Ter Steege, Lucas

**Working Paper**

# Variational inference for Bayesian panel VAR models

ECB Working Paper, No. 2991

**Provided in Cooperation with:**
European Central Bank (ECB)

*Suggested Citation:* Ter Steege, Lucas (2024) : Variational inference for Bayesian panel VAR models, ECB Working Paper, No. 2991, ISBN 978-92-899-6881-2, European Central Bank (ECB), Frankfurt a. M.,
https://doi.org/10.2866/9979266

This Version is available at:
https://hdl.handle.net/10419/311157

EUROPEAN CENTRAL BANK
EUROSYSTEM

# Working Paper Series

Lucas ter Steege

## Variational inference for Bayesian panel VAR models

No 2991

**Abstract**

We study the application of approximate mean field variational inference algorithms to Bayesian panel VAR models in which an exchangeable prior is placed on the dynamic parameters and the residuals follow either a Gaussian or a Student-t distribution. This reduces the estimation time of possibly several hours using conventional MCMC methods to less than a minute using variational inference algorithms. Next to considerable speed improvements, our results show that the approximations accurately capture the dynamic effects of macroeconomic shocks as well as overall parameter uncertainty. The application with Student-t residuals shows that it is computationally easy to include the COVID-19 observations in Bayesian panel VARs, thus offering a fast way to estimate such models.

**Keywords:** Variational Bayes, Panel-VAR, Student-t distribution
**JEL Codes: C18, C32, C33**

## Non-Technical Summary

In this paper we study mean field variational inference (MFVI) algorithms in the context of Bayesian panel VAR models (BPVAR) in which an exchangeable prior is placed on the model's dynamic parameters. These models are attractive for policy analyses and forecasting exercises, especially if the time series dimension of the data is short, because one can infer model dynamics from a richer set of observations across units and increase the effective sample size. However, estimating such models via conventional sampling methods that numerically explore the parameter space is computationally expensive. The aim of this paper is to apply algorithms relying on approximate MVFI techniques that considerably reduce the estimation time. The speed gains result from the fact that MFVI turns the estimation problem into a function maximization problem which is deterministic, thus avoiding the need for stochastic sampling schemes.

This benefit does, however, come at the cost of deliberately ignoring certain posterior correlations among parameters. The joint posterior over the model's parameters is not of a known form, which is why conventionally sampling methods are used to characterize it. MFVI methods, on the other hand, assume a factorization of this posterior distribution that breaks it up into several independent factors, each of which is easy to characterize. Once a given factorization has been assumed, the individual factors are characterized by deterministic sufficient statistics only, and a fast iterative scheme can be employed to find these statistics. Since this convenience sacrifices potentially important features of the true posterior distribution, our aim to study how accurate, and therefore useful, the resulting approximation is.

We present two applications to test the efficacy of MFVI methods to estimate BPVAR models. The first is a replication of the study by Jarociński (2010) in which the VAR residuals are Gaussian, and the model is used to investigate the responses across countries to monetary policy shocks. In the second application the residuals instead follow a Student-t distribution which has heavier tails. This extension generally allows for more robust inference, and is further particularly relevant when dealing with the COVID-19 observations because heavy-tailed residuals can safeguard against parameter bias due to outliers. In both applications we find MVFI methods to produce results almost as accurate as conventional sampling algorithms at a fraction of the computational time. More specifically, both the dynamic responses to shocks as well as the surrounding uncertainty are well captured by the approximations.

We conclude that the usefulness of MVFI approximations, previously established for

conventional VAR models, also extend to more complex time series models. This means that a convenient alternative to computationally burdensome sampling methods is available and could, for instance, fruitfully be applied to recursive forecasting exercises or extensive model testing.

# 1 Introduction

Variational inference (VI) methods are becoming more popular in the VAR model literature as an alternative to conventional Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling. The latter are typically easy to implement as conditional posterior distributions take simple forms, and given a sufficient number of draws, one can characterize the posterior distribution of the models parameters to any desired degree of accuracy. This, however, can become time consuming depending on the convergence properties of the sampler. VI methods on the other hand, especially those based on mean field approximations (MFVI), offer fast algorithms to estimate Bayesian models which have been found to provide results of comparable accuracy to sampling methods. The gain in speed provided by MFVI approximations stems from the fact that these involve deterministic function maximization instead of stochastic sampling. Intuitively, the goal is to replace an intractable posterior distribution with a tractable approximating distribution which is chosen to be close to the true posterior as measured by the Kullback-Leibler divergence.

In this article we apply MFVI methods to Bayesian Panel VAR models with an exchangeable prior and possibly non-Gaussian error terms. In this setup, MCMC methods are cumbersome because the number of parameters not only grows in the number of variables, but also in the number of units included in the panel. Furthermore, there typically exist strong dependencies between unit specific parameters and parameters common to all units, leading to high correlation between subsequent MCMC draws. A large number of draws is then necessary to characterize the posterior accurately, making estimation a numerically daunting task. Our paper investigates the efficacy of MFVI approximations in this setup as a fast estimation alternative.

We find that the MFVI approximation accurately captures both the magnitudes and dynamic properties of the estimated systems as measured in terms of impulse response functions. In addition, error bands constructed under this approximation have similar coverage compared to the MCMC results. At the same time, MFVI estimation takes a fraction of the time necessary for MCMC inference. Therefore, one can reap considerable speed gains while still quantifying estimation uncertainty accurately. This is quiet important because mean field approximations may severely underestimate posterior variances, thus leading to incorrect conclusions from the estimation. Another contribution of our paper is the application of MFVI methods to t-distributed errors in these models. As shown in Hartwig (2022), Lenza and Primiceri (2022) and Bobeica and Hartwig (2023), including the Covid-19 observations in standard Gaussian VAR models can lead to

severely distorted inference, and that heavy-tailed errors can safeguard against this. This can be easily handled using a simple approximation that delivers posterior tractability for MFVI and leads to accurate results also in the extended model.

Relating to the literature, earlier work by Hajargasht (2019) showed that for standard macroeconomic VAR models, VI delivers very accurate results. Recent articles by Chan and Yu (2022), Gefang et al. (2023) and Korobilis and Schröder (2024) also find VI to perform very well both in terms of speed and accuracy as alluded to above, making VI a competitive alternative to MCMC methods. This paper is also related to the literature on VI methods for linear and generalized mixed models such as Lee and Wand (2016), Christmas (2014), Nolan et al. (2020), Menictas et al. (2022), and Hughes et al. (2023). Specifically, the computational insights developed in these papers are directly applicable to the model studied here, leading to an efficient implementation of the MFVI approximation.

The remaining of this article is organized as follows. Section 2 outlines the general model and describes the approximation to the true posterior. Section 3 contains two applications. The first is the original model considered in Jarociński (2010) to show that the proposed method accurately captures the true distribution when errors are normally distributed. The second application considers a dataset that includes the Covid-19 observations and relaxes the normality assumption. Section 4 offers concluding considerations, and the appendix contains details on estimation methods.

## 2 Model setup

The model we consider is the hierarchical panel VAR model considered in Jarociński (2010), extended to include departure from Gaussian errors as outlined in Geweke (1993) and Chan (2020):

$$y_{c,t} = \sum_{l=1}^{L} B_{c,l} y_{c,t-l} + \Delta_c w_t + \Gamma_c z_{c,t} + u_{c,t} \tag{1}$$

$$u_{c,t} \sim \mathcal{N}(0, \omega_{c,t}\Sigma_c) \tag{2}$$

$$\omega_{c,t} \sim \mathcal{IG}\left(\frac{\nu_c}{2}, \frac{\nu_c}{2}\right) \tag{3}$$

where $y_{c,t}$ is a $N \times 1$ vector of endogenous variables for country $c$, $w_t$ is a $W \times 1$ vector of exogenous variables common to all countries, and $z_{c,t}$ contains country specific exogenous variables, including possibly a constant. As shown in Geweke (1993), this model is equivalent to assuming that the error terms follow a multivariate t-distribution with $\nu_c$ degrees

of freedom. We assume that these vary across units, because in the recent use of this error structure in the macroeconomic time series literature, heavy-tailed errors have been used to stabilized inference in the presence of the Covid-19 observations, which likely did not have a homogeneous impact across countries.

To facilitate the derivations, two convenient and equivalent formulations of the above model are usually used. In the first we can stack the equations in the VAR into matrices as:

$$Y_c = X_c B_c + Z_c \Gamma_c + U_c \tag{4}$$

so that $Y_c$ and $U_c$ are $T \times N$ matrices, and $X_c$ is a $T \times K$ that contains lagged endogenous and common exogenous variables, where $K = NL + W$. $\Gamma_c$ has dimensions $T \times N_{z_c}$ and stacks the country specific exogenous variables.

The second representation vectorizes the data matrices above, meaning that we stack the columns of each matrix below each other to obtain:

$$y_c = (I_N \otimes X_c)\beta_c + (I_N \otimes Z_c)\gamma_c + \text{vec}(U_c) \tag{5}$$

$$\text{vec}(U_c) \sim \mathcal{N}(0, \Sigma_c \otimes \Omega_c) \tag{6}$$

with $y_c = \text{vec}(Y_c)$, $\beta_c = \text{vec}(B_c)$, and $\gamma_c = \text{vec}(\Gamma_c)$. $\Omega_c$ is the $T \times T$ matrix with the $\omega_{c,t}$ on the diagonal and zeros elsewhere.

The first formulation is the most convenient when it comes to deriving conditional posterior distributions of the error covariance matrices in VAR models, since the likelihood in this case, conditional on other parameters, has a similar structure as the density of the Inverse-Wishart distribution, which is also the common prior distribution for such parameters. The second representation of the model, instead, is more convenient to derive conditional posterior distributions for the remaining parameters. We also make use of both representations when we derive the approximation to be discussed shortly.

## 2.1 Prior distributions

For the parameters of the model the following priors are specified. First, for the dynamic coefficients an exchangeable prior is used:

$$\beta_c \sim \mathcal{N}(\beta_0, \lambda\Lambda_c) \tag{7}$$

$$\beta_0 \sim \mathcal{N}(\mu_{\beta_0}, V_{\beta_0}) \tag{8}$$

$$\lambda \sim \mathcal{IG}\left(\frac{s_0}{2}, \frac{\nu_0}{2}\right) \tag{9}$$

$$\gamma_c \sim \mathcal{N}(\mu_{\gamma_{c,0}}, V_{\gamma_{c,0}}) \tag{10}$$

This means that the dynamic coefficients for each country are drawn from the same distribution, reflecting the idea that each country is a variation of the same underlying model. The matrices $\Lambda_c$ are taken to be given and parameterized similar to the Minnesota prior distribution exactly as in Jarociński (2010). Non-informative priors can be used by setting $V_{\beta_0}^{-1} = 0$, $V_{\gamma_{c,0}}^{-1} = 0$, $s_0 = -1$ and $\nu_0 = 0$. Setting the latter two parameters implies a uniform prior for $\sqrt{\lambda}$.

For the covariance matrix $\Sigma_c$ generally an Inverse-Wishart distribution is assumed

$$\Sigma_c \sim \mathcal{IW}(R_{c,0}, d_{c,0}) \tag{11}$$

One can set $R_{c,0} = (d_{c,0} - N - 1)diag(\sigma_{c,1}^2, \ldots, \sigma_{c,N}^2)$ and $d_{c,0} = N + 2$, where the diagonal entries are obtained from univariate autoregressive models with a constant and $L$ lags. An often used alternative is the Jeffrey's prior. In the formulas for optimal densities and conditional posteriors, this prior can be implemented by setting $d_{c,0} = 0$ and $R_{c,0} = 0$.

To complete the prior specification we assume a uniform distribution for the degrees of freedom $\nu_c$:

$$\nu_c \sim \mathcal{U}(\underline{\nu}, \overline{\nu}) \tag{12}$$

where we set $\underline{\nu} = 2$ and $\overline{\nu} = 50$ following Chan (2020). The upper bound is chosen to be large enough so that the errors can be normal (approximately), yet small enough that the assumption of Gaussian errors is not enforced through the prior, i.e. that not too much prior probability mass is placed on large degrees of freedom. An alternative that we do not pursue in this article is an exponential distribution for $\nu_c$, which is the approach taken in Geweke (1993).

## 2.2 Posterior approximation

As it stands, the above model does not have a known joint posterior density over its parameters. The usual way around this problem is to implement a Gibbs sampler to sequentially draw from the conditional posterior distributions, which in the above model needs to be augmented with a Metropolis-Hastings step to sample the degrees of freedom parameters of the t-distributions. Draws from such a sampler will converge to the true posterior distribution after a sufficient number of draws, and details on this implementation can be found in Appendix C. However, this is computationally demanding in this model, so we seek a faster albeit approximate way of estimating it.

Before going into the details of the posterior approximation used in this paper, we briefly discuss the general idea behind variational inference (see e.g. Bishop (2006) or Blei et al. (2017) for further details). For a generic model with parameters $\theta$ and observed data $y$ interest lies in characterizing the posterior distribution $p(\theta|y)$. This distribution, however, is typically too complex to be analysed analytically. The VI approach then specifies a tractable joint density function $q(\theta)$ over the models' parameters to best fit the true posterior distribution. To formally illustrate the idea one makes use of the following decomposition of the marginal likelihood:

$$\begin{aligned}\ln(p(y)) &= \int q(\theta) \ln\left(\frac{p(y,\theta)}{q(\theta)}\right) d\theta + \int q(\theta) \ln\left(\frac{q(\theta)}{p(\theta|y)}\right) d\theta \\ &\equiv \mathcal{L}(q) + D_{KL}(q||p)\end{aligned}$$

where the second term is the Kullback-Leibler divergence. Since $D_{KL}(q||p) \geq 0$, we have $\ln(p(y)) \geq \mathcal{L}(q)$, which is why the first term is often referred to as the Evidence Lower Bound (ELBO). The goal is then to minimize the divergence with respect to $q$, which is equivalent to maximizing the ELBO because the marginal likelihood is independent of $q$. Minimizing the Kullback-Leibler divergence is not feasible because it involves the unknown posterior, but evaluating the ELBO is possible because it only involves expectations of the joint density under $q$ and the entropy of the approximating density, which in turn is typically easy to obtain.

To actually obtain a tractable distribution we make the popular *mean field* assumption that the approximating density $q$ factors into $m$ independent parts:

$$q(\theta) = \prod_{m=1}^{M} q_m(\theta_m) \tag{13}$$

Second, given this structure of $q(\theta)$, one can show that minimizing the Kullback-Leibler divergence between $q(\theta)$ and $p(\theta|y)$ means that the individual densities must satisfy

$$\ln\left(q_j^\star(\theta_j)\right) \propto \mathbb{E}_{i=j}\left[\ln(p(\theta, y))\right] \tag{14}$$

Equation (14) provides us with a simple recipe for finding the optimal constituent densities. We first write down the log joint density over parameters and observed data. Then, for any given parameter block $m$, take expectations with respect to everything other than the parameters in block $m$ and inspect the resulting expressions to find simple distributional forms.

The question then is which factorization to assume. One possibility is to use the same blocking scheme as used in the MCMC algorithm, thus factorizing over all parameters. While such an approach seems to work well for VAR models at country level with typical prior distributions, here it can lead to a poor approximation of the posterior. The reason is that likely there exist strong dependencies between $\beta_0$, $\lambda$ and the country coefficients in the true posterior. Therefore, any approximation which breaks this dependency may fail to capture important features of the true distribution. Generally speaking, one wants to retain as much dependency between parameter blocks as possible, but some posterior dependencies must be sacrificed in order to determine $q$ analytically. Here we use the following factorization inspired by Wand (2014), Lee and Wand (2016) and Nolan et al. (2020):

$$q(\theta) = q(\delta)q(\lambda)\prod_{c=1}^{C} q(\Sigma_c)q(\nu_c)\prod_{t=1}^{T} q(\omega_{ct}) \tag{15}$$

where $\delta$ is the $(N^2L + CNK) \times 1$ vector containing the common mean $\beta_0$ and all country coefficients $\beta_c$ and $\gamma_c$. To avoid notational clutter we have suppressed subscripts to indicate that the factors are different densities, but this should be kept in mind. This factorization can model dependencies between the common mean and the country coefficients, but still imposes that the common variance parameter $\lambda$ is independent from the remaining parameters in the posterior. As we show below, this factorization is rich enough to capture the VAR dynamics and still leads to simple q-factors for most parameter blocks. Namely, the optimal density for $\delta$ is a Multivariate Normal distribution, the optimal density for $\lambda$ is an Inverse-Gamma distribution, the error covariances $\Sigma_c$ individually follow Inverse-Wishart distributions, and the $\omega_{c,t}$ follow Inverse-Gamma distributions. Details and derivations are provided in Appendix B.

Let us briefly illustrate the actual steps for the case of the optimal factor for one of the residual covariance matrices $\Sigma_c$. Recall that the optimal density $q(\Sigma_c)$ satisfies (14) which tells us to investigate the expectation under $q$ of log joint density over parameters and data. Writing out the likelihood function and relegating everything unrelated to $\Sigma_c$ to the normalization constant we have:

$$\mathbb{E}_{q(-\Sigma_c)}\left[\ln(p(\theta, y))\right] \propto -\frac{d_{c,0} + T + N + 1}{2} \ln(|\Sigma_c|) - \frac{1}{2} \operatorname{tr}\left\{\Sigma_c^{-1}\left(R_{c,0} + \mu_{U_c' U_c}\right)\right\} \tag{16}$$

where $U_c = Y_c - X_c B_c - Z_c \Gamma_c$. $\mu_{U_c' U_c}$ denotes the expectation over the quadratic form of VAR residuals, which does not simply reduce to a quadratic form of expectations, since one also has to account for the covariance between the elements of $U_c$. The formulas in Hajargasht (2019) show how expectations over quadratic forms are computed, and the details are not important for the discussion at hand. The important point to note here is that, if the optimal factor for the VAR parameters has a convenient and known form, one can easily compute this expectation. In particular, it turns out that $q(\delta)$ is a multivariate normal distribution, so the means and covariance terms corresponding to country $c$ tell us everything we need to know to analytically find $\mu_{U_c' U_c}$. From this it follows that the optimal factor for $\Sigma_c$ is an Inverse-Wishart distribution. The fact that the solution depends on the moments of other optimal factors is not a problem since we can simply iterate over the factors, updating each of them in turn.

The previous steps are repeated for all optimal factors, but when we come to the degrees of freedom parameters $\nu_c$, we run into a problem. Generally, the optimal densities for this parameter given (15) have the form:

$$\log(q(\nu_c)^*) \propto \frac{T\nu_c}{2} \log(\nu_c/2) - T \log(\Gamma(\nu_c/2)) - \frac{\nu_c}{2} \sum_{t=1}^{T} \left(\mu_{\log(\omega_{c,t})} + \mu_{\omega_{c,t}^{-1}}\right) \tag{17}$$

where $\mu_{\log(\omega_{c,t})}$ and $\mu_{\omega_{c,t}^{-1}}$ are the expected values of $\log(\omega_{c,t})$ and $\omega_{c,t}^{-1}$ under $q$, respectively. This is not the kernel of a known density and thus cannot be used as is. Chan and Yu (2022) in their application to models with stochastic volatility also face the issue that in their proposed mean field approximation the relevant approximating factor is of unknown form. They overcome this problem by minimizing the Kullback-Leibler divergence between this unknown density and an approximating Normal distribution, letting them derive an approximate but analytically tractable Gaussian factor. This approach, however, does not work here because minimizing the divergence between the unknown density (17) and some (Gaussian) approximating density $f$ would require us to evaluate

the expectation of the log-gamma term $\mathbb{E}_f \log(\Gamma(\nu_c/2))$ under the approximating density, which is not straightforward. Instead, we follow Christmas (2014) and approximate this term using Stirling's formula:

$$\Gamma(z+1) \sim \sqrt{2\pi z}\left(\frac{z}{e}\right)^z \tag{18}$$

Together with the fact that $\Gamma(z) = \Gamma(z+1)/z$ we arrive at the following expression:

$$\log(q^*(\nu_c)) \approx c + \frac{T}{2}\log(\nu_c) - \frac{\nu_c}{2}\left(\sum_{t=1}^{T}\left(\mu_{\log(\omega_{c,t})} + \mu_{\omega_{c,t}^{-1}}\right) - T\right) \tag{19}$$

which we recognize as the kernel of a Gamma distribution with parameters $a_\nu = T/2 + 1$ and $b_{\nu_c} = 2\left(\sum_{t=1}^{T}\left(\mu_{\log(\omega_{c,t})} + \mu_{\omega_{c,t}^{-1}}\right) - T\right)^{-1}$, and so we use this as the factor for the degrees of freedom. Note that this factor has the intuitive feature that as the errors come close to being Gaussian, the period specific scaling factors $\omega_{c,t}$ are close to one. This causes the mean of the Gamma distribution to become large, as in this case $b_{\nu_c}$ is close to zero.

Two remarks are in order regarding this approximation. First, the support of the Gamma density should ideally be truncated to the interval specified by the prior for $\nu_c$. However, checking convergence of the ELBO requires to evaluate the entropies of the $q$-factors, and this is easier to do using the unrestricted Gamma distribution. Second, the approximation to the Gamma function above works well asymptotically and is actually worst for small values of $\nu_c$. This could pose a serious problem, given that previous results in the literature show that when the data support heavy-tailed error term distributions, virtually all posterior mass is located at small values for the degrees of freedom (see for instance Hartwig (2022) or Bobeica and Hartwig (2023)). However, our results for the second application suggest that the approximation error does not introduce noticeable bias overall because impulse response functions and the latent volatility parameters $\omega_{c,t}$ are accurately recovered in the posterior.

In practice we initialize the moments of these densities at some convenient values and then cycle through them to update the moments individually. This process is repeated until the change in the ELBO is less than $10^{-4}$.[1] For the derivation of the ELBO see Appendix B. Furthermore, we make use of the results in Lee and Wand (2016) who show how to efficiently implement the updates of the optimal factor for $\delta$. Once this procedure has converged, it is easy to obtain any desired moment from the (approximate) posterior

[1] Increasing the convergence tolerance to $10^{-7}$ did not affect the results.

through simulation from the individual q-factors. Note that by definition, the draws thus produced are i.i.d., in contrast to the draws from a MCMC chain.

## 2.3 Accuracy assessment

Once the optimization has converged, we would like to know how accurate the approximation is. Katsevich and Rigollet (2023) are able to obtain bounds on the approximation error when the true posterior is approximated by a joint Gaussian distribution over the parameter vector, but the authors note that the approximation error for mean field approximations in general is presently unknown. Hajargasht (2019) is able to derive the approximation error for his mean field approximation analytically for a conventional VAR model with a natural conjugate prior, but he also notes that once the conjugacy assumption is dropped, no analytical results on the approximation error are available as the posterior is unknown in this case. In general, judging the approximation qualities of mean field distributions is complicated by the fact that there are many possible choices to make about the factorization of the posterior. While any such factorization provides a valid lower bound on the marginal likelihood, some a clearly poorer than others in terms of accuarcy which cannot be judged based on the value of the respective ELBO. Generally speaking, however, one can note that the more independent factors one assumes in the posterior, the poorer the approximation should become, and a factorization that models jointly as many parameters as possible while still being analytically tractable is expected to be the most accurate within the class of mean field approximations.

To assess the accuracy of the approximation, we employ the recommended measure by Faes et al. (2011). For a generic parameter $\theta$, they propose to judge the accuracy of the approximating density by the following statistic:

$$accuracy(q) = 100\left(1 - \frac{1}{2}\int_{-\infty}^{\infty} |q(\theta) - p(\theta|y)|d\theta\right) \tag{20}$$

where the integral denotes the *Integrated Absolute Error (IAE)*. This measure lies between 0 and 100 and is interpreted as a percentage, which 100 meaning that the distributions are the same. The true posterior $p(\theta|y)$ is approximated using the MCMC results. For $q$ we use the sampled parameters as well to base the comparison on actual sampled values for both distributions. Specifically, we first use MATLAB's ksdensity function to fit separate kernel density estimates to the posterior draws for a given parameter. For distributions known to have positive support, we take the natural logarithm of the sampled values in

order to easier approximate the sampled distribution with a kernel estimate. Afterwards, a linear interpolation function is specified to be able to evaluate the densities at all parameter draws. Finally, MATLAB's integral function is used to evaluate the integral in (20). The integration bounds are determined as the smallest and largest value of the samples from both $q(\theta)$ and $p(\theta|y)$.

Additionally, we visually judge the approximation quality by comparing impulse response functions generated from the models. The individual VAR parameters are typically not interpretable, and studying the dynamic effects of shocks through the system is one conventional way of reporting VAR results. As these are highly non-linear functions of the estimated parameters, it is important to know that not only are individual parameters' distributions well approximated, but also that any approximation errors do not compound and distort the results. Furthermore, visual inspection of the results also helps to understand whether parts of the parameter space which are not well captured actually matter for the statistics of interest.

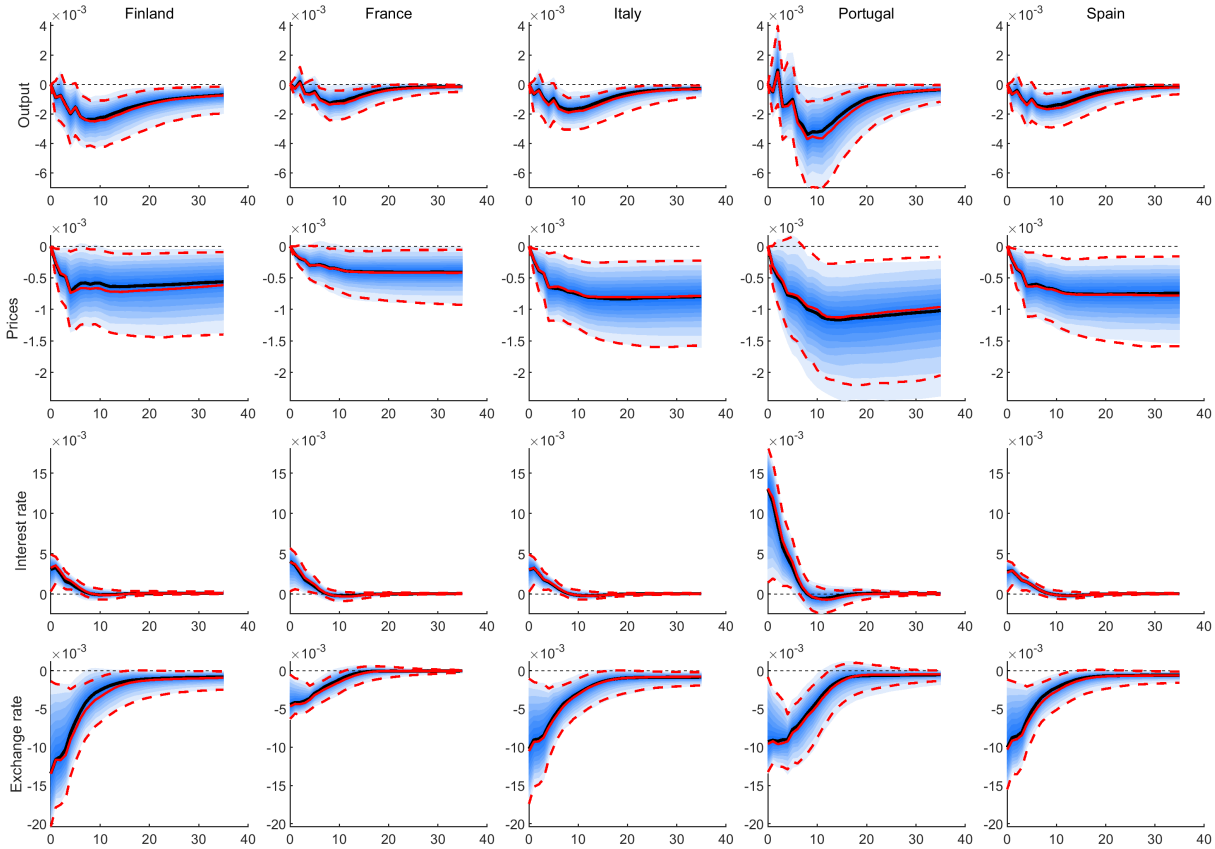
# 3 Applications

## 3.1 Monetary policy shocks in the west of Europe

As the first application we replicate the results from Jarociński (2010) where the error distribution is assumed to be standard Gaussian, to benchmark the method against established results. Each country has four endogenous variables: the log of industrial production, the log of consumer price index, the short-term market interest rate, and the log of the exchange rate in national currency units per euro. The lag length is set to six. Furthermore, the Federal Funds rate, oil prices and non-fuel commodity prices are included as exogenous variables in the vector $w_t$. Lastly, German industrial production (log), the German interbank market interest rate and the D-mark/USD exchange rate are included as exogenous variables to which the exchangeable prior does not apply. Of these exogenous variables lags zero and one are included, except for German industrial production, which enters with lags one and two. The data starts in January of 1987 and ends in December of 1998. The data are taken from the replication files accompanying the article by Jarociński (2010).

In this application the non-informative prior described in section 2 is used. MCMC results were obtained after an initial burn-in period of 10,000 draws. Afterwards, another 1,000,000 draws were obtained, of which every 100th draw was retained, giving 10,000

**Figure 1:** Impulse response functions to monetary policy shocks.




*Notes:* Blue shaded areas denote fancharts of the pointwise MCMC posterior between the 5 and 95 percentiles, with steps of 5%, and the pointwise median shown as solid black lines. Solid red lines are the pointwise medians from VI, and the pointwise 5 and 95 percentiles are shown as red-dashed lines. Monetary policy shocks are identified as in Jarociński (2010).

posterior draws used for inference. The same number of draws were taken from the approximating density $q$ after maximization of the ELBO.[2] For the impulse response functions reported below, we implement the same identification strategy of monetary policy shocks as in Jarociński (2010). For the VI approach, after convergence of the algorithm we draw country parameters for $\beta_c$ and $\Sigma_c$ from their respective distributions (because they are independent by assumption) and check whether the sign restrictions are satisfied as we would during the MCMC iterations.[3]

The major draw for VI methods is the speed with which Bayesian models can be estimated. In this application, running the MCMC chain takes roughly 2.5 hours, whereas

[2]Programs were run in MATLAB on a machine with an Intel Xeon Platinum 8180 @2.50GHz processor.

[3]This step can be parallelized, but it takes less than half a second to sample the IRFs, so we did not implement this for the results presented here.

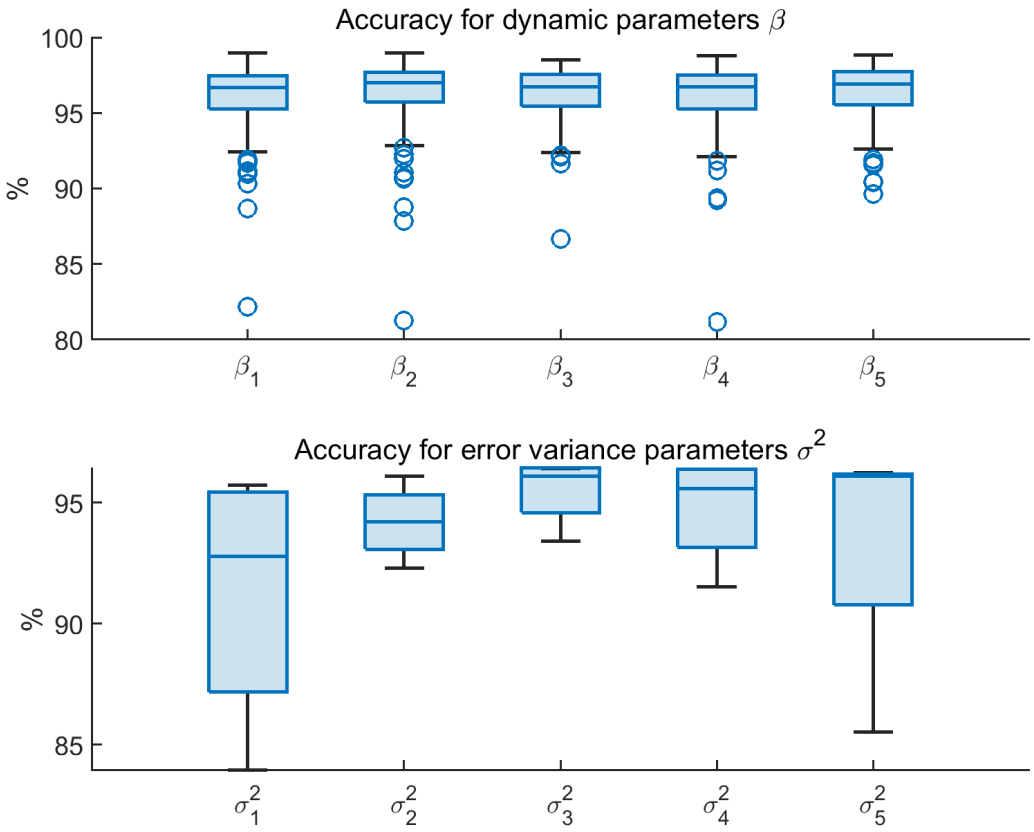
maximizing the ELBO and sampling from the posterior takes less than 30 seconds. As Figure 1 shows, this gain in speed comes at only a small cost when it comes to capturing the dynamic effects of shocks to the system. There, we plot the results from the MCMC algorithm against the results from the approximation. The blue-shaded areas show the pointwise posterior credit sets and the solid black lines denote the pointwise median response for the MCMC results. The red-dashed and red solid lines denote the respective results from the VI approximation. We observe that the pointwise medians match very closely, and the differences one can make out, e.g. the output response for Portugal or the exchange rate response of Finland, are small. Equally important is that the pointwise error bands for the impulse responses are also very close to the simulation results. This is generally not guaranteed because VI methods can severely underestimate posterior variances, depending on how much of the posterior correlations between parameters are sacrificed for tractability. This is not the case in this application, and overall we find that one would not draw different conclusions from the approximate results compared to the MCMC results.

Figure 2 shows accuracy measures for different parameters of the model. The top panel shows, for each country, boxplots for the accuracy measure for the dynamic parameters $\beta_c$. Clearly, the vast majority of parameters are well captured by the approximation with accuracy mostly exceeding 90%, although the accuracy is slightly worse for some parameters. However, as Figure 1 showed, this does not seem to affect overall accuracy for the dynamic relationships in this model. The second panel shows accuracy measures for the diagonal elements of the error covariance matrices. These boxplots show greater heterogeneity across countries, but overall accuracy rarely dips below 85%. One drawback of the approximation, though, is that it does not well capture the posterior distribution for the common variance parameter $\lambda$, and severely understates its posterior variance (not shown). However, for common purposes such as forecasting and impulse response analysis this parameter is not of particular practical interest, because once draws for the VAR parameters and error covariance matrices are obtained from their respective distributions, everything else follows. And given that these can be accurately generated, this inaccuracy seems acceptable.

## 3.2 Estimation with Covid-19 observations

In the second application, we use quarterly data on the five largest Euro Area economies Germany, France, Italy, Spain, and the Netherlands. We include six macroeconomic time

**Figure 2:** Application 1 - accuracy measures for selected parameters.

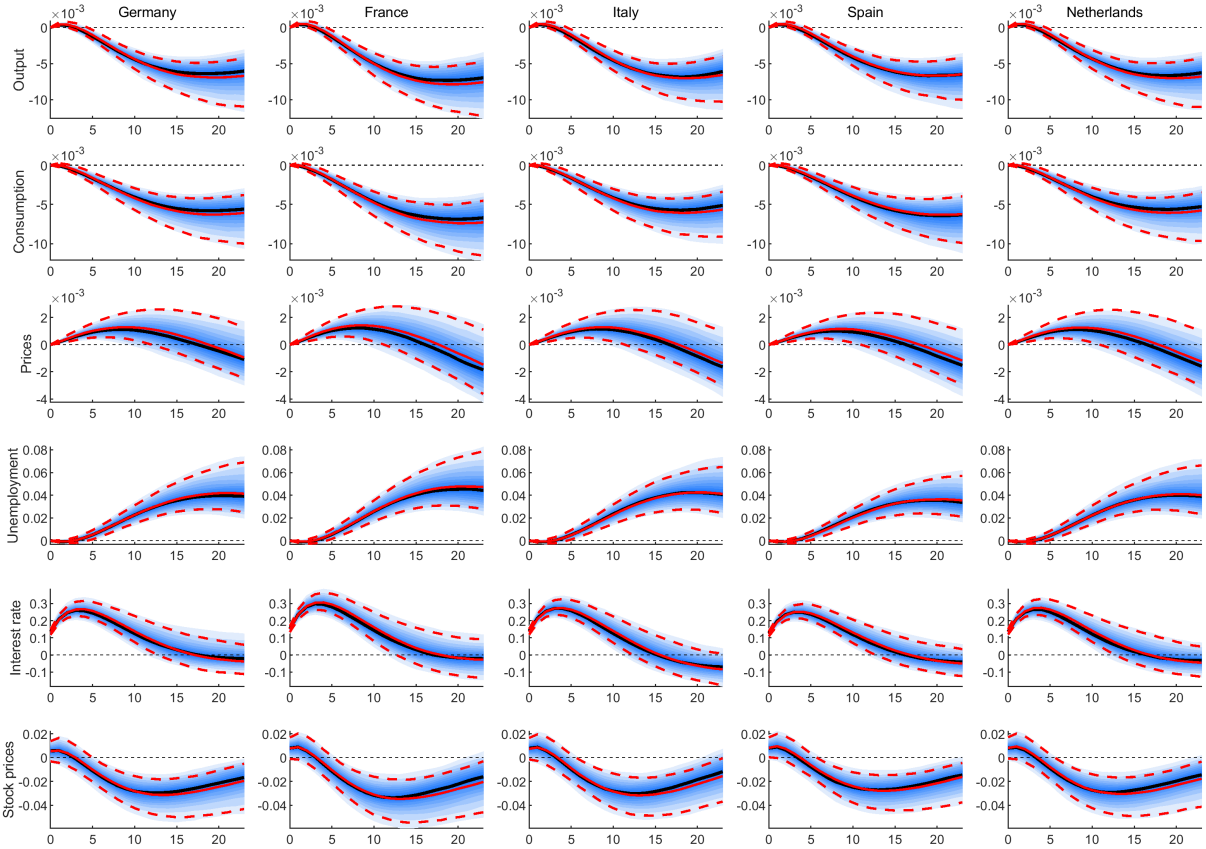



*Notes:* Each boxplot summarizes the distribution of the accuracy measure across the parameters in the respective group by country. For example, the boxplot for $\beta_1$ summarizes the accuracy measures for country one's VAR parameters subject to the exchangeable prior. The second panel show the corresponding statistics for the diagonal elements of the error covariance matrices for each country.

series, namely GDP, consumption expenditures, the harmonized consumer price index, the civilian unemployment rate, a nominal short-term rate for the Euro Area, and country specific stock prices indices. The sample period ranges from 1997Q3 until 2022Q4. All variables except those already expressed in rates are normalized to 1 in Q4 of 2010 to place them on a similar scale, and transformed to natural logarithms. The unemployment rate and the short-term rate are expressed in decimals, such that a value of 0.1 corresponds to a value of 10%. Of the endogenous variables, 2 lags are included. All data are sourced from the ECB Statistical Data Warehouse.[4] A list including series keys and whether the series is publicly available is provided in Appendix A. Following results in Hartwig

[4]https://data.ecb.europa.eu/

**Figure 3:** Impulse response functions - model with t-errors.




*Notes:* Blue shaded areas denote fancharts of the pointwise MCMC posterior between the 5 and 95 percentiles, with steps of 5%, and the pointwise median shown as solid black lines. Solid red lines are the pointwise medians from VI, and the pointwise 5 and 95 percentiles are shown as red-dashed lines. Impulse responses to fifth shock, calculated as the Cholesky factorization of the error covariance matrix, where the variables are ordered as they appear in the figure.

(2022), the prior covariance matrices $\Omega_c$ are calibrated using data until 2019Q4 only to avoid inflated residual variance estimates in influencing the prior. The MCMC results for this application were generated with the same number of draws and burn-in period as in the previous application. Details on the algorithm are provided in Appendix C.

Figure 3 shows that also in the extended model with t-errors the method performs very well as evidenced by the responses to a shock to the short-term rate. For convenience, the impulse responses shown are obtained from the Cholesky decomposition of the error covariance matrix, with the variables ordered as they appear in the figure, and we do not attach any particular structural interpretation to these results. The conclusions for the impulse response functions are the same as above, with only slight differences between the pointwise median IRFs and similar coverage of the error bands. Also here, VI is quiet

fast, taking around 43 seconds to produce the results reported here.

Figure D.1 in the Appendix shows the same accuracy measures as in the previous application. Also here, the dynamic coefficients are well captured and only slightly worse than before. Notably, for the diagonal elements of the error covariance matrix, there is a larger dispersion in accuracy across countries, and also overall the approximation seems to perform somewhat worse. However, as the previous figure showed, this does not imply that the relevant dynamics are not well captured. Also, in Figures D.2 - D.6 in Appendix D we report the estimated error scaling factors for both methods for each country. Shown are posterior quantiles of $\sqrt{\omega_{ct}}$ over time for both methods. Also there the proposed method very accurately captures the latent volatility as well, despite the fact that next to the assumed factorization of the posterior, an additional approximation was needed to arrive at a usable density for the degrees of freedom parameters for each country.

# 4 Concluding remarks

We have presented a fast and accurate mean field variational inference approximation to the true posterior for Panel VAR models with exchangeable priors and Gaussian disturbances. Our results show that one can recover the correct dynamic properties in this setup at greatly reduced computational cost, thus providing an attractive alternative to numerically demanding MCMC methods. Furthermore, we show that the approximation also accurately captures the uncertainty around the point estimates. This is important to establish because a fast algorithm that significantly understates estimation uncertainty would likely only be used for initial exploration exercises, followed by a standard Gibbs sampling algorithm thereafter, limiting the scope of applicability. We then illustrated the extension of the methodology to the case of t-distributed error terms, thus allowing the researcher to use VI methods also for macroeconomic datasets that include, for instance, the Covid-19 observations that can severely bias posterior estimates if not properly modelled. The speed improvements make these methods well suited for applications such as repeated forecasting exercises.

## References


Bishop, C. (2006). Pattern Recognition and Machine Learning. *Springer*.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bobeica, E. and Hartwig, B. (2023). The COVID-19 shock and challenges for inflation modelling. *International Journal of Forecasting*, 39(1):519–539.

Chan, J. C. (2020). Large Bayesian VARs: A flexible Kronecker error covariance structure. *Journal of Business & Economic Statistics*, 38(1):68–79.

Chan, J. C. and Hsiao, C. Y. (2014). Estimation of Stochastic Volatility Models with Heavy Tails and Serial Dependence. *Bayesian Inference in the Social Sciences*, pages 155–176.

Chan, J. C. and Yu, X. (2022). Fast and Accurate Variational Inference for Large Bayesian VARs with Stochastic Volatility. *Journal of Economic Dynamics and Control*, 143:104505.

Christmas, J. (2014). Bayesian spectral analysis with Student-t noise. *IEEE Transactions on Signal Processing*, 62(11):2871–2878.

Faes, C., Ormerod, J. T., and Wand, M. P. (2011). Variational Bayesian Inference for Parametric and Nonparametric Regression with Missing Data. *Journal of the American Statistical Association*, 106(495):959–971.

Gefang, D., Koop, G., and Poon, A. (2023). Forecasting using variational Bayesian inference in large vector autoregressions with hierarchical shrinkage. *International Journal of Forecasting*, 39(1):346–363.

Geweke, J. (1993). Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, 8(S1):S19–S40.

Hajargasht, R. (2019). Approximation properties of variational bayes for vector autoregressions. *arXiv preprint arXiv:1903.00617*.

Hartwig, B. (2022). Bayesian VARs and prior calibration in times of COVID-19. *Studies in Nonlinear Dynamics & Econometrics*.

Hughes, D. M., García-Fiñana, M., and Wand, M. P. (2023). Fast approximate inference for multivariate longitudinal data. *Biostatistics*, 24(1):177–192.

Jarociński, M. (2010). Responses to monetary policy shocks in the east and the west of Europe: a comparison. *Journal of Applied Econometrics*, 25(5):833–868.

Katsevich, A. and Rigollet, P. (2023). On the Approximation Accuracy of Gaussian Variational Inference. *arXiv preprint arXiv:2301.02168*.

Korobilis, D. and Schröder, M. (2024). Monitoring multi-country macroeconomic risk: A quantile factor-augmented vector autoregressive (QFAVAR) approach. *Journal of Econometrics*, page 105730.

Lee, C. Y. Y. and Wand, M. P. (2016). Streamlined mean field variational bayes for longitudinal and multilevel data analysis. *Biometrical Journal*, 58(4):868–895.

Lenza, M. and Primiceri, G. E. (2022). How to estimate a vector autoregression after March 2020. *Journal of Applied Econometrics*, 37(4):688–699.

Menictas, M., Di Credico, G., and Wand, M. P. (2022). Streamlined variational inference for linear mixed models with crossed random effects. *Journal of Computational and Graphical Statistics*, pages 1–17.

Nolan, T. H., Menictas, M., and Wand, M. P. (2020). Streamlined computing for variational inference with higher level random effects. *The Journal of Machine Learning Research*, 21(1):6278–6339.

Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*.

# Appendices

## A Data

The dataset for the first application is taken from the replication files accompanying the article by Jarociński (2010). The data used for the second application are listed in the following table.

**Table 1:** Data for Euro Area application

| Variable | Series Key | Transf. | Public |
|---|---|---|---|
| GDP | MNA.Q.Y.*.W2.S1.S1.B.B1GQ._Z._Z._Z.EUR.LR.N | Index and log-level | Yes |
| Consumption | MNA.Q.Y.*.W0.S1M.S1.D.P31._Z._Z._T.XDC.LR.N | Index and log-level | Yes |
| CPI | ICP.M.*.Y.000000.3.INX | Index and log-level | No |
| | ICP.M.DE.Y.000000.2.INX | Index and log-level | No |
| Unemployment | LFSI.Q.*.S.UNEHRT.TOTAL0.15_74.T | Division by 100 | Yes |
| Short-term rate | FM.Q.U2.EUR.RT.MM.EURIBOR3MD_.HSTA | Division by 100 | Yes |
| Stock prices | FM.Q.*.EUR.DS.EI.TOTMK*.HSTA | Index and log-level | No |
| | FM.Q.DE.EUR.DS.EI.TOTMKBD.HSTA | Index and log-level | No |

*Notes*: Series keys from the ECB Statistical Data Warehouse. $*$ refers to the $ISO3166-2$ country codes. Monthly series are converted to quarterly frequency before transformations are applied. Series key are slightly different for Germany in the case of CPI and Stock prices and are listed separately.

# B Deriving the optimal approximating density

Here we derive the optimal approximating density and show how to obtain its constituent parts mentioned in Section 2. Before stating the joint density over data and parameters, we first state some rather cumbersome definitions. The common regressor matrix that contains all right hand side variables for a given country is defined as $F_c = [X_c,\ Z_c]$. Next, at country level, we define the combined parameter vector $\delta_c = [\beta_c',\ \gamma_c']'$. Further, let the combined parameter vector be defined as $\delta = [\beta_0', \delta_1', \ldots, \delta_C']'$. Lastly, define the selection matrix $S_0$ such that $\beta_0 = S_0\delta$ and the selection matrices $S_{\delta_c}$, $S_{\beta_c}$ and $S_{\gamma_c}$ such that $\delta_c = S_{\delta_c}\delta$, $\beta_c = S_{\beta_c}\delta$ and $\gamma_c = S_{\gamma_c}\delta$. Note that the elements in $\delta_c$ are not in the correct order to correspond to the matrix $F_c$. Hence, we also introduce the permutation matrix $P_c$ that reorders the elements of $\delta_c$ to be compatible with $F_c$.

With these definitions the joint density function over data and parameters for the model is (up to proportionality):

$$
\begin{aligned}
p(y,\theta) \propto\ & \lambda^{-\left(\frac{s_0}{2}+1\right)} \exp\left\{-\frac{\nu_0}{2}\frac{1}{\lambda}\right\} \exp\left\{-\frac{1}{2}(S_0\delta-\mu_{\beta_0})'V_{\beta_0}^{-1}(S_0\delta-\mu_{\beta_0})\right\} \\
& \left[\prod_{c=1}^{C} |\Sigma_c|^{-\frac{T}{2}} |\Omega_c|^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\left(y_c-(I_N\otimes F_c)P_cS_{\delta_c}\delta\right)'\left(\Sigma_c^{-1}\otimes\Omega_c^{-1}\right)\left(y_c-(I_N\otimes F_c)P_cS_{\delta_c}\delta\right)\right\}\right. \\
& \lambda^{-\frac{NK}{2}} \exp\left\{-\frac{1}{2}(S_{\beta_c}\delta-S_0\delta)'(\lambda\Lambda_c)^{-1}(S_{\beta_c}\delta-S_0\delta)\right\} \\
& \exp\left\{-\frac{1}{2}(S_{\delta_c}\delta-\mu_{\gamma_{c,0}})'V_{\gamma_{c,0}}^{-1}(S_{\delta_c}\delta-\mu_{\gamma_{c,0}})\right\} \\
& |\Sigma_c|^{-\frac{1}{2}(d_{c,0}+N+1)} \exp\left\{-\frac{1}{2}\operatorname{tr}\left\{\Sigma_c^{-1}R_{c,0}\right\}\right\} \\
& \left.\prod_{t=1}^{T} \omega_{c,t}^{-\left(\frac{\nu_c}{2}+1\right)} \exp\left\{-\frac{\nu_c}{2}\frac{1}{\omega_{c,t}}\right\}\right]
\end{aligned}
\tag{B.1}
$$

The first line contains the kernel of the Inverse-Gamma density for the overall tightness parameter as well as the kernel of the normal distribution for the exchangeable mean. The following lines are the product over the individual countries' likelihood, the prior on the VAR coefficients, and the prior on the error covariance matrix, and the prior on the latent volatility factors, respectively.

To derive the optimal densities according to our assumed factorization in (15) we apply the result in (14) to each factor and we will discuss them in turn. Note that it is sufficient to consider only the terms that have a dependence on the parameter in question, and

anything else can be relegated to a constant and thus ignored when deriving the optimal densities.

## B.1 Optimal density for $\delta$

Taking logs of (B.1) and dropping terms that do not involve $\delta$ gives

$$\begin{aligned}\ln\left(q^{\star}(\delta)\right) \propto -\frac{1}{2}\sum_{c=1}^{C}\mathbb{E}_{q(-\delta)}\Big[&\frac{1}{C}(S_0\delta-\mu_{\beta_0})'V_{\beta_0}^{-1}(S_0\delta-\mu_{\beta_0})+\\ &\delta'(S_{\beta_c}-S_0)'(\lambda\Lambda_c)^{-1}(S_{\beta_c}-S_0)\delta+\\ &(S_{\gamma_c}\delta-\mu_{\gamma_{c,0}})'V_{\gamma_{c,0}}^{-1}(S_{\gamma_c}\delta-\mu_{\gamma_{c,0}})+\\ &(y_c-(I_N\otimes F_c)P_cS_{\delta_c}\delta)'\left(\Sigma_c^{-1}\otimes\Omega_c^{-1}\right)(y_c-(I_N\otimes F_c)P_cS_{\delta_c}\delta)\Big]\end{aligned} \tag{B.2}$$

where the expectation is taken over all parameters except $\delta$. By assumption (14), we can simply replace the inverses of $\lambda$, $\Sigma_c$ and $\Omega_c$ by the expectations over these quantities. Also, define for convenience $\mathcal{S}=\sum_{c=1}^{C}(S_{\beta_c}-S_0)'\Lambda_c^{-1}(S_{\beta_c}-S_0)$. If we then expand and group together terms, we find the following normal distribution as the optimal factor for $\delta$:

$$q^{\star}(\delta)=\mathcal{N}(\mu_{\delta},V_{\delta}) \tag{B.3}$$

$$V_{\delta}=\left[S_0'V_{\beta_0}^{-1}S_0+\mu_{\lambda^{-1}}\mathcal{S}+\sum_{c=1}^{C}S_{\gamma_{c,0}}'V_{\gamma_{c,0}}^{-1}S_{\gamma_{c,0}}+\sum_{c=1}^{C}S_{\delta_c}'P_c'\left(\mu_{\Sigma_c^{-1}}\otimes F_c'\mu_{\Omega_c^{-1}}F_c\right)P_cS_{\delta_c}\right]^{-1} \tag{B.4}$$

$$\mu_{\delta}=V_{\delta}\left[S_0'V_{\beta_0}^{-1}\mu_{\beta_0}+\sum_{c=1}^{C}S_{\gamma_c}'V_{\gamma_{c,0}}^{-1}\mu_{\gamma_{c,0}}+\sum_{c=1}^{C}S_{\delta_c}'P_c'\left(\mu_{\Sigma_c^{-1}}\otimes F_c'\mu_{\Omega_c^{-1}}\right)y_c\right] \tag{B.5}$$

As noted in the main text, we use the insights in Lee and Wand (2016) to facilitate the computation at this and the following step. The problem is that, generally, solving for $V_{\delta}$ involves the inversion of a large matrix which is time consuming. They show that a hierarchical model such as the one considered here has a structure that leads to the term in square brackets being block diagonal. Importantly, they also recognize that only certain parts of this inverse matrix are actually needed for the algorithm, so that the matrix inversion can be broken down into several smaller blocks. For even more details we refer to their expositions for even more details. Their algorithm also means that it makes sense to compute the update for $\lambda$ (next subsection) within the same updating function. This

is because this update relies on $V_\delta$ and has to therefore be adjusted slightly.

## B.2 Optimal density for $\lambda$

Moving on to the common scaling factor we again only retain terms that functionally depend on $\lambda$, giving:

$$\ln\left(q^\star(\lambda)\right) \propto \mathbb{E}_{q(-\lambda)}\left[-\left(\frac{s_0+CNK}{2}+1\right)\ln(\lambda)-\frac{1}{2}\lambda^{-1}\left(\nu_0+\delta'\mathcal{S}\delta\right)\right] \tag{B.6}$$

Evaluating the expectation we find the following Inverse-Gamma distribution as the optimal factor for $\lambda$:

$$q^\star(\lambda) = \mathcal{IG}\left(\frac{\bar{s}}{2},\frac{\bar{\nu}}{2}\right) \tag{B.7}$$

$$\bar{s} = s_0 + CNK \tag{B.8}$$

$$\bar{\nu} = \nu_0 + \mu_\delta'\mathcal{S}\mu_\delta + \operatorname{tr}\left(\mathcal{S}V_\delta\right) \tag{B.9}$$

from which we find that $\mu_{\lambda^{-1}} = \bar{s}/\bar{\nu}$.

## B.3 Optimal density for $\Sigma_c$

For the optimal covariance matrix density we get

$$\ln\left(q^\star(\Sigma_c)\right) \propto \mathbb{E}_{q(-\Sigma_c)}\left[-\frac{1}{2}(d_{c,0}+T+N+1)\ln\left(|\Sigma_c|\right)-\frac{1}{2}\operatorname{tr}\left\{\Sigma_c^{-1}\left(R_{c,0}+\left(Y_c-F_cG_c\right)'\Omega_c^{-1}\left(Y_c-F_cG_c\right)\right)\right\}\right. \tag{B.10}$$

where $P_c\delta_c = \operatorname{vec}(G_c)$. Again, after taking the expectation over everything inside the trace operator we find that the optimal density for $\Sigma_c$ is Inverse-Wishart

$$q^\star(\Sigma_c) = \mathcal{IW}(d_{c,0}+T,\bar{S}_c) \tag{B.11}$$

$$\bar{S}_c = R_{c,0}+\left(Y_c-F_c\mu_{G_c}\right)'\mu_{\Omega_c^{-1}}\left(Y_c-F_c\mu_{G_c}\right)+\Omega_{G_c} \tag{B.12}$$

where the $i,j$ element of $\Omega_{G_c}$ is given by $\Omega_{G_c}^{[i,j]} = \operatorname{tr}\left(F_c'\mu_{\Omega_c^{-1}}F_cCov(G_{c,i},G_{c,j})\right)$ and $G_{c,i},G_{c,j}$ denote the $i$-th and $j$-th column of $G_c$ (see e.g. Hajargasht (2019)). This implies that $\mu_{\Sigma_c^{-1}} = T\bar{S}_c^{-1}$.

## B.4 Optimal density for $\omega_{c,t}$

For the optimal densities we note that as long as the latent volatility factors $\omega_{c,t}$ are assumed independent from the remaining parameters, the likelihood together with the prior on $\omega_{c,t}$ implies that the optimal density for $q(\omega_c)$ factors into $t$ independent terms even if we did not initially assume this independence between the elements of $\omega_c$. Hence, for each element we get

$$\ln\left(q^{\star}(\omega_{c,t})\right) \propto \mathbb{E}_{q(-\omega_{c,t})}\left[-\left(\frac{\nu_c+N}{2}+1\right)\log(\omega_{c,t}) - \omega_{c,t}^{-1}\frac{\nu_c + u_t'\Sigma_c^{-1}u_t}{2}\right] \tag{B.13}$$

Evaluating the expectation we have that the individual factors are Inverse-Gamma distributions:

$$q^{\star}(\omega_{c,t}) = \mathcal{IG}\left(\frac{\bar{s}_{\omega_{c,t}}}{2}, \frac{\bar{v}_{\omega_{c,t}}}{2}\right) \tag{B.14}$$

$$\bar{s}_{\omega_{c,t}} = \mu_{\nu_c} + N \tag{B.15}$$

$$\bar{v}_{\omega_{c,t}} = \mu_{\nu_c} + d_{c,t} \tag{B.16}$$

$$d_{c,t} = \left[(Y_c - F_c\mu_{G_c})\mu_{\Sigma_c^{-1}}(Y_c - F_c\mu_{G_c})' + F_cR_cF_c'\right]_{t,t} \tag{B.17}$$

where again $P_c\delta_c = \text{vec}(G_c)$ and $d_{c,t}$ denotes the t-th diagonal element of the matrix to its right. The entries of the matrix $R_c$ are given by $R_c^{[i,j]} = \text{tr}(\mu_{\Sigma_c^{-1}}Cov(G_{c,i}', G_{c,j}'))$ and the covariance is taken over rows $i$ and $j$ of the matrix $G_c$. It then follows that $\mu_{\omega_{c,t}^{-1}} = \bar{s}_{\omega_{c,t}}/\bar{v}_{\omega_{c,t}}$ and $\mu_{\log(\omega_{c,t})} = \log(\bar{v}_{\omega_{c,t}}/2) - \psi(\bar{s}_{\omega_{c,t}}/2)$. $\psi(.)$ denotes the digamma function.

## B.5 Optimal density for $\nu_c$

The optimal density for $\nu_c$ was derived in the main text as a Gamma distribution:

$$q^{\star}(\omega_{c,t}) = \mathcal{G}\left(a_{\nu}, b_{\nu_c}\right) \tag{B.18}$$

$$a_{\nu} = T/2 + 1 \tag{B.19}$$

$$b_{\nu_c} = \frac{2}{\sum_{t=1}^{T}\left(\mu_{\log(\omega_{c,t})} + \mu_{\omega_{c,t}^{-1}}\right) - T} \tag{B.20}$$

which implies $\mu_{\nu_c} = a_{\nu}b_{\nu_c}$.

## B.6 Maximizing the Evidence Lower Bounds (ELBO)

Our goal is to maximize $\mathcal{L}(q)$ which is achieved by the optimal densities (B.3), (B.7), (B.11), (B.13), and (B.18). As these are coupled, we initialize $\mu_{\lambda^{-1}} = 10^4$ and $\mu_{\Sigma_c^{-1}} = T \cdot I_N$ and $\mu_{\omega_{c,t}^{-1}} = 1$, and then cycle through the moments above in turn.

To check convergence, we need the ELBO, which for this model is given by (up to constant terms):

$$\begin{aligned}\mathcal{L}(q) \propto \ln\left(|V_\delta|\right) - \frac{\bar{s}}{2}\ln(\bar{\nu}) - \sum_{c=1}^{C} T \ln\left(|\bar{S}_c|\right) + \sum_{c=1}^{C}\log(b_{\nu_c}) \\ - \sum_{c=1}^{C}\sum_{t=1}^{T}\left(\frac{\bar{s}_{\omega_{c,t}}}{2}\mu_{\log(\omega_{c,t})} + \frac{\bar{v}_{\omega_{c,t}}}{2}\mu_{\omega_{c,t}^{-1}}\right) \\ + \sum_{c=1}^{C}\sum_{t=1}^{T}\left(\frac{\bar{s}_{\omega_{c,t}}}{2} + \log(\bar{v}_{\omega_{c,t}}) + \log(\Gamma(\bar{s}_{\omega_{c,t}}/2)) - (1+\bar{s}_{\omega_{c,t}}/2)\psi(\bar{s}_{\omega_{c,t}}/2)\right)\end{aligned} \tag{B.21}$$

We stop the iterations once $\mathcal{L}(q)^{(n)} - \mathcal{L}(q)^{(n-1)} <= 10^{-4}$. In the first application without t-distributed errors, densities (B.13) and (B.18) are not considered and $\Omega_c = I_T$. The ELBO in this case consists of only the first three terms.

# C Details on the Gibbs sampler

Deriving the conditional posterior distributions for the Gibbs sampler is straightforward. Let $\theta = \left\{\beta_0, \lambda, \{\beta_c, \gamma_c, \Sigma_c, \Omega_c, \nu_c\}_{c=1}^{C}\right\}$ be the full set of parameters. Below we outline a particular sequence of updating steps, although others are possible. In the following, we denote by $\theta_{-x}$ the parameter vector with element $x$ taken out. To avoid cluttering the notation we drop indices to denote the iteration number but keep in mind that any given conditional distribution for a given parameter depends on the most recent values for the other ones.

Using the joint density over parameters and data in (B.1) we have the following results. The conditional posterior for $\beta_0$ is:

$$p(\beta_0|y, \theta_{-\beta_0}) = \mathcal{N}(\mu_{\beta_0}, V_{\beta_0}) \tag{C.1}$$

$$V_{\beta_0} = \lambda\left(\sum_{c=1}^{C}\Lambda_c^{-1}\right)^{-1} \tag{C.2}$$

$$\mu_{\beta_0} = V_{\beta_0}\frac{1}{\lambda}\sum_{c=1}^{C}\Lambda_c^{-1}\beta_c \tag{C.3}$$

The conditional posterior for $\lambda$ is:

$$p(\lambda|y,\theta_{-\lambda}) = \mathcal{IG}\left(\frac{\bar{s}}{2},\frac{\bar{\nu}}{2}\right) \tag{C.4}$$

$$\bar{s} = s_0 + CNK \tag{C.5}$$

$$\bar{\nu} = \nu_0 + \sum_{c=1}^{C}(\beta_c - \beta_0)'\Lambda_c^{-1}(\beta - \beta_0) \tag{C.6}$$

The conditional posterior for $\beta_c$ is:

$$p(\beta_c|y,\theta_{-\beta_c}) = \mathcal{N}(\mu_{\beta_c}, V_{\beta_c}) \tag{C.7}$$

$$V_{\beta_c} = \left[\frac{1}{\lambda}\Lambda_c^{-1} + \Sigma_c^{-1}\otimes X_c'\Omega_c^{-1}X_c\right]^{-1} \tag{C.8}$$

$$\mu_{\beta_c} = V_{\beta_c}\left[\lambda\Lambda_c^{-1}\beta_0 + \left(\Sigma_c^{-1}\otimes X_c'\Omega_c^{-1}\right)r_c\right] \tag{C.9}$$

$$r_c = y_c - (I_N\otimes Z_c)\,\gamma_c \tag{C.10}$$

The conditional posterior for $\gamma_c$ is:

$$p(\gamma_c|y,\theta_{-\gamma_c}) = \mathcal{N}(\mu_{\gamma_c}, V_{\gamma_c}) \tag{C.11}$$

$$V_{\gamma_c} = \left[\Sigma_c^{-1}\otimes Z_c'\Omega_c^{-1}Z_c\right]^{-1} \tag{C.12}$$

$$\mu_{\gamma_c} = V_{\gamma_c}\left(\Sigma_c^{-1}\otimes Z_c'\Omega_c^{-1}\right)r_c \tag{C.13}$$

$$r_c = y_c - (I_N\otimes X_c)\,\beta_c \tag{C.14}$$

The conditional posterior for $\Sigma_c$ is:

$$p(\Sigma_c|y,\theta_{-\Sigma_c}) = \mathcal{N}(T,\bar{S}) \tag{C.15}$$

$$\bar{S} = (Y_c - X_cB_c - Z_c\Gamma_c)'\Omega_c^{-1}(Y_c - X_cB_c - Z_c\Gamma_c) \tag{C.16}$$

The conditional posterior for $\Omega_c$ breaks into independent parts for each element $\omega_{c,t}$ given by:

$$p(\omega_{c,t}|y,\theta_{-\omega_{c,t}}) = \mathcal{IG}\left(\frac{\nu_c + N}{2},\frac{\nu_c + u_t'\Sigma_c^{-1}u_t}{2}\right) \tag{C.17}$$

The conditional posterior for $\nu_c$ is given by:

$$p(\nu_c|y,\theta_{-\nu_c}) \propto \frac{T\nu_c}{2}\log(\nu_c/2) - T\log(\Gamma(\nu_c/2)) - \frac{\nu_c}{2}\sum_{t=1}^{T}\left(\log(\omega_{c,t}) + \omega_{c,t}^{-1}\right) \tag{C.18}$$

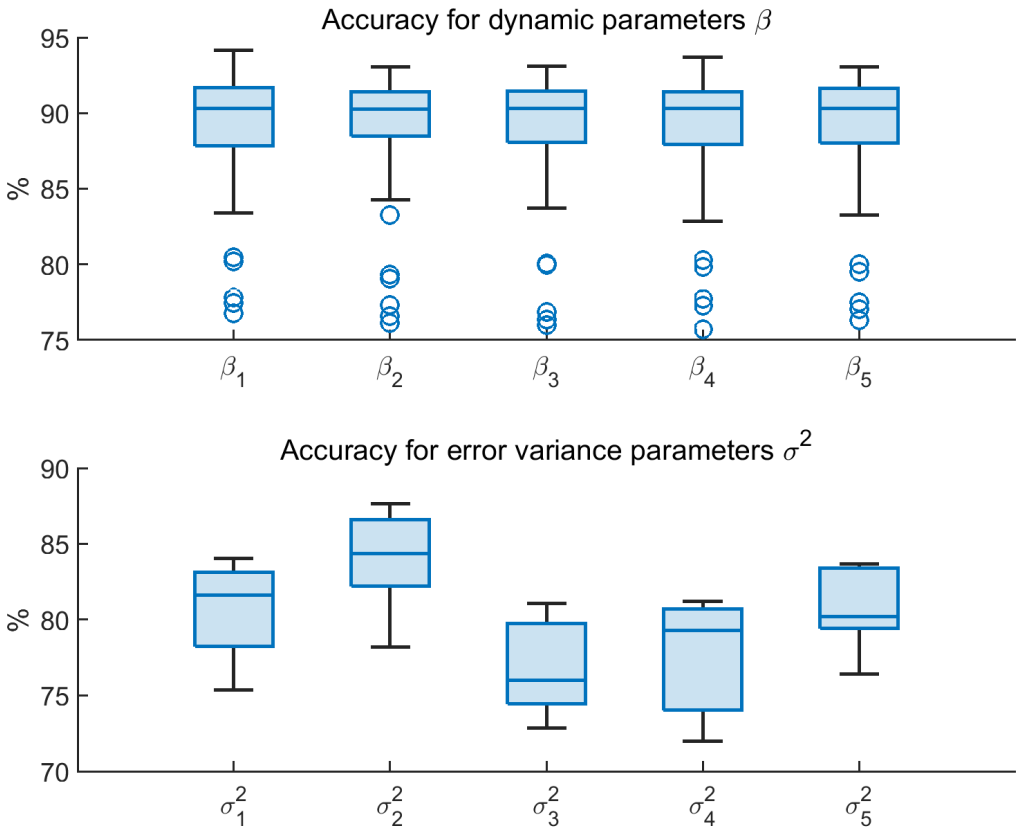
which is not the kernel of a known distribution. Therefore, the independent-MH step as explained in Chan and Hsiao (2014) is used to draw from this distribution. This amounts to first calculating the first and second derivative of the above expression w.r.t. $\nu_c$. Then, the first-order condition is solved numerically to find the mode $\hat{\nu}_c$ of the distribution. Next, the second derivative is evaluated at the mode and multiplied by -1 to obtain the negative Hessian evaluated at the mode, denoted by $\hat{K}_{\nu_c}$. Then, the proposal density is specified as $\mathcal{N}(\hat{\nu}_c, \hat{K}_{\nu_c}^{-1})$.

As above, for Application 1 the last two conditionals are ignored and $\Omega_c = I_T$.

# D Additional results

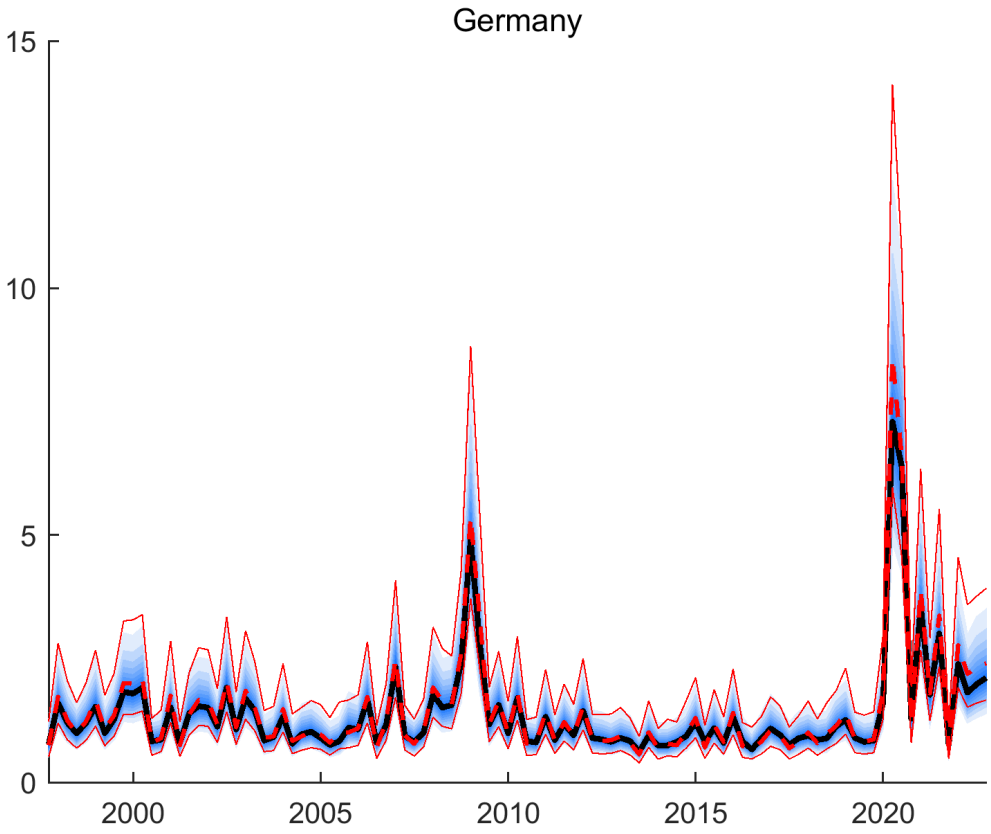
Here we provide supplementary figures to the results in the main text. Figure D.1 shows accuracy measures for the second application from the main text. Figures D.2 - D.6 show latent volatility estimates for each country individually for better visibility of the results.

**Figure D.1:** Application 2 - accuracy measures for selected parameters.



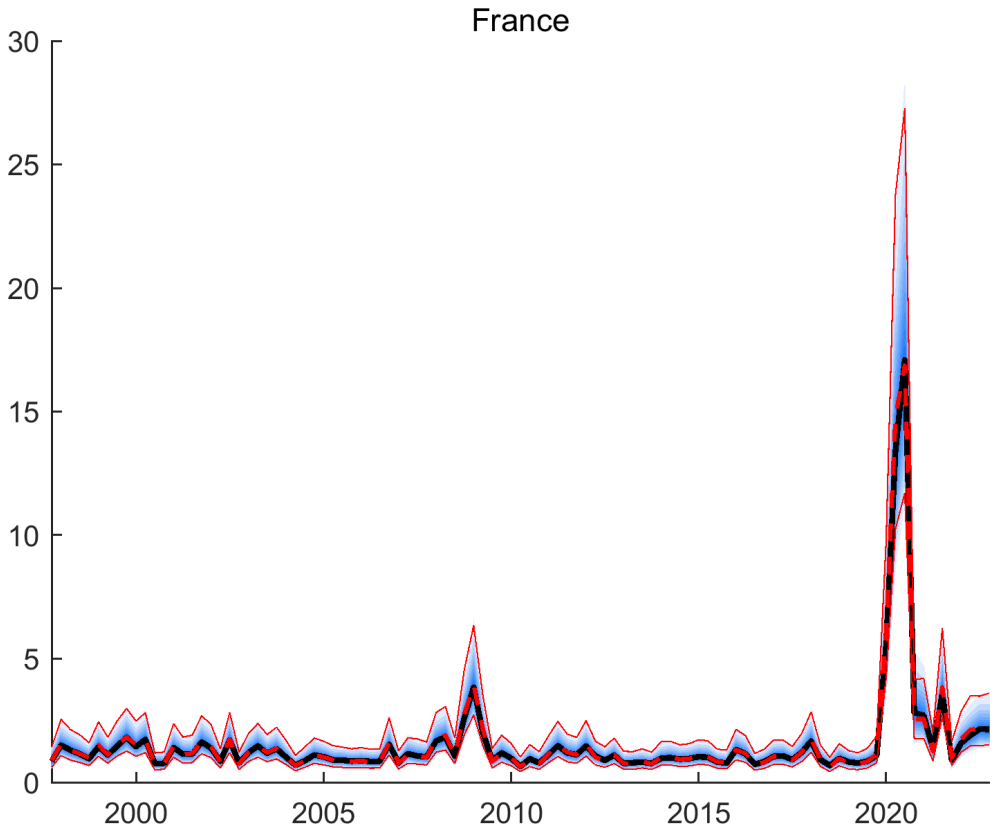

*Notes:* Each boxplot summarizes the distribution of the accuracy measure across the parameters in the respective group by country. For example, the boxplot for $\beta_1$ summarizes the accuracy measures for country one's VAR parameters subject to the exchangeable prior. The second panel show the corresponding statistics for the diagonal elements of the error covariance matrices for each country.

**Figure D.2:** Latent volatility for Germany - MCMC vs. VI



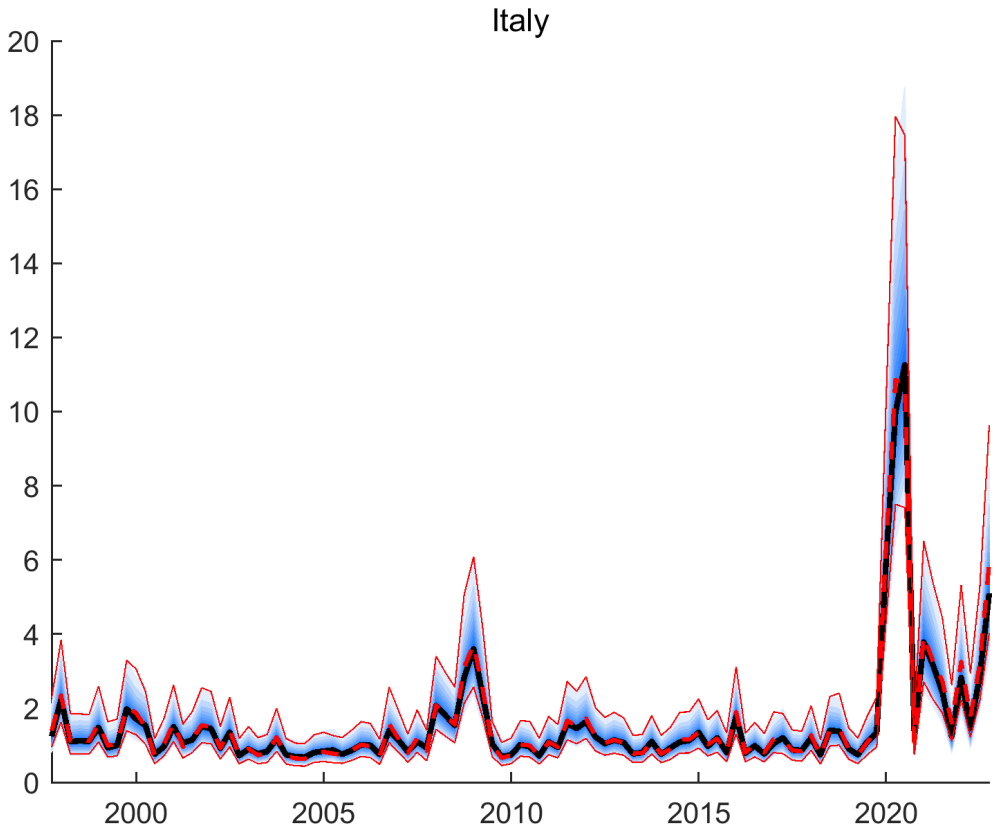

*Notes:* Latent volatility estimates for Germany. Blue shaded areas denote fancharts of the pointwise MCMC posterior for $\sqrt{\omega_{c,t}}$ between the 5 and 95 percentiles, with steps of 5%, and the pointwise median shown as solid black lines. Dashed red lines are the pointwise medians from VI, red solid lines the corresponding 5% and 95% bands.

**Figure D.3:** Latent volatility for France - MCMC vs. VI



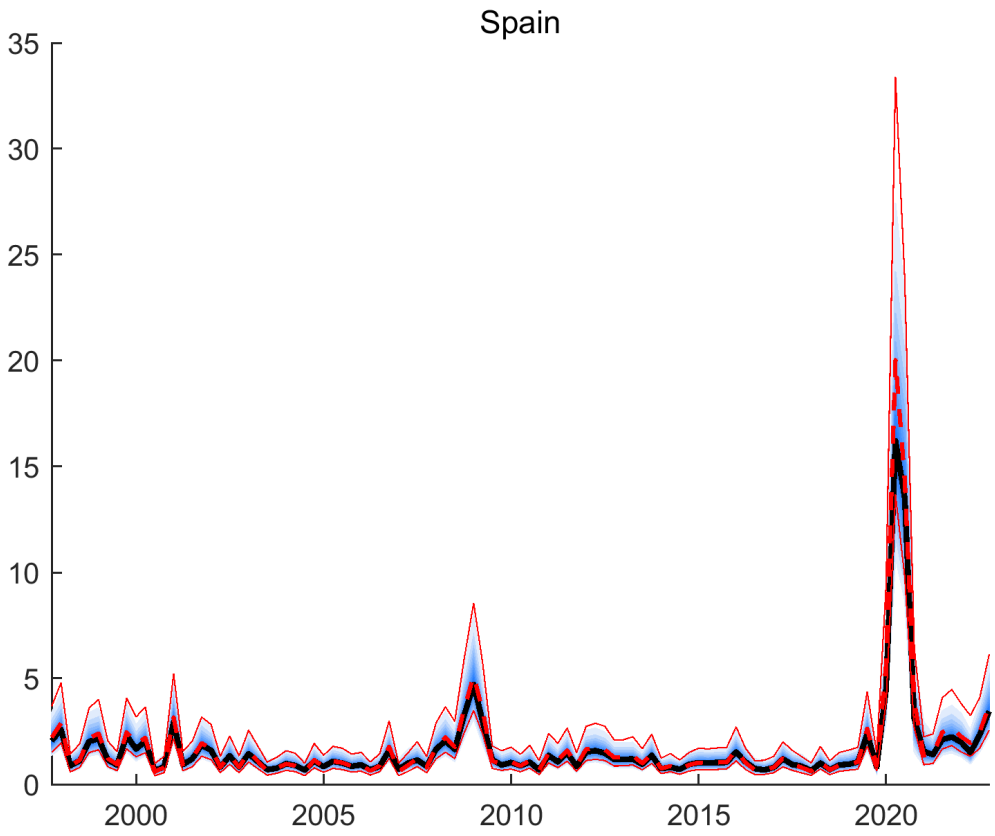

*Notes:* Latent volatility estimates for France. Blue shaded areas denote fancharts of the pointwise MCMC posterior for $\sqrt{\omega_{c,t}}$ between the 5% and 95% percentiles, with steps of 5%, and the pointwise median shown as solid black lines. Dashed red lines are the pointwise medians from VI, red solid lines the corresponding 5% and 95% bands.

**Figure D.4:** Latent volatility for Italy - MCMC vs. VI



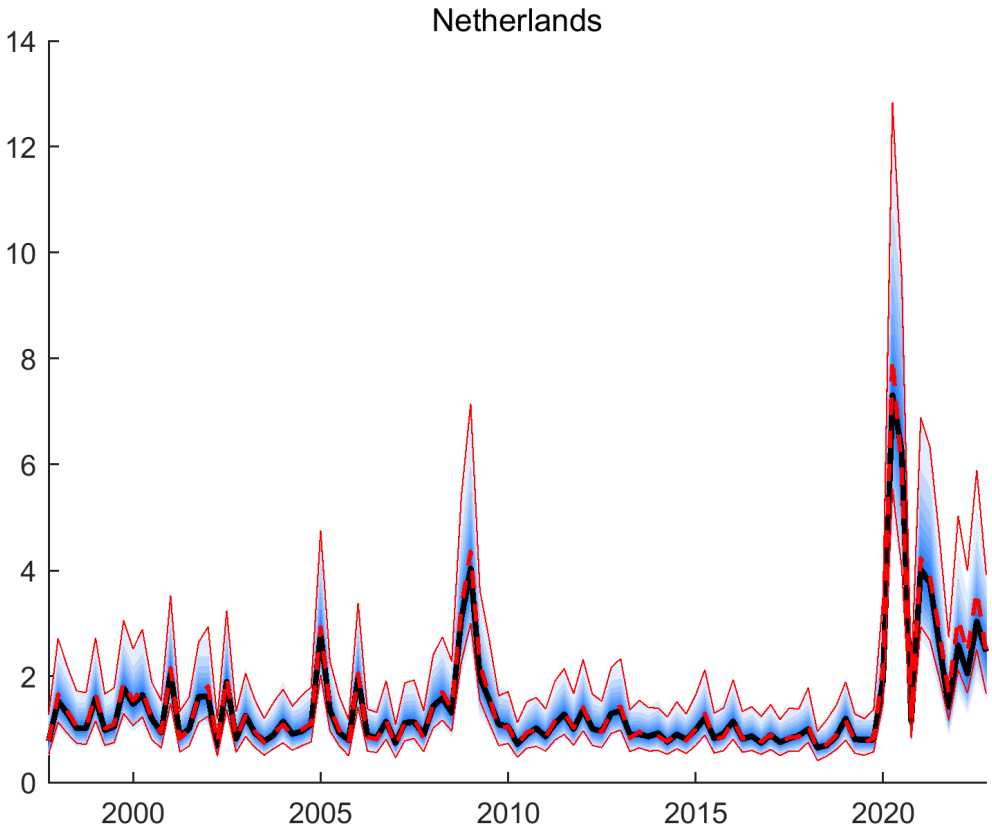

*Notes:* Latent volatility estimates for Italy. Blue shaded areas denote fancharts of the pointwise MCMC posterior for $\sqrt{\omega_{c,t}}$ between the 5 and 95 percentiles, with steps of 5%, and the pointwise median shown as solid black lines. Dashed red lines are the pointwise medians from VI, red solid lines the corresponding 5% and 95% bands.

**Figure D.5:** Latent volatility for Spain - MCMC vs. VI




*Notes:* Latent volatility estimates for Spain. Blue shaded areas denote fancharts of the pointwise MCMC posterior for $\sqrt{\omega_{c,t}}$ between the 5 and 95 percentiles, with steps of 5%, and the pointwise median shown as solid black lines. Dashed red lines are the pointwise medians from VI, red solid lines the corresponding 5% and 95% bands.

**Figure D.6:** Latent volatility for for Netherlands - MCMC vs. VI




*Notes:* Latent volatility estimates for the Netherlands. Blue shaded areas denote fancharts of the pointwise MCMC posterior for $\sqrt{\omega_{c,t}}$ between the 5 and 95 percentiles, with steps of 5%, and the pointwise median shown as solid black lines. Dashed red lines are the pointwise medians from VI, red solid lines the corresponding 5% and 95% bands.

**Acknowledgements**

I thank Juan Manuel Figueres and an anonymous referee of the ECB working paper series for valuable comments and suggestions. The views expressed in the paper do not reflect those of the ECB, the Eurosystem and should not be reported as such.

**Lucas ter Steege**
European Central Bank, Frankfurt am Main, Germany; email: Lucas.ter_steege@ecb.europa.eu



PDF ISBN 978-92-899-6881-2 ISSN 1725-2806 doi:10.2866/9979266 QB-01-24-008-EN-N