Journal of Econometrics 190 (2016) 1–17

Contents lists available at ScienceDirect

Journal of Econometrics

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

Series estimation under cross-sectional dependence

Jungyoon Lee [a], Peter M. Robinson [b,*]

[a] *Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK*
[b] *London School of Economics, Houghton Street, London WC2A 2AE, UK*

ARTICLE INFO

Article history:
Received 19 June 2013
Received in revised form
2 August 2015
Accepted 3 August 2015
Available online 11 September 2015

JEL classification:
C12
C13
C14
C21

Keywords:
Series estimation
Nonparametric regression
Semiparametric regression
Spatial data
Cross-sectional dependence
Mean square rate of convergence
Functional central limit theorem
Data-driven studentization

ABSTRACT

An asymptotic theory is developed for series estimation of nonparametric and semiparametric regression models for cross-sectional data under conditions on disturbances that allow for forms of cross-sectional dependence and heterogeneity, including conditional and unconditional heteroscedasticity, along with conditions on regressors that allow dependence and do not require existence of a density. The conditions aim to accommodate various settings plausible in economic applications, and can apply also to panel, spatial and time series data. A mean square rate of convergence of nonparametric regression estimates is established followed by asymptotic normality of a quite general statistic. Data-driven studentizations that rely on single or double indices to order the data are justified. In a partially linear model setting, Monte Carlo investigation of finite sample properties and two empirical applications are carried out.

1. Introduction

Economic agents are typically interdependent, due for example to externalities, spill-overs or the presence of common shocks. Such dependence is often overlooked or ignored in cross-sectional or panel data analysis. In order to account for possible cross-sectional dependence, one needs first to establish a framework under which its structure can be suitably formalized, and which permits an asymptotic statistical theory that is useful in statistical inference, in particular a central limit theorem for estimates of functions or parameters of interest. Several approaches to modelling cross-sectional dependence prominent in recent literature can accomplish this. One class of models postulates unobserved common factors that affect some or all individual units, see e.g Andrews (2005), Pesaran (2006) and Bai (2009), and entail persistent cross-sectional dependence. Two other classes involve a concept of "economic location" or "economic distance". In economic data, cross-sectional units correspond to economic agents such as individuals or firms, envisaged as positioned in some socio-economic (even geographical) space, whereby their relative locations underpin the strength of dependence, see e.g. Conley (1999) and Pinkse et al. (2002). The spatial autoregressive (SAR) model of Cliff and Ord (1968, 1981), see e.g. Arbia (2006), Lee (2002, 2004) and Kelejian and Prucha (1998, 1999), employs spatial weight matrices whose elements consist of inverse pairwise economic distances between agents, whence the dependent variable or disturbance for a given unit is affected by a weighted average of the other sampled units' variables. The weights are presumed known and reflect the proximity between agents, leaving a small number of parameters to be estimated. Alternatively, mixing conditions extending ones familiar from the time series literature, have been employed. Conley (1999) and Jenish and Prucha (2012), for example, develop spatial mixing and functions-of-mixing conditions in terms of economic distance between agents, under a suitable stationarity assumption, while an alternative type of condition was proposed by Pinkse et al. (2007). Another approach, of Robinson (2011), employs a possibly non-stationary, linear process for disturbances, with dependence in regressors expressed in terms of the departure of joint densities from the product of marginals; a degree of heterogeneity across

[*] Corresponding author. Tel.: +44 20 7955 7516; fax: +44 20 7955 6592.
E-mail address: p.m.robinson@lse.ac.uk (P.M. Robinson).

http://dx.doi.org/10.1016/j.jeconom.2015.08.001

units is permitted, as well as strong dependence analogous to long memory in time series, which is ruled out by mixing conditions, as well as weak dependence, and the model can also accommodate economic distances, as well as lattice or irregularly-spaced data.

On the other hand, nonparametric and semiparametric estimation has become well established in econometric analysis, enabling assumptions of a known parametric functional form, that are frequently not warranted by economic theory, to be dropped or relaxed. There are many theoretical results on nonparametric kernel estimation under temporal dependence, see e.g. Robinson (1983). Jenish (2012), Robinson (2011) and Robinson and Thawornkaiwong (2012) have considered kernel estimation in nonparametric regression and partially linear regression, under forms of cross-sectional dependence. The asymptotic behaviour of series estimation under independence has been studied in Andrews (1991) and Newey (1997). For weakly dependent time series data, Chen and Shen (1998) and Chen et al. (2012) offer a rather complete treatment of asymptotic theory and robust inference of general sieve M-estimation, while Chen and Christensen (2015) shows that spline and wavelet series regression estimation obtains the optimal uniform convergence rate of Stone (1982).

This paper presents an asymptotic theory for series estimation of nonparametric and semiparametric regression models that covers fairly general cross-sectional heterogeneity and dependence, mainly of a weak form. The conditions of the paper may be relevant to cross-sectional, spatial, time series and panel data, and follow the framework of Robinson (2011), with modifications necessitated by the nature of series estimates relative to kernel ones. Our asymptotic results can easily be modified to cover linear and nonlinear parametric regression. Our other main contribution is establishing a theoretical background for a studentization method that offers an alternative to the existing variance estimation literature. In the spatial context, an extension of heteroscedasticity and autocorrelation consistent (HAC) estimation of the covariance matrix in the limiting normal distribution familiar from the time series literature, see e.g. Hannan (1957), is possible if additional information is available, such as the locations or geographical or economic distances between units. Conley (1999) considered HAC estimation under a stationary random field with measurement error in distances, as did Kelejian and Prucha (2007) for SAR-type models, and Robinson and Thawornkaiwong (2012) in a semiparametric regression set-up. However, the small sample performance of HAC estimation can be poor and an alternative studentization that can produce more accurately sized tests was suggested by Kiefer et al. (2000) in a time series setting. The present paper provides theoretical justification for employing a version of such a studentization in spatial or spatio-temporal data, though it critically relies on an assumption that the practitioner can suitably order the data across one or two dimensions, as may be relevant when geographical locations are known, or there are one or two characteristics believed to be strongly associated with dependence between individuals.

The paper is structured as follows. In Section 2, the model setting is outlined. In Section 3, series estimation is introduced and a mean square rate of convergence for the nonparametric component is established. Section 4 contains asymptotic normality results, covering slower-than-$\sqrt{n}$, as well as $\sqrt{n}$, rate of convergence. In the latter setting, Section 5 presents data-driven studentizations in one and two dimensions. Using the semiparametric partially linear regression model, Section 6 presents a Monte Carlo study of finite sample performance and two empirical examples. Section 7 concludes. Two appendices contain proofs.

2. Model setting

The paper commences from the nonparametric regression model

$$Y_i = m(X_i) + U_i, \quad i = 1, 2, \ldots, \tag{1}$$

relating observable random variables $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, for some set $\mathcal{X} \subset \mathbb{R}^q$, where $m : \mathcal{X} \to \mathbb{R}$ and U_i satisfies

$$U_i = \sigma(X_i)e_i, \tag{2}$$

where $\sigma : \mathcal{X} \to \mathbb{R}$ and $e_i \in \mathbb{R}$ are unobservable random variables with zero mean and finite variance, independent of $\{X_i\}_{i=1}^{\infty}$ and $\sigma : \mathcal{X} \to \mathbb{R}$. We regard m and σ as nonparametric functions. The factor $\sigma(X_i)$ allows for conditional heteroscedasticity in Y_i, and the factor e_i for dependence and unconditional heteroscedasticity. We observe (X_i, Y_i) for $i = 1, \ldots, n$, while the e_i (and thence the U_i and Y_i) can form triangular arrays, so $e_i = e_{in}$, etc., but this feature will be suppressed in our notation; triangular arrays enable coverage of a wide range of models for spatial dependence, including SAR models with normalized weight matrices, and stationary models for panel data or multi-dimensional lattice or irregularly-spaced data where the single index i in (1) and (2) requires a re-labelling of multiple indices which is liable to change as n increases, as discussed by Robinson (2011), who considered kernel estimation of m. In some of our work e_i can have semi-strong dependence analogous to that found in long memory time series models. In Sections 3 and 4 we qualify (1) and (2) by detailed regularity conditions, including also restrictions on the dependence of X_i.

Under the preceding conditions $m(x) = E(Y_i|X_i = x)$ for $x \in \mathcal{X}$. We will estimate m by a series nonparametric regression estimate $\hat{m}$, constructed as a linear combination of pre-specified approximating functions. More generally, we are interested in estimating a $d \times 1$ vector functional $a(m)$ of m, as in Andrews (1991) and Newey (1997), where $a(m)$ can be estimated by $a(\hat{m})$. Simple nonparametric examples of $a(m)$ include the value of m, $a(m) = (m(x_1), \ldots, m(x_d))'$, and the value of the partial derivative, $a(m) = \left(\partial/\partial \mathrm{x}_\ell m(\mathrm{x})\big|_{x_1}, \ldots, \partial/\partial \mathrm{x}_\ell m(\mathrm{x})\big|_{x_d}\right)'$, at multiple fixed points $(x_1, \ldots, x_d) \in \mathcal{X}^d$, where x_ℓ is the ℓth element of x. Of semiparametric examples of $a(\cdot)$, the partially linear regression model, will be discussed in detail in Section 5. Other $a(\cdot)$, including nonlinear functionals, can be found in Newey (1997). Andrews (1991) established asymptotic normality for series estimates of a vector-valued linear $a(\hat{m})$, with X_i and U_i independent and non-identically distributed, and indicated that his proof can be extended to cover strong mixing time series regressors. Newey (1997) established uniform and integrated mean square rates of consistency for $\hat{m}(x)$ and asymptotic normality of $a(\hat{m})$ when X_i and U_i are independent and identically distributed (iid) and $a(m)$ is a possibly nonlinear scalar functional, also describing conditions under which $a(\hat{m})$ converges to $a(m)$ at parametric rate. Chen and Shen (1998) and Chen et al. (2012) considered these issues for sieve extreme estimates with weakly dependent time series, with rules of inference that are robust to weak dependence. Chen et al. (2012) also indicated that for certain cases of slower-than-$\sqrt{n}$ rate of convergence such as when $a(m) = m$, the asymptotic variance of the estimate $a(\hat{m})$ coincides with that obtained under independence, as found for kernel estimation by Robinson (1983), for example.

3. Estimation of *m* and mean square convergence rate

The estimation of m is based on user-chosen approximating functions $p_s(\cdot) : \mathcal{X} \to \mathbb{R}$, $s = 1, 2, \ldots$, and a data-free integer κ denoting the number of $p_s(\cdot)$ employed. Denote

$$\begin{aligned} &p^k(\cdot) = (p_1(\cdot), \ldots, p_k(\cdot))', \quad k \geq 1; \\ &P = [p^\kappa(X_1), \ldots, p^\kappa(X_n)]', \quad \kappa \geq 1; \end{aligned} \tag{3}$$

$$Y = (Y_1, \ldots, Y_n)'; \qquad \hat{\beta} = (P'P)^- P'Y, \tag{4}$$

where A^- denotes the Moore–Penrose pseudo-inverse of a matrix A, and a series estimate of $m(x)$ by

$$\hat{m}(x) = p^\kappa(x)'\hat{\beta}. \tag{5}$$