Journal of Econometrics 195 (2016) 104–119

Contents lists available at ScienceDirect

Journal of Econometrics

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

# Econometric estimation with high-dimensional moment equalities

Zhentao Shi *

*Department of Economics, the Chinese University of Hong Kong, Hong Kong Special Administrative Region*

A R T I C L E I N F O

*Article history:*
Received 4 September 2014
Received in revised form
18 July 2016
Accepted 19 July 2016
Available online 11 August 2016

*JEL classification:*
C13
C55
C61

*Keywords:*
Bias correction
Empirical likelihood
Greedy algorithm
High-dimensional estimation

A B S T R A C T

We consider a nonlinear structural model in which the number of moments is not limited by the sample size. The econometric problem here is to estimate and perform inference on a finite-dimensional parameter. To handle the high dimensionality, we must systematically choose a set of informative moments; in other words, delete the uninformative ones. In nonlinear models, a consistent estimator is a prerequisite for moment selection. We develop in this paper a novel two-step procedure. The first step achieves consistency in high-dimensional asymptotics by relaxing the moment constraints of empirical likelihood. Given the consistent estimator, in the second step we propose a computationally efficient algorithm to select the informative moments from a vast number of candidate combinations, and then use these moments to correct the bias of the first-step estimator. We show that the resulting second-step estimator is $\sqrt{n}$-asymptotic normal, and achieves the lowest variance under a sparsity condition. To the best of our knowledge, we provide the first asymptotically normally distributed estimator in such an environment. The new estimator is shown to have favorable finite sample properties in simulations, and it is applied to estimate an international trade model with massive Chinese datasets.



## 1. Introduction

The unprecedented accessibility of large micro-datasets at the individual agent-level opens up opportunities to investigate a multitude of new economic problems as well as old problems from new perspectives. Respecting a long tradition of building parsimonious, often nonlinear, structural models to characterize complex economic phenomena, empirical econometric analysis typically concentrates on a few key parameters that bear economic meaning. Structural models involving moment conditions are widely used in many areas of econometrics. While moment restrictions arise naturally from the economic context, theory usually gives little guidance about the choice of the moment restrictions. In modern empirical work it is nothing extraordinary to estimate with hundreds or even thousands of moments. The fundamental challenge in such empirical work is to develop a theory of estimation and inference that allows for high-dimensionality in the moment conditions relative to the sample size.

Several latest empirical applications take advantage of large datasets and large models. For instance, Altonji et al. (2013) examine a joint model of earnings, employment, job changes, wage rates, and work hours over a career with a full specification of 2429 moments. Eaton et al. (2011) explore the sales of French manufacturing firms in 113 destination countries with 1360 moments. Han et al. (2005) investigate the cost efficiency of the Spanish saving banks in a time-varying coefficient model with 872 moments. All these empirical examples estimate a finite-dimensional parameter of interest in structural models with many nonlinear moments. The underlying models take the following form. A "true" parameter $\beta_0$ satisfies the unconditional moment restriction

$$\mathbb{E}\left[g\left(Z_i,\beta\right)\right]=0_m,\quad \text{for } i=1,\ldots,n,$$

where $\{Z_i\}_{i=1}^n$ is the observed data, $\beta_0 \in \mathcal{B} \subset \mathbb{R}^D$ is finite-dimensional, $g$ is an $\mathbb{R}^m$-valued moment function, and $0_m$ is an $m \times 1$ vector of zeros.

It is known from the literature that the relative magnitude of $m$ and $n$ shapes the asymptotic properties of generalized method of moments (GMM) and empirical likelihood (EL). Under the usual assumptions, consistency demands $m = o(n)$ and $\sqrt{n}$-asymptotic normality demands $m^3 = o(n)$ (Koenker and Machado, 1999; Donald et al., 2003). Though the sample sizes in these cited empirical examples can run to thousands or even millions, valid asymptotic statistical inference may require many more observations still.

In this paper, we consider a nonlinear structural model in which $m$ can be much larger than $n$, and the econometric problem is to estimate and perform inference on a finite-dimensional

* Correspondence to: 912 Esther Lee Building, the Chinese University of Hong Kong, Shatin, New Territories, Hong Kong Special Administrative Region. Fax: +852 2603 5805.

*E-mail address:* zhentao.shi@cuhk.edu.hk.

http://dx.doi.org/10.1016/j.jeconom.2016.07.004
0304-4076/

parameter of interest. The phrase *high-dimensional* in the title refers specifically to the $m > n$ case.

Han and Phillips (2006) explain the challenge when $m$ is large in the simple equally-weighted GMM (EW-GMM). They find that the consistency of EW-GMM depends on the relative strength of main signal against signal variability. The former contains information about the true parameter in population, while the latter arises from the sampling error of moment functions. Despite a small contribution from each moment component, the signal variability induced by a large number of moments accumulates in the quadratic-form of the GMM criterion and can overwhelm the main signal. Their work was followed by Newey and Windmeijer (2009), who provide a new variance estimator of the generalized empirical likelihood (GEL), which is consistent under many weak moments. These two papers do not consider the estimation with high-dimensional moments.

Bai and Ng (2009, 2010) and Belloni et al. (2012) investigate the estimation of the high-dimensional linear instrumental variable (IV) model. In their setting, a large number of IVs all meet the orthogonality condition (zero correlation with the structural error), while only a small handful of the IVs satisfy the relevance condition (non-zero correlation with the endogenous regressors), but the identities of the relevant IVs are unknown. Belloni et al. (2012) use Lasso as an IV selector, and then plug in the predicted endogenous regressors or the post-Lasso two-stage least squares (2SLS) to efficiently estimate the structural parameter. Bai and Ng (2009) utilize boosting to select IVs in the first stage reduced-forms, and Bai and Ng (2010) develop a factor IV estimator to achieve optimality.

The relevant moments for $\beta$ is characterized by $\mathbb{E}\left[\frac{\partial}{\partial\beta}g_j(Z_i, \beta_0)\right]$,[1] the derivative of the moment function at $\beta_0$. The derivative, as a local property, explicitly depends on $\beta_0$ in nonlinear models, but is independent of $\beta$ in the linear IV model. Therefore, in nonlinear models we cannot apply these methods that select relevant IVs in the first stage of 2SLS. In the meantime, we also assume that the identities of the relevant, or informative, moments are unknown. As a result, we cannot randomly pick a small subset of moments, as those moments may weakly or completely fail to identify the parameter. We must choose a set of informative moments in a systematic manner.

We address the estimation problem via a novel two-step procedure adapted to nonlinear models. As we have explained, it is necessary to locate the true $\beta_0$ before evaluating which moment is informative. We propose an estimator called *relaxed empirical likelihood* (REL), which is consistent in high-dimensional asymptotics. Consistency, as a global property, demands the involvement of all moments. REL is constructed to tolerate a slight violation of the equality constraints in the standard EL. The magnitude of tolerance is specified by the user as a tuning parameter, which controls the maximal sampling error of the moments. Under regularity conditions, we show that REL achieves nearly-optimal rate of convergence to the true parameter.

Given REL's consistency, we can then proceed to the second step to select informative moments. With too many possible combinations, the effectiveness of moment selection hinges on computational feasibility. We propose a *boosting-type greedy algorithm* that forms an increasing sequence of selected moments. In each iteration, the algorithm adds only one moment, namely the one that maximizes an information criterion given the moments selected in the preceding iterations. With the selected moments, we can further correct REL's bias incurred in the relaxation. We call the second-step estimator the *bias-corrected REL* (BC-REL).

In high-dimensional statistics, sparsity is often critical for theoretical development. In our context, sparsity means that the number of informative moments is much smaller than the sample size. Without sparsity, BC-REL is $\sqrt{n}$-asymptotic normal after variance standardization; with sparsity, BC-REL obtains the smallest asymptotic variance. To the best of our knowledge, this paper is the first to establish asymptotic normality in high-dimensional nonlinear models.

Our methodology follows a vast literature. The use of many linear IVs originated in Angrist (1990). It motivated intensive econometric research, for example Hahn (2002), Chao and Swanson (2005) and Chao et al. (2011), to name a few. Bai and Ng (2010) and Belloni et al. (2012) resolve the problem of $m > n$ in the linear IV setting. Gautier and Tsybakov (2013) propose a new IV estimator based on the Dantzig selector (Candes and Tao, 2007) in a more general setting that allows for a high-dimensional parameter in the linear structural equation; their focus is different from the other two papers. In another line of research, Carrasco and Florens (2000, 2014) develop GMM theory for many moments or a continuum of moments with $g$ in a Hilbert space. The Hilbert space setting can be restrictive when the moments are generated from detailed observations in large datasets, for example a large number of mutually orthogonal non-degenerate IVs.

Our first step estimation REL is built on EL (Owen, 1988; Qin and Lawless, 1994; Kitamura, 1997). As an information-theoretic alternative to GMM, Kitamura (2001), Kitamura and Stutzer (1997), Newey and Smith (2003) and Otsu (2010) find theoretical advantages of EL. Latest developments of EL to cope with an infinite-dimensional parameter include Otsu (2007) and Lahiri and Mukhopadhyay (2012).

In terms of estimation in high-dimensional models, Caner (2009) and Caner and Zhang (2014) introduce a Lasso-type penalty to GMM for variable selection. Belloni et al. (2010, 2011), Belloni and Chernozhukov (2011), and Belloni et al. (2011) contribute to various aspects of estimation and inference in high-dimensional econometric and statistical models. Fan and Liao (2014) propose the *penalized focused generalized method of moments* in which the sparsity of the high-dimensional parameter is directly associated with the moments.

Regarding the selection step, Andrews and Lu (2001) propose several information criteria in GMM, and Hong et al. (2003) give the EL counterpart. Breusch et al. (1999) discuss the problem of redundant moments, and Hall and Peixe (2003) and Hall et al. (2007) develop a selection criterion. Liao (2013) and Cheng and Liao (2015) use a Lasso-type penalty to remove, in one step, possibly misspecified as well as redundant moments. During the revision of this manuscript, Luo (2014) proposes an alternative selection method via Lasso, which obtains the same asymptotic normality under similar assumptions.

Through this paper, we adopt the following notations. $\|\cdot\|_\infty$ is the sup-norm – the largest element in absolute value – of a matrix. $\|\cdot\|_2$ and $\|\cdot\|_1$ are the vector $l_2$-norm and $l_1$-norm, respectively. $\|\cdot\|_0$ is the cardinality of a set. $C \in (1, \infty)$ represents a fixed finite constant independent of the sample size. $\phi_{\min}(\cdot)$ and $\phi_{\max}(\cdot)$ are the minimal and maximal eigenvalues, respectively. $\mathcal{M} := \{1, \ldots, m\}$ is the index set of all moments, and $\mathcal{D} := \{1, \ldots, D\}$ is the index set of all components of the parameter $\beta$. $I_n$ is an $n \times n$ identity matrix. $(\cdot)^-$ is the Moore–Penrose pseudo-inverse.

The rest of the paper is organized as follows. Section 2 introduces the idea of REL, derives its consistency and rate of convergence. Section 3 constructs BC-REL and establishes its asymptotic normality based on the boosting-type greedy algorithm. Section 4 provides a simulation example, and Section 5

[1] Indeed the criterion function for a combination of moments is the Fisher information that we will introduce in Section 3. Here we temporarily use $\mathbb{E}\left[\frac{\partial}{\partial\beta}g_j(Z_i, \beta_0)\right]$ to keep the discussion non-technical.