Relevant parameter changes in structural break models

Arnaud Dufays¹ and Jeroen V.K. Rombouts²

March 13, 2018

Abstract

Structural break time series models, which are commonly used in macroeconomics and finance, capture unknown structural changes by allowing for abrupt changes to model parameters. However, most specifications suffer from an over-parametrization issue, since all parameters have to change when a break occurs. We introduce a sparse change-point model to detect which parameters change over time. We propose a shrinkage prior distribution, which controls model parsimony by limiting the number of parameters that change from one structural break to another. We develop a Bayesian sampler for inference on the sparse change-point model. An extensive simulation study based on AR, ARMA and GARCH processes highlights the excellent performance of the sampler. We then propose three empirical applications. First, we revisit the US three-month Treasury bill modelling proposed in Pesaran, Pettenuzzo, and Timmermann (2006) to emphasize that many of the detected breaks are due to a change in only a subset of the model parameters. A second application consists of studying the 22 most-used macroeconomic time series for breaks in parameters. As a final exercise, we summarize GARCH parameter breaks detected in 384 financial returns for companies in the S&P 500 index.

Keywords: Shrinkage prior, change-point model, linear model, non-linear model, relevant parameter change, Bayesian inference.

JEL Classification: C11, C15, C22, C58.

¹ Université Laval, Département d’économique and CRREP research associate. Website: https://sites.google.com/site/websiteofarnauddufays/.
Arnaud Dufays gratefully acknowledges financial support from the Fonds de Recherche du Québec - Société et Culture.
² ESSEC Business School, 3 Avenue Bernard Hirsch, Cergy Pontoise, France 95021. E-mail: rombouts@essec.edu; CREST
1 Introduction

It is well-known that lengthy time series can be subject to structural breaks. Evidence from macroeconomic series can be found in Stock and Watson (1996) and in Bauwens, Koop, Korobilis, and Rombouts (2015), while Pastor and Stambaugh (2001) provides evidence for financial series. Since break dates are unknown, models that allow for the possibility of changing their structure or parameters have been developed over the last two decades. Nevertheless, these models do not identify the precise parameter that changes, and they generally don’t perform much better than simple models.

This paper proposes a new approach to model time series that are subject to multiple unknown structural breaks. We extend the literature in two directions. First, we relax the assumption that all parameters must change when automatically detecting structural breaks. This is done by shrinking irrelevant break parameters towards zero in our estimation procedure, which is based on a sampling scheme tailored to structural break or change-point (CP) models. In addition, our shrinkage methodology eliminates the need to search for an optimal number of regimes and only relies on one estimation output. We propose a new shrinkage prior based on a finite mixture of uniform distributions. In contrast to existing shrinkage priors, which are typically used for high-dimensional regression problems, our new prior has desirable properties for structural break models.

Our second main contribution is that we allow for rich and parsimonious within-regime dynamics. Our framework includes flexible processes such as autoregressive (AR), autoregressive moving average (ARMA) and generalized autoregressive conditional heteroskedastic (GARCH) models. The current literature mostly uses an AR structure for the conditional mean of each regime. ARMA-type (resp. GARCH-type) dynamics in the conditional mean (resp. variance) introduce more complexity in the estimation and model selection procedure because of the path dependence problem. Therefore, our approach also allows for the estimation of break dates and relevant parameter changes in time-varying variance and correlation models, which are frequently used in empirical economics and finance.

Hamilton (1989) is a groundbreaking paper that initiated much of the research on regime models. He considers a linear AR model with parameters that can switch according to a
Markov-switching discrete process. The latent states are recurrent, as the process can move from one state to any other state at any date. This approach for regime switches in a stationary time series has been adapted to the structural break or change-point specification. As in this paper, the latter specification has non-recurrent states so that the process can only stay in the same state or move to the next one. This implies a restricted probability transition matrix and is in this sense a particular case of the classical switching model.

Chib (1998) introduces inference on structural break or change-point time series models. Apart from parameter estimation, he also tackles the computation of the marginal likelihood, which is required to estimate the number of breaks. More recently, new specifications and inference for univariate time series have been proposed by Pesaran, Pettenuzzo, and Timmermann (2006); Koop and Potter (2007); Giordani and Kohn (2008); Maheu and Gordon (2008); Geweke and Jiang (2011) and Maheu and Song (2014). These models adopt a hierarchical prior for the regime coefficients, which allows for the coefficients of one regime to be informative about coefficients of other regimes.

While the recent proposed papers are quite general in their model specification, they have the feature that a break triggers abrupt changes in all parameters indexed by the latent state variable. This is also true for studies that first detect the break dates by testing (see Andrews (1993) and Bai and Perron (1998)) or that implement a penalty criterion (such as a least absolute shrinkage and selection operator-type (LASSO-type) penalty as proposed by Tibshirani, Saunders, Rosset, Zhu, and Knight (2005) and Chan, Yip, and Zhang (2014) or the minimum distance length, see Davis, Lee, and Rodriguez-Yam (2006a), Yau and Zhao (2016) or Ng, Pan, and Yau (2017)). Then, given the detected break dates, independent locally stationary models (typically from the same model class) are fitted. When subsets of parameters are invariant between regimes, such a time-series segmentation approach loses efficiency.

Using our new shrinkage prior, we allow for changes in subsets of model parameters that are virtually zero. We state clear rules with respect to the choice of the hyperparameters of the prior distribution. Shrinkage priors, proposed by Mitchell and Beauchamp (1988) and George and McCulloch (1993) in discrete form, are popular in high-dimensional regressions.
to select relevant explanatory variables (e.g., Inoue and Kilian (2008) and Stock and Watson (2012) for time-series applications). Continuous shrinkage priors, such as the spike-and-slab (S&S) priors of Ishwaran and Rao (2005) or the normal-gamma priors of Griffin and Brown (2010), consist of continuous mixtures of parametric distributions that exhibit a high density around zero in order to shrink non-relevant parameters to this value. Although applicable beyond linear regressions (see Scheipl, Fahrmeir, and Kneib (2012), for example), we argue that these standard shrinkage priors are less suitable for structural break models.

The existing Bayesian literature mostly considers models with an AR structure in the conditional mean of each regime. Inferring ARMA or GARCH processes is a major obstacle because of the path dependence problem. This occurs because the conditional mean (or variance) at time \( t \) depends on the entire sequence of regimes visited up to time \( t \). When computing the likelihood function, one needs to integrate over all possible regime paths that grow exponentially with \( t \); see Billio, Monfort, and Robert (1999) for an example of a two-regime ARMA model. Bauwens, Dufays, and Rombouts (2013) address inference and the computation of the marginal likelihood in the context of switching GARCH models using the particle Markov-chain Monte Carlo (MCMC) sampler of Andrieu, Doucet, and Holenstein (2010).

In this paper, we use a Metropolis-type MCMC sampler proposed in Bauwens, Dufays, and De Backer (2011) to sample the break dates. Acknowledging that shrinkage priors lead to a highly multi-modal posterior distribution, we include our MCMC algorithm in a sequential Monte Carlo sampler (see Del Moral, Doucet, and Jasra (2006)). A simulation study based on AR, ARMA and GARCH processes illustrates the excellent performance of the sampler.

We provide three applications. First, we compare the sparse AR model of the monthly three-month US Treasury bill rate with the change-point AR model proposed in Pesaran, Pettenuzzo, and Timmermann (2006). While the break dates detected by the two approaches are very similar, the sparse model shows that most of the structural breaks are caused by changes in the time series’ variance rather than its mean. Second, we estimate the sparse model on 22 macroeconomic time series used in Bauwens, Koop, Korobilis, and
Rombouts (2015). We show that most series experience change points, but only few parameters vary when the breaks occur. Third, we fit a sparse change-point GARCH model on stock returns for companies in the S&P 500 index. The results show that leaving the persistence parameters of the GARCH model constant over time is a restrictive assumption for most of the studied series.

The rest of the paper is organized as follows. Section 2 defines the general sparse change-point framework. Section 3 introduces the desirable properties of shrinkage priors dedicated to CP settings. We present our new prior designed for CP time-series models, discuss its theoretical properties and give details on how hyperparameters can be chosen. Section 4 provides some analytical in-sample results on the size of the breaks that can be detected with our shrinkage approach. Section 5 sketches the algorithms used for estimation and discusses prior parameter elicitation. An extensive simulation study is documented in Section 6. Three empirical applications are presented in Section 7. Section 8 concludes.

2 Change-point time series models

Let \( y_{1:T} = \{y_1, \ldots, y_T\} \) be a univariate real valued time series of \( T \) observations that is modeled as

\[
y_t = f(\bar{\alpha}, z_{1t}) + g(\bar{\beta}, z_{2t})\varepsilon_t,
\]

where \( \varepsilon_t \sim i.i.d. N(0, 1) \), in which the functions \( f(\cdot) \) and \( g(\cdot) \) are deterministic given their arguments, \( \{\bar{\alpha}, \bar{\beta}\} \) are vectors of model parameters and \( X_t = \{z_{1t}, z_{2t}\} \) denotes explanatory variables observed at time \( t \). To cover two popular cases, we assume that \( \bar{\alpha} \in \mathbb{R}^{K_1} \) while \( \bar{\beta} \in \mathbb{R}^{K_2}_+ \) where \( K_1 \) and \( K_2 \) denote the dimensionality of the parameter vectors. In many economic applications, time series span a long period and the assumption about constant parameters is likely violated. To accommodate this, the model can be adapted to allow for \( K \) breaks in the parameters as follows:

\[
y_t = f(\bar{\alpha}_k, z_{1k}) + g(\bar{\beta}_k, z_{2t})\varepsilon_t, \quad \text{for} \ t \in [\tau_{k-1}, \tau_k],
\]
where \( \tau_0 = 1 < \tau_1 < \ldots < \tau_{K+1} = T \) denote the break times when the model parameters switch from one value to another (i.e. \( k = 1, \ldots, K + 1 \)).

**Remark 1.** As special cases, the setting includes standard AR processes, ARMA models and GARCH processes. See the empirical illustrations below for concrete examples.

**Remark 2.** While we illustrate our approach with a Gaussian distribution, non-Gaussian white noise innovations \( \epsilon_t \) can be introduced.

Given that all the model parameters are indexed by \( k \), the CP model defined above assumes that all parameters change when a break occurs. This “change all” assumption is used in most of the literature (e.g. Fryzlewicz and Subba Rao (2014) and Yau and Zhao (2016), who provide theory for consistent estimation of the number and locations of the change-points). These procedures allow time series to be segmented in locally stationary processes that are at most weakly dependent. The last estimated segmented can be used for forecasting.

Although investigating which parameters in \( \{\bar{\alpha}, \bar{\beta}\} \) are affected by a break seems intuitive, this is rarely done in practice. Our main contribution is to provide a procedure that can automatically handle the detection of relevant parameter breaks in one estimation. The key idea to achieve sparsity is the use of a new shrinkage prior that is tailored to CP models. In order to apply the new shrinkage prior that is tailored to CP models. In order to apply the new shrinkage prior, we write the CP model relative to the first (starting) regime as follows:

\[
y_t = f(\alpha_1 + \sum_{i=2}^{k} \alpha_i, z_{1t}) + g(\beta_1 \odot_{i=2}^{k} \beta_i, z_{2t})\epsilon_t \quad \text{for} \quad t \in [\tau_{k-1}, \tau_k]
\]  

(1)

where \( \alpha_i = \bar{\alpha}_i - \bar{\alpha}_{i-1} \) and \( \alpha_1 = \bar{\alpha}_1 \). Therefore, the parameters in levels can simply be obtained by \( \bar{\alpha}_k = (\bar{\alpha}_1 + \sum_{i=2}^{k} \alpha_i) \) for \( k \in [2, K + 1] \). The parameters impacting the variance are directly obtained by \( \bar{\beta}_k = \beta_1 \odot_{i=2}^{k} \beta_i \) for \( k \in [2, K + 1] \), where the operator \( \odot \) denotes the element-wise product or the Hadamard product. The multiplicative transformation of the \( \beta \) parameters makes it easier to satisfy their positivity constraint.

To shorten the notation when discussing inference, we define the parameter sets as \( \alpha = (\alpha_{1:K+1}) = (\alpha_1, \alpha_2, \ldots, \alpha_{K+1}), \beta = (\beta_{1:K+1}) = (\beta_1, \beta_2, \ldots, \beta_{K+1}), \Theta = (\alpha, \beta) \) and \( \tau_{1:K} = \)
\((\tau_1, \ldots, \tau_K)\) as the set of unknown break dates.

3 Shrinkage priors for CP models

Assuming a set of break dates \(J_1 = \{\tilde{\tau}_1, \ldots, \tilde{\tau}_{K_{\text{max}}}\}\) such that \(K_{\text{max}} \geq K\) and the true break dates are \(\tau_{1:K} \in J_1\), the joint density from Equation (1) can be rewritten as

\[
y_{1:T}|J_1 \sim N_T \left( f(\alpha, Z), \text{diag}(g(\beta_1, z_{21})^2, \ldots, g(\beta_1 \odot_{i=2}^{K_{\text{max}}} \beta_i, z_{2T})^2) \right),
\]

where \(f(\alpha, Z) = [f(\alpha_1, z_{11}), \ldots, f(\alpha_1 + \sum_{i=2}^{K_{\text{max}}} \alpha_i 1_{\{t > \tilde{\tau}_{i-1}\}}, z_{1t}), \ldots, f(\alpha_1 + \sum_{i=2}^{K_{\text{max}}} \alpha_i, z_{1T})]'\).

In this case, determining which parameters change after a break is a high-dimensional shrinkage problem. In particular, one must choose an appropriate parameter penalty function \(l(\alpha, \beta)\) and solve the optimization problem:

\[
\{\alpha, \beta\} = \arg\max_{\{\alpha, \beta\}} \ln f_N(y_{1:T}, f(\alpha, Z), \Sigma) + l(\alpha, \beta),
\]

where \(f_N(x, \mu, \Sigma)\) denotes the multivariate normal density function with expectation \(\mu\) and covariance matrix \(\Sigma\). Selecting a continuous concave penalty function, such as a LASSO-type function (see Tibshirani (1994)), simplifies the numerical burden of optimizing the penalized likelihood function of Equation (3). Another intuitive type of penalty stands for an information criterion such as the minimum description (distance) length (see Davis, Lee, and Rodriguez-Yam (2006b)), or the Bayesian information criterion (BIC). In the latter case, the function \(l(\alpha, \beta)\) reduces to \(-\sum_{k=1}^{K_{\text{max}}} \frac{1}{2} \ln T [\sum_{i=1}^{K_1} 1_{\{\alpha_{ki} \neq 0\}} + \sum_{j=1}^{K_2} 1_{\{\beta_{kj} \neq 0\}}]\). This type of penalty achieves two out of the three desirable shrinkage properties: unbiasedness and sparsity (Fan and Li (2001)).

Interestingly, Theorem 3.1 in Ng, Pan, and Yau (2017) provides a way to build the set of break dates, \(J_1\), that asymptotically cover the true break dates. From a frequentist perspective, we can therefore first compute the set \(J_1\) using the method described in Ng, Pan, and Yau (2017) and, in a second step, maximize the function (3) given a consistent penalty that achieves the oracle property. However, this procedure is only asymptotically valid and
does not provide any confidence interval for the break dates. In addition, optimization is challenging because of the multi-modal shape of the penalized likelihood function. The latter issue is exacerbated if the BIC penalization is used due to the lack of continuity of the penalty function. In such a case, the optimization procedure is bound to fail using standard optimization techniques. As the Bayesian paradigm offers numerical simulation methods that can handle multi-modal posterior distributions, we carry out the estimation in the Bayesian framework.

To check which parameters actually change after a break, it is often burdensome to consider all possibilities for many regimes and to find the best specification among them following a criterion such as the marginal likelihood (e.g. Eo (2012)) or the penalized likelihood function of Equation (3). This is because the possibilities grow exponentially with the number of regimes. For instance, in our US Treasury bill application, this approach would have required $4^{10}$ model estimations if we consider an upper bound of 10 regimes and a standard ARMA(1,1) process. Our method of using shrinkage priors achieves the same goal in one estimation.

Shrinkage priors in sparse CP models should have three properties. First, the prior should be able to distinguish between a clear change to a parameter and a small variation. Second, when there is a break, the prior should allow to detect it, regardless of whether it is a small or large break. Third, the hyperparameters of the prior distributions should depend on the user’s preferences for including breaks and should be derived from a simple rule. The first two properties correspond to the sparsity and unbiasedness properties in Fan and Li (2001), translated to a CP context. As emphasized above, this suggests an $L_0$ type of penalty, such as the one implied by the BIC. For the third property, we work with a penalty in terms of log-likelihood. For example, a penalty of -3 implies that any new regime has to improve the log-likelihood function of at least 3 to be detected. As the penalty decreases, the model will detect fewer breaks and become more parsimonious. To implement this simple rule, we choose a threshold value that defines a significant change in the parameter value. A deviation above this threshold is evidence of a new regime. The hyperparameters of the prior distributions are derived such that the difference of the
prior log-density evaluated at the threshold minus the prior log-density evaluated at zero is exactly equal to the user-defined penalty.

**Remark 3.** The penalty parameter acts as the tuning parameter (i.e. "\( \lambda \)") in standard shrinkage regressions. While the latter is usually selected by cross-validation, which implies many model estimations, our penalty parameter has a direct interpretation and can therefore be chosen using an economic theory or a statistical procedure, as we do below.

### 3.1 The 2MU prior

We introduce the 2MU shrinkage prior tailored to CP models. The 2MU distribution is given by a finite mixture of two components

\[
\omega U\left[\frac{-a}{2}, \frac{a}{2}\right] + (1 - \omega)U\left[\frac{-b}{2}, \frac{b}{2}\right]
\]

where \( U\left[\frac{-a}{2}, \frac{a}{2}\right] \) denotes uniform distribution with bounds that depend on \( a \). We assume that \( a \) is small and \( b \) is large (see Section 3.2 for details).

**Remark 4.** For parameters with positive support, the distribution is transformed as \( \omega U\left[1 - \frac{a}{2}, 1 + \frac{a}{2}\right] + (1 - \omega)U[0, b] \).

We define \( P \) as the penalty on the log-likelihood function for detecting a new regime, and \( x \) is any point in the wider uniform component. Then, \( \log f(x) - \log f(0) = P \) yields

\[
P = \log \frac{(1 - \omega)/b}{\omega/a + (1 - \omega)/b}.
\]

Two observations are worth noting. First, if we set \( \omega = 1 \), \( P \to -\infty \), which means that a break is too heavily penalized to occur. Second, setting \( \omega = 0 \) gives the maximum value of the penalty (i.e. \( P^* = \log 1 = 0 \)). In the latter case, the 2MU prior reduces to a uniform distribution with support \([-\frac{b}{2}, \frac{b}{2}]\), which amounts to the standard CP model with diffuse priors.
Re-arranging (5) to obtain an expression for $\omega$ given a chosen penalty $P$ gives

$$\omega = \frac{a(1-e^P)}{be^P + a(1-e^P)},$$

which belongs to $[0,1]$ for any positive values of $a$ and $b$, and any negative value of $P$. In the following, we denote the shrinkage prior by $2MU(a, b, P)$. Figure 1 displays examples of the density function for different values of $P$, $a$ and $b$.

(a) $P = -1, a = 0.1, b = 10$

(b) $P = -3, a = 0.1, b = 10$

(c) $P = -3, a = 1, b = 10$

(d) $P = -5, a = 1, b = 10$

Figure 1 – Mixture of two uniform distributions with different values of $P$, $a$ and $b$.

In terms of the model specification in (1), the prior distribution of the mean parameters
simplifies as follows:

\[
f(\alpha_2, \ldots, \alpha_{K_{\max}+1}) = \prod_{k=2}^{K_{\max}+1} \prod_{i=1}^{K_1} f(\alpha_{ki}),
\]

\[
f(\alpha_{ki}) = \frac{\omega}{a} I_{\{|\alpha_{ki}| < \frac{a}{2}\}} + \frac{1 - \omega}{b} I_{\{|\alpha_{ki}| < \frac{b}{2}\}},
\]

\[
= [b e^P + a(1 - e^P)]^{-1} [(1 - e^P) I_{\{|\alpha_{ki}| < \frac{a}{2}\}} + e^P I_{\{|\alpha_{ki}| < \frac{b}{2}\}}],
\]

\[
= c[I_{\{|\alpha_{ki}| < \frac{a}{2}\}} + e^P I_{\{\frac{b}{2} < |\alpha_{ki}| < \frac{b}{2}\}}],
\]

where \(c = [b e^P + a(1 - e^P)]^{-1}\). Similarly, the prior distribution for the variance parameters can be simplified as

\[
f(\beta_2, \ldots, \beta_{K_{\max}}) = \prod_{k=2}^{K_{\max}+1} \prod_{j=1}^{K_2} f(\beta_{kj}),
\]

\[
f(\beta_{kj}) = \frac{\omega}{a} I_{\{|\beta_{kj}-1| < \frac{a}{2}\}} + \frac{1 - \omega}{b} I_{\{|\beta_{kj} < \frac{b}{2}\}},
\]

\[
= [b e^P + a(1 - e^P)]^{-1} [(1 - e^P) I_{\{|\beta_{kj}-1| < \frac{a}{2}\}} + e^P I_{\{|\beta_{kj} < \frac{b}{2}\}}],
\]

\[
= c[I_{\{|\beta_{kj}-1| < \frac{a}{2}\}} + e^P I_{\{0 < \beta_{kj} < 1 - \frac{a}{2} \text{ or } 1 + \frac{a}{2} < \beta_{kj} < \frac{b}{2}\}}],
\]

Applying Bayes’ rule, the logarithm of the joint posterior density is proportional to

\[
f(\Theta, \tau | y_{1:T}) \propto \ln f_N(y_{1:T}, f(\alpha, Z), \Sigma) + \ln f(\alpha_1, \beta_1, \tau_1; K_{\max}) + \ln f(\alpha_2; K_{\max}+1) + \ln f(\beta_2; K_{\max}+1)
\]

\[
\times \ln f_N(y_{1:T}, f(\alpha, Z), \Sigma) + \ln f(\alpha_1, \beta_1, \tau_1; K_{\max}) + \ldots
\]

\[
\sum_{k=1}^{K_{\max}+1} \sum_{i=1}^{K_1} P I_{\{\frac{a}{2} < |\alpha_{ki}| < \frac{a}{2}\}} + \sum_{j=1}^{K_2} P I_{\{|\beta_{kj} < B\}}.
\]

The maximum a posteriori (MAP) will therefore coincide with the solution to the optimization problem for Equation (3) when

\[
l(\alpha, \beta) = \ln f(\alpha_1, \beta_1, \tau_1; K_{\max}) + \sum_{k=1}^{K_{\max}+1} \sum_{i=1}^{K_1} P I_{\{\frac{a}{2} < |\alpha_{ki}| < \frac{a}{2}\}} + \sum_{j=1}^{K_2} P I_{\{|\beta_{kj} < B\}}.
\]

A model parameter \(|\alpha_{ki}| \) (resp. \(\beta_{kj}\)) will be greater than \(\frac{a}{2}\) (resp. \(\beta_{kj} \notin B\)) if and only
if the likelihood function $\ln f_N(y_{1:T}, \mathbf{f}(\alpha, \mathbf{Z}), \Sigma)$ is decreased by $P$. We recognize an $L_0$ type of penalty. As a consequence, the 2MU prior implies both sparsity and unbiasedness properties.

The previous calculation provides a way to set the penalty parameter $P$. A standard choice would be a penalty that detects a break only if the Bayes factor of the model with a break (and no shrinkage prior) is higher than a model without the break. As this quantity is not computable, one can rely on the BIC to set $P$.

**Remark 5.** Equations (7) and (8) reveal that the 2MU prior leads to a highly multi-modal posterior distribution. When we add parameters to the model, the shape of the posterior distribution exhibits more modes, making its exploration by simulation more difficult and even more so by deterministic optimization methods. This computational issue is well documented within the shrinkage prior literature (see e.g. Garcia-Donato and Martinez-Beneito (2013)), and is also present when using popular shrinkage priors, including the continuous S&S and normal-gamma priors. Since the posterior distribution is highly multi-modal, inference is based on a sophisticated Monte Carlo method rather than standard Markov-chain Monte Carlo methods.

### 3.2 Choosing the threshold, penalty and maximum number of breaks

We have developed a prior distribution for shrinking non-relevant parameters toward zero in a CP context. In practice, we need to select the maximum number of breaks $K_{\max}$, the bound $a$ of the narrow uniform component, and the penalty $P$. We propose ways to choose these parameters.

#### 3.2.1 Maximum number of breaks

To set the maximum number of breaks, $K_{\max}$, our approach is to first estimate a CP model exhibiting a high number of regimes and shrink the parameters of non-relevant regimes to zero. This saves computational resources and coding efforts, since only one estimation is required and computation of the marginal likelihood is avoided. In fact, the standard
procedure is to maximize the marginal likelihood by estimating CP models with different numbers of regimes.

In order to choose $K_{\text{max}}$, we rely on Theorem 3.1 in Ng, Pan, and Yau (2017) and construct a set of potential break dates $J = \{\tilde{\tau}_1, \ldots, \tilde{\tau}_{K_{\text{max}}}\}$ as follows:

1. Choose a minimum regime size $h \geq d (\log T)^3$, where $d$ is a positive constant.

2. For each $t \in [h+1, T-h]$, compute the following likelihood ratio scan statistic, $S_h(t)$:

$$S_h(t) = \frac{1}{h} [S_{K=1,h}(t) - S_{K=0,h}(t)], \quad \text{with}$$

$$S_{K=1,h}(t) = \ln f_N(y_{t-h+1:t}, f(\hat{\alpha}_1, Z), \hat{\Sigma}_1) + \ln f_N(y_{t+1:t+h}, f(\hat{\alpha}_2, Z), \hat{\Sigma}_2), \quad \text{(9)}$$

$$S_{K=0,h}(t) = \ln f_N(y_{t-h+1:t+h}, f(\hat{\alpha}, Z), \hat{\Sigma}), \quad \text{(11)}$$

where the parameters $\{\hat{\alpha}_1, \hat{\Sigma}_1\}$, $\{\hat{\alpha}_2, \hat{\Sigma}_2\}$ and $\{\hat{\alpha}, \hat{\Sigma}\}$ are the maximum likelihood estimates of the model with no breaks over the sample windows $y_{t-h+1:t}$, $y_{t+1:t+h}$ and $y_{t-h+1:t+h}$, respectively.

3. Define the set of change points, $J$, as follows:

$$J = \left\{m \in \{h, h+1, \ldots, T-h\} : S_h(m) = \max_{t \in (m-h, m+h)} S_h(t) \right\}, \quad \text{(12)}$$

where $S_h(t) \equiv 0$ for $t < h$ and $t > T-h$.

Under assumptions detailed in Ng, Pan, and Yau (2017), any true change point $\tau_i$, for $i \in [1, K]$, is in an $h$–neighborhood of a break in the set $J$. In particular, we have that

$$P(\max_{i=1,\ldots,K} \min_{k=1,\ldots,K_{\text{max}}} |\tau_i - \tilde{\tau}_k| < h) \to 1,$$

in which $K_{\text{max}} = |J|$. As a rule of thumb, Ng, Pan, and Yau (2017) suggest using $h = \max(100, \log(T)^4/25)$, since it gives good empirical results for AR, ARMA and GARCH processes. Note that this rule of thumb implies a minimum regime duration of 100 observations. As we are able to detect finite-sample breaks in the Bayesian framework on the basis that model assumptions are correct, we extend the number of break dates in two directions.
Firstly, we add one break point at both time periods \( t \in [1, h - 1] \) and \( t \in [T - h + 1, T] \).
Secondly, we define the set of potential break points, \( J_1 \), as

\[
\begin{align*}
J_1 &= \left\{ m \in \{h, h + 1, \ldots, T - h\} : S_h(m) = \max_{t \in (m - h/2, m + h/2)} S_h(t) \right\}. 
\end{align*}
\] (14)

The maximum number of breaks is set to \( K_{\text{max}} = 2 + |J_1| \). Since \( J \subset J_1 \), therefore \( K_{\text{max}} > \tilde{K}_{\text{max}} \).

### 3.2.2 Bound of the narrow uniform component

The bound \( a/2 \) of the narrow uniform component is the amount a parameter must deviate from its previous value in order to be considered as a new regime. Above this value, the parameter is no longer reduced toward zero and a new regime is allowed. Since the scale of the model parameters as well as their standard deviations are data dependent, we use one value of \( a \) per parameter. For instance, in the case of an ARMA(1,1) process, four bounds are needed. Choosing an appropriate value requires knowledge of the possible variation of the model parameters within a regime. To this end, we first perform inference of the \( J_1 + 1 \) single-regime model on the data segmented by the break dates found in \( J_1 \). Then we use the parameter draws to define the \( a/2 \) bound as the minimum over the \( J_1 + 1 \) inference of the difference between the median and the 2.5\textsuperscript{th} percentile divided by two. This takes into account the scale of the model parameters while ensuring the bound (\( a \)) is small enough to only capture large breaks.

### 3.2.3 Penalty

To account for the choice uncertainty of the penalty parameter, we consider \( P \) as a random variable and use an informative prior. As a natural prior choice, we consider a uniform distribution, since this approach is standard in the S&S literature in which a uniform prior is directly set on the weight \( \omega \). Proposition 1 illustrates the shape of the penalty posterior distribution assuming a uniform prior on \( \rho = e^P \) that implies \( \omega \in [0, 1] \) for \( \rho \in [0, 1] \).

**Proposition 1.** If \( \rho = e^P \sim U[u_1, u_2] \) and \( 0 \leq u_1 \leq u_2 \leq 1 \), the posterior distribution
function of the parameter $\rho$ given a parameter $\alpha|\rho \sim 2MU(a,b,\ln \rho)$ is given by

- if no break is detected (i.e. $|\alpha| \leq \frac{a}{2}$):

$$P(\rho \leq x|\alpha) = \left[ \ln \frac{a + u_2(b - a)}{a + u_1(b - a)} \right]^{-1} \ln \frac{a + x(b - a)}{a + u_1(b - a)}, \quad \text{or} \quad (15)$$

- if a break is detected (i.e. $|\alpha| > \frac{a}{2}$):

$$P(\rho \leq x|\alpha) = \frac{(b - a)(x - u_1) + a \ln \frac{a + u_1(b - a)}{a + x(b - a)}}{(b - a)(u_2 - u_1) + a \ln \frac{a + u_1(b - a)}{a + u_2(b - a)}}. \quad (16)$$

**Remark 6.** Proposition 1 remains unchanged when the parameters lie in a positive support.

In our empirical applications, we use an informative prior centered at the BIC, $\rho \sim U[\frac{1}{2}\rho_{\text{BIC}}, \frac{3}{2}\rho_{\text{BIC}}]$ where $\rho_{\text{BIC}} = e^{P_{\text{BIC}}}$. By doing so, the uncertainty on the parameter $P$ is reflected in the posterior distribution of the model parameters, whereas its fluctuation is under control. Sampling $\rho$ from its conditional distribution is done by applying the inverse probability integral transform on Equations (15) and (16).

**Remark 7.** Since $\rho$ is a random variable, our shrinkage prior exhibits the same appealing property as the continuous S&S prior of Ishwaran and Rao (2005). In fact, when $\rho$ increases, the weight of the narrow uniform component $\omega$ declines. Consequently, it facilitates the move from the spike to the slab component and vice versa, and therefore improves the mixing of the sampler.

Finally, it turns out that one cannot choose the same parameter $\rho$ for all the parameters of the model since its posterior distribution would be shifted towards zero. This feature is not desirable, as we would lose the interpretation of the penalty parameter. More precisely, if we assume that all parameters $\alpha_{2:K_{\text{max}}+1}$ follow the same $2MU(a,b,\ln \rho)$ with $\rho \sim U[0,1]$, then the posterior distribution of the parameter $\rho$ when there are no breaks in all parameters (i.e. $|\alpha_{2:K_{\text{max}}+1}| \leq \frac{a}{2}$) is given by

$$P(\rho \leq x|\alpha_{2:K_{\text{max}}+1}) = \frac{(a + (b - a)x)^{1-N} - a^{1-N}}{b^{1-N} - a^{1-N}}. \quad (17)$$
where $N = K_1(K_{\text{max}} + 1)$.

Equation (17) tells us that most of the probability mass is assigned to small values of $\rho$ (i.e. $\rho \leq \frac{1-a}{b-a}$) and the issue is exacerbated when $N$ increases. In fact, the distribution simplifies to $\frac{1-a^{1-N}}{b^{1-N}-a^{1-N}}$ when $x = \frac{1-a}{b-a}$. For instance, if $a = 0.1$ and $b = 10$, we have $x = 0.091$ (giving a penalty of $-2.4$), but $P(\rho \leq x|\alpha_{2,K_{\text{max}}+1})$ is 0.5, 0.99 and 0.9999 for $N = 1$, $N = 3$ and $N = 5$ respectively. Consequently, as long as more parameters do not experience a break, the penalty $P$ that is imposed to detect a break decreases. By including one penalty variable per model parameter, we avoid these issues as the marginal posterior distribution of each penalty is less sensitive to breaks in the parameters. This is standard practice in the literature on shrinkage priors, including the normal-gamma and the S&S priors.

4 Size of the breaks that can be detected

Asymptotically, any size of parameter breaks can be detected under the set of assumptions presented in Ng, Pan, and Yau (2017). We provide results on the minimum size of breaks that can be detected in a finite sample setting. We consider the following process:

$$y_t = \bar{\mu} + \sum_{j=1}^{K} \mu_j 1_{t>\tau_j} + \sigma_0 \varepsilon_t,$$  (18)

in which $\varepsilon_t \sim i.i.d.N(0,1)$, $\sigma_0^2 \sim U[0, \bar{u}]$, $\mu_j \sim 2\text{MU}(a,b,P)$ and $\tau_j \sim U[\bar{\tau}_{j-1}+1, \bar{\tau}_j]$ for $j = 1, ..., K$, where $\bar{\tau}_j \in [2, T]$ represents a fixed hyperparameter with $\tau_0 = 0$ and $\tau_K = T$. Let us denote $\sigma^2_{t_1:t_2}$ as the empirical variance over the sample $[t_1, t_2]$. We assume that $\max(\sigma^2_{\tau_{j-1}:\tau_j}) < \bar{u}$ for $j = 1, ..., K$.

**Proposition 2.** When $K = 1$, a break is detected in model (18) if

$$\sigma^2_{1:T} > e^{-2P}[\lambda \sigma^2_{1:\tau_1} + (1-\lambda)\sigma^2_{\tau_1+1:T}],$$  (19)

15
where $\lambda = \frac{\eta}{\tau}$. When the BIC penalty is used, the rule simplifies to

$$
\tilde{\sigma}_{1:T}^2 > \left( \frac{0.95T}{0.05} \right)^2 \left[ \lambda \tilde{\sigma}_{1:T}^2 + (1 - \lambda) \tilde{\sigma}_{\tau_1 + 1:T}^2 \right].
$$

(20)

When $K = 2$, a break is detected in model (18) if the following two inequalities hold:

$$
\frac{\tilde{\sigma}_{1:T}^2}{\lambda_1 \tilde{\sigma}_{1:T}^2 + (\lambda_2 - \lambda_1) \tilde{\sigma}_{\tau_1 + 1:T}^2 + (1 - \lambda_2) \tilde{\sigma}_{\tau_2 + 1:T}^2} > e^{-\frac{2}{\lambda_1}}, \text{ and }
$$

$$
\frac{\tilde{\sigma}_{1:T}^2}{\lambda_1 \tilde{\sigma}_{1:T}^2 + (\lambda_2 - \lambda_1) \tilde{\sigma}_{\tau_1 + 1:T}^2 + (1 - \lambda_2) \tilde{\sigma}_{\tau_2 + 1:T}^2} > e^{-\frac{2}{\lambda_2}} \text{ for } \tilde{\tau} \in [2, T - 1],
$$

where $\tau_1 = \lambda_1 T$, $\tau_2 = \lambda_2 T$ and $\tilde{\tau} = \tilde{\lambda} T$.

**Proof. Case of $K=1$**

The posterior density is proportional to

$$
\ln f(\tau_1, \Theta | y_{1:T}) \propto \ln f(y_{1:T} | \tau_1, \Theta) + \ln f(\tau_1, \mu) + P1_{\{\Delta \sigma^2 \in B_1\}}
$$

$$
= -\frac{T}{2} \ln(2\pi \sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{t=1}^{T} (y_t - \mu - \Delta \mu_1 1_{\{t > \tau_1\}})^2 + \ln f(\tau_1, \mu) + P1_{\{\Delta \mu > \frac{\sigma_0^2}{2}\}}.
$$

When the break is detected, the MAP is given by $\tilde{\mu} = \frac{1}{\tau_1} \sum_{t=1}^{\tau_1} y_t = \tilde{\mu}_{1:1} \text{ (i.e. the empirical mean of the first regime)}$ and $\Delta \mu_1 = \tilde{\mu}_{1:T} - \tilde{\mu}_{1:1}$, while the variance is $\tilde{\sigma}_0^2 = \lambda \sum_{t=1}^{\tau_1} (y_t - \tilde{\mu}_{1:1})^2/\tau_1 + (1 - \lambda) \sum_{t=\tau_1+1}^{T} (y_t - \tilde{\mu}_{1:T} + 1/T - \tau_1) = \lambda \tilde{\sigma}_{1:1}^2 + (1 - \lambda) \tilde{\sigma}_{\tau_1 + 1:T}^2$. The posterior density simplifies to

$$
\ln f_1(\tau_1, \Theta | y_{1:T}) = -\frac{T}{2} \ln(\lambda \tilde{\sigma}_{1:1}^2 + (1 - \lambda) \tilde{\sigma}_{\tau_1 + 1:T}^2) - \frac{T}{2} + \ln f(\tau_1, \mu) + P. \quad (21)
$$

When no break is detected, the MAP is given by $\tilde{\mu} = \tilde{\mu}_{1:T}$, $\Delta \mu \approx 0$ and $\tilde{\sigma}_0^2 = \tilde{\sigma}_{1:T}^2$. The posterior density leads to

$$
\ln f_0(\tau_1, \Theta | y_{1:T}) = -\frac{T}{2} \ln(\tilde{\sigma}_{1:T}^2) - \frac{T}{2} + \ln f(\tau_1, \mu). \quad (22)
$$

A break is detected only if Equation (21) is greater than Equation (22). Consequently, the
size of a break that can be detected is

\[
\ln f_1(\tau_1, \sigma | y_{1:T}) - \ln f_0(\tau_1, \sigma | y_{1:T}) > 0,
\]

\[-\frac{T}{2} \ln(\lambda \bar{\sigma}^2_{1:\tau_1} + (1 - \lambda) \bar{\sigma}^2_{\tau_1+1:T}) + P + \frac{T}{2} \ln(\bar{\sigma}^2_{1:T}) > 0, \]

\[
\ln \frac{\bar{\sigma}^2_{1:T}}{\lambda \bar{\sigma}^2_{1:\tau_1} + (1 - \lambda) \bar{\sigma}^2_{\tau_1+1:T}} > -\frac{2}{T} P.
\]

Using \( P_{\text{BIC}} = -\ln \frac{0.95T}{0.05} \), we have

\[
\bar{\sigma}^2_{1:T} > \left( \frac{0.95T}{0.05} \right)^2 [\lambda \bar{\sigma}^2_{1:\tau_1} + (1 - \lambda) \bar{\sigma}^2_{\tau_1+1:T}].
\]

**Case of K=2**

Let us define the set \( \tau = \{\tau_1, \tau_2\} \). When two breaks are detected, the MAP is given by

\[
\tilde{\mu} = \frac{1}{\tau_1} \sum_{t=1}^{\tau_1} y_t = \bar{\mu}_{1:\tau_1} \text{ and } \tilde{\Delta} \mu_1 = \bar{\mu}_{\tau_1+1:\tau_2} - \bar{\mu}_{1:\tau_1} \text{ and } \tilde{\Delta} \mu_2 = \bar{\mu}_{\tau_2+1:T} - \bar{\mu}_{1:\tau_1} - \bar{\mu}_{\tau_1+1:\tau_2},
\]

while the variance is \( \bar{\sigma}^2_0 = \lambda_1 \bar{\sigma}^2_{1:\tau_1} + (\lambda_2 - \lambda_1) \bar{\sigma}^2_{\tau_1+1:\tau_2} + (1 - \lambda_2) \bar{\sigma}^2_{\tau_2+1:T} \). The estimate of the posterior density at the MAP simplifies to

\[
\ln f_K(\tau, \Theta | y_{1:T}) = -\frac{T}{2} \ln(\lambda \bar{\sigma}^2_{1:\tau_1} + (\lambda_2 - \lambda_1) \bar{\sigma}^2_{\tau_1+1:\tau_2} + (1 - \lambda_2) \bar{\sigma}^2_{\tau_2+1:T}) - \frac{T}{2} + \ln f(\tau, \mu) + 2P.
\]

When one break is detected, the change-point \( \tilde{\tau} \) could lie anywhere in the sample. The posterior density for this case is

\[
1^{\text{st}} \text{ break: } \ln f_K(\tau, \Theta | y_{1:T}) = -\frac{T}{2} \ln(\tilde{\lambda} \bar{\sigma}^2_{1:\tilde{\tau}} + (1 - \tilde{\lambda}) \bar{\sigma}^2_{\tilde{\tau}+1:T}) - \frac{T}{2} + \ln f(\tau, \mu) + P.
\]

When no break is identified, we have a posterior density maximized at

\[
\ln f_K(\tau, \Theta | y_{1:T}) = -\frac{T}{2} \ln(\bar{\sigma}^2_{1:T}) - \frac{T}{2} + \ln f(\tau, \mu).
\]
The rules for detecting the two breaks are

\[
\ln f_{K=2}(\tau, \sigma | y_{1:T}) - \ln f_{K=0}(\tau, \sigma | y_{1:T}) > 0,
\]
\[
\frac{\sigma^2_{1:T}}{\lambda_1 \sigma^2_{1,\tau_1} + (\lambda_2 - \lambda_1) \sigma^2_{\tau_1+1,\tau_2} + (1 - \lambda_2) \sigma^2_{\tau_2+1:T}} > e^{-\frac{P}{T}}.
\]

\[
\ln f_{K=2}(\tau, \sigma | y_{1:T}) - \ln f_{K=1}(\tau, \sigma | y_{1:T}) > 0,
\]
\[
\frac{\tilde{\lambda} \sigma^2_{1,\tilde{\tau}} + (1 - \tilde{\lambda}) \sigma^2_{\tilde{\tau}+1:T}}{\lambda_1 \sigma^2_{1,\tau_1} + (\lambda_2 - \lambda_1) \sigma^2_{\tau_1+1,\tau_2} + (1 - \lambda_2) \sigma^2_{\tau_2+1:T}} > e^{-\frac{P}{T}} \text{ for } \tilde{\tau} \in [2, T - 1].
\]

Proposition 2 can be useful for selecting the penalty parameter. From inequality (19), assuming that the break location grows with the sample size (i.e. \( \lambda \) is fixed with respect to \( T \)), we can show that \( \bar{\sigma}_{1,\tau_1} \rightarrow_p \sigma_0^2, \bar{\sigma}_{\tau_1+1:T} \rightarrow_p \sigma_0^2 \) and \( \tilde{\sigma}_{1,\tilde{\tau}} \rightarrow_p \mu_1^2 \lambda(1 - \lambda) + \sigma_0^2 \) (see Appendix A.4 for a proof). Applying these limits, the inequality becomes

\[
|\mu_1| > \sigma_0 \sqrt{\frac{(e^{-\frac{2P}{T}} - 1)}{\lambda(1 - \lambda)}} \geq 2\sigma_0 \sqrt{(e^{-\frac{2P}{T}} - 1)}.
\]  

(23)

As expected, as \( T \rightarrow \infty \), the term \( e^{-\frac{2P}{T}} - 1 \) approaches zero for any penalty parameter that increases slower than \( T \). In particular, the penalty parameter based on the BIC value allows us to detect any break size asymptotically. Fixing \( \sigma_0 \) to one, Figure 2b documents how the term \( 2\sigma_0 \sqrt{e^{-\frac{2P}{T}} - 1} \) evolves in finite sample with \( T \) for different penalty values. Given a consistent estimator of \( \sigma_0 \), one can investigate the minimum break size in the unconditional mean that can be detected and can choose the penalty parameter appropriately. Moreover, one can also simulate the probability of detecting a specific break in the unconditional mean given several break locations, as is done in Figure 2a.

We can also derive the size of the break for another simple process that is given by

\[
y_t = \mu + \epsilon_t, \text{ where } \epsilon_t \sim N(0, \sigma_0^2 \prod_{j=1}^{K} \Delta \sigma_j^2 1_{\{t>\tau_j\}}),
\]  

(24)

in which \( \epsilon_t \sim N(0, 1), \mu \sim U[-\bar{u}_\mu, \bar{u}_\mu], \sigma_0^2 \sim U[0, \bar{u}], \Delta \sigma_j^2 \sim 2MU^+(a,b,P) \) and \( \tau_j \sim \)
(a) Probability of detecting $\mu_1$ when $\sigma_0 = 1, T = 500$ and $\bar{\mu} = 0$: the size of the break in relation to break location $\lambda$.

(b) Minimum value of $|\mu|$ that can be detected given the sample size when $\sigma_0 = 1$ for different penalty parameters. The solid line represents the BIC value, the dotted line represents $P = -5$ and the dashed line represents $P = -15$.

$U[\bar{\tau}_j - 1, \bar{\tau}_j]$ for $j = 1, ..., K$, where $\bar{\tau}_j \in [2, T]$ is a fixed hyperparameter with $\tau_0 = 0$ and $\tau_K = T$. We assume that $|\bar{u}| > \bar{\mu}_{1: T}$ and $\max(\tilde{\sigma}^2_{\tau_j-1: \tau_j}) < \bar{u}$ for $j = 1, ..., K$, where $\tilde{\sigma}^2_{t_1:t_2}$ is the empirical variance of the variable $\{y_t - \bar{\mu}_{1: T}\}$ over the sample $[t_1, t_2]$.

**Proposition 3.** When $K = 1$, a break is detected in model (24) if

$$
\ln[e^{2\mu} (\lambda \tilde{\sigma}^2_{1: \tau_1} + (1 - \lambda) \tilde{\sigma}^2_{\tau_1+1: T})] > \lambda \ln \tilde{\sigma}^2_{1: \tau_1} + (1 - \lambda) \ln \tilde{\sigma}^2_{\tau_1+1: T},
$$

in which $\lambda = \frac{\tau_1}{T}$. When the BIC penalty is used, the rule simplifies to

$$
\ln(\frac{0.05}{0.95T})^2 [\lambda \tilde{\sigma}^2_{1: \tau_1} + (1 - \lambda) \tilde{\sigma}^2_{\tau_1+1: T}] > \lambda \ln \tilde{\sigma}^2_{1: \tau_1} + (1 - \lambda) \ln \tilde{\sigma}^2_{\tau_1+1: T}.
$$

**Proof.** The posterior density leads to

$$
\ln f(\tau_1, \mu, \sigma | y_{1:T}) \propto \ln f(y_{1:T} | \tau_1, \sigma) + \ln f(\tau_1, \mu) + P1_{\{\Delta \sigma^2_1 \in B_1\}},
$$

$$
= \frac{T}{2} \ln(2\pi \sigma^2_0) - \frac{T - \tau_1}{2} \ln \Delta \sigma^2_1 - \frac{1}{2\sigma^2_0} \sum_{t=1}^{\tau_1} \epsilon_t^2 - \frac{1}{2\sigma^2_0 \Delta \sigma^2_1} \sum_{t=\tau_1+1}^{T} \epsilon_t^2 + \ln f(\tau_1) + P1_{\{\Delta \sigma^2_1 \in B_1\}}
$$

When the break is detected, the MAP is given by $\mu = \bar{\mu}_{1: T}$, $\tilde{\sigma}^2_0 = \frac{\sum_{t=1}^{\tau_1} (y_t - \bar{\mu}_{1: T})^2}{\tau_1} = \tilde{\sigma}^2_{1: \tau_1}$ and
\( \Delta \sigma^2_1 = \frac{\sum_{t=\tau_1+1}^{T} e_t^2}{\frac{T}{T-\tau_1}} = \frac{\tilde{\sigma}^2_{\tau_1:T}}{\sigma^2_{\tau_1}}. \) In this case, the posterior density reduces to

\[
\ln f_1(\tau_1, \mu, \sigma|y_{1:T}) = -\frac{T}{2} \ln(2\pi) - \frac{\tau_1}{2} \ln \sigma^2_{\tau_1} - \frac{T-\tau_1}{2} \ln \tilde{\sigma}^2_{\tau_1:T} - \frac{T}{2} + \ln f(\tau_1, \mu) \tag{27}
\]

If the break is not detected, the MAP is given by \( \tilde{\sigma}^2_0 = \frac{\sum_{t=1}^{T} e_t^2}{T} = \lambda \tilde{\sigma}^2_{\tau_1} + (1-\lambda) \tilde{\sigma}^2_{\tau_1:T}, \) where \( \lambda = \tau_1/T \) and \( \Delta \sigma^2_1 \approx 1 \) since \( a \) is small. In this case, the posterior density is

\[
\ln f_0(\tau_1, \mu, \sigma|y_{1:T}) = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\lambda \tilde{\sigma}^2_{\tau_1} + (1-\lambda) \tilde{\sigma}^2_{\tau_1:T}) - \frac{T}{2} + \ln f(\tau_1, \mu) \tag{28}
\]

A break is detected only if Equation (27) is greater than Equation (28). Consequently, the size of a break that can be detected is

\[
\ln f_1(\tau_1, \mu, \sigma|y_{1:T}) - \ln f_0(\tau_1, \sigma|y_{1:T}) > 0,
\]

\[
\ln(\lambda \sigma^2_{\tau_1} + (1-\lambda) \tilde{\sigma}^2_{\tau_1:T}) + \frac{2}{T} P > \lambda \ln \tilde{\sigma}^2_{\tau_1} + (1-\lambda) \ln \tilde{\sigma}^2_{\tau_1:T}
\]

Using \( P_{\text{BIC}} = -\ln \frac{0.95}{0.05} = \ln \frac{0.05}{0.95}, \) the size of the break that can be detected is

\[
\ln(\left[\frac{0.05}{0.95T}\right]^2 [\lambda \sigma^2_{\tau_1} + (1-\lambda) \tilde{\sigma}^2_{\tau_1:T}]) > \lambda \ln \tilde{\sigma}^2_{\tau_1} + (1-\lambda) \ln \tilde{\sigma}^2_{\tau_1:T}.
\]

\[
\square
\]

**Remark 8.** When \( T \to \infty, \) the penalty \( \left[\frac{0.05}{0.95T}\right]^2 \to 1. \) Since the logarithmic function is strictly concave, inequality (26) is always true asymptotically, meaning that any size of break can be asymptotically identified. In addition, inequality (25) can be simplified as

\[
e^{\frac{2P}{T}} (\lambda + (1-\lambda) \frac{\tilde{\sigma}^2_{\tau_1:T}}{\sigma^2_{\tau_1}}) > [\frac{\tilde{\sigma}^2_{\tau_1:T}}{\sigma^2_{\tau_1}}]^{(1-\lambda)}, \tag{29}
\]

Note that \( \left(\frac{\tilde{\sigma}^2_{\tau_1:T}}{\sigma^2_{\tau_1}}\right) \to_p \Delta \sigma^2_1. \) Figure 2b shows the detected values of \( \Delta \sigma^2_1 \) for different values of \( \lambda \) given the BIC penalty. Another interpretation is provided by Figure 2a, which documents the probability of detecting a break in the unconditional variance for several
values of $\lambda$ and $\Delta \sigma_1^2$ based on 10,000 simulated series of size $T = 4,000$ and $\sigma_0^2 = 3$. This data-generating process (DGP) is based on the GARCH DGP proposed in the simulation exercise, in which the unconditional variance is divided by 3 at $\lambda = 0.31$. In both figures, we observe that when $\Delta \sigma_1^2 = 1/3$, the probability of detecting a break amounts to one, regardless of where the break occurs in the sample.

5 Prior elicitation and estimation of the model

We present the prior distributions of the model parameters in the next subsection and then cover the simulation of the posterior distribution in Section 5.2.

5.1 Prior elicitation

The set of parameters can be decomposed into five categories: the parameters without the shrinkage prior $\alpha_1, \beta_1$, the shrinkage parameters with real line support $\alpha_{2:K_{\text{max}}+1}$ and with positive support $\beta_{2:K_{\text{max}}+1}$, the penalty parameters $\rho_{ik}$ and $\rho_{jk}$ for $i = 1, ..., K_1$, $j = 1, ..., K_2$ and $k = 1, ..., K_{\text{max}}$ and the break dates $\tau_{1:K_{\text{max}}}$.

The distributional choice of the first group of parameters is model dependent, while the two sets of shrinkage parameters are distributed according to $\alpha_{ik} \sim 2\text{MU}(a_i, b_i, \ln \rho_{ik})$ and
\( \beta_{jk} \sim 2\text{MU}(a_j, b_j, \ln \rho_{jk}) \) for \( i = 1, \ldots, K_1, \ j = 1, \ldots, K_2 \) and \( k = 1, \ldots, K_{\max} \), respectively. Note that the 2MU hyperparameters \( a \) and \( b \) are different for each parameter but do not depend on \( k \), while there is one random penalty variable per model parameter. The penalty parameters \( \rho_{ik} \) and \( \rho_{jk} \) are uniformly distributed with a lower bound of \( \frac{1}{2} \phi_{\text{BIC}} \) and an upper bound of \( \frac{3}{2} \phi_{\text{BIC}} \).

For the break dates \( \tau_{1:K_{\max}} \), we assume uniform distributions with informative bounds based on the set \( J_1 = \{ \bar{\tau}_1, \ldots, \bar{\tau}_{K_{\max} - 2} \} \), defined in Equation (14), as follows:

1. The first and last break date aim to capture breaks outside the time period \([h, T - h] \).

   Consequently, we set their prior distributions to \( \tau_1 \sim U[1, h - 1] \) and \( \tau_{K_{\max}} \sim U[T - h + 1, T] \).

2. For \( j = 2, \ldots, K_{\max} - 1 \), \( \tau_j \sim U[l_j, u_j] \) with \( l_j = u_{j-1} + 1 \) and \( u_j = \arg\min_{t \in [\bar{\tau}_{j-1}, \bar{\tau}_j - 1]} S_h(t) \).

### 5.2 Model estimation

For a given number of breaks \( K_{\max} \), MCMC inference is typically done by drawing sequentially from the posterior distributions \( \pi(\Theta|y_{1:T}, \tau_{1:K_{\max}}, \rho) \), \( \pi(\tau_{1:K_{\max}}|y_{1:T}, \Theta, \rho) \) and \( \pi(\rho|y_{1:T}, \tau_{1:K_{\max}}, \Theta) \), where \( \rho = (\rho_1, \ldots, \rho_{K_{\max}(K_1 + K_2)}) \) represents the penalty parameters of the 2MU prior distributions. We simulate the posterior distribution of the parameters by combining MCMC and sequential Monte Carlo (SMC) methods, a technique called “SMC sampler” (see Del Moral, Doucet, and Jasra (2006)). This method exhibits several advantages compared to the standard MCMC approach. First, as the realizations of the SMC (referred to as “particles” hereafter) evolve independently, the algorithm can be easily parallelized (e.g. Durham and Geweke (2014)). Second, as shown by Jasra, Stephens, and Holmes (2007) and Herbst and Schorfheide (2014), MCMC based on a single Markov chain are less appropriate than SMC methods for simulating multi-modal distributions, which is the case with shrinkage priors. Furthermore, the SMC sampler delivers an estimate of the marginal likelihood, while additional computation is required for MCMC algorithms. The sampler is also less sensitive to initial values and doesn’t require the choice of burn-in sample size.
Naturally, the SMC sampler also comes with its own disadvantages, the most important being the number of user-specified parameters to tune in order to run the algorithm. Therefore, inference is carried out with the tempered and time (TNT) algorithm (see Dufoys (2016)), a variant of the SMC sampler (Del Moral, Doucet, and Jasra (2006)) that automates the choice of the SMC parameters. The algorithm sequentially iterates by producing outcomes from the prior distribution to the posterior distribution by combining many importance sampling and MCMC simulations. The SMC sketch is detailed in Algorithm 1.

Algorithm 1 SMC algorithm with a fixed number of observations (offline)

Sample M particles from the prior distribution: \( \{\Theta^i\}_{i=1}^M, \{\tau^i\}_{i=1}^M, \{\rho^i\}_{i=1}^M \)

Set the tempered function \( \phi = 0 \), the normalized weights \( W_i = 1/M \) \( \forall i \in [1, M] \) and \( ESS = M \)

while \( \phi < 1 \) do

A - Correction step:

Find \( \tilde{\phi} > \phi \) such that \( M/(\sum_{i=1}^N \tilde{w}_i^2) = 0.95\text{ESS} \), where \( \tilde{w}_i \propto W_i[f(y_{1:T}|\Theta^i, \tau^i)]^{\tilde{\phi} - \phi} \)

\( \forall i \in [1, M] \); Set \( w_i = W_i[f(y_{1:T}|\Theta^i, \tau^i)]^{\tilde{\phi} - \phi} \)

\( \forall i \in [1, M] \); Compute the normalized weights \( W_i = w_i/\sum_{j=1}^N w_j \) and \( ESS = M/(\sum_{i=1}^N W_i^2) \)

Set \( \phi = \min(\tilde{\phi}, 1) \)

B - Resample step if \( ESS < 0.75M \)

Resample the particles by stratified sampling (Carpenter, Clifford, and Fearnhead (1999))

C - MCMC step with targeted distribution: \( \pi_\phi(\Theta, \tau, \rho|y_{1:T}) \propto [f(y_{1:T}|\Theta, \tau)]^{\phi} f(\Theta, \tau, \rho) \)

for \( j = 1 \) to \( N \) do

Apply \( N \) iterations of the MCMC sampler given in Section 5.3 on the M particles.

end for

end while

5.3 MCMC sampler

Algorithm 1 requires an MCMC sampler that targets the posterior distribution \( \pi_\phi(\Theta, \tau, \rho|y_{1:T}) \).

The sampler iterates sequentially over the posterior distributions \( \pi_\phi(\Theta|y_{1:T}, \tau_{1:K_{max}}, \rho) \), \( \pi_\phi(\tau_{1:K_{max}}|y_{1:T}, \Theta, \rho) \) and \( \pi_\phi(\rho|y_{1:T}, \tau_{1:K_{max}}, \Theta) \). We first detail how we sample from the first two distributions. Sampling from the third distribution (the penalty parameters) is done using the inverse transformation method as explained in Section 3.2.3.
5.3.1 Sampling from $\pi(\Theta|y_{1:T}, \tilde{r}_{1:K_{\text{max}}}: \rho)$

Interestingly, $M$ particles evolve simultaneously in the SMC sampler. To rejuvenate one particle, we randomly sample from one of the 10 proposal distributions in Dufays (2016) and accept or reject the candidate according to a Metropolis-Hastings acceptance ratio. The model-free proposal distributions are built from the location of the other particles and are adapted from the differential evolution optimization literature (for a review, see Das and Suganthan (2011)). In particular, at the $j$th MCMC iteration, updating the $i$th particle $\Theta^i$ is done as follows:

1. Choose uniformly and apply one of the following three types of mutation:
   - Standard mutation:
     \[
     \Theta^{\text{Mut}} = \Theta^{r_1} + F_{\text{DE}}(\Theta^{r_2} - \Theta^{r_3}),
     \]
     in which $F_{\text{DE}}$ is a fixed constant and $r_1, r_2, r_3$ are taken without replacement in the $M - 1$ remaining particles.
   - Trigonometric mutation:
     \[
     \Theta^{\text{Mut}} = \frac{3}{3} \sum_{i=1}^3 \Theta^{r_i} / 3 + (p_2 - p_1)(\Theta^{r_1} - \Theta^{r_2}) + (p_3 - p_2)(\Theta^{r_2} - \Theta^{r_3}) + (p_1 - p_3)(\Theta^{r_3} - \Theta^{r_1}),
     \]
     in which $p_i \propto f(y_{1:t} | \Theta^{r_i}, \tau^{r_i})f(\Theta^{r_i}, \tau^{r_i})$ for $i \in [1, 3]$ are probabilities such that $\sum_{i=1}^3 p_i = 1$ and the index $r_i$, for $i \in [1, 3]$, are taken without replacement in the $M - 1$ remaining particles.
   - Firefly mutation:
     \[
     \Theta^{\text{Mut}} = \Theta^{r_1} + F_{\text{FF}}(\Theta^{r_1} - \Theta^{r_2}),
     \]
     where $F_{\text{FF}}$ is a chosen constant and $r_1, r_2$ are taken without replacement in the $M - 1$ remaining particles.

2. Select with equal probability one of the three different moves and propose a new
candidate \( \Theta' \) for the parameters \( \Theta^i \):

- **DREAM proposal:**
  
  - If standard move was selected:
    
    \[
    \Theta' = \Theta^i + F(\delta, d)(\sum_{g=1}^{\delta} \Theta^{r_1(g)} - \sum_{h=1}^{\delta} \Theta^{r_2(h)}) + \zeta,
    \]
    
    where \( i \neq r_1(g), r_2(h) \); \( r_1(.) \) and \( r_2(.) \) are random integers uniformly distributed on the support \([1, M]_i\); and it is required that \( r_1(g) \neq r_2(h) \) when \( g = h \) and \( \zeta \sim N(0, \eta^2_x I) \); \( \delta \sim U[1, 3] \), \( F(\delta, d) = 2.38/\sqrt{2\delta d} \).
  
  - Otherwise:
    
    \[
    \Theta' = \Theta^i + Z_{\text{Dir}} F(\delta = 1, d)(\Theta^{\text{Mut}} - \Theta^{r_1}) + \zeta,
    \]
    
    in which \( \zeta \sim N(0, \eta^2_x I) \); \( \delta \sim U[1, 3] \), \( F(\delta, d) = 2.38/\sqrt{2\delta d} \) and \( Z_{\text{Dir}} = 1 \) with probability 0.5 and -1 otherwise.

- **Stretch move proposal:**

  \[
  \Theta' = \Theta^{\text{Mut}} + Z_{\text{Stretch}} (\Theta^i - \Theta^{\text{Mut}}),
  \]

  in which \( Z_{\text{Stretch}} \sim Z_S \) and \( Z_S \) is a random variable with a cumulative density function given by \( F_{Z_S}(x) = \frac{\sqrt{a_S^2 x - 1}}{a_S - 1} \), with \( a_S = 2.5 \).

- **Walk move:**

  \[
  \Theta' = \Theta^i + Z_{\text{RW}} (\Theta^i - x^{\text{Mut}}),
  \]

  in which \( Z_{\text{RW}} \sim Z_W \) and \( Z_W \) is a random variable with a cumulative density function given by \( F_{Z_W}(x) = 1 - \left(\frac{(a_W + 1)^{1/2} - (x + 1)^{1/2}}{(a_W + 1)^{1/2} - (a_W + 1)^{-1/2}}\right)^{1/2} \), with \( a_W = 2 \).
3. Accept or reject the candidate $\Theta'$ according to the Metropolis-Hastings ratio,

$$\min \left\{ \frac{q(\Theta') \pi(\Theta'|y_{1:T}, \tau, \rho)}{\pi(\Theta|y_{1:T}, \tau, \rho)}, 1 \right\},$$

in which $q(\Theta') = 1$ if the DREAM proposal is chosen, $q(\Theta') = |1 + Z_{W}|^{K_1-1}$ for the walk move and $q(\Theta') = |Z_{S}|^{K_1-1}$ for the stretch move.

The three types of moves have been proposed independently in the MCMC literature (for the DREAM algorithm, see Vrugt, ter Braak, Diks, Robinson, Hyman, and Higdon (2009); for the walk and the stretch moves, see Christen and Fox (2010) and Foreman-Mackey, Hogg, Lang, and Goodman (2013)).

5.3.2 Sampling from $\pi_{\phi}(\tilde{\tau}_{1:K^{\text{max}}}|y_{1:T}, \Theta, \rho)$

Chib (1998) develops models for efficient Bayesian inference of CP that rely on the forward-backward algorithm (see also Rabiner (1989)). The algorithm exhibits a computational complexity of $O(K_{\text{max}}^2 T)$ and must be applied at each MCMC iteration. As a consequence, the forward-backward algorithm is computationally demanding compared to the state-of-the-art frequentist approach, which achieves a computational complexity of $O(T \log T)$. Moreover, the forward-backward algorithm does not work on models with path dependence, such as the ARMA or GARCH processes considered in this paper. To remain generic, we use the D-DREAM Metropolis algorithm developed for CP-GARCH processes in Bauwens, Dufays, and De Backer (2011). In particular, at the $j^{\text{th}}$ MCMC iteration, updating the $i^{\text{th}}$ particle $\tilde{\tau}_{1:K_{\text{max}}}^{i}$ is done as follows:

1. Sample $\mathbf{x} \sim U[0, 1]^{K_{\text{max}}}$ and set $d = \sum_{i=1}^{K_{\text{max}}} 1_{\{x_i > \text{CR}\}}$.

2. Obtain a new break date candidate $\tilde{\tau}_{1:K_{\text{max}}}^{i}$ using the proposal

$$\tilde{\tau}_{1:K_{\text{max}}}^{i} = \tilde{\tau}_{1:K_{\text{max}}}^{i} + \text{round}[F(\delta, d)(\sum_{g=1}^{\delta} \tilde{\tau}_{1:K_{\text{max}}}^{r_1(g)} - \sum_{h=1}^{\delta} \tilde{\tau}_{1:K_{\text{max}}}^{r_2(h)}) + \zeta],$$

with $\forall g, h = 1, 2, ..., \delta$ and $i \neq r_1(g), r_2(h)$. We use $r_1(.)$ and $r_2(.)$ to denote random integers that are uniformly distributed on the $M - 1$ remaining particles, and it is
required that \( r_1(g) \neq r_2(h) \) when \( g = h \); round[.] means that we take the nearest integer and \( \zeta \sim N(0, \eta^2_\Theta I) \).

3. For each \( j = 1, 2, ..., K_{\text{max}} \), set \( \tilde{\tau}'_j = \tilde{\tau}_j \) if \( x_j < CR \).

4. Accept or reject the candidate \( \tilde{\tau}'_1:K_{\text{max}} \) according to the Metropolis ratio,

\[
\min\{ \frac{\pi(\tilde{\tau}'_1:K_{\text{max}} | y_{1:T}, \Theta, \rho)}{\pi(\tilde{\tau}_1:K_{\text{max}} | y_{1:T}, \Theta, \rho)}, 1 \}. \tag{36}
\]

All the empirical results are obtained with \( \delta = 3, \eta_\Theta = 0.003, CR = 0.2 \) and \( F(\delta, d) = \frac{2.38}{\sqrt{2d}} \), i.e. the optimal value for normal posterior distributions (see Vrugt, ter Braak, Diks, Robinson, Hyman, and Higdon (2009)). Note that the D-DREAM algorithm has the advantage of achieving a computational complexity of \( O(T) \), which is the lowest bound if the log-likelihood function needs to be estimated.

6 Simulation study

We carry out a simulation exercise, similar to Yau and Zhao (2016), to illustrate the performance of our approach. We use 12 DGPs based on AR, ARMA and GARCH processes with various breaks in levels, persistence and variance. Table 1 defines the DGPs from which we sample 100 time series each. The DGP without breaks has a sample size of 500 observations for the AR and ARMA models, and 2,000 observations for the GARCH model.

DGPs with breaks for the AR and ARMA models have sample sizes of 1,024, and 4,000 observations for the GARCH models.

Table 2 shows the frequency of selecting the different model specifications for the AR DGPs after having estimated the 100 time series for each of the six DGPs. The overall performance of the sparse CP-AR sampler is excellent. Specifically, when we run the sampler on the model without breaks in any of the mean and variance parameters (DGP 1), no false breaks are detected. One of DGP 2’s two breaks is a relatively small break in the first-order autoregressive coefficient, which the sampler is able to detect in 87 percent of cases. The changes in the second-order autoregressive coefficient and the variance parameter
are correctly detected in 94 and 99 percent of cases, respectively. DGP 3 has two breaks in the first-order autoregressive coefficient, which are detected correctly in 100 percent of cases. The detection rates of zero breaks in the constant parameter and the variance are 98 and 100 percent, respectively. The results for DGP 4 are similar to DGP 3. DGP 5, consisting of an AR process with two breaks in the variance, is detected correctly in 99 percent of cases, with detection rates of zero breaks in the constant parameter and the variance of 95 and 98 percent, respectively. The AR(4) process in DGP 6 has two breaks in the first-order autoregressive coefficient, which are correctly detected in 65 percent of cases, and has detection rates for the other parameters ranging between 78 and 93 percent.

Table 2 also shows averages and standard deviations of the posterior means of the break dates and model parameters when the correct specification is selected. In terms of break point detection, the average break dates for all DGPs are very close to their true values, with relatively small standard deviations.

Table 3 shows the frequency of selecting different model specifications for the ARMA and GARCH DGPs. The no-break ARMA and GARCH models (DGP 7 and DGP 10) have a correct detection rate of at least 96 percent. DGP 8 consists of an ARMA model with one break in the MA parameter and one break in the variance. While the correct detection rate for the parameters is 83 and 96 percent, respectively, the sampler also incorrectly detects breaks in the constant and autoregressive parameters in 20 and 24 percent of the cases, respectively. DGP 9 has an autoregressive coefficient that becomes insignificant after the second break and detects the first break more frequently (48 percent of the cases) than the correct two-break configuration (35 percent of the cases). The single break in the MA parameter is correctly detected in 67 percent of the cases, and the no break detection rate in the constant and variance parameters is 77 and 94 percent, respectively. The CP sampler also works well for the GARCH models specified in DGP 11 and DGP 12, with both specifications having two breaks, which are correctly detected in 67 and 78 percent of the cases, respectively. Finally, in terms of estimated break dates, we find (as expected) from Table 3 that the accuracy of estimated dates deteriorates for some of the parameters of the more complex ARMA (DGP 9) and GARCH (DGP 11) model specifications.
### Table 1 – DGPs used to simulate time series

#### AR Model: \( y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \epsilon_t \) where \( \epsilon_t \sim N(0, \sigma^2) \)

<table>
<thead>
<tr>
<th>DGP 1 - AR(1)</th>
<th>DGP 2 - AR(2)</th>
<th>DGP 3 - AR(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breaks</td>
<td>Sample size</td>
<td>Breaks</td>
</tr>
<tr>
<td>{\cdot}</td>
<td>500</td>
<td>{514,768}</td>
</tr>
<tr>
<td>Values</td>
<td>Regimes</td>
<td>Values</td>
</tr>
<tr>
<td>{0}</td>
<td>1</td>
<td>{0}</td>
</tr>
<tr>
<td>{-0.7}</td>
<td>1</td>
<td>{0.9,1.69,1.32}</td>
</tr>
<tr>
<td>{-}</td>
<td>(-)</td>
<td>{0,-0.81,-0.81}</td>
</tr>
<tr>
<td>{1}</td>
<td>1</td>
<td>{1,1}</td>
</tr>
</tbody>
</table>

#### DGP 4 - AR(1)  
| Breaks | Sample size | Values | Regimes |
| \{50\} | 1024 | \{400,750\} | 1024 |
| Values | Regimes |
| \{0\} | 1 | \{0\} | 1 | \{0\} | 1 |
| \{0.75,-0.5\} | 2 | \{0.999\} | 1 | \{0.6,-0.6,0.5\} | 3 |
| \{-\} | (-) | \{\cdot\} | (-) | \{-0.3\} | 1 |
| \{-\} | (-) | \{-\} | (-) | \{-0.4\} | 1 |
| \{-\} | (-) | \{-\} | (-) | \{0.2\} | 1 |
| \{1\} | 1 | \{1.225,1\} | 3 | \{1\} | 1 |

#### ARMA Model: \( y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 \epsilon_{t-1} + \epsilon_t \) where \( \epsilon_t \sim N(0, \sigma^2) \)

<table>
<thead>
<tr>
<th>DGP 7</th>
<th>DGP 8</th>
<th>DGP 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breaks</td>
<td>Sample size</td>
<td>Breaks</td>
</tr>
<tr>
<td>{\cdot}</td>
<td>500</td>
<td>{514,768}</td>
</tr>
<tr>
<td>Values</td>
<td>Regimes</td>
<td>Values</td>
</tr>
<tr>
<td>{0.5}</td>
<td>1</td>
<td>{0}</td>
</tr>
<tr>
<td>{0.8}</td>
<td>1</td>
<td>{0.9}</td>
</tr>
<tr>
<td>{0.4}</td>
<td>1</td>
<td>{0.7,\cdot,0.7,\cdot}</td>
</tr>
<tr>
<td>{1}</td>
<td>1</td>
<td>{1,1,2.5}</td>
</tr>
</tbody>
</table>

#### GARCH Model: \( y_t \sim N(0, \sigma_t^2) \) where \( \sigma_t^2 = \omega + \alpha(y_{t-1}^2 - \sigma_{t-1}^2) + \beta(\sigma_{t-1}^2 - \omega) \)

<table>
<thead>
<tr>
<th>DGP 10</th>
<th>DGP 11</th>
<th>DGP 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breaks</td>
<td>Sample size</td>
<td>Breaks</td>
</tr>
<tr>
<td>{\cdot}</td>
<td>2000</td>
<td>{1250,2000}</td>
</tr>
<tr>
<td>Values</td>
<td>Regimes</td>
<td>Values</td>
</tr>
<tr>
<td>{0.1}</td>
<td>1</td>
<td>{3.1,1}</td>
</tr>
<tr>
<td>{0.1}</td>
<td>1</td>
<td>{0.1,0.1,0.4}</td>
</tr>
<tr>
<td>{0.85}</td>
<td>1</td>
<td>{0.5,0.5,0.9}</td>
</tr>
</tbody>
</table>
Table 2 – AR DGPs: Frequency of selecting model specifications and date estimation

<table>
<thead>
<tr>
<th># Regimes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>True dates</th>
<th>Estimated dates</th>
<th>St. dev. dates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DGP 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DGP 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>6</td>
<td>6</td>
<td>87</td>
<td>1</td>
<td>0</td>
<td>${514,768}$</td>
<td>${498.81,770.82}$</td>
<td>${35.26,18.71}$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0</td>
<td>94</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>${514}$</td>
<td>${515.98}$</td>
<td>${73.01}$</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>99</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DGP 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>98</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>${400,612}$</td>
<td>${398.63,610.65}$</td>
<td>${6.10,4.13}$</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DGP 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>98</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>1</td>
<td>98</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>${50}$</td>
<td>${49.58}$</td>
<td>${6.17}$</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DGP 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>95</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>96</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0</td>
<td>1</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>${400,750}$</td>
<td>${394.71,754.29}$</td>
<td>${17.29,21.15}$</td>
</tr>
<tr>
<td><strong>DGP 6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>93</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>9</td>
<td>11</td>
<td>65</td>
<td>13</td>
<td>2</td>
<td>${400,612}$</td>
<td>${381.78,627.44}$</td>
<td>${44.62,50.44}$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>78</td>
<td>17</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>87</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>83</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>83</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Frequency is based on the posterior mode of the number of regimes per parameter over 100 simulated time series from the AR DGPs. Bolded values depict the true number of regimes. Estimated dates (st. dev. dates) are averages (standard deviations) of the dates for correctly detected model configurations.
Table 3 – ARMA and GARCH DGPs: Frequency of selecting model specifications and date estimation

<table>
<thead>
<tr>
<th># Regimes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>True dates</th>
<th>Estimated dates</th>
<th>St. dev. dates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DGP 7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>97</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>99</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>98</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DGP 8</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>80</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>76</td>
<td>21</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>12</td>
<td>83</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0</td>
<td>96</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DGP 9</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>77</td>
<td>9</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>16</td>
<td>48</td>
<td>35</td>
<td>1</td>
<td>2</td>
<td>{514,768}</td>
<td>{481.30,747.35}</td>
<td>{78.47,63.8}</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>24</td>
<td>67</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td></td>
<td>{674.55}</td>
<td>{137.07}</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>94</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DGP 10</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DGP 11</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega$</td>
<td>31</td>
<td>67</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>{1250}</td>
<td>{1340.21}</td>
<td>{228.12}</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>5</td>
<td>93</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>{2000}</td>
<td>{1923.04}</td>
<td>{350.67}</td>
</tr>
<tr>
<td>$\beta$</td>
<td>23</td>
<td>73</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>{2000}</td>
<td>{1379.01}</td>
<td>{398.77}</td>
</tr>
<tr>
<td><strong>DGP 12</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega$</td>
<td>22</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>{1250}</td>
<td>{1321.37}</td>
<td>{247.23}</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0</td>
<td>95</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>{2000}</td>
<td>{1903.28}</td>
<td>{254.05}</td>
</tr>
<tr>
<td>$\beta$</td>
<td>89</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ARMA and GARCH DGPs: Posterior mode of the number of regimes per parameter over 100 simulated time series. Bolded values correspond to the true number of regimes. Estimated dates (st. dev. dates) are averages (standard deviations) of the dates for correctly detected model configurations.
7 Applications

7.1 Treasury bill rates

We revisit the empirical application of Pesaran, Pettenuzzo, and Timmermann (2006) (PPT). They estimate a CP-AR model of monthly US three-month Treasury bill (T-Bill) rates from August 1947 to December 2002. In their model, all parameters change after a structural break. After estimating break models with different numbers of regimes, PPT find that the marginal log-likelihood is highest for a model with seven regimes. This leads to a model with 21 parameters for an AR(1) specification.

Table 4 displays the posterior means and standard deviations of the parameters. We observe that the intercept parameters $\mu_k$ have large posterior standard deviations and that the AR parameter $\beta_k$ indicates a unit root for each regime. Therefore, the principal regime changes appear to be due to the variance, which fluctuates from 0.015 up to 2.558.

Table 4 – Posterior medians and 80% confidence interval for the regimes of the most likely sparse CP-AR(1) model in relation to posterior means and standard deviations for the CP-AR(1) model of PPT

<table>
<thead>
<tr>
<th>Dates</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1957</td>
<td>0.021</td>
<td>1.002</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.020)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>1960</td>
<td>0.252</td>
<td>0.895</td>
<td>0.256</td>
</tr>
<tr>
<td></td>
<td>(0.208)</td>
<td>(0.071)</td>
<td>(0.068)</td>
</tr>
<tr>
<td>1966</td>
<td>0.017</td>
<td>1.006</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>(0.067)</td>
<td>(0.020)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>1979</td>
<td>0.220</td>
<td>0.969</td>
<td>0.260</td>
</tr>
<tr>
<td></td>
<td>(0.161)</td>
<td>(0.026)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>1982</td>
<td>0.412</td>
<td>0.958</td>
<td>2.558</td>
</tr>
<tr>
<td></td>
<td>(0.521)</td>
<td>(0.045)</td>
<td>(0.071)</td>
</tr>
<tr>
<td>1989</td>
<td>0.246</td>
<td>0.968</td>
<td>0.161</td>
</tr>
<tr>
<td></td>
<td>(0.211)</td>
<td>(0.27)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>2002</td>
<td>-0.004</td>
<td>0.992</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>(0.054)</td>
<td>(0.011)</td>
<td>(0.005)</td>
</tr>
</tbody>
</table>

To confirm this preliminary result, we estimate an AR(1) process via ordinary least
squares (OLS) and test inequality (25) on the residuals. In particular, we use a rolling window of 150 observations and set $\lambda = 0.5$ to test the inequality.¹ Figure 2 documents the periods when the inequality holds for different penalty parameters. Even with a very strong penalty, there is evidence of breaks in the variance. In addition, when the penalty is set to the BIC value, we recover most of the periods detected by the PPT process. To understand what the BIC penalty implies for possible break detection, Figure 3 shows the values of $\Delta \sigma^2 \in (0, 4)$ that can be detected for different values of break location $\lambda$. These results are based on 10,000 simulated series in which $\sigma_0^2$ is set to the unconditional variance of the OLS residuals with a sample size identical to the three-month T-Bill series. We observe that a large change in the variance is required to detect a break when $\lambda < 0.1$. This is also true to a lesser extent when $\lambda > 0.9$. Figure 2 highlights that small deviations in the variance ranging from 0.75 to 1.5 (in a multiplicative sense) cannot be detected.

Figure 2 – US three-month T-Bill rate: Detection of breaks in the variance of the OLS residuals. Gray denotes periods when the inequality (25) holds and therefore represents possible break points. The first graphic corresponds to the BIC penalty value for a rolling window of 150 observations.

¹ For the beginning and the end of the sample, we test the inequality by adapting $\lambda$ using a fixed series of 150 observations.
Figure 3 – US three-month T-Bill rate: Size of the breaks ranging from 0 to 4 that can be detected in the variance given the break location $\lambda \in (0, 1)$ when the penalty is set to the BIC value.

Next, we estimate the sparse CP-AR model exhibiting one or two lags, as well as a sparse CP-ARMA(1,1) process. Unlike the standard CP model, the sparse model can accommodate multiple breaks during the estimation process and detect the number of regimes for each parameter in one model estimation. Table 5 shows the posterior probabilities for the number of regimes per parameter. It provides clear evidence that breaks mainly occur in the variance parameters, regardless of the model specification. Interestingly, if we focus on the posterior mode of the number of regimes per parameter, the sparse CP-AR model exhibits nine or 10 parameters depending on the number of lags in the AR process, which is substantially lower than 21 parameters of the PPT CP-AR process. Table 4 provides the break dates and parameters corresponding to the sparse CP-AR(1) model, and Figure 4 shows the parameter variation over time.

We also compare our results with the recent frequentist change-point detection procedure proposed by Yau and Zhao (2016). We use the same hyperparameters as in their paper. The parameter $d$, which is used to determine the window length of $h$, is set to 2; the
Figure 4 – US three-month T-Bill rate: Posterior medians of the parameters over time and their corresponding 80% confidence intervals (in dashed lines) and posterior medians of the break dates given the best configuration.
Table 5 – US three-month T-Bill: Posterior probabilities (in %) of the number of regimes per model parameter

<table>
<thead>
<tr>
<th># Regimes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>87</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>46</td>
<td>50</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>71</td>
<td>29</td>
</tr>
<tr>
<td>AR(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>90</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>95</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>92</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>58</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ARMA(1,1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>93</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>35</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>97</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>72</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

number of maximum lags, $p_{max}$, is 10; the minimum distance between two change points is 50 observations; and the minimum distance between the relative position of two change points is 0.01.

Applying their methodology, we find three change points occurring in December 1957, November 1969 and in April 1983. The estimated 95% confidence intervals of these change points are very narrow for the two last breaks (+/- eight months and +/- two months, respectively). However, a limitation of the approach is highlighted by the confidence interval of the first structural break, since it ranges from -45 observations (i.e. 45 months before the start of the data set) and February 1972, which is much later than the second break date estimate.

7.2 Macro data

We estimate the sparse CP-AR(1) model for the 22 macroeconomic monthly time-series data from Bauwens, Koop, Korobilis, and Rombouts (2015). The authors use CP-AR(1) models from Pesaran, Pettenuzzo, and Timmermann (2006) and Koop and Potter (2007)
which, according to their definition, imply a full new set of parameters after each structural
break. Table 6 summarizes the sparse CP-AR(1) results by counting, for each of the 22 time
series, the number of estimated regimes for each of the three AR parameters. Examining
the mean parameters, 12 series have a single regime and the remaining 10 have either
two or three regimes. In contrast, every time series experiences at least one break in the
variance parameter, e.g. eight series have five estimated regimes. Estimating a sparse CP-
AR(4) model for the series provides a similar picture for the mean parameters, although
the number of breaks in the variance parameters tend to decrease.

<table>
<thead>
<tr>
<th># Regimes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_0)</td>
<td>12</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>12</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(\sigma^2)</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7 compares, for each of the macroeconomic series, the estimated break dates of
the sparse CP-AR(1) with the CP-AR(1) model of Pesaran, Pettenuzzo, and Timmermann
(2006). While estimated break dates in both models are often similar, the sparse CP
model is much easier to compute for many time series. For example, industrial production
(INDPRO) has four estimated breaks for the standard CP model, whereas the sparse CP
model detects that those for breaks are only in the variance parameters. We also observe
that the sparse CP model has more breaks than the standard CP model for several of the
time series. For example, non-borrowed reserves of depository institutions (BOGNONBR)
has four breaks for the standard CP model, and eight breaks for the sparse CP model.

7.3 S&P 500 data

We use the sparse CP-GARCH(1,1) model to estimate parameters for stocks in the S&P
500 index. We focus on the 384 stocks that are in the index for the full January 2000 to
**Table 7 – Estimated break dates for monthly macroeconomic times series**

<table>
<thead>
<tr>
<th>Series</th>
<th>Model</th>
<th>β0</th>
<th>β1</th>
<th>σ²</th>
<th># Breaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMCP</td>
<td>Shrink.</td>
<td></td>
<td>2005-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>2004-12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TB3MS</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>1960-06</td>
<td>1966-04</td>
<td>1979-07</td>
<td>1982-08</td>
</tr>
<tr>
<td>INDPRO</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>1960-01</td>
<td>1894-01</td>
<td>2008-06</td>
<td>2009-04</td>
</tr>
<tr>
<td>UTL11</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>1960-03</td>
<td>1984-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOUST</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AHEMAN</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>1968-07</td>
<td>1981-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP500</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GS10</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAA</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPIAUCSL</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEDFUNDS</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOGNONBR</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNRATE</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>2000-01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNRATE</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>2000-01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPIFCG</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>1972-10</td>
<td>1975-01</td>
<td>1981-06</td>
<td>2000-11</td>
</tr>
<tr>
<td>EXUSUK</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>1967-09</td>
<td>1967-12</td>
<td>1971-06</td>
<td></td>
</tr>
<tr>
<td>PAYEMS</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAPMNOI</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>1960-04</td>
<td>1983-06</td>
<td>1983-10</td>
<td></td>
</tr>
<tr>
<td>OILPRICE</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>1962-04</td>
<td>1962-04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUSLOANS</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>1989-06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTALSL</td>
<td>Shrink.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPT</td>
<td>2008-08</td>
<td>2008-10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: The table shows posterior medians of the break dates per mean parameter, given the most likely break date configuration in relation to break dates found with an AR(1) and the PPT approach in Bauwens, Koop, Korobilis, and Rombouts (2015). The series are as follows: PMCP (NAPM Commodity Prices Index), TB3MS (three-month Treasury bill: secondary market rate), INDPRO (Industrial Production Index), UTL11 (capacity utilization: manufacturing), HOUST (new privately owned housing units started), AHEMAN (average hourly earnings: manufacturing), M1 (money stock), SP500 (S&P 500 Index), GS10 (10-year Treasury constant maturity rate), AAA (Moody’s seasoned Aaa corporate bond yield), CPIAUCSL (Consumer Price Index: all items), FEDFUNDS (effective federal funds rate), BOGNONBR (non-borrowed reserves of depository institutions), M2 (M2 money stock), UNRATE (civilian unemployment rate), PPIFCG (Producer Price Index: all commodities), EXUSUK (U.S./U.K foreign exchange rate), PAYEMS (total non-farm payroll: all employees), NAPMNOI (ISM manufacturing: New Orders Index), BUSLOANS (commercial and industrial loans at all commercial banks), and TOTALSL (total consumer credit outstanding).*
October 2017 period, which includes 4,464 daily returns per stock.

Table 8 provides estimated probabilities of the most likely number of regimes per GARCH parameter. None of the parameters stay constant over the set of time series, though the estimated probability of having more than three regimes in the parameters is less than six percent. The parameter $\alpha$ rarely experiences a break, and when it does, two regimes is sufficient for most series. While the long-term variance often has one, two or three regimes, the persistence parameter $\beta$ exhibits two regimes in the majority of the series. Since the persistence parameter $\beta$ and (to a lesser extent) the $\alpha$ parameter experience breaks in many return series, change-point GARCH models where only the long-term variance changes over time, such as in Engle and Rangel (2008) and He and Maheu (2010), lack flexibility.

Table 8 – Sparse CP-GARCH model results for returns of stocks in the S&P 500 index: Estimated probabilities (in %) of the most likely number of regimes per model parameter over the 384 series between January 2000 and October 2017.

<table>
<thead>
<tr>
<th># Regimes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega$</td>
<td></td>
<td></td>
<td>38.54</td>
<td>35.68</td>
<td>20.05</td>
<td>5.21</td>
</tr>
<tr>
<td>$\alpha$</td>
<td></td>
<td></td>
<td></td>
<td>73.18</td>
<td>22.66</td>
<td>3.65</td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37.24</td>
<td>55.21</td>
</tr>
</tbody>
</table>

The red lines in Figure 5 show the average of the posterior median parameters over time for the 384 stocks. There appears to be no abrupt changes over time, meaning that the break dates are quite uniformly distributed over the stocks. Restricting our sample to the 28 financial sector stocks in the index, the black line shows that the unconditional variance abruptly changes around the financial crisis, signalling the start of a highly volatile and uncertain period. Interestingly, this abrupt change does not appear in the two other GARCH parameters. On the other hand, we observe that the persistence starts to slowly decrease in 2010 onward, indicating the end of the financial crisis. The alpha parameter (for all stocks as well as for financial sector stocks only) does not change much over time.
Figure 5 – Sparse CP-GARCH model: The red lines show the average of the posterior medians of parameters over time for all 384 stocks. The black lines show the average of the posterior medians of parameters over time for the 28 financial sector stocks.
8 Conclusion

This paper introduces the sparse change-point time series model that detects which parameters of a model change when a structural break occurs. The proposed approach has several advantages compared to the current methodology.

It solves the over-parametrization issue exhibited by CP models, which assume that all parameters vary when a break is detected. Moreover, our method only requires one estimation of the model to select the number of regimes and does not rely on the marginal likelihood. This substantially reduces computational time and coding efforts. Finally, the detection of breaks depends on a control parameter that penalizes the log-likelihood function, the advantage being that it is simple to interpret compared to a selection based on the marginal likelihood. These achievements are obtained by using shrinkage priors on the standard CP parameters. As the popular existing priors are not appropriate for our methodology, we introduce a discrete mixture of uniform components, detail its properties and discuss selection of its hyperparameters. An extensive simulation study based on AR, ARMA and GARCH processes illustrates the excellent performance of our sampler.

The empirical exercises first revisit the Treasury bill application of Pesaran, Pettenuzzo, and Timmermann (2006). They use a CP-AR(1) process that changes the value of all parameters when a break is detected. We show that our sparse results obtain very similar break dates. However, the sparse CP-AR process also highlights that most of the breaks occur in the variance parameter, making the CP-AR(1) considered in Pesaran, Pettenuzzo, and Timmermann (2006) over-parametrized.

We next compare the sparse CP-AR(1) model with results obtained by Bauwens, Koop, Korobilis, and Rombouts (2015) for 22 macroeconomic time series. While several time series exhibit similar break dates for the two different specifications, more regimes are detected with the sparse model. Nevertheless, the sparse process remains parsimonious, since typically only one or two parameters change when a break occurs.

As a final application, we focus on the returns for 384 stocks in the S&P 500 index from January 2000 to October 2017. We show that all the GARCH parameters experience structural changes and that the persistence parameter appears to change most frequently.
We conclude that GARCH models where only the long-term variance slowly changes over time are insufficient to capture all aspects of volatility.

This paper uses the new shrinkage priors to model univariate time series. Given the recent advances in time-varying parameter vector autoregressive models, such as Koop and Korobilis (2013) and Chan, Eisenstat, and Koop (2016), it would be interesting to extend the approach to multivariate time series.

APPENDIX

A  Comparison with existing shrinkage priors

For ease of notation, we work with a generic $\mu$ model parameter in this section. Shrinkage means that a substantial part of the prior density on $\mu$ is concentrated close to zero, implying no structural break.

Two popular choices of continuous mixture shrinkage priors have emerged from research on variable selection in high-dimensional linear regression. The first is the normal-gamma (N-G) prior of Griffin and Brown (2010), extending the double exponential prior of Park and Casella (2008), i.e. the Bayesian counterpart of the least absolute shrinkage and selection operator (LASSO). See also the scale mixture of normals in West (1987). The N-G prior is defined in hierarchical form as

$$
\begin{align*}
\mu | \Psi_\mu & \sim N(0, \Psi_\mu) \\
\Psi_\mu & \sim G
\end{align*}
$$

where $G$ is a gamma distribution $G(\lambda, \frac{1}{2\nu^2})$. The marginal density of $\Delta \mu$ has a variance of $2\lambda \nu^2$ and kurtosis of $\frac{3}{\lambda}$. The hyperparameters $\lambda, \nu^2$ allow us to generate sparsity when most of the density mass is close to zero, as shown in panel (a) of Figure 6. Panel (b) shows the variance distribution $\Psi_\mu$.

Second, the continuous spike-and-slab (S&S) prior, as advocated by Ishwaran and Rao
Figure 6 – N-G shrinkage prior. The solid line is $\lambda = 0.5$, and the dotted line is $\lambda = 3$ ($\nu = 1$).

(2005), is hierarchically defined as

$$
\Delta \mu | \Psi_{\mu}, \sigma_{\mu}^2 \sim N(0, \Psi_{\mu}\sigma_{\mu}^2)
$$

$$
\sigma_{\mu}^2 \sim \omega \delta_{\sigma_{\mu}^2=c} + (1 - \omega)\delta_{\sigma_{\mu}^2=1}
$$

$$
\Psi_{\mu} \sim IG(\nu/2, \eta/2)
$$

with mixing probability $\omega$. The point mass $c$ is chosen close to but different from zero, with appropriate hyperparameters to generate sparsity. Figure 7 presents an illustration of the marginal distribution as well as the distribution of the variance ($\Psi_{\mu}\sigma_{\mu}^2$).

The above shrinkage distributions have typically been used to generate sparsity in high-dimensional regression models (see Kalli and Griffin (2014) for an example for models with time-varying sparsity). However, we argue that these priors are not as well suited to CP models than our 2MU prior. The following three differences are discussed in the subsequent subsections:

1. The difficulty in relating the hyperparameters to the interpretation as a penalty on the log-likelihood function.

2. The interpretation of the maximum a posteriori formula.

3. The bias of the posterior expectation.
Figure 7 – S&S prior. The solid line is \( \omega = 0.4 \), and the dotted line is \( \omega = 0.8 \) (\( c = 0.01, \nu = 6, \eta = 10 \)).

A.1 Difficulty in setting the penalty for the log-likelihood function

The hierarchical definitions of the N-G and the S&S priors make their dependence on the penalty parameter difficult to interpret. To highlight this, we derive the marginal distribution of \( \mu \) implied by the two priors.

First, from Griffin and Brown (2010), the marginal distribution of the N-G prior is given by

\[
f_{NG}(\mu) = \frac{|\mu|^{\lambda-1/2}K_{\lambda-1/2}(|\mu|/\gamma)}{\sqrt{\pi 2^{\lambda-1/2} \gamma^{\lambda+1/2}\Gamma(\lambda)}},
\]

where \( \Gamma(.) \) denotes the gamma function and \( K(.) \) denotes the modified Bessel function of the third kind. As before, we can set the hyperparameters by relying on the penalty rule \( \log f_{NG}(\bar{x}) - \log f_{NG}(0) = P \). Two issues arise, the first being that we need to use two threshold values \( \bar{x} \) in order to set the two hyperparameters \( (\gamma \text{ and } \lambda) \). The second issue is that a detected break cannot be interpreted as an improvement in the chosen penalty in terms of the log-likelihood function.
Second, the marginal distribution implied by the continuous S&S priors is given by

\[ f_{S&S}(\mu) = \omega \left( \frac{\eta}{\nu c} \right)^{1/2} f_{\nu}^t((\frac{\eta}{\nu c})^{1/2} \mu) + (1 - \omega) \left( \frac{\eta}{\nu} \right)^{1/2} f_{\nu}^t((\frac{\eta}{\nu})^{1/2} \mu), \quad (38) \]

where \( f_{\nu}^t(.) \) denotes the student density function with \( \nu \) degrees of freedom. We can set the weight \( (\omega) \) according to the penalty value \( (P) \) as

\[ \omega = \frac{f_{\nu}^t(\bar{x}(\eta/\nu)^{1/2}) - f_{\nu}^t(0)e^P}{f_{\nu}^t(\bar{x}(\eta/\nu)^{1/2}) - f_{\nu}^t(0)e^P + c^{-1/2}(f_{\nu}^t(0)e^P - f(\bar{x}(\eta/\nu c))^{1/2})}. \quad (39) \]

Equation (39) highlights that \( \omega \) does not necessarily belong to \([0, 1]\) for any value of \( P, \nu \) and \( \eta \). In addition, \( \omega \) is a function of the threshold \( \bar{x} \). Therefore, its value will have an impact on break selection. Note that the same holds for the N-G prior, but not for the 2MU distribution; see (6), which does not depend on \( \bar{x} \). Specifically, the density function of the existing continuous shrinkage priors evaluated at any point between the threshold value and zero may drastically depend on the user’s break sensitivity. Figure 8 makes clear that different penalty values affect the density function between zero and the threshold value for the S&S prior as well as outside this interval, whereas they do not impact the 2MU prior. Consequently, the S&S prior implies that the model parameters are allowed to smoothly change from one regime to another for some penalties. In such instances, the detection and interpretation of abrupt switches in penalty values are more delicate.

### A.2 Interpretation of the maximum a posteriori formula

Without loss of generality, we consider a simple location model in which the intercept has a break at the \( \tau \)th observation given by

\[ y_t = \tilde{\mu} + \mu \mathbb{1}_{t>\tau} + \epsilon_t, \quad (40) \]

where \( \epsilon_t \sim i.i.d.N(0, \sigma^2) \). For each shrinkage prior, we derive the objective function, which is minimized when evaluated at the mode of the posterior distribution \( \mu|y_{1:T}, \sigma^2, \tilde{\mu}, \tau \). These objective functions are:
Figure 8 – (a) and (c) are student distributions with different penalties, centered at zero with $\eta = \nu = 50$ and $c = 0.1$. (b) and (d) are 2MU priors with different penalties, with $a = 2$ and $b = 8$. The threshold value is fixed at $\bar{x} = 1$ for the student distributions.

- The 2MU(P,a,b) prior (see Equations (7)-(8)):

$$\mu_{MAP}^{2MU} = \text{Arg.Min}_{|\mu| < \frac{b}{2}} \left\{ \frac{1}{2\sigma^2} \sum_{t=\tau+1}^{T} (y_t - \bar{\mu} - \mu)^2 - P \mathbb{1}_{\{\frac{a}{2} < |\mu| \leq \frac{b}{2}\}} \right\}$$

- The N-G prior:

$$\mu_{MAP}^{N-G} = \text{Arg.Min}_{\mu} \left\{ \frac{1}{2\sigma^2} \sum_{t=\tau+1}^{T} (y_t - \bar{\mu} - \mu)^2 - (\lambda - 1/2) \log(|\mu|) - \log(K_{\lambda-1/2}(\frac{|\mu|}{\gamma})) \right\}$$
The S&S prior:

\[
\mu_{MAP}^{S&S} = \text{Arg.Min}_{\mu} \left\{ \frac{1}{2\sigma^2} \sum_{t=\tau+1}^{T} (y_t - \tilde{\mu} - \mu)^2 - \log(f_{S&S}(\mu)) \right\}
\]

The penalties of the popular shrinkage priors are continuous for any \( \mu \neq 0 \), while the 2MU prior implies a discontinuous objective function. Consequently, the 2MU prior is related to hard-thresholding penalization \((L_0\) type of penalization) and the non-convex penalization literature (see Dicker, Baosheng, and Lin (2013) and Fan and Li (2001), for example). While an objective function with a hard-thresholding penalty can be justified by information criteria, it suffers from the curse of dimensionality problem, which makes minimization difficult (see Fan and Lv (2010) for details). For high-dimensional regressions, a LASSO-type penalty \((L_1)\) is typically advocated as optimal using the least angle regression algorithm of Efron, Hastie, Johnstone, and Tibshirani (2004). In a Bayesian setting, where the focus is on the entire posterior distribution, this feature comes at the cost of losing the tractability of the LASSO, as we have to simulate from a highly multi-modal posterior distribution.

A.3 Posterior expectation bias

In a Bayesian context, standard statistics summarize the posterior distribution and the Bayesian counterpart of the maximum likelihood estimator is typically the easy-to-compute posterior expectation of the parameters (or the maximum a posteriori, but the two quantities coincide for the three shrinkage priors in this Section). From a frequentist viewpoint, we highlight that the posterior expectations of the N-G and S&S priors are related to the ridge regression estimator and are biased, while the 2MU prior is unbiased. In a CP context, which typically has more breaks when the sample size increases, unbiasedness of the estimator of the local parameter is particularly appealing. To compute the bias for the different shrinkage priors, we continue to consider the simple location model specified by (40).
• The posterior distribution of the N-G prior is given by

\[ \mu | y_{1:T}, \sigma^2, \bar{\mu}, \tau, \Psi_\mu \sim N(\bar{\mu}, \bar{\sigma}^2), \]

where \( \bar{\sigma}^2 = \left[ \frac{T-\tau}{\sigma^2} + \frac{1}{\Psi_\mu} \right]^{-1} \) and \( \bar{\mu} = \bar{\sigma}^2 \sum_{t=\tau+1}^{T} \frac{(y_t - \tilde{\mu})}{\sigma^2} \). Therefore, \( E(\tilde{\mu}|\sigma^2, \bar{\mu}, \tau, \Psi_\mu) = \frac{(T-\tau)\Psi_\mu}{(T-\tau)\Psi_\mu + \sigma^2} \bar{\mu} \).

• Assuming that there is a break (i.e. \( \sigma^2_\mu = 1 \)), the posterior distribution of the S&S prior is also a normal distribution:

\[ \mu | y_{1:T}, \sigma^2, \mu, \tau, \Psi_\mu, \sigma^2_\mu = 1 \sim N(\bar{\mu}, \bar{\sigma}^2), \]

in which \( \bar{\sigma}^2 = \left[ \frac{T-\tau}{\sigma^2} + \frac{1}{\Psi_\mu} \right]^{-1} \) and \( \bar{\mu} = \bar{\sigma}^2 \sum_{t=\tau+1}^{T} \frac{(\mu - \tilde{\mu})}{\sigma^2} \). The bias is identical to the N-G prior.

• Assuming that there is a break (i.e. \( z = 1 \)), the posterior distribution of the 2MU prior is

\[ \mu | y_{1:T}, \sigma^2, \bar{\mu}, \tau, z = 1 \sim N(\bar{\mu}, \bar{\sigma}^2), \]

where \( \bar{\sigma}^2 = \sigma^2/(T - \tau) \) and \( \bar{\mu} = \sum_{t=\tau+1}^{T} (y_t - \tilde{\mu})/(T - \tau) \).

Only the posterior expectation of the 2MU is an unbiased estimator. Note that the posterior expectations of the N-G and S&S priors are obtained by minimizing the objective function

\[ \mu_{MAP}^{S,S,N-G} = \text{Arg.Min}_\mu \left\{ \frac{1}{2\bar{\sigma}^2} \sum_{t=\tau+1}^{T} (y_t - \tilde{\mu} - \mu)^2 + \frac{1}{2\Psi_\mu} \mu^2 \right\}, \]

which is the penalty on the ridge regression estimator.
A.4 Additional proof

In this Section, we show that \( \hat{\sigma}^2_{1:T} \to_p \mu_1^2 \lambda (1 - \lambda) + \sigma_0^2 \), a result that leads to inequality (23). By definition of \( \hat{\sigma}^2_{1:T} \), we have

\[
\hat{\sigma}^2_{1:T} = \frac{1}{T} \sum_{t=1}^{T} (y_t - \frac{1}{T} \sum_{s=1}^{T} y_s)^2,
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} (\mu_1 (1_{\{t > \tau_1\}} - \frac{T - \tau_1}{T}) + \sigma_0 (\varepsilon_t - \frac{1}{T} \sum_{s=1}^{T} \varepsilon_s))^2,
\]

\[
= \mu_1^2 \sum_{t=1}^{T} (1_{\{t > \tau_1\}} - \frac{T - \tau_1}{T})^2 + \sigma_0^2 \sum_{t=1}^{T} (\varepsilon_t - \bar{\varepsilon})^2 + 2 \sigma_0 \mu_1 \sum_{t=1}^{T} (1_{\{t > \tau_1\}} - \frac{T - \tau_1}{T})(\varepsilon_t - \bar{\varepsilon}),
\]

\[
= \mu_1^2 (T - \tau_1) - 2 \mu_1 (T - \tau_1)^2 T + \sigma_0^2 \sum_{t=1}^{T} (\varepsilon_t - \bar{\varepsilon})^2 T + o_p(1),
\]

\[
= \mu_1^2 (1 - \lambda - (1 - \lambda)^2) + \sigma_0^2 \sum_{t=1}^{T} (\varepsilon_t - \bar{\varepsilon})^2 T + o_p(1),
\]

\[
= \mu_1^2 \lambda (1 - \lambda) + \sigma_0^2 \sum_{t=1}^{T} (\varepsilon_t - \bar{\varepsilon})^2 T + o_p(1).
\]

Since \( \sum_{t=1}^{T} \frac{(\varepsilon_t - \bar{\varepsilon})^2}{T} \to_p 1 \), we have that \( \hat{\sigma}^2_{1:T} \to_p \mu_1^2 \lambda (1 - \lambda) + \sigma_0^2 \).

References


