TI 2011-023/4
Tinbergen Institute Discussion Paper

An Alternative Bayesian Approach to Structural Breaks in Time Series Models

Sjoerd van den Hauwe
Richard Paap
Dick J.C. van Dijk

Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

An Alternative Bayesian Approach to Structural Breaks in Time Series Models

Sjoerd van den Hauwe[1,2*] Richard Paap[1,2] Dick van Dijk[1,2]

[1] *Econometric Institute*

Erasmus University Rotterdam

[2] *Tinbergen Institute*

February 7, 2011

Abstract

We propose a new approach to deal with structural breaks in time series models. The key contribution is an alternative dynamic stochastic specification for the model parameters which describes potential breaks. After a break new parameter values are generated from a so-called baseline prior distribution. Modeling boils down to the choice of a parametric likelihood specification and a baseline prior with the proper support for the parameters. The approach accounts in a natural way for potential out-of-sample breaks where the number of breaks is stochastic. Posterior inference involves simple computations that are less demanding than existing methods. The approach is illustrated on nonlinear discrete time series models and models with restrictions on the parameter space.

Keywords: Structural breaks, Bayesian analysis, forecasting, MCMC methods, nonlinear time series.

JEL Classification: C11, C22, C51, C53, C63.

*We thank John Geweke and other participants of the first European Seminar on Bayesian Econometrics in 2010 in Rotterdam for helpful comments. Corresponding author: Sjoerd van den Hauwe, Tinbergen Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands. Tel.: +31-10-40811298, Fax: +31-10-4089162. *E-mail address*: vandenhauwe@ese.eur.nl

1 Introduction

Over the last two decades, empirical evidence showing that macroeconomic and financial time series are subject to occasional structural breaks in their statistical properties has mounted, see Stock and Watson (1996) and Andreou and Ghysels (2009), among many others. A prominent example in macroeconomics is the Great Moderation, referring to the large decline in volatility experienced by many macroeconomic time series in the first half of the 1980s, see McConnell and Perez-Quiros (2000); Stock and Watson (2002); Sensier and van Dijk (2004) and Kim et al. (2008), among others. In finance, the presence of structural breaks in predictive regression models for asset returns is by now well documented, see Pesaran and Timmermann (2002); Paye and Timmermann (2006); Rapach and Wohar (2006); Lettau and Van Nieuwerburgh (2008); Ravazzolo et al. (2008) and Pettenuzzo and Timmermann (forthcoming), among others.

Many empirical studies reporting evidence for structural changes in macroeconomic and financial time series make use of frequentist methods for detecting and dating such breaks, as developed by Andrews (1993); Andrews and Ploberger (1994); Bai and Perron (1998); Bai et al. (1998) and Qu and Perron (2007), among others; see Perron (2006) for a recent survey. These methods can be classified as 'historical' testing procedures (Andreou and Ghysels; 2009), in the sense that they are designed for testing for structural change and the identification of potential break dates ex-post for time series observations spanning a given historical, in-sample period.[1] Out-of-sample forecasting in the presence of structural breaks has presented a much bigger challenge when relying upon frequentist methods, see the survey of Clements and Hendry (2006). A Bayesian approach would much better suit this problem, in the sense that structural change can be made an inherent part of the statistical time series model, in particular including the possibility that breaks occur in the out-of-sample period. Surprisingly then, accounting for possible future breaks when constructing out-of-sample forecasts has not received much attention in the Bayesian literature on structural breaks, with the notable exceptions of Pesaran et al. (2006), Koop and Potter (2007), Maheu and Gordon (2008), and Geweke and

[1] A different strand of literature concerns testing for structural change 'in real time', i.e. monitoring whether new, incoming observations are consistent with a previously specified model, see Chu et al. (1996) and Zeileis et al. (2005), among others.

Jiang (2010).

In this paper we propose a new Bayesian approach to deal with structural breaks in time series models, with an explicit focus on the implications for out-of-sample forecasting. Following the previous literature, we define a structural break as a permanent change in the value of a parameter of the model or, in the Bayesian framework, of a likelihood function. We propose a new stochastic specification to describe the dynamic behavior of the parameter, which has a simple and intuitively appealing interpretation. In each period, with a particular probability a structural break occurs and in that case the new parameter value is generated by a so-called baseline prior distribution. If a break does not occur, the parameter value is equal to the value in the previous period. Put differently, the (conditional) distribution of the model parameter is a two-component mixture, where one component is the baseline prior distribution and the other component is degenerate at the parameter value in the previous period. The mixing probability for the first component is the probability of a structural break. The key advantage of this specification lies in the Bayesian procedures for estimation and forecasting. For estimation purposes, we derive a Markov chain Monte Carlo [MCMC] based algorithm for simulating from the posterior distribution of the model parameters. The posterior simulator boils down to straightforward sampling from three-component mixture distributions, where most weight is put on degenerate components. Our sampler is a single-move algorithm, for which it is well-known that convergence may be problematic (or at least slow). To solve this issue, we introduce a remix step in our sampler which bears similarities to the remixing step in Dirichlet process prior models. For forecasting purposes, the predictive distributions of future observations are also of the mixture type, with one component being the model under the no-break scenario and the other being the model integrated over the baseline prior in case of a break. If the forecast horizon grows, the probability of a break in the out-of-sample period increases and the latter mixture component gets more weight.

The baseline prior and its hyperparameters form a key component in forecasting exercises. Our model specification is such that in the case of a structural break the new parameter value is independently from the past drawn from this baseline prior distribution. However, by including a third layer in the model, this independence assumption may be relaxed and we can train the hyperparameters of the baseline

distribution. Such a strategy is common in marketing (see for example Rossi et al.; 2005) and applied to structural break models in Carlin et al. (1992), Pesaran et al. (2006) and Geweke and Jiang (2010). By doing so, regimes from the past do reveal information for future parameter values that is properly absorbed by the predictive distribution.

Our methodology to accommodate structural breaks in time series models is closely related to the independent, contemporary research by Geweke and Jiang (2010) and the methods by Maheu and Gordon (2008), who propose essentially a similar specification. However, we employ a different submodel representation for the dynamic behavior of the model parameters with favorable computational implications. The simulator of Geweke and Jiang (2010) requires that the regime parameters can be marginalized analytically. But, this requirement restricts the combinations of model and baseline prior distribution that can be considered. Moreover, their sampler requires potentially cumbersome tuning of the Metropolis–Hastings proposal distributions. Our simulator does not have these restrictions and can in principle be applied to any combination of model and baseline prior distribution. Maheu and Gordon (2008) also restrict their analysis to models in which the posterior distribution is of known form and, moreover, their estimation procedures require computationally intensive marginal likelihood evaluations and continuously updating of posterior model probabilities over time.

Our approach to structural breaks in fact offers two key advantages compared to other existing methods. Both are closely related to the desirable properties of structural break models as formulated by Koop and Potter (2007). The first advantage is that our specification allows for an *a priori* unknown number and timing of breaks. In particular, our approach naturally allows for the possibility that breaks may occur beyond the in-sample period. It is commonly recognized that allowing for future breaks is a necessary ingredient for realistic out-of-sample forecasting. Previous attempts to do so have certain limitations and drawbacks. Pesaran et al. (2006), for example, propose an out-of-sample extension of the Markovian model of Chib (1998). In this approach, structural breaks are modeled by means of a non-recurring Markov process, which requires the specification of the number of breaks that occur, both in- and out-of-sample, see also Koop and Potter (2007). Pesaran et al. (2006) circumvent this issue by applying Bayesian model averaging over distinct scenarios,

each with a specific number of breaks in the out-of-sample period. However, this procedure is computationally cumbersome, and still requires a specific plausible choice of the maximum number of breaks to happen over the forecast horizon, which may be difficult to set.[2] Our approach does not suffer from these problems by specifying the number of breaks to be stochastic both in- and out-of-sample.

The second main advantage of our specification is its ability to deal with structural breaks in various types of models. Previous approaches are confined to linear regression models (e.g. Maheu and Gordon; 2008; Geweke and Jiang; 2010) or models that can, at least conditionally, be written in Gaussian state-space form, as in the dynamic mixture models advocated by Gerlach et al. (2000); Giordani et al. (2007) and Giordani and Kohn (2008). By contrast, our set-up can be applied straightforwardly to different types of models (or likelihood functions) as well, including models for limited dependent variables, models for count data, and to copula models for describing the dependence between different time series. This flexibility is mostly due to the computational advantages offered by the proposed posterior simulator for our specification of structural breaks. This efficient sampling scheme is the result of analytically integrating out the break indicator variables. If the sample size is T, then each run of the simulator is of order $\mathcal{O}(T)$ and only requires evaluations of one-observation likelihoods and sampling from simple mixtures. Other simulators first integrate with respect to the regime-specific parameters to improve convergence, see Geweke and Jiang (2010) or Gerlach et al. (2000) for a similar solution in a (conditional) Gaussian state-space specification. Hence, the feasibility of these approaches relies on the computational ease of this integration step.

The outline of the remainder of this paper is as follows. Section 2 introduces the dynamic specification of the breaking process, analyzes its implications for out-of-sample forecasting and describes issued related to the choice of an appropriate baseline prior distribution and the probability of a structural break. Section 3 deals with the methods to simulate from the posterior distribution. Section 4 demonstrates the usefulness and wide applicability of our methods both for descriptive in-sample analysis and for constructing out-of-sample forecasts that incorporate potential future parameter change. This is done by means of four applications involving

[2] In the most extreme, but also unlikely case, this number is equal to the length of the forecasting horizon.

different types of models, including a Poisson count data model, a copula model, a probit model and an autoregressive model. A conclusion and discussion are given in Section 5. The appendices elaborate on issues related to the theoretical results and posterior simulation.

2 Modeling structural breaks

In this section we develop our modeling framework to deal with structural breaks. In Section 2.1 we discuss the model specification in detail and we compare our approach to related alternatives. In Section 2.2 we focus on the implications of our model specification for out-of-sample forecasting. The role of the baseline prior distribution and the probability of structural change are discussed in Sections 2.3 and 2.4.

2.1 Model specification

Let y_t be the time series variable of interest, which is observed in periods $t = 1, \ldots, T$, and let $\mathbf{y}^{k,l} = (y_k, y_{k+1}, \ldots, y_l)'$, $(1 \leq k < l \leq T)$. Hence, $\mathbf{y}^{1,T}$ denotes the complete set of time series observations in the in-sample period, which we will denote by $\mathbf{y}$ for notational convenience. Suppose the time series in period t is characterized by a distribution with probability density function [pdf] $p(y_t|\mathbf{y}^{1,t-1}, \theta_t)$, such that y_t may depend on its own past and a possibly time-varying parameter θ_t.[3]

At the outset, it is useful to remark that we consider the case of a single parameter θ_t solely to facilitate the exposition. Our specification can easily be extended to a multiple parameter setting. In that case, we may impose simultaneous breaks in all parameters or we may allow individual parameters to break independently while, of course, intermediate cases are possible as well. Similarly, although we restrict ourselves to univariate time series here, the modeling framework can easily be extended to a multivariate setting; see Section 4 for an illustration of both issues.

To allow for infrequent structural breaks in the model parameter we propose a stochastic process for θ_t. Specifically, the distribution of the model parameter in

[3]Of course it may depend on explanatory variables $\mathbf{x}_t$ as well, but to keep notation clear we do not mention this explicitly in the conditioning set.

period t is specified by the conditional density

$$p(\theta_t|\boldsymbol{\theta}^{1,t-1}) = p(\theta_t|\theta_{t-1}) = \pi f_0(\theta_t;\boldsymbol{\lambda}) + (1-\pi)\mathbb{I}_{\{\theta_t=\theta_{t-1}\}}, \qquad (t=2,\ldots,T), \quad (1)$$

and $\theta_1 \sim f_0(\cdot;\boldsymbol{\lambda})$, where $0 \leq \pi \leq 1$, f_0 is the pdf of a distribution that we call the 'baseline prior' for θ, which is characterized by hyperparameters $\boldsymbol{\lambda}$, and $\mathbb{I}_{\{A\}}$ is an indicator function that is equal to one if statement A is true and zero otherwise.[4] Hence, the conditional distribution of θ_t is a mixture of two components. With probability π a structural break occurs such that the parameter value changes, with the new value being sampled according to the baseline prior f_0, while with probability $1-\pi$ no break occurs and the distribution of θ_t is degenerate at the value from the previous period. Note that the conditional distributions in (1) result in a joint distribution for $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_T)'$, which we denote by $p(\boldsymbol{\theta})$.

Geweke and Jiang (2010) independently propose a similar approach to deal with structural breaks. A subtle difference (yet crucial for the estimation procedure) is that they explicitly introduce binary dummy variables s_t, $(t=2,\ldots,T)$, indicating the occurrence of a break ($s_t = 1$) or not ($s_t = 0$). Their model for the time-dependent parameters can then be written as

$$p(\theta_t|\boldsymbol{\theta}^{1,t-1},\mathbf{s}^{2,t}) = p(\theta_t|\theta_{t-1},s_t) = f_0(\theta_t)^{\mathbb{I}_{\{s_t=1\}}}\left(\mathbb{I}_{\{\theta_t=\theta_{t-1}\}}\right)^{1-\mathbb{I}_{\{s_t=1\}}},$$

where the break indicators s_t are assumed to be independent and $\mathcal{B}er(\pi)$. This auxiliary variable can be integrated out, which results in the same specification as in (1):

$$\begin{aligned} p(\theta_t|\theta_{t-1}) &= \sum_{s_t=0,1} p(\theta_t|\theta_{t-1},s_t)p(s_t) \\ &= \pi f_0(\theta_t) + (1-\pi)\mathbb{I}_{\{\theta_t=\theta_{t-1}\}}. \end{aligned}$$

Similarly, our suggested approach to structural breaks is related to the mixture innovation models of Giordani et al. (2007) and Giordani and Kohn (2008). The framework in these papers crucially depends on the assumption that the model can be written in Gaussian state-space form (at least conditionally) where the parameters

[4] In a statistical context where we use $\mathbb{I}_{\{\theta=\theta^*\}}$ as a distribution for θ, this means that θ is degenerate in θ^*, that is, $\mathsf{Pr}[\theta=\theta^*]=1$. Our notation has the same meaning as the Dirac delta $\delta_{\theta^*}(\theta)$.

are treated as the states. The state equations are specified such that the parameter values are sampled from a mixture of a degenerate and a Gaussian component. Specifically, the state equation is given by

$$\theta_t = \theta_{t-1} + K_t \eta_t, \qquad \eta_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\eta^2), \tag{2}$$

where the break indicators K_t have the same statistical properties as the s_t above. The state equation (2) can be written in terms of conditional density functions,

$$p(\theta_t|\theta_{t-1}, K_t) = K_t f_\eta(\theta_t - \theta_{t-1}) + (1 - K_t)\, \mathbb{I}_{\{\theta_t = \theta_{t-1}\}},$$

where f_η is the pdf of η_t. If we again analytically integrate out the indicator variable we can straightforwardly see the relation to our approach:

$$\begin{aligned} p(\theta_t|\theta_{t-1}) &= \sum_{K_t=0,1} p(\theta_t|\theta_{t-1}, K_t) p(K_t) \\ &= \pi f_\eta(\theta_t - \theta_{t-1}) + (1 - \pi)\mathbb{I}_{\{\theta_t = \theta_{t-1}\}}. \end{aligned} \tag{3}$$

In case of a break (3) implies that the *change* in the parameter value comes from f_η. In our approach θ will be a new value from the baseline prior f_0.

For computational reasons, the conditional Gaussian state-space approach requires f_η to be the pdf of a (mixed) normal distribution, otherwise the relevant sampling methods developed by Giordani and Kohn (2008) cannot be applied. Theoretically this would not be too restrictive as long as the support for the parameter is $(-\infty, \infty)$. If, however, the support is a subset of the real line or prior beliefs restrict the region (e.g. by truncation), this approach cannot be used anymore. Our framework is much more flexible with respect to distributional assumptions of the parameters, as we can simply opt for a baseline prior f_0 that has the appropriate features. We can even impose that θ can only take discrete values. Apart from this pro, our approach has additional computational advantages, which result from working with the 'reduced' form where the break indicators are marginalized, as will be explained in more detail in Section 3.

Furthermore, the approach of Giordani and Kohn (2008) not only requires the (mixed) Gaussian assumption of the state equation but also of the observation equation. A second major advantage of our approach is that it can be applied to any kind of parametric likelihood function $p(y_t|\mathbf{y}^{1,t-1}, \theta_t)$. Hence, the time series y_t may be

continuous, discrete, or even a combination of both. Moreover, any choice of baseline prior distribution for the parameters and likelihood function can be analyzed, as we will demonstrate in the discussion of the estimation procedure in Section 3 and the illustrations in Section 4.

In order to get a better understanding of the behavior implied by our chosen model specification, it is insightful to examine $p(\boldsymbol{\theta})$ by means of simulation. This is a form of prior predictive analysis as advocated by Lancaster (2004) and Geweke (2005). For initialization we should pick a baseline prior distribution with density f_0 and a breaking probability π. Two routes can be followed. In the first one we simulate a path $\{\theta_t\}_{t=1}^T$ by starting with $\theta_1 \sim f_0$ and subsequently using the conditional distributions $p(\theta_t|\theta_{t-1})$, $(t = 2, \ldots, T)$ as in (1). Alternatively, we may initialize a path $\boldsymbol{\theta}$ and then simulate iteratively from the full conditionals $p\left(\theta_t|\boldsymbol{\theta}_{[-t]}\right)$ for $t = 1, \ldots, T$, where $\boldsymbol{\theta}_{[-t]} = (\theta_1, \ldots, \theta_{t-1}, \theta_{t+1}, \ldots, \theta_T)'$. The first procedure is based on the decomposition $p(\boldsymbol{\theta}) = p(\theta_1) \prod_{t=2}^T p(\theta_t|\theta_{t-1})$, while the second one applies the Gibbs sampling principle. Because the latter also provides the basis for the posterior simulation scheme as described in Section 3, we discuss this approach in more detail.[5]

The model in (1) for the stochastic behavior of the parameters shows that $\{\theta_t\}_{t=1}^T$ is a first-order Markov chain, implying that the full conditional distribution of θ_t only depends on its two immediate neighbors. For θ_t, $(t = 2, \ldots, T-1)$, collecting terms from $p(\boldsymbol{\theta})$ gives

$$\begin{aligned} p\left(\theta_t|\boldsymbol{\theta}_{[-t]}\right) &\propto p(\theta_t|\theta_{t-1})p(\theta_{t+1}|\theta_t) \\ &\propto \pi^2 f_0(\theta_t) f_0(\theta_{t+1}) \\ &\quad + \pi(1-\pi) f_0(\theta_t) \mathbb{I}_{\{\theta_{t+1}=\theta_t\}} + \pi(1-\pi) f_0(\theta_{t+1}) \mathbb{I}_{\{\theta_t=\theta_{t-1}\}} \\ &\quad + (1-\pi)^2 \mathbb{I}_{\{\theta_{t-1}=\theta_t=\theta_{t+1}\}}. \end{aligned} \tag{4}$$

[5]Note that by comparing the two simulation strategies we can also check the validity of the Gibbs sampler. That is, we can check that it traverses the entire support of $p(\boldsymbol{\theta})$ and thereby also retrieves the marginal distributions for the θ_t's, which are given by the baseline prior f_0 (see Proposition A.1 in Appendix A).

Two scenarios are possible:

Scenario 1: $\theta_{t-1} = \theta_{t+1} \equiv \theta^*$**.** In this case θ_t comes from a mixture with two components:

$$\begin{aligned} \theta_t = \theta^* \quad & \text{with probability} \quad \propto (1-\pi)^2 + 2\pi(1-\pi) f_0(\theta^*), \\ \theta_t \sim f_0 \quad & \text{with probability} \quad \propto \pi^2 f_0(\theta^*). \end{aligned}$$

The first component in this mixture corresponds with the situation that the value of θ_t is equal to both its neighbors' value θ^*, which is the case if no breaks occur at t and $t+1$, or a single break occurs at either t or $t+1$ but the new parameter value after the break is identical to the value before. The second component captures the possibility that breaks occur at both t and $t+1$, in which case the value of θ_t is obtained from the baseline prior distribution f_0 (and by construction the value after the break at $t+1$ is again equal to the value at $t-1$).

Scenario 2: $\theta_{t-1} \neq \theta_{t+1}$**.** In this case θ_t comes from a mixture with three components:

$$\begin{aligned} \theta_t = \theta_{t+1} \quad & \text{with probability} \quad \propto \pi(1-\pi), \\ \theta_t = \theta_{t-1} \quad & \text{with probability} \quad \propto \pi(1-\pi), \\ \theta_t \sim f_0 \quad & \text{with probability} \quad \propto \pi^2. \end{aligned}$$

In this case, the three components correspond with the possibilities that (i) no break occurs at time t and (necessarily) a break occurs at $t+1$, (ii) a break occurs at time t and no break occurs at $t+1$, (iii) breaks occur at both t and $t+1$.

If we compare these situations, it shows that when both neighbors are the same, the probability of no break gets an intuitively expected extra 'interaction' weight via the last term in (4). As θ_1 and θ_T only have one neighbor their full conditionals indicate that with probability π they come from the baseline prior and with probability $1-\pi$ they equal their respective neighbors, that is, θ_2 and θ_{T-1}.

In sum, the full conditional distributions of the parameters are a mixture of the baseline prior f_0 and one or two – depending on the scenario – degenerate distributions. Simulating from these distributions is therefore straightforward and fast, also because the degenerate components get most weight.

2.2 Forecasting implications

One of the main reasons why times series models may perform poorly in terms of (out-of-sample) forecasting is the often incorrect assumption that model parameters are stable over time. As shown by Clements and Hendry (2001, 2006), among others, neglecting structural breaks that occur during the in-sample period may yield biased forecasts. As discussed in the introduction, various (frequentist and Bayesian) methods are available for detecting and modeling in-sample breaks, which may be used to annihilate the bias. However, if breaks have occurred in the past, it is likely that further structural breaks may occur during the out-of-sample period as well. Not accounting for this possibility will result in density forecasts that are tighter – as preferred by practioners –, though an essential type of uncertainty is simply neglected. In this section we demonstrate the implications of our modeling framework for out-of-sample forecasting, by examining how this uncertainty with regard to the possibility of future structural breaks affects the resulting density forecasts.

In a Bayesian context, density forecasts are given by the posterior predictive distribution, which combines the model structure, prior considerations and information revealed by the data. At time τ, the posterior predictive density $p(y_{\tau+1}|\mathbf{y}^{1,\tau})$ of $y_{\tau+1}$ can be demarginalized as

$$\begin{aligned} p(y_{\tau+1}|\mathbf{y}^{1,\tau}) &= \int p(y_{\tau+1}|\boldsymbol{\theta}^{1,\tau+1},\mathbf{y}^{1,\tau})p(\boldsymbol{\theta}^{1,\tau+1}|\mathbf{y}^{1,\tau})d\boldsymbol{\theta}^{1,\tau+1} \\ &= \int p(y_{\tau+1}|\theta_{\tau+1},\mathbf{y}^{1,\tau})p(\theta_{\tau+1}|\theta_\tau)p(\boldsymbol{\theta}^{1,\tau}|\mathbf{y}^{1,\tau})d\boldsymbol{\theta}^{1,\tau+1}, \end{aligned} \tag{5}$$

by using the first-order Markov property of $\{\theta_t\}$ and the conditional independence assumptions.[6] The first two densities of the integrand in (5) are given by the (hierarchical) model, while the third component is the posterior density of the model parameters based on the data up to and including time τ. If we apply the dynamic specification of the model parameters (1), this expression further breaks down to

$$\begin{aligned} p(y_{\tau+1}|\mathbf{y}^{1,\tau}) &= \pi \int p(y_{\tau+1}|\theta_{\tau+1},\mathbf{y}^{1,\tau})f_0(\theta_{\tau+1})p(\boldsymbol{\theta}^{1,\tau}|\mathbf{y}^{1,\tau})d\boldsymbol{\theta}^{1,\tau+1} \\ &\quad + (1-\pi)\int p(y_{\tau+1}|\theta_{\tau+1},\mathbf{y}^{1,\tau})\mathbb{I}_{\{\theta_{\tau+1}=\theta_\tau\}}p(\boldsymbol{\theta}^{1,\tau}|\mathbf{y}^{1,\tau})d\boldsymbol{\theta}^{1,\tau+1} \\ &= \pi p_0(y_{\tau+1}|\mathbf{y}^{1,\tau}) + (1-\pi)\int p(y_{\tau+1}|\theta_\tau,\mathbf{y}^{1,\tau})p(\boldsymbol{\theta}^{1,\tau}|\mathbf{y}^{1,\tau})d\boldsymbol{\theta}^{1,\tau}, \end{aligned} \tag{6}$$

[6] Conditional on θ_t, y_t is independent of the previous parameters $\boldsymbol{\theta}^{1,t-1}$.

where $p_0(y_t|\mathbf{y}^{1,t-1})$ is defined to be the marginal likelihood of y_t (possibly conditional on the past $\mathbf{y}^{1,t-1}$) under prior f_0. This result shows that the predictive distribution is a mixture of two components: (i) with probability π a structural break occurs, and we integrate over the baseline prior f_0 that generates the new but unknown parameter value, and (ii) with probability $1-\pi$ no break occurs, and we account for the uncertainty in θ_τ by integrating over the posterior distribution.

In the nested situation in which we do not allow for a structural break at $\tau+1$ ($\pi=0$) the predictive distribution reduces to the second part of the sum in (6). If there is a positive probability that a break occurs in the next period, the predictive probability mass is shifted in the direction of the marginal likelihood $p_0(y_\tau|\mathbf{y}^{1,\tau})$, resulting in a more dispersed density forecast. This mechanism becomes even more clear if we investigate longer forecast horizons, as shown next.

The predictive distribution for h periods ahead is given by[7]

$$p(y_{\tau+h}|\mathbf{y}^{1,\tau}) = \int p(y_{\tau+h}|\theta_{\tau+h})p(\boldsymbol{\theta}^{\tau+1,\tau+h}|\theta_\tau)p(\boldsymbol{\theta}^{1,\tau}|\mathbf{y}^{1,\tau})d\boldsymbol{\theta}^{1,\tau+h}.$$

The intermediate parameters $\boldsymbol{\theta}^{\tau,\tau+h-1}$ can be integrated out analytically by applying Proposition A.2 in Appendix A with the marginal posterior of θ_τ until time τ as initial distribution, i.e., take $g(\theta_\tau) = \int p(\boldsymbol{\theta}^{1,\tau}|\mathbf{y}^{1,\tau})d\boldsymbol{\theta}^{1,\tau-1}$. This results in

$$\begin{aligned} p(\theta_{\tau+h}|\mathbf{y}^{1,\tau}) &= \int p(\boldsymbol{\theta}^{\tau+1,\tau+h}|\theta_\tau)g(\theta_\tau)d\boldsymbol{\theta}^{\tau,\tau+h-1} \\ &= \left[1-(1-\pi)^h\right] f_0(\theta_{\tau+h}) + (1-\pi)^h g(\theta_{\tau+h}). \end{aligned}$$

Therefore, $\theta_{\tau+h}|\mathbf{y}^{1,\tau} \xrightarrow{D} f_0$ if the forecast horizon h becomes large. For very large h the parameter comes approximately from the baseline prior, which makes that $y_{\tau+h}|\mathbf{y}^{1,\tau}$ has the marginal likelihood under f_0 as its limiting distribution: $y_{\tau+h}|\mathbf{y}^{1,\tau} \xrightarrow{D} p_0$. Temporal dependence can be dealt with in a straightforward way by successively simulating intermediate y_t's $\left(\mathbf{y}^{\tau+1,\tau+h-1}\right)$ from the likelihood conditional on the most recently sampled parameter value.[8]

To summarize, the longer the forecast horizon the more the predictive probability mass gets spread according to the marginal likelihood, and possibly shifted away

[7]For notational convenience, here we suppress the (direct) temporal dependencies between the dependent variable in this expression, i.e., we write $p(y_t|\mathbf{y}^{1,t-1},\theta_t) = p(y_t|\theta_t)$.

[8]In case of stationary processes for y_t, this direct temporal dependence introduces a second convergence issue and the limiting distribution is not the 'one-observation' marginal likelihood p_0, but the unconditional distribution of y_t mixed over f_0.

from the constant parameter setting where $\theta_{\tau+h} = \theta_\tau$. This process is illustrated in the following simple example.

Example (Forecasting issues)**:** *Consider a simple normal linear regression model* [9] *that allows for structural breaks in the intercept and the variance:*

$$y_t|\mu_t, \sigma_t^2 \overset{i.i.d.}{\sim} \mathcal{N}(\mu_t, \sigma_t^2), \tag{7}$$

$$f_0(\mu, \sigma^2) = f_{\mathcal{N}}(\mu; b, \sigma^2 B) f_{\mathcal{IG}2}(\sigma^2; \nu, S), \tag{8}$$

where we assume that any breaks in the intercept and variance occur simultaneously. The baseline prior consists of a normal-inverted Gamma–2 distribution with hyperparameters b, B, ν and S. We examine the posterior predictive distributions for different horizons. We start forecasting at time τ where we assume we know $\mu_\tau = 5$ and $\sigma_\tau^2 = 1$. Figure 1 displays the forecasting characteristics in this model. The graphs in Figure 1(a) show the predictive densities for the different horizons. The solid line shows the likelihood under μ_τ and σ_τ^2, which is the pdf of a normal. The one-period ahead predictive distribution is depicted by the dashed line: we can already see the shift of probability mass due to the potential break. The dashed-dotted line is associated with $h = 20$. Obviously, the larger horizon h the closer the marginal likelihood (dotted line) is approximated; for $h = 100$ we are near the dotted line. Note that in this case the 'limiting' distribution is given by

$$p_0(y_{\tau+h}|\mathbf{y}^{1,\tau}) = \int p(y_{\tau+h}|\mu, \sigma^2) f_0(\mu, \sigma^2; b, B, \nu, S) d\mu d\sigma^2.$$

The integral can be evaluated analytically yielding $y_{\tau+h}|\mathbf{y}^{1,\tau} \sim \mathcal{T}(b, (B+1)/(S\nu), \nu)$, for h large. Figures 1(d)–(f) show the evolution of the distributions of the dependent variable, the intercept and the variance, respectively, over time. For the two parameters we can see that ultimately the theoretical marginals as plotted in Figures 1(b)–(c) are approximated. The solid line for μ indicates the marginal Student's t (by integrating out σ^2). For comparison, the dashed line is the pdf of a normal with variance σ^2 fixed at the Student's t's. ∎

[9] The example regression model includes only a constant for purposes of illustration; it can easily be augmented with explanatory variables without changing the argument.

2.3 Baseline prior choice

The baseline prior distribution is a key element in our modeling approach and, as just shown, it plays a crucial role in out-of-sample forecasting. It thus warrants further discussion. Two important considerations when choosing the baseline prior $f_0(\cdot; \boldsymbol{\lambda})$ are (i) the type of distribution and (ii) its hyperparameters $\boldsymbol{\lambda}$.

In our modeling framework, the baseline prior distribution gives birth to new parameter values in case of a structural break. As such it is one of the advantages of our approach, in the sense that restrictions on the model parameters, like positive support for variances, can easily be implemented through the specification of the baseline prior. Furthermore, the effect of the prior specification on forecasting can easily be analyzed. Our model specification does not put any restrictions on the prior. The prior distribution can either be conjugate or non-conjugate. The advantage of a conjugate prior is that it usually facilitates posterior simulation, but non-conjugate priors can also be dealt with easily, as discussed in Section 3.1.2.

Not less important than the type of the baseline prior distribution is the setting of the hyperparameters $\boldsymbol{\lambda}$. This crucially depends on the ultimate goal of the research. If it is mostly exploratory, that is, if we merely want to check for the possibility of structural breaks in the past, choosing an uninformative baseline prior makes sense, where we should ensure that it covers regions with plausible values sufficiently. If, however, the primary interest lies in constructing accurate forecasts, $\boldsymbol{\lambda}$ plays a major role.

As shown before, the predictive distribution is constructed by mixing the likelihood over the posterior and the baseline prior, where the latter gets more weight as the forecast horizon grows. Clearly, choosing a particular value for $\boldsymbol{\lambda}$ means that we fix the long run predictive distribution. This forms no problem when we have leading prior information to be imposed. However, if our prior knowledge is diffuse, this will result in relatively wide-spread predictive distributions. To circumvent the latter situation we may exploit the hierarchical model structure and introduce a third layer, that is, we may put a prior $p(\boldsymbol{\lambda})$ on these hyperparameters. This is a common strategy in Bayesian modeling, see, for example, Geweke (2005) for general comments and Pesaran et al. (2006) and Geweke and Jiang (2010) for a forecasting application.

This additional hierarchical layer in fact turns out to have several advantages.

To illustrate this, suppose there turn out to be $K-1$ structural breaks during the in-sample period, which implies we have K different regimes with parameters θ_k^*, $(k = 1, \ldots, K)$. Each element of $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_K^*)'$ is generated by the baseline f_0. Moreover, conditional on $\boldsymbol{\lambda}$ these K unique parameters are statistically independent (see Proposition A.1 in Appendix A). The first advantage of this extra model layer is that after marginalizing out $\boldsymbol{\lambda}$ the regime parameters *do* show dependence, which Koop and Potter (2007) list as a requirement for any structural breaks model. Second, and perhaps more important, it allows for a data-updating step to learn about $\boldsymbol{\lambda}$. Both advantages combined have the desirable effect that parameter values from the past provide information relevant for future regimes, properly assimilated in the predictive distributions.

In most hierarchical settings a conjugate prior for $\boldsymbol{\lambda}$ is implemented. Integrating out $\boldsymbol{\lambda}$, possibly through simulation, provides the marginal baseline prior $\int f_0(\theta_t; \boldsymbol{\lambda}) p(\boldsymbol{\lambda}) d\boldsymbol{\lambda}$. This marginal baseline prior provides insights in what values for θ_t are *a priori* covered, see Section 4 for an example. It is important to note that in general there will be a limited number of breaks and, hence, a limited number of unique θ_k^* values. Since these contain all the information in the data relevant for $\boldsymbol{\lambda}$, there may be little updating. Hence, $p(\boldsymbol{\lambda}|\boldsymbol{\theta}^*)$ may be close to $p(\boldsymbol{\lambda})$.

2.4 Probability of a structural break

Finally, some remarks concerning our specification of the structural break process are in order. In our set-up, the probability of a structural break, π, is constant over time. This implies that the duration of a regime (that is, the period of time a particular parameter value prevails, in between two consecutive breaks) has a geometric distribution. A theoretical drawback of this implication is that short durations get highest probability *a priori*. That is, if the duration $d \sim \mathcal{G}eo(\pi)$ then $\mathsf{Pr}[d = j|\pi] = \pi(1-\pi)^{j-1}$, $(j = 1, 2, \ldots)$. Koop and Potter (2007) argue for alternatives that do not impose this restriction, for example by opting for Poisson or history-dependent durations. Note that mixing the distribution of d over a prior $p(\pi)$ does not change the form of the marginal for d and its mode remains at $d = 1$.

Two sidemarks are in place with regard to this supposed drawback. First, because π is usually (very) small the dispersion of the resulting geometric distribution is large, assigning different durations pretty much an equal probability. This is in contrast

to the Poisson case where most of the mass is concentrated around its mean ω. However, there is no obvious, neither theoretical nor empirical, argument for why there would be a break every ω periods on average. Instead, empirical research shows that breaks seem to come in at arbitrary points in time instead of obeying a cyclical pattern.

Second, suppose we are about to enter time period t and define d_t to be the duration of the current regime, i.e., the period of time expired since the previous break. If this regime already lasted for j periods, the probability that it will die at time t in the geometric case is $\mathsf{Pr}\left[d_t = j | d_t \geq j\right] = \pi$, for all $j = 1, 2, \ldots$. In the Poisson case $\mathsf{Pr}\left[d_t = j | d_t \geq j\right] \longrightarrow 1$ if j becomes large, which means that occurrence of a break will eventually be enforced due to this regime duration specification. A similar problem arises during forecasting: if a regime already lasts a relatively (compared to ω) long time we will forecast a break with probability close to one. The apparent lack of predictability of the occurrence of structural breaks (Maheu and Gordon; 2008) pleads in favor of geometric durations.

As final remark, note that we may fix π to a specific value or, alternatively, treat it as an unknown model parameter (for which we then have to specify a prior distribution). From a non-statistical point of view we can interpret π as a smoothing parameter: the closer it is to zero the more bumpy behavior is penalized resulting in increased smoothness (less breaks). Because we model infrequent structural change it should take a small number. Often, prior thoughts give a hunch for the expected number of breaks and together with the sample size T we can fix π. Note that despite such a fixation the actual number of breaks is still random. As a full Bayesian alternative we can put a prior on this parameter. A Beta prior appears to be convenient, but any other distribution restricted to $[0, 1]$ is allowed. In case of a prior specification on π, we should take into account the danger of 'overfitting', which may occur because both the number of breaks and π are not set in advance. Giordani et al. (2007) provide examples of prior parameter settings of a (Beta) prior concentrated around small values to avoid this danger. Another alternative is to link π to covariates in a probit fashion to make it time-varying. This would simply introduce an additional hierarchical layer to the model and posterior simulation is straightforward.

3 Posterior simulation

In this section we discuss our procedure to simulate from the posterior distribution. We use simulation techniques from the class of MCMC methods, see, for example, Robert and Casella (2004). Section 3.1 deals with sampling of the time-varying parameters in $\boldsymbol{\theta}$ while Sections 3.2 and 3.3 discuss simulation of the baseline hyper-parameters $\boldsymbol{\lambda}$ and the breaking probability π, respectively. Our simulation approach is different from Geweke and Jiang (2010) and Gerlach et al. (2000), who first integrate with respect to the regime-specific parameters to improve convergence. Their estimation algorithms rely on the analytical tractability of these integrals which limits the combinations of model and baseline prior specifications that can be considered. Our simulator does not require this analytical integration step but instead uses a remixing step to improve convergence. Hence, it is not restricted to conjugate prior settings or to linear regression models or models which can be written in a (mixed) Gaussian state-space representation.

3.1 Time-dependent parameters

We simulate the time-dependent parameters $\{\theta_t\}_{t=1}^T$ conditional on the parameters $\boldsymbol{\lambda}$ and π. To facilitate notation, assume without loss of generality for the moment that $\boldsymbol{\lambda}$ and π are known or fixed, such that inference involves determining the characteristics of $p(\boldsymbol{\theta}|\mathbf{y})$. We start with analyzing the situation of a conjugate baseline prior distribution and likelihood function. In this setting we propose to employ a Gibbs sampler to sequentially sample from the full conditional posteriors $p\left(\theta_t|\mathbf{y}, \boldsymbol{\theta}_{[-t]}\right)$, for $t = 1, \ldots, T$. The non-conjugate setting is examined thereafter.

3.1.1 Conjugate setting

In Section 2.1 we have derived the full conditional *prior* distributions of θ_t in (4). Combining these with the likelihood $p(y_t|\mathbf{y}^{1,t-1}, \theta_t)$ makes that applying Bayes' rule results in the full conditional posterior distributions

$$p\left(\theta_t|\mathbf{y}, \boldsymbol{\theta}_{[-t]}\right) \quad \propto \quad p(y_t|\mathbf{y}^{1,t-1}, \theta_t) p\left(\theta_t|\boldsymbol{\theta}_{[-t]}\right). \tag{9}$$

Hence, the full conditional posteriors are also of the mixture form just like (4). Again we can consider the two possible scenarios for $t = 2, \ldots, T-1$ (the posteriors for

$t = 1$ and $t = T$ are again straightforward special cases):

Scenario 1: $\theta_{t-1} = \theta_{t+1} \equiv \theta^*$**.** In this case θ_t comes from a mixture with two components:

$$\theta_t = \theta^* \qquad \text{with probability} \quad \propto \left[2 + \frac{1-\pi}{\pi f_0(\theta^*)}\right] p(y_t|\mathbf{y}^{1,t-1}, \theta^*),$$

$$\theta_t \sim p(\theta_t|\mathbf{y}^{1,t}) \quad \text{with probability} \quad \propto \left[\frac{\pi}{1-\pi}\right] p_0(y_t|\mathbf{y}^{1,t-1}),$$

where, in order to get the appropriate mixture components and their respective weights, we use the identity

$$p(y_t|\mathbf{y}^{1,t-1}, \theta_t) f_0(\theta_t) = p_0(y_t|\mathbf{y}^{1,t-1}) p(\theta_t|\mathbf{y}^{1,t}), \tag{10}$$

where $p_0(y_t|\mathbf{y}^{1,t-1}) = \int p(y_t|\mathbf{y}^{1,t-1}, \theta_t) f_0(\theta_t) d\theta_t$ is the marginal likelihood of y_t under the baseline prior f_0 and $p(\theta_t|\mathbf{y}^{1,t}) \propto p(y_t|\mathbf{y}^{1,t-1}, \theta_t) f_0(\theta_t)$ is the posterior of θ_t conditional on data up to and including time t. The two components in this mixture again correspond with the situations that (i) no breaks occur at t and $t+1$ or a single break occurs at either t or $t+1$ but the new parameter value after the break is identical to the value before, and (ii) breaks occur at both points in time. In the former case θ_t is set equal to its neighboring values θ^*, whereas in the latter case its value is obtained from the posterior.

Scenario 2: $\theta_{t-1} \neq \theta_{t+1}$**.** In this case θ_t comes from a mixture with three components:

$$\begin{aligned}
\theta_t &= \theta_{t-1} && \text{with probability} && \propto p(y_t|\mathbf{y}^{1,t-1}, \theta_{t-1}), \\
\theta_t &= \theta_{t+1} && \text{with probability} && \propto p(y_t|\mathbf{y}^{1,t-1}, \theta_{t+1}), \\
\theta_t &\sim p(\theta_t|\mathbf{y}^{1,t}) && \text{with probability} && \propto \left[\frac{\pi}{1-\pi}\right] p_0(y_t|\mathbf{y}^{1,t-1}).
\end{aligned}$$

Now the possibilities comprise of (i) no break at time t and a break at $t+1$ (such that $\theta_t = \theta_{t-1}$), (ii) a break at t no break at $t+1$ (such that $\theta_t = \theta_{t+1}$), and (iii) breaks at both t and $t+1$ (such that θ_t is obtained from the 'one-observation' posterior).

One iteration of this Gibbs sampling scheme is performed in $\mathcal{O}(T)$ computing time. Moreover, sampling from the mixture distributions is straightforward. The only

part that may consume considerable computing time is formed by the observation-specific likelihood evaluations to get the mixture weights. Also, vectorization of the marginal likelihood evaluations (which is often possible) is computationally efficient (see Conley et al.; 2008). This contrasts with the methods proposed by Gerlach et al. (2000), which involve time-consuming matrix inversions and decompositions in every iteration and Kalman filter computations that are of order $\mathcal{O}(T^2)$.

As the sampler we propose is of the single-move type, it may suffer from slow convergence. In order to enhance convergence of the Markov chain, we implement a so-called remix step comparable to remixing in Dirichlet process prior models as described by Escobar and West (1995). After running one iteration of the above Gibbs sampler we obtain a particular value for $\boldsymbol{\theta}$. Conditional on this value we can construct subsamples (regimes), according to the break dates $S = \{t \mid \theta_t \neq \theta_{t-1}, t = 2, \ldots, T\}$. In case of $K - 1 = |S|$ breaks, we form K subsamples such that all observations within each subsample are characterized by the distribution $p(y_t|\mathbf{y}^{1,t-1}, \theta_t)$ with the same parameter value $\theta_t = \theta_k^*$, $(k = 1, \ldots, K)$. The index k follows the time order, i.e., θ_k^* is the parameter value of the regime that comes in time immediately after the regime with value θ_{k-1}^*. Suppose t_k is the time index of the last observation in regime k (just prior to the k-th structural break), and $t_0 = 0$, $t_K = T$ and $t_{k-1} < t_k$. Then, the subsamples are denoted $\mathbf{y}^{(k)} = (y_{t_{k-1}+1}, \ldots, y_{t_k})'$ such that $\mathbf{y} = \left(\mathbf{y}^{(1)'}, \ldots, \mathbf{y}^{(K)'}\right)'$. We know from Section 2.3 that every unique parameter value is an independent realization from f_0 (conditional on $\boldsymbol{\lambda}$), enabling us to rewrite the likelihood function and resample θ_k^* from

$$p\left(\theta_k^*|\mathbf{y}^{(k)}\right) \quad \propto \quad f_0(\theta_k^*) \prod_{t=t_{k-1}+1}^{t_k} p(y_t|\mathbf{y}^{1,t-1}, \theta_k^*), \qquad (k = 1, \ldots, K). \tag{11}$$

This is just the 'multi-observation' version of the previously discussed posterior mixture component $p(\theta_t|\mathbf{y}^{1,t})$, and hence it has a known form. These resampled $(\theta_1^*, \ldots, \theta_K^*)'$ are used in the next iteration of the Gibbs sampler. To demonstrate the efficacy of our Gibbs sampler we return to the example from the previous section.

Example (*continued*) (Estimation issues): *The model in* (7)–(8) *shows that the Gaussian likelihood combined with a normal-inverted Gamma–2 baseline prior forms a conjugate setting. Integrating out both* μ *and* σ^2 *provides the one-observation*

marginal likelihoods for $t = 1, \ldots, T$ (Student's t densities),

$$p_0(y_t) = \pi^{-\frac{1}{2}} \frac{\Gamma\left(\frac{1+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} (B+1)^{-\frac{1}{2}} \left(\frac{(y_t - b)^2}{B+1} + S \right)^{-\frac{1+\nu}{2}} S^{\frac{\nu}{2}}.$$

Of which computation can easily be vectorized. The one-observation posteriors for the mixture components have the familiar form:

$$\sigma_t^2 | y_t \sim \mathcal{IG}2(w, W), \qquad \mu_t | y_t, \sigma_t^2 \sim \mathcal{N}(a, A).$$

The parameters of these are

$$w = 1 + \nu, \qquad W = S + \frac{(y_t - b)^2}{B+1},$$
$$a = \frac{y_t B + b}{B+1}, \qquad A = \sigma_t^2 \frac{B}{B+1}.$$

We simulate a time series of $T = 200$ observations from a process with three regimes in both mean and variance, where the structural breaks occur at $t = 40$ and 100. Figure 2(a) shows the simulated time series. After employing the Gibbs sampler with the remix step for 2,000 runs, we obtain the posterior time paths for μ and σ^2 as depicted in Figures 2(c)–(d), where the first 1,000 runs are discarded as burn-in and only the last 1,000 runs are used for constructing the posterior distributions. The chain converges quickly and this only requires 1-2 minutes computing time on a modern personal computer. We see that the data generating process is accurately retrieved and the imposed simultaneous breaking of the two parameters is not restrictive. Figure 2(b) shows the marginal baseline prior for σ^2. Since we have a regime with variance equal to 4 and the baseline prior has only modest support for values larger than 2, it is interesting to note that the variance of this volatile regime is still properly estimated. However, this remark is certainly something to be aware of while choosing the baseline distribution. We further address this issue in the illustrations in Section 4. ∎

The above derived Gibbs sampler can be applied to conjugate and conditional conjugate settings. That is, in case of independent breaks in a vector of time-varying parameters, where we have multiple layers as in (1), we can condition on other time-varying parameters and still employ this procedure.

3.1.2 Non-conjugate setting

In the case of non-conjugate baseline priors and likelihoods we propose to implement a Metropolis–Hastings [MH] sampler. Instead of direct sampling from the consecutive full conditional posteriors, we now use the full conditional priors as candidate distributions to obtain the following algorithm:

Step 1. Initialize the vector of time-varying parameters[10] at $\boldsymbol{\theta}^{(1)}$; set $m = 1$ and repeat Step 2 for $m = 2, \ldots, M$ (= number of simulation runs);

Step 2. For $t = 1, \ldots, T$ sample from the full conditional prior as in (4) and use this proposal value $\theta_t^{\#}$ as a sample from the candidate distribution. The result in (9) determines the MH-steps:

- Compute the proposal acceptance probability (which is the ratio of one-observation likelihoods)

$$\alpha(\theta_t^{(m)}, \theta_t^{\#}) = \min \left\{ \frac{p\left(y_t | \mathbf{y}^{1,t-1}, \theta_t^{\#}\right)}{p\left(y_t | \mathbf{y}^{1,t-1}, \theta_t^{(m)}\right)}, 1 \right\};$$

- Set $\theta_t^{(m+1)} = \theta_t^{\#}$ with probability $\alpha(\theta_t^{(m)}, \theta_t^{\#})$ and $\theta_t^{(m+1)} = \theta_t^{(m)}$ otherwise.

Because of the assumption that structural breaks occur only infrequently, the full conditional prior is the dominant part in (9). Exactly this makes the chosen candidate distribution a well-performing option. Moreover, for the large majority of the observations there will be no break and in iteration m it will hold that $\theta_{t-1}^{(m)} = \theta_t^{(m)} = \theta_{t+1}^{(m)} \equiv \theta^*$. In this case the proposal value $\theta_t^{\#}$ will very likely be θ^* and no likelihood evaluations at all are needed as the acceptance probability obviously equals one. Hence, this MH-sampler requires even less computations compared to the previous Gibbs sampler and is still $\mathcal{O}(T)$. However, because π is small, convergence may take longer. Starting with a no-breaks situation it may take a while before the non-degenerate component f_0 is sampled from.

For the remixing in (11) we can sample from a close candidate and perform an MH-step or, for low-dimensional cases, implement a griddy-Gibbs step (see Ritter and Tanner; 1992).

[10] The easiest way to do so is simply starting in a case of no breaks at all, that is, set $\theta_t^{(1)} = \theta_0$, $(t = 1, \ldots, T)$, with θ_0 somewhere in the support of f_0.

3.2 Baseline parameters

In case of a prior on the hyperparameters of the baseline distribution, we can update by extending the discussed simulation scheme as in any hierarchical model. Conditional on $\boldsymbol{\theta}$ we can construct the vector of the K unique parameter values $\boldsymbol{\theta}^*$ that are independent draws from $f_0(\theta_k^*; \boldsymbol{\lambda})$. Therefore, updating $\boldsymbol{\lambda}$ means sampling from

$$p(\boldsymbol{\lambda}|\mathbf{y}, \boldsymbol{\theta}^*) \quad \propto \quad p(\boldsymbol{\lambda}) \prod_{k=1}^{K} f_0(\theta_k^*; \boldsymbol{\lambda}). \tag{12}$$

Clearly, a conjugate prior distribution for $\boldsymbol{\lambda}$ usually facilitates this simulation step. We refer to Section 4 for examples.

3.3 Breaking probability

In case we treat the probability of a break π as an unknown parameter, we can include it in the MCMC simulation scheme and update by simulating from its full conditional posterior, which can be written as

$$p(\pi|\mathbf{y}, \boldsymbol{\theta}) \quad \propto \quad p(\pi) \prod_{t=2}^{T} p(\theta_t|\theta_{t-1}),$$

because the conditional densities of the parameters are the only parts that involve π. Conditional on a sampled value of the parameter vector, the transition densities reduce to

$$p(\theta_t|\theta_{t-1}) = \begin{cases} \pi f_0(\theta_t), & \text{if} \quad \theta_t \neq \theta_{t-1}, \\ \pi f_0(\theta_t) + (1-\pi), & \text{if} \quad \theta_t = \theta_{t-1}. \end{cases}$$

This shows that a Beta prior does not automatically lead to a full conditional distribution which is also Beta, as the term $\pi f_0(\theta_t)$ in case $\theta_t = \theta_{t-1}$ does not cancel out.[11] However, we can augment the parameter vector with a vector of indicator variables $\mathbf{s} = (s_2, \ldots, s_T)'$, such that conditional on these indicators π *can* be sampled from a Beta distribution, see Geweke and Jiang (2010) and their specification in Section 2.1.

Sampling $\mathbf{s}$ conditional on $\boldsymbol{\theta}$ is simple and fast. Importantly, given $\boldsymbol{\theta}$ the s_t's are non-degenerate. We refer to Appendix B for details of this step and for a proof that it leads to the proper invariant distribution. Note that we actually twist the procedure

[11] Because $\pi f_0(\theta_t)$ is small it is very close to a Beta distribution. Applying an MH-step with as candidate this close Beta distribution turns out to be a good simulator, see Appendix B for details.

proposed by Giordani and Kohn (2008). Instead of simulating indicators and states in one block by integrating out the states first, we sample in one block by first analytically integrating out the indicator variables. This results in a computationally more attractive way to do inference.

If we now take a Beta prior for π the full conditional posterior (conditional on $\mathbf{s}$) of π is also Beta:

$$\pi \sim \mathcal{B}e(r_1, r_2) \qquad \Longrightarrow \qquad \pi|\mathbf{y}, \mathbf{s} \sim \mathcal{B}e(K^* + r_1, T - 1 - K^* + r_2),$$

with $K^* = \sum_{t=2}^{T} \mathbb{I}_{\{s_t=1\}}$ which is larger than or equal to the number of in-sample breaks $K - 1$.

4 Illustrations

In this section we demonstrate the practical usefulness of our approach by presenting four illustrative applications. As we want to highlight the general applicability of our approach to different types of time series models, the illustrations involve a Poisson count data model, a copula model, a probit model and an autoregressive model. These four examples will touch on issues relevant with respect to the modeling process and estimation, including prior specification and the computational ease of our approach in nonlinear models.

4.1 A Poisson count data model for earthquake data

In this example we investigate possible structural instability in a count data model, that is, a model for a time series that takes only discrete values within a limited range. Specifically, we consider a Poisson model for describing the worldwide annual counts of extreme earthquakes (larger than 7.0) for the period 1900-2009.[12] The time series is displayed in Figure 3(a), showing that it ranges between a minimum of 6 in 1986 and a maximum of 41 in 1943. It also appears that the series may be subject

[12]Taken from the *Time Series Data Library* by Rob Hyndman: Hyndman, R.J. (2010), `http://robjhyndman.com/TSDL` accessed on July 23, 2010. Originally collected by the National Earthquake Information Center, which source (`http://earthquake.usgs.gov/earthquakes/eqarchives/year/eqstats.php`) we also have used to extend the sample with data up to 2009.

to occasional level shifts. We examine this possibility by allowing for time-variation in the mean parameter of the Poisson model. The complete model is given by

$$\begin{aligned} y_t|\psi_t &\overset{i.i.d.}{\sim} \mathcal{P}oi(\psi_t), \\ p(\psi_t|\psi_{t-1}) &= \pi f_0(\psi_t) + (1-\pi)\mathbb{I}_{\{\psi_t=\psi_{t-1}\}}, \\ f_0(\psi) &= f_{\mathcal{G}a}(\psi; a, b). \end{aligned}$$

where we opt for a Gamma baseline prior for the parameter ψ to obtain a conjugate setting.

This conjuage setting implies that we can implement a straightforward Gibbs sampler for this model. The necessary marginal likelihoods become

$$p_0(y_t) = \frac{b^a}{(b+1)^{a+y_t}} \frac{\Gamma(a+y_t)}{\Gamma(a)y_t!},$$

where the ratio on the right reduces to $\binom{a+y_t}{a}$ if a is integer. The one-observation posterior for the continuous mixture component is

$$\psi_t|y_t \sim \mathcal{G}a(a+y_t, b+1).$$

Naturally, the posterior for the remix step, conditional on the breaking dates is

$$\psi_k^*|\mathbf{y}^{(k)} \sim \mathcal{G}a\left(a + \sum\nolimits_{t=t_{k-1}+1}^{t_k} y_t, \quad b + (t_k - t_{k-1})\right),$$

where k denotes the regime.

Posterior results for this model are depicted in Figure 3. The results are based on 5,000 iterations of the Gibbs sampler of which the first 2,000 iterations serve as burn-in. This takes about 2-3 minutes computing time. In Figures 3(b)–(c), we display the posterior marginal distributions of the ψ_t's for two different parameterizations of the Gamma baseline prior. The density functions of these two baseline priors are given in Figure 3(d). If we choose for a relatively uninformative prior with wide support, three distinct kinds of geophysical activity seem to occur. If we choose for the more restrictive prior with support concentrated under 20, the 'high-activity' type (values around 25-30) is not present anymore, see Figure 3(c). This demonstrates that it is important that the prior on ψ has enough support to capture all possible regimes in the data. Figures 3(e)–(f) show the marginal posterior break probabilities for each point in time. These are $\mathsf{Pr}[\psi_t \neq \psi_{t-1}|\mathbf{y}] = \mathsf{E}\left[\mathbb{I}_{\{\psi_t \neq \psi_{t-1}\}}|\mathbf{y}\right]$, $(t = 2, \ldots, T)$, which

can easily be computed using the Gibbs output by simply counting the number of breaks given a sample of $\boldsymbol{\psi}^{1,T}$ from the posterior. The probabilities in Figures 3(e) indicate that the structural breaks may either occur almost instantaneously (as in 1905 and 1951) or gradually (during the 1910s, the late 1930s and early 1940s, and around 1980).

4.2 Breaks in copula model parameters

To illustrate the usefulness of our approach in non-conjugate settings we examine a copula model, which is becoming increasingly popular in empirical finance to capture non-standard cross-sectional dependence (see, for example, McNeil et al.; 2005; Jondeau and Rockinger; 2006). We simulate 400 observations $\mathbf{u}_t = (u_{1t}, u_{2t})'$, $(t = 1, \ldots, 400)$, from a bivariate Clayton copula, given by $C(\mathbf{u}_t; \theta_t) = (u_{1t}^{-\theta_t} + u_{2t}^{-\theta_t} - 1)^{-1/\theta_t}$, with $\theta_t > 0$. The parameter θ_t determines the strength of dependence between u_{1t} and u_{2t}, with higher values indicating stronger dependence. For example, Kendall's τ is equal to $\theta_t/(\theta_t + 2)$. Furthermore, the Clayton copula is characterized by lower tail dependence and upper tail independence, in the sense that

$$\lim_{q \downarrow 0} \mathsf{Pr}\left[u_{2t} \leq q | u_{1t} \leq q\right] = C((q,q)'; \theta_t)/q = 2^{-1/\theta_t},$$
$$\lim_{q \uparrow 1} \mathsf{Pr}\left[u_{2t} \geq q | u_{1t} \geq q\right] = \left[1 - 2q + C((q,q)'; \theta_t)\right]/(1-q) = 0.$$

We impose three regimes with structural breaks occurring at observations 101 and 301. The copula parameter values for these three regimes are 0.1, 2 and 5, respectively. Figure 4 displays some characteristics of the simulated data. Figure 4(a) shows a scatter of the bivariate data over the whole sample period in the unit square. No structural change is visible at first sight. Figure 4(c) displays the same data but here we distinguish between the three regimes by using different marker types. For example, the grey bullet data correspond to the most recent regime in which $\theta = 5$; the large parameter value implies stronger (left-tail) dependence.

We use the following model to estimate the parameters of the Clayton copula for the simulated series:

$$\begin{aligned} \mathbf{u}_t | \theta_t &\overset{i.i.d.}{\sim} C^{\mathrm{Cl}}(\theta_t), \\ p(\theta_t | \theta_{t-1}) &= \pi f_0(\theta_t) + (1-\pi)\mathbb{I}_{\{\theta_t = \theta_{t-1}\}}, \\ f_0(\theta) &= f_{log\mathcal{N}}(\theta; a, A), \qquad\qquad \theta \in (0, \infty). \end{aligned}$$

The log-normal baseline prior exhibits desirable properties as it can be used for any parameter which is bounded from below/above. It is easy to see that the baseline prior f_0 and the likelihood $c^{\text{Cl}}(\mathbf{u}_t|\theta_t) = \partial^2 C(\mathbf{u}_t;\theta_t)/(\partial u_{1t}\partial u_{2t})$ are non-conjugate, so we have to use the Metropolis-Hastings sampler to simulate from the full conditional posteriors $p\left(\theta_t|\mathbf{U},\boldsymbol{\theta}_{[-t]}\right)$ with $\mathbf{U} = (\mathbf{u}_1,\ldots,\mathbf{u}_{400})'$. For the remixing step we use a griddy-Gibbs step where we take into account the f_0-prior for the regime parameter: $\theta_k^* \sim log\mathcal{N}(a,A)$, $(k = 1,\ldots,K)$.

If $\theta \downarrow 0$ the Clayton copula becomes an independent copula. The first regime is close to this situation. Therefore we consider a quite uninformative baseline prior ($a = 0.5$ and $A = 1$) which also covers values close to zero. The density of the prior is depicted in Figure 4(b). As we expect only a few breaks we set π equal to 0.01. Posterior results are shown in Figure 4(d). It turns out that the marginal posteriors of the parameters θ_t resemble the data generating process closely. We see very sharp and sudden shifts in the parameter value at times $t = 101$ and $t = 301$. Posterior results turn out to be quite robust with respect to prior parameters settings and specification of the baseline prior. The computational burden of our approach is small as it takes only three minutes computing time to obtain 3,000 draws from the posterior distribution.

4.3 Size spread sign prediction

In the third illustration we apply our method to a probit model to forecast the sign of the size spread in monthly U.S. stock returns. The size spread is defined as the difference between the returns on portfolios consisting of the 20% smallest stocks and portfolios consisting of the 20% largest stocks over the period July 1962 - October 2010.[13] Hence, the data correspond to binary random variables y_t, which equal 1 if the difference is positive and zero otherwise. We model these binary variables using a probit specification:

$$y_t|\mathbf{x}_t,\boldsymbol{\beta}_t \overset{i.i.d.}{\sim} \mathcal{B}er\left(\Phi(\mathbf{x}_t'\boldsymbol{\beta}_t)\right), \tag{13}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. For the explanatory variables $\mathbf{x}_t$ we use a number of series that are typically considered for predicting

[13]The data were obtained from Kenneth French's website data library `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/`.

(relative) stock returns. A preliminary analysis suggests to use the following five variables: credit spread, term spread, market return, and the growth in the Conference Board's leading index.[14] We also include an intercept and the one-month lagged size spread. As empirical studies indicate that the relation between some of the explanatory variables and stock returns change (see, for example, Pesaran and Timmermann; 2002), we allow for breaks in the $\boldsymbol{\beta}$ parameters.

Since we are dealing with a simple 0/1-series, we must be careful not to demand too much from the data. For example, allowing all six parameters to vary leads to too much flexibility corresponding to perfect fit in some time periods.[15] Therefore we focus on possible changes in the effect of the two spread variables and we only allow their coefficients, β_{CS} and β_{TS}, to change simultaneously over time. Thus, we extend the model in (13) with the conditional distribution

$$p(\boldsymbol{\beta}_{\mathrm{S},t}|\boldsymbol{\beta}_{\mathrm{S},t-1}) = \pi f_0(\boldsymbol{\beta}_{\mathrm{S},t}) + (1-\pi)\mathbb{I}_{\{\boldsymbol{\beta}_{\mathrm{S},t}=\boldsymbol{\beta}_{\mathrm{S},t-1}\}},$$

and we propose to use the following Gaussian baseline prior:

$$\boldsymbol{\beta}_{\mathrm{S}} = (\beta_{\mathrm{CS}}, \beta_{\mathrm{TS}})'|\boldsymbol{\mu}, \boldsymbol{\Sigma} \;\sim\; \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{14}$$

For the time-invariant part of $\boldsymbol{\beta}_t$ we apply an uninformative conjugate Gaussian prior. As discussed before, especially for forecasting purposes it would make sense to update the hyperparameters of the baseline prior. We consider a matricvariate normal-inverted Wishart prior for the hyperparamters:

$$\boldsymbol{\mu}'|\boldsymbol{\Sigma} \;\sim\; \mathcal{MN}(\mathbf{p}', q \cdot \boldsymbol{\Sigma}), \tag{15}$$

$$\boldsymbol{\Sigma} \;\sim\; \mathcal{IW}(\mathbf{S}, u). \tag{16}$$

For the breaking probability π we take a Beta prior with parameters r_1 and r_2.

[14] To be more precise: (1) credit spread: the difference between Moody's Baa corporate bond rate and the 10-year Treasury constant maturity rate, in deviance from its one-year moving average; (2) term spread: the difference between the 3-month Treasury bill secondary market rate and the effective Federal funds rate, in deviance from its one-year moving average; (3) stock market return: level of the S&P500 index relative to a two-year moving average; (4) growth in leading index: growth rate of The Conference Board's Composite Leading index over the six most recent months. All explanatory variables are available at the actual time the forecast is constructed, that is, some of them are appropriately lagged to take into account publication delays.

[15] This issue becomes even more relevant when the data show persistent clustering of zeros or ones, for example, in case of an indicator for the business cycle regime.

The prior hyperparameters are set to $\mathbf{p}' = (0, 0)$, $q = 2$, $\mathbf{S} = 2 \cdot \mathbf{I}_2$ and $u = 6$. Following the suggestions in Giordani et al. (2007) the parameters of the prior for π are set to $r_1 = 5$, $r_2 = 3000$, which corresponds to a prior assumption of one to two breaks. Because of the non-conjugate setting of (13) and (14) we have to rely on the MH-procedures described in Section 3.1 to sample from the posterior distribution.[16] To speed up convergence of the chain we employ a tailored remix step. For the remix step we sample latent variables from truncated normals just as in an MCMC sampler for a probit model based on data augmentation (see, for example, Albert and Chib; 1993). Conditional on these latent variables we resample the $\boldsymbol{\beta}_{\mathrm{S}}$ parameters. Note that we only use these variables for the remixing step.

Figure 5 shows the posterior results of this sampler based on 7,000 iterations of the MCMC sampler of which the first 2,000 serve as burn-in. Figures 5(a) and (c) show the posteriors of the two parameters that may be time-variant. Initially the credit spread has no impact, but since the late 1970s its effect becomes positive. After the end of the 1990s the effect becomes negative. The term spread has a positive impact from the beginning of the sample which becomes even stronger in the early 1980s, though, its posterior uncertainty also increases.

The marginal properties of the posterior distribution of the model parameters are reported in Table 1. The posterior median of the $\boldsymbol{\mu}$ parameters is larger than the median of the prior although the increase is small due to the fact that we only have a small amount of breaks in the sample. Figure 5(b) shows the posterior of the two marginal baseline densities implied by $f_0(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ integrated over the posterior $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{y})$. The posterior baseline belonging to the term spread is slightly more shifted to the right due to the positive effect of the term spread on the size spread sign.

In Figure 5(d) we show the 'fitted' in-sample probabilities; $\Phi(\mathbf{x}_t'\boldsymbol{\beta}_t)$ integrated over the posterior distribution. These probabilities do not show an outspoken pattern which is inherent to these models. The hitrate is 63% based on a cut-off of 0.5.

[16] If we use data augmentation for the probit part we can rely on Gibbs steps but this extends the MCMC sampler with more simulation steps, see Albert and Chib (1993)

4.4 Forecasting U.S. quarterly GDP growth

Our final illustration examines structural breaks in an AR(4) model for quarterly growth in U.S. gross domestic product and the implications with regard to forecasting. Similar exercises have been employed by McConnell and Perez-Quiros (2000); Clark (2009) and Geweke and Jiang (2010), for example. If y_t is the annualized quarterly growth rate for the sample period 1960Q1-2010Q3 and the random shocks are assumed to be Gaussian, then the standard AR(4) representation can we written as

$$y_t = \alpha_t + \xi_t y_{t-1} + \sum_{j=1}^{3} \varphi_j^* \Delta y_{t-j} + \varepsilon_t, \qquad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_t^2\right),$$

where we condition on the observations before 1960Q1.

First, we allow for infrequent intercept shifts and changing persistence through the conditional mean parameters α_t and ξ_t, respectively. We impose simultaneous changes in these two parameters to control for the fact that the unconditional expectation of y_t is determined by both in a positive way. A shift in the unconditional mean, either through the intercept or the persistence parameter occurs with probability π_1. To impose unit root stationarity ξ_t should take values smaller than 1. In our framework this truncation is easily dealt with. Further (conjugate) considerations lead to a truncated multivariate Gaussian baseline prior distribution for the time-varying mean parameters $\boldsymbol{\theta}_t = (\alpha_t, \xi_t)'$:

$$\begin{aligned} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) &= \pi_1 f_0(\boldsymbol{\theta}_t) + (1-\pi_1)\mathbb{I}_{\{\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}\}}, \\ \boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \times \mathbb{I}_{\{\xi<1\}}. \end{aligned} \tag{17}$$

Shifts in the volatility of the random shocks ($\sigma_t^2 = \mathsf{Var}\left[\varepsilon_t|\sigma_t^2\right]$) are modeled independently from the previous regression parameters. Therefore we specify a separate layer as in (1) with a break in variance occurring with probability π_2. For this parameter we opt for an inverted Gamma–2 baseline prior:

$$\begin{aligned} p(\sigma_t^2|\sigma_{t-1}^2) &= \pi_2 f_0(\sigma_t^2) + (1-\pi_2)\mathbb{I}_{\{\sigma_t^2 = \sigma_{t-1}^2\}}, \\ \sigma^2|\Omega, \nu &\sim \mathcal{IG}2(\Omega, \nu). \end{aligned}$$

In order to update the baseline prior parameters, we augment the model with a third level. For the baseline parameters of the truncated Gaussian in (17) we use the same conjugate choice as in the sign prediction model of Section 4.3, see

(15)–(16). Following Clark (2009), we use the pre-sample data to set its parameters $\mathbf{p} = (2, 0.4)'$, $q = 10$, $u = 6$ and $\mathbf{S} = 2 \cdot \mathbf{I}_2$.

We impose an inverted Gamma–2 prior for the parameter Ω of the baseline distribution of the conditional variance. This allows us to learn with respect to the distribution of σ_t^2:

$$\Omega \sim \mathcal{IG}2(W, z).$$

To simulate Ω during the MCMC scheme we implement an independence MH-simulator with a Gamma distribution to generate proposal values. Again we use historical data and take $W = 50$, $z = 6$ and $\nu = 9$.

Further prior settings involve a conjugate Gaussian prior on the time-constant parameters

$$(\varphi_1^*, \varphi_2^*, \varphi_3^*)' \sim \mathcal{N}(\mathbf{b}, \mathbf{B}),$$

where we set its hyperparameters such that it is close to a flat prior. To complete, since we have two layers that account for structural breaks in the mean parameters and the conditional variance, respectively, we have to set two priors for the associated break probabilities π_1 and π_2. We use two independent Beta priors:

$$\pi_1 \sim \mathcal{B}e(r_{11}, r_{12}) \qquad \text{and} \qquad \pi_2 \sim \mathcal{B}e(r_{21}, r_{22}),$$

and we set $r_{11} = 5, r_{12} = 1000, r_{21} = 1$ and $r_{22} = 100$. This way the expected probability of a break in either the mean or the conditional variance is approximately equal to 0.02.

The assumption of independence between the two layers has the following implications for estimation. Conditional on the standard deviations $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_T)'$, we have a conjugate setting[17] and therefore we can employ the Gibbs sampler to simulate $(\alpha_t, \xi_t)'$, $(t = 1, \ldots, T)$. Vice versa we also have a conjugate setting and can simulate the conditional variances.

Figures 6(a) and (d) show the data and the posterior unconditional mean of y_t, and the posterior path of σ_t^2, respectively. The unconditional mean is given by $\alpha_t/(1-\xi_t)$. Figure 6(a) shows that no shifts have occurred during the sample period; α is in the range $[1, 1.9]$ and the persistence parameter ξ covers values in $[0.4, 0.65]$.

[17]Even with the truncation of the Gaussian baseline prior for ξ, the marginal likelihoods and posterior distributions can be obtained analytically. Of course, the MH-routines can be applied equally well.

The variance does show significant changes, in line with previous empirical findings. The Great Moderation corresponds to the large decline in volatility in the early 1980s. Recent negative growth rates suspect this decline is being offset, see Clark (2009). However, more data are needed to provide more strong evidence in favor of this hypothesis.

In Figures 6(b) and (c) we display the evolution of the marginal posterior predictive distributions $p(y_{\tau+h}|\mathbf{y}^{1,\tau})$ for $h = 1, \ldots, 40$, for two cases: no breaks at all and the previously described structural break model, respectively. Forecasting starts at $\tau =$ 2002Q4. Figures 6(e) and (f) show the marginal posterior predictive densities for horizions one quarter ahead (solid) and ten years ahead (dashed). Clearly, if we assume parameter stability the Great Moderation is not accounted for and the current variance is heavily overestimated leading to too wide density forecasts. The structural break model starts with tighter forecast densities due to the smaller estimated σ_τ^2. If the forecasting horizon grows, we see that incorporating future structural breaks leads to a predictive distribution that is more peaked than the Gaussian in Figure 6(e). This heavy-tailedness assigns more probability mass to more extreme values as realized in 2010.

5 Conclusion and discussion

In this paper we have proposed a dynamic stochastic specification to model infrequent sudden changes in model parameters over time. The specification is simple and has many nice desirable properties.

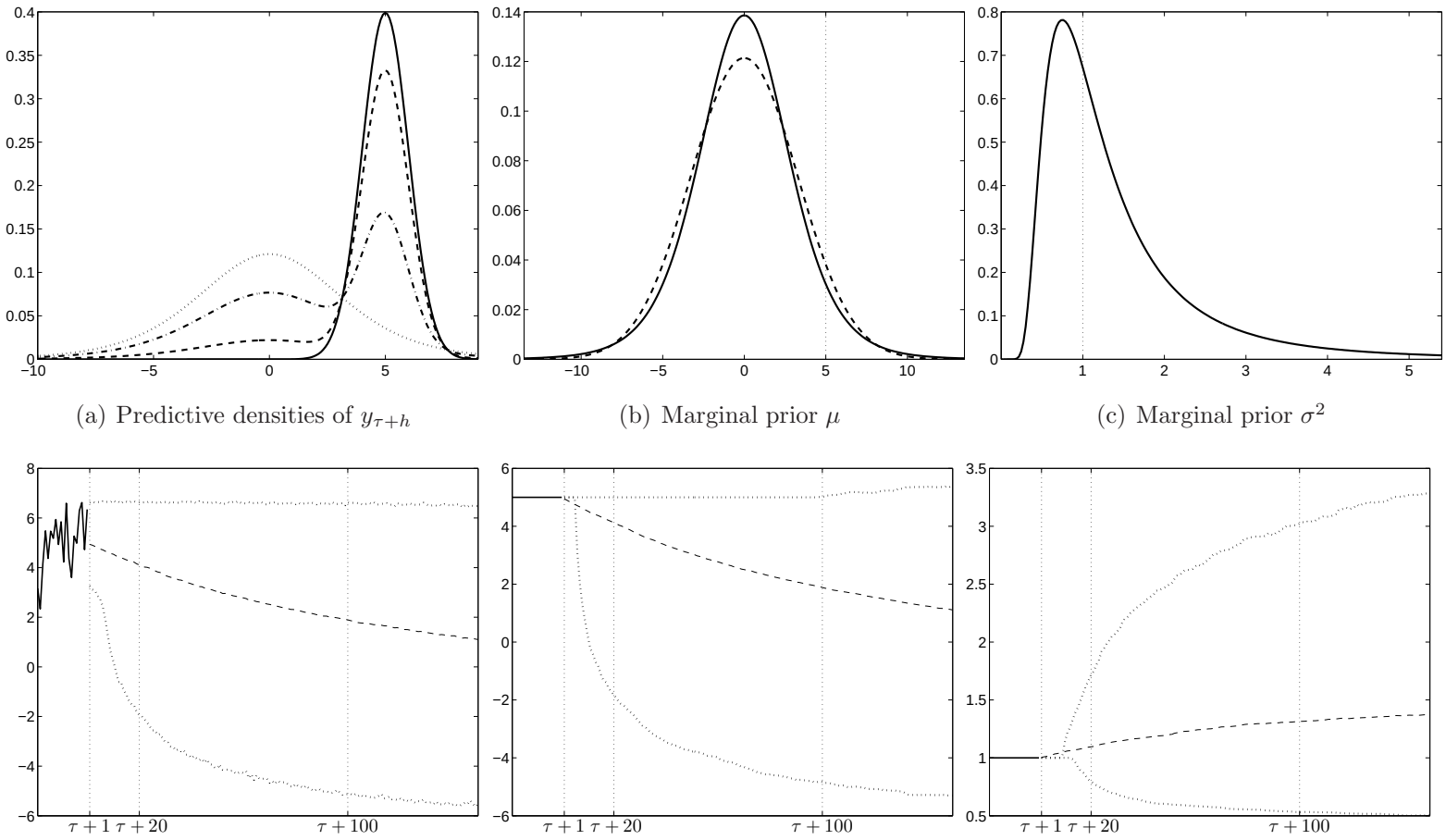
First of all, the number of in-sample and out-of-sample breaks and the break dates are *a priori* unknown. The dynamic specification contains natural implications in terms of out-of-sample forecasting. In existing models, future parameter breaks are neglected or require (computationally demanding) extensions. Our approach implies a random number of out-of-sample breaks where its distribution depends on the forecasting horizon and the breaking probability. The risk of future breaks is assimilated in the posterior predictive distributions according to the rules of probability.

Second, our approach is flexible in the sense that the posterior simulator does not impose any restrictions on the model under consideration. Hence, we do not have to limit ourselves to linear regression models or models which can be written in

a (mixed) Gaussian state-space representation. The modeling part only involves the choice of a likelihood specification and a baseline prior distribution that generates new parameter values if a break occurs.

Third, the proposed posterior simulator is computationally less complex and intensive than existing methods which usually need Kalman recursions or filtering techniques. Our simulator is a single-move sampler and only requires sampling from three-component mixtures followed by a remix step to enhance the convergence of the sampler. This remix step is case-specific and needs to be tailored to the model/prior specification under consideration if it is non-conjugate.

We have illustrated our approach using four examples. Both in real data situations and simulated data sets the methods perform well and the computational burden is relatively small. The parameterization of the baseline prior turns out to be important. To prevent that breaks are not detected, we have to ensure that the baseline priors do not exclude plausible parameter values. Furthermore, the baseline prior plays a key role in multi-step ahead forecasting as it determines the size of out-of-sample parameters. A sensible strategy to obtain a plausible baseline prior is to put a prior on its hyperparameters.

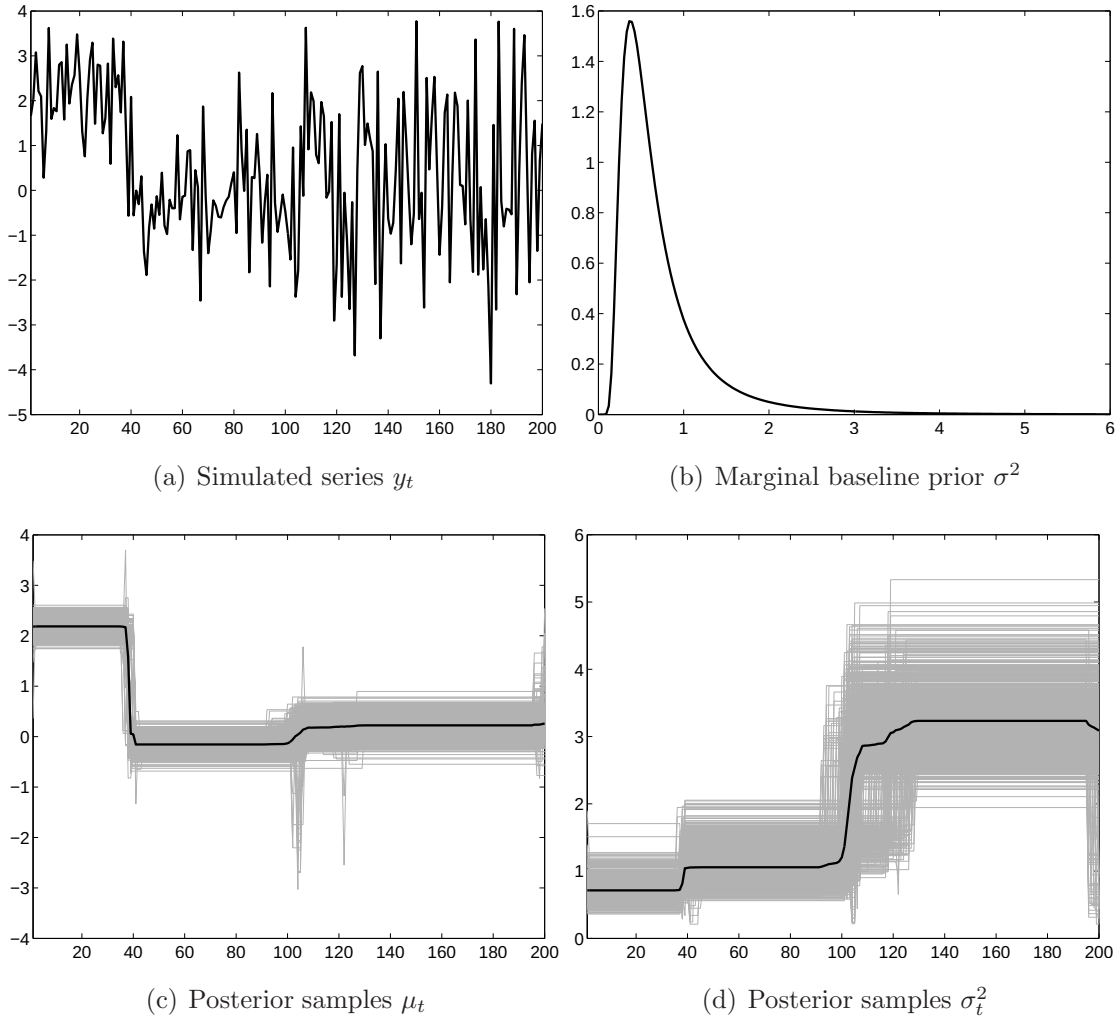


(a) Predictive densities of $y_{\tau+h}$ (b) Marginal prior μ (c) Marginal prior σ^2

(d) Evolution predictive distributions (e) Evolution μ_t (f) Evolution σ_t^2

Figure 1: *Forecasting implications for the example model with potential simultaneous breaks in mean and variance.*

Notes: (a) Pdf's of posterior predictive distributions: Gaussian likelihood with $\mu = 5$ and $\sigma^2 = 1$ (solid) and $p(y_{\tau+h}|\mathbf{y}^{1,\tau})$ for $h = 1$ (dashed), $h = 20$ (dashed-dotted) and $h = 100$ (dotted); (b) Marginal Student's t prior of μ (solid) and Gaussian with variance fixed at the Student's t's (dashed); (c) Marginal inverted Gamma–2 prior of σ^2; (d) Evolution of posterior predictive distributions for horizons $h = 1, \ldots, 150$, median (dashed) and 5th- and 95th-percentiles (dotted); (e) Evolution over time of μ_t; (f) Evolution over time of σ_t^2. The results in (d)–(f) are obtained through simulation (10,000 runs) with hyperparameters set as follows: $b = 0$, $B = 9$, $\nu = S = 6$ and break probability $\pi = 0.01$. Forecasting starts at time τ where $\mu_\tau = 5$ and $\sigma_\tau^2 = 1$.

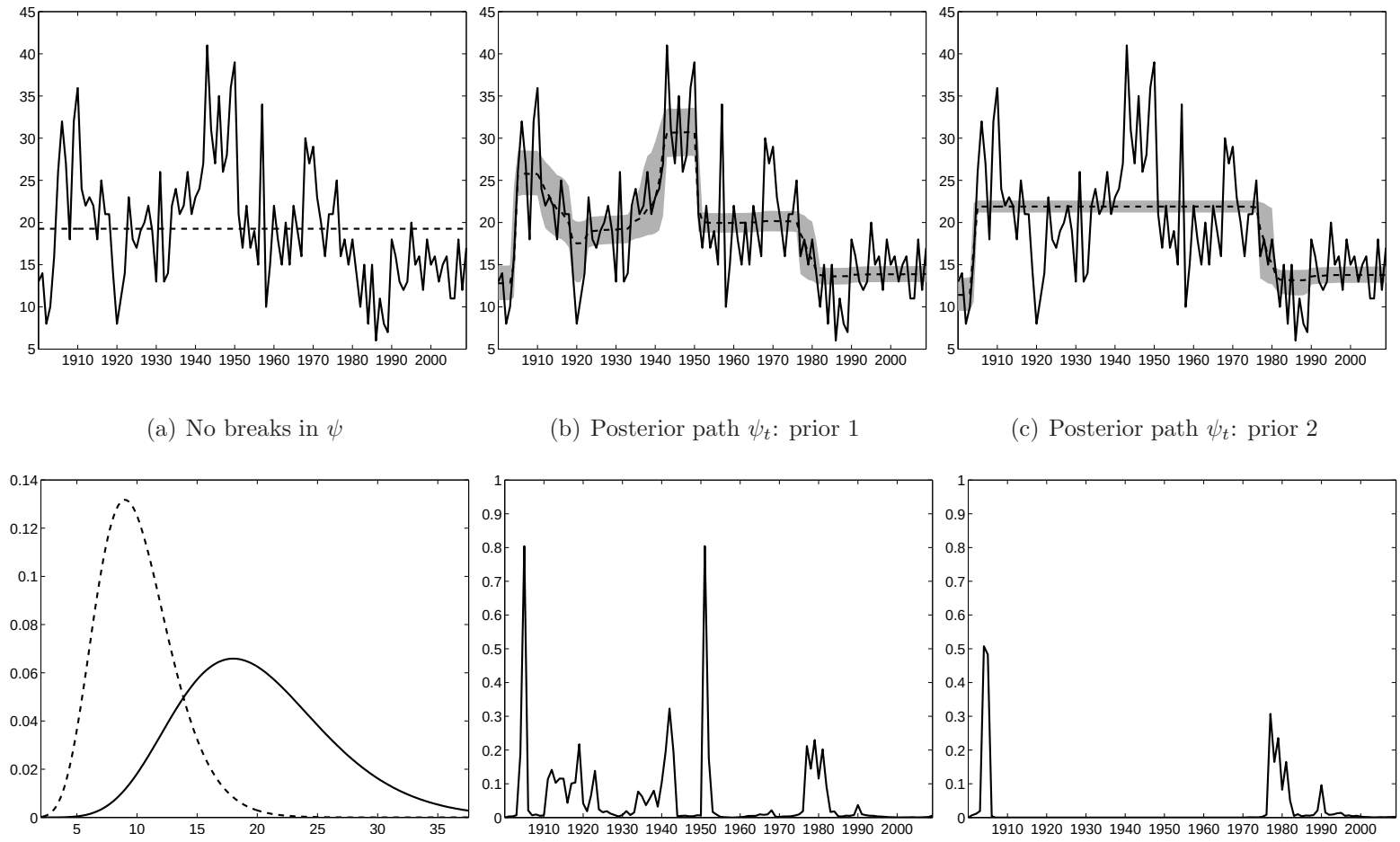


(a) Simulated series y_t (b) Marginal baseline prior σ^2

(c) Posterior samples μ_t (d) Posterior samples σ_t^2

Figure 2: *Estimation results for the example model with potential simultaneous breaks in mean and variance.*

Notes: (a) Simulated series $y_t|\mu_t, \sigma_t^2 \overset{i.i.d.}{\sim} \mathcal{N}(\mu_t, \sigma_t^2)$ with $\mu_t = 2$, $(t \leq 40)$, $\mu_t = 0$, $(t > 40)$ and $\sigma_t^2 = 1$, $(t \leq 100)$, $\sigma_t^2 = 4$, $(t > 100)$; (b) Pdf of marginal inverted Gamma–2 baseline prior for σ_t^2; (c) Posterior mean (solid) and Gibbs samples from posterior $\mu_t|\mathbf{y}$; (d) Posterior mean (solid) and Gibbs samples from posterior $\sigma_t^2|\mathbf{y}$. Results are obtained with prior hyperparameters $\pi = 0.01$, $b = 0$, $B = 16$, $\nu = 6$, $S = 3$ and 2,000 simulation runs of which 1,000 serve as burn-in (1-2 minutes computing time).

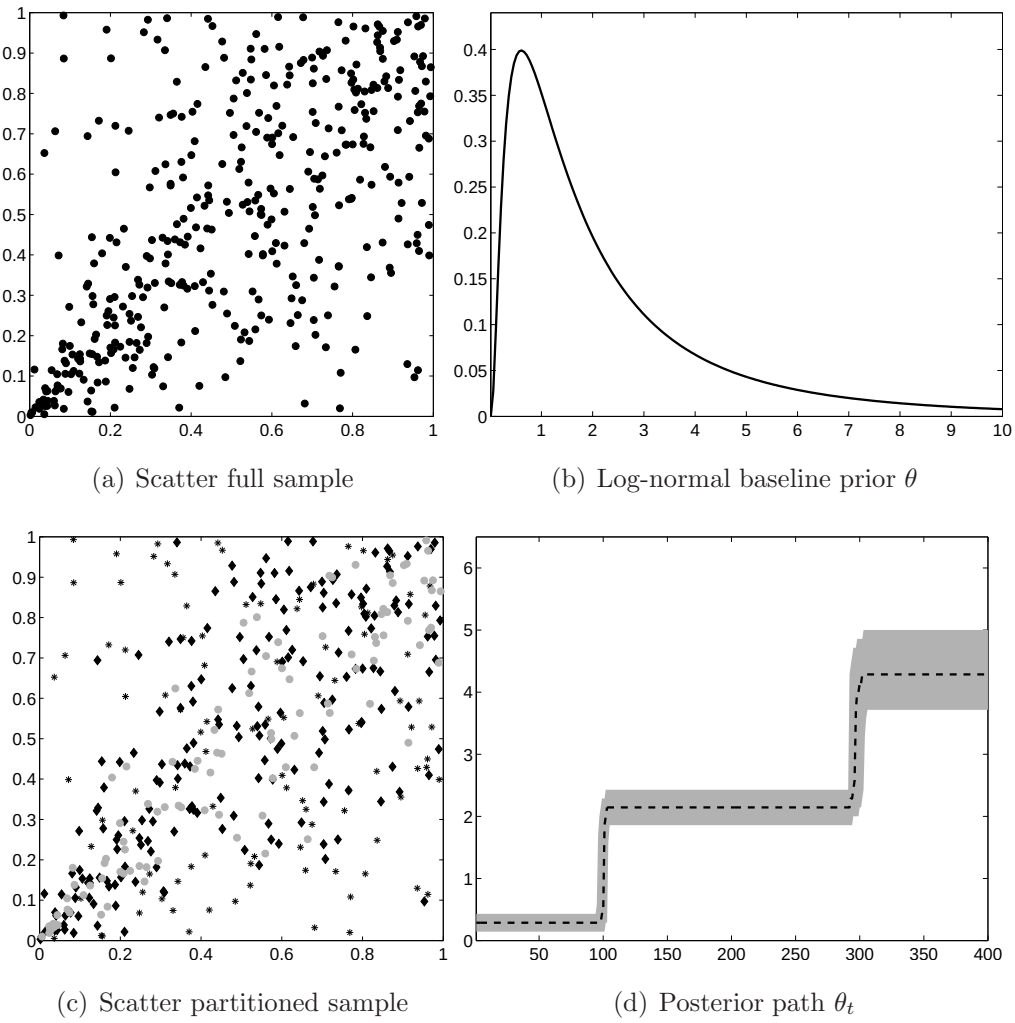


(a) No breaks in ψ (b) Posterior path ψ_t: prior 1 (c) Posterior path ψ_t: prior 2

(d) Gamma baseline priors 1 and 2 (e) Posterior break probability: prior 1 (f) Posterior break probability: prior 2

Figure 3: *Results for the earthquake data with potential breaks in the Poisson parameter.*

Notes: (a) Data series, dashed line represents the posterior mean of the $\mathcal{G}a(a+\sum y_t, b+T)$ when no breaks are allowed; (b) Posterior distribution (mean and 10th- and 90th-percentiles) of ψ_t under 'uninformative' prior 1; (c) Posterior distribution (mean and 10th- and 90th-percentiles) of ψ_t under 'restrictive' prior 2; (d) Solid graph: density of prior 1: $\mathcal{G}a(10, 0.5)$; dashed graph: density of prior 2: $\mathcal{G}a(10, 1)$; (e) Marginal posterior break probabilities under prior 1; (f) Marginal posterior break probabilities under prior 2.



(a) Scatter full sample

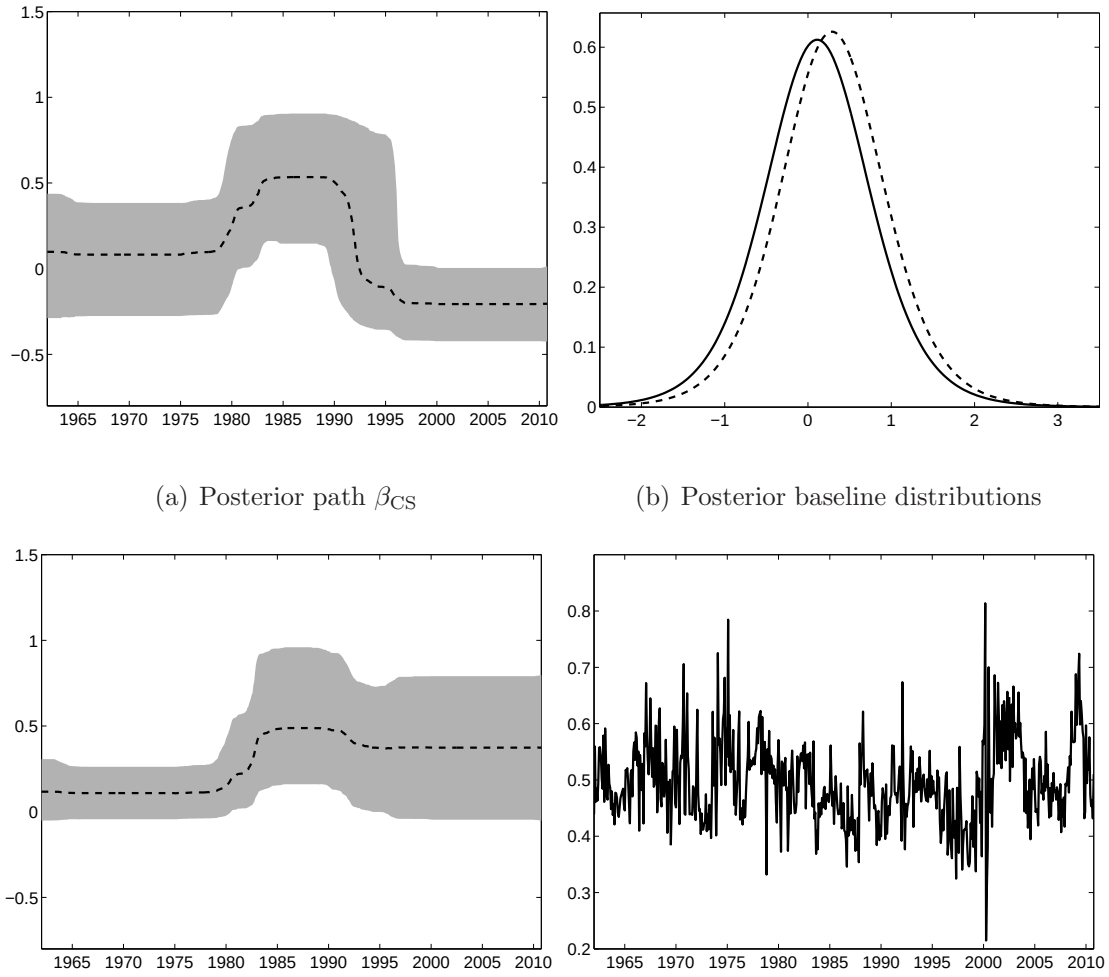
(b) Log-normal baseline prior θ

(c) Scatter partitioned sample

(d) Posterior path θ_t

Figure 4: *Results for the Clayton copula model.*

Notes: (a) Simulated series $\mathbf{u}_t|\theta_t \overset{i.i.d.}{\sim} C^{\mathrm{Cl}}(\theta_t)$, $(t = 1, \ldots, 400)$, with parameter values as follows: $(t \leq 100)$: $\theta_t = 0.1$, $(100 < t \leq 300)$: $\theta_t = 2$ and $\theta_t = 5$, $(t > 300)$; (b) Marginal prior pdf for θ_t for which it holds $\theta_t \sim log\mathcal{N}(0.5, 1)$; (c) Sample partitioned according to break events: asterisk: $t \leq 100$, diamond: $100 < t \leq 300$ and bullet: $t > 300$; (d) Posterior median and 10th- and 90th-percentiles of marginal posteriors $\theta_t|\mathbf{y}$. Results are obtained with prior hyperparameters $\pi = 0.01$, $a = 0.5$, $A = 1$ and 3,000 iterations of the MCMC sampler of which the first 1,000 serve as burn-in (3 minutes computing time).



(a) Posterior path β_{CS}

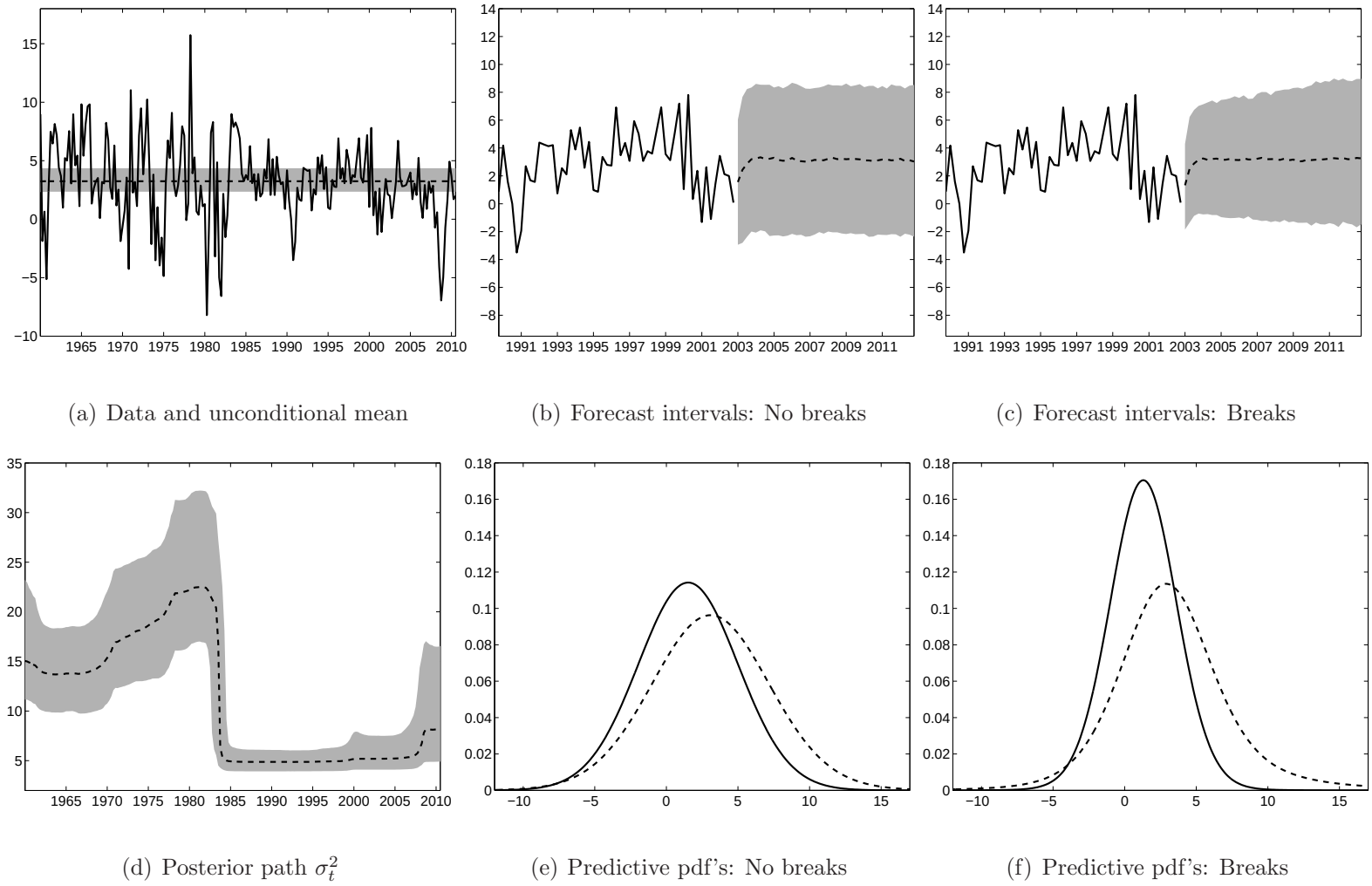
(b) Posterior baseline distributions

(c) Posterior path β_{TS}

(d) Posterior in-sample probabilities

Figure 5: *Results for the sign prediction model.*

Notes: (a) Posterior time-path of the parameter associated with the credit spread (β_{CS}), grey area indicates the 10th- and 90th-percentiles; (b) Posterior densities of the marginal baseline distributions of the two time-varying parameters: β_{CS} (solid) and β_{TS} (dashed); (c) Posterior time-path of the parameter associated with the term spread (β_{TS}); (d) In-sample probit probabilities obtained by integrating $\Phi(\mathbf{x}_t'\boldsymbol{\beta}_t)$ over the full-sample posterior. Results are obtained with prior hyperparameters $r_1 = 5$, $r_2 = 3000$, $\mathbf{p}' = (0, 0)$, $q = 2$, $\mathbf{S} = \text{diag}(2, 2)$ and $u = 6$ and 7,000 iterations of the MCMC sampler of which the first 2,000 serve as burn-in (8 minutes computing time).



(a) Data and unconditional mean

(b) Forecast intervals: No breaks

(c) Forecast intervals: Breaks

(d) Posterior path σ_t^2

(e) Predictive pdf's: No breaks

(f) Predictive pdf's: Breaks

Figure 6: *Results for the AR(4) model for U.S. quarterly GDP growth.*

Notes: (a) Data series and posterior of the unconditional mean of y_t (median and 10th- and 90th-percentiles); (b) Posterior predictive distributions for horizons $h = 1, \ldots, 40$ when forecasting starts in $\tau =$ 2002Q4 and no breaks are allowed for; (c) Same as (b) but now breaks are modeled; (d) Posterior of conditional variance σ_t^2; (e) Densities of predictive distributions for horizons one quarter (solid) and 40 quarters (dashed) in the no-breaks model; (f) Same as (e) but now breaks are modeled.

Table 1: *Posterior properties of the sign prediction model.*

Parameter	Posterior median	Posterior percentiles	
Time-invariant regression parameters			
β_0	0.025	−0.049	0.099
β_{AC}	0.036	0.021	0.051
β_{SP}	−0.826	−1.357	−0.305
β_{LI}	0.348	−0.869	1.534
Third level parameters			
$\boldsymbol{\mu}$	0.112	−0.323	0.550
	0.291	−0.130	0.744
$\boldsymbol{\Sigma}_{(\text{CS,CS})}$	0.327	0.178	0.690
$\boldsymbol{\Sigma}_{(\text{TS,TS})}$	0.306	0.166	0.650
$\boldsymbol{\Sigma}_{(\text{CS,TS})}$	0.012	−0.152	0.184
π	0.002	0.001	0.003

Notes: The table reports the median and the 10th- and 90th-percentile of the marginal posterior distributions of the time-invariant parameters. The first panel depicts properties of the regression parameters that are restricted to be constant over time: intercept (β_0), autocorrelation term (AC), S&P500 (SP) and leading index growth (LI). See the notes of Figure 5 for settings of the prior distributions.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis for binary and polychotomous response data, *Journal of the American Statistical Association* **88**: 901–921.

Andreou, E. and Ghysels, E. (2009). Structural breaks in financial time series, *in* T. Andersen, R. Davis, J.-P. Kreiss and T. Mikosch (eds), *Handbook of Financial Time Series*, Springer-Verlag, Berlin, pp. 839–870.

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

Andrews, D. W. K. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative, *Econometrica* **62**: 1383–1414.

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

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

Carlin, B. P., Gelfand, A. E. and Smith, A. F. (1992). Hierarchical Bayesian analysis of changepoint problems, *Applied Statistics* **41**: 389–405.

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

Chu, C., Stinchcombe, M. and White, H. (1996). Monitoring structural change, *Econometrica* **64**: 10451065.

Clark, T. E. (2009). Is the Great Moderation over?, An empirical analysis, *Economic review*, Federal Reserve Bank of Kansas City.

Clements, M. and Hendry, D. (2006). Forecasting with breaks, *in* G. Elliott, C. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Elsevier Science, Amsterdam, pp. 605–657.

Clements, M. P. and Hendry, D. F. (2001). *Forecasting Non–stationary Economic Time Series*, MIT Press, Cambridge, MA.

Conley, T. G., Hansen, C. B., McCulloch, R. E. and Rossi, P. E. (2008). A semi-parametric Bayesian approach to the instrumental variable problem, *Journal of Econometrics* **144**: 276–305.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association* **90**: 577–588.

Gerlach, R., Carter, C. and Kohn, R. (2000). Efficient Bayesian inference for dynamic mixture models, *Journal of the American Statistical Association* **95**: 819–828.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*, John Wiley & Sons, Hoboken, NJ.

Geweke, J. and Jiang, Y. (2010). Inference and prediction in a multiple structural break model. Working paper, University of Technology, Sydney.

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

Giordani, P., Kohn, R. and van Dijk, D. (2007). A unified approach to nonlinearity, structural change and outliers, *Journal of Econometrics* **137**: 112–133.

Jondeau, E. and Rockinger, M. (2006). The copula-GARCH model of conditional dependencies: An international stock market application, *Journal of International Money and Finance* **25**: 827–853.

Kim, C.-J., Nelson, C. R. and Piger, J. (2008). The less-volatile U.S. economy: A Bayesian investigation of timing, breadth, and potential explanations, *Journal of Business and Economic Statistics* **22**: 80–93.

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

Lancaster, T. (2004). *Introduction to Modern Bayesian Econometrics*, Blackwell Publishing, Malden, MA.

Lettau, M. and Van Nieuwerburgh, S. (2008). Reconciling the return predictability evidence, *Review of Financial Studies* **21**: 1607–1652.

Maheu, J. M. and Gordon, S. (2008). Learning, forecasting and structural breaks, *Journal of Applied Econometrics* **23**: 553–583.

McConnell, M. and Perez-Quiros, G. (2000). Output fluctuations in the United States: What has changed since the early 1980s?, *American Economic Review* **90**: 1464–1476.

McNeil, A. J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management, Concepts, Techniques and Tools*, Princeton University Press, Princeton, NJ.

Paye, B. S. and Timmermann, A. (2006). Instability of return prediction models, *Journal of Empirical Finance* **13**: 274–315.

Perron, P. (2006). Dealing with structural breaks, *in* K. Patterson and T. Mills (eds), *Palgrave Handbook of Econometrics, Vol. 1: Econometric Theory*, Palgrave Macmillan, Hampshire, pp. 278–352.

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

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

Pettenuzzo, D. and Timmermann, A. (forthcoming). Predictability of stock returns and asset allocation under structural breaks, *Journal of Econometrics* .

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

Rapach, D. E. and Wohar, M. E. (2006). Structural breaks and predictive regression models of aggregate U.S. stock returns, *Journal of Financial Econometrics* **4**: 238–274.

Ravazzolo, F., van Dijk, D., Paap, R. and Franses, P. H. (2008). Bayesian model averaging in the presence of structural breaks, *in* M. Wohar and D. E. Rapach (eds), *Frontiers of Economics and Globalization Series, Vol.3*, Emerald Group Publishing, Bingley, pp. 561–594.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler, *Journal of the American Statistical Association* **87**: 861–868.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd edn, Springer-Verlag, New York, NY.

Rossi, P. E., Allenby, G. M. and McCulloch, R. (2005). *Bayesian Statistics and Marketing*, John Wiley & Sons, Chichester.

Sensier, M. and van Dijk, D. (2004). Testing for volatility changes in U.S. macroeconomic time series, *Review of Economics and Statistics* **86**: 833–839.

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

Stock, J. and Watson, M. (2002). Has the business cycle changed and why?, *in* M. Gertler and K. Rogoff (eds), *NBER Macroeconomics Annual 2002*, MIT Press, Cambridge, MA, pp. 159–230.

Zeileis, A., Leisch, F., Kleiber, C. and Hornik, K. (2005). Monitoring structural change in dynamic econometric models, *Journal of Applied Econometrics* **20**: 99–121.

A Derivation of prior marginal distribution θ_t

Proposition A.1. *If the joint of* $(\theta_1, \ldots, \theta_T)'$ *is constructed by the conditionals*

$$p(\theta_t|\boldsymbol{\theta}^{1,t-1}) = p(\theta_t|\theta_{t-1}) = \pi f_0(\theta_t) + (1-\pi)\mathbb{I}_{\{\theta_t=\theta_{t-1}\}}, \quad (t = 2, \ldots, T),$$

and the initialization $\theta_1 \sim f_0$*, then it holds that all* θ_t*'s are marginally* f_0 *distributed.*

Proof: For θ_1 it holds by definition. Now, for θ_2 it follows

$$\begin{aligned} p(\theta_2) &= \int p(\theta_2|\theta_1)p(\theta_1)d\theta_1 \\ &= \pi f_0(\theta_2)\int f_0(\theta_1)d\theta_1 + (1-\pi)\int f_0(\theta_1)\mathbb{I}_{\{\theta_2=\theta_1\}}d\theta_1 \\ &= \pi f_0(\theta_2) + (1-\pi)f_0(\theta_2). \end{aligned}$$

Now suppose the acclaimed holds for arbitrary t, i.e., $\theta_t \sim f_0$, then following the same structure as before it is obvious that θ_{t+1} has f_0 as marginal as well. ∎

Proposition A.2. *Suppose we have two probability density functions* f_0 *and* g*, both defined on the same sample space. If the stochastic process* $\{\theta_t\}$ *has transition density as in* (1) *and the initial state* $\theta_0 \sim g$*, then* $\theta_t \xrightarrow{D} f_0$ *when* $t \to \infty$*.*

Proof: By mathematical induction we iteratively solve for the marginal distributions and show that the influence of g dies out.

First, we show that the exact marginal distribution is determined by the following mixture pdf:

$$p(\theta_t) = \pi\sum_{j=0}^{t-1}(1-\pi)^j \times f_0(\theta_t) + (1-\pi)^t g(\theta_t).$$

Take $t = 1$, then the marginal of θ_1 is derived as follows:

$$\begin{aligned} p(\theta_1) &= \int p(\theta_1|\theta_0)p(\theta_0)d\theta_0 \\ &= \int \pi f_0(\theta_1)g(\theta_0)d\theta_0 + \int (1-\pi)g(\theta_0)\mathbb{I}_{\{\theta_1=\theta_0\}}d\theta_0 \\ &= \pi f_0(\theta_1) + (1-\pi)g(\theta_1). \end{aligned}$$

Hence, the acclaimed holds for $t = 1$. Now assume the equation holds for arbitrary t, then for $t + 1$ it holds that

$$\begin{aligned} p(\theta_{t+1}) &= \int p(\theta_{t+1}|\theta_t)p(\theta_t)d\theta_t \\ &= \int \left[\pi f_0(\theta_{t+1}) + (1-\pi)\mathbb{I}_{\{\theta_{t+1}=\theta_t\}}\right] \\ &\quad \times \left[(1-\pi)^t g(\theta_t) + \pi \sum_{j=0}^{t-1}(1-\pi)^j \times f_0(\theta_t)\right] d\theta_t \\ &= \pi(1-\pi)^t f_0(\theta_{t+1}) + (1-\pi)^{t+1} g(\theta_{t+1}) + \pi \sum_{j=0}^{t-1}(1-\pi)^j \times f_0(\theta_{t+1}). \end{aligned}$$

Second, because $\sum_{j=0}^{t-1}(1-\pi)^j \longrightarrow \frac{1}{\pi}$ and $(1-\pi)^t \longrightarrow 0$ if $t \to \infty$, the marginal pdf of θ_t converges to f_0 for every point in the sample space. ∎

B Introducing indicator variables

In order to clarify the differences and advantages with respect to other approaches it is helpful to augment the parameter vector with indicator variables s_t, $(t = 2, \ldots, T)$ as in Geweke and Jiang (2010). The model for the time-dependent parameters can then be written as (conditional on the break probability π):

$$s_t = \begin{cases} 1, & \text{with probability} \quad \pi, \\ 0, & \text{with probability} \quad 1-\pi, \end{cases}$$

$$p(\theta_t|\boldsymbol{\theta}^{1,t-1}, \mathbf{s}^{2,t}) = f_0(\theta_t)^{\mathbb{I}_{\{s_t=1\}}} \left(\mathbb{I}_{\{\theta_t=\theta_{t-1}\}}\right)^{1-\mathbb{I}_{\{s_t=1\}}}, \qquad (t = 2, \ldots, T).$$

Here s_t indicates which of the two mixture components θ_t is sampled from: from the baseline prior f_0 or the degenerate at θ_{t-1}. The densities above describe the joint $p(\boldsymbol{\theta}, \mathbf{s})$, with $\mathbf{s} = \mathbf{s}^{2,T}$. If we integrate out the auxiliary variables, we get the same specification as in (1):

$$\begin{aligned} p(\theta_t|\boldsymbol{\theta}^{1,t-1}) &= \sum_{\mathbf{s}^{2,t}} p(\theta_t|\boldsymbol{\theta}^{1,t-1}, \mathbf{s}^{2,t})p(\mathbf{s}^{2,t}) \\ &= \sum_{s_t=0,1} p(\theta_t|\theta_{t-1}, s_t)p(s_t) \\ &= \pi f_0(\theta_t) + (1-\pi)\mathbb{I}_{\{\theta_t=\theta_{t-1}\}}, \end{aligned}$$

where we use (i) that given θ_{t-1}, the current θ_t is independent from $\boldsymbol{\theta}^{1,t-2}$ and (ii) the temporal independence of the s_t's.

B.1 Posterior sampling

We have to traverse the space of $p(\boldsymbol{\theta}, \mathbf{s}, \pi|\mathbf{y})$. Analytical integration of $\mathbf{s}$ makes that we can set up a Gibbs sampler to simulate from $p(\boldsymbol{\theta}|\mathbf{y}, \pi)$ as described in Section 3.1. Denote its transition kernel $p^*(\boldsymbol{\theta}|\boldsymbol{\theta}^c, \pi)$. Conditional on a draw of $\boldsymbol{\theta}$, we simulate from $p(\mathbf{s}|\mathbf{y}, \boldsymbol{\theta}, \pi)$ – which is not degenerate. To update π we sample from $p(\pi|\mathbf{s}, \boldsymbol{\theta})$. We have to show that such a Markov chain has the required posterior as its invariant distribution. With γ^c we denote the current state of random variable γ. The invariant distribution is derived as follows:

$$
\begin{aligned}
&\int_{\mathbf{s}^c, \boldsymbol{\theta}^c, \pi^c} p(\boldsymbol{\theta}^c, \mathbf{s}^c, \pi^c) p^*(\boldsymbol{\theta}|\boldsymbol{\theta}^c, \pi^c) p(\mathbf{s}|\boldsymbol{\theta}, \pi^c) p(\pi|\mathbf{s}, \boldsymbol{\theta}) \\
&\quad = \int_{\mathbf{s}^c, \boldsymbol{\theta}^c, \pi^c} p(\mathbf{s}^c|\boldsymbol{\theta}^c, \pi^c) p(\boldsymbol{\theta}^c|\pi^c) p(\pi^c) p^*(\boldsymbol{\theta}|\boldsymbol{\theta}^c, \pi^c) p(\mathbf{s}|\boldsymbol{\theta}, \pi^c) p(\pi|\mathbf{s}, \boldsymbol{\theta}) \\
&\quad = \int_{\boldsymbol{\theta}^c, \pi^c} p(\boldsymbol{\theta}^c|\pi^c) p^*(\boldsymbol{\theta}|\boldsymbol{\theta}^c, \pi^c) p(\pi^c) p(\mathbf{s}|\boldsymbol{\theta}, \pi^c) p(\pi|\mathbf{s}, \boldsymbol{\theta}) \\
&\quad = \int_{\pi^c} p(\boldsymbol{\theta}|\pi^c) p(\pi^c) p(\mathbf{s}|\boldsymbol{\theta}, \pi^c) p(\pi|\mathbf{s}, \boldsymbol{\theta}) \\
&\quad = \int_{\pi^c} p(\boldsymbol{\theta}, \mathbf{s}, \pi^c) p(\pi|\mathbf{s}, \boldsymbol{\theta}) \\
&\quad = p(\pi|\boldsymbol{\theta}, \mathbf{s}) p(\boldsymbol{\theta}, \mathbf{s}) = p(\boldsymbol{\theta}, \mathbf{s}, \pi),
\end{aligned}
$$

where we drop the notational conditioning on $\mathbf{y}$ for purposes of exposition. In this derivation (third equality) we use that we have a valid Gibbs sampler to simulate from $p(\boldsymbol{\theta}|\pi)$. That is, its transition kernel is

$$
p^*(\boldsymbol{\theta}|\boldsymbol{\theta}^c, \pi) = p(\theta_1|\theta_2^c) \prod_{t=2}^{T-1} p(\theta_t|\theta_{t-1}, \theta_{t+1}^c) \times p(\theta_T|\theta_{T-1}),
$$

and

$$
\int_{\boldsymbol{\theta}^c} p(\boldsymbol{\theta}^c|\pi) p^*(\boldsymbol{\theta}|\boldsymbol{\theta}^c, \pi) = p(\boldsymbol{\theta}|\pi).
$$

The complete posterior simulation scheme breaks down into three steps.

Step 1. Sample a new $\boldsymbol{\theta}$ by simulating from $p(\theta_t|\theta_{t-1}, \theta_{t+1}, \pi)$ for $t = 1, \ldots, T$. This means sampling from a three-component mixture as in Section 3.1.

Step 2. Sample a new $\mathbf{s}$ by simulating from $p(s_t|\boldsymbol{\theta}, \pi)$ for $t = 2, \ldots, T$. In this step we use that the s_t's are independent and that

$$p(s_t|\boldsymbol{\theta}, \pi) \quad \propto \quad [\pi f_0(\theta_t)]^{\mathbb{I}_{\{s_t=1\}}} \left[(1-\pi)\mathbb{I}_{\{\theta_t=\theta_{t-1}\}}\right]^{\mathbb{I}_{\{s_t=0\}}}.$$

If $\theta_t \neq \theta_{t-1}$, then s_t is degenerate; $\mathsf{Pr}\left[s_t = 1|\boldsymbol{\theta}, \pi\right] = 1$.

If $\theta_t = \theta_{t-1}$, then the indicator is sampled from

$$p(s_t|\boldsymbol{\theta}, \pi) \propto \begin{cases} 1-\pi, & s_t = 0, \\ \pi f_0(\theta_t), & s_t = 1. \end{cases}$$

Step 3. Sample the breaking probability from $p(\pi|\boldsymbol{\theta}, \mathbf{s}) = p(\pi|\mathbf{s})$. For this full conditional it holds that

$$p(\pi|\mathbf{s}) \propto p(\mathbf{s}|\pi)p(\pi) = \pi^{\sum_{t=2}^{T} \mathbb{I}_{\{s_t=1\}}} (1-\pi)^{\sum_{t=2}^{T} \mathbb{I}_{\{s_t=0\}}} \times p(\pi).$$

Therefore, taking a Beta prior yields a conjugate full conditional posterior distribution:

$$\pi \sim \mathcal{B}e(r_1, r_2) \qquad \Longrightarrow \qquad \pi|\mathbf{y}, \mathbf{s} \sim \mathcal{B}e(K^* + r_1, T - 1 - K^* + r_2),$$

with $K^* = \sum_{t=2}^{T} \mathbb{I}_{\{s_t=1\}}$. We note that $K^* \geq \sum_{t=2}^{T} \mathbb{I}_{\{\theta_t \neq \theta_{t-1}\}} = K - 1$; the sum of the indicators is larger than or equal to the number of parameter breaks. This shows the reason why we need the augmentation step to be able to sample from a Beta distribution. Since K^* will be close to the number of breaks, a Metropolis–Hastings sampler with a Beta proposal with $K^* = K - 1$ is a good alternative simulator.