Journal of Econometrics 186 (2015) 465–476

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Instrumental variable estimation in functional linear models[☆]

Jean-Pierre Florens[a], Sébastien Van Bellegem[b,*]

[a] Toulouse School of Economics, France
[b] Center for Operations Research and Econometrics, Belgium

ARTICLE INFO

Article history:
Available online 3 March 2015

JEL classification:
C26
C14

Keywords:
High dimensional model
Penalized least squares
Instrumental variable
Functional data
Fertility rate
Growth

ABSTRACT

In an increasing number of empirical studies, the dimensionality measured e.g. as the size of the parameter space of interest, can be very large. Two instances of large dimensional models are the linear regression with a large number of covariates and the estimation of a regression function with many instrumental variables. An appropriate setting to analyze high dimensional problems is provided by a functional linear model, in which the covariates belong to Hilbert spaces. This paper considers the case where covariates are endogenous and assumes the existence of instrumental variables (that are functional as well). The paper shows that estimating the regression function is a linear ill-posed inverse problem, with a known but data-dependent operator. The first contribution is to analyze the rate of convergence of the penalized least squares estimator. Based on the result, we discuss the notion of "instrument strength" in the high dimensional setting. We also discuss a generalized version of the estimator, when the problem is premultiplied by an instrument-dependent operator. This extends the technology of Generalized Method of Moments to high dimensional, functional data. A central limit theorem is also established on the inner product of the estimator. The studied estimators are easy and fast to implement, and the finite-sample performance is discussed through simulations and an application to the impact of age-specific fertility rate curves on yearly economic growth in the United Kingdom.

1. Introduction

In an increasing number of economic studies, the dimensionality of the statistical model, measured as the size of the parameter space of interest, can be very large. One canonical example is provided by the multiple linear regression model

$$y = z'\varphi + \varepsilon \tag{1.1}$$

where y is a real random variable, z is a vector of p covariates and ε is an exogenous error term. To reflect that the number of covariates p is large, asymptotic inference is usually derived under the condition that p increases with the sample size (e.g. Fan et al., 2008; Belloni and Chernozhukov, 2009; Belloni et al., 2011). Standard estimators such as Generalized Least Squares (GLS) thus need to be modified because of the unboundedness of the parameter space. The present paper analyzes linear models with large or infinite dimension under endogeneity in the presence of instrumental variables. Since the dimension of the endogenous variables is large, so is the number of instrumental variables. Our study is therefore tightly connected to the recent studies on estimation in the presence of "many instrumental variables" (Morimune, 1983; Carrasco and Florens, 2000; Chao and Swanson, 2005; Hansen et al., 2008; Swanson et al., 2011). In these studies, the dimension of the model is kept fixed (not depending on the sample size) while the dimension of the instrumental vector can be very large or depends on the sample size. Our study below considers the situation where both the dimension of the covariates and the instruments are large.

Estimation with many instrumental variable has a long history in econometrics, see e.g. the above references and the references therein. It is recognized that the empirical performance of inference can be imprecise in this context, because of the presence of weak instruments (see e.g. the survey paper by Stock et al., 2002). This issue naturally arises in the high dimensional setting as well. This study provides a new characterization of strong

☆ We thank Rytis Bagdziunas, Éric Gauthier, Frédérique Fève, Maik Schwarz and Anna Simoni for helpful comments on a preliminary version of this work. Various parts of this work have been presented during the workshop in nonparametric econometrics in Banff (2009), the CIREQ workshop on high dimensional econometrics in Montréal (2012), the Join Statistical Meeting in San Diego (2012), the LATAM workshop in Sao Paulo (2013) and during seminars at IHS (Wien), ECORE (Brussels) and the University of Luxemburg. Comments from the participants of those meetings were greatly appreciated.

* Correspondence to: Voie du Roman Pays, 34, L1.03.01, B-1348 Louvain-la-Neuve, Belgium.
E-mail address: sebastien.vanbellegem@uclouvain.be (S. Van Bellegem).

http://dx.doi.org/10.1016/j.jeconom.2015.02.020
0304-4076/

or weak instruments in the statistical model. The strength of an instrumental vector results from an arbitrage among two indices of the statistical model. The first index (denoted β below) is a measure of the regularity of the unknown parameter (which is infinite dimensional) *with respect to* the eigenvalues of the covariance between the endogenous parameters and the instruments. The second index (denoted γ) can be understood as a measure of the signal-to-noise ratio, that is a ratio between the later eigenvalues onto the eigenvalues of the covariance between the instruments and the error term in the linear model. Strong instruments are such that the regularity of the target parameter (β) is large enough compared to the signal-to-noise ratio (γ). Below we show that, under this condition, the penalized least squares estimator of the parameter has the optimal minimax rate of convergence, that has been derived in Chen and Reiss (2011) in a similar context. On the contrary if the instruments are weak, the optimal rate of convergence is not reached.

Because we are interested in the asymptotic behavior of the estimator, this paper explicitly relates the sampling model to a population model that is infinite dimensional. A convenient framework to make the analysis is to rewrite and generalize model (1.1) using an inner product, that is

$$y = \langle z, \varphi\rangle_{\mathcal{F}} + \varepsilon \tag{1.2}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ stands for the inner product between two elements of a Hilbert space $\mathcal{F}$. This model contains the finite dimensional linear model as a particular case if we consider $\mathcal{F} = \mathbb{R}^K$ with a fixed, finite dimension K. However it allows to accommodate high dimensional parameters, for instance when $\mathcal{F}$ is the space of squared summable sequences, ℓ^2.

The statistical model (1.2) is also appropriate when covariates are not random vectors but random *curves* as it is the case in functional regression. In such a case the Hilbert space $\mathcal{F}$ is for instance the space of real, square integrable functions over [0, 1], denoted $L^2[0, 1]$, equipped with the inner product defined as

$$\langle\varphi_1, \varphi_2\rangle_{L^2[0,1]} = \int_0^1 \varphi_1(t)\varphi_2(t)dt.$$

Examples of regression with functional covariates can be found in various studies (e.g. Kunitomo, 1980; Ramsay and Silvermann, 2005; Hall and Horowitz, 2007; Ferraty and Vieu, 2006; Cardot and Johannes, 2010; Kokoszka and Horvath, 2012). We provide below an application to growth theory, in which the dependent variable is the annual growth rate per capita of United Kingdom and the independent variable is the annual fertility rate.

Our study starts from the general model (1.2) in which z is endogenous and it supposes that we observe instruments w allowing to identify the parameter φ. The study of inference shows that the problem of estimating φ in this statistical model is an ill-posed inverse problem, and inference thus relies on regularization techniques such as the penalized least squares estimation. Two recent surveys on regularization for ill-posed inverse problems in econometrics on which this paper relies on are Carrasco et al. (2007) and Johannes et al. (2011).

The organization of the paper is as follows. In Section 2 the asymptotic properties of the penalized least-square estimator is studied and the notion of instrument strength is discussed. The estimator can be seen as an extension of the parametric instrumental variable estimator to the high dimensional setting. The two key indices (β and γ) are defined. Section 3 studies the numerical and finite sample properties of the estimator. It shows that how the estimator is computed in practice, using a linear algorithm. This section also contains two discussions on the penalty. The first discussion shows how this penalty can be computed in a data-driven way and the second discussion shows how it is possible to impose smoothness priors such as the existence of derivatives on the solution.

Analogously to the finite-dimensional theory of the efficient estimation by the Generalized Method of Moments (GMM), Section 4 questions the efficiency of the estimator. A generalized version of the estimator is studied, when the model is premultiplied by a data-driven operator. It is proved that for a certain class of operators and under some restrictions on the statistical model, the premultiplication does not affect the rate of convergence of the estimator. A Central Limit Theorem is then derived for the inner product $\langle\hat\varphi, g\rangle_{\mathcal{F}}$, if $\hat\varphi$ denoted the generalized penalized least squares estimator and g is a test function in $\mathcal{F}$. The Theorem also provides lower and upper bounds for the asymptotic variance.

Section 5 evaluates the performance of the estimators in an application in growth theory. Recent developments in economics have studied the impact of fertility on growth rate. The dependent variable is the age-specific fertility rate that is the density of birth with respect to mother's age. It is a continuous variable by definition, observed yearly and the framework of functional regression allows to study the marginal impact of the age of fertility onto the growth rate per capita. An Appendix collects the technical proofs of results that are stated in the paper.

2. The high dimensional model and the regularized estimator

2.1. Definition of the estimator

We start by defining the statistical model in the two following assumptions.[1]

Assumption 1. y is a real random variable, z is a random vector in a Hilbert space $\mathcal{F}$ and w is a random vector in a Hilbert space $\mathcal{W}$ such that there exists a unique $\varphi \in \mathcal{F}$ that satisfies $\mathbb{E}(\{y - \langle z, \varphi\rangle_{\mathcal{F}}\}w) = 0$. If we set $u = y - \langle z, \varphi\rangle$, we also denote by Ω the covariance operator of uw, that is $\Omega = \mathbb{E}(u^2\langle w, \bullet\rangle_{\mathcal{W}} w)$.

Assumption 2. We observe (y_i, z_i, w_i) for $i = 1, \ldots, n$ that are independent and with the same distribution as (y, z, w).

In order to derive the normal equations and to find a nonparametric estimator for $\varphi \in \mathcal{F}$ it is useful to translate the above assumptions in a notation with operators. The above assumptions imply that the unknown function $\varphi \in \mathcal{F}$ satisfies:

$$Y_n = K_n\varphi + U_n \tag{2.1}$$

where $U_n = (u_1, \ldots, u_n)'$, $Y_n = (y_1, \ldots, y_n)'$ and $K_n : \mathcal{F} \to \mathbb{R}^n$ is a linear operator defined by $K_n\varphi = (\langle z_1, \varphi\rangle_{\mathcal{F}}, \ldots, \langle z_n, \varphi\rangle_{\mathcal{F}})'$. By definition, the inner product in $\mathbb{R}^n$ is set to

$$\langle a, b\rangle_{\mathbb{R}^n} = \frac{1}{n}\sum_{i=1}^n a_i b_i$$

for any $a, b \in \mathbb{R}^n$.

We also introduce the operator $W_n^\star : \mathbb{R}^n \to \mathcal{W} : \theta \to n^{-1}\sum_{i=1}^n \theta_i w_i$, where $W_n^\star$ stands for the adjoint of W_n. Applied to model (2.1), we get

$$W_n^\star Y_n = W_n^\star K_n\varphi + W_n^\star U_n$$

[1] The assumption includes an identification condition. A primitive condition is found if we consider the covariance operator $\Sigma_{ZW} = \mathbb{E}\{\langle Z, \cdot\rangle_{\mathcal{F}} W\}$. Identification of φ follows if the kernel of this covariance operator is the null set. This can be seen as the infinite dimensional extension of the common rank condition on the covariance matrices appearing in the identification theory of finite dimensional linear models.