Journal of Econometrics 188 (2015) 327–345

Contents lists available at ScienceDirect

Journal of Econometrics

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

A semiparametric model for heterogeneous panel data with fixed effects

Lena Boneva[a,b,1], Oliver Linton[c,2], Michael Vogt[d,*]

[a] London School of Economics, United Kingdom
[b] Bank of England, United Kingdom
[c] University of Cambridge, United Kingdom
[d] University of Konstanz, Germany

ARTICLE INFO

Article history:
Available online 24 March 2015

Keywords:
Heterogeneous panel data
Kernel smoothing
Semiparametric estimation
Factor structure

ABSTRACT

This paper develops methodology for semiparametric panel data models in a setting where both the time series and the cross section are large. Such settings are common in finance and other areas of economics. Our model allows for heterogeneous nonparametric covariate effects as well as unobserved time and individual specific effects that may depend on the covariates in an arbitrary way. To model the covariate effects parsimoniously, we impose a dimensionality reducing common component structure on them. In the theoretical part of the paper, we derive the asymptotic theory for the proposed procedure. In particular, we provide the convergence rates and the asymptotic distribution of our estimators. In the empirical part, we apply our methodology to a specific application that has been the subject of recent policy interest, that is, the effect of trading venue fragmentation on market quality. We use a unique dataset that reports the location and volume of trading on the FTSE 100 and FTSE 250 companies from 2008 to 2011 at the weekly frequency. We find that the effect of fragmentation on market quality is nonlinear and non-monotonic. The implied quality of the market under perfect competition is superior to that under monopoly provision, but the transition between the two is complicated.

1. Introduction

In this paper, we develop estimation methodology for semiparametric panel models in a setting where both the time series and the cross section dimension are large. Such settings have become increasingly common over the last couple of years. In particular, they are frequently encountered in finance and various areas of economics such as industrial organization or labour economics. Cheng Hsiao has been a pioneer in the development of panel data techniques and his monograph (1986, 2003) contains the main methodological background for our work.

We investigate a regression model which has a nonparametric covariate effect along with individual and time specific fixed effects. The covariate effect is allowed to be heterogeneous across individuals, which is feasible given the long time series we are assuming. To restrict the heterogeneity to be of low dimension, we propose a common component structure on the model. In particular, we assume the individual covariate effects to be composed of a finite number of unknown functions that are the same across individuals but loaded up differently for each cross-sectional unit. The covariate effects are thus modelled as linear combinations of a small number of common functions. The individual and time specific effects of the model are allowed to be related to the covariate in quite a general way. This allows a potential channel for endogeneity, which is important in many applications. We recognize that the endogeneity that is permitted is rather limited, but we remark that this type of restriction is extremely widely exploited in empirical microeconomics, see Angrist and Pischke (2009, Chapter 5). A rigorous formulation of the model together with a detailed description of its components is given in Section 2. The issue of identifying the various model components is discussed in Section 3.

---

[*] Correspondence to: Department of Mathematics and Statistics, 78457 Konstanz, Germany.

E-mail addresses: l.m.koerber@lse.ac.uk (L. Boneva), obl20@cam.ac.uk (O. Linton), michael.vogt@uni-konstanz.de (M. Vogt).

[1] Address: Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom.

[2] Address: Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom.

http://dx.doi.org/10.1016/j.jeconom.2015.03.003
0304-4076/

Our model can be regarded as an intermediate case between two extremes. The one extreme is the homogeneous model, where the covariate effect is the same for each cross-sectional unit. This is a very common framework which has been investigated in various parametric and semiparametric studies, see for example Hsiao (1986). In a wide range of applications, it is however rather unrealistic to assume that the covariate effect is the same for all individuals. On the other extreme end, there is the fully flexible model without any restrictions on the covariate effects. One example is the classical SURE model. More recently, Chen et al. (2012) among others have studied a semiparametric version of this very general framework. Even though it is highly flexible, it is however not well suited to some applications. In particular, if the number of individuals is in the hundreds or thousands, the estimation output consists of a huge number of individual functions. This makes the model hardly interpretable. Furthermore, the estimation precision may be very low. Our model lies between these two extremes and allows the user to select the degree of flexibility appropriate for the given application.

Our setting falls in the class of semiparametric panel data models for large cross-section and long time series. Most of the models proposed in the literature for this type of panel data are essentially parametric. Some important papers include Phillips and Moon (1999), Bai and Ng (2002), Bai (2003, 2004), and Pesaran (2006). These authors have addressed a variety of issues including nonstationarity, estimation of unobserved factors, and model selection. Most of the work on semiparametric panel models is in the context of short time series, see for example Kyriazidou (1997). Nonparametric additive models have been considered for instance in Porter (1996). More recent articles include Mammen et al. (2009), Qian and Wang (2011), and Hoderlein et al. (2011).

Only recently, there have been a number of contributions to the non- and semiparametric literature on panels with large cross-section and time series dimension. Linton et al. (2009) consider estimation of a fixed effect time series. Atak et al. (2011) are concerned with seasonality and trends in a panel setting; see also Chen et al. (2013a). Connor et al. (2012) consider a semiparametric additive panel model for stock returns driven by observable covariates and unobservable "factor returns". They allow weak dependence in both time and cross-section direction, but the covariates are not time-varying and there is no individual effect. This model is suited for their application but does not allow a channel for endogeneity. The estimation method is made simpler by the fact that each additive term has a different covariate, whereas the common functions in our model all have the same covariate. Finally, Kneip et al. (2012) consider a model similar to ours except that they focus on time as the key nonparametric covariate.

In Section 8, we apply our methods to an empirical question of recent interest for policy makers and in academic research, that is, the effect of trading venue fragmentation on market quality. In 2007, the monopoly of primary European exchanges such as the London stock exchange was ended by the "Markets in Financial Instruments Directive". Since then, various new trading platforms have emerged and competed for trading volume. We investigate whether this competition has led to improved market quality for participants. It has been argued that High Frequency Trading has been a major beneficiary of the market fragmentation, and that this affects both the amount of fragmentation as well as the quality of the market outcomes.[3] Our model allows for this endogeneity channel by treating this unobservable as part of the individual and time effects. It also allows for heterogeneous nonlinear covariate effects of fragmentation on market quality, which we think are important for capturing the relationship of interest in an adequate way. We use a unique weekly dataset on the location and volume of trading for FTSE 100 and FTSE 250 companies over the period from 2008 to 2011, as well as publicly available measures of market quality. To summarize the results, we find that the effect of fragmentation on market quality is nonlinear and non-monotonic. The implied quality of the market under perfect competition is superior to that under monopoly provision, but the transition between the two regimes is complicated. Our model and procedures may also be applied in many other contexts in economics and finance.

Our method to estimate the common functions and the parameter vectors which constitute the individual covariate effects is introduced in Section 4. The asymptotic properties of the estimators are described in Section 5. In Section 5.2, we derive the uniform convergence rates as well as an asymptotic normality result for our estimators of the common functions. Importantly, the estimators can be shown to converge to the true functions at the uniform rate $\sqrt{\log nT/nTh}$ which is based on the pooled number of data points nT with n being the cross-section dimension and T the length of the time series. Intuitively, this fast rate is possible to achieve because the functions are the same for all individuals. This allows us to base our estimation procedure on information from the whole panel rather than on a single time series corresponding to a specific individual. In Section 5.3, we investigate the asymptotic behaviour of our parameter estimators. In particular, we show that they are asymptotically normal. As will turn out, the parameters are estimated with the same precision as in the case where the common functions are known. In particular, our estimators have the same asymptotic distribution as the oracle estimators constructed under the assumption that the functions are observed. To investigate the small sample performance of our estimation procedures, we conduct a series of simulation experiments. Overall, our procedures work well even for quite small sample sizes. For reasons of brevity, the detailed results are reported in the supplementary material (see Appendix C).

To keep the arguments and discussion as simple as possible, we derive our estimation procedure as well as the asymptotic results under the simplifying assumption that the number of common functions is known. In Sections 6 and 7, we explain how to dispense with this assumption. In particular, we provide a simple rule to select the number of unknown common functions. This complements our estimation procedure and makes it ready to apply to real data.

2. The model

In this section, we provide a detailed description of our model framework. We observe a sample of panel data $\{(Y_{it}, X_{it}) : i = 1, \dots, n,\ t = 1, \dots, T\}$, where i denotes the ith individual and t is the time point of observation. To keep the notation as simple as possible, we assume that both the variables Y_{it} and X_{it} are real-valued and focus on the case of a balanced panel.

The data are assumed to come from the model

$$Y_{it} = \mu_0 + \alpha_i + \gamma_t + m_i(X_{it}) + \varepsilon_{it}, \tag{1}$$

where $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$. Here, m_i are nonparametric functions which capture the covariate effect, μ_0 is the model constant (which may be deterministic or stochastic) and the variables ε_{it} are idiosyncratic error terms. The expressions α_i and γ_t are unobserved individual and time specific effects, respectively, which may depend on the regressors in an arbitrary way, e.g., $\alpha_i = G_i(X_{i1}, \dots, X_{iT}; \eta_i)$ and $\gamma_t = H_t(X_{1t}, \dots, X_{nt}; \delta_t)$ for some deterministic functions G_i, H_t and random errors η_i, δ_t that are independent of the covariates. As usual there is an identification shortfall here, and to identify the

---

[3] See the UK government project "The future of computer based trading in financial markets" for a full description of High Frequency Trading and related concepts. www.bis.gov.uk/foresight/our-work/projects/current-projects/computer-trading.