Generalized Empirical Likelihood for a Continuum of Moment Conditions

Pierre Chaussé [*][†]

March 2017

Abstract

This paper extends the generalized empirical likelihood method to the case in which the moment conditions are defined on a continuum (CGEL). We show, for the iid case, that CGEL is asymptotically equivalent at the first order to the generalized method of moments for a continuum (CGMM) developed by Carrasco and Florens (2000). Because the system of equations that we need to solve becomes singular when the number of moment conditions converges to infinity, we treat CGEL as a nonlinear ill-posed problem and obtain the solution using the regularized Gauss-Newton method. This numerical algorithm is a fast and relatively easy way to compute the regularized Tikhonov solution to nonlinear ill-posed problems in function spaces. In order to compare the properties of CGEL and CGMM, we then perform a numerical study in which we estimate the parameters of a stable distribution using moment conditions based on the characteristic function. The results show that CGEL outperforms CGMM in most cases according to the root mean squared error criterion.

Classification JEL: C13, C30

1 Introduction

When estimating models based on moment conditions, it is often the case that the number of conditions is so large that selecting the right ones becomes an issue. For example, in the case of linear models with endogenous regressors as considered by Carrasco (2011), the set of possible instruments can be countably infinite or defined on a continuum. Moment conditions can also be naturally based on a continuum when, for example, they are defined by characteristic functions or spectral densities. In these cases, methods such as instrumental variables (IV) cannot be based on the whole set of moment conditions because the system of equations implied by the first order conditions becomes singular as the number of conditions increases beyond the sample size. Because it may reduce the quality of the estimators when

[*] Department of Economics, University of Waterloo, Ontario, Canada (pchausse@uwaterloo.ca)

[†] I would like to thank Thomas Parker, Alain Guay, Marine Carrasco, Éric Renault, Douglas Hodgson, Jean-Pierre Florens, Jeroen Rombouts, Steve Ambler and anonymous referees for valuable comments and suggestions.

weak instruments are chosen and inversely for strong instruments, we have to be careful in the selection. Donald and Newey (2001) present a method for selecting the optimal number of instruments but it requires a certain ordering so that the stronger are selected and the weaker are dropped. On the other hand, Carrasco (2011) apply the generalized method of moment for a continuum (CGMM) of Carrasco and Florens (2000) in which the whole set of instruments can be used without imposing any ordering. The method is based on a Tikhonov regularization technique which is comparable to a principal component selection procedure. The most influential moment conditions are therefore automatically selected. In this paper, we extend the generalized empirical likelihood method (GEL) of Smith (1997) so that it can also deal with a continuum of moment conditions (CGEL). The second order asymptotic results obtained by Newey and Smith (2004) and Anatolyev (2005) suggest that CGEL may be a good alternative to CGMM. The CGEL estimator is defined as the solution to a constrained optimization problem in which the number of constraints is infinite. Such a problem cannot easily be solved using a finite number of observations. The main contribution of the paper is to show both theoretically and practically how we can obtain a stable solution to such problems. The method can even be applied to cases in which the number of conditions is finite but large enough so that the problem becomes ill-conditioned. It offers a way to deal with the selection of moment conditions using a Tikhonov type approach similar to CGMM. Furthermore, we present the algorithms in matrix notation to simplify its implementation.

When defining the objective function of the efficient CGMM, we need the regularized solution to a linear ill-posed problem, because the optimal weighting operator cannot be continuously inverted. On the other hand, the objective function of CGEL is well defined. However, the system of equations from which we compute the Lagrange multiplier associated with the moment conditions becomes singular when the number of conditions goes to infinity. As a result, we present CGEL as a nonlinear ill-posed problem in the sense that a unique and stable solution cannot be obtained directly from the first order conditions. The literature in applied mathematics offers several ways to deal with nonlinear ill-posed problems. As a first procedure, we apply the regularized Gauss-Newton algorithm which can be compared to using ridge regression techniques to estimate a poorly conditioned nonlinear regression. We also present an alternative regularized method which is based on the singular value decomposition of the first order Taylor approximation of the solution. This method has the advantage of being less computationally demanding and asymptotically equivalent to the first procedure. We present the algorithms for the exponential tilting (CET), the empirical likelihood (CEL) and the Euclidean empirical likelihood (CEEL) for a continuum by using a matrix notation as in Carrasco et al. (2007a) for CGMM. Moreover, in order to test the over-identifying restrictions, we present a normalized version of the three tests proposed by Smith (2004) so that they are all asymptotically distributed as a standardized normal distribution. We conclude the theoretical part with a brief discussion on how to implement the exponentially tilted empirical likelihood of Schennach (2007) for a continuum (CETEL).

We perform a numerical study in which we compare the finite sample properties of the three CGEL methods using the two proposed algorithms with CGMM. We use the example of estimating the parameters of a stable distribution using the marginal characteristic function as in Carrasco and Florens (2002) and Garcia et al. (2006). We also compare the empirical sizes of the three tests of over-identifying restrictions. All the results are computed for different values of the regularization parameter because no data-driven method is available to select its optimal value. What we get suggests that CGEL may outperform CGMM according to the root mean squared error criterion. We also compare CGEL with the maximum likelihood estimator and find that for a good selection of the regularization parameter, we can get comparable properties.

The paper is organized as follows. Section 2 gives an overview of GEL while section 3 presents the CGEL method and section 5 the three tests of over-identifying restrictions. Section 4 describes the two numerical algorithms, section 6 presents the numerical study and section 7 concludes.

2 GEL

This section presents an overview of the GEL method when there are a finite number of moment conditions. It serves as an introduction to the next section, which generalizes the method to the case of a continuum of conditions. Therefore we express the function defining the moment conditions in a way that facilitates the transition from GEL to CGEL.

We suppose that the vector $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ is uniquely identified by a vector of q moment conditions. Instead of writing these conditions in the usual way as $E[g_\tau(X;\theta_0)] = 0$ for $\tau = 1, ..., q$, we incorporate the index in the function as follows:

$$E^{P_0} g(X, \tau_i; \theta_0) = 0 \ \ \forall i = 1, \cdots, q, \tag{1}$$

where the index i implies that τ belongs to a countable set (finite in this section), and P_0 is the true probability distribution associated with the random variable X. For example, if we are estimating a linear model using instruments, $\tau_i = i$ and defines the condition associated with the ith instrument. But it could also be an element of the function if, for example, the vector of parameters is estimated using characteristic functions. In this case, τ_i would be equal to some selected points of the function which are the most susceptible of producing good estimates.

We suppose that we can estimate the moment function from a vector of n i.i.d. realizations of the random variable X, $\{x_1, x_2, \cdots, x_n\}$. In general, we can write the $q \times 1$ vector of sample moment conditions as follows:

$$\tilde{g}(\theta) = \sum_{t=1}^{n} p_t g(x_t; \theta),$$

where p_t is the probability associated with the realization x_t.

The GMM estimator is defined as the vector of parameters that minimizes the norm of the sample moment $\bar{g}(\theta)$, which is based on the empirical density of the observations $f_n(x_t) = (1/n) \ \forall t$. Going through this optimization problem if necessary because when $q > p$, there is no solution to the sample moment conditions $\bar{g}(\theta) = 0$, in which p_t is restricted to $1/n$. On the other hand, the GEL method consists in finding the implied probabilities p_t, which are as close as possible to $1/n$ according to a certain family of discrepancies $h_n(p_t)$, that satisfy the conditions exactly. This interpretation represents the primal problem which defines the GEL estimators:

$$\hat{\theta}_{gel} = \arg\min_{\theta, p_t} \sum_{t=1}^{n} h_n(p_t) \tag{2}$$

subject to

$$\begin{aligned} \sum_{t=1}^{n} p_t g(x_t, \tau_i; \theta) &= 0 \quad \forall i = 1, \cdots q \text{ and} \\ \sum_{t=1}^{n} p_t &= 1, \end{aligned} \qquad \begin{aligned} (3) \\ (4) \end{aligned}$$

as long as $h_n(p_t)$ belongs to the following Cressie-Read family of discrepancies:

$$h_n(p_t) = \frac{[\gamma(\gamma+1)]^{-1}[(np_t)^{\gamma+1} - 1]}{n}.$$

Smith (1997) shows that the empirical likelihood method (EL) of Owen (2001) ($\gamma = -1$) and the exponential tilting of Kitamura and Stutzer (1997) ($\gamma = 0$) belong to the GEL family of estimators while Newey and Smith (2004) show that it is also the case for the continuous updated estimator (CUE) of Hansen et al. (1996) ($\gamma = 1$). They all have in common that we can express their dual problem as:

$$\hat{\theta}_{gel} = \arg\min_{\theta \in \Theta} \left[\max_{\lambda \in \Lambda_n} \frac{1}{n} \sum_{t=1}^{n} \rho\left(\lambda' g(x_t; \theta)\right) \right], \tag{5}$$

where $\rho(v)$ is a strictly concave function that depends on $h_n(p_t)$ and is normalized so that $\rho'(0) = \rho''(0) = -1$. We can show that $\rho(v) = \ln(1 - v)$ corresponds to EL, $\rho(v) = -\exp(v)$ to ET and to CUE if $\rho(v)$ is quadratic. We assume that Θ is a compact set and λ, which is the $q \times 1$ vector representing the Lagrange multiplier associated with the constraint (3), belongs to $\Lambda_n = \{\lambda : \lambda' g(x_t; \theta) \in \mathcal{D} \; \forall \; x_t\}$, where $\mathcal{D}$ is the domain of $\rho(v)$.

Newey and Smith (2004) and Anatolyev (2005) show that the EL estimator has a lower second order asymptotic bias than ET and CUE and that its bias corrected version is higher order efficient. This performance is, to some extent, due to the fact that EL's estimators of the Jacobian and second moment matrices, as opposed to the other GEL methods, are based on the implied probabilities which carry more information than $1/n$ (see Antoine et al. (2007)). However, because of the non negativity constraint that we need to impose on these implied probabilities, ET offers a natural way to meet this requirement, which makes it numerically more stable than EL especially in presence of model misspecification. In response to this, Schennach (2007) combines ET and EL in a method called the exponentially tilted empirical likelihood (ETEL). This method shares the same second order properties of EL and the stability of ET in presence of model misspecification. Although it does not belong to the GEL family, we offer below a brief discussion because its computational stability is appealing especially in the case of a continuum of conditions.

We can easily verify the equivalence of the primal and dual problems by showing that they share the same following first order conditions:

$$\sum_{t=1}^{n} p_t g(x_t, \tau_i; \theta) = 0 \quad \forall i = 1, \cdots, q,$$

$$\sum_{t=1}^{n} p_t \lambda' \left(\frac{\partial g(x_t; \theta)}{\partial \theta} \right) = 0,$$

with

$$p_t = \frac{1}{n} \rho' \left(\lambda' g(x_t; \theta) \right).$$

The following asymptotic properties of GEL are proved by Newey and Smith (2004). The assumptions that are required for consistency of $\hat{\lambda}_{gel}$ and $\hat{\theta}_{gel}$ are the same as for GMM plus some additional ones associated with the Lagrange multipliers. There is an identification assumption for θ_0, some boundness conditions on higher moments of $\|g(x_t; \theta)\|$ and a non-singularity assumption of the covariance matrix Ω. The latter guarantees that the numerical solution is unique and computable at least with probability approaching one. They show that under these assumptions, $\hat{\theta}_{gel} \xrightarrow{P} \theta_0$ and $\hat{\lambda}_{gel} \xrightarrow{P} 0$. Furthermore, under some additional assumptions which allow to apply a central limit theorem, the estimators are asymptotically distributed as

$$\sqrt{n}(\hat{\theta}_{gel} - \theta_0) \xrightarrow{d} N(0, (G'\Omega^{-1}G)^{-1})$$

and

$$\sqrt{n}\hat{\lambda}_{gel} \xrightarrow{d} N \left\{ 0, \Omega^{-1} - \Omega^{-1} G \left[G'\Omega^{-1} G \right]^{-1} G' \Omega^{-1} \right\},$$

where $G = E(\partial g(X; \theta_0)/\partial \theta)$ and Ω is the asymptotic covariance matrix of $n^{-1/2} \sum_t g(x_t; \theta_0)$. Therefore, GEL shares the same asymptotic properties as GMM.

3 CGEL

In order to illustrate how we can extend the previous results to the case in which the moment conditions are defined on a continuum, and how it affects the stability and existence of the solution, we start by assuming that τ_i, for $i = 1, ..., q$, lies in the fixed interval $[a, b]$ and is defined as $\tau_i = a + i(b - a)/q$. The space in which τ_i lies is therefore $\mathcal{T}_q \subseteq \mathbb{Q} \cap [a, b]$. As q goes to infinity, the space converges to $\mathcal{T}_\infty \equiv \mathcal{T} = [a, b]$. This representation makes sense only if τ_i is an argument of the function defining the moment conditions. For example, if we want to estimate the linear model $y_t = W_t \theta + \epsilon_t$, where $W_t = e^{-x_t^2} + u_t$ and $Cov(\epsilon_t, u_t) \neq 0$, as in Carrasco (2011) with $p = 1$, we can base our estimation on the following moment conditions (In the following, these three expressions will be used interchangeably: $g(x_t; \theta)$, $g_t(\theta)$ or g_t, when we refer to the function from $\mathcal{T}$ to $\mathbb{C}$. The form $g(x_t, \tau; \theta)$ or $g_t(\tau; \theta)$ will be used only when we need to specify the moment condition.):

$$E\Big[g_t(\tau_j; \theta)\Big] = E\Big[(y_t - W_t \theta) e^{(i\tau_j x_t)}\Big] = 0, \text{ for } j = 1, ..., q \text{ and } \tau_j \in \mathcal{T}_q,$$

where the points τ_j are chosen arbitrarily unless some selection methods are used (see Carrasco (2011)). In the simulation below, we estimate the parameters of a stable distribution using the marginal characteristic function for which the same kind of discretization can be applied.

The objective is to define CGEL estimators as the solution to the GEL optimization problem when q goes to infinity. Therefore, we assume that the function $g_t(\tau_i;\theta)$ belongs to an Hilbert space $\mathcal{H}_q$ with inner products defined as $< g,f >_q = \sum_{i=1}^q g(\tau_i)f(\tau_i)\pi(\tau_i)\Delta\tau_i$, where $\pi(\tau)$ is an integrating density as the one introduced by Carrasco et al. (2007a). For GEL, the integrating density is the one from the uniform distribution and $\Delta\tau_i = \Delta\tau_{i-1}$, so that $< g,f >_q$ is the Euclidean inner product. If all $f()$ and $g()$ in $\mathcal{H}_q$ are square-integrable, then $\mathcal{H}_q$ converges to the Hilbert space $\mathcal{H}$ of square-integrable functions on $[a,b]$ with inner product $< f,g >= \int_a^b f(\tau)g(\tau)\pi(\tau)d\tau$. This structure[1] implies that the estimators of GEL is defined by the primal problem:

$$\{\hat{\theta}_q, \hat{\lambda}_q, \hat{\mu}_q, \hat{p}_{qt}\} = \arg\min_{\theta,\mu,p_t,\lambda} \mathcal{L} = \sum_{t=1}^n h_n(p_t) + \left\langle \lambda, \sum_{t=1}^n p_t g_t(\theta) \right\rangle_q + \mu\left(\sum_{t=1}^n p_t - 1\right), \tag{6}$$

where the subscript q means that the estimates are based on q moment conditions. This problem converges to the following primal problem of CGEL when q goes to infinity:

$$\{\hat{\theta}, \hat{\lambda}, \hat{\mu}, \hat{p}_t\} = \arg\min_{\theta,\mu,p_t,\lambda} \mathcal{L} = \sum_{t=1}^n h_n(p_t) + \sum_{t=1}^n p_t \int_a^b \lambda(\tau) g_t(\tau;\theta)\pi(\tau)d\tau + \mu\left(\sum_{t=1}^n p_t - 1\right), \tag{7}$$

In the same way, the dual of GEL and CGEL are respectively:

$$\hat{\theta}_q = \arg\min_{\theta\in\Theta}\left[\max_{\lambda\in\Lambda_{q,n}} P_q(\lambda,\theta) = \frac{1}{n}\sum_{t=1}^n \rho\left(< \lambda, g_t(\theta) >_q\right)\right] \tag{8}$$

and

$$\hat{\theta} = \arg\min_{\theta\in\Theta}\left[\max_{\lambda\in\Lambda_n} P(\lambda,\theta) = \frac{1}{n}\sum_{t=1}^n \rho\left(\int_a^b \lambda(\tau) g_t(\tau;\theta)\pi(\tau)d\tau\right)\right], \tag{9}$$

where $\Lambda_{q,n} = \{\lambda :< \lambda, g(x_t,\theta) >_q \in \mathcal{D}\ \forall\ x_t\}$ and $\Lambda_n = \{\lambda : \int_a^b \lambda(\tau)g(x_t,\tau,\theta)\pi(\tau)d\tau \in \mathcal{D}\ \forall\ x_t\}$. Notice that the continuous updated GMM for a continuum (CCUE) is not a special case of CGEL. When $\rho(v)$ is quadratic, we will refer to the CEEL of Antoine et al. (2007). We only have asymptotic equivalence between CCUE and CEEL as opposed to the case in which the number of conditions is finite (see Appendix D.1).

It follows that we can obtain the solution of GEL by solving the following first order conditions:

$$\sum_{t=1}^n \frac{1}{n}\rho'\left(< \lambda, g_t(\theta) >_q\right) g_t(\tau_i;\theta) = 0 \quad \forall i = 1,\cdots,q, \tag{10}$$

[1] Notice that we focus on a continuum of conditions. However, the case of a countably infinite set of conditions is implicit in this setup if we properly define $\pi(\tau)$. A typical element of the space would be $f(\tau_i)$ for $i \in \mathbb{N}$ and we would get $< f,g >= \sum_i f(\tau_i)g(\tau_i)$. Some times, in the case of a finite number of conditions, GEL fails to produce results due to poorly conditioned first order conditions. In such cases, we could use the CGEL setup to get a more stable solution. We would simply need to set $\pi(\tau_i) = 0$ for $i > q$.

$$\sum_{t=1}^{n}\frac{1}{n}\rho'\left(<\lambda, g_t(\theta)>_q\right)\left\langle\lambda,\left(\frac{\partial g_t(\theta)}{\partial\theta}\right)\right\rangle_q=0, \tag{11}$$

For a given λ, solving the system of p equations (11) is not an issue, even if q goes to infinity, as long as the system is not singular. The problem with GEL arises when we try solving conditions (10) for a given θ. As q increases for a given n, the system becomes more and more poorly conditioned. Indeed, based on the Taylor expansion, we can obtain the solution by using this iterative procedure:

$$\lambda_l=\lambda_{l-1}-\left\langle\left[\frac{1}{n}\sum_{t=1}^{n}\rho''\left(\langle\lambda_{l-1},g_t\rangle_q\right)g_tg_t'\right]^{-1},\left[\frac{1}{n}\sum_{t=1}^{n}\rho'\left(\langle\lambda_{l-1},g_t\rangle_q\right)g_t\right]\right\rangle_q,$$

starting with $\lambda_0=0$. The second term of the right hand side of this procedure is the solution to a system of q linear equations. As q increases, the system becomes singular. As a result, $\lambda(\theta)$ becomes not computable. It is like trying to estimate a model using too many instruments. Therefore, the limiting case of equation (10), which implies the following continuum of conditions

$$\sum_{t=1}^{n}\frac{1}{n}\rho'\left(\int_a^b\lambda(\tau)g_t(\tau;\theta)\pi(\tau)d\tau\right)g_t(\tau,\theta)=0\quad\forall\tau\in[a,b], \tag{12}$$

is ill-posed in the sense that we cannot find a unique solution without imposing a penalty on its instability (Appendix D.2 shows the ill-posedness of equation (12) for the CEEL case even if the right-hand side is not random as required for linear ill-posed problems). The problem arises whether we are dealing with a continuum of conditions, an infinite number of countable conditions or simply when there are a finite but large number of conditions. The last two cases constitute special cases of CGEL simply by selecting the proper integrating density as suggested by Carrasco (2011) for CGMM.

The ill-posedness aspect of GEL in such cases was implicit in the empirical likelihood version of Kitamura et al. (2004) and Donald et al. (2003) since they both require a smoothing parameter. In the first paper, they use a bandwidth parameter while in the second they restrict the number of instruments which also constitute a way of smoothing the problem. Carrasco (2011) deals with the problem by using CGMM, which imposes a Tikhonov's type of penalization in order to make the system solvable (see appendix A.1 for an overview of the CGMM method). It can be seen as a method which automatically selects the most influential moment conditions among the whole set, much like a principal component procedure. A similar approach can be used to solve the ill-posedness of CGEL. Notice, however, that CGMM requires a penalization in order to define its objective function, while CGEL requires it in order to solve it. It is like a nonlinear ridge regression in the sense that the problem, which consists in minimizing $\sum_t u_t^2$ with $u_t=y_t-x(\beta)$, is well defined, but we cannot obtain a stable solution because the columns of $X(\beta)=dx(\beta)/d\beta$ are nearly collinear. In this case, as suggested by Dagenais (1983), we can apply the ridge regression technique to the iterative procedure and substitutes the poorly conditioned matrix $X(\beta)'X(\beta)$ by $[X(\beta)'X(\beta)+\alpha_nI]$ for some $\alpha_n>0$. Even in Theorem 3.1 of Newey and Smith (2004), the uniqueness and existence of the solution are only satisfied with probability approaching one. It may not be the case in small samples, in which case a penalization as in ridge regressions would be required to obtain a stable solution.

In the rest of the paper, we use the notation from the literature on nonlinear and linear operators as the articles from which the numerical procedures used below come from. This also offers a nice and compact way to present the results, especially when working in function spaces. For example, we can rewrite the problem of ill-posedness using the notation of Seidman and Vogel (1989). Their definition of ill-posedness is much like the one we are facing here. Indeed, we can present the first order condition associated with the Lagrange multiplier as the problem of solving the nonlinear operator equation $L(\lambda) = 0$, where

$$L(\lambda) = E\left[\rho'\left(\int_a^b \lambda(\tau)g_t(\tau;\theta_0)\pi(\tau)d\tau\right)g_t(\theta_0)\right],$$

using the disturbed system $\hat{L}(\lambda) = 0$ defined by equation (12) in which θ_0 is replaced by an estimate and $E()$ by the sample mean. The solution is $\lambda = 0$ and is unique given some identification assumptions. It is however ill-posed in the sense that we cannot compute a stable and unique solution to the disturbed system without smoothing it. It is ill-posed even if the right hand side is not random as required by linear ill-posed problems. In the nonlinear case, the ill-posedness appears in the iterative procedure in which a linear ill-posed problem is solved at each iteration. Therefore, in what follows, we regard CGEL as a nonlinear ill-posed problem in function space, which implies that we need a regularized method for computing the solution.

When it is clear, we will use the linear operator notation instead of explicit integrations or inner products. For example, if we have two square-integrable functions $f(x), g(x)$: $\mathcal{T} \to \mathbb{C}$, we will write $fg = \int_{\mathcal{T}} f(x)g(x)\pi(x)dx$ in which case, f is an operator from $L^2(\pi) \to \mathbb{C}$. If furthermore we have a function $A(x,y)$: $\mathcal{T}^2 \to \mathbb{C}$, we will write $(Af)(x) = \int_{\mathcal{T}} A(x,y)f(y)\pi(y)dy$, in which case A is an operator from $L^2(\pi)$ to $L^2(\pi)$ with kernel $A(x,y)$. Using this notation, we can rewrite equation (9) as $P(\lambda,\theta) \equiv (1/n)\sum_{t=1}^n \rho(\lambda g_t)$, where λ is presented as a linear operator from $L^2(\pi)$ to $\mathbb{C}$. Solving the saddle point problem using this notation gives the following first order conditions [2]:

$$F_{n1}(\lambda) \equiv \frac{1}{n}\sum_{t=1}^n \rho'(\lambda g_t)g_t = 0 \tag{13}$$

and

$$F_{n2}(\theta) \equiv \frac{1}{n}\sum_{t=1}^n \rho'(\lambda g_t)\left[\lambda G_t\right] = 0, \tag{14}$$

where

$$G_t \equiv \frac{\partial g_t(\theta)}{\partial \theta}.$$

where F_{n1} is written as a function only of λ to emphasize the fact that it is the system that produces the solution $\lambda(\theta)$ for a given θ and inversely for F_{n2}. $F_{n2}(\theta)$ is the derivative of

[2] For a good review of optimization in function spaces, see Luenberger (1997)

$P(\lambda,\theta)$ with respect of θ. It is therefore a vector with the same dimension as θ, which is $p\times 1$. However, $F_{n1}(\lambda)$ is the Fréchet derivative of $P(\lambda,\theta)$ with respect to the function λ. It is an operator from $L^2(\pi)\rightarrow\mathbb{C}$. That is, if $f\in L^2(\pi)$, then $F_{n1}(\lambda)f=\int_{\mathcal{T}}F_{n1}(\lambda,\tau)f(\tau)\pi(\tau)d\tau$. Since the Fréchet derivative is a generalization of the conventional derivatives for any vector space, we will say that $F_{n2}()$ is also a Fréchet derivative. It is an operator from $\mathbb{R}^p\rightarrow\mathbb{C}$. As a result, for a $p\times 1$ vector y, $F_{n2}(\theta)y=\sum_{i=1}^p F_{n2}(\theta)_i y_i$. Finally, G_t is a $p\times 1$ vector of square-integrable functions. If $h\in L^2(\pi)$, $G_th=\int_{\mathcal{T}}h(\tau)G_t(\tau)\pi(\tau)d\tau$ while if $h\in\mathbb{R}^p$ then $G_th=\sum_{i=1}^p h_iG_{ti}$ (See appendix A.2 from an overview of Fréchet derivatives.).

If we consider a linear ill-posed problem such as the Fredholm integral equation of the first kind $Kg=y$, we can obtain a stable and unique solution by using a Tikhonov approach which consists in solving the following (See Carrasco et al. (2007b) for details on how to solve linear ill-posed problems):

$$\min_g \|Kg-y\|^2+\alpha_n\|g\|^2,$$

where the second term imposes a penalty on the instability of the solution. The regularization parameter α_n determines the degree of penalty. We need to choose it carefully because if it is too small, the solution is more accurate but less stable and inversely if it is too large. The system $Kg=y$ is then replaced by the first order condition of the minimization problem which is:

$$K(Kg-y)+\alpha_n g=0.$$

The system is now well-posed, given certain regularity conditions, and gives the solution $g_\alpha=(K^2+\alpha_nI)^{-1}Ky$, where I is the identity operator and $(K^2+\alpha_nI)^{-1}K$ is a generalized inverse of K. The ill-posedness is caused by the fact that K is a compact bounded operator and is not invertible.

When we deal with a nonlinear system such as $F(g)=y$, ill-posedness is characterized by the non-invertibility of the Fréchet derivative operator. The Fréchet derivative of $F_{n1}(\lambda)$ is an operator with kernel defined as:

$$DF_{n1}(\lambda,\tau_1,\tau_2)=\frac{1}{n}\sum_{t=1}^n\rho''(\lambda g_t)g_t(\tau_1)g_t(\tau_2).$$

Instead of $F_{n1}(\lambda)=0$, we then need to solve the following minimization problem:

$$\min_\lambda \|F_{n1}(\lambda)\|^2+\alpha_n\|\lambda\|^2.$$

In general, the penalty function can be any non-negative function satisfying certain conditions. For example, the Sobolev norm satisfies the conditions required. However, the choice of the penalty function affects only the speed of convergence of the numerical algorithms used. It does not affect the speed of convergence of the estimator to its true value as n goes to infinity, as long as we assume that the numerical solution as been reached. The choice made here is to simplify the presentation.

The first order condition of the minimization problem is

$$DF_{n1}(\lambda)F_{n1}(\lambda)+\alpha_n\lambda=0, \tag{15}$$

which cannot be solved analytically, as for the linear case, because of the nonlinearity. We present the numerical method that we use for solving this system in section 6. The feasible CGEL is therefore defined as the vector $\hat{\theta}$ and the function $\hat{\lambda}$ which solve equations (14) and (15).

We need some assumptions for deriving the asymptotic properties of CGEL. The first set is similar to Assumption A.2 of Carrasco et al. (2007a) but for i.i.d observations.

Assumption 1. *a) The observations $\{x_1, x_2, \cdots, x_n\}$ are i.i.d, b) $L^2(\pi)$ is the Hilbert space of square integrable complex functions in which the inner product $< f, g >$ is defined as $\int f(\tau)g(\tau)\pi(\tau)d\tau$, where $\pi(\tau)$ is a density function which is absolutely continuous with respect to the Lebesgue measure, c) $g(x_t, \tau; \theta) \in L^2(\pi)$, $\forall x_t$ and θ, and d) $g(x_t, \tau, \theta)$ is continuously differentiable with respect to θ for all τ and x_t.*

The second set is similar to Assumption 1 of Newey and Smith (2004). However, we will need to be more specific about ν for consistency. In fact, it will depend on the speed at which α_n goes to zero. What we need is the existence of higher moments when the regularization parameter goes to zero faster. The exact condition is given in Theorem 1.

Assumption 2. *a) $\theta_0 \in \Theta$ is the unique solution to $E^{P_0} g(X;\theta) = 0$, where Θ is a compact subset of $\mathbb{R}^p$, and b) $E^{P_0}[\sup_\theta \|g(X;\theta)\|^\nu] < \infty$ for some $\nu > 2$*

The space in which τ belongs is defined by $\mathcal{T}$ instead of $[a, b]$. For example, if the moment conditions are based on the characteristic function as in Carrasco et al. (2007a), $\mathcal{T}$ is either $\mathbb{R}^2$ or $\mathbb{R}$. It is $[0, \pi]^s$, for some integer s, if the conditions are based on a spectral density as in Berkowitz (2001).

Assumption 1 and 2 imply that:

$$\sqrt{n}\sum_{t=1}^{n} g(x_t, \theta_0) \equiv n^{1/2}\bar{g}(\theta_0) \xrightarrow{L} N(0, K),$$

where K is a covariance operator with the following kernel[3]:

$$k(\tau_1, \tau_2) = E^{P_0}\left[g(X, \tau_1; \theta_0)g(X, \tau_2; \theta_0)\right].$$

The following assumptions replace the full rank properties of Ω that is imposed by Newey and Smith (2004). It implies that the solution of $Kf = g$ exists and is unique as long as $g \in R(K)$, where $R(K)$ is the range of K. It also implies that K can be expressed as the limit of a sequence of linear operator K_n, which is important when K needs to be estimated.

Assumption 3. *a) K is a Hilbert-Schmidt operator, which implies that it is bounded and compact. b) K has only strictly positive eigenvalues. This assumption implies that the null space of K, N(K), is $\{0\}$. c) The skewness operator S with kernel*

$$s(\tau_1, \tau_2, \tau_3) = E[g_t(\tau_1; \theta_0)g_t(\tau_2; \theta_0)g_t(\tau_3; \theta_0)]$$

[3] For a good review of linear operators such as covariance operators applied to econometrics, see Carrasco et al. (2007b)

is bounded and compact.

The following conditions on $G_t \equiv \partial g_t/\partial\theta$ are also required for asymptotic normality and the boundness of $\|E^{P_0}[g_t]\|_3$ guarantees that the remainder term of the Taylor expansion of the first order condition vanishes as n goes to infinity.

Assumption 4. *a) $rank(G_t) = p \ \forall t$, b) $E[\sup_\theta \|G_t\|] < \infty$, c) $E(g(\theta)) \in \mathcal{D}(K^{-1})$ for all θ on a neighborhood of θ_0 and d) $E^{P_0}\|g_t(\theta)\|_3 < \infty$ for all θ.*

The last set of assumptions defines the properties of $\rho(v)$ that we need for the asymptotic theory.

Assumption 5. *a) $\rho(v)$ is strictly concave and twice continuously differentiable. b) $\rho''(v)$ is Lipschitz continuous at least in the neighborhood of 0, c) $\rho'''(v)$ is continuous in the neighborhood of 0 and d) $\rho(v)$ is normalized in such way that $\rho'(0) = \rho''(0) = \rho'''(0) = -1$*

These requirements are satisfied by $\rho(v)$ associated with CEL, CET and CEEL. Assumption 4 b) could be replaced by $\rho''(v)$ being everywhere differentiable since it implies Lipschitz continuity. But it is not necessary. This condition is important in order for the regularized Gauss-Newton method presented in the next section to be locally convergent as explained by Blaschke et al. (1997). The proofs of the following theorems can be found in the appendix.

Theorem 1. *If Assumptions 1 to 4 are satisfied, $\alpha_n = O(n^{-\chi})$ with $0 < \chi < 1/2$, and $\nu > 2/(1-2\chi)$ in Assumption 2b), then $\hat{\theta}_n \xrightarrow{p} \theta_0$ and $\hat{\lambda}_n \xrightarrow{p} 0$, with $\|\hat{\lambda}_n\| = O_p(1/(\alpha_n\sqrt{n}))$, where $\hat{\theta}_n$ and $\hat{\lambda}_n$ are the solutions to the equations (14) and (15).*

Theorem 2. *If assumptions 1, to 5 are satisfied, then:*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{L} N(0, [GK^{-1}G]^{-1})$$

and

$$\sqrt{n}\hat{\lambda} \xrightarrow{L} N\left(0, \left[K^{-1} - K^{-1}G(GK^{-1}G)^{-1}GK^{-1}\right]\right)$$

as n goes to infinity, α_n goes to zero and $n\alpha_n^4$ goes to infinity.

In both theorems, α_n needs to converge to 0 not too quickly because it is necessary for the system to stay stable as n increases. As a result, CGEL shares the same asymptotic properties as CGMM.

We conclude this section by defining the exponentially tilted empirical likelihood method of Schennach (2007) for a continuum (CETEL). It is the vector $\hat{\theta}_{cetel}$ and the function $\hat{\lambda}_{cetel}$ which solve the following conditions:

$$\frac{1}{n}\sum_{t=1}^{n}\rho'_{EL}(\lambda g_t)\left[\lambda G_t\right]=0 \tag{16}$$

and

$$DF_{nET}(\lambda)F_{nET}(\lambda)+\alpha_n\lambda=0, \tag{17}$$

where $\rho_{EL}(v)=\log(1-v)$ and F_{nET} is F_{n1} with $\rho(v)=-e^v$. Since the proofs can easily be derived from the ones from theorems 1 and 2, the asymptotic results are expressed in the following corollary.

Corollary 1. *If the assumptions of the theorems 1 and 2 are satisfied, the CETEL estimator shares the same asymptotic properties as CGEL.*

The idea is that the first order asymptotic properties of CGEL depend only on the behavior of $\rho'(v)$ and $\rho''(v)$ around zero. Since they are all the same, it does not matter which $\rho(v)$ we uses for λ and which one for θ.

4 Estimation procedures

In this section, we present two different estimation procedures which compete in terms of computation time and we express them in matrix form as done by Carrasco et al. (2007a) for CGMM. The first is based on the first order Taylor approximation of the solution $\lambda(\theta)$, while the second solves equation (15) using an iterative procedure. For the GEL case, Guggenberger and Hahn (2005) offer an argument for using what they call the two step empirical likelihood estimator, which is nothing more than the solution obtained from a Newton algorithm after two iterations. They show that increasing the number of iterations does not affect the third order asymptotic bias. Our first procedure approximates the solution $\lambda(\theta)$ which is then used by the numerical optimizer to compute $\hat{\theta}$. Because the second procedure is computationally demanding, it may represent a good alternative. We analyze the properties of both procedures in section 6 through a numerical experiment.

4.1 Taylor approximation and singular value decomposition

The first method follows Carrasco and Florens (2000) who present the singular value decomposition as a way of solving linear ill-posed problems (see also Groetsch (1993)). The ill-posedness arises in the first order Taylor approximation of the solution $\lambda(\theta)$ of equation (15), which implies (see appendix A.2):

$$\hat{K}\hat{\lambda}=-\bar{g}(\theta)+o_p(1),$$

where $\hat{K}$ is the estimated covariance operator of g_t with kernel

$$k_n(\tau_1, \tau_2) = \frac{1}{n}\sum_{t=1}^{n} g(x_t, \tau_1; \theta) g(x_t, \tau_2; \theta).$$

Notice that this approximation is the exact solution of CEEL because in this case, $\rho(v)$ is quadratic and then $F_{n1}(\lambda)$ is linear.

The covariance operator K, is a self-adjoint operator with infinite dimensional range $R(K)$. If we want the solution to $Kx = y$, for $x, y \in L^2(\pi)$, we can use the singular system (ν_i, μ_i) of K, where ν_i is an orthonormal eigenfunction and μ_i the associated singular value. Because the dimension of $R(K)$ is infinite, there are infinitely many singular values. Furthermore, these eigenfunctions are complete in $R(K^2) = N(K)^{\perp}$, where $N(K)$ is the null space of K. It implies that for any $f \in R(K)$:

$$f = \sum_{i=1}^{\infty} < f, \nu_i > \nu_i.$$

We can easily see that any solution $\tilde{x}$ of $Kx = y$ has the following form:

$$\tilde{x} = \sum_{i=1}^{\infty} \frac{1}{\mu_i} < y, \nu_i > \nu_i + \varphi,$$

where $\varphi \in N(K)$. Since $N(K) = \{0\}$, if $y \in R(K)$, the unique solution is:

$$\tilde{x} = \sum_{i=1}^{\infty} \frac{1}{\mu_i} < y, \nu_i > \nu_i.$$

We can obtain a stable solution from the following regularized system:

$$(K^2 + \alpha_n I)x_\alpha = Ky,$$

which implies the following solution:

$$x_\alpha = \sum_{i=1}^{\infty} \left(\frac{\mu_i}{\mu_i^2 + \alpha_n}\right) < y, \nu_i > \nu_i.$$

Therefore, the solution requires an infinite number of eigenfunctions. However, when K is unknown and is replaced by $\hat{K}$, the solution is much simpler. As Carrasco and Florens (2000) show, the dimension of $R(\hat{K})$ is finite:

$$\begin{aligned}
(\hat{K}f)(\tau_2) &= \int_{\mathcal{T}} k_n(\tau_1, \tau_2) f(\tau_1)\pi(\tau_1)d\tau_1 \\
&= \int_{\mathcal{T}} \frac{1}{n}\sum_{t=1}^{n} g(x_t, \tau_1; \theta) g(x_t, \tau_2; \theta) f(\tau_1)\pi(\tau_1)d\tau_1 \\
&= \sum_{t=1}^{n} g(x_t, \tau_2; \theta) \left(\int_{\mathcal{T}} \frac{1}{n} g(x_t, \tau_1; \theta) f(\tau_1)\pi(\tau_1)d\tau_1\right) \\
&\equiv \sum_{t=1}^{n} \delta_t g(x_t, \tau_2; \theta).
\end{aligned}$$

Therefore, $R(\hat{K})$ is spanned by $\{g(x_1;\theta),\cdots,g(x_n;\theta)\}$. It follows that the singular system of $\hat{K}$ is composed of n eigenfunctions $\nu_i^{(n)}$ and n singular values $\mu_i^{(n)}$. We can extend the previous result to our case and show that the regularized solution to $\hat{K}\tilde{\lambda} = -\bar{g}(\theta)$ is

$$\tilde{\lambda} = -\sum_{i=1}^{n}\left(\frac{\mu_i^{(n)}}{\mu_i^{(n)^2}+\alpha_n}\right) < \bar{g}(\theta), \nu_i^{(n)} > \nu_i^{(n)},$$

where the tilde stands for approximated solution. Because $\nu_i^{(n)} \in R(\hat{K})$, we can write $\nu_i^{(n)} = 1/n\sum_j \beta_{ij} g(x_j,\theta)$. Carrasco and Florens (2000) show that the vectors β_i, for $i = 1, ...n$, are the eigenvectors of an $n \times n$ matrix C with typical element

$$c_{ij} = \frac{1}{n}\int_{\mathcal{T}} g(x_i,\tau;\theta)g(x_j,\tau;\theta)\pi(\tau)d\tau$$

and that its eigenvalues are in fact the $\mu_i^{(n)}$ we need. We can therefore obtain the estimator using the following procedure:

1. We construct the $n \times n$ matrix C

2. We compute the eigenvectors β_i and eigenvalues $\mu_i^{(n)}$ for $i = 1, ..., n$

3. We compute the eigenfunctions of $\hat{K}$ as follows:

$$\nu_i^{(n)} = \frac{1}{n}\sum_{j=1}^{n}\beta_{ji}g(\theta,x_j) \qquad i = 1,...,n$$

4. We compute $\tilde{\lambda}$:

$$\tilde{\lambda}(\theta) = -\sum_{i=1}^{n}\left(\frac{\mu_i^{(n)}}{\mu_i^{(n)^2}+\alpha_n}\right) < \bar{g}(\theta), \nu_i^{(n)} > \nu_i^{(n)}$$

5. We estimate θ_0 by solving the following problem:

$$\tilde{\theta} = arg\min_{\theta}\frac{1}{n}\sum_{t=1}^{n}\rho\Big(\tilde{\lambda}(\theta)g(x_t,\theta)\Big). \tag{18}$$

Because the solution to CGEL includes also an estimate of the probability distribution p_t for $t = 1,\cdots,n$ with $\sum_t p_t = 1$, which depends on λ, and that we did not obtain the exact solution to equation (15), we may, if we intend for example to use the implied probabilities to obtain efficient estimates of higher moments of $g_t(\theta)$, have to normalize $p_t(\tilde{\lambda})$ as follows

$$\tilde{p}_t = \frac{p_t(\tilde{\lambda})}{\sum_{t=1}^{n} p_t(\tilde{\lambda})}, \tag{19}$$

where

$$p_t(\tilde{\lambda}) = \frac{1}{n}\rho'\Big(\tilde{\lambda}g(x_t;\tilde{\theta})\Big).$$

For the case in which $\rho(v)$ is quadratic, which does not guarantee the non-negativity of $p_t(\tilde{\lambda})$, the latter can be transformed according to Antoine et al. (2007).

In order to apply this method, it is convenient to rewrite the objective function in matrix notation as in Carrasco et al. (2007a). Let us define the $n \times m$ matrix β which contains the eigenvectors of C associated with its m eigenvalues different from 0, and the $m \times m$ diagonal matrix D with typical element D_{jj}:

$$D_{jj} = \frac{\mu_i^{(n)}}{\mu_i^{(n)2} + \alpha_n}.$$

The following optimization problem is equivalent to the one given by equation (18):

$$\tilde{\theta} = arg \min_{\theta} \frac{1}{n} \sum_{t=1}^{n} \rho\left(-\frac{1}{n} \iota' C[\beta D \beta'] C_{\bullet t} \right), \tag{20}$$

where ι is an $n \times 1$ vector of ones and $C_{\bullet t}$ is the t^{th} column of C (see Appendix C.1 for the proof).

In practice, we need to select a tolerance level in order to determine whether the eigenvalues are considered to be zero or not. Indeed, none of them will be exactly equal to zero. However, the presence of α_n in the denominator of D_{jj} makes it possible to choose $m = n$.

4.2 Solving a nonlinear operator equation

When we want the solution to a nonlinear problem such as $f(x) = 0$, we usually construct an iterative procedure of the form

$$x_i = g(x_{i-1}),$$

which converges to the fix point $g(x) = x$, where x is the solution to the initial problem. The simplest method sets $g(x) = x + \omega f(x)$. If the algorithm converges, then we have $f(x) = 0$ as required. However, this method, if it converges, is slow if we do not select a proper ω. The Newton method sets $\omega = -[f'(x)]^{-1}$ so that the algorithm becomes:

$$x_i = x_{i-1} - f'(x_{i-1})^{-1} f(x_{i-1}).$$

In order for this method to work, the inverse of the first derivative needs to be bounded. When $f'(x_{i-1})^{-1}$ is not bounded, it has to be replaced by a generalized inverse. This is similar to the problem we are facing in this section but with the exception that the solution x is a function from $L^2(\pi)$.

In the case of CGEL, we need to solve equation(15) which we rewrite as follows ($F_{n1}()$ has been replaced by $F()$ for simplicity):

$$F(\lambda) \equiv \frac{1}{n} \sum_{t=1}^{n} \rho'(\lambda g_t) g(x_t; \theta) = 0. \tag{21}$$

So we need the solution to the general nonlinear operator equation $F(\lambda) = 0$, where F is a nonlinear operator from $L^2(\pi)$ to $L^2(\pi)$. As discussed in Section 3, this problem needs to be regularized in order to compute a stable solution. The second estimation procedure is therefore an iterative algorithm which converges to the solution of:

$$\min_{\lambda} \|F(\lambda)\|^2 + \alpha_n \|\lambda - \lambda_0\|^2,$$

so that $F(\lambda)$ is close to 0 and the solution $\hat{\lambda}$ sufficiently smooth. The value of α_n determines how important we consider the smoothness of the solution. If it is set too high, the solution $\hat{\lambda}$ would most likely be a constant (i.e. $\hat{\lambda}(\tau_1) = \hat{\lambda}(\tau_2)\ \forall \tau_1, \tau_2$) with $F(\hat{\lambda})$ not too close to zero, since the first term would become negligible. Conversely, a small α_n would create an unstable solution for which $F(\hat{\lambda})$ is almost zero. It is the same tradeoff that we face when solving linear ill-posed problems.

There are many algorithms that reflect this tradeoff. Ramm (2004a) and Ramm (2004b) present the continuous version of such methods and give the conditions under which they converge to the solution. The discrete algorithm presented by Airayetpan and Ramm (2000) is a regularized Newton method. It is a Newton method applied to a transformed equation. Indeed, the Newton method solves $F(\lambda) = 0$ with the algorithm $\lambda_i = \lambda_{i-1} - DF(\lambda_{i-1})^{-1} F(\lambda_{i-1})$ while the regularized Newton method solves $F(\lambda) + \alpha_n \lambda = 0$ which implies the following algorithm:

$$\lambda_i = \lambda_{i-1} - \omega_i \left[DF(\lambda_{i-1}) + \alpha_n I \right]^{-1} \left(F(\lambda_{i-1}) + \alpha_n \lambda_{i-1} \right)$$

where ω_i is a sequence that we need to choose to control the speed of convergence. Another method which uses a regularized inverse which is closer to the one used for the linear case has been analyzed by Jin (2000). It is a regularized Gauss-Newton method which is defined as follows:

$$\lambda_i = \lambda_{i-1} - \left[\alpha_n I + DF(\lambda_{i-1})^2 \right]^{-1} \left\{ DF(\lambda_{i-1}) F(\lambda_{i-1}) + \alpha_n \lambda_{i-1} \right\},$$

where the initial value λ_0 has been set equal to its asymptotic value, 0. It is the usual starting value for λ when the parameters are estimated by GEL (see for example Guggenberger (2008)). If the algorithm converges, the condition of equation (15) is satisfied. Blaschke et al. (1997) show that the conditions that we impose are sufficient for the convergence of the algorithm.

In order to apply this algorithm, we will present it in matrix form. Because λ enters equation (15) only through $\lambda g_t(\theta) = \int_{\mathcal{T}} \lambda(\tau) g_t(\tau;\theta) \pi(\tau) d\tau$, we only need to solve for λg_t. Therefore we can obtain the result from the following iterative procedure:

$$g_t \lambda_i = g_t \lambda_{i-1} - g_t \left[\alpha_n I + DF(\lambda_{i-1})^2 \right]^{-1} \left\{ DF(\lambda_{i-1}) F(\lambda_{i-1}) + \alpha_n \lambda_{i-1} \right\}. \qquad (22)$$

Let us define the $n \times n$ diagonal matrix V as:

$$V_{tt} = \rho''(\lambda g_t),$$

the $n \times 1$ vector P as:

$$P_t = \rho'(\lambda g_t),$$

and the $n \times n$ matrix C as usual. The following theorem is demonstrated in Appendix C.2.

Theorem 3. *If the conditions of theorem 1 and 2 are satisfied, than CGEL, which is defined by the conditions of equations (14) and (15), is equivalent to the following procedure: We first iterate the following until convergence:*

$$[\lambda_i g] = \left\{[CV]^2 + \alpha_n I\right\}^{-1} \left\{[CV]^2[\lambda_{i-1}g] - [CV][CP]\right\}, \tag{23}$$

with the initial value:

$$\lambda_0 g = -\left\{C^2 + \alpha_n I\right\}^{-1} C^2 \iota,$$

where ι is an $n \times 1$ vector of ones. We then solve the following minimization problem:

$$\frac{1}{n}\sum_{t=1}^{n} \rho(\widehat{\lambda g_t}(\theta)),$$

where $\widehat{\lambda g_t}(\theta)$ is the value to which has converged the algorithm (23).

Notice that we don't need this iterative procedure for CEEL because in that case, $\rho(v) = \rho_0 - v - v^2/2$, which implies that the first order condition for λ becomes:

$$\frac{1}{n}\sum_{t=1}^{n}(-1 - \lambda g_t)g_t = 0,$$

which implies that $\hat{\lambda}$ is the solution to the linear operator equation $\hat{K}\lambda = -\bar{g}(\theta)$. We can then derive the analytical solution of $\hat{\lambda}$ by using the results from the previous section and we obtain:

$$\hat{\lambda}_{CEEL}\bar{g}(\theta) = -\frac{1}{n}\iota' C[\beta D \beta']C_{\bullet t}. \tag{24}$$

That expression can easily be used in $\rho(v)$ to estimate θ or in $\rho'(v)$ to compute the implied probabilities. It is not like CCUE as shown in Appendix D.1, but it is equivalent in large samples. In fact, it should be easier numerically to implement than CCUE as it is argued by Antoine et al. (2007) for the discrete case.

5 Over-identification tests for C-GEL

GEL offers three ways of testing the validity of the moment conditions $E^{P_0}(g_t(\theta_0)) = 0$. Smith (2004) summarizes them and shows that they are first order equivalent and asymptotically chi-square with $(n-q)$ degrees of freedom, where q is the number of moment conditions. The first is the J-test developed by Hansen (1982) which is based on the GMM criterion:

$$J_{gmm} = n\bar{g}(\hat{\theta}_{gmm})'\hat{K}^{-1}\bar{g}(\hat{\theta}_{gmm}) \equiv \|\hat{K}^{-1/2}\sqrt{n}\bar{g}(\hat{\theta}_{gmm})\|^2.$$

In the context of CGEL, two problems arise from this test. First, we need to replace $\hat{K}^{-1}$ by the generalized inverse $(\hat{K}^{\alpha_n})^{-1}$, and second, the test diverges since the number of moment conditions is infinite. Carrasco and Florens (2000) offer a normalized version of this test which is asymptotically $N(0,1)$. We can apply the same normalization for CGEL since it is asymptotically equivalent to CGMM. However, the tests will differ in finite sample since CGEL evaluates $\hat{K}^{\alpha_n}$ at $\hat{\theta}$ while CGMM uses a first step estimate.

The test is based on the singular value representation of the CGMM criterion described in the previous section:

$$\|(\hat{K}^{\alpha_n})^{-1/2}\sqrt{n}\bar{g}(\hat{\theta})\|^2 = \sum_{i=1}^{n}\left(\frac{\mu_i^{(n)}}{\mu_i^{(n)^2}+\alpha_n}\right) < \bar{g}(\hat{\theta}), \phi_i^{(n)} >^2,$$

where $\phi_i^{(n)}$ is the orthonormalized eigenfunction $\nu_i^{(n)}/\|\nu_i^{(n)}\|$ with $\|\nu_i^{(n)}\|^2 = \mu_i^{(n)}/n$. Let us define the following variables:

$$p_n = \sum_{i=1}^{n}\frac{\mu_i^{(n)^2}}{(\mu_i^{(n)^2}+\alpha_n)}$$

$$q_n = 2\sum_{i=1}^{n}\frac{\mu_i^{(n)^4}}{(\mu_i^{(n)^2}+\alpha_n)^2}.$$

Then the first test is defined as:

$$\tilde{J} = \frac{\|(\hat{K}^{\alpha_n})^{-1/2}\sqrt{n}\bar{g}(\hat{\theta})\|^2 - p_n}{\sqrt{q_n}} \Longrightarrow N(0,1),$$

under the null that the over-identifying moment conditions are satisfied. The proof is given by Carrasco and Florens (2000).

The second is the Lagrange multiplier test (LM). In the dual problem, λ is the Lagrange multiplier associated with the sample moment conditions $\sum_t p_t g_t(\theta)=0$. It should therefore be zero if the constraint is not binding. For GEL, the test is defined as follows:

$$LM = n\hat{\lambda}_{gel}'\hat{K}\hat{\lambda}_{gel}.$$

For CGEL, the same normalization is required. The second test is therefore[4]:

$$\widetilde{LM} = \frac{\|\hat{K}^{1/2}\sqrt{n}\hat{\lambda}\|^2 - p_n}{\sqrt{q_n}}.$$

The third test is based on the GEL criterion function $P_q(\lambda,\theta)$ (see equation (8)). We can use it for constructing a likelihood ratio test (LR) for the null hypothesis that $\lambda=0$. It is defined as follows for GEL:

$$LR = 2n(P_q(\hat{\lambda}_{gel},\hat{\theta}_{gel}) - \rho(0)),$$

which implies the following for CGEL:

$$\widetilde{LR} = \frac{2n(P(\hat{\lambda},\hat{\theta}) - \rho(0)) - p_n}{\sqrt{q_n}}.$$

[4] To make sure that the reader is not confused with the notation, notice that $\hat{K}$ is not the same in the LM and $\widetilde{LM}$ tests. For the former, $\hat{K}\hat{\lambda}$ is $\sum_i \hat{K}_{\bullet i}\hat{\lambda}_i$, while for the latter it is $\int_{\mathcal{T}}\hat{k}(\tau,\tau_1)\hat{\lambda}(\tau_1)\pi(\tau_1)d\tau_1$.

The following theorem shows that the three tests are first order equivalent. Moreover, it gives a way to compute them using the same matrix notations that we use above for the estimators.

Theorem 4. *If Assumptions 1 to 4 are satisfied then $\tilde{J}$, $\widetilde{LM}$ and $\widetilde{LR}$ are first order equivalent and asymptotically distributed as $N(0,1)$. Furthermore they can be computed as follows:*

$$\tilde{J} = \frac{\iota'(\beta D \beta' - D)\iota}{\sqrt{2\iota' D^2 \iota}},$$

$$\widetilde{LM} = \frac{\sum_{t=1}^n (g_t(\hat{\theta})\hat{\lambda})^2 - \iota' D \iota}{\sqrt{2\iota' D^2 \iota}}$$

and

$$\widetilde{LR} = \frac{2\sum_{t=1}^n \rho(g_t(\hat{\theta})\hat{\lambda}) - 2n\rho(0) - \iota' D \iota}{\sqrt{2\iota' D^2 \iota}}.$$

where $(g_t(\hat{\theta})\hat{\lambda})$ comes from equation (23), β is the $n\times n$ matrix containing the n eigenvectors of C, ι is a vector of ones and D is an $n\times n$ diagonal matrix with typical element

$$D_{ii} = \frac{\mu_i^{(n)^2}}{(\mu_i^{(n)^2} + \alpha_n)}.$$

The proof is given in the Appendix B.3. We can prove the asymptotic normality of the three tests by showing the first order equivalence of $\|(\hat{K}^{\alpha_n})^{-1/2}\sqrt{n}\bar{g}(\hat{\theta})\|^2$, $\|\hat{K}^{1/2}\sqrt{n}\hat{\lambda}\|^2$ and $2n(P(\hat{\lambda},\hat{\theta}) - \rho(0))$ since Carrasco and Florens (2000) show the result for $\tilde{J}$.

In small samples, we may want to consider alternative methods since it is unlikely that the exact distribution of the three tests are symmetric. As suggested by Arelanno et al. (2011), since the J-test is approximately a quadratic form in normal variables, the Imhof (1961) approach can be used to compute the p-values. In fact, if we assume that $\sqrt{n}\bar{g}(\hat{\theta})$ is asymptotically $N(0,K)$, then $< \sqrt{n}\bar{g}(\hat{\theta}), \phi_i^{(n)} >$ is asymptotically $N(0,\mu_i)$ which implies that:

$$\|(\hat{K}^{\alpha_n})^{-1/2}\sqrt{n}\bar{g}(\hat{\theta})\|^2 \approx \sum_{i=1}^n \left(\frac{\mu_i^{(n)^2}}{\mu_i^{(n)^2} + \alpha_n}\right)\chi_1^2$$

Imhof (1961) shows how to compute the CDF of the above expression. This is of course not valid as n goes to infinity but it may improve the size in small samples. We investigate it in Section 6.

6 A numerical study

As suggested by Nolan (2005), the family of stable distributions offers a good alternative for modeling heavy-tailed and skewed data such as stock returns. We say that a random variable follows a stable distribution if linear combinations preserve the shape of the distribution up to scale and shift, which determine respectively the variance and the expected value when they are well defined. Therefore, the normal distribution is stable because the sum of two normal random variables is also normally distributed. The Cauchy and Lévy distributions are special cases for which moments are either infinite or undefined. The notation used in this section follows Nolan (2009) who presents in details the properties of stable distributions.[5]

These three special cases are the only stable distributions for which the density has a closed form expression. As a result, the maximum likelihood estimation of the parameters can only be performed through numerical computation of the likelihood function. However, there is an analytical representation of its characteristic function. We can therefore base our estimation on the following continuum of moment conditions:

$$E\left[e^{i\tau x_t} - \Psi(\theta;\tau)\right] = 0 \ \ \forall\ \tau \in \mathbb{R} \tag{25}$$

where i is the imaginary number, $\Psi(\theta;\tau)$ is the characteristic function and $\theta = \{\omega, \beta, \gamma, \delta\}$. The elements of θ are respectively the characteristic exponent[6] and the skewness, the scale and the location parameters. They are restricted to the parameter space $]0,2]\times[-1,1]\times]0,\infty[\times\mathbb{R}$. Garcia et al. (2006) estimate the parameters using indirect inference and perform a numerical study to compare it with some other methods. One of them is CGMM and was suggested by Carrasco and Florens (2002). We therefore use this example to compare the performance of CGEL with CGMM in small samples. We want to compare the mean-bias, median-bias and root mean squared errors (RMSE) of the estimators for different choices of α_n.

We need to be careful when working with stable distributions because there are more than one parametrization which implies different analytical forms for the characteristic function. In order to avoid confusions Nolan (2009) defines the distribution by $S(\omega, \beta, \gamma, \delta, pm)$, where $pm = 0, 1, 2$ or 3 defines the type of parametrization used. In this experiment, we follow Garcia et al. (2006) and Carrasco and Florens (2002) by choosing $pm = 1$. Notice that when the moments exist and are finite, γ and δ are note necessarily the variance and the mean of the distribution. For example, we can represent a $N(\mu, \sigma^2)$ by $S(2, 0, \sigma/\sqrt{2}, \mu, 1)$. This parametrization implies the following characteristic function:

$$\Psi(\theta;\tau) = \left\{ \begin{array}{lll} \exp\left(-\gamma^\omega|\tau|^\omega[1 - i\beta(\tan\frac{\pi\omega}{2})(sign(\tau))] + i\delta\tau\right) & \text{for} & \omega \neq 1 \\ \exp\left(-\gamma|\tau|[1 + i\beta\frac{2}{\pi}(sign(\tau))\log|\tau|] + i\delta\tau\right) & \text{for} & \omega = 1 \end{array} \right. ,$$

where $sign(\tau) = 1$ if $\tau > 0$, -1 if $\tau < 0$ and 0 otherwise. Notice that β can be poorly identified when ω is close to 2 as the term $\tan(\pi\omega/2)$ becomes close to zero. That should reflect on the properties of the estimators of β.

We compute $\int_{\mathbb{R}} f(\tau)g(\tau)\pi(\tau)d\tau$ with $\pi(\tau)$ defined as the density of a standardized normal distribution as for Carrasco and Florens (2002). However, no $\tan()$ transformation is done as they do in order to transform the integrals over $\mathbb{R}$ into integrals over a finite interval. The integrals are computed directly over the interval $[-2, 2]$. Because of the integrating density, it makes almost no difference to integrate over a wider interval. Furthermore, it allows a better approximation of the integrals without being too much computationally demanding.

[5] See also his web site on stable distributions http://academic2.american.edu/ jpnolan/stable/stable.html

[6] In general, the characteristic exponent is defined by α_n instead of ω. But in this paper, α_n represents the regularization parameter.

The regularized iterative procedure which computes the solution of $\lambda(\theta)$ is called by the numerical optimizer each time θ is updated. For some values of θ, it happens that α_n is too small to make the system well-posed. In such case, we have to increase it temporarily. More precisely, if the inverse of the condition number of $([CV]^2 + \alpha_n I)$ is less than 9.9×10^{-15}, α_n is raised by 50%. Once the procedure converges, α_n returns to its initial value for the next value of θ. The algorithm is much more stable this way.

The simulations are carried out using **R** and the random variables are generated by the **rstable** generator from the **fBasics** package of Wuertz and Rmetrics (2010). The starting values are obtained by CGMM using the identity matrix starting at the initial guess $\{\omega_0, \beta_0, \gamma_0, \delta_0\} = \{1.1, 0.1, 0.1, 0\}$. The true values are not used so that we can analyze how the methods behave when little is known about the distribution[7]. Finally, instead of reparametrizing ω and β as in Garcia et al. (2006) to restrict their parameter spaces, we use the optimizer **nlminb** which allows inequality constraints. Another possibility is to rescale the parameters ω and β (done inside the optimizer) to make them move slower to the solution. We have found that it helps to prevent them from going outside the parameter space.

We perform the Monte Carlo experiment by generating 1000 samples of size equal to 100[8]. The true distribution is $S(1.7, 0.5, 0.5, 0, 1)$, which is one of the models studied by Garcia et al. (2006). The parameters are estimated using CGMM and CGEL, using both the iterative and singular value decomposition methods, with $\alpha_n = \{0.0001, 0.001, 0.005, 0.01, 0.05, 0.1\}$. Notice however that CEEL is only computed once since the iterative procedure for this method is identical to the one step approximation. We also compute the three tests of over-identifying restrictions for CGEL and compare their empirical sizes with the J-test of CGMM.

The properties of the estimators are presented in tables (1) to (4), in Appendix E [9]. We can see that the relative performance of CGMM and CGEL depends on which parameter is estimated and on the value of α_n. As explained by Carrasco (2011), we can interpret the value of α_n as a way of selecting the number of moment conditions. As the parameter goes down, more information contained in the continuum of conditions is used. Following the second order asymptotic results of Newey and Smith (2004), that should increase the bias of CGMM. This conclusion is verified except for γ. Its impact on the mean squared errors is more ambiguous. It seems consistent with the numerical experiment of Carrasco and Florens (2002) who find that the average α_n which minimizes the RMSE is around 0.05.

The impact of α_n on the bias of the different CGEL estimators using the iterative procedure is very similar. In general, the bias seems to increase when α_n is smaller except for γ . Also, we can always find a value for which the bias of CGEL is lower than the bias of CGMM. As opposed to the bias, the RMSE is very stable and almost always smaller than the one of CGMM estimator. This is explained by the standard deviations (not reported for

[7] We intentionally start far away from the true values in order to analyze the ability of the algorithm to reach the solution. It is of course not recommended in practice, especially with empirical likelihood which requires in general good starting values to converge.

[8] Results for $n = 200$ and the R codes are available upon request.

[9] Notice that the results cannot be compared directly with the ones obtained by Carrasco and Florens (2002) even though we use the same sample size because the true distribution estimated is not the same and because they fix the parameter δ to zero. For example, the larger RMSE that we obtained is explained by the fact that β is poorly identified by the characteristic function as ω approaches 2.

compactness) which are smaller and mostly not affected by α_n. When it is affected, it tends to be positively related which suggests that CGEL uses the extra information more efficiently than CGMM. If we compare the iterative procedure with either CEEL or the approximation method based on the singular value decomposition (sv-CGEL), the latter is most of the time less biased but at the cost of a slightly higher RMSE, except for β. However, the RMSE of both CEEL and sv-CGEL is most of the time lower than CGMM which makes it a good alternative since it is less computationally demanding than the iterative CGEL.

Most of the results suggest that CGEL may outperform CGMM according to the RMSE and some times to the bias. It is not consistent with numerical studies on GEL like the one by Guggenberger (2008) who finds that GEL is much more volatile than GMM. However, his analysis is based only on linear models estimated by moment conditions constructed from weak instruments. Besides, the difference between GMM and GEL does not necessarily apply to CGEL and CGMM. A more complete numerical experiment should be performed in order to support the conclusions obtained in this section with more confidence. Furthermore, the optimal α_n for CGMM does not seem to be the same for CGEL. It would therefore be an interesting extension to derive a data-driven method based on higher order expansions to select α_n for CGEL as done by Carrasco and Florens (2002). We could then more easily compare the methods using their respective optimal α_n.

We can also compare our estimators with the maximum likelihood estimator (MLE) based on numerical computation of the density function. The MLE is the best method for estimating the parameters of a distribution when the latter is known. We can therefore see whether the continuum spans enough information to reach the efficiency of MLE. The result is presented in Table (5). Notice first that the method is more computationally demanding than our methods because we need to simulate the likelihood each time it is evaluated. It is also very unstable in the sense that the parameters often tend to go outside the admissible parameter space. In order to obtain 1000 results we had to generate about 1800 series[10]. The results show that CGEL performs almost as well as MLE for some choices of α_n. It is even better then MLE in some cases. However, this result depends on the algorithm used to compute the likelihood numerically. We may obtain better results with better algorithms. But since CGEL does no rely on the choice of algorithms to compute the likelihood, it may be more suitable for such estimations.

The size of the three tests and the one for the J-test with p-values based on Imhof (1961) are shown in table (6). We can see that α_n is negatively related to the size of all tests. They are above 50% for $\alpha_n = 0.0001$ and close to zero $\alpha_n = 0.1$. It may reflect the instability of the solution $\hat{\lambda}$, on which are based the tests, when α_n is small. As for the properties of the estimators, the value of α_n is very important. In General, the J-test performs better for all iterative GEL methods if we exclude very small α_n. Since these small values are also associated with large bias for the vector of coefficients, a good selection of α_n should give us reliable tests.

Using the Imhof (1961) approach to compute the p-values of the J-test does not make much difference. Indeed, the rejection rates are almost identical. We could improve the size

[10] We could probably have obtained a smaller rejection rate by playing with the parameters of the optimizer but it would have been too time consuming.

of the tests by using an alternative method based on Anatolyev and Gospodinov (2008) who suggest to modify the distribution of the tests by using a parameter that depends on the ratio number of instruments to sample size. They show that it improves the properties of the tests when the ratio is close to one. It is therefore relevant for CGEL. It could also be improved by using some bootstrap procedures to compute the finite sample critical values.

7 Conclusion

The CGEL method, which we apply using either the regularized Gauss-Newton algorithm or the regularized singular value decomposition of the approximated solution, is shown to be asymptotically equivalent to CGMM. However, we show through a Monte Carlo experiment that the small samples properties of CGEL mostly outperform those of CGMM at least according to the root mean squared errors. The results are not necessarily consistent with what we find in the literature which compares GMM and GEL. In particular, studies often find that EL is less biased but produces higher mean squared errors than GMM. Our experiment show the opposite which suggests that the relative performance of GEL and GMM may not be preserved when the moment conditions are defined on a continuum. As suggested by the standard deviation, CGEL seems more efficient in using the additional information implied by a smaller α_n. It may therefore constitute an excellent alternative to CGMM

However, the properties of the estimators and the three tests seem to depend heavily on the value of the regularization parameter. As a result, future studies should include a data-driven method to select α_n which would require either to derive the higher order properties of the estimators or to develop an adapted bootstrap method such as the one proposed by Carrasco and Kotchoni (2010). The same type of expansion or Monte Carlo method should also be developed to improve the properties of the three tests.

Appendix

A Overview of some concepts

A.1 CGMM

We present here a brief overview of CGMM. It is developed by Carrasco and Florens (2000), summarized by Carrasco et al. (2007b) and Carrasco et al. (2007a) show how it can be implemented by expressing the objective function in matrix form. The estimator is defined as:

$$\hat{\theta}_{cgmm} = \arg\min_{\Theta} \|B_n \bar{g}(\theta)\|,$$

where B_n is a sequence of random operators from $L^2(\pi) \rightarrow L^2(\pi)$ which converges to the linear bounded operator B. It plays the same role as the weighting matrix of GMM. In order to achieve efficiency, the operator B must be defined as the inverse of the square root of the asymptotic covariance operator of $\sqrt{n}\bar{g}(\theta_0)$, K. Because of its properties, the inverse of K is unbounded. As a result, the objective function is ill-posed because it can be written as $< \bar{g}(\theta), K^{-1}\bar{g}(\theta) >$, where the second term of the inner product is the solution to $Kx = \bar{g}(\theta)$.

A stable and unique solution can be computed using the Tikhonov approach for which the inverse of the linear operator is substituted by the regularized inverse

$$(K^{\alpha_n})^{-1} = (\alpha_n I + K^2)^{-1} K.$$

The feasible optimal CGMM estimator, in which K is replaced by the consistent estimate $\hat{K}$, is therefore defined as:

$$\hat{\theta}_{cgmm} = \arg\min_{\Theta} \|(\hat{K}^{\alpha_n})^{-1/2} \bar{g}(\theta)\|.$$

In order for $\hat{\theta}_{cgmm}$ to be consistent, certain conditions are required. One of them imposes a rate of convergence for α_n which must satisfy $n\alpha_n^{3/2} \to \infty$ as α_n goes to zero, which implies that $\alpha_n = O(n^{-2/3+\eta})$ for $0 < \eta < 2/3$. The condition on α_n is required in order for $\|(\hat{K}^{\alpha_n})^{-1/2} f_n - K^{-1/2} f\|$ to be $o_p(1)$ for any f_n converging to f. To prove asymptotic normality, the required rate of convergence of α_n is different. We need $n\alpha_n^3 \to \infty$ as α_n goes to zero because we need $\|(\hat{K}^{\alpha_n})^{-1} f_n - K^{-1} f\|$ to be $o_p(1)$. The latter implies that $\alpha_n = O(n^{-1/3+\eta})$ for $0 < \eta < 1/3$. Given these conditions Carrasco and Florens (2000) show that $\sqrt{n}(\hat{\theta}_{cgmm} - \theta_0)$ is asymptotically distributed as $N(0, [GK^{-1}G]^{-1})$.

In order to compute the two step CGMM estimator, we first solve:

$$\tilde{\theta} = \arg\min_{\Theta} \|\bar{g}(\theta)\|,$$

in which the identity operator has been used instead of $(\hat{K}^{\alpha_n})^{-1}$, and then:

$$\hat{\theta}_{cgmm} = \arg\min_{\Theta} \tilde{v}' \left[I - C(\alpha_n I + C^2)^{-1} C \right] \tilde{v},$$

where C is the same matrix defined in Section 4.1 and $\tilde{v} = \{\tilde{v}_1, ..., \tilde{v}_n\}$ with $\tilde{v}_t = g_t(\tilde{\theta})\bar{g}(\theta)$.

A.2 Fréchet derivative

We summarize in this section the concept of Fréchet derivatives and show some results that are used in the proofs bellow (results from this section can be found in Zeidler (1995)).

Consider two Hilbert spaces X and Y, and the functional F: $X \to Y$. For h and $x \in X$, let define the function $\phi : \mathbb{R} \to Y$ as:

$$\phi(s) = F(x + sh)$$

Assuming that $\phi'(0)$ exists for all h, the Gâteaux derivative of F at $x \in X$, $DF(x)$, is defined by the following:

$$\phi'(0) = DF(x)h$$

Furthermore, $DF(x)$ is the Fréchet derivative of F evaluated at $x \in X$ if and only if

$$\lim_{h \to 0} \frac{\|F(x+h) - F(x) - DF(x)h\|}{\|h\|} = 0. \tag{26}$$

As an example, we can compute the linear approximation of the first order condition (13). First, let us consider equation (13) for any $\lambda \in L^2(\pi)$ with (λg_t) being in the domain of $\rho(v)$ for all t (subscripts have been omitted for clarity):

$$F(\lambda) \equiv \frac{1}{n}\sum_{t=1}^{n} \rho'(\lambda g_t) g_t.$$

We first need to find the Gâteaux derivative and then show that it satisfies condition (26):

$$\phi(s) = \frac{1}{n}\sum_{t=1}^{n} \rho'\Big((\lambda + sh) g_t\Big) g_t$$

which implies:

$$\phi'(0) = \frac{1}{n}\sum_{t=1}^{n} \rho''(\lambda g_t) g_t [g_t h].$$

Since g_t and $h \in L^2(\pi)$, $[g_t h] = \int g_t h d\pi$ exists, which implies that the above expression exists for all $h \in L^2(\pi)$. It follows that the Gâteaux derivative is

$$DF(\lambda) = \frac{1}{n}\sum_{t=1}^{n} \rho''(\lambda g_t) g_t g_t.$$

We can see that it is also a Fréchet derivative:

$$\begin{aligned}
\frac{\|F(\lambda+h) - F(\lambda) - DF(\lambda)h\|}{\|h\|} &= \frac{\|\frac{1}{n}\sum_{t=1}^{n} g_t \left(\rho'(\lambda g_t + h g_t) - \rho'(\lambda g_t) - \rho''(\lambda g_t) h g_t\right)\|}{\|h\|} \\
&\leq \frac{1}{n}\sum_{t=1}^{n} \frac{\|g_t \left(\rho'(\lambda g_t + h g_t) - \rho'(\lambda g_t) - \rho''(\lambda g_t) h g_t\right)\|}{\|h\|} \\
&= \frac{1}{n}\sum_{t=1}^{n} \left|\rho'(\lambda g_t + h g_t) - \rho'(\lambda g_t) - \rho''(\lambda g_t) h g_t\right| \frac{\|g_t\|}{\|h\|} \\
&= \frac{1}{n}\sum_{t=1}^{n} \left|\rho'''(\xi)(h g_t)^2/2\right| \frac{\|g_t\|}{\|h\|} \\
&= \frac{1}{n}\sum_{t=1}^{n} O(h^2) \frac{\|g_t\|}{\|h\|} \\
&\longrightarrow 0
\end{aligned}$$

as h goes to zero, where $\xi \in [\lambda g_t, h g_t]$. We have used the mean value Theorem for $\rho(v)$ which is justified by Assumption 5. Using similar arguments, we can show that the second order Fréchet derivative of $F(\lambda)$ is:

$$D^2F(\lambda) = \frac{1}{n}\sum_{t=1}^{n} \rho'''(\lambda g_t) g_t g_t g_t$$

If we come back to the first order condition (13), we can expand it around $\hat{\lambda} = 0$, using the Taylor formula in function space with the integral version of the remainder term (by Theorem 4.C. of Zeidler (1995), Assumption 5 is enough for the existence of such representation):

$$F_{n1}(\hat{\lambda}) = F_{n1}(0) + DF_{n1}(0)\hat{\lambda} + \int_0^1 (1-\delta) D^2 F_{n1}(\delta\hat{\lambda})\hat{\lambda}^2 d\delta,$$

where

$$\begin{aligned} D^2F_{n1}(\delta\hat{\lambda}) &= \frac{1}{n}\sum_{t=1}^n \rho'''(\delta\hat{\lambda}g_t)g_tg_tg_t \\ &= \rho_3\frac{1}{n}\sum_{t=1}^n g_tg_tg_t + \frac{1}{n}\sum_{t=1}^n [\rho'''(\delta\hat{\lambda}g_t) - \rho_3]g_tg_tg_t \\ &\equiv \rho_3\hat{S} + o_p(1) \end{aligned}$$

by the continuity of $\rho'''(v)$, $\rho'''(0) = \rho_3$, $\hat{\lambda} = O_p(n^{-1/2})$ and Lemma 1. $\hat{S}$, which represents the estimator of the skewness operator of g_t with kernel $\hat{s}(\tau_1, \tau_2, \tau_3) = 1/n\sum_t g_t(\tau_1)g_t(\tau_2)g_t(\tau_3)$, is bounded by assumption. Therefore,

$$\begin{aligned} D^2F_{n1}(\delta\hat{\lambda})\hat{\lambda}^2 &= \rho_3\frac{1}{n}\sum_{t=1}^n \Big[g_tg_tg_t\Big]\hat{\lambda}^2 + o_p(\|\hat{\lambda}\|^2) \\ &= \rho_3\frac{1}{n}\sum_{t=1}^n g_t[g_t\hat{\lambda}]^2 + o_p(\|\hat{\lambda}\|^2) \\ &= \rho_3\|\hat{\lambda}\|^2\hat{S}\phi^2 + o_p(\|\hat{\lambda}\|^2) \\ &= O_p(\|\hat{\lambda}\|^2) + o_p(\|\hat{\lambda}\|^2) \\ &= O_p(n^{-1}), \end{aligned}$$

where $\phi = \hat{\lambda}/\|\hat{\lambda}\|$. It follows that:

$$0 = F_{n1}(\hat{\lambda}) = F_{n1}(0) + DF_{n1}(0)\hat{\lambda} + O_p(n^{-1}).$$

which implies

$$0 = -\bar{g}(\theta) - \left(\frac{1}{n}\sum_{t=1}^n g_t(\theta)g_t(\theta)\right)\hat{\lambda} + O_p(n^{-1}) = -\bar{g}(\theta) + -\hat{K}\hat{\lambda} + O_p(n^{-1}).$$

B Proofs

B.1 Theorem 1

The steps are similar to the proof of Theorem 3.1 of Newey and Smith (2004). The following lemma is almost identical to their Lemma A1.

Lemma 1. *If Assumption 2 is satisfied, then for any ζ with $1/\nu < \zeta < (1-2\chi)/2$ and $\Lambda_\zeta = \{\lambda : \|\lambda\| < n^{-\zeta}\}$*

$$\sup_{t,\lambda\in\Lambda_\zeta,\theta\in\Theta} |\lambda g_t(\theta)| \xrightarrow{P} 0,$$

and $\Lambda_\zeta \subseteq \Lambda_n$ w.p.a.1, where $\nu > 2/(1-2\chi)$ and $0<\chi<1/2$.

Proof. Using Cauchy-Schwarz inequality and Lemma D.2 of Kitamura et al. (2004), $\sup_{t,\lambda\in\Lambda_\zeta,\theta\in\Theta} |\lambda g_t(\theta)|$ is bounded by $\|\lambda\| = O(n^{-\zeta})$ times $[\max_t \sup_\theta |g_t(\theta)|] = O_p(n^{1/\nu})$. It goes to zero by construction because $\zeta > 1/\nu$. Also, 0 is in the domain of $\rho(v)$ which implies that Λ_ζ will eventually be in the domain of $\rho(v)$. □

The only difference here is that Newey and Smith (2004) assume $\nu > 2$ and $1/\nu < \zeta < 1/2$, which is just a special case. It allows us to control the speed of convergence of λ.

The proof of the following lemma is different from the one in Newey and Smith (2004) because $\hat{\lambda}$ is a regularized solution. Therefore, we cannot define it as $\arg\max P(\lambda,\bar{\theta})$ for some convergent $\bar{\theta}$. However, we do use the fact that $\alpha_n = o(1)$ which implies that $P(\hat{\lambda},\theta) \geq P(\lambda,\theta)$ w.p.a.1 for any λ and θ.

Lemma 2. *If Assumptions 1 to 4 are satisfied, α_n goes to zero, $n\alpha_n^2$ goes to infinity and if $\bar{\theta} \xrightarrow{P} \theta_0$ and $\bar{g}(\bar{\theta}) = O_p(n^{-1/2})$, then*

$$\bar{\lambda} = \arg\min_{\Lambda_n} V(\lambda,\bar{\theta}) \equiv \left(\|F_{n1}(\lambda,\bar{\theta})\|^2 + \alpha_n\|\lambda\|^2\right)$$

exists w.p.a.1, $\bar{\lambda} = O_p(1/(\alpha\sqrt{n}))$ and $P(\bar{\lambda},\bar{\theta}) \leq \rho_0 + O_p(1/(\alpha_n^2 n))$, where $\rho_0 = \rho(0)$.

Proof. Let us define $\tilde{\lambda} = \arg\min_{\Lambda_\zeta} V(\lambda,\bar{\theta})$. Then:

$$\begin{aligned} V(0,\bar{\theta}) &\geq V(\tilde{\lambda},\bar{\theta}) \\ &= V(0,\bar{\theta}) + V'(0,\bar{\theta})\tilde{\lambda} + \left[\int_0^1 (1-\delta)V''(\delta\tilde{\lambda},\bar{\theta})d\delta\right]\tilde{\lambda}^2, \end{aligned}$$

where

$$V(0,\bar{\theta}) = \|F_{n1}(0,\bar{\theta})\|^2 = \|\bar{g}(\bar{\theta})\|^2,$$
$$V'(0,\bar{\theta}) = 2DF_{n1}(0,\bar{\theta})F_{n1}(0,\bar{\theta}) = 2\hat{K}(\bar{\theta})\bar{g}(\bar{\theta})$$

and

$$\begin{aligned} V''(\delta\tilde{\lambda},\bar{\theta}) &= 2\left[D^2F_{n1}(\delta\tilde{\lambda},\bar{\theta})F_{n1}(\delta\tilde{\lambda},\bar{\theta}) + DF_{n1}(\delta\tilde{\lambda},\bar{\theta})^2 + \alpha_n I\right] \\ &= 2\left[\frac{1}{n}\sum_{t=1}^n \rho'''(\delta\tilde{\lambda}g_t(\bar{\theta}))g_t(\bar{\theta})g_t(\bar{\theta})g_t(\bar{\theta})\right]\left[\frac{1}{n}\sum_{t=1}^n \rho'(\delta\tilde{\lambda}g_t(\bar{\theta}))g_t(\bar{\theta})\right] \\ &\quad +2\left[\sum_{t=1}^n \rho''(\delta\tilde{\lambda}g_t(\bar{\theta}))g_t(\bar{\theta})g_t(\bar{\theta})\right]^2 + 2\alpha_n I \\ &= 2\left[(-\rho_3)\hat{S}(\bar{\theta})\bar{g}(\bar{\theta}) + \hat{K}(\bar{\theta})^2 + \alpha_n I + o_p(1)\right], \end{aligned}$$

by Lemma 1 since it implies, with the properties of $\rho(v)$, that $\rho'(\delta\tilde{\lambda}g_t(\bar{\theta}))$, $\rho''(\delta\tilde{\lambda}g_t(\bar{\theta}))$ and $\rho'''(\delta\tilde{\lambda}g_t(\bar{\theta}))$ converge in probability to -1, -1 and ρ_3 respectively. The term $\hat{S}(\bar{\theta})\bar{g}(\bar{\theta})$ is a linear operator with kernel $\int_{\mathcal{T}}\hat{s}(\tau_1,\tau_2,\tau)\bar{g}(\tau;\bar{\theta})\pi(\tau)d\tau$, where $\hat{s}(\tau_1,\tau_2,\tau)=1/n\sum_t g_t(\tau_1;\bar{\theta})g_t(\tau_2;\bar{\theta})g_t(\tau;\bar{\theta})$. It is $O_p(n^{-1/2})$ by the assumption on $\bar{g}(\bar{\theta})$ and the boundness of the skewness operator. It follows that

$$V''(\delta\tilde{\lambda},\bar{\theta})=2(\hat{K}(\bar{\theta})^2+\alpha_nI)+o_p(1),$$

and then,

$$\begin{aligned}\int_0^1(1-\delta)V''(\delta\tilde{\lambda},\bar{\theta})d\delta &= 2\left[\hat{K}(\bar{\theta})^2+\alpha_nI+o_p(1)\right]\int_0^1(1-\delta)d\delta\\ &= \left[\hat{K}(\bar{\theta})^2+\alpha_nI+o_p(1)\right],\end{aligned}$$

where $\hat{K}(\bar{\theta})^2+\alpha_nI$ is a strictly positive definite linear operator since $\alpha_n>0$. It follows that

$$\begin{aligned}\|\bar{g}(\bar{\theta})\|^2 &\geq \|\bar{g}(\bar{\theta})\|^2+2\hat{K}(\bar{\theta})\bar{g}(\bar{\theta})\tilde{\lambda}+[\hat{K}(\bar{\theta})^2+\alpha_nI+o_p(1)]\tilde{\lambda}^2\\ 0 &\geq 2\hat{K}(\bar{\theta})\bar{g}(\bar{\theta})\tilde{\lambda}+[\hat{K}(\bar{\theta})^2+\alpha_nI+o_p(1)]\tilde{\lambda}^2\\ 0 &\leq -2\hat{K}(\bar{\theta})\bar{g}(\bar{\theta})\tilde{\lambda}-[\hat{K}(\bar{\theta})^2+\alpha_nI+o_p(1)]\tilde{\lambda}^2\\ &\leq C_1\|\bar{g}(\bar{\theta})\|\|\tilde{\lambda}\|-[\hat{K}(\bar{\theta})^2+\alpha_nI+o_p(1)]\tilde{\lambda}^2\\ &\leq C_1\|\bar{g}(\bar{\theta})\|\|\tilde{\lambda}\|-C_2\|\tilde{\lambda}\|^2-o_p(\|\tilde{\lambda}\|^2),\end{aligned}$$

The second last inequality comes from:

$$\begin{aligned}\left|\hat{K}(\bar{\theta})\bar{g}(\bar{\theta})\tilde{\lambda}\right| &\leq \|\hat{K}(\bar{\theta})\bar{g}(\bar{\theta})\|\|\tilde{\lambda}\|\\ &= \|\tilde{\lambda}\|\left(\sum_{t=1}^n\mu_t^2<\phi_t,\bar{g}(\bar{\theta})>^2\right)^{1/2}\\ &\leq [\max_t\mu_t)]\|\tilde{\lambda}\|\left(\sum_{t=1}^n<\phi_t,\bar{g}(\bar{\theta})>^2\right)^{1/2}\\ &= [\max_i(\mu_i)]\|\tilde{\lambda}\|\|\bar{g}(\bar{\theta})\|\end{aligned}$$

and

$$\begin{aligned}-[\hat{K}(\bar{\theta})^2+\alpha_nI]\tilde{\lambda}^2 &= -\sum_{t=1}^n(\mu_t^2+\alpha_n)<\phi_t,\tilde{\lambda}>^2\\ &\leq -[\min_t(\mu_t^2+\alpha_n)]\sum_{t=1}^n<\phi_t,\tilde{\lambda}>^2\\ &= -[\min_t(\mu_t^2+\alpha_n)]\|\tilde{\lambda}\|^2\end{aligned}$$

where μ_t and ϕ_t are the eigenvalues and eigenfunctions of $\hat{K}$ (we have omitted their dependence on the sample size for simplicity). Therefore, $C1 = \max_t \mu_t$, which is bounded away from zero, and $C_2 = \min_t(\mu_t + \alpha_n)$ with is $O_p(\alpha_n)$ since the smallest eigenvalue of $\hat{K}$ is not bounded away from zero. Therefore, we have

$$C_2\|\tilde{\lambda}\|^2 + o_p(\|\tilde{\lambda}\|^2) \leq C_1\|\bar{g}(\bar{\theta})\|\|\tilde{\lambda}\|$$

$$C_2\|\tilde{\lambda}\| + o_p(\|\tilde{\lambda}\|) \leq C_1\|\bar{g}(\bar{\theta})\| = O_p(n^{-1/2}),$$

which implies $\|\tilde{\lambda}\| = O_p(n^{-1/2}/C_2) = O_p(1/(\alpha_n\sqrt{n}))$. Notice that without α_n, the rate of convergence of $\tilde{\lambda}$ is undefined since nothing guarantees that the smallest eigenvalue of $\hat{K}^2$ is strictly positive. However, the eigenvalues of K are strictly positive. Therefore, α_n must go to zero at a speed slower than $O_p(\|\hat{K}^2 - K^2\|) = O_p(n^{-1/2})$ (see Carrasco and Florens (2000)). This is satisfied since we required that $n\alpha_n^2$ goes to infinity. Because $\Lambda_\zeta \subseteq \Lambda_n$ w.p.a.1. The second result follows.

Notice that if it was not for the restriction imposed by the domain of $\rho(v)$, the solution would exist in small sample as well because α_n guarantees that the problem is well-posed. This restriction applies only to CEL because in this case the domain of $\rho(v)$ is $]-\infty, 1[$. For the other CGEL methods considered here, the solution exists always. This is shown using Theorem 1 of Seidman and Vogel (1989).

If we substitute the solution in the objective function of CGEL we obtain:

$$\begin{aligned}
P(\bar{\lambda}, \bar{\theta}) &= \rho_0 - \bar{g}(\bar{\theta})\bar{\lambda} + \left[\int_0^1 (1-\delta)\left(\frac{1}{n}\sum_{t=1}^n \rho''(\delta\bar{\lambda}g_t(\bar{\theta}))g_t(\bar{\theta})g_t(\bar{\theta})\right)d\delta\right]\bar{\lambda}^2 \\
&\leq \rho_0 + \|\bar{g}(\bar{\theta})\|\|\bar{\lambda}\| + C\|\bar{\lambda}\|^2 \\
&= \rho_0 + O_p(1/(\alpha_n n)) + O_p(1/(\alpha_n^2 n)) \\
&= \rho_0 + O_p(1/(\alpha_n^2 n))
\end{aligned}$$

by Lemma 1 and the above results. □

Lemma 3. *If Assumptions 1 to 4 are satisfied, and* $\alpha_n = O(n^{-\chi})$ *with* $0 < \chi < 1/2$ *and* $\nu > 2/(1-2\chi)$ *in Assumption 2b), then* $\bar{g}(\hat{\theta}) = O_p(1/(\alpha_n\sqrt{n})) = o_p(1)$.

Proof. All we need is to show that we can obtain the same inequality as in Lemma A3 of Newey and Smith (2004). Let $\tilde{\lambda} = -n^{-\zeta}\bar{g}(\hat{\theta})/\|\bar{g}(\hat{\theta})\|$, where $1/\chi < \zeta < 1/2 - \nu$, then by Lemma A3 of Newey and Smith (2004):

$$P(\tilde{\lambda}, \hat{\theta}) \geq \rho_0 + n^{-\zeta}\|\bar{g}(\hat{\theta})\| - Cn^{-2\zeta}$$

Because $\hat{\lambda}$ solves the regularized first order condition, we cannot say that $P(\hat{\lambda}, \hat{\theta}) \geq P(\tilde{\lambda}, \hat{\theta})$. But it holds w.p.a.1 because as α_n goes to zero, the first order condition for λ converges to zero. Therefore, we have, w.p.a.1,

$$\rho_0 + n^{-\zeta}\|\bar{g}(\hat{\theta})\| - Cn^{-2\zeta} \leq P(\tilde{\lambda}, \hat{\theta}) \leq P(\hat{\lambda}, \hat{\theta}) \leq P(\hat{\lambda}, \theta_0) \leq \rho_0 + O_p(1/(\alpha_n^2 n))$$

It follows that

$$\begin{aligned}\rho_0 + n^{-\zeta}\|\bar{g}(\hat{\theta})\| - Cn^{-2\zeta} &\leq \rho_0 + O_p(1/(\alpha_n^2 n)) \\ &= \rho_0 + O_p(n^{-1+2\chi})\end{aligned}$$

Therefore

$$\|\bar{g}(\hat{\theta})\| \leq O_p(n^{-1+2\chi+\zeta}) + CO_p(n^{-\zeta}) = O_p(n^{-\zeta})$$

because $\zeta < 1/2 - \chi$ which implies $-1+2\chi+\zeta < \chi - 1/2 < -\zeta$. The rest is identical to Newey and Smith (2004). We pick $\bar{\lambda} = -\epsilon_n \bar{g}(\hat{\theta})$ with $\epsilon_n = o_p(1)$, which is in Λ_n w.p.a.1 and use the same saddle point argument as above. We then obtain $\epsilon_n \|\bar{g}(\hat{\theta})\|^2 = O_p(1/(\alpha_n^2 n))$ which implies $\|\bar{g}(\hat{\theta})\|^2 = O_p(1/(\alpha_n^2 n)) = o_p(1)$. □

Proof of Theorem 1. The proof is straightforward using Lemma 1 to 3 and using the same arguments as Newey and Smith (2004) for Theorem 3.1: (i) First we know from Lemma 3 that $\bar{g}(\hat{\theta}) = o_p(1)$ and by continuity of $g(\theta)$ that $\sup_\theta \|E[g(\theta)] - \bar{g}(\theta)\|$ converges to zero in probability. It follows that:

$$\begin{aligned}\|E[g(\hat{\theta})]\| &\leq \|\bar{g}(\hat{\theta})\| + \|E[g(\hat{\theta})] - \bar{g}(\hat{\theta})\| \\ &\leq \|\bar{g}(\hat{\theta})\| + \sup_\theta \|E[g(\theta)] - \bar{g}(\theta)\| \\ &\xrightarrow{p} 0\end{aligned}$$

By uniqueness of the condition $E[g(\theta_0)] = 0$, $\hat{\theta}$ must converge to θ_0. The estimator $\hat{\theta}$ can then be used in Lemma 2 to show the convergence of $\hat{\lambda}$.

□

B.2 Theorem 2

In order to prove asymptotic normality we first recall the regularized first order conditions:

$$DF(\lambda,\theta)F(\lambda,\theta) + \alpha_n\lambda = 0, \tag{27}$$

$$\frac{1}{n}\sum_{t=1}^{n}\rho'\left(\lambda g_t\right)\lambda G_t = 0, \tag{28}$$

where θ has been explicitly included in $F()$ because we will have to expand it around $\lambda = 0$ and θ_0. Notice that the subscript of $F_{n1}()$ as been omitted for notational convenience. $DF()$ is the Fréchet derivative of $F()$. It is an integral operator with kernel:

$$DF(\tau_1,\tau_2) = \frac{1}{n}\sum_{t=1}^{n}\rho''(\lambda g_t)g_t(\tau_1)g_t(\tau_2).$$

It follows that:

$$[DF(\lambda,\theta)F(\lambda,\theta)](\tau) = \int_{\mathcal{T}}\left\{\frac{1}{n}\sum_{t=1}^{n}\rho''(\lambda g_t)g_t(\tau)g_t(\tau_2)\right\}\left\{\frac{1}{n}\sum_{s=1}^{n}\rho'(\lambda g_s)g_s(\tau_2)\right\}\pi(\tau_2)d\tau_2.$$

We will denote $F'()$ as the derivative of $F()$ with respect to θ. It is an operator from $L^2(\pi)$ to $\mathbb{R}^p$ or from $\mathbb{R}^p$ to $L^2(\pi)$ depending on what turns out to be in front of it. It should always be clear from the context. We first expand equation (27) about $\lambda = 0$ and $\theta = \theta_0$. We denote F_0, DF_0, F_0' and so on, as the operators evaluated at the true value:

$$\begin{aligned} 0 &= DF(\hat{\lambda},\hat{\theta})F(\hat{\lambda},\hat{\theta}) + \alpha_n\hat{\lambda} \\ &= DF_0F_0 + \left[D^2F_0F_0 + DF_0DF_0 + \alpha_n I\right]\hat{\lambda} \\ &\quad + \left[DF_0'F_0 + DF_0F_0'\right](\hat{\theta}-\theta_0) + O_p(\|\hat{\lambda}\|^2 + \|\hat{\theta}-\theta_0\|^2), \end{aligned}$$

where D^2F is the Fréchet derivative of DF and DF' is the derivative of DF with respect to θ. Let us develop each term one by one (recall that $\rho'(0) = \rho''(0) = -1$ and define $\rho_3 = \rho'''(0)$):

$$DF_0 = \frac{1}{n}\sum_{t=1}^n \rho''(0)g_tg_t = -\hat{K}(\theta_0)$$

$$F_0 = \frac{1}{n}\sum_{t=1}^n \rho'(0)g_t = -\bar{g}(\theta_0).$$

It follows that:

$$DF_0F_0 = \hat{K}(\theta_0)\bar{g}(\theta_0)$$

and

$$D^2F_0 = \frac{1}{n}\sum_{t=1}^n \rho'''(0)g_tg_tg_t = \rho_3\hat{S}(\theta_0),$$

where $\hat{S}$ if the estimated skewness operator. It follows that:

$$D^2F_0F_0 = -\rho_3\hat{S}(\theta_0)\bar{g}(\theta_0).$$

The other terms can be obtained in the same way. The expansion of the regularized first order conditions is then:

$$\begin{aligned} 0 &= \hat{K}_0\bar{g}_0 + \left\{\hat{K}_0^2 - \rho_3\hat{S}_0\bar{g}_0 + \alpha_n I\right\}\hat{\lambda} \\ &\quad + \left\{[2\bar{G}g_0]\bar{g}_0 + \hat{K}_0\bar{G}_0\right\}(\hat{\theta}-\theta_0) + O_p(n^{-1}). \end{aligned}$$

The second equation can be expanded in the same way:

$$0 = -\overline{G}_0\hat{\lambda} + O_p(n^{-1}).$$

We can rewrite the above equations in the following compact representation:

$$0 = -B + A_1\hat{\lambda} + A_2(\hat{\theta}-\theta_0) + O_p(n^{-1}) \tag{29}$$

$$0 = A_3\hat{\lambda} + O_p(n^{-1}), \tag{30}$$

where:

$$B = -\hat{K}_0\bar{g}_0,$$

$$A_1 = (\hat{K}_0^2 + \alpha_n I) + O_p(n^{-1}),$$

because $\hat{S}_0\bar{g}_0$ and $\hat{\lambda}_n$ are $O_p(n^{-1/2})$,

$$A_2 = \hat{K}_0\bar{G}_0 + O_p(n^{-1/2})$$

and

$$A_3 = \bar{G}_0.$$

We can solve the system to obtain the following:

$$\sqrt{n}\hat{\lambda} = [I - A_1^{-1}A_2(A_3A_1^{-1}A_2)^{-1}A_3]A_1^{-1}\sqrt{n}B + o_p(1)$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) = (A_3A_1^{-1}A_2)^{-1}A_3A_1^{-1}\sqrt{n}B + o_p(1).$$

If we analyze the last term of each equation, we have:

$$\begin{aligned}
-A_1^{-1}\sqrt{n}B &= (\hat{K}_0^2 + \alpha_n I)^{-1}\hat{K}_0[\sqrt{n}\bar{g}_0] + o_p(n^{-1}) \\
&= (\hat{K}_0^{\alpha_0})^{-1}[\sqrt{n}\bar{g}_0] + o_p(n^{-1}) \\
&= K^{-1}g + \left\{(\hat{K}_0^{\alpha_n})^{-1}[\sqrt{n}\bar{g}_0] - K^{-1}g\right\} + o_p(n^{-1}) \\
&= K^{-1}g + o_p(1)
\end{aligned}$$

as n goes to infinity, α_n goes to zero and $n\alpha_n^3 \to \infty$ by theorem 7 (ii) of Carrasco and Florens (2000), where $g \sim N(0, K)$. Therefore, $A_1^{-1}\sqrt{n}B$ converges to $N(0, K^{-1})$ (See appendix A.1 for details on $\hat{K}^{\alpha_n}$). Using the convergence properties of A_1, A_2 and A_3, we obtain:

$$\sqrt{n}\hat{\lambda} \xrightarrow{L} \left[I - K^{-1}G(GK^{-1}G)^{-1}G\right] N(0, K^{-1})$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{L} \left[(GK^{-1}G)^{-1}G\right] N(0, K^{-1}).$$

The rest of the proof follows by simple manipulations.

B.3 Theorem 4

In Appendix A.2, it is shown that

$$\hat{K}\hat{\lambda} = -\bar{g}(\hat{\theta}) + o_p(1).$$

It follows that

$$(\hat{K}^{\alpha_n})^{-1/2}\hat{K}\hat{\lambda} = -(\hat{K}^{\alpha_n})^{-1/2}\bar{g}(\hat{\theta}) + o_p(1).$$

Theorem 7 (i) of Carrasco and Florens (2000) implies that

$$(\hat{K}^{\alpha_n})^{-1/2}\hat{K}f_n = \hat{K}^{1/2}f_n + o_p(1).$$

which proves the first order equivalence of $\|\sqrt{n}\hat{K}^{1/2}\hat{\lambda}\|$ and $\|\sqrt{n}(\hat{K}^{\alpha_n})^{-1/2}\bar{g}(\hat{\theta})\|$.

In order to show the first order equivalence of $\widetilde{LR}$, we expand the CGEL objective function about $\lambda = 0$:

$$\begin{aligned}
2nP(\hat{\lambda}, \hat{\theta}) &= 2nP(0, \hat{\theta}) + 2nP_{\lambda}(0, \hat{\theta})\hat{\lambda} + 2n\hat{\lambda}P_{\lambda\lambda}(\tilde{\lambda}, \hat{\theta})\hat{\lambda} \\
&= 2\rho(0) - 2n\bar{g}(\hat{\theta})\hat{\lambda} + n\hat{\lambda}\left(\frac{1}{n}\sum_{t=1}^{n} \rho''(\tilde{\lambda}g_t(\hat{\theta}))g_t(\hat{\theta})g_t(\hat{\theta})\right)\hat{\lambda} \\
&= 2\rho(0) - n\hat{\lambda}\left(\frac{1}{n}\sum_{t=1}^{n} g_t(\hat{\theta})g_t(\hat{\theta})\right)\hat{\lambda} + o_p(1) \\
&= 2\rho(0) - n\hat{\lambda}\hat{K}\hat{\lambda} + o_p(1)
\end{aligned}$$

by Lemma 1, where $\tilde{\lambda} \in [0, \hat{\lambda}]$.

The second part of the theorem follows by simple manipulation using the singular value representation of the inverse problem solution. The CGMM objective function can be written as:

$$\|(\hat{K}^{\alpha_n})^{-1/2}\sqrt{n}\bar{g}(\hat{\theta})\|^2 = \sum_{i=1}^{n}\left(\frac{\mu_i^{(n)2}}{\mu_i^{(n)2} + \alpha_n}\right)\frac{< \bar{g}(\theta), \phi_i^{(n)} >^2}{\mu_i^{(n)}},$$

where

$$\phi_i^{(n)} = \frac{\nu_i^{(n)}}{\|\nu_i^{(n)}\|},$$

and

$$\begin{aligned}
\|\nu_i^{(n)}\|^2 &= \left\langle \frac{1}{n}\sum_{j=1}^{n} \beta_{ji}g_j, \frac{1}{n}\sum_{j=1}^{n}\beta_{ji}g_j \right\rangle \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{l=1}^{n}\beta_{ji}\beta_{li}\int g_j(\tau)g_j(\tau)\pi(\tau)d\tau \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\beta_{ji}\beta_{li}C_{jl} \\
&= \frac{1}{n}\sum_{j=1}^{n}\beta_{ji}\left(C_{j\bullet}\beta_i\right) \\
&= \frac{1}{n}\sum_{j=1}^{n}\beta_{ji}\left(\mu_i^{(n)}\beta_{ji}\right) \\
&= \frac{\mu_i^{(n)}}{n}\|\beta_i\|^2 = \frac{\mu_i^{(n)}}{n}.
\end{aligned}$$

It follows that

$$
\begin{aligned}
< \sqrt{n}\bar{g}(\hat{\theta}), \phi_i^{(n)} > &= \left\langle \frac{1}{\sqrt{n}} \sum_{t=1}^{n} g_t, \frac{1}{\sqrt{n\mu_i^{(n)}}} \sum_{j=1}^{n} \beta_{ji} g_j \right\rangle \\
&= \frac{1}{n\sqrt{\mu_i^{(n)}}} \sum_{t=1}^{n} \sum_{j=1}^{n} \beta_{ji} \int g_t(\tau) g_j(\tau) \pi(\tau) d\tau \\
&= \frac{1}{\sqrt{\mu_i^{(n)}}} \sum_{t=1}^{n} \sum_{jk=1}^{n} \beta_{ji} C_{tj} \\
&= \frac{1}{\sqrt{\mu_i^{(n)}}} \sum_{t=1}^{n} C_{t\bullet} \beta_i = \frac{1}{\sqrt{\mu_i^{(n)}}} \sum_{t=1}^{n} \mu_i^{(n)} \beta_{ti} \\
&= \sqrt{\mu_i^{(n)}} \iota' \beta_i.
\end{aligned}
$$

Therefore, the CGMM objective function becomes:

$$
\|(\hat{K}^{\alpha})^{-1/2} \sqrt{n} \bar{g}(\hat{\theta})\|^2 = \sum_{i=1}^{n} \left(\frac{\mu_i^{(n)^2}}{\mu_i^{(n)^2} + \alpha_n} \right) (\iota' \beta_i)^2,
$$

which concludes the proof for $\tilde{J}$. The proof of the $\widetilde{LM}$ representation is much simpler:

$$
\begin{aligned}
\hat{\lambda} \hat{K} \hat{\lambda} &= \int \int \hat{\lambda}(\tau_1) \hat{\lambda}(\tau_2) \left(\frac{1}{n} \sum_{t=1}^{n} g_t(\tau_1) g_t(\tau_2) \right) \pi(\tau_1) \pi(\tau_2) d\tau_1 d\tau_2 \\
&= \frac{1}{n} \sum_{t=1}^{n} \int \hat{\lambda}(\tau_1) g_t(\tau_1) \pi(\tau_1) d\tau_1 \int \hat{\lambda}(\tau_2) g_t(\tau_2) \pi(\tau_2) d\tau_2 \\
&= \frac{1}{n} \sum_{t=1}^{n} (\hat{\lambda} g_t)^2.
\end{aligned}
$$

The result follows.

C Computation of CGEL

C.1 Computation using the singular value decomposition.

We suppose that C has m eigenvalues different from zero. We define β as the $n \times m$ matrix containing the m eigenvectors associated with the eigenvalues. We can therefore write the solution as:

$$
\tilde{\lambda} = -\sum_{i=1}^{m} \left(\frac{\mu_i^{(n)}}{\mu_i^{(n)^2} + \alpha_n} \right) < \hat{g}, \nu_i^{(n)} > \nu_i^{(n)}.
$$

Numerically, the truncation parameter m can be set equals to the rank of C. This will allow m to increase with the sample size since, as n goes to infinity and $\hat{K}$ converges to K, the rank goes to infinity. Because it is not λ but $< g_t, \lambda >$ which enters the objective function, we only need to compute the latter (θ has been omitted for simplicity):

$$< g_t, \tilde{\lambda} >= -\sum_{i=1}^{m} \left(\frac{\mu_i^{(n)}}{\mu_i^{(n)^2} + \alpha_n} \right) < \hat{g}, \nu_i^{(n)} >< g_t, \nu_i^{(n)} >,$$

where:

$$\begin{aligned} < g_t, \nu_j^{(n)} > &= \int_{\mathcal{T}} g_t(\tau) \left(\frac{1}{n} \sum_{i=1}^{m} \beta_{ij} g_i(\tau) \pi(\tau) d\tau \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \beta_{ij} \int_{\mathcal{T}} g_t(\tau) g_i(\tau) \pi(\tau) d\tau \\ &= \sum_{i=1}^{n} \beta_{ij} C_{ti} \\ &= C_{t\bullet} \beta_j, \end{aligned}$$

where $C_{t\bullet}$ is the t^{th} line of C. We can do the same for the other inner product:

$$\begin{aligned} < \hat{g}, \nu_j^{(n)} > &= \int_{\mathcal{T}} \hat{g}(\tau) \left(\frac{1}{n} \sum_{i=1}^{m} \beta_{ij} g_i(\tau) \pi(\tau) d\tau \right) \\ &= \int_{\mathcal{T}} \left(\frac{1}{n} \sum_{t=1}^{n} g_t(\tau) \right) \left(\frac{1}{n} \sum_{i=1}^{m} \beta_{ij} g_i(\tau) \pi(\tau) d\tau \right) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{t=1}^{n} \beta_{ij} \int_{\mathcal{T}} g_t(\tau) g_i(\tau) \pi(\tau) d\tau \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{n} \beta_{ij} C_{ti} \\ &= \frac{1}{n} \sum_{t=1}^{n} C_{t\bullet} \beta_j \\ &= \frac{1}{n} \iota' C \beta_j, \end{aligned}$$

where ι is a $n \times 1$ vector of ones. Therefore we can write:

$$\begin{aligned} < g_t, \tilde{\lambda} > &= -\sum_{i=1}^{m} \left(\frac{\mu_i^{(n)}}{\mu_i^{(n)^2} + \alpha_n} \right) \left[\frac{1}{n} \iota' C \beta_i \right] [C_{t\bullet} \beta_i] \\ &= -\frac{1}{n} \left[\iota' C \right] \sum_{i=1}^{m} \left(\frac{\mu_i^{(n)}}{\mu_i^{(n)^2} + \alpha_n} \right) \left[\beta_i \beta_i' \right] C_{\bullet t} \\ &= -\frac{1}{n} \iota' C \left(\beta D \beta' \right) C_{\bullet t}, \end{aligned}$$

where D is the diagonal matrix defined in the text. The objective function is therefore:

$$\frac{1}{n}\sum_{t=1}^{n}\rho\left(-\frac{1}{n}\iota' C\left(\beta D\beta'\right)C_{\bullet t}\right).$$

C.2 Computation using the regularized Gauss-Newton method

To simplify the notation, we will set $g_t \equiv g_t(\theta)$, $\lambda \equiv \lambda_{i-1}$, $\lambda' \equiv \lambda_i$, $p_t = \rho'(\lambda g_t)$, $p'_t = \rho'(\lambda' g_t)$, $p_t^2 = \rho''(\lambda g_t)$ and $p_t'^2 = \rho''(\lambda' g_t)$. We want to rewrite the following algorithm:

$$\lambda' = \lambda - \left\{DF(\lambda)^2 + \alpha_n I\right\}^{-1}\left\{DF(\lambda)F(\lambda) + \alpha_n\lambda\right\}$$

which can be written as:

$$\begin{aligned}\left\{DF(\lambda)^2 + \alpha_n I\right\}\lambda' &= \left\{DF(\lambda)^2 + \alpha_n I\right\}\lambda - DF(\lambda)F(\lambda) - \alpha_n\lambda \\ &= DF(\lambda)^2\lambda - DF(\lambda)F(\lambda).\end{aligned}$$

What we want to do is to rewrite each term, multiply them by $g_s(\tau_1)\pi(\tau_1)$ and integrate. The first term of the left hand side is:

$$\begin{aligned}DF(\lambda)^2\lambda' &= \int_{\mathcal{T}}\left\{\int_{\mathcal{T}}\left(\frac{1}{n}\sum_{t=1}^{n}p_t^2 g_t(\tau_1)g_t(\tau_2)\right)\left(\frac{1}{n}\sum_{l=1}^{n}p_l^2 g_l(\tau_2)g_l(\tau_3)\right)\pi(\tau_2)d\tau_2\right\}\lambda'(\tau_3)\pi(\tau_3)d\tau_3 \\ &= \frac{1}{n^2}\sum_t\sum_l p_t^2 p_l^2 g_t(\tau_1)\left(\int_{\mathcal{T}} g_t g_l \pi d\tau_2\right)\left(\int_{\mathcal{T}} g_l\lambda'\pi d\tau_3\right) \\ &= \frac{1}{n}\sum_t\sum_l p_t^2 p_l^2 g_t(\tau_1)C_{tl} < g_l, \lambda' > .\end{aligned}$$

Once we apply the transformation, the term becomes:

$$\begin{aligned}\left[DF(\lambda)^2\lambda'\right]g_s &= \int_{\mathcal{T}}\frac{1}{n}\sum_t\sum_l p_t^2 p_l^2 g_t(\tau_1)C_{tl} < g_l, \lambda' > g_s(\tau_1)\pi(\tau_1)d\tau_1 \\ &= \frac{1}{n}\sum_t\sum_l p_t^2 p_l^2 C_{tl} < g_l, \lambda' > \int_{\mathcal{T}} g_t(\tau_1)g_s(\tau_1)\pi(\tau_1)d\tau_1 \\ &= \sum_t\sum_l p_t^2 p_l^2 C_{tl} < g_l, \lambda' > C_{ts} \\ &= \sum_t C_{ts}p_t^2\left[C_{t\bullet}V < g, \lambda' >\right] \\ &= C_{s\bullet}VCV < g, \lambda' >,\end{aligned}$$

where V is defined in the text and $< g, \lambda >$ is the $n\times 1$ vector with typical element $< g_t, \lambda >$. Since it has to be valid for all $s = 1, \cdots, n$, The first term on the left hand side can be written as follows:

$$\left[DF(\lambda)^2\lambda'\right]g = (CV)^2 < g, \lambda' > .$$

It follows that the first term of the right hand side is:

$$\left[DF(\lambda)^2\lambda\right]g = (CV)^2 < g, \lambda > .$$

Clearly, the second term of the left hand side is simply $\alpha_n < g_t, \lambda' >$. The left hand side can therefore be written as:

$$\left\{(CV)^2 + \alpha_n I\right\} < g, \lambda' > .$$

The second term on the right hand side is:

$$\begin{aligned} DF(\lambda)F(\lambda) &= \int_{\mathcal{T}} \left[\frac{1}{n}\sum_t p_t^2 g_t(\tau_1)g_t(\tau_2)\right]\left(\frac{1}{n}\sum_l p_l g_l(\tau_2)\right)\pi(\tau_2) \\ &= \frac{1}{n^2}\sum_t\sum_l p_t^2 p_l^2 g_t(\tau_1)\int_{\mathcal{T}} g_t(\tau_2)g_l(\tau_2)\pi(\tau_2)d\tau_2 \\ &= \frac{1}{n}\sum_t\sum_l p_t^2 p_l^2 g_t(\tau_1)C_{tl}. \end{aligned}$$

If we apply the transformation it becomes:

$$\begin{aligned} \left[DF(\lambda)F(\lambda)\right]g_s &= \int_{\mathcal{T}} \frac{1}{n}\sum_t\sum_l p_t^2 p_l^2 g_t(\tau_1)C_{tl}g_s(\tau_1)\pi(\tau_1)d\tau_1 \\ &= \sum_t\sum_l p_t^2 p_l^2 C_{tl}C_{ts}. \end{aligned}$$

For all $s = 1, \cdots, n$ the term can be written as:

$$CVCP,$$

where P is defined in the text. We can therefore rewrite the iterative procedure as follows:

$$\left\{(CV)^2 + \alpha_n I\right\} < g, \lambda' >= (CV)^2 < g, \lambda > -CVCP,$$

which implies

$$< g, \lambda' >= \left\{(CV)^2 + \alpha_n I\right\}^{-1}\left\{(CV)^2 < g, \lambda > -CVCP\right\}.$$

If we start with $\lambda_0 = 0$, then $V = I$ and $P = \iota$ which gives us the starting value:

$$< g, \lambda' >= \left\{C^2 + \alpha_n I\right\}^{-1}\left\{-C^2\iota\right\} = -\left\{C^2 + \alpha_n I\right\}^{-1}C^2\iota.$$

D CCUE and CEEL

D.1 Note on CCUE

In Section 4.1, we argue that the exact solution of $\hat{\lambda}(\theta)$, in the case of CEEL, can be obtained from the linear ill-posed problem $\hat{K}\hat{\lambda} = -\bar{g}(\theta)$. In this case, the iterative procedure stops after the first iteration and the solution is $\hat{\lambda}(\theta) = -(\hat{K}^{\alpha_n})^{-1}\bar{g}(\theta)$. Because $\rho()$ is quadratic, we can write the objective function as:

$$\begin{aligned} P(\hat{\lambda}(\theta), \theta) &= \rho(0) + \bar{g}(\hat{K}^{\alpha_n})^{-1}\bar{g} - \bar{g}\hat{K}(\hat{K}^{\alpha_n})^{-2}\bar{g}/2 \\ &= \rho(0) + \frac{1}{2}\bar{g}(\hat{K}^{\alpha_n})^{-1}\bar{g} + o_p(1), \end{aligned}$$

because $\hat{K}(\hat{K}^{\alpha_n})^{-2} = (K_n^{\alpha_n})^{-1} + o_p(1)$. Therefore, CEEL is equivalent to CCUE, defined as CGMM in which $\hat{K}^{\alpha_n}(\tilde{\theta})$ is replaced by $\hat{K}^{\alpha_n}(\theta)$, only asymptotically.

D.2 CEEL and the ill-posedness of CGEL

The case in which $\rho(v)$ is quadratic offers a way to show that linear and nonlinear ill-posed problems are very different. If we consider the following system of n linear equations $Ax = y$, in which the matrix A is poorly conditioned, the stability of the solution is an issue only if the right-hand side is random. In CGMM, we need the solution to $Kx = \bar{g}$ in order to compute the objective function. Because $\bar{g}$ is random, the properties of K imply that the system is ill-posed. In nonlinear system of equations, the problem can be ill-posed even if the right-hand side is not random as in equation (13). For example, when the number of conditions is finite and $\rho(v) = -v - 0.5v^2$, the equation is:

$$F_{n1}(\lambda) \equiv \frac{1}{n}\sum_{t=1}^{n}(-1 - g_t'\lambda)g_t = 0,$$

which implies that $\hat{\lambda}(\theta)$ is the solution to the following system of linear equations:

$$\hat{K}\hat{\lambda}(\theta) = -\bar{g}$$

Since $\bar{g}$ is random, the solution is unstable if $\hat{K}$ is poorly conditioned. For the case of a continuum, it is ill-posed by the properties of the covariance operator. The randomness of the left-hand side $F_{n1}(\lambda)$ is therefore as important as the randomness of the right-hand side for the stability of the solution, as opposed to linear ill-posed problem. Equation (13) is therefore ill-posed.

E Results from the numerical experiment

Statistics		CGMM	iter-CEL	iter-CET	sv-CEL	sv-CET	CEEL
Mean-bias	$\alpha = 0.1$	0.00511	0.00224	0.00515	0.05656	0.06127	0.06306
	$\alpha = 0.05$	0.00151	0.01083	0.02337	0.05853	0.06555	0.05454
	$\alpha = 0.01$	0.00748	0.04408	0.06147	0.03742	0.02964	0.03033
	$\alpha = 0.005$	0.00967	0.04968	0.07106	0.03386	0.03620	0.03503
	$\alpha = 0.001$	0.00975	0.09151	0.11687	0.02441	0.02269	0.02397
	$\alpha = 0.0001$	0.01401	0.14470	0.16625	0.00847	0.00815	0.01009
Median-bias	$\alpha = 0.1$	0.00127	0.01266	0.01561	0.05474	0.06141	0.06142
	$\alpha = 0.05$	0.01026	0.01683	0.03613	0.05584	0.06792	0.04752
	$\alpha = 0.01$	0.01957	0.05064	0.06343	0.02915	0.02676	0.02703
	$\alpha = 0.005$	0.02058	0.05137	0.07142	0.02358	0.03363	0.02544
	$\alpha = 0.001$	0.02185	0.09419	0.12366	0.02003	0.02316	0.02217
	$\alpha = 0.0001$	0.00995	0.15933	0.18253	0.00213	0.00212	0.00020
RMSE	$\alpha = 0.1$	0.15966	0.14768	0.14067	0.16790	0.17389	0.17482
	$\alpha = 0.05$	0.16460	0.13782	0.13709	0.17343	0.17451	0.17486
	$\alpha = 0.01$	0.16853	0.14213	0.13702	0.15941	0.15865	0.15640
	$\alpha = 0.005$	0.16584	0.13527	0.13566	0.16840	0.16627	0.16078
	$\alpha = 0.001$	0.17807	0.14354	0.15421	0.15181	0.15599	0.15628
	$\alpha = 0.0001$	0.19515	0.17119	0.18734	0.14400	0.14480	0.15229

Table 1: Properties of the estimator of ω for a sample size of 100

Statistics		CGMM	iter-CEL	iter-CET	sv-CEL	sv-CET	CEEL
Mean-bias	$\alpha = 0.1$	0.14000	0.02898	0.03098	0.04843	0.02750	0.02947
	$\alpha = 0.05$	0.12867	0.03820	0.01306	0.02096	0.02668	0.02518
	$\alpha = 0.01$	0.18330	0.01677	0.02680	0.03701	0.03237	0.03472
	$\alpha = 0.005$	0.19821	0.03615	0.02610	0.00630	0.01785	0.01742
	$\alpha = 0.001$	0.22838	0.04349	0.07008	0.00132	0.00548	0.00047
	$\alpha = 0.0001$	0.27198	0.02963	0.00750	0.02892	0.03156	0.00761
Median-bias	$\alpha = 0.1$	0.06998	0.04613	0.04240	0.04521	0.03573	0.04324
	$\alpha = 0.05$	0.01829	0.03456	0.07287	0.04187	0.03385	0.02262
	$\alpha = 0.01$	0.08276	0.08898	0.07669	0.04032	0.01835	0.02850
	$\alpha = 0.005$	0.08022	0.08246	0.09298	0.00346	0.00385	0.00238
	$\alpha = 0.001$	0.14274	0.09316	0.05920	0.03139	0.02833	0.02550
	$\alpha = 0.0001$	0.21524	0.15297	0.18696	0.07053	0.07520	0.05335
RMSE	$\alpha = 0.1$	0.54745	0.48857	0.48443	0.39627	0.36578	0.37339
	$\alpha = 0.05$	0.56286	0.48177	0.48472	0.38051	0.38536	0.37783
	$\alpha = 0.01$	0.57676	0.51385	0.52319	0.39518	0.38935	0.39701
	$\alpha = 0.005$	0.58441	0.52163	0.53106	0.38610	0.38887	0.39672
	$\alpha = 0.001$	0.57955	0.54570	0.56422	0.39252	0.40272	0.41305
	$\alpha = 0.0001$	0.56002	0.57135	0.55940	0.38983	0.38423	0.41655

Table 2: Properties of the estimator of β for a sample size of 100

Statistics		CGMM	iter-CEL	iter-CET	sv-CEL	sv-CET	CEEL
Mean-bias	$\alpha = 0.1$	0.01207	0.01280	0.01102	0.01002	0.01365	0.01423
	$\alpha = 0.05$	0.01239	0.01187	0.01141	0.01371	0.01445	0.01142
	$\alpha = 0.01$	0.01061	0.01198	0.01003	0.00687	0.00949	0.00946
	$\alpha = 0.005$	0.00617	0.01277	0.01252	0.00945	0.01138	0.00893
	$\alpha = 0.001$	0.00620	0.00629	0.00545	0.00803	0.00584	0.00616
	$\alpha = 0.0001$	0.00250	0.00397	0.00584	0.00704	0.00686	0.00506
Median-bias	$\alpha = 0.1$	0.01410	0.01344	0.01254	0.00928	0.01307	0.01260
	$\alpha = 0.05$	0.01419	0.01277	0.01267	0.01221	0.01477	0.01221
	$\alpha = 0.01$	0.01002	0.01119	0.01194	0.00685	0.01046	0.01046
	$\alpha = 0.005$	0.00668	0.01338	0.01240	0.00917	0.01227	0.00929
	$\alpha = 0.001$	0.00608	0.00678	0.00518	0.00916	0.00588	0.00610
	$\alpha = 0.0001$	0.00735	0.00589	0.00813	0.00713	0.00691	0.00710
RMSE	$\alpha = 0.1$	0.05102	0.04821	0.04806	0.05432	0.05274	0.05499
	$\alpha = 0.05$	0.05293	0.04758	0.04857	0.05429	0.05434	0.05459
	$\alpha = 0.01$	0.05026	0.04727	0.04803	0.05186	0.05239	0.05215
	$\alpha = 0.005$	0.04927	0.04889	0.04825	0.05123	0.05080	0.05250
	$\alpha = 0.001$	0.05048	0.04695	0.04777	0.05022	0.05016	0.05008
	$\alpha = 0.0001$	0.13785	0.04977	0.04909	0.04868	0.04871	0.04955

Table 3: Properties of the estimator of γ for a sample size of 100

Statistics		CGMM	iter-CEL	iter-CET	sv-CEL	sv-CET	CEEL
Mean-bias	$\alpha = 0.1$	0.00037	0.00337	0.00359	0.00701	0.01638	0.01658
	$\alpha = 0.05$	0.00265	0.00067	0.00393	0.02034	0.01526	0.01499
	$\alpha = 0.01$	0.00108	0.00771	0.01653	0.00850	0.00123	0.00173
	$\alpha = 0.005$	0.00277	0.01275	0.02290	0.01874	0.01233	0.01401
	$\alpha = 0.001$	0.00692	0.02253	0.03106	0.01047	0.01361	0.01306
	$\alpha = 0.0001$	0.01282	0.03692	0.03869	0.01328	0.01363	0.00760
Median-bias	$\alpha = 0.1$	0.01578	0.00196	0.00006	0.00128	0.00413	0.00468
	$\alpha = 0.05$	0.00959	0.00344	0.00860	0.00786	0.00019	0.00670
	$\alpha = 0.01$	0.01437	0.01001	0.01543	0.00345	0.00962	0.00938
	$\alpha = 0.005$	0.02217	0.01697	0.02435	0.00831	0.00227	0.00404
	$\alpha = 0.001$	0.02914	0.02270	0.03527	0.00654	0.00627	0.00613
	$\alpha = 0.0001$	0.02096	0.04269	0.04271	0.00943	0.01028	0.00112
RMSE	$\alpha = 0.1$	0.12803	0.10768	0.10649	0.11880	0.14077	0.14123
	$\alpha = 0.05$	0.13424	0.10424	0.10373	0.13247	0.12813	0.12116
	$\alpha = 0.01$	0.13363	0.09945	0.09385	0.11651	0.11659	0.11627
	$\alpha = 0.005$	0.13073	0.09717	0.09963	0.12913	0.11941	0.14442
	$\alpha = 0.001$	0.15350	0.10349	0.09614	0.11262	0.14056	0.14065
	$\alpha = 0.0001$	0.12668	0.09667	0.09333	0.11083	0.11075	0.12064

Table 4: Properties of the estimator of δ for a sample size of 100

	Mean-bias	Median-bias	RMSE	SD
ω	-0.05759	-0.01835	0.14810	0.13654
β	-0.07979	0.10357	0.38982	0.38184
γ	0.01576	0.02135	0.04233	0.03932
δ	-0.00047	0.00135	0.12413	0.12422

Table 5: Properties of the estimators using MLE for a sample size of 100

Tests	CGMM	iter-CEL	iter-CET	sv-EL	sv-ET	CEEL
J-Test for $\alpha=$ 0.1	0.0010	0.0400	0.0300	0.0420	0.0580	0.0590
J-Test for $\alpha=$ 0.05	0.0030	0.0230	0.0300	0.0830	0.0760	0.0690
J-Test for $\alpha=$ 0.01	0.0260	0.0780	0.0680	0.1290	0.1170	0.1130
J-Test for $\alpha=$ 0.005	0.0360	0.0950	0.0840	0.1850	0.1840	0.1960
J-Test for $\alpha=$ 0.001	0.1420	0.1570	0.1530	0.3220	0.3360	0.3370
J-Test for $\alpha=$ 0.0001	0.5400	0.5820	0.5300	0.7820	0.7840	0.7830
LM-Test for $\alpha=$ 0.1		0.0040	0.0000	0.0040	0.0000	0.0010
LM-Test for $\alpha=$ 0.05		0.0040	0.0000	0.0100	0.0000	0.0000
LM-Test for $\alpha=$ 0.01		0.0530	0.0110	0.0610	0.0090	0.0020
LM-Test for $\alpha=$ 0.005		0.1290	0.0200	0.0800	0.0090	0.0120
LM-Test for $\alpha=$ 0.001		0.3660	0.0870	0.2560	0.0880	0.0710
LM-Test for $\alpha=$ 0.0001		0.5170	0.2150	0.4970	0.3040	0.3080
LR-Test for $\alpha=$ 0.1		0.0000	0.0000	0.0080	0.0080	0.0100
LR-Test for $\alpha=$ 0.05		0.0000	0.0000	0.0260	0.0180	0.0230
LR-Test for $\alpha=$ 0.01		0.0020	0.0030	0.1070	0.0890	0.0950
LR-Test for $\alpha=$ 0.005		0.0060	0.0010	0.1600	0.1490	0.1680
LR-Test for $\alpha=$ 0.001		0.0130	0.0070	0.4030	0.3580	0.3670
LR-Test for $\alpha=$ 0.0001		0.0410	0.0130	0.7260	0.6940	0.7180
J-Imhof for $\alpha=$ 0.1	0.0010	0.0400	0.0310	0.0420	0.0580	0.0590
J-Imhof for $\alpha=$ 0.05	0.0030	0.0240	0.0310	0.0830	0.0770	0.0730
J-Imhof for $\alpha=$ 0.01	0.0260	0.0790	0.0680	0.1320	0.1200	0.1160
J-Imhof for $\alpha=$ 0.005	0.0370	0.0980	0.0840	0.1870	0.1890	0.1980
J-Imhof for $\alpha=$ 0.001	0.1460	0.1610	0.1560	0.3280	0.3440	0.3450
J-Imhof for $\alpha=$ 0.0001	0.5430	0.5860	0.5380	0.7870	0.7890	0.7870

Table 6: Sizes of tests of overidentifying rectrictions (level=0.05,sample size=100)

References

R.G. Airayetpan and A.G. Ramm. Dynamical systems and discrete methods for solving nonlinear ill-posed problem. *Applied Mathematics Review*, 1:491–536, 2000.

S. Anatolyev. Gmm, gel, serial correlation, and asymptotic bias. *Econometrica*, 73:983–1002, 2005.

S. Anatolyev and N. Gospodinov. Specification testing in models with many instruments. 2008.

B. Antoine, H. Bonnal, and E. Renault. On the efficient use of the informational content of estimating equations: Implied probabilities and euclidean empirical likelihood. *Journal od Econometrics*, 138:461–487, 2007.

M. Arelanno, L.P. Hansen, and E. Sentana. Underidentification? *Mimeo*, 2011.

J. Berkowitz. Generalized spectral estimation of the consumption based pricing model. *Journal of Econometrics*, 104:269–288, 2001.

B. Blaschke, A. Neubauer, and O. Scherzer. On the convergence rates for the iteratively regularized gauss-newton method. *IMA Journal of Numerical Analysis*, 17:421–436, 1997.

M. Carrasco. A regularization approach to the many instruments problem. *Working Paper: Université de Montréal*, 2011.

M. Carrasco and J.P. Florens. Generalization of gmm to a continuum of moment conditions. *Econometric Theory*, 16:655–673, 2000.

M. Carrasco and J.P. Florens. Efficient gmm estimation using the empirical characteristic function. 2002.

M. Carrasco and R. Kotchoni. Efficient estimation using the characteristic function. *Working Paper: Université de Montréal*, 2010.

M. Carrasco, M. Chernov, J.P. Florens, and E. Ghysels. Efficient of general dynamic models with continuun of moment conditions. *Journal of Econometrics*, (140):529–573, 2007a.

M. Carrasco, J.P. Florens, and E. Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6B:5633–5751, 2007b.

M.G. Dagenais. Extension of the ridge regression technique to non-linear models with additive errors. *Economics Letters*, 12(2):169–174, 1983.

S. Donald and W. Newey. Choosing the number of instruments. *Econometrica*, 69:1161–1191, 2001.

S. Donald, G. Imbens, and W. Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117:55–93, 2003.

R. Garcia, E. Renault, and D. Veredas. Estimation of stable distribution by indirect inference. *Working Paper: UCL and CORE*, 2006.

C. Groetsch. Inverse problems in mathematical sciences. *Wiesbaden: Vieweg*, 1993.

P. Guggenberger. Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator. *Econometric Reviews*, 26:526–541, 2008.

P. Guggenberger and J. Hahn. Finite sample properties of the two-step empirical likelihood estimator. *Econometric Reviews*, 24(3):247–263, 2005.

L.P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

L.P. Hansen, J. Heaton, and A. Yaron. Finit-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, 14:262–280, 1996.

J.P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48:419–426, 1961.

Q. Jin. On the iteratively regularized gauss-newton method for solving nonlinear ill-posed problems. *Mathematics of Computation*, 69(232):1603–1623, 2000.

Y. Kitamura and M. Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(5):861–874, 1997.

Y. Kitamura, G. Tripathi, and H. Ahn. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72:1667–1714, 2004.

D.G. Luenberger. *Optimization by Vector Space Methods*. Wiley and Sons, 1997.

W.K. Newey and R.J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72:219–255, 2004.

J.P. Nolan. Modeling financial data with stable distributions. *Math/Stat Department, American University*, 2005.

J.P. Nolan. Stable distributions. *Math/Stat Department, American University*, 2009.

A.B. Owen. *Empirical Likelihood*. Chapman and Hall, 2001.

A.G. Ramm. Dynamical systems method for solving operator equations. *Comm. Nonlinear Sci. and Numer. Simul.*, 9(N2):383–402, 2004a.

A.G. Ramm. Dynamical systems method for solving nonlinear operator equations. *Int. J. Appl. Math. Sci.*, 1:97–110, 2004b.

S.M. Schennach. Point estimation with exponentially tilted empirical likelihood. *Econometrica*, 35(2):634–672, 2007.

T.I. Seidman and C.R. Vogel. Well posedness and convergence of some regularisation methods for non-linear ill-posed problems. *Inverse Problems*, 5:227–238, 1989.

R.J. Smith. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *The Economic Journal*, 107:503–519, 1997.

R.J. Smith. Gel criteria for moment condition models. 2004.

Diethelm Wuertz and core team members Rmetrics. *fBasics: Rmetrics - Markets and Basic Statistics*, 2010. URL `http://CRAN.R-project.org/package=fBasics`. (Code builtin from the following R contributed packages are used: gmm from Pierre Chaussé and gld from Robert King and gss from Chong Gu and nortest from Juergen Gross and HyperbolicDist from David Scott and sandwich from Thomas Lumley and Achim Zeileis and and fortran/C code from Kersti Aas.).

E. Zeidler. *Applied Functional Analysis: Main Principles and Their Applications*. Springer-Verlag, 1995.