Received: 10 October 2023 | Revised: 7 May 2024 | Accepted: 29 June 2024

DOI: 10.1002/jae.3087

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

Fast and order-invariant inference in Bayesian VARs with nonparametric shocks

Florian Huber[1] | **Gary Koop**[2]

[1]Department of Economics, University of Salzburg, Salzburg, Austria

[2]Department of Economics, University of Strathclyde, Glasgow, UK

Correspondence
Florian Huber, Department of Economics, University of Salzburg, Mönchsberg 2A, 5020 Salzburg, Austria.
Email: florian.huber@plus.ac.at

Summary

The shocks that hit macroeconomic models such as Vector Autoregressions (VARs) have the potential to be non-Gaussian, exhibiting asymmetries and fat tails. This consideration motivates the VAR developed in this paper that uses a Dirichlet process mixture (DPM) to model the reduced-form shocks. However, we do not follow the obvious strategy of simply modeling the VAR errors with a DPM as this would lead to computationally infeasible Bayesian inference in larger VARs and potentially a sensitivity to the way the variables are ordered in the VAR. Instead, we develop a particular additive error structure inspired by Bayesian nonparametric treatments of random effects in panel data models. We show that this leads to a model that allows for computationally fast and order-invariant inference in large VARs with nonparametric shocks. Our empirical results with nonparametric VARs of various dimensions show that nonparametric treatment of the VAR errors often improves forecast accuracy and can be used to analyze the changing transmission of US monetary policy.

KEYWORDS

Bayesian VARs, fast estimation, infinite mixtures, Markov chain Monte Carlo

1 | INTRODUCTION

Bayesian vector autoregressions (VARs) are now routinely used with large numbers of dependent variables. The use of nonconjugate priors or non-Gaussian error distributions typically requires the use of Markov chain Monte Carlo (MCMC) methods, which leads to a large computational burden. This means full system estimation of the reduced form VAR is difficult or infeasible with large VARs. This has led many researchers to avoid full system estimation and instead work with a structural VAR with a diagonal error covariance matrix. The structural VAR allows for estimating one equation at a time that greatly reduces the computational burden making Bayesian estimation of large VARs practical. However, standard specifications for the structural VAR that allow for equation-by-equation estimation (e.g., Carriero et al., 2019) suffer from order dependence (i.e., posterior and predictive densities depend on the manner in which the variables are ordered in the VAR). The importance of order dependence, and in particular, its impact on predictive variances in larger VARs, is discussed in papers such as Arias et al. (2023) and Chan et al. (2021). There have been some order invariant approaches proposed that do allow for equation-by-equation estimation, including Chan et al. (2021) and Wu and Koop (2023), but these assume Gaussian errors and the former relies on the presence of stochastic volatility to identify the model. However, the presence of large VAR shocks that imply sudden shifts in variances and/or asymmetries in predictive densities means

more flexibility is required. These considerations motivate the present paper where we develop a VAR with a nonparametric non-Gaussian error distribution for the shocks. The MCMC algorithm we derive for this VAR is computationally efficient and order invariant.

There is a growing VAR literature that wishes to develop flexible models for the VAR errors. However, our model differs from this literature by its use of nonparametric methods and its focus on computational efficiency and order invariance. For instance, several papers, including Chiu et al. (2017) and Karlsson et al. (2023), work with VARs with parametric, non-Gaussian error distributions (e.g., Student's t distributions), but do not allow for equation-by-equation estimation and, thus, are computationally slower than our approach. A VAR with a nonparametric distribution for the structural shocks is Braun (2021), which does allow for equation-by-equation estimation conditional on the impact matrix. But this paper uses departures from normality to identify the structural shocks. Moreover, the MCMC algorithm becomes more involved due to the fact that the contemporaneous relations need to be sampled using Metropolis–Hastings (MH) updates.

To develop our model, we borrow ideas from the literature on semiparametric and nonparametric estimation of random effects in panel data models, see Frühwirth-Schnatter et al. (2004) or Dunson and Xing (2009). The key insight is that we can exploit the random effect representation of the covariance matrix of the system to enable equation-by-equation estimation, see Fox and Dunson (2015). We decompose the shock vector of the VAR into two components. The first component is a vector of random effects that feature an unknown multivariate distribution, which exhibits correlation between the errors in different equations. The second component is a vector of Gaussian random shocks that are uncorrelated across equations. Conditional on the random effects, the model becomes a system of uncorrelated regression models. But after integrating out the random effects, the resulting shock distribution features cross-sectional dependence. The key implication is that fast estimation is possible as the VAR coefficients can be drawn one equation at a time conditional on the random effects. By assuming a Dirichlet process mixture (DPM) for the vector of random effects, we achieve great flexibility as the joint distribution of the shocks can be skewed, feature heavy tails, be heteroskedastic, or be multimodal. In the multiple equation VAR context, this flexibility is potentially of great benefit as it allows for the errors in the different equations to have different properties.

In an exercise using artificial data, we show how our nonparametric VAR can automatically uncover a variety of departures from Gaussianity. In an empirical exercise involving a large data set of US macroeconomic variables, we demonstrate the advantages of our model both for forecasting and for structural economic analysis.

The remainder of the paper is structured as follows. The next section shows how a random effects representation of the VAR can be used to facilitate equation-by-equation estimation and how we treat nonparametric shocks in the VAR. Section 3 discusses the prior setup, sketches the MCMC algorithm and discusses some computational details. Sections 4 and 5 apply the model to synthetic and real data, respectively. The final section gives a summary and concludes the paper. Data S1 provides details on the dataset used and includes additional empirical results.

2 | VARS WITH NONPARAMETRIC SHOCKS

2.1 | A linear VAR with an additive Gaussian error structure

Before introducing our nonparametric specification for the shocks to the VAR, it is instructive to begin with a parametric version of our model. This involves an M-dimensional vector of dependent variables, $\{\boldsymbol{y}_t\}_{t=1}^{T}$, which evolves as

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{X}_t + \boldsymbol{\epsilon}_t + \boldsymbol{\nu}_t,\ \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{\Sigma}),\ \boldsymbol{\nu}_t \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{\Omega}_t) \tag{1}$$

where $\boldsymbol{X}_t = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p})'$ is a $K(= Mp)$ vector of lagged endogenous variables and $\boldsymbol{A}$ denotes an $M \times K$ matrix of VAR coefficients. This specification differs from a conventional VAR in that it has two M-dimensional errors, $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\nu}_t$, which are assumed to be independent over time and of one another at all leads and lags. The only restriction on $\boldsymbol{\Sigma}$ is that it is positive definite whereas $\boldsymbol{\Omega}_t$ is restricted to be a diagonal matrix with individual error variances $\omega_{1t}, \ldots, \omega_{Mt}$. We assume that the logarithms of ω_{jt} evolve according to independent random walk processes, leading to a standard SV specification.

The covariance matrix of the VAR errors, $\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_t + \boldsymbol{\nu}_t$ is $\boldsymbol{\Xi}_t = \boldsymbol{\Sigma} + \boldsymbol{\Omega}_t$. Notice that the SV assumption on the idiosyncratic shocks implies that the main diagonal elements of $\boldsymbol{\Xi}_t$ are given by $[\boldsymbol{\Xi}_t]_{ii} = \sigma_{ii}^2 + \omega_{it}$, where σ_{ii}^2 is the (i,i)th element of $\boldsymbol{\Sigma}$. We note that, without further restrictions, this specification is not identified as $[\boldsymbol{\Xi}_t]_{ii} = (\sigma_{ii}^2 + \Delta) + (\omega_{it} - \Delta)$ where Δ is any scalar. The key point to note is that this only relates to the main diagonal elements of $\boldsymbol{\Xi}_t$ as $\boldsymbol{\Omega}_t$ is diagonal.

In general, this lack of identification poses no problem for Bayesian estimation and prediction with this model provided a proper prior is used. Furthermore, if the researcher wishes to impose identification, this can easily be done in standard cases. For instance, beginning with Cogley and Sargent (2005), many popular Bayesian VARs have assumed the ω_{it} to follow stochastic volatility (SV) processes. Fixing the SV initial conditions to be zero suffices to identify the model without restricting the initial values for the error covariance matrix.

The reasons why we adopt this additive error structure relate to efficient computation and order invariance. To explain these points, we first summarize some key issues in the Bayesian VAR literature that are particularly acute with larger VARs. The traditional reduced form VAR is given by (1) but with $\boldsymbol{\nu}_t$ set to zero.[1] Direct Bayesian estimation of this reduced form VAR is computationally challenging when M is large due to the need to carry out matrix manipulations involving the very high dimensional posterior covariance matrix of $\boldsymbol{A}$. Accordingly, it is common to carry out Bayesian estimation in a structural form involving the use of the Cholesky decomposition of the reduced form error covariance matrix. That is, decomposing $\boldsymbol{\Sigma} = \boldsymbol{BDB}'$ where $\boldsymbol{B}$ is lower triangular with ones on the diagonal and $\boldsymbol{D}$ is diagonal, the structural VAR is obtained by multiplying both sides of the VAR by $\boldsymbol{B}^{-1}$. The structural VAR has a diagonal error covariance matrix, which means estimation can be carried out one equation at a time. This leads to huge computational improvements. For instance, in the specification used in Carriero et al. (2019), the MCMC algorithm based on the reduced form VAR requires $O(M^6)$ elementary operations to take one draw of the VAR coefficients but only $O(M^4)$ with the structural VAR. Thus, there are enormous computational benefits from working with VARs specified in such a way as to allow equation-by-equation estimation.

However, Bayesian results using the structural VAR based on the Cholesky decomposition are order dependent (i.e., posterior and predictive results depend on the way the variables are ordered in the VAR). This contrasts with the reduced form VAR for which standard implementations (e.g., use of an inverse Wishart prior for $\boldsymbol{\Sigma}$) are order invariant. The empirical importance of ordering issues has been investigated in papers such as Arias et al. (2023) and Chan et al. (2021) and found to be substantive, particularly in the case of large VARs and particularly for predictive variances and higher order predictive moments. Thus, most Bayesian VAR papers either work in reduced form, and face computational challenges unless the VAR dimension is very low, or work in structural form and produce empirical results that depend on the way the variables are ordered in the VAR.

If we now return to our VAR with additive errors in (1), it is straightforward to show that it suffers neither of these drawbacks. Computationally efficient MCMC algorithms can be developed that exploit the fact that (conditional on $\boldsymbol{\epsilon}_t$) the equations are independent of one another. But marginally (i.e., after integrating out $\boldsymbol{\epsilon}_t$), the shocks to $\boldsymbol{y}_t$ are cross-sectionally correlated. In relation to order-invariance, it is worth emphasizing that the ordering issue does not relate to the likelihood (i.e., the structural and reduced form VARs lead to the same likelihood function) but rather relates to the prior that is placed on the error variances and covariances. Indeed Carriero et al. (2019) refer to it as the "prior ordering issue." In our additive error setup, order invariance can be achieved by retaining an inverse Wishart prior for $\boldsymbol{\Sigma}$. As $\boldsymbol{\Omega}_t$ is diagonal any conventional set of priors will lead to order invariance. For instance, in the homoskedastic version of the model, assuming ω_i for $i = 1 \ldots M$ to have inverse-Gamma priors that are independent across i leads to order invariance. In the heteroskedastic case, assuming ω_{it} to have independent SV processes leads to order invariance as well.

This additive specification, however, also has a drawback. As opposed to a standard Cholesky-type decomposition of the reduced-form covariance matrix of the shocks, our specification implies that $\boldsymbol{\Omega}_t$ only impacts the main diagonal elements of $\boldsymbol{\Xi}_t$ and thus does not scale up the covariances between shocks accordingly. Hence, in periods where $\boldsymbol{\Omega}_t$ becomes large (such as during a recession), the covariances between shocks are not scaled up accordingly and thus decrease in relative importance. As empirical evidence shows that shocks tend to co-move in turbulent times this feature could be detrimental for forecasting accuracy. Below we will discuss how our nonparametric model can solve this shortcoming.

In subsequent sections, we will work with a nonparametric version of this model and provide full details of the priors we use and our computationally efficient MCMC algorithm. We stress that the issues discussed in this subsection also hold with nonparametric VARs. However, this subsection provides the basic insights into how these computational benefits are achieved and why our prior is order invariant. In addition, it may be found useful by Bayesian VAR researchers who are happy to remain parametric and work with linear VARs with Gaussian shocks. Specifying the VAR as we have done, with

[1]Typically, the assumption of homoskedasticity is relaxed, but allowing for this does not affect the arguments in this subsection.

additive errors, is an attractive way of achieving fast, order invariant inference even in conventional VARs. Although the heteroskedastic version of this specification has the possibly undesirable feature that error covariances decline in relative importance if $\boldsymbol{\Omega}_t$ becomes large.

2.2 | Allowing for shocks of unknown form

The model in the preceding subsection had a linear conditional mean and Gaussian error structure. In many contexts, linearity and Gaussianity can be restrictive, and this is particularly so in extreme times such as the recent Covid-19 pandemic. In this paper, we focus on relaxing the Gaussianity assumption relating to the shocks hitting the model. We will maintain the assumption of a linear conditional mean and assume a standard SV process for $\boldsymbol{\Omega}_t$.[2] In particular, our nonparametric VAR is the same as the one specified in the preceding section except for assumptions relating to $\boldsymbol{\epsilon}_t$. The assumption of Gaussianity will be replaced by a Dirichlet process mixture (DPM) of Gaussians. We will show that by this simple extension, we will achieve great gains in empirically relevant flexibility. But because the error process remains conditionally Gaussian, the benefits discussed in the preceding subsection (i.e., equation-by-equation estimation and order invariance) will be retained.

The ideas underlying our treatment of $\boldsymbol{\epsilon}_t$ are inspired by papers such as Frühwirth-Schnatter et al. (2004) and Dunson and Xing (2009), which develop parametric and nonparametric Bayesian treatments of random effects in panel data models, and accordingly, we refer to $\boldsymbol{\epsilon}_t$ as a vector of random effects. We model the random effects by introducing a base measure $\mathcal{G}$ and defining a parametric family of component densities f with unknown parameters $\boldsymbol{\vartheta}$:

$$p(\boldsymbol{\epsilon}_t) = \int f(\boldsymbol{\epsilon}_t|\boldsymbol{\vartheta})\mathcal{G}(d\boldsymbol{\vartheta}) = \sum_{j=1}^{\infty} \eta_j f(\boldsymbol{\epsilon}_t|\boldsymbol{\vartheta}_j),$$

where the weights $\sum_{j=1}^{\infty} \eta_j$ sum to 1. We assume that the component densities f are Gaussian with $M \times 1$ mean vector $\boldsymbol{\mu}_j$ and $M \times M$ variance-covariance matrix $\boldsymbol{\Sigma}_j$, which implies

$$p(\boldsymbol{\epsilon}_t) = \sum_{j=1}^{\infty} \eta_j f_{\mathcal{N}}(\boldsymbol{\epsilon}_t|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

with $f_{\mathcal{N}}$ denoting the density of the multivariate Gaussian distribution.

DPMs can be written (see, e.g., Escobar & West, 1995) in terms of a discrete latent random variable $\delta_t \in \{1, 2 \ldots\}$, with $\text{Prob}(\delta_t = j) = \eta_j$, that indicates which mixture component to adopt at time t. Thus, the DPM assumption implies the VAR errors are $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}_M, \boldsymbol{\Xi}_t)$ and time-varying error covariance matrix $\boldsymbol{\Xi}_t = \boldsymbol{\Sigma}_{\delta_t} + \boldsymbol{\Omega}_t$ where $\boldsymbol{\Sigma}_{\delta_t} = \boldsymbol{\Sigma}_j$ if $\delta_t = j$. Thus,

$$\boldsymbol{y}_t = \boldsymbol{\mu}_{\delta_t} + \boldsymbol{A}\boldsymbol{X}_t + \underbrace{\boldsymbol{Q}_{\delta_t}\boldsymbol{w}_t + \boldsymbol{v}_t}_{\boldsymbol{\epsilon}_t} \tag{2}$$

where $\boldsymbol{Q}_{\delta_t}$ is any matrix with the property that $\boldsymbol{\Sigma}_{\delta_t} = \boldsymbol{Q}_{\delta_t}\boldsymbol{Q}'_{\delta_t}$. Various choices for this decomposition of a covariance matrix are possible (e.g., eigendecomposition or Cholesky decomposition).[3]

This model has several properties that make it attractive for use in empirical macroeconomics. It retains a conditionally (i.e., conditional on δ_t) Gaussian structure that leads to simplicity of computation and structural economic interpretation. However, it is extremely flexible as infinite mixtures of Gaussians can approximate any distribution. Notice that $\boldsymbol{\Xi}_t$ varies

[2]Adopting nonparametric approaches for either of these can easily be done. For instance, Huber and Rossini (2020), Huber et al. (2020) and Clark et al. (forthcoming, Clark et al. (forthcoming) model the conditional mean of a VAR nonparametrically using regression trees. The last of these papers also uses regression trees to model the conditional variance. Approaches such as these could be added to the model of the present paper if extra flexibility is desired. However, as will be demonstrated below our model is already very flexible and can model any of the empirical regularities common with macroeconomic data.

[3]Note that if we use the Cholesky decomposition in this manner it does not undermine the order invariance of our model. It is merely used as a step in writing the likelihood function (which is order invariant) to facilitate interpretation and estimation. See subsection 3.1 of Carriero et al. (2019) for a discussion of this issue and why VAR ordering issues relate to the prior used on the error covariance matrix.

across components in the DPM. Depending on the estimated values for δ_t this allows for a wide variety of behavior (i.e., structural breaks, regime switching, and outliers) in the contemporaneous relationships between the elements in $\boldsymbol{y}_t$.

By additionally allowing for SV (via our specification for $\boldsymbol{\Omega}_t$), we have an error process of great flexibility allowing for impulse responses and other structural features to differ over time in a way that is estimated from the data. That is, both δ_t and $\boldsymbol{\Omega}_t$ allow for different types of parameter change, the latter only changing the error variances and being of a smooth nature, with the former additionally relating to the error covariances and allowing for more abrupt types of regime change or structural break. To see this feature, notice that a typical main diagonal element of $\boldsymbol{\Xi}_t$ is

$$[\boldsymbol{\Xi}_t]_{ii} = \sigma^2_{ii,\delta_t} + \omega_{it}.$$

In this equation, σ^2_{ii,δ_t}, the (i,i)th element of $\boldsymbol{\Sigma}_{\delta_t}$, changes abruptly over time and is thus capable of handling large outliers whereas ω_{it} changes smoothly and thus captures slowly varying trends in the error variances. Using only the latter implies that in the presence of large shocks, the SV model would only slowly adapt and would thus imply a higher variance when the large shock has already faded out. This model resembles the SV with outliers model of Carriero et al. (2022).

These properties mean the restrictions on the relations between volatilities and error covariances of the specification that involves a single Gaussian distribution for $\boldsymbol{\nu}_t$ noted in the previous subsection (i.e., that covariances between shocks would not scale up when volatility increases) are relaxed. To see this more clearly, suppose there is an economic event that causes both volatilities and the covariances between shocks to increase. Our model would react by selecting a Gaussian distribution with a larger covariance matrix. If we were to use a single Gaussian distribution, the corresponding covariances would remain as is, and we would thus systematically underestimate the correlations between the shocks. Under our more flexible mixture specification, which implies a larger choice for $\boldsymbol{\Sigma}_{\delta_t}$ and thus a matrix $\boldsymbol{\Xi}_t$ that would imply larger error variances but also stronger correlations across shocks.

The conditional representation of the model also gives insights on how the DPM handles location shifts in the shocks. To see this, note that we allow for the intercept to change over time in a nonparametric manner. Models with time-varying intercepts are common in macroeconomics (see, e.g., Stock & Watson, 2007 and Antolin-Diaz et al., 2017). Traditionally, intercepts have assumed to follow a random walk. However, our nonparametric treatment allows us to uncover the form of parameter change from the data (see also Hauzenberger et al., 2022, for a related but parametric treatment of time-varying parameter regressions).

It is also worth noting that an alternative model for nonparametric shocks would omit the additive error structure by setting $\boldsymbol{\nu}_t = 0$ and simply have one vector of errors that is modeled using a DPM. But this apparently simpler form would both be more computationally demanding (i.e., because order invariant equation-by-equation estimation would be difficult to achieve) and would omit the SV process. That is, a DPM model for the errors on its own would be very flexible at modeling structural breaks and outliers, but the assumption that the DPM errors are independent over time means that it is less able to model gradual changes in volatility. In contrast, our model combines the benefits of a very flexible shock distribution with the gradual volatility change of an SV process.

3 | BAYESIAN INFERENCE IN THE VAR WITH NONPARAMETRIC SHOCKS

3.1 | Prior

In this subsection, we describe our prior. We emphasize that the innovations in this paper relate to the parameters in the random effects. For the remaining parameters, any standard Bayesian priors can be used. In this paper, we use the Normal-Gamma prior of Griffin and Brown (2010) for the VAR coefficients (see, e.g., Huber & Feldkircher, 2019) although any other common Bayesian VAR prior could be used (e.g., the Minnesota prior or a global local shrinkage prior such as the Horseshoe).

For the error variances of the log-volatilities $\log \omega_{it}$, $\sigma^2_{\omega,i}$, we use Gamma priors. The prior implies shrinkage toward homoskedasticity. This can be seen by noting that

$$\sigma^2_{\omega,i} \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2c_{\omega,i}}\right) \Leftrightarrow \pm\sigma_{\omega,i} \sim \mathcal{N}(0, c_{\omega,i}).$$

We then use yet another Gamma prior on $c_{\omega,i} \sim \mathcal{G}(\vartheta_\omega, \vartheta_\omega \lambda_\omega/2)$. This prior features M equation-specific shrinkage parameters, a global shrinkage factor $\lambda \sim \mathcal{G}(0.01, 0.01)$ that pulls all state innovation variances toward zero and a hyperparameter $\vartheta_\omega = 0.6$. It allows to decide, in a data-based manner, whether heteroskedastic measurement errors are relevant or not.

We also consider a homoskedastic version of our model where $\boldsymbol{\Omega}_t = \boldsymbol{\Omega}$ for all t and in this case assume priors $\omega_i \sim \mathcal{G}^{-1}(a_\omega, b_\omega)$ for $i = 1 \ \dots \ M$ with $a_\omega = b_\omega = 10^{-3}$.

Our prior for the mean of the random effects is similar to one developed in Malsiner-Walli et al. (2017). It is a Gaussian prior on $\boldsymbol{\mu}_j$ that shrinks the different elements of $\boldsymbol{\mu}_j$ toward a common location:

$$\boldsymbol{\mu}_j \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{B}_0) \text{ for } j = 1, \ \dots \ , \infty,$$

with $\boldsymbol{B}_0 = \text{diag}(b_1, \ \dots \ , b_M)$ being a diagonal prior variance matrix with

$$b_j \sim \mathcal{G}(c_b, d_b).$$

The hyperparameters c_b, d_b are greater than zero. This is the Normal-Gamma prior proposed in Griffin and Brown (2010). If $c_b = 1$, we obtain the LASSO (Park & Casella, 2008). However, the LASSO is known to overshrink significant signals and undershrink irrelevant ones. Hence, we set $c_b = d_b = 0.6$. This leads to a model that implies more shrinkage and flexible tail behavior. As our prior is fully hierarchical, we also require another prior on $\boldsymbol{\mu}_0$. This is assumed to be $\mathcal{N}(\boldsymbol{0}, c^{-1}\boldsymbol{I}_M)$ with $c \to 0$, yielding a noninformative prior. In all our empirical work, we set $c = 10^{-3}$ to render the prior relatively noninformative but proper.

The combination of a flexible shrinkage prior that forces the component-specific means toward a common location has implications on the clustering behavior of the mixture model. To illustrate this, let μ_{ji} denote the jth element of $\boldsymbol{\mu}_i$ for $i = k$ or $\tilde{k}$ (i.e., these are the two intercepts in the jth equation for two different components of the infinite mixture $k \neq \tilde{k}$). The prior above implies the following in terms of the distance between μ_{jk} and $\mu_{j\tilde{k}}$ (see Yau & Holmes, 2011):

$$\frac{(\mu_{jk} - \mu_{j\tilde{k}})}{\sqrt{2}} \sim \mathcal{N}(0, b_j).$$

Thus, our prior is centered over intercept homogeneity and b_j controls the strength of this belief. If b_j is close to zero, the intercepts collapse to a common value (which is μ_{0j}, the jth element of $\boldsymbol{\mu}_0$). For larger b_j, we allow for more heterogeneity in the intercepts. This feature is crucial as the presence of the nonzero location parameter allows us to capture skewness in the shocks. If we use a noninformative prior on the component means, we would risk overfitting the data. Our shrinkage prior effectively enables us to investigate how much asymmetries are in the data in a fully automatic manner.

For the covariance matrices for each component in the DPM, we use a conjugate Wishart prior on $\boldsymbol{\Sigma}^{-1}$ as this leads to order invariance within each component that implies order invariance in the VAR as a whole. Thus, we assume

$$\boldsymbol{\Sigma}_k^{-1} \sim \mathcal{W}(c_0, \boldsymbol{\Sigma}_0^{-1}).$$

Note that we parameterize the Wishart such that the prior mean equals $c_0\boldsymbol{\Sigma}_0^{-1}$ with c_0 being its degrees of freedom. The prior hyperparameters can be chosen in any way. In our empirical work, we use a relatively noninformative data-based prior inspired by the Minnesota prior. In particular, we set the prior degrees of freedom as $c_0 = M + 4$. The prior scaling matrix is estimated from the data and set as $\boldsymbol{\Sigma}_0 = \text{diag}(\hat{\sigma}_1^2, \ \dots \ , \hat{\sigma}_M^2)$ where $\hat{\sigma}_j^2$ are the OLS error variances obtained by running an AR(p) model for y_{jt}.

In terms of the weights in the DPM, we use a stick breaking process (SBP) prior on η_j. The SBP prior introduces additional auxiliary random variables ν_j (called sticks) such that the weights are obtained sequentially:

$$\eta_1 = \nu_1, \ \eta_j = \nu_j \prod_{i=1}^{j-1}(1 - \nu_j), \ \nu_j \sim \mathcal{B}(1, \alpha).$$

The parameter α determines the clustering behavior of the mixture model. To see this, notice that the prior probability of forming a new cluster when assigning $\boldsymbol{\varepsilon}_t$ conditional on all $\boldsymbol{\varepsilon}_\tau (\tau \neq t)$ is (Frühwirth-Schnatter & Malsiner-Walli, 2019; Lau & Green, 2007)

$$\frac{\alpha}{T-1+\alpha}, \tag{3}$$

and thus decreases in T. This implies that the DPM has the potential to create few larger clusters and then has a low probability of opening up new clusters that are populated by relatively few observations. In macroeconomic data, this behavior might be necessary to single out events that are different from those produced by the DGP in normal times such as the Covid-19 pandemic. As α crucially impacts this behavior, we treat it as an unknown parameter and estimate it from the data. We assume that α arises from a Gamma distribution a priori, that is, $\alpha \sim \mathcal{G}(2,4)$, which implies a prior mean 0.5 and a prior variance 0.125. This choice was originally suggested by Escobar and West (1995) and encourages clustering behavior of the mixture model.

3.2 | Posterior simulation of the VAR coefficients and random effects

In this subsection, we describe how we simulate the VAR coefficients and random effects in more detail. The other steps are relatively standard, and we sketch them in the next subsection. A key theme of this paper is computational efficiency and, to achieve this end, we need to draw the VAR coefficients one equation at a time. This requires knowledge of the random effects that serve to establish correlations across the shocks. Accordingly, we describe how we draw the VAR coefficients and the random effects in more detail.

Conditional on the random effects $\{\boldsymbol{\epsilon}_t\}_{t=1}^T$, we draw the equation-specific VAR coefficients $\boldsymbol{A}_i$ from

$$\boldsymbol{A}_i|\bullet \sim \mathcal{N}(\boldsymbol{m}_i, \boldsymbol{V}_{a,i}),\ i = 1, \ldots, M,$$

where $\bullet$ denotes all arguments necessary to define the full conditional posterior distribution and

$$\begin{aligned}\boldsymbol{V}_{a,i} &= (\tilde{\boldsymbol{X}}_i'\tilde{\boldsymbol{X}}_i + \underline{\boldsymbol{V}}_{a,i}^{-1})^{-1},\\ \boldsymbol{m}_i &= \boldsymbol{V}_{a,i}\tilde{\boldsymbol{X}}_i'\tilde{\boldsymbol{Y}}_i,\end{aligned}$$

denote the posterior moments with $\tilde{\boldsymbol{X}}_i$ being a $T \times K$ matrix with typical tth row $\boldsymbol{X}_t/\sqrt{\omega_{it}}$, $\underline{\boldsymbol{V}}_{a,i}$ denotes the prior variance matrix and $\tilde{\boldsymbol{Y}}_i$ denotes a $T \times 1$ vector with typical element given by $(y_{it} - \epsilon_{it})/\sqrt{\omega_{it}}$ where ϵ_{it} is the ith element of $\boldsymbol{\epsilon}_t$. For large models (characterized by $T \ll K$), the inversion of the posterior covariance matrix becomes computationally cumbersome. Accordingly, one can use the algorithm of Bhattacharya et al. (2016). This has computational complexity of $O(T^2K)$ instead of $O(K^3)$, which speeds up computation enormously.[4]

Next we describe the sampling steps involved in simulating the random effects in more detail. The key point to notice is that the random effects are conditionally independent over time and hence $p(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T|\bullet) = \prod_{t=1}^T p(\boldsymbol{\epsilon}_t|\bullet)$ with time t posteriors given by

$$\boldsymbol{\epsilon}_t|\bullet \sim \mathcal{N}(\overline{\boldsymbol{\epsilon}}_t, \overline{\boldsymbol{V}}_{\epsilon,t}), \tag{4}$$

and moments given by

$$\overline{\boldsymbol{V}}_{\epsilon,t} = \boldsymbol{\Sigma}_{\delta_t} - \boldsymbol{\Sigma}_{\delta_t}(\boldsymbol{\Sigma}_{\delta_t} + \boldsymbol{\Omega}_t)^{-1}\boldsymbol{\Sigma}_{\delta_t},\ \overline{\boldsymbol{\epsilon}}_t = \boldsymbol{\mu}_{\delta_t} + \boldsymbol{\Sigma}_{\delta_t}(\boldsymbol{\Sigma}_{\delta_t} + \boldsymbol{\Omega}_t)^{-1}(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{X}_t - \boldsymbol{\mu}_{\delta_t}).$$

These moments directly follow from the fact that $\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{X}_t$ and $\boldsymbol{\epsilon}_t$ are jointly Gaussian and the conditional distribution of $\boldsymbol{\epsilon}_t|\hat{\boldsymbol{y}}_t$ is given by Equation (4). An interesting special case arises if we set $\boldsymbol{\Omega}_t = 0 \forall t$. This results in a degenerate Gaussian distribution for the error with posterior covariance $\overline{\boldsymbol{V}}_{\epsilon,t} = 0$ and the posterior mean reduces to $\overline{\boldsymbol{\epsilon}}_t = (\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{X}_t)$. Hence, we would end up with a standard VAR.

[4]This algorithm requires that the prior covariance matrix is simple to invert. This is the case in our framework.

3.3 | Full conditional MCMC sampling

Our MCMC sampler is relatively straightforward and contains several steps that involve standard full conditional distributions. Here, we summarize the different updating steps necessary to sample from the joint distribution of the latent states and coefficients of the model.

Our sampler iterates between the following steps:

1. Sample $\boldsymbol{A}_i|\bullet$ for each equation from $p(\boldsymbol{A}_i|\bullet) = \mathcal{N}(\boldsymbol{m}_i, \boldsymbol{V}_{a,i})$ as described in the previous subsection.
2. Sample $\boldsymbol{\Sigma}_k^{-1}$ from $p(\boldsymbol{\Sigma}_k^{-1}|\bullet) = \mathcal{W}(\bar{c}_k, \overline{\boldsymbol{\Sigma}}_k^{-1})$. The posterior degrees of freedom are $\bar{c}_k = c_0 + T_k/2$ where $T_k = \sum_{t=1}^{T} \mathbb{I}(\delta_t = k)$ denotes the number of observations allocated to cluster k. The posterior scaling matrix is given by

$$\overline{\boldsymbol{\Sigma}}_k = \frac{1}{2} \sum_{t:\delta_t=k} (\tilde{\boldsymbol{y}}_t - \boldsymbol{\mu}_{\delta_t})(\tilde{\boldsymbol{y}}_t - \boldsymbol{\mu}_{\delta_t})' + \boldsymbol{\Sigma}_0^{-1},$$

   where $\tilde{\boldsymbol{y}}_t = \boldsymbol{y}_t - \boldsymbol{A}'\boldsymbol{x}_t - \boldsymbol{v}_t$.
3. Sample $\boldsymbol{\mu}_k$ from $p(\boldsymbol{\mu}_k|\bullet) = \mathcal{N}(\overline{\boldsymbol{\mu}}_k, \overline{\boldsymbol{V}}_{\mu,k})$. The posterior moments of the random intercept terms are given by

$$\overline{\boldsymbol{V}}_{\mu,k} = \left(\boldsymbol{\Sigma}_k^{-1} T_k + \boldsymbol{B}_0^{-1}\right)^{-1},$$
$$\overline{\boldsymbol{\mu}}_k = \overline{\boldsymbol{V}}_{\mu k} \left(\boldsymbol{\Sigma}_k^{-1} \sum_{t:\delta_t=k} \tilde{\boldsymbol{y}}_t + \boldsymbol{B}_0^{-1} \boldsymbol{\mu}_0\right).$$

4. Sample $\boldsymbol{\mu}_0$ from $p(\boldsymbol{\mu}_0|\bullet) = \mathcal{N}(\overline{\boldsymbol{\mu}}_0, \overline{\boldsymbol{V}}_{\mu,0})$ with

$$\overline{\boldsymbol{V}}_{\mu,0} = \frac{1}{J}\boldsymbol{B}_0, \ \overline{\boldsymbol{\mu}}_0 = \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{\mu}_j.$$

5. Sample $\{\boldsymbol{\epsilon}_t\}_{t=1}^{T}$ from $p(\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_T|\bullet) = \prod_{t=1}^{T} \mathcal{N}(\overline{\boldsymbol{\epsilon}}_t, \overline{\boldsymbol{V}}_{\epsilon t})$ with moments given below Equation (4).
6. Sample b_i (for $i = 1, \dots, M$) from $p(b_j|\bullet) = \text{GIG}(p_0, p_K, z_j)$ with $p_0 = 2d_b$, $p_K = c_b - J/2$ and $z_j = \sum_{j=1}^{J} (\mu_{j,i} - \mu_{0,i})^2$ and GIG denoting the generalized inverse Gaussian distribution.[5]
7. Sample α using a Metropolis–Hastings step. As proposal distribution, we use the log-normal distribution centered on the previously accepted value of $\log \alpha$, denoted by $\log \alpha^{(a)}$. More formally, the proposal distribution reads: $\log \alpha^* \sim \mathcal{N}(\log \alpha^{(a)}, c_\alpha^2)$ with c_α denoting a scaling parameter. We set this scaling parameter to achieve an acceptance probability between 40% and 60%. This is done by using the first 25% of the burn-in stage to tune c_α accordingly.
8. Sample $\omega_{1t}, \dots, \omega_{Mt}$ from $p(\omega_{1t}, \dots, \omega_{Mt}|\bullet)$. For the SV case, $\omega_{j1}, \dots, \omega_{jT}$ (for $j = 1, \dots, M$) and the parameters of the state equation are drawn using the algorithm outlined in Kastner and Frühwirth-Schnatter (2014) and implemented in the `R` package `stochvol`. When we assume homoskedasticity, we simply draw ω_j from an inverse Gamma posterior.
9. Sample $v_1, \dots, v_{J-1}$ from $p(v_1, \dots, v_{J-1}|\bullet) = \prod_{j=1}^{J} \mathcal{B}\left(1 + T_j, \alpha + \sum_{l=j+1}^{J} T_l\right)$.
10. Sample $\delta_1, \dots, \delta_T$ from $p(\delta_1, \dots, \delta_T|\bullet) = \prod_{t=1}^{T} p(\delta_t|\bullet)$ using the Slice sampler (Kalli et al., 2011) in two steps. First, let $u_t|\delta_t \sim \mathcal{U}(0, \zeta_{\delta_t})$ denote a set of auxiliary random variables with $\zeta_k = (1-w)w^{k-1}$ and $w = 0.8$. Then, conditional on u_t, we simulate δ_t from its discrete distribution as follows:

$$\text{Prob}(\delta_t = k|\bullet) \propto \frac{\mathbb{I}(u_t < \zeta_k)}{\zeta_k} \eta_k f_{\mathcal{N}}(\tilde{\boldsymbol{y}}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Notice that J is a truncation parameter that determines the effective number of regimes that is obtained by solving $1 - \sum_{j=1}^{J} \zeta_j < \min(u_1, \dots, u_T)$. This implies that our infinite mixture model becomes effectively finite dimensional and thus computationally tractable. As an important special case, we obtain the VAR described in Section 2.1 by fixing $J = 1$. We will refer to this model as BVAR–J1.

[5]The density of the GIG is given by $f(x) = x^{\lambda-1} e^{-\frac{1}{2}(\chi/x + \psi x)}$.

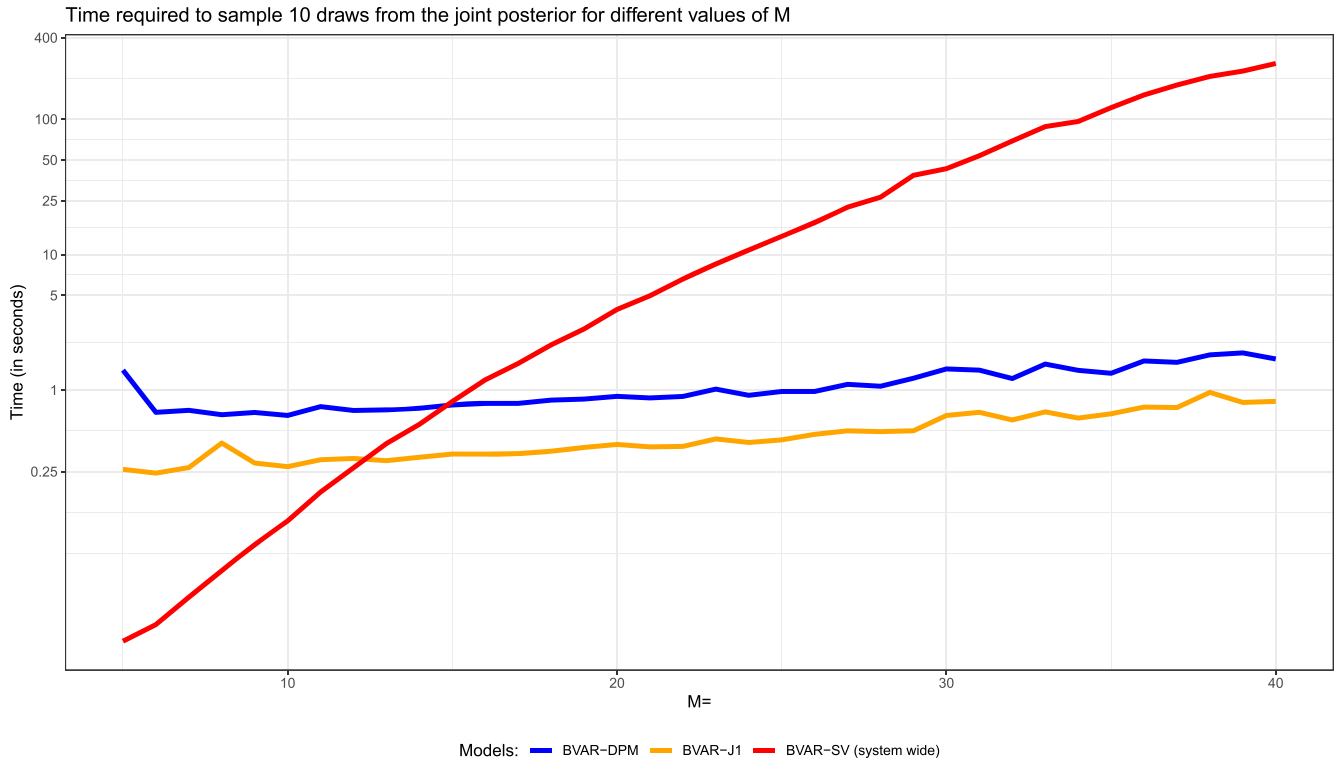

FIGURE 1 Comparison of computation times between BVAR-DPM, BVAR-J1 and BVAR-SV (full system) for $M \in \{5, \ldots, 40\}$. The y-axis is in log scale.

In most of our applications, we repeat this algorithm 20,000 times and discard the first 10,000 draws as burn-in. For the actual data application and in simulations, we did not encounter any mixing issues when we consider functions of the parameters such as impulse responses or forecast densities.

Before applying our model to the data, it is worth stressing that the mixture model is not identified with respect to relabeling the latent discrete indicators. This does not cause any issues if interest is exclusively on either unconditional impulse responses or forecast distributions. In our structural application, we discuss impulse responses conditional on a given cluster. To avoid label switching in this case, we introduce restrictions that ensure a unique ordering ex-post. Specifically, we post process the posterior draws by sorting the mixture probabilities in a descending order and then relabeling the corresponding mixture model accordingly. This implies that the first cluster is always the one with the most observations.

3.4 | Computational aspects of our algorithm

The algorithm described in the previous subsection is efficient and has the same complexity as the original (but incorrect) version of the algorithm proposed in Carriero et al. (2019) and its corrected variant proposed in Carriero et al. (2022).[6] This is because we sample from the equation-specific posteriors. This implies that the posterior covariance matrices are $K \times K$ and inversion of such matrices has computational complexity $O(K^3)$ (or $O(T^2K)$ if $T \ll K$). A similar computational advantage can be obtained by applying the algorithm outlined in Kastner and Huber (2021) that uses a factor model to render equation-by-equation estimation possible. In any case, all these algorithms are much more efficient than the one based on treating the VAR as a full system of equations in large dimensions.

Figure 1 illustrates the magnitude of efficiency improvements of equation-by-equation estimation relative to system wide estimation and the increase in computation required to add the DPM. It compares our model (BVAR-DPM) to a restricted version of it, which is Gaussian and obtained by setting $J = 1$ (BVAR-J1) and a BVAR with SV but estimated

[6]Notice that Carriero et al. (2019) do not use the efficient sampling algorithm of Bhattacharya et al. (2016) to simulate from the equation-specific coefficient posteriors if $T < K$.

TABLE 1 Empirical distribution of inefficiency factors based on 1000 (retained) draws from the posterior of $\boldsymbol{A}$ using system wide estimation and our proposed approach.

	Minimum	25%	50%	75%	Maximum
BVAR-DPM	1	1.15	1.40	1.96	8.85
BVAR-SV (system wide)	1	1.74	2.20	2.76	8.77

using system wide estimation (labeled BVAR-SV (system wide)). All models feature a single lag. Figure 1 compares the computation times necessary to generate 10 draws from the posterior. Note that the y-axis of the figure is in log scale.

The figure shows that BVAR-DPM and BVAR-J1 are faster than system wide estimation when M exceeds 12 (in the case of BVAR-J1) and when M exceeds 14 (in the case of the BVAR-DPM). In larger panels, full system estimation of the BVAR-SV becomes much slower and the computational burden increases substantially. Differences between the blue and orange line reflect the additional computational burden from adding the DPM piece to the model. In most empirical work, the effective number of regimes is small (i.e., the infinite mixture reduces to a four-components mixture of Gaussians). This implies that if the true number of regimes is small, the computational burden would also decrease under the DPM specification.

The algorithms of Carriero et al. (2019) and Carriero et al. (2022) have the same computational complexity as our approach that sets $J = 1$. In this case, computation times would be very similar (and the shape of the curves would, in fact, be identical) and our algorithm thus scales as well as theirs. However, it is worth stressing that ours is order invariant, can have nonparametric errors, and is applicable to any conditional mean function.

One question pertaining to our algorithm is whether it has favorable mixing properties. We illustrate this using the large US macroeconomic dataset we describe in Section 5.1 and using $p = 5$ lags. As our method is much faster than system wide estimation, we follow Carriero et al. (2019) and start by producing 1000 draws from the posterior of $\boldsymbol{A}$ using the system wide algorithm. This takes around 1.2 h. We then run our proposed algorithm for 1.2 h as well. This produces around 17,000 draws from the posterior distribution. As the chain from our algorithm is much longer, we thin the chain to have 1000 draws for both estimation methods. We then compare inefficiency factors (IFs) between both methods. The empirical quantiles of IFs over the elements of $\boldsymbol{A}$ are shown in Table 1.

This analysis reveals that both algorithms produce draws from $p(\boldsymbol{A}|\bullet)$ that mix well. The maximum IF is 8.85 in the case of the BVAR-DPM and around 8.77 for the BVAR-SV estimated using the system wide algorithm. Considering the quantiles of the empirical distribution shows that the IFs of the BVAR-DPM are always lower than the ones obtained from the BVAR-SV. Because IFs below 30 are generally viewed as acceptable (Primiceri, 2005), we take this as evidence that using our algorithm and adding the DPM piece yields an algorithm that mixes better than the system wide algorithm.

4 | ARTIFICIAL DATA EXERCISE

We illustrate the merits of our approach by simulating data from a set of different DGPs. These DGPs differ in terms of model size, the error distributions and the number of observations. With respect to model size, we consider three sizes that capture typical situations in applied macroeconomic work. The smallest DGP has $M = 5$ endogenous variables, the medium-sized DGP features $M = 10$ and the largest includes $M = 20$ endogenous variables.

With respect to the error distributions, we consider three different shock assumptions. Two of these three feature substantial departures from homoskedasticity and normality while one assumes Gaussian and homoskedastic shocks. All of these DGPs assume that the conditional mean of the process is given by

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t,\ \boldsymbol{y}_0 = \boldsymbol{0},\ t = 1, \dots, T,$$

where $\boldsymbol{A}$ has diagonal elements $A_{ii} = 0.75$ and off-diagonal elements sampled from a Gaussian distribution $A_{ij} \sim \mathcal{N}(0, 0.1^2)(i \neq j)$. To ensure stationarity, we reject draws of $\boldsymbol{A}$, which imply an unstable model.[7]

The errors $\boldsymbol{\epsilon}_t$ differ across the DGPs in the following four ways.

1. For the first DGP, we assume that $\boldsymbol{\epsilon}_t$ is multivariate Student t with three degrees of freedom and covariance matrix $\boldsymbol{W}$. The matrix $\boldsymbol{W}$ is created as follows. We specify a lower uni-triangular matrix $\boldsymbol{U}$ with $u_{ij} \sim \mathcal{N}(0, 0.1^2)$ for $i = 2, \dots, M; j = 1, \dots, M-1$ and then set $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{U}'$.

[7]This happens only in very rare cases.

2. The second DGP has a common stochastic volatility specification (Carriero et al., 2016). We assume that the shocks are Gaussian distributed with zero mean and time-varying covariance matrix $\boldsymbol{W}_t = e^{\mathfrak{s}_t} \times \boldsymbol{W}$. The log-volatility process $\mathfrak{s}_t$ evolves according to a random walk with innovation variance 0.25^2.
3. Finally, we also consider a homoskedastic DGP that assumes that $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W})$.

For all these DGPs, we simulate time series of three different lengths. The first one is a short sample of $T = 100$ observations. This reflects quarterly macroeconomic data commonly observed in, for example, the Euro Area. Then, we focus on a sample of $T = 250$ observations. This resembles a situation commonly faced when working with US quarterly macroeconomic data. Finally, we also consider longer time series with $T = 750$ observations. The long sample serves to analyze how the model would behave if the researcher uses monthly data.

For comparison, we consider the BVAR-DPM and the BVAR-DPM with $J = 1$ and set the number of lags equal to 5. All models are estimated with homoskedastic and SV measurement errors. This combination allows us to analyze whether the BVAR-DPM is capable of recovering non-Gaussian features in the DGP without overfitting. In particular, the final DGP can be used to focus on the question whether adding the nonparametric component to the model leads to a deterioration in estimation accuracy or whether the DPM can recover the simple Gaussian case without overfitting.

To rank models in terms of estimation accuracy, we focus on the mean absolute error (MAE) between the true set of coefficients and the posterior median of the estimated VAR coefficients. Table 2 shows the accuracy of the BVAR-DPM relative to the BVAR with Gaussian errors in terms of their estimation of the VAR coefficients. All results are means across MAEs from 50 replications from each of the DGPs.

In general, our results indicate that if the Gaussian-errored model is mis-specified, the DPM model is consistently more accurate. These accuracy gains increase with the sample size. This result is not surprising given that for, for example, t-distributed shocks, using a smaller sample implies a lower probability of observing outliers over time. Hence, in short samples, the series we simulate look like being generated with Gaussian shocks and the parameter estimates thus do not profit much from having a DPM specification in the shocks. With larger samples, more outliers mechanically show up and the increased sample size does not overweight the adverse effect these outliers have on parameter estimates.

When we consider differences across model sizes, we find that accuracy gains from using the full DPM mixture that potentially involves an infinite number of Gaussians relative to simply setting $J = 1$ decline with larger information sets. This finding points toward the fact that large models can soak up non-Gaussian features in the data.

In cases where the BVAR $J = 1$ is correctly specified (i.e., with the homoskedastic model and homoskedastic DGP and with the SV model with the SV DGP), relative MAEs are very close to one. In fact, for the homoskedastic and Gaussian DGP, both the DPM and the model with $J = 1$ produce identical parameter estimates.

This short discussion illustrates that the BVAR-DPM is capable of capturing model features such as skewness and fat tails, but when the DGP does not have such features it can successfully uncover the underlying Gaussian model with very little overfitting.

TABLE 2 Simulation results.

		Homoskedastic						**Stochastic volatility**					
		t		**SV**		**Homosk.**		**t**		**SV**		**Homosk.**	
$T \downarrow$	$M \downarrow$	**DPM**	$J = 1$	**DPM**	$J = 1$	**DPM**	$J = 1$	**DPM**	$J = 1$	**DPM**	$J = 1$	**DPM**	$J = 1$
100	5	0.97	1.50	0.91	2.39	1.00	1.55	1.03	1.12	0.67	0.69	1.03	1.09
	10	0.98	1.59	1.01	1.85	1.00	1.63	1.00	1.03	0.90	0.91	1.00	1.02
	20	0.99	2.02	1.00	2.26	1.00	2.07	1.00	1.01	0.93	0.93	0.99	0.99
250	5	0.77	1.39	0.74	4.49	1.00	1.59	0.80	1.03	0.36	0.37	0.99	1.12
	10	0.82	1.39	0.90	2.21	1.00	1.60	0.86	1.03	0.72	0.73	1.03	1.08
	20	0.97	1.35	1.00	1.86	1.00	1.60	0.93	1.00	0.85	0.85	1.01	1.03
750	5	0.47	1.28	0.66	7.61	1.00	1.48	0.47	0.86	0.21	0.22	1.02	1.21
	10	0.55	1.28	0.73	4.01	1.00	1.44	0.63	0.90	0.37	0.38	1.02	1.15
	20	0.70	1.18	0.75	3.49	1.00	1.23	0.72	0.92	0.42	0.43	1.01	1.11

Note: This table shows mean absolute error (MAE) ratios between BVAR-DPM and a BVAR that sets $J = 1$ and has homoskedastic measurement errors. Shaded columns are raw MAEs multiplied by 100. The MAEs are computed as the difference between the posterior median of the coefficients and the true VAR coefficients. All results are based on averaging over 50 replications from each DGP.

TABLE 3 Effective number of clusters.

		Homoskedastic			Stochastic volatility		
$T \downarrow$	$M \downarrow$	t	SV	Homosk.	t	SV	Homosk.
100	5	3.0	1.5	1.0	2.8	1.2	1.0
	10	2.6	1.4	1.0	2.6	1.3	1.0
	20	2.6	2.0	1.0	2.5	1.8	1.0
250	5	2.9	2.1	1.0	2.9	1.8	1.0
	10	2.8	1.9	1.0	2.8	2.0	1.0
	20	3.0	2.5	1.0	2.9	2.3	1.0
750	5	3.4	2.6	1.0	3.3	2.0	1.0
	10	2.9	2.4	1.0	2.8	2.3	1.0
	20	2.7	2.3	1.0	2.9	2.2	1.0

When we use the DPM, we can also infer the effective number of clusters. This is achieved as follows. For the ith run of our MCMC algorithm, we compute

$$J^{(i)} = \sum_{j}^{J} \mathbb{I}\left(T_j^{(i)} > 0\right).$$

We can then compute the posterior median of these runs to obtain an estimate of the number of clusters. Table 3 shows the mean over these posterior median estimates across the different realizations from the DGP. Note that, when the DGP is Gaussian and the DPM extension is unnecessary, our algorithm is correctly selecting $J = 1$.

When the DGP involves a Student's t distribution, we are finding an interesting pattern where the number of clusters is inversely related to the VAR dimension (this holds for $T \in \{100,750\}$). For $T = 250$, we find that the number of clusters is close to three for all values of M. This is consistent with our conjecture that, as the VAR dimension increases and more explanatory variables appear on the right hand side of each equation, the extra variables can fit some of the fat tailed behavior of the DGP.

When we use a DGP that features SV, we find a slightly lower number of clusters and a rather mixed pattern when it comes to the relationship between size and the number of clusters. Comparing our different specifications that either use SV on the measurement error variances or assume them to be homoskedastic reveals that if the former specification is adopted, the effective number of clusters declines slightly.

5 | EMPIRICAL APPLICATION USING US DATA

5.1 | Data and specifications

We use US quarterly macroeconomic data from 1960Q1 to 2023Q4 taken from the FRED database, see McCracken and Ng (2020). A full list of variables is given in Section A in Data S1. In our forecasting exercise, we evaluate forecast performance beginning in 1977Q1 and rely on the iterated method of forecasting and consider forecast horizons of one-quarter and 1 year.

We work with small ($M = 4$), medium ($M = 7$) and large ($M = 26$) dimensional VARs with the variables included in each being given in Data S1. All variables are transformed to be approximately stationary (with detailed information provided in Data S1). In the forecasting exercise, we evaluate the performance of the models by focusing on the variable-specific performance for GDP growth, the unemployment rate and inflation. These three variables form our set of focus variables. We choose a long lag length, $p = 5$, and trust our shrinkage prior to prevent overfitting.

In Section 5.4, we carry out a forecasting exercise to assess whether adding the DPM to the VAR improves predictive accuracy. To this end, we compare the performance of the model with nonparametric shocks (BVAR-DPM) to models with Gaussian errors that we obtain by taking our BVAR-DPM and setting $J = 1$. This is the model described in Section 2.1. All other specification details, including prior choice, are the same in all of our models. In this way, we can focus on the specific issue of what the use of the DPM adds to the model. Note that we are not also including a BVAR with full error covariance because, as established previously, it is computationally much more burdensome in larger models and is expected to give results very similar to the BVAR with $J = 1$. We consider two versions of every model, one with SV and one homoskedastic.

TABLE 4 Probability of a given effective number of clusters.

$J =$	1	2	3	4	5	6	7	8	9
	0.00	0.03	0.77	0.19	0.01	0.00	0.00	0.00	0.00

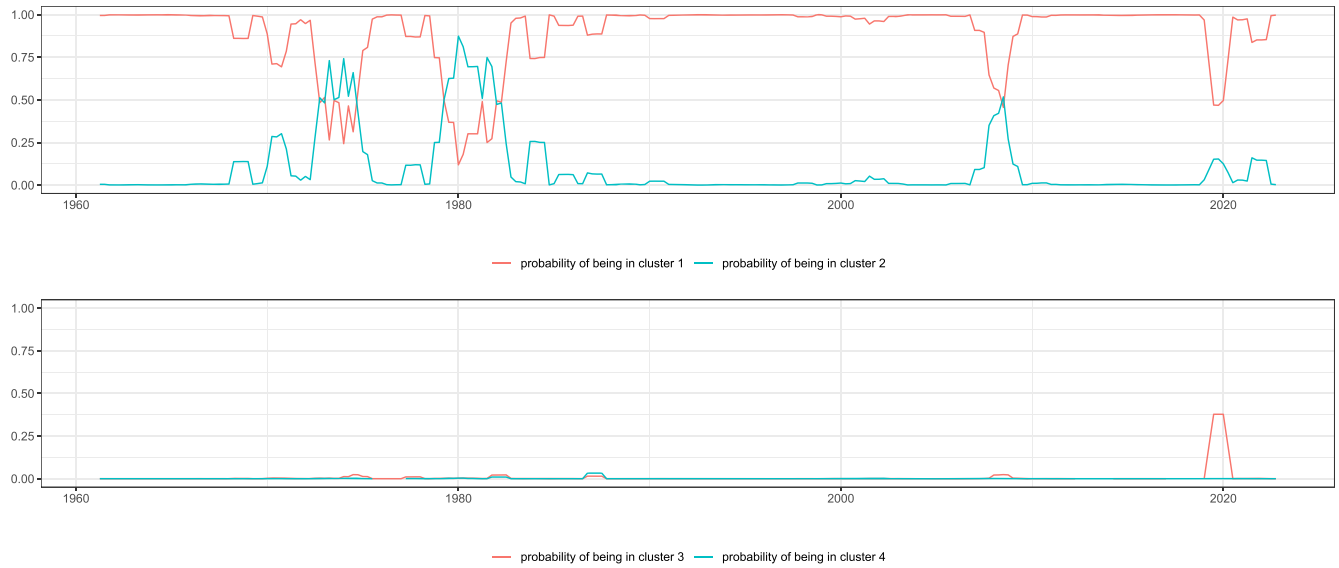

FIGURE 2 Probability of being in a given cluster over time.

5.2 | Full sample analysis

In order to illustrate the properties of our model, we begin by carrying out full sample estimation using a single model: the BVAR-DPM without SV using the medium data set. We use the homoskedastic version of the model because, as we will show in our forecasting experiment, this model performs well and simplifies interpretation of the impulse responses discussed in the next subsection.

An advantage of the DPM is that it can be used to estimate the effective number of components in the Gaussian mixture and our simulation results show that it does so accurately. Table 4 presents evidence relating to this. It provides quantitative information on the posterior distribution of the effective number of clusters. The table suggests that $J = 3$ is the most likely number of clusters and there is no probability associated with $J = 6$ or more clusters. We are finding no evidence in favor of the conventional Gaussian VAR that has $J = 1$.

The probabilities of being in one of the first four clusters estimated by the DPM over time can be seen in Figure 2. This figure shows the posterior probability that $\delta_t = j$ for $j = 1, \ldots, 4$. The probability of the fifth cluster will be one minus the sum of these lines and which is effectively zero so is not plotted here. To introduce persistence and simplify the discussion, we report yearly rolling averages of these probabilities. Recall that we solve the label switching issue by sorting the posterior draws ex post so that the mixture weights are descending.

Cluster 1 is predominant and holds with high probability in most periods. Cluster 2 is associated with times of high volatility such as periods in the mid 1970s and early 1980s as well as the financial crisis and, to a lesser extent, the pandemic. Cluster 3 is largely associated with the pandemic but plays a small, brief role in other times of high volatility. These are the three main clusters. The tiny amount of probability associated with Cluster 4 is in the early 1980s. Another interesting point to note is that, much of the time Cluster 1 applies with probability near one. But in more volatile times like the financial crisis and the pandemic, no single cluster holds with probability near one. For instance, early in the Covid-19 pandemic each of the main three clusters receives roughly equal probability. Thus, in normal times, a single Gaussian distribution suffices to model the error distribution, but in less stable times a mixture of two or more Gaussians is required.

A deeper understanding of the properties of the clusters can be obtained by looking at the posteriors of $\boldsymbol{\Sigma}_k$ for $k = 1 \ldots 5$. Figure 3 contains box and whisker plots of the posterior of the log of the determinant of each $\boldsymbol{\Sigma}_k$. The red horizontal line is the (log) determinant of $\boldsymbol{\Sigma}_0$.

It can be seen that Cluster 1 is the low volatility cluster whereas Clusters 2 and 3 have much higher volatility (as evidenced by much larger log-determinants). Clusters 2 through 3 also have much more uncertainty (e.g., wider credible

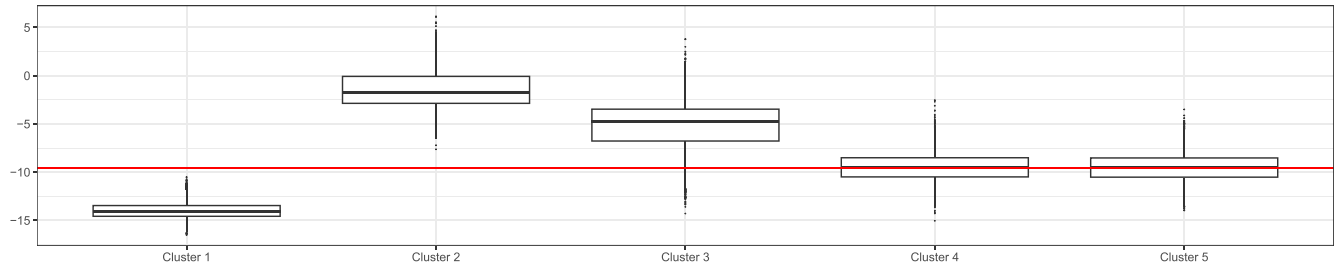

FIGURE 3 Boxplots of the posterior of the log-determinant of $\boldsymbol{\Sigma}_k$ and the log-determinant of $\boldsymbol{\Sigma}_0$ (red line)

intervals) than Cluster 1. Clusters 4 and 5 (which play a role very rarely and appear to be similar to one another) lie between the low volatility Cluster 1 and higher volatility Clusters 2 and 3.

Comparing the posterior of the determinants to the prior reveals that particularly in Clusters 1, 2, and 3, our model departs substantially from the prior while in the other cases, posterior medians are much closer to the prior. This is not surprising because Cluster 4 and 5 include only relatively few observations and are thus strongly influenced by the prior. Notice that the prior, being calibrated using the error variances from AR(5) regressions, implies variances that are too high in normal periods (i.e., the observations allocated to regime 1) and too low in turbulent times (i.e., the observations allocated to regimes 2 and 3).

5.3 | Structural impulse responses to a monetary policy shock

Next we use the medium-scale BVAR-DPM without SV to investigate the dynamic effects of monetary policy shocks using a standard identification scheme with a Cholesky ordering where the Federal Funds rate is ordered above financial market variables and below real quantities. This ordering implies that real quantities belong to the "slow-moving" block whereas financial markets are "fast-moving" (Bernanke et al., 2005). To economize on space, we focus on the IRFs of the three focus variables (output growth, unemployment, and inflation).

A feature of our approach is that, conditional on the mixture indicators δ_t, our model can be interpreted as a constant parameter VAR with a fully time-varying covariance matrix. This implies that the structural form of the model features time-varying parameters. To see this, multiply (2) by $\boldsymbol{\Psi}_t^{-1}$, the inverse of the lower Cholesky factor of $\boldsymbol{\Xi}_t$, from the left. This yields

$$\boldsymbol{\Psi}_t^{-1}\boldsymbol{y}_t = \tilde{\boldsymbol{A}}_t\boldsymbol{X}_t + \tilde{\boldsymbol{\varepsilon}}_t. \tag{5}$$

Here, we let $\tilde{\boldsymbol{A}}_t = \boldsymbol{\Psi}_t^{-1}\boldsymbol{A}$ and $\tilde{\boldsymbol{\varepsilon}}_t \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{I}_M)$. The key implication is that the structural coefficients of the model are time-varying and all parameters in the structural form of the model change if the regime shifts.

We start our analysis by considering median impulse responses across clusters. Because within each regime we have a standard VAR with constant coefficients, we compute the IRFs per regime. This gives a set of J dynamic responses to a monetary policy shock. To analyze differences in shapes and magnitudes, we focus on the posterior median in the main body of the paper. Results that include posterior credible intervals and the reactions of the other variables in $\boldsymbol{y}_t$ are provided in Section B in Data S1.

Figure 4 presents posterior medians of the impulse responses implied by each cluster. In general, the model produces impulse responses that are consistent with our economic intuition. In response to unexpected increases in the policy rate, output growth declines and the unemployment rate increases. Inflation, unexpectedly, increases, pointing toward a price puzzle. It is worth stressing that this price puzzle is most pronounced in the second cluster, which is the high volatility cluster.

When we compare differences across clusters, it can be seen that Cluster 2 stands out as implying very different impulse responses but in ways that are different for the different variables. For instance, all of the clusters are very similar for long-run impulse responses (e.g., greater than 10 quarters), Cluster 2 differs greatly from other clusters at short horizons (e.g., less than 1 year) for most of the variables. In particular, short-run responses in Cluster 2 appear to be much more pronounced. This is driven by the fact that the variances of the structural shocks are much larger and the monetary shock implies a stronger impact reaction of the Federal funds rate (see panel (d) of Figure B.1 in Data S1). For medium-run responses, the second cluster also yields responses of GDP growth and unemployment rates, which differ from the remaining clusters. For GDP growth, our results indicate an overshoot in real activity after around 2 years whereas

the unemployment reaction appears to be much more persistent. This points toward differences in the transmission of monetary policy to real activity in turbulent periods (i.e., periods that are allocated to Cluster 2).

Next we turn to the question about the overall effects of monetary shocks on the three focus variables. The overall impulse responses (i.e., averaged over the different clusters) for our core variables (i.e., the ones common to all VARs) are given in Figure 5. These are calculated by taking the impulse responses for cluster i and weighting it by the posterior mean of η_i. This figure also contains credible intervals for both our BVAR-DPM and the BVAR with $J = 1$ so that the reader can gage whether the differences in impulse responses between our model and the Gaussian-errored equivalent are substantial in a statistical sense.

The figure suggests that differences between the model the DPM and the model that sets $J = 1$ are not substantial in the sense that the credible intervals of the BVAR-DPM include the ones of the BVAR in almost all cases. The main exception is the reaction of inflation. Here, we observe a much stronger immediate increase (i.e., more evidence for a price puzzle) but also a slightly more pronounced one-year-ahead decline in inflation. It is also worth stressing that short-run unemployment reactions of the BVAR with $J = 1$ suggest substantial posterior mass of the IRFs are located below zero, suggesting a decline in the unemployment rate to a monetary tightening. The BVAR-DPM allocates appreciably less posterior evidence to declines in unemployment rates.

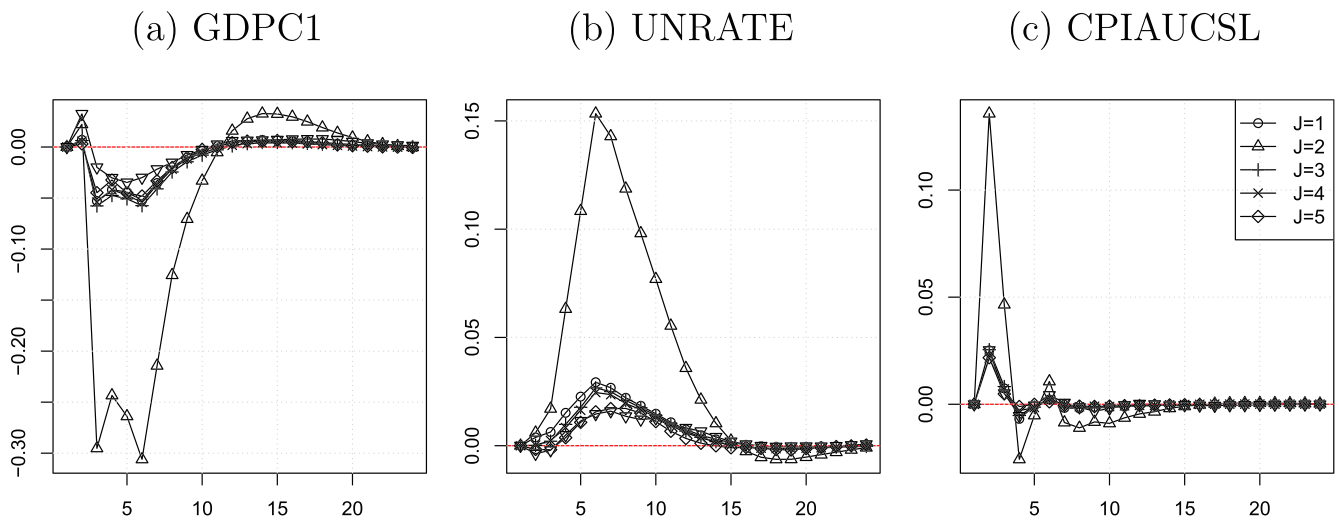

FIGURE 4 Median impulse responses to a monetary policy shock for the different clusters. *Note*: The black lines show the posterior median of the impulse responses for the different clusters

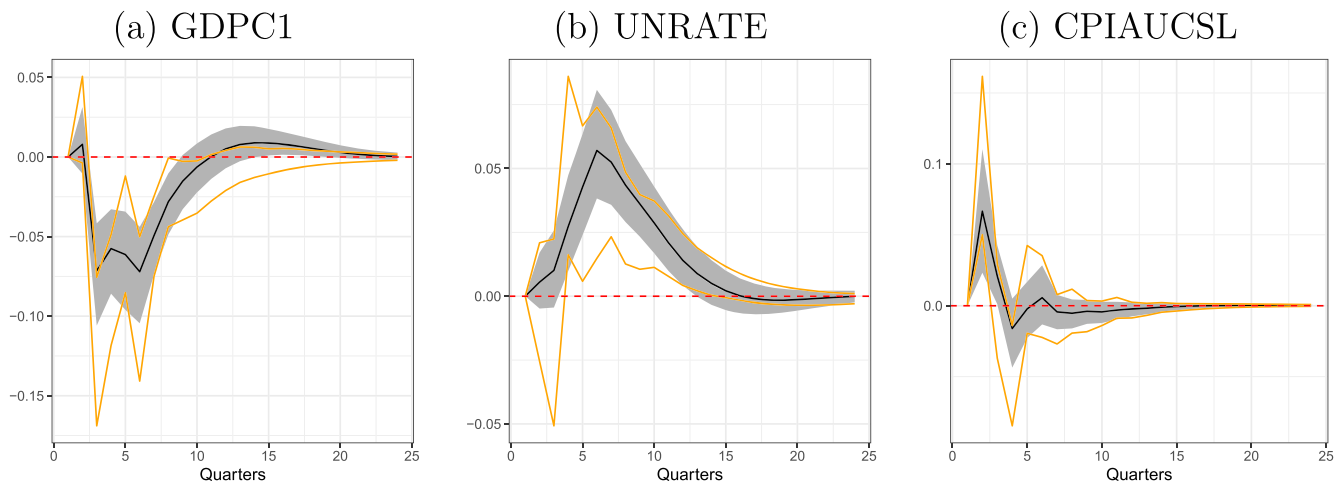

FIGURE 5 Median impulse responses to a monetary policy shock, average across clusters. *Note*: The black lines show the posterior median of the impulse responses for the BVAR-DPM. The shaded region is a credible interval. The orange lines denote credible intervals for the Gaussian BVAR. Both intervals cover 16th to 84th percentiles

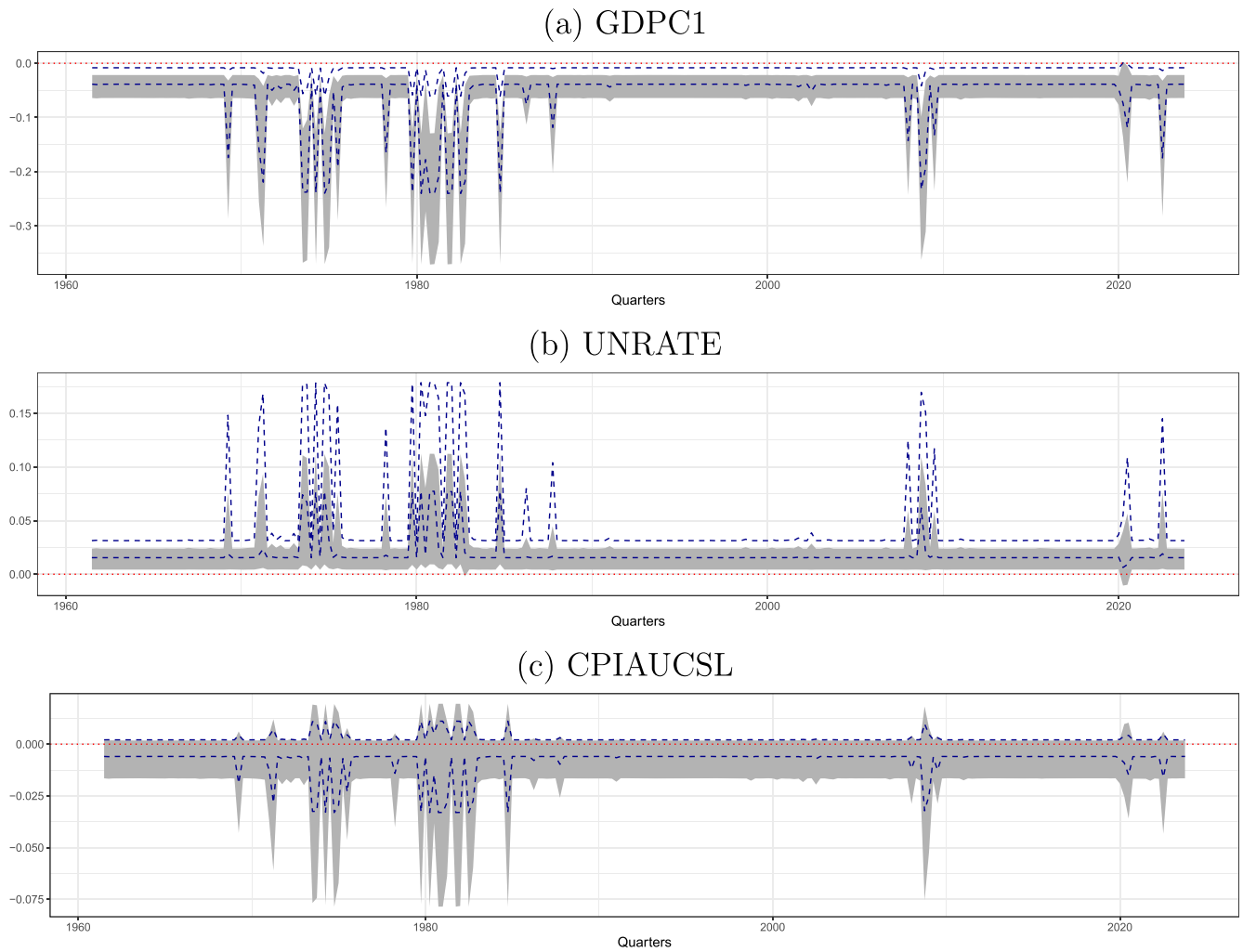

FIGURE 6 Impulse responses over time. *Note*: The gray shaded area denotes the 16th to 84th percentiles of the posterior of the four-step-ahead responses. The blue dashed lines represent the 16th and 84th posterior percentiles of the eight-step-ahead responses

Finally, we turn to the question about how impulse responses evolve over time. The discussion of the cluster probabilities over time in combination with the cluster-specific IRFs provides some information on how IRFs change across clusters and in which periods the model assigns observations to these clusters. However, the actual time t IRF is a convex combination of the IRFs across the different regimes. Hence, when viewed at a particular point in time, substantial differences can arise to the cluster-specific IRFs.

In Figure 6, we show the 16th and 84th posterior percentiles of the four and eight-step-ahead responses for each point in time. This is achieved by computing the IRFs for $t = 1, \dots, T$ using Equation (5).

This exercise tells a similar story to the cluster specific IRFs but links these differences to particular time periods. Substantial declines in output, inflation and increases in unemployment in response to monetary shocks can be observed in the mid 1970s, early 1980s and during the financial crisis. Notice, however, that for both horizons, we find reactions that look (slightly) different from the ones observed for the individual clusters. This mostly relates to the fact that even in turbulent times, the first regime gets some weight (between 10% and 25%), and this gives rise to different marginal posterior distributions of the h-step-ahead IRFs.

5.4 | Forecasting performance

In this subsection, we present the results of our forecasting exercise. Our measure of point forecast performance is the mean squared forecast error (MSE) and our measure of density forecast performance is the average of log predictive

TABLE 5 Forecasting performance: 1977:Q1 to 2023:Q4.

		GDPC1		UNRATE		CPIAUCSL	
		DPM	$J = 1$	**DPM**	$J = 1$	**DPM**	$J = 1$
One-quarter-ahead							
Homosk.	S	0.900	1.096	0.823	1.140	0.939	0.977
		(−0.041)	(−0.009)	(0.161)	(−0.153)	(0.035)	(0.032)
	M	0.873	1.015	0.748	1.104	0.915	0.955
		(−0.064)	(−0.014)	(0.240)	(−0.545)	(0.032)	(0.023)
	L	0.979	0.964	0.967	0.983	1.064	1.039
		(−0.199)	(−0.145)	(−0.575)	(−0.334)	(−0.056)	(−0.049)
SV	S	0.933	0.935	0.850	0.859	0.973	1.015
		(0.000)	(0.146)	(0.043)	(0.073)	(0.111)	(0.109)
	M	0.875	0.874	0.805	0.795	0.911	0.958
		(−0.038)	(0.118)	(−0.306)	(0.087)	(0.124)	(0.101)
	L	1.027	1.381	0.965	2.296	0.999	1.217
		(0.022)	(−1.463)	(0.123)	(−2.638)	(−0.002)	(−1.539)
One-year-ahead							
Homosk.	S	0.980	0.989	0.922	1.000	1.056	1.049
		(0.807)	(0.817)	(1.091)	(0.659)	(0.538)	(0.506)
	M	0.985	1.005	0.909	1.018	1.028	1.014
		(0.787)	(0.772)	(1.115)	(0.577)	(0.537)	(0.521)
	L	1.005	1.012	0.972	1.003	0.991	0.991
		(0.661)	(0.636)	(0.917)	(0.806)	(0.528)	(0.511)
SV	S	0.988	0.988	0.933	0.941	1.039	1.038
		(0.546)	(0.524)	(0.811)	(0.696)	(0.579)	(0.542)
	M	0.973	0.978	0.913	0.910	1.003	1.026
		(0.436)	(0.489)	(0.972)	(0.659)	(0.554)	(0.506)
	L	0.995	1.091	0.965	1.430	0.981	1.277
		(0.101)	(−2.420)	(0.252)	(−2.908)	(0.114)	(−2.168)

Note: The table reports mean squared forecast errors relative to the large BVAR with SV and average log predictive likelihood differences (in parentheses) between a given model and the large BVAR with SV. The BVAR-SV is estimated by setting $J = 1$. The blue shaded cells are absolute MSEs and LPLs. Bold numbers indicate the best performing model for a given variable and horizon.

likelihoods (LPL). Tail forecast performance is measured using absolute quantile scores (QSs). Results are reported relative to a benchmark model, which is the large BVAR that sets $G = 1$ and has heteroskedastic measurement errors.

We start our discussion by focusing on MSE ratios and LPL differences between a given model and the BVAR with $J = 1$ and SV. Table 5 presents the relative MSEs and LPLs (in parentheses) for the three variables being forecast for both homoskedastic and heteroskedastic models, allowing for a variable-specific examination of forecast performance.

With some exceptions, it can be seen that DPM models perform well for all variables and both forecast horizons. The gains are particularly pronounced for one-step-ahead point forecasts of the unemployment rate. In this case, gains in predictive accuracy reach over 20%. For one-step-ahead GDP growth and inflation forecasts, we find gains that are slightly more muted (over 12% for GDP growth and around 9% for inflation). The gains in point forecasting accuracy often carry over to increases in density forecast performance. In this respect, adding the DPM piece often improves predictive accuracy for unemployment and inflation rate forecasts. Only for GDP growth, we find values close to zero and hence little evidence that adding a DPM improves one-step-ahead density forecasts. It is also worth stressing that having SV on the measurement errors only plays a limited role for point forecasts, but for density forecasts, we find that for unemployment forecasts, the DPM with homoskedastic measurement errors produces the most accurate density forecasts. For inflation, we find that models with SV and the DPM produce slightly superior density forecasts.

The improvements in forecast performance deteriorate slightly if we move to the one-year-ahead horizon. In this case, DPM models still produce more precise density forecasts for all three focus variables but gains are smaller (around 3% for GDP growth and 9% for the unemployment rate and 2% for inflation). One-year-ahead density forecasts, however, appear to profit more from using a DPM. Interestingly, in this case, we also find that adding SV to the measurement errors often harms predictive performance. This motivates our choice of using a homoskedastic BVAR-DPM in the previous subsections.

TABLE 6 Tail forecasting performance: 1977:Q1 to 2023:Q4.

		GDPC1		UNRATE		CPIAUCSL	
		DPM	$J=1$	DPM	$J=1$	DPM	$J=1$
One-quarter-ahead							
Homosk.	S	0.914	0.940	1.151	1.372	0.930	0.930
		(0.959)	(1.010)	(0.877)	(0.929)	(1.076)	(1.060)
	M	0.938	0.948	1.047	1.356	0.950	0.958
		(0.952)	(0.996)	(0.847)	(0.903)	(1.012)	(1.024)
	L	0.988	0.990	1.241	1.254	1.042	1.033
		(1.019)	(1.021)	(0.858)	(0.886)	(1.167)	(1.130)
SV	S	0.929	0.969	1.136	0.684	0.961	1.001
		(0.921)	(0.742)	(0.868)	(0.963)	(1.012)	(0.987)
	M	0.952	0.960	1.186	0.630	0.947	0.986
		(0.946)	(0.721)	(0.923)	(0.946)	(0.930)	(0.945)
	L	0.995	0.128	0.930	0.095	0.992	0.139
		(0.996)	(0.125)	(0.971)	(0.127)	(1.060)	(0.102)
Homosk.	S	0.604	0.616	0.561	0.526	0.821	0.829
		(0.549)	(0.555)	(0.630)	(0.632)	(0.755)	(0.754)
	M	0.607	0.612	0.538	0.558	0.831	0.819
		(0.572)	(0.575)	(0.617)	(0.634)	(0.730)	(0.716)
	L	0.687	0.694	0.587	0.586	0.804	0.792
		(0.671)	(0.672)	(0.616)	(0.627)	(0.724)	(0.734)
SV	S	0.632	0.681	0.545	0.609	0.825	0.839
		(0.551)	(0.608)	(0.638)	(0.695)	(0.766)	(0.768)
	M	0.664	0.677	0.614	0.630	0.809	0.838
		(0.593)	(0.602)	(0.678)	(0.705)	(0.788)	(0.771)
	L	0.937	0.211	0.859	0.189	0.985	0.175
		(0.922)	(0.202)	(0.902)	(0.219)	(0.936)	(0.178)

Note: The table reports relative quantile scores (5% and 95 %, in parentheses) to the large BVAR with SV. The BVAR-SV is estimated by setting $J = 1$. The blue shaded cells are absolute quantile scores. Bold numbers indicate the best performing model for a given variable and horizon.

To analyze whether the DPM component improves forecasts, we can compare each of the BVAR-DPM models with the equivalent model with $J = 1$. For all three variables, two forecast horizons and for both forecast metrics, we usually find that one of the DPM versions of the model forecasts better than the Gaussian one. It is worth stressing that there are several cases where adding SV to the model, which sets $J = 1$ improves predictive performance substantially relative to the homoskedastic version. As stated above, this does not necessarily hold for the DPM models. This corroborates the findings in the forecasting literature that points toward the necessity for using flexible assumptions on the shocks to improve forecasts (Clark, 2011; Carriero et al., 2022; Huber & Feldkircher, 2019) but that there is a trade-off between flexibility in error distribution and SV. That is, once you allow for fat tailed error distributions there is less need to additionally add SV to the model.

In terms of VAR dimension, it is interesting to note that the strongest performance for the DPM models arises with small and medium VARs. With large VARs, the evidence is more mixed both with regards to the need for DPM and with regards to the need for SV. We conjecture that in the large VAR, the explanatory power of the right hand side variables can mop up some (but not all) of the need to allow for non-Gaussianity or volatility change.

Table 6 presents evidence on tail forecasting performance. The table shows 5% and 95% QSs relative to the BVAR with $J = 1$ and SV so that numbers smaller than one indicate that a given model produces more accurate tail forecasts than the benchmark.

We would expect the DPM to be particularly good in capturing tail behavior in a way that the Gaussian model cannot. And, with some exceptions, Table 6 confirms this expectation. The pattern of results is similar to those found in Table 5, but is slightly stronger in favor of BVAR-DPM models. More specifically, we again find gains of the different models relative to the benchmark. Depending on the model size, we also find accuracy gains from using the DPM specification relative to setting $J = 1$. These gains are more pronounced for one-year-ahead tail forecasts of output growth and the unemployment rate and smaller-sized models.

6 | CONCLUSIONS

In this paper, we propose a new specification for the errors in a VAR that takes a particular additive form involving two components. The first is a homoskedastic error and a conventional inverse Wishart prior can be used for its covariance matrix. The second is a diagonal error covariance matrix with diagonal elements following SV processes. We show that, by adopting this additive form, we gain two major advantages. First, computation is much faster as it allows for equation-by-equation estimation. Second, posterior and predictive inference does not depend on the way the variables are ordered in the VAR. We then extend this model to allow the first error to follow a DPM. We discuss, both theoretically and empirically, the great flexibility that is obtained by doing so. In addition, we develop a computationally fast MCMC algorithm that allows for posterior and predictive inference in high dimensional nonparametric VARs. Our empirical results, using artificial and real data, show that our approach produces more accurate parameter estimates and forecast distributions. Moreover, when it comes to structural analysis, we can leverage the flexibility of our model to focus on how the effects of monetary policy shocks have changed over time.

ACKNOWLEDGEMENTS

Huber gratefully acknowledges financial support from the Austrian Science Fund (FWF, grant no. ZK 35). We thank two anonymous reviewers and the editor, Herman van Dijk, for helpful comments and suggestions that improved the paper.

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at https://doi.org/10.15456/jae.2024191.2007840176.

DATA AVAILABILITY STATEMENT

The dataset and estimation codes for the BVAR-DPM can be downloaded from the JAE's data archive.

ORCID

Florian Huber https://orcid.org/0000-0002-2896-7921
Gary Koop https://orcid.org/0000-0002-6091-378X

REFERENCES

Antolin-Diaz, J., Drechsel, T., & Petrella, I. (2017). Tracking the slowdown in long-run GDP growth. *The Review of Economics and Statistics*, *99*(2), 343–356. https://doi.org/10.1162/REST_a_00646

Arias, J. E., Rubio-Ramirez, J. F., & Shin, M. (2023). Macroeconomic forecasting and variable ordering in multivariate stochastic volatility models. *Journal of Econometrics*, *235*, 1054–1086.

Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, *120*(1), 387–422.

Bhattacharya, A., Chakraborty, A., & Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, *103*, 985–991.

Braun, R. (2021). The importance of supply and demand for oil prices: Evidence from non-Gaussianity. (*957*): Bank of England https://ideas.repec.org/p/boe/boeewp/0957.html

Carriero, A., Chan, J., Clark, T. E., & Marcellino, M. (2022). Corrigendum to large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, *227*(2), 506–512.

Carriero, A., Clark, T. E., & Marcellino, M. (2016). Common drifting volatility in large Bayesian VARs. *Journal of Business & Economic Statistics*, *34*(3), 375–390.

Carriero, A., Clark, T. E., & Marcellino, M. (2019). Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, *212*(1), 137–154. https://ideas.repec.org/a/eee/econom/ v212y2019i1p137-154.html

Carriero, A., Clark, T. E., Marcellino, M., & Mertens, E. (2022). Addressing Covid-19 outliers in BVARs with stochastic volatility. *Review of Economics and Statistics*, 1–38.

Chan, J. C. C., Koop, G., & Yu, X. (2021). Large order-invariant Bayesian VARs with stochastic volatility. arXiv preprint arXiv:2111.07225.

Chiu, C.-W. J., Mumtaz, H., & Pintr, G. (2017). Forecasting with VAR models: Fat tails and stochastic volatility. *International Journal of Forecasting*, *33*(4), 1124–1143. https://www.sciencedirect.com/science/article/pii/ S016920701730033X

Clark, T. E. (2011). Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics*, *29*(3), 327–341.

Clark, T. E., Huber, F., Koop, G., Marcellino, M., & Pfarrhofer, M. (forthcoming). Tail forecasting with multivariate Bayesian additive regression trees.

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.

Dunson, D. B., & Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, *104*(487), 1042–1051.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588.

Fox, E. B., & Dunson, D. B. (2015). Bayesian nonparametric covariance regression. *Journal of Machine Learning Research*, *16*(77), 2501–2542. http://jmlr.org/papers/v16/fox15a.html

Frühwirth-Schnatter, S., & Malsiner-Walli, G. (2019). From here to infinity: Sparse finite versus dirichlet process mixtures in model-based clustering. *Advances in data analysis and classification*, *13*, 33–64.

Frühwirth-Schnatter, S., Tüchler, R., & Otter, T. (2004). Bayesian analysis of the heterogeneity model. *Journal of Business & Economic Statistics*, *22*(1), 2–15.

Griffin, J. E., & Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, *5*(1), 171–188.

Hauzenberger, N., Huber, F., Koop, G., & Onorante, L. (2022). Fast and flexible Bayesian inference in time-varying parameter regression models. *Journal of Business & Economic Statistics*, *40*(4), 1904–1918.

Huber, F., & Feldkircher, M. (2019). Adaptive shrinkage in Bayesian vector autoregressive models. *Journal of Business & Economic Statistics*, *37*(1), 27–39.

Huber, F., Koop, G., Onorante, L., Pfarrhofer, M., & Schreiner, J. (2020). Nowcasting in a pandemic using non-parametric mixed frequency VARs. *Journal of Econometrics*, *232*, 52–69.

Huber, F., & Rossini, L. (2020). Inference in Bayesian additive vector autoregressive tree models. https://arxiv.org/abs/2006.16333

Kalli, M., Griffin, J. E., & Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, *21*(1), 93–105.

Karlsson, S., Mazur, S., & Nguyen, H. (2023). Vector autoregression models with skewness and heavy tails. *Journal of Economic Dynamics and Control*, *146*, 104580. https://www.sciencedirect.com/science/article/pii/S0165188922002834

Kastner, G., & Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, *76*, 408–423.

Kastner, G., & Huber, F. (2021). Sparse Bayesian vector autoregressions in huge dimensions. *Journal of Forecasting*, *39*, 1142–1165.

Lau, J. W., & Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, *16*(3), 526–558.

Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grün, B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, *26*(2), 285–295.

McCracken, M., & Ng, S. (2020). Fred-QD: A quarterly database for macroeconomic research: National Bureau of Economic Research.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, *72*(3), 821–852.

Stock, J. H., & Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*(S1), 3–33.

Wu, P., & Koop, G. (2023). Fast, order-invariant Bayesian inference in VARs using the eigendecomposition of the error covariance matrix. manuscript.

Yau, C., & Holmes, C. (2011). Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Analysis*, *6*(2), 329–351. https://doi.org/10.1214/11-BA612

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of the article.

How to cite this article: Huber, F., & Koop, G. (2024). Fast and order-invariant inference in Bayesian VARs with nonparametric shocks. *Journal of Applied Econometrics*, 1-20. https://doi.org/10.1002/jae.3087