Journal of Statistical Planning and Inference 167 (2015) 69–89

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Panel data segmentation under finite time horizon

L. Torgovitski

Mathematical Institute, University of Cologne, Weyertal 86-90, 50931, Cologne, Germany

ARTICLE INFO

Article history:
Received 5 January 2015
Received in revised form 20 May 2015
Accepted 21 May 2015
Available online 16 June 2015

Keywords:
Panel data
Change point estimation
Segmentation
Nonparametric
CUSUM
Total variation denoising
LASSO
Serial dependence

ABSTRACT

We study the nonparametric change point estimation for common changes in the means of panel data. The consistency of estimates is investigated when the number of panels tends to infinity but the sample size remains finite. Our focus is on weighted denoising estimates, involving the group fused LASSO, and on the weighted CUSUM estimates. Due to the fixed sample size, the common weighting schemes do not guarantee consistency under (serial) dependence and most typical weightings do not even provide consistency in the i.i.d. setting when the noise is too dominant.

Hence, on the one hand, we propose a consistent covariance-based extension of existing weighting schemes and discuss straightforward estimates of those weighting schemes. The performance will be demonstrated empirically in a simulation study. On the other hand, we derive sharp bounds on the change to noise ratio that ensure consistency in the i.i.d. setting for classical weightings.

1. Introduction

The aim of this paper is to study the estimation of changes in the context of *panel data*. We focus on *common changes*, i.e. changes that occur simultaneously in many panels (but not necessarily in all) at the same time points and we consider an asymptotic framework where the number d of panels tends to infinity but the panel sample size n is fixed.

The analysis of change point estimation in panel data is subject of intensive research (in particular in econometrics) and, as discussed in Bai (2010), dates back at least to the works of Joseph and Wolfson (1992, 1993). However, the setting $d \to \infty$, which we are looking at, is generally not studied much in the literature concerning change point analysis and the settings $n \to \infty$ or $n, d \to \infty$ are far more established.

For the classical setting of $n \to \infty$ we refer to Csörgő and Horváth (1997). In the context of panel data especially the setting $n, d \to \infty$ is quite popular (cf., e.g., Bai (2010), Horváth and Hušková (2012) and Kim (2014)). Nevertheless, the assumption $d \to \infty$ and n fixed is also quite natural (cf., e.g., Bai (2010), Bleakley and Vert (2011a), Hadri et al. (2012) and also Peštová and Pešta (2015)). It reflects the situation where the amount of panels, i.e. the dimensionality, is much larger than the sample size.

Bai (2010) and Bleakley and Vert (2011a) mention important applications in finance, biology and medicine where in particular the framework of common changes is appropriate: In finance such changes may occur simultaneously across many stocks e.g. due to a credit crisis or due to tax policy changes. In biology and medicine relevant applications are in the study of genomic profiles within classes of patients. As mentioned in Bleakley and Vert (2011a) the latter example fits particularly well in the n fixed and $d \to \infty$ framework because the length of panels in genomic studies is fixed but the amount of panels can be increased by raising the number of patients.

E-mail address: ltorgovi@math.uni-koeln.de.

http://dx.doi.org/10.1016/j.jspi.2015.05.007
0378-3758/

The body of literature related to change point estimation (and detection) is huge. Hence, we do not attempt to summarize it here and refer the reader instead to the reviews in Jandhyala et al. (2013), Aue and Horváth (2013), Frick et al. (2014), and Horváth and Rice (2014). Change point analysis in the $d \to \infty$ and n fixed setting goes at least back to the (aforementioned) papers by Bleakley and Vert (2010, 2011a) and by Bai (2010). Therein estimation of common changes is studied independently from different perspectives. However, as we will see, the setups of Bleakley and Vert (2010, 2011a) and of Bai (2010) are closely related.[1]

Bai (2010) considered a *least squares* estimate for independent panels of linear time series under a single change point assumption and Bleakley and Vert (2011a) developed a weighted *total variation denoising* approach for the multiple change point scenario. Furthermore, Bleakley and Vert (2011a) proposed a computationally efficient algorithm and implemented it in a convenient MATLAB package GFLseg[2] which we also used in some of our simulations.

In this article we study consistency properties, in particular what we define as *perfect estimation*,[3] for the denoising estimate and for the weighted CUSUM (cumulative sums) estimate under weak dependence. Both types of estimates depend on certain weighting schemes w. Two schemes, w^{simple} and w^{standard}, were already considered by Bleakley and Vert (2011a) for the denoising approach in the n fixed and $d \to \infty$ setting (cf. Section 2.2.2 for the precise definition). They showed that w^{standard} ensures perfect estimation and therefore has better consistency properties for $d \to \infty$ than w^{simple} does. (Notice that Bai (2010) showed perfect estimation for the least squares estimate, which corresponds to the weighted CUSUM estimate with w^{standard}.)

We pick up the ideas of Bleakley and Vert (2011a) and extend them in various directions which will shed some new light on weighting schemes in general. First, we will emphasize the connection between the total variation denoising approach and the weighted CUSUM estimates. Notice that Bleakley and Vert (2011a) assumed independent panels of independent Gaussian observations. We continue by showing that their consistency results hold true under much weaker distributional assumptions, e.g. for panels of non-Gaussian time series with common factors. This is important since many datasets are neither Gaussian nor independent. An implication of our results is that w^{standard} generally does not provide consistency for panels of time series and therefore does not ensure perfect estimation under dependence.

As a solution, we propose a modified weighting scheme w^{exact}, which is a generalization of w^{standard}, that takes the covariance structure within panels into account. We show that this is the only choice that may generally ensure perfect estimation and derive quite mild conditions under which w^{exact} indeed ensures this property. In a detailed simulation study we confirm our results and demonstrate the gain in accuracy of w^{exact}. Moreover, we show that our approach outperforms the classical schemes even in random change point settings and for rather moderate dimensions. In practice, the weights w^{exact} have to be estimated. Therefore, we discuss feasible approaches and show their applicability in simulations.

Complementary to the study of perfect estimation, we investigate consistent estimation for a further class of weights w^{weighted}, which contains w^{simple} and w^{standard} as special cases, and characterize changes which are (not) correctly estimated as $d \to \infty$.

1.1. Basic setup

We observe d panels $\{Y_{i,k}\}_{i=1,\dots,n}$ for $k = 1, \dots, d$ in a *signal plus noise* model where

$$Y_{i,k} = m_{i,k} + \left(\varepsilon_{i,k} + \gamma_k \zeta_i\right).$$

Here, $\{m_{i,k}\}_{i,k\in\mathbb{N}}$ is an array of deterministic signals and $\{\varepsilon_{i,k}\}_{i,k\in\mathbb{N}}$ is an array of random centered noises. The $\{\zeta_i\}_{i\in\mathbb{N}}$ are the so-called *common factors* which are assumed to be random, centered and independent of $\{\varepsilon_{i,k}\}_{i,k\in\mathbb{N}}$. Their effect on the kth panel is quantified via the deterministic *factor loadings* $\gamma_k \in \mathbb{R}$.

We assume a (multiple) common change points scenario given by

$$m_{i,k} = \begin{cases} \mu_{1,k}, & i = 1, \dots, u_1, \\ \mu_{2,k}, & i = u_1 + 1, \dots, u_2, \\ \dots, & \dots, \\ \mu_{P+1,k}, & i = u_P + 1, \dots, n, \end{cases} \tag{1.1}$$

where we call $u_1, \dots, u_P \in \mathbb{N}$ change points. The $\mu_{j,k} \in \mathbb{R}$, $j = 1, \dots, P+1$, describe the piecewise constant signals in each panel, i.e. the means of the observations. In other words the means jump simultaneously from levels $m_{u,k}$ to levels $m_{u+1,k}$ in all panels $k = 1, \dots, d$ at change points $u \in \{u_1, \dots, u_P\}$. However, we do not require $m_{u,k} \neq m_{u+1,k}$ to hold for all $k = 1, \dots, d$, i.e. the changes do not have to occur in all panels. Later on we will impose more specific conditions on the average magnitude of changes.

Subsequently, we assume that $n \geq 3$ since otherwise the model (1.1) is not reasonable because for $n = 1$ the model may not contain any change and for $n = 2$ it holds trivially that $P = 1$ with $u_1 = 1$.

---

1 Notice that Bleakley and Vert (2011a) is a revised version of Bleakley and Vert (2010). Hence, we will mostly refer to the more recent article.

2 Download is available at http://cbio.ensmp.fr/GFLseg and is licensed under the GNU General Public License.

3 See Section 2.2.3 and (2.16).