Journal of Econometrics 191 (2016) 86–109

Contents lists available at ScienceDirect

Journal of Econometrics

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

Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso[☆]

Junhui Qian[a], Liangjun Su[b,*]

[a] Antai College of Economics and Management, Shanghai Jiao Tong University, China

[b] School of Economics, Singapore Management University, Singapore

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form
28 September 2014
Accepted 24 September 2015
Available online 27 October 2015

JEL classification:
C13
C23
C33
C51

Keywords:
Adaptive Lasso
Change point
Group fused Lasso
Panel data
Penalized least squares
Penalized GMM
Structural change

ABSTRACT

In this paper we consider estimation and inference of common breaks in panel data models via adaptive group fused Lasso. We consider two approaches—penalized least squares (PLS) for first-differenced models without endogenous regressors, and penalized GMM (PGMM) for first-differenced models with endogeneity. We show that with probability tending to one, both methods can correctly determine the unknown number of breaks and estimate the common break dates consistently. We establish the asymptotic distributions of the Lasso estimators of the regression coefficients and their post Lasso versions. We also propose and validate a data-driven method to determine the tuning parameter used in the Lasso procedure. Monte Carlo simulations demonstrate that both the PLS and PGMM estimation methods work well in finite samples. We apply our PGMM method to study the effect of foreign direct investment (FDI) on economic growth using a panel of 88 countries and regions from 1973 to 2012 and find multiple breaks in the model.

1. Introduction

Recently there has been a growing literature on the estimation and tests of common breaks in panel data models in which there are N individual units and T time series observations for each individual. Depending on whether T is allowed to pass to infinity, the model is called "short" for fixed T and "large" (or of large dimension) if T passes to infinity. Implicitly, one usually allows N to pass to infinity in panel data models.[1] Most of the literature falls into two categories depending on whether the parameters of interest are allowed to be heterogeneous across individuals or not. The first category focuses on homogeneous panel data models and includes De Watcher and Tzavalis (2005), Baltagi et al. (2016), and De Watcher and Tzavalis (2012). De Watcher and Tzavalis (2005) compare the relative performance of two model and moment selection methods in detecting breaks in short panels; Baltagi et al. (2016) consider the estimation and identification of change points in large dimensional panel models with either stationary or nonstationary regressors and error terms; De Watcher and Tzavalis (2012) develop a testing procedure for common breaks in short linear dynamic panel data models. The second category considers estimation and inference of common breaks in heterogeneous panel data models; see Bai (2010), Kim (2011, 2014), Hsu and Lin (2012), Baltagi et al. (2014), among

☆ The authors would like to thank the co-editor Jianqing Fan, an associate editor, and three anonymous referees for their constructive comments and suggestions. They also express their sincere appreciation to Xiaohong Chen, Han Hong, and Chihwa Kao for discussions on the subject matter and participants in the 10th Symposium on Econometric Theory and Applications (SETA 2014, Taipei) and the 4th Shanghai Econometrics Workshop for helpful comments. Su gratefully acknowledges the Singapore Ministry of Education for Academic Research Fund under the Tier 2 grant number MOE2012-T2-2-021.

* Correspondence to: School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903, Singapore. Tel.: +65 6828 0386.
E-mail address: ljsu@smu.edu.sg (L. Su).

[1] Bai (1997a), Bai et al. (1998) and Qu and Perron (2007) extend the estimation of single-time series models to multiple-ones with simultaneous structural breaks where the number of equations is fixed.

http://dx.doi.org/10.1016/j.jeconom.2015.09.004
0304-4076/

others. Bai (2010) establishes the asymptotic properties of the estimated break point in a location–scale heterogeneous panel data model with either fixed or large T; Kim (2011) extends Bai's (2010) method and develops an estimation procedure for a common deterministic time trend break in large heterogeneous panels with a multi-factor error structure; Kim (2014) continues the study by estimating the common break date and common factors jointly; Hsu and Lin (2012) extends Bai's (2010) theory to nonstationary panel data models where the error terms follow an I(1) process; Baltagi et al. (2014) study the estimation of large dimensional static heterogeneous panels with a common break by extending Pesaran's (2006) common correlated effects (CCE) estimation procedure. In addition, Chan et al. (2008) extend the testing procedure of Andrews (2003) from time series to heterogeneous panels where the breaks may occur at different time points across individuals; Liao and Wang (2012) study the estimation of individual-specific structural breaks that exhibit a common distribution in a location–scale panel data model; Yamazaki and Kurozumi (2014) develop an LM-type test for slope homogeneity along the time dimension in fixed-effects panel data models with fixed N and large T.[2]

A common feature of all of the above works is that a one-time break, common or not, is assumed in the estimation procedure. Although the assumption of a single break greatly facilitates the estimation and inference procedure, inferences based on it could be misleading if the underlying model has an unknown number of multiple breaks. For this reason, a large literature on the estimation and inference of models with multiple structural changes has been developed in the single or multiple time series framework; see, e.g., Bai (1997a,b), Bai and Perron (1998), Qu and Perron (2007), Su and White (2010), Kurozumi (2015), and Qian and Su (2014, in press). In view of the fact that the conventional *avg*- and *exp*-type test statistics for multiple structural changes requires all permissible partitions of the sample which could be prohibitively large, Qian and Su (in press) propose shrinkage estimation of regression models with multiple structural changes by extending the fused Lasso of Tibshirani et al. (2005) to the time series regression framework.

In this paper we propose a shrinkage-based methodology for estimating panel data models with an unknown number of structural changes. The new methodology is most suitable for the vision that the regression coefficients in a panel data model may be time-varying but at the same time exhibit certain sparseness in abrupt changes or breaks. This vision seems pertinent in many applied studies using panel data that have a long time span measured in decades. During such a long time span, shocks to technologies, preferences, policies, and so on, may result in the change of a statistical relation applied economists seek to discover; but the shocks tend to be small over a relatively short time interval so that it does not alter the statistical relationship in short time. In this case, one has to allow the parameters in the model to change over time in an unknown way and recognize that parameters do not always alter from one time period to another one. Multiple structural breaks may occur during the whole time span but the number of breaks is generally small in comparison with the total number of time periods in the data, resulting in the sparseness of the breaks.

In terms of econometrics methodology, this paper extends the Lasso-type shrinkage approach in Qian and Su (in press) to panel data settings. To the best of our knowledge, this is the first in the literature to deal with panel data models with possibly multiple structural changes explicitly.[3] To stay focused, we consider homogeneous linear panel data models with an unknown number of common breaks and we do not allow cross section dependence. The extension to heterogeneous panel data models and to panel data models with cross section dependence will be discussed at the end of Section 7. For the advantage of the use of panel data to study common breaks, we refer the readers directly to Bai (2010) and De Watcher and Tzavalis (2012). Despite the fact that the Lasso-type shrinkage estimation has a long history and wide applications in statistics (see, e.g. Tibshirani, 1996; Knight and Fu, 2000; Fan and Li, 2001), the application of Lasso-type shrinkage techniques in econometrics has a relatively short history. But the number of applications in econometrics has been increasing very fast in the last few years. For example, Caner (2009) and Fan and Liao (2014) consider covariate selection in GMM estimation; Belloni et al. (2012) and García (2011) consider selection of instruments in the GMM framework; Liao (2013) provides a shrinkage GMM method for moment selection and Cheng and Liao (2015) consider the selection of valid and relevant moments via penalized GMM; Liao and Phillips (2015) apply adaptive shrinkage techniques to cointegrated systems; Kock (2013) considers Bridge estimators of static linear panel data models with random or fixed effects; Caner and Knight (2013) apply Bridge estimators to differentiate a unit root from a stationary alternative; Caner and Han (2014) proposes a Bridge estimator for pure factor models and shows the selection consistency; Lu and Su (in press) apply adaptive group Lasso to choose both regressors and the number of factors in panel data models with factor structures; Cheng et al. (2015) provide an adaptive group Lasso estimator for pure factor structures with a one-time structural break. This paper adds to the literature by applying the shrinkage idea to panel data models with an unknown number of breaks.

We propose two approaches, penalized least squares (PLS) and penalized general method of moments (PGMM), for the estimation of the panel data model with an unknown number of breaks. We apply first differencing to remove the fixed effects in the equation and focus on the first-differenced equation. When there is no endogeneity issue in the first-differenced equation, we propose to apply PLS to estimate the unknown number of break points and the regime-specific regression coefficients jointly where the penalty term is imposed through the adaptive group fused Lasso (AGFL) component. In the presence of endogeneity in the first-differenced equation, which may arise from endogenous regressors or lagged dependent variables in the original fixed-effects equation, we propose to apply PGMM to estimate the unknown number of break points and the regime-specific regression coefficients jointly where, again, the penalty term is imposed through the AGFL component. Unlike Qian and Su (in press) who can only establish the claim that the group fused Lasso cannot under-estimate the number of breaks in a time series regression and that all the *break fractions* (but not the break dates) can be consistently estimated as in Bai and Perron (1998), we show that with probability approaching one (w.p.a.1) both of our PLS and PGMM methods can correctly determine the unknown number of breaks and estimate the common *break dates* consistently. We obtain estimates of the regression coefficients via both the Lasso and post Lasso procedures and establish their asymptotic distributions. We also

[2] Pesaran and Yamagata (2008) and Su and Chen (2013) propose LM-type tests for slope homogeneity along the cross section dimension in large dimensional linear panel data models with additive fixed effects and interactive fixed effects, respectively.

[3] Bai (2010, Section 6) discusses the case of multiple breaks. As he remarks, if the number of breaks is given, the one-at-a-time approach of Bai (1997b) can be used to estimate the break dates, and if the number of breaks is unknown, a test for existence of break point can be applied to each subsample before estimating a break point. Alternatively, one can use information criteria to determine the number of breaks in the latter case, but further investigation is called for.