



Asymptotic properties of Bayesian inference in linear regression with a structural break

Kenichi Shimizu*

Adam Smith Business School, University of Glasgow, United Kingdom



ARTICLE INFO

Article history:

Received 18 September 2020

Received in revised form 30 March 2022

Accepted 31 March 2022

Available online 30 April 2022

Keywords:

Structural break

Bernstein–von Mises theorem

Sensitivity check

Model averaging

ABSTRACT

This paper studies large sample properties of a Bayesian approach to inference about slope parameters γ in linear regression models with a structural break. In contrast to the conventional approach to inference about γ that does not take into account the uncertainty of the unknown break date, the Bayesian approach that we consider incorporates such uncertainty. Our main theoretical contribution is a Bernstein–von Mises type theorem (Bayesian asymptotic normality) for γ under a wide class of priors, which essentially indicates an asymptotic equivalence between the conventional frequentist and Bayesian inference. Consequently, a frequentist researcher could look at credible intervals of γ to check robustness with respect to the uncertainty of the break date. Simulation studies show that the conventional confidence intervals of γ tend to undercover in finite samples whereas the credible intervals offer more reasonable coverages in general. As the sample size increases, the two methods coincide, as predicted from our theoretical conclusion. Using data from Paye and Timmermann (2006) on stock return prediction, we illustrate that the traditional confidence intervals on γ might underrepresent the true sampling uncertainty.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

We consider the linear regression with a structural break, following the notations of Bai (1997):

$$y_t = \begin{cases} w_t' \alpha + z_t' \delta_1 + \epsilon_t, & \text{for } t = 1, \dots, \lfloor \tau T \rfloor \\ w_t' \alpha + z_t' \delta_2 + \epsilon_t, & \text{for } t = \lfloor \tau T \rfloor + 1, \dots, T, \end{cases} \quad (1)$$

where w_t and z_t are $d_w \times 1$ and $d_z \times 1$ vectors of covariates, and the random variable ϵ_t is a regression error. $\lfloor a \rfloor$ is the largest integer that is strictly smaller than a . The relationship between the outcome y_t and the covariate z_t , measured by δ 's, changes across regimes, which are defined by the break location parameter $\tau \in (0, 1)$. There can be another set of covariates w_t whose relationship with y_t , measured by α , stays unchanged across the regimes. The unknown parameters include the break location τ as well as the slope parameters $\gamma = (\alpha, \delta_1, \delta_2)$. The focus of the current study is on inference about the slope parameters γ .¹

* Correspondence to: University Ave, Glasgow, G12 8QQ, United Kingdom.

E-mail address: Kenichi.Shimizu@glasgow.ac.uk.

¹ For an extensive review of important aspects in structural break models such as estimation and inference of the number of breaks as well as break locations, see Perron (2006).

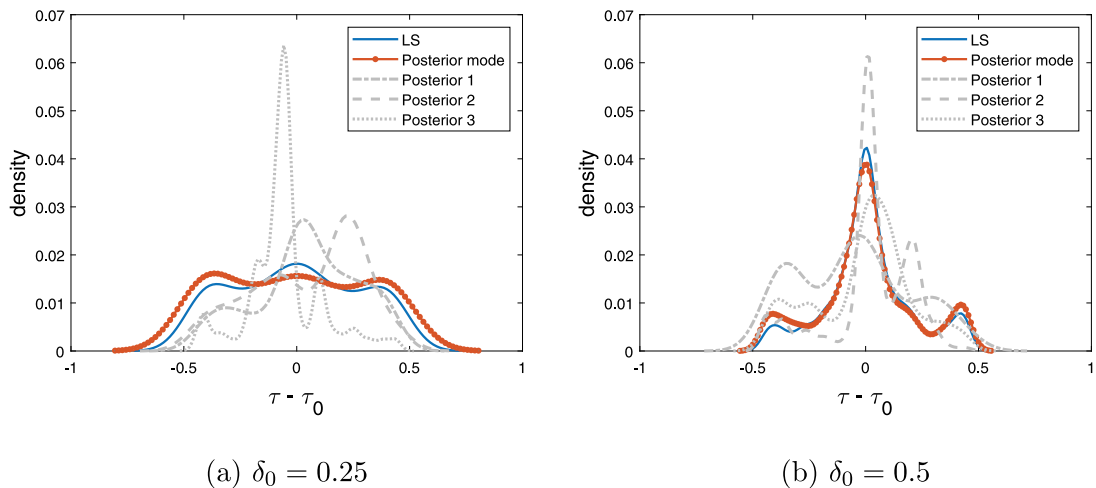


Fig. 1. Finite-sample distribution of $\hat{\tau}_{LS}$ (blue solid curve) based on the model $y_t = \delta_0 1(t > \lfloor \tau_0 T \rfloor) + \epsilon_t$, $\epsilon_t \sim \text{i.i.d.}N(0, 1)$, $\tau_0 = 0.5$, $T = 100$, with 1000 repeated experiments. The horizontal axis is $\tau - \tau_0$. We also show finite-sample distribution of the posterior mode of τ (red solid curve with small circles). In addition, we randomly chose 3 data realizations out of the 1000 repetitions to plot posterior densities of τ in gray dashed curves (hence each of them represents a realization of one data set).

1.1. The classic literature

In the literature, the conventional least-squares estimators $(\hat{\tau}_{LS}, \hat{\gamma}_{LS})$ for (τ, γ) are computed as follows: for each candidate τ , compute the sum of squared residuals of the regression and denote the minimizing choice by $\hat{\tau}_{LS}$. Plug in the value $\tau = \hat{\tau}_{LS}$ in the model and define $\hat{\gamma}_{LS} = \hat{\gamma}(\hat{\tau}_{LS})$, where $\hat{\gamma}(\tau)$ is the usual OLS estimator of γ assuming the break location τ . Bai (1997) assumes that the true jump size δ_0 is either fixed or shrinks to zero as $T \rightarrow \infty$, but at a rate slower than $\sqrt{T} \rightarrow \infty$. Bai shows that $\hat{\tau}_{LS}$ converges at the rate T^{-1} in the former case and, in the latter case, finds an asymptotic distribution of $\hat{\tau}_{LS}$ that can be used for constructing confidence intervals for τ . In both cases, Bai proves that the asymptotic distribution of $\hat{\gamma}_{LS}$ is the same as that of $\hat{\gamma}(\tau_0)$, where τ_0 is the true value of τ . This means that one can ignore the very problem of unknown τ when making inference on γ .

Fig. 1 shows examples of finite-sample distributions of $\hat{\tau}_{LS}$ (blue solid curves).² Note that despite the T -consistency, $\hat{\tau}_{LS}$ displays significant variation, especially when the true break size δ_0 is small.³ In practice, the conventional approach to inference on the slope parameters γ would ignore this uncertainty, neglecting all possible values of τ other than $\hat{\tau}_{LS}$. As a consequence, the corresponding confidence intervals on γ tend to undercover since it might not be the case that $\hat{\tau}_{LS} = \tau_0$ in a given sample.

1.2. Bayesian perspective

For a Bayesian, this non-standard estimation problem⁴ can be dealt with by placing prior on both τ and γ and by computing corresponding posterior probabilities. The uncertainty of τ will be automatically reflected on the marginal posterior probability of γ . This is because the posterior distribution of γ given the data \mathbf{D}_T can be written as a mixture where the weights correspond to the marginal posterior density $\pi_T(\tau)$ of τ :

$$p(\gamma | \mathbf{D}_T) = \int p(\gamma | \tau, \mathbf{D}_T) \pi_T(\tau) d\tau, \tag{2}$$

where $p(\gamma | \tau, \mathbf{D}_T)$ is the conditional posterior distribution of γ given τ . The posterior density $\pi_T(\tau)$ reflects the uncertainty of τ given the data set. Fig. 1 shows three realizations of $\pi_T(\tau)$ (gray dashed curves) which are randomly chosen out of the 1000 repetitions. Compared to the conventional approach, the key difference is that the Bayesian approach (2) incorporates all possibilities of τ (not just $\hat{\tau}_{LS}$) and weights them according to the posterior density, resulting in longer lengths of credible intervals of γ which consequently tend to have more reasonable coverages. Note that, unlike conventional frequentist methods, Bayesian inference has a valid interpretation even in finite samples as it does not rely on asymptotic theory.

² Fig. 1 also shows distributions (red solid curves with small circles) of a Bayesian point estimator of τ , the posterior mode. We later show that the posterior mode converges to the same limiting distribution as $\hat{\tau}_{LS}$.

³ In addition, the distributions exhibit three modes as reported in the literature (e.g., Baek, 2018; Casini and Perron, 2021).

⁴ Estimation of structural break models is considered non-standard in a sense that there is a non-regular parameter (e.g., break location) whose point estimator converges faster than $T^{-1/2}$, the rate at which the regular parameters (e.g., slope coefficients) converge.

In this study, we examine the asymptotic properties of the Bayesian approach under the fixed jump size framework. Specifically, we prove a Bernstein–von Mises type theorem for the slope parameters γ which validates a frequentist interpretation of Bayesian credible regions. A frequentist researcher could look at the credible interval of γ to check robustness with respect to the uncertainty of the break location. Such sensitivity analysis is reasonable as our result guarantees the credible interval to converge to the conventional confidence interval. We first establish theoretical results under normal likelihood and natural conjugate prior. We further extend the results to non-conjugate priors using Laplace approximation.

The literature on theoretical properties of Bayesian approaches in non-regular models such as (1) is very scarce despite their popularity in applications. To our knowledge, frequentist properties of Bayesian approaches for linear regression models with structural breaks have not been studied in the literature. Ghosal and Samanta (1995) consider a general non-regular estimation problem from a Bayesian perspective and establish conditions under which the Bernstein–von Mises theorem holds for the regular part of the parameter. However, their assumptions are difficult to verify in regard to our model in consideration.

Recently, Casini and Perron (2020) propose a Quasi-Bayes estimator of the break location τ , which is defined by an integration rather than an optimization. Their approach provides a better approximation about the uncertainty of τ than the conventional method. Although our focus of the current paper is on inference about the slope coefficients γ and not τ , our Bayesian approach toward inference shares the same spirit; any probabilistic statement about γ is expressed as a weighted average (2) over the marginal posterior density of τ .

The paper is organized as follows. Section 2 introduces the model and lists a set of assumptions. Section 3 describes a Bayesian approach based on normal likelihood and conjugate prior. The section then establishes frequentist properties of the approach. Section 4 extends the results to non-conjugate priors. Section 5 presents simulation evidence to assess the adequacy of the asymptotic theory and to illustrate that conventional confidence intervals on the slope parameters tend to undercover. Section 6 reports an empirical application to the stock return prediction model of Paye and Timmermann (2006). Section 7 concludes the paper. The mathematical proofs are listed in Appendix A. Proofs of the intermediate results and additional tables for simulation studies are provided in the online appendix.

2. The model and data generating process

2.1. The model

Using the reparametrization $x_t = (w'_t, z'_t)'$, $\beta = (\alpha', \delta'_1)'$, and $\delta = \delta_2 - \delta_1$, the Eqs. (1) can be rewritten as

$$y_t = \begin{cases} x'_t \beta + \epsilon_t, & \text{for } i = 1, \dots, \lfloor \tau T \rfloor \\ x'_t \beta + z'_t \delta + \epsilon_t, & \text{for } i = \lfloor \tau T \rfloor + 1, \dots, T. \end{cases} \tag{3}$$

Note that z_t is a subvector of x_t . More generally, let $z_t = R'x_t$, where R is a $d_x \times d_z$ known matrix with full column rank and hence z_t is defined as a linear transformation of x_t . For $R = (0_{d_z \times d_w}, I_{d_z})'$, we obtain model (3). For $R = I_{d_x}$, a pure change model is obtained. To rewrite the model in matrix form, we introduce further notations. Define $Y = (y_1, \dots, y_T)'$, $\epsilon = (\epsilon_1, \dots, \epsilon_T)'$, $X = (x_1, \dots, x_T)'$, $X_{1\tau} = (x_1, \dots, x_{\lfloor \tau T \rfloor}, 0, \dots, 0)'$, and $X_{2\tau} = (0, \dots, 0, x_{\lfloor \tau T \rfloor + 1}, \dots, x_T)'$. Define Z , $Z_{1\tau}$, and $Z_{2\tau}$ similarly. Then, $Z = XR$, $Z_{1\tau} = X_{1\tau}R$, and $Z_{2\tau} = X_{2\tau}R$. Now, Eqs. (3) can be written as

$$Y = X\beta + Z_{2\tau}\delta + \epsilon = \chi_\tau \gamma + \epsilon, \tag{4}$$

where $\chi_\tau = (X, Z_{2\tau})$ and $\gamma = (\beta', \delta')'$. Let $\mathcal{H} \subset (0, 1)$ be the space of the break locations. For a given $\tau \in \mathcal{H}$, define $S_T(\tau) = Y'Y - Y'\chi_\tau (\chi'_\tau \chi_\tau)^{-1} \chi'_\tau Y$. The least-squares estimators of τ and γ are defined as $\hat{\tau}_{LS} = \operatorname{argmin}_{\tau \in \mathcal{H}} S_T(\tau)$ and $\hat{\gamma}_{LS} = \hat{\gamma}(\hat{\tau}_{LS})$, respectively, where $\hat{\gamma}(\tau)$ denotes the usual OLS estimator given the value of τ .

2.2. Data generating process

The data are assumed to include T observations on a response and a vector of covariates: $\mathbf{D}_T = (\mathbf{Y}_T, \mathbf{X}_T) = (y_1, \dots, y_T, x_1, \dots, x_T)$ where $y_t \in \mathbb{R}$ and $x_t \in \mathbb{R}^{d_x}$, $t = 1, \dots, T$. Conditional on \mathbf{X}_T , the response is generated according to model (4) with the true parameters $(\gamma'_0, \sigma_0^2, \tau_0)'$. We let $\theta = (\gamma', \sigma^2)'$ denote the regression parameters. We make the following assumptions about the true data-generating-process (DGP):

Assumption 1.

- (i) $\delta_0 \neq 0$.
- (ii) ϵ_t is i.i.d. with $E(\epsilon_t|x_t) = 0$, $E(\epsilon_t^2|x_t) = \sigma_0^2$, where σ_0^2 is unknown to the econometrician.
- (iii) $\Sigma_X = E[x_t x'_t] = \operatorname{plim} \frac{1}{T} \sum_{t=1}^T x_t x'_t$ exists and is positive definite.
- (iv) For all $\tau_1, \tau_2 \in (0, 1)$ with $\tau_1 < \tau_2$, $\frac{1}{T} \sum_{\lfloor \tau_1 T \rfloor + 1}^{\lfloor \tau_2 T \rfloor} x_t \epsilon_t = O_p(T^{-1/2})$ and $\frac{1}{T} \sum_{\lfloor \tau_1 T \rfloor + 1}^{\lfloor \tau_2 T \rfloor} x_t x'_t = (\tau_2 - \tau_1)\Sigma_X + O_p(T^{-1/2})$

Under the above assumptions, the classical theoretical results apply. Bai (1997) shows that the convergence rate of $\hat{\tau}_{LS}$ is T^{-1} if δ_0 is fixed with respect to the sample size: $\hat{\tau}_{LS} = \tau_0 + O_p(T^{-1})$, and that the least-squares estimator for γ is asymptotically normal with the asymptotic covariance matrix being the same as if τ_0 is known:

$$\sqrt{T}(\hat{\gamma}_{LS} - \gamma_0) \xrightarrow{d} N_{(d_x+d_z)}(0, \sigma_0^2 V^{-1}), \tag{5}$$

where

$$V = \text{plim } T^{-1} \begin{pmatrix} \sum_{t=1}^T x_t x_t' & \sum_{t=\lfloor \tau_0 T \rfloor + 1}^T x_t z_t' \\ \sum_{t=\lfloor \tau_0 T \rfloor + 1}^T z_t x_t' & \sum_{t=\lfloor \tau_0 T \rfloor + 1}^T z_t z_t' \end{pmatrix} = \text{plim } T^{-1} \chi'_{\tau_0} \chi_{\tau_0}.$$

This means that τ can be treated as known for the purpose of inference about γ . In other words, the uncertainty of the break location is essentially ignored, and thus the confidence interval for γ tends to undercover in finite samples (see Section 5 for simulation).

There are several comments on Assumption 1. In threshold regression models (see Hansen, 2000), the threshold variable is often one of the regressors. In this case, sorting the threshold variable leads to a trend in the regressors, which requires an alternative approach for the asymptotic analysis. We do not consider the case with one of the regressors being the threshold variable in this paper. In addition, we require the regression errors to be i.i.d. with variance σ^2 . Adding more flexibility such as heteroscedasticity and serial correlation would be an important future direction.

3. A Bayesian approach under normal likelihood and conjugate prior

The distribution of covariates is assumed to be ancillary and it is not modeled. Throughout this paper, we assume the normal likelihood function⁵

$$p(\mathbf{Y}_T | \mathbf{X}_T, \theta, \tau) = \prod_{t=1}^T \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \chi'_{\tau,t}\gamma)^2}{2\sigma^2}\right), \tag{6}$$

where $\chi'_{\tau,t}$ is the t th row of the matrix χ_τ . Note that the normality is not assumed for the true DGP, so the model can be mis-specified.

The break location τ and the regression parameters θ are independent a-priori and the prior on θ is the natural conjugate prior. That is, $\pi(\gamma, \sigma^2, \tau) = \pi(\gamma | \sigma^2) \pi(\sigma^2) \pi(\tau)$ where the prior on γ conditional on σ^2 is normal $N_{(d_x+d_z)}(\underline{\mu}, \sigma^2 \underline{H}^{-1})$ and the prior on σ^2 is inverse-gamma $InvGamma(\underline{a}, \underline{b})$. Note that by taking $\underline{H} \rightarrow 0$, $\underline{a} \rightarrow -(d_x + d_z)/2$, and $\underline{b} \rightarrow 0$, we have the uninformative improper prior $\pi(\gamma, \sigma^2) \propto \sigma^{-2}$ as a special case. The prior on τ can be of any form as long as it is positive at τ_0 , and $\pi(\tau)$ is finite for all $\tau \in \mathcal{H}$.

The conjugate prior is a popular choice in the Bayesian estimation of linear regression models. Our restriction on the prior for the break location τ is very mild. For example, the uniform distribution on \mathcal{H} satisfies the requirement. Recently, Baek (2018) investigates the same model (1). As the distribution of $\hat{\tau}_{LS}$ might exhibit tri-modality for small jumps, Baek proposes a new point estimator for τ based on a modified objective function. The proposed modification can be regarded as equivalent to specifying a certain type of prior for τ , and indeed such prior satisfies our restriction.

Under the normal likelihood function and the prior defined above, the posterior distributions are

$$\pi_T(\tau) \propto [\det(\bar{H}_\tau)]^{-0.5} \bar{b}_\tau^{-\bar{a}} \times \pi(\tau), \tag{7}$$

$$\gamma | \tau, \mathbf{D}_T \sim t_{(d_x+d_z)}(2\bar{a}, \bar{\mu}_\tau, (\bar{b}_\tau/\bar{a})\bar{H}_\tau^{-1}), \tag{8}$$

$$\sigma^2 | \tau, \mathbf{D}_T \sim InvGamma(\bar{a}, \bar{b}_\tau), \tag{9}$$

where $\bar{H}_\tau = \underline{H} + \chi'_\tau \chi_\tau$, $\bar{\mu}_\tau = \bar{H}_\tau^{-1} [\underline{H}\underline{\mu} + \chi'_\tau Y]$, $\bar{b}_\tau = \underline{b} + 0.5 [\underline{\mu}' \underline{H} \underline{\mu} + Y'Y - \bar{\mu}'_\tau \bar{H}_\tau \bar{\mu}_\tau]$, and $\bar{a} = \underline{a} + T/2$, and $t_k(v, \mu, \Sigma)$ is the k -dimensional t-distribution with v degrees of freedom, a location vector $\mu \in \mathbb{R}^k$, and a $k \times k$ shape matrix Σ . See the online appendix for a derivation.

Due to the availability of the closed-forms for the posterior distributions conditional on τ , the posterior sampling is simple and fast. One can first draw $\tau_{(1)}, \dots, \tau_{(S)}$ from the marginal posterior of τ as in (7) via, for example, the Metropolis–Hastings algorithm, where S is the number of posterior draws. For each $\tau_{(s)}$, one can sample posterior draws of $\sigma^2_{(s)}$ from the posterior conditional on $\tau = \tau_{(s)}$, namely (9). Conditional on τ and σ^2 , one can draw γ from $p(\gamma | \sigma^2, \tau, \mathbf{D}_T)$.⁶ For example, a laptop with a 2.2 GHz processor and 8 GB RAM takes about 4.1 s to sample 10,000 posterior draws in an empirical example in Section 6 that has ten slope coefficients in total.

⁵ Similarly, Qu and Perron (2007) propose a quasi-maximum likelihood estimator assuming normal errors.

⁶ It can be shown that $\gamma | \sigma^2, \tau, \mathbf{D}_T \sim N_{(d_x+d_z)}(\bar{\mu}_\tau, \sigma^2 \bar{H}_\tau^{-1})$.

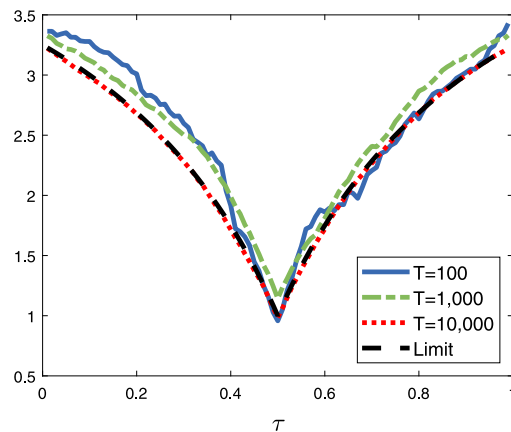


Fig. 2. Example of $T^{-1}S_T(\tau)$ with $T = 100$ (solid, blue), $T = 1000$ (dash-dotted, green), and $T = 10,000$ (dotted, red) and $\text{plim} T^{-1}S_T(\tau)$ (dashed, black). $\tau_0 = 0.5$.

3.1. Asymptotic theory

We investigate the asymptotic behavior of the Bayesian method under the normal likelihood and the conjugate prior defined above. We do so in two steps. Section 3.1.1 shows that the marginal posterior of the break location τ contracts to the true value τ_0 at the rate of T^{-1} , the same rate at which the least-squares estimator $\hat{\tau}_{LS}$ converges. The proof is based on studying the behavior of the log ratio of the marginal posterior densities of τ . In addition, we establish the limiting distribution of the posterior mode of τ . Section 3.1.2 establishes a Bernstein–von Mises type theorem for the regression slope coefficients γ . The proof is based on the T -consistency of the marginal posterior of τ and the fact that the conditional posterior for $\sqrt{T}(\gamma - \hat{\gamma}_{LS})$ given τ is asymptotically normal. Proofs of the theorems can be found in Appendix A.

3.1.1. Marginal posterior of τ

The first step for proving the Bernstein–von Mises theorem is the marginal posterior consistency of τ at rate T^{-1} . Marginal posteriors have not been studied extensively or systematically in the literature. Here, we directly analyze the form of the marginal posterior of τ . Let $L_T(\tau)$ be the marginal likelihood conditional on τ , that is $L_T(\tau) = \int p(\mathbf{Y}_T | \mathbf{X}_T, \theta, \tau) \pi(\theta, \tau) d\theta$, which is available up to a multiplicative constant under the normal likelihood and the conjugate prior as can be seen in (7). The marginal posterior density $\pi_T(\tau)$ of τ is defined as

$$\pi_T(\tau) = \frac{L_T(\tau)}{\int L_T(\tau) d\tau}.$$

The following theorem establishes the first step for proving the Bernstein–von Mises theorem, the T -consistency of the marginal posterior of τ . It states that the posterior mass outside of a ball around τ_0 with radius proportional to T^{-1} will be asymptotically negligible.

Theorem 1 (Marginal Posterior Consistency of τ at Rate T^{-1}). Suppose Assumption 1 holds. Then, under the normal likelihood and the conjugate prior described above, $\forall \xi > 0, \epsilon > 0, \exists M > 0$ and $k > 0$ such that $T \geq k \implies$

$$P_{\theta_0, \tau_0} \left(\int_{B_{M/T}^c(\tau_0)} \pi_T(\tau) d\tau < \xi \right) > 1 - \epsilon,$$

where for any constant $d > 0$, $B_d^c(\tau_0)$ denotes the set difference $\mathcal{H} \setminus (\tau_0 - d, \tau_0 + d)$.

The proof of Theorem 1 is built on some intermediate steps, Propositions 1–4. It can be shown that $\int_{B_{M/T}^c(\tau_0)} \pi_T(\tau) d\tau$ is bounded by the product of $\int_{B_{M/T}^c(\tau_0)} \frac{L_T(\tau)}{L_T(\tau_0)} d\tau$ and the inverse of $\int_{B_{M_0/T}^c(\tau_0)} \frac{L_T(\tau')}{L_T(\tau_0)} d\tau'$ for each T and for any $M_0 > 0$. Proposition 1 shows that under the normal likelihood and the conjugate prior, due to the availability of the marginal likelihood conditional on τ up to a normalization constant as in (7), studying the log marginal likelihood ratio boils down to comparing the sum of squared residuals $S_T(\tau)$. Proposition 2 establishes the probability limit of $T^{-1}S_T(\tau)$, for which we show an example in Fig. 2. We then show that the limit of $T^{-1}S_T(\tau)$ achieves a unique minimum at τ_0 (Proposition 3), and study the modulus of continuity of an appropriate empirical process (Proposition 4) in order to derive bounds. The detail of the proof of Theorem 1 can be found in Appendix A.1.

The Bayesian counterpart of the least-squares estimator $\hat{\tau}_{LS}$ would be the posterior mode: $\hat{\tau}_{Bayes} = \operatorname{argmax}_{\tau \in \mathcal{H}} \pi_T(\tau)$. Bai (1997) shows that $\operatorname{argmax}_m W^*(m)$ is the asymptotic distribution of $\hat{\tau}_{LS}$.⁷ A consequence of the proof of Theorem 1 is that $\hat{\tau}_{Bayes}$ converges to the same limiting distribution. See Appendix A.2 for a proof.

Corollary 1 (Limiting Distribution of the Posterior Mode of τ). Suppose Assumption 1 holds. Then, under the normal likelihood and the conjugate prior described above,

$$\lfloor T(\hat{\tau}_{Bayes} - \tau_0) \rfloor \xrightarrow{d} \operatorname{argmax}_m W^*(m).$$

3.1.2. Bernstein–von Mises theorem for γ

The marginal posterior of γ is a mixture with weights corresponding to the marginal posterior density $\pi_T(\tau)$. Furthermore, due to Theorem 1, we can focus our attention on the values of τ in a T^{-1} neighborhood of τ_0 :

$$\int p(\gamma|\tau, \mathbf{D}_T)\pi_T(\tau)d\tau = \int_{B_{M/T}(\tau_0)} p(\gamma|\tau, \mathbf{D}_T)\pi_T(\tau)d\tau + o_p(1).$$

We are now ready to establish the Bernstein–von Mises type result.

Theorem 2 (Bernstein–von Mises Theorem for the Slope Coefficients). Suppose Assumption 1 holds. Then, under the normal likelihood and the conjugate prior described above,

$$d_{TV} \left(\pi \left[\sqrt{T}(\gamma - \hat{\gamma}_{LS}) \middle| \mathbf{D}_T \right], N_{(d_x+d_z)}(0, \sigma_0^2 V^{-1}) \right) \rightarrow 0,$$

in P_{θ_0, τ_0} – probability where d_{TV} is the total variation distance.

The proof of Theorem 2 exploits the fact that the conditional posterior for $\sqrt{T}(\gamma - \hat{\gamma}_{LS})$ given τ is asymptotically normal, which is close to the asymptotic distribution of $\hat{\gamma}_{LS}$ when τ is close to τ_0 . A bound on the Kullback–Leibler (KL) divergence between two normal densities together with the T -consistency is used to make the argument precise. The proof is presented in Appendix A.3.

4. An extension to non-conjugate priors

The previous section establishes the asymptotic properties of the posterior distributions under the conjugate prior. A natural question is whether these results can be extended to other priors. For example, an independent prior between the slope coefficients γ and the error variance σ^2 , e.g., $\pi(\gamma, \sigma^2) = \pi(\gamma)\pi(\sigma^2)$ with $\gamma \sim N_{(d_x+d_z)}(\underline{\mu}, \underline{\Sigma})$ and $\sigma^2 \sim \operatorname{InvGamma}(\underline{a}, \underline{b})$, is a popular choice for Bayesian estimation of regression models in practice. Under the normal likelihood and the conjugate prior, the analytical expressions of the marginal posterior of τ up to a normalization constant (7) and the conditional posterior of γ given τ (8) facilitate the theoretical analysis. They are not available, for instance, under the independent prior mentioned above. In this section, we extend the theoretical results by keeping the normal likelihood (6) but without requiring the conjugate prior on θ . In order to study the asymptotic behavior of the posterior distributions without having their closed-form expressions, we employ a Laplace approximation type result of Hong and Preston (2012). To do so, we make an additional assumption as shown below. Let $\hat{\theta}(\tau)$ be the maximum likelihood estimator of θ conditional on $\tau \in \mathcal{H}$, i.e., $\hat{\theta}(\tau) = \operatorname{arg sup}_{\theta \in \Theta} \log p(\mathbf{Y}_T | \mathbf{X}_T, \theta, \tau)$. Denote by $\theta^*(\tau)$ the corresponding pseudo true parameter value that minimizes the KL divergence between the model $p(\mathbf{Y}_T | \mathbf{X}_T, \theta, \tau)$ and the DGP.

Assumption 2.

- (i) There is a compact convex subset Θ of $\mathbb{R}^{d_x+d_z+1}$ such that $\theta^*(\tau) \in \operatorname{int}(\Theta)$ for all $\tau \in \mathcal{H}$.
- (ii) The prior $\pi(\theta, \tau)$ is supported on $\Theta \times \mathcal{H}$. It is continuous in θ and bounded away from 0 and ∞ around $(\theta^*(\tau), \tau)$ for all $\tau \in \mathcal{H}$.

⁷ $W^*(m)$ is a stochastic process defined on the set of integers as follows: $W^*(0) = 0$, $W^*(m) = W_1(m)$ for $m < 0$, and $W^*(m) = W_2(m)$ for $m > 0$, with

$$W_1(m) = -\delta_0 \sum_{i=m+1}^0 z_i z_i' \delta_0 + 2\delta_0 \sum_{i=m+1}^0 z_i \epsilon_i, \text{ for } m = -1, -2, \dots$$

$$W_2(m) = -\delta_0 \sum_{i=1}^m z_i z_i' \delta_0 - 2\delta_0 \sum_{i=1}^m z_i \epsilon_i, \text{ for } m = 1, 2, \dots$$

Under the normal likelihood and Assumption 1, together with Assumption 2, we can invoke the Laplace approximation result, Theorem 3 of Hong and Preston (2012). Note that, under the normal likelihood and Assumption 1, $\theta^*(\tau)$ exists and is a function of parameters in the DGP. In this section, we no longer assume the natural conjugate prior on θ . For instance, the independent prior $\pi(\gamma, \sigma^2, \tau) = \pi(\gamma)\pi(\sigma^2)\pi(\tau)$ mentioned above satisfies the conditions in (ii) of Assumption 2 as long as they are truncated on Θ and $\pi(\tau)$ is positive and finite at all τ .

Theorem 3 establishes the T -consistency of the marginal posterior of τ for non-conjugate priors under the additional conditions.

Theorem 3 (Marginal Posterior Consistency of τ at Rate T^{-1} , Non-Conjugate Priors). Suppose Assumptions 1 and 2 hold. Then, under the normal likelihood, $\forall \xi > 0, \epsilon > 0, \exists M > 0$ and $k > 0$ such that $T \geq k \implies$

$$P_{\theta_0, \tau_0} \left(\int_{B_{M/T}^c(\tau_0)} \pi_T(\tau) d\tau < \xi \right) > 1 - \epsilon,$$

where for any constant $d > 0$, $B_d^c(\tau_0)$ denotes the set difference $\mathcal{H} \setminus (\tau_0 - d, \tau_0 + d)$.

Recall that while proving the T -consistency under the conjugate prior (i.e., Theorem 1), we utilize the closed-form expression of the marginal posterior of τ up to a multiplicative constant (7) in order to study the behavior of the marginal likelihood ratio conditional on τ . Under non-conjugate priors, such expression is not available in general. For this reason, we invoke Theorem 3 of Hong and Preston (2012) to approximate the quantity $\int p(\mathbf{Y}_T | \mathbf{X}_T, \theta, \tau) \pi(\theta, \tau) d\theta$ to prove our Theorem 3. See Appendix A.4 for the detail.

As in the previous section, an implication of the T -consistency of the marginal posterior of τ is that the posterior mode converges to the limiting distribution of $\hat{\tau}_{LS}$. Proof is in Appendix A.5.

Corollary 2 (Limiting Distribution of the Posterior Mode of τ , Non-Conjugate Priors). Suppose Assumptions 1 and 2 hold. Then, under the normal likelihood,

$$\lfloor T (\hat{\tau}_{Bayes} - \tau_0) \rfloor \xrightarrow{d} \operatorname{argmax}_m W^*(m),$$

where the stochastic process $W^*(m)$ is defined in Section 3.1.1.

Theorem 4 establishes our main theoretical result, the Bernstein–von Mises theorem for γ , under the prior defined in Assumption 2 (ii).

Theorem 4 (Bernstein–von Mises Theorem for the Slope Coefficients, Non-Conjugate Priors). Suppose Assumptions 1 and 2 hold. Then, under the normal likelihood,

$$d_{TV} \left(\pi \left[\sqrt{T} (\gamma - \hat{\gamma}_{LS}) \middle| \mathbf{D}_T \right], N_{(d_x + d_z)}(0, \sigma_0^2 V^{-1}) \right) \rightarrow 0,$$

in P_{θ_0, τ_0} – probability where d_{TV} is the total variation distance.

When proving the corresponding result under the conjugate prior (i.e., Theorem 2), we utilize the closed-form expression of the marginal posterior of γ given τ (8). As this is not available under the prior in this section, we again use an approximation to study the asymptotic behavior of the marginal posterior. See Appendix A.6 for a proof.

5. Simulation

The main purpose of the simulation studies below is to compare inference on the slope parameters between the two methods: the conventional least-squares method in Bai (1997) and the Bayesian approach described in our paper. For the Bayesian approach, we use the uniform prior for τ and the conjugate prior for the regression parameters with $\underline{H} = 0.1I_{(d_x + d_z)}$, $\underline{\mu} = 0_{(d_x + d_z)}$, and $\underline{a} = \underline{b} = 1$. The findings are similar even when we use the uninformative improper prior. Following the literature (e.g., Casini and Perron, 2021), we set the range of the candidate values of τ to be $(\epsilon, 1 - \epsilon)$ with $\epsilon = 0.05$ for all methods.⁸

We consider the following model: $y_t = \delta_0 1(t > \lfloor \tau_0 T \rfloor) + \epsilon_t$. We let $\tau_0 = 0.5$ and consider different values of the jump size $\delta_0 \in \{0.25, 0.5, 1.0, 2.0\}$ and the sample size $T \in \{20, 50, 100, 250, 500, 1000\}$. In order to compare the methods in repeated experiments, for each combination of δ_0 and T , we generate 1000 data sets. The error ϵ_t is independently and identically generated from $N(0, 1)$. In the online appendix, we present robustness checks under $\tau_0 = 0.3$ as well as the errors generated from a non-normal distribution, and illustrate that the overall findings are similar to what we present in this section.

⁸ It prevents the break location estimator from being in the first and last 100% of the sample. The trimming parameter ϵ should not be chosen too high otherwise it might introduce bias in the break location estimate. Casini and Perron (2021) find the choice $\epsilon = 0.05$ performs well in general, which we also confirm in our simulation exercises.

Table 1
Simulation results for δ .

$\delta_0 =$	Least-squares				Bayesian				Bayesian (τ fixed at $\hat{\tau}_{LS}$)			
	0.25	0.50	1.00	2.00	0.25	0.50	1.00	2.00	0.25	0.50	1.00	2.00
Coverage												
$T = 20$	0.68	0.77	0.88	0.95	0.96	0.97	0.96	0.95	0.65	0.70	0.88	0.95
$T = 50$	0.67	0.84	0.94	0.96	0.96	0.97	0.96	0.95	0.64	0.82	0.93	0.94
$T = 100$	0.69	0.90	0.96	0.95	0.96	0.97	0.96	0.95	0.69	0.89	0.95	0.95
$T = 250$	0.83	0.94	0.94	0.96	0.96	0.96	0.94	0.96	0.83	0.94	0.95	0.94
$T = 500$	0.91	0.95	0.95	0.96	0.97	0.96	0.96	0.96	0.88	0.96	0.95	0.95
$T = 1000$	0.93	0.94	0.95	0.96	0.96	0.95	0.95	0.96	0.94	0.95	0.95	0.96
Length												
$T = 20$	3.87	3.60	3.20	2.82	4.85	4.59	4.20	3.16	3.46	3.30	2.99	2.76
$T = 50$	2.31	2.07	1.82	1.76	2.91	2.67	2.13	1.79	2.21	2.04	1.79	1.74
$T = 100$	1.61	1.38	1.26	1.24	2.10	1.78	1.34	1.25	1.56	1.36	1.23	1.24
$T = 250$	0.93	0.81	0.78	0.78	1.21	0.92	0.80	0.79	0.93	0.80	0.78	0.78
$T = 500$	0.61	0.56	0.55	0.55	0.76	0.58	0.56	0.56	0.61	0.56	0.55	0.55
$T = 1000$	0.41	0.39	0.39	0.39	0.46	0.40	0.39	0.40	0.41	0.39	0.39	0.39
MSE												
$T = 20$	3.85	2.79	1.75	0.60	1.13	0.91	0.86	0.58	3.00	2.59	1.39	0.47
$T = 50$	1.35	0.78	0.26	0.20	0.42	0.33	0.25	0.20	1.28	0.79	0.28	0.20
$T = 100$	0.67	0.28	0.11	0.10	0.21	0.15	0.11	0.10	0.65	0.27	0.10	0.10
$T = 250$	0.18	0.05	0.04	0.04	0.08	0.05	0.04	0.04	0.18	0.05	0.04	0.04
$T = 500$	0.05	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.06	0.02	0.02	0.02
$T = 1000$	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

The columns 2–5 and 6–9 of Table 1 show the simulation results concerning δ for the least-squares estimator and the Bayesian estimator, respectively. The top panel “Coverage” shows empirical coverages of the true jump size δ_0 by the 95% confidence and credible intervals. The frequentist confidence intervals are computed based on the conventional asymptotic theory (5). For the Bayesian approach, we report the equal-tailed credible intervals. The middle panel “Length” presents the average lengths of the aforementioned intervals. The bottom panel “MSE for δ ” shows the mean-squared-errors for the point estimator, which is the least-squares estimator $\hat{\delta}_{LS}$ for the conventional method and the posterior mean for the Bayesian approach.

There are several significant findings. First, for small T and/or small δ_0 , the conventional confidence intervals significantly undercover. Meanwhile, the Bayesian credible intervals have relatively reasonable coverages. Second, the Bayesian intervals tend to be longer than the conventional confidence intervals for small T and/or δ_0 . Third, as T increases, the discrepancy between the two methods decreases, as expected from the Bernstein–von Mises theorem that we establish.

Table 2 shows the results of estimation and inference of the break location τ . Although the main focus of the current paper is on inference about the slope parameters and not on inference about τ , we report the empirical coverage and the length of the 95% confidence interval of Bai (1997) and the highest posterior density (HPD) set.⁹ We also report the inverted likelihood ratio (ILR) confidence set suggested by Eo and Morley (2015).

Overall, the HPD set and the ILR confidence set of the break location τ behave similarly although the HPD set slightly undercovers relative to the ILR confidence set for small T and/or δ_0 . We confirm several findings of Eo and Morley (2015). First, when T is large, the confidence interval of Bai has longer lengths than the ILR confidence set and the HPD set.¹⁰ Second, when T and δ_0 are small, the confidence interval of Bai tends to severely undercover compared to the ILR confidence set and the HPD set. The interval of Bai indeed has a shorter length than the other two sets for small T , but its undercoverage raises concerns for small samples in practice.¹¹ The bottom panels of Table 2 show the mean-absolute-error (MAE) of the point estimator of τ which is $\hat{\tau}_{LS}$ for the conventional method and the posterior mode $\hat{\tau}_{Bayes}$ for the Bayesian approach. It is known that the finite-sample distribution of the least-squares estimator $\hat{\tau}_{LS}$ tends to be trimodal (see Baek, 2018) when the jump size is relatively small. The same seems to be true for the Bayesian point estimator (see Fig. 1).

To better understand the importance of the uncertainty of the break location τ for inference on the slope parameters, we conduct a hypothetical experiment. We repeat the simulation exercise but now fixing τ at the least-squares estimate

⁹ Note that although we prove that the posterior mode of the break location τ converges to the limiting distribution of the least-squares estimator, whether the posterior distribution of τ converges to the same limit or not is still an open question. The Bayesian literature on the Bernstein–von Mises-like result for non-regular parameters is very scarce. To our best knowledge, the only available work is that by Kleijn and Knapik (2012) whose results do not seem to be applicable to the model in consideration in this paper. Hence, it is not guaranteed that a credible set of τ has frequentist coverage even asymptotically. However, we emphasize that credible sets on τ still have a statistically valid interpretation even in finite samples.

¹⁰ Eo and Morley (2015) explain that the likelihood ratio test is more powerful than the Wald-type test used to construct the confidence interval of Bai, which results in a shorter length of the ILR confidence set.

¹¹ In addition, as also reported by Eo and Morley (2015), the ILR confidence set tends to slightly overcover even in large sample.

Table 2
Simulation results for τ .

$\delta_0 =$	Least-squares				Bayesian				ILR			
	0.25	0.50	1.00	2.00	0.25	0.50	1.00	2.00	0.25	0.50	1.00	2.00
Coverage												
$T = 20$	0.50	0.58	0.75	0.93	0.83	0.87	0.94	0.97	0.91	0.92	0.92	0.95
$T = 50$	0.51	0.68	0.87	0.97	0.83	0.92	0.96	0.97	0.93	0.93	0.96	0.98
$T = 100$	0.53	0.78	0.91	0.96	0.85	0.96	0.95	0.97	0.93	0.96	0.96	0.98
$T = 250$	0.67	0.87	0.94	0.97	0.91	0.94	0.94	0.97	0.94	0.95	0.96	0.98
$T = 500$	0.75	0.93	0.96	0.98	0.93	0.95	0.94	0.96	0.95	0.96	0.97	0.98
$T = 1000$	0.85	0.92	0.96	0.97	0.92	0.90	0.91	0.95	0.95	0.96	0.97	0.98
Length ($\times 100$)												
$T = 20$	47.9	50.0	50.1	32.1	83.1	80.5	68.1	29.8	83.5	80.9	62.8	22.9
$T = 50$	48.3	50.6	40.2	14.0	76.1	69.3	40.0	9.01	80.9	71.4	37.0	9.18
$T = 100$	46.8	47.5	24.8	6.67	73.8	58.6	19.2	3.92	78.9	58.9	18.0	4.23
$T = 250$	47.6	34.3	9.78	2.66	64.2	30.6	5.83	1.50	67.7	28.9	6.37	1.68
$T = 500$	44.0	18.9	4.76	1.31	50.1	12.4	2.60	0.73	51.2	12.7	3.03	0.81
$T = 1000$	33.8	9.49	2.35	0.64	28.4	5.02	1.21	0.38	28.9	5.99	1.51	0.42
MAE ($\times 10$)												
$T = 20$	2.63	2.27	1.43	0.37	3.13	2.72	1.64	0.38				
$T = 50$	2.45	1.84	0.69	0.14	2.88	2.12	0.78	0.14				
$T = 100$	2.32	1.24	0.38	0.06	2.76	1.50	0.39	0.06				
$T = 250$	1.70	0.60	0.13	0.03	2.02	0.69	0.13	0.03				
$T = 500$	1.20	0.27	0.06	0.01	1.31	0.28	0.06	0.01				
$T = 1000$	0.63	0.13	0.03	0.01	0.70	0.14	0.03	0.01				

$\hat{\tau}_{LS}$ in the Bayesian approach. The columns 10–13 of Table 1 display the results. We see that the credible intervals undercover for small T and/or small δ_0 as is the case with the conventional confidence intervals of δ . They also have similar lengths in general. Importantly, the credible intervals when τ is fixed at $\hat{\tau}_{LS}$ have shorter lengths compared to the full Bayesian intervals. On average, the full Bayesian credible intervals are 17.1% longer¹² than the credible intervals produced by fixing the value of τ at $\hat{\tau}_{LS}$. Note that a Bayesian equivalent of the conventional approach to inference on the slope parameters would be to fix the value of τ at the posterior mode (whose value is very similar to $\hat{\tau}_{LS}$ as we can see from Fig. 1 and deduce from Corollary 1). We can see in Fig. 1 that both $\hat{\tau}_{LS}$ and the posterior mode of τ display significant amount of variations in finite samples. Fixing τ at a point estimate forces the Bayesian approach to ignore this uncertainty of τ ; as a result, the credible interval on δ becomes shorter and hence undercovers. The full Bayesian approach takes into account such uncertainty via the marginal posterior of τ (see examples of the density in Fig. 1), which results in longer lengths of the credible intervals on the slope parameters that help them avoid undercoverages. In contrast, by construction (i.e., Eq. (5)), the conventional confidence intervals do not have this feature.

In summary, the simulation exercises demonstrate that (1) the credible intervals on the slope coefficient tend to have more reasonable coverages than the conventional confidence intervals because of longer lengths, (2) the longer length of the credible intervals is a reflection of the uncertainty of the unknown¹³ break location τ , and (3) the two intervals converge to each other asymptotically as expected from our Bernstein–von Mises theorem.

6. Application

In this section, we illustrate difference in estimation and inference of the regression parameters in linear regression models with a structural break between the conventional approach and the Bayesian approach that we consider in this paper. Paye and Timmermann (2006) consider the problem of ex-post prediction in stock returns under a structural break in the coefficients of state variables. Their multivariate model with a structural break is

$$Ret_t = \begin{cases} \delta_1^{(1)} + \delta_1^{(2)}Div_{t-1} + \delta_1^{(3)}Tbill_{t-1} + \delta_1^{(4)}Spread_{t-1} + \delta_1^{(5)}Def_{t-1} + \epsilon_t, & \text{if } t \leq \lfloor \tau T \rfloor, \\ \delta_2^{(1)} + \delta_2^{(2)}Div_{t-1} + \delta_2^{(3)}Tbill_{t-1} + \delta_2^{(4)}Spread_{t-1} + \delta_2^{(5)}Def_{t-1} + \epsilon_t, & \text{if } t > \lfloor \tau T \rfloor, \end{cases}$$

where Ret_t is the excess return for the stock index during month t , Div_{t-1} is the lagged dividend yield, $Tbill_{t-1}$ is the lagged local country short interest rate, $Spread_{t-1}$ is the lagged local country term spread, and Def_{t-1} is the lagged U.S. default premium. The authors estimate the model using the conventional frequentist approach: they first compute $\hat{\tau}_{LS}$

¹² The difference is larger when T and/or δ_0 are/is smaller.

¹³ When τ_0 is known, the two intervals behave very similarly. To illustrate this point, we conduct another hypothetical experiment by repeating the simulation exercise as before but now fixing the value of τ at the true value τ_0 in both conventional and Bayesian approaches. Table 3 summarizes the results. In this case, we see that both confidence and credible intervals have coverages quite close to 95% in all cases. They also have similar lengths. Note that when the true value τ_0 is given, the usual asymptotic normality and the regular Bernstein–von Mises theorem apply. As a consequence, both frequentist and Bayesian intervals seem to converge faster to the limit compared to the case with unknown τ .

Table 3
Simulation results for δ , τ fixed at τ_0 .

$\delta_0 =$	Least-squares				Bayesian			
	0.25	0.50	1.00	2.00	0.25	0.50	1.00	2.00
Coverage								
$T = 20$	0.93	0.95	0.96	0.95	0.94	0.95	0.95	0.94
$T = 50$	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94
$T = 100$	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.96
$T = 250$	0.95	0.95	0.95	0.94	0.94	0.96	0.94	0.94
$T = 500$	0.96	0.95	0.95	0.96	0.96	0.95	0.95	0.96
$T = 1000$	0.95	0.96	0.96	0.95	0.95	0.96	0.96	0.95
Length								
$T = 20$	2.77	2.77	2.77	2.77	2.72	2.74	2.73	2.74
$T = 50$	1.75	1.75	1.75	1.75	1.75	1.74	1.74	1.75
$T = 100$	1.24	1.24	1.24	1.24	1.23	1.24	1.24	1.23
$T = 250$	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
$T = 500$	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55
$T = 1000$	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39
MSE								
$T = 20$	0.53	0.50	0.47	0.52	0.51	0.47	0.45	0.50
$T = 50$	0.21	0.20	0.21	0.21	0.21	0.20	0.20	0.20
$T = 100$	0.10	0.09	0.10	0.09	0.10	0.09	0.10	0.09
$T = 250$	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
$T = 500$	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
$T = 1000$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

and then obtain point estimates as well as confidence intervals for the slope coefficients by fixing τ at $\hat{\tau}_{LS}$. We examine whether the Bayesian method performs differently from the conventional approach.

Monthly series are collected from Global Financial Data and Federal Reserve Economic Data (FRED). In this paper, we consider estimating the model for the United Kingdom and Japan.¹⁴ The indices used to define the total return and the dividend yield are the FTSE All-share for the U.K. and Nikko Securities Composite for Japan. The dividend yield is expressed as an annual rate and is constructed as the sum of dividends over the preceding 12 months, divided by the current price. For each country, a 3-month Treasury bill rate is used as a measure of the short interest rate while the yield on a long-term government bond is used as a measure of the long interest rate. Excess returns are defined as the difference between the total return on stocks in the local currency and the local short rate. A term spread is the difference between the long and short local country interest rates. The U.S. default premium is defined as the difference in yields between Moody's Baa and Aaa rated bonds. For each country, the sample spans between January 1970 and December 2003.

For both approaches, we set the range of the candidate values of τ to be $(\epsilon, 1 - \epsilon)$ with $\epsilon = 0.05$ as we do in the simulation studies in the previous section. For the Bayesian approach, we use the uniform prior on $(\epsilon, 1 - \epsilon)$ for τ and the conjugate prior for the regression parameters with $\underline{H} = 0.1I_{(d_x+d_z)}$, $\underline{\mu} = 0_{(d_x+d_z)}$, and $\underline{a} = \underline{b} = 1$. The findings are similar even when we use the uninformative improper prior. For the break date, we compute the least-squares estimator $\hat{\tau}_{LS}$ and the posterior mode $\hat{\tau}_{Bayes}$ of τ as well as the 95% confidence interval of Bai (1997), the highest posterior density (HPD) set, and the inverted likelihood ratio (ILR) confidence set of Eo and Morley (2015). For the slope parameters, we compute $\hat{\gamma}_{LS}$ and the posterior mean of γ as well as the 90% confidence intervals of Bai (1997) based on the asymptotic result (5) and the equal-tailed credible intervals.

When the uncertainty about τ is small, estimation and inference of the slope parameters roughly match between the conventional least-squares approach and the Bayesian approach, as illustrated by our simulation studies and indicated by our proven Bernstein–von mises theorem. See Table 4 for the results for the U.K. Both methods estimate a break at 1975:01. The confidence interval of Bai (1997), the Bayesian highest posterior density (HPD) set, and the inverted likelihood ratio (ILR) confidence set by Eo and Morley (2015) are all similar and narrow, indicating that the uncertainty about τ is small. This can be seen also from the posterior density on the break date in Panel (a) of Fig. 3, which has a sharp peak around 1975:01.¹⁵ Paye and Timmermann (2006) explain that the break in the mid-1970s might be related to the large macroeconomic shocks reflecting oil price increases. As a result of the small uncertainty about τ , the point estimates of the slope parameters as well as the corresponding confidence/credible intervals are similar between the conventional and the Bayesian approach. Importantly, when the confidence interval of a given slope parameter includes (or does

¹⁴ Paye and Timmermann (2006) conduct the sequential method suggested by Bai and Perron (1998), Bai and Perron (2003), and Perron (2006) for determining the number of breaks and find multiple breaks for some countries. They find single breaks for the U.K. and Japan, but, for example, two breaks for the U.S. A fully Bayesian approach would be to place a prior on the number of breaks and use a trans-dimensional estimation method such as a reversible jump MCMC, which is beyond the scope of this paper.

¹⁵ The mean and the standard deviation of the excess return of the FTSE All-share index during the sample period are -1.53 and 6.94 respectively. At $t = 1974:12$, we have $Ret_t = -9.9$ while at $t = 1975:01$, $Ret_t = 43.75$, where the change is about 7.7 standard deviations. Therefore, the change in the dependent variable is large enough for the break point to be detected with small uncertainty.

Table 4
Estimation results for the U.K. stock return.

Slopes	Least-squares			Bayesian		
	Estimate	LB	UB	Estimate	LB	UB
$\delta_1^{(1)}$	-21.2	-28.1	-14.2	-18.6	-25.2	-11.9
$\delta_1^{(2)}$	-0.35	-3.01	2.30	0.07	-2.51	2.64
$\delta_1^{(3)}$	-0.77	-1.52	-0.03	-0.96	-1.70	-0.23
$\delta_1^{(4)}$	0.80	-0.57	2.19	0.80	-0.56	2.14
$\delta_1^{(5)}$	19.4	11.8	27.0	16.5	9.18	23.8
$\delta_2^{(1)}$	19.1	11.9	26.4	16.5	9.62	23.3
$\delta_2^{(2)}$	1.42	-1.43	4.29	0.99	-1.77	3.77
$\delta_2^{(3)}$	-0.36	-1.20	0.47	-0.16	-0.99	0.64
$\delta_2^{(4)}$	-0.98	-2.42	0.45	-0.98	-2.39	0.41
$\delta_2^{(5)}$	-19.4	-27.2	-11.6	-16.5	-23.9	-8.95
τ	Least-squares			Bayesian		
	Estimate	LB	UB	Estimate	LB	UB
	0.150	0.145	0.155	0.150	0.149	0.152
	(75:01)	(74:11)	(75:03)	(75:01)	(74:12)	(75:02)
ILR	Estimate	LB	UB			
		0.149	0.151			
		(74:12)	(75:02)			

LB = lower bound and UB = upper bound.

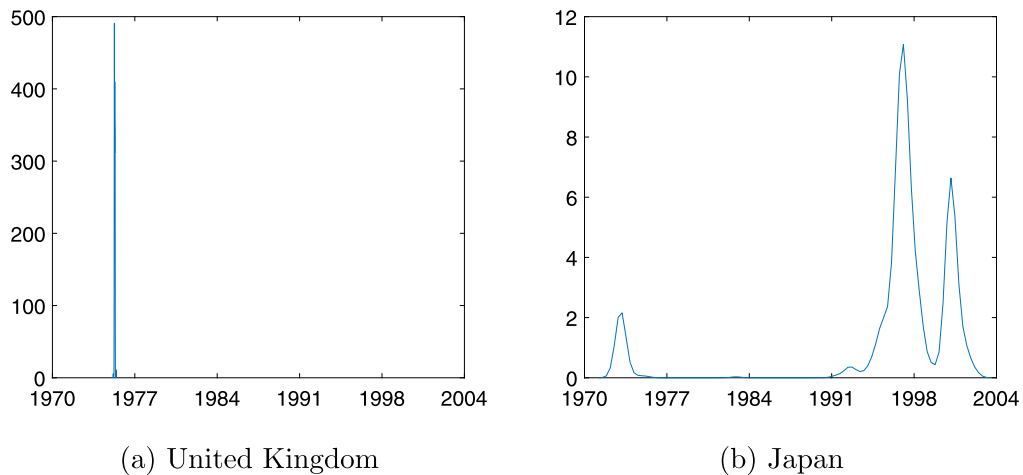


Fig. 3. Posterior density of the break date.

not include) zero, the corresponding credible interval also includes (or does not include) zero. Hence, the conventional approach to inference about the slope parameters for the U.K. sample seems to be robust with respect to the uncertainty of the break date.

In contrast, when the uncertainty of τ is large, the conventional and the Bayesian results on inference about the slope parameters might disagree. Table 5 shows the results for Japan. Although both $\hat{\tau}_{LS}$ and the posterior mode of τ are at 1996:05, the HPD set and the ILR confidence set are much wider than the confidence interval of Bai (1997), indicating a large uncertainty of the break date. The posterior density on τ in Fig. 3 also illustrates that the uncertainty of the break date is much larger for Japan than for the U.K. during the sample period.¹⁶ The large uncertainty of τ is reflected on Bayesian inference about the slope parameters. In the upper panel of Table 5, we see that in general the Bayesian credible intervals are wider than the confidence intervals. Importantly, this can have a qualitative consequence on statistical importance of

¹⁶ In addition, the posterior on τ for Japan exhibits tri-modality, which would be similar to the tendency of a finite-sample distribution of $\hat{\tau}_{LS}$ to have three modes as reported in the literature (e.g., Baek, 2018; Casini and Perron, 2021).

Table 5
Estimation results for the Japanese stock return.

Slopes	Least-squares			Bayesian		
	Estimate	LB	UB	Estimate	LB	UB
$\delta_1^{(1)}$	2.41	0.24	4.57	1.49	−1.89	5.27
$\delta_1^{(2)}$	1.17	0.44	1.89	1.31	0.29	3.26
$\delta_1^{(3)}$	−1.98	−2.45	−1.50	−2.00	−3.53	−1.23
$\delta_1^{(4)}$	−0.85	−1.55	−0.16	−0.29	−1.43	3.59
$\delta_1^{(5)}$	2.76	1.39	4.12	2.29	0.15	4.11
$\delta_2^{(1)}$	−16.8	−29.6	−4.07	−9.39	−23.4	8.77
$\delta_2^{(2)}$	12.4	2.35	22.4	6.22	−5.91	16.8
$\delta_2^{(3)}$	−7.11	−14.1	−0.04	−3.86	−13.1	5.32
$\delta_2^{(4)}$	4.76	1.25	8.27	1.70	−5.86	6.59
$\delta_2^{(5)}$	−8.35	−14.9	−1.78	−4.65	−12.1	4.98
τ	Least-squares			Bayesian		
	Estimate	LB	UB	Estimate	LB	UB
	0.780	0.777	0.782	0.779	0.068	0.934
	(96:05)	(96:04)	(96:06)	(96:05)	(72:03)	(01:10)
	ILR					
	Estimate	LB	UB			
		0.080	0.890			
		(72:08)	(00:02)			

LB = lower bound and UB = upper bound.

some parameters. For seven of the ten slope coefficients, the confidence intervals do not include zero while the Bayesian credible intervals do. Hence, the conventional approach to inference on the slope parameters might not be robust with respect to the uncertainty of the break date, for the Japanese sample.

7. Conclusion and future direction

In this paper, we establish a Bernstein–von Mises type theorem for the slope coefficients in linear regression with a structural break. By doing so, we bridge the gap between the frequentist and the Bayesian approaches for inference on this model. On the one hand, a frequentist researcher can look at Bayesian credible intervals for the slope coefficients as a robustness check to see whether the uncertainty of the break location affects inference on the slope parameters. Such sensitivity analysis is reasonable as our theoretical result guarantees the credible interval to converge to the conventional confidence interval that the frequentist researcher would use otherwise. On the other hand, Bayesian inference can be conveyed to frequentists via our proven result.

Potential extensions include several directions. First, the homoscedasticity assumption could be too strong in some applications, and hence extending the results to the case of heteroscedasticity and autocorrelation would be of interest. Second, a popular Bayesian method of Chib (1998) is different from the approach we took in this paper in that we place an explicit prior on τ and that Chib's framework can be naturally extended to the case of multiple breaks. It would be interesting to study frequentist properties of Chib's approach.

Acknowledgments

The paper is based on the first chapter of my 2021 doctoral dissertation at Brown University. Earlier versions were circulated under the titles “Structural break in linear regression models: Bayesian asymptotic analysis” and “Bayesian inference in linear regression with a structural break”. I am grateful for Andriy Norets, my dissertation advisor, for guidance and encouragement through the work on this project. I thank for valuable comments and suggestions from Eric Renault, Susanne Schennach, Dimitris Korobilis, Jesse Shapiro, Kenneth Chay, Toru Kitagawa, Adam McCloskey, Siddhartha Chib, and Florian Gunsilius. I thank to audiences at the 2020 NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics (SBIES) conference for helpful discussions. I am grateful for the editor, the associate editor, and two anonymous referees for their helpful comments. All remaining errors are mine.

Appendix A. Proofs of theorems and corollaries

Here, we provide proofs of Theorems 1–4 and Corollaries 1–2. Proofs of intermediate results, Propositions 1–4, can be found in the online appendix.

A.1. Proof of Theorem 1

Proof of Theorem 1. Note that

$$\pi_T(\tau) = \frac{L_T(\tau)}{\int L_T(\tau')d\tau'} = \frac{L_T(\tau_0)}{\int L_T(\tau')d\tau'} \frac{L_T(\tau)}{L_T(\tau_0)} = \pi_T(\tau_0) \frac{L_T(\tau)}{L_T(\tau_0)},$$

$$\pi_T(\tau_0) = \frac{L_T(\tau_0)}{\int L_T(\tau')d\tau'} \leq \frac{L_T(\tau_0)}{\int_{B_{M_0/T}^c(\tau_0)} L_T(\tau')d\tau'} = \left[\int_{B_{M_0/T}^c(\tau_0)} \frac{L_T(\tau')}{L_T(\tau_0)} d\tau' \right]^{-1},$$

for any $M_0 > 0$. Hence for each T and for any $M_0 > 0$,

$$\int_{B_{M_0/T}^c(\tau_0)} \pi_T(\tau)d\tau = \pi_T(\tau_0) \int_{B_{M_0/T}^c(\tau_0)} \frac{L_T(\tau)}{L_T(\tau_0)} d\tau \leq \left[\int_{B_{M_0/T}^c(\tau_0)} \frac{L_T(\tau')}{L_T(\tau_0)} d\tau' \right]^{-1} \int_{B_{M_0/T}^c(\tau_0)} \frac{L_T(\tau)}{L_T(\tau_0)} d\tau. \tag{A.1}$$

Therefore, we want to find

1. an upper bound for $\int_{B_{M_0/T}^c(\tau_0)} \frac{L_T(\tau)}{L_T(\tau_0)} d\tau$ and
2. a lower bound for $\int_{B_{M_0/T}^c(\tau_0)} \frac{L_T(\tau')}{L_T(\tau_0)} d\tau'$ for some $M_0 > 0$

We can write the marginal likelihood ratio as

$$\frac{L_T(\tau)}{L_T(\tau_0)} = \exp \left[T \left\{ \frac{1}{T} \log \left(\frac{L_T(\tau)}{L_T(\tau_0)} \right) \right\} \right].$$

The proof of Theorem 1 is built on some intermediate steps, Propositions 1–4. Proposition 1 shows that, under the normal likelihood and the conjugate prior, studying this ratio boils down to comparing the sum of squared residuals $S_T(\tau)$.

Proposition 1. Suppose Assumption 1 holds. Then, with the normal likelihood and the conjugate prior described above, under P_{θ_0, τ_0} , for all τ ,

$$\frac{1}{T} \log \left(\frac{L_T(\tau)}{L_T(\tau_0)} \right) = \frac{1}{2} \log \left(\frac{S_T(\tau_0)}{S_T(\tau)} \right) + O_p(T^{-1}).$$

Let us first examine the limit of the quantity $Q_T(\tau) = T^{-1}S_T(\tau)$. Proposition 2 states that $Q_T(\tau)$ converges in probability to some deterministic function $Q(\tau)$. See Fig. 2 for examples of $Q_T(\tau)$ and $Q(\tau)$.

Proposition 2. Suppose Assumption 1 holds. Then, under P_{θ_0, τ_0} , for all τ ,

$$Q_T(\tau) = Q(\tau) + O_p(T^{-1/2}),$$

where

$$Q(\tau) = \sigma_0^2 + \begin{cases} (\tau_0 - \tau) \frac{(1-\tau_0)}{(1-\tau)} \delta_0' R' \Sigma_X R \delta_0, & \text{if } \tau \leq \tau_0 \\ (\tau - \tau_0) \frac{\tau_0}{\tau} \delta_0' R' \Sigma_X R \delta_0, & \text{if } \tau > \tau_0 \end{cases} \equiv \sigma_0^2 + \Delta(\tau).$$

Define $G_T(\tau) = g(Q_T(\tau))$ and $G(\tau) = g(Q(\tau))$ where $g(x) = -\frac{1}{2} \log(x)$. Due to Proposition 1, we can write

$$T^{-1} \log \left(\frac{L_T(\tau)}{L_T(\tau_0)} \right) = G_T(\tau) - G_T(\tau_0) + O_p(T^{-1}). \tag{A.2}$$

Proposition 3 says that the limit $G(\tau)$ of $G_T(\tau)$ attains its maximum at τ_0 .

Proposition 3. $G(\tau)$ attains its unique maximum at τ_0

Proposition 4 establishes the modulus of continuity of the empirical process $\{G_T(\tau) - G_T(\tau_0)\} - \{G(\tau) - G(\tau_0)\}$ outside of a ball around τ_0 with radius proportional to T^{-1} .

Proposition 4. Suppose Assumption 1 holds. Then, under $\forall \xi > 0, \forall \epsilon > 0, \exists M > 0$ and $k > 0$ such that $T \geq k \implies$

$$P_{\theta_0, \tau_0} \left(\inf_{\tau \in B_{M/T}^c(\tau_0)} \frac{|\{G_T(\tau) - G_T(\tau_0)\} - \{G(\tau) - G(\tau_0)\}|}{|\tau - \tau_0|} < \xi \right) > 1 - \epsilon.$$

By Proposition 3, $G(\cdot)$ attains its unique max at τ_0 . Note that the convex function $G(\tau)$ is not differentiable at τ_0 . Hence we have,

$$\begin{aligned} G(\tau) - G(\tau_0) &< |\tau - \tau_0|B_1, \\ G(\tau) - G(\tau_0) &> |\tau - \tau_0|B_2, \end{aligned}$$

for some $B_1, B_2 < 0$. By Proposition 4, given $\xi_1 > 0, \exists M > 0$: with $P_{\theta_0, \tau_0} \rightarrow 1$,

$$G_T(\tau) - G_T(\tau_0) < \xi_1|\tau - \tau_0| + G(\tau) - G(\tau_0) < |\tau - \tau_0|\{\xi_1 + B_1\}, \tag{A.3}$$

for all $\tau \in B_{M/T}^c(\tau_0)$. Similarly, given $\xi_2 > 0, \exists M_0 > 0$: with $P_{\theta_0, \tau_0} \rightarrow 1$,

$$G_T(\tau) - G_T(\tau_0) > -\xi_2|\tau - \tau_0| + G(\tau) - G(\tau_0) > |\tau - \tau_0|\{-\xi_2 + B_2\}, \tag{A.4}$$

for all $\tau \in B_{M_0/T}^c(\tau_0)$. Recall, by Eq. (A.2), we have

$$\frac{L_T(\tau)}{L_T(\tau_0)} = \exp\left[T\left(G_T(\tau_0) - G_T(\tau)\right) + O_p(1)\right].$$

Hence, from Eq. (A.3), given $\xi_1 > 0$, small compared to $-B_1$, there is $B'_1 < 0$, which is independent of M : we have with $P_{\theta_0, \tau_0} \rightarrow 1$,

$$\frac{L_T(\tau)}{L_T(\tau_0)} \leq \exp\left[T|\tau - \tau_0|B'_1 + O_p(1)\right] = \exp\left[T|\tau - \tau_0|B'_1\right]O_p(1), \tag{A.5}$$

for all $\tau \in B_{M/T}^c(\tau_0)$. Note that the statement above still holds with a larger value of $M > 0$ as the area outside of the ball will be contained by that for the original M . Similarly, from Eq. (A.4), there is $B'_2 < 0$ and $M_0 > 0$: with $P_{\theta_0, \tau_0} \rightarrow 1$,

$$\frac{L_T(\tau)}{L_T(\tau_0)} \geq \exp\left[T|\tau - \tau_0|B'_2 + O_p(1)\right] = \exp\left[T|\tau - \tau_0|B'_2\right]O_p(1), \tag{A.6}$$

for all $\tau \in B_{M_0/T}^c(\tau_0)$. Now, by Inequality (A.5) and the fundamental theorem of calculus,

$$\int_{B_{M/T}^c(\tau_0)} \frac{L_T(\tau)}{L_T(\tau_0)} d\tau \leq \int_{B_{M/T}^c(\tau_0)} \exp\left[T|\tau - \tau_0|B'_1\right] d\tau O_p(1) = \frac{1}{TB'_1} \left(e^{TB'_1} - e^{B'_1M}\right) O_p(1).$$

Similarly, by Inequality (A.6),

$$\int_{B_{M_0/T}^c(\tau_0)} \frac{L_T(\tau)}{L_T(\tau_0)} d\tau \geq \int_{B_{M_0/T}^c(\tau_0)} \exp\left[T|\tau - \tau_0|B'_2\right] d\tau O_p(1) = \frac{1}{TB'_2} \left(e^{TB'_2} - e^{B'_2M_0}\right) O_p(1).$$

This means, together with the bound (A.1),

$$\int_{B_{M/T}^c(\tau_0)} \pi_T(\tau) d\tau \leq \left[\int_{B_{M_0/T}^c(\tau_0)} \frac{L_T(\tau')}{L_T(\tau_0)} d\tau'\right]^{-1} \int_{B_{M/T}^c(\tau_0)} \frac{L_T(\tau)}{L_T(\tau_0)} d\tau \leq \frac{B'_2}{B'_1} \frac{e^{B'_1T} - e^{B'_1M}}{e^{B'_2T} - e^{B'_2M_0}} O_p(1),$$

which can be made arbitrarily small by choosing $M > 0$ and T sufficiently large. \square

A.2. Proof of Corollary 1

Proof of Corollary 1. The main structure of the proof follows Proposition 2 of Bai (1997) and relies on an implication of our Theorem 1. First, note that we have

$$\begin{aligned} \hat{\tau}_{Bayes} &= \operatorname{argmax}_{\tau \in \mathcal{H}} \pi_T(\tau) \\ &= \operatorname{argmax}_{\tau \in \mathcal{H}} L_T(\tau) \\ &= \operatorname{argmax}_{\tau \in \mathcal{H}} \frac{1}{T} \log \left(\frac{L_T(\tau)}{L_T(\tau_0)} \right), \end{aligned}$$

which converges in distribution to $\operatorname{argmax}_{\tau \in \mathcal{H}} \log \left(\frac{S_T(\tau_0)}{S_T(\tau)} \right)$ by Proposition 1. We have

$$\begin{aligned} \operatorname{argmax}_{\tau \in \mathcal{H}} \log \left(\frac{S_T(\tau_0)}{S_T(\tau)} \right) &= \operatorname{argmin}_{\tau \in \mathcal{H}} S_T(\tau) \\ &= \operatorname{argmax}_{\tau \in \mathcal{H}} V_T(\tau) - V_T(\tau_0), \end{aligned}$$

where $V_T(\tau) = \hat{\delta}(\tau)'(Z_2'MZ_2)\hat{\delta}(\tau)$. Bai (1997) shows that $V_T(\tau) - V_T(\tau_0)$ converges in distribution to $W^*(\lfloor T(\tau - \tau_0) \rfloor)$ uniformly on any bounded interval around τ_0 . Let $m^* = \operatorname{argmax}_m W^*(m)$, which is $O_p(1)$. Hence, $\forall \epsilon > 0, \exists R_1 > 0 : P(|m^*| > R_1) < \epsilon$. Our Theorem 1 implies that $\hat{\tau}_{Bayes} = \tau_0 + O_p(T^{-1})$. In other words, $\forall \epsilon > 0, \exists R_2 > 0 : P(T|\hat{\tau}_{Bayes} - \tau_0| > R_2) < \epsilon$. Take $R = \max\{R_1, R_2\}$.

Define $\hat{\tau}_R = \operatorname{argmax}_{T|\tau - \tau_0| \leq R} V_T(\tau) - V_T(\tau_0)$ and $m_R^* = \operatorname{argmax}_{|m| \leq R} W^*(m)$. Then we have $T|\hat{\tau}_R - \tau_0| \xrightarrow{d} m_R^*$. In other words, $|P(\lfloor T(\hat{\tau}_R - \tau_0) \rfloor = j) - P(m_R^* = j)| < \epsilon$ as $T \rightarrow \infty \forall |j| \leq R$.

Note that if $T|\hat{\tau}_{Bayes} - \tau_0| < R$, then $\hat{\tau}_R = \hat{\tau}_{Bayes}$. Similarly, if $|m^*| < R$, then $m_R^* = m^*$. Hence, $|P(\lfloor T(\hat{\tau}_{Bayes} - \tau_0) \rfloor = j) - P(m^* = j)|$ is bounded by $|P(\lfloor T(\hat{\tau}_R - \tau_0) \rfloor = j) - P(m_R^* = j)| + P(T|\hat{\tau}_{Bayes} - \tau_0| \geq R) + P(|m^*| \geq R) < 3\epsilon$. As ϵ can be made arbitrarily small, the desired result holds. \square

A.3. Proof of Theorem 2

Proof of Theorem 2. Define $z = \sqrt{T}(\gamma - \hat{\gamma}_{LS})$ and let $\phi(x; \mu, \Sigma)$ be the multivariate normal density with mean μ and covariance matrix Σ evaluated at x .

$$\begin{aligned} d_{TV}\left(\pi(z|\mathbf{D}_T), N_{(d_x+d_z)}(0, \sigma_0^2 V^{-1})\right) &= \int |\pi(z|\mathbf{D}_T) - \phi(z; 0, \sigma_0^2 V^{-1})| dz \\ &\leq \int \int |\pi(z|\tau, \mathbf{D}_T) - \phi(z; 0, \sigma_0^2 V^{-1})| dz d\pi(\tau|\mathbf{D}_T) \\ &= \int d_{TV}\left(\pi(z|\tau, \mathbf{D}_T), \phi(z; 0, \sigma_0^2 V^{-1})\right) d\pi(\tau|\mathbf{D}_T) \\ &= \int_{B_{M/T}(\tau_0)} d_{TV}\left(\pi(z|\tau, \mathbf{D}_T), \phi(z; 0, \sigma_0^2 V^{-1})\right) d\pi(\tau|\mathbf{D}_T) + o_p(1), \end{aligned}$$

where the last equality is due to Theorem 1.

From (8), asymptotically, the posterior of γ conditional on τ is normal:

$$\begin{aligned} \gamma|\tau, \mathbf{D}_T &\overset{a}{\sim} N_{(d_x+d_z)}(\bar{\mu}_\tau, (\bar{b}_\tau/\bar{a})\bar{H}_\tau^{-1}) \\ \implies z|\tau, \mathbf{D}_T &\overset{a}{\sim} N_{(d_x+d_z)}(\sqrt{T}(\bar{\mu}_\tau - \hat{\gamma}_{LS}), (T\bar{b}_\tau/\bar{a})\bar{H}_\tau^{-1}). \end{aligned}$$

The total variation distance is bounded above by 2 times square root of the KL divergence. In general, the KL divergence between two p -dimensional normal distributions $N_p(\mu_1, \Sigma_1)$ and $N_p(\mu_2, \Sigma_2)$ is bounded above by

$$\underbrace{\frac{|\det(\Sigma_2^{-1}) - \det(\Sigma_1^{-1})|}{\min(\det(\Sigma_1^{-1}), \det(\Sigma_2^{-1}))}}_I + \underbrace{p\|\Sigma_2^{-1} - \Sigma_1^{-1}\|_\infty \|\Sigma_1\|_\infty}_{II} + \underbrace{\|\mu_1 - \mu_2\|_2^2 \|\Sigma_2^{-1}\|_2}_{III}, \tag{A.7}$$

where $\|\Sigma\|_\infty = \max_{ij} |\Sigma_{ij}|$ is the largest element of Σ in the absolute value, and $\|\Sigma\|_2 = \sup_\mu \|\Sigma\mu\|_2 / \|\mu\|_2$ is a matrix norm induced by the standard norm on \mathbb{R}^p , $\|\mu\|_2 = \sum_{i=1}^p \mu_i^2$. We can bound the total variation distance between the posterior density of z conditional on τ and that of $N_{(d_x+d_z)}(0, \sigma_0^2 V^{-1})$ using the bound (A.7), with $\mu_1 = \sqrt{T}(\bar{\mu}_\tau - \hat{\gamma}_{LS})$, $\Sigma_1 = (T\bar{b}_\tau/\bar{a})\bar{H}_\tau^{-1}$, $\mu_2 = 0$, and $\Sigma_2 = \sigma_0^2 V^{-1}$.

To show $III = o_p(1)$, we write

$$\sqrt{T}(\bar{\mu}_\tau - \hat{\gamma}_{LS}) = \sqrt{T}(\bar{\mu}_\tau - \hat{\gamma}(\tau)) + \sqrt{T}(\hat{\gamma}(\tau) - \hat{\gamma}_{LS}). \tag{A.8}$$

By definition,

$$\bar{\mu}_\tau = \left[\frac{1}{T}H + \frac{1}{T}\chi'_\tau \chi_\tau \right]^{-1} \left[\frac{1}{T}H\bar{\mu} + \frac{1}{T}\chi'_\tau Y \right] = \hat{\gamma}(\tau) + O_p(T^{-1}),$$

so the first term in (A.8) is $o_p(1)$. To show that the second term in (A.8) is $o_p(1)$ for $\tau \in B_{M/T}(\tau_0)$, write $\sqrt{T}(\hat{\gamma}(\tau) - \hat{\gamma}_{LS}) = \sqrt{T}(\hat{\gamma}(\tau) - \gamma_0) - \sqrt{T}(\hat{\gamma}_{LS} - \gamma_0)$. Note that $Y = X\beta_0 + Z_{2\tau_0}\delta_0 + \epsilon = X\beta_0 + Z_{2\tau}\delta_0 + \epsilon_\tau^*$, where $\epsilon_\tau^* = (Z_{2\tau_0} - Z_{2\tau})\delta_0 + \epsilon$. This implies

$$\begin{aligned} \sqrt{T}(\hat{\gamma}(\tau) - \gamma_0) &= \left[\frac{1}{T} \begin{pmatrix} X'X & X'Z_{2\tau} \\ Z'_{2\tau}X & Z'_{2\tau}Z_{2\tau} \end{pmatrix} \right]^{-1} \frac{1}{\sqrt{T}} \begin{pmatrix} X'\epsilon_\tau^* \\ Z'_{2\tau}\epsilon_\tau^* \end{pmatrix} \\ &= \left[\frac{1}{T} \begin{pmatrix} X'X & X'Z_{2\tau} \\ Z'_{2\tau}X & Z'_{2\tau}Z_{2\tau} \end{pmatrix} \right]^{-1} \frac{1}{\sqrt{T}} \begin{pmatrix} X'\epsilon + X'(Z_{2\tau_0} - Z_{2\tau})\delta_0 \\ Z'_{2\tau}\epsilon + Z'_{2\tau_0}(Z_{2\tau_0} - Z_{2\tau})\delta_0 \end{pmatrix}. \end{aligned}$$

For $|\tau - \tau_0| < \frac{M}{T}$, we have

$$\begin{aligned} \frac{1}{T}X'Z_{2\tau} - \frac{1}{T}X'Z_{2\tau_0} &= o_p(1), & \frac{1}{T}Z'_{2\tau}Z_{2\tau} - \frac{1}{T}Z'_{2\tau_0}Z_{2\tau_0} &= o_p(1), \\ \frac{1}{\sqrt{T}}X'(Z_{2\tau_0} - Z_{2\tau}) &= o_p(1), & \frac{1}{\sqrt{T}}Z'_{2\tau_0}(Z_{2\tau_0} - Z_{2\tau}) &= o_p(1), \\ \frac{1}{\sqrt{T}}Z'_{2\tau}\epsilon - \frac{1}{\sqrt{T}}Z'_{2\tau_0}\epsilon &= o_p(1), \end{aligned}$$

which implies

$$\sqrt{T}(\hat{\gamma}(\tau) - \gamma_0) = \left[\frac{1}{T} \begin{pmatrix} X'X & X'Z_{2\tau_0} \\ Z'_{2\tau_0}X & Z'_{2\tau_0}Z_{2\tau_0} \end{pmatrix} \right]^{-1} \frac{1}{\sqrt{T}} \begin{pmatrix} X'\epsilon \\ Z'_{2\tau_0}\epsilon \end{pmatrix} + o_p(1).$$

Similarly, since the least-square estimator $\hat{\tau}_{LS} \in B_{M/T}(\tau_0)$ for sufficiently large T , we can show

$$\begin{aligned} \sqrt{T}(\hat{\gamma}_{LS} - \gamma_0) &= \sqrt{T}(\hat{\gamma}(\hat{\tau}_{LS}) - \gamma_0) \\ &= \left[\frac{1}{T} \begin{pmatrix} X'X & X'Z_{2\hat{\tau}_{LS}} \\ Z'_{2\hat{\tau}_{LS}}X & Z'_{2\hat{\tau}_{LS}}Z_{2\hat{\tau}_{LS}} \end{pmatrix} \right]^{-1} \frac{1}{\sqrt{T}} \begin{pmatrix} X'\epsilon + X'(Z_{2\tau_0} - Z_{2\hat{\tau}_{LS}})\delta_0 \\ Z'_{2\hat{\tau}_{LS}}\epsilon + Z'_{2\tau_0}(Z_{2\tau_0} - Z_{2\hat{\tau}_{LS}})\delta_0 \end{pmatrix} \\ &= \left[\frac{1}{T} \begin{pmatrix} X'X & X'Z_{2\tau_0} \\ Z'_{2\tau_0}X & Z'_{2\tau_0}Z_{2\tau_0} \end{pmatrix} \right]^{-1} \frac{1}{\sqrt{T}} \begin{pmatrix} X'\epsilon \\ Z'_{2\tau_0}\epsilon \end{pmatrix} + o_p(1). \end{aligned}$$

Hence, $\sqrt{T}(\hat{\gamma}(\tau) - \hat{\gamma}_{LS}) = o_p(1)$.

To show *I* and *II* are $o_p(1)$, note that $\Sigma_1 - \Sigma_2$ equals

$$(T\bar{b}_\tau/\bar{a})\bar{H}_\tau^{-1} - \sigma_0^2V^{-1} = \left[(T\bar{b}_\tau/\bar{a})\bar{H}_\tau^{-1} - \frac{TS_T(\tau)}{T - (d_x + d_z)}(\chi_\tau\chi_\tau)^{-1} \right] + \left[\frac{TS_T(\tau)}{T - (d_x + d_z)}(\chi_\tau\chi_\tau)^{-1} - \sigma_0^2V^{-1} \right]. \tag{A.9}$$

For the first term in (A.9), we have

$$(T\bar{b}_\tau/\bar{a})\bar{H}_\tau^{-1} = \frac{\bar{b}_\tau}{\underline{a} + T/2} \left[\frac{1}{T}H + \frac{1}{T}\chi'_\tau\chi_\tau \right]^{-1} = \frac{\frac{1}{T}\bar{b}_\tau}{\underline{a}/T + 1/2} \left[\frac{1}{T}H + \frac{1}{T}\chi'_\tau\chi_\tau \right]^{-1}.$$

Note that $(1/T)\bar{b}_\tau = \frac{1}{2T}S_T(\tau) + O_p(T^{-1})$, so we have $(T\bar{b}_\tau/\bar{a})\bar{H}_\tau^{-1} = S_T(\tau)[\chi'_\tau\chi_\tau]^{-1} + O_p(T^{-1})$. Therefore, the term in the first square brackets in (A.9) is $o_p(1)$. For the second term in (A.9), we have that for $|\tau - \tau_0| < \frac{M}{T}$,

$$\begin{aligned} \frac{TS_T(\tau)}{T - (d_x + d_z)}(\chi_\tau\chi_\tau)^{-1} - \sigma_0^2V^{-1} &= \underbrace{(Q_T(\tau) - Q_T(\tau_0))\hat{V}_T^{-1}(\tau)}_{=o_p(1)} + \underbrace{Q_T(\tau_0)(\hat{V}_T^{-1}(\tau) - \hat{V}_T^{-1}(\tau_0))}_{=o_p(1)} \\ &\quad + \underbrace{(Q_T(\tau_0) - \sigma_0^2)}_{=o_p(T^{-1/2})}\hat{V}_T^{-1}(\tau_0) + \underbrace{\sigma_0^2(\hat{V}_T^{-1}(\tau_0) - V^{-1})}_{=o_p(1)} + o_p(1) = o_p(1), \end{aligned}$$

where $\hat{V}_T(\tau) = \frac{1}{T}\chi'_\tau\chi_\tau$.

This implies that $\Sigma_2^{-1} - \Sigma_1^{-1} = o_p(1)$. Hence *II* = $o_p(1)$. By continuity of determinants, we also have that *I* = $o_p(1)$ for $\tau \in B_{M/T}(\tau_0)$.

Finally, for $\tau \in B_{M/T}(\tau_0)$

$$d_{TV}(\pi(z|\tau, \mathbf{D}_T), N_{(d_x+d_z)}(0, \sigma_0^2V)) \leq 2\sqrt{o_p(1)} = o_p(1).$$

Therefore $d_{TV}(\pi(z|\mathbf{D}_T), N_{(d_x+d_z)}(0, \sigma_0^2V^{-1}))$ is bounded above by

$$\int_{B_{M/T}(\tau_0)} d_{TV}(\pi(z|\tau, \mathbf{D}_T), \phi(z; 0, \sigma_0^2V^{-1}))d\pi(\tau|\mathbf{D}_T) + o_p(1) = o_p(1). \quad \square$$

A.4. Proof of Theorem 3

Proof of Theorem 3.

Recall that the proof of Theorem 1 is an implication of Propositions 1–4. Assumption 1 implies Propositions 2–4. Proposition 1 establishes that under the normal likelihood and the conjugate prior, Assumption 1 implies

$$\frac{1}{T} \log \left(\frac{L_T(\tau)}{L_T(\tau_0)} \right) = \frac{1}{2} \log \left(\frac{S_T(\tau_0)}{S_T(\tau)} \right) + O_p(T^{-1}). \tag{A.10}$$

Therefore, **Theorem 3** can be proved if we establish the above equality under the normal likelihood together with **Assumptions 1–2**.

As we do not have a closed-form expression (up to a normalization constant) for $L_T(\tau) = \int p(\mathbf{Y}_T|\mathbf{X}_T, \theta, \tau)\pi(\theta, \tau)d\theta$ under the non-conjugate priors defined in **Assumption 2(ii)**, we utilize a Laplace approximation type result to investigate this integral. For a given τ , denote by $F_T(\theta, \tau) = \log p(\mathbf{Y}_T|\mathbf{X}_T, \theta, \tau)$ the log likelihood function conditional on τ . Under the normal likelihood and **Assumption 1**, **Assumptions 2 and 3** of **Hong and Preston (2012)** are satisfied. Now with our **Assumption 2**, we can invoke **Theorem 3** of **Hong and Preston (2012)** (see their page 361) which establishes that

$$\log \int e^{F_T(\theta, \tau) - F_T(\hat{\theta}(\tau), \tau)} \pi(\theta, \tau) d\theta = \log [\pi(\theta^*(\tau), \tau)(2\pi)^{(d_x+d_z+1)/2} \det(-TA_\theta(\tau))^{-1/2}] + o_p(1),$$

for each τ , where $-A_\theta(\tau)$ is the probability limit of $-\frac{1}{T} \frac{\partial^2}{\partial \theta \partial \theta'} F_T(\hat{\theta}(\tau), \tau)$ and is positive definite.

Note that

$$\frac{L_T(\tau)}{L_T(\tau_0)} = \frac{e^{F_T(\hat{\theta}(\tau), \tau)} \int e^{F_T(\theta, \tau) - F_T(\hat{\theta}(\tau), \tau)} \pi(\theta, \tau) d\theta}{e^{F_T(\hat{\theta}(\tau_0), \tau_0)} \int e^{F_T(\theta, \tau_0) - F_T(\hat{\theta}(\tau_0), \tau_0)} \pi(\theta, \tau_0) d\theta},$$

which implies that

$$\begin{aligned} \frac{1}{T} \log \left(\frac{L_T(\tau)}{L_T(\tau_0)} \right) &= \frac{1}{T} F_T(\hat{\theta}(\tau), \tau) - \frac{1}{T} F_T(\hat{\theta}(\tau_0), \tau_0) \\ &\quad + \frac{1}{T} \log [\pi(\theta^*(\tau), \tau)(2\pi)^{(d_x+d_z+1)/2} \det(-TA_\theta(\tau))^{-1/2}] \\ &\quad - \frac{1}{T} \log [\pi(\theta^*(\tau_0), \tau_0)(2\pi)^{(d_x+d_z+1)/2} \det(-TA_\theta(\tau_0))^{-1/2}] + O_p(T^{-1}) \\ &= \frac{1}{T} \log \left[\frac{p(\mathbf{Y}_T|\mathbf{X}_T, \hat{\theta}(\tau), \tau)}{p(\mathbf{Y}_T|\mathbf{X}_T, \hat{\theta}(\tau_0), \tau_0)} \right] + \frac{1}{T} \log \left[\frac{\pi(\theta^*(\tau), \tau)}{\pi(\theta^*(\tau_0), \tau_0)} \right] \\ &\quad - \frac{1}{2T} \log \left[\frac{\det(-A_\theta(\tau))}{\det(-A_\theta(\tau_0))} \right] + O_p(T^{-1}). \end{aligned}$$

Note that we assumed that $\pi(\theta^*(\tau), \tau)$ and $\pi(\theta^*(\tau_0), \tau_0)$ are finite and non-zero. Hence, the term involving the ratio of the priors is $O_p(T^{-1})$. Also, $-A_\theta(\tau)$ is a positive definite matrix hence its determinant is a finite positive number. We have

$$p(\mathbf{Y}_T|\mathbf{X}_T, \hat{\theta}(\tau), \tau) \propto \left(\frac{1}{\hat{\sigma}^2(\tau)} \right)^{T/2} \exp \left[-\frac{1}{2\hat{\sigma}^2(\tau)} \underbrace{\sum_{t=1}^T (y_t - \chi_{\tau,t} \hat{\gamma}(\tau))^2}_{=S_T(\tau)} \right] = \left(\frac{1}{\hat{\sigma}^2(\tau)} \right)^{T/2} \exp(-T/2),$$

where the last equality is due to the fact that $\hat{\sigma}^2(\tau) = S_T(\tau)/T$. This implies the desired result i.e., **(A.10)**. Note that **Propositions 2–4** hold under **Assumption 1**. Therefore, given **(A.10)**, the rest of the proof of **Theorem 3** follows the same argument in the proof of **Theorem 1** in **A.1**. \square

A.5. Proof of Corollary 2

Proof of Corollary 2. By definition, we have

$$\hat{\tau}_{Bayes} = \operatorname{argmax}_{\tau \in \mathcal{H}} \pi_T(\tau) = \operatorname{argmax}_{\tau \in \mathcal{H}} \frac{1}{T} \log \left(\frac{L_T(\tau)}{L_T(\tau_0)} \right),$$

In the proof of **Theorem 3**, we have shown that Eq. **(A.10)** holds under the normal likelihood and **Assumptions 1–2**. Therefore, $\hat{\tau}_{Bayes}$ converges in distribution to $\operatorname{argmax}_{\tau \in \mathcal{H}} \log \left(\frac{S_T(\tau_0)}{S_T(\tau)} \right) = \operatorname{argmax}_{\tau \in \mathcal{H}} V_T(\tau) - V_T(\tau_0)$.

Furthermore, **Theorem 3** implies that $\hat{\tau}_{Bayes} = \tau_0 + O_p(T^{-1})$. Based on these two facts, the rest of the proof follows the same argument as in the proof of **Corollary 1** in **A.2**. \square

A.6. Proof of Theorem 4

Proof of Theorem 4.

Under the normal likelihood and **Assumptions 1–2**, the proof of **Theorem 3** of **Hong and Preston (2012)** (see their page 367) implies that the posterior of $\sqrt{T}(\gamma - \hat{\gamma}(\tau))$ conditional on τ converges in total variation in probability to the multivariate normal distribution $N(0, -A_\gamma^{-1}(\tau))$, where $A_\gamma^{-1}(\tau)$ is the sub-matrix of $A_\theta^{-1}(\tau)$ obtained by deleting the last row and the last column, $-A_\theta(\tau)$ is the probability limit of $-\frac{1}{T} \frac{\partial^2}{\partial \theta \partial \theta'} F_T(\hat{\theta}(\tau), \tau)$, and $F_T(\theta, \tau) = \log p(\mathbf{Y}_T|\mathbf{X}_T, \theta, \tau)$ is the

log likelihood function conditional on τ . This means that the total variation between the posterior of $z = \sqrt{T}(\gamma - \hat{\gamma}_{LS})$ given τ and $N(\sqrt{T}(\hat{\gamma}(\tau) - \hat{\gamma}_{LS}), -A_{\gamma}^{-1}(\tau))$ converges to 0 in probability.

The bound (A.7) on the KL divergence between two normal densities can be used again now with $\mu_2 = 0$, $\Sigma_2 = \sigma_0^2 V^{-1}$, $\mu_1 = \sqrt{T}(\hat{\gamma}(\tau) - \hat{\gamma}_{LS})$, and $\Sigma_1 = -A_{\gamma}^{-1}(\tau) = \text{plim } \hat{\sigma}^2(\tau) \hat{V}_T^{-1}(\tau)$ where $\hat{V}_T(\tau) = \frac{1}{T} \chi_{\tau}' \chi_{\tau}$. Note that from the proof of Theorem 2 in A.3, we know that $\mu_1 = o_p(1)$ for $|\tau - \tau_0| < \frac{M}{T}$.

For $|\tau - \tau_0| < \frac{M}{T}$,

$$\begin{aligned} \Sigma_1 - \Sigma_2 &= \underbrace{-A_{\gamma}^{-1}(\tau) - Q_T(\tau) \hat{V}_T^{-1}(\tau)}_{=o_p(1)} \\ &+ \underbrace{(Q_T(\tau) - Q_T(\tau_0)) \hat{V}_T^{-1}(\tau)}_{=o_p(1)} \\ &+ Q_T(\tau_0) \underbrace{(\hat{V}_T^{-1}(\tau) - \hat{V}_T^{-1}(\tau_0))}_{=o_p(1)} \\ &+ \underbrace{(Q_T(\tau_0) - \sigma_0^2) \hat{V}_T^{-1}(\tau_0)}_{=o_p(T^{-1/2})} \\ &+ \sigma_0^2 \underbrace{(\hat{V}_T^{-1}(\tau_0) - V^{-1})}_{=o_p(1)} + o_p(1) = o_p(1), \end{aligned}$$

which implies

$$\Sigma_2^{-1} - \Sigma_1^{-1} = o_p(1).$$

The rest of the proof can be done similarly as in the proof of Theorem 2 in A.3 by applying the bound (A.7). \square

Appendix B. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jeconom.2022.03.006>.

References

- Baek, Y., 2018. Estimation of a structural break point in linear regression models. arXiv preprint [arXiv:1811.03720](https://arxiv.org/abs/1811.03720).
- Bai, J., 1997. Estimation of a change point in multiple regression models. *Rev. Econ. Stat.* 79 (4), 551–563.
- Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 47–78.
- Bai, J., Perron, P., 2003. Computation and analysis of multiple structural change models. *J. Appl. Econometrics* 18 (1), 1–22.
- Casini, A., Perron, P., 2020. Generalized Laplace inference in multiple change-points models. *Econom. Theory* 1–31.
- Casini, A., Perron, P., 2021. Continuous record Laplace-based inference about the break date in structural change models. *J. Econometrics* 224 (1), 3–21.
- Chib, S., 1998. Estimation and comparison of multiple change-point models. *J. Econometrics* 86 (2), 221–241.
- Eo, Y., Morley, J., 2015. Likelihood-ratio-based confidence sets for the timing of structural breaks. *Quant. Econ.* 6 (2), 463–497.
- Ghosal, S., Samanta, T., 1995. Asymptotic behaviour of Bayes estimates and posterior distributions in multiparameter nonregular cases. *Math. Methods Statist.* 4 (4), 361–388.
- Hansen, B.E., 2000. Sample splitting and threshold estimation. *Econometrica* 68 (3), 575–603.
- Hong, H., Preston, B., 2012. Bayesian averaging, prediction and nonnested model selection. *J. Econometrics* 167 (2), 358–369.
- Kleijn, B., Knapik, B., 2012. Semiparametric posterior limits under local asymptotic exponentiality. arXiv preprint [arXiv:1210.6204](https://arxiv.org/abs/1210.6204).
- Paye, B.S., Timmermann, A., 2006. Instability of return prediction models. *J. Empir. Financ.* 13 (3), 274–315.
- Perron, P., 2006. Dealing with structural breaks. In: *Palgrave Handbook of Econometrics*, vol. 1, (2), pp. 278–352.
- Qu, Z., Perron, P., 2007. Estimating and testing structural changes in multivariate regressions. *Econometrica* 75 (2), 459–502.