becomes an important issues. Indeed, the objective function must be evaluated a large number of times until the convergence of a numerical optimization algorithm. Two approaches are proposed here: quadrature methods and Monte Carlo integration.

6.1.1 Quadrature Method

Let us first assume that $d = 1$ and consider a $f(\tau, \theta), \tau \in \mathbb{R}$ that is continuously differentiable at any order. Then $f(\tau, \theta)$ can be expanded as:

$$f(\tau, \theta) = \sum_{k=0}^{2n-1} a_k(\theta) \tau^k + R(\tau, \theta) \tag{46}$$

where the remainder $R(\tau, \theta)$ is negligible for n large enough. Our objective is to approximate $\int_{\mathbb{R}} f(\tau, \theta) \pi(\tau) d\tau$ with high accuracy. Assume that the weighting function is Gaussian up to a transformation: $\pi(s) = \exp(-s^2)$. This function puts little weight on extreme values of $\widehat{h}_T(s, \theta)$, which is a desirable feature in finite samples where the behavior of the variance of \widehat{h}_t is not under control.

Our choice of weighting function allows us to approximate the objective function using Gauss-Hermite quadratures, which amounts to finding n points (τ_1, \dots, τ_n) and weights $(\omega_1, \dots, \omega_n)$ such that:

$$\int P(\tau) \exp\{-\tau^2\} d\tau = \sum_{k=1}^n \omega_k P(\tau_k), \quad (47)$$

for any polynomial function $P(\cdot)$ of order smaller or equal to $2n - 1$. The quadrature points and weights are determined by solving the system of equations:

$$\int s^l \exp\{-s^2\} ds = \sum_{k=1}^n \omega_k s_k^l \text{ for all } l = 0, \dots, 2n - 1. \quad (48)$$

This is a nonlinear system of $2n$ equations with $2n$ unknowns. Ultimately, the quadrature points and weights depend on the integration domain, the weighting function and the order of accuracy of the quadrature rule (controlled by n).

For the function $f(\tau, \theta)$ that admits an expansion of the form (46):

$$\int f(\tau) \exp\{-\tau^2\} d\tau - \sum_{k=1}^n \omega_k f(\tau_k) = \int R(\tau, \theta) \exp(-\tau^2) d\tau. \quad (49)$$

If $f(\tau)$ is analytic on \mathbb{R} , the approximation error $\int R(\tau, \theta) \exp(-\tau^2) d\tau$ can be made arbitrarily small by increasing n . However, this result has a more general scope as it applies also to functions that are analytic per sub-intervals of \mathbb{R} and bounded functions in particular.

Let us now consider the computation of $K_T \widehat{h}_T(\tau, \theta)$. We have:

$$\begin{aligned} K_T \widehat{h}_T(\tau, \theta) &= \int \widehat{k}_T(\tau, s) \widehat{h}_T(s, \theta) \exp(-s^2) ds. \\ &= \sum_{k=1}^n \omega_k \widehat{k}_T(\tau, s_k) \widehat{h}_T(s_k, \theta) \end{aligned}$$

Hence, we have the matrix approximation:

$$K_T \widehat{h}_T(\theta) \simeq \widehat{W}_T \widehat{h}_T(\theta), \quad (50)$$

where \widehat{W}_T is the matrix with (j, k) elements

$$W_{jk} = \omega_k \widehat{k}_T(s_j, s_k), \quad (51)$$

and $\widehat{h}_T(\theta) = \left(\widehat{h}_T(s_1, \theta), \dots, \widehat{h}_T(s_n, \theta) \right)'$.

The matrix \widehat{W}_T is the best finite dimensional approximation of the operator K_T . The regularized inverse of K_T is therefore approximated as:

$$K_{T,\lambda}^{-1} \approx \left(\widehat{W}_T^2 + \lambda_T I \right)^{-1} \widehat{W}_T, \quad (52)$$

so that:

$$K_{T,\lambda}^{-1} \widehat{h}_T(\theta) \approx \left(\widehat{W}_T^2 + \lambda_T I \right)^{-1} \widehat{W}_T \widehat{h}_T(\theta). \quad (53)$$

Substituting into the objective function of the CGMM yields:

$$\widehat{Q}_{T,\lambda} \approx \sum_{k=1}^n \omega_k \left[K_{T,\lambda}^{-1} \widehat{h}_T(s_k, \theta) \right] \left[\widehat{h}_T(s_k, \theta) \right], \quad (54)$$

where $K_{T,\lambda}^{-1} \widehat{h}_T(s_k, \theta)$ is the k^{th} element of the vector $K_{T,\lambda}^{-1} \widehat{h}_T(\theta)$.

In theory, the extension of the Gauss-Hermite quadrature method to the multivariate case is straightforward. When $\tau \in \mathbb{R}^d$, the d -dimensional set of multivariate quadrature points is given by the Cartesian product:

$$D = \left\{ \tau = (\tau_{(1)}, \dots, \tau_{(d)}) : \tau_{(i)} \in \{s_1, \dots, s_n\} \text{ for all } i = 1 \text{ to } d \right\},$$

where $\{s_1, \dots, s_n\}$ is the set of quadrature points in a one dimension integration and $\tau_{(i)}$ is the i^{th} coordinate of τ . Associated with each $\tau \in D$ is the weight:

$$\bar{\omega}(\tau_{(1)}, \dots, \tau_{(d)}) = \omega(\tau_{(1)}) \dots \omega(\tau_{(d)}),$$

where $\omega(\tau_{(i)}) = \omega_k$ if and only if $\tau_{(i)} = s_k, i = 1, \dots, d$. Finally:

$$\int_{\mathbb{R}^d} P(\tau) \exp\{-\tau^2\} d\tau = \sum_{k_1=1}^n \dots \sum_{k_d=1}^n \omega_{k_1} \dots \omega_{k_d} P(s_{k_1}, \dots, s_{k_d}), \quad (55)$$

The multivariate Gauss-Hermite quadrature has the undesirable feature that its complexity increases exponentially with the number of quadrature points in a one dimension. The method is therefore subject to a ‘‘curse of dimensionality’’. Indeed, the size of the matrix \widehat{W}_T is n^d in the representation $K_T \widehat{h}_T(\theta) \simeq \widehat{W}_T \widehat{h}_T(\theta)$. As \widehat{W}_T must be inverted at each iteration of the optimization algorithm, the CGMM becomes virtually infeasible by quadrature methods when $d \geq 3$. An alternative approach that circumvents this difficulty is suggested below.

6.1.2 Monte Carlo Integration

This approach relies on the alternative formula of the CGMM objective function provided in Carrasco Chernov, Florens, and Ghysels (2007):

$$\widehat{Q}_{T,\lambda} = v(\theta)' \left[\lambda_T I_T + \widehat{C}_T^2 \right]^{-1} \overline{v(\theta)}, \quad (56)$$

where $\widehat{C}_T \equiv \widehat{C}_T(\widehat{\theta}^1)$ is the square matrix of size T with (t, l) element $c_{t,l}/(T - \dim(\theta))$, I_T is the identity matrix of size T , and $v(\theta) = (v_1, \dots, v_T)'$ with:

$$v_t = \int \overline{h_t(\tau, \widehat{\theta}^1)} h_T(\tau, \theta) \pi(\tau) d\tau \text{ and} \quad (57)$$

$$c_{t,l} = \int \overline{h_t(\tau, \widehat{\theta}^1)} h_l(\tau, \widehat{\theta}^1) \pi(\tau) d\tau. \quad (58)$$

The main drawback of the above expressions lies in that it involves the inverse of the matrix \widehat{C}_T which has size T . However, this must be balanced by the fact that the objective function can be evaluated by Monte Carlo integration whatever the dimensionality of τ as explained below.

Let $\pi(\tau)$ denote the multivariate standard normal density and $(\tau^{(1)}, \dots, \tau^{(M)})$ be M values of τ simulated according to $\pi(\tau)$. The Monte Carlo approximations of v_t and

$c_{t,l}$ are:

$$\tilde{v}_t \approx \frac{1}{M} \sum_{k=1}^M \overline{h_t(\tau^{(k)}, \hat{\theta}^1)} h_T(\tau^{(k)}, \theta) \text{ and} \quad (59)$$

$$\tilde{c}_{t,l} \approx \frac{1}{M} \sum_{k=1}^M \overline{h_t(\tau^{(k)}, \hat{\theta}^1)} h_l(\tau^{(k)}, \hat{\theta}^1). \quad (60)$$

In order to guarantee the convergence of the optimization algorithm, it is crucial to simulate the set $(\tau^{(1)}, \dots, \tau^{(M)})$ once and for all at the beginning of the procedure and supply this as a fixed array to the code that evaluates the objective function of the CGMM.

An exercise that compares the performance of the Monte Carlo integration to that of the Hermitian quadrature would not be trivial as it requires to define an objective criteria to balances the computing cost against the statistical efficiency. Focusing on statistical efficiency alone, an intuitive reasoning suggests that Hermitian quadrature should be preferred when the dimensionality of τ is small while Monte Carlo integration is better otherwise.

6.2 Computing the Variance of the CGMM Estimator

The asymptotic covariance matrix of the second step CGMM estimator is given by:

$$\begin{aligned} AVar(\hat{\theta}) &= Var\left[\sqrt{T}(\hat{\theta} - \theta_0)\right] \\ &= \left\langle K^{-1/2}E\left(\hat{G}_t(\cdot, \theta)\right), K^{-1/2}E\left(\hat{G}_t(\cdot, \theta)\right) \right\rangle^{-1}, \end{aligned} \quad (61)$$

where $\hat{G}_t(\tau, \theta) = \frac{\partial \hat{h}_t(\tau, \theta)}{\partial \theta}$ is a column vector of length q with element $\hat{G}_{t,i}(\tau, \theta) = \frac{\partial \hat{h}_t(\tau, \theta)}{\partial \theta_i}$, and for any vectors f and g , $\langle f, g \rangle$ is a matrix with elements $\langle f, g \rangle_{i,j} = \langle f_i, g_j \rangle$, see Carrasco and Florens (2000). This covariance matrix is consistently estimated by:

$$\widehat{AVar}(\hat{\theta}) = \left\langle K_{T,\lambda}^{-1/2} \hat{G}_T(\tau, \hat{\theta}), K_{T,\lambda}^{-1/2} \hat{G}_T(\tau, \hat{\theta}) \right\rangle^{-1}, \quad (62)$$

where $\hat{G}_T(\tau, \hat{\theta}) = \frac{1}{T} \sum_{t=1}^T \hat{G}_t(\tau, \hat{\theta})$.

This formula is convenient when the scalar products are evaluated by quadrature

methods. Indeed, let us define:

$$\begin{aligned}\widehat{G}_{T,i}(\theta) &= \left(\widehat{G}_{T,i}(\tau_1, \theta), \dots, \widehat{G}_{T,i}(\tau_N, \theta) \right)' \text{ and} \\ K_{\lambda T}^{-1/2} \widehat{G}_{T,i}(\theta) &= \left(K_{T,\lambda}^{-1/2} \widehat{G}_{T,i}(\tau_1, \theta), \dots, K_{T,\lambda}^{-1/2} \widehat{G}_{T,i}(\tau_N, \theta) \right)',\end{aligned}$$

where $N = n^d$ and $\widehat{G}_{T,i}(\tau, \widehat{\theta}) = \frac{1}{T} \sum_{t=1}^T \widehat{G}_{t,i}(\tau, \widehat{\theta})$. Then we have:

$$K_{T,\lambda}^{-1/2} \widehat{G}_{T,i}(\theta) = \left(\widehat{W}_T^2 + \lambda_T I \right)^{-1/2} \widehat{W}_T^{1/2} \widehat{G}_{T,i}(\theta),$$

where \widehat{W}_T is defined in (51). The (i, j) element of $\widehat{AVar}(\widehat{\theta})^{-1}$ can then be computed as:

$$\left(\widehat{AVar}(\widehat{\theta})^{-1} \right)_{i,j} = \sum_{k=1}^N \omega_k \left(K_{T,\lambda}^{-1/2} \widehat{G}_{T,i}(\theta) \right)_k \left(\overline{K_{T,\lambda}^{-1/2} \widehat{G}_{T,j}(\theta)} \right)_k, \quad (63)$$

where $\left(K_{T,\lambda}^{-1/2} \widehat{G}_{T,i}(\theta) \right)_k$ is the k^{th} coordinate of $K_{T,\lambda}^{-1/2} \widehat{G}_{T,i}(\theta)$.

Alternatively, Carrasco Chernov, Florens, and Ghysels (2007) establish the expression:

$$\widehat{AVar}(\widehat{\theta}) = \left(\frac{1}{T - \dim(\theta)} V(\widehat{\theta})' \left[\lambda_T I_T + \widehat{C}_T^2 \right]^{-1} \overline{V(\widehat{\theta})} \right)^{-1}, \quad (64)$$

where \widehat{C}_T is the same as in (56), $V(\widehat{\theta})$ is the (T, q) matrix with (t, i) element:

$$V_{t,i} = \int \overline{h_t(\tau, \widehat{\theta}) \widehat{G}_{T,i}(\tau, \widehat{\theta})} \pi(\tau) d\tau,$$

Formula (64) is best suited when Monte Carlo integration is used to evaluate the objective function. In this case, $V_{t,i}$ is estimated as:

$$\widetilde{V}_{t,i} \approx \frac{1}{M} \sum_{k=1}^M \overline{h_t(\tau^{(k)}, \widehat{\theta}) \widehat{G}_{T,i}(\tau^{(k)}, \widehat{\theta})},$$

where $(\tau^{(1)}, \dots, \tau^{(M)})$ are M values of τ that are drawn from the multivariate normal density $\pi(\tau)$.

Finally, recall that the set $(\tau^{(1)}, \dots, \tau^{(M)})$ must be simulated once and for all at the beginning of the estimation procedure and supplied as a fixed array to the code that evaluates the objective function of the CGMM as well as to those that compute post-estimation outputs.

7 Application to the USD/GBP exchange rate

For this application we use the USD/GBP daily exchange rates (1 GBP = x USD) from 1980-01-01 to 2017-05-01, which are publicly available on the web site of the Federal Reserve Board of St Louis (the FRED database). We obtain the exchange rates at monthly frequency by subsampling the daily data.⁵ The descriptive statistics, modeling choices and estimation results are presented in separate subsections below.

7.1 Descriptive Analysis

Table 1 presents standard descriptive statistics for the time series of daily and monthly increments of the USD/GBP exchange rates. The sample averages are negative as a result of an overall appreciation of the USD vis-à-vis the GBP over the period targeted by our study (see Figure 1). The sample means are lower than the medians, which indicates that the empirical distributions are negatively skewed.

Table 1: Moments of the increments of the USD/GBP exchange rates

	Daily frequency	Monthly frequency
Mean	-0.0001	-0.0021
Median	0.0000	-0.0018
Minimum	-0.1161	-0.2620
Maximum	0.0707	0.1567
Standard deviation	0.0103	0.0501
Skewness	-0.3899	-0.6765
Excess kurtosis	4.9847	2.7232
First order autocorrelation	0.0459	0.0833

Unlike the excess kurtosis, the magnitude of the skewness increases with the level of aggregation. At first glance, the first order autocorrelation suggests that the monthly increments of the USD/GBP exchange rates are more predictable than their daily analogues but significance tests are needed in order to confirm this claim. Figure 1 suggests that exchange rates obey a mean reverting process.

⁵The monthly exchange rate data therefore consist of one observation at the beginning of every month, which is distinct from the monthly average of daily exchange rates.

Figure 1. Trajectory of the daily USD/GBP exchange rates

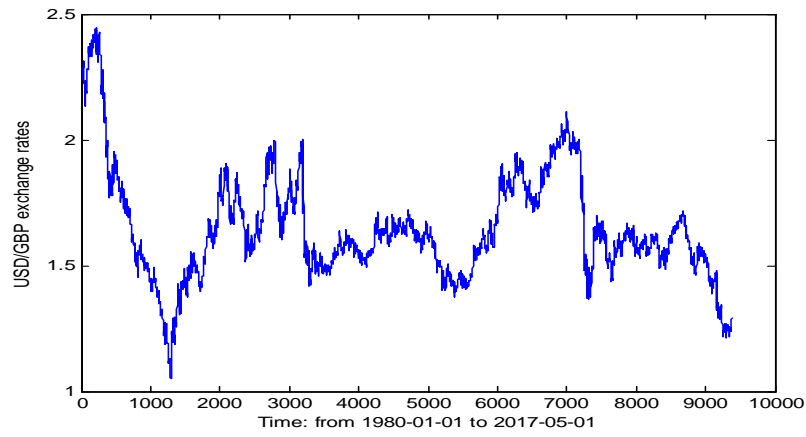


Figure 2. Increments of the USD/GBP exchange rate

Figure 2a. Daily frequency

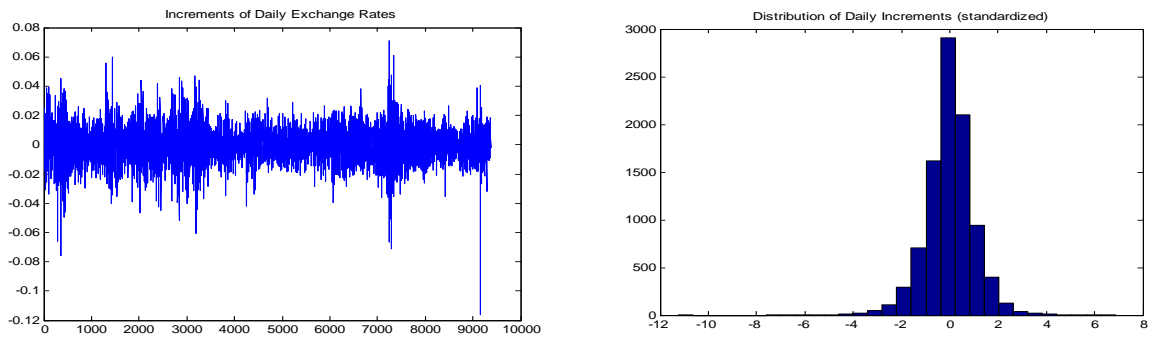


Figure 2b. Monthly frequency

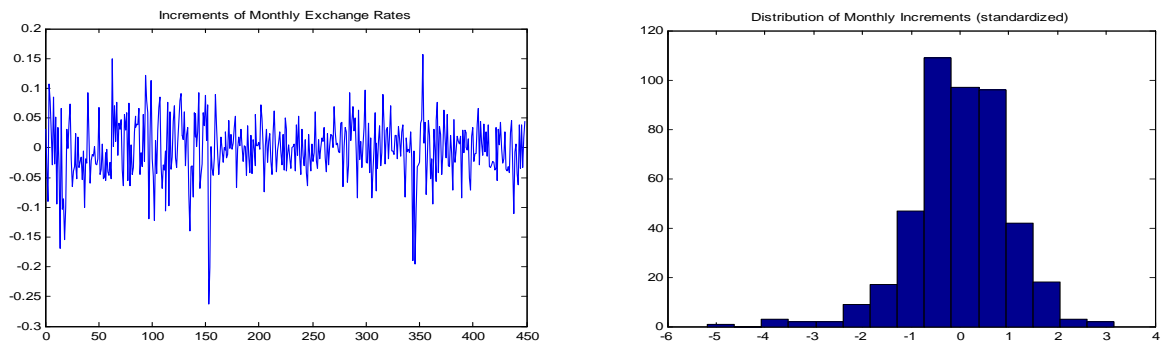


Figure 2 shows the trajectory and distribution of the daily and monthly increments of the USD/GBP exchange rates. It seems that the daily series may be subject to time varying volatility and rare jumps (Figure 2a, left). At monthly frequency (Figure 2b, left), time varying volatility is still visible but the jumps could probably be ignored at a modest cost. The histograms of the increments of the USD/GBP exchange rates (Figures 2a and 2b, right) indicate the presence of a modest negative skewness, significant excess kurtosis are rare extreme values. Indeed, a Jarque-Bera test for the null hypothesis that the increments of the USD/GBP exchange rate follow a normal distribution is strongly rejected at both frequencies (see Table 2).

Table 2: Jarque-Bera normality test for the increments of the USD/GBP exchange rates

	Daily frequency	Monthly frequency
Test statistic	9946.7	172.6
Critical value (5%)	5.986	5.842
P-value	< 0.001	< 0.001

Next, we use the daily data to compute the realized variance, skewness and kurtosis at monthly frequency using the following formulas:

$$\begin{aligned} \text{Realized Variance:} \quad RV_t &= \sum_{j=1}^{m_t} (x_{t,j} - x_{t,j-1})^2 \\ \text{Realized Skewness:} \quad RS_t &= \frac{\sqrt{m_t}}{RV_t^{3/2}} \sum_{j=1}^{m_t} (x_{t,j} - x_{t,j-1})^3 \\ \text{Realized Kurtosis:} \quad RK_t &= \frac{m_t}{RV_t^2} \sum_{j=1}^{m_t} (x_{t,j} - x_{t,j-1})^4 \end{aligned}$$

where $x_{t,j}$ is the exchange rate that has been observed during day j of month t and m_t is the number of observations during that month.

As realized measures are computed on intervals of time, the first month ($t = 1$) is January 1980 and the last month ($T = 448$) is April 2017. Figure 3 presents the trajectories and distributions of RV_t , RS_t and RK_t for the period 01/1980 - 04/2017. It is seen that all three realized moments exhibit substantial time variation. The histogram of the realized variance is reminiscent of a Gamma probability distribution function with a modest shape parameter that is contaminated with a few jumps. The histogram of the realized skewness is rather symmetric on the support $[-3,3]$. The empirical distribution

of the realized kurtosis has a Gamma shape as well, but it is less subject to extreme values than the realized variance.

Figure 3. Monthly Realized Measures

Figure 3a. Monthly Realized Variance

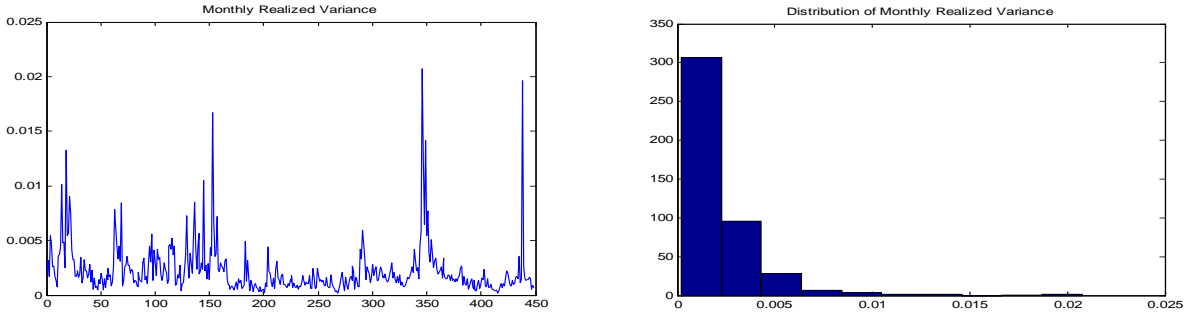


Figure 3b. Monthly Realized Skewness

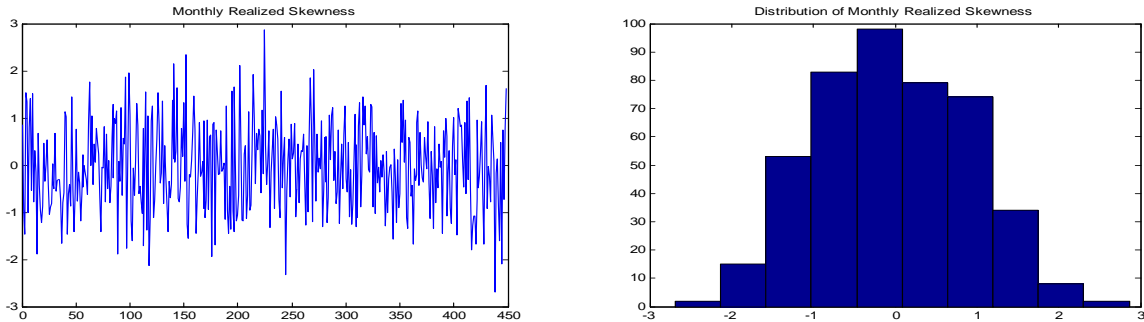
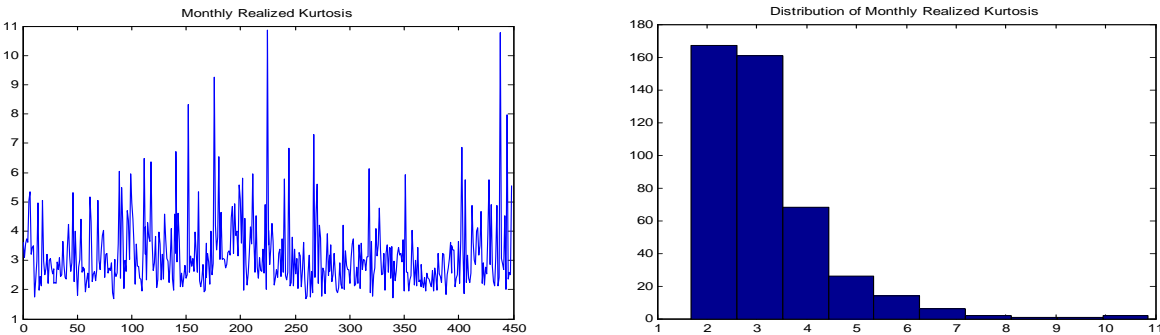


Figure 3c. Monthly Realized Kurtosis



Many authors have investigated theoretically and empirically whether exchange rates follow a random walk (see Rossi, 2013). Table 3 shows the results of Ljung-Box tests for the null hypothesis that the increments of exchange rates has no autocorrelation. We have performed the test at selected lags between 2 and 120 using daily data and between 2 and 30 using monthly data. The null hypothesis is rejected at all lags with daily data and never rejected with monthly data. These results are consistent with the presence of a low order moving average (MA) type noise contaminating the exchange rates.

Table 3: Ljung-Box tests for the null hypothesis that the increments of exchange rates are uncorrelated over time

Number of lags	Daily frequency			Monthly frequency		
	Test statistic	Critical Value (5%)	P-value	Test statistic	Critical Value (5%)	P-value
2	19.76	5.99	0.0001	3.55	5.99	0.1695
3	23.90	7.81	0.0000	4.00	7.81	0.2605
4	25.82	9.48	0.0000	4.13	9.48	0.3878
5	25.99	11.07	0.0001	4.96	11.07	0.4203
10	36.95	18.30	0.0001	11.81	18.30	0.2976
20	52.74	31.41	0.0001	24.76	31.41	0.2107
30	72.26	43.77	0.0000	35.46	43.77	0.2262
40	87.84	55.75	0.0000	–	–	–
60	122.41	79.08	0.0000	–	–	–
80	144.27	101.87	0.0000	–	–	–
100	172.93	124.34	0.0000	–	–	–
120	202.05	146.56	0.0000	–	–	–

Table 4 shows the results of variance ratio tests for the null hypothesis that the exchange rates obey a random walk.⁶ The null hypothesis is rejected at all lags for monthly data. For daily data, the hypothesis is not rejected at lags between 20 and 70 but it is rejected at all other lags. This is quite surprising given that the Ljung-Box tests support that the daily data violate the random walk assumption while the monthly data are uncorrelated. Together, the results of the Ljung-Box and Variance

⁶For details on the Variance Ratio test used here, see Lo and MacKinlay (1988, 1989).

Ratio tests suggest that the monthly increments of exchange rates are uncorrelated and conditionally heteroskedastic.

Table 4: Variance Ratio tests for the null hypothesis that the increments of exchange rates follow a random walk

Number of lags	Daily frequency			Monthly frequency		
	Test statistic	Critical Value (5%)	P-value	Test statistic	Critical Value (5%)	P-value
2	3.35	1.96	0.0008	4.70	1.96	0.0000
3	3.02	1.96	0.0025	4.65	1.96	0.0000
4	2.29	1.96	0.0216	4.80	1.96	0.0000
5	2.09	1.96	0.0361	4.89	1.96	0.0000
10	1.64	1.96	0.1006	3.42	1.96	0.0006
20	1.34	1.96	0.1801	2.15	1.96	0.0011
30	1.44	1.96	0.1497	2.07	1.96	0.0084
40	1.58	1.96	0.1120	–	–	–
60	1.67	1.96	0.0940	–	–	–
80	1.93	1.96	0.0529	–	–	–
100	2.13	1.96	0.0327	–	–	–
120	2.13	1.96	0.0330	–	–	–

7.2 Fitting a Variance Gamma Model to the USD/GBP Exchange Rate

We specify a mean reverting dynamics with Gamma variance for the monthly increment of the USD/GBP exchange rates:

$$e_t = x_t - x_{t-1} = \kappa (\mu - x_{t-1}) + \sqrt{y_t} u_t,$$

where x_t is the monthly USD/GBP exchange rate, $y_t \sim \text{Gamma}(\delta, \sigma)$ and $u_t \sim N(0, 1)$ is an IID noise. Conditionally on y_t and x_{t-1} , e_t follows a normal distribution with mean $\kappa (\mu - x_{t-1})$ and variance y_t . Therefore:

$$E[\exp(i\tau_1 e_t) | x_{t-1}, y_t] = \exp\{i\tau_1 \kappa (\mu - x_{t-1}) - \tau_1^2 y_t / 2\}, r \in \mathbb{R}.$$

The variance process (y_t) is allowed to be autocorrelated to the extent that its marginal distribution is a Gamma with the following density:

$$f(y) = \frac{1}{\Gamma(\delta)\sigma^\delta} y^{\delta-1} \exp(-y/\sigma).$$

The dynamic structure of y_t is left unspecified.

The CF of e_t conditional on x_{t-1} is given by:

$$\begin{aligned} \varphi(\tau_1, \theta, x_{t-1}) &= E[\exp(i\tau_1 e_t) | x_{t-1}] \\ &= \int_0^\infty \frac{1}{\Gamma(\delta)\sigma^\delta} y^{\delta-1} \exp(-y/\sigma) \exp\{i\tau_1 \kappa(\mu - x_{t-1}) - \tau_1^2 y/2\} dy \\ &= \frac{\exp\{i\tau_1 \kappa(\mu - x_{t-1})\}}{\Gamma(\delta)\sigma^\delta} \int_0^\infty y^{\delta-1} \exp\{-(1/\sigma + \tau_1^2/2)y\} dy \\ &= (1 + \tau_1^2 \sigma/2)^{-\delta} \exp\{i\tau_1 \kappa(\mu - x_{t-1})\}. \end{aligned}$$

where $\theta = (\kappa, \mu, \delta, \sigma)$ collects all the parameters of the model.

The following moment function may be used for the estimation of θ :

$$h_t(\tau, \theta) = \{\exp(i\tau_1 e_t) - \varphi(\tau_1, \theta, x_{t-1})\} \exp(i\tau_2 x_{t-1}), \quad \tau = (\tau_1, \tau_2)' \in \mathbb{R}^2.$$

The first step estimator is obtained as:

$$\hat{\theta}^{(1)} = \arg \min_{\theta} \int h_T(\tau, \theta) \overline{h_T(\tau, \theta)} \pi(\tau) d\tau,$$

where $h_T(\tau, \theta) = \frac{1}{T-1} \sum_{t=2}^T h_t(\tau, \theta)$ and $\pi(\tau) = \exp(-\tau_1^2 - \tau_2^2)$. As explained in Section 6, the objective function above is approximated with high precision by Gauss-Hermite quadrature.

We use this first step estimator to estimate the asymptotic covariance operator K by assuming that $h_t(\tau, \theta)$ is uncorrelated over time. The second step CGMM estimator is given by:

$$\hat{\theta}_T(\lambda) = \arg \min_{\theta} \int h_T(\tau, \theta) \overline{K_{\lambda T}^{-1} h_T(\tau, \theta)} \pi(\tau) d\tau,$$

where $K_{\lambda T}^{-1} = (K_T^2 + \lambda I)^{-1} K_T$, $K_T = K_T(\hat{\theta}^{(1)})$ is the estimate of K and $\lambda \in (0, 1)$ is the regularization parameter.⁷ It is important to make the parameter space bounded

⁷The MATLAB codes used in this chapter are available for sharing.

by imposing some restrictions during the numerical optimization. Here we imposed:

$$\theta = (\kappa, \mu, \delta, \sigma) \in (0, 1) \times (0, 2) \times (0, 2) \times (0, 2).$$

These restrictions are useful because the parameters of the Gamma distribution are difficult to identify even when the process is observed.

The ideal value of the regularization parameter is the one that minimizes the MSE of $\widehat{\theta}_T(\lambda)$. We estimate this MSE using the two naive bootstrap methods described in Subsection 5.1. In both cases, we use $M = 1000$ bootstrap samples of size $T - 1 = 447$. In the first approach (parametric bootstrap), $\widehat{\theta}^{(1)}$ is used to simulated the bootstrap sample whereas in the second approach (resampling), each bootstrap sample consists of random draws with replacement from $\{h_2(\tau, \theta), \dots, h_T(\tau, \theta)\}$. We compute the second step CGMM estimator from each sample and for each λ on the following grid:

$$\lambda \in \{10^{-9}; 5 \times 10^{-9}; 10^{-8}; 5 \times 10^{-8}; 10^{-7}; 5 \times 10^{-7}; 10^{-6}; 5 \times 10^{-6}; 10^{-5}; 5 \times 10^{-5}; 10^{-4}\}.$$

Let $\widehat{\theta}_T^{(m)}(\lambda_j)$ denote the second step estimator obtained using the m^{th} bootstrap sample and using λ_j as regularization parameter. The MSE of $\widehat{\theta}_T^{(m)}(\lambda_j)$ is estimated as:

$$MSE(\lambda_j) = \frac{1}{M} \sum_{m=1}^M \left\| \widehat{\theta}_T^{(m)}(\lambda_j) - \widehat{\theta}^{(1)} \right\|^2, j = 1, \dots, J.$$

The optimal regularization parameter is estimated as the minimizer $\widehat{\lambda}^{opt}$ of $MSE(\lambda_j)$, $j = 1, \dots, J$.

Figure 4 plots $MSE(\lambda)$ as a function of $\ln(\lambda)$ for the two naive bootstrap approaches. Interestingly, the two MSE curves have similar shapes. The MSE tends to be smaller under the parametric bootstrap than under the naive resampling. Both curves increase slowly and rather erratically between $\lambda = 10^{-9}$ and $\lambda = 5 \times 10^{-6}$, and then increase faster above this range. None of the two curves is strictly convex, which is possibly due to the fact that $MSE(\lambda)$ is too noisy for the number of bootstrap replications considered (i.e., M is not large enough), or to the fact that $MSE(\lambda)$ does not approximate $AMSE(\lambda, \theta_0)$ well enough (i.e., T is not large enough). Given the shape of the MSE curves, $\widehat{\lambda}_{TM} = 10^{-6}$ is a good choice of regularization parameter for this application. Indeed, the sensitivity of $MSE(\lambda)$ to λ is low on the range $\lambda \leq 10^{-6}$ while

$MSE(\lambda)$ increases fast on the range $\lambda \geq 10^{-6}$.⁸

Figure 4: MSE of the second step CGMM estimator as a function of $\ln(\lambda)$

Number of bootstrap replications: $M = 1000$.

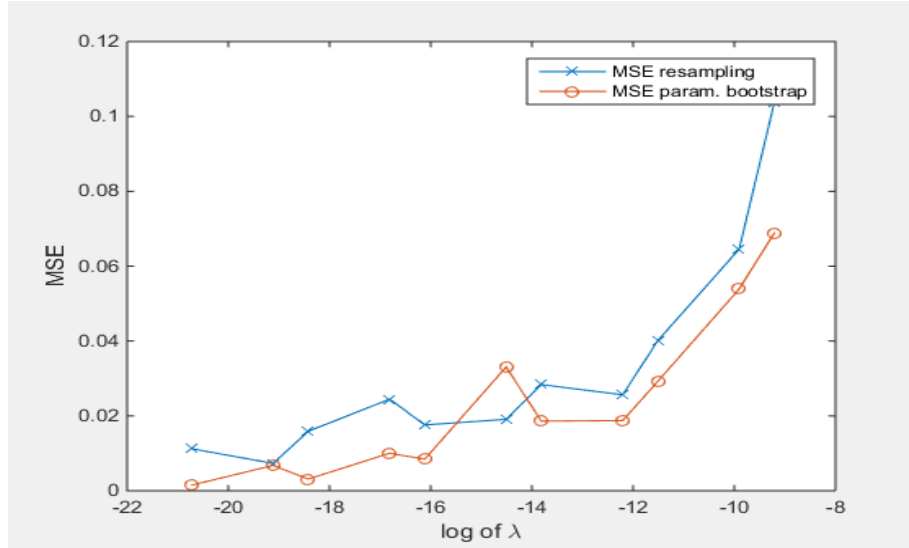


Table 5: Estimated parameters for the Variance Gamma model

These results are computed using $\hat{\lambda}_{TM} = 10^{-6}$

		$\hat{\kappa}$	$\hat{\mu}$	$\hat{\delta}$	$\hat{\sigma}$
Naive Resampling	Mean	0.0296	1.5604	0.0763	0.0366
	Median	0.0308	1.5814	0.0754	0.0320
	Standard Dev.	0.0069	0.1265	0.0878	0.0644
	Percentile 2.5%	0.0133	1.2740	0.0404	0.0209
	Percentile 97.5%	0.0440	1.7266	0.1037	0.0586
Parametric Bootstrap	Mean	0.0323	1.5810	0.0786	0.0353
	Median	0.0316	1.5814	0.0732	0.0294
	Standard Dev.	0.0094	0.0649	0.1088	0.0498
	Percentile 2.5%	0.0145	1.4307	0.0224	0.0158
	Percentile 97.5%	0.0567	1.7053	0.1068	0.0899

The second step CGMM estimator from the m^{th} sample is $\hat{\theta}_T^{(m)}(\hat{\lambda}_{TM})$. We use $\hat{\theta}_T^{(m)}(\hat{\lambda}_{TM}), m = 1, \dots, M$ to compute the summary statistics that are presented in

⁸Ultimately, the optimal choice of λ is intended to jointly control two risks, which are the unbound-
edness of $K_{T,\lambda}^{-1}$ as λ becomes smaller and smaller, and the inefficiency of the CGMM estimator as λ
becomes larger and larger.

Table 5. The rows of the table are respectively the mean, median, standard deviation, 2.5th and 97.5th percentiles of the empirical distribution of $\hat{\theta}_T^{(m)}(\hat{\lambda}_{TM})$. The results are rather similar for both estimation methods although the medians are slightly closer to each other than the means. The standard deviations are often higher under the parametric bootstrap than under the naive resampling, which is not surprising given that the sampling errors associated with the latter approach are constrained by the available data. The low value of $\hat{\kappa}$ suggests that the process in level (x_t) is close to a random walk.

We compute the second step CGMM estimator $\hat{\theta}_T(\hat{\lambda}_{TM})$ using the actual data and use it to filter the latent variance process (y_t). As y_t is assumed to be an IID *Gamma*(δ, σ) process, the only means to obtain time varying predictions is by conditioning on (e_t, x_{t-1}) . The resulting prediction, $E(y_t|e_t, x_{t-1})$ is a backcast rather than a forecast as it updates the constant prior mean $E(y_t) = \delta\sigma$ into the posterior means $E(y_t|e_t, x_{t-1}), t = 2, \dots, T$ upon observing the realizations of e_t and x_{t-1} .

This posterior mean is given by:

$$E(y_t|e_t, x_{t-1}) = \sigma \frac{\int_0^\infty v^{\delta-1/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2\sigma v} - v\right) dv}{\int_0^\infty v^{\delta-3/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2\sigma v} - v\right) dv}.$$

We approximate this expression by Monte Carlo integration as follows:

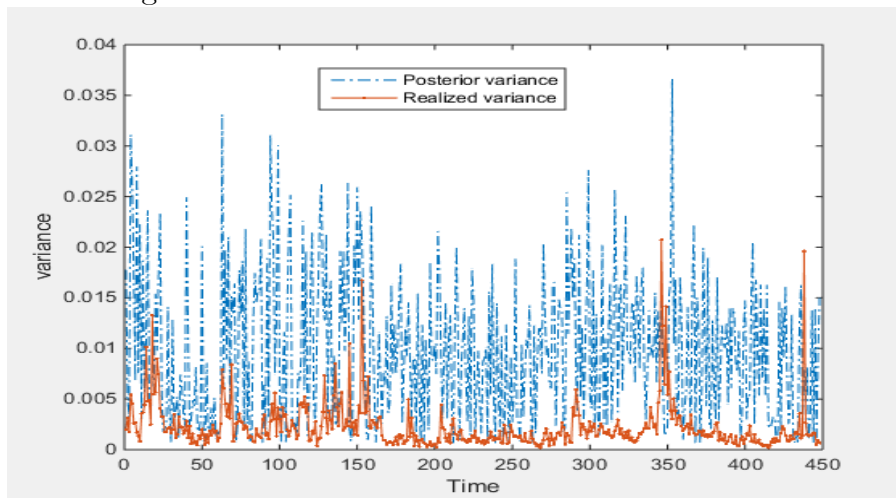
$$\hat{E}(y_t|e_t, x_{t-1}) \simeq \hat{\sigma} \frac{\frac{1}{1000} \sum_{m=1}^{1000} v_m^{\delta-1/2} \exp\left(\frac{-(e_t - \hat{\kappa}\hat{\theta} + \hat{\kappa}x_{t-1})^2}{2\hat{\sigma}v_m}\right)}{\frac{1}{1000} \sum_{m=1}^{1000} v_m^{\delta-3/2} \exp\left(\frac{-(e_t - \hat{\kappa}\hat{\theta} + \hat{\kappa}x_{t-1})^2}{2\hat{\sigma}v_m}\right)},$$

where $v_m, m = 1, \dots, 1000$ are independent draws from the exponential distribution $Exp(1)$.

Figure 5 shows the time series plot of $\hat{E}(y_t|e_t, x_{t-1})$ superimposed to RV_t . Clearly, the two processes do not capture the same information about the exchange rate variability over time. Indeed, $\hat{E}(y_t|e_t, x_{t-1})$ is much volatile due to the conditioning on e_t . Given that the model only specifies the unconditional distribution for the latent variance process, the conditioning on e_t is the price to pay in order to obtain time varying predictions for y_t . The main lesson learned from this empirical application is that a model that specifies the latent variance of the exchange rate increments as IID will in general deliver poor forecasts of that variance. A model that specifies the autocorrela-

tion structure of the variance (e.g., autoregressive Gamma) can produce time varying predictions that are not conditioned on e_t and is expected to produce better variance predictions.

Figure 5: Posterior Variance vs. Realized Variance



8 Conclusion

This chapter is aimed at popularizing the CGMM to applied researchers. First, we discuss several empirically relevant situations where the CGMM procedure can be successfully used to estimate a finite dimensional parameter. Although most of our examples are based on the characteristic function, we underscore that the CGMM can also be used to deal with conditional moment restrictions. Second, we present the construction of the objective function of the CGMM procedure and review the asymptotic properties of the CGMM estimator. Third, we discuss the selection of the regularization parameter and present numerical algorithms that are useful for the implementation of the CGMM estimator. Finally we present an empirical application where a Variance Gamma model is fitted to the USD/GBP exchange rate.

This model assumes that a monthly increment of exchange rate follows a Gaussian distribution conditional on its variance while the variance itself follows IID Gamma process. The only means to obtain time varying variance predictions within this model is by computing its posterior means, that is, its expectation conditional on the ex post realizations of the exchange rate increments. We find that the posterior mean of the variance process is a poor predictor of the monthly realized variance inferred from daily

data. This suggests that model that specify the variance of the exchange rate increments as IID should be avoided in empirical applications.

By avoiding the discretization of a continuum of moment condition, the CGMM permits its efficient use and therefore permits to achieve the maximum likelihood efficiency when the whole procedure is based on the characteristic function. When the moment restrictions are deduced from a characteristic function, the CGMM has the potential to achieve the maximum likelihood efficiency. In the case where a conditional moment restriction is converted into a continuum of moment condition, the CGMM permits to avoid the arbitrariness of the choice of the instruments and to efficiently exploit the information content of this restriction. However, the efficiency gain of the CGMM over the GMM comes with a slightly higher computation burden.

When the dimensionality of the model is large, quadrature methods are inefficient. In this case, one may resort to the Indirect CGMM procedure proposed in Kotchoni (2014), which involves solving a large number of one-dimensional models and optimally combining the solutions into a final estimator.

References

- [1] Andrew, D.W.K. and J.C. Monahan (1992) "An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator", *Econometrica*, vol. 60, 953-966.
- [2] Rossi, B. (2013) "Exchange Rate Predictability", *Journal of Economic Literature* 51, No. 4 (DECEMBER 2013), pp. 1063-1119. Stable URL: <http://www.jstor.org/stable/23644817>
- [3] Bierens, H. (1982), "Consistent Model Specification Tests," *Journal of Econometrics* 20, 105–134.
- [4] Buse, A. (1992), "The Bias of Instrumental Variable Estimators", *Econometrica* 60:1, 173–180.
- [5] Carrasco M., Chernov M., Florens J. P. and E. Ghysels (2007) "Efficient estimation of general dynamic models with a continuum of moment conditions", *Journal of Econometrics*, vol. 140, 529-573.
- [6] Carrasco, M. and J. P. Florens (2000) "Generalization of GMM to a continuum of moment conditions", *Econometric Theory*, vol. 16, 797-834.
- [7] Carrasco, M., J. P. Florens, and E. Renault (2007) "Linear Inverse Problems in Structural Econometrics: Estimation based on spectral decomposition and regularization", in the *Handbook of Econometrics*, vol. 6, edited by J.J. Heckman and E.E. Leamer.
- [8] Carrasco, M. and R. Kotchoni (2017) "Efficient Estimation Using the Characteristic Function", *Econometric Theory*, vol 33, Issue 2 April 2017, pp. 479-526
- [9] Chacko, G. and L. Viceira (2003), "Spectral GMM estimation of continuous-time processes," *Journal of Econometrics*, 116, 259-292.
- [10] Cochrane, J.H. (2009) "*Asset Pricing*" (Revised Edition), Princeton University Press.
- [11] Cox, J.C., J.E. Ingersoll and S.A. Ross (1985), "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, 385–407.

- [12] Devroye, L. (1986), "Non-Uniform Random Variate Generation," Springer-Verlag.
- [13] Dominguez, M. A. and I. N. Lobato (2004) "Consistent Estimation of Models defined by Conditional Moment Restriction," *Econometrica* 72 (5), 1601–1615.
- [14] Donald, S. and W. Newey (2001), "Choosing the number of instruments," *Econometrica*, 69, 1161-1191.
- [15] Duffie, D. and K. Singleton (1993), "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica*, 61, 929-952.
- [16] Feuerverger, A. and P. McDunnough (1981a), "On Efficient Inference in Symmetry Stable Laws and Processes," Csorgo, M., ed. *Statistics and Related Topics*. New York: North Holland, 109-122.
- [17] Feuerverger, A. and P. McDunnough (1981b), "On some Fourier Methods for Inference," *J. R. Statist. Assoc.* 76:379-387.
- [18] Feuerverger, A. and P. McDunnough (1981c), "On the Efficiency of Empirical Characteristic Function Procedures," *J. R. Statist. Soc. B*, 43, 20-27.
- [19] Feuerverger, A. and R. Mureika (1977), "The empirical characteristic function and its applications," *The annals of statistics*, 5, No 1, 88-97.
- [20] Gouriéroux, C. and J. Jasiak (2005) "Autoregressive Gamma Processes," *Journal of Forecasting*, vol. 25, 129–152. DOI: 10.1002/for.978.
- [21] Hansen, L. (1982) "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, vol. 50, 1029-1054.
- [22] Jiang, G. and J. Knight (2002), "Estimation of continuous time processes via the empirical characteristic function," *Journal of Business and Economic Statistics*, 20, 198-212.
- [23] Knight, J. L. and J. Yu (2002) "The empirical characteristic function in time series estimation", *Econometric Theory*, vol. 18, 691-721.
- [24] Koenker, R., Machado, J. A. F., Skeels, C. L. and A. H. I. Welsh (1994), "Momentary Lapses: Moment Expansions and the Robustness of Minimum Distance Estimators," *Econometric Theory*, 10, 172-197.

- [25] Kotchoni, R. (2012) "Applications of the characteristic function based continuum GMM in finance," *Computational Statistics and Data Analysis* vol. 56 (11), 3599–3622.
- [26] Kotchoni, R. (2014) "The indirect continuous-GMM estimation," *Computational Statistics and Data Analysis*, vol. 76, August 2014, Pages 464-488
- [27] Lavergne, P. and V. Patilea (2013), "Smooth minimum distance estimation and testing with conditional estimating equations: uniform in bandwidth theory," *Journal of Econometrics*, vol. 177, Issue 1, November 2013, Pages 47-59.
- [28] Lo, A. W. and A. C. MacKinlay (1988) "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test." *Review of Financial Studies*. Vol. 1, 1988, pp. 41-66.
- [29] Lo, A. W. and A. C. MacKinlay (1989) "The Size and Power of the Variance Ratio Test." *Journal of Econometrics*. Vol. 40, 1989, pp. 203-238.
- [30] Madan, D. B. and E. Senata (1990) "The Variance Gamma Model for share market returns", *Journal of Business*, vol. 63, no. 4.
- [31] Mandelbrot, B. (1963) "The variations of certain speculative prices", *Journal of Business*, vol. 36, 394-419.
- [32] McCulloch, J.H. (1986) "Simple Consistent Estimators of Stable Distribution Parameters", *Communications in Statistics - Simulation and Computation*, vol. 15, 1109-1136.
- [33] McCulloch, J. H. (1998) "Numerical approximation of the symmetric stable distribution and density" In R. J. Adler, R. E. Feldman, & M. S. Taqqu (eds.), *A Practical Guide to Heavy Tails*, 489–500. Boston: Birkhäuser.
- [34] Mittnik, S., Rachev, S. T., Doganoglu, T. and D. Chenyao (1999) "Maximum likelihood estimation of stable paretian models", *Mathematical and Computer Modelling*, vol. 29, 276-293.
- [35] Nagar, A. L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 573-595.

- [36] Newey, W. K. and D. McFadden (1994), "Large Sample Estimation and Hypotheses Testing," *Handbook of Econometrics*, Vol IV, ed. by R.F. Engle and D.L. McFadden.
- [37] Newey, W. K. and R. J. Smith (2004), "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72:1, 219–255.
- [38] Newey, W. and K. West (1987) "A simple positive definite, heteroskedasticity and autocorrelation consistent covariance matrix", *Econometrica*, vol. 55, 703-708.
- [39] Nolan, J. P. (1997) "Numerical calculation of stable densities and distribution functions", *Communication Statistics: Stochastic Models*, vol. 13, 759-774.
- [40] Paoletta, M. (2007) *“Intermediate Probability: A Computational Approach”*, Wiley, Chichester.
- [41] Paulson, A. S., Holcomb W. E. and R. A. Leitch (1975), "The estimation of the parameters of the stable laws," *Biometrika*, 62:1, 163-170.
- [42] Phillips, P. C. B. and H. R. Moon (1999), "Linear Regression Limit Theory for Nonstationary Panel Data," *Econometrica*, 67:5, 1057–1112.
- [43] Press, S. J. (1972) "Estimation in univariate and multivariate stable distributions." *J. Am. Statist. Assoc.* 67:842–846.
- [44] Rilstone, P., Srivastava, V. K., and A. Ullah (1996), "The Second-Order Bias and Mean-Squared Error of Nonlinear Estimators," *Journal of Econometrics*, 75, 369-395.
- [45] Rothenberg, T. J. (1983), "Asymptotic properties of Some Estimators in Structural Models," *Studies in Econometrics, Time Series and Multivariate Statistics*, edited by S. Karlin, T. Amemiya and L. A. Goodman, New York: Academic Press.
- [46] Rothenberg, T. J. (1984), "Approximating the Distributions of Econometric Estimators and Test Statistics," *Handbook of Econometrics*, Vol. 2, ed. by Z. Griliches and M. D. Intriligator. New York: North-Holland.
- [47] Singleton, K. J. (2001), "Estimation of Affine Pricing Models Using the Empirical Characteristic Function," *Journal of Econometrics*, 102, 111-141.

- [48] Thorton, J. C. and A. S. Paulson (1977) "Asymptotic distribution of CF-based estimator for the stable laws," *Sankhya* 39:341–354.
- [49] Weron, R. (1996) "On the Chambers-Mallows-Stuck method for simulating skewed stable random variables", *Statistics & Probability Letters*, vol. 28, 165-171.
- [50] Yu, J. (2004) "Empirical Characteristic Function Estimation and Its Applications", *Econometric Reviews*, vol. 23, No 2, 93-123.
- [51] Zolotarev, V. (1986), "One-Dimensional Stable Distributions", *American Mathematical Society, Providence, RI*. Russian original, 1983.

Appendix: Prediction of the volatility process

The joint distribution of (e_t, y_t) given x_{t-1} is given by:

$$\begin{aligned}
 f(e_t, y_t | x_{t-1}) &= f(e_t | x_{t-1}, y_t) f(y_t | x_{t-1}) \\
 &= f(e_t | x_{t-1}, y_t) f(y_t) \\
 &= \frac{y_t^{\delta-3/2}}{\sqrt{2\pi}\Gamma(\delta)\sigma^\delta} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2y_t} - \frac{y_t}{\sigma}\right).
 \end{aligned}$$

The distribution of e_t given x_{t-1} is obtained by marginalizing $f(e_t, y_t | x_{t-1})$:

$$\begin{aligned}
 f(e_t | x_{t-1}) &= \int_0^\infty f(e_t, y_t | x_{t-1}) dy_t \\
 &= \frac{1}{\sqrt{2\pi}\Gamma(\delta)\sigma^\delta} \int_0^\infty y_t^{\delta-3/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2y_t} - \frac{y_t}{\sigma}\right) dy_t \\
 &= \frac{1}{\sqrt{2\pi}\sigma\Gamma(\delta)} \int_0^\infty v^{\delta-3/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2\sigma v} - v\right) dv.
 \end{aligned}$$

The distribution of y_t given e_t and x_{t-1} is:

$$\begin{aligned}
 f(y_t | e_t, x_{t-1}) &= \frac{f(e_t, y_t | x_{t-1})}{f(e_t | x_{t-1})} \\
 &= \frac{1}{\sigma^{\delta-1/2}} \frac{y_t^{\delta-3/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2y_t} - \frac{y_t}{\sigma}\right)}{\int_0^\infty v^{\delta-3/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2\sigma v} - v\right) dv}.
 \end{aligned}$$

Hence, the expectation of y_t given e_t and x_{t-1} is given by:

$$\begin{aligned}
 E(y_t | e_t, x_{t-1}) &= \frac{1}{\sigma^{\delta-1/2}} \frac{\int_0^\infty y_t^{\delta-1/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2y_t} - \frac{y_t}{\sigma}\right) dy_t}{\int_0^\infty v^{\delta-3/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2\sigma v} - v\right) dv} \\
 &= \frac{\sigma \int_0^\infty v^{\delta-1/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2\sigma v} - v\right) dv}{\int_0^\infty v^{\delta-3/2} \exp\left(\frac{-(e_t - \kappa\theta + \kappa x_{t-1})^2}{2\sigma v} - v\right) dv}.
 \end{aligned}$$

■