Bank of Russia

TRUNCATED PRIORS FOR TEMPERED HIERARCHICAL DIRICHLET PROCESS VECTOR AUTOREGRESSION

No. 47 / October 2019

WORKING PAPER SERIES

Sergei Seleznev

Sergei Seleznev

Bank of Russia. Email: SeleznevSM@cbr.ru

The author is grateful to Valeriy Charnovoky, Sylvia Kaufmann, Ramis Khabibullin, Dmitriy Kreptsev, Mariam Mamedli, Alexey Ponomarenko and the participants in the 12th RCEA Bayesian Workshop and the 25th CEF International Conference for their helpful comments and suggestions.

Bank of Russia Working Paper Series is anonymously refereed by members of the Bank of Russia Research Advisory Board and external reviewers.

Cover image: Shutterstock.com

Address: 12 Neglinnaya street, Moscow, 107016
Tel.: +7 495 771-91-00, +7 495 621-64-65 (fax)
Website: www.cbr.ru

Abstract

We construct priors for the tempered hierarchical Dirichlet process vector autoregression model (tHDP-VAR) that in practice do not lead to explosive forecasting dynamics. Additionally, we show that tHDP-VAR and its variational Bayesian approximation with heuristics demonstrate competitive or even better forecasting performance on US and Russian datasets.

Keywords: Bayesian nonparametrics, forecasting, hierarchical Dirichlet process, infinite hidden Markov model.

JEL: C11, C32, C53, E37.

1. Introduction

Despite the fact that linear Bayesian vector autoregression (BVAR) models show relatively good forecasting performance (De Mol, Giannone & Reichlin, 2008; Giannone, Lenza & Primiceri, 2015), the existence of changes in the structure of the economy (a great recession, ZLB, monetary and fiscal policy changes in emerging markets, etc.) allows us to believe in non-linearity (time-varying parameters) in the data-generating process. Taking into account this time variation for BVARs is one of the most promising ways to increase the forecasting properties of these models.

In general, (B)VAR models with time-varying parameters can be written in the following form[1]:

$$y_t = b_t + A_{1,t}y_{t-1} + \cdots + A_{p,t}y_{t-p} + e_t \quad (1)$$

$$e_t \sim N(0; \Sigma_t) \quad (2)$$

where y_t is an $N \times 1$ vector of endogenous variables, e_t is an $N \times 1$ vector of shocks, and $b_t(N \times 1), A_{1,t}(N \times N), \dots, A_{p,t}(N \times N)$ and $\Sigma_t(N \times N)$ are time-varying parameters of the model. The process for the evolution of the parameters completes the model.

There is a vast amount of literature that incorporates different evolutions of the parameters in VAR models. This literature includes models with random walk coefficients (Cogley & Sargent, 2005; Primiceri, 2005), Markov-switching (MS) models (Sims, Waggoner & Zha, 2008; Bognanni & Herbst, 2017), threshold (Galvao & Marcellino, 2010) and latent threshold models (Nakajima & West, 2013), smooth transition models (Auerbach & Gorodnichenko, 2013), nonparametric models (Kapetanios, Marcellino & Venditti, 2016), and score-driven models (Gorgi, Koopman & Schaumburg, 2017), among others.

This paper concentrates on choosing a prior distribution for the Markov-switching VAR model with an infinite number of regimes, namely the tempered hierarchical Dirichlet process VAR (tHDP-VAR)[2], which might be helpful for short- and medium-term forecasting. The model with an infinite number of regimes has several advantages. It is not necessary to build a set of models with different numbers of regimes and choose between them (or weight them). In the tHDP-VAR model this is done automatically. Additionally, the tHDP-VAR model allows for the appearance of new regimes in the forecasting horizon, which might be especially useful in the case of a conditional forecasting procedure.

[1] For simplicity we assume Gaussian errors.
[2] We choose tHDP as a class representative because of its popularity, but any other model with an infinite number of regimes might be used.

This model originates from the Bayesian nonparametric literature (Ghosh & Ramamoorthi, 2003; Hjort, Holmes, Muller & Walker, 2010; Ghosal & Van der Vaart, 2017) and is based on the (tempered) hierarchical Dirichlet process (Beal, Ghahramani & Rasmussen, 2002; Teh, Jordan, Beal & Blei, 2006; Fox, Sudderth, Jordan & Willsky, 2007), which provides an elegant way to introduce an infinite number of switching regimes.

Bayesian nonparametric models are used for part-of-speech tagging (Van Gael & Ghahramani, 2011), topic modelling (Teh, Jordan, Beal & Blei, 2006; Wang, Paisley & Blei, 2011), speaker diarization (Stephenson & Raphael, 2015), human motion capture (Stephenson & Raphael, 2015) and description of visual scenes (Sudderth, Torralba, Freeman & Willsky, 2008), among other uses. However, to the best of our knowledge there are only a few works that apply a similar methodology for macroeconomic problems. Jochmann (2015) models inflation using hierarchical Dirichlet processes, and Song (2014) concentrates on real interest rates. Hou (2016) is the closest to our paper. This author predicts the dynamics of GDP inflation, GDP growth and the effective federal fund rate in the US. Although the specification proposed by Hou (2016) works well for US data, it does not guarantee adequate performance on other datasets (for example, on datasets with a larger number of series and/or "less stable" data) or longer horizons.

In this paper, we show that imposing traditional prior distributions for the tHDP-VAR model may lead to the occurrence of explosive forecasts. To mitigate this problem, we propose a procedure that assumes that the explosive roots of the VAR model are truncated. In addition to theoretical considerations, the paper describes a sampling algorithm and several heuristics that can be useful for accelerating and stabilizing the algorithm in practice.

To demonstrate the properties of the algorithm for real data, we compare the predictive performance of the model with the VAR model and the BVAR model in the spirit of Giannone, Lenza and Primiceri (2015) for US and Russian data. We show that the proposed algorithm and a number of heuristics work better than alternative models on these data.

The rest of the paper is organized as follows: in Section 2 we describe hierarchical Dirichlet processes; Markov-switching VAR with an infinite number of regimes is explained in Section 3; Sections 4 and 5 are devoted to the prior distributions and the estimation algorithm; the applications of algorithms are shown in Section 6; and Section 7 concludes.

2. Hierarchical Dirichlet processes

As its name suggests, one part of the hierarchical Dirichlet process (HDP) is the Dirichlet process (DP), so first we have to define the DP. A formal definition can be found in Ferguson (1973), but for a better understanding we use the stick-breaking construction definition given by Sethuraman (1994). The Dirichlet process with parameters α and G_0, $DP(G_0, \alpha)$, defines the distribution on distributions ($G \sim DP(G_0, \alpha)$):

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \, \delta(\theta - \theta^k) \tag{3}$$

where δ is the Dirac delta function, θ is the set of parameters of the distribution G and

$$\beta'_k \sim Beta(1, \alpha) \qquad k = 1, 2, \ldots \tag{4}$$

$$\pi_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) \qquad k = 1, 2, \ldots \tag{5}$$

$$\theta^k \sim G_0(\theta^k) \qquad k = 1, 2, \ldots \tag{6}$$

The hierarchical Dirichlet process (Teh, Jordan, Beal & Blei, 2006) defines a set of Dirichlet distributions, $G_1, \ldots, G_n$, with the common base distribution ($G_0 \sim DP(H, \gamma)$):

$$G_i \sim DP(G_0, \alpha) \qquad k = 1, 2, \ldots \tag{7}$$

$$G_0 \sim DP(H, \gamma) \tag{8}$$

3. Hierarchical Dirichlet process VAR

The hierarchical Dirichlet process can be used to construct a Markov-switching vector autoregression with an infinite number of regimes. The parameters of the model $\theta_t = \{b_t, A_{1,t}, \ldots, A_{p,t}, \Sigma_t\}$ follow the Markov process, with the probability of transition from state i to state j (π_{ij}) given by the substitution of (4) and (5) into (7) and (8):

$$\beta'_{0j} \sim Beta(1, \gamma) \qquad j = 1, 2, \ldots \tag{9}$$

$$\pi_{0j} = \beta'_{0j} \prod_{l=1}^{j-1} (1 - \beta'_{0l}) \qquad j = 1, 2, \ldots \tag{10}$$

$$\beta'_{ij} \sim Beta\left(\alpha \pi_{0j}, \alpha\left(1 - \sum_{l=1}^{j} \pi_{0l}\right)\right) \qquad i, j = 1, 2, \ldots \tag{11}$$

$$\pi_{ij} = \beta'_{ij} \prod_{l=1}^{j-1} (1 - \beta'_{il}) \qquad i, j = 1, 2, \ldots \tag{12}$$

Combining equations (1), (2), (9)-(12), and adding the prior distribution, H, for the parameters of each state, we obtain a Markov-switching vector autoregression with an infinite number of regimes. We call this model the hierarchical Dirichlet process VAR (HDP-VAR) in the spirit of Bayesian nonparametrics.

Following Fox, Sudderth, Jordan and Willsky (2007), we introduce persistence of states (a high probability of transition to the same state), which is traditionally used in time series modelling via DP. Equation (12) in this case is replaced by:

$$\pi'_{ij} = \beta'_{ij} \prod_{l=1}^{j-1}(1-\beta'_{il}) \qquad i,j = 1,2,\dots \qquad (13)$$

$$\pi_{ij} = \frac{\alpha}{\alpha+\kappa}\pi'_{ij} + \frac{\kappa}{\alpha+\kappa} I(i=j) \qquad i,j = 1,2,\dots \qquad (14)$$

where κ determines the state persistence, and $I(i=j)$ is an indicator that is 1 if $i=j$ and 0 otherwise. As can be seen from equation (14), the transition probability is shifted to the current state, and the persistence coefficient determines the degree of the shift. It can also be noted that for $\kappa = 0$, equations (13) and (14) are equivalent to equation (12), which shows that hierarchical Dirichlet processes are a special case of the tempered hierarchical Dirichlet process (tHDP).

4. Prior distributions

One should be cautious before applying non-truncated prior distributions for the coefficients of the BVAR based model, such as BVAR, MS-VAR or tHDP-VAR.

We illustrate this with the help of the Bayesian AR(1) model

$$y_t = A_1 y_{t-1} + e_t \qquad (15)$$

$$e_t \sim N(0,1) \qquad (16)$$

with Gaussian prior distribution

$$A_1 \sim N(0,1) \qquad (17)$$

Let us assume that there are no observations. It is easy to see from Figure 1 that in this case the probability of being outside the bounds of stationarity is nonzero. Hence, it is easy to obtain the following result, which is exactly the existence of explosive forecasts:

$$\lim_{h\to\infty}\left(\int y_{t+h}^2 p(y_{t+h}|y_t, A_1)p(A_1)dA_1\right) = \infty \qquad (18)$$

If the prior distribution is replaced with the posterior, then (18) still holds. The probability of being outside the bounds of stationarity becomes smaller but remains positive (see Figure 1). This usually just leads to a decreased influence of the explosive forecasts for a fixed horizon.

For the tHDP-VAR model this implies that there are three important sources[3] of unrealistic (explosive) forecasts: historically non-observed regimes, regimes with a small number of observations, and regimes with unusual data such as crises. The first two sources are similar in nature and are the result of the prior dominance. The third source contains large jumps in the observed variables that often shift a probability mass toward regions of non-stationarity. The problem of explosiveness is also relevant for BVAR, but in practice the probability of being outside the bounds of stationarity is reduced by the data. Explosiveness is therefore not a problem in BVARs for the horizons of interest. Even in those cases for which the explosiveness has to be excluded in BVAR, this can easily be done by applying simple criteria for bubble roots (Blake & Mumtaz, 2017[4]).

It is also easy to derive the stability conditions for the Markov-switching model with a finite number of regimes. Costa, Fragoso and Marques (2005) show that the stability of this class of model is closely related to the spectral radius of the model. So for the MS model, the stability of the model can be verified. The same conditions cannot be directly transferred to the model with an infinite number of regimes, and we just exclude the explosive roots of the coefficients of each regime. This does not ensure the stability of the model, but we find that this approach significantly mitigates the problem of explosive forecasts in practice. As will be demonstrated later, the forecasts for models with truncated priors lie in reasonable ranges and in most cases outperform BVAR forecasts.

Truncation of priors is done explicitly by multiplying the traditional prior distribution of each state by an indicator that the coefficients are not explosive. The Minnesota prior and the dummy-initial-observation prior from Giannone, Lenza and Primiceri (2015) are called “traditional prior distributions” in this paper.

The first and second moments of the Minnesota prior distribution[5] are:

$$E\left((A_s)_{ij}\middle|\Sigma\right) = \begin{cases} \delta_i & \text{if } i = j \text{ and } s = 1 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

$$cov\left((A_s)_{ij}, (A_r)_{hm}\middle|\Sigma\right) = \begin{cases} \frac{\lambda^2}{s^2}\frac{\Sigma_{ih}}{\psi_{jj}/(d-n-1)} & \text{if } m = j \text{ and } r = s \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

It is also assumed that the covariance matrix has an inverse Wishart distribution:

$$\Sigma \sim IW(\Psi, d) \tag{21}$$

[3] An explosive forecast may appear in all regimes, but in practice these three sources cover all cases for the horizons of interest.

[4] These authors use a similar truncation strategy for TVP-VAR, but do not prove its validity. As in the case described in this paper, it is just a heuristic.

[5] To simplify the notation, we omit the index of the regime.

where Ψ, d, λ are hyperparameters of the prior distribution, which additionally depend on hyperparameters. The matrix Ψ is chosen to be diagonal, with hyperparameters ψ_{jj} having an inverse gamma distribution with parameters 0.0004 and 0.0004. The hyperparameter d is set to be equal to the number of variables in the model plus 2 ($d = N + 2$). The hyperparameter λ has a gamma distribution with mode 0.2 and standard deviation 0.4.

The dummy-initial-observation prior can be implemented with dummy observations:

$$y^{++} = \frac{\bar{y}_0'}{\delta_0} \tag{22}$$

$$x^{++} = \left[\frac{1}{\delta_0}, y^{++}, \dots, y^{++}\right] \tag{23}$$

where $\bar{y}_0'$ is the average of the initial p observations and δ_0 is a hyperparameter having a gamma prior distribution with mode and standard deviation equal to 1.

The prior distributions (19), (20), (22) and (23) are truncated as follows:

$$vec\left(\left[A_1, \dots, A_p\right]'\right) \sim N\left(m_A(\lambda_H), \Sigma_A(\Sigma, \lambda_H)\right) I(l < 1 - \varepsilon) \tag{24}$$

$$b \sim N\left(m_b(\lambda_H), \Sigma_b(\Sigma, \lambda_H)\right) I(l < 1 - \varepsilon) \tag{25}$$

where $0 < \varepsilon < 1$, l is the largest absolute VAR root, λ_H is the set of hyperparameters[6], and $m_A(\lambda_H), \Sigma_A(\Sigma, \lambda_H), m_b(\lambda_H), \Sigma_b(\Sigma, \lambda_H)$ are the first and second moments of $\left[A_1, \dots, A_p\right]$ and b.

Note that the constraints (24) and (25) also implicitly transform a prior distribution "raising" in the region of stationarity bounds and its vanishing outside these bounds. Thus, in regions where only a small part of the distribution has become obscured, the probability of hyperparameters is multiplied by a large number.

Like Giannone, Lenza and Primiceri (2015), we set upper and lower bounds for the distributions of the hyperparameters[7], which also helps to mitigate the problem of explosive forecasts. For example, for the BVAR model, conditions (24) and (25) restrict explosive forecasts given the covariance matrix and hyperparameters, while by integrating out the covariance matrix and hyperparameters one can obtain unbounded moments. For the prior distributions from Giannone, Lenza and Primiceri (2015), the parameter constraints are set to limit the moments of the forecasts. To avoid tedious mathematics, we illustrate this problem by using an AR(0) model:

$$y_t = e_t \tag{26}$$

[6] It is easy to show that the Minnesota prior and the dummy-initial-observation prior can be written as:

$$vec\left(\left[A_1, \dots, A_p\right]'\right) \sim N\left(m_A(hyp), \Sigma_A(\Sigma, hyp)\right)$$
$$b \sim N\left(m_b(hyp), \Sigma_b(\Sigma, hyp)\right)$$

[7] The restrictions are not described in the text but occur in the code.

$$e_t \sim N(0, \sigma) \quad (27)$$

$$\sigma^2 \sim IW(\psi, 3) \quad (28)$$

$$\psi \sim IG(0.0004, 0.0004) \quad (29)$$

Conditions (24) and (25) are satisfied. However, as can be seen from (30), the mean of y_t^2 does not exist because of the absence of a mean of ψ:

$$\int y_t^2 p(y_t|\sigma^2, \psi) p(\sigma^2|\psi) p(\psi) dy_t d(\sigma^2) d\psi = \int \sigma^2 p(\sigma^2|\psi) p(\psi) d(\sigma^2) d\psi = \int \psi p(\psi) d\psi \quad (30)$$

Setting upper and lower bounds for ψ restricts the moments of y_t^2.

The tempered hierarchical Dirichlet process has a number of hyperparameters: α, γ and κ. We choose prior distributions for the hyperparameters that are the same as in Hou (2016), except for the distribution of $\alpha + \kappa$, which is slightly biased to the right:

$$\alpha + \kappa \sim Gamma(a_{\alpha+\kappa}, b_{\alpha+\kappa}) \quad (31)$$

$$\gamma \sim Gamma(a_\gamma, b_\gamma) \quad (32)$$

$$\rho = \frac{\kappa}{\alpha+\kappa} \sim Beta(a_\rho, b_\rho) \quad (33)$$

where $a_{\alpha+\kappa} = 10, b_{\alpha+\kappa} = 1, a_\gamma = 1, b_\gamma = 1, a_\rho = 10, b_\rho = 1$.

Also note that tHDP-VAR, like other MS-VAR models, should in fact be trained from scratch when a new regime appears because of the problem of a short sample. This disadvantage can be partially reduced in several ways, if necessary. For example, one can choose hierarchical prior distributions like those chosen by Hou (2016), who suggests using more flexible hyperparameters for mean and covariance matrices than Giannone, Lenza and Primiceri (2015) do. This allows new regimes to be learnt from the previous ones. Another way is to share a part of the parameters for all regimes (for example, Sims, Waggoner and Zha (2008) share the mean), which helps to reduce the number of estimated parameters. Dynamic stochastic general equilibrium (DSGE) priors can also be used (Del Negro & Schorfheide, 2004) to add information from the DSGE model before the data occurrence. All these methods can be applied and, depending on the task, may be better or worse than their alternatives. In this paper, another simple method is used to add information to the new regimes. In addition to the priors from Giannone, Lenza and Primiceri (2015), we add a number of real observations from pre-sample data or from the sample for another country. In fact, these are dummy observations without learned hyperparameters. These dummy observations are added to help the model choose coefficients, as it might do, for example, with DSGE priors, but with real instead of artificial data.

5. Model estimation

A mixture of the traditional algorithm from Fox, Sudderth, Jordan and Willsky (2007) and the beam sampler from Van Gael, Saatci, Teh and Ghahramani (2008) is used for the estimation of tHDP. This algorithm is similar to the algorithms presented by Song (2014) and Hou (2016), and it is described in Appendix B (we call it the basic algorithm).

In addition to small differences between our algorithm and the algorithm described by Hou (2016), such as the use of the adaptive Metropolis‑Hastings (MH) algorithm (Roberts & Rosenthal, 2009) for sampling VAR hyperparameters or the scheme for sampling auxiliary variables, there is a step for the coefficients for the truncated VAR that is largely determined by computational time. Coefficients for the truncated model cannot be sampled from the normal-inverse-Wishart distribution like in the case of non-truncated priors. To solve this problem, the accept-reject algorithm is applied to the non-truncated model for each regime separately. The number of iterations in the accept-reject algorithm can be large, especially for regimes where a small proportion of the non-truncated distribution lies in the stationary region ($I(l < 1 - \varepsilon)$), which may lead to a significantly increased computational time.

In order to mitigate the problem, which usually occurs in models with a large number of series and/or "less stable" series, we offer a heuristic that helps to reduce the computational time for forecasting models. Even for a model with three variables (see below), this heuristic helps to reduce computational time by several times. This algorithm does not draw exactly from the posterior, but it is expected that its forecasting performance might be close to the exact algorithm. At the first step, the model with non-truncated priors is estimated using the variational Bayes algorithm[8] (see Appendix C). At the forecasting stage, the state of the last period, transition probabilities and VAR coefficients are sampled from the approximate posterior density, and are accepted if the VAR coefficients lie in the stationary region. We should note that the VAR coefficient has to be drawn only for the regime of the last period and the regimes that are sampled in the forecasting period. It helps to avoid sampling coefficients for "rare" regimes, which are often "less stable". Alternatively, we check the stability of the system for each new regime using the criteria from Costa, Fragoso and Marques (2005) and ensure the stability of the forecasts.

[8] See Beal (2003) and Wainwright and Jordan (2008) for an introduction to variational Bayes approximation.

6. Applications

We apply the proposed algorithms to two datasets: US data (change of logarithm of GDP, change of logarithm of GDP deflator, and effective federal fund rate), and data for the Russian Federation (change of logarithm of GDP, change of logarithm of CPI, and MIACR interest rate (weighted average actual rates on Moscow banks' credits)). The first dataset (from 1959Q2 to 2008Q4) is taken from Giannone, Lenza and Primiceri (2015) and is more or less standard for papers on VAR and MS-VAR models. The second dataset is of interest because of the possible presence of several structural breaks in a short period of time (we use data from 1997Q2 to 2016Q4).

For each dataset, we estimate several models: a VAR model, a BVAR model, and a tHDP-VAR model estimated by basic and variational algorithms. We also check the usefulness of adding several real points into each regime for the variational algorithm, and we check for stability using the algorithm from Costa, Fragoso and Marques (2005). All models are estimated for lags from 1 to 5.

For each model, we run recursive forecasts (estimate the model, run the forecasting procedure, add one point to the sample, estimate the model, etc.). The VAR model is trained using the maximum likelihood method, and then coefficients are used for point forecasting. The BVAR model is trained using the MH algorithm (5,000 initial iterations and 20,000 main iterations) in a similar way to the method in Giannone, Lenza and Primiceri (2015). The tHDP-VAR model is estimated using the algorithm from Appendix B. For the preliminary estimation, 110,000 iterations of the algorithm are run for starting points (without inclusion of the first forecast point). Then we add one point and train the model using the last values of the sampled parameters and variables as the initial ones, and also setting the regime for the new period equal to the previous one[9]. As for the BVAR model, only 20,000 of the 25,000 iterations are saved. For simplicity, we fix the hyperparameters for HDP in the variational algorithm, setting $\alpha = 1$, $\gamma = 1$, $\kappa = 10$, which is an additional model constraint; this constraint may influence the forecasting performance, but in our experience does not do so for a wide range of hyperparameters. Taking into account that the variational approximation can be used without initial draws (see Appendix C), only 20,000 trajectories are sampled.

[9] The retraining procedure helps to reduce computational time and is successfully applied in machine learning (Graves, 2012).

We concentrate on five issues: 1) does the non-truncated model produce unrealistic forecasts? 2) are the MS model forecasting properties better than the forecasting properties of VAR and BVAR? 3) is variational approximation useful? 4) are additional observations useful? and 5) does our simple truncation strategy perform in a similar way to more complex strategies?

Does the non-truncated model produce unrealistic forecasts? We demonstrate the existence of explosive forecasts for the non-truncated model by plotting the US GDP deflator forecast for tHDP-VAR with 3 lags at the first forecasting point. Figure 2 shows that the change of logarithm of the GDP deflator is explosive, which confirms this. Using this type of forecast obviously leads to a large mean square forecast error, so we do not demonstrate the results of non-truncated MS models later.

Are the MS model forecasting properties better than forecasting properties of VAR and BVAR? We start the forecasting procedure from 1974Q4 and 2004Q2 and predict values for horizons from 1 to 12. RWMSFE is used (root weighted mean square forecast error) as a measure of forecasting performance:

$$RWMSFE = \sqrt{\sum_{i=1}^{N} \frac{MSFE_i}{var(y_i)}}$$

where $MSFE_i$ is the mean square forecast error for the ith variable and $var(y_i)$ is the empirical variance of the ith variable.

Figures 3–7 demonstrate the relative RWMSFEs (VAR(1) is the benchmark) for the US dataset depending on the starting point for calculation (Table 1 contains RWMSFE for the full test period), because the results might be sensitive to the period of testing. The same results for the Russian data are shown in Figures 8–12 and Table 2. The ranking of the models depends on the starting points, but for almost all the models and forecasting horizons tHDP-VAR outperforms VAR and BVAR. BVAR and VAR rarely produce competitive results, and only do so for large horizons (8, 12) or large lags (3–5)[10]. Moreover, for both datasets the most preferable tHDP-VAR model usually contains 1–2 lags, so the worsening of the forecasting properties for the larger lag lengths it is not a problem. These results help to show that for both datasets the MS model is useful for forecasting.

[10] Note also that for the US data such results appear only for starting points after 1980.

Is variational approximation useful? We are basically concerned with the forecasting properties of the models, so we measure usefulness as an improvement/deterioration of RWMSFE. Figures 3–12 demonstrate that RWMSFE is usually larger for variational Bayesian (VB) approximation. There are many possible reasons for this behaviour of VB approximation: initialization (or optimization procedure), having fixed hyperparameters, approximating the non-truncated model, and using a poor approximation family, among others. We ran a number of additional experiments and found that, for the Russian dataset, the optimizer fell into a local optimum for some points, so the results for this dataset are partially related to this. For some different sets of hyperparameters, the RWMSFEs are similar to those plotted, which is a sign of the small impact of this factor[11]. We failed to split other effects and leave this for further research.

Despite the fact that the VB approximation is sometimes not good, this model might still be useful for forecasting. For example, for short horizons on the Russian dataset it performs better than VAR and BVAR. We relate this to large outliers in the Russian dataset and the fact that the MS model does not use these explicitly for other regimes. Additionally, as expected, the VB algorithm reduces computational time by a factor of 2–4.

Are additional observations useful? To investigate this question, we add ten additional observations for each regime in the VB model for both datasets. For US data, we choose ten starting points as additional observations. The starting points for the Russian dataset are extremely volatile and cannot be used as information for regimes. We found that, in practice, the same ten starting points from the US dataset are a good choice.

For the US dataset, there is no difference between the previously estimated BVAR and the BVAR estimated with the additional observations. This is the consequence of the fact that we exclude starting points from the data and add them again as additional observations. By contrast, the VB model demonstrates changes in forecasting performance. Depending on the lag and forecasting horizon, it performs better or worse than the model without additional observations (usually better), but in all cases except for one (5 lags, 1 period to predict) it is no worse than the BVAR and VAR models. For the Russian dataset, the additional observations do the same work (slightly changing the results for BVAR). For this dataset, the VB model with additional observations is always better than VAR and BVAR.

In some cases, additional observations decrease the performance for shorter horizons, but in all cases they improve RWMSFE for longer horizons. As mentioned above,

[11] We probably did not choose enough competitors because of computational constraints.

the VB model with additional observations improves the forecasts of BVAR and VAR, so, of course, this model is useful. For the "more unstable" Russian dataset, additional observations help in two other ways. First, they decrease computational times by a factor of 30–50 by stabilizing the regimes and producing fewer samples with bubble roots. Secondly, the VB algorithm with additional observations falls into poor local optima less often.

Does our simple truncation strategy perform in a similar way to more complex strategies? We ran the VB model with additional observations with the alternative truncation strategy that ensures non-explosive forecasts. To save space, only the results for 3 lags are plotted. Figures 13–14 demonstrate that the model with alternative truncation has similar RWMSFEs, so our truncation strategy works well for both datasets.

7. Conclusion

We investigate the occurrence of explosive forecasts in tHDP-VAR models. Our analysis demonstrates that one of the sources of explosive forecasts is the nonzero probability of large roots for each regime. Despite the fact that this is not the unique source of this type of forecast (and is not always the source), we show that excluding bubble roots from each regime helps to mitigate the problem for reasonable forecasting horizons for US and Russian datasets. We also find that tHDP-VAR and VB approximations (with additional observations) might be useful for forecasting.

References

Auerbach, A.J. & Gorodnichenko, Y. (2013). Output spillovers from fiscal policy. American Economic Review, 103(3), 141-146.

Beal, M.J. (2003). Variational algorithms for approximate Bayesian inference. PhD thesis, University of London.

Beal, M.J., Ghahramani, Z. & Rasmussen, C. (2002). The infinite hidden Markov model. In: T.G. Dietterich, S. Becker & Z. Ghahramani (Eds.) Advances in neural information processing systems. Cambridge: MIT Press, pp. 577-584.

Bishop, C.M. (2006). Pattern recognition and machine learning. New York: Springer-Verlag.

Blake, A.P. & Mumtaz, H. (2017). Applied Bayesian econometrics for central bankers. Centre for Central Banking Studies, Bank of England, Technical Books.

Bognanni, M. & Herbst, E. (2017). A sequential Monte Carlo approach to inference in multiple-equation Markov-switching models. Journal of Applied Econometrics, 33(1), 126-140.

Cogley, T. & Sargent, T.J. (2005). Drifts and volatilities: Monetary policies and outcomes in the post WWII US. Review of Economic Dynamics, 8(2), 262-302.

Costa, O., Fragoso, M. & Marques, R. (2005). Discrete-time Markov jump linear systems. New York: Springer-Verlag.

De Mol, C., Giannone, D. & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? Journal of Econometrics, 146(2), 318-328.

Del Negro, M. & Schorfheide, F. (2004). Priors from general equilibrium models for VARS. International Economic Review, 45(2), 643-673.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1, 209-230.

Fox, E.B., Sudderth, E.B., Jordan, M.I. & Willsky, A.S. (2007). Developing a tempered HDP-HMM for systems with state persistence. MIT Laboratory for Information and Decision Systems, Technical Report P-2777.

Galvao, A.B. & Marcellino, M. (2010). Endogenous monetary policy regimes and the great moderation. European University Institute Working Paper, 2010/22.

Ghosal, J.K. & Van Der Vaart, A. (2017). Fundamentals of nonparametric Bayesian inference. Cambridge: Cambridge University Press.

Ghosh, J.K. & Ramamoorthi, R.V. (2003). Bayesian nonparametrics. New York: Springer-Verlag.

Giannone, D., Lenza, M., & Primiceri, G.E. (2015). Prior selection for vector autoregressions. The Review of Economics and Statistics, 97, 436-451.

Gorgi, P., Koopman, S.J. & Schaumburg, J. (2017). Time-varying vector autoregressive models with structural dynamic factors. Manuscript.

Graves, A. (2012). Supervised sequence labelling with recurrent neural networks. New York: Springer-Verlag.

Hjort, N.L., Holmes, C., Muller, P. & Walker, S.G. (2010). Bayesian nonparametrics. Cambridge: Cambridge University Press.

Hou, C. (2016). Infinite hidden Markov switching VARs with application to macroeconomic forecast. Manuscript.

Hughes, M.C., Kim D.I. & Sudderth, E.B. (2015). Reliable and scalable variational inference for the hierarchical Dirichlet process. JMLR, W&CP, 38, 370‑378.

Jochmann, M. (2015). Modeling US inflation dynamics: A Bayesian nonparametric approach. Econometric Reviews, 34(5), 537-558.

Kapetanios, G., Marcellino, M. & Venditti, F. (2016). Large time-varying parameter VARs: A non-parametric approach. CEPR Discussion Papers 11560.

Nakajima, J. & West, M. (2013). Bayesian analysis of latent threshold dynamic models. Journal of Business and Economic Statistics, 31(2), 151-164.

Primiceri, G.E. (2005). Time varying structural vector autoregressions and monetary policy. Review of Economic Studies, 72, 821-852.

Roberts, G. & Rosenthal, J. (2009). Examples of adaptive MCMC. Journal of Computational and Graphical Statistics, 18, 349-367.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica, 4, 639-650.

Sims, C.A., Waggoner, D.F. & Zha, T. (2008). Methods for inference in large multiple-equation Markov-switching models. Journal of Econometrics, 146(2), 255-274.

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

Stephenson, W. & Raphael, B.J. (2015). Variational inference for hierarchical Dirichlet process based nonparametric models. Manuscript.

Sudderth, E.B., Torralba, A., Freeman, W.T. & Willsky, A.S. (2008). Describing visual scenes using transformed objects and parts. International Journal of Computer Vision, 77, 291-330.

Teh, Y.W., Jordan, M.I., Beal, M.J. & Blei, D.M. (2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476), 1566-1581.

Van Gael, J. & Ghahramani, Z. (2011). Nonparametric hidden Markov models. In: D. Barber, A.T. Cemgil and S. Chiappa (Eds.), Bayesian time-series models, Cambridge: Cambridge University Press, pp. 317-340.

Van Gael, J., Saatci, Y., Teh Y.W. & Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. Proceedings of the 25th International Conference on Machine Learning, 1088-1095.

Wainwright, M. & Jordan, M. (2008). Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1(1–2), 1-305.

Wang, C., Paisley, J. & Blei, D. (2011). Online variational inference for the hierarchical Dirichlet process. In Artificial Intelligence and Statistics.

Yedidia, S.J., Freeman, W.T. & Weiss, Y. (2001). Understanding belief propagation and its generalizations. Mitsubishi Electric Research Laboratories, TR2001-22.

Appendix A

model	lag	1	2	4	8	12
VAR		1	1	1	1	1
BVAR		0,97	0,98	0,98	1,00	1,02
tHDP-VAR	**1**	**0,94**	**0,91**	**0,88**	**0,81**	**0,77**
VB		0,96	0,93	0,93	0,86	0,82
VB with obs		0,95	0,94	**0,88**	0,84	0,79
VAR		1,03	1,03	1,00	0,94	0,89
BVAR		0,98	0,99	0,98	0,96	0,94
tHDP-VAR	**2**	**0,97**	**0,94**	**0,91**	**0,80**	**0,75**
VB		0,98	0,98	0,96	0,85	0,79
VB with obs		0,98	1,00	0,92	0,83	0,77
VAR		1,02	1,03	1,02	1,04	1,05
BVAR		0,97	0,99	0,97	0,97	0,97
tHDP-VAR	**3**	**0,96**	**0,94**	**0,90**	**0,81**	**0,78**
VB		0,99	1,00	0,98	0,86	0,79
VB with obs		**0,96**	0,98	0,91	0,83	**0,78**
VAR		1,07	1,07	1,04	1,03	0,97
BVAR		0,98	0,99	0,98	0,97	0,96
tHDP-VAR	**4**	**0,95**	**0,95**	0,91	**0,80**	**0,75**
VB		1,00	1,01	1,00	0,88	0,80
VB with obs		0,96	0,99	**0,90**	0,83	0,77
VAR		1,09	1,09	1,03	1,01	0,92
BVAR		0,97	0,99	0,97	0,96	0,93
tHDP-VAR	**5**	**0,96**	**0,94**	**0,89**	**0,80**	**0,74**
VB		0,97	0,98	0,96	0,86	0,80
VB with obs		0,99	0,98	0,90	0,81	0,75

Table 1. Relative RWMSFEs (VAR(1) is benchmark) for the US dataset

model	lag	1	2	4	8	12
VAR		1	1	1	1	1
BVAR		0,87	0,87	0,85	0,99	0,98
tHDP-VAR	**1**	**0,61**	**0,58**	**0,52**	0,68	0,76
VB		0,67	0,71	0,75	1,08	1,27
VB with obs		0,65	0,61	**0,52**	**0,63**	**0,71**
VAR		1,11	1,16	0,74	0,88	1,01
BVAR		0,93	0,93	0,75	0,85	0,93
tHDP-VAR	**2**	**0,60**	0,60	0,56	0,79	0,96
VB		0,70	0,74	0,77	1,34	1,67
VB with obs		**0,60**	**0,55**	**0,46**	**0,62**	**0,71**
VAR		1,24	1,50	0,73	0,72	0,85
BVAR		1,06	1,18	0,75	0,78	0,94
tHDP-VAR	**3**	**0,60**	**0,60**	0,58	0,89	1,12
VB		0,75	0,77	0,79	1,21	1,48
VB with obs		0,80	**0,60**	**0,49**	**0,67**	**0,78**
VAR		1,19	1,42	1,07	1,19	0,90
BVAR		1,06	1,21	0,90	0,92	0,98
tHDP-VAR	**4**	**0,60**	**0,60**	0,54	0,79	0,93
VB		0,71	0,76	0,77	1,16	1,31
VB with obs		0,79	0,64	**0,50**	**0,66**	**0,75**
VAR		1,24	0,89	0,58	0,77	0,93
BVAR		1,08	0,85	0,59	0,75	0,93
tHDP-VAR	**5**	**0,59**	**0,58**	0,52	0,72	0,80
VB		0,70	0,74	0,75	1,07	1,18
VB with obs		0,63	0,59	**0,49**	**0,68**	**0,80**

Table 2. Relative RWMSFEs (VAR(1) is benchmark) for the Russian dataset

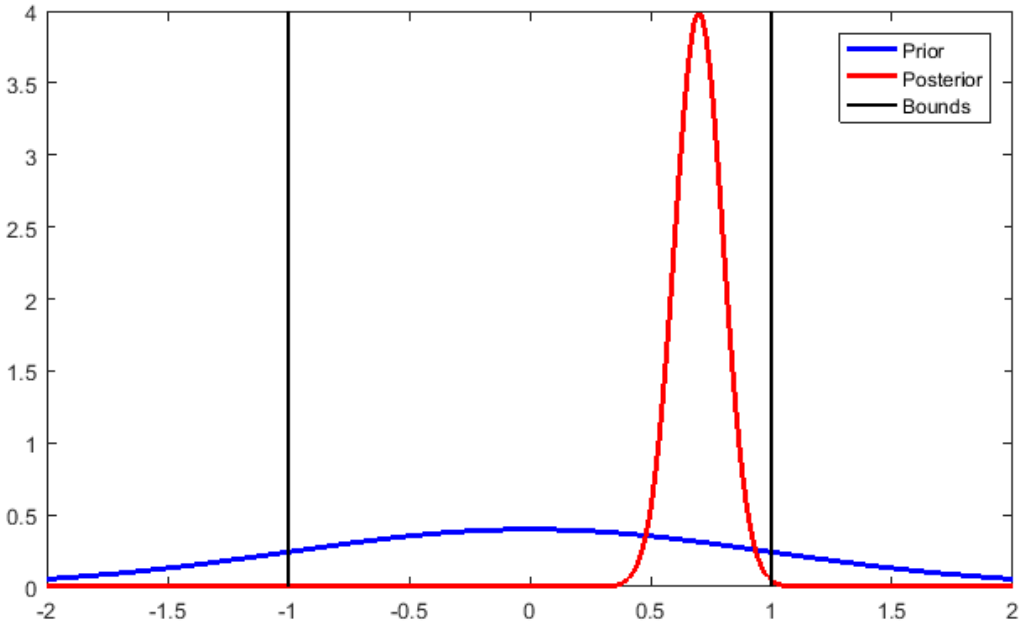

Figure 1. Prior, posterior and bounds of stationarity for the AR(1) model

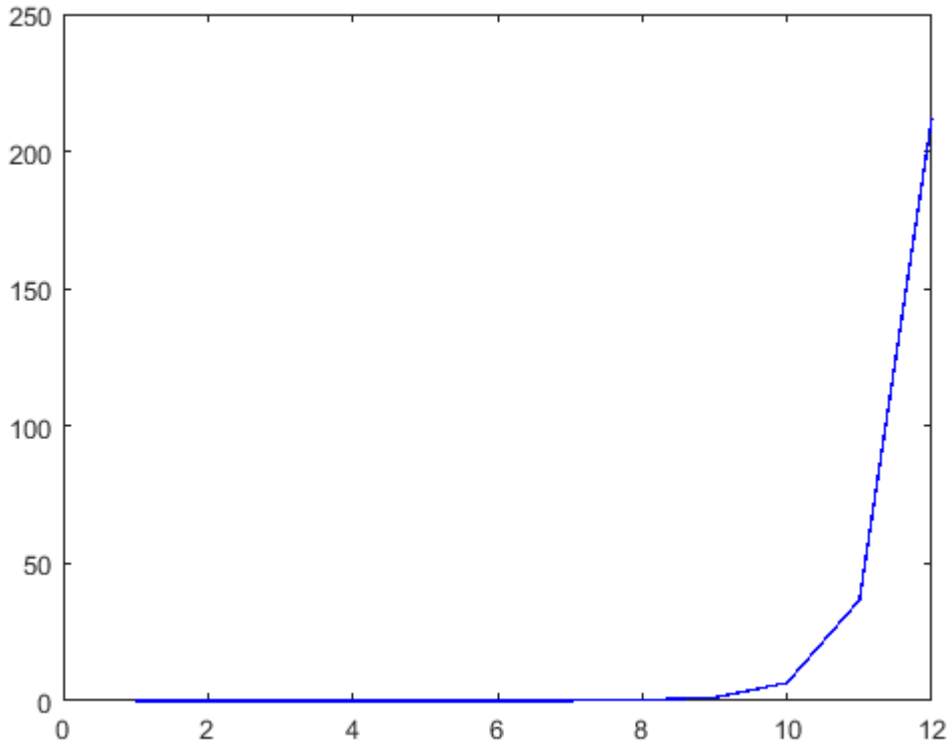

Figure 2. US GDP deflator forecast for tHDP-VAR with 3 lags at the first point, log-change

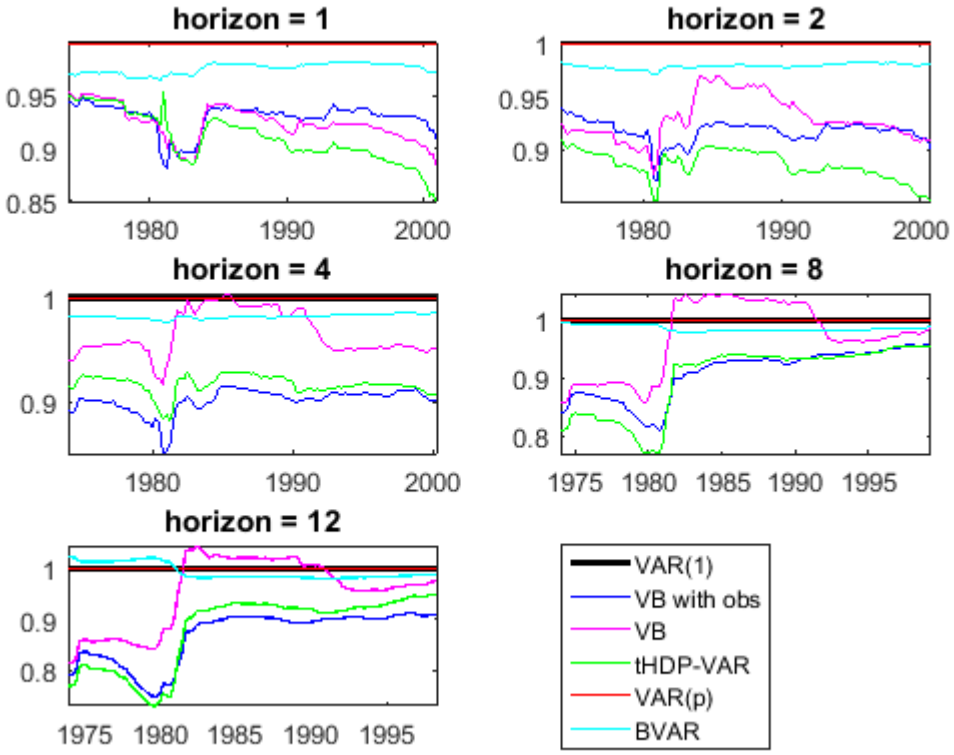

Figure 3. Relative RWMSFEs (VAR(1) is benchmark) for the US dataset depending on the starting point for calculation, 1 lag

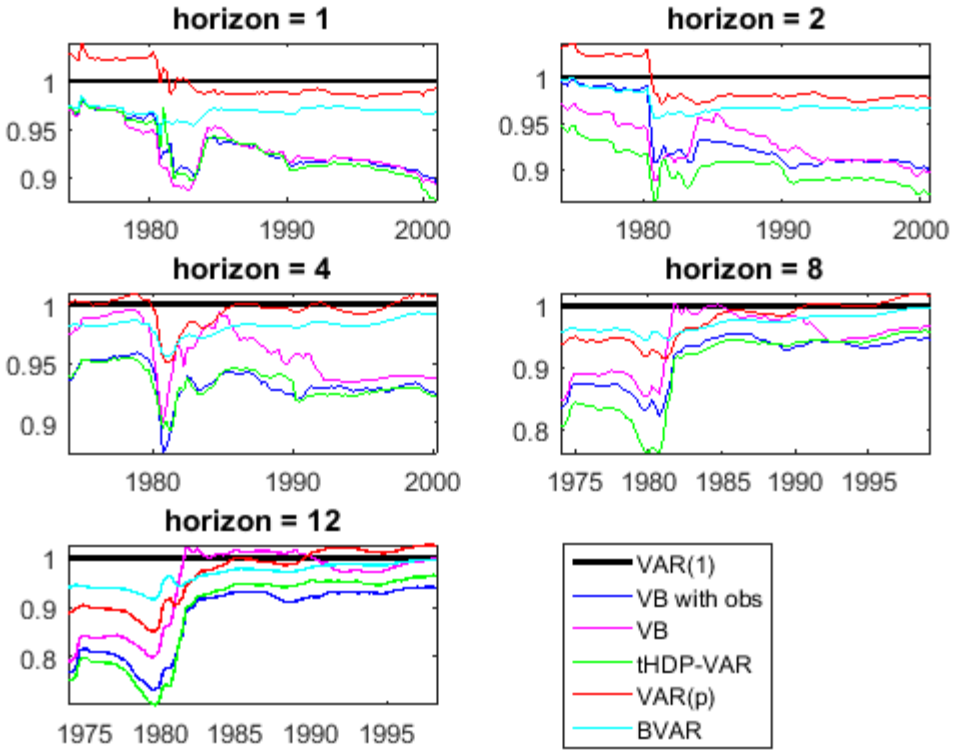

Figure 4. Relative RWMSFEs (VAR(1) is benchmark) for the US dataset depending on the starting point for calculation, 2 lags

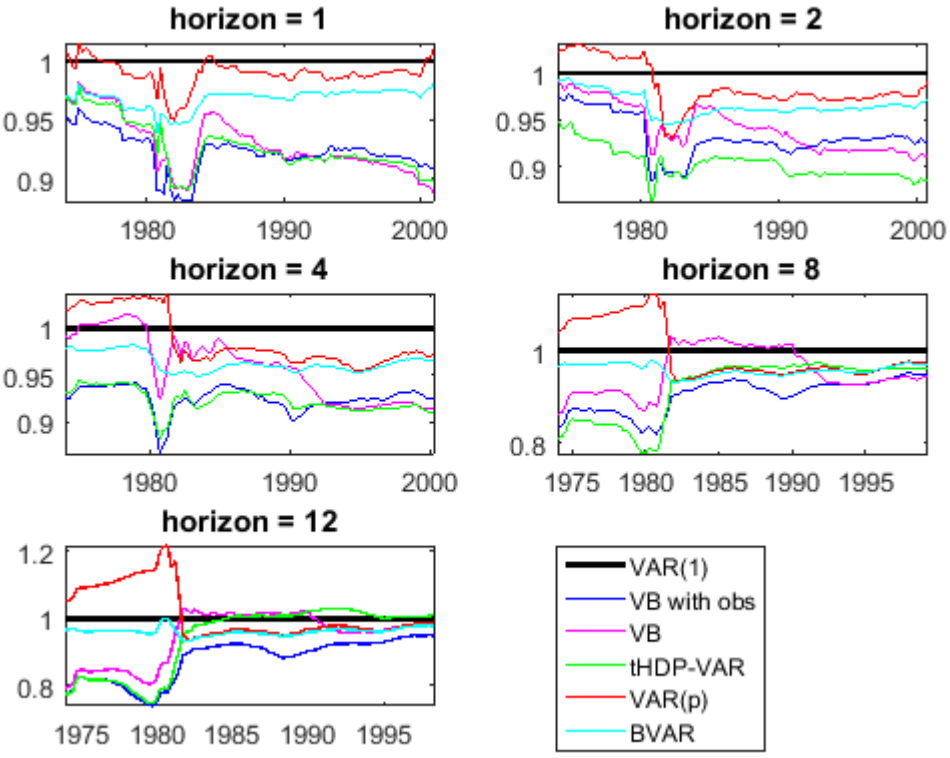

Figure 5. Relative RWMSFEs (VAR(1) is benchmark) for the US dataset depending on the starting point for calculation, 3 lags

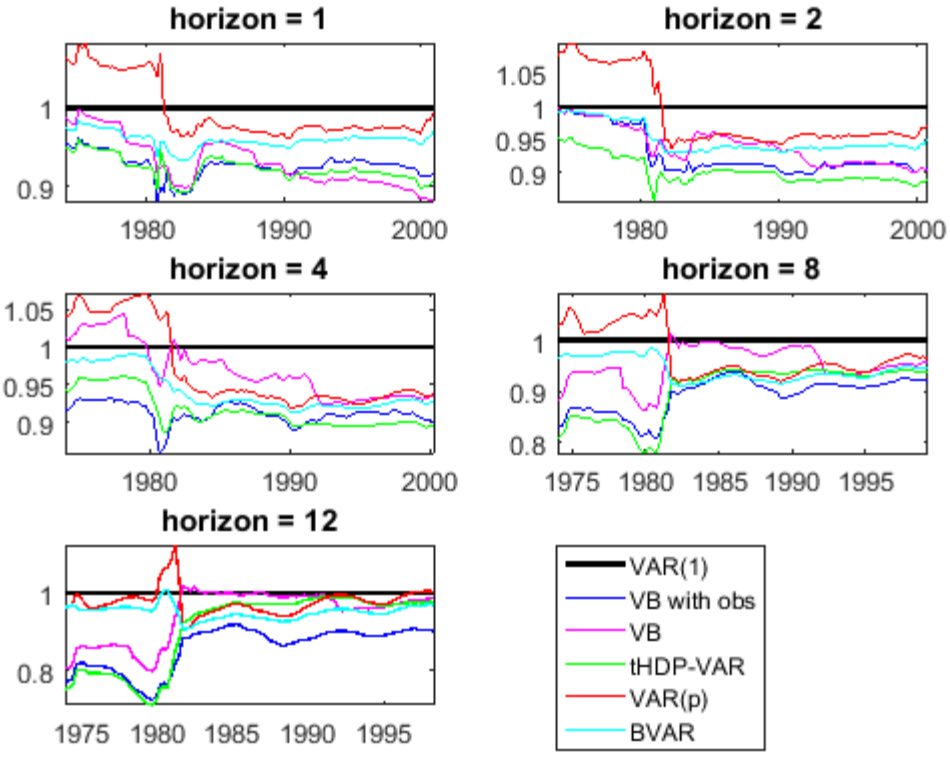

Figure 6. Relative RWMSFEs (VAR(1) is benchmark) for the US dataset depending on the starting point for calculation, 4 lags

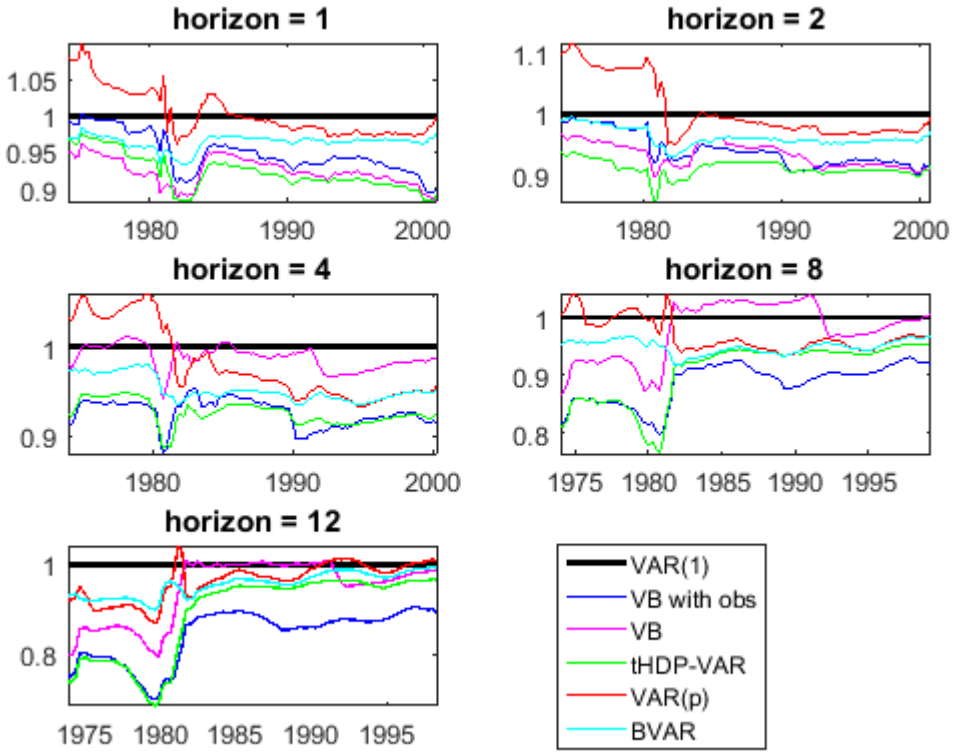

Figure 7. Relative RWMSFEs (VAR(1) is benchmark) for the US dataset depending on the starting point for calculation, 5 lags

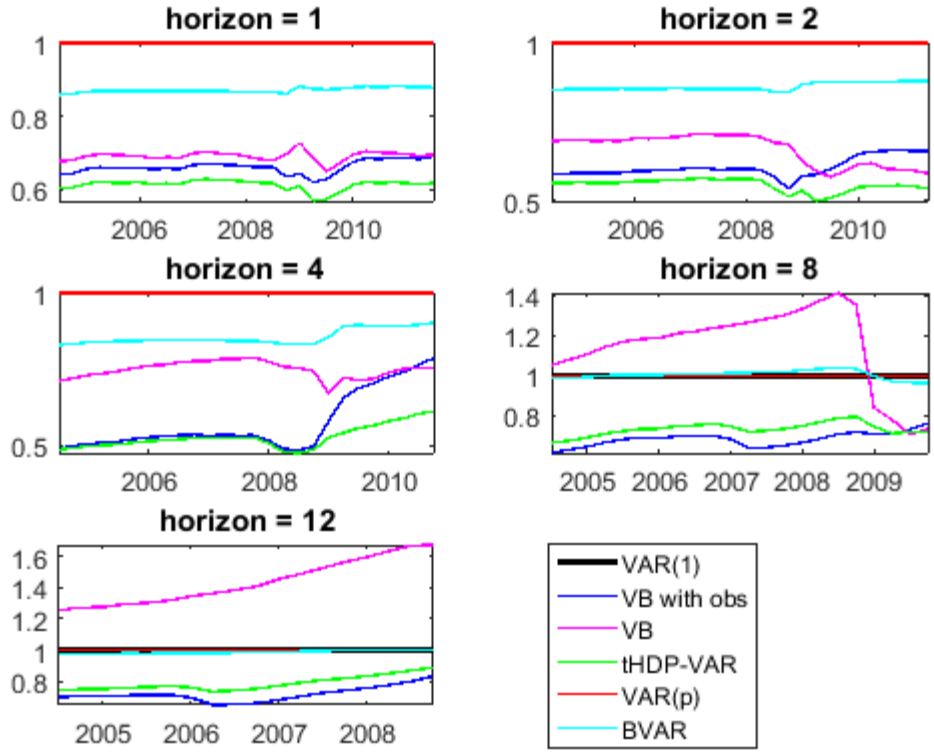

Figure 8. Relative RWMSFEs (VAR(1) is benchmark) for the Russian dataset depending on the starting point for calculation, 1 lag

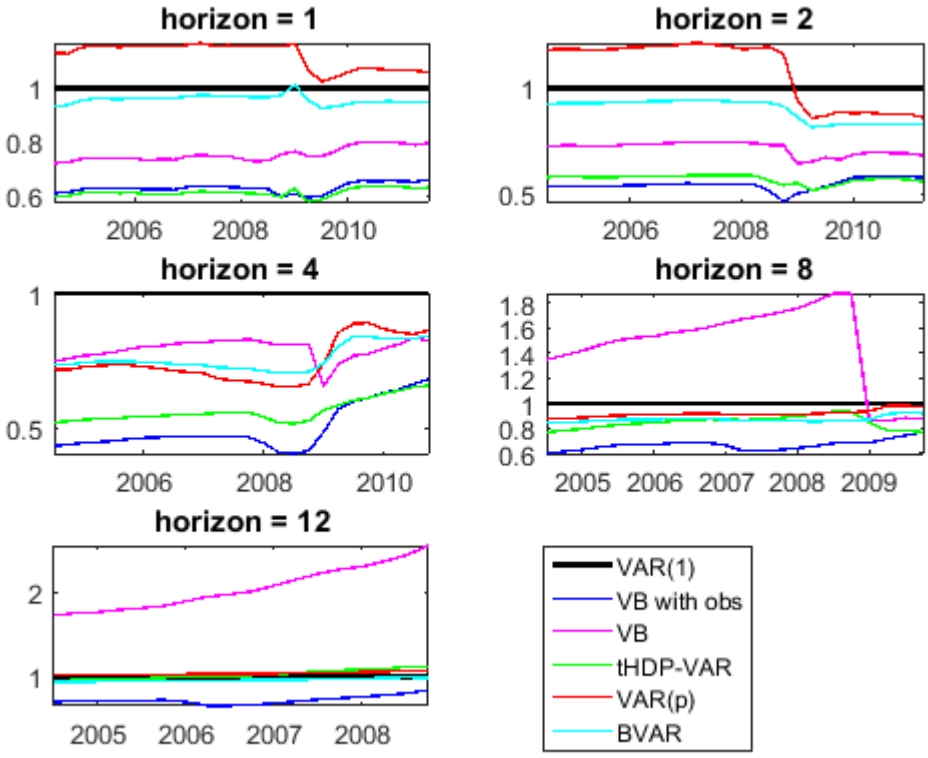

Figure 9. Relative RWMSFEs (VAR(1) is benchmark) for the Russian dataset depending on the starting point for calculation, 2 lags

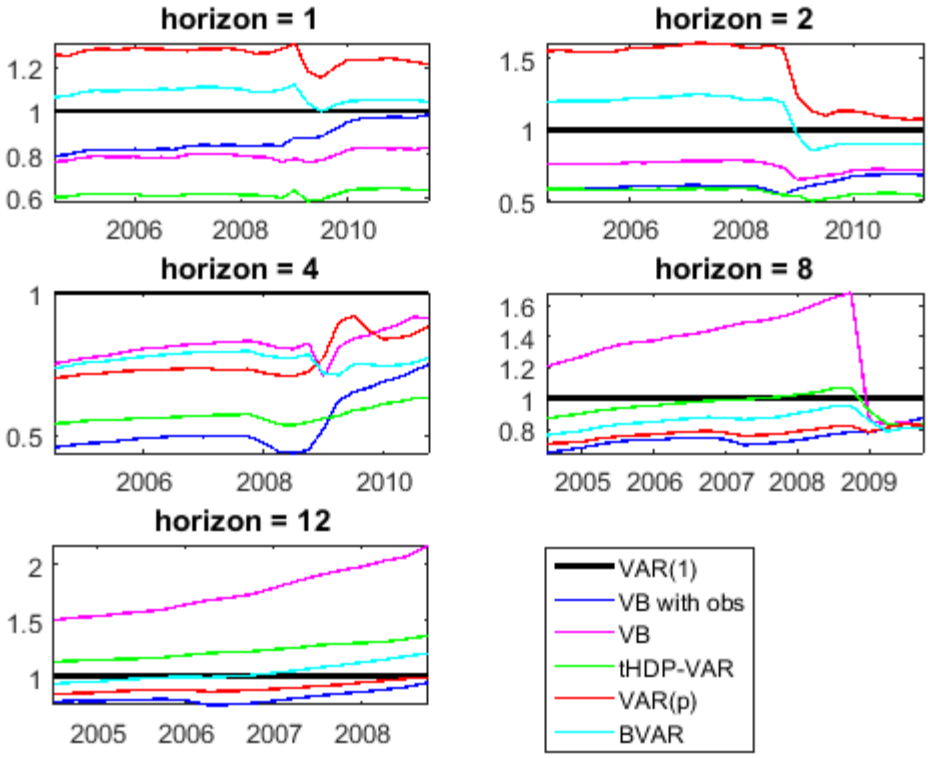

Figure 10. Relative RWMSFEs (VAR(1) is benchmark) for the Russian dataset depending on the starting point for calculation, 3 lags

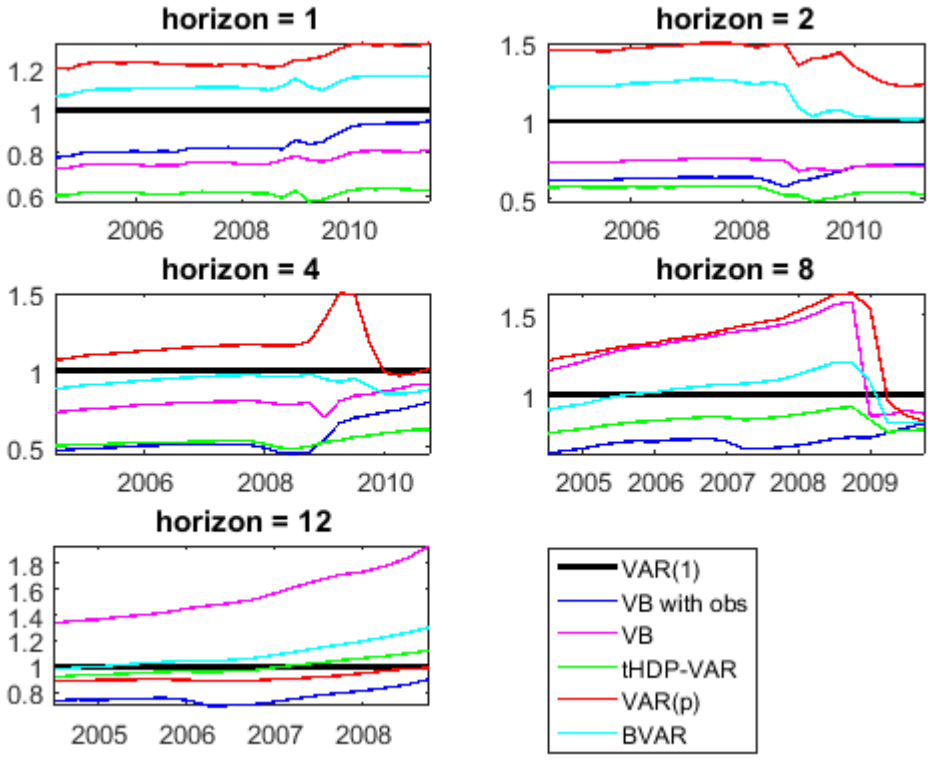

Figure 11. Relative RWMSFEs (VAR(1) is benchmark) for the Russian dataset depending on the starting point for calculation, 4 lags

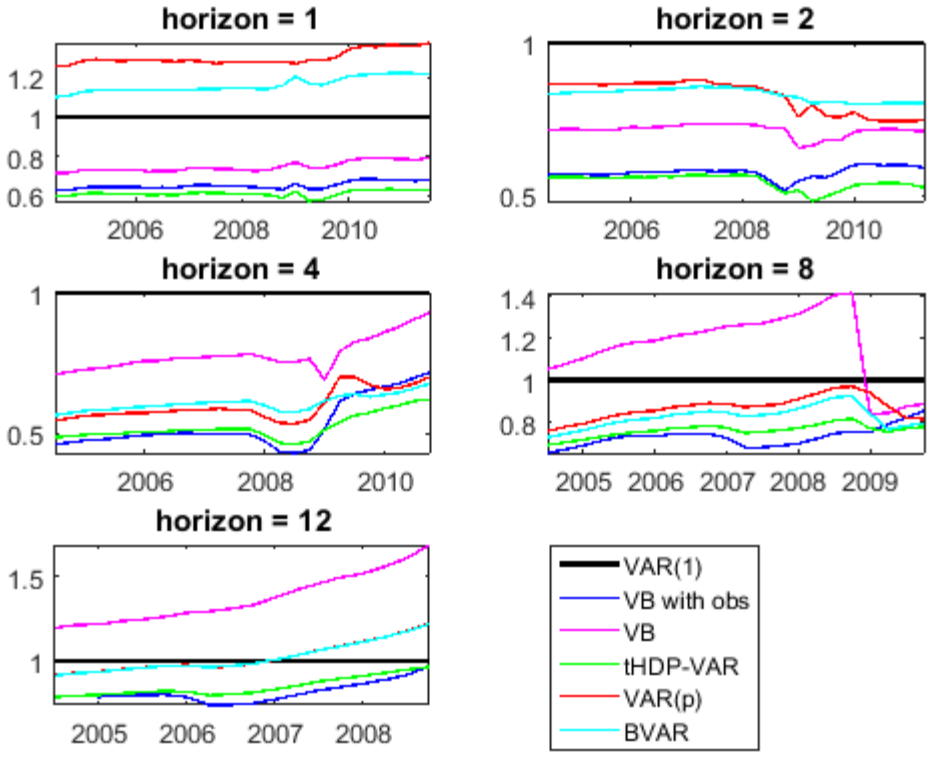

Figure 12. Relative RWMSFEs (VAR(1) is benchmark) for the Russian dataset depending on the starting point for calculation, 5 lags

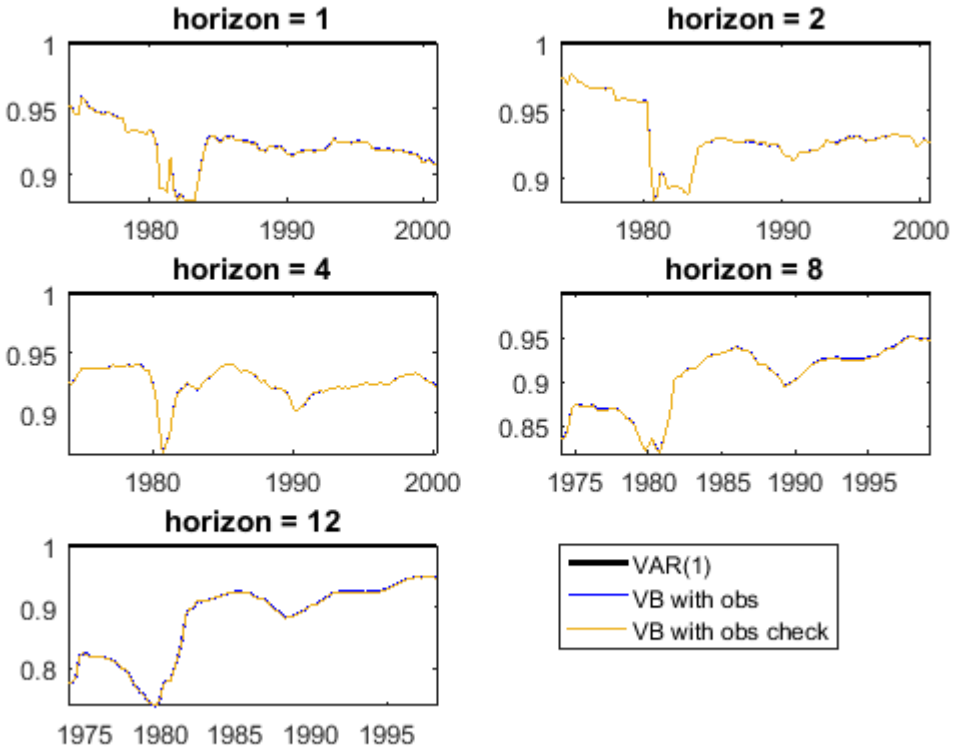

Figure 13. Relative RWMSFEs (VAR(1) is benchmark) for the US dataset depending on starting point for calculation, 3 lags

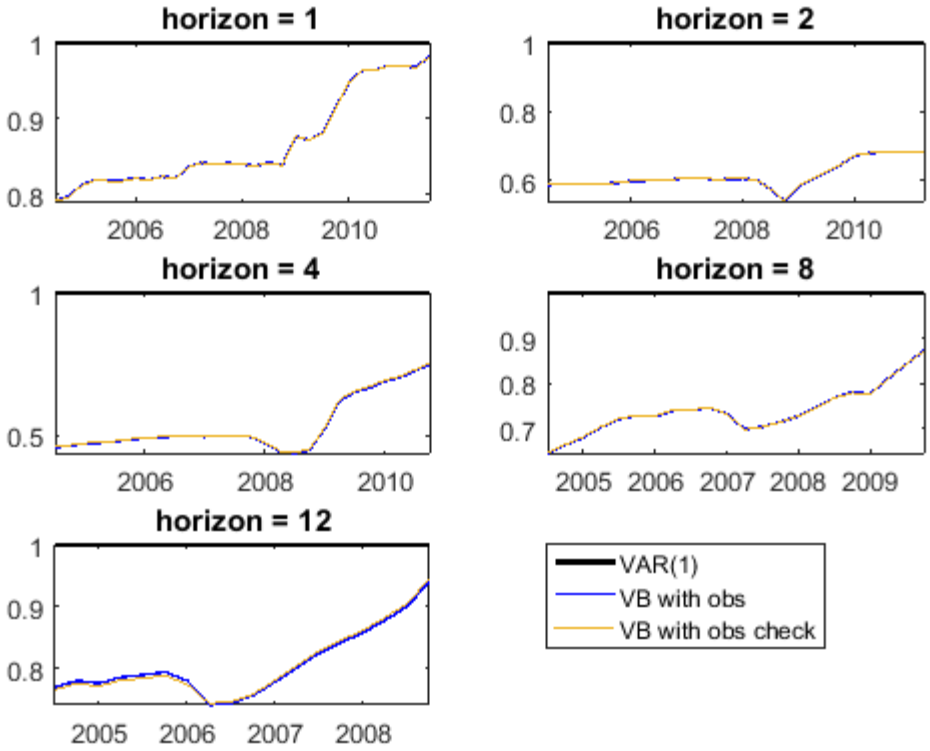

Figure 14. Relative RWMSFEs (VAR(1) is benchmark) for Russian dataset depending on starting point for calculation, 3 lags

Appendix B

The scheme of sampling can be written as follows:

For $n_{sim} = 1, \dots, N_{sim}$:

1. Sample auxiliary variables w, m, m_h and $\bar{m}$:

1.1. For each active state (i, j), define $J_{ij} = \{\tau | c_{\tau-1} = i, c_\tau = j\}$. Set $m_{ij} = 0$, $n = 0$, and for each $\tau \in J_{ij}$ sample

$$x \sim Ber\left(\frac{\alpha\beta'_j + \kappa I(i=j)}{n + \alpha\beta_j + \kappa I(i=j)}\right)$$

Increment n, and if $x = 1$ increment m_{ij}.

1.2. For each active j, sample:

$$w_j \sim Binomial\left(m_{jj}, \frac{\rho}{\rho + \beta'_j(1-\rho)}\right)$$

1.3. For each active j, set $m_{h,j} = 0$. For $j = c_0$, set $m_{h,j} = 1$.

1.4. Set

$$\bar{m} = m - diag(w) + diag(m_h)$$

2. Sample α, κ, γ:

2.1. Sample

$$\eta \sim Beta(\gamma + 1, \bar{m}_{..})$$

and for each active j, sample

$$r_j \sim Beta\big(\alpha + \kappa + 1, n_{j.}\big)$$

$$s_j \sim Binomial\left(1, \frac{n_{j.}}{\alpha + \kappa + n_{j.}}\right)$$

2.2. Sample

$$\gamma \sim \frac{a_\gamma + \bar{K} - 1}{\bar{m}_{..}\big(b_\gamma - log(\eta)\big)} Gamma\big(a_\gamma + \bar{K}, b_\gamma - log(\eta)\big) + \left(1 - \frac{a_\gamma + \bar{K} - 1}{\bar{m}_{..}\big(b_\gamma - log(\eta)\big)}\right) Gamma\big(a_\gamma + \bar{K} - 1, b_\gamma - log(\eta)\big)$$

and

$$\alpha + \kappa \sim Gamma\left(a_{\alpha+\kappa} + m_{..} - \sum s_j, b_{\alpha+\kappa} - \sum log\big(r_j\big)\right)$$

2.3. Sample

$$\rho \sim Beta\left(\sum w_j + a_\rho, m_{..} - \sum w_j + b_\rho\right)$$

3. For K active states, sample

$$\left(\pi_{01},\ldots,\pi_{0K},\sum_{k=K+1}^{\infty}\pi_{0k}\right)\sim Dir(\overline{m}_{.1},\ldots,\overline{m}_{.K},\gamma)$$

4. For K active states, sample

$$\pi_k=\left(\pi_{k1},\ldots,\pi_{kK},\sum_{l=K+1}^{\infty}\pi_{kl}\right)\sim Dir\left(\alpha\pi_{01}+n_{k1},\ldots,\alpha\pi_{0k}+\kappa+n_{kk},\ldots,\alpha\pi_{0K}+n_{kK},\alpha\sum_{l=K+1}^{\infty}\pi_{0l}\right)$$

5. Sample VAR hyperparameters (λ_H) using the adaptive MH algorithm from Roberts and Rosenthal (2009) with proposal density

$$q\left(\lambda_H'\middle|\lambda_H^{(i-1)}\right)$$

and acceptance rate

$$\alpha'=min\left(1,\frac{\prod_{k=1}^{K}p(\theta^k|\lambda_H')}{\prod_{k=1}^{K}p\left(\theta^k|\lambda_H^{(i-1)}\right)}\frac{p(\lambda_H')}{p\left(\lambda_H^{(i-1)}\right)}\frac{q\left(\lambda_H^{(i-1)}\middle|\lambda_H'\right)}{q\left(\lambda_H'\middle|\lambda_H^{(i-1)}\right)}\right)$$

For active states, sample θ^k with the accept-reject algorithm.

6. Sample auxiliary variables U_t for $t=0,\ldots,T$

$$U_t\sim I(t=0)Uniform\left(0,\pi_{0c_0}\right)+I(t>0)Uniform\left(0,\pi_{c_{t-1}c_t}\right)$$

7. Set $K'=K$ and check that for each t

$$U_0>\sum_{l=K'+1}^{\infty}\pi_{0l}\text{ and }U_t>\sum_{l=K'+1}^{\infty}\pi_{c_{t-1}l}$$

If it is not true, set $K'=K'+1$ and expand π_0 and π in the following way:

7.1. Set $\beta_{K'}=bb\sum_{l=K'}^{\infty}\pi_{0l}$ and $\pi_{0K'+1}=\frac{\pi_{0K'}}{bb}(1-bb)$, where $bb\sim Beta(1,\gamma)$.

7.2. For $k=1,\ldots,K'-1$, set $\pi_{kK'}=bb\sum_{l=K'}^{\infty}\pi_{kl}$ and $\pi_{kK'+1}=\frac{\pi_{0K'}}{bb}(1-bb)$, where $bb\sim Beta(\alpha\pi_{0K'},\alpha\pi_{0K'+1})$.

7.3. Sample

$$\pi_{K'}\sim Dir(\alpha\pi_{01},\ldots,\alpha\pi_{0K'}+\kappa,\alpha\pi_{0K'+1})$$

Repeat until convergence.

8. For non-active states, sample θ^k using the accept-reject algorithm.

9. Sample $c_{0:T}$ in the following way:

9.1. Set

$$p(c_0=k|U,\theta,\pi_0,\pi)\sim I(U_0<\pi_{0k})$$

$$p(c_1=k|U,\theta,\pi_0,\pi)\sim\sum_{i=1}^{K'}I(U_1<\pi_{ik})p(c_0=i|U,\theta,\pi_0,\pi)$$

9.2. For $t = 1, \dots, T$:

Set

$$p(c_t = k|y_{1:t}, U, \theta, \pi_0, \pi) \sim p(c_t = k|y_{1:t-1}, U, \theta, \pi_0, \pi) p\left(y_t|\theta_{c_t}, y_{1:t-1}\right)$$

If $t < T$, set

$$p(c_{t+1} = k|y_{1:t}, U, \theta, \pi_0, \pi) \sim \sum_{i=1}^{K'} I(U_{t+1} < \pi_{ik}) p(c_t = i|y_{1:t}, U, \theta, \pi_0, \pi)$$

9.3. Sample c_T from $p(c_T = k|y_{1:T}, \theta, \pi_0, \pi)$.

9.4. For $t = T - 1, \dots, 0$, sample c_t from

$$p(c_t = k|c_{t+1} = i, y_{1:T}, U, \theta, \pi_0, \pi) \sim p(c_t = k|y_{1:t}, U, \theta, \pi_0, \pi) I(U_{t+1} < \pi_{ki})$$

10. Remove empty states.

Appendix C

For variational approximation, we rewrite the system in the following form:

$$\beta'_k \sim Beta(1,\gamma) \qquad k = 1,2,\dots$$
$$\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l) \qquad k = 1,2,\dots$$
$$\pi_{i1},\dots,\pi_{iN},\dots \sim DP\left(\alpha+\kappa, \frac{\alpha\beta+\kappa I(i=j)}{\alpha+\kappa}\right) i = 0,1,2,\dots$$
$$p(c_t = k | c_{t-1} = k', \pi_{1*},\dots,\pi_{N*},\dots) = \pi_{k'k}$$
$$y_t \sim p\left(\theta^k, y_{t-1},\dots,y_{t-p}\right)$$

For brevity, we omit the dependence on hyperparameters. For simplicity of calculations (and for comparability with Stephenson and Raphael (2015)), we additionally assume that π_0 has a different distribution from that described earlier.

For approximation of the model, we follow Stephenson and Raphael (2015) in representing the approximated density as:

$$q(\beta',\pi,\theta,c) = q(\beta')q(\pi)q(\theta)q(c)$$

where β',π,θ,c are sets of variables with respective indexes. In this case, the objective function has the following form:

$$\begin{aligned} L &= E_q\left[log\left(p(y,\beta',\pi,\theta,c)\right) - log\left(q(\beta',\pi,\theta,c)\right)\right] \\ &= E_q\left[log\left(p(y|\theta,c)\right) + log\left(p(\theta)\right) - log\left(q(\theta)\right) + log\left(p(\beta'|\gamma)\right) - log\left(q(\beta')\right)\right. \\ &\left. + log\left(p(\pi|\beta',\alpha,\kappa)\right) - log\left(q(\pi)\right) + log\left(p(c|\pi)\right) - log\left(q(c)\right)\right] \end{aligned}$$

where E_q is an operator of expectation with respect to q, and y are the observations of the model $(y_1,\dots,y_T)$.

C.1. Optimizing $q(\pi)$

The objective function with respect to $q(\pi)$ is

$$L = const + E_q\left[log\left(p(\pi|\beta'_k,\alpha,\kappa)\right) - log\left(q(\pi)\right) + log\left(p(c|\pi)\right)\right]$$

The first-order derivative is

$$\frac{\partial L}{\partial q(\pi)} = E_{q(\beta')}\left[log\left(p(\pi|\beta',\alpha,\kappa)\right)\right] - log\left(q(\pi)\right) - 1 + E_{q(c)}\left[log\left(p(c|\pi)\right)\right]$$

Setting the first-order derivative equal to zero, we have

$$log\left(q(\pi)\right) = E_{q(\beta')}\left[log\left(p(\pi|\beta',\alpha,\kappa)\right)\right] - 1 + E_{q(c)}\left[log\left(p(c|\pi)\right)\right]$$

$$log\big(q(\pi)\big) = log(q(\pi_0) \dots q(\pi_K) \dots) = log\big(q(\pi_0)\big) + \cdots log\big(q(\pi_K)\big) + \cdots$$
$$= E_{q(\beta')}\big[log\big(p(\pi_0|\beta', \alpha, \kappa)\big)\big] + \cdots + E_{q(\beta')}\big[log\big(p(\pi_K|\beta', \alpha, \kappa)\big)\big] + \cdots - 1$$
$$+ E_{q(c)}\left[log\left(p\big(c_0|\pi_{c_0}\big)\right)\right] + \sum_{t=1}^{T} E_{q(c)}\left[log\left(p\big(c_t|\pi_{c_t}, c_{t-1}\big)\right)\right]$$

Finally, it is possible to get the closed form solution

$$q(\pi_0) = Dir(\hat{\vartheta}_0)$$
$$\hat{\vartheta}_{0k} = \alpha E_{q(\beta')}[\beta_k] + E_{q(c)}[I(c_0 = k)]$$
$$q(\pi_i) = Dir(\hat{\vartheta}_i) \qquad i = 1, 2, \dots$$
$$\hat{\vartheta}_{ik} = \alpha E_{q(\beta')}[\beta_k] + \kappa I(i = j) + \sum_{t=1}^{T} E_{q(c)}[I(c_{t-1} = i) I(c_t = k)]$$

C.2. Optimizing $q(\theta)$

The objective function with respect to $q(\theta)$ is

$$L = const + E_q\big[log\big(p(y|\theta, c)\big) + log\big(p(\theta)\big) - log\big(q(\theta)\big)\big]$$

The first-order derivative is

$$\frac{\partial L}{\partial q(\theta)} = E_{q(c)}\big[log\big(p(y|\theta, c)\big)\big] + log\big(p(\theta)\big) - 1 - log\big(q(\theta)\big) = 0$$

The logarithm of the kth coordinate of the distribution of $q(\theta)$ is

$$log\Big(q(\theta^k)\Big) = const + log\Big(p(\theta^k)\Big) + E_{q(c)}\left[\sum_{t=1}^{T} I(c_t = k)\, log\Big(p(y_t|\theta^k)\Big)\right]$$

Taking into account the fact that we are building the approximation for the tHDP-VAR model, we get:

$$log\left(p\big(b, A_1, \dots, A_p, \Sigma|\lambda_H\big)\right) =$$
$$= const - \frac{1}{2} log|\Sigma \otimes \Omega| - \frac{1}{2}(\beta - \beta_0)^T (\Sigma \otimes \Omega)^{-1} (\beta - \beta_0) - \frac{1}{2} log|\Sigma|$$
$$- \frac{1}{2}(Y^{++} - X^{++}\beta)^T \Sigma^{-1} (Y^{++} - X^{++}\beta) + \frac{d}{2} log|\Psi| - \frac{N + d + 1}{2} log|\Sigma|$$
$$- \frac{1}{2} tr(\Psi \Sigma^{-1})$$

where $\beta = vec([b, A_1, \dots, A_p]')$, β_0 is the mean of the Minnesota prior coefficients, Ω is the matrix of hyperparameters which is responsible for the variance of the Minnesota prior coefficients, and Y^{++} and X^{++} are dummy observation variables.

$$log\Big(p(y|\theta^k)\Big) = -\frac{1}{2} log|\Sigma_k \otimes I_T| - \frac{1}{2}(Y - X\beta_k)^T (\Sigma_k \otimes I_T)^{-1} (Y - X\beta_k)$$

where $y = [y_1, \dots, y_T]'$, $Y = vec(y)$, $x_t = [1, y'_{t-1}, \dots, y'_{t-p}]$, $x_t = I_N \otimes x'_t$, $x = [x_1, \dots, x_T]'$, $x_t = I_N \otimes x$.

$$
\begin{aligned}
log\big(q(\theta_k)\big) = const &- \frac{1}{2} log|\Sigma_k \otimes \Omega| - \frac{1}{2}(\beta_k - \beta_0)^T(\Sigma \otimes \Omega)^{-1}(\beta_k - \beta_0) - \frac{1}{2} log|\Sigma_k| \\
&- \frac{1}{2}(Y^{++} - X^{++}\beta_k)^T \Sigma_k^{-1}(Y^{++} - X^{++}\beta_k) + \frac{d}{2} log|\Psi| - \frac{N+d+1}{2} log|\Sigma_k| \\
&- \frac{1}{2} tr(\Psi \Sigma_k^{-1}) - \frac{1}{2} E_{q(c)}\left[\sum_{t=1}^{T} I(c_t = k)\, log|\Sigma_k|\right] \\
&- \frac{1}{2} E_{q(c)}\left[\sum_{t=1}^{T} I(c_t = k)(y_t - x_t\beta_k)^T {\Sigma_k}^{-1}(y_t - x_t\beta_k)\right] \\
&= const - \frac{Np + 1 + N + d + 1 + 1 + \sum_{t=1}^{T} I(c_t = k)}{2} log|\Sigma_k| - \frac{N}{2} log|\Omega| \\
&+ \frac{d}{2} log|\Psi| - \frac{1}{2} tr(\Psi \Sigma_k^{-1}) \\
&- \frac{1}{2}\Big((\beta_k - \beta_0)^T(\Sigma_k \otimes \Omega)^{-1}(\beta_k - \beta_0) + (Y^{++} - X^{++}\beta_k)^T \Sigma_k^{-1}(Y^{++} - X^{++}\beta_k) \\
&+ (Y - X\beta_k)^T\big(\Sigma_k^{-1} \otimes E_{q(c)}[I_T(c_t = k)]\big)(Y - X\beta_k)\Big)
\end{aligned}
$$

This can be rewritten in the following form:

$$q(\beta_k|\Sigma_k) = N\left(vec(\tilde{b}_k); \Sigma_k \otimes \left(\Omega^{-1} + x^T E_{q(c)}[I_T(c_t = k)]x + (x^{++})^T x^{++}\right)^{-1}\right)$$

$$
\begin{aligned}
q(\Sigma_k) = IW\Bigg(&\Psi + \left(\tilde{b}_k - \beta_0^m\right)^T \Omega^{-1}\left(\tilde{b}_k - \beta_0^m\right) + \tilde{e}^{++T}\tilde{e}^{++} + \tilde{e}^T E_{q(c)}[I_T(c_t = k)]\tilde{e}; d + 1 \\
&+ \sum_{t=1}^{T} I(c_t = k)\Bigg)
\end{aligned}
$$

where $\tilde{b}_k = \left(\Omega^{-1} + x^T E_{q(c)}[I_T(c_t = k)]x + (x^{++})^T x^{++}\right)^{-1}\left(\Omega^{-1}\beta_0^m + (x^{++})^T y^{++} + x^T E_{q(c)}[I_T(c_t = k)]y\right)$.

Furthermore, we will use

$$\Psi_k = \Psi + \left(\tilde{b}_k - \beta_0^m\right)^T \Omega^{-1}\left(\tilde{b}_k - \beta_0^m\right) + \tilde{e}^{++T}\tilde{e}^{++} + \tilde{e}^T E_{q(c)}[I_T(c_t = k)]\tilde{e}$$

$$d_k = d + 1 + \sum_{t=1}^{T} I(c_t = k)$$

C.3. Optimizing $q(c)$

The objective function with respect to $q(c)$ is

$$L = const + E_q\big[log\big(p(y|\theta, c)\big) + log\big(p(c|\pi)\big) - log\big(q(c)\big)\big]$$

The first-order derivative is

$$\frac{\partial L}{\partial q(c)} = E_{q(\theta)}\big[log\big(p(y|\theta,c)\big)\big] + E_{q(\pi)}\big[log\big(p(c|\pi)\big)\big] - log\big(q(c)\big) - 1$$

Following the traditional variational approximation of DP, we assume only K nonzero regimes for the variable c. The probability of occurrence of the sequence c is

$$\begin{aligned} log\big(q(c)\big) &= const + E_{q(\theta)}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} I(c_t = k)\, log\Big(p(y_t|\theta^k)\Big)\right] \\ &+ E_{q(\pi)}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{k=1}^{K} I(c_t = k)I(c_{t-1} = i)\, log(\pi_{ik})\right] \\ &+ E_{q(\pi)}\left[\sum_{k=1}^{K} I(c_0 = k)\, log(\pi_{0k})\right] \\ &= const + \sum_{t=1}^{T}\sum_{k=1}^{K} I(c_t = k)E_{q(\theta)}\Big[log\Big(p(y_t|\theta^k)\Big)\Big] \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{k=1}^{K} I(c_t = k)I(c_{t-1} = i)E_{q(\pi)}[log(\pi_{ik})] + \sum_{k=1}^{K} I(c_0 = k)E_{q(\pi)}[log(\pi_{0k})] \end{aligned}$$

Similarly to Beal (2003), we use the belief propagation algorithm (Yedidia, Freeman & Weiss, 2001) for calculating $E_{q(c)}[I(c_{t-1} = i)I(c_t = k)]$ and $E_{q(c)}[\sum_{t=1}^{T} I(c_t = k)]$, which are needed for $q(\theta)$.

Forward and backward messages can be expressed as:

$$m_{0,1}(c_1 = k) = \sum_{i=1}^{K} \tilde{\pi}_{0i}\tilde{\pi}_{ik}$$

$$m_{t-1,t}(c_t = k) = \sum_{i=1}^{K} \tilde{p}(y_{t-1}|c_{t-1} = i)\tilde{\pi}_{ik}m_{t-2,t-1}(c_{t-1} = i) \qquad t = 2, \dots, T$$

$$m_{T,T-1} = (c_{T-1} = k) = \sum_{i=1}^{K} \tilde{p}(y_T|c_T = i)\tilde{\pi}_{ki}$$

$$m_{t,t-1}(c_{t-1} = k) = \sum_{i=1}^{K} \tilde{p}(y_t|c_t = i)\tilde{\pi}_{ki}m_{t+1,t}(c_t = i) \qquad t = T-1, \dots, 1$$

The required densities are

$$q(c_0 = k) \propto \tilde{\pi}_{0k}m_{1,0}(c_0 = k)$$

$$q(c_t = k) \propto \tilde{p}(y_t|c_t = k)m_{t-1,t}(c_t = k)m_{t+1,t}(c_t = k) \qquad t = 1, \dots, T-1$$

$$q(c_T = k) \propto \tilde{p}(y_T|c_T = k)m_{T-1,T}(c_T = k)$$

$$q(c_0 = k, c_1 = i) \propto \tilde{\pi}_{0k}\tilde{p}(y_1|c_1 = i)\tilde{\pi}_{ki}m_{2,1}(c_1 = i)$$

$$q(c_{t-1} = k, c_t = i) \propto \tilde{p}(y_{t-1}|c_{t-1} = i)\tilde{p}(y_t|c_t = i)\tilde{\pi}_{ki}m_{t+1,t}(c_t = i)m_{t-2,t-1}(c_{t-1} = k)$$

$$t = 2, \dots, T-1$$

$$q(c_{T-1} = k, c_T = i) \propto \tilde{p}(y_{T-1}|c_{T-1} = i)\tilde{p}(y_T|c_T = i)\tilde{\pi}_{ki}m_{T-2,T-1}(c_{T-1} = k)$$

where

$$\tilde{\pi}_{ik} = e^{E_{q(\pi)}[log(\pi_{ik})]} = e^{E_{q(\pi)}[log(\pi_{ik})]} = e^{\psi(\hat{\vartheta}_{ik}) - \sum_{m=1}^{K+1}\psi(\hat{\vartheta}_{im})}$$

$$\tilde{p}(y_t|c_t = k) = e^{E_{q(\theta)}\left[log\left(p(y_t|\theta^k)\right)\right]}$$

As Bishop (2006) does, we calculate:

$$E_{q(\theta)}\left[log\left(p(y_t|\theta^k)\right)\right] = E_{q(\theta_k)}\left[log\left(p(y_t|\theta^k)\right)\right]$$

$$= -\frac{1}{2}E_{q(\theta^k)}[N\,log(2\pi) + log|\Sigma_k| + (Y_t - X_t\beta_k)^T\Sigma_k^{-1}(Y_t - X_t\beta_k)]$$

$$E_{q(\theta^k)}[log|\Sigma_k|] = \int log|\Sigma_k|\, IW(\Sigma_k|\,\Psi_k, d_k)d\Sigma_k = -\int log|\Sigma_k|\, W(\Sigma_k|\,\Psi_k^{-1}, d_k)d\Sigma_k$$

$$= -\left(\sum_{i=1}^{N}\psi\left(\frac{d_k + 1 - i}{2}\right) + N\,log\,2 - log|\Psi_k|\right)$$

$$E_{q(\theta^k)}[(Y_t - X_t\beta_k)^T\Sigma_k^{-1}(Y_t - X_t\beta_k)] = Y_t^TY_t - 2E_{q(\theta^k)}[Y_t^T\Sigma_k^{-1}X_t\beta_k] + E_{q(\theta^k)}[\beta_k^TX_t^T\Sigma_k^{-1}X_t\beta_k]$$

$$= d_kY_t^T\Psi_k^{-1}Y_t - 2Y_t^TE_{q(\theta^k)}[\Sigma_k^{-1}]X_t\tilde{b}_k$$

$$+ E_{q(\theta^k)}[\beta_k^T(I_N \otimes x_t^T)(\Sigma_k^{-1} \otimes 1)(I_N \otimes x_t)\beta_k]$$

$$= d_kY_t^T\Psi_k^{-1}Y_t - 2d_kY_t^T\Psi_k^{-1}X_t\tilde{b}_k + E_{q(\theta^k)}[\beta_k^T(\Sigma_k^{-1} \otimes x_t^Tx_t)\beta_k]$$

To calculate the last term, we introduce the variable:

$$\beta_k^{ch} = \left(\Sigma_k^{-\frac{1}{2}} \otimes \left(\Omega^{-1} + x^TE_{q(c)}[I_T(c_t = k)]x + (x^{++})^Tx^{++}\right)^{\frac{1}{2}}\right)\left(\beta_k - vec\left(\tilde{b}_k\right)\right)$$

$$= (L \otimes M)\left(\beta_k - vec\left(\tilde{b}_k\right)\right) \sim N\left(0, I_{N(Np+1)}\right)$$

then

$$E_{q(\theta^k)}[\beta_k^T(\Sigma_k^{-1} \otimes x_t^Tx_t)\beta_k]$$

$$= d_k\tilde{b}_k^TX_t^T\Psi_k^{-1}X_t\tilde{b}_k + E_{q(\theta_k)}\left[\left(\beta_k^{ch}\right)^T(L^T \otimes M^T)^{-1}(\Sigma_k^{-1} \otimes x_t^Tx_t)(L \otimes M)^{-1}\beta_k^{ch}\right]$$

$$= d_k\tilde{b}_k^TX_t^T\Psi_k^{-1}X_t\tilde{b}_k + E_{q(\theta_k)}\left[\left(\beta_k^{ch}\right)^T(I_N \otimes (M^T)^{-1}x_t^Tx_t(M)^{-1})\,\beta_k^{ch}\right]$$

$$= d_k\tilde{b}_k^TX_t^T\Psi_k^{-1}X_t\tilde{b}_k + N\,tr((M^T)^{-1}x_t^Tx_t(M)^{-1})$$

C.4. Optimizing $q(\beta')$

The objective function with respect to $q(\beta')$ is

$$L = const + E_q\left[log\left(p(\beta'|\gamma)\right) - log\left(q(\beta')\right) + log\left(p(\pi|\beta', \alpha, \kappa)\right)\right]$$

Following Hughes, Kim and Sudderth (2015) and Stephenson and Raphael (2015), we assume $q(\beta')$ of the form:

$$q(\beta') = \prod_{k=1}^{\infty} Beta(\beta'_k|\hat{\rho}_k\hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k)$$

The objective function is rewritten as:

$$L = const + E_q\left[log\left(p(\beta'|\gamma)\right) - log\left(q(\beta')\right) + log\left(p(\pi|\beta', \alpha, \kappa)\right)\right] = const + \sum_{k=1}^{K}\Big((\gamma - 1)\, log(1-\beta'_k) - log\left(B(1,\gamma)\right) - (\hat{\rho}_k\hat{\omega}_k - 1)\, log\,\beta'_k - \left((1-\hat{\rho}_k)\hat{\omega}_k - 1\right) log(1-\beta'_k) + log\left(B(\hat{\rho}_k\hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k)\right)\Big) Beta(\beta'_k|\hat{\rho}_k\hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k)d\beta'_k + E_q[\sum_{i=0}^{K}(-log(D(\alpha\beta + \kappa I(i = k))) + (\alpha\beta_{k>K} - 1)\, log(\pi_{ik>K}) + \sum_{k=1}^{K}(\alpha\beta_k + \kappa I(i=k) - 1)\, log(\pi_{ik}))] \geq const + \sum_{k=1}^{K}\Big(-log\left(B(1,\gamma)\right) + log\left(B(\hat{\rho}_k\hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k)\right) + (\gamma - (1-\hat{\rho}_k)\hat{\omega}_k)\Big(\psi\left((1-\hat{\rho}_k)\hat{\omega}_k\right) - \psi(\hat{\omega}_k)\Big) + (1-\hat{\rho}_k\hat{\omega}_k)\left(\psi(\hat{\rho}_k\hat{\omega}_k) - \psi(\hat{\omega}_k)\right)\Big) + E_q[(K^2 + K)\, log(\alpha) + K(log(\kappa) - log(\alpha + \kappa)) + (log(\alpha+\kappa) - log(\kappa))\sum_{k=1}^{K}\beta_k + log(\beta_{K+1}) + K\sum_{k=1}^{K+1} log(\beta_k)] + \alpha\sum_{k=1}^{K} E_q[\beta_k]\left(\sum_{i=0}^{K} E_q[log(\pi_{ik})]\right) + \sum_{k=1}^{K}\left(\sum_{i=0}^{K}(\kappa I(i=k) - 1)E_q[log(\pi_{ik})]\right) + \sum_{i=0}^{K}\Big(\left(\alpha E_q[\beta_{k>K}] - 1\right)E_q[log(\pi_{ik>K})]\Big) = const + \sum_{k=1}^{K}\Big(log\left(B(\hat{\rho}_k\hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k)\right) + (K + 1 - \hat{\rho}_k\hat{\omega}_k)\left(\psi(\hat{\rho}_k\hat{\omega}_k) - \psi(\hat{\omega}_k)\right) + (K(K+1-k) + 1 + \gamma - (1-\hat{\rho}_k)\hat{\omega}_k)\Big(\psi\left((1-\hat{\rho}_k)\hat{\omega}_k\right) - \psi(\hat{\omega}_k)\Big)\Big) + \sum_{k=1}^{K} E_q[\beta_k]\left(log(\alpha+\kappa) - log(\kappa) + \alpha\sum_{i=0}^{K} E_q[log(\pi_{ik})]\right) + \alpha E_q[\beta_{k>K}]\sum_{i=0}^{K}\left(E_q[log(\pi_{ik>K})]\right)$$

To find the optimum, we introduce a change of variables ($0 < \hat{\rho} < 1, \hat{\omega} > 0$) to make the problem unconstrained:

$$\hat{\omega}_{sub} = log\,\hat{\omega}$$

$$\hat{\rho}_{sub} = -\,log\left(\frac{1}{\hat{\rho}} - 1\right)$$

The objective function has the form:

$$L' = \sum_{k=1}^{K}\Big(log\big(B(\hat{\rho}_k\hat{\omega}_k,(1-\hat{\rho}_k)\hat{\omega}_k)\big) + (K+1-\hat{\rho}_k\hat{\omega}_k)\big(\psi(\hat{\rho}_k\hat{\omega}_k) - \psi(\hat{\omega}_k)\big) + (K(K+1-k)+1+\gamma-(1-\hat{\rho}_k)\hat{\omega}_k)\Big(\psi\big((1-\hat{\rho}_k)\hat{\omega}_k\big) - \psi(\hat{\omega}_k)\Big)\Big) + \sum_{k=1}^{K} E_q[\beta_k]\big(log(\alpha+\kappa) - log(\kappa) + \alpha\sum_{i=0}^{K} E_q[log(\pi_{ik})]\big) + \alpha E_q[\beta_{k>K}]\sum_{i=0}^{K}\big(E_q[log(\pi_{ik>K})]\big)$$

The first-order derivative is

$$\frac{\partial L'}{\partial \hat{\omega}_{sub,k}} = \frac{\partial L'}{\partial \hat{\omega}_k}\frac{\partial \hat{\omega}_k}{\partial \hat{\omega}_{sub,k}}$$

$$= \Big((K+1-\hat{\rho}_k\hat{\omega}_k)\big(\hat{\rho}_k\psi'(\hat{\rho}_k\hat{\omega}_k) - \psi'(\hat{\omega}_k)\big) + (K(K+1-k)+1+\gamma-(1-\hat{\rho}_k)\hat{\omega}_k)\Big((1-\hat{\rho}_k)\psi'\big((1-\hat{\rho}_k)\hat{\omega}_k\big) - \psi'(\hat{\omega}_k)\Big)\Big)\hat{\omega}_k$$

$$\frac{\partial L'}{\partial \hat{\rho}_{sub,k}} = \frac{\partial L'}{\partial \hat{\rho}_k}\frac{\partial \hat{\rho}_k}{\partial \hat{\rho}_{sub,k}}$$

$$= \Bigg(\hat{\omega}_k(K+1-\hat{\rho}_k\hat{\omega}_k)\psi'(\hat{\rho}_k\hat{\omega}_k) - \hat{\omega}_k(K(K+1-k)+1+\gamma-(1-\hat{\rho}_k)\hat{\omega}_k)\psi'\big((1-\hat{\rho}_k)\hat{\omega}_k\big) + \sum_{m=1}^{K}\Delta_{km}\Bigg(log(\alpha+\kappa) - log(\kappa) + \alpha\sum_{i=0}^{K}E_q[log(\pi_{im})]\Bigg) + \alpha\Delta_{k,K+1}\sum_{i=0}^{K}\big(E_q[log(\pi_{ik>K})]\big)\Bigg)\hat{\rho}_k(1-\hat{\rho}_k)$$

where

$$\Delta_{km} = \begin{cases} -\dfrac{1}{1-\hat{\rho}_k}E_q[\beta_m] & k < m \\ \dfrac{1}{\hat{\rho}_k}E_q[\beta_m] & k = m \\ 0 & k > m \end{cases}$$

C.5. Choosing hyperparameters

We also optimize λ_H every 100 iterations by introducing a loss function, which is determined by prior distributions.

In addition, the initial number of nonzero components K in $q(c)$ is chosen to be as large as possible. If there exist components with a sum of probabilities less than 0.0001 after convergence, we remove these components and restart the algorithm.