Journal of Econometrics 180 (2014) 158–173

Contents lists available at ScienceDirect

Journal of Econometrics

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

Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter

Joel L. Horowitz *

Department of Economics, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA

ARTICLE INFO

Article history:
Received 19 September 2012
Received in revised form
3 October 2013
Accepted 24 March 2014
Available online 29 March 2014

JEL classification:
C13
C14
C21

Keywords:
Ill-posed inverse problem
Regularization
Series estimation
Nonparametric estimation

ABSTRACT

In nonparametric instrumental variables estimation, the mapping that identifies the function of interest, g, is discontinuous and must be regularized to permit consistent estimation. The optimal regularization parameter depends on population characteristics that are unknown in applications. This paper presents a theoretically justified empirical method for choosing the regularization parameter in series estimation. The method adapts to the unknown smoothness of g and other unknown functions. The resulting estimator of g converges at least as fast as the optimal rate multiplied by $(\log n)^{1/2}$. The asymptotic integrated mean-square error (AIMSE) of the estimator is within a specified factor of the optimal AIMSE.

1. Introduction

This paper is about estimating the unknown function g in the model

$$Y = g(X) + U; \qquad E(U|W = w) = 0 \tag{1.1}$$

for almost every w or, equivalently,

$$E[Y - g(X)|W = w] = 0 \tag{1.2}$$

for almost every w. In this model, g is a function that satisfies regularity conditions but is otherwise unknown, Y is a scalar dependent variable, X is a continuously distributed explanatory variable that may be correlated with U (that is, X may be endogenous), W is a continuously distributed instrument for X, and U is an unobserved random variable. The data are an independent random sample of (Y, X, W). The paper presents a theoretically justified, empirical method for choosing the regularization parameter that is needed for estimation of g.

Existing nonparametric estimators of g in (1.1)–(1.2) can be divided into two main classes: sieve (or series) estimators and kernel estimators. Sieve estimators have been developed by Ai and Chen (2003), Newey and Powell (2003), Blundell et al. (2007), and Horowitz (2012). Kernel estimators have been developed by Hall and Horowitz (2005) and Darolles et al. (2011). Florens and Simoni (2010) describe a quasi-Bayesian estimator based on kernels. Hall and Horowitz (2005) and Chen and Reiss (2011) found the optimal rate of convergence of an estimator of g. Horowitz (2007) gave conditions for asymptotic normality of the estimator of Hall and Horowitz (2005). Horowitz and Lee (2012) showed how to use the sieve estimator of Horowitz (2012) to construct uniform confidence bands for g. Newey et al. (1999) present a control function approach to estimating g in a model that is different from (1.1)–(1.2) but allows endogeneity of X and achieves identification through an instrument. The control function model is non-nested with (1.1)–(1.2) and is not discussed further in this paper. Chernozhukov et al. (2007), Horowitz and Lee (2007), and Gagliardini and Scaillet (2012) have developed methods for estimating a quantile-regression version of model (1.1)–(1.2). Chen and Pouzo (2009, 2012) developed a method for estimating a large class of nonparametric and semiparametric conditional moment models with possibly non-smooth moments. This class includes the quantile-regression version of (1.1)–(1.2).

As is explained further in Section 2 of this paper, the relation that identifies g in (1.1)–(1.2) creates an ill-posed inverse problem. That is, the mapping from the population distribution of (Y, X, W) to g is discontinuous. Consequently, g cannot be estimated consistently by replacing unknown population quantities in the identifying relation with consistent estimators. To achieve a consistent

* Tel.: +847 491 8253; fax: +847 491 7001.
E-mail address: joel-horowitz@northwestern.edu.

http://dx.doi.org/10.1016/j.jeconom.2014.03.006
0304-4076/

estimator it is necessary to regularize (or modify) the mapping that identifies g. The amount of modification is controlled by a parameter called the regularization parameter. The optimal value of the regularization parameter depends on unknown population characteristics and, therefore, cannot be calculated in applications. Although there have been proposals of informal rules-of-thumb for choosing the regularization parameter in applications, theoretically justified empirical methods are not yet available.

This paper presents an empirical method for choosing the regularization parameter in sieve or series estimation, where the regularization parameter is the number of terms in the series approximation to g. The method consists of optimizing a sample analog of a weighted version of the integrated mean-square error of a series estimator of g. The method does not require *a priori* knowledge of the smoothness of g or of other unknown functions. It adapts to their unknown smoothness. The estimator of g based on the empirically selected regularization parameter also adapts to unknown smoothness. It converges in probability at a rate that is at least as fast as the asymptotically optimal rate multiplied by $(\log n)^{1/2}$, where n is the sample size. Moreover, its asymptotic integrated mean-square error (AIMSE) is within a specified factor of the optimal AIMSE. The paper does not address the question of whether the factor of $(\log n)^{1/2}$ can be removed or is an unavoidable price that must be paid for adaptation. This question is left for future research.

Section 2 provides background on the estimation problem and the series estimator that is used with the adaptive estimation procedure. This section also reviews the relevant mathematics and statistics literature. The problems treated in that literature are simpler than (1.1)–(1.2). Section 3 describes the proposed method for selecting the regularization parameter. Section 4 presents the results of Monte Carlo experiments that explore the finite-sample performance of the method. Section 5 presents an empirical example, and Section 6 presents concluding comments. All proofs are in the Appendix.

2. Background

This section explains the estimation problem and the need for regularization, outlines the sieve estimator that is used with the adaptive estimation procedure, and reviews the statistics literature on selecting the regularization parameter.

2.1. The estimation problem and the need for regularization

Let X and W be continuously distributed random variables. Assume that the supports of X and W are $[0, 1]$. This assumption does not entail a loss of generality, because it can be satisfied by, if necessary, carrying out monotone increasing transformations of X and W. Let f_{XW} and f_W, respectively, denote the probability density functions of (X, W) and W. Define

$$m(w) = E(Y|W = w)f_W(w).$$

Let $L_2[0, 1]$ be the space of real-valued, square-integrable functions on $[0, 1]$. Define the operator A from $L_2[0, 1] \to L_2[0, 1]$ by

$$(Ah)(w) = \int_{[0,1]} h(x)f_{XW}(x, w)dx,$$

where h is any function in $L_2[0, 1]$. Then g in (1.1)–(1.2) satisfies $Ag = m$.

Assume that A is one-to-one, which is necessary for identification of g. Then $g = A^{-1}m$. If f_{XW}^2 is integrable on $[0, 1]^2$, then zero is a limit point (and the only limit point) of the singular values of A. Consequently, the singular values of A^{-1} are unbounded, and A^{-1} is a discontinuous operator. This is the ill-posed inverse problem. Because of this problem, g could not be estimated consistently by replacing m in $g = A^{-1}m$ with a consistent estimator, even if A were known. To estimate g consistently, it is necessary to regularize (or modify) A so as to remove the discontinuity of A^{-1}. A variety of regularization methods have been developed. See, for example, Engl et al. (1996), Kress (1999), and Carrasco et al. (2007), among many others. The regularization method used in this paper is series truncation, which is a modification of the Petrov–Galerkin method that is well-known in the theory of integral equations. See, for example, Kress (1999, pp. 240–245). It amounts to approximating A with finite-dimensional matrix. The singular values of this matrix are bounded away from zero, so the inverse of the approximating matrix is a continuous operator. The details of the method are described further in Section 2.2.

2.2. Sieve estimation and regularization by series truncation

The adaptive estimation procedure uses a two-stage estimator that is a modified version of Horowitz's (2012) sieve estimator of g. The estimator is defined in terms of series expansions of g, m, and A. Let $\{\psi_j : j = 1, 2, \ldots\}$ be a complete, orthonormal basis for $L_2[0, 1]$. The expansions are

$$g(x) = \sum_{j=1}^{\infty} b_j\psi_j(x),$$

$$m(w) = \sum_{k=1}^{\infty} m_k\psi_k(w),$$

and

$$f_{XW}(x, w) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} c_{jk}\psi_j(x)\psi_k(w),$$

where

$$b_j = \int_{[0,1]} g(x)\psi_j(x)dx,$$

$$m_k = \int_{[0,1]} m(w)\psi_k(w)dw,$$

and

$$c_{jk} = \int_{[0,1]^2} f_{XW}(x, w)\psi_j(x)\psi_k(w)dxdw.$$

To estimate g, we need estimators of m_k, c_{jk}, m, and f_{XW}. Denote the data by $\{Y_i, X_i, W_i : i = 1, \ldots, n\}$, where n is the sample size. The estimators of m_k and c_{jk}, respectively, are $\hat{m}_k = n^{-1}\sum_{i=1}^n Y_i\psi_k(W_i)$ and $\hat{c}_{jk} = n^{-1}\sum_{i=1}^n \psi_j(X_i)\psi_k(W_i)$. The estimators of m and f_{XW}, respectively, are $\hat{m}(w) = \sum_{k=1}^{J_n}\hat{m}_k\psi_k(w)$ and $\hat{f}_{XW}(x, w) = \sum_{j=1}^{J_n}\sum_{k=1}^{J_n}\hat{c}_{jk}\psi_j(x)\psi_k(w)$, where J_n is a pilot series truncation point that, for now, is assumed to be non-stochastic. It is assumed that as $n \to \infty, J_n \to \infty$ at a rate that is specified in Section 3.1. Section 3.3 describes an empirical method for choosing J_n. Define the operator $\hat{A}$ that estimates A by

$$(\hat{A}h)(w) = \int_{[0,1]} h(x)\hat{f}_{XW}(x, w)dx.$$

The first-stage estimator of g is defined as[1]

$$\tilde{g} = \hat{A}^{-1}\hat{m}. \tag{2.2}$$

[1] $\hat{A}$ and $\hat{A}^{-1}$ are defined on the subspace spanned by $\{\psi_j : j = 1, \ldots, J_n\}$. Under the assumptions of this paper, $\hat{A}$ can be represented by a square, non-singular matrix, and (2.2) is equivalent to $\hat{g} = (\hat{A}^*\hat{A})^{-1}\hat{A}^*\hat{m}$. Eq. (2.2) and this equivalence do not hold if $\hat{A}$ is non-square, as can happen if X and W have different dimensions. The row and column dimensions of a non-square $\hat{A}$ can be chosen separately, thereby requiring the choice of two regularization parameters. The treatment of this case is beyond the scope of this paper.