Change Point Detection in Time Series Using Mixed Integer Programming

Artem Prokhorov[a,d,e] Peter Radchenko[a] Alexander Semenov[b]
Anton Skrobotov[c,d]

[a] University of Sydney Business School

[b] Industrial and Systems Engineering, University of Florida

[c] Russian Presidential Academy of National Economy and Public Administration

[d] Center for Econometrics and Business Analytics (CEBA), St. Petersburg State University

[e] Centre Interuniversitaire de Recherche en Économie Quantitative (CIREQ), University of Montreal

February 28, 2024

CONFERENCE DRAFT PAPER – PLEASE DO NOT CITE OR CIRCULATE WITHOUT THE AUTHOR'S PERMISSION

Abstract

We use cutting-edge mixed integer optimization (MIO) methods to develop a framework for identification and estimation of structural breaks in time series regressions. The framework requires a transformation of the problem into quadratic programming, which opens up significant advantages for joint estimation. When restated as an l_0-penalized regression problem, the method can be compared to the popular l_1-penalized regression (LASSO). We show that MIO is capable of finding provably optimal solutions using a well-known optimization solver. The framework permits simultaneous estimation of the number and location of structural breaks as well as regression coefficients, while it accommodates the option of specifying a given or minimal number of breaks. We demonstrate the effectiveness of our approach through extensive numerical experiments obtaining much more accurate estimation of the number of breaks in comparison to popular non-MIO methods. Two empirical applications using US macroeconomic data provide insights into economic dynamics that were not available before.

Keywords: Structural breaks, l_0-penalization, l_1-penalization, mixed integer quadratic programming

JEL Classification Codes: C12, C22

1 Introduction

Identifying structural breaks in time series, also known as change points, regime shifts and concept drifts, is a major area of interest within theoretical and applied statistics, going back at least to the 1960s (see, e.g., Shiryaev, 1963; Roberts, 1966). In modern econometrics the focus has been on statistical approaches that estimate break points by minimizing the regression sum of squares (see, e.g., Bai and Perron, 1998, 2003) or l_1-penalized sum of squares (see, e.g., Qian and Su, 2016; Kaddoura and Westerlund, 2023). As one of the most highly-cited examples, Bai and Perron (1998) proposed a specific-to-general testing strategy for estimating the number of breaks in linear regression models with potential heterogeneity in the errors. The method requires testing the null hypotheses of m breaks against the alternative of $m+1$ break starting with $m = 0$. The estimated number of breaks then is that for which the null hypothesis is not rejected.

Such approaches have been criticized for not offering a consistent estimator of break dates and for the tendency to overestimate the true number of breaks with a positive probability, equal to the tests' significance level asymptotically. Bai and Perron (2003) suggested using information criteria to choose the number of breaks, providing a consistent estimator of the break number. Still, the approaches inevitably put restrictive assumptions on the minimal length of a regime to be set by the researcher, while the results and hence critical values crucially depend on this length both in large and in small samples.

Penalized methods have been proposed to circumvent the restriction on the minimal length of a regime. The LASSO (Least Absolute Shrinkage and Selection Operator) of Tibshirani (1996) has been extremely effective at selecting the number of regression parameters with a simultaneous estimation of the non-zero parameters in linear models. For the structural break model, Harchaoui and Lévy-Leduc (2010) and Bleakley and Vert (2011) considered the estimation of break locations in one-dimensional piecewise constant signals, under the assumption of independence. Chan et al. (2014) extended their approach to dependent data allowing the number of breaks to grow with the sample size. They also provided a justification for using a second step in the selection procedure in order to avoid an overestimation of the number of breaks. The model of Chan et al. (2014) is a structural break autoregressive model without any exogenous regressors.

Behrendt and Schweikert (2021) proposed using adaptive group LASSO to select the number of breaks consistently, instead of the two-step procedure of Chan et al. (2014). The two-step procedure of Chan et al. (2014) is easier to apply, but it is less efficient than adaptive group LASSO of Behrendt and Schweikert (2021). Qian and Su (2016) also considered a linear regression model and estimated the number of regimes and model parameters by using adaptive fused LASSO. Their approach is also two-step due to an overestimation of the true break date in the first step.

In the context of these developments, it has been a common belief that mixed integer optimization (MIO) was not suitable for such problems due what is known as "combinatorial explosion", that is, the explosive growth in the number of combinations to consider and associated insurmountable computational task. However, recent remarkable advances in computational and algorithmic methods of optimization have shown attractive properties of integer and mixed integer programming as a means of obtaining efficient solutions in a wide range of statistical problems (see, e.g., Bertsimas et al., 2016; Mazumder et al., 2023; Hazimeh et al., 2023; Gómez and Prokopyev, 2021; Rebennack and Krasko, 2020).

For example, the problem of subset selection has become feasible and even standard for datasets much larger than statisticians previously thought possible (see, e.g., Bertsimas et al., 2020). Moreover, it is often suggested that subset selection using integer programming outperforms LASSO in many situations. The MIO challenge was famously picked by statisticians resulting in a discussion with rejoinders in a recent issue of Statistical Science (see, e.g., Hastie et al., 2020). Examples like this raise the prospects of applying MIO in other settings of interest to econometricians.

This paper proposes an MIO-based methodology for simultaneously estimating the number and location of structural breaks as well as the parameters of a time series regression model. We design a formulation that transforms the classical structural break detection problem into a mixed integer quadratic programming problem for which we can obtain a provably optimal solution. Importantly, our framework permits estimation of the unknown number of structural breaks while accommodating the option of specifying a required or minimal number of breaks. No requirements on the time between breaks is needed. As we show, the new estimator enjoys theoretical optimality properties under assumptions that are weaker than in the alternatives available in the literature.

To demonstrate the effectiveness of our approach, we conduct Monte-Carlo simulations. The proposed MIO representation is solved using a well-known optimization solver. We examine optimal and sub-optimal solutions of the problem, and the effect of tuning the parameters. We show how to choose the tuning parameters and compare our results with established methods such as those proposed by Bai and Perron (1998) and Qian and Su (2016).

The paper is organized as follows. Section 2 formulates the model and assumptions. In Section 3, we state the additional assumption and main asymptotic results on consistency and asymptotic normality of the proposed estimators. Numerical experiments are discussed in Section 4. Empirical applications are provided in Section 5. Section 7 concludes. All proofs are collected in the Appendix.

2 Model

We assume that the data is generated by the following process:

$$y_t = {\beta_t^*}^\top x_t + u_t, \tag{1}$$

where x_t is a $p \times 1$ vector of regressors, u_t is the error term, and the $(p \times 1)$ vector β_t^* takes distinct vector values $\alpha_j^*, j = 1, \ldots, m^*+1$, in the time interval $T_{j-1}^* \leq t < T_j^*$, where we use the convention that $T_0^* = 1$ and $T_{m^*+1}^* = T$. In this model, the indices $(T_1^*, \ldots, T_{m^*}^*)$, or break points, are assumed to be unknown, and the number of regimes is $m^* + 1$.

The goal is to find the unknown number m^* of unknown break dates $(T_1, \ldots, T_{m^*})$ as well as the regression coefficients $\alpha = ({\alpha_1^*}^\top, \ldots, {\alpha_{m+1}^*}^\top)^\top$. It is clear that with no penalty, an in-sample prediction error minimization for (1) gives $m^* = T - 1$ breaks and a perfect fit. This solution is unlikely to generalize well out-of-sample. To avoid overfitting, it is natural to impose a penalty that counteracts the reduction in prediction error for adjacent values of β_t^* that are not too far from one another. A common way of doing this utilizes various forms of l_1-norm of the difference $\beta_t^* - \beta_{t-1}^*$; see, e.g., Group Fused Lasso of Qian and Su (2016), Grouped Lasso of Kaddoura and Westerlund (2023).

Remark 1. Model (1) can be considered as pure structural change model in the Bai and Perron (1998) terminology. At the same time, we can consider the so-called partial structural change model where some elements of the β_t do not sustain structural changes. In others words, the β_t can be decomposed as $((p_1+p_2)\times 1)$ vector $\beta_t = (\beta_{1t}', \beta_{2t}')'$ with $(p_2 \times 1)$ sub-vector $\beta_{2t} = \beta_2$ which does not depend on t. The pure structural change model is considered for exposition purpose and brevity, but all results for pure structural change model can be extended for partial structural change model with more tedious proofs.

Qian and Su (2016) proposed to estimate unknown β_t by using the so-called fused Lasso with minimization a l_1 penalized least squares objective function. The problem can be formulated as follows:

$$\min_{\beta_t, z_t} \sum_{t=1}^{T} (y_t - \beta_t' x_t)^2 + \lambda \sum_{t=2}^{T-1} ||\beta_t - \beta_{t-1}||, \tag{2}$$

with $\lambda = \lambda_T$ being positive tuning parameter.

Instead of using Lasso, we can find the unknown β_t by solving the following l_0-penalized optimization problem:

$$\min_{\beta_t} \quad \sum_{t=1}^{T} (y_t - \beta_t' x_t)^2 \tag{3a}$$

$$\text{s. t.} \quad -Mz_t \leq \beta_{t+1} - \beta_t \leq Mz_t, \quad \text{for all } t = 1, \ldots, T-1, \tag{3b}$$

$$\sum_{t=1}^{T-1} z_t = m \tag{3c}$$

$$z_t + z_{t+1} \leq 1 \quad \text{for all } t = 1, \ldots, T-2, \tag{3d}$$

$$z_t \in \{0, 1\} \quad \text{for all } t = 1, \ldots, T-1, \tag{3e}$$

where the binary variable z_t is equal to 1 if there is a break at time t.

Remark 2. Constraints (3b)-(3c) have either computational or conceptual meaning, or both. Constraint (3b) ensures that β_t does not change in between breaks, which is a conceptual constraint; and it ensures that when a break occurs, β_t does not jump by more than M, where M is a large number, which is a computational constraint. Constaint (3c) restricts the maximum number of

breaks, $m_{\max}$, which is a computational constraint unless it is dictated by the context of a problem. Constraint (3d) prevents the breaks from happening consecutively, which is also a computational constraint. We can also set it as $z_t + z_{t+1} + z_{t+2} \leq 1$ to force intervals be 4 points. Constraint (3e) defines z_t which takes integer values 0 or 1.

Remark 3. Qian and Su (2016) rely on block partitions, while MIO is exact optimization.

3 Asymptotic properties

In this section we study the asymptotic properties of our proposed approach. We start by introducing some notation. Let $I_j^* = T_j^* - T_{j-1}^*$ for $j = 1, ..., m^* + 1$ and define

$$I_{\min} = \min_{1 \leq j \leq m^*+1} I_j^*, \quad J_{\min} = \min_{1 \leq j \leq m^*} \|\alpha_{j+1}^* - \alpha_j^*\|, \quad \text{and} \quad J_{\max} = \max_{1 \leq j \leq m^*} \|\alpha_{j+1}^* - \alpha_j^*\|.$$

We note that $I_{\min}$ is the smallest interval length among the $m^* + 1$ regimes of the true data generating process, while $J_{\min}$ and $J_{\max}$ measure the smallest and largest jump sizes, respectively, in the true vector of coefficients.

The main result in this section establishes consistency of our approach in estimating the true number of breaks, breakpoints, and regression coefficients, and also derives the corresponding rates of convergence. This result corresponds to the combination of the following two theorems in Qian and Su (2016): Theorem 3.4 (on correctly estimating the true number of breaks using the information criterion) and Theorem 3.1 (on the rate of convergence for the breakpoints and the coefficients when the correct number of breaks is used). We impose the same assumptions (A1 and A2) on the $\{(x_t, u_t)\}$ process as Qian and Su (2016) do in their theoretical analysis. These assumptions are formally stated in the appendix. We also impose the following additional requirements.

Assumption A3.

(i) $J_{\max} = O(1)$ and $T\delta_T J_{\min}^2/(\log T)^{c_\delta} \to \infty$ as $T \to \infty$, where $c_\delta = 6$ if A1(ii.a) is satisfied and $c_\delta = 1$ if A1(ii.b) is satisfied.

(ii) $\delta_T = O\big(I_{\min}^{1/2}/T\big)$ and $T^{1/2} m^* \big(I_{\min} J_{\min}^2\big)^{-1} \to 0$ as $T \to \infty$.

This assumption is a weaker version of Assumption A3 in Qian and Su (2016), which is required for their Theorem 3.4. More specifically, we do not impose their conditions $Tm^*\big[(\log I_{\min})^{c_\delta/2} T^{-1/2} I_{\min}^{-1/2} + I_{\min}^{-1}\big]\big(I_{\min} J_{\min}^2\big)^{-1} \to 0$ and $m^* = O\big(\log T\big)$.

Theorem 1. *Suppose that Assumptions A1-A2 in the appendix are satisfied, Assumption A3 holds,* $\lambda/\big[m^* T \delta_T\big] \to \infty$, $\lambda/\big[J_{\min}^2 I_{\min}\big] \to 0$ *as* $T \to \infty$. *Then, we have*

$$P(\widehat{m} = m^*) \to 1 \quad as \quad T \to \infty;$$

$$P\Big(\max_{1 \le j \le m^*} \big|\widehat{T}_j - T_j^*\big| \le T\delta_T\Big) \to 1 \quad as \quad T \to \infty;$$

$$\widehat{\alpha}_j - \alpha_j^* = O_p\Big(\big[I_j^*\big]^{-1/2}\Big) \quad for\ each \quad j = 1, ..., m^* + 1.$$

Remark 4. *Our assumptions are weaker than the corresponding assumptions imposed by Qian and Su (2016). Here is a summary of the differences:*

- *As mentioned in the paragraph before Theorem 1, we don't impose two bounds involving* m^* *that Qian and Su (2016) do in their Assumption 3. In particular, we allow* m^* *to grow faster than* $\log T$ *as* $T \to \infty$.

- *We don't impose the bound* $\widehat{m} \le m_{\max}$, *where* $m_{\max} \le C \log T$, *as Qian and Su (2016) do on page 1386 – they use this bound in the proof of their Theorem 3.4 on recovering the correct number of breaks (see the statement and proof of their Lemma E.1 on page 1424). Thus, we do not restrict the range of* $\widehat{m}$ *in our optimization problem to take advantage of the upper bound* $m^* \le C \log T$; *such a bound would typically be unknown in practice.*

Remark 5. *Qian and Su (2016) impose conditions on tuning parameter* ρ_T, *which controls the penalty on the total number of breaks in the information criterion that they use to determine the final estimator. We also impose conditions on* λ, *which controls our penalty on the number of breaks. As the two estimators use these penalties differently, we cannot directly compare the conditions on* ρ_T *and* λ. *However, both sets of conditions are standard – they are used to ensure that the penalty is neither too large nor too small, so that the correct number of breaks can be*

recovered with high probability. We note that the width of the λ-range considered in Theorem 1, i.e., $m^*T\delta_T \ll \lambda \ll J^2_{\min}I_{\min}$, grows without bound as $T \to \infty$, because $J^2_{\min}I_{\min} \to \infty$ and $J^2_{\min}I_{\min}/[m^*T\delta_T] \to \infty$ under the conditions imposed in Assumption A3(ii).

In the above remarks, we compare our estimator to the following two-stage procedure of Qian and Su (2016). First, a base GFL estimator is obtained for a range of values of the tuning parameter corresponding to the group fused Lasso penalty; second, the final estimator is determined by selecting the tuning parameter using an information criterion that penalizes the number of breaks. In contrast, our approach does everything in one go, and avoids the estimation bias that comes from Lasso penalty. We note that the aforementioned two-stage approach still has a tuning parameter, ρ_T, corresponding to the penalty on the number of breaks, which can be seen as the counterpart of our tuning parameter λ.

In comparison to the base GFL estimator of Qian and Su (2016), which uses a group fused Lasso penalty in place of our penalty that simply counts the number of breaks, we observe that our estimator has better asymptotic properties. In particular, while Qian and Su (2016) show that the GFL estimator has at least as many breaks as the true model (see their Theorem 3.3), they do not establish a complimentary upper bound result. In contrast, we show that our estimator recovers the correct number of breaks with probability tending to one.

Next, we establish the asymptotic normality of our estimated regression coefficients. To state a clean result, we assume that m^* is *fixed* and nonzero. However, we note that this result can be extended to the general case as in Qian and Su (2016), by imposing additional assumptions on m^* and stating the central limit theorem for pre-specified fixed-dimensional sub-vectors of coefficients.

We impose the following additional conditions, which are also required by Qian and Su (2016) in the analogous result for their estimator.

Assumption A4.

(i) $\delta_T^{-1}I_{\min}^{-1}\left[I_{\min}^{1/2}T^{-1/2}(\log I_{\min})^{c_\delta/2}+1\right] = O(1)$;

(ii) $T\delta_T/I_{\min}^{1/2} \to 0$ as $T \to \infty$.

Because our estimator recovers the correct number of breakpoints with probability tending to one, we follow the approach of Qian and Su (2016) and establish asymptotic normal-

ity for the estimator that solves optimization problem (3) with the restriction that the total number of breakpoints is exactly m^*, i.e., $\sum_{t=1}^{T} z_t = m^*$. We write $\widehat{\boldsymbol{\alpha}}_{m^*}$ for the corresponding vector of estimated regression coefficients and observe that $\widehat{\boldsymbol{\alpha}}_{m^*} = \big(\widehat{\mathbb{X}}^\top \widehat{\mathbb{X}}\big)^{-1} \widehat{\mathbb{X}}^\top Y$, where $\widehat{\mathbb{X}} = \text{diag}\big((x_1, ..., x_{\widehat{T}_1 - 1})^\top, ..., (x_{\widehat{T}_{m^*}}, ..., x_T)^\top\big)$. Let $\Psi = \text{plim} D^{-1} \mathbb{X}^\top \mathbb{X} D^{-1}$ and $\Phi = \text{plim} D^{-1} \mathbb{X}^\top U U^\top \mathbb{X} D^{-1}$, where $D = \text{diag}\big(I_1^{*1/2} \mathbb{I}_p, ..., I_{m^*+1}^{*\,1/2} \mathbb{I}_p\big)$ and $\mathbb{X}$ is defined analogously to $\widehat{\mathbb{X}}$ but using the true rather than the estimated breakpoints.

Theorem 2. *Let* $\widehat{D} = \text{diag}\big([\widehat{T}_1 - \widehat{T}_0]^{1/2} \mathbb{I}_p, ..., [\widehat{T}_T - \widehat{T}_{m^*}]^{1/2} \mathbb{I}_p\big)$ *and suppose that Assumptions A1-A4 hold. Then,* $\widehat{D}\big(\widehat{\boldsymbol{\alpha}}_{m^*} - \boldsymbol{\alpha}^*\big) \xrightarrow{d} N\big(\mathbf{0}, \Psi^{-1} \Phi \Psi^{-1}\big)$.

We note that Theorem 2 is a direct consequence of Theorem 3.6 in Qian and Su (2016) on the asymptotics of their post-LASSO estimator of regression coefficients.

4 Monte-Carlo simulations

In this section, we investigate the finite sample properties of the number of breaks selection by the MIP algorithm proposed in Section 2.

In order to compare the new estimator to Grouped Fused Lasso approach (GFL), we follow Qian and Su (2016) and use the same data generating process as in (1) with the following cases of interest: the case of no breaks, the case of one break, and the case of many breaks. We also compare MIP and GFL methods with classical approaches used in Bai and Perron (2003), namely, BIC, LWZ information criteria, and sequential method of Bai and Perron (1998) (SEQ).

5 The case of no breaks

The Monte Carlo simulations reported in this section are based on data generated by the following DGP:

$$y_t = 1 + x_t + u_t, \tag{4}$$

where

1. $x_t \sim i.i.d.N(0,1)$, $u_t \sim i.i.d.N(0, \sigma_u^2)$

2. $x_t = 0.5x_{t-1} + \eta_t$, $\eta_t \sim i.i.d.N(0, 0.75)$, $u_t \sim i.i.d.N(0, \sigma_u^2)$

3. $x_t \sim i.i.d.N(0, 1)$, $u_t = \sigma_u v_t$, $v_t = 0.5v_{t-1} + \varepsilon_t$, $\varepsilon_t \sim i.i.d.N(0, 1)$

4. $x_t = 0.5x_{t-1} + \eta_t$, $\eta_t \sim i.i.d.N(0, 0.75)$, $u_t = \sigma_u\sqrt{h_t}\varepsilon_t$, $h_t = 0.05 + 0.05u_{t-1}^2 + 0.9h_{t-1}$, $\varepsilon_t \sim i.i.d.N(0, 1)$

5. $x_t = 0.5x_{t-1} + \eta_t$, $\eta_t \sim i.i.d.N(0, 0.75)$, $u_t \sim i.i.d.N(0, \sigma_1^2)$ for $t \in \{1, 2, \ldots, T/2\}$ and $u_t \sim i.i.d.N(0, \sigma_2^2)$ for $t \in \{T/2, T/2 + 1, \ldots, T\}$

6. $y_t = \alpha y_{t-1} + \varepsilon_t$, $x_t = y_{t-1}$, $\varepsilon_t \sim i.i.d.N(0, 1 - \alpha^2)$

In these DGP, different types of serial correlation and conditional heteroskedasticity are assumed. The parameters $\sigma_u \in \{0.5, 1, 1.5\}$, $\sigma_1 = 0.1$, $\sigma_2 \in \{0.2, 0.3, 0.5\}$, $\alpha \in \{0.2, 0.5, 0.9\}$.

Table 5 presents the results of correct detection of 0 breaks (in persentages) for DGP-1–DGP3, while Table 5 presents the results of correct detection of 0 breaks (in percentages) for DGP-4–DGP6. For all GDPs and settings, MIP and GFL give very close percentage points of correct detection of no breaks. BIC and SEQ work uniformly worse, while LWZ gives almost 100% correct detection of no breaks in all GDPs. One can observe, that correct detection increases if the variance σ_u increases and/or sample size increases.

6 The case of one break

The Monte Carlo simulations reported in this section are based on data generated by the following DGP:

$$y_t = \beta_t x_t + u_t, \tag{5}$$

where

1. $\beta_t = \mathbf{1}\{T/2 < t \leq T\}$, $x_t \sim i.i.d.N(0, 1)$, $u_t \sim i.i.d.N(0, \sigma_u^2)$

2. $\beta_t = \mathbf{1}\{T/2 < t \leq T\}$, $x_t \sim i.i.d.N(0, 1)$, $u_t = \sigma_u v_t$ with $v_t = 0.5v_{t-1} + \epsilon_t$, $\varepsilon_t \sim N(0, 0.75)$

3. $\beta_t = \mathbf{1}\{T/2 < t \leq T\}$, $x_t = 0.5x_{t-1} + \eta_t$, $\eta_t \sim i.i.d.N(0, 0.75)$, $u_t \sim i.i.d.N(0, \sigma_u^2)$

Table 1: Percentage of correct detection $m = 0$ breaks

	σ_u	T	MIP	GFL	BIC	LWZ	SEQ
DGP-1	0.5	100	97.6	98.2	96.4	100.0	84.4
		200	99.6	99.6	98.4	100.0	89.2
		500	100.0	100.0	99.2	100.0	94.2
	1	100	96.8	97.8	97.2	100.0	84.6
		200	99.6	99.8	97.8	100.0	92.2
		500	100.0	100.0	99.2	100.0	94.2
	1.5	100	96.6	98.2	97.8	100.0	85.2
		200	99.6	99.8	98.2	99.8	92.2
		500	100.0	100.0	99.2	100.0	93.2
DGP-2	0.5	100	66.8	70.8	65.6	94.0	71.0
		200	82.0	85.2	67.8	97.4	79.4
		500	95.6	96.2	73.0	99.4	83.0
	1	100	80.4	84.6	80.0	97.6	75.8
		200	93.4	93.8	85.0	99.0	81.4
		500	99.2	99.6	87.4	100.0	84.0
	1.5	100	87.6	91.6	88.0	99.2	79.6
		200	95.2	96.0	92.0	99.6	83.6
		500	99.6	99.8	94.6	100.0	85.2
DGP-3	0.5	100	97.0	98.2	96.6	100.0	86.6
		200	99.8	99.8	97.8	100.0	89.2
		500	100.0	100.0	98.8	100.0	92.2
	1	100	93.0	97.2	96.2	100.0	89.2
		200	98.0	99.2	98.2	100.0	91.8
		500	100.0	100.0	99.0	100.0	92.2
	1.5	100	91.4	96.4	97.0	100.0	89.6
		200	97.8	99.2	98.4	100.0	91.6
		500	100.0	100.0	99.4	100.0	91.8

Table 2: Percentage of correct detection $m = 0$ breaks

	σ_u	T	MIP	GFL	BIC	LWZ	SEQ
DGP-4	0.5	100	56.6	61.8	56.8	91.2	65.6
		200	80.2	84.6	64.0	98.0	76.4
		500	94.0	94.8	67.6	98.4	84.0
	1	100	76.0	80.0	76.2	97.6	70.8
		200	91.2	94.0	83.0	99.2	79.4
		500	98.0	98.2	87.2	99.6	82.8
	1.5	100	88.2	90.8	91.8	98.6	83.0
		200	92.6	92.8	92.0	98.8	90.2
		500	97.8	95.4	93.0	99.6	94.4
DGP-5	$\sigma_2 = 0.2$	100	55.8	60.2	56.8	91.2	67.4
		200	71.4	77.6	58.0	96.0	74.2
		500	95.0	96.2	67.0	99.4	87.2
	$\sigma_2 = 0.3$	100	57.2	62.8	58.4	91.4	68.6
		200	73.6	79.0	58.8	96.2	73.4
		500	94.8	96.2	69.0	99.4	86.8
	$\sigma_2 = 0.5$	100	59.4	65.6	60.8	92.2	68.8
		200	77.2	81.2	62.0	96.8	73.4
		500	95.8	96.8	73.2	99.6	87.0
DGP-6	$a = 0.2$	100	97.2	98.0	97.2	100.0	81.2
		200	99.8	99.8	98.0	100.0	88.6
		500	100.0	100.0	99.4	100.0	91.4
	$a = 0.5$	100	96.8	98.2	97.4	99.6	83.0
		200	99.4	99.6	98.4	100.0	90.2
		500	100.0	100.0	99.2	100.0	91.6
	$a = 0.9$	100	96.2	97.6	96.0	99.6	81.4
		200	99.8	100.0	97.6	100.0	88.6
		500	100.0	100.0	98.2	100.0	88.6

4. $\beta_t = \mathbf{1}\{T/2 < t \leq T\}$, $x_t = 0.5x_{t-1} + \eta_t$, $\eta_t \sim i.i.d.N(0, 0.75)$, $u_t = \sigma_u\sqrt{h_t}\varepsilon_t$, $h_t = 0.05 + 0.05u_{t-1}^2 + 0.9h_{t-1}$, $\varepsilon_t \sim i.i.d.N(0, 1)$

5. $\beta_t = \mathbf{1}\{T/2 < t \leq T\}$, $x_t = 0.5x_{t-1} + \eta_t$, $\eta_t \sim i.i.d.N(0, 0.75)$, $u_t = \sigma_u v_t$ with $v_t = \varepsilon_t + 0.5\varepsilon_{t-1}$, $\varepsilon_t \sim i.i.d.N(0, 0.8)$

6. $\beta_t = 0.2\mathbf{1}\{1 < t \leq T/2\} + 0.8\mathbf{1}\{T/2 < t \leq T\}$, $x_t = y_{t-1}$, $u_t \sim i.i.d.N(0, \sigma_u^2)$

Table 6 presents the results of correct detection of 0 breaks (in persentages) for DGP-1–DGP3, while Table 6 presents the results of correct detection of 0 breaks (in percentages) for DGP-4–DGP6. In all tables, the columns pce demonstrate the percentage of correct detection of one break for some method, and columns hd/T denotes the Hausdorff distance divided by T between estimated break date and true break date (conditional on the correct estimation of $\hat{m} = 1$) and measures the accuracy of break date estimation.

It can be seen that again MIP and GFL give the similar results: in some cases MIP detect one break more often, while in other cases GFL detect one break more often. Also, unreported results demonstrate that MIP tend to overestimate the number of breaks in small samples while GFL tend to underestimate the number of breaks in small samples. The percentage of correct detection decreases if the variance σ_u increases. However, the results (correct detection and accuracy) are improved if the sample size T increases.

7 The case of many breaks

For many breaks, we follow the setups of Qian and Su (2016) and consider

$$y_t = \beta_t x_t + u_t, \tag{6}$$

where $x_t \sim i.i.d.N(0, 1)$, $u_t \sim i.i.d.N(0, \sigma_u^2)$,

$$\beta_t = \begin{cases} 0, & \delta(2i) + 1 \leq t < \delta(2i+1) \\ 1, & \delta(2i+1) + 1 \leq t < \delta(2i+2) \end{cases}, \ i = 0, 1, \ldots, R/2.$$

For the first setup (DGPn-1), we fix the length of the regime $\delta = 30$ and allow different number of regimes $R \in \{6, 10, 20\}$. For the second setup (DGPn-1), we fix the number of the regimes $R = 10$

Table 3: Percentage of correct detection of one break and accuracy of changepoint estimation when $m = 1$, DGP-1

			MIP		GFL		BIC		LWZ		SEQ	
	σ_u	T	pce	hd/T	pce	hd/T	pce	hd/T	pce	hd/T	pce	hd/T
DGP-1	0.5	100	94.2	1.2	98.6	1.7	96.4	1.3	100	1.3	88.6	1.2
		200	99.4	0.6	99.2	0.8	98.8	0.6	100	0.6	95.2	0.6
		500	100	0.2	99.8	0.3	99.4	0.2	100	0.2	96.8	0.2
	1	100	92	4.4	95	3.6	96	4.1	80	3.8	90.2	4.2
		200	99	1.9	99.4	1.9	98.2	1.9	98.8	1.9	96	1.9
		500	100	0.8	99.2	0.8	99.4	0.8	100	0.8	96.6	0.7
	1.5	100	63.6	7.6	64.8	5.3	70.6	6.7	28	5.7	71.4	7.0
		200	86.2	3.7	86.2	3.1	94	3.8	55.6	3.3	93.8	3.9
		500	99.6	1.6	98.6	1.4	99.4	1.6	97.8	1.6	97	1.6
DGP-2	0.5	100	91.2	1.0	98.4	1.5	97.8	1.0	99.8	1.0	88.8	1.1
		200	97	0.6	98.6	0.8	97.2	0.6	100	0.6	93.4	0.6
		500	99.8	0.2	99.4	0.3	98.8	0.2	100	0.2	95.6	0.2
	1	100	88.2	3.7	95.2	3.0	96	3.6	81.2	3.2	88.6	3.5
		200	96.2	1.7	99.2	1.6	97.2	1.7	98.8	1.6	93.2	1.6
		500	99.6	0.7	99.4	0.7	98.6	0.7	100	0.7	95.4	0.7
	1.5	100	60.8	7.0	63.4	4.9	69.8	6.4	27.4	4.8	72.6	6.4
		200	83.4	3.5	84	3.0	93.2	3.5	55.8	2.8	91.8	3.7
		500	99.4	1.5	98.8	1.4	99.2	1.6	96.4	1.5	96	1.6
DGP-3	0.5	100	93.8	1.2	98.8	1.6	96.6	1.2	100	1.2	88.4	1.2
		200	99.2	0.6	100	0.8	98.4	0.6	100	0.6	94.8	0.6
		500	100	0.2	99.8	0.3	99	0.2	100	0.2	94.4	0.2
	1	100	89.2	4.5	93.4	3.7	94	4.1	77.2	3.9	86.4	4.2
		200	99.4	2.0	99.4	1.8	97.4	2.0	98.6	2.0	94.4	1.9
		500	100	0.7	99.8	0.8	99	0.7	100	0.7	94.2	0.7
	1.5	100	60.6	8.2	61.2	5.6	67	7.4	22.4	5.8	69.8	8.0
		200	87.8	4.3	87.2	3.1	93.4	4.4	58	3.7	93.8	4.4
		500	99.6	1.7	99	1.4	99.2	1.7	98.8	1.7	95.6	1.7

Table 4: Percentage of correct detection of one break and accuracy of changepoint estimation when $m = 1$, DGP-4

			MIP		GFL		BIC		LWZ		SEQ	
	σ_u	T	pce	hd/T	pce	hd/T	pce	hd/T	pce	hd/T	pce	hd/T
DGP-4	0.5	100	95.4	0.8	99.6	1.4	97.2	0.8	99.6	0.8	90.4	0.8
		200	99.6	0.5	99.6	0.7	97.6	0.5	100	0.5	93.4	0.5
		500	100	0.2	99.8	0.3	99.6	0.2	100	0.2	96.6	0.2
	1	100	90	3.8	95.4	3.3	93.8	3.7	84.4	3.5	87.6	4.0
		200	97.6	1.9	99.4	1.8	97.2	2.0	99	1.9	94	2.0
		500	99.8	0.8	98.8	0.8	99.6	0.8	100	0.8	96.6	0.8
	1.5	100	23.6	21.7	19.8	12.6	24	13.7	4.4	14.5	31.8	12.9
		200	15.4	16.8	15	14.1	23.6	11.8	1.2	8.0	31.8	10.0
		500	3.6	29.3	4.4	27.5	16.2	15.8	0.4	32.6	21.4	10.6
DGP-5	0.5	100	95.6	1.1	97.8	1.5	97	1.1	100	1.2	86.4	1.1
		200	98.6	0.6	99.4	0.9	98.6	0.6	100	0.6	92.8	0.6
		500	100	0.2	99.2	0.3	99.2	0.2	100	0.2	96.4	0.2
	1	100	93.6	4.1	96.2	3.7	95.8	4.2	80.6	4.1	87.8	4.4
		200	98.4	1.9	99.6	1.9	98.8	1.8	99.6	1.9	92.4	1.8
		500	100	0.7	99.2	0.7	99	0.7	100	0.7	96	0.7
	1.5	100	63.4	7.9	64.2	5.9	70.2	7.4	27.6	6.6	72.2	8.1
		200	87	4.4	87.4	3.4	95	4.1	58.4	3.6	92	4.1
		500	99.6	1.5	99	1.4	98.8	1.5	98.8	1.5	96.2	1.4
DGP-6	0.5	100	65	8.1	64.6	8.4	69.8	7.0	28.8	5.9	70	7.6
		200	93	4.2	93	5.8	97	4.2	71.2	4.1	91.8	4.2
		500	100	1.5	97.6	2.5	99.2	1.5	99.6	1.5	93.4	1.5
	1	100	65	8.1	64.6	8.4	69.8	7.0	28.8	5.9	70	7.6
		200	93	4.2	93	5.8	97	4.2	71.2	4.1	91.8	4.2
		500	100	1.5	97.6	2.5	99.2	1.5	99.6	1.5	93.4	1.5
	1.5	100	65	8.1	64.6	8.4	69.8	7.0	28.8	5.9	70	7.6
		200	93	4.2	93	5.8	97	4.2	71.2	4.1	91.8	4.2
		500	100	1.5	97.6	2.5	99.2	1.5	99.6	1.5	93.4	1.5

and allow different values of the regimes lengths by varying $T \in \{150, 300, 600\}$.

Table 7 demonstrates the results of correct detection of specific number of breaks denoted by $R-1$ and indicated in the 3rd column. This table also indicate the sampel size T. The percentage of correct detection of multiple breaks are compared only for MIP and GFL because classical methods of Bai and Perron (1998, 2003) do not allow 9 or 19 breaks for corresponding sample sizes. Th pces for MIP are uniformly better in all cases, and the better performance is highly valuable in case of high variance ($\sigma_u = 0.5$). Also, the accuracy of MIP is also better for most cases in terms of Hausdorff distance.

Table 5: Share of correct detections of one break and accuracy of change-point estimation in case of many breaks

				MIP		GFL	
DGPn-1	σ_u	R	T	pce	hd/T	pce	hd/T
	0.2	6	180	98.8	0.6	86.6	0.6
		10	300	98.6	0.5	76.4	0.5
		20	600	100.0	0.4	56.4	0.3
	0.5	6	180	99.2	1.9	37.8	2.0
		10	300	94.8	1.4	26.8	3.5
		20	600	27.0	1.0	1.8	1.6
				MIP		GFL	
DGPn-2	σ_u	R	T	pce	hd/T	pce	hd/T
	0.2	10	150	95.8	1.1	66.6	1.0
		10	300	99.2	0.5	78.0	0.5
		10	600	100.0	0.2	82.8	0.3
	0.5	10	150	43.2	2.8	12.4	3.8
		10	300	94.4	1.5	19.6	4.4
		10	600	100.0	0.8	23.0	2.1

8 Empirical application

8.1 Level shifts in the US real interest rate

In this section, we consider U.S. real interest rate series from 1961Q1 to 1986Q3 used by Garcia and Perron (1996) and Bai and Perron (2003). Our model is a simple level shift model which can be written as follows

$$y_t = \mu_j + u_t, \ j = 1, \ldots, m+1. \tag{7}$$

The results are presented in Table 6. The MIP and GFL methods estimate 4 breaks while classical methods detect onnly 2, 0 and 3 breaks for BIC, LWZ and SEQ, respectively. The most common break date is 1972Q4 which can be identified with oil crisis, while the breaks in 1980th can be connected with Paul Volker period and Great Moderation. We can find that the level breaks identified by MIP and GFL around 1980th are very close to each other which violates the assumption of BIC, LWZ and SEQ about the minimal length of the regime.

The break dates identified by MIP are depicted in Figure 1.

Table 6: Estimated break dates

	$\hat{m}$	Dates					
MIP	4		1972Q4	1980Q1		1981Q3	1983Q1
GFL	4		1972Q4	1980Q1	1980Q4	1981Q3	
BIC	2		1972Q4		1980Q4		
LWZ	0						
SEQ (trim=0.1)	3	1967Q1	1972Q4		1980Q4		

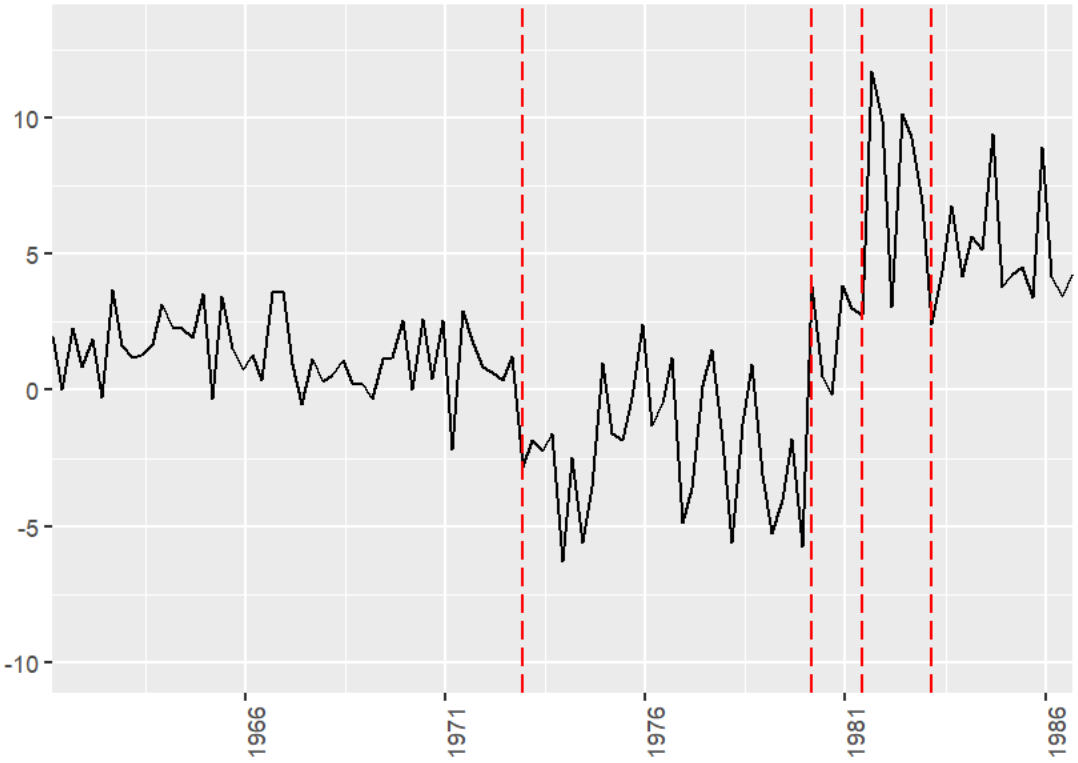

Figure 1: US real interest rate series (1961Q1–1986Q3)

8.2 New Keynesian Phillips curve

In this section, we consider structural breaks in New Keynesian Phillips curve model proposed Galı and Gertler (1999) and studied in Perron and Yamamoto (2015). Their model is

$$r_t = c_t + \beta_t gap_t + \gamma_t \pi_t + u_t, \tag{8}$$

where r_t – US federal funds rate, gap_t – GDP gap, π_t – inflation is the inflation rate (the sample period is 1966Q1–2015Q2).

The results are presented in Table 7. The MIP identifies 4 breaks while GFL identifies 3 breaks.

Table 7: Percentage of correct detection of one break and accuracy of changepoint estimation when in case of many breaks

	$\hat{m}$	Dates				
MIP	4	1969Q2			1980Q4	
GFL	3			1979Q4	1980Q4	
BIC	5		1970Q4		1980Q4	
LWZ	4		1970Q4		1980Q4	
Seq (trim=0.1)	1				1980Q4	
	$\hat{m}$	Dates				
MIP	4		1989Q2		2001Q1	
GFL	3				2001Q1	
BIC	5	1987Q3		1999Q1		2005Q3
LWZ	4		1989Q2		2001Q1	
Seq (trim=0.1)	1					

9 Conclusion

We propose a new way of handling change-points in econometrics based on computational advances in mixed integer optimization and we work out statistical properties for the estimator of the number of breaks, break locations and the regression coefficient in one step. The approach shows remarkable adaptivity and versatility in that it dominates the LASSO-based alternatives for common computational settings and dimensionality. In two empirical simulations, the proposed method provides additional insights, offering evidence of a larger number of breaks in New Keynesian Phillips curve in the USA during 1966-2015 and distinct and more realistic break locations for real interest rates in 1961-1986.

References

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

Behrendt, S. and Schweikert, K. (2021). A note on adaptive group lasso for structural break time series. *Econometrics and Statistics*, 17:156–172.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2):813–852.

Bertsimas, D., Pauphilet, J., and Parys, B. V. (2020). Sparse Regression: Scalable Algorithms and Empirical Performance. *Statistical Science*, 35(4):555 – 578.

Bleakley, K. and Vert, J.-P. (2011). The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*.

Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2014). Group lasso for structural break time series. *Journal of the American Statistical Association*, 109(506):590–599.

Galı, J. and Gertler, M. (1999). Inflation dynamics: A structural econometric analysis. *Journal of monetary Economics*, 44(2):195–222.

Garcia, R. and Perron, P. (1996). An analysis of the real interest rate under regime shifts. *The review of economics and statistics*, pages 111–125.

Gómez, A. and Prokopyev, O. A. (2021). A mixed-integer fractional optimization approach to best subset selection. *INFORMS Journal on Computing*, 33(2):551–565.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, 35(4):579 – 592.

Hazimeh, H., Mazumder, R., and Radchenko, P. (2023). Grouped variable selection with discrete optimization: Computational and statistical perspectives. *The Annals of Statistics*, 51(1):1–32.

Kaddoura, Y. and Westerlund, J. (2023). Estimation of panel data models with random interactive effects and multiple structural breaks when t is fixed. *Journal of Business & Economic Statistics*, 41(3):778–790.

Mazumder, R., Radchenko, P., and Dedieu, A. (2023). Subset selection with shrinkage: Sparse linear modeling when the snr is low. *Operations Research*, 71(1):129–147.

Perron, P. and Yamamoto, Y. (2015). Using ols to estimate and test for structural changes in models with endogenous regressors. *Journal of Applied Econometrics*, 30(1):119–144.

Qian, J. and Su, L. (2016). Shrinkage estimation of regression models with multiple structural changes. *Econometric Theory*, 32(6):1376–1433.

Rebennack, S. and Krasko, V. (2020). Piecewise linear function fitting via mixed-integer linear programming. *INFORMS Journal on Computing*, 32(2):507–530.

Roberts, S. (1966). A comparison of some control chart procedures. *Technometrics*, pages 411–430.

Shiryaev, A. (1963). On optimum methods in quickest detection problems. *Probability Theory Application*, 8:22–46.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Appendix

Theoretical assumptions

We now state Assumptions A1 and A2 from Qian and Su (2016). We denote by $\mu_{\max}$ and $\mu_{\min}$ the largest and the smallest eigenvalues, respectively, of a symmetric matrix A.

Assumption A1.

(i) $\{(x_t, u_t)\}$ is a strong mixing process with mixing coefficients $\alpha(\cdot)$ satisfying $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $\rho \in (0,1)$. $E(x_t u_t) = 0$ for each t.

(ii) Either one of the following two conditions is satisfied: (a) $\sup_{t\geq 1} E\|x_t\|^{4q} < \infty$ and $\sup_{t\geq 1} E|u_t|^{4q} < \infty$ for some $q > 1$;(b) There exist some constants c_{xx}, c_{xu} c_{uu} such that $\sup_{t\geq 1} E\big[\exp(c_{xx}\|x_t\|^{2\gamma})\big] \leq C_{xx} < \infty$, $\sup_{t\geq 1} E\big[\exp(c_{xu}\|x_t u_t\|^{\gamma})\big] \leq C_{xu} < \infty$ and $\sup_{t\geq 1} E\big[\exp(c_{uu}|u_t|^{2\gamma})\big] \leq C_{uu} < \infty$ for some $\gamma \in (0, \infty]$. The case $\gamma = \infty$ is understood as uniform boundedness of $\|(x_t, u_t)\|$.

Assumption A2.

(i) There exist two positive constants $\underline{c}_{xx}$ and $\overline{c}_{xx}$ and a positive sequence $\{\delta_T\}$ declining to zero as $T \to \infty$ such that

$$\begin{aligned} \underline{c}_{xx} &\leq \inf_{r-s\geq T\delta_T} \mu_{\min}\Big(\frac{1}{r-s}\sum_{t=s}^{r-1} E(x_t x_t^\top)\Big) \\ &\leq \sup_{r-s\geq T\delta_T} \mu_{\max}\Big(\frac{1}{r-s}\sum_{t=s}^{r-1} E(x_t x_t^\top)\Big) \leq \overline{c}_{xx}. \end{aligned}$$

(ii) $T\delta_T$ satisfies one of the following two conditions: (a) $T\delta_T \geq c_v T^{1/q}$ for some $c_v > 0$ is A1(ii.a) is satisfied; (b) $T\delta_T \geq c_v (\log T)^{(2+\gamma)/\gamma}$ for some $c_v > 0$ if A1(ii.b) is satisfied.

We note that inequality $E\big[\exp(c_{uu}|u_t|^{2\gamma})\big] \leq C_{uu} < \infty$ is omitted from assumption A1(ii.b) of Qian and Su (2016); however, it is needed in their proofs. This inequality is the counterpart of inequality $\sup_{t\geq 1} E|u_t|^{4q} < \infty$ in assumption A1(ii.a), and is required to control the tail behavior of the error term u_t.

Preliminaries

In this subsection, we state some existing results that we will use in the proof of Theorem 1.

Lemma 1. *Suppose that Assumptions A1 and A2 hold. Then,*

(i) $\sup_{r-s\geq T\delta_T} \mu_{\max}\big(\frac{1}{r-s}\sum_{t=s}^{r-1} x_t x_t^\top\big) \leq \overline{c}_{xx} + o_p(1)$;

(ii) $\sup_{r-s\geq T\delta_T} \mu_{\min}\big(\frac{1}{r-s}\sum_{t=s}^{r-1} x_t x_t^\top\big) \geq \underline{c}_{xx} + o_p(1)$;

(iii) $\sup_{r-s\geq T\delta_T} \big\|\frac{1}{\sqrt{r-s}}\sum_{t=s}^{r-1} x_t u_t\big\| = O_p\big([\log T]^{c_\delta/2}\big)$;

(iv) $\sup_{0<r-s<T\delta_T} \sum_{t=s}^{r-1} u_t^2 = O_p(T\delta_T)$.

Parts (i) and (ii) of Lemma 1 are established in Lemma A.3 of Qian and Su (2016); part (iii) is established in their Lemma A.4; part (iv) – in the proof of their Lemma E1 (page 1425).

Proof of Theorem 1

Proof. To simplify the presentation, we will write $\gtrsim$ and $\lesssim$ to indicate that inequalities $\geq$ and $\leq$, respectively, hold up to positive universal multiplicative factors. Given a vector $\boldsymbol{\beta} = (\beta_1^\top, ..., \beta_T^\top)^\top$, with $\boldsymbol{\beta}_j \in \mathbb{R}^p$, we define

$$Q(\boldsymbol{\beta}) = \sum_{t=1}^{T} \big(y_t - \beta_t^\top x_t\big)^2.$$

Note that

$$Q(\widehat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}^*) = \sum_{t=1}^{T} \Big[(\widehat{\beta}_t - \beta_t^*)^\top (x_t x_t^\top)(\widehat{\beta}_t - \beta_t^*) - 2(\widehat{\beta}_t - \beta_t^*)^\top x_t u_t\Big]. \tag{9}$$

We will prove the three claims of Theorem 1 in sequence.

Claim 1: $\mathbf{P}(\widehat{\mathbf{m}} = \mathbf{m}^*) \to \mathbf{1}$. Using the combined set of the true and estimated breakpoints, $\{T_j^*\} \cup \{\widehat{T}_j\}$, we can divide the time interval index set into a collection of consecutive time intervals, $\{1, ..., T\} = \cup_k C_k$, so that on each such interval both the estimated and the true regression coefficients stay constant, i.e., neither $\widehat{\beta}_t$ nor β_t^* change their values for $t \in C_k$, provided that

the interval C_k is fixed. We let $\widehat{h}_k = \widehat{\beta}_t - \beta_t^*$ for $t \in C_k$, noting that this definition does not depend on the specific t as long as $t \in C_k$. Thus, we can rewrite equation (9) as follows:

$$Q(\widehat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}^*) = \sum_k \left[\widehat{h}_k^\top \Big(\sum_{t \in C_k} x_t x_t^\top \Big) \widehat{h}_k - 2 \widehat{h}_k^\top \Big(\sum_{t \in C_k} x_t u_t \Big) \right]. \tag{10}$$

We write L_k for the length of the interval C_k and define

$$\xi_T = \sup_{r-s \geq T\delta_T} \Big\| \frac{1}{\sqrt{r-s}} \sum_{t=s}^{r-1} x_t u_t \Big\|; \qquad \nu_T = \sup_{0 < r-s < T\delta_T} \sum_{t=s}^{r-1} u_t^2.$$

When $L_k \geq T\delta_T$, we have

$$\widehat{h}_k^\top \Big(\sum_{t \in C_k} x_t x_t^\top \Big) \widehat{h}_k \gtrsim L_k \|\widehat{h}_k\|^2 \tag{11}$$

by Lemma 1(ii).We also have $\widehat{h}_k^\top (\sum_{t \in C_k} x_t u_t) \lesssim \sqrt{L_k} \|\widehat{h}_k\| \xi_T$, which implies

$$\widehat{h}_k^\top \Big(\sum_{t \in C_k} x_t u_t \Big) \leq c_1 L_k \|\widehat{h}_k\|^2 + c_2 \xi_T^2 \tag{12}$$

for some constants c_1, c_2, where we can choose an arbitrarily small c_1 by increasing c_2.

When $L_k < T\delta_T$, we have $\widehat{h}_k^\top (\sum_{t \in C_k} x_t u_t) \leq \big[\widehat{h}_k^\top (\sum_{t \in C_k} x_t x_t^\top) \widehat{h}_k \big]^{1/2} \big[\sum_{t \in C_k} u_t^2 \big]^{1/2}$, and hence

$$\widehat{h}_k^\top \Big(\sum_{t \in C_k} x_t u_t \Big) \leq c_1 \widehat{h}_k^\top \Big(\sum_{t \in C_k} x_t x_t^\top \Big) \widehat{h}_k + \tilde{c}_2 \nu_T, \tag{13}$$

where we can again choose an arbitrarily small c_1 by increasing $\tilde{c}_2$.

Combining inequalities (10) - (13), and using a sufficiently small c_1, we derive

$$Q(\widehat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}^*) \gtrsim -(\xi_T^2 + \nu_T)(\widehat{m} + m^*).$$

Noting that $Q(\widehat{\boldsymbol{\beta}}) + \lambda \widehat{m} \leq Q(\boldsymbol{\beta}^*) + \lambda m^*$, we then deduce that

$$\lambda \widehat{m} \leq \lambda m^* + O_p\Big([\xi_T^2 + \nu_T][\widehat{m} + m^*] \Big). \tag{14}$$

Observe that $\xi_T^2 = (\log T)^{c_\delta}$ and $\nu_T = O_p(T\delta_T)$ by Lemma 1(iii) and Lemma 1(iv), respectively. Hence, $\xi_T^2+\nu_T = O_p(T\delta_T)$ by Assumption 2(ii), and thus $(\xi_T^2+\nu_T)(m^*+1) = o_p(\lambda)$ by the assumed lower bound on λ, which, in turn, implies $[\xi_T^2 + \nu_T][\widehat{m} + m^*] = o_p(\lambda|\widehat{m} - m^*| + \lambda)$. Consequently, we can rewrite inequality (14) as

$$\widehat{m} \leq m^* + o_p(|\widehat{m} - m^*|) + o_p(1),$$

and hence $\widehat{m} \leq m^*$ with probability tending to one.

We will now argue by contradiction to establish that, with probability tending to one, within $I_{\min}/5$ of each true breakpoint lies an estimated breakpoint. Suppose that this is false, and hence, with positive non-vanishing probability, there exists a (randomly selected) true breakpoint $T^*_{\tilde{k}}$, such that no estimated breakpoints are within $I_{\min}/5$ of $T^*_{\tilde{k}}$.

We define $\tilde{C}_- = \{1, ..., T\} \cap (T^*_{\tilde{k}} - I_{\min}/5\,,\, T^*_{\tilde{k}})$ and $\tilde{C}_+ = \{1, ..., T\} \cap [T^*\,,\, T^*_{\tilde{k}} + I_{\min}/5)$. We set $\tilde{\boldsymbol{\beta}}$ equal $\widehat{\boldsymbol{\beta}}$ for all t except the ones falling in $\tilde{C}$, where we set $\tilde{\beta}_t = \beta^*_t$. Note that

$$Q(\widehat{\boldsymbol{\beta}}) - Q(\tilde{\boldsymbol{\beta}}^*) = \sum_{t \in \tilde{C}_- \cup \tilde{C}_+} \Big[(\tilde{\beta}_t - \beta^*_t)^\top (x_t x_t^\top)(\tilde{\beta}_t - \beta^*_t) - 2(\tilde{\beta}_t - \beta^*_t)^\top x_t u_t\Big]. \tag{15}$$

Observing that the vector of estimated regression coefficients stays constant in the interval $\tilde{C}_- \cup \tilde{C}_+$, we denote this vector by $\widehat{\gamma}$. We write γ^*_1 and γ^*_2 for the true regression coefficient vectors in the intervals $\tilde{C}_-$ and $\tilde{C}_+$, respectively. Noting that the lengths of the intervals $\tilde{C}_-$, $\tilde{C}_+$ are of order $I_{\min}$ and applying inequalities (11)-(12) with $L_k \gtrsim I_{\min}$ and a sufficiently small c_1, we derive

$$\begin{aligned} Q(\widehat{\boldsymbol{\beta}}) - Q(\tilde{\boldsymbol{\beta}}) &\geq c_3 I_{\min}\big(\|\widehat{\gamma} - \gamma^*_1\|^2 + \|\widehat{\gamma} - \gamma^*_2\|^2\big) - c_4\xi_T^2 \\ &\geq (c_3/2) I_{\min} J^2_{\min} - c_4\xi_T^2, \end{aligned}$$

for some positive constants c_3 and c_4. Let $\tilde{m}$ be the number of breakpoints corresponding to $\tilde{\boldsymbol{\beta}}$, and note that $\tilde{m} \leq \widehat{m} + 2$. Because $Q(\widehat{\boldsymbol{\beta}}) + \lambda\widehat{m} \leq Q(\tilde{\boldsymbol{\beta}}) + \lambda\tilde{m}$, we can then deduce that inequality

$$I_{\min} J^2_{\min} \lesssim \xi_T^2 + \lambda$$

holds with positive non-vanishing probability. Because $\xi_T^2 = O_p([\log T]^{c_\delta})$ and $\lambda/\big[J_{\min}^2 I_{\min}\big] \to 0$, we conclude that $J_{\min}^2 = O\big([\log T]^{c_\delta}/I_{\min}\big)$. Assumption A3(ii) and the lower bounds on δ_T in Assumption A2(ii) imply that $I_{\min} \to \infty$ as $T \to \infty$. Hence, applying Assumption A2(ii) again, we derive $T\delta_T = o(I_{\min})$. Consequently, the derived bound on $J_{\min}^2$ implies $J_{\min}^2 = O\big([\log T]^{c_\delta}/[T\delta_T]\big)$, which constitutes a contradiction with the lower bound imposed on $J_{\min}^2$ in Assumption A3(i).

Thus, we have established that the following two statements hold with probability tending to one: (a) $\widehat{m} \leq m^*$; and (b) within $I_{\min}/5$ of each true breakpoint $T_1^*, ..., T_{m^*}^*$ lies an estimated breakpoint. It follows directly that $\widehat{m} = m^*$, which completes the proof of claim 1.

Claim 2: $\mathbf{P}\Big(\mathbf{max_{1 \leq j \leq m^*}} \big|\widehat{\mathbf{T}}_\mathbf{j} - \mathbf{T}_\mathbf{j}^*\big| \leq \mathbf{T}\boldsymbol{\delta}_\mathbf{T}\Big) \to \mathbf{1}$. We restrict our attention to the set of probability tending to one where statements (a) and (b) in the paragraph above are satisfied. Because $\widehat{m} = m^*$ and $|\widehat{T}_j - T_j^*| \leq I_{\min}/5$ for each $j = 1, ..., m^*$, the length of the interval where the estimated coefficient vector is $\widehat{\boldsymbol{\beta}}_j$ while the true coefficient vector is $\boldsymbol{\beta}_j^*$ is at least $3I_{\min}/5$. Applying inequalities (11)-(13), collecting the terms, and taking the constants c_1 sufficiently small, we derive

$$Q(\widehat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}^*) \geq c_5 I_{\min}\Big(\sum_{j=1}^{m^*} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|^2\Big) - c_6\big(\xi_T^2 + \nu_T\big)m^*$$

for some positive constants c_5 and c_6. Recall that $\xi_T^2 + \nu_T = O_p(T\delta_T)$. Thus, taking into account $Q(\widehat{\boldsymbol{\beta}}) \leq Q(\boldsymbol{\beta}^*)$, we can conclude that

$$\sum_{j=1}^{m^*} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|^2 = O_p\left(\frac{m^* T\delta_T}{I_{\min}}\right).$$

By the assumptions imposed on λ, the right hand side of the above inequality is $o_p(J_{\min}^2)$. Consequently, and because $\min_{j \neq k} \|\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_k^*\|^2 \geq J_{\min}^2$, we arrive at $\min_{j \neq k} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_k^*\|^2 \geq J_{\min}^2/5$.

We will now argue by contradiction to establish that $\max_{1 \leq j \leq m^*} |\widehat{T}_j - T_j^*| \leq T\delta_T$ with probability tending to one. Suppose that this is false, and hence, with positive non-vanishing probability, there exists a (randomly selected) true breakpoint $T_{\tilde{k}}^*$, such that $|\widehat{T}_{\tilde{k}} - T_{\tilde{k}}^*| > T\delta_T$. For concreteness, suppose that $\widehat{T}_{\tilde{k}} > T_{\tilde{k}}^*$. The complimentary case can be handled by nearly identical arguments with minor notational modifications. To simplify the presentation, we will write $\widehat{\gamma}_1$ for the estimated

regression coefficient vector in the interval $(\widehat{T}_{\tilde{k}-1}, \widehat{T}_{\tilde{k}})$ and write $\widehat{\gamma}_2$ for the estimated regression coefficient vector in the interval $(\widehat{T}_{\tilde{k}}, \widehat{T}_{\tilde{k}+1})$. Similarly, we use γ_1^* for the true regression coefficients in $(T^*_{\tilde{k}-1}, T^*_{\tilde{k}})$ and γ_2^* for the ones in $(T^*_{\tilde{k}-1}, T^*_{\tilde{k}+1})$. Let $\tilde{\boldsymbol{\beta}}$ equal $\widehat{\boldsymbol{\beta}}$ for all t except the ones in the interval $(T^*_{\tilde{k}}, \widehat{T}_{\tilde{k}})$, where we set $\tilde{\beta}_t = \widehat{\gamma}_2$. Note that the number of breakpoints corresponding to $\tilde{\boldsymbol{\beta}}$ is still m^*, and hence $Q(\widehat{\boldsymbol{\beta}}) \leq Q(\tilde{\boldsymbol{\beta}})$. Consequently, applying inequalities (11)-(13) with a sufficiently small c_1 once again, and collecting the terms, we deduce that inequality

$$L_T \|\widehat{\gamma}_1 - \gamma_2^*\|^2 = O_p\left([\log T]^{c_\delta} + L_T \|\widehat{\gamma}_2 - \gamma_2^*\|^2\right), \tag{16}$$

where $L_T = |\widehat{T}_{\tilde{k}} - T^*_{\tilde{k}}|$, holds with positive non-vanishing probability. We showed earlier that $\|\widehat{\gamma}_2 - \gamma_2^*\|^2 = o_p(J^2_{\min})$ and $\|\widehat{\gamma}_1 - \gamma_2^*\|^2 \geq J^2_{\min}/5$. Hence, inequality (16) gives

$$J^2_{\min} = O\left(\frac{[\log T]^{c_\delta}}{L_T}\right),$$

which contradicts the lower bound imposed on $J^2_{\min}$ in Assumption A3(i) because $L_T \geq T\delta_T$.

Claim 3: $\widehat{\boldsymbol{\alpha}}_{\mathbf{j}} - \boldsymbol{\alpha}^*_{\mathbf{j}} = \mathbf{O_p}\big([\mathbf{I}^*_{\mathbf{j}}]^{-\mathbf{1/2}}\big)$. It is only left to establish the stated rate of convergence for the regression coefficients $\widehat{\boldsymbol{\alpha}}_j$. This result follows directly from Theorem 3.1(ii) in Qian and Su (2016) after setting the ℓ_1 penalty weight (their λ parameter) to zero and recalling that $\delta_T = O\big(I^{1/2}_{\min}/T\big)$. While the result in Qian and Su (2016) has an additional assumption $m^* = O(\log T)$, an analysis of their proof reveals that this assumption is not required as long as $\max_{1\leq j\leq m^*} \left|\widehat{T}_j - T^*_j\right| \leq T\delta_T$ with probability tending to one, which is a property that we established in the previous paragraph. □

Proof of Theorem 2

Proof. Let $\widehat{\boldsymbol{\alpha}}^* = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top Y$. In the proof of their Theorem 3.6, Qian and Su (2016) show that $\widehat{\boldsymbol{\alpha}}^*$ has the asymptotic distribution specified in the statement of Theorem 2, i.e., $D(\widehat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}^*) \xrightarrow{d} N\big(\mathbf{0}, \Psi^{-1}\Phi\Psi^{-1}\big)$. Consequently, to complete the proof, it is sufficient to establish

$$\widehat{D}(\widehat{\boldsymbol{\alpha}}_{m^*} - \boldsymbol{\alpha}^*) - D(\widehat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}^*) = o_p(1). \tag{17}$$

The above stochastic bound is derived by Qian and Su (2016) for their post-Lasso estimator in the proof of their Theorem 3.6. However, an analysis of the proof reveals that, under our imposed assumptions, bound (17) holds for any estimator $\tilde{\boldsymbol{\alpha}}$ of the form $\tilde{\boldsymbol{\alpha}} = \left(\tilde{\mathbb{X}}^\top \tilde{\mathbb{X}}\right)^{-1} \tilde{\mathbb{X}}^\top Y$, where $\tilde{\mathbb{X}} = \text{diag}(\tilde{\mathbb{X}}_1, ..., \tilde{\mathbb{X}}_{m^*+1})$ and $\tilde{\mathbb{X}}_j = (x_{\tilde{T}_{j-1}}, ..., x_{\tilde{T}_j - 1}^\top)$, such that $P(\max_{1 \leq j \leq m^*} \left|\tilde{T}_j - T_j^*\right| \leq T\delta_T) \to 1$ as $T \to \infty$. By Theorem 1, this condition is satisfied for our estimator $\widehat{\boldsymbol{\alpha}}_{m^*}$. □