Iterative Estimation for Conditional Estimating Equations

Weiyu Li[1,2] & Valentin Patilea[1]

[1] CREST-Ensai & IRMAR, Campus de Ker Lann, rue Blaise Pascal, BP 37203, 35172 Bruz cedex, France.
[2] Corresponding author. Email: weiyu.li@ensai.fr

Abstract

Many statistical and econometric models could be written under the form of conditional estimating equations, also called of conditional moment equations. In the classical approach for estimating parameters identified by such restrictions, one replaces the conditional moments by a sufficiently rich finite set of unconditional moments and applies the generalized method of moments (GMM). However, the GMM approach does not guarantee consistency since the parameters are not necessarily identified by a finite set of marginal moments. Motivated by this aspect, several recent articles proposed alternative approaches that preserve consistency. Herein we consider an estimation approach for conditional estimating equations that is called smooth minimum distance (SMD) and is based on the optimization of a nonlinear contrast. We introduce an iterative version of SMD based on a quadratic approximation of the contrast. At any step of the iteration, the estimate has an explicit form and therefore the new method could be easily implemented. We present en extensive empirical study of the new method. In particular we compare it with classical methods (least squares, maximum likelihood, GMM).

Keywords: conditional moment equations, Newton-Kantorovich method, quadratic forms

1 Introduction

Many statistical en econometric models could be written under the form

$$\mathbb{E}[g(Y, X; \theta) \mid X] = 0 \quad p.s. \quad \Leftrightarrow \quad \theta = \theta_0, \tag{1}$$

where $Y \in \mathbb{R}^d$, $X \in \mathbb{R}^q$, g is a given function, $\theta \in \Theta \subset \mathbb{R}^p$ is the parameter of the model and θ_0 is the 'true' unknown value of the parameter that corresponds to the data generating process. See, for instance, Kitamura *et al.* (2004) and Lavergne & Patilea (2008) for examples and a revue of the literature.

Suppose that the observations $(Y_1^\top, X_1^\top)^\top, \cdots, (Y_n^\top, X_n^\top)^\top$ represent an independent sample from the random vector $(Y^\top, X^\top)^\top$; (here and in the following $A^\top$ denotes the transposed of a matrix A). Let

$$g_i(\theta) = g(Y_i, X_i; \theta), \quad 1 \le i \le n, \qquad \theta \in \Theta.$$

Let K be a symmetric function defined on $\mathbb{R}^q$ such that its Fourier Transform $\mathcal{F}[K]$ is strictly positive and let

$$K_{ij} = K(X_i - X_j).$$

Following the idea of Lavergne & Patilea (2008), one could estimate the parameter θ_0 by

$$\widetilde{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta), \tag{2}$$

where

$$Q_n(\theta) = \sum_{1\leq i,j\leq n} g_i(\theta) g_j(\theta) K_{ij}.$$

This estimation idea is related with a presmoothing idea in parametric inference, see Cristobal Cristobal *et al.* (1987).

If model (1) is correct, for any n

$$\mathbb{E}[Q_n(\theta)] \geq 0 \qquad \text{et} \qquad \mathbb{E}[Q_n(\theta)] = 0 \quad \text{if and only if} \quad \theta = \theta_0.$$

This property, combined with mild technical conditions, guarantees the consistence of the estimator $\widetilde{\theta}_n$. Lavergne & Patilea (2008) derived the asymptotic behavior of this estimator under general conditions. Herein we propose an iterative version of their estimator that avoids nonlinear optimization.

2 An iterative approach

Let

$$\nabla_\theta g_i(\theta) = \frac{\partial g}{\partial \theta}(Y_i, X_i; \theta) \in \mathbb{R}^p.$$

For θ close to θ' we can write $g_i(\theta) \approx g_i(\theta') + \nabla_\theta g_i(\theta')^\top (\theta - \theta')$ and define

$$Q_n(\theta, \theta') = \sum_{1\leq i,j\leq n} \Big[g_i(\theta') + \nabla_\theta g_i(\theta')^\top (\theta - \theta') \Big] \Big[g_j(\theta') + \nabla_\theta g_j(\theta')^\top (\theta - \theta') \Big] K_{ij}.$$

Note that for a fixed θ' the quantity $Q_n(\theta, \theta')$ is a positive semi-definite quadratic form. Then a simple idea would be to consider the iterations

$$\theta_n^{(k)} = \arg\min_\theta Q_n(\theta, \theta^{(k-1)}), \qquad k = 1, 2, \cdots,$$

where $\theta_n^{(0)}$ is some initial value. This leads us to consider the following iterations

$$\begin{aligned} \theta_n^{(k)} \;=\; \theta_n^{(k-1)} - & \left[\alpha_n I_p + \sum_{1\leq i,j\leq n} \nabla_\theta g_i(\theta_n^{(k-1)}) \nabla_\theta g_j(\theta_n^{(k-1)})^\top K_{ij} \right]^{-1} \\ & \times \left[\sum_{1\leq i,j\leq n} \nabla_\theta g_i(\theta_n^{(k-1)}) g_j(\theta_n^{(k-1)}) K_{ij} \right], \quad k = 1, 2, \cdots, \end{aligned}$$

where $\alpha_n > 0$ is a regularization parameter that avoids the inversion of ill-conditioned matrices. The estimator we propose is $\widehat{\theta}_n = \theta_n^{(k^\star)}$ for some value $k^\star$ obtained from a stopping rule for the iterations.

Let us note that our iterative method could be interpreted as a Newton-Kantorovich method for solving a nonlinear equation.

3 Empirical evidence

In this section we present some simulation experiments that we performed to study the small sample size properties of our estimator.

3.1 Endogeneity in nonlinear regressions

Consider the model

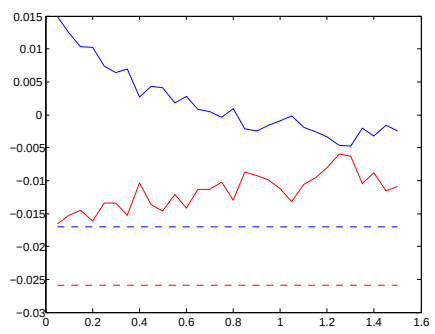
$$Y = (X^\top \theta)^2 + U \tag{3}$$

where

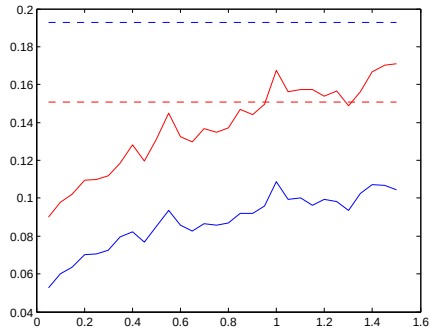
$$\begin{bmatrix} U \\ V \\ Z \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)$$

and $X = (X_1, X_2)^\top$ with $X_1 = Z + V/2$ and $X_2 = Z^2 + V$. Hence, $\mathbb{E}[Y - (X^\top\theta)^2 \mid X] \neq 0$ but $\mathbb{E}[Y - (X^\top\theta)^2 \mid Z] = 0$ a.s. We separately simulate 1000 samples of size $n = 20$ of independent realizations of Y and X using $\theta_0 = (0.5, 0.9)^\top$. We compare our new estimation method with the classical GMM based on the instrumental variables Z and Z^2. In Figure 1 we provide the bias and the standard deviation of the estimates obtained using the two methods. For our method, the bandwidth h is chosen by a cross-validation method.

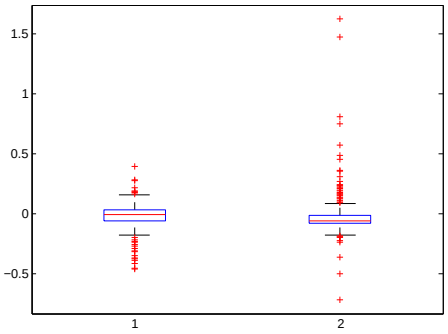
The simulation results reveal a good performance of our new method. The bias are low compared with the classical method. The results are quite stable with respect to the bandwidth. Similar conclusions were obtained with other small sample sizes, like $n = 10$ and $n = 30$.



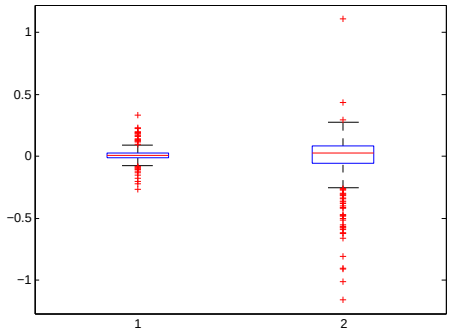
(a) Bias



(b) Standard deviation

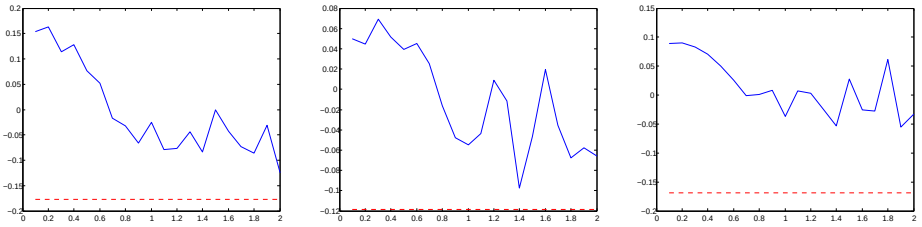


(c) Box-plot of the bias of $\hat{\theta}_1$



(d) Box-plot of the bias of $\hat{\theta}_2$

Figure 1: Results for Model (3) when sample size $n = 20$: in (a) and (b) the solid lines are the result of our iterative method for different bandwidths h and the dashed lines are the result of GMM, the red lines represent θ_1 and the blue lines represent θ_2; in (c)and (d) the first box-plot is the result of our method and the second one is the result of GMM



(a) Bias of $\hat{\theta}_1$ (b) Bias of $\hat{\theta}_2$ (c) Bias of $\hat{\theta}_3$

Figure 2: Bias of parameters estimates in Model (4) when sample size $n = 20$: solid lines for the results of our method, dashed lines for the results of MLE.

3.2 Logistic regression

In this example we consider the popular logistic regression with i.i.d. data $(Y, X^\top)^\top$. Let

$$Y \sim Logistic(\mu), \tag{4}$$

where $\mu = X^\top\theta$ and $\theta_0 = (1.2, 0.6, 0.8)^\top$. Here X has a multivariate normal distribution

$$N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{bmatrix}\right).$$

We compare our approach with the most popular estimation method, that is the maximum likelihood. The conditional moment equations we use for our method are the score equations, that is we take

$$g(Y, X, \theta) = \left[Y - \frac{\exp(X^\top\theta)}{1 + \exp(X^\top\theta)}\right] X.$$

Here the data-driven selection of h is done by minimization of the criterion used for estimation, that is the bandwidth is solution of the problem

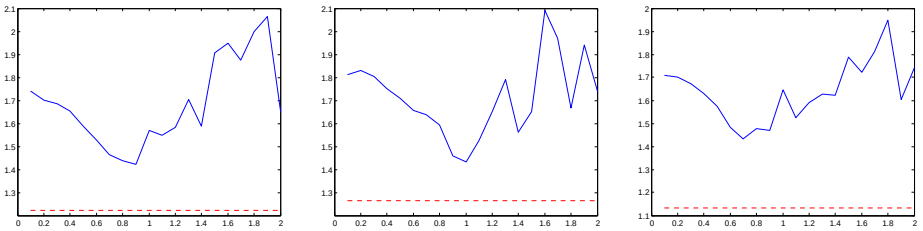
$$\min_h \min_\theta \sum_{1\le i,j\le n} g_i(\theta)^\top g_j(\theta) K_{i,j}.$$

To avoid the inversion of ill-conditioned matrices, a regularization parameter α_n was used and its value was set $\alpha_n = 0.01$.

Several sample sizes were considered, we only report the case $n = 20$. The experiment was repeated 1000 times. In Figure 2 we report the biases for our method with different bandwidths and the bias of the maximum likelihood estimator (MLE). For most of the values h, our estimates are less biased. In Figure 3 we present the corresponding standard deviations. Overall, MLE perform better, but for many bandwidths our estimators have less than 25% extra variability. In Figure 4 we present the result obtained with the data-driven bandwidth rule. The box-plots reveal that our estimators behave well, at least as well as the MLE.

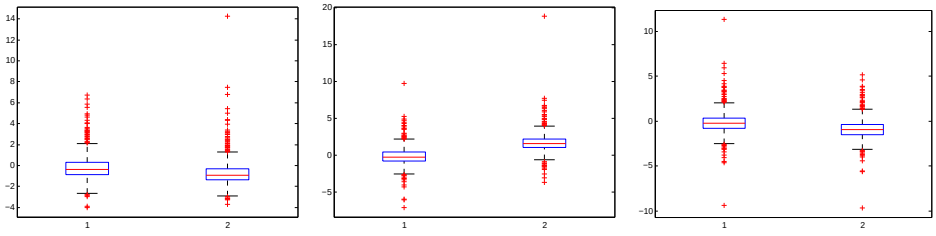
References

Cristobal Cristobal, J.A., Faraldo Roca, P. and Gonzalez Manteiga, W. (1987), "A Class of Linear Regression Parameter Estimators Constructed by Nonparametric Estimation", *Ann. Statist.* 15(2), 603–609.



(a) Standard deviation of $\hat{\theta}_1$ (b) Standard deviation of $\hat{\theta}_2$ (c) Standard deviation of $\hat{\theta}_3$

Figure 3: Standard deviation of Model (4) when sample size $n = 20$: solid lines for the results of our method, dashed lines for the result of MLE.



(a) Box-plot of the bias of $\hat{\theta}_1$ (b) Box-plot of the bias of $\hat{\theta}_2$ (c) Box-plot of the bias of $\hat{\theta}_3$

Figure 4: Box-plot of parameters estimates in Model (4) with $n = 20$: the first box-plot is the result of our method and the second one is the result of MLE.

Kantorovich, L.V. and Akilov, G.P. (1964), *Functional Analysis in Normed Spaces*, Pergamon Press.

Kitamura, Y., Tripathi, G. and Ahn, H. (2004), "Empirical likelihood-based inference in conditional moment restriction models", *Econometrica* 72(6), 1667–1714.

Lavergne, P. and Patilea, V. (2008), "Smooth Minimum Distance Estimation and Testing with Conditional Estimating Equations: Uniform in Bandwidth Theory", Working Paper, Simon Fraser University.