Variable Selection in Panel Models with Breaks

Simon C. Smith[a], Allan Timmermann[b], Yinchu Zhu[c]

[a] *USC Dornsife INET, Department of Economics, USC*
[b] *University of California, San Diego*
[c] *University of Oregon*

Draft: November 6, 2018

Abstract

We develop a Bayesian approach that performs variable selection in panel regression models affected by breaks. Our approach enables deactivation of pervasive regressors and activation of weak regressors for short periods (regimes). We establish theoretical results on the concentration properties of the posterior as well as the rate of convergence for estimating the break dates. Our methodology is demonstrated in simulations and in an empirical application to firms' choice of capital structure. We find that ignoring breaks can lead to overestimating the number of relevant regressors, but also a failure to activate regressors that are informative only in short-lived regimes.

Keywords: Variable selection, Structural breaks, Panel data, Bayesian analysis, High-dimensional modeling, Firms' Choice of Capital Structure

JEL classifications: G10, C11, C15

Email address: atimmermann@ucsd.edu. Correspondence to: UC San Diego, Rady School of Management, 9500 Gilman Drive, La Jolla, CA 92093-0553, United States. Phone: (858) 5340894. Fax: (858) 5340745. (Allan Timmermann)

We are grateful to the editor (Eric Ghysels) and two anonymous referees for constructive comments.

1. Introduction

Variable selection procedures are in widespread use throughout economics, being employed for many applications such as selecting which variables to include in a predictive regression (Pesaran and Timmermann 2000; Jochmann et al. 2010), choice of lag length in autoregressive models (Marcellino et al. 2006), selecting variables in a Vector Autoregressive model (Korobilis 2013), or determining the number of factors (Bai and Ng 2002). Much of the literature on variable selection assumes that the underlying data generating process is stable and uses this assumption to establish properties of the selected model such as asymptotic consistency, see, e.g., Leeb and Pötscher (2005). Far less is known about model selection in the type of unstable environment found empirically to characterize many economic time series Stock and Watson (1996); Pesaran and Timmermann (2002); Rossi (2013).[1]

Ignoring model instability could adversely affect variable selection with consequences in areas such as economic forecasting. For example, previously strong predictor variables may no longer have predictive power over outcomes or, conversely, new predictors may gain strength following a structural break. Failing to account for either of these scenarios could reduce forecasting performance.

To address these issues requires having a procedure that locates breaks and performs regime-specific variable selection between breaks. However, conducting variable selection in the presence of breaks quickly leads to the dimension of the model space becoming very large. Estimating breaks in regression models involves a complex search across possible breakpoint locations (Bai and Perron 1998), and introducing variable selection compounds the complexity.

This paper develops a new Bayesian panel regression approach that jointly estimates an unknown number of structural breaks and performs regime-specific variable

[1] See also Andrews (1993), Bai and Perron (1998), Chib (1998), Primiceri (2005), Elliott and Müller (2006), Pesaran et al. (2006), Koop and Potter (2007), and Giordani and Kohn (2012).

selection.[2] Our approach accounts for model uncertainty. This is important because uncertainty about which variables have explanatory power at a given point in time tends to be greater, the more often a process is affected by breaks and, thus, the fewer time-series observations are available for variable selection and parameter estimation.

We provide theoretical results on the frequentist properties of posteriors of Bayesian panel break models for a large class of priors and specifications. We show that with high probability, the posterior concentrates on a small neighborhood of the true parameter and characterize the rate at which this neighborhood shrinks. Our results are different from typical Bayesian analysis in that we do not assume that the likelihood is correctly specified. Although we adopt a Gaussian likelihood, our theoretical results only require mild moment conditions on the errors which are allowed to have weak cross-sectional and serial dependence. However, when the likelihood is misspecified, we recommend that frequentist methods be used for inference.[3]

Our theoretical analysis also does not impose restrictions on the growth rate of n (the cross-sectional dimension) and T (the time-series dimension). The main theoretical result (Theorem 1) does not require restrictions on the duration of the regimes or the magnitude of any breaks. The relative size of n and T and lengths of the regimes directly enter the conclusions of the theoretical results and thus have implications for the rate of convergence.[4] We show that in the classical setup, the rate of convergence for detecting the break dates matches the optimal rate up to a logarithm factor.

[2] Papers that estimate breaks in panels includes Bai et al. (1998), Bai (2010), Baltagi et al. (2016), and Smith and Timmermann (2017b) (see also Smith and Timmermann (2017a)).

[3] When the likelihood is misspecified, Bayesian credible sets based on the posterior are not valid confidence sets in the frequentist sense and fail to have the right coverage, see, e.g., Royall and Tsou (2003); Kleijn and Van der Vaart (2012); Müller (2013); Bissiri et al. (2016). For this reason, the literature (e.g., Müller (2013)) has recommended that frequentist methods be used unless misspecification of the likelihood can definitely be ruled out.

[4] In empirical applications, a pre-filtering step can be used to deal with cross-sectional dependencies in the data. This introduces additional parameter estimation error and can lead to a non-vanishing bias which could be important in applications with relatively short panels. For further discussion of biases in dynamic panel models see, e.g., Nickell (1981); Anderson and Hsiao (1982); Arellano and Bond (1991); Ahn and Schmidt (1995); Phillips and Moon (1999); Hahn and Kuersteiner (2002); Alvarez and Arellano (2003); Gouriéroux et al. (2010); Han et al. (2014); Dhaene and Jochmans (2015); Liu et al. (2017).

When conditions ensuring consistent identification of the break dates hold, the estimation error for each regime depends on its duration. For long regimes, we can expect the rate $\sqrt{(nT)^{-1}\log(nT)}$ for the time-specific slope and variance parameters. For regimes that last for only one period, we can still obtain a rate of convergence of $\sqrt{n^{-1}\log(nT)}$. We expect average estimation errors to be small in long regimes and larger in short-lived regimes.

We next develop a framework that preserves conjugacy and demonstrate how to estimate the resulting model using a four-step procedure. First, we estimate the parameters conditional on the existing estimates of the breakpoint vector and selected variables. Second, conditional on the existing parameter and breakpoint estimates, variable selection is performed within each regime. Third, conditional on the parameter estimates and selected variables, the existing breakpoint estimates are perturbed to help them converge to the true break dates. The fourth step jointly estimates the number and timing of breakpoints and performs regime-specific variable selection.

Our approach employs the reversible jump Markov chain Monte Carlo algorithm of Green (1995). We specify conjugate priors on the regression parameters which enables them to be integrated out of the posterior, enhancing the mixing and considerably reducing the computational burden. Computational efficiency is crucial due to the high-dimensional nature of our search problem.

The usefulness of our methodology is demonstrated on a simulated data set in which, by construction, we know the underlying break dates and the informative variables in each regime. We show how regime-specific variable selection is successfully implemented in the presence of breaks. We also illustrate our method in an empirical application in the field of corporate finance that provides new insights into which variables can help explain variation in the leverage ratio of corporations.

Our paper is related to a number of previous studies. Bai (2010) considers common breaks in the mean and variance of panel models while Baltagi et al. (2016) study the effect of common breaks on frequentist estimation and inference in heterogeneous panel regression models. Smith and Timmermann (2017b) introduce the idea of using a reversible jump Markov chain Monte Carlo estimation approach to panel models with common break dates but heterogeneous slope coefficients and apply this approach to forecasting inflation in the European Union. They do not consider variable selection, nor perform any theoretical analysis of the asymptotic properties of the Bayesian panel estimator.[5]

The remainder of the paper is set out as follows. Section 2 conducts our theoretical analysis. Section 3 lays out the methodology, model and prior distributions, while Section 4 details model estimation and variable selection. Sections 5 and 6 cover the simulation study and the empirical application, and Section 7 concludes. Technical proofs and additional material are covered in supplemental material appendices.

2. Breaks in Bayesian Panel Models: A Theoretical Framework

This section provides theoretical results on the frequentist properties of a broad class of Bayesian panel break estimators. Our main theoretical results hold in finite samples and thus no assumptions are imposed on the relative size of the cross-sectional and time-series dimensions (n and T).

We introduce the following notations that will be used in the rest of the paper. For a vector $x = (x_1, ..., x_p)'$ and $r \geq 1$, we define $\|x\|_r = (\sum_{j=1}^{p} |x_j|^r)^{1/r}$. We use $\circ$ to denote the Hadamard product (entry-wise product) and $\otimes$ to denote Kronecker product. Indicator functions are denoted by $\mathbf{1}\{\cdot\}$. $\|\cdot\|_F$ denotes the Frobenius norm.

[5] Papers that provide test statistics or estimation approaches to identify breaks for multivariate time series or panels include Bai et al. (1998), Qu and Perron (2007), Bai and Carrion-I-Silvestre (2009), Kim (2011), and Baltagi et al. (2017).

2.1. Setup and notations

Consider the panel regression model

$$y_{it} = X_{it}'\beta_t + \varepsilon_{it} \qquad \text{for } 1 \leq i \leq n,\ 1 \leq t \leq T, \tag{1}$$

where $X_{it} \in \mathbb{R}^r$. ε_{it} is uncorrelated with X_{it} and satisfies $E\varepsilon_{it} = 0$ and $E\varepsilon_{it}^2 = \sigma_t^2$.

We assume that the model parameters $\{(\beta_t, \sigma_t)\}_{t=1}^T$ exhibit structural breaks. We capture such breaks through mappings $t \mapsto \beta_t$ and $t \mapsto \sigma_t$ which are piecewise constant functions in $\mathbb{R}^r$ and $\mathbb{R}$, respectively and parameterize these functions as follows. Let $\theta \in \mathbb{R}^{d_\Theta}$ with $d_\Theta = r(K+1) + 2K + 1$ denote the parameter vector

$$\theta = (\alpha_1, \alpha_2, ..., \alpha_{K+1}, \omega_1, \omega_2, ..., \omega_{K+1}, \lambda_1, \lambda_2, ..., \lambda_K),$$

where $\alpha_1, ..., \alpha_{K+1} \in \mathbb{R}^r$ denote the values for β_t in the $K+1$ regimes, $\omega_1, ..., \omega_{K+1} > 0$ denote the standard deviations of the residuals in the $K+1$ regimes and $\lambda_1, ..., \lambda_K \in [0, 1]$ denote the break dates as a fraction of T:

$$\beta_t(\theta) = \sum_{j=1}^{K+1} \alpha_j \mathbf{1}\{\lambda_{j-1}T < t \leq \lambda_j T\} \quad \text{and} \quad \sigma_t(\theta) = \sum_{j=1}^{K+1} \omega_j \mathbf{1}\{\lambda_{j-1}T < t \leq \lambda_j T\},$$

where $\lambda_0 = 0$ and $\lambda_{K+1} = 1$. The log-likelihood takes the following form:

$$\ell_n(\mathbf{Z}, \theta) = -\frac{nT}{2}\log(2\pi) - n\sum_{t=1}^T \log\sigma_t(\theta) - \sum_{i=1}^n \sum_{t=1}^T \frac{(y_{it} - X_{it}'\beta_t(\theta))^2}{2\sigma_t^2(\theta)}, \tag{2}$$

where $\mathbf{Z}$ denotes the observed data $\{(y_{it}, X_{it})\}_{1\leq i\leq n,\ 1\leq t\leq T}$. The likelihood then takes the form $p(\mathbf{Z} \mid \theta) = \exp(\ell_n(\mathbf{Z}, \theta))$. We do not impose any distributional assumption on the error terms although we adopt a Gaussian likelihood specification in equation (2). In Assumption 2 that follows, we only impose mild conditions on moments and weak dependence for the errors. Our theoretical results show that the proposed

Bayesian method retains certain frequentist properties even if the likelihood model is misspecified. This differs from the usual Bayesian setup that assumes correct specification of the likelihood. To properly describe the prior and posterior, we introduce the parameter space Θ, which is assumed to be a compact subset of $\mathbb{R}^{d_\Theta}$. The true value of the parameter θ is denoted by θ_*. The prior, denoted by Π, is a non-random probability measure on Θ. Our theory allows for a large class of priors and we do not assume that the prior probability measure Π admits a density. We follow the standard framework of analyzing frequentist properties of Bayesian procedures and consider a general prior which is only assumed to be a probability distribution; see, e.g., Ghosal et al. (2000); Shen and Wasserman (2001); Ghosal and van der Vaart (2007); Lian (2010).

Combining the likelihood and the prior, we obtain the posterior as a random probability measure on the parameter space Θ. For any subset $A \subset \Theta$, we define

$$\Pi(A \mid \mathbf{Z}) = \frac{\int_A p(\mathbf{Z} \mid \theta) d\Pi(\theta)}{\int_\Theta p(\mathbf{Z} \mid \theta) d\Pi(\theta)}.$$

If Π has a density $d\Pi(\theta) = \pi(\theta) d\theta$, the posterior also has a density $d\Pi(\theta \mid \mathbf{Z}) = \mathrm{C} \times p(\mathbf{Z} \mid \theta)\pi(\theta)d\theta$, where $\mathrm{C} = 1/\int_\Theta p(\mathbf{Z} \mid \theta)\pi(\theta)d\theta$ so the posterior integrates to one.

2.2. Theoretical results

We next establish theoretical properties for the posterior $\Pi(\cdot \mid \mathbf{Z})$. Our analysis is quite different from the Bernstein-von-Mises results for regular models in which the likelihood is smooth in its parameters. First, we do not assume that the likelihood is correctly specified, whereas the Bernstein-von-Mises results, which state that the posterior distribution is close to the asymptotics of the maximum likelihood estimator (see, e.g., Theorem 10.1 of Van der Vaart (2000)), require correct specification of the likelihood. Since the functional form of the likelihood in equation (2) can be misspecified, the information equalities used in proving Bernstein-von-Mises results might not hold. Second, the likelihood model we consider is not smooth in θ since breaks lead to a non-differentiable likelihood. Conventional methods for analyzing regular likelihoods therefore do not apply.

Our general results do not impose restrictions on the break size and the length of the shortest regime. Thus, we should not expect a standard asymptotic distribution even for frequentist estimators and so we focus on the concentration properties of the posterior. As a random probability measure on Θ, $\Pi(\cdot \mid \mathbf{Z})$ has mass equal to one on the entire Θ. We show that with probability close to one, this random probability measure assigns almost all mass on a shrinking neighborhood of θ_* and characterize the size of this neighborhood. First, we introduce some regularity conditions.

Assumption 1. *Let Θ be a compact set with K being fixed. There exists a constant $M_0 > 0$ such that $\left\{\theta: \max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\sigma_t(\theta)-\sigma_t(\theta_*)\right)^2, \frac{1}{T}\sum_{t=1}^{T}\|\beta_t(\theta)-\beta_t(\theta_*)\|_2^2\right\} \leq M_0\right\} \subset \Theta$. Moreover, there exists a constant $M_1 > 0$ such that $\sigma_t(\theta) \geq M_1$ for any $\theta \in \Theta$ and $1 \leq t \leq T$.*

Assumption 1 states that Θ is compact and contains a neighborhood of θ_*. The assumption of a compact parameter space is routinely imposed in the frequentist analysis of parametric models and allows us to derive a uniform law of large numbers. Here, Assumption 1 serves a similar role in bounding the empirical process for Kullback-Leibler divergence. Before introducing our moment conditions on the residuals, recall the strong mixing coefficient: for $s \geq 1$,

$$\alpha_{mixing}(s) = \sup\Big\{|P(A\bigcap B) - P(A)P(B)|:\ A \in \sigma\left(\{(\varepsilon_{i\tau}, X_{i\tau})\}_{\tau\leq j,\ 1\leq i\leq n}\right), \\ B \in \sigma\left(\{(\varepsilon_{i\tau}, X_{i\tau})\}_{\tau\geq j+s,\ 1\leq i\leq n}\right) \text{ and } j \in \mathbb{N}\Big\},$$

where $\sigma(\cdot)$ denotes the σ-algebra generated by random variables.

Assumption 2. *There exist constants $M_2, M_3 > 0$ and $M_4, M_5 > 2$ such that $M_3 \geq$*

$M_4(M_4+2)/(M_4-2)$, $\alpha_{mixing}(t) \leq M_2 t^{-M_3}$ for all $t \geq 1$, $E|n^{-1/2}\sum_{i=1}^n(\varepsilon_{it}^2 - E\varepsilon_{it}^2)|^{M_4} \leq M_5$ and $E\|n^{-1/2}\sum_{i=1}^n X_{it}\varepsilon_{it}\|_2^{M_4} \leq M_5$.

Assumption 2 imposes very mild conditions on the moments and weak dependence of the error terms in both the cross-sectional and time-series dimensions. The assumption also allows for misspecification of the likelihood. Although the likelihood specification in equation (2) is borrowed from the setting of independent Gaussian ε_{it}, we require neither normality, nor independence of the residuals which are allowed to have fatter tails than normal distributions and can have weak dependence.

Moreover, we allow for lagged dependent variables. As long as $EX_{it}\varepsilon_{it} = 0$, we should expect $n^{-1/2}\sum_{i=1}^n X_{it}\varepsilon_{it} = O_P(1)$ under weak cross-sectional dependence.[6] The following result provides a simple sufficient condition for the moment bounds in Assumption 2 in a setup with weak factors.

Lemma 1. *Suppose that $X_{it} = \Phi_i' f_t + v_{it}$ and $\varepsilon_{it} = \phi_i' f_t + u_{it}$. Assume that $f_t \in \mathbb{R}^{d_f}$ satisfies $E\|f_t\|_2^{C_1} \leq C_2$ and $\{(\Phi_i, \phi_i)\}_{i=1}^n$ is non-random with $\max_{1\leq i\leq n}\|\Phi_i\|_F \leq C_3$, $\max_{1\leq i\leq n}\|\phi_i\|_2 \leq C_3$, $\|n^{-1/2}\sum_{i=1}^n \phi_i\phi_i'\|_F \leq C_3$ and $\|n^{-1/2}\sum_{i=1}^n \phi_i \otimes \Phi_i\|_2 \leq C_3$ for some constants $C_1, C_2, C_3 > 0$. Moreover, assume that $\{(v_{it}, u_{it})\}_{i=1}^n$ is independent across i with $Eu_{it} = 0$, $Ev_{it} = 0$, $Ev_{it}u_{it} = 0$ and satisfies that $E|u_{it}|^{2C_4} \leq C_5$, $E\|v_{it}\|_2^{2C_4} \leq C_5$ and $E\|v_{it}u_{it}\|_2^{C_4} \leq C_5$. Assume that $C_1 \geq 2C_4 > 1$. Then there exists a constant $C_6 > 0$ depending only on $C_1, ..., C_5$ such that $E|n^{-1/2}\sum_{i=1}^n(\varepsilon_{it}^2 - E\varepsilon_{it}^2)|^{C_4} \leq C_6$ and $E\|n^{-1/2}\sum_{i=1}^n X_{it}\varepsilon_{it}\|_2^{C_4} \leq C_6$.*

Lemma 1 assumes the weak factor setup considered by Kleibergen (2009) and Onatski (2012) among others.[7] We next impose a regularity condition for identification which says that, even in the absence of breaks in the volatility of the residuals, changes in the slope coefficients can be identified through the likelihood.

[6] In this case, unit root processes for the dependent variable are ruled out since $n^{-1/2}\sum_{i=1}^n X_{it}\varepsilon_{it}$ is required to have bounded moments.

[7] Identification assumptions similar to Assumption 3 are routinely imposed in the literature; see, for example, Assumption A2 of Bai and Perron (1998) and Assumption A1 of Oka and Perron (2018).

Assumption 3. *For $\theta \in \Theta$, define $\delta_t = \delta_t(\theta) = \beta_t(\theta) - \beta_t(\theta_*)$. Assume that with probability at least $1 - \gamma_n$,*

$$M_6 \leq \inf_{\theta \in \Theta} \frac{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it}'\delta_t)^2}{T^{-1}\sum_{t=1}^{T}\|\delta_t\|_2^2} \leq \sup_{\theta \in \Theta} \frac{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it}'\delta_t)^2}{T^{-1}\sum_{t=1}^{T}\|\delta_t\|_2^2} \leq M_7,$$

where $M_6, M_7 > 0$ are constants and $\gamma_n = o(1)$.

Finally, we introduce a regularity condition which requires that the prior should allocate non-trivial mass on neighborhoods of the true parameter. Intuitively, if the prior puts zero weights around the true parameter, the posterior cannot concentrate on the true parameter even in large samples:

Assumption 4. *There exist constants $M_8, M_9, M_{10} > 0$ such that for any $x \in (0, M_8)$, $\Pi\left(\left\{\theta \in \Theta : \max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\sigma_t(\theta) - \sigma_t(\theta_*)\right)^2, \frac{1}{T}\sum_{t=1}^{T}\|\beta_t(\theta) - \beta_t(\theta_*)\|_2^2\right\} \leq x\right\}\right) \geq M_9 x^{M_{10}}$.*

Assumption 4 is satisfied by many priors and is also imposed by Lian (2010). A sufficient condition for Assumption 4 is that in a neighborhood of θ_*, Π admits a density that is bounded away from zero. If the prior of $(\alpha_1, ..., \alpha_{K+1})$ given $(\lambda_1, ..., \lambda_K)$ has density bounded away from zero, then Assumption 4 holds.

We can now establish the first theoretical result which is a finite-sample bound that characterizes the concentration properties of the posterior in estimating the entire paths of the slope coefficients $\{\beta_t\}_{t=1}^{T}$ and the volatility $\{\sigma_t\}_{t=1}^{T}$:

Theorem 1. *Let Assumptions 1, 2, 3 and 4 hold. Then for any $c \in (0, 1)$, there exists a constant $D_c > 0$ such that with probability at least $1 - c - \gamma_n$,*

$$\Pi\left(B_n^c \mid \mathbf{Z}\right) \geq 1 - D_c\left(nT\right)^{-M_{10}/2},$$

where $B_n^c = \Theta \backslash B_n$ and

$$B_n = \left\{\theta \in \Theta : \max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\sigma_t(\theta) - \sigma_t(\theta_*)\right)^2, \frac{1}{T}\sum_{t=1}^{T}\|\beta_t(\theta) - \beta_t(\theta_*)\|_2^2\right\} \geq \tilde{M}\frac{\log(nT)}{nT}\right\}$$

and $\tilde{M} > 0$ *are constants depending only on* $K, \Theta, M_0, M_1, ..., M_{10}$.

Theorem 1 says that, with high probability, the posterior allocates almost all mass on the set B_n^c. Essentially, the average squared estimation error converges to zero at rate $(nT)^{-1}\log(nT)$. The rate of convergence is optimal up to the logarithm factor. Even for panel models with constant slope coefficient and errors that are i.i.d across time and units, the pooled estimator converges at the rate $1/(nT)$. The logarithm factor is present in many frequentist studies of Bayesian procedures, e.g., van der Vaart and van Zanten (2008); Lian (2010); Rousseau (2010); Giné and Nickl (2011).

Theorem 1 holds regardless of the relative magnitude of n and T. In practice, this means that the average (across t) estimation error decays to zero if at least one of n and T tends to infinity. In this sense, Theorem 1 states that information from both dimensions of the panel data helps in identifying the model parameter that can potentially have finitely many breaks.

Theorem 1 establishes the near-optimal rate without any assumption regarding the duration of regimes or the break size. This is different from usual analyses of frequentist procedures which typically establish the asymptotic properties of estimators for the slope coefficients under the assumption that regime durations are large enough to consistently identify the break date; see e.g., Bai and Perron (1998); Baltagi et al. (2016); Cheng et al. (2016). Many frequentist methods have been proposed for inference of break dates; these methods typically have a requirement on the break size, such as Bai and Perron (1998); Qu and Perron (2007); Bai (2010); Oka and Perron (2018). Instead, we study estimation of the time trajectories $t \mapsto \beta_t$ and $t \mapsto \sigma_t$ and derive the rate of convergence without knowledge of regime lengths or magnitude of the breaks.

We note that Theorem 1 is a statement on estimating the vectors $\beta_* = (\beta_1(\theta_*)', ..., \beta_T(\theta_*)')' \in \mathbb{R}^{rT}$ and $\sigma_* = (\sigma_1(\theta_*)', ..., \sigma_T(\theta_*)')' \in \mathbb{R}^T$ in terms of the Euclidean norm, rather than the sup norm. Therefore, the result does not imply that, for any t, the posterior would concentrate on the set $\{\theta : \|\beta_t(\theta) - \beta_t((\theta_*)\|\leq$

$C(nT)^{-1}\log(nT)\}$ for some constant $C > 0$. To derive properties for the slope and variance parameters for a specific value of t, we rely on the piecewise constant property of the path $t \mapsto \beta_t$ (and $t \mapsto \sigma_t$).

When the number of breaks is known or can be consistently estimated, we can further characterize the identification of the break dates.[8] In particular, we can translate the bound on the estimation error of the trajectory of β_t and σ_t into a bound on the estimation error of the break dates which takes into account the length of the regimes and the size of the breaks. To this end, let $\lambda(\theta) = (\lambda_1, ..., \lambda_K)$ denote the break dates (as fractions of T). Let $\theta_* = (\alpha_{*,1}, ..., \alpha_{*,K+1}, \omega_{*,1}, ..., \omega_{*,K+1}, \lambda_{*,1}, ..., \lambda_{*,K})$ denote the components of the true parameter value θ_*. We have a general result on piecewise constant functions:

Theorem 2. *Let $G = \min_{1\leq j\leq K+1}|\lambda_{*,j} - \lambda_{*,j-1}|$. For any $\theta \in \Theta$, we then have*

$$T^{-1}\sum_{t=1}^{T}\|\beta_t(\theta)-\beta_t(\theta_*)\|_2^2 \geq \min_{1\leq j\leq K+1}\|\alpha_{*,j}-\alpha_{*,j-1}\|_2^2 \times \min\{0.16G,\ 0.28\,\|\lambda(\theta)-\lambda(\theta_*)\|_\infty\}$$

and

$$T^{-1}\sum_{t=1}^{T}\|\sigma_t(\theta)-\sigma_t(\theta_*)\|_2^2 \geq \min_{1\leq j\leq K+1}\|\omega_{*,j}-\omega_{*,j-1}\|_2^2 \times \min\{0.16G,\ 0.28\,\|\lambda(\theta)-\lambda(\theta_*)\|_\infty\}.$$

Using Theorem 2, we next characterize the rate of convergence for the breaks.

Corollary 1. *Suppose the assumptions of Theorem 1 hold. Assume that $\gamma_n \to 0$ and the number of breaks is equal to K. Let $\Delta_\lambda = \min_{1\leq k\leq K+1}|\lambda_k - \lambda_{k-1}|$, $\Delta_\alpha = \min_{1\leq j\leq K+1}\|\alpha_{*,j} - \alpha_{*,j-1}\|_2^2$ and $\Delta_\sigma = \min_{1\leq j\leq K+1}\|\omega_{*,j} - \omega_{*,j-1}\|_2^2$. We have*

[8] The literature on estimation of structural break models focuses on the situation with a known number of breaks; see Bai and Perron (1998); Qu and Perron (2007); Baltagi et al. (2016) among others. Determining the number of breaks is typically done by sequentially testing for the number of breaks or by applying information criteria. The consistency of these procedures is usually proved under strong assumptions, such as large break sizes with long regimes. Within the Bayesian framework, how to test the number of breaks is still largely unknown.

1. *If* $0.16nT\Delta_\alpha\Delta_\lambda \geq \tilde{M}\log(nT)$*, then*

$$\Pi\left(\left\{\theta \in \Theta : \ \|\lambda(\theta_*) - \lambda(\theta)\|_\infty < \tilde{M}\frac{\log(nT)}{0.28\Delta_\alpha nT}\right\} \mid \mathbf{Z}\right) = 1 - o_P(1),$$

   where $\tilde{M}$ *is the constant in Theorem 1.*

2. *If* $0.16nT\Delta_\sigma\Delta_\lambda \geq \tilde{M}\log(nT)$*, then*

$$\Pi\left(\left\{\theta \in \Theta : \ \|\lambda(\theta_*) - \lambda(\theta)\|_\infty < \tilde{M}\frac{\log(nT)}{0.28\Delta_\sigma nT}\right\} \mid \mathbf{Z}\right) = 1 - o_P(1).$$

Corollary 1 characterizes how the rate of convergence for estimating $\lambda(\theta_*)$ depends on the duration of regimes (Δ_λ) and the break size (Δ_α and Δ_σ). For long regimes and large break sizes, the rate of convergence for $\lambda(\theta_*)$ is $(nT)^{-1}\log(nT)$ which is again optimal up to the logarithm factor.[9] Thus, when $n = 1$, the rate of convergence of our procedure is $T^{-1}\log T$. Having multiple time series makes the rate of convergence faster. This is particularly valuable when regimes can be short. Since the break dates can only take integer values, the shortest regime can last for only one period, i.e., $\Delta_\lambda = T^{-1}$. Consistently recovering the corresponding break date (i.e., $\|\lambda(\theta_*) - \lambda(\theta)\|_\infty < T^{-1}$) requires that $\max\{\Delta_\alpha, \Delta_\sigma\} > \tilde{M}(0.16n)^{-1}\log(nT)$. This highlights the benefit of having panel data since $\tilde{M}(0.16n)^{-1}\log(nT)$ decreases in n.

Corollary 1 highlights both the mechanisms and tradeoffs for identifying the break dates. First, the length of the regimes and the break size complement each other in detecting break dates. When the minimum length of each regime is large, i.e., Δ_λ is bounded away from zero, the break size does not need to be very large for consistently estimating $\lambda(\theta_*)$; we only need Δ_α or Δ_σ to be larger than $O\left((nT)^{-1}\log(nT)\right)$. Corollary 1 also allows the length of the shortest regime to be small. Consistency is achievable for $\lambda(\theta_*)$ whenever the product of the shortest regime duration and the smallest break size is larger than $O\left((nT)^{-1}\log(nT)\right)$.

[9] For a single time series, Bai and Perron (1998) obtain the rate of convergence T^{-1}.

Second, our setup uses both the slope coefficients and volatility parameters to detect break locations. This is an attractive feature since we do not need explicit knowledge of which channel will have more identification power and so could improve the empirical appeal of our procedure.

Corollary 1 provides a sufficient condition for consistent identification of the break locations. Since $\{\lambda_j(\theta)\}_j$ only takes values on a grid with increments equal to $1/T$, exact recovery of the break locations occurs whenever estimation errors of $\lambda(\theta)$ is smaller than $1/(3T)$. For the same reason, $\Delta_\lambda \geq 1/T$. This yields the next result:

Corollary 2. *Let* $\mathcal{Q} = \{\theta \in \Theta : \ \|\lambda(\theta) - \lambda(\theta_*)\|_\infty = 0\}$. *Suppose the assumptions of Theorem 1 hold. Assume that* $\gamma_n \to 0$ *and the number of breaks is equal to* K. *If* $\max\{\Delta_\sigma, \Delta_\alpha\} \geq 11\tilde{M}\log(nT)/n$, *then* $\Pi_n(\mathcal{Q} \mid \mathbf{Z}) = 1 - o_P(1)$.

Corollary 2 provides sufficient conditions for consistently recovering break locations, which in turn has direct implications for the estimation error each time period. Since the estimation error is constant within each regime, the average estimation error in Theorem 1 can be written as a weighted sum of K different estimation errors:

Corollary 3. *Suppose the assumptions of Theorem 1 hold. Assume that* $\gamma_n \to 0$ *and the number of breaks is equal to* K. *Also assume that* $\max\{\Delta_\sigma, \Delta_\alpha\} \geq 11\tilde{M}\log(nT)/n$. *Then* $\Pi_n(\bigcap_{1\leq j\leq K} G_{n,j} \mid \mathbf{Z}) = 1 - o_P(1)$, *where*

$$G_{n,j} = \left\{\theta \in \Theta : \max_{\lambda_{j-1}T<t\leq\lambda_j T} \max\left\{|\sigma_t(\theta) - \sigma_t(\theta_*)|\,, \|\beta_t(\theta) - \beta_t(\theta_*)\|_2\right\} \leq \frac{\bar{M}\log(nT)}{nT(\lambda_j - \lambda_{j-1})}\right\}$$

and $\bar{M} > 0$ *is a constant.*

Corollary 3 states that when sufficient conditions for consistent identification of the break dates are imposed, the estimation error in each regime depends on the duration of that regime. For long regimes ($\lambda_j - \lambda_{j-1}$ is bounded away from zero), we can expect the rate $\sqrt{(nT)^{-1}\log(nT)}$ for the slope and variance parameters. For regimes that last for only one period (i.e., $\lambda_j - \lambda_{j-1} = 1/T$), we can still obtain a rate of convergence of $\sqrt{n^{-1}\log(nT)}$. In this sense, loosely speaking, we would expect average estimation errors to be small in long regimes and larger in short-lived regimes.

2.3. Effect of Prefiltering the Data

When there is strong cross-sectional or serial dependence in the error terms, Assumption 2 could be violated. In these situations, it can be beneficial to prefilter the data such that the weak dependence assumption in the errors approximately holds. A concern that arises with prefiltered data is whether estimation errors in the prefiltering stage poses a problem for the previous theoretical results.

Example 1 (Common Time Trend)**.** Suppose we observe data $\{(\tilde{y}_{it}, \tilde{X}_{it})\}_{1\leq i\leq n,\ 1\leq t\leq T}$ from the following model: $\tilde{y}_{it} = \tilde{X}_{it}'\beta_t + \varepsilon_{it}$, where $\varepsilon_{it} = \mu_t + u_{it}$. In this case, we do not, in general, have weak cross-sectional dependence because at each time t, all the cross-sectional units are partly driven by μ_t. An obvious fix would be to remove the common time trend μ_t by subtracting $\hat{\mu}_t = n^{-1}\sum_{j=1}^{n}\tilde{y}_{jt}$ to get $\hat{y}_{it} = \tilde{y}_{it} - \hat{\mu}_t$. After some algebra, we can show that $\hat{y}_{it} = \hat{X}_{it}'\beta_t + u_{it} - e_t$, where $\hat{X}_{it} = \tilde{X}_{it} - \bar{\mu}_{X,t}$, $e_t = n^{-1}\sum_{j=1}^{n}u_{jt}$ and $\bar{\mu}_{X,t} = n^{-1}\sum_{j=1}^{n}\tilde{X}_{jt}$. In this case, although we estimate the model in (1), the prefiltered data is $\{(\hat{y}_{it}, \hat{X}_{it})\}_{1\leq i\leq n,\ 1\leq t\leq T}$. Hence, we will compute the log-likelihood in equation (2) using the prefiltered data.

Example 2 (Factor structure)**.** Suppose we observe data $\{(\check{y}_{it}, \check{X}_{it})\}_{1\leq i\leq n,\ 1\leq t\leq T}$ from the following model: $\check{y}_{it} = \check{X}_{it}'\beta_t + \varepsilon_{it}$, where the error term is given by $\varepsilon_{it} = \phi_i' f_t + u_{it}$ and f_t and ϕ_i are unobserved factors and factor loadings, respectively.

When $n^{-1}\sum_{i=1}^{n}\phi_i\phi_i'$ and $T^{-1}\sum_{t=1}^{T}f_tf_t'$ have eigenvalues bounded away from zero, we have a strong factor structure and might want to prefilter the data using principal component analysis (Bai 2003, 2009) or common correlated effects (Pesaran 2006).

To formally study the problem of prefiltering the data, consider again the model in equation (1), where the unobserved variables $\mathbf{Z} = \{(y_{it}, X_{it})\}_{1\leq i\leq n,\ 1\leq t\leq T}$ satisfy Assumptions 2 and 3. However, suppose we use the prefiltered data $\hat{\mathbf{Z}} =$

$\{(\hat{y}_{it}, \hat{X}_{it})\}_{1\leq i\leq n,\ 1\leq t\leq T}$. The log-likelihood for the prefiltered data is given by

$$\ell_n(\hat{\mathbf{Z}}, \theta) = -\frac{nT}{2}\log(2\pi) - n\sum_{t=1}^{T}\log\sigma_t(\theta) - \sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\left(\hat{y}_{it} - \hat{X}_{it}'\beta_t(\theta)\right)^2}{2\sigma_t^2(\theta)}.$$

Let $p(\hat{\mathbf{Z}} \mid \theta) = \exp(\ell_n(\hat{\mathbf{Z}}, \theta))$. We compute the posterior using the prefiltered data

$$\Pi(A \mid \hat{\mathbf{Z}}) = \frac{\int_A p(\hat{\mathbf{Z}} \mid \theta)d\Pi(\theta)}{\int_\Theta p(\hat{\mathbf{Z}} \mid \theta)d\Pi(\theta)} \qquad \text{for} \qquad A \subseteq \Theta.$$

The theoretical properties of $\Pi(\cdot \mid \hat{\mathbf{Z}})$, the posterior computed using prefiltered data can be analyzed in two ways. First, we can apply our previous results by showing that the prefiltered data satisfies the weak dependence condition in Assumption 2. Consider Example 1 with non-random regressors. In this case, $\hat{X}_{it} = X_{it}$ and $\hat{y}_{it} = X_{it}'\beta_t + \varepsilon_{it}$, where $\varepsilon_{it} = u_{it} - e_t$. Notice that $n^{-1/2}\sum_{i=1}^n(\varepsilon_{it}^2 - E\varepsilon_{it}^2) = n^{-1/2}\sum_{i=1}^n(u_{it}^2 - Eu_{it}^2) + (\sqrt{n}e_t^2 - \sqrt{n}Ee_t^2) - 2e_t n^{-1/2}\sum_{i=1}^n u_{it} + 2Ee_t n^{-1/2}\sum_{i=1}^n u_{it}$ and $n^{-1/2}\sum_{i=1}^n X_{it}\varepsilon_{it} = n^{-1/2}\sum_{i=1}^n X_{it}u_{it} - (\sqrt{n}e_t)n^{-1}\sum_{i=1}^n X_{it}$. Suppose u_{it} is i.i.d across i and independent across t and $Eu_{it} = 0$ with u_{it} being sub-Gaussian. Also assume that $n^{-1}\sum_{i=1}^n \|X_t\|_2^2$ is bounded. Then Hoeffding's inequality and Bernstein's inequality imply that all the moments of $\sqrt{n}e_t$, $n^{-1/2}\sum_{i=1}^n u_{it}$ and $\|n^{-1/2}\sum_{i=1}^n X_{it}u_{it}\|_2$ are bounded. Since we also have serial independence of u_{it}, Assumption 2 is satisfied and Theorem 1 and Corollary 1 still hold if we replace $\Pi(\cdot \mid \mathbf{Z})$ with $\Pi(\cdot \mid \hat{\mathbf{Z}})$.

Second, we can explicitly incorporate the estimation error in $\hat{\mathbf{Z}}$ in the proof. We resort to this method when formally verifying Assumption 2 for $\hat{\mathbf{Z}}$ is difficult. However, when Assumption 2 does not hold, the rate of convergence might be worse than $(nT)^{-1}\log(nT)$ (the claim in Theorem 1). This is not surprising since strong dependence in the (prefiltered) data is expected to distort the likelihood by causing a problem for the law of large numbers and the central limit theorem. In these cases, frequentist estimators are typically easier to analyze; nonetheless, we provide a theoretical result for $\Pi(\cdot \mid \hat{\mathbf{Z}})$. We first describe the quality of $\hat{\mathbf{Z}}$ as an estimate for

$\mathbf{Z}$.

Assumption 5. *With probability at least* $1-\psi_n$, $\sup_{\theta\in\Theta}|\ell_n(\hat{\mathbf{Z}},\theta)-\ell_n(\mathbf{Z},\theta)|/(nT)\leq\rho_n$, *where* $\psi_n,\rho_n>0$ *are sequences of positive numbers.*

Assumption 5 is a high-level condition for the estimation error in $\hat{\mathbf{Z}}$. It is flexible enough to allow for a general theoretical framework. A case-by-case analysis exploiting the dependence structure and the prefiltering algorithm is needed to verify Assumption 5.[10] The following result states the concentration properties of $\Pi(\cdot\mid\hat{\mathbf{Z}})$.

Theorem 3. *Let Assumptions 1, 2, 3, 4 and 5 hold. Then for any* $c\in(0,1)$, *there exists a constant* $D_c>0$ *such that with probability at least* $1-c-\gamma_n-\psi_n$, $\Pi\left(B_n\mid\hat{\mathbf{Z}}\right)\geq 1-D_c\,(nT)^{-M_{10}/2}$, *where* $B_n^c=\Theta\backslash B_n$ *and*

$$B_n=\left\{\theta\in\Theta:\ \max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\sigma_t(\theta)-\sigma_t(\theta_*)\right)^2,\frac{1}{T}\sum_{t=1}^{T}\|\beta_t(\theta)-\beta_t(\theta_*)\|_2^2\right\}\right.$$
$$\left.\geq\tilde{M}_1\frac{\log(nT)}{nT}+\tilde{M}_2\rho_n\right\},$$

and $\tilde{M}_1,\tilde{M}_2>0$ *are constants depending only on* $K,\Theta,M_0,M_1,...,M_{10}$.

Theorem 3 generalizes Theorem 1 to prefiltered data. If the data satisfy Assumption 2 without prefiltering, then $\rho_n=\psi_n=0$ and Theorem 3 becomes Theorem 1. By Theorem 3, the quality of the prefiltering stage in general affects the rate of concentration of the posterior.

Instead of using the density of the prefiltered data in (2), in some cases we can work with the density for the prefiltered data.[11] Consider Example 1 with non-random regressors and assume that u_{it} is independent across i and t and $u_{it}\sim N(0,\sigma_t^2(\theta_*))$. It is not difficult to see that $E\varepsilon_{it}^2=(1-n^{-1})\sigma_t^2$ and $E\varepsilon_{it}\varepsilon_{jt}=-\sigma_t^2/n$ for $i\neq j$. Let Σ_n be the $n\times n$ matrix with $1-n^{-1}$ on the diagonal and $-n^{-1}$ on the off diagonal. Let

[10] Notice that biases can arise in dynamic panel models, especially with short T. This issue has been discussed in the studies listed in footnote 4.

[11] We thank a referee for bringing up this point.

$\hat{y}_t = (\hat{y}_{1t}, ..., \hat{y}_{nt})' \in \mathbb{R}^n$ and $\hat{X}_t = (\hat{X}_{1t}, ..., \hat{X}_{nt})' \in \mathbb{R}^{n\times r}$. Then the density of $(\hat{y}_t, \hat{X}_t)$ is $-\frac{1}{2}\log[\det(\sigma_t^2(\theta)\Sigma_n)] - \frac{1}{2}(\hat{y}_t - \hat{X}_t\beta_t(\theta))'(\sigma_t^2(\theta)\Sigma_n)^{-1}(\hat{y}_t - \hat{X}_t\beta_t(\theta))$. Since $(\hat{y}_t, \hat{X}_t)$ is independent across t, we have

$$\ell_n(\hat{\mathbf{Z}}, \theta) = -\frac{1}{2}\sum_{t=1}^{T}\log[\det(\sigma_t^2(\theta)\Sigma_n)] - \frac{1}{2}\sum_{t=1}^{T}(\hat{y}_t - \hat{X}_t\beta_t(\theta))'(\sigma_t^2(\theta)\Sigma_n)^{-1}(\hat{y}_t - \hat{X}_t\beta_t(\theta)).$$

Since the likelihood for the prefiltered data is exact in this case, prefiltering will not introduce any distortions. More generally, if one is willing to explicitly model the dependence structure in the data, one can modify the likelihood for the data and, under the assumption that the model is correctly specified, a separate prefiltering step is not necesssary.

2.4. Bayesian Inference under Misspecification

A growing literature in statistics shows that under a misspecified likelihood, Bayesian credible sets are not asymptotically valid frequentist confidence sets; see, e.g., Royall and Tsou (2003); Kleijn and Van der Vaart (2012); Müller (2013); Bissiri et al. (2016). For example, assuming i.i.d. Gaussian errors across i and t within regimes is likely to lead to a misspecified likelihood. This means that the frequentist validity of Bayesian credible sets is questionable regardless of any prefiltering .

To demonstrate that the frequentist validity of Bayesian inference can be invalidated by even very small forms of misspecification of the likelihood, consider panel data from a location model with no breaks: $y_{it} = \mu_* + \varepsilon_{it}$, where $\varepsilon_{it} = u_{it} + (nT)^{-1/2}\psi v$, u_{it} and v are i.i.d. $N(0, 1)$, and ψ is a nonrandom constant which captures the degree of cross-sectional dependence. Suppose the likelihood is based on an i.i.d. $N(\mu, 1)$ distribution, so ψ becomes a gauge for the level of misspecification. For any nonzero ψ, the Bayesian credible sets do not provide correct frequentist inference. To see this, consider the log-likelihood $\ell_n(\mu) = -\frac{nT}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \mu)^2$. Under regularity conditions, we can apply Theorem 2.1 of Kleijn and Van der Vaart (2012) to obtain

$$\sup_{x\in\mathbb{R}}\left|\Pi_n((-\infty,x])-\Phi\left(\sqrt{nT}(x-\mu_*)+\Delta_n\right)\right|=o_P(1), \tag{3}$$

where $\Phi(\cdot)$ is the cdf of $N(0,1)$, $\Pi_n(\cdot)$ is the posterior and $\Delta_n = (nT)^{-1/2}\sum_{i=1}^n\sum_{t=1}^T u_{it}+\psi v$. For $\alpha\in(0,1)$, let $B_{n,1-\alpha}=[\mu_*-\Delta_n/\sqrt{nT}-z_{1-\alpha/2}/\sqrt{nT},\mu_*-\Delta_n/\sqrt{nT}+z_{1-\alpha/2}/\sqrt{nT}]$, where $z_{1-\alpha/2}=\Phi^{-1}(1-\alpha/2)$. By (3), we notice that $\Pi_n(B_{n,1-\alpha})=1-\alpha+o_P(1)$. Hence, $B_{n,1-\alpha}$ is a Bayesian credible set. However, it is easy to see that $P(\mu_*\in B_{n,1-\alpha})=P(|\Delta_n|\leq z_{1-\alpha/2})=2\Phi(z_{1-\alpha/2}/\sqrt{1+\psi^2})-1$. Clearly, for any $\psi\neq 0$, $\lim_{n\to\infty}P(\mu_*\in B_{n,1-\alpha})<1-\alpha$. Therefore, even when the misspecification for the error term is only of the order $O((nT)^{-1/2})$, Bayesian inference based on the i.i.d $N(0,1)$ model would be invalid in the frequentist sense.

Next, we show that it is impossible in practice to clearly detect such misspecification even when μ_* is known to be zero.

Lemma 2. *Let $\{u_{it}\}_{1\leq i\leq n,\ 1\leq t\leq T}$ be i.i.d $N(0,1)$ random variables and let v be another $N(0,1)$ variable that is independent of u_{it}. Suppose we observe $\{w_{it}\}_{1\leq i\leq n,\ 1\leq t\leq T}$ where $w_{it}=u_{it}+(nT)^{-1/2}\psi v$, and $\psi\in\mathbb{R}$ is a constant. Consider the problem of testing $H_0:\ \psi=0$ versus $H_1:\ \psi=c_0$, where $c_0>0$ is a constant. Then for any $\alpha\in(0,1)$, the power of the likelihood ratio test of nominal size α is*

$$1-F_1\left(\frac{\chi^2_{1,1-\alpha}}{1+c_0^2}\right),$$

where $F_1(\cdot)$ denotes the cdf of the χ^2_1 distribution and $\chi^2_{1,1-\alpha}=F_1^{-1}(1-\alpha)$.

Lemma 2 is a statement about the likelihood-ratio test of a point null hypothesis and a point alternative hypothesis. By the Neyman-Pearson lemma, the likelihood-ratio test is the uniformly most powerful test. Since Lemma 2 holds for any sample size, as n,T tend to infinity at any rate, no test can guarantee that in testing H_0 versus H_1 both the Type I and Type II errors will go to zero.

Hence, in practice, one should not expect to reliably detect tiny misspecification (of the order $O(1/\sqrt{nT})$). However, by the previous analysis, the frequentist validity of Bayesian credible sets is sensitive to such small misspecification. Since barely detectable misspecification can invalidate Bayesian inference (in a frequentist sense), we do not think Bayesian methods should be used for inference in applied research regardless of prefiltering.

Of course, prefiltering introduces additional misspecification for the likelihood model due to its own estimation error and thus has an effect on the frequentist properties of Bayesian credible sets. Theoretical results might be obtained for a specific prefiltering procedure, assuming that the likelihood model without prefiltering can be correctly specified. However, in practice the likelihood model is likely to be misspecified from the beginning which causes problems for Bayesian methods as an inference tool even under very small misspecification. We therefore think it is more appropriate to use frequentist inference procedures, which is the general recommendation for Bayesian methods; see Remark 4 of Müller (2013).

3. A Bayesian Panel Break Model

We next develop a linear panel regression model with pooled parameters that is a special case of the class of models analyzed above. We first introduce the panel regression model, then present the priors and the resulting posterior distribution. Next, we show how to jointly estimate an unknown number of structural breaks and perform variable selection within regimes. For generality, we assume that the model is fitted to prefiltered data so that the dependent variable for the ith cross-sectional unit at time t, $\hat{y}_{it}$, is regressed on a set of r independent variables, $\hat{X}_{it}$.

3.1. A panel model with pooled parameters

Our panel model for the kth regime takes the following form:

$$\hat{y}_{it} = \hat{X}_{it}'\beta_k + \hat{\varepsilon}_{it}, \qquad i = 1, \ldots, n, \qquad t = \Lambda_{k-1} + 1, \ldots, \Lambda_k \tag{4}$$

in which X_{it} is an $(r \times 1)$ vector of the r covariates for the ith series at time t and β_k is an $(r \times 1)$ vector containing the pooled regression coefficients on the r covariates for the kth regime. We collect the list of K breakpoints into a vector, $\Lambda = (\Lambda_1, \ldots, \Lambda_K)$.

To induce the variable selection properties of the posterior, we rewrite β_t (and σ_t) to be the product of an indicator and the magnitude of these quantities; this way, the prior on the indicator provides a shrinkage property. The frequentist tradition of model selection often involves thresholding slope coefficients that are not significantly different from zero and inference will be asymptotically valid as long as $r/(nT) \to 0$ (Cattaneo et al. 2018a,b).

Let $\boldsymbol{\beta}$ denote the $(r \times (K{+}1))$ matrix of coefficients on each of the covariates in each of the regimes. Assume the error terms are independently and identically Normally distributed $\hat{\varepsilon}_{it} \sim \mathcal{N}(0, \sigma_k^2)$ for $t = \Lambda_{k-1} + 1, \ldots \Lambda_k$. Hence, the coefficients β and error-term variances $\sigma^2 = (\sigma_1^2, \ldots, \sigma_{K+1}^2)$ are allowed to shift to new values following each break. Finally, let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ and $l_k = \Lambda_k - \Lambda_{k-1}$ denote the duration of the kth regime which consists of the observations $(\Lambda_{k-1} + 1, \ldots, \Lambda_k)$, while $l = (l_1, \ldots, l_{K+1})$ denotes a vector of regime durations.

The likelihood for our model is[12]

$$\begin{aligned} p(\hat{\boldsymbol{y}} \mid \hat{\boldsymbol{X}}, \theta, \Lambda) &= \prod_{k=1}^{K+1} \prod_{t=\Lambda_{k-1}+1}^{\Lambda_k} p(\hat{\boldsymbol{y}}_t \mid \hat{\boldsymbol{X}}_t, \theta_t) \\ &= \prod_{k=1}^{K+1} (2\pi\sigma_k^2)^{-l_k n/2} \exp\left[-\frac{1}{2\sigma_k^2} \sum_{k=1}^{K+1} \sum_{t=\Lambda_{k-1}+1}^{\Lambda_k} (\hat{y}_t - \hat{X}_t'\beta_k)'(\hat{y}_t - \hat{X}_t'\beta_k) \right], \end{aligned} \tag{5}$$

[12] Recall that variables with a hat superscript have been prefiltered.

in which $\hat{y}_t$ denotes the $(n \times 1)$ vector of observations on the dependent variable for the cross-sectional units $i = 1, \ldots, n$ at time t, $\hat{\boldsymbol{X}}_t$ is the $(r \times n)$ matrix of observations on the r covariates for the cross-sectional units $i = 1, \ldots, n$ at time t, $\theta_t = (\beta_t, \sigma_t^2)$ denotes the parameter vector at time t, $\hat{\boldsymbol{y}}$ denotes the $(n \times T)$ matrix of observations on the dependent variable, and $\hat{\boldsymbol{X}}$ denotes the $(r \times n \times T)$ three-dimensional array of observations on the r covariates.

The aim is to estimate the number of breakpoints K, their locations $\Lambda = (\Lambda_1, \ldots, \Lambda_K)$ and the parameter vector θ in each regime.

3.2. Prior Distributions

We place a Poisson prior over the regime durations

$$p(l_k \mid \zeta_k) = \frac{\zeta_k^{l_k} e^{-\zeta_k}}{l_k!}, \qquad k = 1, \ldots, K+1, \tag{6}$$

where ζ_k has a conjugate Gamma prior

$$p(\zeta_k) = \frac{d^c}{\Gamma(c)} \lambda_k^{c-1} e^{-d\lambda_k}, \qquad k = 1, \ldots, K+1, \tag{7}$$

and c and d are the hyperparameters of $\zeta = (\zeta_1, \ldots, \zeta_{K+1})$. The Poisson intensity parameter ζ_k captures the expected duration of regime k and thus the probability of a break to the parameters in the kth regime. Marginalising ζ leaves the prior on the breakpoints

$$p(\Lambda) = \prod_{k=1}^{K+1} p(l_k \mid \zeta_k) p(\zeta_k) = \prod_{k=1}^{K+1} \frac{1}{l_k!} \frac{\Gamma(c + l_k)}{(d+1)^{c+l_k}} \frac{d^c}{\Gamma(c)}. \tag{8}$$

To perform regime-specific variable selection, we introduce an indicator vector for the kth regime ι_k. Each element of this vector can take a value of either zero or one and therefore assigns positive prior mass at zero for the corresponding coefficients β_k.

The indicator vector is specified as a binomial distribution with hyperparameter ξ_k

$$\begin{aligned} p(\iota_k \mid \xi_k) &= \binom{r}{m_k} \xi_k^{m_k} (1-\xi_k)^{r-m_k}, \\ m_k &= \sum_{\kappa=1}^{r} \iota_{\kappa,k}, \qquad k = 1, \ldots, K+1. \end{aligned} \tag{9}$$

The hyperparameter ξ_k represents the average probability across the r regressors that each of the regressors should be selected in the kth regime.

We assume that ξ_k follows a conjugate Beta distribution

$$p(\xi_k) = \xi_k^{e-1}(1-\xi_k)^{f-1} \frac{\Gamma(e+f)}{\Gamma(e)\Gamma(f)}, \qquad k = 1, \ldots, K+1. \tag{10}$$

Noting that the binomial coefficient for the kth regime is equal to $r!\,(m_k!\,(r-m_k)!\,)^{-1}$ and multiplying and dividing by $\Gamma(e+f+r)(\Gamma(e+m_k)\Gamma(f+r-m_k))^{-1}$, we can marginalise ξ_k and discard it, leaving

$$p(\boldsymbol{\iota}) = \prod_{k=1}^{K+1} \left(\frac{r!}{m_k!\,(r-m_k)!} \frac{\Gamma(e+f)}{\Gamma(e)\Gamma(f)} \frac{\Gamma(e+m_k)\Gamma(f+r-m_k)}{\Gamma(e+f+r)} \right). \tag{11}$$

We set $e = f = 1$ and thus assign equal probability to including or omitting each regressor in every regime. However, if the user wanted to supply prior information, the values of e and f can simply be adjusted such that the probability of including a variable is equal to $e/(e+f)$. [13]

To stay as close to conventional practice as possible, we specify an inverse gamma prior over the regime-specific variances

$$p(\sigma^2) = \prod_{k=1}^{K+1} \frac{b^a}{\Gamma(a)} \sigma_k^{2^{-(a+1)}} \exp\left(-\frac{b}{\sigma_k^2}\right), \tag{12}$$

[13] For a more detailed discussion of prior choices for variable selection see Giannone et al. (2017).

and a Normal prior on the slope coefficients conditional on the variances

$$p(\beta \mid \sigma^2) = \prod_{k=1}^{K+1} 2\pi^{-r/2}(\sigma_k^2)^{-r/2} \mid V_\beta \mid^{-1/2} \exp\left(-\frac{1}{2\sigma_k^2}\beta_k' V_\beta^{-1}\beta_k\right),$$
$$V_\beta = I_r \sigma_\beta^2. \tag{13}$$

3.3. Posterior Distribution

To reduce the computational burden, which can be critical when using large panels, we integrate out the parameters from the posterior distribution, obtaining[14]

$$p(\hat{\boldsymbol{y}} \mid \hat{\boldsymbol{X}}, \Lambda) = \prod_{k=1}^{K+1} (2\pi)^{-l_k n/2} \frac{b^a}{\Gamma(a)} \frac{\Gamma(\tilde{a}_k)}{\tilde{b}_k^{\tilde{a}_k}} |\Sigma_k|^{1/2} | V_\beta |^{-1/2} \tag{14}$$

where, for $k = 1, \ldots, K+1$

$$\Sigma_k^{-1} = V_\beta^{-1} + \sum_{t=\Lambda_{k-1}+1}^{\Lambda_k} \hat{X}_t \hat{X}_t',$$
$$\mu_k = \Sigma_k \left(\sum_{t=\Lambda_{k-1}+1}^{\Lambda_k} \hat{X}_t' \hat{y}_t\right),$$
$$\tilde{a}_k = a + (l_k n)/2,$$
$$\tilde{b}_k = \frac{1}{2}\left(2b + \sum_{t=\Lambda_{k-1}+1}^{\Lambda_k} \hat{y}_t' \hat{y}_t - \mu_k' \Sigma_k^{-1} \mu_k\right). \tag{15}$$

3.4. Modeling Cross-sectional Dependencies

One way to account for (strong) cross-sectional correlations in panel regressions is by modeling this through a set of unobserved common factors f_t which affect both the

[14] The values of X that correspond to omitted variables are set equal to zero in equation (15).

residuals and the regressors and use the common correlated effects (CCE) approach of Pesaran (2006) to prefilter the data using cross-sectional averages of the independent and dependent variables as proxies for any unobserved common factors. This is the approach we propose to take here. In particular, we modify equation (4) as follows:

$$\begin{aligned} y_{it} &= X_{it}'\beta_k + \varepsilon_{it}, \qquad i = 1, \ldots, n, \qquad t = \Lambda_{k-1} + 1, \ldots, \Lambda_k \\ \varepsilon_{it} &= \phi_i f_t + \nu_{it} \end{aligned} \tag{16}$$

in which ν_{it} are idiosyncratic errors and ϕ_i are factor loadings. We assume ν_{it} are independent of X_{it}, but f_t and X_{it} may be correlated

$$X_{it} = \Phi_i' f_t + v_{it}. \tag{17}$$

The existence of unobserved common factors f_t and their possible correlation with X_{it} may render OLS estimation of (16) inconsistent. These correlations can be removed by transforming the data with cross-sectional averages of the dependent variable and the regressors.[15]

Combining (16) and (17), we have

$$\psi_{it} = \begin{pmatrix} y_{it} \\ X_{it} \end{pmatrix} = C_{ik}' f_t + u_{it}, \qquad k = 1, \ldots, K+1, \qquad t = \Lambda_{k-1} + 1, \ldots, \Lambda_k \tag{18}$$

in which

$$C_{ik} = (\phi_i, \Phi_i') \begin{pmatrix} 1 & 0 \\ \beta_k & I_r \end{pmatrix}$$

[15] Because v_{it} is i.i.d., it may be possible to apply the grouped fixed effects approach of Bonhomme and Manresa (2015), but finding breaks in the time series rather than cross-sectional dimension. However, the computational burden of an iterative method in which each iteration implements a k-mean clustering could be high. Moreover, there is no guarantee of global convergence under this algorithm. We think this is one of the situations where MCMC algorithms might deliver a more stable numerical performance.

and

$$u_{it} = \begin{pmatrix} \nu_{it} + \beta_k v_{it} \\ v_{it} \end{pmatrix}$$

where C_{ik} also shifts following a break. Letting $\boldsymbol{\beta} = (\beta_1, \dots, \beta_T)$, $Y_i = (y_{i1}, \dots, y_{iT})'$, $\boldsymbol{X}_i = (x_{i1}, \dots, x_{iT})'$, and $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT})'$, we can write (16) in matrix form as

$$Y_i = \boldsymbol{X}_i' \beta + \varepsilon_i. \tag{19}$$

Let $\bar{\psi}_t = \frac{1}{n} \sum_{i=1}^{n} \psi_{it}$, which assumes equal cross-sectional weights, such that

$$\bar{\psi}_t = \bar{C}_k' f_t + \bar{u}_t, \qquad k = 1, \dots, K+1, \qquad t = \Lambda_{k-1} + 1, \dots, \Lambda_k.$$

Let $\bar{\psi} = (\bar{\psi}_1, \dots, \bar{\psi}_T)$ and define the orthogonal projection matrix $\boldsymbol{M}_{\boldsymbol{\psi}} = I_T - \bar{\psi}(\bar{\psi}'\bar{\psi})^{-1}\bar{\psi}'$ such that premultiplying the regression model in (19) by $\boldsymbol{M}_{\boldsymbol{\psi}}$ obtains

$$\hat{Y}_i = \hat{\boldsymbol{X}}_i' \boldsymbol{\beta} + \hat{\varepsilon}_i, \tag{20}$$

in which $\hat{\varepsilon}_i = \boldsymbol{M}_{\boldsymbol{\psi}} F \phi_i + \boldsymbol{M}_{\boldsymbol{\psi}} \nu_i$, $\hat{Y}_i = \boldsymbol{M}_{\boldsymbol{\psi}} Y_i$, $\hat{\boldsymbol{X}}_i = \boldsymbol{M}_{\boldsymbol{\psi}} \boldsymbol{X}_i$, and $F = (f_1, \dots, f_T)$.

Because of the time variation in β_t, the first component of $\bar{\psi}_t$ is related to a mixture of β_t and f_t; hence only the remaining components of $\bar{\psi}_t$ provide useful information on f_t. For this reason, the above prefiltering procedure only works provided that $E(\Phi_i)$ has full rank and $m \leq \kappa$.[16]

4. Estimation

This section provides details of the four steps in our algorithm for variable selection, breakpoint detection and parameter estimation.

Step 1: **Estimating the parameter vector**. We sample β_k and σ_k^2 for the $K+1$

[16] We are grateful to a referee for pointing this out.

regimes from their full conditionals

$$\begin{aligned} \sigma_k^2 \mid \Lambda &\sim IG(\tilde{a}_k, \tilde{b}_k), \qquad k = 1, \ldots, K+1, \\ \beta_k \mid \Lambda, \sigma_k^2 &\sim MVN(\mu_k, \Sigma_k \sigma_k^2), \qquad k = 1, \ldots, K+1, \end{aligned} \tag{21}$$

in which $\mu, \Sigma, \tilde{a}$, and $\tilde{b}$ are computed using (15).

Step 2 : **Performing regime-specific variable selection**. Conditional on the estimated breakpoints and parameters, ι_k is updated for regimes $k = 1, \ldots, K+1$. Specifically, we sample with equal probability a value of zero or one for each of the $\kappa = 1, \ldots, r$ covariates in the kth regime to obtain the proposal indicator vector ι_k^*. Using ι_k^*, we compute Σ_k^*, μ_k^*, and $\tilde{b}_k^*$ from (15), while m_k^* is computed from (9). Using (11) and (14), the proposed ι_k^* is accepted with probability min(1, α), where

$$\alpha = \frac{\tilde{b}_k^{\tilde{a}_k} \mid \Sigma_k^* \mid^{1/2}}{\tilde{b}_{k^*}^{\tilde{a}_k} \mid \Sigma_k \mid^{1/2}} \frac{m_k!\,(r - m_k)!}{m_k^*!\,(r - m_k^*)!} \frac{\Gamma(e + m_k^*)\Gamma(f + r - m_k^*)}{\Gamma(e + m_k)\Gamma(f + r - m_k)}.$$

Here $q(\iota_k, \iota_k^*)$ denotes the proposal density of ι_k^* given the existing indicator vector ι_k. If the proposal is accepted, then ι_k^*, m_k^*, Σ_k^* μ_k^* and $\tilde{b}_k^*$ are substituted for ι_k, m_k, Σ_k, μ_k, and $\tilde{b}_k$. We repeat the procedure for each of the $K+1$ regimes in turn.

Step 3: **Estimating the breakpoint locations**. The birth and death moves detailed below enable the introduction or removal of breaks and so a simple perturbation to the existing breakpoint locations is all we need to help them converge to their true values. This perturbation is provided through a random-walk Metropolis-Hastings step. Each breakpoint Λ_k for $k = 1, \ldots, K$ is perturbed by u sampled uniformly from the interval $[-s, s]$ to give the new break date $\Lambda_{k^*} = \Lambda_k + u$. The proposed breakpoint vector Λ^* is the same as Λ, but Λ_{k^*} has been substituted for Λ_k. The proposed regime durations are computed as $l_{k^*} = \Lambda_{k^*} - \Lambda_{k-1}$ and $l_{k^*+1} = \Lambda_{k+1} - \Lambda_{k^*}$. Using (15), we compute $\Sigma_{k^*}^{-1}$, $\Sigma_{k^*+1}^{-1}$, μ_{k^*}, μ_{k^*+1}, $\tilde{a}_{k^*}$, $\tilde{a}_{k^*+1}$, $\tilde{b}_{k^*}$, and $\tilde{b}_{k^*+1}$. The proposal is accepted with probability $\min(1, \alpha)$ where

$$\alpha = \frac{|\Sigma_{k^*}|^{1/2}}{|\Sigma_k|^{1/2}} \frac{|\Sigma_{k^*+1}|^{1/2}}{|\Sigma_{k+1}|^{1/2}} \frac{\Gamma(\tilde{a}_{k^*})}{\tilde{b}_{k^*}^{\tilde{a}_{k^*}}} \frac{\tilde{b}_k^{\tilde{a}_k}}{\Gamma(\tilde{a}_k)} \frac{\Gamma(\tilde{a}_{k^*+1})}{\tilde{b}_{k^*+1}^{\tilde{a}_{k^*+1}}} \frac{\tilde{b}_{k+1}^{\tilde{a}_{k+1}}}{\Gamma(\tilde{a}_{k+1})} \frac{\Gamma(l_{k^*}+c)}{\Gamma(l_k+c)} \frac{\Gamma(l_{k^*+1}+c)}{\Gamma(l_{k+1}+c)} \frac{l_k!}{l_{k^*}!} \frac{l_{k+1}!}{l_{k^*+1}!}.$$

If the proposal is accepted, we substitute Λ_{k^*}, l_{k^*}, l_{k^*+1}, $\tilde{a}_{k^*}$, $\tilde{a}_{k^*+1}$, $\tilde{b}_{k^*}$, $\tilde{b}_{k^*+1}$, μ_{k^*}, μ_{k^*+1}, $\Sigma_{k^*}^{-1}$, and $\Sigma_{k^*+1}^{-1}$ for Λ_k, l_k, l_{k+1}, $\tilde{a}_k$, $\tilde{a}_{k+1}$, $\tilde{b}_k$, $\tilde{b}_{k+1}$, μ_k, μ_{k+1}, Σ_k^{-1}, and Σ_{k+1}^{-1}.

Step 4: **Estimating the number of breakpoints**. The sampler begins at a sensible number of breakpoints. On every iteration of the Markov chain Monte Carlo run with equal probability the sampler attempts to either add (birth move) or remove (death move) one breakpoint. The proposed move is accepted with the probability that ensures detailed balance is maintained across the entire parameter space including the number of breakpoints and selected variables. If the move is accepted, the corresponding breakpoint, indicator and parameter vectors are updated; otherwise they are discarded and the algorithm continues.

Birth move. Each iteration either a birth or death move is entered with equal probability.[17] A birth move proposes to introduce one new breakpoint denoted Λ_{k^*} and hence increase K to $K+1$. We draw Λ_{k^*} uniformly from the time series $\Lambda_{k^*} \sim U[1, T]$ and if $\Lambda_{k^*} \in \Lambda$ the proposal is rejected. If $\Lambda_{k^*} \notin \Lambda$ then Λ_{k^*} is added to Λ to construct the proposed breakpoint vector Λ^*. The durations of the proposed regimes are $l_{k^*} = \Lambda_{k^*} - \Lambda_{k^c-1}$ and $l_{k^*+1} = \Lambda_{k^c} - \Lambda_{k^*}$, respectively.

We next propose a value of zero or one for each of the $\kappa = 1, \ldots, r$ covariates in the two new regimes to construct the new indicator vectors ι_{k^*} and ι_{k^*+1}. Using ι_{k^*} and ι_{k^*+1} where appropriate, we calculate $\Sigma_{k^*}^{-1}$, $\Sigma_{k^*+1}^{-1}$, μ_{k^*}, μ_{k^*+1}, $\tilde{a}_{k^*}$, $\tilde{a}_{k^*+1}$, $\tilde{b}_{k^*}$, and $\tilde{b}_{k^*+1}$ from (15), while m_{k^*} and m_{k^*+1} are computed from (9). Since the parameter vector θ has been marginalised from the posterior it does not need to be proposed.

[17] If $K = T - 1$, any proposed birth move is immediately rejected and likewise for the proposal of a death move if $K = 0$.

The proposed birth move is accepted with probability equal to $\min(1, \alpha)$, where

$$\begin{aligned}\alpha = {} & \frac{\Gamma(\tilde{a}_{k^*})}{\tilde{b}_{k^*}^{\tilde{a}_{k^*}}} \frac{b^a}{\Gamma(a)} \frac{\Gamma(\tilde{a}_{k^*+1})}{\tilde{b}_{k^*+1}^{\tilde{a}_{k^*+1}}} \frac{|\Sigma_{k^*}|^{1/2}|\Sigma_{k^*+1}|^{1/2}}{|\Sigma_{k^c}|^{1/2}|\ V_\beta\ |^{1/2}} \frac{\tilde{b}_{k^c}^{\tilde{a}_{k^c}}}{\Gamma(\tilde{a}_{k^c})} \frac{\Gamma(e+f)}{\Gamma(e)\Gamma(f)} \\ & \times \frac{r!\,\Gamma(e+m_{k^*})\Gamma(f+r-m_{k^*})}{m_{k^*}!\,(r-m_{k^*})!\,\Gamma(e+f+r)} \frac{T2^r}{K+1} \frac{\Gamma(e+m_{k^*+1})\Gamma(f+r-m_{k^*+1})}{\Gamma(e+m_{k^c})\Gamma(f+r-m_{k^c})} \\ & \times \frac{m_{k^c}!\,(r-m_{k^c})!}{m_{k^*+1}!\,(r-m_{k^*+1})!} \frac{\Gamma(l_{k^*}+c)}{(d+1)^c} \frac{d^c}{\Gamma(c)} \frac{l_{k^c}!}{l_{k^*}!\,l_{k^*+1}!} \frac{\Gamma(l_{k^*+1}+c)}{\Gamma(l_{k^c}+c)}.\end{aligned}$$

If the move is accepted, Λ is replaced by Λ^* and l, m, $\boldsymbol{\iota}$, Σ^{-1}, μ, $\tilde{a}$, and $\tilde{b}$ are updated by substituting their values for the new regimes k^* and k^*+1 for their values in the existing regime k^c. If the move is rejected, the proposals are discarded.

Death move. If a death move is entered, an attempt is made to reduce K to $K-1$ by removing an existing breakpoint Λ_{k^c} which is sampled uniformly from the existing breakpoint vector $\Lambda_{k^c} \sim U[\Lambda_1, \Lambda_K]$. Removing Λ_{k^c} from Λ constructs the proposed breakpoint vector Λ^*. This move attempts to merge two existing shorter regimes k^c and k^c+1 separated by Λ_{k^c} into one new longer regime k^*, and thus $l_{k^*} = l_{k^c} + l_{k^c+1}$.

Next, the regime-specific indicator vector for the new regime ι_{k^*} must be proposed. Specifically, each of the $\kappa = 1, \ldots, r$ indicators for the r covariates in the newly proposed regime k^* are sampled with equal probability as either zero or one. Using ι_{k^*}, we compute $\Sigma_{k^*}^{-1}$, μ_{k^*}, $\tilde{a}_{k^*}$ and $\tilde{b}_{k^*}$ from (15), and m_{k^*} from (9).

The death move is accepted with probability $\min(1, \alpha)$, where

$$\begin{aligned}\alpha = {} & \frac{\Gamma(\tilde{a}_{k^*})}{\tilde{b}_{k^*}^{\tilde{a}_{k^*}}} \frac{\tilde{b}_{k^c}^{\tilde{a}_{k^c}}}{\Gamma(\tilde{a}_{k^c})} \frac{|\Sigma_{k^*}|^{1/2}|\ V_\beta\ |^{1/2}}{|\Sigma_{k^c}|^{1/2}|\Sigma_{k^c+1}|^{1/2}} \frac{\tilde{b}_{k^c+1}^{\tilde{a}_{k^c+1}}}{\Gamma(\tilde{a}_{k^c+1})} \frac{\Gamma(a)}{b^a} \frac{l_{k^c}!}{l_{k^*}!} \frac{\Gamma(l_{k^*}+c)}{\Gamma(l_{k^c}+c)} \frac{(d+1)^c}{\Gamma(l_{k^c+1}+c)} \\ & \times l_{k^c+1}! \frac{\Gamma(c)}{d^c} \frac{K}{T} \frac{1}{2^r} \frac{m_{k^c}!\,(r-m_{k^c})!}{m_{k^*}!\,(r-m_{k^*})!} \frac{\Gamma(e)\Gamma(f)}{\Gamma(e+f)} \frac{\Gamma(e+m_{k^*})\Gamma(f+r-m_{k^*})}{\Gamma(e+m_{k^c})\Gamma(f+r-m_{k^c})} \\ & \times \frac{m_{k^c+1}!\,(r-m_{k^c+1})!\,\Gamma(e+f+r)}{r!\,\Gamma(e+m_{k^c+1})\Gamma(f+r-m_{k^c+1})}.\end{aligned}$$

If the move is rejected, the proposals are discarded; otherwise Λ is replaced by Λ^*, while the values of l, $\boldsymbol{\iota}$, m, Σ^{-1}, μ, $\tilde{a}$, and $\tilde{b}$ for the newly proposed regime k^* are substituted for the corresponding values in the existing regimes k^c and $k^c + 1$.

5. Simulation Study

We next conduct a simulation study to demonstrate the importance of the increased power derived from the cross-section when performing variable selection in a panel with pooled parameters and breaks. First, we show how, in the absence of cross-sectional dependencies, the ability of the algorithm to converge to the true underlying data generating process (DGP) increases with the size of the cross-section. We then study the robustness of the method to (i) increasing the number of regressors; (ii) including fat tailed errors; (iii) adding dynamic effects (lagged dependent variables); (iv) allowing for cross-sectional dependencies with and without persistent factors; and (v) changing the volatility of the residual in the regression model.

5.1. Performance of Algorithm in the Baseline Scenario

We first explore the ability of our algorithm to converge to the true underlying DGP in which the coefficients of a subset of $r = 8$ covariates change across regimes as displayed in Table 1. Specifically, four panels with $T = 100$ and n increasing in size from 1 through 25, 50, and 100 are considered for the regression

$$y_{it} = X_{it}'\beta_k + \varepsilon_{it}, \qquad t = \Lambda_{k-1} + 1, \ldots, \Lambda_k, \qquad k = 1, \ldots, K+1, \qquad i = 1, \ldots, n. \tag{22}$$

The first of the covariates is a unit vector so $\beta_{1,k}$ is an estimate of the intercept term in the kth regime. The remaining seven covariates are drawn from the process $X_{it} \sim N(0, 1)$ for $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$.

The intercept and regression coefficients break at $t = 40$ and $t = 65$: $\beta_{1:r,t} = 1$ for $t \leq 40$, $\beta_{1:r,t} = 1.25$ for $40 < t \leq 65$, and $\beta_{1:r,t} = 1.5$ for $65 < t \leq 100$. The error is assumed to be cross-sectionally independent and normally distributed $\varepsilon_{it} \sim N(0, \sigma^2)$ for $i = 1, 2, \dots, n$ and $t = 1, 2, \dots, T$, where σ breaks at $t = 50$: $\sigma_t = 1$ for $t \leq 50$, and $\sigma_t = 1.5$ for $50 < t \leq 100$. Within regimes in which the covariates have no explanatory power, that is, when the corresponding value of the indicator vector ι_k is equal to zero, the product of the indicator and coefficient will equal zero.

The prior hyperparameter values of a and b are set equal to 2 and 1.25, respectively. Moreover, we set $c = 2$, and compute $d = 0.08$ such that the prior expected value of ζ in each regime is equal to the mean of the simulated regime durations, 25.[18]

We begin the algorithm at one breakpoint at time $t = 5$ with every element of the indicator vector set equal to zero in both regimes. When $n = 1$, a time series version of our algorithm is only able to find two of the three break points at $t = 40$ and $t = 50$ as displayed in Figure 1a, both of which are estimated with considerable uncertainty. Figure 1b displays how as the size of the cross-section increases to 5 the algorithm detects the third break at $t = 65$. Figures 1c and 1d show how increasing the cross-section further to 25 and 100 gradually removes the uncertainty surrounding the estimated break dates.

Table 1 displays the posterior estimates of the indicator vector and the sum of the absolute estimation error of the indicator variable across the 8 regressors within each of the four regimes as the cross-section increases. When $n = 1$ the sum of the absolute estimation error of the indicator variable across the regressors in regimes 1, 2, 3, and 4, respectively, are 0.61, 1.72, 1.32, and 1.07. The total sum across the four regimes is therefore 4.72.[19] As the cross-section increases to 25 and 100, respectively, the sum of the absolute estimation error of the indicator variable across the four regimes gradually reduces to 2.04 and 0.06. The increased power derived from the expanding size of the cross-section enables more accurate variable selection.

[18] Both samplers were run for 2,500 iterations beyond a burn-in period of 5,000 iterations and were thinned at an interval of 2. Computation was fast with both algorithms taking less than one minute to run on a Windows XP based laptop with an Intel core i5 processor.

[19] Note the wrong inclusion of x_2 in the second regime and of x_5 in the fourth regime, and the wrong exclusion of x_7 in the third regime.

Table 2 displays simulated values and posterior estimates obtained from the models that perform variable selection with and without breaks when $n = 100$. Variable selection procedures that ignore breaks are susceptible to omitting covariates that are only informative for short periods (x_8) and likewise including (for the entire sample) covariates that are active for only a short period (x_7). The variable selection procedure that ignores breaks defines β_7 as being active for the full sample and β_8 as never being active. The regime-specific variable selection approach, however, has the flexibility to deactivate β_7 and activate β_8 in only the second regime as required by the underlying DGP.

5.2. Extensions

Increasing the number of regressors. One might expect that as $(K+1)r$ grows large relative to nT, the ability of our methodology to converge to the true DGP will break down. We explore this through a simulation exercise in which n is fixed at 20, T is fixed at 100, the number of breaks K is fixed at 3 with break dates at $t = 40$, $t = 50$ and $t = 65$, and r increases from 5 to 250. Table 3 displays the mean absolute estimation error computed across the r estimated indicator values and their true values $MAE_{\iota_k} = \frac{1}{r}\sum\limits_{\kappa=1}^{r} \mid \hat{\iota}_{\kappa,k} - \iota_{\kappa,k} \mid$ as r increases from 5 to 250.

For relatively small numbers of covariates ($r = 5$ and $r = 12$), the posterior estimates of the indicator variables on the covariates are very accurate with mean absolute errors less than 0.05. Increasing the number of covariates to 25 induces a high dimensional search, with $2^{25} \approx 33.5$ million combinatorial possibilities ignoring the presence of breaks which exacerbates the search problem even further. The benefit of the Bayesian stochastic search algorithm is displayed here with the MAE of the indicator vector being less than 0.1 in all four regimes. Increasing the number of covariates further, however, leads to the method breaking down as the nT data set is not large enough relative to $r(K+1)$ to infer the correct values of ι_k in the kth regime. Mean absolute errors increase to 0.3 when $r = 50$ and as high as 0.491 when $r = 250$.

Fat-tailed errors. In many economic applications, errors have heavy tails. We therefore pursue a simulation exercise in which the true DGP is identical to the one detailed in Section 5.1, except the errors are now assumed to follow a Student t-distribution with 6 degrees of freedom. We standardise the errors so they have a standard deviation which is the same as in the baseline scenario. Our methodology precisely estimates all three breaks.

Dynamic models. Panel models that include lagged dependent variables as regressors may compromise inference, see Bai (2009), Chudik and Pesaran (2015), Moon and Weidner (2015), and Moon and Weidner (2017).[20] However, for a model with pooled parameters like the one presented here, Everaert and Groote (2016) report that biases are likely to be modest for time series of at least moderate dimension. In any event, most of any induced bias will apply to the autoregressive coefficient.

To evaluate whether including an autoregressive term reduces the ability of our method to correctly identify the breakpoints, we conduct a simulation study in which the true DGP now includes a lagged dependent variable as a regressor which is active in each of the four regimes

$$y_{it} = X_{it}'\beta_k + \epsilon_{it}, \qquad t = \Lambda_{k-1}, \ldots, \Lambda_k, \qquad k = 1, \ldots, K+1, \qquad i = 1, \ldots, n.$$

where the final element of X_{it} now equals y_{it-1} and its coefficient equals 0.9. The setup is otherwise unchanged from the baseline scenario.

The method performs strongly, precisely estimating the three breaks at their true dates (not shown). The breakpoint estimates are unaffected by further including a second and third lagged dependent variable. In our pooled parameter framework, including autoregressive terms as regressors in the model does not appear to hinder

[20] Hansen (2003) develops a generalised reduced rank regression framework to allow for breaks in a cointegrating Vector Autoregression.

our ability to accurately estimate breaks.

Common correlated effects. We next introduce correlated effects into the simulated panel data set and explore the ability of our algorithm to converge to the true underlying data generating process with and without prefiltering the data as described in Section 3.4.

The data generating process for series $i = 1, \ldots, n$ is assumed to take the form

$$\begin{aligned} y_{it} &= X_{it}'\beta_k + \varepsilon_{it}, \qquad t = \Lambda_{k-1} + 1, \ldots, \Lambda_k, \qquad k = 1, \ldots, K+1, \\ \varepsilon_{it} &= \phi_{1i} f_t + u_{it}, \end{aligned}$$

in which $\phi_{1i} \sim iid\ N(1, 0.1)$ and the idiosyncratic errors are generated as $u_{it} \sim iid\ N(0, \sigma_t^2)$. The first of the covariates is a unit vector so that $\beta_{1,k}$ is an estimate of the intercept term in the kth regime. In contrast with Section 5.1, the common factor f_t now presides in both the error-term ε_{it} and the covariate X_{it} as $X_{it} = \phi_{2i} f_t + \nu_{it}$, where $\phi_{2i} \sim iid\ N(0.5, 0.1)$ and $\nu_{it} \sim iid\ N(0, 0.75)$. We generate the factor f_t as $f_t = 0.5 f_{t-1} + \nu_{ft}$, where $\nu_{ft} \sim iid\ N(0, 0.75)$.

We run our regime-specific variable selection procedure with and without prefiltering the data with cross-sectional averages. The time-series dimension is $T = 100$ and the cross-sectional dimension is $n = 200$. We display in Figure 2 the results of our algorithm with (bottom panel) and without (top panel) prefiltering. When prefiltering the data, we detect the three breaks with 100% certainty. Without prefiltering, however, the performance of the algorithm is poor. The three true breaks are correctly identified but at the expense of spuriously detecting an additional three breaks, showing the importance of prefiltering the data.

Moderate cross-sectional dependencies. We next consider the performance of our methodology to detect breaks when mild or moderate cross-sectional dependencies remain. The DGP is the same as described in Section 5.1 except the error terms are now drawn from a multivariate Normal distribution with cross-sectional correlations $\varepsilon_t \sim MVN(0, S\sigma_t^2)$, where $\sigma_t = 1$ for $t = 1, \ldots, 50$ and $\sigma_t = 1.5$ there-after. All elements on the main diagonal of the covariance matrix S equal 1, and all off diagonals equal ρ.

For mild correlations ($\rho \leq 0.20$), the ability of the method to detect the correct breaks is unaffected. Increasing ρ beyond 0.2, the correct breaks are still identified, but additional spurious breaks are also detected. The higher the value of ρ, the higher the number of spurious breaks detected.

Highly persistent factors. Baltagi et al. (2016) report that the increased power to detect breaks obtained from the cross-section is lost once the factors become highly persistent. We therefore perform a simulation in which the persistence of the common factor is increased from 0.5 to 0.7 and 0.9, respectively. When the persistence of the factor is increased from 0.5 to 0.7 increasing the cross-sectional dimension to 100 still improves the precision of the breakpoint estimation but to a lesser degree than before. Once the persistence is increased to 0.9, however, the breakpoints are estimated no more precisely when the cross-sectional dimension is increased to 100.

Volatility and variable selection. We finally investigate whether the procedure that ignores breaks can activate (deactivate) regressors that are only informative (noninformative) in regimes that exhibit high volatility even though such regimes are longer relative to the analysis carried out in Section 5.1. We explore this issue with a simulation that has high volatility in some regimes but not in others. Specifically, the DGP now sets $\sigma_t = 5$ in regimes 3 and 4. We only consider an intercept and two (instead of seven) regressors. The results (available in the supplementary material) show that the procedure that ignores breaks is unable to activate regressor 2 which only becomes informative in the fourth regime which has high volatility. Similarly, it is unable to deactivate regressor 3 which is only deactivated in the fourth regime. The procedure that allows for breaks, however, is able to estimate the correct breakpoints and select the informative and noninformative variables within regimes even if they are characterised by high volatility.

6. Empirical Application

A large literature in corporate finance explores firms' choice of capital structure, i.e., their decisions on how much debt and equity to issue to finance operations. Many studies consider how firms should exploit their debt structure to optimally trade off bankruptcy costs against tax savings on interest payments as more debt is issued and firms become more levered. In a comprehensive study, Frank and Goyal (2009) attempt to uncover which of the many variables that have been proposed to explain firms' capital structure are actually informative. They consider 25 covariates and report that six core variables account for 27% of the variation in leverage; the remaining 19 covariates only explain an additional 2% of the variation.

We next revisit the study of Frank and Goyal (2009), but performing variable selection on the covariates *and* allowing for structural breaks. This may reveal that some of the six core covariates are actually superfluous. Alternatively, covariates that are informative only during short regimes may be incorrectly omitted altogether while covariates that lose explanatory power during short regimes will fail to be omitted. We focus on 22 of the covariates with complete data records and investigate whether any of the core (other) covariates are omitted (included) in some regimes.

6.1. Data

We generalize the regression of Frank and Goyal (2009) to allow for an unknown number of structural breaks K occurring at unknown times and perform variable selection within each regime $k = 1, \dots, K+1$

$$LV_{it} = \alpha_k + X'_{it-1}\beta_k + \varepsilon_{it}, \qquad t = \Lambda_{k-1}+1, \dots, \Lambda_k, \qquad \varepsilon_{it} \sim N(0, \sigma_k^2) \tag{23}$$

LV_{it} is the market-based leverage ratio measure - total debt to market assets (TDM) - for firm i at time t and X_{it-1} contains for firm i at time t the 22 covariates listed in Table A.4 in the supplementary Appendix.[21] We drop the first 8 years of data due to missing observations and all remaining firms that have 40 observations or fewer on the dependent variable. This leaves an unbalanced panel consisting of n=175 firms and T=46 annual observations of data from 1958 through 2003.

6.2. Cross-sectional and serial dependencies

We first explore the ability of our pre-filtering procedure to remove cross-sectional dependencies in the data. We also consider what proportion of cross-sectional dependencies are driven by co-movements between firms belonging to the same industry. Finally, we test for serial dependence.

We prefilter the data using cross-sectional averages of the dependent and all independent variables as proxies for any unobserved common factors as described in Section 3.4. Without prefiltering, the CD statistic of Pesaran (2004) is 90.02. At the 5% level, we therefore conclusively reject the null of no cross-sectional dependencies. After prefiltering, the CD statistic is reduced to 1.46 and so we are unable to reject the null on the prefiltered data. This demonstrates that prefiltering has successfully removed most of the cross-sectional dependencies in the data.

We further test for serial dependence in the data using the Durbin-Watson statistic. For each series in the cross-section we use the residuals to compute a Durbin-Watson statistic. The mean of the DW statistics across all n series is 1.501 with a spread between 0.588 and 2.444. Although some series may experience negative serial correlation, on average, positive serial correlation is more common. For models with an intercept and 20 regressors and for a time-series dimension of 40 (we have 22 regressors and on average 41 observations on the dependent variable), Savin and White

[21] A detailed description of the dataset is provided in Appendix B of Frank and Goyal (2009).

(1977) report lower and upper Durbin Watson critical values of 0.338 and 2.838, respectively. All series have DW statistics that lie between these bounds and thus the test is inconclusive. The estimated autocorrelations are also small for the majority of series and so we conclude that serial dependence in the data is unlikely to affect in a major way our inference concerning the presence of breaks in the panel model.

Priors. Our empirical application sets $d = 1$ and $c = 10$, thus assuming that a break is expected to occur every ten years for the models that allow for breaks. We set the hyperparameters $e = 1$ and $f = 1$ which implies that the prior mean of ξ_k for regimes $k = 1, \ldots, K + 1$ is 0.5 so each variable is equally likely to be selected or omitted. We set $a = 2$ and $b = 0.23$ to give a prior expected error-term variance equal to 0.23, which is the variance of the dependent variable across the entire data set. Finally, we set the prior standard deviation of β, σ_β, equal to 0.5.

6.3. Results

We find evidence of three breaks occurring at 1962, 1973, and 1999 when estimating the regime-specific variable selection model. The combination of annual data and increased power obtained from the cross-sectional dimension of $n = 175$ results in highly concentrated posterior estimates of the break dates as displayed in Figure 3 (top window). The break in 1962 corresponds to the increase in corporate leverage (Graham et al. 2015). Leverage jumped from 16% in the early 1960s to 27% in 1972. The break at 1973 corresponds to this peak value after which corporate leverage fell relatively quickly before rising slowly again. Finally, the break in 1999 is capturing the fall from 28% in 1997 to 23% in 2003.

Figure 3 (bottom window) shows that the first short regime (1958-1962), has only four core predictors with more than equal probability of being active, with the market-to-book ratio being the strongest (see Table 4). On the other hand, some of the other factors such as uniqueness of product are active in the first regime with a posterior probability estimate of 0.818 (not shown). A similar scenario occurs in the final regime (2000-2003) in which Tangibility and Inflation are deactivated but SGA is activated. Accounting for breaks may thus result in short regimes in which (i) strong factors are omitted and (ii) weak factors are included. The insight of Frank and Goyal (2009) that of 25 potential predictors only six core predictors are required to explain almost all the variation in leverage therefore appears to be an overestimate which results from ignoring structural breaks. Ignoring breaks also results in some other predictors being overlooked when they are active only for short regimes.

7. Conclusion

This article develops a new Bayesian approach for jointly estimating an unknown number of structural breaks and performing regime-specific variable selection in a panel regression framework with pooled coefficients and common breaks.

Our procedure for variable selection in the presence of breaks is likely to prove useful in many empirical applications such as forecasting where the relevance of individual predictors undergoes change. In such settings, it can be important to eliminate predictors that are no longer relevant and, conversely, introduce new variables that, at least for a period of time, possess predictive power over the outcome.

References

Ahn, S. C. and Schmidt, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of econometrics*, 68(1):5–27.

Alvarez, J. and Arellano, M. (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica*, 71(4):1121–1159.

Anderson, T. W. and Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of econometrics*, 18(1):47–82.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, pages 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. (2010). Common breaks in means and variances for panel data. *Journal of Econometrics*, 157(1):78–92.

Bai, J. and Carrion-I-Silvestre, J. L. (2009). Structural changes, common stochastic trends, and unit roots in panel data. *The Review of Economic Studies*, 76(2):471–501.

Bai, J., Lumsdaine, R. L., and Stock, J. H. (1998). Testing for and dating common breaks in multivariate time series. *The Review of Economic Studies*, 65(3):395–432.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Baltagi, B. H., Feng, Q., and Kao, C. (2016). Estimation of heterogeneous panels with structural breaks. *Journal of Econometrics*, 191(1):176–195.

Baltagi, B. H., Kao, C., and Liu, L. (2017). Estimation and identification of change points in panel models with nonstationary or stationary regressors and error term. *Econometric Reviews*, 36(1-3):85–102.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018a). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34(2):277–301.

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018b). Inference in linear regression models with many covariates and heteroskedasticity. *Journal of the American Statistical Association*, (forthcoming).

Cheng, X., Liao, Z., and Schorfheide, F. (2016). Shrinkage estimation of high-dimensional factor models with structural instabilities. *The Review of Economic Studies*, 83(4):1511–1543.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241.

Chudik, A. and Pesaran, M. H. (2015). Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics*, 188(2):393–420.

Dhaene, G. and Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3):991–1030.

Elliott, G. and Müller, U. K. (2006). Efficient tests for general persistent time variation in regression coefficients. *The Review of Economic Studies*, 73(4):907–940.

Everaert, G. and Groote, T. (2016). Common correlated effects estimation of dynamic panels with cross-sectional dependence. *Econometric Reviews*, 35(3):428–462.

Frank, M. Z. and Goyal, V. K. (2009). Capital structure decisions: which factors are reliably important? *Financial Management*, 38(1):1–37.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.

Giannone, D., Lenza, M., and Primiceri, G. E. (2017). Economic predictions with big data: The illusion of sparsity. *Unpublished Manuscript, Northwestern University*.

Giné, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in l^r-metrics, $1 \leq r \leq \infty$. *The Annals of Statistics*, 39(6):2883–2911.

Giordani, P. and Kohn, R. (2012). Efficient bayesian inference for multiple change-point and mixture innovation models. *Journal of Business & Economic Statistics*, 26(1):66–77.

Gouriéroux, C., Phillips, P. C., and Yu, J. (2010). Indirect inference for dynamic panel models. *Journal of Econometrics*, 157(1):68–77.

Graham, J. R., Leary, M. T., and Roberts, M. R. (2015). A century of capital structure: The leveraging of corporate america. *Journal of Financial Economics*, 118(3):658–683.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Hahn, J. and Kuersteiner, G. (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica*, 70(4):1639–1657.

Han, C., Phillips, P. C., and Sul, D. (2014). X-differencing and dynamic panel model estimation. *Econometric Theory*, 30(1):201–251.

Hansen, P. R. (2003). Structural changes in the cointegrated vector autoregressive model. *Journal of Econometrics*, 114(2):261–295.

Jochmann, M., Koop, G., and Strachan, R. W. (2010). Bayesian forecasting using stochastic search variable selection in a var subject to breaks. *International Journal of Forecasting*, 26(2):326–347.

Kim, D. (2011). Estimating a common deterministic time trend break in large panels with cross sectional dependence. *Journal of Econometrics*, 164(2):310–330.

Kleibergen, F. (2009). Tests of risk premia in linear factor models. *Journal of econometrics*, 149(2):149–173.

Kleijn, B. and Van der Vaart, A. (2012). The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Koop, G. and Potter, S. M. (2007). Estimation and forecasting in models with multiple breaks. *The Review of Economic Studies*, 74(3):763–789.

Korobilis, D. (2013). Var forecasting using bayesian variable selection. *Journal of Applied Econometrics*, 28(2):204–230.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Lian, H. (2010). Posterior convergence and model estimation in bayesian change-point problems. *Electronic Journal of Statistics*, 4:239–253.

Liu, L., Moon, H. R., and Schorfheide, F. (2017). Forecasting with dynamic panel data models. *arXiv:1709.10193*.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1):499–526.

Moon, H. R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579.

Moon, H. R. and Weidner, M. (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory*, 33(1):158–195.

Müller, U. K. (2013). Risk of bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, pages 1417–1426.

Oka, T. and Perron, P. (2018). Testing for common breaks in a multiple equations system. *Journal of Econometrics*, 204(1):66 – 85.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Pesaran, M. H. (2004). General diagnostic tests for cross section dependence in panels. *Unpublished Manuscript, University of Cambridge.*

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *The Review of Economic Studies*, 73(4):1057–1084.

Pesaran, M. H. and Timmermann, A. (2000). A recursive modelling approach to predicting uk stock returns. *The Economic Journal*, 110(460):159–191.

Pesaran, M. H. and Timmermann, A. (2002). Market timing and return prediction under model instability. *Journal of Empirical Finance*, 9(5):495–510.

Phillips, P. C. and Moon, H. R. (1999). Linear regression limit theory for nonstationary panel data. *Econometrica*, 67(5):1057–1111.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Qu, Z. and Perron, P. (2007). Estimating and testing structural changes in multivariate regressions. *Econometrica*, 75(2):459–502.

Rossi, B. (2013). Advances in forecasting under instability. *Handbook of Economic Forecasting (Elliott, G and Timmermann, A. (Eds.))*, Chapter 21:1203–1324.

Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *The Annals of Statistics*, 38(1):146–180.

Royall, R. and Tsou, T.-S. (2003). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):391–404.

Savin, N. E. and White, K. J. (1977). The durbin-watson test for serial correlation with extreme sample sizes or many regressors. *Econometrica*, 45(8):1989–1996.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of statistics*, 29(3):687–714.

Smith, S. C. and Timmermann, A. (2017a). Break risk. *Unpublished Manuscript, University of California, San Diego.*

Smith, S. C. and Timmermann, A. (2017b). Detecting breaks in real time: A panel forecasting approach. *Unpublished Manuscript, University of California, San Diego.*

Stock, J. H. and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1):11–30.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463.

Table 1: Variable selection as the cross-section increases

Regressor	Regime				Regime			
	1	2	3	4	1	2	3	4
		True values				n=1		
1	1	1	1	1	1	1	1	1
2	0	0	1	1	0.17	0.99	0.99	1
3	0	1	1	1	0.08	0.99	0.99	1
4	0	1	1	0	0.12	1	1	0.13
5	0	1	1	0	0.16	1	1	0.88
6	1	1	1	1	1	0.99	0.99	1
7	1	0	1	1	1	0.15	0.15	1
8	0	1	0	0	0.08	0.44	0.44	0.06
$\sum_{\kappa=1}^{r} \mid \iota_{\kappa,k} - \hat{\iota}_{\kappa,k} \mid$					0.61	1.72	1.32	1.07
		n=25				n=100		
1	1	1	1	1	1	1	1	1
2	0.05	0.02	1	1	0.02	0	1	1
3	0.02	1	1	1	0.01	1	1	1
4	0.01	1	1	0.09	0	1	1	0
5	0.08	1	1	0.02	0	1	1	0
6	1	1	1	0.41	1	1	1	1
7	1	0.14	0.98	1	1	0.01	1	1
8	0.01	0.07	0	0.06	0.02	1	0	0
$\sum_{\kappa=1}^{r} \mid \iota_{\kappa,k} - \hat{\iota}_{\kappa,k} \mid$	0.17	1.09	0.02	0.76	0.05	0.01	0	0

Table 1: Variable selection as the cross-section increases. This table displays the simulated values of the indicator vector on the 8 covariates across regimes and the corresponding posterior estimates as the size of the cross-section increases. We also display the sum of the absolute estimation error of the indicator vector across the 8 covariates within each regime. The estimates are derived using the hyperparameter values and simulated data detailed in Section 5.1.

Table 2: Posterior estimates of variables

Regime	Time	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	σ
Indicator ($\iota_{\kappa,k}$)										
Variable selection without breaks										
1	$0 < t \leq 100$	1	1	1	1	1	1	1	0	
Parameters ($\iota_{\kappa,k} \times \beta_{\kappa,k}$)										
Simulated values										
1	$t \leq 40$	1	0	0	0	0	1	1	0	1
2	$40 < t \leq 50$	1.25	0	1.25	1.25	1.25	1.25	0	1.25	1
3	$50 < t \leq 65$	1.25	1.25	1.25	1.25	1.25	1.25	1.25	0	1.5
4	$65 < t$	1.5	1.5	1.5	0	0	1.5	1.5	0	1.5
Variable selection without breaks										
1	$0 < t \leq 100$	1.32	0.75	1.35	0.82	0.84	1.22	1.10	0	1.28
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0)	(0.05)
Regime-specific Variable Selection										
1	$t \leq 40$	1.02	-0.00	-0.00	0	0	1.00	0.98	-0.00	0.99
		(0.02)	(0.03)	(0.00)	(0)	(0)	(0.02)	(0.02)	(0.02)	(0.02)
2	$40 < t \leq 50$	1.27	0	1.22	1.21	1.31	1.31	-0.00	1.28	1.00
		(0.03)	(0)	(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	(0.02)	(0.04)
3	$50 < t \leq 65$	1.31	1.22	1.23	1.23	1.28	1.26	1.29	0	1.45
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0)	(0.08)
4	$65 < t$	1.56	1.51	1.50	0	0	1.48	1.49	0	1.45
		(0.02)	(0.02)	(0.02)	(0)	(0)	(0.03)	(0.02)	(0.00)	(0.05)

Table 2: Posterior estimates of variables. This table displays the posterior estimates of the regression coefficients and the indicator variable on the 8 covariates across the 4 regimes obtained from the variable selection procedure with and without breaks. The estimates are derived using the hyperparameter values detailed in Section 5.1.

Table 3: Variable selection as the number of regressors increases

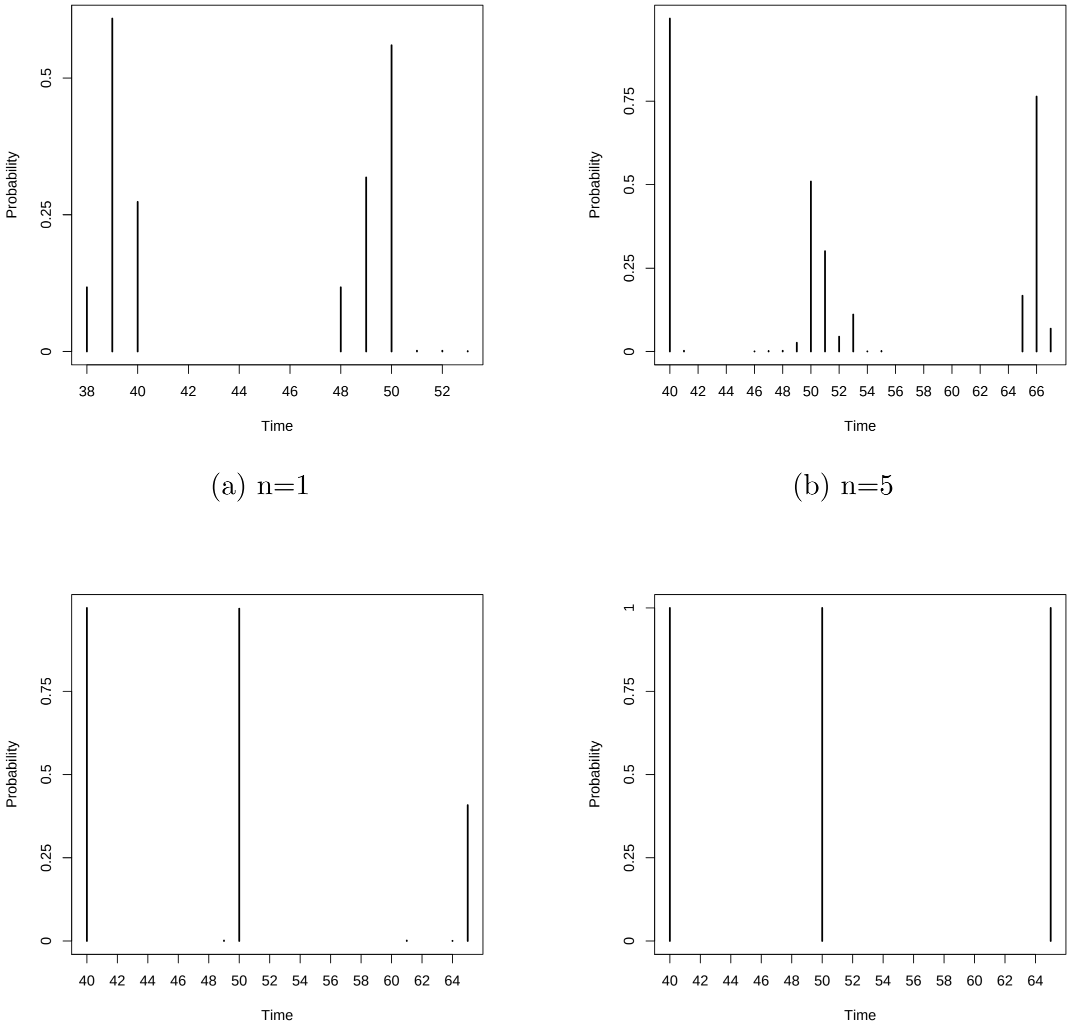
r	Regime			
	1	2	3	4
5	0.038	0.039	0.049	0.030
12	0.044	0.038	0.047	0.036
25	0.094	0.098	0.092	0.085
50	0.302	0.297	0.263	0.350
125	0.349	0.462	0.420	0.466
250	0.399	0.484	0.491	0.478

Table 3: Variable selection as the number of regressors increases. This table displays in regime k the mean absolute error MAE_{ι_k} of the posterior estimates of the inclusion indicator variable relative to the true simulated value across all r covariates when n, T and K are fixed at 20, 100 and 3, respectively. The MAE_{ι_k} is therefore bounded between 0 and 1 with 0 reflecting perfect estimation of the r indicator variables in the kth regime and 1 reflecting the worst possible estimation.

Table 4: Parameter estimates for core regressors

Predictor		Selected Variables					Slope Coefficient ($\iota_{\kappa,k} \times \beta_{\kappa,k}$)		
		Regime			Average		Regime		
	1	2	3	4		1	2	3	4
Profitability	0.457	0.998	0.997	0.953	0.851	-0.005	-0.633	-1.016	-0.502
	(0.498)	(0.044)	(0.049)	(0.109)	(0.175)	(0.031)	(0.046)	(0.042)	(0.151)
Total assets	0.225	0.024	0.989	0.951	0.547	0.012	0.001	0.011	0.023
	(0.311)	(0.161)	(0.101)	(0.170)	(0.186)	(0.017)	(0.008)	(0.005)	(0.007)
Market-to-book	0.878	0.996	0.986	0.954	0.954	0.033	0.009	0.024	0.050
	(0.327)	(0.059)	(0.119)	(0.209)	(0.179)	(0.039)	(0.015)	(0.007)	(0.010)
Industry Leverage	0.705	0.998	0.996	0.965	0.916	0.040	0.356	0.412	0.492
	(0.424)	(0.038)	(0.066)	(0.183)	(0.178)	(0.057)	(0.019)	(0.022)	(0.060)
Tangible assets	0.546	0.998	0.981	0.262	0.697	0.012	0.193	0.165	0.007
	(0.497)	(0.042)	(0.137)	(0.496)	(0.293)	(0.027)	(0.011)	(0.014)	(0.033)
Inflation	0.506	0.697	0.833	0.172	0.552	0.002	0.084	0.022	0.019
	(0.499)	(0.458)	(0.471)	(0.459)	(0.472)	(0.044)	(0.102)	(0.076)	(0.129)
Break dates	1962, 1973, 1999								

Table 4: Parameter estimates for core regressors. This table reports estimates of the slope coefficients multiplied by their corresponding indicator variables ($\iota_{\kappa,k} \times \beta_{\kappa,k}$) on the six core regressors identified by Frank and Goyal (2009) and their standard deviations (reported in brackets below) when estimating the model that performs regime-specific variable selection using the full sample. We also report the posterior modes of the estimated break dates and the posterior means of the indicator variables in each regime and the average across regimes.



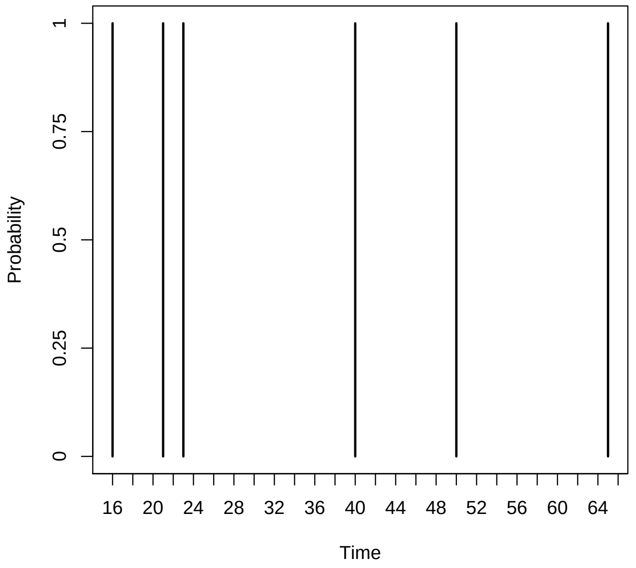
(a) n=1

(b) n=5

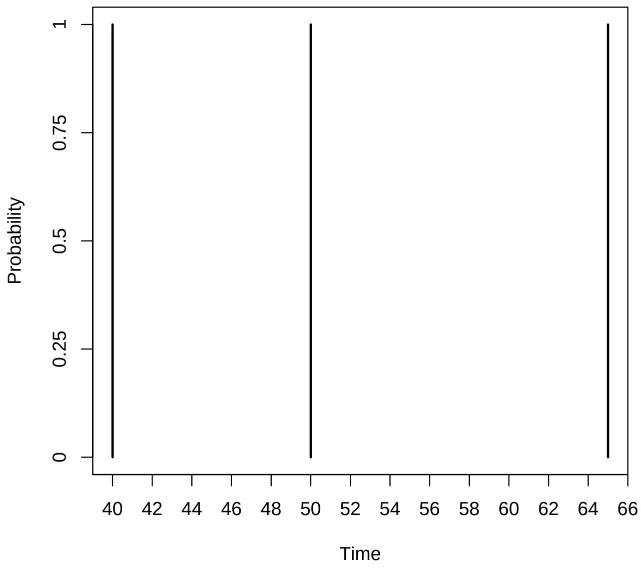
(c) n=25

(d) n=100

Figure 1: This figure displays the posterior probabilities of the estimated break dates as the size of the cross-section increases from 1 (top left) to 100 (bottom right) using the hyperparameters and simulated data detailed in Section 5.1.

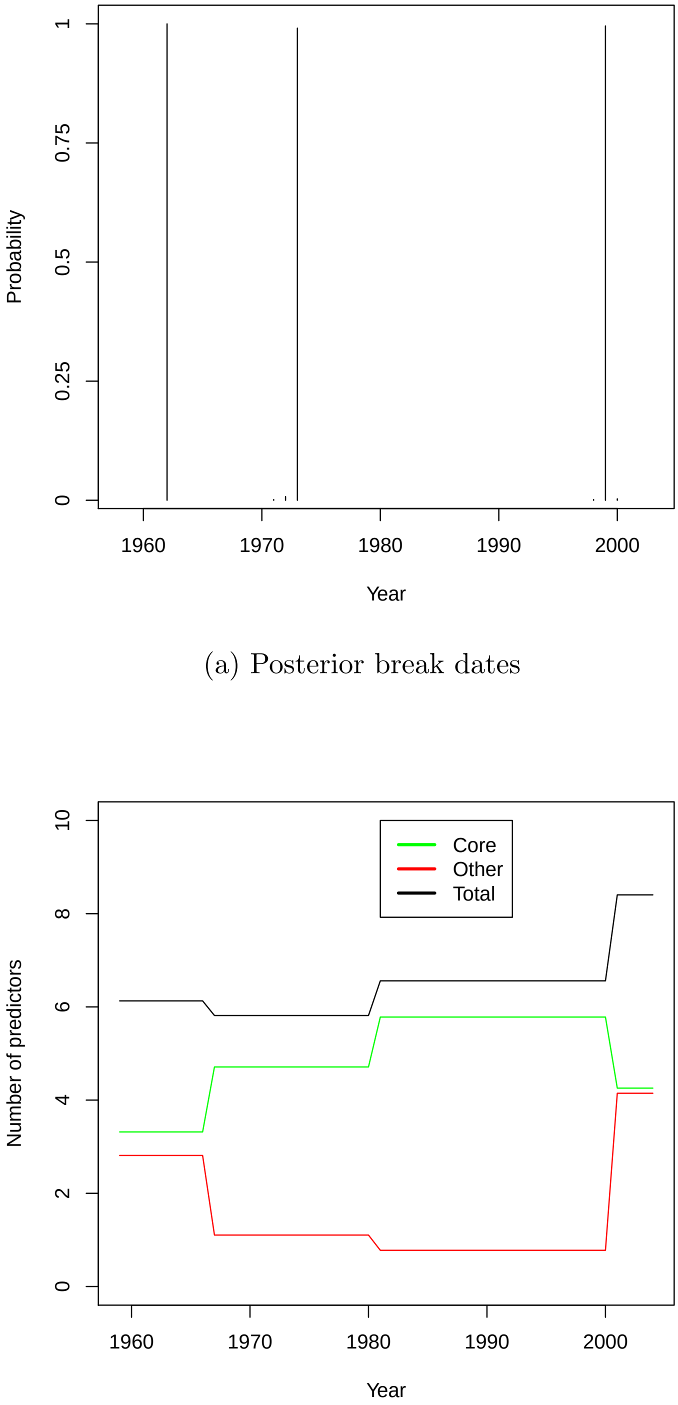


(a) No prefiltering



(b) Prefiltering

Figure 2: This figure displays the posterior probabilities of the estimated break dates when the cross-sectional dimension is $n = 200$ using the hyperparameters and the simulated data detailed in Section 5.2 with (bottom panel) and without (top panel) prefiltering the data with cross-sectional averages.



(a) Posterior break dates

(b) Number of predictors

Figure 3: The top panel of this Figure plots the posterior probabilities of the breaks dates estimated from the model that allows for regime-specific variable selection. The bottom panel graphs the total number of predictors that explain variation in the leverage ratio through the sample. Specifically, the green line graphs the total number of core predictors, the red line graphs the total number of other predictors, and the black line graphs their sum.