Restricted Large Bayesian Vector Autoregressions

Eva F. Janssens[†] and Robin L. Lumsdaine[‡]

March 20, 2025

Abstract

Without regularization, vector autoregressions (VAR) with many variables suffer from large standard errors and poor out-of-sample forecasting. Regularization is typically viewed as requiring a bias-variance trade-off, but we show in finite samples that imposing parameter equality restrictions in a VAR – akin to clustering parameters along both rows and columns of the coefficient matrix – can reduce *both* bias and variance by lowering second-order bias, even if the restrictions are mildly misspecified. We specify the Matrix Dirichlet Process (MDP) prior for VARs, which remains agnostic ex-ante about the appropriate restrictions by jointly sampling both the clustering structure and parameter values, within a computationally efficient Gibbs sampling framework. Simulations and two applications – price spillovers across 45 industries and personal income growth across all U.S. states – demonstrate that the MDP prior recovers intuitive economic relationships and improves forecasting performance.

Keywords: Clustering, large vector autoregression, shrinkage, bias-variance trade-off, Bayesian analysis

JEL classification codes: C11, C22, C53, C55, E37

* Disclaimer: The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of any of the institutions with which the authors are affiliated.

* Acknowledgements: We are grateful for helpful comments from seminar participants at the Board of Governors of the Federal Reserve System, Michigan State University and Princeton University. Any errors are our own.

Contact information:
† Eva F. Janssens: Department of Economics, University of Michigan, e-mail: evafj@umich.edu
‡ Robin L. Lumsdaine: Kogod School of Business, American University; Erasmus University Rotterdam; National Bureau of Economic Research; Tinbergen Institute, email: robin.lumsdaine@american.edu

1 Introduction

One of the fundamental challenges in modeling economic systems is that as specifications grow richer, capturing interlinkages via systems of equations, the number of parameters in the model increases quadratically with the number of equations. This issue arises in a range of applications, including Vector Autoregressions (VARs), transition probabilities in (hidden) Markov models, network-based models, and the estimation of covariance matrices in multivariate systems. As the number of parameters grows, models suffer from inflated standard errors and poor out-of-sample fit at best, and at worst, they may lose identification entirely as the number of parameters exceeds the number of observations.

To address these challenges, researchers have developed dimension-reduction techniques that impose structure while preserving flexibility. One approach is low-rank factorization, as in Factor-Augmented VARs (FAVARs) (Bernanke, Boivin, and Eliasz, 2005), which summarize high-dimensional data using a small number of common factors. Alternatively, regularization methods—including LASSO and Ridge regression, Bayesian VARs with Minnesota priors (Litterman, 1979, 1980) or spike-and-slab priors (Giannone, Lenza, and Primiceri, 2021) – address overfitting by shrinking or setting certain coefficients to zero. Economic theory can also motivate dimension-reduction, as in economic networks, imposing factorization or sparsity based on production or financial linkages (Foerster, Sarte, and Watson, 2011; Mlikota, 2023), or spatial econometric models, where entities closer in space are assumed to be more strongly connected.

These methods are often viewed through the lens of a bias-variance trade-off, where reducing variance through additional structure comes at the cost of introducing bias. This paper aims to challenge that common view and suggests that it is possible to think of dimension-reduction methods that simultaneously reduce bias and variance. The key insight is that time series models like VARs often exhibit poor finite-sample properties and suffer from substantial second-order bias, as shown in Bao and Ullah (2007). We show that imposing parameter equality restrictions – when they are correctly specified – reduces second-order bias, improving inference and estimation uncertainty. Conversely, incorrect restrictions introduce misspecification bias. However, as long as the reduction in second-order bias dominates the bias introduced by misspecification, the additional structure from these restrictions serves as a net improvement in both finite-sample bias and variance.

While economic theory may justify certain restrictions, we take an agnostic approach, letting the data determine the structure. In frequentist univariate settings, clustering methods like k-means (Bonhomme and Manresa, 2015; Hahn and Moon, 2010; Lumsdaine,

Okui, and Wang, 2022) effectively impose parameter equality restrictions but are limited to single-dimensional clustering. When parameters span multiple dimensions – such as time and cross-sections in panels or lags and endogenous variables in VARs – co-clustering becomes NP-hard, making standard frequentist methods impractical for high-dimensional economic models where restrictions are inherently multidimensional.

The Bayesian counterpart to clustering can be achieved using Dirichlet Process priors, which can be reinterpreted as sampling equality restrictions alongside parameters. This approach explicitly accounts for uncertainty about which restrictions should hold, allowing the data to simultaneously determine both the clustering structure and the parameter estimates. A key advantage of this framework is its natural extension to multidimensional clustering. To this end, we build on the Matrix Dirichlet Process (MDP) framework of Dunson, Xue, and Carin (2008) used for meta-analysis, to allow for simultaneous clustering along two dimensions in a VAR setting. We further develop a computationally efficient Gibbs sampler, facilitating the joint estimation of parameter values and restriction structures in high-dimensional models. The MDP prior can easily be combined with other priors, enabling researchers to impose sign restrictions and other identification constraints.

We first assess the performance of the Matrix Dirichlet Process (MDP) prior for large VAR models through simulations based on five realistic data-generating processes (DGPs). These settings include variations of a Minnesota-like structure, as well as spatial and network-based dependencies. We compare bias in the posterior mean, posterior width, and root mean squared forecast error (RMSFE) across three priors: the MDP prior, an uninformative prior, and the Minnesota prior. The results demonstrate that the MDP prior dominates across all DGPs, producing lower bias and tighter posterior distributions. While the Minnesota prior shrinks variance, it introduces substantial bias, whereas the MDP prior achieves both bias reduction and more precise inference. The forecasting results further confirm this advantage, with the MDP prior consistently outperforming alternative priors in terms of RMSFE.

Next, the paper considers two empirical applications. The first examines price spillovers across 45 U.S. industries, using monthly Producer Price Index (PPI) data. As such, it contributes to the literature on sectoral shock spillovers (Long and Plosser, 1983, Foerster et al., 2011, and Atalay, 2017), where structural models are commonly used to quantify sectoral interactions and shock propagation. In contrast to earlier literature, a Bayesian VAR model with the MDP prior provides a flexible reduced-form approach, allowing the data to determine sectoral interdependencies without imposing strong structural assumptions. We estimate a Bayesian VAR (BVAR) under the MDP prior, an uninformative prior, and the

Minnesota prior, comparing posterior distributions and impulse response functions (IRFs). The uninformative prior leads to excessively wide posteriors, preventing meaningful inference. The Minnesota prior tightens credible intervals but often yields counterintuitive signs in the IRFs. In contrast, the MDP prior yields economically plausible results, such as correctly identifying the positive spillover of oil and gas price shocks to utility prices, which the model estimated with the Minnesota prior fails to capture.

The second application examines personal income growth across all U.S. states, focusing on spillover effects and forecasting accuracy. As in the first application, the uninformative prior produces excessively wide posteriors, while the Minnesota prior fails to capture meaningful economic relationships, often yielding implausible cross-state spillovers. In contrast, the MDP prior identifies economically sensible patterns, with stronger spillovers between neighboring states, consistent with regional economic linkages. Since forecasting is the primary objective, we compare the RMSFE across priors and find that the MDP prior consistently outperforms the Minnesota prior, with RMSFEs occasionally an order of magnitude smaller, underscoring its effectiveness in high-dimensional forecasting.

Related Literature Bayesian priors are widely used in large VARs to manage overparameterization. Bańbura, Giannone, and Reichlin (2010) demonstrate the effectiveness of the Litterman (1979) prior when shrinkage is calibrated to the cross-sectional dimension. Their hyperparameter selection ensures consistent in-sample fit across models. Koop (2013) also finds that the Minnesota prior performs well in forecasting. Chan (2022) refines the Minnesota prior by allowing asymmetric shrinkage, penalizing lags of other variables more than own lags. Giannone, Lenza, and Primiceri (2019) propose a prior specifically to regularize the long-run behavior of the VAR.

Other approaches aim to refine shrinkage priors through selective shrinkage. George, Sun, and Ni (2008) introduce hierarchical priors that stochastically determine which parameters to shrink, building on George and McCulloch (1993). Similarly, Giannone et al. (2021) apply a spike-and-slab prior (though not in a VAR setting), which sets some parameters exactly to zero rather than merely shrinking them. Related to our paper is also Billio, Casarin, and Rossini (2019) who exogenously split the parameters of a VAR model in M blocks (e.g., a block is a lag), and then use Bayesian LASSO for each block, where they use independent Dirichlet Process Priors on the Lasso parameters across blocks. The main difference is that the MDP prior identifies similarities between parameters to impose restrictions and form clusters, moving beyond simple zero restrictions or shrinkage. In addition, unlike many priors, the MDP prior does not assume parameters are i.i.d.; instead, it exploits the matrix structure of these models.

Another way to handle high-dimensional models is through dimensionality reduction techniques. These include dynamic factor models (Bernanke et al., 2005; Forni, Hallin, Lippi, and Reichlin, 2000), principal component analysis (PCA) (De Mol, Giannone, and Reichlin, 2008; Stock and Watson, 2002), and Bayesian compressed regression (Koop, Korobilis, and Pettenuzzo, 2019). However, these methods impose more rigid structures than the clustering framework of the MDP prior, which flexibly groups parameters across equations and variables rather than assuming all variables load on the same set of factors.

Outline In Section 2, we analyze second-order bias in the context of large VAR models, and show how parameter restrictions reduce this bias. Section 3 proposes our MDP priors and discusses a Gibbs Sampling algorithm. In Section 4, we show performance for five realistic data generating processes. Sections 5 and 6 present our applications: PPI spillovers across industries, and forecasting state-level personal income. Section 7 concludes.

2 Trading Off Small Sample Bias with Misspecification Bias

Consider a Vector Autoregression of order one (VAR(1)) with p variables, and let $y_t = (y_{1t}, y_{2t}, ..., y_{pt})$ be a $p \times 1$ vector observed at time t. We assume a stationary process

$$y_t = \Phi y_{t-1} + u_t, \quad u_t \sim N(0, \Omega). \tag{1}$$

Rilstone, Srivastava, and Ullah (1996) derive expressions for the second-order bias of estimators that solve moment equations of the form

$$\psi_T(\hat{\beta}) = \frac{1}{T} \sum_{t=1}^{T} q_t(\hat{\beta}) = 0, \tag{2}$$

with $q_t(\beta)$ a moment function, and ψ_T the sample moment condition. This framework nests the first-order conditions of many maximum likelihood estimators, the orthoginality condition of least squares estimators, and the moment conditions of method of moment estimators, but also can be related to Generalized Method of Moments. To derive the finite-sample moments of such an estimator, Rilstone et al. (1996) find an analytically tractable stochastic approximation for $\hat{\beta}$ of the form

$$\hat{\beta} - \beta_0 = a_{-1/2} + a_{-1} + o_P(T^{-3/2}), \tag{3}$$

where the notation $a_{-s/2}$ denotes a term of order $O_p(T^{-s/2})$ and the second-order bias follows from the expectation of $a_{-1/2} + a_{-1}$. Second-order bias is also sometimes referred to as small-sample bias, as it vanishes as $T \to \infty$. Expressions for this expansion are derived

by implementing a Taylor series expansion of $\psi_T(\hat{\beta}) = 0$ around β_0 and using a Nagar (1959)-type expansion of the inverse of the gradient of the moment function to solve for $\hat{\beta} - \beta_0$ recursively.

Bao and Ullah (2007) extend this framework for time series models where data are not i.i.d., and, among other things, derive expressions for the second-order bias of the VAR(1) model. Assume that $\hat{\beta}$ is $\sqrt{T}$-consistent and that Assumptions (A)-(C) in Rilstone et al. (1996) hold, which relate to the existence, boundedness and validity of a Lipschitz-condition for the derivatives of q_t around β_0 up to order two and $\psi_T(\beta)$ having an invertible Jacobian that is stochastically bounded around β_0. Let H_1, H_2 denote the matrices of partial derivatives $H_i = \nabla_i \psi_T$ of order $i = 1, 2$ evaluated at β_0, with ∇_i the symbol for the partial derivative, and denote $\bar{H}_i = \mathbb{E}(H_i)$. Bao and Ullah (2007) show (in equation (3) of their paper, using their notation)

$$a_{-1/2} = -Q\psi_T(\beta_0), \quad \text{with } Q = \bar{H}_1^{-1}, \tag{4}$$

$$a_{-1} = -QVa_{-1/2} - \frac{1}{2}Q\bar{H}_2[a_{-1/2} \otimes a_{-1/2}]. \tag{5}$$

Because $\mathbb{E}(a_{-1/2}) = 0$, the second-order bias of $\hat{\beta}$ is given by $E(a_{-1})$.

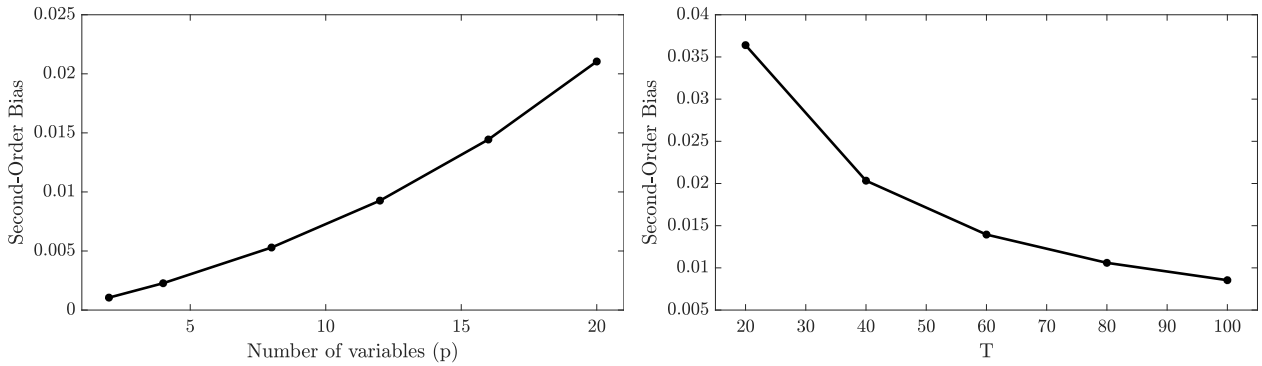
Theorem 3 in Bao and Ullah (2007) provides expressions for the second-order bias of a VAR(1) model as in Equation (1). While these expressions are not so easy to interpret, we can use them to compute the theoretically implied second-order bias for specific examples. We visualize how the second-order bias changes, as the data length T and number of variables p vary, in Figure 1. For this particular data-generating process, the bias grows faster than linearly in p, and decreases at a slower-than-linear rate in T. Hence, not only does estimating a VAR with a large number of variables rapidly increase the variance of our estimators, estimating a VAR with many variables will also result in substantial small-sample bias.

One possible bias correction approach is to use the expressions for second-order bias to construct a plug-in estimator. This is typically suboptimal from a mean squared error (MSE) efficiency perspective, however, as it increases variance. Bias correction is only desirable when the researcher places a lot of weight on bias reduction relative to variance control, as discussed by Li, Plagborg-Møller, and Wolf (2024) in the context of VARs and Local Projections (Jordà, 2005).

Now, consider the possibility of imposing a set of linear restrictions $R\beta = r$. In particular, restrictions that two or more coefficients are equal across different variables, e.g. $\phi_{1i} = \phi_{2j}$,

Figure 1: Second-Order Bias of a VAR(1) in element $\hat{\phi}_{12}$

(a) Varying number of variables p *(b) Varying T*



Note: Figure (a) assumes $T = 50$, Ω has 3 on the diagonal and 1 on the off-diagonal, Φ has 0.7 on the diagonal and 0.3/21 on the off-diagonal. For Figure (b), p is set to 8, Ω is the same and Φ has 0.7 on the diagonal and 0.3/9 on the off-diagonal. y_0 is assumed to be zero.

assuming that these restrictions are in fact correct. It is well known that such restrictions reduce the variance of the estimator, e.g., one can find a textbook treatment of this result in Hansen (2022). It is also possible to show that such restrictions reduce the second-order bias, as we will do below in Theorem 1. Intuitively, adding linear equality restrictions increases the effective sample size, because we can use more information to estimate the same parameter, hence decreasing the second-order bias. To formalize this, Theorem 1 considers what happens to a_{-1} of the expansion in Equation (3) when imposing a restriction $R\beta = r$ to the estimator. Under the assumption that the restriction holds, this results in a decrease in a_{-1}.

Theorem 1. *Maintaining assumptions (A)-(C) as in Rilstone et al. (1996), the difference between the second-order bias of an unrestricted moment condition-based estimator as in* (2) *and one with the additional restriction $R\beta = r$ is given by the expectation of*

$$\Delta a_{-1} = (Q - Q_r)\left(V a_{-1/2} + \frac{1}{2}\bar{H}_2[a_{-1/2} \otimes a_{-1/2}]\right). \tag{6}$$

where

$$Q - Q_r = \bar{H}_1^{-1}R'(R\bar{H}_1^{-1}R')^{-1}R\bar{H}_1^{-1}. \tag{7}$$

is positive semi-definite.

Proof: See Appendix A.

Theorem 1 tells us restrictions of the form $R\beta = r$ can reduce second-order bias if the restrictions hold true. On the other hand, if the restriction is false, a restriction of the form $\phi_{1i} = \phi_{2i}$ can introduce misspecification bias proportional to $|\phi_{1i} - \phi_{2j}|$, and, unlike second-order bias, misspecification bias does not vanish with the sample size. However, as long as the misspecification bias is smaller than the reduction in second-order bias, introducing restrictions can be bias-reducing in finite samples.

In practice, the true parameters are unknown, making it difficult to determine the optimal restrictions needed to reduce bias and variance. This is where the Matrix Dirichlet Process Prior proves useful. The Gibbs Sampling algorithm identifies parameters that are similar and can be clustered without compromising model fit. In doing so, the algorithm effectively discovers restrictions that are nearly valid. As outlined above, this can lead to improvements in both bias and variance, a result we will demonstrate in our simulations.

3 A Matrix Dirichlet Process Prior for VAR models

Dirichlet Process (DP) priors (Ferguson, 1973) are the Bayesian counterpart of parameter clustering.[1] Denote the parameters of interest by ϕ_i, $i = 1, ..., N$. In the one-dimension clustering problem:

$$\phi_i | p \sim \sum_{l=1}^{L_{\max}} p_l \delta_{\bar{\phi}_l}(\phi_i), \quad p \sim DP(\alpha, G_0), \quad \bar{\phi}_l \sim G_0$$

where G_0 is the underlying prior on the parameters ϕ_l, and $\delta_{\bar{\phi}_l}$ is an indicator function, equal to one when ϕ_i is in cluster l, and zero else. Parameters in the same cluster ($l = 1, ..., L_{\max}$) take on the same value, which we denote by $\bar{\phi}_l$. $L_{\max}$ is a truncation, as the untruncated DP prior has $L_{\max} \to \infty$.[2] The probability of being assigned to cluster l is denoted by p_l,

$$p_l = \beta_l \prod_{k=1}^{l-1} (1 - \beta_k), \; l = 1, ..., L_{\max}, \;\; \beta_l \sim \text{Beta}(1, \alpha)$$

A lower value of α results in fewer expected clusters, meaning that more parameters are likely to be shared. This can be intuitively understood through the stick-breaking representation of the Dirichlet Process (DP) prior. In this framework, the probability of an observation being assigned to cluster l, denoted as p_l, is determined by β_l, which

[1] Not to be confused with clustering standard errors, which is a method used to adjust standard errors in the presence of dependence among grouped observations.

[2] Careful choice of $L_{\max}$ and hyperprior probabilities ensures that this truncation approximation is sufficiently accurate.

represents the length of the lth piece in a sequential stick-breaking process. When α is small, the initial stick segments tend to be longer, leading to fewer total breaks. Consequently, more observations are assigned to the same cluster, reinforcing the tendency toward fewer distinct clusters.

We wish to apply this prior to a VAR model, that is, our parameters of interest now come from a matrix Φ with elements ϕ_{ij}. The DP prior can be applied directly to matrices ***if*** we want to assume all elements are independent. However, if we want to take the matrix structure of our model into account, Matrix Dirichlet Process priors are a more appropriate choice.

We build on Dunson et al. (2008), who propose a Matrix Dirichlet Process Prior in the context of meta-analysis. A key distinction between their framework and ours lies in the clustering structure. In meta-analysis, their approach allows clustering across equations but not within them, meaning that while some parameters may be shared across studies, there is no reason why one would want to cluster parameters of different explanatory variables within a single study. In contrast, in our VAR setting, we aim to allow clustering in both the row and column dimensions of the parameter matrix, enabling a more flexible dependence structure that is appropriate for VAR models. We follow the notation in Dunson et al. (2008) closely.

For simplicity, consider that we only have one lag in our VAR model, with parameter matrix Φ. Our Matrix Dirichlet Process (MDP) prior is given by:

$$\phi_{ij}|G \sim \sum_{l=1}^{L_{\max}} G_{ijl}\delta_{\bar{\phi}_l}(\phi_{ij}), \quad G \sim MDP(\alpha, \gamma, G_0), \quad \bar{\phi}_l \sim G_0$$

where the probability that ϕ_{ij} is in cluster l is given by

$$G_{ijl} = \beta_{ijl}\left\{\prod_{l<h}(1-\beta_{ijh})\right\} \tag{8}$$

with $\beta_{ijh} = \alpha_{ih}\gamma_{jh}$, where $\alpha_{ih} \sim \text{Beta}(1, \alpha)$ and $\gamma_{jh} \sim \text{Beta}(1, \gamma)$. The parameters α_{ih} and γ_{jh} measure the tendency to allocate i to cluster h and the tendency to allocate j to cluster h, respectively. The MDP prior is in principle generalizable to n dimensions, with β equal to the product of n parameters drawn from n separate Beta distributions.

As in the univariate DP, the equation for G_{ijl} follows the stick-breaking process, where lower values of α and γ lead to fewer expected clusters. While α_{ih} and γ_{jh} are independent, the variables β_{ijh} are not, as they determine the prior probabilities of parameter clustering.

An attractive feature of this prior framework is its flexibility in allowing asymmetry between prior and posterior clustering strength. Specifically, one can set $\alpha = \gamma$ or allow them to differ. In a VAR setting, this means controlling whether clustering is stronger across rows or columns of the coefficient matrix. Setting $\alpha < \gamma$ implies the researcher expects greater similarity across equations than of parameters within individual equations. As a researcher may want to remain agnostic about the prior clustering strength, one can also put priors on the distributional parameters α and γ themselves, and sample them along with the other parameters and cluster assignments.

As in Dunson et al. (2008), the Matrix Dirichlet Process Prior has the following prior probabilities for two parameters ϕ_{ik} and ϕ_{jk} being clustered together

$$P(\phi_{ik} \in l \text{ and } \phi_{jk} \in l) = \frac{1}{(\alpha+1)(\gamma+2)-1}$$

and has the desirable property that

$$P(\phi_{ik} \in l \text{ and } \phi_{jk} \in l | \phi_{im} \in q \text{ and } \phi_{jm} \in q) > P(\phi_{ik} \in l \text{ and } \phi_{jk} \in l),$$

that is, if i and j are similar in one dimension/attribute (k), it is also more likely they are similar in another dimension/attribute (m). This is particularly attractive in the context of VARs, because parameters are likely dependent; if variable i and j are similarly exposed to variable k, then they are also more likely similarly exposed to variable l. This dependence can be captured with the MDP prior.

Another advantage of this approach is that it is not necessary to abandon other priors that one might want to use – for example, for the identification of shocks – in order to reap the benefits from clustering. The MDP prior can be combined with other popular priors, like the Minnesota prior, or with priors that impose sign restrictions, by careful choice of G_0.

For ease of interpretation, the notation above focuses on the case where clustering is performed across two dimensions with a single lag (VAR(1)). However, various approaches can be used to extend this framework to multiple lags, such as in a VAR(k). Our approach clusters over pairs (i, j) rather than individual parameters, meaning that each cluster represents a vector of coefficients across the k different lags. Alternatively, one could apply independent MDP priors to each lag separately. A third approach introduces a dependency

structure in the MDP prior, ensuring that variable pairs (i, j) clustered together at one lag are also more likely to be clustered together at other lags. We favor the first approach due to its parsimony.

3.1 A Computationally Efficient Gibbs Sampler

When estimating large Bayesian VARs, it is important to have an algorithm that is computationally efficient. One can imagine two-dimensional clustering can become computationally burdensome, and, in fact, infeasible. For our version of the MDP prior and VAR models we adapt the sampling algorithm in Dunson et al. (2008), built on Ishwaran and James (2001), to allow for a fast Gibbs sampling algorithm.

Jointly drawing coefficients and clusters is challenging, but the iterative process simplifies this by leveraging conditional posteriors. Given a known form for the conditional posterior, we first draw cluster assignments conditional on the parameters, then sample parameters conditional on the clusters. Iterating avoids the need for acceptance/rejection criteria, unlike approximate samplers such as random walk Metropolis-Hastings, which rely on such schemes and can have slow convergence. The closed-form conditional posteriors and Gibbs sampling algorithm make our approach tractable and computationally efficient.

A VAR(k) model for p variables is:

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots + \Phi_k y_{t-k} + u_t, \quad t = 1, \ldots, T \tag{9}$$

with $u_t \sim \mathcal{N}(0, \Omega)$. We can stack our Vector Autoregression as: $Y = XB + \varepsilon$, with

$$y = \begin{bmatrix} y_k' \\ y_{k+1}' \\ \vdots \\ y_T' \end{bmatrix}, \qquad X = \begin{bmatrix} y_{k-1}' & y_{k-2}' & \cdots & y_1' \\ y_k' & y_{k-1}' & \cdots & y_2' \\ \vdots & \vdots & \ddots & \vdots \\ y_{T-1}' & y_{T-2}' & \cdots & y_{T-k}' \end{bmatrix}, \text{ and } \qquad B = \begin{bmatrix} \Phi_1' \\ \Phi_2' \\ \vdots \\ \Phi_k' \end{bmatrix},$$

where

$$y_t' = \begin{bmatrix} y_{1,t} & y_{2,t} & \ldots & y_{p,t} \end{bmatrix}.$$

and $\varepsilon \sim N(0, \Sigma)$ where Σ is a block diagonal matrix with blocks Ω.

For the regression parameters B, we use the conjugate normal prior and for Σ,

$$p(\Sigma) \propto |\Sigma|^{-(p+1)/2}.$$

This prior for Σ is the matrix generalization of the prior $p(\sigma^2) \propto \sigma^{-2}$, commonly used in univariate settings.

For the Gibbs sampler, we introduce three sets of additional latent variables that simplify the sampling algorithm. Denote $C_{ij} = h$ as the cluster assignment indicator, taking values $h \in \{1, ..., L_{\max}\}$, indicating that variable pair (i, j) has been assigned to cluster number h. We have

$$C_{jm} = \min\{l : S_{jml} = T_{jml} = 1\},$$

with $S_{jml} \sim$ i.i.d. Bernoulli(α_{ml}) and $T_{jml} \sim$ i.i.d. Bernoulli(γ_{jl}), for $l = 1, \ldots, L_{\max} - 1$, and $S_{jmL_{\max}} = T_{jmL_{\max}} = 1$.

The Gibbs sampler iterates between the following four steps:

(1) Sample latent variables $S_{jml}, T_{jml}, \alpha_{mh}, \gamma_{jh}$ and the cluster assignments C_{ij} conditional on the previous draw of Σ, B.

(2) Sample hyperparameters α and γ conditional on cluster assignments $\{C_{ij}\}$.

Note that this step is optional, and only needed if one uses a hyperprior on α and γ. Sampling α and γ is standard if one uses the conjugate Gamma prior (West, 1992). Throughout all applications, we use hyperprior Gamma(1,1) for both α and γ, which is fairly diffuse and lets the data speak on how much to cluster. As in West (1992), after having sampled α_{mh} and γ_{jh} in the previous step, we can update the posteriors and draw a new α and γ.

(3) Sample B conditional on the last draw of Σ and the cluster assignments C_{ij}.

(4) Sample Σ conditional on the last draw of B.

This is standard under the prior we assume; the full conditional posterior density of Σ is given by:

$$p(\Sigma|B, Y) \propto |\Sigma|^{-(T-k+p+1)/2} \exp\left(-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1}(Y - XB)'(Y - XB)\right]\right),$$

hence, we sample Σ from an inverted Wishart distribution with parameters:

$$S = (Y - XB)'(Y - XB) \quad \textit{and} \quad T - k \textit{ degrees of freedom.}$$

Next, we elaborate on steps (1) and (3) in detail. We assume conjugate priors for B and Σ.

3.1.1 Step (1): Sampling cluster assignments C_{ij}

We sample S_{mjh} and T_{mjh} by setting them to 1 for $h = C_{mj}$, and sampling them with probabilities $\kappa_{st} = P(S_{mjh} = s,\ T_{mjh} = t)$ for $h = 1, ..., C_{mj} - 1$, where

$$\kappa_{00} = \frac{(1 - \alpha_{mh})(1 - \gamma_{jh})}{(1 - \alpha_{mh}\gamma_{jh})}$$
$$\kappa_{10} = \frac{\alpha_{mh}(1 - \gamma_{jh})}{(1 - \alpha_{mh}\gamma_{jh})}$$
$$\kappa_{01} = \frac{(1 - \alpha_{mh})\gamma_{jh}}{(1 - \alpha_{mh}\gamma_{jh})},$$

and $\kappa_{11} = 1 - \kappa_{00} - \kappa_{10} - \kappa_{01}$. After sampling S_{mjh} and T_{mjh}, we can sample α_{mh} and γ_{jh} from their conditional posteriors:

$$\alpha_{mh}|... \sim \text{Beta}\left(1 + \sum_{j:\ C_{mj} \geq h} S_{mjh},\ \alpha + \sum_{j:\ C_{mj} \geq h} (1 - S_{mjh})\right)$$
$$\gamma_{jh}|... \sim \text{Beta}\left(1 + \sum_{j:\ C_{mj} \geq h} T_{mjh},\ \gamma + \sum_{j:\ C_{mj} \geq h} (1 - T_{mjh})\right)$$

With the sampled α_{mh}'s and γ_{jh}'s in hand, one can compute $\beta_{ijh} = \alpha_{ih}\gamma_{jh}$, and G_{ijl} follows from Equation (8). At last, cluster assignments C_{ij} follow a multinomial distribution, with the conditional probability that pair (i, j), with cluster assignment variable C_{ij}, is in cluster h computed as:

$$P(C_{ij} = h|\cdot) = \frac{G_{ijh}\pi(Y|X, B_{(ij)} = B_{(h)}, \cdot)}{\sum_{k=1}^{L_{\max}} G_{ijk}\pi(Y|X, B_{(ij)} = B_{(k)}, \cdot)}$$

with $\pi(\cdot)$ the posterior of the VAR which element-wise evaluates the fit if pair (i, j) is assigned to clusters $h = 1, ..., L_{\max}$, where the notation $B_{(ij)} = B_{(k)}$ indicates that the elements of B that correspond to the pair (i, j) are replaced by the cluster values in $B_{(k)}$. All other parameters in this element-wise comparison are kept fixed at their previous draws.

Note that this final step can be computationally intensive: if $L_{\max}$ is large and the number of variables is high, it requires numerous likelihood evaluations. A practical approximation to this step involves evaluating the expression only for clusters where G_{ijh} is relatively high. This can be done stochastically, for example, by randomly selecting M clusters without replacement from $h = 1, \ldots, L_{\max}$ for each (i, j) pair, with selection probabilities proportional to G_{ijh}. In this iteration, the other G_{ijm} values are set to zero, ensuring that

(i, j) cannot be assigned to cluster m in this step. However, in subsequent iterations, (i, j) retains a positive probability of being assigned to cluster m again.

3.1.2 Step (3): Sampling B conditional on cluster assignment and previous Σ

First, assume there are no clusters, and assume a conjugate prior for B. In that case, the Gibbs sampler for the VAR is completely standard, and we sample from a matricvariate normal distribution. More specifically, if we use a conjugate prior $B|\Sigma \sim MN(BP, \Sigma \otimes Q)$, we need to sample from the matricvariate normal distribution with mean $(X'X + Q^{-1})^{-1}(X'Y + Q^{-1}BP)$ and covariance $\Sigma \otimes (X'X + Q^{-1})^{-1}$. To make more notation, write

$$B \sim MN\left(M, \Sigma \otimes V\right)$$

where:

$$\begin{aligned} M &= (X'X + Q^{-1})^{-1}(X'Y + Q^{-1}BP), \\ V &= (X'X + Q^{-1})^{-1}. \end{aligned}$$

Next, we introduce the cluster assignments, drawn in step (1) of the algorithm as C_{ij}'s. These cluster assignments can be written as a set of linear equality constraints: $RB = r$. This representation is convenient, because we can use the following. The conditional posterior of B is still matricvariate normal, but now with mean $M + VR'(RVR')^{-1}(r - RM)$ and covariance $(\Sigma \otimes V) - (\Sigma \otimes VR')(RVR')^{-1}(\Sigma \otimes RV)$.

4 Performance Evaluation on Simulated Data

To assess the performance of our Bayesian framework and compare it to existing methods, we construct five realistic data-generating processes (DGPs) for VAR(1) models. The analysis considers two sample sizes: (i) $(N = 15, T = 50)$, for which we generate 50 synthetic datasets, and (ii) $(N = 30, T = 150)$, for which we generate 10 synthetic datasets. Across these experiments, we systematically vary Φ, the coefficient matrix of the VAR, and Ω, the variance-covariance matrix of the innovations. The choice of moderate N combined with small T allows the models to mimic the behavior and estimation challenges typically encountered in large VARs.

For each synthetic dataset, we simulate data of length $T + T_{\text{out}}$ based on the corresponding VAR model. We evaluate both the bias of the posterior mean and the variance of the

posterior estimates, alongside the forecasting performance on $T_{\text{out}} = 100$ out-of-sample forecasts. We use 5,000 draws for each sampler, discarding the first 500 as burn-in.

We compare our Bayesian clustering framework with two alternatives: (i) a sampler with an uninformative prior and (ii) a sampler with a conjugate variation of the Minnesota prior. We follow the approach in Bańbura et al. (2010) to select the hyperparameter of the Minnesota prior which governs the tightness of the own and cross-lags. For the prior mean of the diagonal elements, we provide the Minnesota prior with the *correct* DGP-implied mean, reflecting that researchers often have a reasonable prior guess for this parameter.

The first three of the five DGPs are based on plausible variations of the Minnesota prior, which we use to draw Φ (the coefficient matrix) and Ω (the variance-covariance matrix) for the synthetic datasets. For all three DGPs, Ω is drawn from the Inverse Wishart distribution, with scale matrix I and degrees of freedom $N + 10$.

(i) **DGP 1: Minnesota VAR –** Φ and Ω are drawn from a Minnesota prior with $\lambda = 0.1$, diagonal mean 0.5, and a decay parameter of 2, corresponding to a standard specification of the Minnesota prior.

(ii) **DGP 2: Minnesota VAR with Block Structure –** This is a variation of DGP 1, with again diagonal mean of 0.5 for Φ, but imposing a block structure on the coefficient matrix. Specifically, the matrix is divided into blocks of size $N/3$. The prior tightness of the off-diagonal elements is set to 1 within each block and 0.01 for cross-lags outside the blocks.

(iii) **DGP 3: Minnesota VAR (lower diagonal zero) –** Similar to DGP 1, but the Minnesota prior is used only for the upper-diagonal elements of Φ, while the lower-diagonal elements are set to zero. This setup mimics a hierarchical structure among the variables.

The motivation to consider processes that (closely) mimic the structure of a Minnesota prior is that one would expect a Bayesian framework that uses the Minnesota prior to result in the best inference. The results presented below show that even for DGP 1, however, the Matrix Dirichlet Process Prior performs better in terms of the statistics we consider, partly due to the uncertainty around the regularization hyperparameter of the Minnesota prior. As soon as we deviate from the standard Minnesota framework in small but empirically realistic ways, as in DGP 2 and 3, the performance difference between the Matrix Dirichlet Process prior and the Minnesota prior further diverges in favor of the MDP prior.

The remaining two DGPs are based on spatial and network structures, for which we choose Ω to have 3's on the diagonal and 1's as off-diagonal elements, and for the coefficient matrix Φ:

(iv) **DGP 4: Spatial VAR** – This DGP represents a spatial VAR model in which the entries of Φ reflect decayed influence based on the spatial proximity of variables. For each pair of variables i and j, the coefficients are determined as:

$$\phi_{ij} = c \cdot \exp(-b \cdot |i - j|),$$

where $b = 0.3$ controls the decay rate, $c = 0.5$ scales the overall influence, and $|i - j|$ is the absolute difference between indices i and j. This structure captures the notion that variables closer in index exert a stronger influence on one another, while the influence decays exponentially with distance.

(v) **DGP 5: Network VAR** – This DGP is informed by our empirical application and is based on a network VAR structure, where the coefficient matrix Φ is constructed from the 15-sector Input-Output Table published by the Bureau of Economic Analysis (2023), or a 30 sector version constructed by merging sectors in the 71-sector version. The adjacency matrix derived from this dataset is scaled by total flows to normalize the influence across nodes. To ensure stationarity, the resulting matrix is further multiplied by α=0.9 for 15 sectors, and $\alpha = 0.7$ for 30 sectors. This setup is inspired by the network VAR (NVAR) framework in Mlikota (2023), which models spillovers between variables as proportional to the strength of their connections in the network. In particular, as in Mlikota (2023):

$$y_t = \alpha_1 A y_{t-1} + ... + \alpha_p A y_{t-p} + u_t$$

with A an observed or estimated network and $\alpha_i, i = 1, \ldots, p$ a scalar.

Intuitively, we expect the MDP prior to be better suited for data structures such as these, as it accounts for dependencies among parameters – an aspect the Minnesota prior neglects by assuming parameter independence. While alternative priors specifically tailored to the respective DGPs may also perform well, a key appeal of the MDP prior is its consistent performance across a diverse range of DGPs including when the assumed DGP is misspecified.

Table 1 compares the bias of the posterior means and the posterior standard deviations for the diagonal and non-diagonal elements of the coefficient matrix across the uninformative prior, Minnesota prior, and MDP prior estimates for the five DGPs and two sample

sizes, ($N = 15, T = 50$) and ($N = 30, T = 150$). Separating the diagonal and non-diagonal elements provides valuable insights into how the methods perform under different dependency structures, highlighting their respective strengths and limitations.

The results indicate that, compared to the uninformative prior, the Minnesota prior increases bias in the diagonal elements but reduces the posterior variance for both diagonal and non-diagonal elements across all DGPs. The effect of the Minnesota prior on the bias of the non-diagonal elements varies depending on the specific DGP, but, for the set-up with $N = 30$, on averages increases bias relative to the uninformative prior.

In contrast, the MDP prior consistently reduces both average bias and variance of both diagonal and non-diagonal elements, compared to the uninformative prior and the Minnesota prior, across all DGPs, often substantially so. This advantage arises from the prior's ability to cluster coefficients, effectively sharing information to better estimate the coefficients. In small samples, where VAR models are prone to significant bias, this clustering reduces the small-sample bias, even when the true coefficients are not exactly equal. The trade-off involves a small misspecification bias cost, which is outweighed by the reduction in small-sample bias, resulting in an overall improvement, consistent with the theoretical results in Section 2.

Next, we evaluate the out-of-sample forecast performance in comparison to the oracle forecast (i.e., using the true values of Ω and Φ) at forecast horizons $h = 1, 2, 4$ via point forecasts using the Root Mean-Squared Forecast Error (RMSFE), averaging across all variables i, based on the posterior means. We compute the distance relative to the oracle forecast using the statistic $\log(\hat{M}/M - 1)$, where M denotes the statistic from the oracle forecast, and $\hat{M}$ the statistic from the different model posteriors. Table 2 summarizes the results and shows that, across all experiments, the MDP Prior on average results in better forecast performance than either the Minnesota prior or the uninformative prior. The gain in forecasting performance from using the MDP prior is largest in the third, fourth and fifth DGP, where the prior can exploit symmetries in the Spatial VAR of DGP 4, and sparsity in the network VAR of DGP 5 and hierarchical Minnesota VAR in DGP 3.

Table 1: Average bias (based on posterior mean) and standard deviation of posteriors

N	T	Statistic	Uninformative Prior	Minnesota Prior	MDP Prior
		DGP 1 – Minnesota VAR			
15	50	Ave. bias diag.	0.068	0.203	0.018
		Ave. bias non-diag.	0.051	0.026	0.007
		Ave. std. diag.	0.259	0.153	0.132
		Ave. std. non-diag.	0.265	0.155	0.065
30	150	Ave. bias diag.	0.021	0.213	0.009
		Ave. bias non-diag.	0.016	0.022	0.006
		Ave. std. diag.	0.131	0.103	0.066
		Ave. std. non-diag.	0.132	0.104	0.035
		DGP 2 – Minnesota VAR with Block Structure			
15	50	Ave. bias diag.	0.044	0.204	0.024
		Ave. bias non-diag.	0.037	0.031	0.009
		Ave. std. diag.	0.234	0.145	0.122
		Ave. std. non-diag.	0.239	0.147	0.056
30	150	Ave. bias diag.	0.019	0.155	0.011
		Ave. bias non-diag.	0.013	0.025	0.005
		Ave. std. diag.	0.121	0.097	0.071
		Ave. std. non-diag.	0.122	0.098	0.033
		DGP 3 – Minnesota VAR (lower diagonal zero)			
15	50	Ave. bias diag.	0.075	0.210	0.017
		Ave. bias non-diag.	0.057	0.026	0.007
		Ave. std. diag.	0.271	0.155	0.119
		Ave. std. non-diag.	0.278	0.158	0.054
30	150	Ave. bias diag.	0.025	0.221	0.005
		Ave. bias non-diag.	0.018	0.024	0.004
		Ave. std. diag.	0.136	0.108	0.059
		Ave. std. non-diag.	0.138	0.109	0.031
		DGP 4 – Spatial VAR			
15	50	Ave. bias diag.	0.043	0.066	0.002
		Ave. bias non-diag.	0.042	0.063	0.019
		Ave. std. diag.	0.262	0.168	0.144
		Ave. std. non-diag.	0.264	0.170	0.047
30	150	Ave. bias diag.	0.013	0.029	0.007
		Ave. bias non-diag.	0.012	0.018	0.001
		Ave. std. diag.	0.124	0.104	0.095
		Ave. std. non-diag.	0.124	0.104	0.017
		DGP 5 – Network VAR			
15	50	Ave. bias diag.	0.042	0.118	0.022
		Ave. bias non-diag.	0.041	0.066	0.005
		Ave. std. diag.	0.252	0.162	0.144
		Ave. std. non-diag.	0.254	0.164	0.049
30	150	Ave. bias diag.	0.037	0.235	0.005
		Ave. bias non-diag.	0.011	0.016	0.001
		Ave. std. diag.	0.110	0.095	0.060
		Ave. std. non-diag.	0.111	0.096	0.012

Table 2: Distance of average RMSFE to the oracle forecast, computed as $\log(\hat{M}/M_0 - 1)$*, where* $\hat{M}$ *is the RMSFE following from the three different posteriors, and* M_0 *the RMSFE from the oracle forecast based on the true DGP parameters. The lower* $\log(\hat{M}/M_0 - 1)$*, the closer the model performance to the oracle forecast.*

N	T	Statistic	Horizon	Uninformative Prior	Minnesota Prior	MDP Prior
DGP 1 – Minnesota VAR						
15	30	RMSFE	1	-1.21	-1.70	-2.13
			2	-1.53	-1.93	-2.39
			4	-1.98	-2.37	-2.86
30	150	RMSFE	1	-1.80	-1.79	-1.79
			2	-1.95	-1.95	-2.13
			4	-2.02	-2.11	-2.22
DGP 2 – Minnesota VAR with Block Structure						
15	50	RMSFE	1	-1.23	-1.42	-2.04
			2	-1.50	-1.63	-2.25
			4	-1.85	-1.91	-2.42
30	150	RMSFE	1	-1.90	-1.79	-2.17
			2	-2.18	-2.02	-2.34
			4	-2.40	-2.23	-2.58
DGP 3 – Minnesota VAR (lower diagonal zero)						
15	50	RMSFE	1	-1.20	-1.70	-2.56
			2	-1.54	-1.92	-2.91
			4	-2.00	-2.42	-3.71
30	150	RMSFE	1	-1.83	-1.78	-2.80
			2	-2.03	-1.96	-3.39
			4	-2.35	-2.37	-4.33
DGP 4 – Spatial VAR						
15	50	RMSFE	1	-1.51	-1.15	-3.39
			2	-2.13	-1.66	-4.08
			4	-2.78	-2.16	-4.16
30	150	RMSFE	1	-2.05	-1.76	-3.42
			2	-2.85	-2.42	-3.70
			4	-3.52	-3.15	-4.17
DGP 5 – Network VAR						
15	50	RMSFE	1	-1.47	-1.10	-2.99
			2	-2.02	-1.55	-3.65
			4	-2.76	-2.07	-3.93
30	150	RMSFE	1	-1.63	-1.29	-5.35
			2	-1.94	-1.66	-5.08
			4	-1.63	-1.29	-5.35

5 First Application: PPI Spillovers Across Industries

To illustrate the MDP Prior for VAR models further, we consider two empirical applications with large-dimensional systems of equations. In both cases, we are interested in the propagation of shocks across units and analyze the performance of the MDP Prior relative to the Minnesota prior, a popular tool for the estimation of large BVARs.

The first empirical application examines industry-level monthly producer price indices (PPI) from the Bureau of Labor Statistics (2007–2024).[3] Network theory suggests that shocks propagate through inter-industry linkages, motivating a Network-VAR (NVAR) representation (Mlikota, 2023) or a dynamic factor structure (Foerster et al., 2011) to reduce the number of estimated parameters. However, these approaches are restrictive, as spillovers may arise from multiple networks (e.g., banking linkages and production networks), and sectoral responses can be heterogeneous due to buffers that affect shock sensitivity. An unrestricted vector autoregression provides greater flexibility but suffers from high dimensionality, leading to noisy estimates. In fact, in this particular setting, a VAR of order higher than four would be infeasible as the number of parameters exceeds the number of observations. MDP Priors provide a flexible framework for imposing parameter restrictions, as such aiding inference and reducing noise.

5.1 Data and implementation details

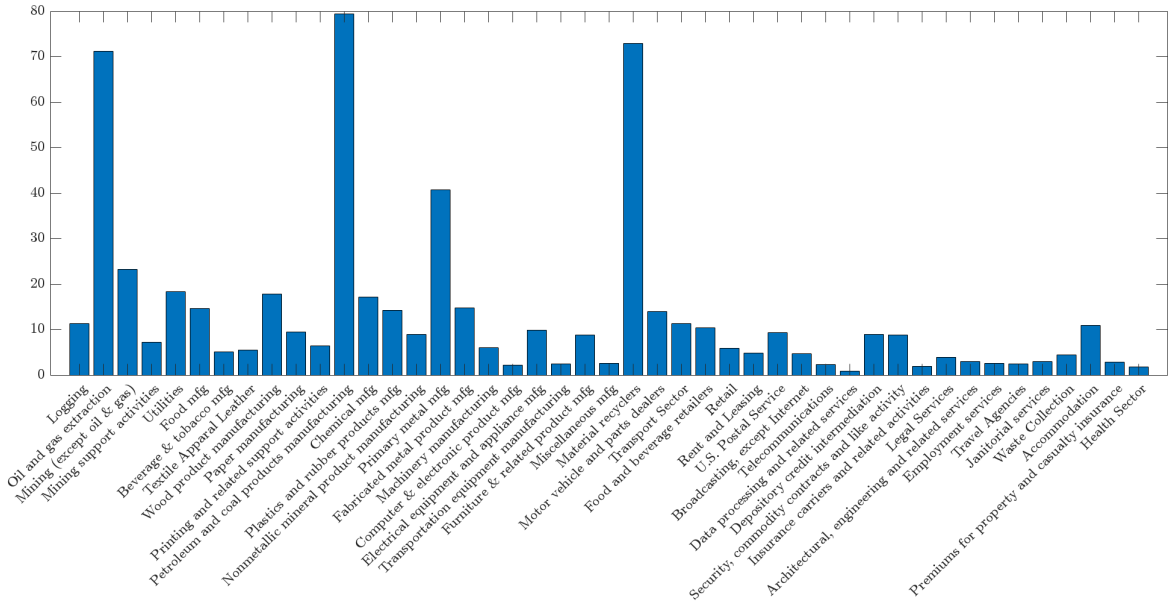
The data are aggregated into 45 sectors by computing weighted averages of the merged sectors, where weights are based on industry-level output in 2007, obtained from the Bureau of Economic Analysis.[4] The aggregation procedure involves merging: (i) smaller transportation industries (e.g., air, rail, water, truck, and transit) into a single transportation sector; (ii) smaller healthcare industries into a comprehensive healthcare sector; (iii) various real estate categories into a single real estate sector; (iv) most retail industries, except for food and beverages; and (v) textile, apparel, and leather products into one category. To remove long-run trends, we apply a Hamilton filter at a monthly frequency to the aggregated PPI series.

The series are ordered from upstream to downstream to facilitate the interpretation of impulse response functions, computed using Cholesky decomposition, where shocks originate in downstream industries and propagate upstream. The ordering follows: (i) resource extraction (logging, oil and gas extraction, mining, and mining support); (ii) utilities; (iii)

[3] For estimation, we exclude the last 24 observations (two years) to reserve them for out-of-sample forecast evaluation.

[4] This aggregation follows the approach in Mlikota (2023).

Figure 2: Standard deviations of industry PPI



manufacturing (food, beverages, textiles, apparel, leather, wood, paper, chemicals, plastics, minerals, metals, machinery, computers, and transportation equipment); (iv) motor vehicle and parts dealers; (v) transportation; (vi) retail; and (vii) services.

Figure 2 presents the standard deviations of industry prices, ordered from downstream to upstream. The results indicate substantial variation in volatility across industries, with upstream sectors exhibiting greater price fluctuations on average compared to downstream sectors (e.g., retail and services), which tend to be relatively more stable.

To implement the MDP prior, cluster assignment follows the priors $\alpha_{ih} \sim \text{Beta}(1, \alpha)$ and $\gamma_{jh} \sim \text{Beta}(1, \gamma)$. The number of clusters is determined by the data through a hierarchical prior structure, where α and γ follow conjugate Gamma priors, $\text{Gamma}(\eta_1, \eta_2)$, both set to 1 to remain uninformative. We use uninformative priors for the VAR lag coefficients and the Jeffrey's prior for the covariance matrix of the error terms as in the methodology section. To determine the lag length, we sample $p \in 1, 2, 3, 4$ with equal priors in our Gibbs-sampling routine, assuming a multinomial posterior proportional to the likelihoods under different lag specifications. The posterior favors $p = 2$, which is therefore used in the analysis. For $p = 2$, we generate 20,000 draws, discarding the first 25% as burn-in, and present the results below.

For comparison, we also estimate the model under a noninformative prior, which coincides with the maximum likelihood estimator, and a conjugate Minnesota-like prior, where the first lag's diagonal elements have prior mean 0.5, and higher-order lags' prior mean is set to zero, using the standard Matlab implementation. The shrinkage parameter is selected as in Bańbura et al. (2010).

5.2 Posteriors

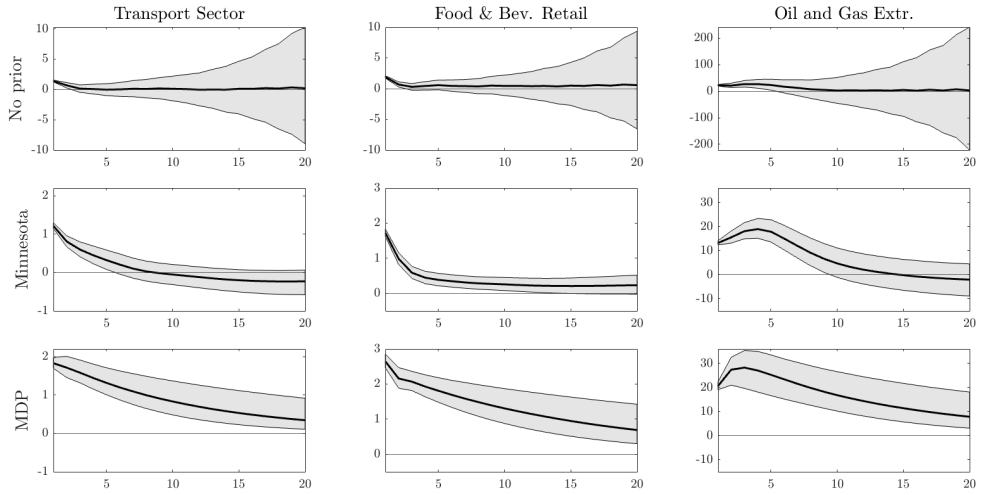
Figures B.1 and B.2 in the Appendix present selected posterior distributions of the estimated coefficients under the three prior specifications. Figure B.1 particularly highlights the effectiveness of the MDP prior (shown in red); by clustering coefficients, it sharpens identification, setting many to precisely zero, whereas the other priors yield noisier inference. Figure B.2 focuses on the Primary Metal Manufacturing and Fabricated Metal Product Manufacturing sectors, where coefficients are clearly nonzero. The MDP prior produces tighter posteriors, allowing it to detect significant positive first-lag effects and negative second-lag effects, enhancing interpretability compared to the other priors.

Appendix Section B.2 provides further insights into the clusters formed by the MDP prior over the last 500 draws of the sampler, revealing the presence of eight sizeable clusters. The largest cluster, a zero cluster, contains more than 950 parameters, indicating strong regularization. The remaining clusters have nonzero cluster means but have fewer members; for instance, one cluster (row 7 in Table B.1) has a mean coefficient of approximately 0.1 for the first lag and -0.05 for the second lag, grouping 40 coefficients together. Figure B.3 illustrates which parameters tend to be clustered together, showing that many service sectors are assigned to the same clusters, highlighting commonalities in their estimated dynamics.

5.3 Impulse Response Functions

Figure 3 presents the impulse response functions (IRFs) to an industry's own shocks, along with 80% credible intervals, for the Transportation, Food and Beverage Retail, and Oil and Gas Extraction sectors, under the three prior choices. Under the uninformative prior (top row), the own-shock IRFs are not significantly different from zero after five months, and the error bands widen substantially, making inference imprecise beyond a few periods. Both the Minnesota prior (middle row) and MDP prior (bottom row) result in more narrow credible intervals. However, the Minnesota prior, by construction, shrinks lagged effects for $p > 1$ toward zero, resulting in less persistent IRFs compared to the MDP prior. The MDP prior, in contrast, identifies statistically significant IRFs beyond a year, consistent with economic intuition, as price levels are known to be highly persistent.

Figure 3: Selected Impulse Response Functions: Own Shocks



Turning to shock propagation across industries, Figure 4 presents impulse response functions (IRFs) for upstream-to-downstream transmission across five industry pairs. Specifically, we examine the impact of positive shocks from Oil and Gas Extraction on Utilities (first column) and Petroleum and Coal Manufacturing (second column), from Mining to Primary Metal Manufacturing (third column), from Transportation Equipment Manufacturing to Motor Vehicles and Parts Dealers (fourth column), and from Wood Product Manufacturing to Paper Manufacturing (fifth column). Estimation under the uninformative prior produces wide posteriors, making inference noisy (as reflected in the larger y-axis scale relative to the other priors). The Minnesota prior leads to counterintuitive results, notably suggesting that a positive shock to Oil and Gas Extraction lowers Utility prices and that an increase in Wood Product Manufacturing prices reduces Paper Manufacturing prices. In contrast, the MDP prior yields economically intuitive IRFs, many of them statistically significant.

5.4 Forecast Error Variance Decomposition

We next analyze forecast error variance decompositions (FEVDs) to determine which sectors contribute most to each other's forecast uncertainty. Tables 3 and 4 present these decompositions under the MDP prior, Minnesota prior, and an uninformative prior, for two different subsets of sectors.

Figure 4: Selected Impulse Response Functions based on Upstream to Downstream Cholesky Decomposition

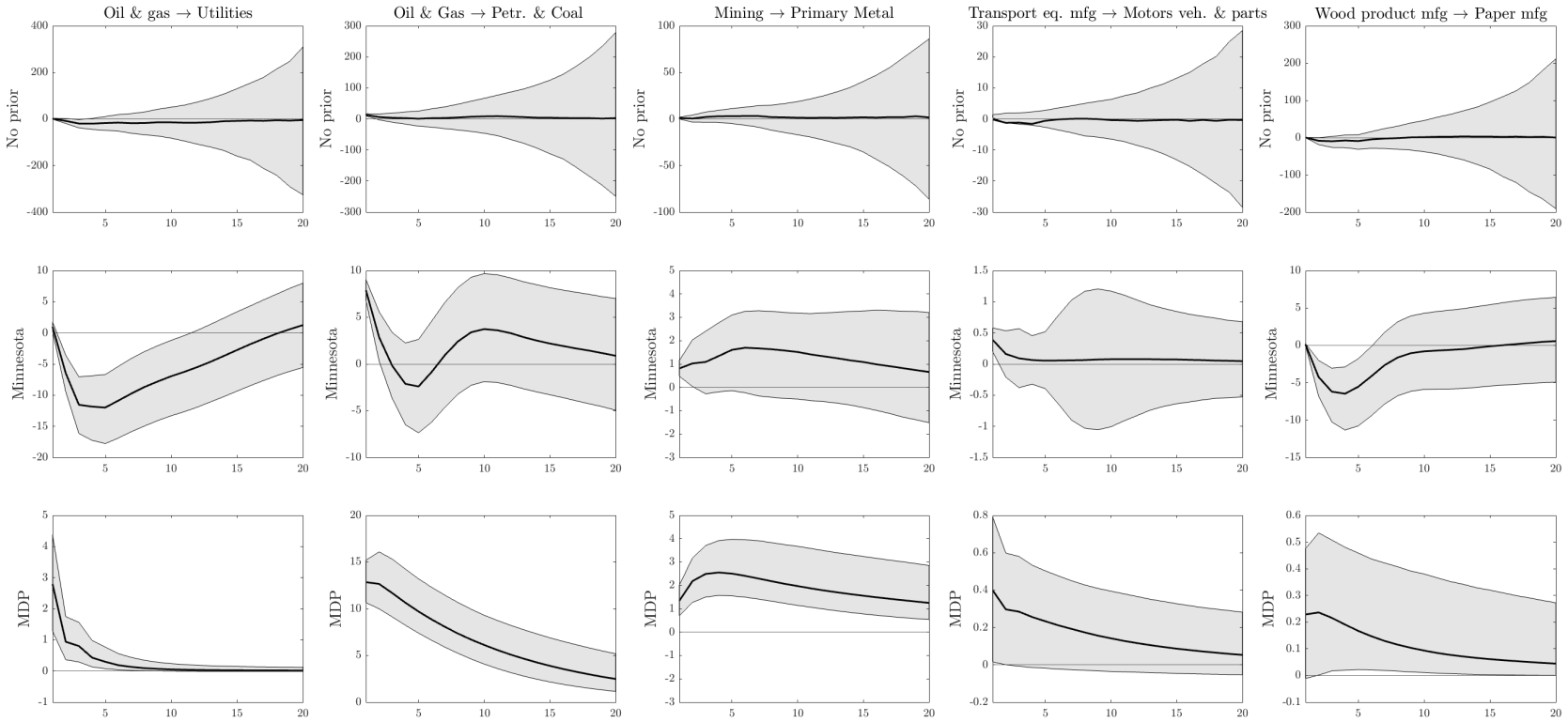


Table 3 examines the Oil and Gas Extraction sector, Petroleum and Coal Products Manufacturing, the Transport sector, and the Retail sector. Both the Minnesota and uninformative priors assign a dominant role to Oil and Gas Extraction, attributing to it a large share of forecast error variance in the other sectors. This result is largely driven by the volatility patterns in Figure 2 which dominate the estimation and result in wide posteriors, as shown in Figure B.1 for the Printing sector. In contrast, the MDP prior provides a more economically meaningful decomposition. In addition to identifying Oil and Gas prices, Petroleum and Coal Manufacturing, and Utilities prices as key drivers for prices in the Transport sector and Retail sector, it also preserves a significant role for each industry's own shocks, which aligns with the well-documented persistence in price levels.

Table 4 further illustrates the advantages of the MDP prior, as it correctly attributes the forecast error variance from both Metal Manufacturing industries to the Mining industry, and from the Fabricated Metal Product Manufacturing to the Primary Metal Manufacturing. The other priors fail to capture these expected relationships, reinforcing the MDP prior's ability to uncover meaningful sectoral dynamics.

Table 3: Forecast Error Variance Decomposition (after 6 months: oil, transport and retail)

Sector (contribution of row to column)	Oil and gas extr.	Petroleum & Coal Products mfg	Transport Sector	Retail
		No Prior		
Oil and gas extr.	0.794	0.313	0.389	0.306
Petroleum & Coal Prod. mfg	0.099	0.474	0.320	0.167
Transport Sector	<0.001	<0.001	<0.001	<0.001
Retail	<0.0011	<0.001	<0.001	<0.001
		Minnesota Prior		
Oil and gas extr.	0.795	0.291	0.519	0.602
Petroleum & Coal Prod. mfg	0.089	0.567	0.277	0.102
Transport Sector	<0.001	<0.001	0.001	<0.001
Retail	<0.001	<0.001	<0.001	<0.001
		MDP prior		
Oil and gas extr.	0.934	0.385	0.092	0.132
Petroleum & Coal Prod. mfg	0.023	0.525	0.107	0.047
Transport Sector	<0.001	<0.001	0.501	0.013
Retail	<0.001	<0.001	<0.001	0.487

Table 4: Forecast Error Variance Decomposition (after 6 months: mining and metal)

Sector (contribution of row to →)	Mining extr.	Primary Metal mfg	Fabricated Metal mfg
No Prior			
Mining (except oil & gas)	0.083	0.012	0.010
Primary Metal mfg	0.007	0.040	0.008
Fabricated Metal Product mfg	<0.001	<0.001	<0.001
Minnesota Prior			
Mining (except oil & gas)	0.348	0.008	0.012
Primary Metal mfg	0.348	0.008	0.012
Fabricated Metal Product mfg	0.003	<0.001	<0.001
MDP Prior			
Mining (except oil & gas)	0.960	0.096	0.111
Primary Metal mfg	<0.001	0.684	0.100
Fabricated Metal Product mfg	<0.001	0.003	0.199

5.5 Forecasting Performance

We evaluate forecasting performance under the three prior choices using the most recent 24 months of data. Figure 5 presents one-step-ahead fitted values from the Minnesota prior and MDP estimation for eight selected sectors, compared to the realized values (in red). Figure 6 reports the Root Mean Squared Forecast Errors (RMSFEs) for the first four forecast horizons.

The MDP prior consistently produces fitted values that closely track the realized data, outperforming both alternatives. This dominant forecasting performance is further reflected in the substantially lower RMSFEs relative to the other two priors. While the Minnesota prior improves over the uninformative prior by imposing shrinkage, the MDP prior's clustering structure enhances forecast accuracy even further. For several sectors, the MDP prior leads to RMSFEs that are several orders of magnitude lower than those of the Minnesota prior or flat prior, in particular, for the Food sector, Plastics and Rubber Product Manufacturing, Chemical Manufacturing, the Transport sector, and Retail.

Figure 5: Fitted values for one-step ahead forecast, selected sectors

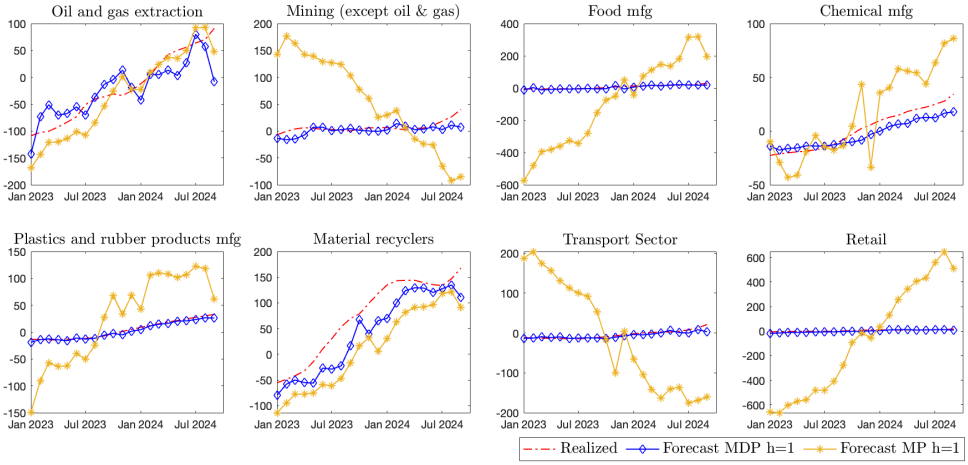
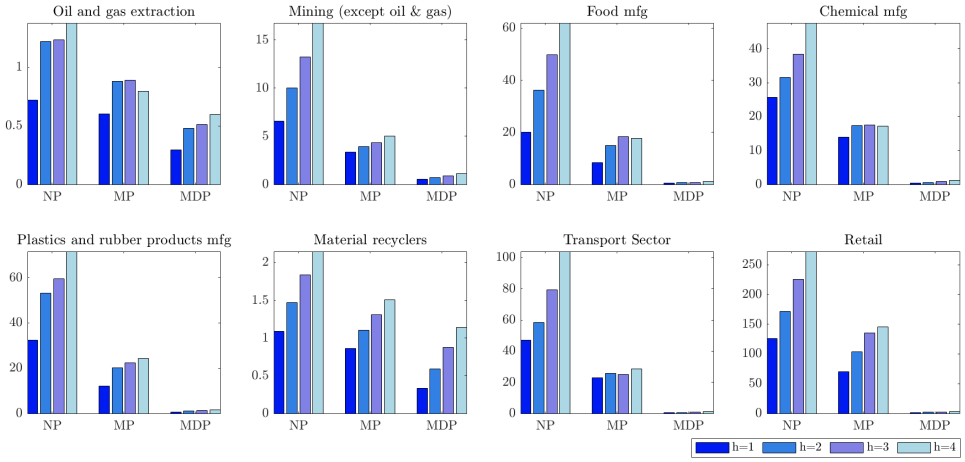


Figure 6: RMSFE at four horizons, selected sectors



6 Second application: Forecasting State-Level Personal Income

In the second application, we analyze quarterly state-level personal income data for the United States from 1948 to 2023.[5] Forecasting state-level income is key for, among other things, tax revenue projections and federal funding allocation. Because states are economically interconnected, we account for cross-state dependencies by estimating a Bayesian Vector Autoregression (BVAR) model for the demeaned growth rates of state-level personal income, comparing three different prior specifications.

[5]Data obtained from https://www.bea.gov/data/income-saving/personal-income-by-state on November 30, 2023.

The MDP prior is well-suited for this setting, as state economies are interdependent, and the MDP reflects these cross-state dependencies. Because our model treats state-level income growth rates as individual variables, it results in a large VAR with 51 variables; one per state, plus D.C. We compare the forecasting performance of the MDP BVAR against models with a flat prior and the Minnesota prior, tuning the latter following Bańbura et al. (2010). Because the model is in growth rates, we shrink the diagonal elements of the first lag coefficients toward zero under the Minnesota prior, reflecting the limited persistence of income growth rates. With this in mind, we also restrict the VAR model to two lags, balancing parsimony and dynamic accuracy.

6.1 Impulse Response Functions

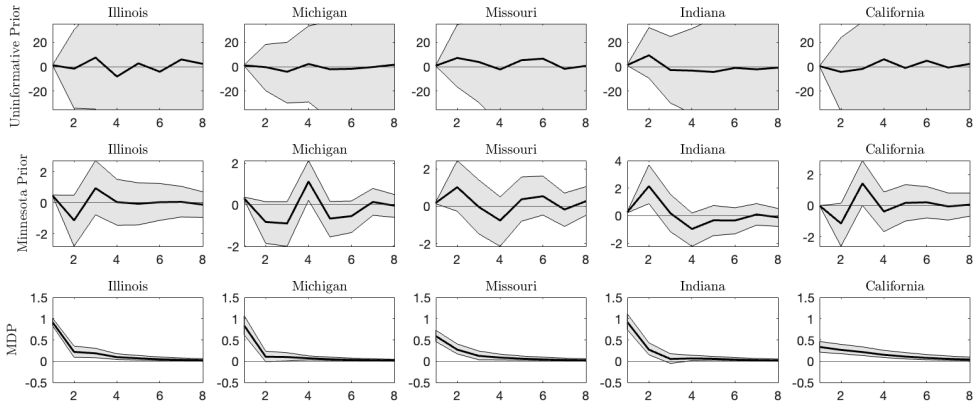
Figure 7 presents selected model-implied Impulse Response Functions (IRFs) along with their credible intervals. We use Generalized Impulse Responses (GIRFs) following Pesaran and Shin (1998), as the model lacks a natural variable ordering. Specifically, we analyze the effects of a shock originating in three states: Illinois, California, and South Dakota.

The results indicate that the MDP BVAR (third row) produces IRFs with tighter credible intervals and a more economically intuitive interpretation. For example, a shock in Illinois has notable spillover effects on neighboring states (e.g., Michigan, Indiana), while its impact on more distant states (e.g., California) is small. Similarly, a shock originating in California has a sizeable impact on Arizona, Nevada, and Colorado, but smaller effects on Michigan.

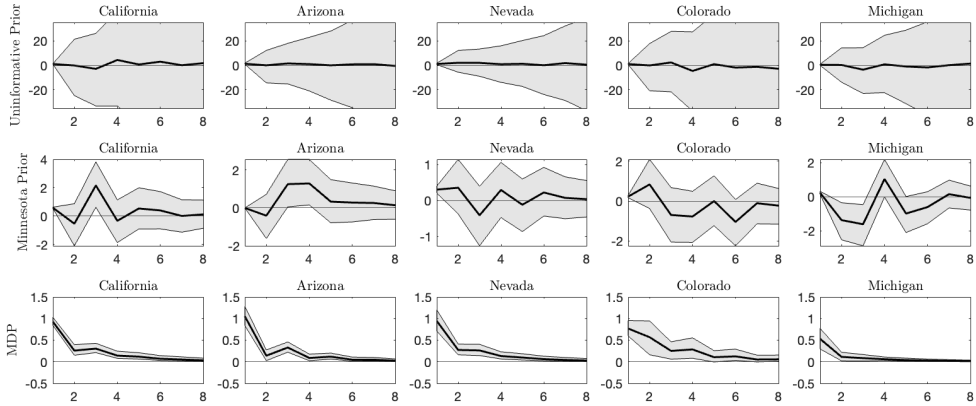
In contrast, a shock originating in South Dakota has large but short-lived effects on North Dakota and modest effects on Montana, but propagates little beyond those neighboring states. Because the model is specified in growth rates, the responses to these shocks show little persistence. However, even when the long-term effects (6 quarters after the shock) are zero, the initial impact of the shock is statistically significant in the first period and up to 2-3 periods after. The Impulse Response Functions obtained from the uninformative prior and Minnesota prior, on the other hand, produce noisier inference, sometimes with counterintuitive signs.

Figure 7: Selected Generalized Impulse Response Functions

(a) Propagation of a Illinois Shock



(b) Propagation of a California shock



(c) Propagation of a South Dakota shock

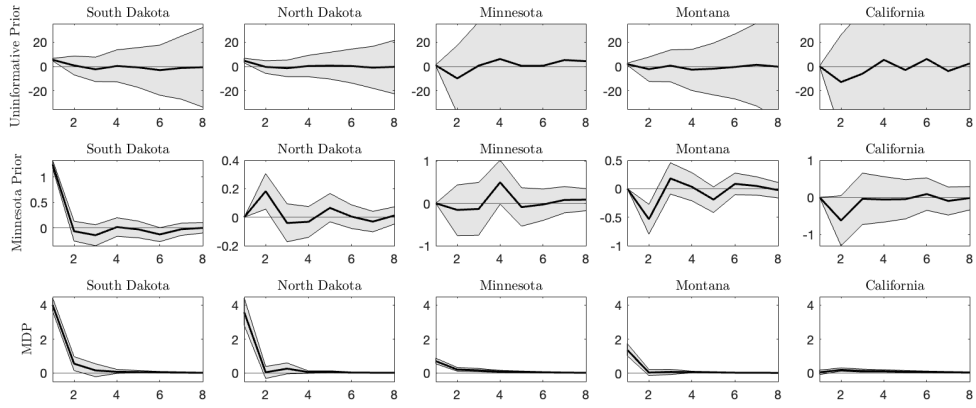
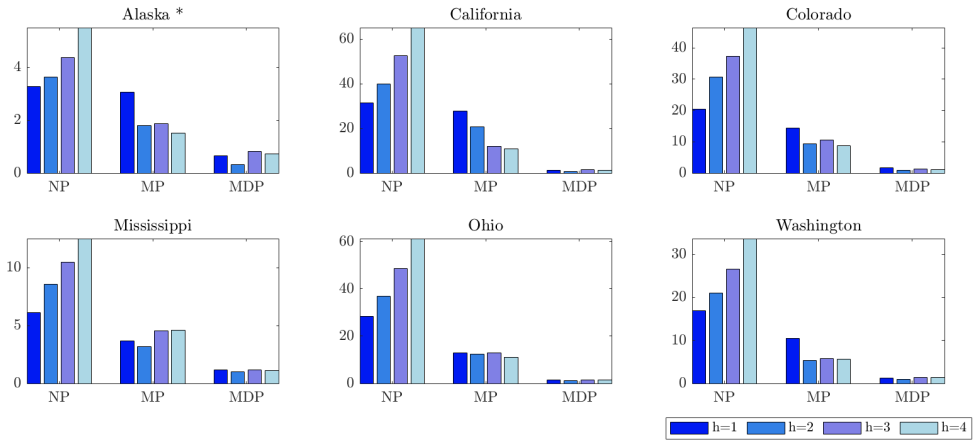


Figure 8: Root Mean Squared Forecast Errors for $h = 1, 2, 3, 4$ step ahead forecasts for No Prior, Minnesota Prior (MP) and MDP prio for out-of-sample forecasts over the last 25 years. RMSFE divided by standard deviation of each respective series. Model estimated using data up to 100 quarters ago (no rolling window).



6.2 Forecasting Performance

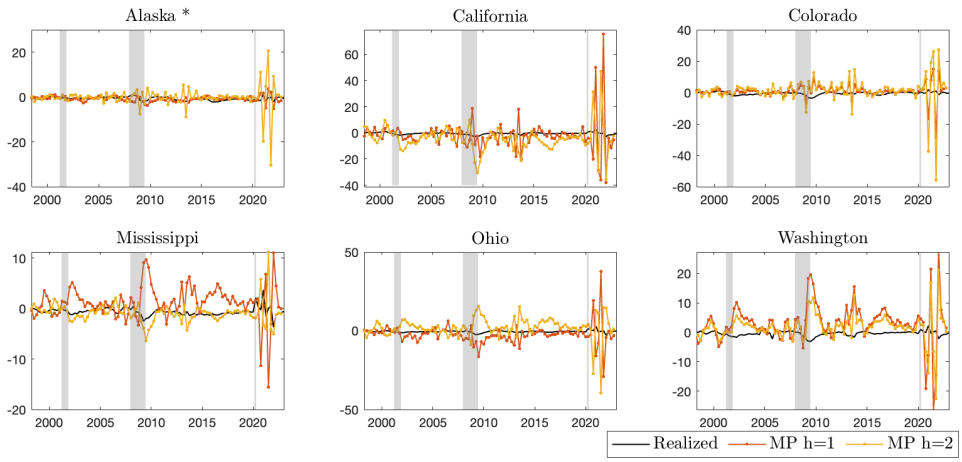
Next, we evaluate the forecasting performance of the state-level BVAR(2) estimated with three different priors. Figure C.1 shows the composition of the importance of all states to the forecasting variance of the other states based on the estimates using the MDP, and the results are intuitive. For example, Midwest states depend closely on other Midwest states, and we observe similar results for the southern states, further motivating the use of a VAR for forecasting.

Figure 8 reports the Root Mean Squared Forecast Errors (RMSFE) for $h = 1, 2, 3, 4$ step-ahead forecasts, normalized by the standard deviation of each respective series, for a selected number of states. The model is estimated using data up to 100 quarters ago (without a rolling window) and forecasts the subsequent 100 quarters. Appendix Figure C.2 shows the RMSFE for all states.

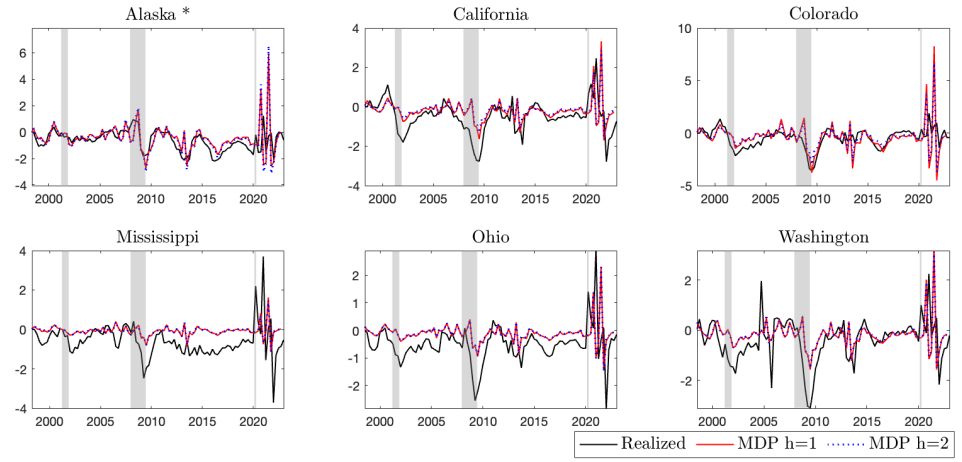
Additionally, Figures 9a and 9b plot the time series of the forecasts alongside the realized data for the Minnesota Prior and MDP prior, respectively. Given the poor performance of the BVAR with uninformative prior, as evident from Figure 8, we do not visualize its forecasts. While the Minnesota Prior clearly outperforms the BVAR with uninformative prior, both the RMSFE results and the forecast time series show how the MDP prior achieves substantially better forecasting performance than both priors, with RMSFEs that are orders of magnitude smaller. This finding is again consistent with our simulation evidence in Section 4.

Figure 9: Twenty Years of Out-of-Sample Forecasting using State-level BVAR(2), $h = 1$ and $h = 2$ steps ahead. Shaded areas are NBER recessions.

(a) Based on Minnesota Prior



(b) Based on MDP Prior



7 Conclusion

This paper proposes the use of the Matrix Dirichlet Process (MDP) prior for estimating large Bayesian VARs. The theoretical motivation for this prior comes from the observation that large VARs can exhibit substantial second-order bias, which can be mitigated by imposing equality restrictions. However, incorrectly specified restrictions may introduce misspecification bias. The MDP prior balances this trade-off by identifying and imposing equality constraints that are well-supported by the data. As a result, the MDP prior reduces both variance and second-order bias, challenging the conventional view that regularization methods necessarily involve a bias-variance trade-off.

To make the MDP prior computationally feasible in a large Bayesian VAR framework, we develop a computationally efficient sampling algorithm. The MDP prior effectively exploits the sparsity often observed in VAR coefficient matrices by clustering coefficients into groups. Zero coefficients are tightly estimated as a single cluster, while clusters of non-zero coefficients are identified and estimated with greater precision by sharing information across parameters. In two empirical applications, we demonstrate that the MDP prior achieves both improved forecasting performance and more interpretable results compared to traditional shrinkage priors, such as the Minnesota prior. These results suggest that the MDP prior provides a promising alternative for modeling large VARs.

Large Bayesian VARs have been extended to incorporate stochastic volatility (Carriero, Clark, and Marcellino, 2016), time variation (Chan, 2023), and structural restrictions (Chan, 2023). While these extensions are beyond the scope of this paper, they can likely be integrated with the MDP prior. Notably, our implementation accommodates other priors on the coefficient matrices and leaves variance-covariance estimation unrestricted, making it compatible with a wide range of Bayesian VAR frameworks.

References

Atalay, E. (2017). "How Important Are Sectoral Shocks?". *American Economic Journal: Macroeconomics*, *9*(4), 254–280. doi: 10.1257/mac.20160353

Bańbura, M., Giannone, D., and Reichlin, L. (2010). "Large Bayesian Vector Auto Regressions". *Journal of Applied Econometrics*, *25*(1), 71–92.

Bao, Y., and Ullah, A. (2007). "The Second-Order Bias and Mean Squared Error of Estimators in Time-Series Models"'. *Journal of Econometrics*, *140*(2), 650–669.

Bernanke, B. S., Boivin, J., and 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. Retrieved from https://academic.oup.com/qje/article-abstract/120/1/387/1931468 doi: 10.1162/0033553053327452

Billio, M., Casarin, R., and Rossini, L. (2019). "Bayesian Nonparametric Sparse VAR Models". *Journal of Econometrics*, *212*(1), 97–115.

Bonhomme, S., and Manresa, E. (2015). "Grouped Patterns of Heterogeneity in Panel Data"'. *Econometrica*, *83*(3), 1147–1184.

Carriero, A., Clark, T. E., and Marcellino, M. (2016). "Common Drifting Volatility in Large Bayesian VARs". *Journal of Business & Economic Statistics*, *34*(3), 375–390.

Chan, J. C. (2022). "Asymmetric Conjugate Priors for Large Bayesian VARs". *Quantitative Economics*, *13*(3), 1145–1169.

Chan, J. C. (2023). "Large Hybrid Time-Varying Parameter VARs". *Journal of Business & Economic Statistics*, *41*(3), 890–905.

De Mol, C., Giannone, D., and Reichlin, L. (2008). "Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?". *Journal of Econometrics*, *146*(2), 318–328.

Dunson, D. B., Xue, Y., and Carin, L. (2008). "The Matrix Stick-Breaking Process". *Journal of the American Statistical Association*, *103*(481), 317–327.

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems". *The Annals of Statistics*, 209–230.

Foerster, A. T., Sarte, P.-D. G., and Watson, M. W. (2011). "Sectoral versus Aggregate Shocks: A Structural Factor Analysis of Industrial Production"'. *Journal of Political Economy*, *119*(1), 1–38.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). "The Generalized Dynamic-Factor Model: Identification and Estimation". *Review of Economics and Statistics*, *82*(4), 540–554.

George, E. I., and McCulloch, R. E. (1993). "Variable Selection via Gibbs Sampling". *Journal of the American Statistical Association*, *88*(423), 881–889.

George, E. I., Sun, D., and Ni, S. (2008). "Bayesian Stochastic Search for VAR Model Restrictions". *Journal of Econometrics*, *142*(1), 553–580.

Giannone, D., Lenza, M., and Primiceri, G. E. (2019). "Priors for the Long Run". *Journal of the American Statistical Association*, *114*(526), 565–580.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). "Economic Predictions with Big Data: The Illusion of Sparsity"'. *Econometrica*, *89*(5), 2409–2437.

Hahn, J., and Moon, H. R. (2010). "Panel Data Models with Finite Number of Multiple Equilibria"'. *Econometric Theory*, *26*(3), 863–881.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Ishwaran, H., and James, L. F. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors". *Journal of the American Statistical Association*, *96*(453), 161–173.

Jordà, Ò. (2005). "Estimation and Inference of Impulse Responses by Local Projections". *American Economic Review*, *95*(1), 161–182.

Koop, G. (2013). "Forecasting with Medium and Large Bayesian VARs". *Journal of Applied Econometrics*, *28*(2), 177–203.

Koop, G., Korobilis, D., and Pettenuzzo, D. (2019). "Bayesian Compressed Vector Autoregressions". *Journal of Econometrics*, *210*(1), 135–154.

Li, D., Plagborg-Møller, M., and Wolf, C. K. (2024). "Local Projections vs. VARs: Lessons from Thousands of DGPs". *Journal of Econometrics*, 105722.

Litterman, R. B. (1979). *"Techniques of Forecasting Using Vector Autoregressions"* (Working Paper No. 115). Federal Reserve Bank of Minneapolis. Retrieved from https://www.minneapolisfed.org/research/working-papers/techniques-of-forecasting-using-vector-autoregressions

Litterman, R. B. (1980). *"A Bayesian Procedure for Forecasting with Vector Autoregressions"* (Working Paper). Massachusetts Institute of Technology (MIT), Department of Economics. Retrieved from https://dspace.mit.edu/handle/1721.1/63593

Long, J. B., and Plosser, C. I. (1983). "Real Business Cycles"'. *Journal of Political Economy*, *91*(1), 39–69. doi: 10.1086/261128

Lumsdaine, R. L., Okui, R., and Wang, W. (2022). "Estimation of a Break Point in Group Membership Structure"'. *Journal of Econometrics*, *233*, 45–65.

Mlikota, M. (2023). "Cross-Sectional Dynamics Under Network Structure: Theory & Macroeconomic Applications"'.

Nagar, A. L. (1959). "The Bias and Moment Matrix of the General k-class Estimators of the Parameters in Simultaneous Equations". *Econometrica: Journal of the Econometric Society*, 575–595.

Pesaran, H. H., and Shin, Y. (1998). "Generalized Impulse Response Analysis in Linear Multivariate Models"'. *Economics letters*, *58*(1), 17–29.

Rilstone, P., Srivastava, V. K., and Ullah, A. (1996). "The Second-Order Bias and Mean Squared Error of Nonlinear Estimators". *Journal of Econometrics*, *75*(2), 369–395.

Stock, J. H., and Watson, M. W. (2002). "Forecasting using Principal Components from a Large Number of Predictors". *Journal of the American Statistical Association*, *97*(460), 1167–1179.

West, M. (1992). *"Hyperparameter Estimation in Dirichlet Process Mixture Models"*. Duke University ISDS Discussion Paper# 92-A03.

A Proof of Theorem 1

Consider the restriction $R\beta = r$ where R is a $q \times p$ matrix and r is a $q \times 1$ vector of known restrictions, we assume R is of full row rank. Denote by β_r the constrained estimator that satisfies $R\beta = r$. In addition, we assume the restriction is true, hence β_0 also satisfies $R\beta_0 = r$. Next, introduce a Lagrange multiplier λ and define the Lagrangian function:

$$\mathcal{L}(\beta, \lambda) = \psi_T(\beta)' Q \psi_T(\beta) + \lambda'(R\beta - r). \tag{A.1}$$

with $Q = \bar{H}_1^{-1}$ Differentiating with respect to β and λ, we obtain:

$$\nabla_\beta \mathcal{L} = \nabla_\beta(\psi_T(\beta)' Q \psi_T(\beta)) + R'\lambda = 0, \tag{A.2}$$

$$\nabla_\lambda \mathcal{L} = R\beta - r = 0. \tag{A.3}$$

Expanding $\psi_T(\beta)$ around β_0, where β_r is the restricted β:

$$\psi_T(\beta_r) \approx \psi_T(\beta_0) + H_1(\beta_r - \beta_0). \tag{A.4}$$

and substituting this in the first-order condition:

$$\bar{H}_1(\beta_r - \beta_0) + R'\lambda = -\psi_T(\beta_0). \tag{A.5}$$

Combining this with the constraint and our assumption $R\beta_0 = r$ holds for β_0, we obtain

$$\begin{bmatrix} \bar{H}_1 & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \beta_r - \beta_0 \\ \lambda \end{bmatrix} = \begin{bmatrix} -\psi_T(\beta_0) \\ 0 \end{bmatrix} \tag{A.6}$$

Solving gives:

$$\beta_r - \beta_0 \approx -Q_r \psi_T(\beta_0), \tag{A.7}$$

where:

$$Q_r = \bar{H}_1^{-1} - \bar{H}_1^{-1} R'(R\bar{H}_1^{-1}R')^{-1} R\bar{H}_1^{-1}. \tag{A.8}$$

It follows that the difference between the second-order bias with and without restriction is given by

$$\Delta a_{-1} = (Q - Q_r)\left(V a_{-1/2} + \frac{1}{2}\bar{H}_2[a_{-1/2} \otimes a_{-1/2}] \right). \tag{A.9}$$

where we use that $a_{-1/2} = a^r_{-1/2}$ because β_0 satisfies the restrictions $R\beta = r$ and

$$Q - Q_r = \bar{H}_1^{-1} - \left(\bar{H}_1^{-1} - \bar{H}_1^{-1}R'(R\bar{H}_1^{-1}R')^{-1}R\bar{H}_1^{-1}\right) \tag{A.10}$$

$$= \bar{H}_1^{-1}R'(R\bar{H}_1^{-1}R')^{-1}R\bar{H}_1^{-1}. \tag{A.11}$$

Given the positive semi-definiteness of $Q - Q_r$, it follows that introducing the restrictions $R\beta = r$ results in a decrease in the second-order bias a_{-1}.

B First Application: Supplemental Tables and Figures

B.1 Selected Posteriors

Figure B.1: Selected posteriors

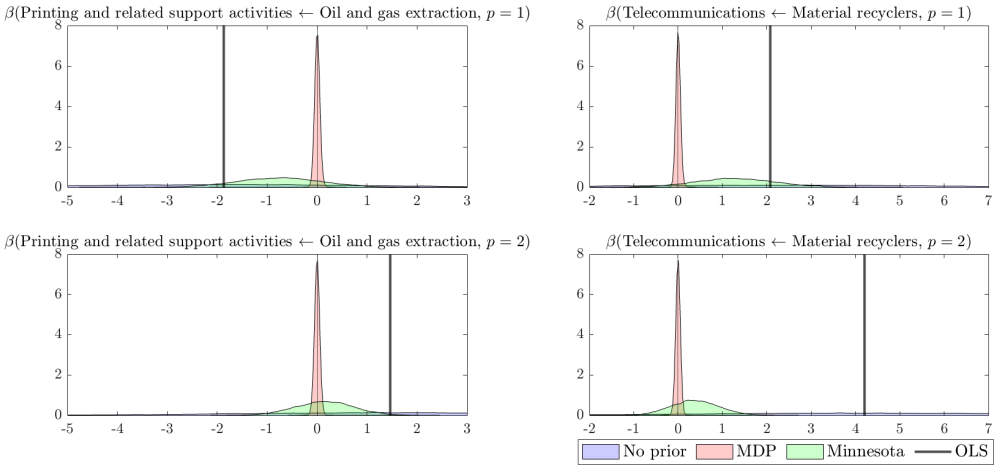
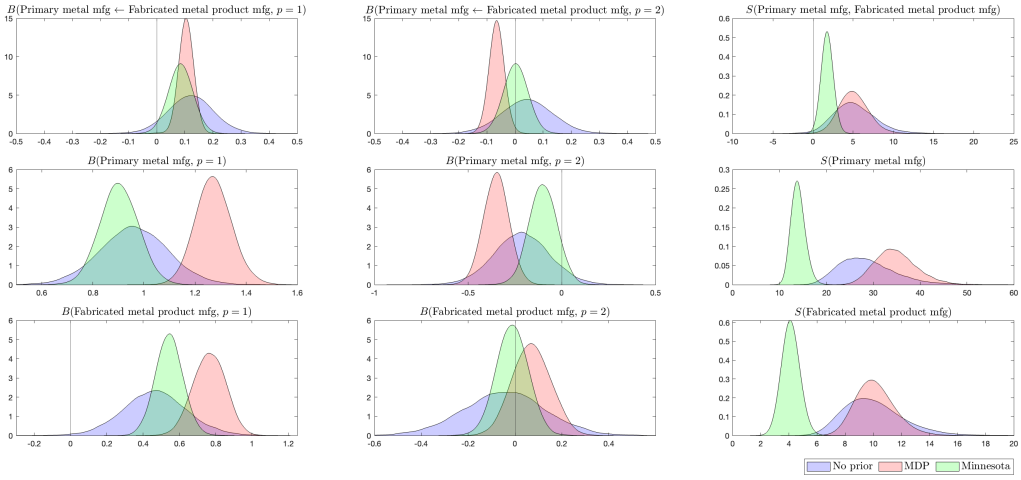


Figure B.2: Selected posteriors



B.2 Insights in clusters

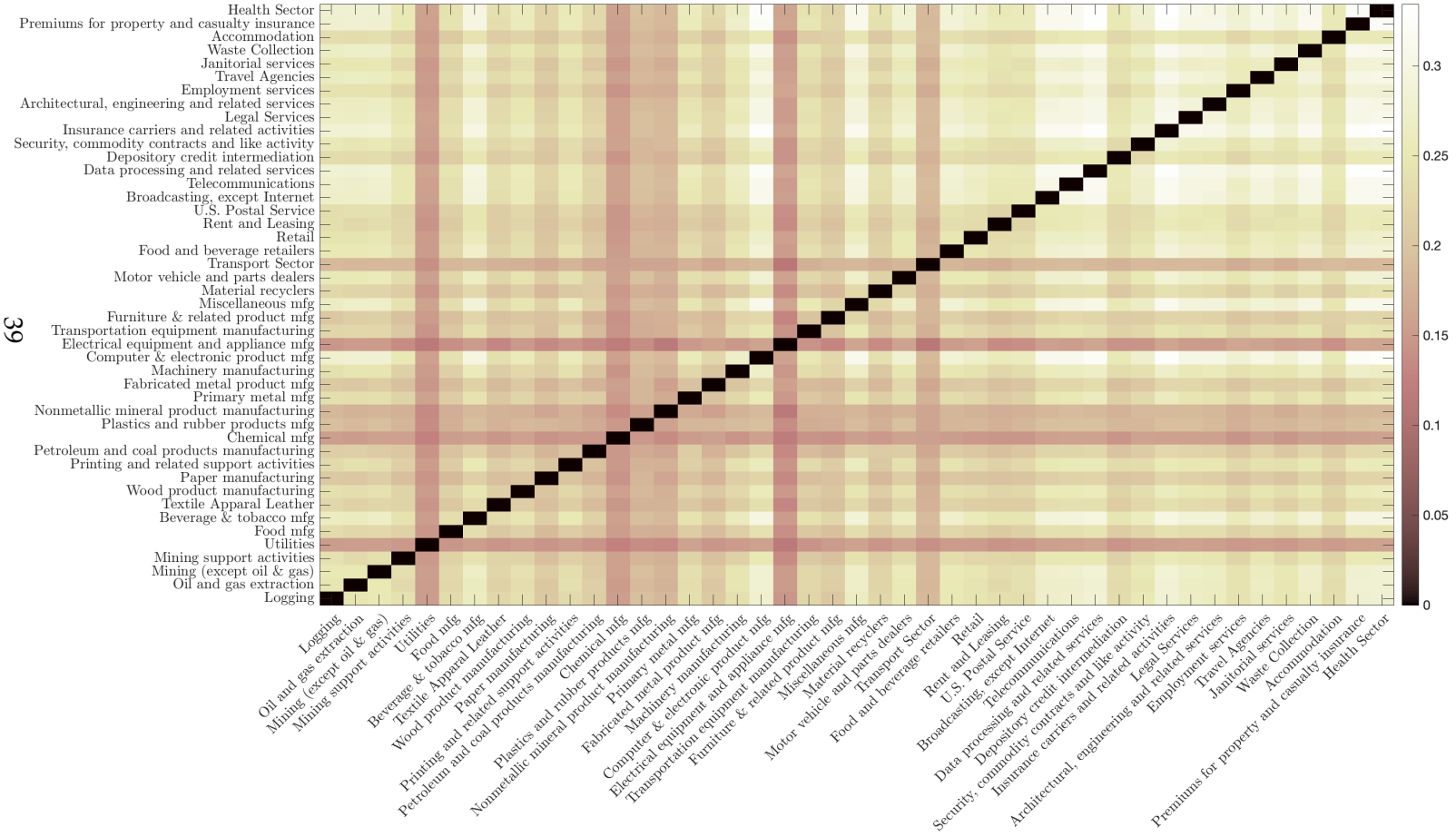
Averaged over the last 500 draws, Table B.1 summarizes the resulting eight clusters that contain more than one member. It is important to emphasize that our interest is not in the clusters *per se*, and that there is nothing specific about the cluster numbering (for legibility we have ordered them according to the number of parameters in each cluster, from most to least). The second column in the table illustrates that there are some sizable clusters. The third and fourth columns are the estimated parameters on the first and second lags. Although the first cluster is large with cluster means close to zero, there are also several sizable non-zero clusters.

Table B.1: Cluster sizes and coefficient means

	Total in cluster	Cluster mean (lag 1)	Cluster mean (lag 2)
1	957	0.0000	-0.0001
2	461	0.0005	-0.0002
3	259	0.0132	-0.0133
4	125	-0.0135	0.0140
5	73	0.0049	0.0000
6	40	0.0036	0.0041
7	40	0.0937	-0.0515
8	16	0.2139	-0.1868

To highlight exactly which parameters are clustered together, we focus our attention on industry pairs. Each (x, y)-pair in Figure B.3 illustrates the average probability that the effect of the x-axis industry on all other sectors and the effect of the y-axis industry on all other sectors are clustered together. Note that the figure is symmetric around the diagonal. This Figure illustrates, for example, that many of the service sectors are likely to be clustered together (as can be seen from the lighter color in the top right). In contrast, the dark lines in both the column and row direction indicate that, for example, the Utilities, Transport, Chemical mfg sectors are unlikely to be clustered with any other sector, suggesting they have clearly distinguishable effects on other sectors in the network.

Figure B.3: What parameters are clustered together? This figure shows the average frequency that the effect that the sector on the x-axis has on the other sectors is clustered together with the effect that the sector on the y-axis has on the other sectors. If this value is high, it means the sectors on the x and y axis have similar effects on other sectors. If the value is low, they have different effects and are not often clustered together.



C Second Application: Supplemental Figures

Figure C.1: Forecast Error Variance Decomposition at $h = 6$ quarters. States are sorted by geographical region, and color coded as such (e.g., green is the Midwest).

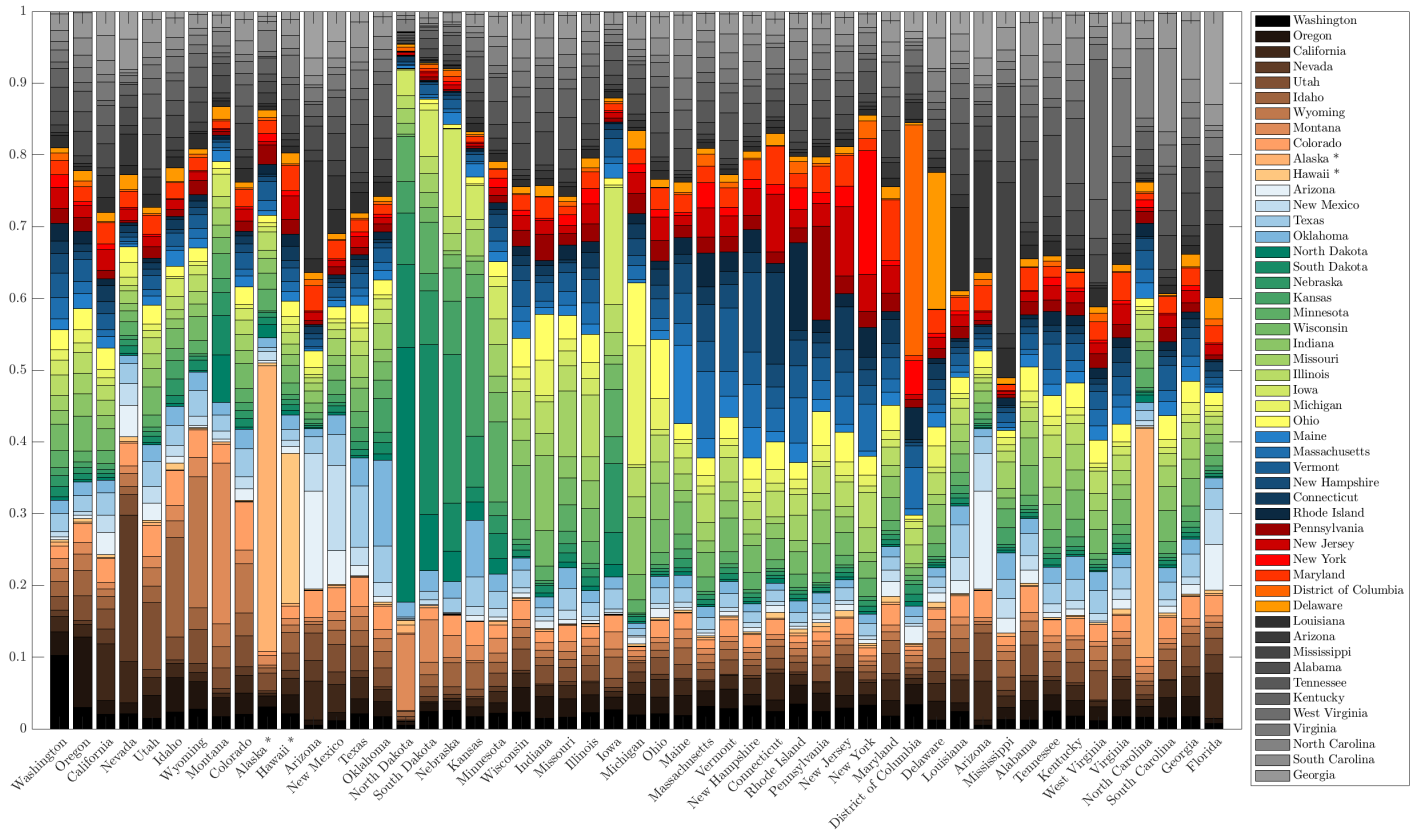


Figure C.2: Root Mean Squared Forecast Errors for $h = 1, 2, 3, 4$ step ahead forecasts for No Prior, Minnesota Prior (MP) and MDP prior for out-of-sample forecasts over the last 25 years. RMSFE divided by standard deviation of each respective series. Model estimated using data up to 100 quarters ago (no rolling window).

