University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Dufays, A;Li, Z;Rombouts, JVK;Song, Y

Title:
Sparse change-point VAR models

Date:
2021-09-01

Citation:
Dufays, A., Li, Z., Rombouts, J. V. K. & Song, Y. (2021). Sparse change-point VAR models. Journal of Applied Econometrics, 36 (6), pp.703-727. https://doi.org/10.1002/jae.2844.

Persistent Link:
https://hdl.handle.net/11343/302998

Check for updates

Author Manuscript

Journal of Applied Econometrics MS 12690

Sparse Change-point VAR models

Arnaud Dufays[1] and Zhuo Li[2] and Jeroen V.K. Rombouts[3] and Yong Song[4]

November 10, 2020

Abstract

Change-point (CP) VAR models face a dimensionality curse due to the proliferation of parameters that arises when new breaks are detected. We introduce the Sparse CP-VAR model which determines which parameters truly vary when a break is detected. By doing so, the number of new parameters to be estimated at each regime is drastically reduced and the break dynamics becomes easier to be interpreted. The Sparse CP-VAR model disentangles the dynamics of the mean parameters and the covariance matrix. The former use CP dynamics with shrinkage prior distributions while the latter are driven by an infinite hidden Markov framework. An extensive simulation study is carried out to compare our approach with existing ones. We provide applications to financial and macroeconomic systems. It turns out that many off-diagonal VAR parameters are zero for the entire sample period and that most break activity is in the covariance matrix. We show that this has important consequences for portfolio optimisation, in particular when future instabilities are included in the predictive densities. Forecasting-wise, the Sparse CP-VAR model compares favorably to several time-varying parameter models in terms of density and point forecast metrics.

Keywords: Shrinkage prior, Structural breaks, Change-point model, Time-varying parameters, Forecasting, Relevant parameter change, Buy-and-hold portfolio strategy.

JEL Classification: C11, C15, C22, C58.

[1]Université de Laval, 2216 Pavillon J.-A.DeSève, G1V0A6, Québec, Canada.https://arnauddufays.com/
[2]University of Melbourne
[3]ESSEC Business School, CREST
[4]University of Melbourne

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jae.2844

This article is protected by copyright. All rights reserved.

Author Manuscript

1 Introduction

It is well known that economic and financial time series observed over a long period are subject to structural breaks, the dates of which are unknown. Evidence can be found in Stock and Watson (1996), Clark and McCracken (2010), Bauwens, Koop, Koribilis, and Rombouts (2015) and in Aastveit, Carriero, Clark, and Massimiliano (2017) for macroeconomic series, and in Pastor and Stambaugh (2001) and Liu and Maheu (2007) for financial series. Ignoring breaks when modelling long time series creates bias in parameter estimates, gives misleading interpretation, and may generate severe forecasting errors (see, e.g., Pesaran and Timmermann (2004)). As a consequence, structural break or change-point (CP) time series models have been developed to identify substantial but infrequent parameter changes. Seminal contributions on such univariate structural break models are proposed in Chib (1998), Pesaran, Pettenuzzo, and Timmermann (2006), Koop and Potter (2007), Giordani and Kohn (2008), Maheu and Gordon (2008) and Geweke and Jiang (2011).

The modelling of multivariate time series makes the problem of structural breaks more complex, since for example the break dates potentially differ among the time series. In terms of cross-relationships, a variable might be significantly impacting another variable until a structural break happens, or a variable might become significant only after a break. Direct extensions from the univariate to the multivariate structural break model have been implemented by Pettenuzzo and Timmermann (2011), and Chib and Kang (2013), though the lack of sparsity makes it difficult to estimate the models in practice given the large number of parameters even for standard low dimensional models with few lags. Extending these two frameworks, Maheu and Song (2018) offer an efficient sampler to estimate multiple structural breaks in multivariate time series.

This paper proposes a sparse change-point vector autoregressive model, the Sparse CP-VAR, to model multivariate time series subject to multiple unknown structural breaks. We relax the assumption that all parameters need to change when detecting automatically for structural breaks. The approach is general such that breaks can happen in either the autoregressive parameters, off-diagonal VAR parameters or covariance matrix parameters. We achieve this by shrinking irrelevant break-parameters towards zero using an estimation procedure which is based on a sampling scheme tailored to CP models. The parameters of the VAR model can be shrunk to zero over the entire sample period. Our shrinkage methodology alleviates the search of the number of regimes and relies only on one estimation output rather than estimating different models and comparing their marginal likelihoods. For generality, breaks in the mean are separated from covariance matrix breaks which are modeled using a hierarchical Dirichlet process. The latter feature has an important impact on the parsimony

This article is protected by copyright. All rights reserved.

of our CP-VAR model in our applications.

We tackle the over-parametrization issue of using continuous spike and slab shrinkage priors by calibrating the considerable amount of hyper-parameters to the uniform mixture shrinkage prior of Dufays and Rombouts (2020). Our proposed solution tailored to CP models has the following three advantages. First, although all the parameters are allowed to vary from one regime to another, we can assess which ones are actually time varying. Second, the estimation accuracy of the parameters is improved compared to the standard CP process since only relevant parameters have to be estimated when a new regime occurs. In addition, when break dates are common among parameters, which is frequently the case for macroeconomic series, the accuracy of their identification increases with the number of series in the system. Third, shrinkage priors allow inferring directly the number of regimes without relying on marginal likelihood computations.

The paper that comes closest to our sparse CP-VAR framework is Huber, Kastner, and Feldkircher (2019). They introduce the estimation of large VARs with mixture innovation components, as originally proposed by Giordani and Kohn (2008), for each parameter coefficient in the system. They propose a threshold approach which approximates the sampling of state indicators. Computationally attractive, their Markov-chain Monte Carlo (MCMC) sampler efficiently explores the posterior support but does not necessarily converge to the posterior of interest as shown in our Internet Appendix.

We illustrate several features of the Sparse CP-VAR model on simulated data. We detail the ability of detecting the number of regimes for each model parameter and compare with alternative CP models and Lasso based methods (e.g. Chan, Yip, and Zhang, 2014). We study the precision of the parameter estimates. We perform robustness checks by varying the size of the parameter breaks, by comparing our approach to the standard spike and slab prior, and by estimating a misspecified VAR model. We also investigate the break detection accuracy when some VAR parameters experience the same break dates.

We provide two applications. We first revisit the exercise of Pettenuzzo and Timmermann (2011) on excess return predictability. We find no evidence of breakpoints in the mean parameters while Pettenuzzo and Timmermann (2011) found six regimes. This difference comes from the large number of breakpoints found in the covariance matrix which in our approach do not trigger a change in the mean parameters. Although our results favor constant mean dynamics, a forecasting exercise shows that, in terms of predictive likelihood, we dominate Pettenuzzo and Timmermann's model as well as the CP-VAR model of Maheu and Song (2018) and the TVP model with stochastic volatility of Primiceri (2005). We illustrate the impact of incorporating future breaks on the investor's portfolio share in a risky asset. Our second application illustrates the in- and out-of-sample performance of the Sparse CP-VAR

Author Manuscript

This article is protected by copyright. All rights reserved.

model on 19 US macroeconomic variables. The preferred specification detects few breakpoints, and overall forecasts better than the standard VAR processes as well as the TVP-VAR model with stochastic volatility of Koop and Korobilis (2013).

The rest of the paper is organized as follows. Section 2 defines the Sparse CP-VAR model. Besides the model specification, we explicit the shrinkage prior and the choice of its hyperparameters. Section 3 details the estimation algorithm. Section 4 explains how we deal with time varying covariance dynamics using a hierarchical Dirichlet process. Section 5 provides a detailed simulation study. Section 6 shows the performance of the Sparse CP-VAR model in empirical applicaitons. Section 7 concludes and the Internet Appendix contains technical details and additional simulation results.

2 Sparse Change-point VAR models

2.1 Model specification

We denote a multivariate time series by $y_{1:T} = \{y_1, ..., y_T\}$ where $y_t \in \mathbb{R}^{n\times 1}$ $\forall t$ and a discrete break date vector $\tau_{1:K} = \{\tau_1, ..., \tau_K\}$ such that $\tau_0 \equiv 0 < \tau_1 < \tau_2 < \ldots < \tau_K < \tau_{K+1} \equiv T$. A CP-VAR(p) model with $K+1$ regimes is specified, for $t \in [\tau_j + 1, \tau_{j+1}]$ and $j = 0, ..., K$, as

$$\begin{aligned} y_t &= a_j^{(0)} + \sum_{i=1}^{p} a_j^{(i)} y_{t-i} + \varepsilon_t \;\text{ with }\; \varepsilon_t \sim \mathcal{N}(0, \Sigma_j), \\ &= A_j x_t + \varepsilon_t, \end{aligned} \tag{1}$$

in which $A_j = (a_j^{(0)} \;\; a_j^{(1)} \;\; \ldots \;\; a_j^{(p)}) \in \mathbb{R}^{n\times k}$, $x_t = (1 \;\; y'_{t-1} \;\; \ldots \;\; y'_{t-p})' \in \mathbb{R}^{k\times 1}$ and $k = (n\times p)+1$. Inference on CP models is typically done using a MCMC sampler with the number of regimes determined by the marginal likelihood. Chib (1998) puts forward a self-contained method for inference on CP processes. In the multivariate case however, the number of parameters in the CP-VAR model is prohibitively large even for a low dimensional system with few lags. We propose the implementation of shrinkage priors which allows working in such a high dimensional parameter setting. Its use requires rewriting the CP-VAR(p) in (1) as follows

$$y_t = [A_0 + \sum_{i=1}^{K} \Delta A_i \mathbb{1}_{\{t>\tau_i\}}] x_t + \varepsilon_t \text{ with } \varepsilon_t \sim \mathcal{N}(0, \Sigma_t), \tag{2}$$

where $\Delta A_i = A_i - A_{i-1}$ $\forall i \in [1, K]$ and $\Delta A_i \in \mathbb{R}^{n\times k}$. For interpretation and forecasting, we make use of the original VAR parameters in levels rather than differences. The first regime parameter corresponds to A_0 and subsequent parameters can simply be obtained as

Author Manuscript

This article is protected by copyright. All rights reserved.

$A_j = (A_0 + \sum_{i=1}^{j} \Delta A_i)$ for $j \in [1, K]$. Working with differences in parameters is beneficial for two reasons. First, the prior distributions of the parameters do not have to depend on the previous regimes. Having dependent priors over regimes complicates the estimation, see for example Koop and Potter (2007). Second, a natural choice of the mean of a parameter in first-difference is zero, i.e. no break, which eases the use of shrinkage priors. Note that in (2) the covariance matrix dynamics is now separated from the mean dynamics. As our model allows dissociating mean and covariance parameters, we postpone the discussion of the variance dynamics until Section 4.

2.2 Shrinkage prior

Many Bayesian shrinkage priors that can be expressed as scale mixtures of normals exist in the literature, e.g. the continuous spike and slab of Ishwaran and Rao (2005), the Bayesian Lasso of Park and Casella (2008), the horseshoe prior of Carvalho, Polson, and Scott (2010), the Normal-Gamma of Griffin and Brown (2010), or the Dirichlet-Laplace prior of Bhattacharya, Pati, Pillai, and Dunson (2015). While any of these priors could be employed, we use the spike and slab (S&S) prior for shrinking the parameters because its two component mixture facilitates the interpretation of breaks. The S&S prior is specified by the following hierarchical structure

$$\begin{aligned}
\mathrm{Vec}(A_0)|\Gamma_0 &\sim \mathcal{N}(0, \underbrace{\mathrm{diag}\{\phi_0^{(1)}, \ldots, \phi_0^{(nk)}\}}_{\Gamma_0}), && (3)\\
\mathrm{Vec}(\Delta A_i)|\Gamma_i &\sim \mathcal{N}(0, \underbrace{\mathrm{diag}\{\phi_i^{(1)}, \ldots, \phi_i^{(nk)}\}}_{\Gamma_i}), \quad \forall i \in [1, K] && (4)\\
\phi_i^{(j)}|z_i^{(j)}, (\sigma_i^{(j)})^2 &= c_{z_i^{(j)}}^{(j)} (\sigma_i^{(j)})^2 \quad \forall i \in [0, K],\ \forall j \in [1, nk], && (5)\\
(\sigma_i^{(j)})^2 &\sim \mathcal{IG}(v^{(j)}/2, v^{(j)}/2), \quad \forall i \in [0, K],\ \forall j \in [1, nk], && (6)\\
z_i^{(j)} &\sim \mathrm{Bernoulli}(1 - \omega_i^{(j)}), \quad \forall i \in [0, K]\ \ \forall j \in [1, nk], && (7)\\
\omega_i^{(j)} &\sim \mathcal{B}eta(\alpha_{\omega_i^{(j)}}, \beta_{\omega_i^{(j)}}), \quad \forall i \in [0, K]\ \ \forall j \in [1, nk], && (8)
\end{aligned}$$

where $\mathcal{IG}(\cdot, \cdot)$ denotes the inverse-Gamma. The discrete variable $c_{z_i^{(j)}}^{(j)}$ acts as a spike if $z_i^{(j)} = 0$ or a slab if $z_i^{(j)} = 1$. The spike reduces the variance since $c_0^{(j)} < c_1^{(j)}$. The spike and slab prior is continuous because the random variable $(\sigma_i^{(j)})^2$ makes a bridge between the two components by altering their variances. The shrinkage property of the continuous S&S prior is intuitively understood from the marginal density function of a mean parameter. Direct calculations give

Author Manuscript

This article is protected by copyright. All rights reserved.

that

$$f_{\text{S\&S}}(a_i^{(j)}|v^{(j)},\omega_i^{(j)}) \quad = \quad \frac{\omega_i^{(j)}}{\sqrt{c_0^{(j)}}} f_t(\frac{|a_i^{(j)}|}{\sqrt{c_0^{(j)}}}|v^{(j)}) + \frac{(1-\omega_i^{(j)})}{\sqrt{c_1^{(j)}}} f_t(\frac{|a_i^{(j)}|}{\sqrt{c_1^{(j)}}}|v^{(j)}), \tag{9}$$

where the parameter $a_i^{(j)}$ denotes the jth element of $\text{Vec}(A_0)$ if $i = 0$ and of $\text{Vec}(\Delta A_i)$ for $i \in [1, K]$ and $f_t(\cdot|v)$ denotes the student density function with v degrees of freedom. The continuous S&S prior adds a considerable number of random variables and hyper-parameters into the model. In this paper, we explore two different ways of fixing the hyper-parameters. First, we develop a new calibration procedure in the next section. Second, in Internet Appendix E.1 we document results using the most common specification of the S&S hyper-parameters.

2.2.1 Calibration of the S&S hyper-parameters

The S&S shrinkage priors augment the model specification with a lot of hyper-parameters. Fortunately, many of these hyper-parameters are fixed using standard rules developed in Dufays and Rombouts (2020) and Dufays and Rombouts (2018) such that the implied marginal distribution (9) mimics a mixture of two uniform distributions (hereafter denoted by 2MU) and which turns out to be effective in CP models. More explicitly, the 2MU prior is defined by three parameters: $a^{\text{2MU}}, b^{\text{2MU}}$, with $0 < a^{\text{2MU}} < b^{\text{2MU}}$, which denotes the bounds of the uniform components and P that stands for the penalty parameter. The 2MU distribution on model parameter $a_i^{(j)}$ is given by

$$a_i^{(j)}|a_j^{\text{2MU}}, b_j^{\text{2MU}}, P \quad \sim \quad \omega^{\text{2MU}}(P)\mathcal{U}[\frac{-a_j^{\text{2MU}}}{2}, \frac{a_j^{\text{2MU}}}{2}] + (1-\omega^{\text{2MU}}(P))\mathcal{U}[\frac{-b_j^{\text{2MU}}}{2}, \frac{b_j^{\text{2MU}}}{2}], \tag{10}$$

with $\omega_{\text{2MU}}(P) = \frac{a_j^{\text{2MU}}(1-e^P)}{b_j^{\text{2MU}}e^P + a_j^{\text{2MU}}(1-e^P)}$, obtained from $\log f(a_i^{(j)} = \frac{b_j^{\text{2MU}}}{2}|a_j^{\text{2MU}}, b_j^{\text{2MU}}, P) - \log f(a_i^{(j)} = 0|a_j^{\text{2MU}}, b_j^{\text{2MU}}, P) = P$. In particular, P implies a prior on having a break in the parameter because it controls the probability of being in the wide component. The parameter P penalizes the log-likelihood function by an amount of P when the underlying parameter lies in the region $[\frac{-b_j^{\text{2MU}}}{2}, \frac{b_j^{\text{2MU}}}{2}] \backslash [\frac{-a_j^{\text{2MU}}}{2}, \frac{a_j^{\text{2MU}}}{2}]$, i.e. away from zero. The 2MU prior implies a L_0 penalized likelihood and more negative values of P imply less detected breaks. Ultimately if $P \to -\infty$, the process has no break and becomes a standard VAR model. As shown in Internet Appendix D, we motivate the following hyper-parameters of the 2MU distribution

Author Manuscript

This article is protected by copyright. All rights reserved.

for each parameter $a_i^{(j)} \sim 2\text{MU}(a_j^{2\text{MU}}, b_j^{2\text{MU}}, P)$:

$$
\begin{aligned}
a_j^{2\text{MU}} &= \frac{1}{3}\hat{\sigma}_{a_0^{(j)}}, && (11)\\
b_j^{2\text{MU}} &= \frac{a_j^{2\text{MU}}(19TK+1)}{K}, && (12)\\
P &= P_{0.95}, && (13)\\
\omega^{2\text{MU}}(P_{0.95}) &= \frac{a_j^{2\text{MU}}(1-e^{P_{0.95}})}{b_j^{2\text{MU}}e^{P_{0.95}} + a_j^{2\text{MU}}(1-e^{P_{0.95}})}, && (14)
\end{aligned}
$$

in which $\hat{\sigma}_{a_0^{(j)}}$ denotes the standard error of the OLS estimate of $a_0^{(j)}$ and $P_\pi = -(\ln \frac{\pi}{1-\pi} + \ln T)$ with $\pi \in [0.5, 1)$. The penalty value $P_{0.95}$ is equivalent to strong evidence in favor of a breakpoint in the scale of Kass and Raftery (1995).

To optimally mimic the 2MU distribution, we minimize the Kullback-Leibler divergence between the 2MU prior (see Equation (10)) and the marginal density function of the continuous spike and slab prior given by Equation (9) with respect to c_0, c_1 and v. So, we use the same weight for the continuous spike and slab distribution as for the 2MU prior distributions, i.e. $\omega^{2\text{MU}}(P_{0.95})$, see Equation (14). In particular, we minimize the divergence by maximizing the log-likelihood function of the spike and slab prior given a sample from the 2MU distribution. The hyper-parameters of the beta distribution of the continuous spike and slab given in Equation (8) are computed as, for any $i \in [0, K]$ and $j \in [1, nk]$,

$$
\begin{aligned}
\omega_i^{(j)} &\sim \mathcal{B}eta(a_{\omega_i^{(j)}}, b_{\omega_i^{(j)}}),\\
\{a_{\omega_i^{(j)}}, b_{\omega_i^{(j)}}\} &= \arg\min(P[\omega_i^{(j)} \in [\omega^{2\text{MU}}(P_{0.75}), \omega^{2\text{MU}}(P_{0.99})]] - 0.999)^2.
\end{aligned}
$$

Making $\omega_i^{(j)}$ a random variable has the appealing property of facilitating the movement from the spike to the slab component and vice versa, and therefore improving the mixing of the sampler. This structure is similar to the S&S prior of Ishwaran and Rao (2005).

2.3 Choosing the number of breaks K

Intuitively, the S&S priors will shrink parameters of irrelevant regimes to zero and so the choice of potential breaks K does not matter as long as it is large enough. We set the number of breaks to a small but conservative number as $K = \lceil \ln T \rceil$. While the number of breaks increases with T, it remains bounded by the maximum growth rate of the number of VAR segments in Callot and Kristensen (2015) so their asymptotic results apply. Also, by fixing the number of breaks smaller than the sample size, we stick to the baseline idea of CP models that

Author Manuscript

This article is protected by copyright. All rights reserved.

is detecting few abrupt changes in the parameters. Imposing $K < T$ has relevant practical consequences because it solves the MCMC convergence issues and it drastically decreases the computational burden. To illustrate this, consider our financial application in which we have a trivariate VAR model with 2 lags and a sample size of $T = 949$. Assuming $K = T$ implies 19,929 mean parameters to estimate. Our approach limits this number to 144. With respect to the approach of estimating the number of breaks using the marginal likelihood, our method has the advantage of considerably reducing the computational burden as only one model needs to be estimated.

Because we choose $K < T$, we need a prior for the breaks. We do not directly model the breakpoints $\tau_{1:K}$ since it complicates the sampler and the elicitation of the prior distributions.[1] We make use of the latent variable $s_t \in [1, K+1]$ to model the break dynamics as in Chib (1998). This discrete variable denotes the active regime at time t. The break dynamics $s_{1:T} = \{s_1, s_2, \ldots, s_T\}$ is driven by a first-order Markov chain with transition matrix given by

$$P_{CP} = \begin{pmatrix} p_1 & 1-p_1 & 0 & ... & 0 \\ 0 & p_2 & 1-p_2 & ... & 0 \\ & & ... & & \\ 0 & 0 & ... & 0 & 1 \end{pmatrix}. \tag{15}$$

Note that the break date parameters $\tau_{1:K}$ are directly built from the state vector $s_{1:T} = \{s_1, s_2, \ldots, s_T\}$ given the one-to-one mapping between the two variables.

2.4 Predictive densities subject to future breaks

Pettenuzzo and Timmermann (2011) introduce out-of-sample breaks in an optimal portfolio application and conclude that such breaks have a much larger impact on asset allocation than parameter uncertainty. This is due to the fact that, when a break occurs in the predictive process, all the VAR parameters of the new regime are drawn from a prior calibrated using the in-sample parameter values. However, when the number of in-sample breaks is found to be small empirically, the learning process about the break size is limited and accounting for future breaks drastically increases prediction uncertainty.

In our framework, we use additional information on the break size to calibrate future structural changes. In particular, our Sparse CP-VAR model innovates in two aspects. First, the covariance matrix independently evolves from the mean dynamics. Therefore, the mean parameters do not necessarily experience future breaks when the covariance matrix abruptly

[1] The best sampler based on a single chain, we are aware of, is Stephens (1994). However, it consists of sampling one at a time the break dates in a MCMC context.

Author Manuscript

This article is protected by copyright. All rights reserved.

changes. Second, our process learns which parameters are time-varying over the sample period which is useful information to realistically incorporate future parameter instabilities when computing predictive densities. In particular, we sample the break size of a parameter in a future regime i, that is $a_i^{(j)}, \forall j \in [1, nk]$ and $i > K$, from the prior,

$$a_i^{(j)} | a_{0:K}^{(j)}, z_{0:K}^{(j)} \quad \sim \quad \omega^{(j)} \mathcal{N}(\bar{\mu}^{(j)}, (\bar{\sigma}^{(j)})^2) + (1 - \omega^{(j)}) \mathbb{1}_0, \tag{16}$$

where $\omega^{(j)} = \frac{\tilde{K}^{(j)}}{K}$ (i.e. the probability of an in-sample break) with $\tilde{K}^{(j)} = \sum_{i=1}^{K} z_i^{(j)}$ (i.e., the number of active in-sample breaks), $\bar{\mu}^{(j)} = \mathbb{1}_{\{\tilde{K}^{(j)} > 1\}} \frac{\sum_{i=1}^{K} a_i^{(j)} \mathbb{1}_{\{z_i^{(j)} = 1\}}}{\tilde{K}^{(j)}}$ and $(\bar{\sigma}^{(j)})^2 = \mathbb{1}_{\{\tilde{K}^{(j)} > 1\}} \frac{\sum_{i=1}^{K} (a_i^{(j)} - \bar{\mu}^{(j)})^2 \mathbb{1}_{\{z_i^{(j)} = 1\}}}{\tilde{K}}$. If no in-sample break in a parameter has been detected, this distribution implies a point mass at zero. When at least two breaks in the parameters $a_{0:K}^{(j)}$ occur in-sample, the prior distribution of Equation (16) becomes a mixture distribution with one component being a Normal distribution centered around the average size of the breaks. Regarding the distribution of the probability of staying in the same regime at the next period (i.e. p_{future}), we set a prior in line with Pesaran, Pettenuzzo, and Timmermann (2006) that is specified as $f(p_{\text{future}} | \tau_{1:K}) \sim B(T - K, K)$. We could of course consider more sophisticated prior distributions based on the estimated break dates $\tau_{1:K}$ or on the regime durations as in Koop and Potter (2007).

3 Inference

The MCMC sampler exhibits two innovations compared to standard inference on CP-VAR models. First, we jointly sample the mean parameters which significantly improves the MCMC mixing. Second, we enhance the sampling of the spike and slab parameters. The latter innovation can be adapted to any other spike and slab application. The full MCMC sampler can be found in Internet Appendix A.

3.1 Sampling the mean parameters

Let us introduce the unit vector $e_K(i) = (\underbrace{1 \;\; 1 \;\; \ldots \;\; 1}_{\#i} \;\; 0 \;\; \ldots \;\; 0)' \in \mathbb{R}^{K+1 \times 1}$. To sample the mean parameters, we rewrite the CP-VAR(p) model as follows

$$\begin{aligned} y_t &= [A_0 + \sum_{i=1}^{K} \Delta A_i \mathbb{1}_{\{t > \tau_i\}}] x_t + \epsilon_t \text{ with } \epsilon_t \sim \mathcal{N}(0, \Sigma_t), \\ &= A \tilde{x}_t + \epsilon_t, \end{aligned} \tag{17}$$

Author Manuscript

This article is protected by copyright. All rights reserved.

where $A = (A_0 \;\; \Delta A_1 \;\; \ldots \;\; \Delta A_K) \in \mathbb{R}^{n\times\bar{k}}$, $\tilde{x}_t = e_K(s_t)\otimes x_t = (x_t'\mathbb{1}_{s_t\geq 1} \;\; x_t'\mathbb{1}_{s_t\geq 2} \;\; \ldots x_t'\mathbb{1}_{s_t\geq K+1})' \in \mathbb{R}^{\bar{k}\times 1}$ and $\bar{k} = (K+1)((n\times p)+1)$.
In matrix form, the model reads

$$\begin{aligned} Y &= AX + E, \\ \mathrm{Vec}(Y) &= \mathrm{Vec}(AX) + \mathrm{Vec}(E), \end{aligned} \tag{18}$$

where $Y = (y_1 \;\; y_2 \;\; \ldots \;\; y_T) \in \mathbb{R}^{n\times T}$, $X = (\tilde{x}_1 \;\; \tilde{x}_2 \;\; \ldots \;\; \tilde{x}_T) \in \mathbb{R}^{\bar{k}\times T}$ and $E = (\epsilon_1 \;\; \epsilon_2 \;\; \ldots \;\; \epsilon_T) \in \mathbb{R}^{n\times T}$. The covariance matrix Ω of $\mathrm{Vec}(E)$ is given by

$$\Omega = \begin{pmatrix} \Sigma_1 & 0_{n\times n} & \ldots & 0_{n\times n} \\ 0_{n\times n} & \Sigma_2 & \ldots & 0_{n\times n} \\ & & \ldots & \\ 0_{n\times n} & 0_{n\times n} & \ldots & \Sigma_T \end{pmatrix} \in \mathbb{R}^{nT\times nT}. \tag{19}$$

The likelihood function can be expressed as follows

$$\begin{aligned} f(Y|A, \Sigma_{1:K_\Sigma}, s^{\Sigma}_{1:T}, \tau_{1:K}) &\propto e^{-\frac{1}{2}(\mathrm{Vec}(Y-AX)'\Omega^{-1}\mathrm{Vec}(Y-AX))}, \\ &\propto e^{-\frac{1}{2}(\mathrm{Vec}(A)'(X\otimes I_n)\Omega^{-1}(X'\otimes I_n)\mathrm{Vec}(A) - 2\mathrm{Vec}(Y)'\Omega^{-1}(X'\otimes I_n)\mathrm{Vec}(A))}. \end{aligned}$$

The full conditional posterior distribution is therefore normally distributed since the prior distribution of A is a Normal distribution conditional on the S&S parameters. More specifically, we have that

$$\begin{aligned} &f(A|Y, \Sigma_{1:K_\Sigma}, s^{\Sigma}_{1:T}, \Gamma_{0:K}, \tau_{1:K}) \propto \\ &\quad e^{-\frac{1}{2}(\mathrm{Vec}(A)'((X\otimes I_n)\Omega^{-1}(X'\otimes I_n)+\underline{\Omega}^{-1})\mathrm{Vec}(A) - 2(\mathrm{Vec}(Y)'\Omega^{-1}(X'\otimes I_n)+\mathrm{Vec}(\underline{A})'\underline{\Omega}^{-1})\mathrm{Vec}(A))}, \end{aligned}$$

in which $\underline{A}$ and $\underline{\Omega}$ denote the expectation and covariance matrix of the prior distribution of A given $\Gamma_{0:K}$, respectively (see Equation (3)). We conclude that

$$\mathrm{Vec}(A)|Y, \Sigma_{1:K_\Sigma}, s^{\Sigma}_{1:T}, \Gamma_{0:K} \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma}),$$

where $\bar{\Sigma} = ((X\otimes I_n)\Omega^{-1}(X'\otimes I_n) + \underline{\Omega}^{-1})^{-1}$, $\bar{\mu} = \bar{\Sigma}[(X\otimes I_n)\Omega^{-1}\mathrm{Vec}(Y) + \underline{\Omega}^{-1}\mathrm{Vec}(\underline{A})]$. In practice, exploiting the sparsity of Ω speeds up the computation. In addition, as proposed by Chan (2013), we save time by computing the posterior expectation as follows

This article is protected by copyright. All rights reserved.

Author Manuscript

$\bar{\mu} = (C_{\bar{\Sigma}^{-1}})^{-1}(C'_{\bar{\Sigma}^{-1}})^{-1}[(X \otimes I_n)\Omega^{-1}\text{Vec}(Y) + \underline{\Omega}^{-1}\text{Vec}(\underline{A})]$ in which we use the Cholesky decomposition: $C'_{\bar{\Sigma}^{-1}}C_{\bar{\Sigma}^{-1}} = \bar{\Sigma}^{-1}$.

3.2 Sampling the spike and slab parameters

In the shrinkage literature, e.g. Ishwaran and Rao (2005) and Malsiner-Walli and Wagner (2016a), the spike and slab parameters are usually sampled sequentially as follows, $\forall i \in [0, K]$ and $\forall j \in [1, nk]$,

1. $(\sigma_i^{(j)})^2 | a_i^{(j)}, c_0^{(j)}, c_1^{(j)}, z_i^{(j)}, v^{(j)} \sim \mathcal{IG}(\frac{v^{(j)}+1}{2}, \frac{(a_i^{(j)})^2/(c_{z_i^{(j)}}^{(j)})+v^{(j)}}{2})$
2. $\omega_i^{(j)} | z_i^{(j)} \sim \mathcal{B}eta(a_\omega + \mathbb{1}_{\{z_i^{(j)}=0\}}, b_\omega + \mathbb{1}_{\{z_i^{(j)}=1\}})$
3. $\Pr(z_i^{(j)} = 1 | a_i^{(j)}, (\sigma_i^{(j)})^2, \omega_i^{(j)}) = \frac{f(a_i^{(j)}|(\sigma_i^{(j)})^2, z_i^{(j)}=1)(1-\omega_i^{(j)})}{f(a_i^{(j)}|(\sigma_i^{(j)})^2, z_i^{(j)}=1)(1-\omega_i^{(j)}) + f(a_i^{(j)}|\sigma_i^{(j)})^2, z_i^{(j)}=0)\omega_i^{(j)}}$.

Nevertheless, the variables $\{\sigma_i^{(j)}, z_i^{(j)}\}_{i,j}$ are bound to exhibit high correlation. We propose a joint sampler for these variables. To alleviate the notation, we drop the indices i, j and define $c = (c_0, c_1)'$. The joint density function of $\{\sigma^2, z\}$ can be decomposed as

$$f(\sigma^2, z | v, c, a, \omega) \quad = \quad f(z | v, c, a, \omega) f(\sigma^2 | v, c, a, z),$$

in which the distribution $\sigma^2 | v, c, a, z \sim \mathcal{IG}(\frac{v+1}{2}, \frac{a^2/c_z+v}{2})$ as shown in item 1 above. The challenge is to sample from the marginal distribution of z. To do so, we exploit the discreteness of z. The Bayesian rule gives, for $i = 0, 1$,

$$f(z = i | v, c, a, \omega) \quad = \quad \frac{f(a | v, c, z = i) f(z = i | \omega)}{f(a | v, c, z = 0) f(z = 0 | \omega) + f(a | v, c, z = 1) f(z = 1 | \omega)}.$$

where $f(a | v, c, z = i)$ is a student density as shown by Equation (9). Then the marginal distribution of z is computed as

$$f(a | v, c, z = i) \quad = \quad (2\pi c_i)^{-\frac{1}{2}} \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{(\frac{v}{2})^{\frac{v}{2}}}{(\frac{\frac{a^2}{c_i}+v}{2})^{\frac{v+1}{2}}}, \text{for } i = 0, 1, \tag{20}$$

$$f(z = 0 | v, c, a, \omega) \quad = \quad \frac{(c_0)^{-\frac{1}{2}}(\frac{a^2}{c_0} + v)^{-\frac{v+1}{2}}\omega}{(c_0)^{-\frac{1}{2}}(\frac{a^2}{c_0} + v)^{-\frac{v+1}{2}}\omega + (c_1)^{-\frac{1}{2}}(\frac{a^2}{c_1} + v)^{-\frac{v+1}{2}}(1-\omega)}. \tag{21}$$

To sample from the joint distribution, we first sample z from (21) and then we draw $\sigma^2 | v, c, a, z \sim \mathcal{IG}(\frac{v+1}{2}, \frac{a^2/c_z+v}{2})$.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

4 Including time-varying covariance dynamics

In empirical applications, there is ample evidence of time variation in the variance of the time series. Therefore, we enrich the Sparse CP-VAR model with dynamics for the covariance matrix using an infinite hidden Markov structure based on the hierarchical Dirichlet process of Teh, Jordan, Beal, and Blei (2006). In particular, the prior of the covariance matrix is defined as follows

$$\begin{aligned}
\Sigma_t &\sim G_{\Sigma_{t-1}}, \\
G_{\Sigma_{t-1}}|G_0, \Sigma_{t-1}, \alpha_\Sigma, \kappa_\Sigma &\sim \mathcal{DP}(\alpha_\Sigma + \kappa_\Sigma, \frac{\alpha_\Sigma G_0 + \kappa_\Sigma \delta_{\Sigma_{t-1}}}{\alpha_\Sigma + \kappa_\Sigma}), \\
G_0|H, \eta_\Sigma &\sim \mathcal{DP}(\eta_\Sigma, H), \\
H|S_0, v_0 &\equiv \mathcal{IW}(v_0, S_0), \\
S_0 &\sim \mathcal{W}(\underline{s}, \underline{S}),
\end{aligned}$$

where δ_x denotes a probability measure concentrated at x, $\mathcal{DP}(a, H)$ stands for a Dirichlet process with concentration parameter a and base distribution H and $W(s, S)$ (resp. $\mathcal{IW}(s, S)$) is the Wishart (resp. inverse-Wishart) distribution with scale parameter s and positive definite matrix S. The parameter κ has been introduced by Fox, Sudderth, Jordan, and Willsky (2008) and accounts for the time series persistence. The base distribution H is inverse-Wishart with scale parameter v_0. A hierarchical structure on the base distribution is proposed by Song (2014) to improve the MCMC mixing since births of new regimes are favored. Relying on the stick-breaking representation of the Dirichlet process (Sethuraman (1994)), the model with the hierarchical structure of the covariance matrix can be written as

$$y_t = [A_0 + \sum_{i=1}^{K} \Delta A_i \mathbb{1}_{\{t>\tau_i\}}]x_t + \epsilon_t \quad \text{with} \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_{s_t^\Sigma}), \tag{22}$$

$$s_t^\Sigma | s_{t-1}^\Sigma, \{p_i^\Sigma\}_{i=1}^\infty \sim p_{s_{t-1}^\Sigma}^\Sigma, \tag{23}$$

$$p_i^\Sigma | \pi^\Sigma, \alpha_\Sigma, \kappa_\Sigma \sim \mathcal{DP}(\alpha_\Sigma + \kappa_\Sigma, \frac{\alpha_\Sigma \pi^\Sigma + \kappa_\Sigma \delta_i}{\alpha_\Sigma + \kappa_\Sigma}), \tag{24}$$

$$\pi^\Sigma | \eta_\Sigma \sim Stick(\eta_\Sigma), \tag{25}$$

$$\Sigma_{s_t^\Sigma} | H \sim H, \tag{26}$$

$$H | S_0, v_0 \equiv \mathcal{IW}(v_0, S_0), \tag{27}$$

$$S_0 \sim \mathcal{W}(\underline{s}, \underline{S}), \tag{28}$$

This article is protected by copyright. All rights reserved.

in which the state vector $s^{\Sigma}_{1:T} = \{s^{\Sigma}_1, \ldots, s^{\Sigma}_T\}$ is a discrete variable taking integer values that denotes the current regime of the covariance matrix, the sequence $p^{\Sigma}_i = \{p_{i,1}, p_{i,2}, \ldots\}$ denotes an infinite-dimensional vector of probabilities and the sequence $\pi^{\Sigma} = \{\pi_1, \pi_2, ...\}$ is driven by a stick-breaking process conveniently written as $\pi \sim Stick(\eta)$ and satisfies $\sum_{i=1}^{\infty} \pi_i = 1$ with probability one. The hierarchical structure (23)-(28) makes transparent that the model is Markovian with an unbounded number of variance regimes. The covariance switches are a priori independent of the mean changes, which turns out to provide better in-sample and out-of-sample results, see e.g. Bauwens, Carpantier, and Dufays (2015).

To sample the covariance state vector, we rely on the degree K_{Σ} weak limit approximation (Ishwaran and Zarepour (2002)) which consists of truncating the infinite dimension with an upper bound K_{Σ} number of regimes. As explored in Kurihara, Welling, and Teh (2007) and Fox, Sudderth, Jordan, and Willsky (2011), this truncation does not distort the posterior distribution as long as the number K_{Σ} is large enough. In our applications, we set it to 30. Regarding the other prior distributions of the Dirichlet process parameters, we use $\eta_{\Sigma} \sim \mathcal{G}(5, 2)$, $(\alpha_{\Sigma} + \kappa_{\Sigma}) \sim \mathcal{G}(5, 2)$ and $\frac{\kappa_{\Sigma}}{\alpha_{\Sigma}+\kappa_{\Sigma}} \sim \mathcal{B}eta(100, 1)$ where $\mathcal{G}(g_1, g_2)$ denotes the Gamma distribution with shape and scale given respectively by g_1 and g_2.

5 Simulations

We illustrate several features of the Sparse CP-VAR model on simulated data. First, we detail the ability of the method to detect the number of regimes for each model parameter. We compare with alternative CP models and Lasso based methods. Second, we study the precision of the parameter estimates over two sample sizes. Additional simulations, including robustness checks, can be found in the Internet Appendix. For the estimations in this section, the number of MCMC iterations amounts to 20,000 with a burnin size of 10,000 draws. We keep one out of two draws after the burnin period to build the posterior distribution.

5.1 Detecting the number of regimes

To illustrate the performance of the Sparse CP-VAR model in detecting the true number of regimes, we simulate 500 time series of $T = \{512, 949\}$ observations from three data generating processes (DGPs) given by a VAR(1) model specified as

$$y_t \quad = \quad A_t[1 \; y'_{t-1}]' + \epsilon_t \quad \text{with } \epsilon_t \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_t\right). \tag{29}$$

Author Manuscript

This article is protected by copyright. All rights reserved.

Table 1 displays the different values of $A_t = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{pmatrix}$ and Σ_t.

Table 1: **CP-VAR model parameters for simulation study.**

			DGP A		
Regime	1	2	3	4	#Regimes
Time	$1 \leq t < \lfloor T/4 \rfloor$	$\lfloor T/4 \rfloor \leq t < \lfloor 2T/4 \rfloor$	$\lfloor 2T/4 \rfloor \leq t < \lfloor 3T/4 \rfloor$	$\lfloor 3T/4 \rfloor \leq t$	
A_t	$\begin{pmatrix} 0.4 & 0.6 & 0 \\ 0 & 0.3 & 0.8 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.6 & 0 \\ 0 & 0.3 & 0.8 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.6 & 0 \\ 0 & 0.3 & 0.8 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.6 & 0 \\ 0 & 0.3 & 0.8 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$
Σ_t	$\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$	1
			DGP B		
Regime	1	2	3	4	#Regimes
Time	$1 \leq t < \lfloor T/4 \rfloor$	$\lfloor T/4 \rfloor \leq t < \lfloor 2T/4 \rfloor$	$\lfloor 2T/4 \rfloor \leq t < \lfloor 3T/4 \rfloor$	$\lfloor 3T/4 \rfloor \leq t$	
A_t	$\begin{pmatrix} 0.4 & 0.6 & 0 \\ 0 & 0 & 0.9 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.6 & 0 \\ 0 & 0 & 0.1 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.6 & -0.25 \\ 0 & 0 & 0.9 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.6 & -0.25 \\ 0 & 0 & 0.1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 4 \end{pmatrix}$
Σ_t	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 2.5 & 0.3 \\ 0.3 & 2.5 \end{pmatrix}$	$\begin{pmatrix} 2.5 & 0.3 \\ 0.3 & 2.5 \end{pmatrix}$	$\begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$	3
			DGP C		
Regime	1	2	3	4	#Regimes
Time	$1 \leq t < \lfloor 0.5T \rfloor$	$\lfloor 0.5T \rfloor \leq t < \lfloor 2T/3 \rfloor$	$\lfloor 2T/3 \rfloor \leq t < \lfloor 3T/4 \rfloor$	$\lfloor 3T/4 \rfloor \leq t$	
A_t	$\begin{pmatrix} 0 & 0.5 & 0.1 \\ 0 & 0.1 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0 & -0.5 & 0.1 \\ 0 & 0.1 & -0.5 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.5 & 0.1 \\ 0 & 0.1 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0 & -0.5 & 0.1 \\ 0 & 0.1 & -0.5 \end{pmatrix}$	$\begin{pmatrix} 0 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix}$
Σ_t	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	1

DGP A stands for a VAR(1) model without breaks and with a contemporaneous correlation of 0.3 between the two time series. Two parameters are set to zero in order to assess the shrinkage effect of the first regime. DGPs B and C (also used in Preuß, Puchstein, and Dette (2015)) exhibit four regimes for one parameter of the CP-VAR model. The second dependent variable of DGP B has a strong persistence when the first and the third regimes are active. Moreover, the covariance matrix has two structural breaks both in the variances and in the covariance. The means of the dependent variables of DGP C are set to zero to focus on shifts in dynamics rather than in the level.

The DGPs allow evaluating (i) if the Sparse CP-VAR process detects the four regimes present in the AR(1) coefficient of the second variable (DGP B) and the four regimes in the two AR coefficients presented by DGP C; (ii) if the second intercept as well as the first cross AR coefficient are shrunk towards zero in the posterior distribution because these coefficients are set to zero in DGP A (similar for the parameters set to zero in DGPs B and C); (iii) if regarding DGP B, where the second cross AR parameter experiences a break in the middle of the sample, only one regime is detected since the variable is insignificant before.

Table 2 summarizes the percentage of detected number of regimes, computed as the pos-

Author Manuscript

This article is protected by copyright. All rights reserved.

terior mode of the marginal probabilities of having a specific number of regimes, per model parameter. For the VAR model without breaks in DGP A, the sampler detects minimum 98.4 percent of the times the correct number of regimes in the mean, and 97.2 percent of the times the correct number of regimes in the covariance matrix. Therefore, there is little risk on estimating the Sparse CP-VAR model when there are no breaks.

In the case of multiple breaks, the results of DGP B and C indicate that detecting the four regimes works well with minimum correct detection rates of respectively 99.2 and 87.6 percent when T=512. As expected, the sampler performs less well in identifying the close to zero parameters in DGP C. In fact, for $a_{2,2}$ and $a_{1,3}$ which are both equal to 0.1, the detection rates are respectively 23.2 and 25.6 percent when T=512. Though, the detection rate increases to 50 and 48.6 percent when T=949. The breaks in the covariance matrix of DGP B are detected 88.4 percent of the times, the single regime variance of DGP C 97.4 percent of the times.[2]

We compare our results with the common breaks in mean and variance approaches of Pettenuzzo and Timmermann (2011) (PT) and Maheu and Song (2018) (HSB). PT denotes a bivariate CP-VAR model and is reviewed in Internet Appendix B including a detailed derivation and the replication of their results. HSB denotes the hierarchical structural break VAR model as explained in Internet Appendix C. Regarding the PT simulation results for T=949, DGP A is correctly detected at all times, DGP B 78.6 percent, and DGP C 43.8 percent of the times. HSB outperforms PT with higher detection rates for DGP B and C. The results are worse for T=512 for the PT model, especially DGP C with only a 4.8 percent detection rate.

To compare our results in Table 2 with a frequentist method, we implement the adaptive Lasso of Callot and Kristensen (2015). We observe that the detected number of regimes is quite unreliable when breaks occur. The large variation in the detected number of regimes does not vanish for T=949.

Table 3 shows the percentage of detected number of regimes per parameter for the Sparse CP-VAR by extending the bivariate models of Equation (29) to eight-variate VAR models. Since the number of mean and variance-covariance parameters per regime amount to 72 and 36 respectively we generate time series of length T=719 which is in the range of monthly macroeconomic time series. To save space, we only report the parameters of the first two equations, but the results are representative for the whole system. We find that also in this higher dimensional VAR model the Sparse CP-VAR model works well for the three DGPs. The detection rates of the mean parameters are systematically in the nineties, except for the

[2]Note that the exact identification of the number of regimes in the variance-covariance is not the goal and the real focus of our paper is providing a flexible method to model the variance process.

Author Manuscript

This article is protected by copyright. All rights reserved.

Table 2: **Monte Carlo regime detection results for the CP-VAR model.** Percentage of detected number of regimes per parameter based on 500 simulations from DGPs given by Equation (29) and Table 1. Bold values indicate the true number of regimes.

#Regimes	0	1	2	3	4	5	6	7	#Regimes	0	1	2	3	4	5	6
							DGP A - T=512									
Sparse CP-VAR									Adaptive Lasso							
$a_{1,1}$	0	**98.4**	1.4	0.2	0	0	0	0	$a_{1,1}$	3	**95**	2	0	0	0	0
$a_{2,1}$	**99**	0.4	0.6	0	0	0	0	0	$a_{2,1}$	**100**	0	0	0	0	0	0
$a_{1,2}$	0	**99.8**	0.2	0	0	0	0	0	$a_{1,2}$	0	**95**	5	0	0	0	0
$a_{2,2}$	0	**99.8**	0.2	0	0	0	0	0	$a_{2,2}$	0	**99**	1	0	0	0	0
$a_{1,3}$	**99.8**	0.2	0	0	0	0	0	0	$a_{1,3}$	**97**	3	0	0	0	0	0
$a_{2,3}$	0	**98.6**	1.4	0	0	0	0	0	$a_{2,3}$	0	**95**	5	0	0	0	0
Σ_t	0	**97.2**	2.8	0	0	0	0	0								
PT model	0	**100**	0	0	0	0	0	0								
HSB model	0	**100**	0	0	0	0	0	0								
							DGP A - T=949									
Sparse CP-VAR									Adaptive Lasso							
$a_{1,1}$	0	**99.2**	0.4	0.4	0	0	0	0	$a_{1,1}$	1	**99**	0	0	0	0	0
$a_{2,1}$	**100**	0	0	0	0	0	0	0	$a_{2,1}$	**100**	0	0	0	0	0	0
$a_{1,2}$	0	**99.8**	0.2	0	0	0	0	0	$a_{1,2}$	0	**97**	3	0	0	0	0
$a_{2,2}$	0	**100**	0	0	0	0	0	0	$a_{2,2}$	0	**99**	1	0	0	0	0
$a_{1,3}$	**99.8**	0.2	0	0	0	0	0	0	$a_{1,3}$	**97**	3	0	0	0	0	0
$a_{2,3}$	0	**99.2**	0.6	0.2	0	0	0	0	$a_{2,3}$	0	**97**	3	0	0	0	0
Σ_t	0	**98.2**	1.8	0	0	0	0	0								
PT model	0	**100**	0	0	0	0	0	0								
HSB model	0	**99.8**	0.2	0	0	0	0	0								
							DGP B - T=512									
Sparse CP-VAR									Adaptive Lasso							
$a_{1,1}$	0	**99.4**	0.6	0	0	0	0	0	$a_{1,1}$	25	**74**	1	0	0	0	0
$a_{2,1}$	**99.2**	0	0.8	0	0	0	0	0	$a_{2,1}$	**99**	1	0	0	0	0	0
$a_{1,2}$	0	**100**	0	0	0	0	0	0	$a_{1,2}$	0	**88**	12	0	0	0	0
$a_{2,2}$	**99.8**	0.2	0	0	0	0	0	0	$a_{2,2}$	**84**	16	0	0	0	0	0
$a_{1,3}$	0.2	**99**	0.8	0	0	0	0	0	$a_{1,3}$	3	**80**	13	4	0	0	0
$a_{2,3}$	0	0	0.2	0	**99.2**	0.6	0	0	$a_{2,3}$	0	20	47	22	**9**	1	1
Σ_t	0	0	0	**81.2**	18.8	0	0	0								
PT model	0	0	1	23.6	**72**	3	0.4	0								
HSB model	0	0	0	0.6	**83.6**	12.4	2.4	1								
							DGP B - T=949									
Sparse CP-VAR									Adaptive Lasso							
$a_{1,1}$	0	**100**	0	0	0	0	0	0	$a_{1,1}$	5	**91**	4	0	0	0	0
$a_{2,1}$	**100**	0	0	0	0	0	0	0	$a_{2,1}$	**99**	1	0	0	0	0	0
$a_{1,2}$	0	**99.6**	0.4	0	0	0	0	0	$a_{1,2}$	0	**89**	11	0	0	0	0
$a_{2,2}$	**100**	0	0	0	0	0	0	0	$a_{2,2}$	**78**	20	2	0	0	0	0
$a_{1,3}$	0	**99.8**	0.2	0	0	0	0	0	$a_{1,3}$	0	**78**	20	2	0	0	0
$a_{2,3}$	0	0	0	0	**99.4**	0.6	0	0	$a_{2,3}$	0	5	34	23	**24**	6	8
Σ_t	0	0	0	**88.4**	11	0.6	0	0								
PT model	0	0	0	0.8	**78.6**	14	5	1.6								
HSB model	0	0	0	0	**95.8**	3.8	0.4	0								
							DGP C - T=512									
Sparse CP-VAR									Adaptive Lasso							
$a_{1,1}$	**98.8**	0.4	0.8	0	0	0	0	0	$a_{1,1}$	**97**	3	0	0	0	0	0
$a_{2,1}$	**98.4**	0.2	1.4	0	0	0	0	0	$a_{2,1}$	**97**	3	0	0	0	0	0
$a_{1,2}$	0	0	10.8	0.2	**89**	0	0	0	$a_{1,2}$	22	15	31	27	**3**	1	1
$a_{2,2}$	76	**23.2**	0.8	0	0	0	0	0	$a_{2,2}$	42	**57**	1	0	0	0	0
$a_{1,3}$	74.2	**25.6**	0.2	0	0	0	0	0	$a_{1,3}$	45	**52**	3	0	0	0	0
$a_{2,3}$	0	0	11.8	0.6	**87.6**	0	0	0	$a_{2,3}$	16	21	27	28	**7**	1	0
Σ_t	0	**97**	3	0	0	0	0	0								
PT model	0	0	95.2	0	**4.8**	0	0	0								
HSB model	0	0	9	0	**72.4**	15.8	2.2	0.6								
							DGP C - T=949									
Sparse CP-VAR									Adaptive Lasso							
$a_{1,1}$	**99.6**	0.2	0.2	0	0	0	0	0	$a_{1,1}$	**94**	6	0	0	0	0	0
$a_{2,1}$	**99.8**	0.2	0	0	0	0	0	0	$a_{2,1}$	**95**	5	0	0	0	0	0
$a_{1,2}$	0	0	0.2	0.2	**99.4**	0.2	0	0	$a_{1,2}$	0	0	14	62	**12**	11	1
$a_{2,2}$	49.6	**50**	0.4	0	0	0	0	0	$a_{2,2}$	11	**82**	7	0	0	0	0
$a_{1,3}$	51.2	**48.6**	0.2	0	0	0	0	0	$a_{1,3}$	15	**79**	6	0	0	0	0
$a_{2,3}$	0	0	0	0	**99.8**	0.2	0	0	$a_{2,3}$	0	0	17	56	**16**	7	4
Σ_t	0	**97.4**	2.6	0	0	0	0	0								
PT model	0	0	53.6	0.6	**43.8**	2	0	0								
HSB model	0	0	0	0	**92.4**	6.8	0.8	0								

Author Manuscript

This article is protected by copyright. All rights reserved.

parameters $a_{2,2}$ and $a_{1,3}$ of DGP C, which are harder to distinguish being significant or not. Extending to eight-variate VAR models changes the HSB performance as the regime detection for DGP B greatly deteriorates.

Table 3: **Monte Carlo for an eight variate CP-VAR model.**
Percentage of detected number of regimes per parameter for the Sparse CP-VAR model based on 500 simulations from a 8-variate VAR generated by duplicating four times the bivariate DGPs given by Equation (29) with T=719. Bold values correspond to the true number of regimes. The number of mean parameters per regime amounts to 72. The table documents the parameters of the first two equations.

#Regime	0	1	2	3	4	5	6	≥7		0	1	2	3	4	5	6	≥7		0	1	2	3	4	5	6	≥7
	DGP A									DGP B									DGP C							
$a_{1,1}$	0	**98**	2	0	0	0	0	0	$a_{1,1}$	0	**99**	1	0	0	0	0	0	$a_{1,1}$	**99**	1	0	0	0	0	0	0
$a_{2,1}$	**100**	0	0	0	0	0	0	0	$a_{2,1}$	**99**	0	1	0	0	0	0	0	$a_{2,1}$	**100**	0	0	0	0	0	0	0
$a_{1,2}$	0	**99**	1	0	0	0	0	0	$a_{1,2}$	0	**99**	1	0	0	0	0	0	$a_{1,2}$	0	0	1	0	**98**	1	0	0
$a_{2,2}$	0	**99**	1	0	0	0	0	0	$a_{2,2}$	**100**	0	0	0	0	0	0	0	$a_{2,2}$	65	**34**	0	0	0	0	0	0
$a_{1,3}$	**100**	0	0	0	0	0	0	0	$a_{1,3}$	0	**98**	2	0	0	0	0	0	$a_{1,3}$	62	**37**	1	0	0	0	0	0
$a_{2,3}$	0	**99**	1	0	0	0	0	0	$a_{2,3}$	0	0	0	0	**99**	1	0	0	$a_{2,3}$	0	0	2	0	**97**	1	0	0
$a_{1,4}$	**100**	0	0	0	0	0	0	0	$a_{1,4}$	**100**	0	0	0	0	0	0	0	$a_{1,4}$	**99**	1	0	0	0	0	0	0
$a_{2,4}$	**100**	0	0	0	0	0	0	0	$a_{2,4}$	**100**	0	0	0	0	0	0	0	$a_{2,4}$	**100**	0	0	0	0	0	0	0
$a_{1,5}$	**99**	0	1	0	0	0	0	0	$a_{1,5}$	**100**	0	0	0	0	0	0	0	$a_{1,5}$	**100**	0	0	0	0	0	0	0
$a_{2,5}$	**100**	0	0	0	0	0	0	0	$a_{2,5}$	**100**	0	0	0	0	0	0	0	$a_{2,5}$	**100**	0	0	0	0	0	0	0
$a_{1,6}$	**100**	0	0	0	0	0	0	0	$a_{1,6}$	**100**	0	0	0	0	0	0	0	$a_{1,6}$	**100**	0	0	0	0	0	0	0
$a_{2,6}$	**99**	1	0	0	0	0	0	0	$a_{2,6}$	**100**	0	0	0	0	0	0	0	$a_{2,6}$	**100**	0	0	0	0	0	0	0
$a_{1,7}$	**99**	0	0	0	0	0	0	0	$a_{1,7}$	**100**	0	0	0	0	0	0	0	$a_{1,7}$	**100**	0	0	0	0	0	0	0
$a_{2,7}$	**100**	0	0	0	0	0	0	0	$a_{2,7}$	**99**	1	0	0	0	0	0	0	$a_{2,7}$	**100**	0	0	0	0	0	0	0
$a_{1,8}$	**100**	0	0	0	0	0	0	0	$a_{1,8}$	**99**	0	0	0	0	0	0	0	$a_{1,8}$	**100**	0	0	0	0	0	0	0
$a_{2,8}$	**100**	0	0	0	0	0	0	0	$a_{2,8}$	**100**	0	0	0	0	0	0	0	$a_{2,8}$	**100**	0	0	0	0	0	0	0
$a_{1,9}$	**100**	0	0	0	0	0	0	0	$a_{1,9}$	**99**	1	1	0	0	0	0	0	$a_{1,9}$	**100**	0	0	0	0	0	0	0
$a_{2,9}$	**99**	0	1	0	0	0	0	0	$a_{2,9}$	**99**	1	0	0	0	0	0	0	$a_{2,9}$	**100**	0	0	0	0	0	0	0
Σ_t	0	**90**	10	0	0	0	0	0	Σ_t	0	0	12	**60**	27	1	0	0	Σ_t	0	**99**	1	0	0	0	0	0
HSB Model	0	**100**	0	0	0	0	0	0		0	0	1.8	0	**16.4**	8.6	13.8	59.4		0	0	0	0.2	**95.4**	4.2	0.2	0

5.2 Accuracy of the CP-VAR parameter estimates

Table 4 documents the average RMSE defined as $\text{RMSE} = \frac{1}{500}\sum_{i=1}^{500} T^{-1}\sqrt{\sum_{t=1}^{T}(\beta_t - \hat{\beta}_t^{(i)})^2}$ in which β_t is the true parameter at time t and $\hat{\beta}_t^{(i)}$ indicates the posterior median of the Sparse CP-VAR model or the adaptive Lasso estimate at time t for simulation i. For the true parameters which are fixed to zero (underlined numbers in the table), the adaptive Lasso performs very well which average RMSEs very close or equal to zero. Though they cannot be zero by construction, the Sparse CP-VAR has small RMSEs which decrease substantially when moving to T=949. The results change substantially for the true parameters which are different from zero. Now, the Sparse CP-VAR model delivers systematically smaller RMSEs than the adaptive Lasso except for the parameters $a_{2,2}$ and $a_{1,3}$ of DGP C. The differences in RMSE are large due to the inability of the adaptive Lasso to detect a break in the parameter. This is illustrated in Figure 1 which shows the 500 adaptive Lasso estimates of the coefficient $a_{2,3}$ of DGP C. Most of the estimates do not capture the three breakpoints even though the assumptions of Callot and Kristensen (2015) hold for this DGP. However, the posterior medians of the 500 replications given by the Sparse CP-VAR highlight the three breakpoints.

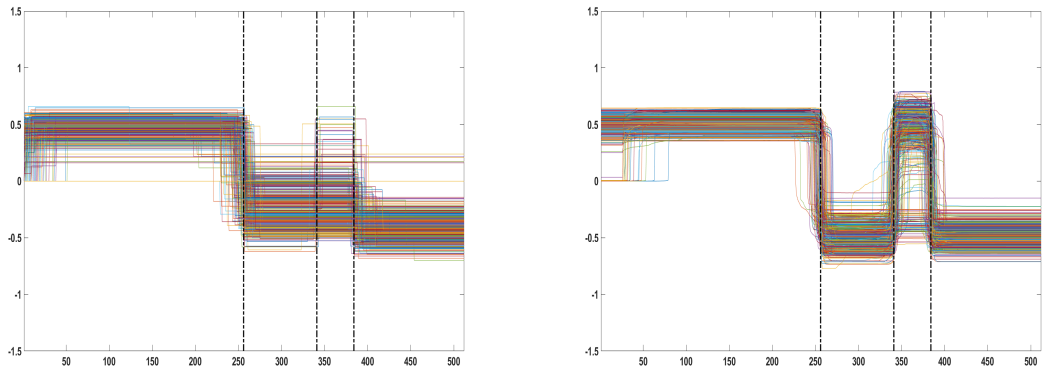
Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Table 4: **Monte Carlo parameter precision results for the CP-VAR model**
Average RMSE of the estimates compared to the true parameter based on 500 simulations from DGPs given by Equation (29) Bold values correspond to the smallest RMSE and underlined values indicate a true parameter equal to zero over the whole sample period (i.e., $\beta_t = 0$ $\forall t \in [1, T]$).

	RMSE					
	$a_{1,1}$	$a_{2,1}$	$a_{1,2}$	$a_{2,2}$	$a_{1,3}$	$a_{2,3}$
			DGP A - T=512			
Ad. Lasso	0.078	**<u>0.000</u>**	0.053	0.043	<u>0.001</u>	0.040
Sparse CP	**0.066**	<u>0.008</u>	**0.037**	**0.040**	**<u>0.001</u>**	**0.020**
			DGP A - T=949			
Ad. Lasso	0.049	**<u>0.000</u>**	0.036	0.032	<u>0.001</u>	0.027
Sparse CP	**0.039**	<u>0.002</u>	**0.024**	**0.025**	**<u>0.001</u>**	**0.013**
			DGP B - T=512			
Ad. Lasso	0.169	**<u>0.001</u>**	0.071	<u>0.013</u>	<u>0.070</u>	0.444
Sparse CP	**0.047**	<u>0.010</u>	**0.028**	**<u>0.002</u>**	**<u>0.042</u>**	**0.120**
			DGP B - T=949			
Ad. Lasso	0.077	**<u>0.000</u>**	0.042	<u>0.013</u>	<u>0.048</u>	0.348
Sparse CP	**0.032**	<u>0.002</u>	**0.020**	**<u>0.001</u>**	**<u>0.029</u>**	**0.088**
			DGP C - T=512			
Ad. Lasso	**<u>0.004</u>**	**<u>0.003</u>**	0.344	**0.077**	**0.077**	0.342
Sparse CP	<u>0.004</u>	<u>0.006</u>	**0.138**	0.090	0.089	**0.141**
			DGP C - T=949			
Ad. Lasso	<u>0.003</u>	<u>0.004</u>	0.241	**0.047**	**0.049**	0.242
Sparse CP	**<u>0.001</u>**	**<u>0.001</u>**	**0.090**	0.069	0.071	**0.089**



(a) Adaptive Lasso (b) Sparse CP-VAR

Figure 1: Adaptive Lasso estimates and posterior medians of the coefficient $a_{2,3}$ over time for 500 replications of DGP C.

This article is protected by copyright. All rights reserved.

Author Manuscript

6 Applications

We provide applications to excess return modelling and a macroeconomic system. The number of MCMC iterations for all estimations amounts to 100,000 with a burnin size of 50,000 draws, and one out of two draws are kept after the burnin period to estimate the posterior distributions. All out-of-sample forecasting results are obtained using cluster computation via High Performance Computing system (Lafayette, Sauter, Vu, and Meade, 2016) at the University of Melbourne.

6.1 Excess return modelling

We revisit the asset return predictability application of Pettenuzzo and Timmermann (2011). They estimate two bivariate first order VAR models for excess returns with either the dividend yield or the 1-month Treasury Bill. Their model assumes no autoregressive lag of the lagged excess returns in the VAR equations. The monthly data ranges between December 1926 and December 2019 and is obtained from the Center for Research in Security Prices. In the next section, we estimate our Sparse CP-VAR model for the three variables jointly which allows investigating which off diagonal VAR parameters are relevant and which are subject to breaks.

6.1.1 In-sample results

Table 5 documents the posterior probability of a specific number of regimes for each of the Sparse CP-VAR(2) parameters.[3] It turns out that most of the mean parameters are invariant over time and often insignificant in terms of posterior probability. Interestingly, while the standard single regime trivariate VAR(2) model exhibits 21 mean parameters, the Sparse CP-VAR process only detects 5 significant parameters based on the mode of the number of regimes per parameter. No mean parameters of the excess return equation are significant. While this was expected for the lagged excess return, it is surprising that both Tbill and the dividend have no predictive power on the excess return. In fact, all variations come from the covariance matrix dynamic as the Sparse CP-VAR(2) process detects between 11 to 18 regimes. The estimation of an univariate CP model of Giordani and Kohn (2008) with AR(1) structure and the lagged dividend and Tbill series, hereafter G&K's model, confirms our results as shown at the bottom of Table 5. In particular, the mean parameters are constant over time while the variance experiences even more changes compared to the Sparse CP-VAR(2) model. This result has to be contrasted with Pettenuzzo and Timmermann (2011) who found at least six regimes on their bivariate models fitted on data until December 2005.

[3]The excess return has been multiplied by 10 while the Tbill series has been multiplied by 1000.

This article is protected by copyright. All rights reserved.

We analyse deeper the nature of the detected breaks in two directions. First, shutting down the capacity for regimes in the covariance matrix does not imply breaks in the mean, although we find a few short regimes (March 1935, Jun 1980 and Nov 1982) trying to capture some of the variance regimes. Second, in the opposite direction, estimating the Sparse CP VAR model without allowing breakpoints in the mean, we find numbers of variance regimes similar to those documented by PT. In particular, the posterior mode of the regimes in the covariance matrix amounts to 8 and 11 for the bivariate model with the Dividend yield and the Tbill series, respectively.

6.1.2 Forecasting performance

We follow PT and take the out-of-sample evaluation period from October 1942 to the end of the sample (here December 2019). We estimate bivariate Sparse CP-VAR, PT-VAR, HSB-VAR and TVPSV-VAR models with excess returns and either the dividends or treasury bill (Tbill) variables. TVPSV-VAR denotes the time varying parameter stochastic volatility approach of Primiceri (2005).

We measure forecast performance using the one-step ahead excess return predictive density which we compute as

$$f\left(r_{t+1} = r_{t+1}^{o} | y_{1:t}\right) \approx \frac{1}{N} \sum_{i=1}^{N} f\left(r_{t+1} = r_{t+1}^{o} | y_{1:t}, \Psi^{i}\right).$$

where r_{t+1}^{o} is the observed value of r_{t+1}, $y_{1:t} = (r_t, x_t)'_{1:t}$, Ψ^i the set of model parameters at the ith MCMC draw, x_t either dividend yield or Tbill and N the effective number of MCMC draws. To gauge the forecasting performance over time, we use the cumulative log-predictive likelihood which at time t is defined as the sum of of log predictive densities from the beginning of the out-of-sample period until t.

Figure 2 plots the difference of cumulative log-predictive likelihoods of excess returns for the Sparse CP-VAR compared to PT-VAR, HSB-VAR and TVPSV-VAR. For the total out-of-sample period, we see that the Sparse CP-VAR provides better one-step ahead density forecasts, except for the TVPSV-VAR with the Tbill variable. While for some periods the forecasting performance is comparable with the PT-VAR model, the decade following 1960, the period 1996-2001, and the beginning of the Global Financial Crisis is generating most of the Sparse CP-VAR superior forecasting performance. The situation is different when comparing the Sparse CP-VAR with the HSB-VAR where most of the Sparse CP-VAR gains come from the period up to 1960 for the model with dividend yields, while for the model with Tbill the gains in favor of Sparse CP-VAR are more linear over the entire out-of-sample

Author Manuscript

This article is protected by copyright. All rights reserved.

Table 5: **Excess return modelling: Sparse CP-VAR(2).**
Posterior probability in percentage of having a number of regimes for each parameter of the trivariate system estimated using a Sparse CP-VAR(2) model. Posterior probabilities for the univariate model of Giordani and Kohn (2008) with excess return as dependent variable are also documented. Bold values indicate the posterior mode.

Post. probability of a specific number of regimes								
#Regime	0	1	2	3	4	5	6	7
	Excess returns Equation							
$a_{1,1}$	**99**	1	0	0	0	0	0	0
$a_{1,2}$	**99**	1	0	0	0	0	0	0
$a_{1,3}$	**99**	1	0	0	0	0	0	0
$a_{1,4}$	**99**	1	0	0	0	0	0	0
$a_{1,5}$	**99**	1	0	0	0	0	0	0
$a_{1,6}$	**99**	1	0	0	0	0	0	0
$a_{1,7}$	**99**	1	0	0	0	0	0	0
	Dividend yield Equation							
$a_{2,1}$	0	35	**63**	2	0	0	0	0
$a_{2,2}$	**100**	0	0	0	0	0	0	0
$a_{2,3}$	0	**99**	1	0	0	0	0	0
$a_{2,4}$	**74**	26	0	0	0	0	0	0
$a_{2,5}$	**100**	0	0	0	0	0	0	0
$a_{2,6}$	**99**	1	0	0	0	0	0	0
$a_{2,7}$	**81**	18	1	0	0	0	0	0
	Treasury Bill Equation							
$a_{3,1}$	**100**	0	0	0	0	0	0	0
$a_{3,2}$	**100**	0	0	0	0	0	0	0
$a_{3,3}$	**100**	0	0	0	0	0	0	0
$a_{3,4}$	0	**100**	0	0	0	0	0	0
$a_{3,5}$	**100**	0	0	0	0	0	0	0
$a_{3,6}$	**100**	0	0	0	0	0	0	0
$a_{3,7}$	0	**100**	0	0	0	0	0	0
	Covariance matrix							
#Regime	10	11	12	13	14	15	16	≥ 17
Prob.	0	4	17	**30**	25	14	6	4
	Excess returns Equation - G&K's model							
#Regime	1	2	3	4	5	6	7	8
$a_{1,1}$	**97**	3	0	0	0	0	0	0
$a_{1,2}$	**97**	3	0	0	0	0	0	0
$a_{1,3}$	**99**	1	0	0	0	0	0	0
$a_{1,4}$	**99**	1	0	0	0	0	0	0
	Variance							
#Regime	10	11	12	13	14	15	16	≥ 17
Prob.	0	1	3	7	11	**13**	13	52

period. Finally, comparing the Sparse CP-VAR with the TVPSV-VAR, we find opposite results depending on the variable that is modelled jointly with the excess-returns. The model with dividend yields is slightly in favor of the Sparse CP-VAR model with mild gains until 1987 and stronger performance afterwards. In contrast, the model with Tbill is in favor of the TVPSV-VAR from the end of the 1960s until the end of the forecasting sample.

Author Manuscript

This article is protected by copyright. All rights reserved.

We do two robustness checks regarding the Sparse CP-VAR models for which results are available upon request. First, the results reported here ignore the possibility of out-of-sample breaks. We also assess if incorporating breaks in the forecasts improves the results as suggested in Pesaran, Pettenuzzo, and Timmermann (2006) but no large differences in the respective log-predictive likelihoods are found. We also do the excess return forecasting exercise with trivariate, rather than bivariate, Sparse CP-VAR models, and we find that including both Tbill rates and dividend yields, only improves the performance marginally.

We finalise the forecasting analysis with two point forecast metrics. First, the Campbell and Thompson (2008) out-of-sample R^2 statistic for the excess return forecasts defined as

$$R^2_{\text{OS}} = \left[1 - \frac{\sum_{t=1}^{T}\left(r_t^o - \hat{r}_t\right)^2}{\sum_{t=1}^{T}\left(r_t^o - \bar{r}_t\right)^2}\right] \times 100\%$$

where r_t^0 is the observed excess return at time t, $\hat{r}_t$ is the 1-step-ahead excess return forecast estimated through period $t-1$ and $\bar{r}_t$ is the historical average excess return estimated through period $t-1$. If R^2_{OS} is positive, then the predictive regression has lower average mean-squared prediction error than the historical average return. Table 6 shows that the Sparse CP-VAR model outperforms the other models, though the TVPSV-VAR is close, and that the R^2 values are close to zero confirming the notoriously difficult task of forecasting the conditional mean of the future excess return distribution no matter the model sophistication.

Table 6: **Out of Sample R^2 for Excess Return Forecasts**

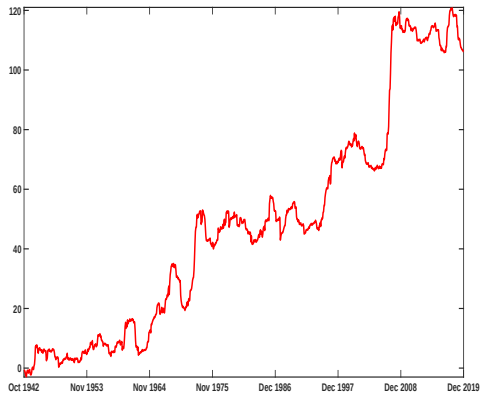
Dividend Yields Equation			
CP-VAR	PT-VAR	HSB-VAR	TVPSV-VAR
0.69%	-15.14%	-6.05%	0.31%
Tbill Equation			
CP-VAR	PT-VAR	HSB-VAR	TVPSV-VAR
-0.18%	-12.47%	-6.16%	-0.29%

As second point forecast metric, we consider the mean squared forecast error (MSFE) for the one month to one year horizon. Table 7 displays the MSFEs relative of the CP-VAR model relative to the TVPSV-VAR, HSB-VAR and PT-VAR models. The CP-VAR and TVPSV-VAR model MFESs are close for all horizons, the two other models are more strongly dominated by the CP-VAR model.

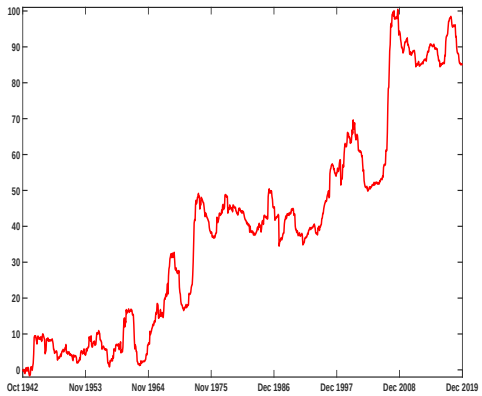
Author Manuscript

This article is protected by copyright. All rights reserved.

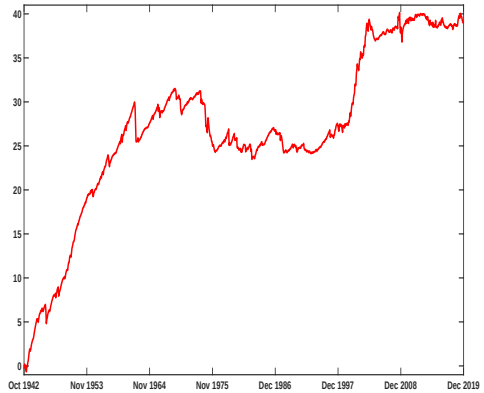
Author Manuscript



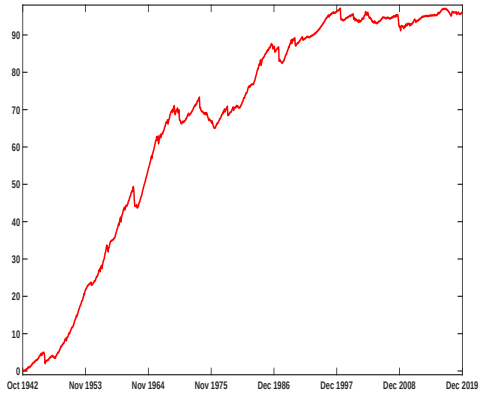
(a) Sparse CP relative to PT - Dividends



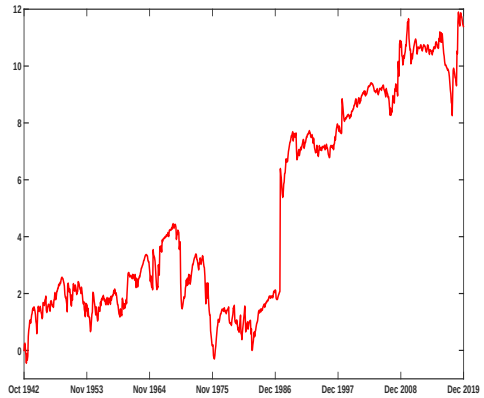
(b) Sparse CP relative to PT - Tbill



(c) Sparse CP relative to HSB-VAR - Dividends



(d) Sparse CP relative to HSB-VAR - Tbill



(e) Sparse CP relative to TVPSV-VAR - Dividends

(f) Sparse CP relative to TVPSV-VAR - Tbill

Figure 2: Cumulative marginal log-predictive likelihood for excess returns from Sparse CP-VAR relative to PT-VAR, HSB-VAR and TVPSV-VAR.

This article is protected by copyright. All rights reserved.

Table 7: **Relative MSFE for Excess Return Forecasts**
The table exposes one-step-ahead forecast results over the out-of-sample period October 1942 - December 2019. Relative MSFE is the ratio of MSFEs of the Sparse CP-VAR and competing models. Values smaller than one are in favor of the Sparse CP-VAR model.

	Dividend Yields Equation											
CP-VAR vs	h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8	h=9	h=10	h=11	h=12
TVPSV-VAR	0.996	0.994	0.995	1.001	0.995	0.997	0.998	0.995	0.992	0.992	0.993	0.991
HSB-VAR	0.936	0.928	0.945	0.944	0.947	0.939	0.956	0.959	0.964	0.965	0.970	0.976
PT-VAR	0.863	0.847	0.889	0.881	0.873	0.866	0.889	0.884	0.898	0.895	0.907	0.916

	Tbill Equation											
CP-VAR vs	h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8	h=9	h=10	h=11	h=12
TVPSV-VAR	0.999	0.999	1.002	1.004	0.998	1.003	1.001	1.000	0.998	0.997	0.997	0.996
HSB-VAR	0.944	0.932	0.959	0.946	0.967	0.959	0.962	0.952	0.955	0.963	0.965	0.978
PT-VAR	0.890	0.873	0.922	0.911	0.890	0.917	0.917	0.903	0.907	0.905	0.890	0.928

6.1.3 Portfolio allocation

Improving the predictive return density is important for portfolio allocation problems. In particular, assuming a Constant Relative Risk Aversion (CRRA) utility function, two financial assets and a buy-and-hold strategy, an investor solves the following optimization program,

$$\omega \quad = \quad \text{argmax}_{\omega} \frac{W_T^{1-\gamma}}{1-\gamma} E_{r_{T+h}|y_{1:T}}[\{(1-\omega)\exp(r_{T+h}^{\text{free}}) + \omega \exp(r_{T+h})\}^{1-\gamma}], \tag{30}$$

where W_T is the investor's wealth at time T, ω is the weight invested in the risky asset, γ denotes the risk aversion parameter and r_{T+h}^{free} and r_{T+h} are the cumulative risk-free rate and the risky excess-returns from $T+1$ to $T+h$, respectively.

Motivated by Spiegel (2008), Dangl and Halling (2012) develop a smooth time-varying parameter model for improving return predictability. Our Sparse CP-VAR model provides a competitive and complementary framework by allowing for parsimonious abrupt changes. In particular, our approach can contribute to the following four topics in the return predictability literature. First, parameter estimation uncertainty increases future risks and thus reduces the long-term allocation in the risky asset (e.g., Barberis, 2000). Our model mitigates parameter estimation uncertainty by shrinking irrelevant parameters to zero. Second, Pettenuzzo and Timmermann (2011) show that model uncertainty, computed by averaging CP models with different number of breakpoints, produces a similar effect as parameter estimation uncertainty. The Sparse CP-VAR process detects which parameters change when a break is detected and therefore mitigates the model uncertainty effect by efficiently estimating

Author Manuscript

This article is protected by copyright. All rights reserved.

the constant parameters using the full sample. Third, it is well known that the plethora of return predictors exhibit noisy signals (e.g., Goyal and Welch, 2008; Rapach, Strauss, and Zhou, 2010; Farmer, Schmidt, and Timmermann, 2019). The Sparse CP-VAR framework can incorporate many predictors without increasing the noisiness while dealing with structural breaks. This is a promising alternative to averaging over univariate predictive regressions (e.g., Rapach, Strauss, and Zhou, 2010; Li and Tsiakas, 2017). Fourth, accounting for future parameter instabilities has a major impact on asset allocation and we elaborate on this in more detail next.

We illustrate the impact of future instabilities produced by the Sparse CP-VAR model on portfolio allocations using the dividend yield. Table 5 shows that the dividend yield exhibits around two regimes in the intercept and one in the AR term coefficient. Therefore, potential future breaks will only appear in those parameters implying that the excess-return dynamic is influenced mainly through the covariance matrix. Figure 3 displays the allocation sensitivity of investing at different horizons from one to fifty months with a risk aversion parameter of $\gamma = 5$. The Sparse CP-VAR model yields similar results when future breaks are ignored or not. In particular, when the dividend yield is higher than its empirical mean, short term investments are profitable while long-term investments becomes less lucrative due to its mean reversion.

Incorporating future breaks into the Sparse CP-VAR model does not eradicate return predictability. In fact, investors keep allocating a minimum of 30% percent of their wealth in the risky asset at long-term horizons. However, as can be seen from the difference between Figure 3(c) and Figure 3(d), incorporating future instabilities has large impact in the PT model, in particular the investor keeps decreasing its allocation in the risky asset with respect to the forecast horizon. Interestingly, Figure 3(c) shows that portfolio allocation is non-linear at long-term horizons because the weight of the risky asset is larger when the dividend yield is set to its empirical mean in December 2019 than to any other values. This non-linear allocation with respect to the dividend yield is a consequence of the standard CP assumption which changes all the model parameters when a break is detected. In fact, the large uncertainty of the dividend yield parameter in the excess returns equation, noting that the size of the last regime is approximately 337 observations, implies that this parameter has a 1% probability of being negative and therefore decreases expected excess-returns.

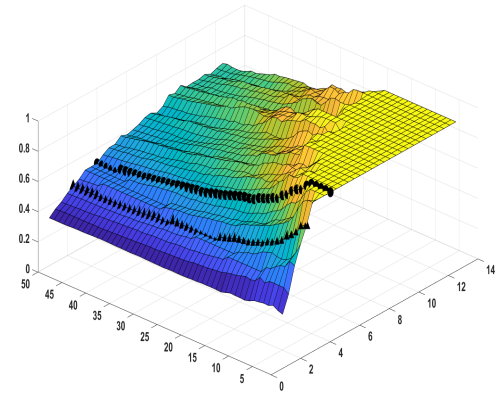
6.2 Macroeconomic application

To illustrate our approach on a larger dimensional system, we focus on a VAR application with 19 major monthly US macroeconomic variables used in Carriero, Clark, and Marcellino

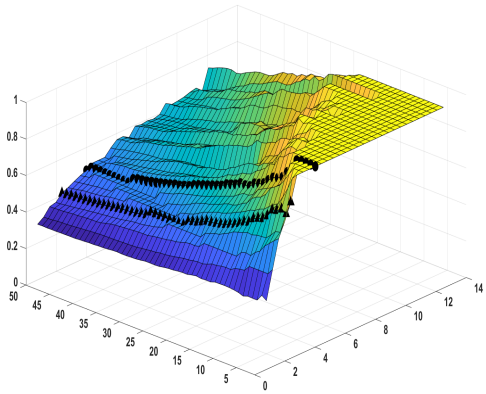
Author Manuscript

This article is protected by copyright. All rights reserved.

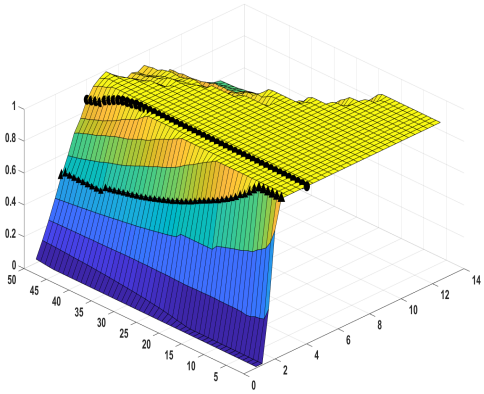
Author Manuscript



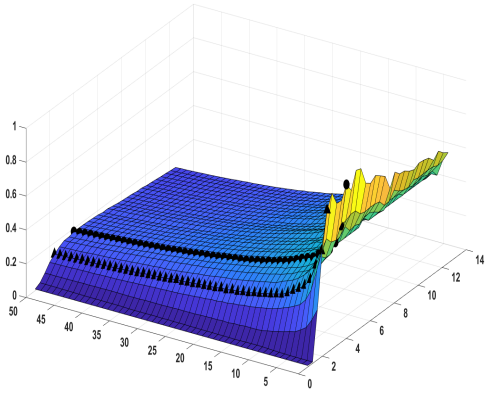
(a) Sparse CP without future breaks



(b) Sparse CP with future breaks



(c) PT without future breaks



(d) PT with future breaks

Figure 3: Excess-returns and dividend yield - Sensitivity analysis of the risky asset allocation proportion (Z-axis) with respect to the dividend yield (Y-axis) of the buy-and-hold portfolio optimized in December 2019 for 1 to 50 months ahead (X-axis) using the CRRA risk aversion parameter of $\gamma = 5$. The weights are optimized on a constraint support with bounds given by (0.01,0.99). The dividend yield ranges from 0 to 14, the optimal investments given by the dividend yield value in December 2019 and by the empirical mean value of the dividend yield are given by black triangles and circles, respectively.

This article is protected by copyright. All rights reserved.

(2019). The series are part of the dataset of McCracken and Ng (2016), downloadable from the Federal Reserve Bank of ST Louis, and span the period February 1959 to May 2020 (736 observations). See Table 8 for the variable names and abbreviations.

Table 8: **Variables names and mnemonics tor the 19 macro series.** Variable transformations between brackets.

Variable Name	**Mnemonic**
Real personal income (Δ ln)	RPI
Real PCE (Δ ln)	DPCERA3M086SBEA
Real manufacturing and trade sales (Δ ln)	CMRMTSPLx
Industrial production (Δ ln)	INDPRO
Capacity utilization in manufacturing	CUMFNS
Civilian unemployment rate	UNRATE
Total nonfarm employment (Δ ln)	PAYEMS
Hours worked: goods-producing (ln)	CES0600000007
Average hourly earnings: goods-producing (Δ ln)	CES0600000008
PPI for finished goods (Δ ln)	WPSFD49207
PPI for commodities (Δ ln)	PPICMM
PCE price index (Δ ln)	PCEPI
Federal funds rate	FEDFUNDS
Total housing starts (ln)	HOUST
S&P 500 price index (Δ ln)	SP500
U.S.-U.K. exchange rate (Δ ln)	EXUSUKx
1 yr. Treasury-FEDFUNDS spread	T1YFFM
10 yr. Treasury-FEDFUNDS spread	T10YFFM
BAA-FEDFUNDS spread	BAAFFM

6.2.1 In-sample results

Table 9 documents the relevant lagged variables when estimating a Sparse CP-VAR(1) model. The estimated maximum number of detected regimes per mean parameter amounts to three, though most of the parameters exhibit up to two regimes. The small number of regimes provides evidence of sparse time-varying system dynamics. In fact, with 71 parameters, the Sparse CP-VAR(1) model exhibits less parameters than a standard VAR(1) process with its 380 mean parameters per regime. Because of the high dimensional system, we only graphically document results for the RPI, PCEPI and FEDFUNDS variables. Figure 4 displays the time variation in the relevant mean parameters. We also report two posterior means using the large VAR model of Koop and Korobilis (2013), TVPSV-VAR hereafter, that differ by the degree of shrinkage in the AR coefficients. As explained in Koop and Korobilis (2013), a large degree of shrinkage is necessary to produce reasonable forecast performance. Consequently, they impose very tight priors for the AR coefficients in their VAR models (prior variances

Author Manuscript

This article is protected by copyright. All rights reserved.

ranging from 1e-10 to 0.1).[4] Figure 4 also provides in gray circles the posterior mean of the coefficients with variances from 0.1 to 1. The latter show similar parameter dynamics for most of the series. As emphasized by the RPI intercept, the smooth time-varying dynamics create additional noise since the intercept is insignificant according to the Sparse CP-VAR model. Regarding the other coefficients, we observe that the smooth parameter dynamic is unable to capture the abrupt changes observed in the PCEPI and FEDFUNDS series. Finally, the number Sparse CP-VAR covariance matrix regimes turns out to be between eight and nine with break dates related to well known macroeconomic slowdowns such as the great moderation and the Global Financial Crisis.

Table 9: **Sparse CP-VAR with 19 macro variables: Lagged significant variables.** For each dependent variable, the table exposes the lagged variables that exhibit at least one significant regime. Two or more significant regimes are in bold.

RPI	RPI	PAYEMS			
DPCERA3M086SBEA	DPCERA3M086SBEA				
CMRMTSPLx	CMRMTSPLx	**INDPRO**	WPSFD49207		
INDPRO	UNRATE	**PAYEMS**			
CUMFNS	CUMFNS	UNRATE	**PAYEMS**		
UNRATE	UNRATE	**PAYEMS**			
PAYEMS	CUMFNS	UNRATE	CES0600000007	HOUST	
CES0600000007	DPCERA3M086SBEA	**CUMFNS**	**CES0600000007**	**FEDFUNDS**	
CES0600000008	CES0600000008	FEDFUNDS			
WPSFD49207	UNRATE	PCEPI			
PPICMM	PPICMM				
PCEPI	WPSFD49207	**PCEPI**	FEDFUNDS		
FEDFUNDS	**INDPRO**	**FEDFUNDS**	**T1YFFM**	T10YFFM	**BAAFFM**
HOUST	**HOUST**				
SP500	SP500				
EXUSUKx	**EXUSUKx**				
T1YFFM	**T1YFFM**				
T10YFFM	**FEDFUNDS**	SP500	T1YFFM	T10YFFM	**BAAFFM**
BAAFFM	**FEDFUNDS**	SP500	T1YFFM	**T10YFFM**	BAAFFM

6.2.2 Forecasting performance

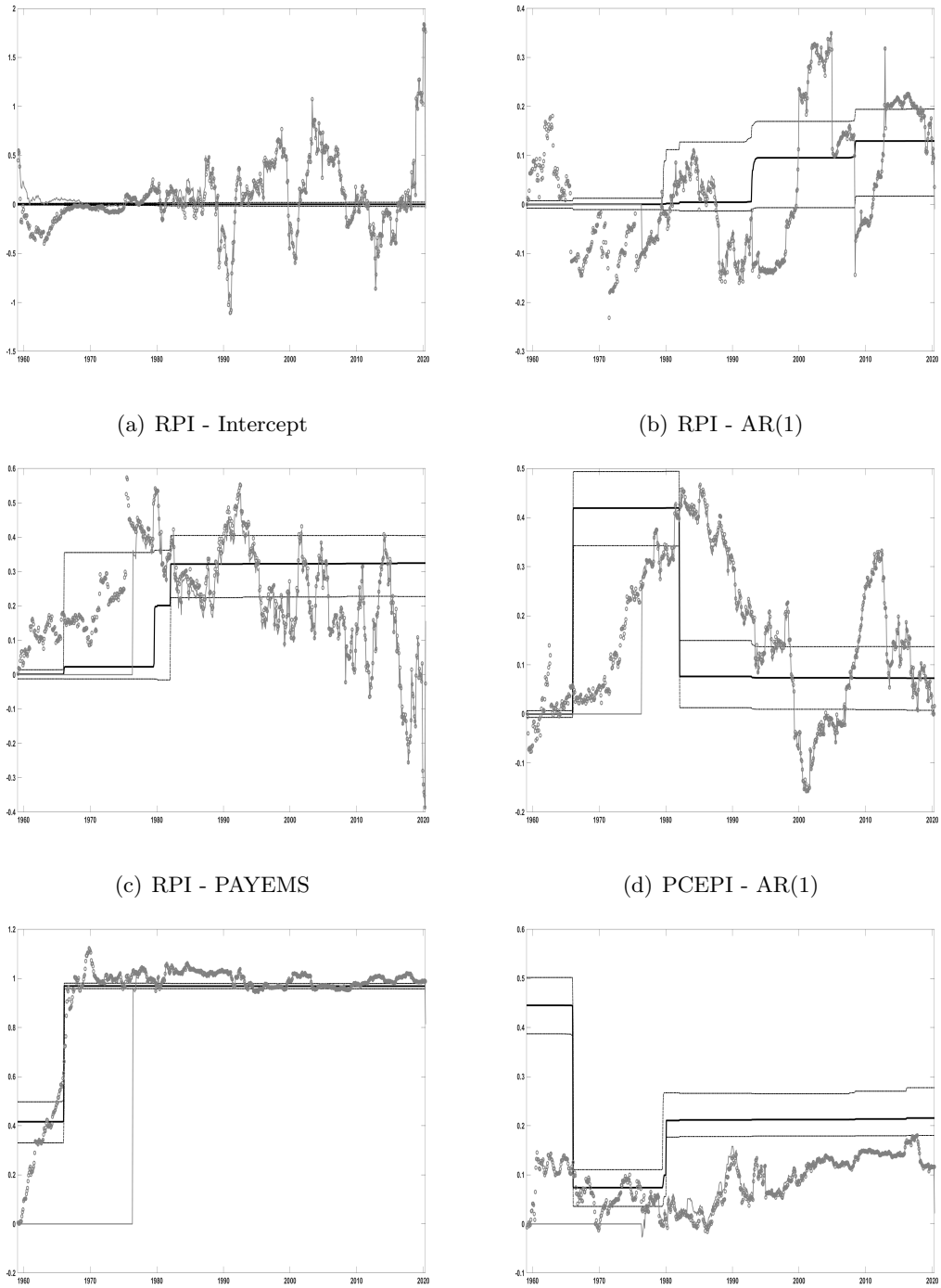
The forecast evaluation period covers the period January 2000 to May 2020. We compute one-step-ahead marginal predictive densities to compute two forecast metrics: the cumulative log predictive likelihood and the mean squared forecast error using as point forecast the median of the predictive density. We compare the Sparse CP-VAR model with the TVPSV-VAR model of Koop and Korobilis (2013) as well as the standard constant parameter VAR model, all specifications have one lag. Note that the HSB-VAR model and the TVP model of Primiceri

[4] These priors are not coefficient-dependent as in our Sparse CP VAR model. Therefore, the AR coefficients are shrunk too much and are estimated as zeros as shown in gray in Figure 4.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript



(a) RPI - Intercept

(b) RPI - AR(1)

(c) RPI - PAYEMS

(d) PCEPI - AR(1)

(e) FEDFUNDS - AR(1)

(f) FEDFUNDS - 1 yr. Treasury-FEDFUNDS

Figure 4: Posterior medians (black) with 95% credible intervals over time (dotted grey) of the Sparse CP-VAR model. TVPSV-VAR posterior means estimated from a dynamic model averaging with strong shrinkage priors for the AR coefficients (grey) and with less informative priors (grey circles).

This article is protected by copyright. All rights reserved.

(2005) are not feasible on this relatively large dimensional setting given the prohibitively large computation time required to run the MCMC samplers.

Table 10 presents the forecast results for each of the 19 macro variables. In terms of relative cumulative log predictive likelihood, we find that the Sparse CP-VAR model always has larger overall predictive likelihood than the TVPSV-VAR process, the differences being particularly large for the employment related variables UNRATE and PAYEMS. Smaller differences between the two models are found for PPI for finished goods (WPSFD49207) and the Federal Funds Rate (FEDFUNDS) variables. The performance of the CP-VAR is typically even better compared to the standard VAR, with the notable exceptions of the exchange rate EXUSUKx and housing starts HOUST, the latter outperforming both the CP-VAR and TVPSV-VAR models. In terms of point forecast performance and measured by MSFE, apart from the variables Capacity utilization in manufacturing CUMFNS and UNRATE, the dominance of the CP-VAR over the TVPSV-VAR prevails with an average relative MSFE over the 19 variables equalling 0.75. The overall performance of the standard VAR model compared to the CP-VAR averages only 0.68, one exception is the PPI for commodities (PPICMM) series with a relative MSFE of 1.08 and which according to Table 9 has no breaks and only autoregressive dynamics.

The natural question to ask is how the beginning of the COVID pandemic has altered the forecast performance of the models on this large set of macro variables. To answer this, we shorten the forecast evaluation period to March 2020, effectively excluding the April and May 2020 observations. The results, see Table 10, are striking for several of the variables under both the predictive likelihood and MSFE loss functions. Not surprisingly, the largest changes are found in the employment related variables. For example for the PAYEMS variable, the Sparse CP-VAR model dominates the TVPSV-VAR model with a log predictive likelihood difference as large as 1481.16 until May 2020. This outperformance drops to 88.87 when excluding the two last two months impacted due to the start of the COVID crisis. Over all variables, we find that the average relative MSFE performance changes from 0.68 to 0.91 and from 0.75 to 0.92 comparing the Sparse CP-VAR with the standard VAR and the TVPSV-VAR respectively when the COVID crisis observations are excluded. These results highlight the performance of the Sparse CP-VAR in its ability to quickly adapt to sudden breaks in time series.

To better understand what the non-parametric specification for the covariance matrix adds to the model, we forecast with a Sparse CP-VAR with constant covariance matrix (results available on request). It turns out that this constant covariance matrix specification is clearly not enough to generate the large superior forecast performance against the benchmark models. In fact, on average over all the series, the relative marginal log likelihood difference between

Author Manuscript

This article is protected by copyright. All rights reserved.

the constant variance Sparse CP-VAR and the TVPSV-VAR (standard VAR, respectively) amounts to -7.92 (13.28) compared to the original 41.97 (63.17) for the Sparse CP-VAR model. The results go in the same direction for the mean squared forecast error metric.

The sum up, forecasting-wise the Sparse CP-VAR model seems to strike a good balance between sparsity and parameter variation over time in this relatively large dimensional setting. We also tested the importance of taking into account out-of-sample breaks in the Sparse CP-VAR model. The differences are very small with an average relative cumulative log predictive likelihood and relative MSFE of respectively -0.206 and 0.999 in favour of the out-of-sample break models over the 19 macro variables. This is not surprising given the one month forecasting horizon where the probability of one out-of-sample break is extremely small given the small amount of breaks found in-sample.

Table 10: **Sparse CP-VAR with 19 macro variables: Forecasting performance**
The table exposes one-step-ahead forecast results over the out-of-sample period January 2000 - March 2020 and January 2000 - May 2020. Relative Marginal logPL is the difference in the cumulative log predictive likelihood between the CP-VAR and competing models (positive is in favor of the Sparse CP-VAR model). Relative MSFE is the ratio of MSFEs of the CP-VAR and competing models (smaller than one is in favor of the Sparse CP-VAR model).

	Relative Marginal logPL CP-VAR vs				Relative MSFE CP-VAR vs			
	TVPSV-VAR		Standard VAR		TVPSV-VAR		Standard VAR	
Forecast periods ends:	March 2020	May 2020	March 2020	May 2020	March 2020	May 2020	March 2020	May 2020
RPI	165.82	311.42	158.76	345.66	0.83	0.59	0.99	0.95
DPCERA3M086SBEA	211.26	250.51	97.05	385.72	1.00	0.91	1.05	0.69
CMRMTSPLx	31.87	97.86	27.26	67.40	0.97	0.74	1.04	0.67
INDPRO	39.16	92.45	31.08	170.60	1.09	0.71	1.12	0.61
CUMFNS	20.38	54.45	24.60	193.24	1.19	1.67	1.15	0.57
UNRATE	25.58	672.92	-5.51	1752.17	0.96	1.23	1.18	0.63
PAYEMS	88.87	1481.16	75.79	3074.22	0.86	0.98	0.85	0.61
CES0600000007	10.94	89.48	8.89	41.13	0.95	0.35	1.00	0.46
CES0600000008	16.08	21.85	58.52	64.99	0.86	0.83	0.69	0.24
WPSFD49207	-1.92	0.33	17.30	20.12	0.94	0.96	1.06	0.98
PPICMM	-6.17	13.34	18.67	18.59	0.97	0.25	1.07	1.08
PCEPI	31.28	43.00	47.47	45.65	0.88	0.38	0.89	0.89
FEDFUNDS	4.46	4.41	254.46	255.35	0.89	0.92	0.25	0.23
HOUST	10.18	14.31	-5.80	-6.85	0.93	0.47	1.08	0.96
S&P 500	41.77	42.81	35.21	38.57	0.84	0.73	0.93	0.75
EXUSUKx	35.82	37.08	4.35	5.30	0.80	0.71	0.97	0.95
T1YFFM	-1.06	23.92	148.01	148.93	0.75	0.36	0.49	0.49
T10YFFM	13.48	14.63	95.83	96.73	0.89	0.73	0.65	0.62
BAAFFM	59.59	62.85	108.24	109.82	0.94	0.78	0.78	0.63
Average	41.97	175.20	63.17	359.33	0.92	0.75	0.91	0.68
Standard Deviation	(13.07)	(81.34)	(15.50)	(175.44)	(0.02)	(0.08)	(0.06)	(0.05)

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

7 Conclusion

We propose the Sparse CP-VAR approach which is able to capture the parameters of the VAR model which vary over time. This opens the door for higher dimensional VAR models than the currently employed models since in our applications the Sparse CP-VAR model is typically far more parsimonious than standard VAR alternatives. To detect which VAR parameters are evolving, we use continuous spike and slab prior distributions calibrated such that the maximum à posteriori of the conditional posterior density given the breakpoints maximizes the Bayesian information criterion. While the dynamic of the VAR mean parameters follows a CP structure, the covariance matrix is driven by the sticky infinite hidden Markov process. An extensive simulation study shows the excellent performance of the Sparse CP-VAR model. Disentangling the mean and the covariance dynamics proves to be highly beneficial in the empirical illustrations to financial and macroeconomic time series. In fact, the existing models with common breaks typically find many breakpoints that are mostly due to changes in covariance matrix parameters. The latter feature makes it hard to include future instabilities into predictive densities which we illustrate in an optimal portfolio exercise. In contrast, the way we handle out-of-sample breaks in the Sparse CP-VAR model does not add artificial uncertainty in the predictive densities and generates more meaningful conclusions.

We recommend the use of the Sparse CP-VAR model for modelling systems over long periods which are potentially affected by substantial but infrequent breaks in the mean parameters and unknown form of heteroskedasticity. In our economic applications, the number of VAR parameters is drastically reduced facilitating the interpretation of the model, and improving forecasting performance. This is in contrast to the typical time-varying parameter VAR model which has new parameters at every observation according to random walk dynamics which imply mostly small parameter changes. Sparsity comes at a cost of including shrinkage priors on parameters the hyper-parameters of which have to be carefully chosen. As with any flexible Bayesian model, inference is not trivial to implement and for this reason code has been made available to estimate the Sparse CP-VAR model.

As future research, we will investigate the empirical difference between the 2MU prior distribution and the Dirichlet-Laplace prior distribution of Bhattacharya, Pati, Pillai, and Dunson (2015) and how shrinkage priors empirically compare with exact sparsity in a CP-VAR environment as in Giannone, Lenza, and Primiceri (2018) for VAR systems. Also, extending the dynamic shrinkage processes of Kowal, Matteson, and Ruppert (2019) to VAR models to investigate if economic data favor rare and abrupt breaks or smooth time-varying parameter dynamics would be interesting.

This article is protected by copyright. All rights reserved.

Author Manuscript

References

Aastveit, K. A., A. Carriero, T. Clark, and M. Massimiliano (2017): "Have Standard VARS Remained Stable Since the Crisis?," *Journal of Applied Econometrics*, 32(5), 931–951.

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

Barberis, N. (2000): "Investing for the long run when returns are predictable," *The Journal of Finance*, 55(1), 225–264.

Bauwens, L., J.-F. Carpantier, and A. Dufays (2015): "Autoregressive Moving Average Infinite Hidden Markov-Switching Models," *Journal of Business and Economic Statistics*, Forthcoming.

Bauwens, L., G. Koop, D. Koribilis, and J. Rombouts (2015): "The Contribution of Structural Break Models to Forecasting Macroecomic Series," *Journal of Applied Econometrics*, 30(4), 596–620.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015): "Dirichlet–Laplace priors for optimal shrinkage," *Journal of the American Statistical Association*, 110(512), 1479–1490.

Callot, L., and J. T. Kristensen (2015): "Vector autoregressions with parsimoniously time varying parameters and an application to monetary policy," *Tinbergen Institute Discussion Paper*, 14-145.

Campbell, J. Y., and S. B. Thompson (2008): "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?," *Review of Financial Studies*, 21(4), 1509–1531.

Carriero, A., T. E. Clark, and M. Marcellino (2019): "Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors," *Journal of Econometrics*, 212(1), 137–154.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010): "The horseshoe estimator for sparse signals," *Biometrika*, 97(2), 465–480.

Chan, J. C. (2013): "Moving average stochastic volatility models with application to inflation forecast," *Journal of Econometrics*, 176(2), 162–172.

This article is protected by copyright. All rights reserved.

Author Manuscript

Chan, N., C. Yip, and R. Zhang (2014): "Group LASSO for Structural Break Time Series," *Journal of the American Statistical Association*, 109, 590–599.

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

Chib, S., and K. H. Kang (2013): "Change-Points in Affine Arbitrage-Free Term Structure Models," *Journal of Financial Econometrics*, 11(2), 302–334.

Ciuperca, G. (2011): "A general criterion to determine the number of change-points," *Statistics & Probability Letters*, 81(8), 1267–1275.

Clark, T. E., and M. W. McCracken (2010): "Averaging forecasts from VARs with uncertain instabilities," *Journal of Applied Econometrics*, 25(1), 5–29.

Dangl, T., and M. Halling (2012): "Predictive regressions with time-varying coefficients," *Journal of Financial Economics*, 106(1), 157–181.

Dufays, A., and J. V. Rombouts (2018): "Sparse Change-point HAR Models for Realized Variance," *Econometric Reviews*, pp. 1–24.

——— (2020): "Relevant parameter changes in structural break models," *Journal of Econometrics*, 217(1), 46–78.

Farmer, L., L. Schmidt, and A. Timmermann (2019): "Pockets of predictability," *Available at SSRN 3152386*.

Fox, E., E. Sudderth, M. Jordan, and A. Willsky (2008): "A sticky HDP-HMM with Application to Speaker Diarization," in *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, ed. by W. W. Cohen, A. McCallum, and S. T. Roweis, vol. 307 of *ACM International Conference Proceeding Series*, pp. 312–319. ACM.

——— (2011): "A sticky HDP-HMM with Application to Speaker Diarization," *Annals of Applied Statistics*, 5(2A), 1020–1056.

George, E. I., U. Makov, and A. Smith (1993): "Conjugate likelihood distributions," *Scandinavian Journal of Statistics*, pp. 147–156.

Geweke, J., and Y. Jiang (2011): "Inference and prediction in a multiple-structural-break model," *Journal of Econometrics*, 163(2), 172 – 185.

This article is protected by copyright. All rights reserved.

Author Manuscript

Giannone, D., M. Lenza, and G. E. Primiceri (2018): "Economic predictions with big data: The illusion of sparsity," *Staff Reports 847, Federal Reserve Bank of New York.*

Giordani, P., and R. Kohn (2008): "Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models," *Journal of Business & Economic Statistics*, 26(1), 66–77.

Goyal, A., and I. Welch (2008): "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21(4), 1455–1508.

Griffin, J., and P. Brown (2010): "Inference with normal-gamma prior distributions in regression problems," *Bayesian Analysis*, 5, 171–188.

Hall, A. R., D. R. Osborn, and N. Sakkas (2013): "Inference on structural breaks using information criteria," *The Manchester School*, 81, 54–81.

Huber, F., G. Kastner, and M. Feldkircher (2019): "Should I stay or should I go? A latent threshold approach to large-scale mixture innovation models," *Journal of Applied Econometrics*, 34(5), 621–640.

Ishwaran, H., and J. S. Rao (2005): "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*, 33(2), 730–773.

Ishwaran, H., and M. Zarepour (2002): "Exact and Approximate Sum Representations for the Dirichlet Process," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30(2), 269–283.

Kass, R., and A. Raftery (1995): "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Koop, G., and D. Korobilis (2013): "Large time-varying parameter VARs," *Journal of Econometrics*, 177(2), 185–198.

Koop, G., and S. Potter (2007): "Estimation and Forecasting with Multiple Breaks," *Review of Economic Studies*, 74, 763–789.

Kowal, D. R., D. S. Matteson, and D. Ruppert (2019): "Dynamic shrinkage processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4), 781–804.

Kurihara, K., M. Welling, and Y. Teh (2007): "Collapsed variational Dirichlet process mixture models," in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07).*

This article is protected by copyright. All rights reserved.

Author Manuscript

Kurozumi, E., and P. Tuvaandorj (2011): "Model selection criteria in multivariate models with multiple structural changes," *Journal of Econometrics*, 164(2), 218–238.

Lafayette, L., G. Sauter, L. Vu, and B. Meade (2016): "Spartan Performance and Flexibility: An HPC-Cloud Chimera.," OpenStack Summit, Barcelona.

Li, J., and I. Tsiakas (2017): "Equity premium prediction: The role of economic and statistical constraints," *Journal of financial markets*, 36, 56–75.

Liu, C., and J. M. Maheu (2007): "Are There Structural Breaks in Realized Volatility?," *Journal of Financial Econometrics*, 6(3), 326–360.

Maheu, J. M., and S. Gordon (2008): "Learning, forecasting and structural breaks," *Journal of Applied Econometrics*, 23(5), 553–583.

Maheu, J. M., and Y. Song (2018): "An Efficient Bayesian Approach to Multiple Structural Change in Multivariate Time Series," *Journal of Applied Econometrics*, 33(2), 251–270.

Malsiner-Walli, G., and H. Wagner (2016a): "Comparing Spike and Slab Priors for Bayesian Variable Selection," *Austrian Journal of Statistics*, 40(4), 241–264.

——— (2016b): "Comparing Spike and Slab Priors for Bayesian Variable Selection," *Austrian Journal of Statistics*, 40(4), 241264.

McCracken, M. W., and S. Ng (2016): "FRED-MD: A Monthly Database for Macroeconomic Research," *Journal of Business & Economic Statistics*, 34(4), 574–589.

Park, T., and G. Casella (2008): "The Bayesian Lasso," *Journal of the American Statistical Association*, 103(482), 681–686.

Pastor, L., and R. F. Stambaugh (2001): "The Equity Premium and Structural Breaks," *Journal of Finance*, 56, 1207–1239.

Pesaran, M. H., D. Pettenuzzo, and A. Timmermann (2006): "Forecasting Time Series Subject to Multiple Structural Breaks," *Review of Economic Studies*, 73, 1057–1084.

Pesaran, M. H., and A. Timmermann (2004): "How costly is it to ignore breaks when forecasting the direction of a time series?," *International Journal of Forecasting*, 20, 411–425.

Pettenuzzo, D., and A. Timmermann (2011): "Predictability of stock returns and asset allocation under structural breaks," *Journal of Econometrics*, 164(1), 60–78.

This article is protected by copyright. All rights reserved.

Preuss, P., R. Puchstein, and H. Dette (2015): "Detection of multiple structural breaks in multivariate time series," *Journal of the American Statistical Association*, 110(510), 654–668.

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

Rapach, D., J. Strauss, and G. Zhou (2010): "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *Review of Financial Studies*, 23, 821–862.

Sethuraman, J. (1994): "A constructive definition of dirichlet priors," *Statistica Sinica*, 4, 639–650.

Song, Y. (2014): "Modelling regime switching and structural breaks with an infinite hidden Markov model," *Journal of Applied Econometrics*, 29(5), 825–842.

Spiegel, M. (2008): "Forecasting the equity premium: Where we stand today," *The Review of Financial Studies*, 21(4), 1453–1454.

Steel, M. F. (2019): "Model averaging and its use in economics," *Journal of Economic Literature*, forthcoming.

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

Stock, J. H., and M. W. Watson (1996): "Evidence on Structural Instability in Macroeconomic Time Series Relations," *Journal of Business & Economic Statistics*, 14, 11–30.

Teh, Y., M. Jordan, M. Beal, and D. M. Blei (2006): "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581.

Yao, Y.-C., et al. (1988): "Estimating the number of change-points via Schwarz'criterion," *Statistics & Probability Letters*, 6(3), 181–189.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

INTERNET APPENDIX

Sparse Change-point VAR Models

A Full MCMC algorithm for the Sparse CP-VAR Model

We detail the entire MCMC algorithm.

1. Sampling the state vector from $s_{1:T}|y_{1:T}, A_0, \Delta A_{1:K}, \Sigma_{1:T}, P, P^{\Sigma}, s^{\Sigma}_{1:T}$ by the Forward-Backward algorithm (see Chib (1998)).

2. Sampling the covariance state vector from $s^{\Sigma}_{1:T}|y_{1:T}, A_0, \Delta A_{1:K}, \Sigma_{1:T}, P, s_{1:T}$ by the Forward-Backward algorithm (see Chib (1998)).

3. Sampling the IHMM parameter $P^{\Sigma} = \{p^{\Sigma}_1, \ldots, p^{\Sigma}_{K_\Sigma}\}, \pi^{\Sigma}, \alpha^{\Sigma}, \kappa^{\Sigma}, \eta^{\Sigma}$, see Fox, Sudderth, Jordan, and Willsky (2011).

4. Sampling the covariance matrices :

$$\Sigma_i|y_{1:T}, A_1, \Delta A_{2:K}, s_{1:T}, s^{\Sigma}_{1:T} \sim \mathcal{IW}(\sum_{t=1}^{T} \mathbb{1}_{s^{\Sigma}_t=i} + v_0, \sum_{t=1}^{T} \epsilon_t \epsilon_t' \mathbb{1}_{s^{\Sigma}_t=i} + S_0) \quad \forall i \in [1, K_\Sigma].$$

5. Sampling the hyper-parameter S_0 of the IW distribution :

$$S_0|\Sigma_{1:K_\Sigma} \sim \mathcal{W}(K_\Sigma v_0 + \underline{s}, (\sum_{i=1}^{K_\Sigma} \Sigma_i^{-1} + \underline{S}^{-1})^{-1})$$

6. Sampling the mean parameters, see Section 3.1.

7. Sampling the spike and slab variables, see Section 3.2.

B Pettenuzzo and Timmermann model (PT-VAR)

We review the posterior sampler proposed in Pettenuzzo and Timmermann (2011) (PT afterwards) since it turns out that their algorithm contains some issues impacting their final results. One issue (on drawing $\psi^{-2}_{k,i}$) is related to the Markov blanket for the observations and hence directly affects the SIC which determines the posterior distribution of the number of regimes. Other issues are related to the hyper parameters.

This article is protected by copyright. All rights reserved.

Table 11 illustrates that the inference on the number of regimes is changed. The rows named "PT" copy the output from the original paper of PT. The rows named "PT replication" are results from our replication of the original algorithm, showing that we successfully replicated the original algorithm. The rows named "PT revision" are the inference from the revised algorithm. The posterior mode now indicates four rather than nine regimes for the model with dividend yields, and eight rather than six regimes for the model with T-bill.

Table 11: Posterior probabilities for the number of regimes

Dividend Yield											
#Regime	1	2	3	4	5	6	7	8	9	10	11
PT	0	0	0	0	0	0	0	0	0.99984	0.00004	0.00005
PT replication	0	0	0	0	0.0021	0	0	0	0.9979	0	0
PT revision	0	0	0	0.5196	0.4804	0	0	0	0	0	0
T-bill											
#Regime	1	2	3	4	5	6	7	8	9		
PT	0	0	0	0	0.00029	0.99971	0	0	0		
PT replication	0	0	0	0	0	0.9999	0	0	0.0001		
PT revision	0	0	0	0	0	0	0	1	0		

The next section defines the PT model and our revised posterior simulation algorithm. We use footnotes to highlight the differences with the original PT sampler.

B.1 PT Model and inference

PT define a bivariate model with structural breaks. Define $z_t = (r_t, x_t)'$, where r_t is the return and x_t is the predictor. The model is given by

$$z_t = B'_{s_t}\tilde{x}_t + u_t, \quad u_t \sim \mathcal{N}(0, \Sigma_{s_t}) \tag{31}$$

where the regressor is $\tilde{x}_t = (1, x_t)$ (i.e the model does not include lagged returns), and where $\Sigma_k = diag(\psi_k)\Lambda_k diag(\psi_k)$ with Λ the correlation matrix and ψ_k the vector of standard deviations. The transition matrix for state $s_t = 1, ..., K+1$ follows Chib (1998) with $K+1$ regimes (K breaks).

B.1.1 Prior

Define $\Theta_k = (B_k, \Sigma_k) = (B_k, \Lambda_k, \psi_k)$ and the collection of time varying parameters as $\Theta = \{\Theta_k\}_{k=1}^{K+1}$.

1. First layer related to the model

Author Manuscript

This article is protected by copyright. All rights reserved.

(a)

$$vec(B_k) \sim \mathcal{N}(b_0, V_0)$$

$B_k = \begin{pmatrix} \mu_r & \mu_x \\ \beta_r & \beta_x \end{pmatrix}$ and $vec(B_k) = (\mu_r, \beta_r, \mu_x, \beta_x)'$, where μ_r and μ_x are the intercepts in return and predictor equation, respectively. And β_r and β_x are the slopes in return and predictor equation, respectively.

- $|\beta_x| < 1$ for stationarity which seems fine in practice.
- $\frac{\mu_x}{1-\beta_x} \geq 0$ from economic motivation.

(b)

$$\psi_{k,i}^{-2} \sim \mathcal{G}(v_{0,i}/2, v_{0,i}d_{0,i}/2)$$

with $E(\psi_{k,i}^2) = \frac{d_{0,i}v_{0,i}}{v_{0,i}-2}$ and $Var(\psi_{k,i}^2) = \frac{2\left(E(\psi_{k,i}^2)\right)^2}{v_{0,i}-4}$. Hence, $sd(\psi_{k,i}^2) = E(\psi_{k,i}^2)\sqrt{\frac{2}{v_{0,i}-4}}$.

(c)

$$\lambda_{k,i,c} \sim \mathcal{N}(\mu_{\rho,i,c}, \sigma_{\rho,i,c}^2)\mathbf{1}(\lambda_{k,i,c} \in (-1,1)).$$

(d)

$$p_{kk} \sim \mathcal{B}eta(a,b)$$

2. Second layer related to the hyper-parameters.

   (a) Related to $vec(B_k) \sim \mathcal{N}(b_0, V_0)$

   - $$b_0 \sim \mathcal{N}(\underline{\mu}_\beta, \underline{\Sigma}_\beta)$$

     with set $\underline{\mu}_\beta = [0,0,0,0.9]'$ and $\underline{\Sigma}_\beta = diag(sc, sc, sc, 1)$, where $sc = 1000$.

   - $$V_0^{-1} \sim \mathcal{W}(\underline{V}_\beta^{-1}, \underline{v}_\beta),$$

     where $E(V_0) = \frac{\underline{V}_\beta}{\underline{v}_\beta - m - 1}$, where $m = 2$ is the dimension of V_0 in the application. $\underline{v}_\beta = 2m + 2$ and $\underline{V}_\beta = diag(0.5, 5, 1e-5, 0.1)$ for the dividend yield specification and $diag(0.1, 1000, 1e-5, 0.1)$ for the T-bill rate specification.

   (b) Related to volatility $\psi_{k,i}^{-2} \sim \mathcal{G}(v_{0,i}/2, v_{0,i}d_{0,i}/2)$

   - $$v_{0,i} \sim Exp(-\underline{\rho}_{0,i})$$

     The prior mean $E(v_{0,i}) = \frac{1}{\rho_{0,i}}$ witht $\rho_{0,i} = 1/\epsilon$, where ϵ is the smallest number that Matlab can interpret. Hence, $E(v_{0,i}) = \epsilon$ is very small.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

- $$d_{0,i} \sim \mathcal{G}(\underline{c}_{0,i}, \underline{d}_{0,i})$$

  with set $\underline{c}_{0,i} = 1$ and $\underline{d}_{0,i} = \epsilon$. Hence, the prior mean $E(d_{0,i}) = \frac{1}{\epsilon}$ is very large.
- Evaluating at the prior means of $v_{0,i}$ and $d_{0,i}$, the conditional mean $E(\psi_{k,i}^2)$ does not exist. But the conditional mean of the precision $E(\psi_{k,i}^{-2}) = d_{0,i}^{-1} = \epsilon$ is very small.

(c) Related to correlation $\lambda_{k,i,c} \sim \mathcal{N}(\mu_{\rho,i,c}, \sigma_{\rho,i,c}^2)\mathbf{1}(\lambda_{k,i,c} \in (-1,1))$. In PT, $\lambda_{k,i,c}$ is a scalar because it is a bivariate model, hence $i = 1$ and $c = 2$.

- $$\mu_{\rho,1,2} \sim \mathcal{N}(\underline{\mu}_{\mu,i,c}, \underline{\tau}_{i,c}^2)$$

  with $\underline{\mu}_{\mu,1,2} = 0$ and $\underline{\tau}_{1,2}^2 = 100$.
- $$\sigma_{\rho,i,c}^{-2} \sim \mathcal{G}(\underline{a}_{\rho,i,c}, \underline{b}_{\rho,i,c}),$$

  where $\underline{a}_{\rho,1,2} = 1$ and $\underline{b}_{\rho,1,2} = 0.01$. The prior mean $E(\sigma_{\rho,i,c}^{-2}) = 100$, while the prior mean $E(\sigma_{\rho,i,c}^2)$ does not exist.

(d) Related to break probability $p_{kk} \sim \mathcal{B}eta(a, b)$

- $$a \sim \mathcal{G}(\underline{a}_0, \underline{b}_0)$$

  with $\underline{a}_0 = 1$ and $\underline{b}_0 = \epsilon$.
- $$b \sim \mathcal{G}(\underline{a}_0, \underline{b}_0)$$

  with $\underline{a}_0 = 1$ and $\underline{b}_0 = \epsilon$.

B.2 MCMC

1. $S = (s_1, ..., s_T)$. Use the method proposed in Chib (1998).
2. $B = (B_1, ..., B_{K+1})$

$$vec(B) \sim \mathcal{N}(\bar{b}_k, \overline{V}_k),$$

where

$$\overline{V}_k = (X_k'(I \otimes \Sigma_k^{-1})X_k + V_0^{-1})^{-1}$$
$$\bar{b}_k = \overline{V}_k(X_k'(I \otimes \Sigma_k^{-1})Z_k + V_0^{-1}b_0).$$

This article is protected by copyright. All rights reserved.

The data Z_k and X_k are the collection of dependent and independent variables in the kth regime. Specifically, Z_k stacks $(r_t, x_t)'$ and X_k stacks $\begin{pmatrix} 1 & x_t & 0 & 0 \\ 0 & 0 & 1 & x_t \end{pmatrix}$.[5]

3. $b_0 \sim \mathcal{N}(\overline{\mu}_\beta, \overline{\Sigma}_\beta)$ with prior $b_0 \sim \mathcal{N}(\underline{\mu}_\beta, \underline{\Sigma}_\beta)$, where

$$\overline{\Sigma}_\beta = ((K+1)V_0^{-1} + \underline{\Sigma}_\beta^{-1})^{-1}$$
$$\overline{\mu}_\beta = \overline{\Sigma}_\beta \left(V_0^{-1} \sum_{k=1}^{K+1} vec(B_k) + \underline{\Sigma}_\beta^{-1} \underline{\mu}_\beta \right)$$

4. $V_0^{-1} \sim \mathcal{W}(\overline{V}_\beta^{-1}, \overline{v}_\beta)$ with prior $V_0^{-1} \sim \mathcal{W}(\underline{V}_\beta^{-1}, \underline{v}_\beta)$, where

$$\overline{v}_\beta = \underline{v}_\beta + (K+1)$$
$$\overline{V}_\beta = \underline{V}_\beta + \sum_{k=1}^{K+1} (vec(B_k) - b_0)(vec(B_k) - b_0)'$$

5. $\psi_{k,i}^{-2} \sim \mathcal{G}\left(\frac{v_{0,i}+n_i}{2}, \frac{v_{0,i}d_{0,i}+\Xi_{i,i}}{2}\right)$ with prior $\psi_{k,i}^{-2} \sim \mathcal{G}(v_{0,i}/2, v_{0,i}d_{0,i}/2)$, where

$$\Xi = (Z_k - X_k B_k)'(Z_k - X_k B_k)$$

is the 2×2 matrix related to the variance of the residual in the ith regime and n_i is the number of observations in the ith regime. Use this as proposal distribution to the Metropolis-Hastings method.[6]

---

[5]Note: $X_k'(I \otimes \Sigma_k^{-1})X_k$ is the summation of form $\begin{pmatrix} 1 & 0 \\ x_t & 0 \\ 0 & 1 \\ 0 & x_t \end{pmatrix} \Sigma_k^{-1} \begin{pmatrix} 1 & x_t & 0 & 0 \\ 0 & 0 & 1 & x_t \end{pmatrix}$ and $X_k'(I \otimes \Sigma_k^{-1})Z_k$ is the summation form of $\begin{pmatrix} 1 & 0 \\ x_t & 0 \\ 0 & 1 \\ 0 & x_t \end{pmatrix} \Sigma_k^{-1} \begin{pmatrix} r_t \\ x_t \end{pmatrix}$. PT formulate $X_k'\Sigma_k^{-1}X_k$ and $X_k'\Sigma_k^{-1}Z_k$ instead of $X_k'(I \otimes \Sigma_k^{-1})X_k$ and $X_k'(I \otimes \Sigma_k^{-1})Z_k$, where the dimension does not match.

[6]PT ignore the covariance structure ($\lambda_{k,i,c} \neq 0$) and draw $\psi_{k,i}^{-2}$ directly from the proposal distribution without adjustment from the Metropolis-Hastings step. Because $\psi_{k,i}^{-2}$ is in the Markov blanket for the observations, this directly affects the SIC which is used in PT to infer the number of regimes.

Author Manuscript

This article is protected by copyright. All rights reserved.

6. The conditional posterior kernel of $v_{0,i}$ is given by

$$p(v_{0,i} \mid \cdot) \propto \prod_{k=1}^{K+1} G(\psi_{k,i}^{-2} \mid v_{0,i}, d_{0,i}) \exp(-\underline{\rho}_{0,i} v_{0,i})$$
$$\propto \exp(-\underline{\rho}_{0,i} v_{0,i}) \prod_{k=1}^{K+1} \left[\Gamma\left(\frac{v_{0,i}}{2}\right)^{-1} \left(\frac{v_{0,i} d_{0,i}}{2}\right)^{\frac{v_{0,i}}{2}} \left(\psi_{k,i}^{-2}\right)^{\frac{v_{0,i}}{2}-1} e^{-\frac{v_{0,i} d_{0,i}}{2} \psi_{k,i}^{-2}} \right]$$

and then is fitted to a Metropolis-Hastings algorithm.[7] We work with the log kernel for the Metropolis-Hastings algorithm:

$$\log(p(v_{0,i} \mid \cdot)) = -\underline{\rho}_{0,i} v_{0,i} - (K+1) \log\left(\Gamma\left(\frac{v_{0,i}}{2}\right)\right) + \frac{(K+1) v_{0,i}}{2} \log\left(\frac{v_{0,i} d_{0,i}}{2}\right)$$
$$+ \left(\frac{v_{0,i}}{2} - 1\right) \sum_{k=1}^{K+1} \log\left(\psi_{k,i}^{-2}\right) - \frac{v_{0,i} d_{0,i}}{2} \sum_{k=1}^{K+1} \psi_{k,i}^{-2} + Const$$
$$= -\underline{\rho}_{0,i} v_{0,i} - (K+1) \log\left(\Gamma\left(\frac{v_{0,i}}{2}\right)\right) + \frac{(K+1) v_{0,i}}{2} \log\left(\frac{v_{0,i} d_{0,i}}{2}\right)$$
$$+ \frac{v_{0,i}}{2} \sum_{k=1}^{K+1} \log\left(\psi_{k,i}^{-2}\right) - \frac{v_{0,i} d_{0,i}}{2} \sum_{k=1}^{K+1} \psi_{k,i}^{-2} + Const$$

7. $d_{0,i} \sim \mathcal{G}\left(\frac{v_{0,i}}{2}(K+1) + \underline{c}_{0,i}, \frac{v_{0,i}}{2} \sum_{k=1}^{K+1} \psi_{k,i}^{-2} + \underline{d}_{0,i}\right)$. [8]

8.

$$f(\mu_{\rho,i,c} \mid \cdot)$$
$$\propto \prod_{k=1}^{K+1} \exp\left\{-\frac{(\lambda_{k,i,c} - \mu_{\rho,i,c})^2}{2\sigma_{\rho,i,c}^2}\right\} \exp\left\{-\frac{(\mu_{\rho,i,c} - \underline{\mu}_{\mu,i,c})^2}{2\underline{\tau}_{i,c}^2}\right\} \mathbf{1}(\text{restriction}) \mathbf{c}(\text{restriction})$$
$$\propto \prod_{k=1}^{K+1} \left[\exp\left\{-\frac{(\lambda_{k,i,c} - \mu_{\rho,i,c})^2}{2\sigma_{\rho,i,c}^2}\right\} \left(\Phi\left(\frac{1 - \mu_{\rho,i,c}}{\sigma_{\rho,i,c}}\right) - \Phi\left(\frac{-1 - \mu_{\rho,i,c}}{\sigma_{\rho,i,c}}\right)\right)^{-1} \right]$$
$$\times \exp\left\{-\frac{(\mu_{\rho,i,c} - \underline{\mu}_{\mu,i,c})^2}{2\underline{\tau}_{i,c}^2}\right\} \mathbf{1}(\text{restriction})$$

The normalization term embedded in the restriction includes a complicated functional

[7] PT use an adaptive rejection sampling step in the Gibbs sampling algorithm by following George, Makov, and Smith (1993). We think that the prior on $\psi_{k,i}^{-2}$ in PT is different from that in George, Makov, and Smith (1993). In addition, PT write $\exp(-\underline{\rho}_{0,i})$ instead of $\exp(-\underline{\rho}_{0,i} v_{0,i})$ as the kernel of $v_{0,i}$'s prior, which seems a typo.

[8] PT use $d_{0,i} \sim \mathcal{G}\left(v_{0,i}(K+1) + \underline{c}_{0,i}, \sum_{k=1}^{K+1} \psi_{k,i}^{-2} + \underline{d}_{0,i}\right)$ though the prior on $\psi_{k,i}^{-2}$ is $\mathcal{G}(v_{0,i}/2, v_{0,i} d_{0,i}/2)$, not a regular $\mathcal{G}(v_{0,i}/2, d_{0,i}/2)$.

Author Manuscript

This article is protected by copyright. All rights reserved.

form of $\lambda_{k,i,c}$, $\mu_{\rho,i,c}$ and $\sigma^2_{\rho,i,c}$. We apply a random-walk Metropolis-Hasting method to explore the posterior density.

9.

$$
\begin{aligned}
& f(\sigma^2_{\rho,i,c} \mid \cdot) \\
& \propto \prod_{k=1}^{K+1} (\sigma^{-2}_{\rho,i,c})^{1/2} \exp\left\{-\frac{(\lambda_{k,i,c}-\mu_{\rho,i,c})^2}{2\sigma^2_{\rho,i,c}}\right\} \mathbf{1}(\text{restriction})\mathbf{c}(\text{restriction}) \\
& \times (\sigma^{-2})^{\underline{a}_{\rho,i,c}-1} \exp\left\{-\underline{b}_{\rho,i,c}\sigma^{-2}_{\rho,i,c}\right\} \\
& \propto \prod_{k=1}^{K+1} \left[(\sigma^{-2}_{\rho,i,c})^{1/2} \exp\left\{-\frac{(\lambda_{k,i,c}-\mu_{\rho,i,c})^2}{2\sigma^2_{\rho,i,c}}\right\} \left(\Phi\left(\frac{1-\mu_{\rho,i,c}}{\sigma_{\rho,i,c}}\right) - \Phi\left(\frac{-1-\mu_{\rho,i,c}}{\sigma_{\rho,i,c}}\right)\right)^{-1}\right] \mathbf{1}(\text{restriction}) \\
& \times (\sigma^{-2})^{\underline{a}_{\rho,i,c}-1} \exp\left\{-\underline{b}_{\rho,i,c}\sigma^{-2}_{\rho,i,c}\right\}
\end{aligned}
$$

We apply a random-walk Metropolis-Hasting method to explore the posterior density.[9]

10. $\lambda_{k,i,c}$. We use a random-walk Metropolis-Hastings algorithm.

11. $p_{kk} \mid \cdot \sim \mathcal{B}eta(a+n_k, b+1)$, where n_k is the duration of regime k minus 1.

12. a and b are drawn using a Metropolis-Hastings step.

C Hierarchical Structural Break Model (HSB-VAR)

The standard model specification of a multivariate structural break (SB) VAR(q) model can be specified as

$$y_t = B_t' x_t + \varepsilon_t \text{ with } \varepsilon_t \sim \mathcal{N}(0, \Sigma_t),$$

where y_t is an $n \times 1$ vector from a multivariate time series, $x_t = \left(1, y_{t-1}', y_{t-2}', \ldots, y_{t-p}'\right)' \in \mathbb{R}^{k\times 1}$ as on VAR model where $k = (n \times p) + 1$. The coefficients are summarised in a $k \times n$ matrix B_t, every $\varepsilon_t \in \mathbb{R}^{n\times 1}$ is independently distributed as a matrix normal ($\mathcal{MN}$) distribution with zero mean and Σ_t as the covariance matrix. Upon any time t, the HSB-VAR(p) as in Maheu and Song (2018) assumes that a structural break occurs with probability $\pi \sim \mathcal{B}eta(\pi_a, \pi_b)$, then the duration measure d_t would have been forced to 1 if change does happen or equals to $d_{t-1}+1$ otherwise. Meanwhile, the new parameters B_t and Σ_t can be obtained from an inverse

[9] PT ignore the $(\sigma^{-2}_{\rho,i,c})^{1/2}$ term in the Gaussian kernel of $\lambda_{k,i,c}$.

Author Manuscript

This article is protected by copyright. All rights reserved.

Wishart matrix normal distribution $\mathcal{MN}-\mathcal{IW}\left(\underline{B},\underline{\Omega},\underline{S},\underline{v}\right)$, where $\left(\underline{B},\underline{\Omega},\underline{S},\underline{v}\right)$ are inferred from a hierarchical prior. Therefore, HSB-VAR(p) is summarized as

$$
\begin{aligned}
d_t &= \begin{cases} d_{t-1}+1, & \text{w.p. } 1-\pi \\ 1, & \text{w.p. } \pi \end{cases} \\
(B_t,\Sigma_t) &\begin{cases} \sim \mathcal{MN}-\mathcal{IW}\left(\underline{B},\underline{\Omega},\underline{S},\underline{v}\right), & \text{if } d_t=1 \\ = \left(B_{t-1},\Sigma_{t-1}\right), & \text{o.w.} \end{cases} \\
\pi &\sim \mathcal{B}eta\left(\pi_a,\pi_b\right) \\
y_t|B_t,\Sigma_t,x_t &\sim \mathcal{N}\left(B_t^{'}x_t,\Sigma_t\right),
\end{aligned}
$$

with the hierarchical priors are assumed as

$$
\begin{aligned}
&\underline{\Omega} \sim \mathcal{IW}\left(\Omega_0,\varpi_0\right), \\
&\underline{B}|\Omega_0 \sim \mathcal{MN}\left(B_0,\Lambda_0,\Omega_0\right) \\
&\underline{S} \sim \mathcal{W}\left(S_0,\varrho_0\right) \\
&\underline{v} \sim \mathcal{G}\left(v_a,v_b\right)\mathbf{1}\left(\underline{v}\geq n+1\right)
\end{aligned}
$$

For consistency, we use the same hyper-parameter setting method for priors and same MCMC algorithms for sampling from Maheu and Song (2018).

D Why using the 2MU distribution ?

Dufays and Rombouts (2020) and Dufays and Rombouts (2018) argue in favor of the 2MU prior distribution based on theoretical arguments. In the CP-VAR framework, the 2MU distribution exhibits two major advantages. First, it mimics the hard-thresholding penalty function which simplifies the choice of the penalty parameter P as many theoretical results exist for such penalties. Second, the 2MU distribution also exhibits a penalty function that can be directly related to Bayesian model selection techniques.

1. *Hard-thresholding penalty function.* Assuming that all the mean parameters of the CP-VAR(p) model are driven by a 2MU prior distribution with hyper-parameters $(a^{2\text{MU}},b^{2\text{MU}},P)$ and setting $b^{2\text{MU}}=\infty$ for easing the notations, the log posterior density is proportional

This article is protected by copyright. All rights reserved.

Author Manuscript

to

$$\log f(A_0, \Delta A_1, ..., \Delta A_K | y_{1:T}, \tau_{1:K}, \Sigma_{1:T}) \quad \propto \quad \log(f(y_{1:T}|A_0, \Delta A_1, ..., \Delta A_K, \tau_{1:K}, \Sigma_{1:T})) + P \sum_{i=0}^{K} \sum_{j=1}^{nk} \mathbb{1}_{\{|a_i^{(j)}| > a^{2\mathrm{MU}}\}},$$

where $f(y_{1:T}|A_0, \Delta A_1, ..., \Delta A_K, \tau_{1:K}, \Sigma_{1:T})$ denotes the likelihood function. The penalty function for one parameter is given by $P\mathbb{1}_{\{|a_i^{(j)}| > a^{2\mathrm{MU}}\}}$ for $i \in [0, K]$ and $j \in [1, nk]$ which stands for any penalty function of information criteria that are linear with respect to the number of parameters. Examples of such criteria are the Aikake information criterion (AIC), the Bayesian information criterion (BIC) and the modified Bayesian information criterion (see Kurozumi and Tuvaandorj (2011)). For instance, maximizing this density function with respect to the mean parameters and setting $P = -2$ is equivalent to maximizing the AIC. Nevertheless, the AIC has been shown to overestimate the number of regimes in CP frameworks because its penalty does not depend on the sample size, see, e.g., Yao et al. (1988) and Hall, Osborn, and Sakkas (2013). The BIC instead has been shown to be a consistent criterion in many CP setups, see Yao et al. (1988) and Ciuperca (2011).

2. *Bayesian model selection technique.* In Bayesian statistics, the marginal likelihood is the standard tool for model selection. Comparing two models M_q and M_z with uniform prior distributions on the models, the Bayes factor gives the posterior probability of each model of generating the data. In fact, when $\ln \frac{f(y_{1:T}|M_q)}{f(y_{1:T}|M_z)} = \ln \frac{\pi}{(1-\pi)}$, the posterior probability of Model q (resp. Model z) is equal to π (resp. $1-\pi$). Assuming that Model q exhibits one break more than Model z (so two additional parameters, say τ_q and $a_i^{(j)}$, i.e. the break location and the magnitude of the change), we want to select Model q, if and only if $\ln \frac{f(y_{1:T}|M_q)}{f(y_{1:T}|M_z)} > \ln \frac{\pi}{1-\pi}$ with $\pi \geq 0.5$. Using the BIC approximation of the marginal likelihood leads to $\ln \frac{f(y_{1:T}|M_q)}{f(y_{1:T}|M_z)} \approx \ln \frac{f(y_{1:T}|\Theta^q, M_q)}{f(y_{1:T}|\Theta^z, M_z)} - \log T \geq \ln \frac{\pi}{1-\pi}$ in which Θ^p denotes the maximum likelihood estimates (MLE) of Model p with $p = q, z$. Given that the ratio of the likelihood function $\ln \frac{f(y_{1:T}|\Theta^q, M_q)}{f(y_{1:T}|\Theta^z, M_z)} = \ln \frac{f(y_{1:T}|\Theta^q, M_q)}{f(y_{1:T}|\Theta^z, a_i^{(j)}=0, \hat{\tau}_q, M_q)}$ where $\hat{\tau}_q$ stands for the MLE of τ_q of Model q, we have that if the MLE increases the likelihood function by at least $\ln \frac{\pi}{1-\pi} + \ln T$, we want that the posterior density verifies

$$f(\Theta^q | y_{1:T}, M_q) \quad > \quad f(\Theta^z, a_i^{(j)} = 0, \hat{\tau}_q | y_{1:T}, M_q). \tag{32}$$

Condition (32) implies that a move from $\{\Theta^z, a_i^{(j)} = 0, \hat{\tau}_q\}$ to Θ^q is accepted with probability one in a standard Metropolis algorithm. Imposing condition (32) leads to a

Author Manuscript

This article is protected by copyright. All rights reserved.

way of fixing the penalty value from a Bayesian perspective:

$$\begin{aligned}
\frac{f(\Theta^q|y_{1:T}, M_q)}{f(\Theta^z, a_i^{(j)}=0, \hat{\tau}_q|y_{1:T}, M_q)} &> 1, \\
\ln \frac{f(y_{1:T}|\Theta^q, M_q)}{f(y_{1:T}|\Theta^z, a_i^{(j)}=0, \hat{\tau}_q, M_q)} &> -\ln \frac{f(\Theta^q)}{f(\Theta^z, a_i^{(j)}=0)}, \\
&> -P, \\
\ln \frac{\pi}{1-\pi} + \ln T &> -P.
\end{aligned}$$

Both Bayesian and frequentist frameworks suggest using a penalty parameter related to the BIC, i.e. $P_\pi = -(\ln \frac{\pi}{1-\pi} + \ln T)$ with $\pi \in [0.5, 1)$. We use the penalty value $P_{0.95}$ which is equivalent to a very strong evidence on the scale of Kass and Raftery (1995).
The choice of the penalty value does not provide any guidance for the hyper-parameters $a^{2\mathrm{MU}}$ and $b^{2\mathrm{MU}}$. The latter parameters are related to the prior probability of the model size. In particular, let us denote by $\tilde{q}$ the prior probability of no break in parameter $a_i^{(j)}$ $\forall i \in [0, K]$ and $\forall j \in [1, nk]$ (i.e. $\tilde{q} = P[|a_i^{(j)}| < \frac{a^{2\mathrm{MU}}}{2}|\omega] = \frac{a^{2\mathrm{MU}}}{b^{2\mathrm{MU}}e^P + a^{2\mathrm{MU}}(1-e^P)}$). A model exhibiting g structural breaks, say M_g has a prior probability given by

$$P[M = M_g] \quad = \quad (\tilde{q})^{(K+1)nk-g}(1-\tilde{q})^g.$$

The prior probability of not detecting breaks geometrically decays with respect to the number of potential breaks (i.e. K), the dimension of the dependent variables (i.e. n) and the number of explanatory variables (i.e. k). To avoid a strong prior in favor of detecting breaks, we suggest setting $\tilde{q} = \frac{K}{K+1}$ which leads to an expected prior model size equivalent to the VAR model without breakpoints. In practice, we need different values of $a^{2\mathrm{MU}}$ and $b^{2\mathrm{MU}}$ per model parameter because the VAR parameters exhibit various scales. Assuming that a parameter scale does not evolve over time, we have to choose $a_j^{2\mathrm{MU}}$ and $b_j^{2\mathrm{MU}}$ for $j \in [1, nk]$. To set these parameters, we use the penalty $P_{0.95}$. In particular, since $a_j^{2\mathrm{MU}}$ for $j \in [1, nk]$ must be small, we estimate a standard VAR by OLS and set $a_j^{2\mathrm{MU}} = \frac{1}{3}\hat{\sigma}_{a_0^{(j)}}$ in which $\hat{\sigma}_{a_0^{(j)}}$ denotes the standard error of the OLS estimate $a_0^{(j)}$. These constraints lead to a value of $b_j^{2\mathrm{MU}}$ given by

$$\begin{aligned}
\tilde{q} &= \frac{a_j^{2\mathrm{MU}}}{b_j^{2\mathrm{MU}}e^{P_{0.95}} + a_j^{2\mathrm{MU}}(1-e^{P_{0.95}})}, \\
b_j^{2\mathrm{MU}} &= a_j^{2\mathrm{MU}}\frac{(1-\tilde{q}(1-e^{P_{0.95}}))}{\tilde{q}e^{P_{0.95}}}, \quad \text{with } e^{P_{0.95}} = (19T)^{-1} \text{ and } \tilde{q} = \frac{K}{K+1}, \\
&= a_j^{2\mathrm{MU}}(\frac{19T}{K}+1) \quad (= O(\frac{\sqrt{T}}{\ln T})).
\end{aligned}$$

This article is protected by copyright. All rights reserved.

Author Manuscript

As shown in Dufays and Rombouts (2020), to keep the penalty interpretation, we should use one weight $\omega_i^{(j)}$ per model parameter $a_i^{(j)}$. Also, to take into account the penalty uncertainty while sticking to our BIC interpretation, we set the following prior distribution on $\omega_i^{(j)} \sim \mathcal{B}eta(l_\omega, u_\omega)$ with $\{l_\omega, u_\omega\} = \arg\min(P[\omega_i^{(j)} \in [\omega^{2\text{MU}}(P_{0.75}), \omega^{2\text{MU}}(P_{0.99})]] - 0.999)^2$ with $\omega^{2\text{MU}}(P_\pi) = \frac{a_j^{2\text{MU}}(1-e^{P_\pi})}{b_j^{2\text{MU}}e^{P_\pi}+a_j^{2\text{MU}}(1-e^{P_{0.95}})}$. Making $\omega_i^{(j)}$ a random variable has the appealing property of facilitating the move from the spike to the slab component and vice versa, and therefore improving the mixing of the sampler. This structure is similar to the S&S prior of Ishwaran and Rao (2005).

To sum up, we have the following hyper-parameters of the 2MU distribution for each parameter $a_i^{(j)} \sim 2\text{MU}(a_j^{2\text{MU}}, b_j^{2\text{MU}}, P)$:

$$\begin{aligned}
a_j^{2\text{MU}} &= \frac{1}{3}\hat{\sigma}_{a_0^{(j)}}, \\
b_j^{2\text{MU}} &= \frac{a_j^{2\text{MU}}(19TK+1)}{K}, \\
P &= P_{0.95}, \\
\omega^{2\text{MU}}(P_{0.95}) &= \frac{a_j^{2\text{MU}}(1-e^{P_{0.95}})}{b_j^{2\text{MU}}e^{P_{0.95}} + a_j^{2\text{MU}}(1-e^{P_{0.95}})}.
\end{aligned}$$

The hyper-parameters of the beta distribution of the continuous spike and slab given in Equation (8) are computed as, for any $i \in [0, K]$ and $j \in [1, nk]$,

$$\begin{aligned}
\omega_i^{(j)} &\sim \mathcal{B}eta(a_{\omega_i^{(j)}}, b_{\omega_i^{(j)}}), \\
\{a_{\omega_i^{(j)}}, b_{\omega_i^{(j)}}\} \equiv \{a_\omega, b_\omega\} &= \arg\min(P[\omega_i^{(j)} \in [\omega^{2\text{MU}}(P_{0.75}), \omega^{2\text{MU}}(P_{0.99})]] - 0.999)^2.
\end{aligned}$$

E Additional results on the simulation study

E.1 Robustness checks

We perform three robustness checks by varying the size of the parameter breaks, by comparing our approach to the standard S&S prior, and by estimating a misspecified VAR model.

Table 12 summarises the results of a small Monte Carlo study which varies the size of the breaks, while these were previously held constant. Without loss of generality, this is done using a univariate AR(1) model where we vary either the intercept or slope parameter after a break in the middle of the sample of 300 observations. The baseline parameters come from

Author Manuscript

This article is protected by copyright. All rights reserved.

the Core inflation series used in Section 6.2 and the break sizes range from small too moderate (i.e. 0.15, 0.3 and 0.45). The results for the Sparse CP-AR model show that for both breaks in the intercept and slope, the smallest breaks are difficult to detect. The Sparse CP approach selects oftentimes a single regime model as best implying detection rates of 5 and 20 percent for the constant and AR parameters respectively. However, these results improve strongly for the breaks with size 0.45, with respective detection rates of 31 and 97 percent. As a comparison to the Sparse CP approach, we do the same exercice with the adaptive Lasso. The detection rates for the intercept turn out to be almost zero for all break sizes. For the slope parameter, the results strongly improve with detection rates between 48, 71 and 66 percent for the three break sizes respectively.

Table 12: **Monte Carlo for a Sparse CP-AR model - Sensitivity to the break size.** Percentage of detected number of regimes per parameter based on 100 simulations for a univariate model given by $y_t = \alpha_0 + \alpha_1 y_{t-1} + \sigma\epsilon_t$, with $\alpha_0 = 0.73$, $\alpha_1 = 0.77$, $\sigma = 1.51$ and $\epsilon_t \sim N(0,1)$. The parameters are the OLS estimates from the Core inflation series used in Section 6.2. A break is assumed at $t = 150$ and the sample size amounts to $T = 300$. Break size is subtracted from the corresponding parameter. For instance, a break size of $\Delta\alpha_0 = 0.15$ means that $\alpha_0 = 0.73$ in the first regime and $\alpha_0 = 0.58$ in the second (which is approximately equivalent to a relative size of $\frac{\Delta\alpha_0}{\hat{\sigma}_{\alpha_0}} \approx 1$). Bold values correspond to the true number of regimes exhibited by the parameter. Regarding the adaptive Lasso approach, a break is detected if an estimate is different from zero.

#Regime	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
	Break size of $\Delta\alpha_0 = 0.15$					Break size of $\Delta\alpha_0 = 0.3$					Break size of $\Delta\alpha_0 = 0.45$				
α_0	0	95	**5**	0	0	6	87	**7**	0	0	11	58	**31**	0	0
α_1	0	99	1	0	0	0	97	3	0	0	0	98	2	0	0
σ^2	0	99	1	0	0	0	99	1	0	0	0	100	0	0	0
							Adaptive Lasso								
α_0	15	83	**2**	0	0	25	75	**0**	0	0	46	54	**0**	0	0
α_1	0	77	21	2	0	0	73	25	1	1	0	72	21	6	1
	Break size of $\Delta\alpha_1 = 0.15$					Break size of $\Delta\alpha_1 = 0.3$					Break size of $\Delta\alpha_1 = 0.45$				
α_0	0	86	14	0	0	0	95	5	0	0	0	93	7	0	0
α_1	0	79	**20**	1	0	0	13	**86**	1	0	0	2	**97**	1	0
σ^2	0	98	2	0	0	0	100	0	0	0	0	100	0	0	0
							Adaptive Lasso								
α_0	6	94	0	0	0	3	95	1	1	0	0	93	7	0	0
α_1	0	37	**48**	11	3	0	6	**71**	18	3	0	0	**66**	20	10

A natural question to ask is how important is the calibration of the S&S prior to the 2MU prior in the context of structural breaks in VAR models. To answer this, we apply a standard

Author Manuscript

This article is protected by copyright. All rights reserved.

S&S prior defined, $\forall i \in [0, K] \;\; \forall j \in [1, nk]$, as follows

$$
\begin{aligned}
a_i^{(j)} &\sim \mathcal{N}(0, r(z_i^{(j)})\phi_i^{(j)}), \\
r(z_i^{(j)}) &= \begin{cases} 10^{-4}, & \text{if } z_i^{(j)} = 0, \\ 1, & \text{otherwise,} \end{cases} \\
\phi_i^{(j)} &\sim \mathcal{IG}(\eta = 5, Q = 4) \\
z_i^{(j)} &\sim \text{Bernoulli}(1 - \omega) \\
\omega &\sim \mathcal{B}eta(\alpha_\omega = K, \beta_\omega = 1).
\end{aligned}
\tag{33}
$$

The choice of the hyper-parameters in Equation (33) are identical to Malsiner-Walli and Wagner (2016b) except for $\{\alpha_\omega, \beta_\omega\}$. As explained in Steel (2019), a Beta prior on $z_i^{(j)}$, $\forall i \in [0, K] \;\; \forall j \in [1, nk]$, leads to a prior distribution on Model p, with $z = \{z_i^{(j)}\}_{i=0,j=1}^{K,nk}$ and $p \in [0, nk(K+1)]$ denoting the number of parameters that break, given by

$$
\begin{aligned}
f(M_p|z) &= \frac{\Gamma(\alpha_\omega + \beta_\omega)}{\Gamma(\alpha_\omega)\Gamma(\beta_\omega)} \int \omega^{(nk(K+1)-p)+\alpha_\omega}(1-\omega)^{p+\beta_\omega} f(\omega) d\omega, \\
&= \frac{\Gamma(\alpha_\omega + \beta_\omega)}{\Gamma(\alpha_\omega)\Gamma(\beta_\omega)} \frac{\Gamma(\alpha_\omega + (nk(K+1) - p))\Gamma(\beta_\omega + p)}{\Gamma(\alpha_\omega + \beta_\omega + nk(K+1))}
\end{aligned}
\tag{34}
$$

Steel (2019), adapted to our framework, advocates for $\alpha_\omega = \frac{(nk(K+1)-m^*)}{m^*}$ and $\beta_\omega = 1$ in which m^* denotes the prior choice on the mean model size, that is the number of breaks. In our setting, we choose $m^* = K + 1$ which implies that we expect at least one VAR parameter per regime to break.[10]

Table 13, in the left panel, documents the break point detection rates for the three DGPs and sample size T=512 when we do not calibrate the spike and slab prior to the 2MU distribution. We observe that the standard S&S over-estimates the number of regimes and does not perform as well as the Sparse CP-VAR (see right panel of Table 13 which are taken from 2 above). For example in DGP B regarding parameter $a_{2,3}$, the standard S&S detects five regimes, rather than four, 43 percent of the times compared to 3 percent for the Sparse CP-VAR.

Table 14 documents the percentage of detected number of regimes for the three DGPs when a CP-VAR(2) model is estimated rather than the correct CP-VAR(1). That is, we add to the model in (29) a term $A_2 y_{t-2}$ with all parameters in $A_2 = \begin{pmatrix} a_{1,4} & a_{1,5} \\ a_{2,4} & a_{2,5} \end{pmatrix}$ equal to zero. We find similar results as for the CP-VAR(1) model. In addition, the four parameters in A_2

[10] We also tested $m^* = nk$ which implies that we expect all the VAR parameters to be significant at least in one period over the sample and the results were quantitatively the same.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Table 13: **Monte Carlo results for the Sparse CP-VAR model with standard spike and slab priors.**
Percentage of detected number of regimes per parameter based on 100 simulations from DGPs given by Equation (29) for T=512 and by the parameters from Table 1. Bold values correspond to the true number of regimes.

#Regime	0	1	2	3	4	5	6		0	1	2	3	4	5	6	7
	Sparse CP-VAR with standard S&S								Sparse CP-VAR with 2MU							
DGP A																
$a_{1,1}$	0	**2**	38	59	1	0	0	$a_{1,1}$	0	**99**	1	0	0	0	0	0
$a_{2,1}$	**48**	1	23	24	4	0	0	$a_{2,1}$	**99**	0	1	0	0	0	0	0
$a_{1,2}$	0	**9**	84	7	0	0	0	$a_{1,2}$	0	**100**	0	0	0	0	0	0
$a_{2,2}$	0	**14**	67	19	0	0	0	$a_{2,2}$	0	**100**	0	0	0	0	0	0
$a_{1,3}$	**94**	3	2	1	0	0	0	$a_{1,3}$	**99**	1	0	0	0	0	0	0
$a_{2,3}$	0	**38**	51	11	0	0	0	$a_{2,3}$	0	**98**	2	0	0	0	0	0
Σ_t	0	**99**	1	0	0	0	0	Σ_t	0	**97**	3	0	0	0	0	0
DGP B																
$a_{1,1}$	0	**50**	37	12	1	0	0	$a_{1,1}$	0	**99**	1	0	0	0	0	0
$a_{2,1}$	**70**	14	13	3	0	0	0	$a_{2,1}$	**99**	0	1	0	0	0	0	0
$a_{1,2}$	0	**75**	22	3	0	0	0	$a_{1,2}$	0	**99**	1	0	0	0	0	0
$a_{2,2}$	**81**	16	3	0	0	0	0	$a_{2,2}$	**100**	0	0	0	0	0	0	0
$a_{1,3}$	0	**68**	32	0	0	0	0	$a_{1,3}$	0	**99**	1	0	0	0	0	0
$a_{2,3}$	0	0	0	0	**56**	43	1	$a_{2,3}$	0	0	0	0	**97**	3	0	0
Σ_t	0	0	0	**82**	18	0	0	Σ_t	0	0	0	**82**	18	0	0	0
DGP C																
$a_{1,1}$	**61**	24	12	3	0	0	0	$a_{1,1}$	**98**	2	0	0	0	0	0	0
$a_{2,1}$	**71**	14	11	4	0	0	0	$a_{2,1}$	**99**	1	0	0	0	0	0	0
$a_{1,2}$	0	0	0	0	**59**	41	0	$a_{1,2}$	0	0	9	0	**90**	1	0	0
$a_{2,2}$	28	**61**	10	1	0	0	0	$a_{2,2}$	81	**19**	0	0	0	0	0	0
$a_{1,3}$	19	**53**	24	4	0	0	0	$a_{1,3}$	74	**23**	2	1	0	0	0	0
$a_{2,3}$	0	0	0	0	**57**	43	0	$a_{2,3}$	0	0	10	2	**88**	0	0	0
Σ_t	0	**100**	0	0	0	0	0	Σ_t	0	**99**	1	0	0	0	0	0

This article is protected by copyright. All rights reserved.

are insignificant 99 percent on average. As expected, increasing the sample size to $T = 949$ yields even better detection rates. The latter results are available on request.

Table 14: **Monte Carlo of the Sparse CP-VAR(2) model.**
Percentage of detected number of regimes per parameter based on 100 simulations from DGPs given by Equation (29) with T=512 and parameters from Table 1. Bold values correspond to the true number of regimes exhibited by the parameter. The Sparse CP-VAR model is compared with the PT model (PT) and HSB model (HSB).

#Regime	0	1	2	3	4	5	6	7		0	1	2	3	4	5	6	7		0	1	2	3	4	5	6	7
	DGP A									DGP B									DGP C							
$a_{1,1}$	0	**98**	1	0	0	0	1	0	$a_{1,1}$	0	**99**	1	0	0	0	0	0	$a_{1,1}$	**98**	0	2	0	0	0	0	0
$a_{2,1}$	**99**	1	0	0	0	0	0	0	$a_{2,1}$	**100**	0	0	0	0	0	0	0	$a_{2,1}$	**100**	0	0	0	0	0	0	0
$a_{1,2}$	0	**99**	1	0	0	0	0	0	$a_{1,2}$	0	**100**	0	0	0	0	0	0	$a_{1,2}$	0	0	13	2	**85**	0	0	0
$a_{2,2}$	0	**100**	0	0	0	0	0	0	$a_{2,2}$	**100**	0	0	0	0	0	0	0	$a_{2,2}$	70	**28**	2	0	0	0	0	0
$a_{1,3}$	**100**	0	0	0	0	0	0	0	$a_{1,3}$	0	**100**	0	0	0	0	0	0	$a_{1,3}$	74	**26**	0	0	0	0	0	0
$a_{2,3}$	0	**100**	0	0	0	0	0	0	$a_{2,3}$	0	0	0	0	**99**	0	1	0	$a_{2,3}$	0	0	14	2	**84**	0	0	0
$a_{1,4}$	**100**	0	0	0	0	0	0	0	$a_{1,4}$	**100**	0	0	0	0	0	0	0	$a_{1,4}$	**99**	1	0	0	0	0	0	0
$a_{2,4}$	**99**	1	0	0	0	0	0	0	$a_{2,4}$	**99**	0	1	0	0	0	0	0	$a_{2,4}$	**100**	0	0	0	0	0	0	0
$a_{1,5}$	**100**	0	0	0	0	0	0	0	$a_{1,5}$	**98**	2	0	0	0	0	0	0	$a_{1,5}$	**99**	1	0	0	0	0	0	0
$a_{2,5}$	**100**	0	0	0	0	0	0	0	$a_{2,5}$	**99**	0	1	0	0	0	0	0	$a_{2,5}$	**99**	0	0	0	0	0	1	0
Σ_t	0	**96**	4	0	0	0	0	0	Σ_t	0	0	0	**84**	16	0	0	0	Σ_t	0	**98**	2	0	0	0	0	0
PT	0	**100**	0	0	0	0	0	0	PT	0	0	3	54	**35**	8	0	0	PT	0	0	**100**	0	0	0	0	0
HSB	0	**100**	0	0	0	0	0	0	HSB	0	0	0	1	**84**	13	2	0	HSB	0	0	10	0	**62**	24	2	2

E.2 Break detection accuracy

In the Sparse CP-VAR model, breaks are shared among the mean parameters. The break location is therefore more precisely estimated if parameters experience a break at the same time. By doing so, the Sparse CP process extends standard CP models since we can detect which parameters are time-varying while preserving the feature of standard CP models consisting in detecting more accurately the break locations if many parameters are subject to the same break. This should be contrasted with TVP models which use mixture distributions in the state equations for detecting breakpoints (see, e.g., Giordani and Kohn (2008)) as the prior distribution of the break dynamic is independent for each model parameter. To illustrate this, we simulate ten independent series ($T = 512$) from the following univariate DGP,

$$y_t = \phi_i y_{t-1} + \sigma_i \epsilon_t, \text{ for } t \in (\tau_{i-1}, \tau_i], \tag{35}$$

$$\tau = \{\tau_0, \tau_1, \tau_2, \tau_3, \tau_4\} = \{0, 128, 256, 384, 512\}, \tag{36}$$

$$\phi = \{\phi_1, \phi_2, \phi_3, \phi_4\} = \{0.9, 0.1, 0.9, 0.1\}, \tag{37}$$

$$\sigma^2 = \{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2\} = \{1, 2.5, 2.5, 0.5\}, \tag{38}$$

with $\epsilon_t \sim \mathcal{N}(0, 1)$. We estimate the TVP-AR(1) model of Giordani and Kohn (2008) (GK, hereafter) on the first simulated series and we compare the results with the Sparse CP-VAR(1)

Author Manuscript

This article is protected by copyright. All rights reserved.

model when we consider the first two series and then the ten generated series together. The univariate model is at its advantage in the comparison since it is correctly specified. The Sparse CP-VAR(1) model, however, allows off diagonal VAR parameters in the mean parameters as well as a non-diagonal covariance matrix. The number of parameters in one regime for the Sparse CP-VAR model with ten series amounts to 165 compared to only 3 for the univariate model.

Figure 5 shows the probability over time of a break for the GK model and the Sparse CP-VAR model.[11] The breaks are correctly detected by the two models. For each true location, GK results deliver a probability at least of 10% detecting a breakpoint.[12] Interestingly, the Sparse CP-VAR model with two dimensions exhibits higher probabilities of detecting the breaks for the second and the last breaks and equivalent probabilities as the GK model for the first break. Although the breaks are shared among the parameters, the off diagonal parameters of the Sparse CP-VAR model have to be estimated and this deteriorates the break estimation. When estimated over 10 series, the Sparse CP-VAR model detects the breakpoints with high probabilities compared to GK and the Sparse CP-VAR model with two dimensions. By stacking time series with the same breaks into a multivariate system, we can therefore improve the detection accuracy of the breaks when they occur simultaneously. This is analogous to Bai, Lumsdaine, and Stock (1998) who show in the frequentist framework that tighter asymptotic confidence intervals of the break estimators can be obtained in multivariate framework when breaks are shared among the series.

F Huber, Kastner, and Feldkircher model (HKF)

The HKF algorithm of Huber, Kastner, and Feldkircher (2019) does not converge to the true posterior distribution and depends on the starting values. One could argue that the algorithm still represents a good approximation from a practical perspective. To illustrate that this is not the case, we generate one time series from DGP B with $T = 512$ and we use the HKF software program in R for simulating ten replications from the posterior distribution. We use the default hyper-parameters and a number of MCMC iterations amounting to 10,000 with a burnin size of 5,000. For this exercise, we only focus on the parameter $a_{2,3}$ which has three breaks over the sample. For every replication, on each of the 5 000 posterior draws, we save as breakpoints the locations of the first-difference parameters being larger than 0.05 in absolute

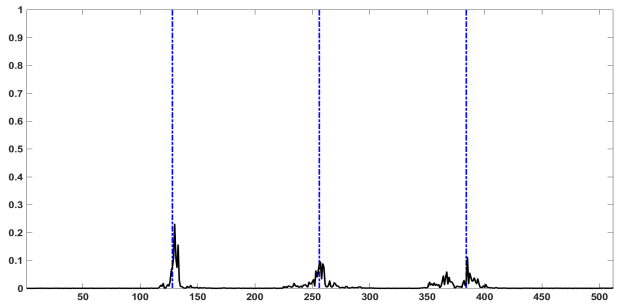
[11] These probabilities are empirical probabilities computed from the posterior draws indicating in which component lies the parameter at time t. For GK results, we compute the frequency that the AR parameter experiences a break at time t. Regarding Sparse CP-VAR models, we compute the empirical probability that the parameters of the first simulated series lies in the wide component at time t.

[12] If the breakpoints are uniformly distributed, only ten locations per break can exhibit a 10% probability.

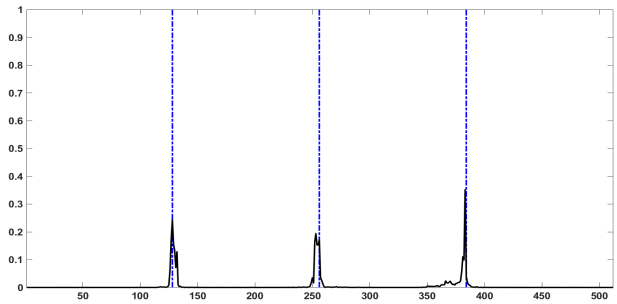
Author Manuscript

This article is protected by copyright. All rights reserved.

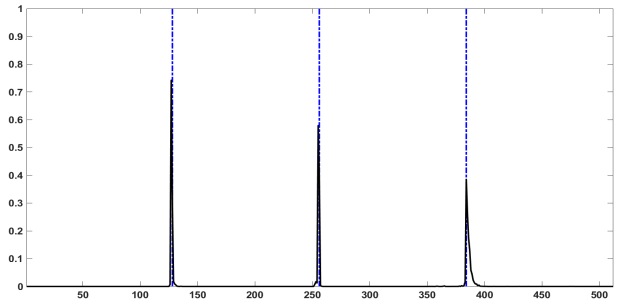
Author Manuscript



(a) Univariate model of GK



(b) Sparse CP model with 2 dependent variables



(c) Sparse CP model with 10 dependent variables

Figure 5: Posterior break probability over time in black. Dotted lines indicate the true location of the break.

This article is protected by copyright. All rights reserved.

value.

Table 15 documents the breakpoints for the ten replications along with the probability of observing these breakpoint locations in the posterior draws. We observe the same issues as pointed out in Appendix D. In particular, the break locations are different at every replication and when a break is detected, it often becomes an absorbing state given that many probabilities are exactly equal to one.

Table 15: **Mixing properties of the HKF algorithm.**
One series from DGP B: Posterior locations of the breaks with their probabilities on 10 replications using the HKF algorithm.

True break locations: $\tau = \{128,\ 256,\ 384\}$									
Sim 1	134	256	313	398	Sim 6	311	0	0	0
Prob	1	1	0.44	1	Prob	0.26	0	0	0
Sim 2	129	244	319	383	Sim 7	127	257	270	380
Prob	1	1	0.33	1	Prob	1	1	0.36	1
Sim 3	119	256	315	393	Sim 8	123	236	281	380
Prob	1	1	0.44	1	Prob	1	1	0.26	1
Sim 4	125	261	305	380	Sim 9	116	258	319	380
Prob	1	1	0.42	1	Prob	1	1	0.47	1
Sim 5	326	0	0	0	Sim 10	320	0	0	0
Prob	0.20	0	0	0	Prob	0.24	0	0	0

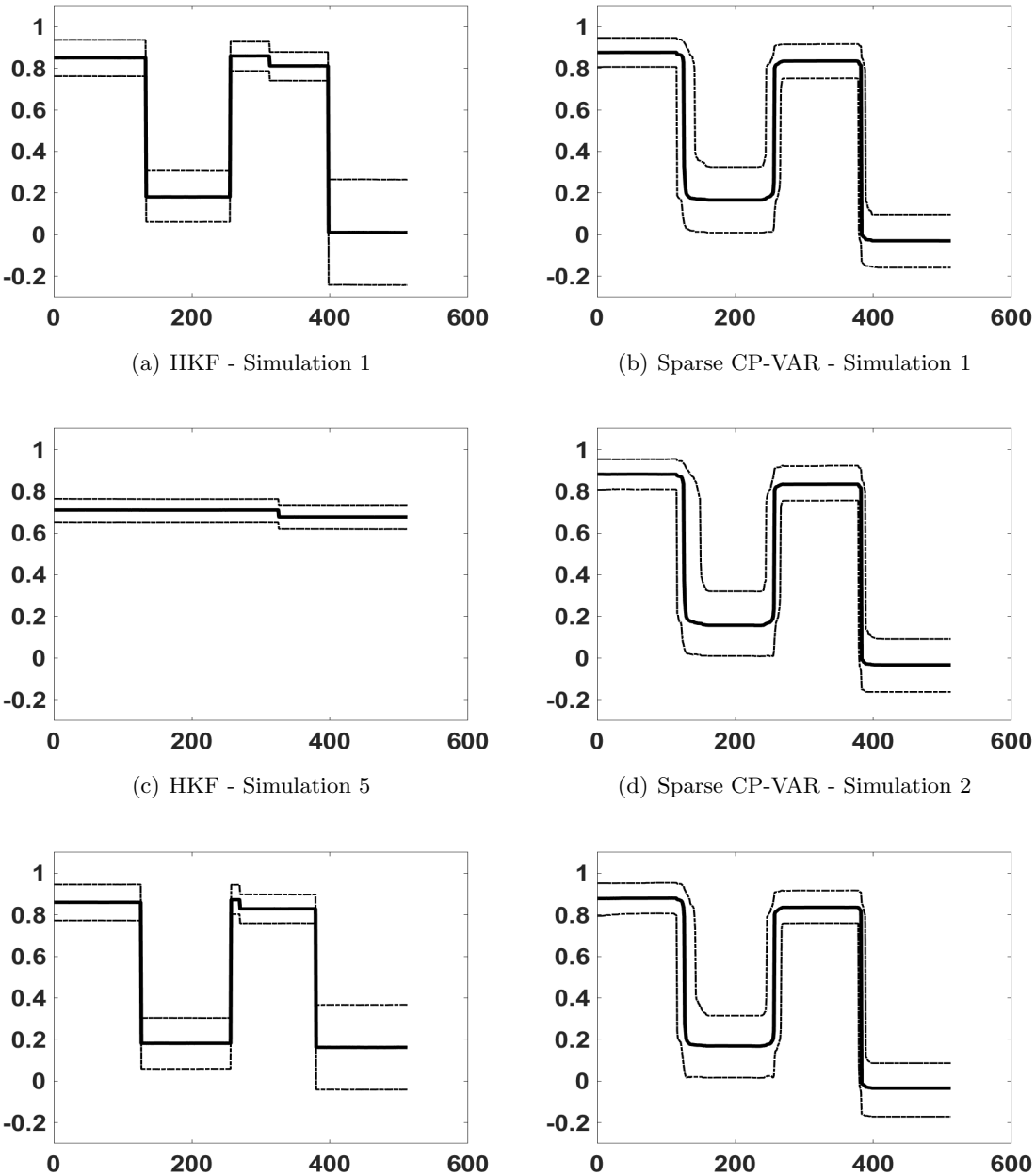
Figure 6 shows the $a_{2,3}$ parameter posterior distributions over time on three representative replications out of the ten. We observe that the break locations depend on the starting values and that they are sharply (but wrongly) estimated. This is evidence of the presence of absorbing states in the MCMC algorithm. In addition, we see that the forecast performance of the fifth simulation will be very different from simulations one and seven although all these results have been obtained using the same data and the same model specification. Figure 6 also documents three replications of the Sparse-VAR model on the same data. All the simulations have converged and no marked differences between the graphics can be pointed out.

To elaborate more theoretically on this MCMC convergence issue, we consider sampling from a mixture of two Normal distributions using the HKF threshold approach. Let us assume that the posterior distribution of a parameter θ is a mixture of two Normal components, i.e. $f(\theta) = \omega f_1(\theta) + (1-\omega) f_2(\theta)$ in which $f_i(\theta) \sim \mathcal{N}(\mu_i, \sigma_i^2)$ (see below for an example). To sample from such a distribution, HKF introduce a threshold variable $d \in [l_d, u_d]$ with $0 < l_d < u_d$ that identifies from which mixture component the parameter θ should be sampled. In particular, given a starting value $\theta^{(0)}$ and for $j = 1, ..., N$, the posterior sampling algorithm operates as:

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript



(a) HKF - Simulation 1

(b) Sparse CP-VAR - Simulation 1

(c) HKF - Simulation 5

(d) Sparse CP-VAR - Simulation 2

(e) HKF - Simulation 7

(f) Sparse CP-VAR - Simulation 3

Figure 6: Posterior distribution of $a_{2,3}$ over time for three representative replications. The solid black line stands for the median while the dotted lines document the 5% and 95% quantiles respectively.

- Sample the threshold variable $d^{(j)} \sim d|\theta^{(j-1)}$ in which

$$f(d|\theta) \quad \propto \quad (f_1(\theta)\mathbb{1}_{\{|\theta|\leq d\}} + f_2(\theta)\mathbb{1}_{\{|\theta|>d\}})f(d).$$

This article is protected by copyright. All rights reserved.

- Set $z^{(j)} = 1 + \mathbb{1}_{\{|\theta^{(j-1)}|>d^{(j)}\}}$.
- Sample $\theta^{(j)} \sim \mathcal{N}(\mu_{z^{(j)}}, \sigma^2_{z^{(j)}})$.

Proposition 1 shows that the invariant distribution of the HKF algorithm depends on the prior distribution of the threshold d.

Proposition 1. *Consider sampling from a two-component distribution $f(\theta) = \omega f_1(\theta) + (1-\omega) f_2(\theta)$ using the HKF algorithm in which $f_1(\theta)$ and $f_2(\theta)$ are proper densities and $\omega \in (0,1)$. Let us denote $q_i = \int_{-l_d}^{l_d} f_i(\theta) d\theta$ for $i = 1, 2$. The stationary distribution of the HKF algorithm, $\pi(\theta) = \bar{\omega} f_1(\theta) + (1-\bar{\omega}) f_2(\theta)$, depends on the prior distribution of the threshold:*

1. *When the prior distribution of d is a Dirac function $f(d) = \mathbb{1}_{\{d=l_d\}}$, the weight of the stationary distribution is given by $\bar{\omega} = \frac{q_2}{(1+q_2-q_1)}$.*
2. *When the prior distribution of d is Uniform with $d = \{l_d, u_d\}$, the weight of the stationary distribution is equal to $\bar{\omega} = \frac{q_2+b_1(1-q_2)}{(1+q_2-q_1-a_1(1-q_1)+b_1(1-q_2))}$ in which*

$$\begin{aligned} a_1 &= 2\int_{l_d}^{u_d} \frac{f_1(\theta)^2}{f_1(\theta)+f_2(\theta)} d\theta, \\ b_1 &= 2\int_{l_d}^{u_d} \frac{f_1(\theta) f_2(\theta)}{f_1(\theta)+f_2(\theta)} d\theta. \end{aligned}$$

3. *When the prior distribution of d is Uniform over the support $[l_d, u_d]$ and if $q_1 = 1$, the weight of the stationary distribution is equal to $\bar{\omega} = 1$.*

Proof. See Appendix F.1 □

Remark 1. In HKF, the prior distribution of the threshold is a discrete uniform distribution over the support $[l_d, u_d]$ with 150 different bins. While the stationary distribution in such a case cannot be derived analytically, the conclusions of Proposition 1 apply. The weight of the stationary distribution using their prior distribution will be a complex function of the density functions $f_1(\theta)$ and $f_2(\theta)$.

Remark 2. The third case of Proposition 1 is relevant since in the HKF framework, $f_1(\theta) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ with σ_1 that is close to zero (and so way smaller than l_d). Using Chernoff's inequality, we can show that $q_1 \geq 1 - \exp\{-\frac{(l_d+\mu_1)^2}{2\sigma_1^2}\} - \exp\{-\frac{(l_d-\mu_1)^2}{2\sigma_1^2}\}$. In particular, when

Author Manuscript

This article is protected by copyright. All rights reserved.

$\mu_1 = 0$ (i.e. the shrinkage is strong), setting $k = \frac{l_d^2}{\sigma_1^2}$, we have that

$$q_1 \quad \geq \quad 1 - 2\exp\{-\frac{k}{2}\},$$

which is very close to one for any $k > 9$ and therefore $f_2(\theta)$ vanishes from the posterior.

Remark 3. The mixing of the HKF algorithm can be slow as it depends on the probabilities $1 - q_1$ and $P_2[|\theta| \leq u_d]$. If one of these probabilities is close to zero, one of the component can virtually never be visited. For instance, if $q_1 = 0.999$ and the starting value $\theta^{(0)}$ of the MCMC algorithm is such that $|\theta^{(0)}| \leq l_d$, the wide component of the mixture will be at most visited on average 1 over 1000 MCMC draws.

For an empirical illustration of Proposition 1, we focus on a model with one intercept:

$$\begin{aligned} y_t &= \mu + \gamma\epsilon_t, \text{ with } \epsilon_t \sim \mathcal{N}(0,1), \\ \mu &\sim \omega\mathcal{N}(0,\sigma_1^2) + (1-\omega)\mathcal{N}(0,\sigma_2^2). \end{aligned}$$

The posterior distribution of $\mu|y_{1:T}$ is given by

$$\begin{aligned} \mu|y_{1:T} &\sim \bar{\omega}\mathcal{N}(\mu_1^*, \sigma_1^{2,*}) + (1-\bar{\omega})\mathcal{N}(\mu_2^*, \sigma_2^{2,*}), \\ \sigma_i^{2,*} &= \frac{\gamma^2\sigma_i^2}{T\sigma_i^2 + \gamma^2}, \quad i = 1,2, \\ \mu_i^* &= \frac{T\sigma_i^2}{T\sigma_i^2 + \gamma^2}\bar{y}_{1:T}, \quad i = 1,2, \\ \bar{\omega} &= \frac{\omega}{1 + (\frac{T\sigma_1^2+\gamma^2}{T\sigma_2^2+\gamma^2})^{\frac{1}{2}}\exp\{\frac{T\bar{y}_{1:T}}{2\gamma^2}(\mu_2^* - \mu_1^*)\}}. \end{aligned}$$

Since the posterior distribution is a mixture of two normal distributions, the conclusion of Proposition 1 holds. The HKF model to estimate the unconditional expectation is as follows

$$\begin{aligned} y_t &= \mu + \gamma\epsilon_t, \text{ with } \epsilon_t \sim \mathcal{N}(0,1), \\ \mu &\sim \omega\mathcal{N}(0,\sigma_1^2) + (1-\omega)\mathcal{N}(0,\sigma_2^2), \\ \sigma_1^2 &= \frac{\gamma^2}{100T}, \\ \sigma_2^2 &\sim \text{IG}(a_\sigma, b_\sigma) \text{ i.e. IG}(3, 33.3), \\ d|\sigma_2^2 &\sim U[l_d, u_d] \text{ i.e. } U[0.1\sigma_2, 1.5\sigma_2]. \end{aligned}$$

Note that $k \equiv \frac{l_d^2}{\sigma_1^2} \sim \mathcal{IG}(a_\sigma, \frac{0.01 b_\sigma}{\sigma_1^2})$. Using the values of their paper with $\gamma = 1$ leads to

Author Manuscript

This article is protected by copyright. All rights reserved.

$k \sim \mathcal{IG}(3, 33.3T)$. For instance, we have $E(k) = \frac{33.3T}{2}$ which is very large even for small time series. Figure 7 shows the density of $\mu|y_{1:T}$ when $T = 10$ or $T = 100$ (with $\mu = 2$ and $\mu = 1$) using the threshold approach (with $\omega = 0.2$) with respect to the true posterior density (also assuming $\omega = 0.2$, $\sigma_1^2 = \frac{1}{100T}$ and $\sigma_2^2 \sim \mathcal{IG}(3, 33.3)$). Even with 100 000 MCMC iterations, we observe that the threshold approach is highly sensitive to initial values. In addition, when the parameter is different but close to zero, the absorbing state phenomenon highlighted in Proposition 1 (see also Remark 2) can occur even when the sample size is at least as large as $T = 100$.

To show the implications of Proposition 1 in a framework close to the Sparse CP-VAR model, we now focus on a multivariate illustrative model specified as,

$$y_t = \mu_1 + \mu_2 \mathbb{1}_{\{t \geq \tau\}} + \epsilon_t, \tag{39}$$

$$\epsilon_t \sim \mathcal{N}(0, \gamma^2), \tag{40}$$

$$\mu_i \sim \omega \mathcal{N}(0, \sigma_1^2) + (1 - \omega)\mathcal{N}(0, \sigma_2^2). \tag{41}$$

Model given by Equations (39) to (41) is a particular case of the HKF model. In fact, they introduce a break at each period of time while the current model only allows for one break to get analytical results. To ease the notation, we redefine the model as $y_t = x_t'\theta + \epsilon_t$ in which $x_t = (1, \mathbb{1}_{\{t \geq \tau\}})'$ and $\theta = (\mu_1, \mu_2)'$. We also introduce two discrete variables $z = \{z_1, z_2\}$ that indicates which component of the spike and slab distribution is active for μ_1 and μ_2. After standard computations, the posterior distributions of the model are given by

$$f(z_1, z_2|y_{1:T}, \tau) = \frac{(\sigma_{z_1}^2 \sigma_{z_2}^2)^{-\frac{1}{2}} |\bar{\Sigma}_{z_1,z_2}|^{\frac{1}{2}} \exp\{0.5 \bar{\mu}_{z_1,z_2}' \bar{\Sigma}_{z_1,z_2}^{-1} \bar{\mu}_{z_1,z_2}\} f(z_1, z_2)}{\sum_{z_1=1}^{2} \sum_{z_2=1}^{2} (\sigma_{z_1}^2 \sigma_{z_2}^2)^{-\frac{1}{2}} |\bar{\Sigma}_z|^{\frac{1}{2}} \exp\{0.5 \bar{\mu}_z' \bar{\Sigma}_z^{-1} \bar{\mu}_z\} f(z_1, z_2)}, \tag{42}$$

$$f(\mu_1, \mu_2|y_{1:T}, \tau, z_i) \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma}_z) \tag{43}$$

$$\bar{\Sigma}_z = [\Sigma_{z_i}^{-1} + \sum_{t=1}^{T} \frac{x_t x_t'}{\gamma^2}]^{-1} \quad (= \begin{pmatrix} \sigma_{z_1}^{-2} + \frac{T}{\gamma^2} & \frac{T-\tau}{\gamma^2} \\ \frac{T-\tau}{\gamma^2} & \sigma_{z_2}^{-2} + \frac{T-\tau}{\gamma^2} \end{pmatrix}^{-1}), \tag{44}$$

$$\bar{\mu}_z = \bar{\Sigma}_z (\sum_{t=1}^{T} \frac{x_t y_t}{\gamma^2}). \tag{45}$$

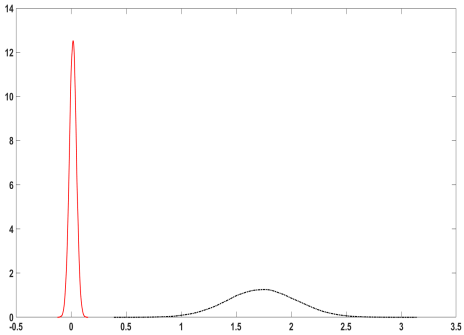
As expected, the posterior distribution of the toy model is a mixture of four components. Proposition 2 details the invariant distribution when we use the threshold algorithm to sample from a mixture of four components.

Proposition 2. *Consider sampling from the four-component distribution* $f(\theta) = \sum_{i=1}^{4} \omega_i f_i(\theta)$ *with* $\sum_{i=1}^{4} \omega_i = 1$ *in which* $f_i(\theta)$ *are proper densities and* $\theta \in \mathbb{R}^2$ *using the HKF algorithm*

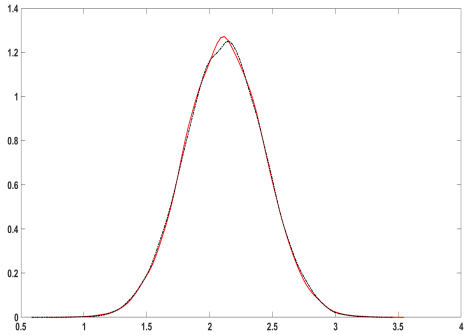
Author Manuscript

This article is protected by copyright. All rights reserved.

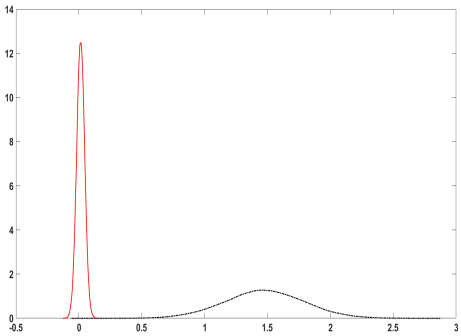
Author Manuscript



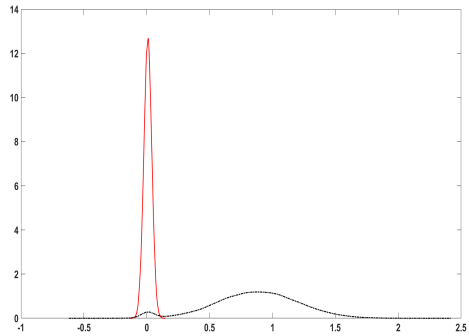
(a) $\mu = 2$, Starting value $\mu = 0$, $T = 10$



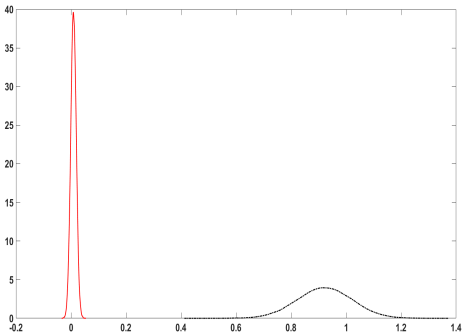
(b) $\mu = 2$, Starting value $\mu = 10$, $T = 10$



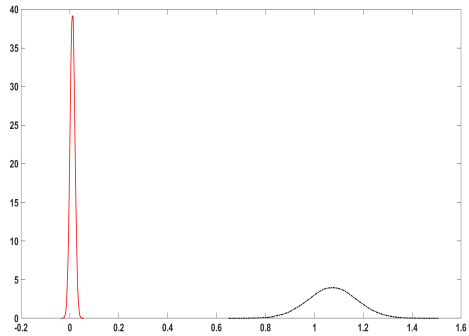
(c) $\mu = 1$, Starting value $\mu = 0$, $T = 10$



(d) $\mu = 1$, Starting value $\mu = 10$, $T = 10$



(e) $\mu = 1$, Starting value $\mu = 0$, $T = 100$



(f) $\mu = 1$, Starting value $\mu = 10$, $T = 100$

Figure 7: Posterior distribution of $\mu|y_{1:T}$: True posterior density in dotted black and density obtained by the HKF algorithm in red.

This article is protected by copyright. All rights reserved.

when the prior distribution of d is a Dirac function $f(d) = \mathbb{1}_{\{d=l_d\}}$. For each $i = 1, ..., 4$, let us denote four different probabilities to cut the support: $q_{i1} = P_i[|\theta_1| \leq l_d, |\theta_2| \leq l_d]$, $q_{i2} = P_i[|\theta_1| > l_d, |\theta_2| \leq l_d]$, $q_{i3} = P_i[|\theta_1| \leq l_d, |\theta_2| > l_d]$ and $q_{i4} = P_i[|\theta_1| > l_d, |\theta_2| > l_d]$. Then, the stationary distribution of the MCMC algorithm is also a mixture of four components, i.e. $\pi(\theta) = \sum_{i=1}^{4} \bar{\omega}_i f_i(\theta)$, with weights verifying the following equalities:

$$\underbrace{\begin{pmatrix} q_{11} & q_{21} & q_{31} & q_{41} \\ q_{12} & q_{22} & q_{32} & q_{42} \\ q_{13} & q_{23} & q_{33} & q_{43} \\ q_{14} & q_{24} & q_{34} & q_{44} \end{pmatrix}}_{P} \begin{pmatrix} \bar{\omega}_1 \\ \bar{\omega}_2 \\ \bar{\omega}_3 \\ \bar{\omega}_4 \end{pmatrix} = \begin{pmatrix} \bar{\omega}_1 \\ \bar{\omega}_2 \\ \bar{\omega}_3 \\ \bar{\omega}_4 \end{pmatrix}.$$

Proof. See Appendix F.2 □

Remark 4. Although this is precluded in a mixture of Normal distributions, the MCMC algorithm can exhibit multiple stationary distributions if some probabilities $q_{ij} = 0$ for $i \in [1, 4]$ and $j \in [1, 4]$.

Remark 5. Note that the indices ij of q_{ij} with $i \in [1, 4]$ and $j \in [1, 4]$ can be interpreted as probabilities of moving from component i to component j. So, the mixing of the HKF algorithm directly depends on these probabilities q_{ij}. For instance, the MCMC mixing will be slow if the probabilities of moving from component i to components j for all $j \neq i$ are very small.

Proposition 2 highlights that the stationary distribution can be very different from the posterior distribution from which we want to sample. To empirically show the problem, we simulate 20 observations from the multivariate toy model with values $\gamma = 1$, $\mu_1 = 2$, $\mu_2 = 0.5$ and $\tau = 5$. The weights of the posterior distribution amount to $\omega_1 = 0.00$, $\omega_2 = 0.669$ (i.e. $z_1 = 2$,$z_2 = 1$), $\omega_3 = 0.001$ and $\omega_4 = 0.33$ (i.e. $z_1 = 2$,$z_2 = 2$). Consequently, the posterior distribution is mostly a mixture implying the second and the last component. Setting $l_d = 0.2$ (i.e. 1% prior quantile of the HKF model), we can compute the probability matrix of Proposition 2 as we know analytically the distribution of each component. For this

Author Manuscript

This article is protected by copyright. All rights reserved.

example, the matrix is given by

$$\begin{pmatrix} q_{11} & q_{21} & q_{31} & q_{41} \\ q_{12} & q_{22} & q_{32} & q_{42} \\ q_{13} & q_{23} & q_{33} & q_{43} \\ q_{14} & q_{24} & q_{34} & q_{44} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.0722 \\ 0 & 0 & 1 & 0.0006 \\ 0 & 0 & 0 & 0.9273 \end{pmatrix}.$$

Note that some probabilities q_{ij} are exactly equal to one due to numerical approximations.[13] According to Proposition 2, the threshold algorithm exhibits three stationary distributions which arise depending on the starting value. In addition, there is no probability of sampling from the fourth component (when the MCMC algorithm has converged) even though the posterior weight of this component amounts to 33%. To illustrate these issues, we use the exact HKF algorithm (so with $\sigma_2^2 \sim \mathcal{IG}(3, 33.3)$ and $d \sim \mathcal{U}[0.1\sigma_2, 1.5\sigma_2]$). Figure 8 documents the posterior draws of the HKF algorithm given three different starting values. Depending on the starting value, we either detect the break with probability one (although the posterior probability of detecting it amounts to 0.331) or do not detect the break with probability one. Even though we do not have a Dirac function for the threshold prior distribution and we use the Uniform distribution of HKF, we observe that Proposition 2 is still valid as we clearly observe the three absorbing states generated by the HKF algorithm on Figure 8. Additional evidence of these problems are given in a Sparse CP-VAR framework in Section 5.

F.1 Proof of Proposition 1

Proof. For each prior distribution, we look for the density $\pi(\theta)$ that satisfies the invariant distribution condition, i.e. $\pi(\theta) = \int\int f(\theta|z = 1 + 1_{|\theta^*|\geq d})f(d|\theta^*)\pi(\theta^*)dd\ d\theta^*$ assuming that the invariant density can be expressed as $\pi(\theta) = \omega f_1(\theta) + (1-\omega)f_2(\theta)$. To ease the notations, we define $P_i[\theta \leq x] = \int_{-\infty}^{x} f_i(\theta)d\theta$ for $i = 1, 2$. Before detailing each prior distribution, note that the HKF algorithm iterates over the following distributions:

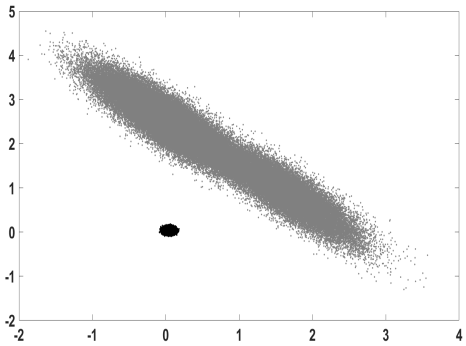
$$\begin{aligned} f(d|\theta) &\propto (f(\theta|z=1)\mathbb{1}_{\{|\theta|\leq d\}} + f(\theta|z=2)\mathbb{1}_{\{|\theta|>d\}})f(d), \\ z &= 1 + \mathbb{1}_{\{|\theta|>d\}}, \quad \text{(a deterministic function)}, \\ \theta|z=i &\sim \mathcal{N}(\mu_i, \sigma_i^2). \end{aligned}$$

[13]The probabilities are computed based on 1,000,000 draws.

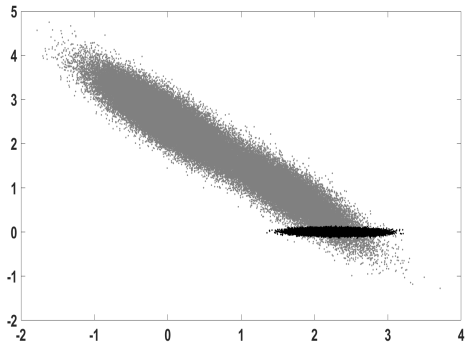
Author Manuscript

This article is protected by copyright. All rights reserved.

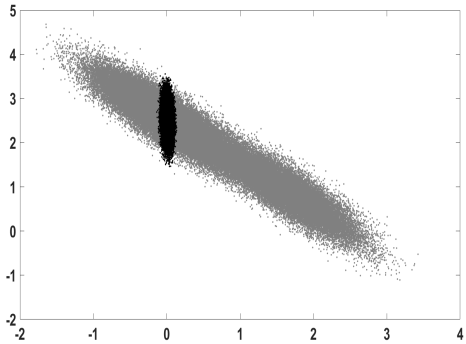
Author Manuscript



(a) Starting values $\mu = (0, 0)'$



(b) Starting values $\mu = (100, 0)'$



(c) Starting values $\mu = (0, 100)'$

Figure 8: Bivariate model: Posterior draws in grey and draws from the HKF algorithm in black.

This article is protected by copyright. All rights reserved.

1. When the prior distribution of d is a Dirac function, we have that

$$\begin{aligned}
\pi(\theta) &= \int f(\theta|z=1+1_{|\theta^*|\geq l_d})\pi(\theta^*)d\theta^*, \\
&= \omega\int f(\theta|z=1+1_{|\theta^*|\geq l_d})f_1(\theta^*)d\theta^* + (1-\omega)\int f(\theta|z=1+1_{|\theta^*|\geq l_d})f_2(\theta^*)d\theta^*, \\
&= \omega[q_1 f_1(\theta)+(1-q_1)f_2(\theta)] + (1-\omega)[q_2 f_1(\theta)+(1-q_2)f_2(\theta)], \\
&= [\omega q_1+(1-\omega)q_2]f_1(\theta) + [\omega(1-q_1)+(1-\omega)(1-q_2)]f_2(\theta), \\
&= \omega f_1(\theta)+(1-\omega)f_2(\theta), \qquad \text{so that } \omega = \frac{q_2}{(1+q_2-q_1)}.
\end{aligned}$$

2. When the threshold d can take uniformly two different values $\{l_d, u_d\}$, we have that

$$\begin{aligned}
\pi(\theta) &= \sum_{j=1}^{2}\underbrace{\int f(\theta|z=1+1_{|\theta^*|\geq d_j})f(d_j|\theta^*)\pi(\theta^*)d\theta^*}_{Q_j}, \\
Q_1 &= \int f(\theta|z=1+1_{|\theta^*|\geq l_d})f(d=l_d|\theta^*)\pi(\theta^*)d\theta^*, \\
&= \omega(\frac{q_1}{2}f_1(\theta)+(1-q_1)f_2(\theta)\underbrace{\int f(d=l_d|\theta^*)f_1(\theta^*||\theta^*|>l_d)d\theta^*)}_{Q_{11}}, \\
&\quad + (1-\omega)\underbrace{\int f(\theta|z=1+1_{|\theta^*|\geq l_d})f(d=l_d|\theta^*)f_2(\theta^*)d\theta^*}_{Q_{12}}, \\
Q_{11} &= \int \frac{(f_1(\theta^*)\mathbb{1}_{\{|\theta^*|\leq l_d\}}+f_2(\theta^*)\mathbb{1}_{\{|\theta^*|>l_d\}})}{\sum_{j=1}^{2}(f_1(\theta^*)\mathbb{1}_{\{|\theta^*|\leq d_i\}}+f_2(\theta^*)\mathbb{1}_{\{|\theta^*|>d_i\}})}f_1(\theta^*||\theta^*|>l_d)d\theta^* \\
&= 2\int_{l_d}^{u_d}\frac{f_2(\theta^*)f_1(\theta^*)}{f_1(\theta^*)+f_2(\theta^*)}d\theta^* + \frac{(1-P_1[|\theta|\leq u_d])}{2}. \\
Q_{12} &= [2\int_{l_d}^{u_d}\frac{f_2(\theta^*)^2}{f_1(\theta^*)+f_2(\theta^*)}d\theta^* + \frac{(1-P_2[|\theta|\leq u_d])}{2}].
\end{aligned}$$

Author Manuscript

This article is protected by copyright. All rights reserved.

In a similar way, we can develop the second term Q_2:

$$
\begin{aligned}
Q_2 &= \omega \underbrace{\int f(\theta|z=1+1_{|\theta^*|\geq u_d}) f(d=u_d|\theta^*) f_1(\theta^*) d\theta^*}_{Q_{21}} \\
&\quad + (1-\omega) \underbrace{\int f(\theta|z=1+1_{|\theta^*|\geq u_d}) f(d=u_d|\theta^*) f_2(\theta^*) d\theta^*}_{Q_{22}}, \\
Q_{21} &= \frac{q_1}{2} f_1(\theta) + (1-q_1)[f_1(\theta) 2 \int_{l_d}^{u_d} \frac{f_1(\theta^*)^2}{f_1(\theta^*)+f_2(\theta^*)} d\theta^* + f_2(\theta) \frac{(1-P_1[|\theta| \leq u_d])}{2}], \\
Q_{22} &= \frac{q_2}{2} f_1(\theta) + (1-q_2)[f_1(\theta) 2 \int_{l_d}^{u_d} \frac{f_1(\theta^*) f_2(\theta^*)}{f_1(\theta^*)+f_2(\theta^*)} d\theta^* + f_2(\theta) \frac{(1-P_2[|\theta| \leq u_d])}{2}]]
\end{aligned}
$$

We conclude that ω should verify:

$$
\begin{aligned}
\omega &= \omega q_1 + (1-\omega) q_2 + \omega(1-q_1) \underbrace{2 \int_{l_d}^{u_d} \frac{f_1(\theta^*)^2}{f_1(\theta^*)+f_2(\theta^*)} d\theta^*}_{a_1} \\
&\quad + (1-\omega)(1-q_2) \underbrace{2 \int_{l_d}^{u_d} \frac{f_1(\theta^*) f_2(\theta^*)}{f_1(\theta^*)+f_2(\theta^*)} d\theta^*}_{b_1}, \\
&= \frac{q_2 + b_1(1-q_2)}{(1+q_2-q_1-a_1(1-q_1)+b_1(1-q_2))}
\end{aligned}
$$

3. When the prior distribution of d is Uniform over the support $[l_d, u_d]$ and $q_1 = 1$, we have that

$$
\begin{aligned}
\pi(\theta) &= \int f(\theta|d) f(d|\theta^*) \pi(\theta^*) d\theta^* dd, \\
&= \omega \int f(\theta|d) f(d|\theta^*) f_1(\theta^*) d\theta^* dd + (1-\omega) \int f(\theta|d) f(d|\theta^*) f_2(\theta^*) d\theta^* dd, \\
&= \omega f_1(\theta) + (1-\omega) \underbrace{\int f(\theta|d) f(d|\theta^*) f_2(\theta^*) d\theta^* dd}_{R_1}, \\
R_1 &= q_2 f_1(\theta) + (1-q_2) \underbrace{\int f(\theta|d) f(d|\theta^*) f_2(\theta^* || \theta^*| > l_d) d\theta^* dd}_{R_2},
\end{aligned}
$$

Note that the distribution $d|\theta^* \sim \mathcal{U}[l_d, \min(\theta^*, u_d)]$ for $\theta^* \geq l_d$ since $f_1(\theta^*) = 0$, $\forall |\theta^*| \geq$

Author Manuscript

This article is protected by copyright. All rights reserved.

l_d, and because we have

$$\begin{aligned}
f(d|\theta^*, |\theta^*| \in (l_d, u_d)) &= f(d)\frac{(f_1(\theta^*)\mathbb{1}_{\{|\theta^*|\leq d\}} + f_2(\theta^*)\mathbb{1}_{\{|\theta^*|>d\}})}{\int (f_1(\theta^*)\mathbb{1}_{\{|\theta^*|\leq d\}} + f_2(\theta^*)\mathbb{1}_{\{|\theta^*|>d\}})f(d)dd}, \\
&= \frac{\mathbb{1}_{\{|\theta^*|>d\}}}{|\theta^*| - l_d}, \\
f(d|\theta^*, |\theta^*| \geq u_d) &= \frac{f_2(\theta^*)\mathbb{1}_{\{|\theta^*|>d\}}}{f_2(\theta^*)\int dd}, \\
&= \frac{\mathbb{1}_{\{d\in(l_d,u_d)\}}}{u_d - l_d}.
\end{aligned}$$

We can now simplify the last term R_2 as follows

$$\begin{aligned}
R_2 &= f_2(\theta)\int_{|\theta^*|\geq l_d}\int_{l_d}^{u_d} \frac{\mathbb{1}_{\{|\theta^*|>d\}}}{\min(|\theta^*|, u_d) - l_d} f_2(\theta^*||\theta^*| > l_d)d\theta^* dd, \\
&= f_2(\theta).
\end{aligned}$$

Combining these results, the weight of the invariant distribution should verify

$$\begin{aligned}
\omega &= \omega + (1-\omega)q_2, \\
&= 1.
\end{aligned}$$

□

F.2 Proof of Proposition 2

Proof. The invariant distribution verifies

$$\begin{aligned}
\pi(\theta) &= \int f(\theta|z_1 = 1 + 1_{|\theta_1^*|\geq l_d}, z_2 = 1 + 1_{|\theta_2^*|\geq l_d})\pi(\theta^*)d\theta^*, \\
&= \sum_{i=1}^{4}\omega_i \underbrace{\int f(\theta|z_1 = 1 + 1_{|\theta_1^*|\geq l_d}, z_2 = 1 + 1_{|\theta_2^*|\geq l_d})f_i(\theta^*)d\theta^*}_{Q_i} \\
Q_i &= \sum_{j=1}^{4} q_{ij} f_j(\theta).
\end{aligned}$$

Author Manuscript

This article is protected by copyright. All rights reserved.

The stationary density simplifies into

$$\pi(\theta) \quad = \quad \sum_{j=1}^{4}[\sum_{i=1}^{4}\omega_i q_{ij}]f_j(\theta),$$

which leads to the system of equations given in Proposition 2. □

Author Manuscript

This article is protected by copyright. All rights reserved.