Estimation of a Mixture via the Empirical Characteristic Function

Marine Carrasco — CREST, INSEE
Jean-Pierre Florens — IDEI, Université Toulouse I

This version: January 2000

1. Introduction

In many circumstances, the likelihood function does not have a simple tractable expression. Examples, that will be developed later, are the convolution and the mixture of distributions. In such instances, estimation using the characteristic function offers a nice alternative to maximum likelihood method. It has been shown by Feuerverger and McDunnough (1981) that the empirical characteristic function yields an efficient estimator when used with a specific weighting function. However, this weighting function depends on the likelihood which is of course unknown. This poses the problem of the implementation of this method. Here we show that the empirical characteristic function yields a continuum of moment conditions that can be handled by the method developed by Carrasco and Florens (1999). We simply estimate the parameters of the model by GMM based on this continuum of moment conditions. We show that this method delivers asymptotically efficient estimators while being relatively easy to implement. A close investigation shows that Carrasco-Florens' results gives a rationale to Feuerverger and McDunnough's approach and is much more general since it applies to any continuum of moments. Using our continuous GMM method avoids the explicit derivation of the optimal weighting function as in Feuerverger and McDunnough. We give a general method to estimate it from the data.

Next, we discuss the efficient estimation based on the conditional characteristic function. As long as identifiability holds, our estimators reach the Cramer Rao efficiency bound for any choice of instruments. The issue on optimal instruments can be completely ignored here. The way we choose the weight in our GMM objective function guarantees efficiency.

In Section 2, we give the principal definitions and two examples. Section 3 reviews the results of Carrasco-Florens (1999). Section 4 explains how to obtain efficient estimators using the (unconditional) characteristic function. In Section 5, we turn our attention to the use of the conditional characteristic function. Section 6 discusses the implementation. Section 7 develops an example on duration of stay in a state. Finally, Section 8 concludes.

2. Definitions and examples

2.1. Definitions

Suppose $X_1, ..., X_n$ are iid realizations of the same random variable X with density $f_\theta(x)$ and c.d.f. $F_\theta(x)$. $\theta \in \mathbf{R}^k$ is the parameter of interest. θ_0 is the true value of θ. Let $\psi_\theta(t)$ denote the characteristic function of X

$$\psi_\theta(t) \equiv \int e^{itx} dF_\theta(x) = E^\theta\left(e^{itX}\right)$$

and $\psi_n(t)$ denote the empirical characteristic function

$$\psi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}.$$

We construct moment conditions of the form

$$h(t, X_j; \theta) = e^{itX_j} - \psi_\theta(t).$$

Obviously h satisfies

$$E^\theta\left[h(t, X_j; \theta)\right] = 0 \text{ for all } t \text{ in } \mathbf{R}.$$

Our aim is to use this continuum of moment conditions to obtain an efficient estimator of θ.

Let μ be a probability on $\mathbf{R}$ and $L^2(\mu)$ be the Hilbert space of complex valued functions such that

$$L^2(\mu) = \left\{f : \mathbf{R} \to \mathbf{C} | \int f(t)^2 \mu(dt) < \infty\right\}.$$

Note that while $\psi_\theta(t)$ is not integrable on $\mathbf{R}$, $\psi_\theta(t)$ belongs to $L^2(\mu)$ for any probability μ because $\psi_\theta(t) \leq \psi_\theta(0) = 1$. Candidates for μ include the density of the standard normal distribution and $\mu(t) = I\{-K \leq t \leq K\}$ where I is the indicator function. These different choices will be investigated by simulation later.

2.2. Examples

In the following, we give two motivating examples.

Example 1: Finite mixture of distributions

Finite mixture models are commonly used to model data from a population composed of a finite number of homogeneous subpopulations. An example of application is the estimation of a cost function in presence of multiple technologies of production (see Beard, Caudill, and Gropper, 1991). Morduch and Stern (1997) recently used a mixture model to detect sex bias in health outcomes in Bangladesh. Ignoring heterogeneity may lead to seriously misleading results.

Consider the case where a population is supposed to be formed of two homogeneous subpopulations. Let $X_1, ..., X_n$ be an iid sample. Let π be the unknown proportion of individuals of type 1. Individuals of type 1 have a density $f\left(x, \theta_1\right)$ and those of type 2 have a density $f\left(x, \theta_2\right)$. The econometrician does not observe the type of the individuals, so that the likelihood for one observation is

$$\pi f\left(x, \theta_1\right) + (1 - \pi) f\left(x, \theta_2\right).$$

Such models can be estimated using the EM algorithm or the method of moments, see Heckman, Robb, and Walker (1990) among others. An alternative way is the use of the characteristic function which is equal to

$$\pi \psi_{\theta 1}\left(t\right) + (1 - \pi) \psi_{\theta 2}\left(t\right)$$

where $\psi_{\theta_j}\left(t\right) = \int e^{itx} dF_{\theta j}\left(x\right)$ with $j = 1, 2$.

Example 2: Convolution of distributions

Assume one observes the sum of two random variables $X = Y + Z$, where Y and Z are independent and individually non-observed. In most cases, the likelihood will have an intractable form whereas the characteristic function of X is easily obtained from

$$\psi_X = \psi_Y \times \psi_Z$$

where ψ_Y and ψ_Z are the characteristic functions of Y and Z respectively.

In a microeconometric model, Y might be a person-specific heterogeneity term while Z is an error term. Our results below will allow for the presence of covariates in the model. Both notions of mixture and convolution are closely related, see Mundlack and Yahav (1981).

3. Brief review of GMM when a continuum of moment is available

Let H be the Hilbert space of reference. Here $H = L^2(\mu)$. In the following $\|.\|$ will denote the norm in $L^2(\mu)$. Let B be a bounded linear operator defined on H or a subspace of H and B_n a sequence of random bounded linear operators converging to B. Let

$$\bar{h}_n(t;\theta) = \frac{1}{n}\sum_{j=1}^{n} h(t, X_j;\theta).$$

The GMM estimator is such that

$$\hat{\theta}_n = \arg\min_{\theta}\left\|B_n \bar{h}_n(.;\theta)\right\|.$$

Under a set of conditions listed in Carrasco and Florens (1999), this estimator is consistent and asymptotically normal. In the class of all weighting operator B, one yields to an estimator with minimal variance. This optimal B is shown to be equal to $K^{-1/2}$ where K is the covariance operator associated with $h(t, X;\theta)$. That is

$$\begin{array}{rcl} K & : & f \in H \to g \in H \\ f(t) & \to & g(s) = \int k(s,t) f(t) \mu(dt) \end{array}$$

where

$$k(s,t) = E^{\theta_0}\left[h(s, X;\theta_0) h(t, X;\theta_0)\right]$$

$K^{-1/2}g$ does not exist on the whole space H but on a subset of it which corresponds to the so-called reproducing kernel Hilbert space (RKHS) associated with K denoted $\mathcal{H}(K)$. We use the notation

$$\left\|K^{-1/2}g\right\| = \|g\|_K$$

where $\|.\|_K$ denotes the norm in $\mathcal{H}(K)$. Since the inverse of K is not bounded, we use a penalization term α_n to guarantee the existence of the inverse. The estimation of K and the choice of α_n will be discussed in Section (6). Let K_n^{α} denote a consistent estimator of K. The optimal GMM estimator of θ is obtained by:

$$\hat{\theta}_n = \arg\min_{\theta}\left\|\left(K_n^{\alpha}\right)^{-1/2}\bar{h}_n(.;\theta)\right\|.$$

Under the assumptions A1 to A11 listed in Appendix, we have the following results:

$$\hat{\theta}_n \to \theta_0 \text{ in probability,}$$

as n and $n\alpha_n^{3/2}$ go to infinity and α_n goes to zero and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, \left(\left\| E^{\theta_0}\left(\frac{\partial h}{\partial \theta'}\right)\right\|_K^2\right)^{-1}\right) \tag{3.1}$$

as n and $n\alpha_n^3$ go to infinity and α_n goes to zero.

4. Estimation using the characteristic function

4.1. A useful result of Parzen

Parzen (1970, page 25) gives a simple formula for the norm of a function g in the RKHS associated with a covariance kernel k on $\mathbf{R}^2$. Assume that k takes the form

$$k(s,t) = \int u(s,x)\, u(t,x)\, P(dx) \tag{4.1}$$

where P is a measure and $\{u(t,.)\,, t \in \mathbf{R}\}$ is a family of functions in $L^2(P)$. If (4.1) holds, the RKHS norm of g is given by

$$\|g\|_K^2 = \|G\|_{L^2(P)}^2 \tag{4.2}$$

where G is solution of the equation

$$g(t) = \int G(x)\, u(t,x)\, P(dx)\,. \tag{4.3}$$

This result will be useful in the following to calculate the inverse of the variance of our GMM estimator $\hat{\theta}_n$, namely:

$$\left\| E^{\theta_0}\left(\frac{\partial h}{\partial \theta'}\right)\right\|_K^2 .$$

4.2. Efficiency

In this section, we check Assumptions A1 to A11 and show that the GMM method described above applies. Next, we determine the asymptotic variance of $\hat{\theta}_n$ and show that it coincides with the Cramer Rao efficiency bound. We have the following relation

$$\begin{aligned} \bar{h}_n(t;\theta) &= \frac{1}{n}\sum_{j=1}^{n}\left(e^{itX_j}-\psi_\theta(t)\right) \\ &= \psi_n(t)-\psi_\theta(t). \end{aligned}$$

Feuerverger and Mureika (1977) showed that $\sqrt{n}\bar{h}_n(t;\theta_0)$ converges weakly to a Gaussian process with mean zero and covariance

$$\begin{aligned} E^{\theta_0}\left[\bar{h}_n(s;\theta_0)\bar{h}_n(t;\theta_0)\right] &= E^{\theta_0}\left[\left(e^{isX}-\psi_{\theta_0}(s)\right)\left(e^{itX}-\psi_{\theta_0}(t)\right)\right] \\ &= \psi_{\theta_0}(t+s)-\psi_{\theta_0}(t)\psi(s). \end{aligned}$$

It follows that Assumption 8 is satisfied with

$$k(s,t)=\psi_{\theta_0}(t+s)-\psi_{\theta_0}(t)\psi_{\theta_0}(s).$$

The next step it to calculate the variance of $\hat{\theta}_n$. We now apply results (4.2), (4.3) in our setting. We have

$$k(s,t)=\int\left(e^{isx}-\psi_{\theta_0}(s)\right)\left(e^{itx}-\psi_{\theta_0}(t)\right)dF_{\theta_0}(x)$$

so that

$$u(t,x)=e^{itx}-\psi_{\theta_0}(t)$$

and

$$g(t)=E^{\theta_0}\left(\frac{\partial h}{\partial\theta'}\right)=-\left.\frac{\partial\psi_\theta}{\partial\theta}\right|_{\theta=\theta_0}.$$

We are looking for the solution of

$$\begin{aligned} \left.\frac{\partial\psi_\theta}{\partial\theta}\right|_{\theta=\theta_0} &= \int G(x)\left(e^{itx}-\psi_{\theta_0}(t)\right)dF_{\theta_0}(x) \\ &= \int G(x)e^{itx}dF_{\theta_0}(x)-\psi_{\theta_0}(t)E^{\theta_0}(G) \qquad (4.4) \end{aligned}$$

Let

$$G(x) = \left. \frac{\partial \ln f_\theta}{\partial \theta'} \right|_{\theta=\theta_0} \tag{4.5}$$

then, we obtain

$$\begin{aligned} \int G(x) e^{itx} dF_{\theta_0}(x) &= \int \frac{\partial f_\theta(x)}{\partial \theta'} e^{itx} dx \\ &= \left. \frac{\partial}{\partial} \psi_\theta(t) \right|_{\theta=\theta_0} \end{aligned}$$

and

$$E^{\theta_0}(G) = E^{\theta_0}\left(\frac{\partial \ln f_\theta}{\partial \theta'} \right) = 0.$$

This shows that G defined in (4.5) is solution to (4.4). Hence, we obtain

$$\left\| E^{\theta_0}\left(\frac{\partial h}{\partial \theta'} \right) \right\|_K^2 = E^{\theta_0}\left[\left(\left. \frac{\partial \ln f_\theta}{\partial \theta'} \right|_{\theta=\theta_0} \right)^2 \right] \equiv I_\theta$$

where I_θ corresponds to the Fisher information matrix. Therefore $\hat{\theta}_n$ is as efficient as the maximum likelihood estimator.

4.3. Comparison with Feuerverger and McDunnough

Feuerverger and McDunnough (1981) (to be referred to as FM) show that a way to reach the efficiency bound is to estimate θ (assumed for simplicity to be scalar) by minimizing

$$\int_{-\infty}^{+\infty} w(t) (\psi_n(t) - \psi_\theta(t)) dt = 0 \tag{4.6}$$

with

$$w(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left. \frac{\partial \ln f_\theta}{\partial \theta} \right|_{\theta=\theta_0} e^{-itx} dx.$$

They obtain this optimal weighting function w by solving (see FM, page 25):

$$Kw(t) = \int k(s,t) w(s) ds = \left. \frac{\partial \psi_\theta}{\partial \theta} \right|_{\theta=\theta_0}.$$

In other words,

$$w(t) = K^{-1} \left. \frac{\partial \psi_\theta}{\partial \theta} \right|_{\theta=\theta_0},$$

assuming that $\frac{\partial \psi_\theta}{\partial \theta}$ belongs to the range of K. (4.6) coincides with our first order condition, indeed our objective function is given by

$$Q_n = \left(K^{-1/2}\bar{h}_n\left(.;\theta\right), K^{-1/2}\bar{h}_n\left(.;\theta\right)\right),$$

here we use K instead of its estimator for comparison with FM. The FOC condition is

$$\left(K^{-1/2}\frac{\partial}{\partial \theta}\bar{h}_n\left(.;\theta\right), K^{-1/2}\bar{h}_n\left(.;\theta\right)\right) = 0. \tag{4.7}$$

Assuming that $\frac{\partial \psi_\theta}{\partial \theta}(= \frac{\partial}{\partial \theta}\bar{h}_n\left(.;\theta\right))$ belongs to the range of K, (4.7) can be rewritten as

$$\begin{aligned} \left(K^{-1}\frac{\partial \psi_\theta}{\partial \theta}, \bar{h}_n\left(.;\theta\right)\right) &= 0, \\ \left(w, \psi_n\left(t\right) - \psi_\theta\left(t\right)\right) &= 0 \end{aligned}$$

which coincides with (4.6). FM are aware that $\frac{\partial \psi_\theta}{\partial \theta}$ is not integrable and propose as solution to take

$$w_m\left(t\right) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \left.\frac{\partial \ln f_\theta}{\partial \theta}\right|_{\theta=\theta_0} I\left\{-m \leq x \leq m\right\} e^{-itx}dx.$$

A major problem is that w depends on the likelihood which is, of course, unknown. FM suggests to discretize an interval of $\mathbf{R}$and to apply the usual GMM on the resulting set of moment conditions. However, discretization means a loss of efficiency. They argue that letting the interval between observations, Δt , go to zero, the GMM estimator will reach the efficiency bound. From Carrasco and Florens, it is clear that the passage at the limit requires a lot of care and that, when Δt goes to zero, the dimension of the covariance matrix increases and its inverse is not bounded.

5. Conditional characteristic function

In practice, models frequently include explanatory variables so that estimation has to rely on the conditional characteristic function (CCF). In this section, we explain how to construct moment conditions without loss of efficiency.

Assume that an iid sample $X_i = (Y_i, Z_i)$ is available. Denote the characteristic function of Y conditional on $Z$by

$$\psi_\theta\left(t|Z\right) \equiv E^\theta\left(e^{itY}|Z\right)$$

For now, Z may be exogenous with respect to θ or not. We want to exploit the knowledge of the CCF to estimate θ. Let $m(.)$ be a function of Z (independent of θ), we know that

$$E\left[\left(e^{itY}-\psi_{\theta}(t|Z)\right)m(Z)\right]=0.$$

Inference can be based on the continuum of moments

$$h(t,X_j,\theta)=\left(e^{itY_j}-\psi_{\theta}(t|Z_j)\right)m(Z_j). \tag{5.1}$$

To assess the efficiency of this method, we follow the same approach as in Section 4.2. The kernel of the covariance is given by:

$$\begin{aligned} k(s,t) &= E^{\theta_0}\left[h(t,X_j,\theta)\,h'(t,X_j,\theta)\right] \\ &= E^{\theta_0}\left[m(Z_j)\,m(Z_j)'\left(\psi_{\theta}(s+t|Z_j)-\psi_{\theta}(s|Z_j)\,\psi_{\theta}(t|Z_j)\right)\right]. \end{aligned}$$

The asymptotic variance of the estimator $\hat{\theta}_n$ given in (3.1) involves the norm in the RKHS associated with k of

$$E^{\theta_0}\left(\frac{\partial h}{\partial\theta'}\right)=E^{\theta_0}\left[-\frac{\partial\psi_{\theta}}{\partial\theta'}m(Z)\bigg|_{\theta=\theta_0}\right]$$

To simplify the notation, we assume m, z, and θ are scalar. We again apply results (4.2), (4.3) with

$$g(t)=E^{\theta_0}\left[-\frac{\partial\psi_{\theta}}{\partial\theta'}m(Z)\bigg|_{\theta=\theta_0}\right].$$

We have

$$\|g\|_K=\|G\|^2_{L^2(P)}$$

where G is solution of

$$g(t)=\int G(y,z)\left(e^{ity}-\psi_{\theta_0}(t|z)\right)m(z)\,dF_{\theta_0}(x). \tag{5.2}$$

Let

$$G(y,z)=-\frac{\partial\ln f_{\theta}}{\partial\theta'}(y|z)\bigg|_{\theta=\theta_0}$$

and replace in (5.2). In the following, we drop θ_0. The first term in the right-hand side of (5.2) equals

$$\begin{aligned} -\int m(z)\,f_{\theta}(z)\left[\int\frac{\partial f_{\theta}}{\partial\theta}(y|z)\,e^{ity}dy\right]dz &= -\int m(z)\,f_{\theta}(z)\,\frac{\partial\psi_{\theta}}{\partial\theta}(t|z)\,dz \\ &= E^{\theta}\left[-\frac{\partial\psi_{\theta}}{\partial\theta'}m(Z)\right]. \end{aligned}$$

The second term on the right hand side of (5.2) equals

$$\begin{aligned}\int m(z)\, G(y,z)\, \psi_\theta(t|z)\, f_\theta(y|z)\, f_\theta(z)\, dydz &= \int m(z)\, \frac{\partial f_\theta}{\partial\theta}(y|z)\, \psi_\theta(t|z)\, f_\theta(z)\, dydz \\ &= \int m(z)\, \psi_\theta(t|z)\, f_\theta(z) \left[\int \frac{\partial f_\theta}{\partial\theta}(y|z)\, dy\right] dz \\ &= 0.\end{aligned}$$

Hence the inverse of the variance of $\hat{\theta}_n$ is given by

$$\|G\|^2_{L^2(P)} = E^{\theta_0}\left[\left(\left.\frac{\partial \ln f_\theta(y|z)}{\partial\theta'}\right|_{\theta=\theta_0}\right)^2\right].$$

If moreover, Z is exogenous with respect to θ, that is its distribution does not depend on θ, we have

$$\|G\|^2_{L^2(P)} = E^{\theta_0}\left[\left(\left.\frac{\partial \ln f_\theta(y,z)}{\partial\theta'}\right|_{\theta=\theta_0}\right)^2\right] = I_\theta.$$

It shows that if the distribution of Z depends on θ then there is a loss of efficiency of using the conditional characteristic function. But if Z is exogenous, the estimator is efficient whatever the choice of the instruments m. This is a very important result. It means that the choice of m is dictated by identification only. Efficiency is guaranteed by the fact that we use an infinity of moments.

Proposition 5.1. *Assume Z is exogenous for θ. Let m be such that θ is identified from a single moment (5.1) indexed by t then the GMM estimator $\hat{\theta}_n$ based on the continuum of moment (5.1) with $t \in \mathbf{R}$ is efficient.*

6. Implementation

6.1. Estimation of the covariance operator

A first step estimator of θ denoted $\hat{\theta}_1$ is obtained by minimizing $\left\|\overline{h}(\theta)\right\|_{(L^2)^q}$ (this corresponds to $\tilde{\theta}$ for $B = I$). The covariance operator K can be estimated by replacing its kernel by the sample covariance using the first step estimator $\hat{\theta}_1$.

$$K_T : f \to \int_0^1 \frac{1}{T}\sum_{t=1}^T h_t\left(\pi_1, \hat{\theta}_1\right) h_t\left(\pi_2, \hat{\theta}_1\right)' f(\pi_2)\, d\pi_2$$

The operator K_T is degenerate and therefore has a finite number of eigenfunctions. The $k \times 1-$eigenfunctions ϕ are solutions of

$$\frac{1}{T}\sum_{t=1}^{T} h_t\left(\pi_1, \hat{\theta}_1\right) \int_0^1 h_t\left(\pi_2, \hat{\theta}_1\right)' \phi\left(\pi_2\right) d\pi_2 = \mu\phi\left(\pi_1\right)$$

$\phi\left(\pi\right)$ is necessarily of the form $\sum_{t=1}^{T} \beta_t h_t\left(\pi, \hat{\theta}_1\right)$. Replacing ϕ by its expression, one gets

$$\frac{1}{T}\sum_{t=1}^{T} h_t\left(\pi_1, \hat{\theta}_1\right)\left[\sum_{s=1}^{T} \beta_s \int_0^1 h_t\left(\pi_2\right)' h_s\left(\pi_2\right) d\pi_2\right] = \mu \sum_{s=1}^{T} \beta_s h_s\left(\pi_1\right)$$

Therefore, the $\{\beta_t\}$ satisfy

$$\frac{1}{T}\sum_{s=1}^{T} \beta_s \int_0^1 h_t\left(\pi_2\right)' h_s\left(\pi_2\right) d\pi_2 = \mu\beta_t, \text{ for } t = 1, ..., T$$

Let C be the matrix with cells $c_{st} = \frac{1}{T}\int_0^1 h_s\left(\pi\right)' h_t\left(\pi\right) d\pi$. Let $\underline{\beta}^j = \left[\beta_1^j, ..., \beta_T^j\right]$ be the orthonormal eigenvectors and μ_j the eigenvalues of C. Hence, the jth eigenfunction of K_T is given by

$$\phi_j\left(\pi\right) = \sum_{t=1}^{T} \beta_t^j h_t\left(\pi, \hat{\theta}_1\right)$$

and is associated to the eigenvalues μ_j.

Remark 1. *Note that even if the $\underline{\beta}^j$ are orthonormal, the ϕ_j are not necessarily orthonormal, so they need to be orthonormalized using Gram-Schmidt process (see for instance Hochstadt p. 47). From now on, $\left\{\phi_j\right\}_j$ denote the orthonormal eigenfunctions of K_T.*

Remark 2. *The dimension of the matrix to diagonalize, C, is $T \times T$. So it increases with the sample size, thus the computational time increases very fast. On the other hand, the dimension of C is $T \times T$ whatever the number of moment conditions is, so adding moment conditions does not complicate the calculation of the eigenvalues and eigenfunctions.*

6.2. Estimation of the inverse of K

The calculation of the objective function involves the calculation of $K^{-1/2}$ where $K^{-1/2}$ can be seen as $(K^{-1})^{1/2}$. We first study the properties of K^{-1}. $K^{-1}f$ is solution of

$$Kg = f$$

This integral equation is called a Fredholm equation of the first kind. This problem is typically ill-posed. The solution does not exist necessarily, when it exists, it is not unique in general, and the solution does not depend continuously on the input f. To guarantee the uniqueness of the solution, we restrict our attention to least-squares solutions of minimum norm. g is solution of $\inf\{\| Ku - f \| : u \in (L^2)^k\}$. The least-squares solution will exist if f lie in $R(K) + N(K)$ where $R(K)$ is the range of K and $N(K)$ is the null space of K. The instability of the solution plagues the estimation method especially when f is observed with an error, which happens in our case since f is estimated. One way to address this problem is to replace the ill-posed problem for a well-posed problem. This can be done by including a regularization parameter λ. The problem

$$(K^2 + \lambda I)g_\lambda = Kf$$

approximates the initial Fredholm equation. The solution g_λ is called Tikhonov approximation to $K^{-1}f$. This method is described in some details by Groetsch (1993, p.84). Tikhonov's method has a nice variational interpretation. Indeed, the solution of

$$\min_g \|Kg - f\|^2 + \lambda \|g\|^2$$

is $(K^2 + \lambda I)^{-1} Kf$. So that the presence of λ penalizes large values of $\|g\|^2$. Let $f = (f_1, f_2, ..., f_q)'$ and $\phi = \left(\phi_1, \phi_2, ..., \phi_q\right)'$. The square root of the generalized inverse of K is

$$\left(K_T^\lambda\right)^{-1/2} f = \sum_{j=1}^{T} \frac{\sqrt{\mu_j}}{\sqrt{\mu_j^2 + \lambda}} \left(f, \phi_j\right) \phi_j = \sum_{j=1}^{T} \frac{\sqrt{\mu_j}}{\sqrt{\mu_j^2 + \lambda}} \left\{ \sum_{i=1}^{k} \left(f_i, \phi_{ji}\right) \phi_{ji} \right\}$$

for f in the reproducing kernel Hilbert space associated to K. Clearly, the solution g_λ should converge to $K^{-1}f$ when λ goes to zero, but for λ close to zero, the solution becomes instable. There is a trade-off between the accuracy of the solution and its stability. Therefore, the right choice of λ is crucial.

6.3. Choice of the regularization parameter λ

The choice of the regularization parameter λ is delicate. From the simulations, it appears that $\hat{\theta}$ is not very sensitive to the choice of λ. However, λ plays a determinant role in the estimation of the variance that enters in the Wald test. λ must converge to zero to guarantee the convergence of $\left(K_T^{\lambda}\right)^{-1/2} f$ to $K^{-1/2}f$, but it should not converge to zero too fast to guarantee the stability of the solution. The theoretical result requires $T\lambda^3 \to \infty$ as T goes to infinity. But this is not very informative on how to choose λ. A satisfying choice of λ should be based on the data. Several approaches are present in the literature, see Groetsch (1993, p.84). We will describe one of these methods: the discrepancy principle which basically a method of cross-validation. This method has been developed to improve the estimation of the solution of a Fredholm equation of the first kind and was not meant to be used in testing. The strategy is the following. Assume that one wants to solve the following equation: $K^{\lambda 1/2} g_T^{\lambda} = f_T$ where $\|f_T - f\| = O(\frac{1}{\sqrt{T}})$. First assume K known. The idea of the discrepancy principle is to choose the regularization parameter so that the size of the residual $d(\lambda) = \left\|K^{1/2}\left(K^{\lambda -1/2} f_T\right) - f_T\right\|$ is the same as the error level $\frac{1}{\sqrt{T}}$. In our case, K is not known but estimated by K_T. Therefore, the problem is

$$K_T^{\lambda 1/2} g_T^{\lambda} = f_T.$$

Assume that $\| f_T - f \| \leq \frac{a}{\sqrt{T}}$, where a is some scalar and that $\| K_T^{1/2} g - K^{1/2} g \|^2 \leq \| g \|^2 \; h_T$ where h_T is some constant, c, times T, and $\| g \|^2 \leq D$. Then, the discrepancy method consists in choosing λ such that

$$\| K_T^{1/2} g_T^{\lambda} - f_T \|^2 = \Delta^2$$

where $\Delta^2 = (cD + a)/T$. In consequence, the approximation error is chosen to be equal to the error level in the data. Indeed $\| K_T^{1/2} g_T^{\lambda} - f_T \|^2$ is decreasing in λ, therefore if $\|f_T\| \geq \frac{a}{\sqrt{T}}$, the problem above has always a solution. Moreover, it can be shown that the resulting $g_T^{\lambda(T)}$ converges to $K^{-1/2}f$ as T goes to infinity, in $(L^2)^k$-norm. The proof is similar to that proposed by Groetsch (1993, p.90). For a detailed discussion of the choice of Δ in the case of inexact data, read Morozov (1984, p.53). To implement the test W_{GMM}, first we need to choose a function f_T and a constant c and then select λ so that $d(\lambda) = \Delta^2$. f_T can not be chosen equal to h because h belongs to the kernel of K. A better choice is $\frac{\partial \bar{h}}{\partial \theta}$ but this is a matrix as soon as the dimension of θ is greater than 1. One could choose a component

$\frac{\partial \bar{h}}{\partial \theta_j}$. Finally, it is worth to notice that the rate of λ obtained by setting $d(\lambda) = \Delta^2$ may differ from the rate stated in Section 3. This last rate is a sufficient condition and not a necessary and sufficient condition.

7. Application to mixture of distributions

7.1. The model

Assume one observes i.i.d realizations of durations T and exogenous variables Z. Let T_0 be a latent duration and ε an unmeasured person-specific heterogeneity such that for an individual i, one observes

$$T_i = \exp\left(\beta' Z_i + \varepsilon_i\right) T_{0i} \tag{7.1}$$

where ε_i and T_{0i} are assumed to be independent. Lancaster's book (1990) gives many examples of (7.1). T_i may be for instance the unemployment spell. Taking the logarithm, we have the regression

$$\ln T_i = \beta' Z_i + \varepsilon_i + \ln T_{0i}. \tag{7.2}$$

Models of this type have been more often specified in terms of hazard than in terms of regression (Lancaster, p.219). While (7.2) gives rise to a convolution problem, specification in terms of hazard gives rise to a mixture problem. Estimation by maximum likelihood where the mixing distribution is integrated out can be performed using the EM algorithm (Lancaster, Chapter 8).

Assume that $\varepsilon \sim iid\mathcal{N}\left(0,\sigma\right)$ and $T_0 \sim Gamma\left(\nu\right)$. Denote $\eta = \ln T_0$ and $\theta = \left(\beta', \nu, \sigma\right)'$. Since ε and η are independent, we have

$$\psi_X = \psi_\varepsilon \times \psi_\eta.$$

Moreover, we have

$$\begin{aligned} \psi_\varepsilon &= e^{-\sigma^2 t^2/2} \\ &= \psi_{R\varepsilon} + i\psi_{I\varepsilon} \end{aligned}$$

and

$$\psi_\eta = \psi_{R\eta} + i\psi_{I\eta}.$$

Let $X = \varepsilon + \eta$. The characteristic function of X is given by:

$$\begin{aligned} \operatorname{Re}\left(\psi_X\right) &\equiv \psi_{RX} = \psi_{R\varepsilon}\psi_{R\eta} - \psi_{I\varepsilon}\psi_{I\eta} \\ \operatorname{Im}\left(\psi_X\right) &\equiv \psi_{IX} = \psi_{I\varepsilon}\psi_{R\eta} + \psi_{R\varepsilon}\psi_{I\eta}. \end{aligned}$$

The conditional characteristic function of $\ln T_i$ given Z_i has for real and imaginary parts

$$\begin{aligned} \operatorname{Re}\left(\psi_{\ln T}\right) &= \cos\left(t\beta' Z\right)\psi_{RX} - \sin\left(t\beta' Z\right)\psi_{IX}, \\ \operatorname{Im}\left(\psi_{\ln T}\right) &= \cos\left(t\beta' Z\right)\psi_{IX} + \sin\left(t\beta' Z\right)\psi_{RX}. \end{aligned}$$

7.2. The method

We pick as instrument a function $m(Z)$ which choice is dictated by identifiability consideration. We could choose $m(Z) = Z$, for instance. The moment conditions are given by $h = (h_1, h_2)'$ with

$$\begin{aligned} h_1\left(t, X_j; \theta\right) &= \left[\cos\left(tX_j\right) - \operatorname{Re}\left(\psi_{\ln T}\left(t, \theta\right)\right)\right] m(Z), \\ h_2\left(t, X_j; \theta\right) &= \left[\sin\left(tX_j\right) - \operatorname{Im}\left(\psi_{\ln T}\left(t, \theta\right)\right)\right] m(Z). \end{aligned}$$

In the following, we will assume Z and β scalar.

First step estimator:

The first step estimator is given by

$$\hat{\theta}_n^1 = \arg\min_\theta \|h\|_{L^2(\mu)} = \arg\min_\theta \int \left(\bar{h}_1\left(t;\theta\right)\right)^2 \mu(dt) + \int \left(\bar{h}_2\left(t;\theta\right)\right)^2 \mu(dt).$$

where $\bar{h}_i\left(t;\theta\right) = \frac{1}{n}\sum_{j=1}^n h_i\left(t, X_j; \theta\right)$, $i = 1, 2$.

Estimation of the covariance:

The kernel can be estimated by

$$\hat{k}(\pi_1, \pi_2) = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T \tilde{h}_1\left(\pi_1\right)\tilde{h}_1\left(\pi_2\right) & \frac{1}{T}\sum_{t=1}^T \tilde{h}_1\left(\pi_1\right)\tilde{h}_2\left(\pi_2\right) \\ \frac{1}{T}\sum_{t=1}^T \tilde{h}_2\left(\pi_1\right)\tilde{h}_1\left(\pi_2\right) & \frac{1}{T}\sum_{t=1}^T \tilde{h}_2\left(\pi_1\right)\tilde{h}_2\left(\pi_2\right) \end{pmatrix}$$

where $\tilde{h} = h - \bar{h}$. We chose to center h because this reduces importantly the sensitivity of the test to the serial correlation. It does not change anything from a theoretical point of view. By the method described in Section 4, one can get estimators of the eigenfunctions $\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}$ and the eigenvalues. The eigenfunctions are 2-dimensional vectors of functions. In the following, we shall denote

$$\|f_1\|_{K_1}^2 = \sum_j \frac{\mu_j}{\mu_j^2 + \lambda}\left(f_1, \phi_{1j}\right)^2$$

for any f_1 in $L^2[0,1]$ such that the sum above converges. Symmetrically, we define

$$\|f_2\|_{K_2}^2 = \sum_j \frac{\mu_j}{\mu_j^2 + \lambda} \left(f_2, \phi_{2j} \right)^2$$

for any f_2 in $L^2[0,1]$ such that the sum above converges. Let

$$(f_1, f_2)_{K_{12}} = \sum_j \frac{\mu_j}{\mu_j^2 + \lambda} \left(f_1, \phi_{1j} \right) \left(f_2, \phi_{2j} \right)$$

These notations may lead to think of K_1 as an operator with eigenfunctions ϕ_{1j} and eigenvalues μ_j and similarly for K_2. Note however that while $\left\{ \phi_j \right\}_j = \left\{ \begin{pmatrix} \phi_{1j} \\ \phi_{2j} \end{pmatrix} \right\}_j$ have been constructed to be orthonormal, the sequences $\left\{ \phi_{1j} \right\}_j$ and $\left\{ \phi_{2j} \right\}_j$ will not be orthonormal. Let $f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$. The inner product in our space will be defined by

$$\left(f, \phi_j \right) = \left(f_1, \phi_{1j} \right) + \left(f_2, \phi_{2j} \right)$$

$$\begin{aligned} \|f\|_K &= \textstyle\sum_j \frac{\mu_j}{\mu_j^2+\lambda} \left\{ \left(f_1, \phi_{1j} \right) + \left(f_2, \phi_{2j} \right) \right\}^2 \\ &= \|f_1\|_{K_1}^2 + \|f_2\|_{K_2}^2 + 2 \left(f_1, f_2 \right)_{K_{12}} \end{aligned}$$

Second step estimator:

$\hat{\theta}$ minimizes $\left\| \bar{h} \right\|_K^2 = \left\| \begin{matrix} \bar{h}_1 \\ \bar{h}_2 \end{matrix} \right\|_K^2$. This minimization does not a closed-form solution, an optimization procedure needs to be used.

7.3. Monte Carlo experiment

Next we want to perform simulations to assess the small sample performance of our estimators. To be completed.

8. Conclusion

We showed how to apply GMM to construct efficient estimation based on the characteristic function. We illustrated our method on a model of duration of stay in a state. This type of models is frequently encountered in microeconometrics.

However, the use of the characteristic function is not restricted to a cross-section setting and has received recently an increasing interest in finance. While the likelihood of a asset pricing model is not easily tractable, its CCF has a closed-form solution and offers a way to estimate the parameters (Singleton, 1999). Moreover, a subordinated or stochastic volatility model can be interpreted as a mixture (Mandelbrot, 1973). We will consider the estimation of stochastic volatility models using GMM in another paper (Carrasco, Chernov, Florens, and Ghysels 2000).

9. Appendix

Let X be a random element (r.e.) defined on a complete probability space $(\Omega, \mathcal{F}, P_0)$ that takes its values in $(S, \mathcal{S})$. Let H be an Hilbert space with the inner product (.,.) that defines a norm $\| \, . \, \|$.

Assumption 1: The observed data $\{x^1, ..., x^n\}$ are independent realizations of the stochastic process X.

Assumption 2: Let h be a function on $S \times \Theta$ that takes its values in H where Θ is a compact subset of $\mathbb{R}^q$. h is a continuous function of θ.

Assumption 3: h is integrable with respect to F_{θ_0} for any θ and the equation

$$E^{\theta_0}(h(X, \theta)) = 0$$

has a unique solution θ_0 which is an interior point of Θ.

Assumption 4: $E^{\theta_0}(h(X, \theta)) \in \mathcal{H}(K) + \mathcal{H}(K)^{\perp}$ for any $\theta \in \Theta$.

Assumption 5: Let $N(K^{-1/2})$ denote the null space of $K^{-1/2}$, $N(K^{-1/2}) = \left\{ f \in H | K^{-1/2} f = 0 \right\}$. We assume that $E^{\theta_0}(h(X, \theta)) \in N(K^{-1/2})$ implies $E^{\theta_0}(h(X, \theta)) = 0$.

Assumption 6: $h(x, \theta)$ is differentiable with respect to $\theta = (\theta_1, ..., \theta_q)$ and $E^{\theta_0}\left(\frac{\partial h(X,\theta)}{\partial \theta_j}\right) = \frac{\partial}{\partial \theta_j} E^{\theta_0}\left(h\left(X, \theta\right)\right) \in \mathcal{D}(K^{-1})$ for any $\theta \in \Theta$.

Moreover the matrix $\left(K^{-1/2} E^{\theta_0}\left[\frac{\partial h}{\partial \theta'}(X, \theta)\right], K^{-1/2} E^{\theta_0}\left[\frac{\partial h}{\partial \theta'}(X, \theta)\right]\right)$ is positive definite and symmetric.

Assumption 7: The inner product satisfies the following differentiation rule

$$\frac{\partial}{\partial \theta'}(u(\theta), v(\theta)) = \left(\frac{\partial}{\partial \theta'} u(\theta), v(\theta)\right) + \left(u(\theta), \frac{\partial}{\partial \theta'} v(\theta)\right)$$

Assumption 8: $\sqrt{n} \bar{h}_n(\theta_0)$ converges in law to Y as n goes to infinity, where $Y \sim \mathcal{N}(0, K)$ in $L^2(\mu)$.

Assumption 9: The covariance kernel $k(t, s)$ is an L^2 kernel.

Assumption 10: $E \|h\|^4 < \infty$.

Assumption 11: $\left\| \bar{h}_n(\theta) - E^{\theta_0} h(\theta) \right\| = O_p\left(\frac{1}{\sqrt{n}}\right)$ uniformly in θ on Θ.

$\left\| \frac{\partial \bar{h}_n}{\partial \theta}(\theta) - E^{\theta_0} \frac{\partial h}{\partial \theta}(\theta) \right\| = O_p\left(\frac{1}{\sqrt{n}}\right)$ uniformly in θ on Θ.

10. References

Beard, R., S. Caudill, and D. Gropper (1991) “Finite Mixture Estimation of Multiproduct Cost Functions”, *Review of Economics and Statistics*, 654-664.

Carrasco, M., M. Chernov, J. P. Florens and E. Ghysels (2000), work in progress.

Carrasco, M. and J. P. Florens (1999) “Generalization of GMM to a continuum of moment conditions”, forthcoming in *Econometric Theory.*

Feuerverger, A. and P. McDunnough (1981) “On the Efficiency of Empirical Characteristic Function Procedures”, *J. R. Statist. Soc.* B, 43, No. 1, 20-27.

Feuerverger, A. and R. Mureika (1977) “The Empirical Characteristic Function and its Applications”, *The Annals of Statistics*, Vol. 5, No. 2, 88-97.

Groetsch, C. (1993) Inverse Problems in the Mathematical Sciences, Vieweg, Wiesbaden.

Heckman, J., R. Robb, and J. Walker (1990) “Testing the Mixture of Exponentials and Estimating the Mixing distribution by the Method of Moments”, Journal of the American Statistical Association, Vol. 85, 582-589.

Hochstadt, H. (1973) *Integral Equations.* Wiley and Sons.

Morozov, V.A. (1984) *Methods for Solving Incorrectly Posed Problems.* Springer Verlag, New York.

Lancaster, T. (1990) *The Econometric Analysis of Transition data*, Cambridge University Press, UK.

Mandelbrot, B. (1973) “Comments on: “A subordinated stochastic process model with finite variance for speculative prices” by Peter Clark”, *Econometrica*, Vol. 41, No. 2, 157-158.

Morduch, J. and H. Stern (1997) “Using mixture models to detect sex bias in health outcomes in Bangladesh”, *Journal of Econometrics*, 77, 259-276.

Mundlak, Y. and J. Yahav (1981) Random Effects, Fixed Effects, Convolution, and Separation”, Econometrica, Vol. 49, 1399-1416.

Parzen, E. (1970) “Statistical Inference on time series by RKHS methods”, *12th Biennial Seminar Canadian Mathematical Congress Proc.*, R. Pyke, ed., Canadian Mathematical Society, Montreal.

Singleton, K. (1999) “Estimation of Affine Pricing Models Using the Empirical Charactersitic Function”, mimeo, Stanford University.