2019-07

Selective linear segmentation for detecting relevant parameter changes

Arnaud Dufays
Elysee Aristide Houndetoungan
Alain Coën

Septembre/ September 2019

**Centre de recherche sur les risques
les enjeux économiques et les politiques publiques**

www.crrep.ca

ABSTRACT

Change-point processes are one flexible approach to model long time series. We propose a method to uncover which model parameter truly vary when a change-point is detected. Given a set of breakpoints, we use a penalized likelihood approach to select the best set of parameters that changes over time and we prove that the penalty function leads to a consistent selection of the true model. Estimation is carried out via the deterministic annealing expectation-maximization algorithm. Our method accounts for model selection uncertainty and associates a probability to all the possible time-varying parameter specications. Monte Carlo simulations highlight that the method works well for many time series models including heteroskedastic processes. For a sample of 14 Hedge funds (HF) strategies, using an asset based style pricing model, we shed light on the promising ability of our method to detect the time-varying dynamics of risk exposures as well as to forecast HF returns.

JEL Classification: C11, C12, C22, C32, C52, C53.

Keywords: change-point, structural change, time-varying parameter, model selection, Hedge funds.

Arnaud Dufays : Département des sciences de gestion, Université Namur et Département d'économique, Université Laval. CRREP et CeReFiM.
Elysee Aristide Houndetoungan : Département d'économique, Université Laval.
Alain Coën : Department of Finance, UQAM.

1 Introduction

Long time series are standard in this period of large publicly available datasets. Care is required when modeling such a time series since many of them span over critical events that may change the series dynamic. At least two statistical solutions exist to take into account these changes. On the one hand, a process with fixed parameters can be used but it needs to exhibit a rich and complex dynamic. This complexity often makes the model difficult to estimate and to interpret (see, for instance, long memory processes such as Geweke and Porter-Hudak (1983)). On the other hand, one can rely on time-varying parameter (TVP) models and in particular Markov-switching and change-point (CP) processes since they allow for abrupt changes in the model parameters when a critical event affects the series dynamic (see Hamilton, 1989; Bauwens, Koop, Korobilis, and Rombouts, 2015). For instance, CP models generally boil down to fitting standard and easy-to-interpret processes to model segments of long time series. This paper deals with CP linear regression models where we allow the mean parameters to change over time.

The CP literature dates back to Chernoff and Zacks (1964) and is nowadays vast. Just focusing on linear regressions, Andrews (1993), Bai and Perron (1998), Killick, Fearnhead, and Eckley (2012), Fryzlewicz et al. (2014) and Yau and Zhao (2016) develop prominent procedures to detect breakpoints. On the Bayesian side, there also exist many ways to estimate structural breaks and important contributions can be found in Stephens (1994), Chib (1998), Fearnhead and Liu (2007), Rigaill, Lebarbier, and Robin (2012) and Maheu and Song (2013). While all these methods differ in the criterion or in the algorithm used to detect the changes, most of them rely on the assumption that, when a break is detected (that may be triggered by the change in only one model parameter), a new segment is created and a new set of parameters needs to be estimated. Although the assumption seems harmless, it creates two important drawbacks:

1. From an interpretation perspective, if all the parameters have to change when a break is detected, it is difficult to assess which parameters have indeed abruptly varied and so it complicates the economic interpretation of the structural break.

2. Forecasting wise, when a parameter does not vary from one regime to another, its estimation is more accurate than if two parameters were considered over these two regimes. This feature could improve the predictions of the model.

In this paper, we propose a method to relax the assumption that a break triggers a change in all the model parameters. To do so, we first estimate the potential break dates exhibited by the series and then we use a penalized likelihood approach to detect which

parameters change. Because some segments in the CP regression can be small, we opt for a (nearly) unbiased penalty function, called the seamless-L0 (SELO) penalty function, recently proposed by Dicker, Huang, and Lin (2013). We prove the consistency of the SELO estimator in detecting which parameters indeed vary over time and we suggest using a deterministic annealing expectation-maximisation (DAEM) algorithm to deal with the multimodality of the objective function (see Ueda and Nakano, 1998). Since the SELO penalty function depends on two tuning parameters, we use a criterion (new in this literature) to choose the best tuning parameters and as a result the best model. This new criterion exhibits a Bayesian interpretation which makes possible to assess the parameters' uncertainty as well as the model's uncertainty. This last feature is determinant when predicting a time series since the Bayesian model averaging technique, that typically improves forecast accuracy, is readily applicable (see, e.g., Raftery, Kárnỳ, and Ettler, 2010; Koop and Korobilis, 2012).

We are aware of four other papers that also relax the assumption on the number of parameters that changes when a break is detected. In the frequentist literature, the influential paper of Bai and Perron (1998) proposes a method that also operates when only a subset of parameters can break. However, prior knowledge of this subset is required as the number of possibilities grows exponentially with the number of breaks as well as with the number of parameters that can break. From a Bayesian perspective, Giordani and Kohn (2008) (and its related empirical paper, Koop, Leon-Gonzalez, and Strachan (2009)) specify a mixture state space model and provide an example of how it can be used to distinguish the parameters that abruptly change over time. While the process is very flexible, the estimation procedure breaks down when the number of parameters is large.[1] Another Bayesian method is proposed by Eo (2016). He suggests estimating all the possible models and choosing the best specification using the marginal likelihood criterion. While the method works well for models exhibiting a small amount of parameters, it is not applicable for large models. Moreover, the approach is time-consuming as many posterior distributions need to be simulated. Recently, Dufays and Rombouts (2018) propose a Bayesian CP model where irrelevant parameters are shrunk to zero using shrinkage priors. The method is generic as it operates for linear and non-linear models. However, as they use a discontinuous shrinkage prior, it renders the estimation time-consuming, complicated to code and not suited for large models. Moreover, while a Monte Carlo study shows that the method works well, it is not supported by asymptotic results.

We believe that our method exhibits several advantages over the existing alternatives. Firstly, it operates for small and large dimensions. It only requires that the amount of

[1] At each iteration of the algorithm, it is required to explore the discrete space of the latent variable that is related to the mixture distributions of the innovations. This discrete space grows exponentially with the number of model parameters. See the argument in Chan, Koop, Leon-Gonzalez, and Strachan (2012) on page 9.

observations per segment is larger than the number of parameters to be estimated in the segment. Secondly, the estimation is very fast compared to the Bayesian alternatives and we provide an R package to disseminate our approach. As a final advantage, we relax the assumption on breakpoints *once* the structural breaks have been detected which makes our approach operating in combination with any existing CP methods. To be specific, in this paper, we illustrate our approach with the CP procedure of Yau and Zhao (2016) but any other CP method could have been used. Also, instead of choosing one breakpoint detection algorithm, we could apply several of these methods and discriminate the detected structural changes using our marginal likelihood criterion once the parameters that truly evolve have been identified.

A final reference close to our framework is Chan, Yau, and Zhang (2014) who propose a penalized regression for segmenting time series in piecewise linear models. The paper uses a group Lasso penalty function (see Yuan and Lin, 2006) to get an overestimated number of segments and in a second phase, an information criterion is used to improve the estimation. By doing so, their procedure stands for another CP detection method based on a penalized regression approach. While the purpose of their paper is different from ours, their framework could be used for detecting the relevant changes in parameters. Nevertheless, we differ from their methods in many aspects. First, we use an almost unbiased penalty function and from a theoretical perspective, as we use the penalized regression on a potential break date set, our assumptions for a consistent estimator are different and in line with the standard penalized regression literature. We also use a Bayesian criterion to select among the promising models uncovered by the penalty function which allows for model uncertainty and for Bayesian model averaging. Also, our estimation procedure is fast compared to Chan, Yau, and Zhang (2014) since we iterate on closed-form expressions and because our model exhibits fewer parameters. As a final difference, we provide break uncertainty.

An extensive Monte Carlo study highlights how our method works in practice. For instance, we show that even when the number of parameters are as high as 100, the approach accurately detects which parameters change when a break is detected.
Eventually, we apply our method on Hedge funds (HF) returns. As reported by Fung and Hsieh (1997) (and hereafter by many others), HF strategies differ significantly from regulated investments like mutual funds. HF managers aims at targeting absolute returns regardless of the market conditions. Therefore, they follow highly dynamic, complex and essentially opaque trading strategies. HF strategies are exposed to time-varying risk sensitivities captured by numerous and changing economic risk factors. This feature has important implications for performance appraisal. As highlighted by Fung, Hsieh, Naik, and Ramodarai (2008), by Meligkotsidou and Vrontos (2008), by Bollen and Whaley (2009), and more recently by Pat-

ton, Ramodarai, and Streatfield (2015), the dynamics of HF risk exposures and the nonlinear generating process of HF returns should be associated with market events and structural breaks. In this context, a CP detection method is particularly relevant for modeling them. For a sample of 14 monthly Credit Suisse HF indices spanning from March 1994 to March 2016, and using the asset based style pricing model introduced by Fung and Hsieh (2001), we illustrate the relevance of our selective linear segmentation model. Specifically, our modeling is particularly appealing to detect time-varying exposures in HF tradings. We also investigate the prediction performance of our approach and it turns out that the selective segmentation approach compares favorably in terms of root mean squared forecast errors and cumulative log-predictive densities with respect to other CP processes. In particular, it almost systematically dominates the CP model which assumes that all the parameters vary when a break is detected.

The paper is organized as follows. Section 2 documents the model specification and the SELO penalty function. Section 3 explains how the DAEM algorithm is applied to our framework. In Section 4, we detail the criterion used to select the SELO tuning parameters and we relate it to the Bayesian paradigm. Section 5 documents the CP method of Yau and Zhao (2016) and discusses how it can be slightly improved. An extensive Monte Carlo study is proposed in Section 6. We end the paper by applying the method on HF returns in Section 7. All the proofs are given in the Appendix.

2 Model specification

We consider a standard linear regression specified as

$$\begin{aligned} y_t &= \beta_1 + \beta_2 x_{t,2} + \ldots + \beta_K x_{t,K} + \epsilon_t \\ &= \mathbf{x}_t' \boldsymbol{\beta}_1 + \epsilon_t \,, \end{aligned} \tag{1}$$

where $\epsilon_t \sim MDS(0, \sigma^2)$ (in which MDS stands for the martingale difference sequence), $\mathbf{x}_t = (1, x_{t,2}, \ldots, x_{t,K})'$ and $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_K)'$. Typically, if a linear model is estimated over a long period, the parameters are subject to abrupt changes over time. To take this time-varying dynamic into account, we allow for $m-1$ structural breaks in the model parameters as follows,

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_i^* + \epsilon_t \,, \text{ for } \tau_{i-1} < t \leq \tau_i, \tag{2}$$

in which $\boldsymbol{\beta}_i^*$, is the true parameter of the explanatory variables over the regime i, $\boldsymbol{\tau}_0 = \{\tau_0, \ldots, \tau_m\} \in \mathbb{N}^{m+1}$ where $\tau_0 = 0$, $\tau_m = T$ and $\tau_i < \tau_{i+1}$ $\forall i \in [0, m-1]$. In this paper, we

are interested in capturing which parameters are subject to breaks and which do not vary over time. To do so, we reframe the model (2) as follows,

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta}_1^* + \mathbf{x}_t'(\sum_{j=2}^{m} \Delta\boldsymbol{\beta}_j^* \mathbf{1}_{\{t>\tau_{j-1}\}}) + \epsilon_t\,, \\ \mathbf{y} &= \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \end{aligned} \tag{3}$$

where $\mathbf{1}_{\{x>a\}} = 1$ if $x > a$ and zero otherwise, $\Delta\boldsymbol{\beta}_j^* = \boldsymbol{\beta}_j^* - \boldsymbol{\beta}_{j-1}^*$, for $j \in [2, m]$, stands for the model parameters in first-difference, $\mathbf{y} = (y_1, \ldots, y_T)'$, $\mathbf{X}_{\boldsymbol{\tau}} = (\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}_{\tau_1}, \ldots, \tilde{\mathbf{X}}_{\tau_{m-1}})$ with $\tilde{\mathbf{X}}_{\tau_i} = (\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, \mathbf{x}_{\tau_i+1}, \ldots, \mathbf{x}_T)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)'$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\prime}, \Delta\boldsymbol{\beta}_2^{*\prime}, \ldots, \Delta\boldsymbol{\beta}_m^{*\prime})' \in \Re^{Km\times 1}$. Note that the matrix $\tilde{\mathbf{X}}_{\tau_0}$ stands for the standard regressors since $\tau_0 = 0$. Regarding the notations, the first-difference parameter in regime j is a K-dimensional vector $\Delta\boldsymbol{\beta}_j^*$ such that $\Delta\boldsymbol{\beta}_j^* = (\Delta\beta_{j1}^*, \ldots, \Delta\beta_{jK}^*)'$. Let us also denote $A = \{(j,k); \Delta\beta_{jk}^* \neq 0, \text{ for } j \in [2, m] \text{ and for } k \in [1, K]\}$, the set of indices defining the true model.

Our strategy to uncover which parameters truly vary over time consists in first finding where are the potential break dates $\boldsymbol{\tau}$, then, in a second phase, in detecting which parameters evolve. Note that even when we know the true break dates $\boldsymbol{\tau}$, the problem of finding which parameters vary when a break occurs is not straightforward as the number of models to consider amounts to $2^{(m-1)K}$. Consequently, it is infeasible to carry out an exhaustive model selection when K or m is large. We propose a penalized likelihood approach to explore this large model space and to select which parameters experience breaks. To focus on our selective segmentation approach, we shall first assume that we have obtained a set of potential break dates $\boldsymbol{\tau}$. We discuss how we estimate this set in Section 5.

Remark 1. In the situation where all the models can be considered (i.e., $(m-1)K \leq 10$), we do not need to rely on the penalized likelihood approach explained in Section 2.1. In particular, we could directly estimate all the model combinations and select the best one according to the marginal likelihood criterion given in Section 4.

2.1 Penalized likelihood and choice of the penalty function

As emphasized by Equation (3), given a set of break dates $\boldsymbol{\tau}$, the problem of finding which parameters abruptly change when a break occurs boils down to a penalized linear regression problem. Specifically, one can solve the following optimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + T\sum_{j=2}^{m}\sum_{k=1}^{K} \text{pen}(\Delta\beta_{jk}), \tag{4}$$

where $||\ .\ ||_p$ denotes the L_p norm and $\text{pen}(\Delta\beta_{jk})$ stands for a penalty function. Popular choices of $\text{pen}(\Delta\beta_{jk})$ are the Lasso penalty function (i.e., $\text{pen}(\Delta\beta_{jk}) = \lambda||\Delta\boldsymbol{\beta}_{jk}||_1$, see Tibshirani (1994)) or the rigde function (i.e., $\text{pen}(\Delta\beta_{jk}) = \lambda\Delta||\beta_{jk}||_2^2$, see, for instance, Ishwaran and Rao (2005)).

Following Fan and Li (2001), standard desirable properties induced by a penalty function are i) unbiasdness, ii) sparsity and iii) continuity. For instance, the ridge function is only continuous while the Lasso penalty function achieves sparsity and continuity (beside at zero). However one standard issue with these popular penalty functions is that they provide biased (but typically consistent) estimators. In our framework, this drawback is problematic since a segment can sometimes contain a small amount of observations that makes consistency results not sufficient. Recently, Dicker, Huang, and Lin (2013) propose a penalty function, called seamless-L_0 (SELO), that exhibits all the desirable properties. For a model parameter denoted ω, the penalty function reads as

$$\mathcal{P}_{\mathbf{SELO}}(\omega|\zeta,\lambda) \quad = \quad \frac{\lambda}{\ln 2}\ln(\frac{2|\omega|+\zeta}{|\omega|+\zeta}),$$

where the parameter ζ controls for the concavity of the function and λ stands for the penalty imposed when $\omega \neq 0$. We slightly modify their function to end up with parameters that are directly interpretable. In fact, we use the following penalty function,

$$\mathcal{P}_{\mathbf{SELO}}(\omega|a,\lambda) \quad = \quad \frac{\lambda}{\ln 2}\ln(\frac{2(\frac{|\omega|}{a})+\zeta}{(\frac{|\omega|}{a})+\zeta}), \tag{5}$$

where $\zeta = \frac{2^y-2}{1-2^y}$ with $y \in (0,1)$, the parameter a can be interpreted as an interval $\omega \in [-a,a]$ in which ω will be biased since $\mathcal{P}_{\mathbf{SELO}}(a) = \lambda y$. In practice, we set $y = 0.99$ so that when $|\omega| > a$, we have $\mathcal{P}_{\mathbf{SELO}}(\omega) \approx \lambda$ and $\frac{d\mathcal{P}_{\mathbf{SELO}}(\omega)}{d\omega}|_{|\omega|>a} \approx 0$. Figure 1 shows the SELO penalty function with $a = 1$ and $\lambda = 0.9$. We observe that the function is almost flat for absolute values greater than a.

2.2 Consistency of the SELO estimator

The interval $[-a,a]$ in which a parameter is biased is likely to change with the variable to which it refers. Furthermore, if we assume that this interval is fixed over time, we should set a new parameter a for each variable on the m segments. Unlike Dicker, Huang, and Lin (2013) who define a single parameter for all the variables, we use K parameters $a_1,\ldots,a_K$,

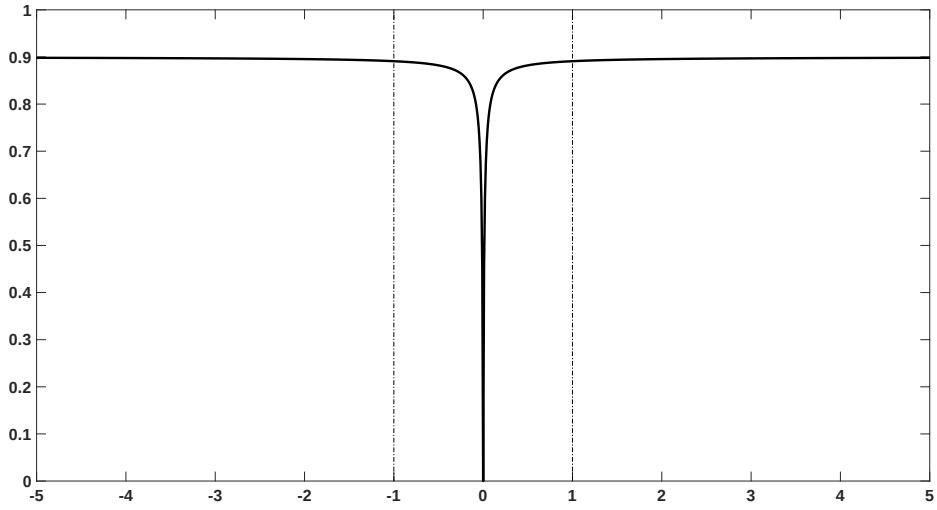


Figure 1: **SELO penalty function.** Penalty function is shown in solid black lines while vertical dotted lines highlight the interval [-a,a]. The SELO parameters are set to $\lambda = 0.9$ and $a = 1$.

that is, one per explanatory variable. Thus, the objective function to minimize is given by

$$f(\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + T\frac{\lambda}{\ln(2)}\sum_{j=2}^{m}\sum_{k=1}^{K}\ln\left(\frac{2(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta}{(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta}\right). \tag{6}$$

Before discussing how to maximize the objective function, we present the main results about the modified SELO estimator. As highlighted in Dicker, Huang, and Lin (2013), the SELO estimator is consistent under reasonable conditions. Proposition 1 shows that this consistency result also applies in our framework. To do so, we consider the following assumptions (in which a sequence $\omega_T \to \omega$ is understood as $\lim_{T\to\infty}\omega_T = \omega$):

A1. $\boldsymbol{\tau} = \boldsymbol{\tau}_0$ and $\forall j \in [1, m]$, we have $\tau_j - \tau_{j-1} = T\delta_{\tau_j} \to \infty$, where $\sum_{j=1}^{m}\delta_{\tau_j} = 1$ and $\delta_{\tau_j} > \epsilon_\tau > 0$ with $\epsilon_\tau = \mathcal{O}(1)$.

A2. $\frac{Km\sigma^2}{T} \to 0$ and $\rho\sqrt{\frac{T}{Km\sigma^2}} \to \infty$, where $\rho = \min_{r,k\in A}\left(|\Delta\beta^*_{r,k}|\right)$.

A3. There exist r_0, $R_0 > 0$ such that $r_0 \leq \lambda_{T,\min} < \lambda_{T,\max} \leq R_0$, where $\lambda_{T,\min}$ and $\lambda_{T,\max}$ are the smallest and largest eigenvalues of $\left(T^{-1}\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}\right)$ respectively.

A4. The process $\{\epsilon_t, \mathbf{x}_t\}_{t\in(\tau_{j-1},\tau_j]}$ is ergodic and stationary for any $j = 1, \ldots, m$. Moreover,

$\forall\ t \in [1, T]$, $\mathbb{E}(\epsilon_t|\mathbf{x}_t) = 0$ and $\mathbb{E}(\epsilon_t^2|\mathbf{x}_t) = \sigma^2$.

A5. $\lambda = \mathcal{O}_p(1)$, $\zeta = \mathcal{O}(1)$ and $a_k = \mathcal{O}_p\left((mK)^{-1}\left(\frac{mK\sigma^2}{T}\right)^{\frac{3}{2}}\right)$, $\forall k \in [1, K]$.

We first discuss the assumptions before detailing our consistency result. Assumptions A1 to A5 are similar to those found in the variable selection literature (see Fan, Peng, et al., 2004; Dicker, Huang, and Lin, 2013) and in the CP literature (see Bai and Perron, 1998; Yau and Zhao, 2016). Condition A1 assumes that the estimated change points are the true locations. However, the SELO estimator maintains the same asymptotic properties with a set of potential breakpoints as long as it contains the true break dates (see the adapted assumption A6 below). In such case, Proposition 1 also ensures that the number of breakpoints is consistently estimated. Note that condition A1 implies that the length of each segment increases linearly with T. As a consequence, the number of regimes is fixed. Although unattractive, this condition is generally made in the CP literature (see, e.g., Perron et al., 2006; Yau and Zhao, 2016). For interested readers, Perron et al. (2006) motivate this assumption in details. Assumption A2 allows the minimum break size to decrease with the sample size but not at a faster rate than $\sqrt{\frac{Km\sigma^2}{T}}$. Conditions A3 are related to the eigenvalues and are standard in the variable selection literature (see, e.g., Zhang, Li, and Tsai, 2010). However, we show in Appendix A.4 that this condition is not innocuous and that it implies a fixed number of regimes in our setting. Avoiding this assumption would imply stronger conditions on the process $\{y_t, \mathbf{x}_t\}$ (see, e.g., Chan, Yau, and Zhang, 2014). The assumption A4 refers to ergodicity and stationarity of each segment and imposes standard exogeneity and homoscedasticity. This assumption ensures that sampled counterparts of the first two moments of $\{\mathbf{x}_t\epsilon_t\}$ are converging to finite values. Eventually, condition A5 defines restrictions on the tuning parameters rate. The same condition applies in Dicker, Huang, and Lin (2013). The consistency of SELO estimator is given by the following Proposition.

Proposition 1. *Assume that A1-A5 hold and let,*

$$f_T(\boldsymbol{\beta}) = \frac{1}{T}||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + \sum_{j=2}^{m}\sum_{k=1}^{K}\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda). \tag{7}$$

There exists a sequence of $\sqrt{\frac{T}{Km\sigma^2}}$*-consistent local minima* $\hat{\boldsymbol{\beta}}$ *of* $f_T(\boldsymbol{\beta})$ *as defined by Equation* (7) *such that:*

(i) $\lim_{T\to\infty} \mathbf{P}\left(\left\{(j,k); \hat{\beta}_{jk} \neq 0\right\} = A\right) = 1$

(ii) $\forall\ \delta > 0$, $\lim_{T\to\infty} \mathbf{P}\left(||\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^*|| > \delta\right) = 0$

Proof. The proof is given in Appendix A □

Remark 2. Proposition 1 also applies when the set of breakpoints contains additional spurious break dates. In particular, Proposition 1 holds if we relax assumption A1 by the less restrictive assumption:

A6. $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{\hat{m}}\}$ with $\hat{m} \geq m$ and $\boldsymbol{\tau}_0 \subseteq \boldsymbol{\tau}$ and $\forall j \in [1, \hat{m}]$, we have $\tau_j - \tau_{j-1} = T\delta_{\tau_j} \to \infty$, where $\sum_{j=1}^{\hat{m}} \delta_{\tau_j} = 1$ and $\delta_{\tau_j} > \epsilon_\tau > 0$ with $\epsilon_\tau = \mathcal{O}(1)$.

3 Estimation

The objective function to minimize is given by

$$\begin{aligned} f(\boldsymbol{\beta}) &= ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + T\frac{\lambda}{\ln(2)} \sum_{j=2}^{m}\sum_{k=1}^{K} \ln\left(\frac{2(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta}{(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta}\right), \\ &= ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + \sum_{j=2}^{m}\sum_{k=1}^{K} \ln q_k(\Delta\beta_{jk}), \end{aligned} \tag{8}$$

in which $q_k(\Delta\beta_{jk}) = \left(\frac{2(\frac{|\Delta\beta_{jk}|}{a_k})+\zeta}{(\frac{|\Delta\beta_{jk}|}{a_k})+\zeta}\right)^{\left(\frac{T\lambda}{\ln(2)}\right)}$. Due to the penalty function, we cannot find any analytical expression of the minimizer. In addition to that, the function likely exhibits many local modes which complicates the optimization. We address the problem of finding the global mode by using a deterministic annealing expectation-minimization (DAEM) algorithm (see Ueda and Nakano, 1998). To do so, we first approximate the penalty function by a mixture of three Normal components (to take into account the large tail of the SELO penalty function), the details of it are given in Appendix A.5. Secondly, since minimizing the sum of squared residuals is identical to maximizing a likelihood function when the error term is normally distributed, we work with the following model

$$\mathbf{y} \quad = \quad \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta} + \eta,$$

where $\eta \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_T)$. The modified model implied the following objective function to maximize with respect to $(\boldsymbol{\beta}, \sigma^2)$:

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= -\frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 - \sum_{j=2}^{m}\sum_{k=1}^{K}\ln g_k(\Delta\beta_{jk}), \\ g_k(\Delta\beta_{jk}) &= \sum_{i=1}^{3}\omega_i^{(k)} f_N(\Delta\beta_{jk}|\mu_i^{(k)}, s_i^{(k)}), \end{aligned} \tag{9}$$

where $f_N(x|\mu, s)$ stands for the normal density function evaluated at x with expectation and variance given by μ and s respectively and $\omega_i^{(k)} \in (0,1)$ with $\sum_{i=1}^{3}\omega_i^{(k)} = 1$. Note that the function $f(\boldsymbol{\beta}, \sigma^2)$ in Equation (9) is proportional to the posterior density of the parameter distribution $\boldsymbol{\beta}, \sigma^2|\mathbf{y}$ from a Bayesian perspective with prior distributions given by $f(\sigma^2, \boldsymbol{\beta}_1) \propto 1$ and $f(\Delta\beta_{jk}) = g_k(\Delta\beta_{jk})$ for $j \in [2, m]$ and $k \in [1, K]$. The optimization is therefore equivalent to finding the mode of $\boldsymbol{\beta}, \sigma^2|\mathbf{y}$. Using a data augmentation approach, we add latent variables $\mathbf{z} = (z_{21}, z_{22}, \ldots, z_{mK})'$ such that $f(z_{jk} = i) = \omega_i^{(k)}$, $\forall j \in [2, m]$, $\forall k \in [1, K]$ and $\forall i \in [1, 3]$. With these latent variables, we can write the prior distribution of $\Delta\beta_{jk}$ in a convenient hierarchical way as follows,

$$\begin{aligned} f(\Delta\beta_{jk}|z_{jk} = i) &= f_N(\Delta\beta_{jk}|\mu_i^{(k)}, s_i^{(k)}), \\ f(z_{jk} = i) &= \omega_i^{(k)}. \end{aligned}$$

By fixing $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$, the EM algorithm (and its DAEM variant) solves the following optimization at iteration n,

$$\text{argmax}_{\boldsymbol{\theta}_n} Q(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1}) = \text{argmax}_{\boldsymbol{\theta}_n}\mathbb{E}_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}_{n-1}}(\ln f(\boldsymbol{\theta}_n, \mathbf{z}|\mathbf{y})|\mathbf{y}, \boldsymbol{\theta}_{n-1}).$$

One can easily show that maximizing $Q(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1})$ implies that $f(\boldsymbol{\theta}_n|\mathbf{y}) \geq f(\boldsymbol{\theta}_{n-1}|\mathbf{y})$.

3.1 Derivation of the DAEM algorithm

To apply the DAEM algorithm, we need to find an expression of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{n-1})$. Given a set of parameter $\boldsymbol{\theta}_{n-1}$, we have that

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{n-1}) &= \mathbb{E}_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}_{n-1}}(\ln f(\boldsymbol{\theta}|\mathbf{y},\mathbf{z})f(\mathbf{z}|\mathbf{y})|\mathbf{y},\boldsymbol{\theta}_{n-1}) \\ &\propto \ln f(\mathbf{y}|\boldsymbol{\beta},\sigma^2) + \ln f(\boldsymbol{\beta}_1,\sigma^2) + \sum_{j=2}^{m}\sum_{k=1}^{K}\sum_{i=1}^{3} -\frac{(\Delta\beta_{kj}-\mu_i^{(k)})^2}{2s_i^{(k)}} f(z_{kj}=i|\mathbf{y},\boldsymbol{\theta}_{n-1}), \\ &\propto \ln f(\mathbf{y}|\boldsymbol{\beta},\sigma^2) - \frac{1}{2}\sum_{i=1}^{3}(\boldsymbol{\beta}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i(\boldsymbol{\beta}-\boldsymbol{\mu}_i), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu}_i &= (\underbrace{0,0,\ldots,0}_{\text{K-dimensional}}, \mu_i^{(1)}, \mu_i^{(2)}, \ldots, \mu_i^{(K)}, \mu_i^{(1)}, \ldots)' \in \Re^{mK\times 1}, \\ \boldsymbol{\Sigma}_i &= \text{diag}(\underbrace{0,0,\ldots,0}_{\text{K-dimensional}}, \frac{p_{21}^{(i)}}{s_i^{(1)}}, \frac{p_{22}^{(i)}}{s_i^{(2)}}, \ldots, \frac{p_{2K}^{(i)}}{s_i^{(K)}}, \frac{p_{31}^{(i)}}{s_i^{(1)}}, \ldots, \frac{p_{mK}^{(i)}}{s_i^{(K)}}), \end{aligned}$$

with $p_{jk}^{(i)} = f(z_{jk}=i|\mathbf{y},\boldsymbol{\theta}_{n-1})$ $\forall i \in [1,3]$,$\forall j \in [2,m]$ and $\forall k \in [1,K]$. Importantly, the difference between the EM algorithm and its DA version only appears in the quantities $p_{jk}^{(i)}$. In fact, the DAEM algorithm introduces an increasing function $\phi(r) : [1,N] \to (0,1]$ such that $0 < \phi(1) \leq 1$ and $\phi(N) = 1$. For each value $r = 1,\ldots,N$, it applies recursively the EM algorithm (that starts with the final estimate of the previous EM algorithm) where the posterior probabilities $p_{jk}^{(i)}$ are denoted $p_{jk}^{(i,\phi(r))}$ and are modified as follows,

$$p_{jk}^{(i,\phi(r))} \propto (f_N(\Delta\beta_{jk}|\mu_i^{(k)},s_i^{(k)})\omega_i^{(k)})^{\phi(r)}. \tag{10}$$

When $r = N$, the increasing function $\phi(r) = 1$ and the standard EM algorithm is run (but with a promising starting point). To find the maximum of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{n-1})$, we sequentially maximize $\boldsymbol{\beta}$ given σ^2 and then σ^2 with respect to $\boldsymbol{\beta}$. This approach, called coordinate iterative ascent, operates in two steps:

1. Compute $\boldsymbol{\beta}_n = \text{argmax}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta},\sigma^2_{n-1}|\boldsymbol{\theta}_{n-1})$.
2. Compute $\sigma^2_n = \text{argmax}_{\sigma^2} Q(\boldsymbol{\beta}_n,\sigma^2|\boldsymbol{\theta}_{n-1})$.

At the end of the two steps, we necessarily have $Q(\boldsymbol{\beta}_{n-1}, \sigma^2_{n-1}|\boldsymbol{\theta}_{n-1}) \leq Q(\boldsymbol{\beta}_n, \sigma^2_{n-1}|\boldsymbol{\theta}_{n-1}) \leq Q(\boldsymbol{\beta}_n, \sigma^2_n|\boldsymbol{\theta}_{n-1})$. The maximisation of $\boldsymbol{\beta}$ given σ^2_{n-1} leads to

$$\boldsymbol{\beta}_n = [\sigma^{-2}_{n-1} X'_{\boldsymbol{\tau}} X_{\boldsymbol{\tau}} + \sum_{i=1}^{3} \boldsymbol{\Sigma}_i]^{-1} [\sigma^{-2}_{n-1} X'_{\boldsymbol{\tau}} \mathbf{y} + \sum_{i=1}^{3} \boldsymbol{\Sigma}_i \boldsymbol{\mu}_i].$$

The update of σ^2 conditional to $\boldsymbol{\beta}_n$ is given by

$$\sigma^2_n = \frac{[(\mathbf{y} - X_\tau \boldsymbol{\beta}_n)'(\mathbf{y} - X_\tau \boldsymbol{\beta}_n)}{T}.$$

We summarize the DAEM procedure in Algorithm 1. In practice, the minimum distance e indicating a convergence of the algorithm is set to 10^{-5} and the number of DAEM iteration N is fixed to 10.

Algorithm 1 DAEM algorithm

Initialize $\boldsymbol{\beta}_0$ using Algorithm 2
Set $\sigma^2_0 = \frac{[(\mathbf{y} - \mathbf{X}_\tau \boldsymbol{\beta}_0)'(\mathbf{y} - \mathbf{X}_\tau \boldsymbol{\beta}_0)}{T}$, $\phi(1) = (\frac{1}{N})^2$, $r = 1$ and dist $= \infty$.
while $r <= N$ **do**
  Set $n = 0$ and $\boldsymbol{\theta}_n = (\boldsymbol{\beta}'_0, \sigma^2_0)'$.
  while dist $> e$ **do**
    Increment $n = n + 1$.
    Compute the posterior probabilities $p_{jk}^{(i,\phi(r))}$ given in Equation (10) for i=1,2,3
    Compute the mean parameters

$$\boldsymbol{\beta}_n = [\sigma^{-2}_{n-1} \mathbf{X}'_{\boldsymbol{\tau}} \mathbf{X}_{\boldsymbol{\tau}} + \sum_{i=1}^{3} \boldsymbol{\Sigma}_i]^{-1} [\sigma^{-2}_{n-1} \mathbf{X}'_{\boldsymbol{\tau}} \mathbf{y} + \sum_{i=1}^{3} \boldsymbol{\Sigma}_i \boldsymbol{\mu}_i].$$

    Compute the variance parameter

$$\sigma^2_n = \frac{[(\mathbf{y} - \mathbf{X}_\tau \boldsymbol{\beta}_n)'(\mathbf{y} - \mathbf{X}_\tau \boldsymbol{\beta}_n)}{T}.$$

    Set $\boldsymbol{\theta}_n = (\boldsymbol{\beta}'_n, \sigma^2_n)'$ and compute the distance value dist $= ||\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}||_2$.
  end while
  Increment $r = r + 1$ and set $\phi(r) = (\frac{r}{N})^2$. Set $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_n$ and $\sigma^2_0 = \sigma^2_n$.
end while

The EM and the DAEM algorithms are sensitive to starting values. Inspired by Zhao, Hautamäki, Kärkkäinen, and Fränti (2012), we mitigate this issue by randomly exploring the model space using a swapping approach before applying the DAEM algorithm. To be specific, we generate N_{init} values as explained in Algorithm 2 and we initialize the DAEM algorithm

with the parameter estimates that minimize the penalized function given in Equation (8). In practice, we set $N_{\text{init}} = \min(2^{(m-1)K-1}, 3000)$.

Algorithm 2 Initialization of the DAEM algorithm

for $n = 1$ to N_{init} **do**
  Set $\hat{A} = \emptyset$ and sample $p \sim U[0,1]$.
  For $j = 2, ..., m$ and for $k = 1, ..., K$, do $\hat{A} = \hat{A} \cup (j,k)$ with probability p .
  $(f_n, \boldsymbol{\beta}_n) = \text{Swap}(\hat{A})$ (see Algorithm 3)
end for
Return the OLS estimates $\boldsymbol{\beta}_{\hat{n}}$ such that $\hat{n} = \arg\min_{n \in [1, N_{\text{init}}]} f_n$.

Algorithm 3 Swap the set of indices - Swap($\hat{A}$)

Given a set of indices $\hat{A}$ defining the parameters $\Delta\boldsymbol{\beta} \neq 0$, for $j = 2, ..., m$ and for $k = 1, ..., K$ **do**
Build the sets $\tilde{A}_{jk} = \hat{A} \cup (j,k)$ if $\hat{A} \cap (j,k) = \emptyset$ or the set $\tilde{A}_{jk} = \hat{A} \backslash (j,k)$ otherwise.
For each set $\tilde{A}_{jk}$, compute the OLS estimates $(\hat{\boldsymbol{\beta}}_{jk})$ and the penalized function $f_{jk} = f(\hat{\boldsymbol{\beta}}_{jk})$ (see (8)).
For the set $\hat{A}$, compute the OLS estimates $(\hat{\boldsymbol{\beta}}_{\hat{A}})$ and the penalized function $f_{\hat{A}} = f(\hat{\boldsymbol{\beta}}_{\hat{A}})$ (see (8)).
Find $(\hat{j}, \hat{k}) = \arg\min_{j,k} f_{jk}$
if $f_{\hat{j}\hat{k}} < f_{\hat{A}}$ **then**
  Return $\hat{\boldsymbol{\beta}}_{\hat{j}\hat{k}}$ and $f_{\hat{j}\hat{k}}$
else
  Return $\hat{\boldsymbol{\beta}}_{\hat{A}}$ and $f_{\hat{A}}$
end if

4 Selection of the penalty parameters and parameter uncertainties

The SELO penalty function exhibits two tuning parameters $\boldsymbol{a}$ and λ. The standard approach to fix them consists in considering a grid of values of these parameters and in selecting the parameters that maximize a (generally consistent) information criterion (e.g., Zhang, Li, and Tsai, 2010). Instead of relying on a standard information criterion and select the tuning parameters $\boldsymbol{a}$ and λ that maximize it, we consider each pair $(\boldsymbol{a}, \lambda)$ as a model to take into account the model uncertainty. For a given value of $(\boldsymbol{a}, \lambda)$, the DAEM algorithm exposed in Section 3.1 provides an estimate $\widehat{\Delta\boldsymbol{\beta}}$ of $\Delta\boldsymbol{\beta}$ which delivers an estimate of $\hat{A}$, i.e., the set of indices with $\Delta\hat{\beta}_{jk} \neq 0$ for $j \in [2, m]$ and for $k \in [1, K]$. This set tells us which covariates should be included in the linear regression and which should not. Let us denote by $\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}$ the covariates related to the first-difference estimates that are different from zero. We use the

following criterion for selecting $\boldsymbol{a}$ and λ:

$$f(\mathbf{y}|\boldsymbol{a},\lambda,\boldsymbol{\tau}) = (\frac{g_{\hat{A}}}{1+g_{\hat{A}}})^{k_{\hat{A}}/2}[\frac{g_{\hat{A}}}{1+g_{\hat{A}}}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g_{\hat{A}})}s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}]^{-\frac{T-K}{2}}, \tag{11}$$

where $s_{\tilde{\mathbf{X}}_{\tau_0}}$ stands for the residual sum of squares (RSS) from the ordinary least squares (OLS) with $\mathbf{X} = \tilde{\mathbf{X}}_{\tau_0}$ (i.e., a regression without break), $s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}$ is the RSS from the OLS with $\mathbf{X} = (\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})$, the value $k_{\hat{A}} = |\hat{A}|$ denotes the number of first-difference parameters different from zero in the model and $g_{\hat{A}}$ is a user parameter. We properly derive the criterion in Appendix B. Fernandez, Ley, and Steel (2001) show that the criterion (11) is consistent in the sense that it selects asymptotically the true subset of regressors when $g_{\hat{A}} = w(T)^{-1}$ as stated in proposition 2.

Proposition 2. *(Adaption of Fernandez, Ley, and Steel, 2001). Conditional on the true break dates, the criterion* (11) *is asymptotically maximized for the true subset of covariate A if the following conditions on the parameter $g_{\hat{A}} = w(T)^{-1}$ holds*

i) $\lim_{T\to\infty} w(T) = \infty$.

ii) $\lim_{T\to\infty} \frac{w'(T)}{w(T)} = 0$.

iii) $\lim_{T\to\infty} \frac{T}{w(T)} \in [0,\infty)$.

Proof. See Appendix C. □

Remark 3. Proposition 2 can be readily adapted when the conditioning set is a potential break date set complying with Assumption A6.

In Fernandez, Ley, and Steel (2001), they advocate for setting $g_{\hat{A}} = \min(T^{-1}, (k_{\hat{A}}+K)^{-2})$ as this prior empirically delivers good results for selecting the true covariates in standard linear regressions. However, we deviate from this benchmark prior by fixing $g_{\hat{A}} = \frac{1}{T^{\alpha}-1}$ with $\alpha = 1$ when $k_{\hat{A}} = 0$ and $\alpha = \frac{k_{\hat{A}}+\hat{m}_{\hat{A}}-1}{k_{\hat{A}}} > 1$ when $k_{\hat{A}} > 0$ in which $\hat{m}_{\hat{A}}$ denotes the number of active segments. When $\alpha > 1$, we show in Appendix C.1 that the criterion in Equation (11) asymptotically converges in probability to

$$\ln f(\mathbf{y}|\boldsymbol{a},\lambda,\boldsymbol{\tau}) \to_p -\frac{T}{2}\ln s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}} - \frac{\alpha k_{\hat{A}}}{2}\ln T. \tag{12}$$

The asymptotic value is equivalent to the Bayesian information criterion (BIC) of a linear regression model exhibiting a number of parameters of $\alpha k_{\hat{A}}$.[2] Consequently, the model penalty takes additionally into account the number of active breakpoints when $\alpha = \frac{k_{\hat{A}}+\hat{m}_{\hat{A}}-1}{k_{\hat{A}}}$. This stronger penalty works empirically well and is motivated by several CP papers advocating for stronger penalties than the BIC as it tends to overfit the number of regimes in finite sample (see, e.g., Liu, Wu, and Zidek, 1997; Zhang and Siegmund, 2007; Kim and Kim, 2016).

Interestingly, criterion (11) stands for a marginal likelihood in the Bayesian paradigm under $\epsilon \sim \mathcal{N}(0, \sigma^2 I_T)$ and the following prior,

$$\begin{aligned} f(\boldsymbol{\beta}_1) &\propto 1, \\ f(\sigma^2) &\propto \sigma^{-2}, \\ f(\Delta\boldsymbol{\beta}_{\hat{A}}|\sigma^2, \boldsymbol{\tau}) &\sim \mathcal{N}(\mathbf{0}, \sigma^2 (g_{\hat{A}} (\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})' \mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}} \tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})^{-1}) \\ f(\Delta\boldsymbol{\beta}_{\hat{A}^c}) &\sim \text{Dirac}(\mathbf{0}), \end{aligned} \tag{13}$$

where $\mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}} = I_T - \tilde{\mathbf{X}}_{\tau_0}((\tilde{\mathbf{X}}_{\tau_0})'\tilde{\mathbf{X}}_{\tau_0})^{-1}(\tilde{\mathbf{X}}_{\tau_0})'$. The prior distributions given by Equations (13) lead to simple posterior inference. The posterior distribution of the model parameters are given by, see Appendix B.1 for derivations,

$$\begin{aligned} \sigma^2|\mathbf{y}, \boldsymbol{\tau} &\sim \mathcal{IG}(\frac{T-K}{2}, \frac{\frac{g_{\hat{A}}}{1+g_{\hat{A}}} s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g_{\hat{A}})} s_{\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}}{2}), \\ \Delta\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \boldsymbol{\tau} &\sim \mathcal{N}((1+g_{\hat{A}})^{-1}[(\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})'\mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}}\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}]^{-1}(\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})'\mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}}\mathbf{y}, \frac{\sigma^2}{(1+g_{\hat{A}})}[(\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})'\mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}}\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}]^{-1}), \\ \boldsymbol{\beta}_1|\mathbf{y}, \sigma^2, \Delta\boldsymbol{\beta}, \boldsymbol{\tau} &\sim \mathcal{N}((\tilde{\mathbf{X}}'_{\tau_0}\tilde{\mathbf{X}}_{\tau_0})^{-1}\tilde{\mathbf{X}}'_{\tau_0}(\mathbf{y} - \tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}\Delta\boldsymbol{\beta}), \sigma^2(\tilde{\mathbf{X}}'_{\tau_0}\tilde{\mathbf{X}}_{\tau_0})^{-1}), \\ \Delta\boldsymbol{\beta}_{\hat{A}^c}|\mathbf{y}, \boldsymbol{\tau} &= \mathbf{0}, \end{aligned}$$

in which $\mathcal{IG}(-,-)$ denotes the Inverse-Gamma distribution. Consequently, we can go beyond selecting the best pair $(\boldsymbol{a}_p, \lambda_p)$ (i.e., the pair that maximizes the criterion (11)) and can take the uncertainty of this selection into account. Given a set of models $M_z = (\boldsymbol{a}_z, \lambda_z)$, with $z = 1, ..., Z$, we can directly assess the posterior probability of a specific model as follows

$$f(M_p|\mathbf{y}, \boldsymbol{\tau}) = \frac{f(\mathbf{y}|\boldsymbol{a}_p, \lambda_p, \boldsymbol{\tau}) f(M_p|\boldsymbol{\tau})}{\sum_{z=1}^{Z} f(\mathbf{y}|\boldsymbol{a}_z, \lambda_z, \boldsymbol{\tau}) f(M_z|\boldsymbol{\tau})}, \forall p \in [1, Z], \tag{14}$$

[2]The BIC of a linear regression model with K parameters is given by $-\frac{T}{2}\ln(\frac{^{s}\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}{T}) - \frac{K}{2}\ln T$. So the marginal likelihood criterion of Equation (11) converges to the BIC up to an additive constant (that is $\frac{T}{2}\ln T$).

where $f(M_z|\boldsymbol{\tau})$ denotes the prior probability of model M_z. In this paper, we assume uninformative prior, so $f(M_z|\boldsymbol{\tau}) = Z^{-1}$. The posterior probability can be used to account for uncertainty on the selected regressors. In fact, we have

$$\begin{aligned} f(\boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}, \sigma^2, M|\mathbf{y}, \boldsymbol{\tau}) = f(\boldsymbol{\beta}_1|\mathbf{y}, \boldsymbol{\tau}, \sigma^2, \Delta\boldsymbol{\beta}, M) f(\Delta\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\tau}, \sigma^2, M) \\ f(\sigma^2|\mathbf{y}, \boldsymbol{\tau}, M) f(M|\mathbf{y}, \boldsymbol{\tau}) \end{aligned} \tag{15}$$

It is worth emphasizing that the consistent property of the criterion (11) does not depend on the normality assumption. Only, the posterior distribution of the model parameters does. We do not see this as a limitation since one can easily extend the model with another distributional assumption and compute the posterior distribution by numerical integrations.

4.1 Prediction using Bayesian model averaging

Equation (15) shows how to take into account the uncertainty of the model parameters with respect to the selection of the SELO parameters. The Bayesian paradigm also provides a simple tool to forecast the series taking this uncertainty into account. In particular, the predictive density $f(y_{T+1:T+h}|\mathbf{y})$, for $h \geq 1$, is related to the posterior density as follows

$$\begin{aligned} f(y_{T+1:T+h}|\mathbf{y}, \boldsymbol{\tau}) &= \sum_{z=1}^{Z} \int f(y_{T+1:T+h}|\mathbf{y}, \boldsymbol{\tau}, \boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}, \sigma^2, M_z) f(\boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}, \sigma^2, M_z|\mathbf{y}, \boldsymbol{\tau}) d\boldsymbol{\beta}_1 d\Delta\boldsymbol{\beta} d\sigma^2, \\ &\approx \frac{1}{N} \sum_{i=1}^{N} f(y_{T+1:T+h}|\mathbf{y}, \boldsymbol{\tau}, \boldsymbol{\beta}_1^{(i)}, \Delta\boldsymbol{\beta}^{(i)}, (\sigma^2)^{(i)}, M^{(i)}), \end{aligned} \tag{16}$$

where $\{\boldsymbol{\beta}_1^{(i)}, \Delta\boldsymbol{\beta}^{(i)}, (\sigma^2)^{(i)}, M^{(i)}\}_{i=1}^{N}$ are independent draws from the posterior distribution (i.e., $\boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}, \sigma^2, M|\mathbf{y}, \boldsymbol{\tau}$). From (16), it is apparent that the predictive density takes the model uncertainty into account.[3] This feature should be contrasted with the standard penalized regression literature in which forecasting is performed using one unique set of parameter estimates; i.e., the estimates given by one penalty parameter selected, for instance, by cross-validation or by an information criterion.

In practice, simulations from the posterior distribution are not required for evaluating the

[3] Using the full marginal likelihood for weighting the models' predictions could raise concerns as only the last segment matters in CP processes. However, as marginal likelihood is frequently used for selecting the number of regimes in the literature and because it is also informative about the fit of the last regime, this average should give large weights to the models exhibiting a good fit at the end of the sample. Nevertheless, we could also weight the models' predictions using the predictive marginal likelihood $f(y_{t_1+1:T}|y_{1:t_1}, \boldsymbol{\tau})$ in which t_1 is a user-defined value.

predictive density. Assuming that the future covariates $\mathbf{x}_{T+1:T+h}$ are observed at time T, the predictive distribution of $y_{T+1:T+h}$ given a model M_z turns out to be a multivariate student distribution. Appendix B.2 documents the analytical expression of $f(y_{T+1:T+h}|\mathbf{y}, M_z)$. Therefore, we can efficiently take into account model uncertainty in the predictive density since Equation (16) simplifies into

$$f(y_{T+1:T+h}|\mathbf{y}, \boldsymbol{\tau}) = \sum_{z=1}^{Z} f(y_{T+1:T+h}|\mathbf{y}, \boldsymbol{\tau}, M_z) f(M_z|\mathbf{y}, \boldsymbol{\tau}). \tag{17}$$

4.2 How to choose the values of λ and $\boldsymbol{a}$

When the number of models to consider is too large to directly explore the model space using the criterion (11) (i.e., when $(m-1)K > 10$, see remark 1), we rely on the SELO penalty function to uncover the promising explanatory variables. While the asymptotic result of Proposition 1 is reassuring, it only applies for the true parameters λ and $\boldsymbol{a}$. Similar to what is generally done in the penalized regression literature, we propose to explore many values of λ and $\boldsymbol{a}$ and consider each couple as a model that would be ultimately discriminated via criterion (11). For the parameter $\boldsymbol{a}$, we use a value of $a_i = \kappa \times \text{std}(\hat{\beta}_{j1})$ for each j of the K parameters per regime where $\text{std}(\hat{\beta}_{j1})$ stands for the standard deviation of the OLS estimate $\hat{\beta}_{j1}$ when we assume no break in the linear regression (i.e., $\mathbf{X} = \mathbf{X}_{\tau_0}$). We test several values for the parameter κ, namely $\kappa \in \{0.1, 1\}$. Regarding the penalty parameter λ, we test 50 different values uniformly spaced in the interval $(0, \bar{\lambda}]$ in which $\bar{\lambda} = 2\ln T$. The penalty imposes by the upper bound $\bar{\lambda}$ is conservative enough as it is stronger than standard information criteria such as the BIC (that corresponds to a penalty of $\frac{1}{2}\ln T$) and the modified BIC.

5 Break date detection

In this Section, we present one approach to obtain a set of potential break dates. Before going into details, it is worth emphasizing that our method for detecting which parameters vary when a break occurs is independent of the segmentation detection procedure used in the first phase. To build the break date set, we could, for instance, adapt the dynamic programming method of Bai and Perron (2003) for the marginal likelihood given by Equation (11) and therefore propose our own CP detection method. We could also detect the locations of the segments using one of the standard segmentation approaches such as Bai and Perron (1998), Killick, Fearnhead, and Eckley (2012) or Korkas and Fryzlewiczv (2017). Even better, we

could apply several CP detection algorithms and discriminate between the sets of breakpoints by comparing their marginal likelihoods *once* the SELO optimization has been carried out on each set. However, as the emphasis of the paper is not on the break detection, we prefer relying on one break detection procedure, the one documented in Yau and Zhao (2016), because i) it delivers a set of potential break dates with a computational complexity of $\mathcal{O}(T\,(\log(T))^2)$ (which is faster than $\mathcal{O}(T^2)$, i.e., the complexity of the dynamic programming method of Bai and Perron (2003)) and because ii) we slightly improve their CP detection procedure. In particular, their estimated breakpoints depend on one tuning parameter, the radius h. Instead of fixing it, we use multiple values of h and we also adapt their approach to end up with a potential breakpoint set.
It is worth noting that, as the paper combines model selection and change point detection methods, our approach only requires a set of potential break dates that includes the correct break dates. By penalizing the parameter variation between two consecutive regimes, the spurious break dates are consistently deleted (see remarks 2 and 3).

5.1 Segmentation procedure

Yau and Zhao (2016) propose a likelihood ratio scan method in three steps for estimating multiple break dates in piecewise stationnary processes. They also establish the consistency of the estimated number and location fractions of the change points. We apply their three steps to detect the break dates but we modify them to reduce the computational burden and to keep at the end of the procedure a potential break date set (that could overestimate the true number of regimes). We now detail the three steps that we use to segment the data.
First step. Fix a window radius $h \in [K+1, T-K]$. For $t = h$ to $T-h$, compute the likelihood ratio scan statistic given by,

$$S_h(t) = \frac{1}{h} L_{t-h+1:t}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) + \frac{1}{h} L_{t+1:t+h}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) - \frac{1}{h} L_{t-h+1:t+h}(\hat{\boldsymbol{\beta}}, \hat{\sigma}), \tag{18}$$

where $\hat{L}_{t_1:t_2}(\hat{\boldsymbol{\beta}}, \hat{\sigma})$ denotes the maximum value of the log-likelihood of model (1) over the segment $t \in [t_1, t_2]$, assuming that $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Then, the set $\Gamma(h)$ of potential break dates is given by,

$$\Gamma(h) = \left\{ j \in \{h+h+1, \ldots, T-h\}; S_h(j) = \max_{t \in [j-h, j+h]} S_h(t) \right\}, \tag{19}$$

where $S_h(t) = 0$ for $t < h$ and $t > T-h$. As the window radius h is crucial, we differ from Yau and Zhao (2016) by using a grid of M values uniformly-spaced in the interval $[\frac{h_{\text{YZ}}}{2}, 2h_{\text{YZ}}]$

in which h_{YZ} denotes their advocated value that is $h_{\text{YZ}} = \max\left\{25, (\log(T))^2\right\}$ when $T < 800$ and $h_{\text{YZ}} = \max\left\{50, 2\,(\log(T))^2\right\}$ otherwise. So, at the end of the first step, we end up with M potential break date sets, i.e., $\Gamma(h_1), \ldots, \Gamma(h_M)$.

Second step. For every $z \in [1, M]$ and $i \in [1, m_{h_z} - 1]$ where $m_{h_z} = |\Gamma(h_z)| + 1$, we re-estimate each break date location $\tau_i^{(z)} \in \Gamma(h_z)$ as follows

$$\hat{\tau}_i^{(z)} \quad = \quad \text{argmax}_{t \in [\tau_i^{(z)} - h_z, \tau_i^{(z)} + h_z]} L_{\tau_i^{(z)} - \lfloor 1.5 h_z \rceil : t}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) + L_{t+1 : \tau_i^{(z)} + \lfloor 1.5 h_z \rceil}(\hat{\boldsymbol{\beta}}, \hat{\sigma}),$$

in which $\lfloor x \rceil$ stands for the nearest integer to x. Gathering all the new locations in the set $\hat{\Gamma}(h_z) = \{\hat{\tau}_1^{(z)}, \ldots, \hat{\tau}_{m_{h_z}}^{(z)}\}$, it is clear from Theorems 1 to 3 in Yau and Zhao (2016) that for any $j \in \{1, \ldots, m-1\}$, there exist $\hat{\tau}_i^{(z)} \in \hat{\Gamma}(h_z)$ with $i \in [1, m_{h_z} - 1]$ such that $\hat{\tau}_i^{(z)} - \tau_j = \mathcal{O}_p(1)$.

Third step. We select the best breakpoints among the M potential break date sets by minimizing the Minimum Description Length (MDL) defined by, for $z \in [1, M]$,

$$\begin{aligned} \text{MDL}(h_z) = & \ln^+(m_{h_z} - 1) + m_{h_z} \ln(T) \\ & + \sum_{j=1}^{m_{h_z}} \left(\frac{K+1}{2} \log(\hat{\tau}_j^{(z)} - \hat{\tau}_{j-1}^{(z)}) - L_{\hat{\tau}_{j-1}^{(z)} + 1 : \hat{\tau}_j^{(z)}}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) \right), \end{aligned} \tag{20}$$

where $\hat{\tau}_0 = 0$, $\hat{\tau}_{m_{h_z}} = T$ and $\{\hat{\tau}_j\}_{j=2,\ldots,m_{h_z}-1} = \hat{\Gamma}(h_z)$. In practice, we fix $M = 30$.

5.2 Break uncertainty

Given a set of break dates obtained either from the procedure described in Section 5.1 or from any other existing break detection method such as the one of Bai and Perron (1998), our method to uncover the partial structural changes can be undertaken. Let us denote by $M_* = (\boldsymbol{a}_*, \lambda_*)$ the SELO parameters maximizing the marginal likelihood criterion (11) and their corresponding break dates $J = \{\bar{\tau}_0 = 0, \bar{\tau}_1, \ldots, \bar{\tau}_{\hat{m}-1}, \bar{\tau}_{\hat{m}} = T\}$. To provide break uncertainty, we shall infer the posterior distribution of the structural breaks; i.e., $\boldsymbol{\tau} \equiv \tau_1, \ldots, \tau_{\hat{m}-1} | \mathbf{y}, M_*$. To do so, we first assume uninformative priors for the break dates using the set J. For $i = 1, ..., \hat{m} - 1$, the break parameter τ_i is driven by a Uniform distribution as follows

$$\tau_i \quad \sim \quad \mathcal{U}[\left\lfloor \frac{\bar{\tau}_{i-1} + \bar{\tau}_i}{2} \right\rfloor + \gamma, \left\lfloor \frac{\bar{\tau}_{i-1} + \bar{\tau}_i}{2} \right\rfloor - \gamma],$$

in which $\lfloor x \rfloor$ stands for the nearest integer less than or equal to x and $\gamma = (K+1)$ is a minimum duration parameter ensuring that the marginal likelihood criterion (11) can be computed for any break parameters complying with the prior distributions given by Equation (13). The posterior density is proportional to

$$\begin{aligned} f(\boldsymbol{\tau}|\mathbf{y}, M_*) &\propto f(\mathbf{y}|M_*, \boldsymbol{\tau}) f(\boldsymbol{\tau}), \\ &\propto f(\mathbf{y}|M_*, \boldsymbol{\tau}) \left(\prod_{i=1}^{\hat{m}-1} \mathbf{1}_{\{\boldsymbol{\tau}_i \in [\lfloor \frac{\bar{\tau}_{i-1}+\bar{\tau}_i}{2} \rfloor + \gamma, \lfloor \frac{\bar{\tau}_{i-1}+\bar{\tau}_i}{2} \rfloor - \gamma]\}} \right). \end{aligned} \tag{21}$$

As shown in Appendix B, the marginal likelihood $f(\mathbf{y}|M_*, \boldsymbol{\tau}) = f(\mathbf{y}|\boldsymbol{a}_*, \lambda_*, \boldsymbol{\tau})$ exhibits a closed form expression. Several solutions exist to sample the break parameters (see, e.g., Stephens, 1994; Liao, 2008). In this paper, we use the D-DREAM algorithm developed in Bauwens, Dufays, and De Backer (2011). It builds a symmetric proposal distribution inspired by the Differential Evolution optimization literature and draws from this proposal distribution are accepted or rejected through a Metropolis step in a Markov-chain Monte Carlo (MCMC) algorithm. As shown in Bauwens, Dufays, and De Backer (2011), the D-DREAM algorithm complexity is $\mathcal{O}(T)$ and leads to a rapidly mixing MCMC algorithm since the break parameters are jointly sampled from the proposal distribution. To infer the break parameters, we apply the following steps:

- Sample $R = 2m$ initial structural break vectors $\{\boldsymbol{\tau}_i\}_{i=1}^R$ from the prior distribution.
- At each MCMC iteration, for each $j = 1, ..., R$, apply the D-DREAM Metropolis move:

  1. Propose a new draw of the break parameter as follows

  $$\hat{\boldsymbol{\tau}}_j = \boldsymbol{\tau}_j + \left\lfloor \gamma(\delta, m)(\sum_{g=1}^{\delta} \boldsymbol{\tau}_{r_1(g)} - \sum_{h=1}^{\delta} \boldsymbol{\tau}_{r_2(h)}) + \xi \right\rceil, \tag{22}$$

  with $\xi \sim \mathcal{N}(0, (0.0001)I)$ and $\forall g, h = 1, 2, ..., \delta$, $j \neq r_1(g)$, $r_2(h)$; $r_1(.)$ and $r_2(.)$ stand for random integers uniformly distributed on the support $[1, R]$. We set $\gamma(\delta, m) = \frac{2.38}{\sqrt{2\delta m}}$ and $\delta \sim \mathcal{U}[1, 3]$.
  2. Accept the proposal $\hat{\boldsymbol{\tau}}_j$ according to the probability

  $$\alpha(\boldsymbol{\tau}_j, \hat{\boldsymbol{\tau}}_j) = \min\{\frac{f(\mathbf{y}|M_*, \hat{\boldsymbol{\tau}}_j)}{f(\mathbf{y}|M_*, \boldsymbol{\tau}_j)}, 1\}.$$

In practice, we set the number of MCMC iterations to 1000 and start collecting the draws

after round$[\frac{M}{2}]$ MCMC iterations.

6 Monte Carlo study

In this Section, we document a Monte Carlo study to assess the accuracy of the SELO approach. We rely on nine different data generating processes (DGPs) that are documented in Table 1. For each DGP, we simulate 1000 series with a sample size equal to $T = 1024$ and we investigate i) the performance of detecting the break dates using the approach in Section 5.1 and ii) the performance of the SELO method for detecting which parameter truly varies when a break occurs. The nine DGPs differ in their mean parameter specifications. For each of them, we study the SELO performance when the innovation is either homoskedastic or driven by a GARCH process.

Regarding the DGPs, the first six DGPs are piecewise stationary AR models directly taken from Yau and Zhao (2016) while the others cover situations with exogenous explanatory variables. DGP A and E do not exhibit any breakpoint. They aim at showing the performance of the SELO approach when only spurious break dates are detected. DGPs B and C are weakly persistent piecewise stationary AR models exhibiting three regimes. Simulated series from DGP D experience a break after 50 observations. This DGP should highlight the performance in a short regime context. DGPs E and F are highly persistent piecewise stationary AR models but DGP F differs by exhibiting breaks in the mean parameters. Eventually, DGPs G, H and I include exogenous variables. While DGPs G only exhibits exogeneous regressors, DGPs H and I stand for ARX processes by mixing the parameters of the DGPs B and G.

Table 1: **Data Generating Processes of sample size amounting to $T = 1024$.**
This Table summarizes the DGPs from which 1000 series are simulated for the Monte Carlo study. The variables V and W stand for exogenous variables such that, $V_t \sim \mathcal{N}(0, 3^2)$ and $W_t \sim \mathcal{N}(0, 4^2)$. For instance, DGP B is an AR(2) model that exhibits two breakpoints at $t = 512$ and $t = 768$. The true values of the first AR term for the first two regimes are equal to 0.9 and 1.69, respectively. The dynamic of the variance is either homoskedastic ('Constant') or heteroskedastic ('GARCH').

	DGP A	**DGP B**	**DGP C**
Breaks	-	[512, 768]	[400, 612]
Intercept	[0]	[0, 0, 0]	[0, 0, 0]
AR_1	[- 0.7]	[0.9, 1.69, 1.32]	[0.4, - 0.6, 0.5]
AR_2	-	[0, - 0.81, - 0.81]	-
	DGP D	**DGP E**	**DGP F**
Breaks	[50]	-	[400, 750]
Intercept	[0, 0]	[0]	[0, 0, 0]
AR_1	[0.75, - 0.5]	[0.999]	[1.399, 0.999, 0.699]
AR_2	-	-	[- 0.4, 0, 0.3]
	DGP G	**DGP H**	**DGP I**
Breaks	[400, 750]	[400,750]	[512, 768]
Intercept	[1, 0, 0]	[0, 0, 0]	[0,0,0]
AR_1	-	[0.9, 1.69, 1.32]	[0.9, 1.69, 1.32]
AR_2	-	[0, -0.81, -0.81]	[0, -0.81, -0.81]
V	[1.5, 0.9, 2.2]	[1.5, 0.9, 2.2]	[1.5, 0.9, 2.2]
W	[- 0.6, - 0.6, - 1]	[- 0.6, - 0.6, - 1]	[- 0.6, - 0.6, - 1]
	Dynamic of the variance of $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$		
Constant	$\sigma_t^2 = 1$, $\forall t \in [1, T]$		
GARCH	$\sigma_t^2 = 0.05 + 0.05\epsilon_{t-1}^2 + 0.9\sigma_{t-1}^2$, $\forall t \in [1, T]$ and $\sigma_0^2 = \frac{0.05}{1-0.95} = 1$		

Table 2 documents the percentage of detecting a number of regimes per model parameter over the 1000 simulated series per DGP for the SELO method. Overall, the detection rates of identifying the true number of regimes per parameter are excellent and besides DGP F, they are at least equal to 86.4%. Interestingly, this detection rate does not deteriorate when the innovation is driven by a GARCH process. The worst detection rates arise for the DGP F. Even though this DGP is highly persistent with an autocorrelation structure that barely varies over time, the SELO method correctly identifies that the intercept does not experience abrupt switches 69.6% of the times. Note that the potential breakpoint sets for this DGP poorly identify the true breakpoints since only 25.5% of the sets exhibit at least one potential CP

Table 2: **Break estimates : SELO approach.**
Based on 1000 replications, this Table presents several metrics for assessing the performance of the SELO method on DGPs detailed in Table 1. **Number of regimes** is the rate of detecting a specific number of regimes per model parameter. Bold values correspond to the true number of regimes. **Break** documents the rate of having at least one breakpoint in the potential CP set located in the neighborhood of 50 observations of the true breakpoints. We use '—' when the DGP exhibits no breakpoint. **Exact** denotes the rate of detecting the true number of breakpoints for all the model parameters with a posterior probability of at least 10%.

		Constant Variance								**GARCH Variance**							
		Number of regimes						Break	Exact	**Number of regimes**						Break	Exact
DGP		1	2	3	4	5	6			1	2	3	4	5	6		
A	Intercept	**99.4**	0.6	0	0	0	0	—	99.9	**99.2**	0.8	0	0	0	0	—	99.2
	AR1	**99.5**	0.5	0	0	0	0			**99.4**	0.6	0	0	0	0		
B	Intercept	**98.6**	1.4	0	0	0	0			**97.3**	2.7	0	0	0	0		
	AR1	0	0	**100**	0	0	0	100	99.7	0	0.2	**99.4**	0.4	0	0	99.3	99.5
	AR2	0	**98.8**	1.2	0	0	0			0	**98.3**	1.7	0	0	0		
C	Intercept	**97.9**	2	0.1	0	0	0	99.8	99.7	**97.6**	2.4	0	0	0	0	99.8	99.1
	AR1	0	0	**100**	0	0	0			0	0	**99.7**	0.3	0	0		
D	Intercept	**97.4**	2.6	0	0	0	0	99.8	99.5	**97.6**	2.2	0.2	0	0	0	99.7	99.1
	AR1	0.1	**99.4**	0.5	0	0	0			0.2	**99.3**	0.4	0.1	0	0		
E	Intercept	**86.4**	12.4	1.2	0	0	0	—	94.6	**84.8**	12.5	2.6	0.1	0	0	—	91.5
	AR1	**93.7**	6	0.3	0	0	0			**91**	8.1	0.5	0.3	0.1	0		
F	Intercept	**69.6**	23.7	6.5	0.1	0.1	0			**65.1**	27.9	6.5	0.5	0	0		
	AR1	0	68.3	**31.4**	0.1	0.2	0	25.5	23.5	0	69.3	**29.7**	0.9	0.1	0	22.4	22.5
	AR2	0	71.4	**28.4**	0.2	0	0			0	73.2	**26.4**	0.4	0	0		
G	Intercept	0	**99.3**	0.7	0	0	0			0	**99.2**	0.8	0	0	0		
	V	0	0	**99.8**	0.2	0	0	100	99.8	0	0	**99.7**	0.3	0	0	100	99.8
	W	0	**99.2**	0.8	0	0	0			0	**99**	0.9	0.1	0	0		
H	Intercept	**88.9**	11	0.1	0	0	0			**92.9**	7	0.1	0	0	0		
	AR1	0	0	**92.7**	7.3	0	0			0	0	**94.8**	5.2	0	0		
	AR2	0	**92.6**	7.4	0	0	0	100	83.1	0	**94.2**	5.8	0	0	0	100	86.9
	V	0	0	**87.7**	12.3	0	0			0	0	**89.7**	10.3	0	0		
	W	0	**88**	12	0	0	0			0	**90.3**	9.7	0	0	0		
I	Intercept	**91.7**	8.3	0	0	0	0			**91**	8.9	0.1	0	0	0		
	AR1	0	0	**94.3**	5.7	0	0			0	0	**95**	4.9	0.1	0		
	AR2	0	**94.6**	5.4	0	0	0	100	85.7	0	**94.9**	5	0.1	0	0	100	85
	V	0	0	**89.8**	10.2	0	0			0	0	**89.4**	10.6	0	0		
	W	0	**88.7**	11.1	0.2	0	0			0	**90**	10	0	0	0		

close to every true breakpoints. Therefore, the SELO detection rate could hardly exceed this bound. As exemplified by DGPs G, H and I, the detection rates of the SELO method remain excellent when exogenous variables kick in even in the presence of heteroscedasticity. The

Table also documents the rate of detecting the true model (i.e. jointly the correct number of regimes) with a posterior probability of at least 10%. For all the DGPs but DGP F, the correct detection amounts to at least 83.1% and 85% for the constant and the GARCH innovation dynamics, respectively. These excellent results highlight that model uncertainty should be taken into account since several models often exhibit high posterior probabilities.

We end this simulation section with a "big data" example motivated by the fact that when the number of explanatory variables is large, the current Bayesian alternatives do not work (see Giordani and Kohn, 2008; Eo, 2016; Dufays and Rombouts, 2018). To do so, we propose the DGP J that is specified by 100 explanatory variables and one change point as follows:

DGP J: piecewise linear model with big data

$$Y_t = \begin{cases} \mathbf{x}_t'\boldsymbol{\beta_1} + \varepsilon_t & \text{if } 1 \leq t \leq 499, \\ \mathbf{x}_t'\boldsymbol{\beta_2} + \varepsilon_t & \text{if } 500 \leq t \leq T, \end{cases}$$

where $T = 1024$, $\forall\ t \in [1, T]$ and for $i = 1, ..., 100$, $\mathbf{x}_{t,i} \sim \mathcal{N}(0, 1)$ and $\varepsilon_t \sim \mathcal{N}(0, 1)$. The parameter values of $\boldsymbol{\beta_1}$ are uniformly and randomly set to -1 or 1. In the second regime, the parameter values of $\boldsymbol{\beta_2}$ are equal to $\boldsymbol{\beta_1}$ except for 10 of them randomly chosen that are set to the opposite value (i.e. $-\boldsymbol{\beta}_1$). Thus, 10 parameters of DGP J does experience a break at observation 500.
We simulate 100 series from DGP J to assess the SELO performance in detecting which parameters experience a breakpoint. For every simulation, the selective segmentation approach identifies 10 parameters that experience one breakpoint in the sample while the others remain constant. In addition, the exact model specification was always among the specification exhibiting a posterior probability of at least 10%.

7 Empirical application

We illustrate the selective segmentation method with 14 monthly Credit Suisse HF indices spanning from March 1994 to March 2016. These indices are the weighted average of HF returns following specific trading strategies. Fung and Hsieh (2004) suggested a risk-based approach to model HF returns and identified seven factors on which HF strategies are generally exposed (see also Fung and Hsieh, 2001). Table 3 documents the fourteen strategies on which we focus as well as the seven factors of Fung and Hsieh (2004).

Table 3: Description of the HF returns and the risk factors.

Credit Suisse Hedge fund indices		Factors of Fung and Hsieh (2004)	
HFI	Hedge Fund Index	PMKT	Market factor (S&P 500)
CNV	Convertible Arbitrage	SMB	Small firm minus big firm
DSB	Dedicated Short Bias	TERM	Change in 10-year treasury yields
EME	Emerging Markets	DEF	Change in the yield spread of
EMN	Equity Market Neutral		10-year treasury and Moody's Baa bonds
EDR	Event Driven	PTFSBD	Lookback options on Bonds
EDD	Event Driven Distressed	PTFSFX	Lookback options on currencies
EDM	Event Driven Multi-Strategy	PTFSCOM	Lookback options on commodities
EDRA	Event Driven Risk Arbitrage		
FIA	Fixed Income Arbitrage		
GMA	Global Macro		
LES	Long/Short Equity		
MFU	Managed Futures		
MUS	Multi-Strategy		

It is well acknowledged in the financial literature that HF strategies (or trading techniques) are time-varying. Their changing risk exposures are directly related to market events and economic fluctuations (see, e.g., Agarwal and Naik (2004), Fung, Hsieh, Naik, and Ramodarai (2008) or Patton, Ramodarai, and Streatfield (2015) among others). Hedge fund time-varying risk dynamics has important implications for performance appraisal. As pointed out by Mitchell and Pulvino (2001), the changes can be in response to arbitrage opportunities. The cycles of mergers and acquisitions in the 1990s and the 2000s and the corresponding level of risk arbitrage led by HF are illustrations of these changing dynamics. In standard linear asset pricing models, the intercept and risk factor loadings are not constant but time-varying. Moreover, HF returns exhibit significant non-linearities. Therefore, there is a need of dynamic models able to capture non-linearities and changes in risk exposures.

Following Meligkotsidou and Vrontos (2008) we suggest the use of CP risk factor models. This class of models is suited for studying the changes in risk exposures and their time-varying parameters. However, instead of directly focusing on the seven factors, we take a slightly different approach since we additionally take into account autocorrelations of the returns.[4] To do so, we first look at the best autoregressive model that fits the returns. In particular, for each HF returns, we estimate ARX(q) models with q ranging from 0 to 4 and in which the explanatory variables are the seven factors (and an intercept) and we select the best AR order using the Bayesian information criterion (BIC). Table 4 documents the best order for

[4] As reported by Getmansky, Lo, and Makarov (2004), the analysis of serial dependence of returns is a reasonable way of assessing the liquidity of hedge fund investments.

each strategy.

Table 4: **Order of the optimal ARX-model for each HF strategy.**
The optimal AR order is chosen by maximizing the Bayesian information criterion over the whole sample. When looking for the best autoregressive lag order, the explanatory variables include the seven factors and an intercept.

Strat.	HFI	CNV	DSB	EME	EMN	EDR	EDD
Lag order	0	1	1	1	0	1	2
Strat.	EDM	EDRA	FIA	GMA	LES	MFU	MUS
Lag order	1	1	1	0	1	0	0

As reported by Fung and Hsieh (2004), composites obtained from the individual funds may be contaminated with severe survivorship, selection and instant history biases. Therefore, to avoid these problems, we use the Credit Suisse indices that provide full transparency about their constituents.

Section 7.1 discusses in-sample results of our selective segmentation method and we compare them to those of standard CP models and time-varying parameter models. We then illustrate the difference of our approach with the CP method of Meligkotsidou and Vrontos (2008) in Section 7.2. Section 7.3 documents a forecasting exercise in which we assess the predictive performance of the selective segmentation approach with respect to flexible alternatives. Importantly, all the subsequent results include the optimal AR order documented in Table 4 as additional explanatory variables.

7.1 Hedge funds strategies evolve over time

Fung and Hsieh (2004) focus on linear models. However, as the period covers critical events such as the Long Term Capital Management (LTCM) collapse, the dot-com crisis and the financial crisis, one could argue that CP models are more appropriate. In this Section, we focus on two specific indices, namely the Hedge Fund Index (HFI) and the HF returns that are applying a Fixed-Income Arbitrage (FIA) strategy. Results for all the other returns are available upon request.

Tables 5 and 6 show how the selective segmentation method can improve the interpretation of CP models. The Tables document how the results evolve from a standard linear risk model to a selective segmentation model passing by a standard CP process. As expected, for the two HF returns, ignoring breakpoints can be misleading as the CP results emphasize that they modify the risk exposition of the returns. Also, although one can study in details the results of the standard CP model, the selective segmentation model offers a straightforward

picture of the relevant risk factors and how the risk exposition evolves. It also estimates more accurately the parameters that do not change when a break occurs. As the CP model detects three breakpoints for the HFI and six abrupt changes for the FIA strategy, the number of models to consider amounts to 2^{24} and 2^{54} respectively. Our selective segmentation strategy explores these large model spaces in less than several minutes on a standard laptop. Let us now discuss in more details the results of the two returns.

Hedge Fund Index

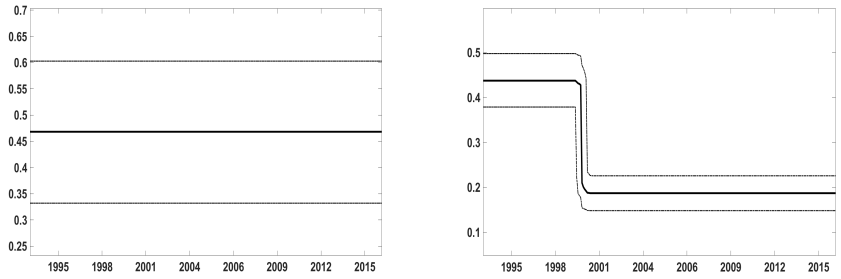
As documented in Table 5, the CP model with breakpoints determined by the approach in Section 5.1 finds four regimes (hereafter CP-YZ). The relevant breakpoints occur in April 2000, in April 2005 and March 2006. Interestingly, these dates coincide with the dot-com crash at the end of March 2000 and when the U.S. housing market bubble reached new heights. It is well acknowledged in the financial literature that the end of the dot-com bubble had important consequences for financial markets in the early 2000s and until the Global Financial Crisis (2008). While all the parameters change for the CP model, the selective segmentation mainly identifies that the factors related to the breaks are the market factor (PMKT), the credit spread factor (TERM) and the default risk factor (DEF). Moreover, it discards a spurious break occurring in March 2006 making the model even more parsimonious. We can notice that the market factor decreases from 0.44 during the first period to 0.19 during the second period. HFI is indeed more conservative during the 2000s and less correlated with the financial markets. We observe the same trend for the credit risk factor increasing from -13.00 to -2.78. These results are consistent with the important variations of interest rates during these two sub-periods and the important increase of credit risk in the 2000s. As a final note, credible intervals of the breakpoints are narrow which indicates a sharp change in the risk exposition at these periods (see also Figure 2 below).

Table 5: **Hedge Fund Index: linear, CP and selective segmentation regression models.**
The Table details the parameter estimates of the linear model, of the CP model and of the selective segmentation process with HFI returns as the dependent variable. Parentheses and brackets indicate standard deviations and 90% credible intervals, respectively. A cell filled with '—' indicates that the parameter does not vary over the related period. The posterior probability of the selective segmentation model amounts to 60%.

Period	Int.	PMKT	SMB	TERM	DEF	PTFSBD	PTFSFX	PTFSCOM
	Standard linear risk model							
03-1994 to 03-2016	0.50	0.24	0.08	-0.96	-3.26	-0.01	0.02	0.01
	(0.09)	(0.02)	(0.03)	(0.46)	(0.59)	(0.01)	(0.01)	(0.01)
	CP-YZ risk model							
03-1994 to 04-2000	0.78	0.42	0.13	-2.93	-12.89	-0.02	0.02	0.03
	(0.22)	(0.05)	(0.05)	(1.29)	(2.59)	(0.01)	(0.01)	(0.02)
05-2000 to 04-2005	0.50	0.15	0.08	-2.02	-1.71	-0.01	0.02	0.01
	(0.25)	(0.05)	(0.08)	(1.24)	(2.42)	(0.02)	(0.01)	(0.02)
05-2005 to 03-2006	0.78	0.92	-0.38	-3.71	15.40	-0.04	-0.15	0.04
	(0.83)	(1.10)	(0.85)	(8.30)	(28.36)	(0.08)	(0.20)	(0.10)
04-2006 to 03-2016	0.25	0.25	-0.09	0.34	-2.46	-0.00	0.01	0.00
	(0.17)	(0.05)	(0.08)	(0.88)	(0.84)	(0.01)	(0.01)	(0.01)
	Selective segmentation risk model (60%)							
03-1994 to 04-2000	0.47	0.44	0.07	-2.76	-13.00	-0.01	0.01	0.01
[12-1999 08-2000]	(0.08)	(0.04)	(0.03)	(0.55)	(1.53)	(0.01)	(0.00)	(0.01)
05-2000 to 04-2005	—	0.19	—	—	-2.78	—	—	—
[07-2003 07-2005]	—	(0.02)	—	—	(0.55)	—	—	—
05-2005 to 03-2016	—	—	—	0.23	—	—	—	—
	—	—	—	(0.60)	—	—	—	—

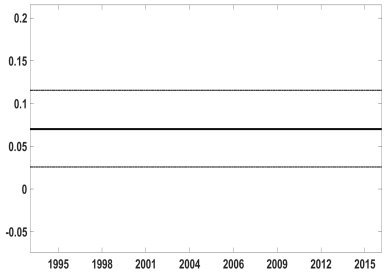
Figures 2 and 3 show the posterior medians over time and their corresponding credible intervals of the mean parameters given by our method (see Section 5.2 for the related Bayesian model and how the breakpoints are integrated out) and the time-varying parameter (TVP) model (see Appendix D for the model specification). As with the CP model, one can easier interpret the time-varying dynamics of the parameters given by the selective segmentation method than those of the TVP model. For instance, while the exposition to the default factor seems fixed over the sample due to the smooth transition of the parameter, it is clear that the exposition is changing when we look at the selective segmentation results. Regarding the market factor, we also observe with the TVP model that the exposition seems different before

and after the dot-com crash but the credible intervals are too wide to confirm the statement.

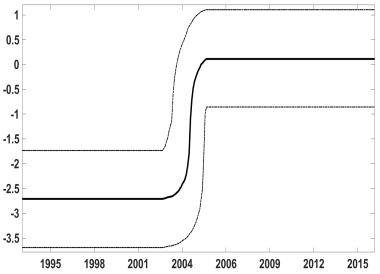


(a) Intercept

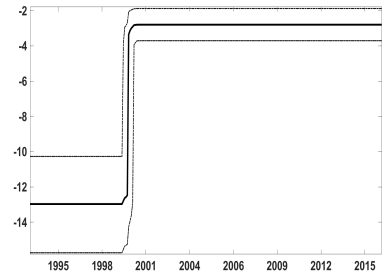
(b) PMKT



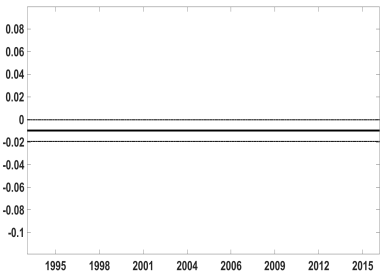
(c) SMB



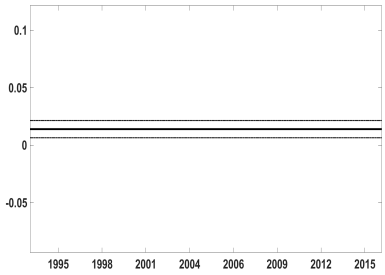
(d) TERM



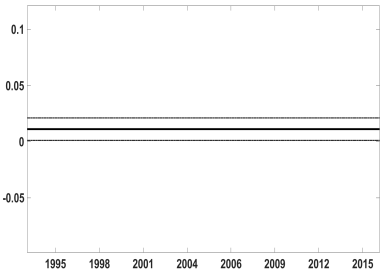
(e) DEF



(f) PTFSBD



(g) PTFSFX



(h) PTFSCOM

Figure 2: **HFI returns - Selective segmentation model.** Posterior medians and the 90% credible intervals of the model parameters over time taking into account break uncertainty as presented in Section 5.2.

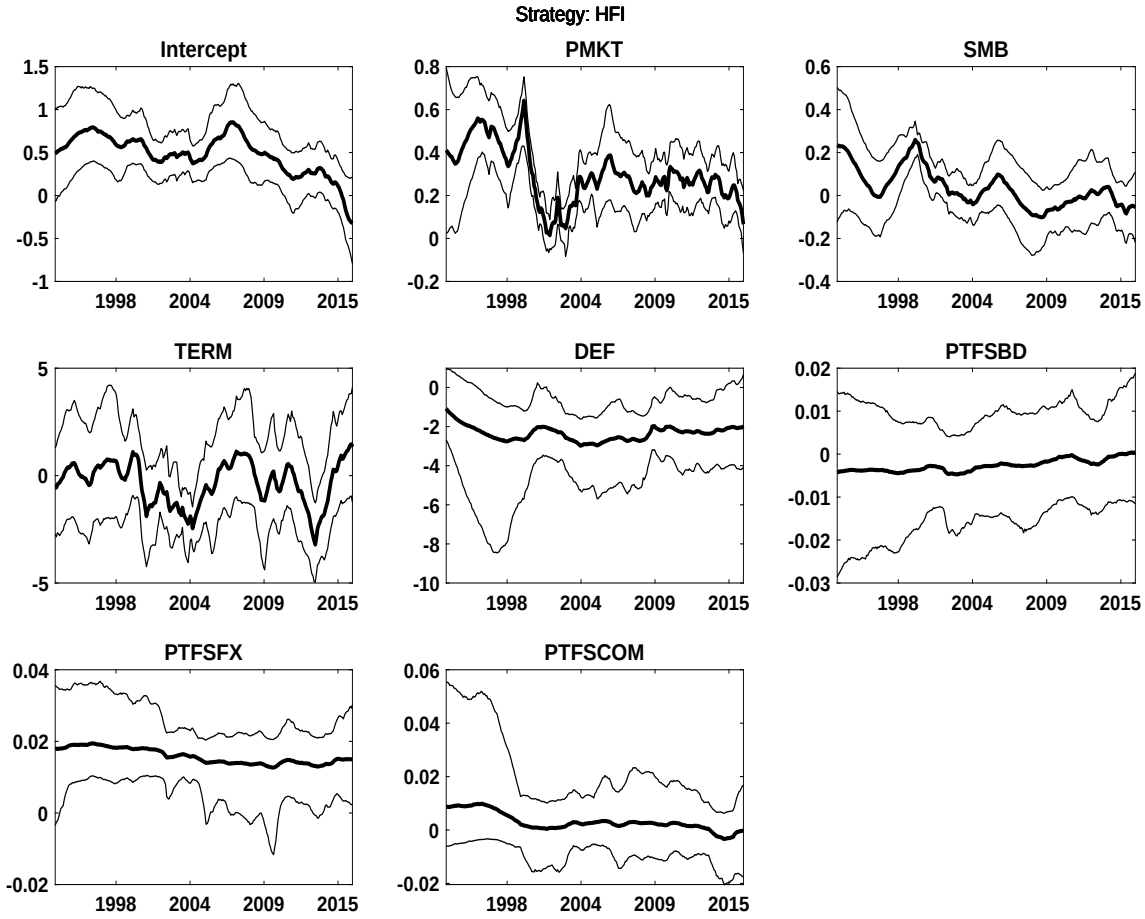


Figure 3: **HFI returns - TVP model.** 90% credible intervals of the model parameters over time in relation with the posterior median, reported in bold black line.

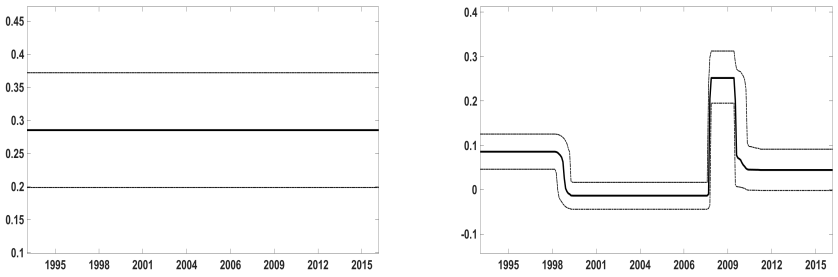
Fixed Income Arbitrage (FIA) strategy

The FIA strategy is based on the exploitation of inefficiencies in the pricing of bonds and interest rate derivatives (including futures, options, swaps and also mortgage back securities). It was very appreciated among hedge fund managers until the collapse of the LTCM fund in September 1998. After this incident, a change of behavior among managers has been observed for this strategy on financial markets.

The results of the FIA returns from a standard linear regression to the selective segmentation model are documented in Table 6. Focusing on the latter model, a change point is detected in August 1999 and three other breakpoints capture the financial crisis. The selective segmentation specification highlights the role played by the market factor (i.e., PMKT, for the first, second, fourth and fifth period with estimates of: 0.09, -0.01, 0.24 and 0.04, respectively), the variation of the size effect, the default risk factor and two trend following factors. In addition, the credit spread factor, TERM is significant and constant over time. The currency trend following factor, PTFSFX, captures the changes in the monetary policy and is significant over the whole period. The selected specification shows that the exposition to this factor dramatically changes when the global financial crisis dawned. We also observe a variation of the size effect, SMB, with a peak during the global financial crisis and an estimate -0.85, and of the commodity trend following factor, PTFSCOM since the first semester 2007. Figures 4 and 5 document the time-varying dynamics of the parameters given the selective segmentation method and the TVP model, respectively. From the TVP dynamics, the relevant factors do not easily pop up.

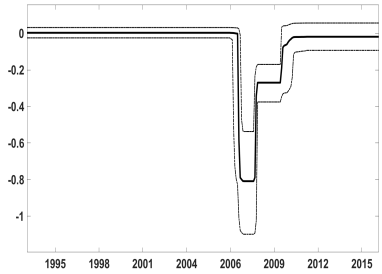
Table 6: **Fixed Income Arbitrage: linear, CP and selective segmentation risk models.**
The Table details the parameter estimates of the linear model, of the CP model and of the selective segmentation process with FIA returns as the dependent variable. Parentheses indicate standard deviations and brackets [-] document the 95% credible intervals of the breakpoints that are computed from the Bayesian model given in Section 5.2. A cell filled with '—' indicates that the parameter does not vary over the related period. The posterior probability of the selective segmentation model amounts to 63%.

Period	Int.	AR1	PMKT	SMB	TERM	DEF	PTFSBD	PTFSFX	PTFSCOM
	Standard linear risk model								
03-1994 to 03-2016	0.31	0.31	0.08	0.01	-1.39	-3.95	-0.01	-0.00	-0.01
	(0.09)	(0.05)	(0.02)	(0.03)	(0.43)	(0.64)	(0.01)	(0.00)	(0.01)
	CP-YZ risk model								
03-1994 to 08-1999	0.40	0.42	0.09	0.11	-2.08	-6.74	-0.01	-0.00	0.01
	(0.15)	(0.09)	(0.03)	(0.04)	(0.79)	(1.91)	(0.01)	(0.01)	(0.01)
09-1999 to 04-2002	0.59	0.56	0.04	0.03	-0.11	0.24	0.01	0.04	-0.01
	(0.26)	(0.15)	(0.03)	(0.04)	(1.16)	(2.05)	(0.01)	(0.01)	(0.01)
05-2002 to 04-2003	0.76	0.30	-0.07	0.19	0.85	-1.68	-0.02	-0.02	-0.01
	(1.08)	(0.39)	(0.14)	(0.13)	(7.21)	(2.77)	(0.09)	(0.02)	(0.07)
05-2003 to 06-2007	0.09	0.55	-0.00	0.12	-1.46	-2.07	-0.00	0.01	-0.00
	(0.17)	(0.14)	(0.07)	(0.07)	(0.79)	(2.21)	(0.01)	(0.01)	(0.01)
07-2007 to 07-2008	0.22	0.63	0.11	-0.90	1.46	-1.89	0.01	-0.15	0.05
	(0.41)	(0.27)	(0.10)	(0.34)	(3.12)	(3.58)	(0.04)	(0.05)	(0.02)
08-2008 to 10-2010	0.75	-0.28	0.30	-0.37	-2.61	-7.71	-0.03	0.02	-0.11
	(0.21)	(0.08)	(0.04)	(0.08)	(0.91)	(0.88)	(0.01)	(0.01)	(0.02)
11-2010 to 03-2016	0.07	0.12	0.10	-0.05	0.30	-2.31	0.00	-0.00	-0.01
	(0.13)	(0.12)	(0.04)	(0.06)	(0.85)	(1.32)	(0.01)	(0.01)	(0.01)
	Selective segmentation risk model (63%)								
03-1994 to 08-1999	0.28	0.30	0.09	0.00	-1.40	-10.48	-0.00	-0.00	0.01
[12-1998 10-1999]	(0.05)	(0.04)	(0.02)	(0.02)	(0.26)	(1.06)	(0.00)	(0.00)	(0.00)
09-1999 to 06-2007	—	—	-0.01	—	—	-3.09	—	—	—
[01-2007 07-2007]	—	—	(0.02)	—	—	(0.36)	—	—	—
07-2007 to 07-2008	—	—	—	-0.85	—	—	—	-0.17	0.05
[05-2008 07-2008]	—	—	—	(0.16)	—	—	—	(0.02)	(0.02)
08-2008 to 10-2010	—	—	0.24	-0.27	—	—	—	-0.00	-0.06
[01-2010 10-2010]	—	—	(0.03)	(0.06)	—	—	—	(0.00)	(0.01)
11-2010 to 03-2016	—	—	0.04	-0.02	—	—	—	—	0.00
	—	—	(0.03)	(0.05)	—	—	—	—	(0.01)

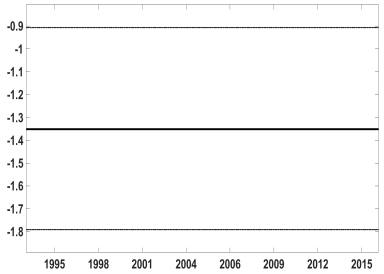


(a) Intercept

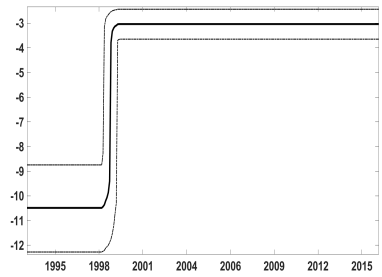
(b) PMKT



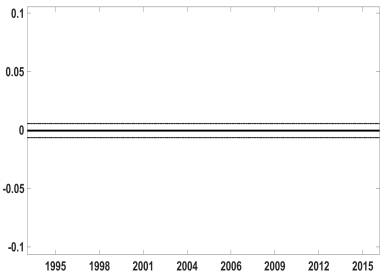
(c) SMB



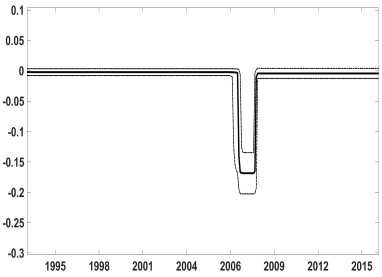
(d) TERM



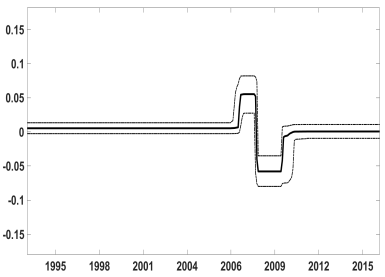
(e) DEF



(f) PTFSBD



(g) PTFSFX



(h) PTFSCOM

Figure 4: **FIA returns - Selective segmentation model.** Posterior medians and the 90% credible intervals of the model parameters over time taking into account break uncertainty as presented in Section 5.2.

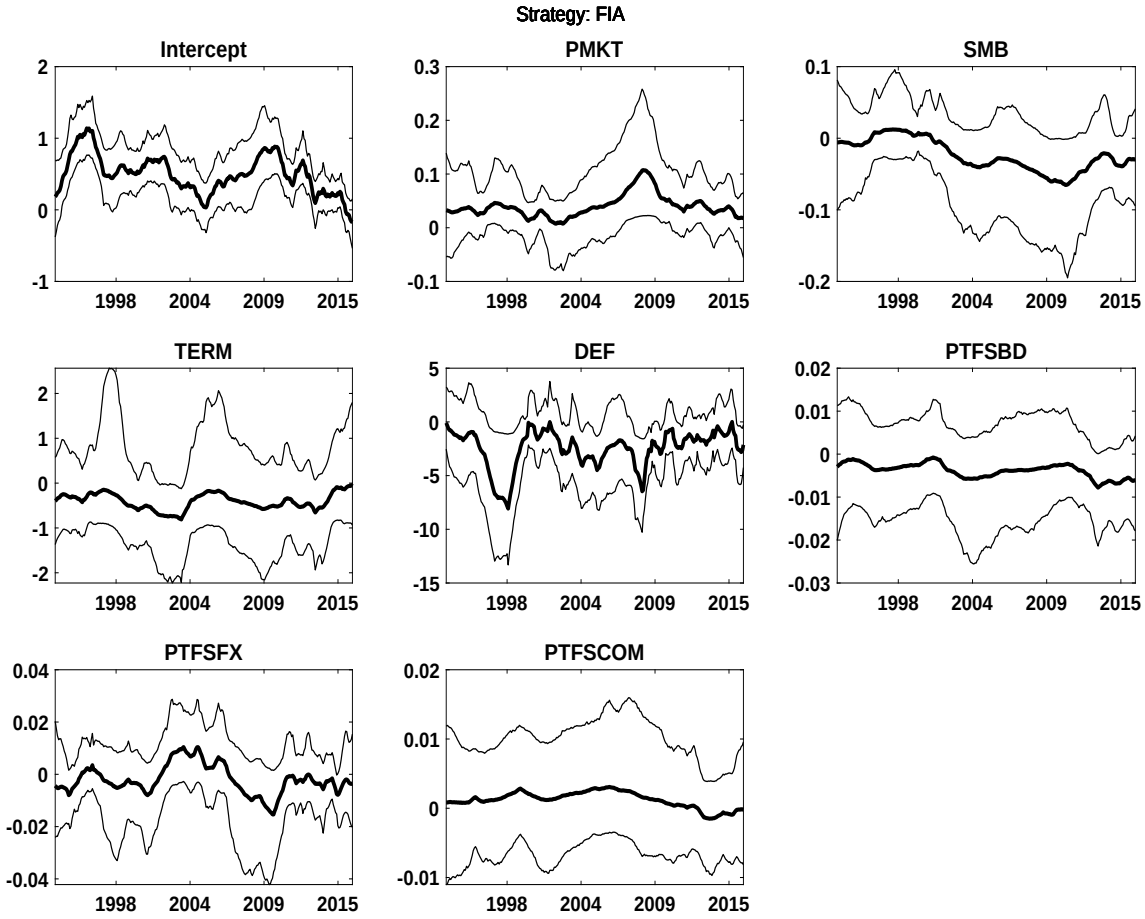


Figure 5: **FIA returns - TVP model.** 90% credible intervals of the model parameters over time in relation with the posterior median, reported in bold black line.

7.2 Comparison with advanced CP models

We now compare our results with those of Meligkotsidou and Vrontos (2008). Meligkotsidou and Vrontos (2008) rely on CP models to capture the risk exposition of HF returns over time. In particular, they consider the 128 distinct combinations of the seven risk factors and for each of them, they estimate a CP model exhibiting several numbers of segments m (from one to ten). Eventually, they use the marginal likelihood to select the best model among the set of $m \times 2^K$ estimated processes (i.e., 1280 models since $m = 10$ and $K = 7$). Their approach consists therefore in first selecting the relevant factors and then, in investigating if the exposition to them is time-varying.

There is a striking difference with our approach since, for each breakpoint, our procedure can detect what are the time-varying factors. In fact, our approach discriminates between $2^{m \times K}$ models; a number of models that exponentially increases with the amount of breaks. Note that we could also search for the best regressors to include by considering all the 128 distinct combinations of the seven factors. In such a case, the number of models to consider would reach $2^{(m+1)\times K}$.

We reproduce the results of Meligkotsidou and Vrontos (2008) on our data by additionally taking the autocorrelation structure into account. Fixing the AR order q to the value given in Table 4, for each possible combination of the factors, we estimate CP-ARX(q) models with different numbers of breaks (ranging from 1 to 10) by (globally) minimizing the MDL criterion. Then, we report the combination of factors exhibiting the best MDL value. Hereafter, we denote this model by CP-MV.[5]

Table 7 documents the factors of the CP-MV model for the two HF strategies. It also reports the factors which exhibit significant parameter estimates at least at one period for the TVP model and the selective segmentation approach. For the HFI, the selected factors by the TVP and the selective segmentation methods are identical and they only differ with respect to the selection of PTFSCOM for the FIA strategy. The CP-MV approach does not select the PTFSX and the PTFSCOM factors for the FIA strategy and the SMB factor for the HFI. Regarding the HFI strategy, it is surprising that SMB, the spread between small-cap and large-cap stock returns, is not selected as a significant factor. As pointed out by Fung

[5]Our approach is slightly different as the one used in Meligkotsidou and Vrontos (2008) since we minimize the MDL criterion instead of maximizing the marginal likelihood of a Bayesian CP model for finding the best combination of the factors and the breakpoints. This is motivated by the fact that the MDL criterion consistently selects the true number of regimes while there is no equivalent proof for the marginal likelihood used in Meligkotsidou and Vrontos (2008). In addition, Ardia, Dufays, and Ordas (2019) show that the MDL criterion is equal to minus the marginal log-likelihood of a CP Bayesian model with particular g-prior distributions. So, our approach can be understood as the method of Meligkotsidou and Vrontos (2008) with different hyper-parameters. We globally minimize the MDL criterion using the dynamic programming of Bai and Perron (2003).

and Hsieh (2004), it is the second most important equity risk factor after PMKT for many hedge fund strategies. The analysis of the single strategy hedge fund index, i.e. Fixed Income Arbitrage (FIA), also highlights some differences as PMKT and TERM are not among the selected factors by the CP-MV process. It may be surprising since the market premium is (almost) always used in linear asset pricing models (even for the analysis of hedge fund returns). We may also note that the look-back straddle on commodities, PTFSCOM, designed to capture non-linearities especially during changes in international economic policies is not selected whereas the phenomenon is observed just after the global financial crisis.

Table 7: **HFI and FIA strategies: Selected factors given several time-varying parameter models.**
Selected factors by the TVP, the selective segmentation process and the CP-MV model of Meligkotsidou and Vrontos (2008). The factors of the latter process are chosen by minimizing the MDL criterion while for the TVP and the selective segmentation model, a factor is selected if its related parameter estimate is significant at least at one period over the sample.

	HFI			FIA		
	TVP	Sel. Seg.	CP-MV	TVP	Sel. Seg.	CP-MV
PMKT	√	√	√	√	√	
SMB	√	√				
TERM	√	√	√	√	√	
DEF	√	√	√	√	√	√
PTFSBD						
PTFSFX	√	√	√	√	√	√
PTFSCOM					√	

Table 7 does not inform on the dynamic of the selected factors by the CP-MV process. Although the preferred specification of the CP-MV model does not include all the factors, the risk exposure of the HF strategies is still abruptly changing over time. Regarding the HFI, four breakpoints are detected and two of them occur in December 2000 and in April 2005. As documented in Table 8, when we apply the selective segmentation approach on top of the four breakpoints, only these two breakpoints are relevant as the mean parameters in the other regimes remain constant. This result is in line with those found by our approach when all the risk factors are used as reported in Table 5. We believe that the extra breakpoints are therefore related to the variance dynamic. The FIA strategy exhibits seven regimes which makes the CP-MV model heavily parametrized (i.e., $K \times m = 35$ parameters). This large number of regimes is perhaps related to the fact that more breakpoints are needed to adequately fit the FIA returns since the CP-MV specification does not include all the risk factors or because the

breakpoints also capture the variance dynamic. Using the selected factors and the breakpoints of the best CP-MV specification, we estimate the selective segmentation model to uncover what are the static and the dynamic parameters. Table 9 shows how the selective segmentation method improves the interpretation of the CP-MV results. First, we must acknowledge that the best specification selected by the CP-MV model is doubtful with regards to the financial literature and the practice. Nevertheless, we compare the CP-MV approach with the selected segmentation to highlight the contribution of the latter. In particular, we observe that the 'alpha' is varying and statistically positive except during the global financial crisis where it is negative but not statistically significant. This is an illustration of the lack of absolute returns during the crisis. As expected, the default risk factor is negative, time varying and very high during crises (-12.17 during the LTCM collapse and –6.47 during the global financial crisis). After the global financial crisis, the default factor is constant, negative and not statistically significant. This result is consistent with the trend observed on financial markets (especially on fixed incomes markets after the global financial crisis). The currency trend following factor, PTFSFX, is very low, time varying and statistically significant during the global financial crisis (as expected) and before the impact of the quantitative easing policies starting in the late 2010. After this date (11/2010) the factor is not statistically significant. This is an illustration of the impact of quantitative easing on fixed income arbitrage.

Table 8: **Hedge Fund Index: Best CP-MV model and best selective segmentation model.**
The Table details the parameter estimates of the preferred CP-MV model and of the selective segmentation process given the selected factors and the breakpoints found by the CP-MV model. Parentheses indicate standard deviations. A cell filled with '—' indicates that the parameter does not vary over the related period. The posterior probability of the selective segmentation model amounts to 59%.

	Preferred CP-MV model					**Selective segmentation (59%)**				
Period	Int.	PMKT	TERM	DEF	PTFSFX	Int.	PMKT	TERM	DEF	PTFSFX
03-1994 to 12-2000	0.85	0.40	-2.90	-13.46	0.02	0.53	0.42	-2.56	-12.46	0.02
	(0.21)	(0.05)	(1.18)	(2.42)	(0.01)	(0.08)	(0.03)	(0.59)	(1.50)	(0.00)
01-2001 to 10-2003	0.51	0.09	-1.46	-2.95	0.01	—	0.20	—	-2.87	—
	(0.33)	(0.08)	(1.71)	(2.85)	(0.02)	—	(0.02)	—	(0.57)	—
11-2003 to 04-2005	0.40	0.27	-2.20	-4.33	0.02	—	—	—	—	—
	(0.54)	(0.21)	(2.96)	(7.59)	(0.03)	—	—	—	—	—
05-2005 to 01-2009	0.77	0.25	1.01	-3.08	0.00	—	—	0.37	—	—
	(0.29)	(0.08)	(1.30)	(1.55)	(0.02)	—	—	(0.62)	—	—
02-2009 to 03-2016	0.16	0.24	-0.17	-2.04	0.01	—	—	—	—	—
	(0.22)	(0.05)	(1.33)	(1.27)	(0.01)	—	—	—	—	—

7.3 Out-of-sample

Sections 7.1 and 7.2 highlight the in-sample advantages of detecting which parameter truly varies when a break is detected. In addition to that, since the selective segmentation method can more accurately estimate parameters that do not change when a break occurs, we could also expect some prediction gains with respect to the standard CP model. In this Section, we investigate this aspect using the root mean squared forecast errors (RMSFE) and the cumulative log predictive density (LPD), two standard loss functions specified as,

$$\begin{aligned} \text{RMSFE} &= \sqrt{\frac{1}{T-\underline{t}}\sum_{t=\underline{t}+1}^{T}(y_t-\hat{y}_t)^2}, \\ \text{CLPD} &= \sum_{t=\underline{t}+1}^{T}\log f(y_t|y_{1:t-1}), \end{aligned}$$

in which $\hat{y}_t$ is the conditional mean of y_t given the information up to period t (i.e., $\mathbb{E}(y_t|y_{1:t-1},\mathbf{x}_t)$), $f(y_t|y_{1:t-1})$ denotes the predictive density of the model and $\underline{t}+1$ denotes the beginning of the out-of-sample forecasting period. In our prediction exercise, the training set is fixed to

Table 9: **Fixed Income Arbitrage: Best CP-MV model and best selective segmentation model.**
The Table details the parameter estimates of the preferred CP-MV model and of the selective segmentation process given the selected factors and the breakpoints found by the CP-MV model. Parentheses indicate standard deviations. A cell filled with '—' indicates that the parameter does not vary over the related period. The posterior probability of the selective segmentation model amounts to 49%.

	Preferred CP-MV model				**Selective segmentation (49%)**			
Period	Int.	AR1	DEF	PTFSFX	Int.	AR1	DEF	PTFSFX
03-1994 to 05-1995	0.40	0.23	7.57	-0.03	0.45	0.48	-0.14	-0.00
	(0.19)	(0.23)	(3.14)	(0.01)	(0.07)	(0.13)	(1.99)	(0.00)
06-1995 to 08-1997	1.28	-0.13	-0.68	0.00	—	—	—	—
	(0.53)	(0.44)	(2.57)	(0.01)	—	—	—	—
09-1997 to 11-1998	-0.23	0.10	-7.52	-0.07	—	0.12	-12.17	—
	(0.22)	(0.08)	(1.47)	(0.02)	—	(0.04)	(1.30)	—
12-1998 to 02-2008	0.38	0.29	-1.46	0.00	—	—	-1.52	—
	(0.08)	(0.09)	(0.48)	(0.00)	—	—	(0.58)	—
03-2008 to 05-2009	-0.41	0.04	-6.97	-0.03	-0.29	—	-6.47	-0.03
	(0.20)	(0.05)	(0.51)	(0.01)	(0.23)	—	(0.54)	(0.01)
06-2009 to 10-2010	1.23	0.06	0.09	-0.05	1.10	—	-0.84	—
	(0.35)	(0.21)	(0.94)	(0.01)	(0.22)	—	(0.56)	—
11-2010 to 03-2016	0.24	0.28	-1.91	-0.01	0.28	—	—	-0.01
	(0.10)	(0.16)	(0.73)	(0.00)	(0.11)	—	—	(0.01)

20% of the sample size and the 80% remaining observations are used to assess the forecast performance (i.e., $\underline{t} = 0.2T$). Since our data comprise 265 monthly returns, the out-of-sample set of observations amounts to 212 months. Each time we move forward by one month, all the considered models are re-estimated and a forecast for the next period is produced.
As competitors to our model, we consider three other processes: i) a linear regression, ii) a standard CP model with breakpoints determined by the modified method of Yau and Zhao (2016) documented in Section 5.1 (hereafter CP-YZ), iii) a CP model with the number and the locations of the breakpoints selected by minimizing the MDL criterion (hereafter CP-MDL).[6] The minimization of the MDL criterion is carried out using the dynamic programming of Bai and Perron (2003).[7] In addition to the factors and an intercept, we also account for

[6]We do not compare with the CP model of Meligkotsidou and Vrontos (2008) as the selected factors change over time. Since this Section aims at highlighting the improvements obtained from capturing which parameters truly evolve when a break occurs, we prefer sticking to model specifications that use jointly the seven risk factors as explanatory variables.

[7]See Eckley, Fearnhead, and Killick (2011) for a discussion on the implementation of the algorithm for the MDL criterion. Minimum regime duration is set to 3(K+1) to avoid capturing outliers. This choice is in favor of the standard CP model as the parameter estimates of the new regimes are based on at least

the autocorrelation of the HF returns by fixing the AR order to the value given in Table 4. Regarding the CLPD metric, we assume a normal distribution for the error term and we also use the prior distributions given in Equation (13) for the linear and the full CP models.

Table 10 documents the RMSFE and the CLPD criteria for all the Credit Suisse HF returns. For both metrics, we observe that the selective segmentation process does not always produce the best predictions. However it can still improve the RMSFE for half of the series and the CLPD for 5 out of 14 HF returns. In addition, Table 10 highlights that the selective segmentation process provides the most robust predictions. In particular, our approach delivers at least the second best predictive performance for twelve series regarding the RMSFE metric. Focusing on the CLPD criterion, the SELO method is at least second best for all the HF returns but one. This is evidence that model averaging stabilizes the forecast by reducing its variance as argued in Rapach, Strauss, and Zhou (2009). Note that the predictions of the CP model are based on the same breakpoints as the selective segmentation model. Therefore, it is remarkable that the latter model systematically dominates the standard CP model with breakpoints determined by the approach in Section 5.1 in terms of either RMSFE or CLPD criterion. Regarding the breakpoints estimated by minimizing the MDL criterion, the SELO method favorably compares for 9 and 11 out of the 14 series regarding the RMSFE and the CLPD respectively. From this small sample of series, we could argue that when CP process is used to produce predictions, detecting which parameters truly vary over time is relevant as it would likely improve the forecast performance.

8 Conclusion

Since the seminal work of Chernoff and Zacks (1964), many CP detection methods for linear models have been proposed. Most of these CP models have in common to assume, at least in practice, that all the model parameters have to change when a break is detected. In this paper, we propose to go beyond this standard framework by capturing which parameters vary when a structural break occurs. Even when conditioning to the break dates, detecting the parameters that vary from one segment to the next is not straightforward since the number of possibilities grows exponentially with the number of breaks and the number of explanatory variables. To solve this dimensional problem, we propose a penalized regression method to explore the model space and we select the best specification by maximizing a criterion that can be interpreted as a marginal likelihood in the Bayesian paradigm.
To carry out the model space exploration, we use an almost unbiased penalty function, a

3(K+1) observations.

Table 10: **RMSFE and CLPD for the fourteen HF strategies ($\underline{t} = 0.2T$).**
The Table details the RMSFE and the CLPD for five processes. Bold values indicate the model that delivers the best prediction performance. A star points out the second best model.

Models				**RMSFE**			
	HFI	CNV	DSB	EME	EMN	EDR	EDD
Linear	1.48	**1.59**	**2.70**	2.76	**2.99**	**1.29**	1.40*
CP-MDL	1.41*	1.79*	2.78	**2.58**	3.96	1.35	1.42
SELO-MDL	1.43	1.85	2.71*	2.65*	3.95*	1.33*	**1.39**
CP-YZ	1.55	3.33	4.23	3.20	4.91	1.77	3.11
SELO-YZ	**1.37**	2.46	3.16	2.66	4.68	1.44	1.88
	EDM	EDRA	FIA	GMA	LES	MFU	MUS
Linear	1.43*	1.01*	1.22*	2.47	1.84	**3.30**	**1.20**
CP-MDL	1.44	1.02	1.22	2.31*	1.80	3.40	1.23*
SELO-MDL	**1.43**	1.02	**1.21**	2.33	1.73*	3.34*	1.30
CP-YZ	1.95	1.34	2.63	3.14	1.97	3.90	1.47
SELO-YZ	1.57	**1.00**	2.35	**2.30**	**1.68**	3.42	1.34
				CLPD			
	HFI	CNV	DSB	EME	EMN	EDR	EDD
Linear	-379.42	**-402.79**	**-509.95**	-521.87	**-560.25**	**-346.20**	-351.16
CP-MDL	**-358.98**	-427.01	-516.63	**-504.94**	-666.98	-358.19	-346.83*
SELO-MDL	-362.38	-424.38*	-510.02*	-508.96*	-660.77*	-356.52*	**-346.20**
CP-YZ	-368.04	-479.76	-555.85	-530.85	-914.44	-392.15	-401.66
SELO-YZ	-360.55*	-467.14	-528.03	-510.18	-902.43	-362.43	-352.77
	EDM	EDRA	FIA	GMA	LES	MFU	MUS
Linear	-366.22*	-298.82*	**-332.62**	-487.74	-429.74	**-555.18**	-334.95*
CP-MDL	-366.23	-301.30	-338.76	-471.55	-410.62	-559.24	**-326.39**
SELO-MDL	**-364.57**	-300.24	-335.81*	-470.80*	-403.54*	-557.64*	-335.97
CP-YZ	-394.03	-337.94	-464.04	-495.44	-443.56	-576.15	-351.87
SELO-YZ	-372.84	**-296.37**	-416.79	**-468.52**	**-400.65**	-562.11	-343.74

desirable property in CP frameworks that is not exhibited by standard penalty functions (e.g., LASSO and Ridge estimators). Also, we prove the consistency of our estimator and we show how to estimate it using the DAEM algorithm. To apply the DAEM algorithm in our context, we transform the penalty function into a mixture of Normal distributions. This simple transformation greatly speeds the estimation as the DAEM algorithm iterates over closed-form expressions.

Once the promising models have been uncovered by the penalized regression approach, selecting the parameters of the penalty function is carried out by maximizing a marginal likelihood. Thanks to the Bayesian interpretation of this consistent criterion, we can take model uncertainty into account and can do Bayesian model averaging, a feature that generally improves forecast performance. A simulation study highlights that our selective segmentation method works well in practice for a range of diversified data generating processes.

We illustrate our approach with HF returns. The selective segmentation model has two main advantages. First, as the standard CP models, it detects the breakpoints and the corresponding regimes. Second, it highlights the time-varying dynamics of the changing risk factors. When we compare our model with previous advanced CP models, we observe that it is particularly appealing to capture the time-varying dynamics of risk exposures. Then, we test the predictive performance of the selective segmentation approach with respect to the linear regression and standard CP processes. We note that our method produces the most robust forecasts and almost systematically dominates the CP processes based on the same breakpoints. These encouraging results suggest promising developments and applications in financial economics.

Importantly, an R package for estimating the model is available on the corresponding author's web page. This package stands for a building block of our future research that will include a dynamic variance and multivariate models.

Acknowledgments

The authors are grateful to seminar participants at Aix-Marseille School of Economics, at Maastricht University, at ShanghaiTech University and at Neuchatel University. They also thank participants for their helpful comments at the 11th CFE conference, at the 8th PhD Student Conference in International Macroeconomics and Financial Econometrics, at the 59e congrès de la Société Canadienne de Science Economique, at R à Québec and at the 53rd Annual Conference of the Canadian Economique Association.

Arnaud Dufays acknowledges financial support from the F.S.R-FNRS via the contract CR 144. He also acknowledges financial supports from the Fonds de recherche du Québec – Société

et culture and from the SSHRC-CRSH via the project 430-2017-00215.

A Proofs of the consistency of the Penalty function

In this appendix, we proof Proposition 1. To do so, we first state and prove two Lemmas.

Lemma 1. *Under the conditions A1-A5 and let,*

$$f_T(\boldsymbol{\beta}) = \frac{1}{T}||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + \sum_{j=2}^{m}\sum_{k=1}^{K}\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) \tag{A.1}$$

Then of every $\nu \in (0,1)$, there exists a constant $C_0 > 0$ such that

$$\liminf_{T\to\infty} \mathbf{P}\left[\underset{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 \leq C\sqrt{\frac{Km\sigma^2}{T}}}{\arg\min} f_T(\boldsymbol{\beta}) \subseteq \left\{\boldsymbol{\beta} \in \Re^{Km\times 1}; ||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 < C\sqrt{\frac{Km\sigma^2}{T}}\right\}\right] > 1-\nu$$

for all $C \geq C_0$.

Proof. The proof is given in Appendix A.1. □

Lemma 2. *Let $C > 0$ and f_T as defined by Equation (7). Under the conditions A1-A5,*

$$\liminf_{T\to\infty} \mathbf{P}\left[\underset{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 \leq C\sqrt{\frac{Km\sigma^2}{T}}}{\arg\min} f_T(\boldsymbol{\beta}) \subseteq \left\{\boldsymbol{\beta} \in \Re^{Km\times 1}; \boldsymbol{\beta}_{A^c} = 0\right\}\right] = 1$$

where $A^c = \{(j,k), j = 1,\ldots,m \text{ and } k = 0,\ldots,K-1\}\backslash A$ is the complement of A in $\{(j,k), j = 1,\ldots,m \text{ and } k = 0,\ldots,K-1\}$, $\boldsymbol{\beta}_{A^c} \in \Re^{|A^c|\times 1}$ is the $|A^c|$-dimensional sub-vector of $\boldsymbol{\beta}$ containing components subscripted by A^c.

Proof. See Appendix A.2 for the proof. □

A.1 Proof of Lemma 1

Proof 1. *We consider the objective function*

$$f_T(\boldsymbol{\beta}) = \frac{1}{T}||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + \sum_{j=2}^{m}\sum_{k=1}^{K}\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda)$$

Let $\alpha_T = \sqrt{\frac{Km\sigma^2}{T}}$ *and* $\nu \in (0,1)$. *To prove the lemma 1, It suffices to show that*

$$\mathbf{P}\left(f_T(\boldsymbol{\beta}^*) < \inf_{||\boldsymbol{u}||_2=1} f_T(\boldsymbol{\beta}^* + C\alpha_T \boldsymbol{u})\right) = 1-\nu$$

for $C>0$ *sufficiently large and for any* T *sufficiently large. In other words, we shall show that* $H_T(\boldsymbol{u}) = f_T(\boldsymbol{\beta}^* + C\alpha_T\boldsymbol{u}) - f_T(\boldsymbol{\beta}^*)$ *is positive for any* T *when* C *is large enough and for all* $||\boldsymbol{u}||_2 = 1$, *where* $\boldsymbol{u} = (u_{11}, u_{12}\ \ldots, u_{1K},\ \ldots,\ u_{mK}) \in \mathbb{R}^{mK\times 1}$.
We can easily show that

$$\begin{aligned}
H_T(\boldsymbol{u}) &= \frac{1}{T}\left(C^2\alpha_T^2||\mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{u}||_2^2 - 2C\alpha_T\boldsymbol{\epsilon}'\mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{u}\right) + \\
&\quad \sum_{j=2}^{m}\sum_{k=1}^{K}\left(\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^* + C\alpha_T u_{jk}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)\right) \\
H_T(\boldsymbol{u}) &\geq \frac{1}{T}\left(C^2\alpha_T^2||\mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{u}||_2^2 - 2C\alpha_T\boldsymbol{\epsilon}'\mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{u}\right) + \\
&\quad \sum_{(j,k)\in\mathbf{D}(\boldsymbol{u})}\left(\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)\right)
\end{aligned}$$

where $\Delta\beta_{jk}^{*+} = \Delta\beta_{jk}^* + C\alpha_T u_{jk}$ *and*
$\mathbf{D}(\boldsymbol{u}) = \left\{(j,k); j\geq 2 \text{ and } \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda) < 0\right\}$.
For any $(j,k)\in\mathbf{D}(\boldsymbol{u})$, *clearly* $\Delta\beta_{jk}^* \neq 0$, *otherwise* $\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda) \geq 0$. *Thus, if* $C>0$ *is sufficiently large and fixed, as* $\lim_{T\to\infty} C\alpha_T = 0$, *we can consider that* $\Delta\beta_{jk}^{*+}$ *and* $\Delta\beta_{jk}^*$ *have the same sign for* T *sufficiently large; that is* $0 \notin (c_T^-, c_T^+)$, *where* $c_T^- = \min(\Delta\beta_{jk}^{*+}, \Delta\beta_{jk}^*)$ *and* $c_T^+ = \max(\Delta\beta_{jk}^{*+}, \Delta\beta_{jk}^*)$. *By the fact that* $\mathcal{P}_{\mathbf{SELO}}(x|a_k,\lambda)$ *is a concave function on* $x\in(-\infty,0]$ *and on* $x\in[0,+\infty)$, *thus also on* (c_T^-, c_T^+), *we can establish the following conditions using the mean value theorem.*

$$\frac{\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)}{C\alpha_T u_{jk}} \leq \max\left(\mathcal{P}'_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda), \mathcal{P}'_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)\right)$$

$$\frac{\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)}{C\alpha_T u_{jk}} \geq \min\left(\mathcal{P}'_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda), \mathcal{P}'_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)\right)$$

where $\mathcal{P}'_{\mathbf{SELO}}$ *stands for the* $\mathcal{P}_{\mathbf{SELO}}$ *first derivative.*

Let us note that $\forall\ (j,k) \in \mathbf{D}(\boldsymbol{u})$, $\Delta\beta^*_{jk} > 0 \implies u_{jk} < 0$ *and* $\Delta\beta^*_{jk} < 0 \implies u_{jk} > 0$ *so that* $\mathcal{P}_{\mathbf{SELO}}(\Delta\beta^{*+}_{jk}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta^*_{jk}|a_k,\lambda) < 0$ *holds. Applying the mean value theorem in both cases, we end up with a common condition given by*

$$\begin{aligned}\mathcal{P}_{\mathbf{SELO}}(\Delta\beta^{*+}_{jk}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta^*_{jk}|a_k,\lambda) &\geq -C\alpha_T|u_{jk}||\mathcal{P}'_{\mathbf{SELO}}(\Delta\beta^*_{jk} + C\alpha_T u_{jk}|a_k,\lambda)| \\ &\geq -\frac{C\lambda\alpha_T a_k\zeta}{\ln(2)(\rho^2 - 2\rho C\alpha_T)} \\ &\geq -\frac{C\lambda\alpha_T a_{max}\zeta}{\ln(2)(\rho^2 - 2\rho C\alpha_T)}\end{aligned}$$

where the last two inequalities come from the $|\mathcal{P}'_{\mathbf{SELO}}(\Delta\beta^*_{jk} + C\alpha_T u_{jk}|a_k,\lambda)|$ *minimization with respect to* $\Delta\beta^*_{jk}$*. Then*

$$H_T(\boldsymbol{u}) \geq \underbrace{\frac{C^2\alpha_T^2||\mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{u}||_2^2}{T}}_{Q_1} - \underbrace{\frac{2C\alpha_T\boldsymbol{\epsilon}'\mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{u}}{T}}_{Q_2} - \underbrace{\frac{CKm\lambda\alpha_T a_{max}\zeta}{\ln(2)(\rho^2 - 2\rho C\alpha_T)}}_{Q_3}$$

Focusing on each term, we can show that

$$Q_1 \equiv \frac{C^2\alpha_T^2||\mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{u}||_2^2}{T} = C^2\alpha_T^2\boldsymbol{u}'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}\boldsymbol{u} \geq C^2\alpha_T^2\lambda_{T,\min}$$

where $\lambda_{T,\min}$ *is the smallest eigenvalue of* $\dfrac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}$.

To show this condition, we can decompose $\dfrac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}$ *into* $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$ *(by A3). Moreover, any vector of* Km *dimension can be decomposed into a linear combination of the eigenvectors (i.e.,* $\boldsymbol{u} = \boldsymbol{U}\boldsymbol{\omega}$*). Note that* $\boldsymbol{u}'\boldsymbol{u} = \boldsymbol{\omega}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{\omega} = \boldsymbol{\omega}'\boldsymbol{\omega} = \sum_{i=1}^{Km}\omega_i^2 = 1$.

Thus $\boldsymbol{u}'\dfrac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}\boldsymbol{u} = \boldsymbol{\omega}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{U}\boldsymbol{\omega} = \boldsymbol{\omega}'\boldsymbol{\Lambda}\boldsymbol{\omega} = \sum_{i=1}^{Km}\omega_i^2\lambda_i \geq \lambda_{T,\min}$.

The second term is given by

$$\begin{aligned} Q_2 \equiv \frac{2C\alpha_T \boldsymbol{\epsilon}' \mathbf{X}_{\boldsymbol{\tau}} \boldsymbol{u}}{T} &\leq \frac{2C\alpha_T |\boldsymbol{\epsilon}' \mathbf{X}_{\boldsymbol{\tau}} \boldsymbol{u}|}{T} \\ &\leq \frac{2C\alpha_T ||\boldsymbol{\epsilon}' \mathbf{X}_{\boldsymbol{\tau}}||_2 ||\boldsymbol{u}||_2}{T} \text{ by Cauchy-Schwartz} \\ &\leq \frac{2C\alpha_T^2}{\sqrt{Km\sigma^2}} \sqrt{\frac{(\boldsymbol{\epsilon}'(\mathbf{X}_{\boldsymbol{\tau}} \mathbf{X}_{\boldsymbol{\tau}}')\boldsymbol{\epsilon})}{T}} \\ &\leq \mathcal{O}_p(C\alpha_T^2) \quad (By\ A3\ and\ A4). \end{aligned}$$

To show that $\sqrt{\frac{(\boldsymbol{\epsilon}'(\mathbf{X}_{\boldsymbol{\tau}} \mathbf{X}_{\boldsymbol{\tau}}')\boldsymbol{\epsilon})}{T}} = \mathcal{O}_p(1)$*, we rely on the spectral theorem to decompose* $\mathbf{X}_{\boldsymbol{\tau}} \mathbf{X}_{\boldsymbol{\tau}}'$ *into two orthogonal matrices and a diagonal matrix of eigenvalues. With this decomposition, we can show that* $\epsilon' \frac{\mathbf{X}_{\boldsymbol{\tau}} \mathbf{X}_{\boldsymbol{\tau}}'}{T} \epsilon \leq max_i \lambda_i \frac{\epsilon'\epsilon}{T}$ *which is* $\mathcal{O}_p(1)$ *under Assumption A3 and the fact that the variance is bounded.*

The last term Q_3 *is defined by*

$$\begin{aligned} Q_3 &\equiv \frac{CKm\lambda\alpha_T a_{max}\zeta}{\ln(2)(\rho^2 - 2\rho C\alpha_T)} \\ &= C\alpha_T^2 \frac{\lambda\zeta \frac{a_{\max}}{(Km)^{-1}\alpha_T^3}}{\ln(2)\left(\left(\frac{\rho}{\alpha_T}\right)^2 - 2C\left(\frac{\rho}{\alpha_T}\right)\right)} \end{aligned}$$

By A2, $\left(\frac{\rho}{\alpha_T}\right)^2 - 2C\left(\frac{\rho}{\alpha_T}\right) \to \infty$ *and by A5,* $\lim_{T\to\infty} \lambda\zeta \frac{a_{\max}}{(Km)^{-1}\alpha_T^3} < \infty$*. Hence* $Q_3 = o(C\alpha_T^2)$*. Combining the conditions on* Q_1*,* Q_2 *and* Q_3 *we establish that*

$$H_T(\boldsymbol{u}) \geq C^2\alpha_T^2\lambda_{T,\min} + \mathcal{O}_p(C\alpha_T^2) + o(C\alpha_T^2)$$

. It follows that there exists $C_0 > 0$ *is large such that for all* $C > C_0$*,* $\mathbf{P}\left(\inf_{||\boldsymbol{u}||_2=1} H_T(\boldsymbol{u}) > 0\right) = 1 - \nu$*, for* T *sufficiently large.*

A.2 Lemma 2

Proof 2. *Let $\boldsymbol{\beta} \in \Re^{Km\times 1}$ such that $||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 < C\sqrt{\frac{Km\sigma^2}{T}}$. We consider $\tilde{\boldsymbol{\beta}} \in \Re^{Km\times 1}$, where $\tilde{\boldsymbol{\beta}}_{A^c} = 0$ and $\tilde{\boldsymbol{\beta}}_A = \boldsymbol{\beta}_A$. We can notice that*

$$||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2 = ||\boldsymbol{\beta}_{A^c} - \tilde{\boldsymbol{\beta}}_{A^c}||_2 = ||\boldsymbol{\beta}_{A^c} - \boldsymbol{\beta}^*_{A^c}||_2$$
$$||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2 < C\alpha_T$$

On the other hand

$$||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2 = ||\boldsymbol{\beta}^*_A - \tilde{\boldsymbol{\beta}}_A||_2 = ||\boldsymbol{\beta}^*_A - \boldsymbol{\beta}_A||_2$$
$$||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2 < C\alpha_T$$

Let us define $G_T(\boldsymbol{\beta}) = f_T(\boldsymbol{\beta}) - f_T(\tilde{\boldsymbol{\beta}})$. Similarly to the proof of the lemma 1, it suffices to show that $G_T(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}) > 0$.

$$\begin{aligned}
G_T(\boldsymbol{\beta}) &= \frac{1}{T}\left(||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 - ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\tilde{\boldsymbol{\beta}}||_2^2\right) + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k,\lambda) \\
&= \frac{1}{T}\left(||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\tilde{\boldsymbol{\beta}} - \mathbf{X}_{\boldsymbol{\tau}}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})||_2^2 - ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\tilde{\beta}||_2^2\right) + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k,\lambda) \\
&= (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - 2(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}(\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\tilde{\boldsymbol{\beta}})}{T} + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k,\lambda) \\
&= (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - 2(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\boldsymbol{\epsilon}}{T} - 2(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}) \\
&\quad + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k,\lambda) \\
&= (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - 2\alpha_T\frac{(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'}{\sqrt{Km\sigma^2}}\frac{\mathbf{X}'_{\boldsymbol{\tau}}\boldsymbol{\epsilon}}{\sqrt{T}} - 2(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}) \\
&\quad + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k,\lambda)
\end{aligned}$$

By A1, A3 and A4, $\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T} = \mathcal{O}_p(1)$ *and* $\frac{\mathbf{X}'_{\boldsymbol{\tau}}\boldsymbol{\epsilon}}{\sqrt{T}} = \mathcal{O}_p(1)$. *Moreover* $||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2 < C\alpha_T$ *and* $||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2 \leq C\alpha_T$. *Then, for any* T *sufficiently large*

$$G_T(\boldsymbol{\beta}) = \mathcal{O}_p\left(||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2\alpha_T\right) + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda)$$

As $\mathcal{P}_{\mathbf{SELO}}(x|a_k, \lambda)$ *is a concave function on* $x \in]-\infty, 0]$ *and on* $x \in [0, +\infty[$, *for any* $\nu_1 < \nu_2 \leq \nu_3 \leq 0$ *(resp.* $0 \leq \nu_1 \leq \nu_2 < \nu_3$*),* $\frac{\mathcal{P}_{\mathbf{SELO}}(\nu_1) - \mathcal{P}_{\mathbf{SELO}}(\nu_3)}{\nu_1 - \nu_3} \geq \frac{\mathcal{P}_{\mathbf{SELO}}(\nu_2) - \mathcal{P}_{\mathbf{SELO}}(\nu_3)}{\nu_2 - \nu_3}$ *(resp.* $\frac{\mathcal{P}_{\mathbf{SELO}}(\nu_3) - \mathcal{P}_{\mathbf{SELO}}(\nu_1)}{\nu_3 - \nu_1} \leq \frac{\mathcal{P}_{\mathbf{SELO}}(\nu_2) - \mathcal{P}_{\mathbf{SELO}}(\nu_1)}{\nu_2 - \nu_1}$*).*
$\forall\ (j,k) \in A^c$, $\Delta\beta^*_{jk} = 0$ *and* $\Delta\beta_{jk}$ *is strictly positive or negative.*
Thus, $-C\alpha_T \leq \beta_{jk} < 0$ *or* $0 < \Delta\beta_{jk} < C\alpha_T$, *since* $|\Delta\beta_{jk}| \leq ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_2 < C\alpha_T$. *In both cases, we end up with*

$$\begin{aligned}
\frac{\mathcal{P}_{\mathbf{SELO}}\left(C\alpha_T|a_k, \lambda\right)}{C\alpha_T} &\leq \frac{\mathcal{P}_{\mathbf{SELO}}\left(\Delta\beta_{jk}|a_k, \lambda\right)}{|\Delta\beta_{jk}|} \\
\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) &\geq \frac{\lambda}{\ln(2)C\alpha_T} \ln\left(\frac{C\alpha_T}{C\alpha_T + a_k\zeta} + 1\right)|\Delta\beta_{jk}| \\
\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) &\geq \frac{\lambda}{\ln(2)C} \ln\left(\frac{C\alpha_T}{C\alpha_T + a_{\max}\zeta} + 1\right)|\Delta\beta_{jk}|
\end{aligned}$$

for any T *sufficiently large. Thus*

$$\sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) \geq \frac{\lambda}{\ln(2)C} \ln\left(\frac{C}{C + a_{\max}\zeta\sqrt{\frac{T}{Km\sigma^2}}} + 1\right)||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2.$$

Furthermore, by A5 $a_{\max} = \mathcal{O}_p\left(\sqrt{\frac{mK\sigma^2}{T}}\frac{\sigma_2}{T}\right)$. *Thereby*

$$a_{\max}\zeta\sqrt{\frac{T}{Km\sigma^2}} \xrightarrow{p} 0 \text{ and } \liminf_{T\to\infty}\left(\frac{C}{C + a_{\max}\zeta\sqrt{\frac{T}{Km\sigma^2}}} + 1\right) > 0$$

It follows that, there exists $\tilde{C} > 0$ *such that*

$$\frac{G_T(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}})}{||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2} \geq \tilde{C}\lambda + \mathcal{O}_p\left(\sqrt{\frac{Km\sigma^2}{T}}\right)$$

Thereby the result follows.

A.3 Proof of the Proposition 1

Proof 3. *The theorem is immediatly given by the lemmas (1) and (2), in the sense that there exists a sequence of local minima* $\hat{\boldsymbol{\beta}}$ *of* $f_T(\boldsymbol{\beta})$ *such that* $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = \mathcal{O}_p\left(\sqrt{\frac{Km\sigma^2}{T}}\right)$ *and* $\hat{\boldsymbol{\beta}}_{A^c} = \mathbf{0} \in \mathbb{R}^{|A^c| \times 1}$. *Thus, as* $\sqrt{\frac{Km\sigma^2}{T}} \to 0$, *it follows that* $||\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}^*_A||_2 = o_P(1)$.

A.4 Consequence of bounded eigenvalues

We show that bounded eigenvalues of the matrix $\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}$ implies a fixed number of regimes. Note first that

$$\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}} = \sum_{t=1}^{T}(\mathbf{1}_{\{t\}} \otimes \mathbf{x}_t)(\mathbf{1}_{\{t\}} \otimes \mathbf{x}_t)', \tag{A.2}$$

$$= \sum_{t=1}^{T}(\mathbf{x}_t\mathbf{x}'_t) \otimes (\mathbf{1}_{\{t\}}\mathbf{1}'_{\{t\}}), \tag{A.3}$$

where we define $\mathbf{1}_{\{t\}} = (\mathbf{1}_{\{t>\tau_0\}}, \mathbf{1}_{\{t>\tau_1\}}, \ldots, \mathbf{1}_{\{t>\tau_{m-1}\}})'$. Let us define $n_i = \sum_{t=1}^{T}\mathbf{1}_{\{t>\tau_{i-1}\}}$, i.e., the number of observations from the beginning of regime i to the end of the sample. Working with $\mathbf{x}_t \equiv 1$, we have that

$$\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{1}_{\{t\}}\mathbf{1}'_{\{t\}}), \tag{A.4}$$

$$= \frac{1}{T}\begin{pmatrix} n_1 & n_2 & n_3 & \ldots & n_m \\ n_2 & n_2 & n_3 & \ldots & n_m \\ n_3 & n_3 & n_3 & \ldots & n_m \\ & & & \ldots & \\ n_m & n_m & n_m & \ldots & n_m \end{pmatrix} \tag{A.5}$$

in which $n_1 = T$. It leads to the following determinant, when $m > 1$,

$$|\frac{\mathbf{X}_{\boldsymbol{\tau}}'\mathbf{X}_{\boldsymbol{\tau}}}{T}| = T^{-m} n_m \prod_{i=1}^{m-1} (n_i - n_{i+1}), \tag{A.6}$$

$$= \frac{n_m}{T} \prod_{i=1}^{m-1} \frac{(n_i - n_{i+1})}{T}. \tag{A.7}$$

If each break increases with T such that $n_i = (1 - \delta_{i-1})T$ with $\delta_0 = 0$ and $\delta_m = 1$, we have

$$|\frac{\mathbf{X}_{\boldsymbol{\tau}}'\mathbf{X}_{\boldsymbol{\tau}}}{T}| = (1 - \delta_{m-1}) \prod_{i=1}^{m-1} (\delta_i - \delta_{i-1}) \geq 0 \quad \forall T. \tag{A.8}$$

It shows that the number of segments cannot increase with T otherwise the determinant tends to zero. In fact, assuming a growing number of regime such that the minimum regime duration is given $\epsilon_\tau = \mathcal{O}(T^{q-1})$ with $q \in [0, 1)$ (which implies $m = \mathcal{O}(T^q)$) leads to a determinant tending to zero because

1. If $m = \mathcal{O}(T^q)$, then $\prod_{i=1}^{m-1}(\delta_i - \delta_{i-1}) \to 0$ since $\delta_i - \delta_{i-1} < 1$.
2. If $\tau_i - \tau_{i-1} > T\epsilon_\tau$ with $\epsilon_\tau = \mathcal{O}(T^{q-1})$ and $q \in [0, 1)$, it implies that $(\delta_i - \delta_{i-1}) = \mathcal{O}(T^{q-1})$ which tends to zero. Consequently, we have $(\delta_i - \delta_{i-1}) \to 0$.

A.5 Approximation of the penalty function with mixture of normal densities

To derive the DAEM algorithm, a mixture of normal densities has been assumed for the mean parameter. We now provide a simple mixture approximation of the SELO penalty. Note that in practice, one can use the output of the DAEM algorithm as a starting point to optimize the function of Equation (8). Due to the mixture approximation and the continuity of the SELO penalty function, the starting point would be in general very close to the value that globally minimizes the function (8). The prior density of one model parameter $\Delta\beta_{jk}$ is as follows

$$\begin{aligned} f(\Delta\beta_{jk}) &= \sum_{i=1}^{3} \omega_i^{(k)} f_N(\Delta\beta_{jk}|\mu_i^{(k)}, s_i^{(k)}), \\ &= \lambda_1^{(k)} f_N(\Delta\beta_{jk}|0, s_1^{(k)}) + \lambda_2^{(k)} f_N(\Delta\beta_{jk}| - a_k, s_2)\mathbf{1}_{\{(\Delta\beta_{jk} \leq -a_k\}} \\ &\quad + \lambda_2^{(k)} f_N(\Delta\beta_{jk}|a_k, s_2)\mathbf{1}_{\{(\Delta\beta_{jk} \geq a_k\}}. \end{aligned}$$

We set the slab variance s_2 to 1000. To fix the other parameters, we use the three next conditions

$$\begin{aligned} \text{i)} \quad \lambda_1^{(k)} + \lambda_2^{(k)} &= 1, \\ \text{ii)} \; C\lambda_1^{(k)} f_N(a_k|0, s_1^{(k)}) &= \lambda_2 f_N(a_k|a_k, s_2), \\ \text{iii)} \quad C\lambda &= \ln f(0) - \ln f(a_k), \end{aligned}$$

where C is a constant that determines the overlapping of the spike and the slab component at a_k. The first condition ensures that the prior density is proper while the third condition imposes a cost of $-\lambda$ when the model parameter is equal to a_k compared to when it is equal to 0 (in log density). In comparison, the SELO penalty gives a penalty of 0.99λ. Under these conditions, we easily find the value $s_1^{(k)}$ and $\lambda_1^{(k)}, \lambda_2^{(k)}$:

$$\begin{aligned} s_1^{(k)} &= \frac{a_k^2}{2(\lambda + \ln(\frac{C}{2} + 1))}, \\ \lambda_1^{(k)} &= \frac{\sqrt{s_1^{(k)}}}{e^{-a_k^2/(2s_1^{(k)})} C\sqrt{s_2} + \sqrt{s_1^{(k)}}} \\ \lambda_2^{(k)} &= \frac{e^{-a_k^2/(2s_1^{(k)})} C\sqrt{s_2}}{e^{-a_k^2/(2s_1^{(k)})} C\sqrt{s_2} + \sqrt{s_1^{(k)}}}. \end{aligned}$$

Note that the constant C can be interpreted as an error of type I. Let us relate $\delta_{jk} = 0$ to the spike component and $\delta_{jk} = 1$ to the slab component when $\Delta\beta_{jk} < 0$. Then the probability that $\delta_{jk} = 0$ when $\Delta\beta_{jk} = a_k$ is given by

$$\begin{aligned} \alpha \equiv P[\delta_{jk} = 0|\Delta\beta_{jk} = a_k] &= \frac{f_N(\Delta\beta_{jk}|0, s_1^{(k)})\lambda_1^{(k)}}{f_N(\Delta\beta_{jk}|0, s_1^{(k)})\lambda_1^{(k)} + f_N(\Delta\beta_{jk}|a_k, s_2)\frac{\lambda_2^{(k)}}{2}}, \\ C &= 2\frac{(1-\alpha)}{\alpha}. \end{aligned}$$

In practice, the error of type I (i.e., α) is set to 5%. Once all the parameters are fixed, we slightly modify the expectation of the slab components to ensure a steady decline in the

parameter penalization. In fact, we use the following mixture of normal densities

$$\begin{aligned}\pi(\Delta\beta_{jk}) &= \lambda_1^{(k)} f_N(\Delta\beta_{jk}|0, s_1^{(k)}) + \lambda_2^{(k)} f_N(\Delta\beta_{jk}| - q_k, s_2)\mathbf{1}_{\{(\Delta\beta_{jk}\leq -q_k\}} \\ &\quad +\lambda_2 f_N(\Delta\beta_{jk}|q_k, s_2)\mathbf{1}_{\{(\Delta\beta_{jk}\geq q_k\}}, \\ q_k &= \min(\sqrt{2\lambda s_1^{(k)}}, a_k).\end{aligned}$$

Figure A.1 documents the mixture approximation (minus the log of the density) with respect to the SELO penalty function for different values of $\boldsymbol{a} = (a_1, \ldots, a_K)$ and λ. We observe that the slab component can generate unbiased estimators for a wide range of values.

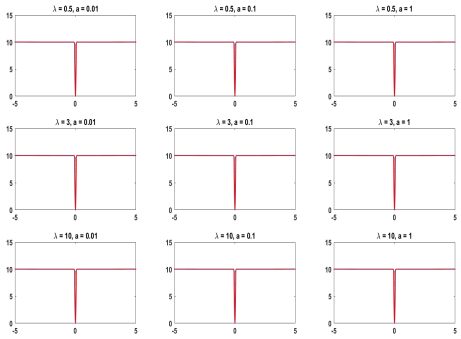


Figure A.1: Mixture approximation (in red) with respect to the SELO penalty function (in blue).

B Marginal likelihood (11) for the linear model

Let us derive the criterion (11). We first define $X_1 = \tilde{\mathbf{X}}_{\tau_0}$, $X_2 = X_{\boldsymbol{\tau}}^{\hat{A}}$ and $M_{X_1} = \mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}}$. Given the prior distributions in Equation (13), the marginal likelihood is given by,

$$\begin{aligned}f(\mathbf{y}|\boldsymbol{a}, \lambda, \boldsymbol{\tau}) &= \int\int (2\pi)^{\frac{-(T+k_{\hat{A}})}{2}} (\sigma^2)^{-\frac{(T+2+k_{\hat{A}})}{2}} |g_{\hat{A}}(X_2)' M_{X_1} X_2)|^{1/2} \\ &\exp\frac{-1}{2\sigma^2}\{\underbrace{(y - X_1\boldsymbol{\beta}_1 - X_2\Delta\boldsymbol{\beta})'(y - X_1\boldsymbol{\beta}_1 - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}}(X_2)' M_{X_1} X_2)\Delta\boldsymbol{\beta}}_{B}\} d(\boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}) d\sigma^2.\end{aligned}$$

Focusing on the expression in the exponential, we have

$$\begin{aligned}
B &= (y - X_1\boldsymbol{\beta}_1 - X_2\Delta\boldsymbol{\beta})'(y - X_1\boldsymbol{\beta}_1 - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}} \underbrace{(X_2' M_{X_1} X_2)}_{\Sigma_X} \Delta\boldsymbol{\beta}, \\
&= (y - X_2\Delta\boldsymbol{\beta})'(y - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}} \underbrace{(X_2' M_{X_1} X_2)}_{\Sigma_X} \Delta\boldsymbol{\beta} + \boldsymbol{\beta}_1' X_1' X_1 \boldsymbol{\beta}_1 - 2\boldsymbol{\beta}_1' X_1'(y - X_2\Delta\boldsymbol{\beta}), \\
&= (y - X_2\Delta\boldsymbol{\beta})'(y - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}} \Sigma_X \Delta\boldsymbol{\beta} + (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1)'\Omega^{-1}(\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1) - \bar{\boldsymbol{\beta}}_1{}'\Omega^{-1}\bar{\boldsymbol{\beta}}_1,
\end{aligned}$$

where $\Omega^{-1} = X_1'X_1$, $\bar{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\Delta\boldsymbol{\beta})$ and $\bar{\boldsymbol{\beta}}_1{}'\Omega^{-1}\bar{\boldsymbol{\beta}}_1 = (y - X_2\Delta\boldsymbol{\beta})'X_1(X_1'X_1)^{-1}X_1'(y - X_2\Delta\boldsymbol{\beta}) = (y - X_2\Delta\boldsymbol{\beta})'P_{X_1}(y - X_2\Delta\boldsymbol{\beta})$. The marginal likelihood can be simplifies as

$$\begin{aligned}
f(\mathbf{y}|\boldsymbol{a}, \lambda, \boldsymbol{\tau}) &= |X_1'X_1|^{-\frac{1}{2}} \int\int (2\pi)^{\frac{-(T+k_{\hat{A}}-K)}{2}} (\sigma^2)^{-\frac{(T+2+k_{\hat{A}}-K)}{2}} |g_{\hat{A}}\Sigma_X|^{1/2} \\
&\quad \exp\frac{-1}{2\sigma^2}\{\underbrace{(y - X_2\Delta\boldsymbol{\beta})'(y - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}}\Sigma_X\Delta\boldsymbol{\beta} - (y - X_2\Delta\boldsymbol{\beta})'P_{X_1}(y - X_2\Delta\boldsymbol{\beta})}_{C}\}d(\Delta\boldsymbol{\beta})d\sigma^2
\end{aligned}$$

Again, focusing on the expression of the exponential, we obtain

$$\begin{aligned}
C &= (y - X_2\Delta\boldsymbol{\beta})'(y - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}}\Sigma_X\Delta\boldsymbol{\beta} - (y - X_2\Delta\boldsymbol{\beta})'P_{X_1}(y - X_2\Delta\boldsymbol{\beta}), \\
&= y'[I_T - P_{X_1}]y + \Delta\boldsymbol{\beta}'[g_{\hat{A}}X_2'M_{X_1}X_2 + X_2'X_2 - X_2'P_{X_1}X_2]\Delta\boldsymbol{\beta} - 2\Delta\boldsymbol{\beta}'X_2'[I_T - P_{X_1}]y, \\
&= y'M_{X_1}y + \Delta\boldsymbol{\beta}'[(1 + g_{\hat{A}})X_2'M_{X_1}X_2]\Delta\boldsymbol{\beta} - 2\Delta\boldsymbol{\beta}'X_2'M_{X_1}y, \\
&= y'M_{X_1}y + (\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})'\bar{\Sigma}^{-1}(\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}}) - \bar{\boldsymbol{\mu}}'\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}},
\end{aligned}$$

where $\bar{\Sigma}^{-1} = (1 + g_{\hat{A}})X_2'M_{X_1}X_2 = (1 + g_{\hat{A}})\Sigma_X$ and $\bar{\boldsymbol{\mu}} = \bar{\Sigma}X_2'M_{X_1}y$, $\bar{\boldsymbol{\mu}}'\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}} = (1 + g_{\hat{A}})^{-1}y'M_{X_1}X_2[X_2'M_{X_1}X_2]^{-1}X_2'M_{X_1}y$.

Eventually, we find the following marginal likelihood

$$
\begin{aligned}
f(\mathbf{y}|\boldsymbol{a},\lambda,\boldsymbol{\tau}) &= (2\pi)^{\frac{-(T-K)}{2}}|X_1'X_1|^{-\frac{1}{2}}|g_{\hat{A}}\Sigma_X|^{1/2}|(1+g_{\hat{A}})\Sigma_X|^{\frac{-1}{2}}\int(\sigma^2)^{-\frac{(T+2-K)}{2}} \\
&\quad \exp\frac{-1}{2\sigma^2}\{y'M_{X_1}y-(1+g_{\hat{A}})^{-1}y'M_{X_1}X_2[X_2'M_{X_1}X_2]^{-1}X_2'M_{X_1}y\}d\sigma^2, \\
&= (\pi)^{\frac{-(T-K)}{2}}\Gamma(\frac{T-K}{2})|X_1'X_1|^{-\frac{1}{2}} \\
&\quad (\frac{g_{\hat{A}}}{1+g_{\hat{A}}})^{k_{\hat{A}}/2}[y'M_{X_1}y-(1+g_{\hat{A}})^{-1}y'M_{X_1}X_2[X_2'M_{X_1}X_2]^{-1}X_2'M_{X_1}y]^{-\frac{T-K}{2}}, \\
&= (\pi)^{\frac{-(T-K)}{2}}\Gamma(\frac{T-K}{2})|X_1'X_1|^{-\frac{1}{2}} \\
&\quad (\frac{g_{\hat{A}}}{1+g_{\hat{A}}})^{k_{\hat{A}}/2}[\frac{g_{\hat{A}}}{1+g_{\hat{A}}}y'M_{X_1}y+\frac{1}{(1+g_{\hat{A}})}[\tilde{y}'\tilde{y}-\tilde{y}'X_2[X_2'M_{X_1}X_2]^{-1}X_2\tilde{y}]]^{-\frac{T-K}{2}}, \\
&= (\pi)^{\frac{-(T-K)}{2}}\Gamma(\frac{T-K}{2})|X_1'X_1|^{-\frac{1}{2}}(\frac{g_{\hat{A}}}{1+g_{\hat{A}}})^{k_{\hat{A}}/2}[\frac{g_{\hat{A}}}{1+g_{\hat{A}}}s_{X_1}+\frac{1}{(1+g_{\hat{A}})}s_{X_1,X_2}]^{-\frac{T-K}{2}},
\end{aligned}
$$

where the penultimate equality comes from the Frisch-Waugh theorem.

B.1 Posterior distribution

$$
\begin{aligned}
f(\boldsymbol{\beta}_1,\Delta\boldsymbol{\beta},\sigma^2|\mathbf{y},\boldsymbol{\tau}) &\propto (2\pi)^{\frac{-(T+k_{\hat{A}})}{2}}(\sigma^2)^{-\frac{(T+2+k_{\hat{A}})}{2}}|g_{\hat{A}}(X_2)'M_{X_1}X_2)|^{1/2} \\
&\quad \exp\frac{-1}{2\sigma^2}\left(\underbrace{(y-X_1\boldsymbol{\beta}_1-X_2\Delta\boldsymbol{\beta})'(y-X_1\boldsymbol{\beta}_1-X_2\Delta\boldsymbol{\beta})+\Delta\boldsymbol{\beta}'g_{\hat{A}}(X_2)'M_{X_1}X_2\Delta\boldsymbol{\beta}}_{\text{Exp}}\right)
\end{aligned}
$$

Focusing on the expression of the exponential, we have

$$
\begin{aligned}
\text{Exp} &= (y-X_2\Delta\boldsymbol{\beta})'(y-X_2\Delta\boldsymbol{\beta})+\Delta\boldsymbol{\beta}'g_{\hat{A}}\Sigma_X\Delta\boldsymbol{\beta}+(\boldsymbol{\beta}_1-\bar{\boldsymbol{\beta}}_1)'\Omega^{-1}(\boldsymbol{\beta}_1-\bar{\boldsymbol{\beta}}_1)-\bar{\boldsymbol{\beta}}_1{}'\Omega^{-1}\bar{\boldsymbol{\beta}}_1, \\
&= y'M_{X_1}y-\bar{\boldsymbol{\mu}}'\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}}+(\Delta\boldsymbol{\beta}-\bar{\boldsymbol{\mu}})'\bar{\Sigma}^{-1}(\Delta\boldsymbol{\beta}-\bar{\boldsymbol{\mu}})+(\boldsymbol{\beta}_1-\bar{\boldsymbol{\beta}}_1)'\Omega^{-1}(\boldsymbol{\beta}_1-\bar{\boldsymbol{\beta}}_1), \\
&= \frac{g_{\hat{A}}}{1+g_{\hat{A}}}s_{X_1}+\frac{1}{(1+g_{\hat{A}})}s_{X_1,X_2}+(\Delta\boldsymbol{\beta}-\bar{\boldsymbol{\mu}})'\bar{\Sigma}^{-1}(\Delta\boldsymbol{\beta}-\bar{\boldsymbol{\mu}})+(\boldsymbol{\beta}_1-\bar{\boldsymbol{\beta}}_1)'\Omega^{-1}(\boldsymbol{\beta}_1-\bar{\boldsymbol{\beta}}_1),
\end{aligned}
$$

where $\bar{\Sigma}^{-1}=(1+g_{\hat{A}})X_2'M_{X_1}X_2=(1+g_{\hat{A}})\Sigma_X$ and $\bar{\boldsymbol{\mu}}=\bar{\Sigma}X_2'M_{X_1}y$, $\bar{\boldsymbol{\mu}}'\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}}=(1+g_{\hat{A}})^{-1}y'M_{X_1}X_2[X_2'M_{X_1}X_2]^{-1}X_2'M_{X_1}y$ and $\Omega^{-1}=X_1'X_1$, $\bar{\boldsymbol{\beta}}_1=(X_1'X_1)^{-1}X_1'(y-X_2\Delta\boldsymbol{\beta})$.

The posterior distribution can be decomposed as

$$\begin{aligned} f(\boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \boldsymbol{\tau}) &= f(\sigma^2|\mathbf{y}, \boldsymbol{\tau}) f(\Delta\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\tau}, \sigma^2) f(\boldsymbol{\beta}_1|\mathbf{y}, \boldsymbol{\tau}, \sigma^2, \Delta\boldsymbol{\beta}) \\ &\propto (\sigma^2)^{-\frac{(T+2-K)}{2}} \exp \frac{-1}{\sigma^2} \{ \frac{\frac{g_{\hat{A}}}{1+g_{\hat{A}}} s_{X_1} + \frac{1}{(1+g_{\hat{A}})} s_{X_1,X_2}}{2} \} \\ & (\sigma^2)^{-\frac{(k_{\hat{A}})}{2}} |g_{\hat{A}}(X_2)' M_{X_1} X_2)|^{1/2} \exp \frac{-1}{2\sigma^2} \{ (\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})' \bar{\Sigma}^{-1} (\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}}) \} \\ & (\sigma^2)^{-\frac{(K)}{2}} \exp \frac{-1}{2\sigma^2} \{ (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1)' \Omega^{-1} (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1) \}. \end{aligned}$$

It gives the following posterior distribution

$$\begin{aligned} \sigma^2|\mathbf{y}, \boldsymbol{\tau} &\sim \mathcal{IG}(\frac{T-K}{2}, \frac{\frac{g_{\hat{A}}}{1+g_{\hat{A}}} s_{X_1} + \frac{1}{(1+g_{\hat{A}})} s_{X_1,X_2}}{2}), \\ \Delta\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\tau}, \sigma^2 &\sim \mathcal{N}((1+g_{\hat{A}})^{-1} [X_2' M_{X_1} X_2]^{-1} X_2' M_{X_1} y, \sigma^2 (1+g_{\hat{A}})^{-1} [X_2' M_{X_1} X_2]^{-1}), \\ \boldsymbol{\beta}_1|\mathbf{y}, \boldsymbol{\tau}, \sigma^2, \Delta\boldsymbol{\beta} &\sim \mathcal{N}((X_1'X_1)^{-1} X_1'(y - X_2 \Delta\boldsymbol{\beta}), \sigma^2 (X_1'X_1)^{-1}). \end{aligned}$$

B.2 Predictive density

In Appendix B.1, we derive the following posterior distributions:

$$\begin{aligned} \sigma^2|\mathbf{y}, \boldsymbol{\tau} &\sim \mathcal{IG}(\underbrace{\frac{T-K}{2}}_{a_{\sigma^2}}, \underbrace{\frac{\frac{g_{\hat{A}}}{1+g_{\hat{A}}} s_{X_1} + \frac{1}{(1+g_{\hat{A}})} s_{X_1,X_2}}{2}}_{b_{\sigma^2}}), \\ \Delta\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\tau}, \sigma^2 &\sim \mathcal{N}(\underbrace{(1+g_{\hat{A}})^{-1} [X_2' M_{X_1} X_2]^{-1} X_2' M_{X_1} y}_{\boldsymbol{\mu}_{\Delta\boldsymbol{\beta}}}, \underbrace{\sigma^2 (1+g_{\hat{A}})^{-1} [X_2' M_{X_1} X_2]^{-1}}_{\boldsymbol{\Sigma}_{\Delta\boldsymbol{\beta}}}), \\ \boldsymbol{\beta}_1|\mathbf{y}, \boldsymbol{\tau}, \sigma^2, \Delta\boldsymbol{\beta} &\sim \mathcal{N}(\underbrace{(X_1'X_1)^{-1} X_1'(y - X_2 \Delta\boldsymbol{\beta})}_{\boldsymbol{\mu}_{\boldsymbol{\beta}}}, \underbrace{\sigma^2 (X_1'X_1)^{-1}}_{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}}). \end{aligned}$$

Given these results, we can derive the joint posterior distribution of the variable $\boldsymbol{\psi} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \Delta\boldsymbol{\beta} \end{pmatrix}$. In particular, a standard algebraic calculus leads to

$$\begin{aligned} \boldsymbol{\psi}|\mathbf{y}, \boldsymbol{\tau}, \sigma^2 &\sim \mathcal{N}\left(\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{B}\mu_{\Delta\boldsymbol{\beta}} \\ \boldsymbol{\mu}_{\Delta\boldsymbol{\beta}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{\Sigma}_{\beta}^{-1}\boldsymbol{B} \\ \boldsymbol{B}'\boldsymbol{\Sigma}_{\beta}^{-1} & [\boldsymbol{\Sigma}_{\Delta\boldsymbol{\beta}}^{-1} + \boldsymbol{B}'\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{B}] \end{pmatrix}^{-1} \right), \\ &\sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\psi}}, \boldsymbol{\Sigma}_{\boldsymbol{\psi}}\right) \end{aligned} \tag{B.1}$$

with $\hat{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'y$ and $\boldsymbol{B} = (X_1'X_1)^{-1}X_2$. Consequently, the predictive density is given by

$$y_{T+1}|\mathbf{y}, \boldsymbol{\tau}, \sigma^2 \sim \mathcal{N}\left(\mathbf{x}_{T+1}'\boldsymbol{\mu}_{\boldsymbol{\psi}}, \sigma^2(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\psi}\mathbf{x}_{T+1} + 1)\right). \tag{B.2}$$

Since $\sigma^2|\mathbf{y}, \boldsymbol{\tau}$ follows an inverse gamma distribution, the predictive distribution of $y_{T+1}|\mathbf{y}$ is a student distribution. Its density is given by

$$\begin{aligned} f(y_{T+1}|\mathbf{y}, \boldsymbol{\tau}) &= \frac{b_{\sigma^2}^{a_{\sigma^2}}}{\Gamma(a_{\sigma^2})}(2\pi(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{x}_{T+1} + 1))^{-\frac{1}{2}} \\ &\quad \int (\sigma^2)^{-(a_{\sigma^2}+1+0.5)} \exp\left(-\frac{1}{\sigma^2}[\frac{(y_{T+1} - \mathbf{x}_{T+1}'\boldsymbol{\mu}_{\boldsymbol{\psi}})^2(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{x}_{T+1} + 1)^{-1} + 2b_{\sigma^2}}{2}] \right) d\sigma^2, \\ &= \frac{b_{\sigma^2}^{a_{\sigma^2}}}{\Gamma(a_{\sigma^2})}(2\pi(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{x}_{T+1} + 1))^{-\frac{1}{2}}\Gamma(a_{\sigma^2} + 0.5) \\ &\quad \left(\frac{(y_{T+1} - \mathbf{x}_{T+1}'\boldsymbol{\mu}_{\boldsymbol{\psi}})^2(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{x}_{T+1} + 1)^{-1} + 2b_{\sigma^2}}{2} \right)^{-(a_{\sigma^2}+0.5)}, \end{aligned} \tag{B.3}$$

The final expression in Equation (B.3) is equivalent to a student density with expectation $\mathbf{x}_{T+1}'\boldsymbol{\mu}_{\boldsymbol{\psi}}$, scale parameter $\frac{b_{\sigma^2}}{a_{\sigma^2}}(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\psi}\mathbf{x}_{T+1} + 1)$ and degree of freedom equal to $2a_{\sigma^2}$.

C Consistency of the criterion

To prove the theorem, we focus on the ratio of the criterion for two different models $s = (a_s, \lambda_s)$ and $j = (a_j, \lambda_j)$ where s is considered as the true model. To simplify the notation, we denote by $\mathbf{X}_z$ the explanatory variable included by model z (i.e., $\mathbf{X}_z = \mathbf{X}_{\boldsymbol{\tau}}^{\hat{A}_z}$) for $z = s, j$ and $g_{\hat{A}} = g = \frac{1}{w(T)}$ and write the marginal likelihood as $f(\mathbf{y}|a_z, \lambda_z)$ instead of $f(\mathbf{y}|a_z, \lambda_z, \boldsymbol{\tau})$. We

need to show that $\frac{f(\mathbf{y}|a_j,\lambda_j)}{f(\mathbf{y}|a_s,\lambda_s)} \to_p 0$ for any $j \neq s$. In particular, we have

$$\frac{f(\mathbf{y}|a_j,\lambda_j)}{f(\mathbf{y}|a_s,\lambda_s)} = \underbrace{\frac{\left(\frac{g}{1+g}\right)^{k_{\hat{A}_j}/2}}{\left(\frac{g}{1+g}\right)^{k_{\hat{A}_s}/2}}}_{C_{js}} \underbrace{\left[\frac{\frac{g}{1+g}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g)}s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}{\frac{g}{1+g}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g)}s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}\right]^{-\frac{T-K}{2}}}_{D_{js}}. \tag{C.1}$$

Focusing on the first term, it is easy to show that

$$\begin{aligned} C_{js} &= \frac{(1+g)^{k_{\hat{A}_s}/2}}{(1+g)^{k_{\hat{A}_j}/2}} g^{\frac{k_{\hat{A}_j}-k_{\hat{A}_s}}{2}} \\ &= \frac{(1+w(T)^{-1})^{k_{\hat{A}_s}/2}}{(1+w(T)^{-1})^{k_{\hat{A}_j}/2}} w(T)^{\frac{k_{\hat{A}_s}-k_{\hat{A}_j}}{2}} \\ &= \mathcal{O}(w(T)^{\frac{k_{\hat{A}_s}-k_{\hat{A}_j}}{2}}). \end{aligned}$$

When $T \to \infty$, we have

$$\begin{aligned} C_{js} &= 0 \text{ when } k_{\hat{A}_s} < k_{\hat{A}_j}, \\ &= 1 \text{ if } k_{\hat{A}_s} = k_{\hat{A}_j}, \\ &\to +\infty \text{ when } k_{\hat{A}_s} > k_{\hat{A}_j}. \end{aligned}$$

We now discuss three possible cases.

1. $k_{\hat{A}_s} < k_{\hat{A}_j}$ and the model j does not nest the model s. In such case, the term $C_{js} \to 0$. The second term also tends to zero since we have

$$\begin{aligned} D_{js} &= \left[\frac{\frac{g}{1+g}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g)}s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{\frac{g}{1+g}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g)}s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}\right]^{\frac{T-K}{2}} \\ &= \left[\frac{gs_{\tilde{\mathbf{X}}_{\tau_0}} + s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{gs_{\tilde{\mathbf{X}}_{\tau_0}} + s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}\right]^{\frac{T-K}{2}} \end{aligned}$$

   Using he fact that M_j does not nest M_s and the Frisch-Waugh theorem (see also Lemma A.1 in Fernandez, Ley, and Steel (2001)), we have that $\lim_{T\to\infty} \frac{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}{T} = \sigma^2 + b_j$ with $b_j > 0$. Combining with the fact that $g \to 0$, we end up with a limit of D_{js} given by

$$\lim_{T\to\infty} D_{js} = \left[\frac{\sigma^2}{\sigma^2 + b_j}\right]^{\frac{T-K}{2}} \to 0.$$

2. The model j does not nest the true model but $K_j < K_s$. In such case, the term $C_{js} \to +\infty$. However, we can show that $\lim_{T\to\infty} C_{js} w(T)^{-\frac{(K_s-K_j)}{2}+\frac{K_s-K_j}{T-K}} \to 1$. Indeed, we have that

$$\lim_{T\to\infty} C_{js} w(T)^{-\frac{(K_s-K_j)}{2}+\frac{K_s-K_j}{T-K}} = \frac{(1+w(T)^{-1})^{k_{\hat{A}_s}/2}}{(1+w(T)^{-1})^{k_{\hat{A}_j}/2}} w(T)^{\frac{K_s-K_j}{T-K}}.$$

Let us define $q_T = w(T)^{\frac{K_s-K_j}{T-K}}$. We can compute the limit as follows $\lim_{T\to\infty} w(T)^{\frac{K_s-K_j}{T-K}} = \lim_{T\to\infty} \exp \ln q_T$. The limit of $\ln q_T$ is given by

$$\begin{aligned}\lim_{T\to\infty} q_T &= \lim_{T\to\infty} \frac{K_s-K_j}{T-K} \ln w(T), \\ &= \lim_{T\to\infty} \frac{w'(T)}{w(T)} \quad (=0 \quad \text{by assumption}).\end{aligned}$$

We conclude that $\lim_{T\to\infty} w(T)^{\frac{K_s-K_j}{T-K}} = 1$. Now, we need to show that $D_{js} w(T)^{\frac{(K_s-K_j)}{2}-\frac{K_s-K_j}{T-K}} \to 0$. In fact, we have

$$\begin{aligned}D_{js} w(T)^{\frac{\frac{(K_s-K_j)}{2}(T-K-2)}{T-K}} &= \lim_{T\to\infty} \underbrace{(\frac{\sigma^2}{\sigma^2+b_j})}_{a<1}^{\frac{T-K}{2}} w(T)^{\frac{(K_s-K_j)}{2}} \\ &= \lim_{T\to\infty} \frac{w(T)^{\frac{(K_s-K_j)}{2}}}{a^{-\frac{T-K}{2}}},\end{aligned}$$

By applying $\left\lceil \frac{(K_s-K_j)}{2} \right\rceil$ times the Hospital's rule, we find that $a^{\frac{T-K}{2}}$ dominates and so $D_{js} w(T)^{\frac{\frac{(K_s-K_j)}{2}(T-K-2)}{T-K}} \to 0$.

3. We now consider the last case in which the model j nests the true model s. Consequently, we have $K_s < K_j$ and the term $C_{js} \to 0$. Regarding the other term, we can express it as

$$\begin{aligned}D_{js} &= [\frac{gs_{\tilde{\mathbf{X}}_{\tau_0}} + s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{gs_{\tilde{\mathbf{X}}_{\tau_0}} + s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}]^{\frac{T-K}{2}}, \\ &= \underbrace{[\frac{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}]^{\frac{T-K}{2}}}_{Q_1} \underbrace{[\frac{A_s+w(T)}{A_j+w(T)}]^{\frac{T-K}{2}}}_{Q2},\end{aligned}$$

where $A_i = \frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_i}}$ for $i = j, s$. It is clear that the first term Q_1 has a limiting distribution related to the likelihood ratio test. In fact, we have that

$$\begin{aligned}\frac{T-K}{2}\ln\frac{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}} &= \underbrace{\frac{T-K}{2T}}_{\to\frac{1}{2}} \underbrace{T\ln\frac{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}}_{\to_d \chi^2(\Delta_{js})}, \\ &\to_d \text{Gamma}(\frac{\Delta_{js}}{2}, 1),\end{aligned}$$

in which $\Delta_{js} = |K_s - K_j|$. Since $Y \sim \text{Gamma}(\frac{\Delta_{js}}{2}, 1)$ is $\mathcal{O}_p(1)$, we have that $C_{js}\exp Y \to_p 0$.

Focusing on the second term Q_2, using assumption (iii), we have that

$$\begin{aligned}\frac{T-K}{2}\ln[\frac{A_s+w(T)}{A_j+w(T)}] &= \frac{T-K}{2}\ln[1+\frac{A_s-A_j}{A_j+w(T)}], \\ &= \mathcal{O}_p(\frac{T}{w(T)}). \\ &\to_p [0,\infty).\end{aligned}$$

Since $C_{js} \to 0$, we conclude that $C_{js}Q_1Q_2 \to_p 0$.

C.1 Convergence to the BIC

In this appendix, we show that when $g_{\hat{A}} = \frac{1}{T^\alpha}$ with $\alpha > 1$, the marginal likelihood given by

$$f(\mathbf{y}|\boldsymbol{a},\lambda,\boldsymbol{\tau}) = (\frac{g_{\hat{A}}}{1+g_{\hat{A}}})^{k_{\hat{A}}/2}[\frac{g_{\hat{A}}}{1+g_{\hat{A}}}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g_{\hat{A}})}s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}}]^{-\frac{T-K}{2}}, \tag{C.2}$$

tends to the same penalty function than the BIC with $\alpha k_{\hat{A}}$ parameters. We have the following results:

$$\begin{aligned}f(\mathbf{y}|\boldsymbol{a},\lambda,\boldsymbol{\tau}) &= T^{-\frac{\alpha k_{\hat{A}}}{2}}[\frac{1}{T^\alpha}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{T^\alpha-1}{T^\alpha}s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}}]^{-\frac{T-K}{2}}, \\ &= T^{-\frac{\alpha k_{\hat{A}}}{2}}[s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}}]^{-\frac{T-K}{2}}[\frac{T^\alpha-1}{T^\alpha}]^{-\frac{T-K}{2}}[\frac{1}{T^\alpha-1}\frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}}}+1]^{-\frac{T-K}{2}}, \\ &= T^{-\frac{\alpha k_{\hat{A}}}{2}}[s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}}]^{-\frac{T-K}{2}}\underbrace{[1-\frac{1}{T^\alpha}]^{-\frac{T-K}{2}}}_{C_1}\underbrace{[\frac{1}{T^\alpha-1}\frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}}}+1]^{-\frac{T-K}{2}}}_{C_2},\end{aligned} \tag{C.3}$$

We now show that the two quantities, i.e. C_1 and C_2, tends to 1 when $T \to +\infty$:

$$\begin{aligned}
C_1 &= \exp\left(-\frac{T-K}{2}\ln[1-\frac{1}{T^\alpha}]\right), \\
&\to_p 1 \text{ since } \alpha > 1, \\
C_2 &= \exp\left(-\frac{T-K}{2}\ln[\frac{1}{T^\alpha-1}\frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}}+1]\right), \\
&\to_p 1 \text{ since } \alpha > 1 \text{ and } \frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}} = \mathcal{O}_p(1), \\
\ln f(\mathbf{y}|\boldsymbol{a},\lambda,\boldsymbol{\tau}) &\to_p -\frac{T}{2}\ln s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}} - \frac{\alpha k_{\hat{A}}}{2}\ln T, \\
&= -\frac{T}{2}\ln s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}} - \frac{k_{\hat{A}}+\hat{m}_{\hat{A}}-1}{2}\ln T,
\end{aligned} \tag{C.4}$$

As a consequence, the log of the criterion tends to

$$\ln f(\mathbf{y}|\boldsymbol{a},\lambda,\boldsymbol{\tau}) \to_p -\frac{T}{2}\ln s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}} - \frac{\alpha k_{\hat{A}}}{2}\ln T. \tag{C.5}$$

D Time-varying parameter model

We also consider a standard time-varying parameter process (see Primiceri (2005)). The model specification is given by

$$
\begin{aligned}
y_t &= \mathbf{x}_t'\boldsymbol{\beta}_0 + \mathbf{x}_t'\text{diag}(\omega_1, ..., \omega_K)\boldsymbol{\beta}_t + \sigma_t\epsilon_t, && \text{(D.1)}\\
\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1} &\sim \mathcal{N}(\boldsymbol{\beta}_{t-1}, I_K), && \text{(D.2)}\\
\ln\sigma_t^2 &= \ln\sigma_{t-1}^2 + \eta_t, && \text{(D.3)}
\end{aligned}
$$

where $\eta_t \sim \mathcal{N}(0, q)$ with $q \sim \mathcal{IG}(4, 1.5)$, $(\boldsymbol{\beta}_0', \omega_1, ..., \omega_K) \sim \mathcal{N}(0, I_{2K})$ and K stands for the number of explanatory variables. The model parameters can be estimated with a standard Markov-chain Monte Carlo (e.g., Bitto and Frühwirth-Schnatter, 2019). In order to take into account the autocorrelation structure, we use the same lag orders as exposed in Table 4. The other explanatory variables consist in an intercept and the seven factors.

References

Agarwal, V., and N. Naik (2004): "Risks and portfolio decisions involving hedge funds," *Review of Financial Studies*, 17, 63–98.

Andrews, D. W. (1993): "Tests for parameter instability and structural change with unknown change point," *Econometrica: Journal of the Econometric Society*, pp. 821–856.

Ardia, D., A. Dufays, and C. Ordas (2019): "Change-point segmentation: the Bayesian bridge," *Mimeo*.

Bai, J., and P. Perron (1998): "Estimating and testing linear models with multiple structural breaks," *Econometrica: Journal of the Econometric Society*, 66(1), 47–48.

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

Bauwens, L., A. Dufays, and B. De Backer (2011): "Estimating and forecasting structural breaks in financial time series," *Journal of Empirical Finance*, Forthcoming, DOI: 10.1016/j.jempfin.2014.06.008.

Bauwens, L., G. Koop, D. Korobilis, and J. Rombouts (2015): "The contribution of structural break models to forecasting macroecomic series," *Journal of Applied Econometrics*, 30(4), 596–620.

Bitto, A., and S. Frühwirth-Schnatter (2019): "Achieving shrinkage in a time-varying parameter model framework," *Journal of Econometrics*, 210(1), 75–97.

Bollen, N., and R. E. Whaley (2009): "Hedge fund risk dynamics: Implications for performance appraisal," *Journal of Finance*, 64, 985–1035.

Chan, J. C., G. Koop, R. Leon-Gonzalez, and R. W. Strachan (2012): "Time varying dimension models," *Journal of Business & Economic Statistics*, 30(3), 358–367.

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

Chernoff, H., and S. Zacks (1964): "Estimating the current mean of a normal distribution which is subjected to changes in time," *The Annals of Mathematical Statistics*, 35(3), 999–1018.

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

Dicker, L., B. Huang, and X. Lin (2013): "Variable selection and estimation with the seamless-L 0 penalty," *Statistica Sinica*, 23, 929–962.

Dufays, A., and J. Rombouts (2018): "Relevant parameter changes in structural break models," *Working paper*.

Eckley, I. A., P. Fearnhead, and R. Killick (2011): *Analysis of changepoint models*chap. 10, pp. 205–224. Cambridge University Press, Cambridge.

Eo, Y. (2016): "Structural changes in inflation dynamics: multiple breaks at different dates for different parameters," *Studies in Nonlinear Dynamics & Econometrics*, 20(3), 211–231.

Fan, J., and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., H. Peng, et al. (2004): "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, 32(3), 928–961.

Fearnhead, P., and Z. Liu (2007): "On-line inference for multiple changepoint problems," *Journal of Royal Statistical Society, Series B*, 69 (4), 589–605.

Fernandez, C., E. Ley, and M. F. Steel (2001): "Benchmark priors for Bayesian model averaging," *Journal of Econometrics*, 100(2), 381–427.

Fryzlewicz, P., et al. (2014): "Wild binary segmentation for multiple change-point detection," *The Annals of Statistics*, 42(6), 2243–2281.

Fung, W., and D. Hsieh (1997): "Empirical characteristics of dynamic trading strategies: The case of hedge funds," *Review of Financial Studies*, 10, 275–302.

Fung, W., and D. Hsieh (2001): "The risk in hedge fund strategies: theory and evidence from trend followers," *The review of financial studies*, 14(2), 313–341.

Fung, W., D. Hsieh, N. Naik, and T. Ramodarai (2008): "Hedge funds: Performance, risk and capital formation," *Journal of Finance*, 63, 1777–1803.

Fung, W., and D. A. Hsieh (2004): "Hedge fund benchmarks: A risk-based approach," *Financial Analysts Journal*, 60(5), 65–80.

Getmansky, M., A. W. Lo, and I. Makarov (2004): "An econometric model of serial correlation and illiquidity in hedge fund returns," *Journal of Financial Economics*, 74(3), 529–609.

Geweke, J., and S. Porter-Hudak (1983): "The Estimation and Application of Long Memory Time Series Models," *Journal of Time Series Analysis*, 4, 221–238.

Giordani, P., and R. Kohn (2008): "Efficient Bayesian inference for multiple change-point and mixture innovation models," *Journal of Business and Economic Statistics*, 26, 66–77.

Hamilton, J. (1989): "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, 57, 357–384.

Ishwaran, H., and J. S. Rao (2005): "Spike and slab variable selection: Frequentist and Bayesian strategies," *The Annals of Statistics*, 33(2), 730–773.

Killick, R., P. Fearnhead, and I. A. Eckley (2012): "Optimal detection of change-points with a linear computational cost," *Journal of the American Statistical Association*, 107(500), 1590–1598.

Kim, J., and H.-J. Kim (2016): "Consistent model selection in segmented line regression," *Journal of statistical planning and inference*, 170, 106–116.

Koop, G., and D. Korobilis (2012): "Forecasting inflation using dynamic model averaging," *International Economic Review*, 53(3), 867–886.

Koop, G., R. Leon-Gonzalez, and R. W. Strachan (2009): "On the evolution of the monetary policy transmission mechanism," *Journal of Economic Dynamics and Control*, 33(4), 997–1017.

Korkas, K. K., and P. Fryzlewiczv (2017): "Multiple change-point detection for non-stationary time series using wild binary segmentation," *Statistica Sinica*, 27, 287–311.

Liao, W. (2008): "Bayesian Inference of Structural Breaks in Time Varying Volatility Models," *Working Paper, New-York University*.

Liu, J., S. Wu, and J. V. Zidek (1997): "On Segmented Multivariate Regressions," *Statistica Sinica*, 7, 497–525.

Maheu, J., and Y. Song (2013): "A new structural break model, with an application to canadian inflation forecasting," *International journal of forecasting*, 30, 144–160.

Meligkotsidou, L., and I. D. Vrontos (2008): "Detecting structural breaks and identifying risk factors in hedge fund returns: A Bayesian approach," *Journal of Banking & Finance*, 32(11), 2471–2481.

Mitchell, M., and T. Pulvino (2001): "Characteristics of risk and return in risk arbitrage," *Journal of Finance*, 56, 2135–2175.

Patton, A., T. Ramodarai, and M. Streatfield (2015): "Change you can believe in? Hedge fund data revisions," *Journal of Finance*, 70, 963–999.

Perron, P., et al. (2006): "Dealing with structural breaks," *Palgrave handbook of econometrics*, 1(2), 278–352.

Primiceri, G.-E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *Review of Economic Studies*, 72, 821–852.

Raftery, A. E., M. Kárnỳ, and P. Ettler (2010): "Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill," *Technometrics*, 52(1), 52–66.

Rapach, D. E., J. K. Strauss, and G. Zhou (2009): "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *Review of Financial Studies*, 23(2), 821–862.

Rigaill, G., É. Lebarbier, and S. Robin (2012): "Exact posterior distributions and model selection criteria for multiple change-point detection problems," *Statistics and computing*, 22(4), 917–929.

Stephens, D. A. (1994): "Bayesian Retrospective Multiple-Changepoint Identification," *Applied Statistics*, 1, 159–178.

Tibshirani, R. (1994): "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Ueda, N., and R. Nakano (1998): "Deterministic annealing EM algorithm," *Neural networks*, 11(2), 271–282.

Yau, C. Y., and Z. Zhao (2016): "Inference for multiple change points in time series via likelihood ratio scan statistics," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4), 895–916.

Yuan, M., and Y. Lin (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

Zhang, N. R., and D. O. Siegmund (2007): "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics*, 63(1), 22–32.

Zhang, Y., R. Li, and C.-L. Tsai (2010): "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, 105(489), 312–323.

Zhao, Q., V. Hautamäki, I. Kärkkäinen, and P. Fränti (2012): "Random swap EM algorithm for Gaussian mixture models," *Pattern Recognition Letters*, 33(16), 2120–2126.