Received: 2 April 2015 | Revised: 18 January 2017

DOI: 10.1002/jae.2566

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Identifying relevant and irrelevant variables in sparse factor models

Sylvia Kaufmann[1] | Christian Schumacher[2]

[1]Study Center Gerzensee, Gerzensee, Switzerland

[2]Deutsche Bundesbank, Frankfurt am Main, Germany

Correspondence
Christian Schumacher, Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main, Germany.
Email: christian.schumacher@bundesbank.de

Summary

This paper considers factor estimation from heterogeneous data, where some of the variables—the relevant ones—are informative for estimating the factors, and others—the irrelevant ones—are not. We estimate the factor model within a Bayesian framework, specifying a sparse prior distribution for the factor loadings. Based on identified posterior factor loading estimates, we provide alternative methods to identify relevant and irrelevant variables. Simulations show that both types of variables are identified quite accurately. Empirical estimates for a large multi-country GDP dataset and a disaggregated inflation dataset for the USA show that a considerable share of variables is irrelevant for factor estimation.

1 | INTRODUCTION

Factor models based on large datasets have received increasing attention in the macroeconomic literature. Policy-relevant applications of these models include, for example, the estimation of composite business cycle indicators, forecasting or structural investigations of shock transmission (Altissimo, Cristadoro, Forni, Lippi, & Veronese, 2010; Bernanke, Boivin, & Eliasz, 2005; Kose, Otrok, & Whiteman, 2003; Stock & Watson, 2002).

A rarely discussed issue in the literature is the composition of the data used to estimate factors. Economists can in principle make use of all the information contained in a large macroeconomic dataset, as, for example, principal components (PC) analysis is available to estimate the factors (Stock & Watson, 2002). Nevertheless, the question as to what extent the whole dataset is informative for factor estimation usually remains open. One of the few papers that addresses this issue is Boivin and Ng (2006). In the context of factor estimation by PC, they show that it generally depends on the information content of data whether their inclusion improves the precision of factor estimates. They illustrate the point with Monte Carlo simulations, where the data-generating process (DGP) includes variables that depend strongly on the factors and others that are noisy. These data thus contain heterogeneous information about the factors and imply a factor loading matrix in which a considerable number of loadings can be restricted to zero.

In the present paper, we propose an estimation framework that takes into account heterogeneous information in the data. We estimate a parametric factor model with Bayesian methods and choose a sparse, parsimonious specification of the factor loading matrix. In particular, the loadings are estimated under a sparse prior, which implies that many elements of the factor loading matrix may equal exactly zero. Such an approach has been pursued in the context of gene expression analysis (Carvalho, 2006; Carvalho et al., 2008; Lucas et al., 2006; West, 2003). In an economic context, sparse priors have hardly been used before. An exception is the paper by Frühwirth-Schnatter and Lopes (2010), who estimate a sparse factor model for exchange rate data. In the majority of macroeconomic applications, normal priors are employed to estimate the loadings (Bernanke et al., 2005; Kose et al., 2003; Kose, Otrok, & Whiteman, 2010; Mackowiak, Mönch, & Wiederholt, 2009). The sparse prior we use here is an extension to priors for variable selection in multiple regression (George & McCulloch, 1993, 1997;

[1]The views expressed in this paper are those of the authors and do not necessarily represent those of the Deutsche Bundesbank.

Geweke, 1996) and is more general than the normal prior. The sparse prior on a factor loading concentrates more prior mass near zero than the normal prior and therefore allows for a sharper distinction between relevant and irrelevant variables.

Based on the posterior distribution of the loadings, we can investigate the relevance or irrelevance of variables in the model. We define irrelevant variables as having only zero factor loadings, whereas relevant variables are related to at least one factor. Thus the irrelevant variables are not related to the factors and, vice versa, do not contain information for estimating the unobserved factors. In the paper, we compare three ways of identifying relevant and irrelevant variables based on the posterior distribution of the loadings. In particular, we compute univariate and multivariate highest posterior density (HPD) regions to check whether a row vector of loadings is significantly different from the zero vector. In addition, we directly compute the probability that a row vector in the loadings matrix equals the zero vector.

The results on the relevance and irrelevance of variables have important implications for empirical applications in which factors are given a structural interpretation. For example, Francis, Owyang, and Savascin (2012), following Kose et al. (2003), investigate international business cycles by estimating international factors from a large set of country-specific gross domestic product (GDP) series. If variables for a certain country are not loaded by the international factors and, vice versa, do not contain information to estimate them, one could conjecture that these countries are not linked to international business cycles. Mackowiak et al. (2009) identify common and sector-specific shocks from a factor model estimated for disaggregated US consumer price index (CPI) inflation data covering many different products and services. In this context, irrelevant variables cannot be affected by common shocks and thus are only driven by sector-specific shocks.

To assess the quantitative performance of the sparse factor model, we provide Monte Carlo simulations. We simulate DGPs with different degrees of sparsity. The results show that the common component is estimated more accurately using the sparse prior rather than the normal prior for the factor loadings. Overall, the performance of the one-layer sparse prior specification used in Frühwirth-Schnatter and Lopes (2010) is similar to the performance of the hierarchical two-layer prior in Carvalho et al. (2008). We also compare two ways of identifying relevant and irrelevant variables. Generally, we identify well relevant and irrelevant variables independently of the degree of sparsity in the DGP and across variations in the simulation design.

To illustrate the empirical properties of the sparse factor model and the methods to find relevant and irrelevant variables, we estimate a sparse factor model for, respectively, the multi-country GDP dataset used by Francis et al. (2012) and the US sectoral inflation data used in Mackowiak et al. (2009). Posterior inference on the factor loadings reveals that a considerable share of time series in both datasets have zero loading rows. In the international data, for example, the vast majority of African countries are not driven by factors. In the US inflation dataset, we find that more than half of the sectoral inflation rates are not explained by factors and thus mainly driven by sector-specific determinants.

The paper proceeds as follows. Section 2 introduces the factor model and the specification of sparsity on the loadings, defines relevant and irrelevant variables, and discusses identification issues. In Section 3, we present the estimation procedure and expose the different ways we identify relevant and irrelevant variables from the posterior output. Section 4 contains the Monte Carlo simulation results, and Section 5 the empirical results. Finally, Section 6 summarizes and concludes.

2 | SPARSE FACTOR MODEL

2.1 | Factor representation and dynamics

Assume that X_t is an $N \times 1$ vector of nontrending series observed in time period $t = 1, \dots, T$. The number of time series N will typically be large, and to condense data information we work with the factor model

$$X_t = \lambda f_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma_\varepsilon), \tag{1}$$

$$\Phi(L) f_t = \eta_t, \quad \eta_t \sim N(0, \Sigma_\eta), \tag{2}$$

where the $(k \times 1)$ factors f_t with $k \ll N$ capture the common dynamics in the series. The factor loading matrix λ has dimension $(N \times k)$. The vector lag polynomial $\Phi(L) = I - \Phi_1 L - \cdots - \Phi_p L^p$ is of order p with $L^j f_t = f_{t-j}$. The idiosyncratic components ε_t are independent of each other, $E(\varepsilon_{it}\varepsilon_{jt}) = 0$, $i \neq j$, that is, the matrix Σ_ε is diagonal with elements σ_i^2 for $i = 1, \dots, N$. Moreover, the factor innovations and the idiosyncratic components are also assumed to be independent, $E(\eta_t \varepsilon_t) = 0$. This ensures that, conditional on the true number of factors, any correlation across data is captured by the factors f_t. Of course, if the common dynamics are not sufficient to capture all (uncorrelated) dynamics in the data, one might specify additional idiosyncratic dynamics; for example, modify Equation 1 to $X_t = \lambda f_t + \xi_t$ with $\Psi(L)\xi_t = \varepsilon_t$, where $\Psi(L)$ is a diagonal vector lag polynomial of order q. In order to simplify the notation, we focus on the specification in Equation 1 in the following sections. However, in the empirical exercises we allow for idiosyncratic dynamics.

2.2 | Sparse loading matrix

A sparse factor model is defined in terms of a sparse factor loading matrix (Bhattacharya & Dunson, 2011; Carvalho et al., 2008; Frühwirth-Schnatter & Lopes, 2010; West, 2003). Assuming a sparse structure on the factor loading matrix λ implies that each variable does not have to associate with all the factors. The variable might even be unrelated to all factors. These parsimonious associations lead to zero elements in the factor loading matrix λ: Some or potentially many of the loadings in columns are zero, and some rows of the loading matrix might include only zero coefficients. However, the positions of nonzero values are unknown and have to be estimated.

Following West (2003), Lucas et al. (2006), and Carvalho et al. (2008), we specify a point mass-normal mixture prior to estimate the sparse structure of the loading matrix:

$$\pi\left(\lambda_{ij}\right) = (1-\beta_{ij})\delta_0\left(\lambda_{ij}\right) + \beta_{ij}N\left(0,\tau_j\right), \quad i=1,\ldots,N, \quad j=1,\ldots k, \tag{3}$$

where $\delta_0(\cdot)$ represents the Dirac function with point mass at zero and β_{ij} is the variable- and factor-specific prior probability of a nonzero loading. A flexible prior probability distribution $\pi(\beta_{ij})$ introduces two hierarchical layers:

$$\pi\left(\beta_{ij}\right) = (1-\rho_j)\delta_0\left(\beta_{ij}\right) + \rho_j B(ab, a(1-b)), \tag{4}$$

$$\pi\left(\rho_j\right) = B(r_0 s_0, r_0(1-s_0)). \tag{5}$$

The prior assumes that some or many loadings will have zero prior probability β_{ij}, which is reflected in the point mass-normal mixture specification (Equation 4), where $B(ab, a(1-b))$ is a Beta distribution with mean $b > 0$ and precision $a > 0$. The factor-specific probabilities ρ_j also follow a Beta distribution, $B(r_0 s_0, r_0(1-s_0))$ with mean s_0 and precision r_0. For each factor, the prior for β_{ij} assigns a variable-specific loading probability, $\beta_{ij} \in [0,1)$ including zero. It implies a common probability $1-\rho_j b$ of a zero loading on factor j across variables. Given b, which is chosen to lie between 0.75 and 0.9 in Lucas et al. (2006) and Carvalho et al. (2008), it is mainly s_0 that determines the degree of sparsity. Very small values of s_0 lead to a high probability $1-\rho_j b$ of a zero loading, and thus a lot of sparsity. Further details on the prior will be provided in Section 3.1.

The prior (Equations 3–5) encompasses a prior often used for variable selection as in George and McCulloch (1993, 1997) and Geweke (1996):

$$\pi\left(\lambda_{ij}\right) = (1-\rho_j)\delta_0\left(\lambda_{ij}\right) + \rho_j N\left(0,\tau_j\right), \tag{6}$$

$$\pi\left(\rho_j\right) = B(r_0 s_0, r_0(1-s_0)). \tag{7}$$

Compared to the two-layer prior (Equations 3–5), the variable selection prior neglects the layer to specify a variable-specific non-zero loading probability, β_{ij}. Frühwirth-Schnatter and Lopes (2010) implement a hierarchical prior, in which the significance of the factor loading is governed by an indicator $\gamma_{ij} \in \{0,1\}$ with $P\left(\gamma_{ij}=1|\rho_j\right) = \rho_j$, where ρ_j also follows a Beta distribution. The expected a priori rate of a nonzero factor loading $\rho_j b$ in the two-layer specification can also be implemented in a one-layer specification by constraining ρ_j to lie in $(0, b)$.

Macroeconomic applications using Bayesian techniques to estimate factor models typically employ a normal prior for the loadings (Bernanke et al., 2005; Kose et al., 2003, 2010; Mackowiak et al., 2009). The sparse prior (Equations 3-5) also encompasses the normal prior when $\beta_{ij}=1$.

By means of Monte Carlo simulations (see Section 4), we will evaluate the performance of the two-layer, one-layer and the normal prior specification in terms of the estimation precision that is obtained for the common component and in terms of their ability to discriminate between relevant and irrelevant variables.

2.3 | Relevant and irrelevant variables for factor analysis

We address the relevance and irrelevance of variables in the factor model by focusing on the loading matrix λ in Equation 1. In particular, we aim at identifying those variables among the N series which are unrelated to the factors. We call these variables irrelevant. In contrast, relevant variables are loaded by at least one of the factors.

For a particular variable $x_{it} \in X_t$, the factor model is defined by $x_{it} = \lambda_{i\cdot} f_t + \varepsilon_{it}$, where $\lambda_{i\cdot}$ is the ith $(1 \times k)$ row of λ. If all elements in the row vector $\lambda_{i\cdot}$ are equal to zero, namely

$$\lambda_{i\cdot} = (\lambda_{i1}\,\lambda_{i2}\,\ldots\,\lambda_{ik}) = (0\;0\;\ldots\;0), \tag{8}$$

the common factors do not have an impact on variable x_{it}, and only the idiosyncratic component matters, $x_{it} = \varepsilon_{it}$. Thus we define a variable as irrelevant if its estimated row vector of factor loadings is not significantly different from the zero vector. A relevant variable has at least one significant nonzero loading element in the corresponding row of the loading matrix.

The results on the relevance of variables have important implications for empirical applications in which factors are given a structural interpretation. For example, Francis et al. (2012) discuss the role of global and regional factors based on economic identification (zero) restrictions. Mackowiak et al. (2009) discuss aggregate versus sectoral factors, where identification is motivated by economic theory. The conclusions in both studies are based on variance decompositions and impulse response functions. The importance of identifying relevant and irrelevant variables in structural analyses like these can be highlighted as follows. Consider, for example, impulse response functions. The factor model (Equations 1–2) has a moving average (MA) representation:

$$X_t = \lambda[\Phi(L)]^{-1}\eta_t + \varepsilon_t, \tag{9}$$

from which structural $(q \times 1)$ shocks $\hat{\eta}_t$ can be identified by $\eta_t = R\hat{\eta}_t$, using the $(k \times q)$ matrix R of contemporaneous identification restrictions (Forni, Giannone, Lippi, & Reichlin, 2009). Note that $\hat{\eta}_t$ are shocks to the common factors, whereas ε_t are idiosyncratic shocks. Obviously, the impulse responses $\lambda[\Phi(L)]^{-1}R$ are convolutions of the loadings, the MA lag polynomials and the identification matrix. Thus a variable associated with a zero loading row will remain unaffected by the structural shocks $\hat{\eta}_t$, irrespective of the structural identification scheme in R. In this respect, the approach we propose extends the inferential toolbox of structural factor analysis. By identifying relevant variables we identify those that are significantly affected by structural shocks $\hat{\eta}_t$.

Note that zero rows in the loading matrix also have implications for the estimation of the unobserved factors, given that factors are typically estimated by solving a signal extraction problem (Carter & Kohn, 1994; Frühwirth-Schnatter, 1994). Koopman and Harvey (2003) express filtered and smoothed states, in our case factors, as linear functions of the observed data. The series-specific weight will equal zero if the corresponding row contains only zero loadings.

In the simulations and empirical exercises we will estimate the factor model using Bayesian techniques (see Sections 3.1 and 3.2). The identification of relevant and irrelevant variables based on the posterior output is described in detail in Section 3.3.

2.4 | Identification

The assumptions that Σ_ε is diagonal and $E(\eta_t\varepsilon_t) = 0$ are not sufficient to identify model (Equations 1-2). For any nonsingular $(k \times k)$ matrix H, there is an observationally equivalent model:

$$\begin{aligned} X_t &= \lambda H H^{-1} f_t + \varepsilon_t & \qquad H^{-1}\Phi(L) H H^{-1} f_t &= H^{-1}\eta_t \\ &= \lambda^* f_t^* + \varepsilon_t & \Phi^*(L) f_t^* &= \eta_t^*. \end{aligned} \tag{10}$$

To identify the model—up to sign identification—we have to choose a nonsingular matrix H. As it contains k^2 elements, we need k^2 identifying restrictions to pin down λ and f_t (Bai & Ng, 2013; Lawley & Maxwell, 1971). A standard identification scheme to provide order identification of the factors sets the upper diagonal elements of the $k \times k$ leading matrix in λ equal to zero, $\lambda_{ij} = 0$ for $j > i$ (Aguilar & West, 2000; Geweke & Zhou, 1996). This provides us with $k(k-1)/2$ restrictions. If we additionally impose Σ_η equal to the identity matrix, $\Sigma_\eta = I$, we additionally have $k(k+1)/2$ restrictions. Thus we end up with k^2 identifying restrictions, of which the latter ones, the unit innovation variances, also scale the factors. As an extension, Frühwirth-Schnatter and Lopes (2010) propose a generalized upper-zero triangular identification scheme, in which $l_1 < \cdots < l_k$, where l_j denotes the row of the top nonzero entry in λ, such that $\lambda_{l_j,j} \neq 0$, and $\lambda_{ij} = 0, \forall i < l_j$.

In the present paper, we work with the identification restrictions $\Sigma_\eta = I$, but we pursue an alternative approach with respect to the zero restrictions on the loading matrix. Given that we work with the assumption of a sparse loading matrix, that is, a setting where the prior puts positive mass on a zero loading coefficient, we do not set a priori $k(k-1)/2$ loadings to zero (Bhattacharya & Dunson, 2011; Knowles & Ghahramani, 2011).[2] If sparsity is present in a large dataset, and given that usually $k \ll N$, the posterior inference will usually identify more than $k(k-1)/2$ loadings to be zero. Identification is obtained by first postprocessing the Gibbs output to identify factor position and factor sign. In a second step, we remove rows and columns of zero loadings in the identified loading matrix and assess whether the sparse structure fulfills sufficient identification restrictions given in Bai and Wang (2014). Details on posterior identification are provided in Section 3.2.

Not imposing a priori zero restrictions on specific loading elements is also justified by the fact that we want to identify irrelevant variables. The rows of zero loadings corresponding to these variables remain zero rows under any rotational scheme. If we partition the variables in X_t into two $(N_1 \times 1)$ and $(N_2 \times 1)$ blocks, $X_t = [X'_{1,t} \quad X'_{2,t}]'$, where the second block would gather

[2] Alternative identification restrictions are proposed in Kaufmann and Schumacher (2013), where the sparse estimation of the factor loading matrix is combined with the identifying restriction of semi-orthogonal loadings, that is, where $\lambda'\lambda$ is a diagonal matrix.

variables with only zero factor loadings, $\lambda_2 = 0$, then the rotation with any invertible matrix H

$$\widetilde{\lambda} = \lambda H = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} H = \begin{pmatrix} \lambda_1 H \\ \lambda_2 H \end{pmatrix} \tag{11}$$

yields $\widetilde{\lambda}_2 = \lambda_2 H = 0H = 0$.

3 | BAYESIAN ESTIMATION AND POSTERIOR IDENTIFICATION

3.1 | Markov chain Monte Carlo (MCMC) estimation

To derive the posterior distribution of the model parameters $\theta = \{\lambda, \Phi, \Sigma_\varepsilon\}$ and the factors $f^T = (f_T, f_{T-1}, \ldots, f_1)$, we combine the prior with the likelihood

$$\pi\left(\theta, f^T | X^T\right) \propto L\left(X^T | f^T, \theta\right) \pi\left(f^T | \theta\right) \pi(\theta). \tag{12}$$

The complete-data likelihood can be factorized as

$$L\left(X^T | f^T, \theta\right) = \prod_{t=1}^{T} \pi\left(X_t | f^t, \theta\right), \tag{13}$$

in which the observation density is multivariate normal $\pi(X_t | f^t, \theta) = N(\lambda f_t, \Sigma_\varepsilon)$. The prior density of the unobserved factors can also be factorized:

$$\pi\left(f^T | \theta\right) = \prod_{t=p+1}^{T} \pi\left(f_t | f^{t-1}, \theta\right) \pi\left(f^p | \theta\right), \tag{14}$$

where f^p contains the initial states $f^p = \left(f_{-p+1}, \ldots, f_{-1}, f_0\right)$.

Finally, we assume independent priors for the parameters

$$\pi(\lambda, \Phi, \Sigma_\varepsilon) = \pi(\lambda)\, \pi(\Phi) \pi(\Sigma_\varepsilon). \tag{15}$$

The prior for λ is specified for each element λ_{ij} either in a hierarchical two-layer or one-layer form (see Equations 3–5 or 6–7), respectively. The hyperparameter τ_j follows an inverse Gamma distribution, $\pi\left(\tau_j\right) = \mathrm{IG}(g_0, G_0)$. In addition, we also work with a normal prior for the factor loadings, $\pi\left(\lambda_{ij}\right) = N\left(0, \tau_j\right)$. The vector autoregressive (VAR) parameters $\Phi = \mathrm{vec}\left([\Phi_1, \ldots, \Phi_p]'\right)$ in Equation 2 obtain a normal prior truncated to the stationary region, $\pi(\Phi) = N(p_0, P_0)\, I\{Z(\Phi) > 1\}$, where $I\{\cdot\}$ is the indicator function and $Z(\Phi) > 1$ means that the roots of the characteristic equation of the process $\Phi(L)$ lie outside the unit circle. Given that Σ_ε is diagonal, we assume independent inverse Gamma prior distributions for the variances, $\pi(\sigma^2_{\varepsilon,i}) = \mathrm{IG}(u_0, U_0)$, for $i = 1, \ldots, N$.

To obtain a sample from Equation 12, we iteratively go through five steps:

1. Simulate f^T from $\pi\left(f^T | X^T, \theta\right)$. We sample f^T in one sweep using the precision-based sampler proposed in Chan and Jeliazkov (2009).
2. Simulate Φ, Σ_ε from $\pi\left(\Phi, \Sigma_\varepsilon | f^T, X^T, \lambda\right)$. Given the conjugate priors, the posterior distributions are multivariate normal and inverse Wishart, respectively.
3. Simulate λ from $\pi\left(\lambda | f^T, X^T, \Sigma_\varepsilon\right)$. Under a sparse prior, the posterior of λ_{ij} will be sparse, too (Carvalho et al., 2008).
4. Update the hyperparameters depending on the chosen prior distribution for factor loadings.
5. Perform a random permutation of factor position and factor sign to explore the whole posterior distribution: For each $j = 1, \ldots, k$, we replace $\{f_{jt}\}_{t=1}^{T}$ and $\{\lambda_{ij}\}_{i=1}^{N}$ by $\{-f_{jt}\}_{t=1}^{T}$ and $\{-\lambda_{ij}\}_{i=1}^{N}$, respectively, with probability 0.5. To permute factor position, define κ to be a random permutation of $\{1, 2, \ldots, k\}$ and substitute $\{f_{\kappa_1 t}, f_{\kappa_2 t}, \ldots, f_{\kappa_k t}\}_{t=1}^{T}$ and $\{\lambda_{i\kappa_1}, \lambda_{i\kappa_2}, \ldots, \lambda_{i\kappa_k}\}_{i=1}^{N}$ for, respectively, $\{f_{1t}, f_{2t}, \ldots, f_{kt}\}_{t=1}^{T}$ and $\{\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ik}\}_{i=1}^{N}$. All factor-specific parameters like the VAR parameters have to be permuted and sign-adjusted accordingly.

The interested reader will find a detailed derivation of the posterior distributions of the factor loadings and the hyperparameters underlying Steps 3 and 4 in Appendix A. The conditional posterior distributions in Steps 1 and 2 are standard and thus not provided in detail here.[3]

[3] Detailed derivations of the posteriors can be found in the appendix of the discussion paper Kaufmann and Schumacher (2012).

3.2 | Posterior identification

Given that we randomly switch sign and order of factors and loadings in each iteration of the sampler, the posterior MCMC output is generally multimodal and not identified. In order to obtain an identified posterior sample, we apply clustering procedures similar to those proposed by Frühwirth-Schnatter (2011) for finite-mixture models. In the sparse factor framework, we use k-medoids clustering to simultaneously estimate the factor-cluster representatives and allocate each draw uniquely to the estimated clusters (Kaufman & Rousseeuw, 1987). k-medoids clustering attempts to minimize the distance between factor draws assigned to a cluster and a factor draw designated to be the center of that cluster. k-medoids chooses observed data points, in our case factor draws, as centers, which are called medoids. As the distance measure, we use the absolute correlation, which accounts for draws that are subject to sign switches. Due to the presence of sign-switching, k-means clustering as employed in Frühwirth-Schnatter (2011) cannot be used in our context. In a final step, sign identification is achieved by switching the sign of each factor and its corresponding loadings to obtain a majority of nonzero positive factor loadings. If of interest, all factor-specific parameter draws have to be reordered and sign-adjusted accordingly, too.

3.3 | Identifying relevant and irrelevant variables from the posterior output

The Gibbs sampler provides us with G^{eff} samples from the posterior distribution of the factor loadings, $\lambda^{(g)}$, and the factors, $f_t^{(g)}$, $g = 1, \dots, G^{\text{eff}}$. We use these posterior samples to assess the relevance and irrelevance of variables. In general, using posterior output for inference on a multivariate object is not a trivial task (Held, 2004; Hanson & McMillan, 2012). We propose three alternative methods:

1. We derive the 95% highest posterior density (HPD) interval (Bauwens, Lubrano, & Richard, 1999, p. 31) for each loading individually and evaluate whether zero is included or not. The variable x_{it} is identified as relevant if at least one HPD interval of its factors loadings $\lambda_{i\cdot}$ excludes zero. Otherwise, we call the variable irrelevant.
2. We also consider a multivariate method to assess jointly the relevance of a variable's loadings. We follow Hanson and McMillan (2012) and estimate a multivariate HPD region based on the Mahalanobis distance of each loading vector:

$$D_{\lambda_{i\cdot}}^{(g)} = (\lambda_{i\cdot}^{(g)} - \overline{\lambda}_{i\cdot})S_{\lambda_{i\cdot}}^{-1}(\lambda_{i\cdot}^{(g)} - \overline{\lambda}_{i\cdot})', \tag{16}$$

with $\lambda_{i\cdot}^{(g)} = (\lambda_{i1}^{(g)}\lambda_{i2}^{(g)} \dots \lambda_{ik}^{(g)})$, posterior mean $\overline{\lambda}_{i\cdot} = (1/G^{\text{eff}})\sum_{g=1}^{G^{\text{eff}}}\lambda_{i\cdot}^{(g)}$ and covariance $S_{\lambda_{i\cdot}} = (1/G^{\text{eff}})\sum_{g=1}^{G^{\text{eff}}}(\lambda_{i\cdot}^{(g)} - \overline{\lambda}_{i\cdot})'(\lambda_{i\cdot}^{(g)} - \overline{\lambda}_{i\cdot})$. For each variable, the joint 95% HPD region of the loadings is defined by the 95% of vectors with the smallest Mahalanobis distance. We denote the largest distance among all distances associated with these vectors by $D_{\lambda_{i\cdot}}^{95\%}$. A variable is called relevant if the zero vector with distance $D_{\lambda_{i\cdot}=0_{1\times k}}$ is outside the HPD region, that is, if $D_{\lambda_{i\cdot}=0_{1\times k}} > D_{\lambda_{i\cdot}}^{95\%}$. Otherwise, it is called irrelevant.
3. Finally, we directly estimate the posterior probability that a loading row is equal to zero using the sampled row vectors of loadings according to

$$P(\lambda_{i\cdot} = 0_{1\times k}|X^T) \approx \frac{1}{G^{\text{eff}}}\sum_{g=1}^{G^{\text{eff}}} I\{\lambda_{i\cdot}^{(g)} = 0_{1\times k}\}. \tag{17}$$

The variable x_{it} is identified as relevant if $P(\lambda_{i\cdot} \neq 0_{1\times k}|X^T) = 1 - P(\lambda_{i\cdot} = 0_{1\times k}|X^T) > 0.95$ and irrelevant otherwise.

We will compare the methods in simulations to check how effectively they can separate relevant and irrelevant variables.

4 | SIMULATIONS

The simulations have two goals: (i) Under different degrees of sparsity in the data, we investigate how precisely we can estimate the common component using different priors for the factor loadings. (ii) We evaluate to what extent we can identify irrelevant and relevant variables using the three methods. The DGP will vary with respect to the information content of the data for the factors, which is implemented by assuming different degrees of sparsity (see Section 4.1). We describe the specification of the estimated models and the evaluation statistics in Section 4.2. The results are discussed in Section 4.3.

4.1 | Factor DGP

We simulate multivariate data with $N = 60$ time series of length $T = 100$, driven by a $k = 2$-dimensional factor process. The dataset is simulated in two blocks: The first block of variables, for $i = 1, \dots, 40$, is simulated with different degrees of sparsity. In this block, there are relevant and irrelevant variables, but it is a priori unclear where they are located. The second block contains variables where the DGP is clear with respect to the location of relevant and irrelevant variables: For $i = 41, \dots, 50$, we simulate 10 irrelevant variables, and for $i = 51, \dots, 60$, we simulate 10 relevant variables. After estimation, we can evaluate which of these variables have been correctly identified.

1. The DGP for the first block of variables, for $i = 1, \dots, 40$, is a two-factor DGP, where the factors follow a VAR model with $p = 1$. The variables are simulated according to

$$x_{it} = \lambda_{i\cdot} f_t + \varepsilon_{it}, \ \varepsilon_{it} \sim N(0, 0.74), \tag{18}$$

$$f_t = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.8 \end{bmatrix} f_{t-1} + \eta_t, \quad \eta_t \sim N(0, I). \tag{19}$$

   To sample the loadings $\lambda_{i\cdot}$, we rely on the two-layer prior distributions in Equations 3–5. The degree of sparsity is successively decreased by increasing the hyperparameter s_0, namely $s_0 = \{0.1, 0.5, 0.9\}$ (see Equation 5). By setting $b = 0.8$ in the first layer, the marginal expected probability of a nonzero factor loading increases from 8% to 72%. As regards precision, we fix $r_0 = 30$ and $a = 30$. The nonzero factor loadings are simulated out of the normal distributions $N(m_j, M)$ with $M = 0.01$ and $m_1 = 0.60, m_2 = 0.40$. The variance $M = 0.01$ is relatively tight in order to sharply define nonzero loadings. For each loading, we additionally sample the sign to be positive or negative with equal probability.
2. The DGP for the second block of variables, $i = 41, \dots, 60$, contains 10 irrelevant and 10 relevant variables:
   (a) Irrelevant variables for $i = 41, \dots, 50$: We simulate 10 completely idiosyncratic variables with DGP $x_{it} \sim N(0, 0.74)$.
   (b) Relevant variables for $i = 51, \dots, 60$: The DGP for relevant variables is defined by Equation 18 with $\lambda_{ij} \sim N(m_j, M)$, with $M = 0.01$ and $j = \{1, 2\}$ randomly drawn. Either $m_1 = 0.60$ or $m_2 = 0.40$, and the other loading equals zero. In case m_1 is drawn, x_{it} will depend on the first factor only, and otherwise on the second only.

Overall, we obtain three specifications for the first block of variables differing in the sparsity degree s_{0j}. The overall sparsity, of course, is also affected by the mix of the 20 additional variables in the second block of the dataset. For each DGP specification, we produce $R = 65$ replications of multivariate data as well as loadings $\lambda_{i\cdot}^{(r)}$ and factors $f^{T(r)}$, $r = 1, \dots, R$. Note that the data are simulated such that identification of the factor structure is possible. The relevant variables for $i = 51, \dots, 60$ can be regarded as factor founders, as they load on one of the factors only. Furthermore, the sparse loading structure in the first block of variables implies additional zero loadings, which also determine factor identification.

4.2 | Model estimation and evaluation statistics

The factor model is estimated for $k = 2$ factors under different sparse prior distributions for the loadings. We compare the results of the sparse factor model estimated with the two-layer prior (Equations 3–5) to the factor models estimated, respectively, with the one-layer prior (Equations 6–7) and the normal prior. To estimate the factor model under the sparse priors, we assume that the true degree of sparsity in the DGP is unknown and choose an intermediate degree of sparsity $s_0 = 0.5$. We additionally fix $r_0 = 3$ and $a = 3$, implying low precision a priori when estimating the model compared to the DGP. We set $b = 0.8$. Furthermore, τ_j is parametrized by an inverse Gamma, $\tau_j \sim \text{IG}(2, 0.5)$. Concerning the variances of the idiosyncratic components, we define $\sigma^2_{\varepsilon,i} \sim \text{IG}(2.0, 1.0)$, $i = 1, \dots, N$. To specify the hyperparameters of the factor VAR parameters, we implement a Minnesota prior with variance equal to $\vartheta_0^2 = 0.09$ on the first autoregressive lag and a shrinkage factor of $\vartheta_1^2 = 0.03$ for off-diagonal parameters (Litterman, 1986).[4] According to the discussion in Section 2.4, we will estimate the model without any additional identifying zero restrictions on the factor loading matrix. To estimate the model for each replication, we draw 6,000 times

[4]This induces the following prior moments for the elements in the VAR parameters $\boldsymbol{\Phi}_l$, $l = 1, \dots, p$:

$$E(\Phi_{lij}) = 0 \quad \text{for all } i, j = 1, \dots, k, \qquad var(\Phi_{lij}) = \begin{cases} \frac{\vartheta_0^2}{l^2} & \text{for } i = j \\ \frac{\vartheta_1^2 \vartheta_0^2}{l^2} & \text{otherwise.} \end{cases} \tag{20}$$

from the posterior, discard the first 2,000 draws, keep every second one, and evaluate the posterior with $G^{\text{eff}} = 2,000$ retained iterations. Posterior identification is then achieved by using k-medoids clustering as outlined in Section 3.2.

To assess the estimation precision of the common components under each setting, we follow Boivin and Ng (2006) and compute the root mean squared error $\text{RMSE}^{(r)}$ by

$$\text{RMSE}^{(r)} = \frac{1}{N}\frac{1}{G^{\text{eff}}}\sum_{i=1}^{N}\sum_{g=1}^{G^{\text{eff}}}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\lambda_{i\cdot}^{(r,g)}f_t^{(r,g)} - \lambda_{i\cdot}^{(r)}f_t^{(r)}\right)^2}, \tag{21}$$

for all R replications with simulated true values $\lambda_{i\cdot}^{(r)}$, $f_t^{(r)}$ and G^{eff} sampled values $\lambda_{i\cdot}^{(r,g)}$ and $f_t^{(r,g)}$. Finally, we average over replications to obtain the mean $\text{RMSE} = \frac{1}{R}\sum_{r=1}^{R}\text{RMSE}^{(r)}$.

To evaluate the performance of the alternative methods in identifying the relevant and irrelevant variables discussed in Section 3.3, we proceed as follows. For each replication, we first assess the relevance of each of the variables $i = 41, \ldots, 60$ based on the univariate and multivariate HPD regions and the zero-row probability. As a summary statistic, we then compute the ratio of the correctly identified number of irrelevant and relevant variables, respectively, to the true number in the DGP (10 each).

4.3 | Simulation results

Table 1 contains the results for the RMSE of the common component. In particular, three statistics are provided: Panel A contains the mean RMSE; Panel B shows the relative RMSE of the model estimated under, respectively, the normal and one-layer prior, to the RMSE of the model estimated with the two-layer sparse prior. Thus, a relative RMSE greater than one indicates a superior performance of the two-layer prior. Panel C contains the average differences of the $\text{RMSE}^{(r)}$ of, respectively, the one-layer and the normal prior to the $\text{RMSE}^{(r)}$ of the two-layer prior. If the average is smaller than zero, the two-layer prior is superior. We also provide the simulated 95% interval of differences to assess the significance.

The results can be summarized as follows. Panel A of Table 1 shows that the differences between the RMSEs of the two-layer and one-layer prior are quite small, whereas both clearly improve over the normal prior; see also the corresponding relative

TABLE 1 Simulation results on the accuracy of estimated common components under alternative priors for the factor loadings

	s_0		
	0.1	**0.5**	**0.9**
A. RMSE, mean over replications of common components			
Normal prior	0.276	0.394	0.486
One-layer prior	0.179	0.357	0.479
Two-layer prior	0.179	0.357	0.478
B. RMSE, normalized to one for two-layer prior			
Normal prior	1.538	1.104	1.017
One-layer prior	1.001	1.001	1.002
Two-layer prior	1.000	1.000	1.000
C. Difference of RMSEs between priors (mean, 95% interval)			
Normal minus two-layer	0.097	0.037	0.008
	(0.084, 0.115)	(0.022, 0.055)	(0.003, 0.014)
One-layer minus two-layer	0.000	0.000	0.001
	(−0.001, 0.001)	(−0.000, 0.002)	(−0.003, 0.005)

Note. Panel A contains the root mean squared error (RMSE) based on the differences between estimated and simulated common components averaged over variables, replications and Gibbs draws according to Equation 21. Results are presented for the one-layer prior of the loadings, the two-layer prior, the normal prior, and different sparsity parameter values s_0 in the DGP. Panel B shows the relative RMSE of the model estimated under, respectively, the normal and one-layer prior, to the RMSE of the model estimated under the two-layer prior. Panel C contains the averaged differences of the RMSEs between the one-layer and two-layer prior as well as the differences of the RMSEs between the normal prior and the two-layer prior. The simulated 95% intervals of differences are presented in parentheses.

TABLE 2 Simulation results on the identification of relevant variables under different priors and methods to assess relevance

	s_0 0.1	0.5	0.9
A. Univariate loadings HPD			
Ratio of correctly identified variables			
Normal prior	0.97	0.99	0.99
One-layer prior	0.93	0.97	0.97
Two-layer prior	0.92	0.97	0.97
Differences between priors (mean, 95% interval)			
Normal minus two-layer	0.05	0.02	0.02
	(0.00, 0.20)	(0.00, 0.10)	(0.00, 0.10)
One-layer minus two-layer	0.00	0.00	0.01
	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.10)
B. Multivariate loadings HPD			
Ratio of correctly identified variables			
Normal prior	0.96	0.99	0.98
One-layer prior	0.95	0.97	0.97
Two-layer prior	0.95	0.97	0.97
Differences between priors (mean, 95% interval)			
Normal minus two-layer	0.01	0.02	0.01
	(−0.10, 0.10)	(0.00, 0.10)	(0.00, 0.10)
One-layer minus two-layer	−0.00	0.00	0.00
	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
C. Zero-row probability			
Ratio of correctly identified variables			
Normal prior	1.00	1.00	1.00
One-layer prior	0.93	0.98	0.99
Two-layer prior	0.93	0.98	0.98
Differences between priors (mean, 95% interval)			
Normal minus two-layer	0.07	0.02	0.02
	(0.00, 0.20)	(0.00, 0.10)	(0.00, 0.10)
One-layer minus two-layer	0.00	0.00	0.00
	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)

Note. Panel A contains results on the performance of the univariate loadings HPD intervals as outlined in Section 3.3 with respect to identifying relevant variables, Panel B contains results for the multivariate loadings HPD regions defined in Equation 16, and Panel C results for zero-row probabilities from Equation 17. For each method, the ratio of correctly identified relevant variables is shown, computed for different priors on the loadings and different sparsity parameters s_0 in the DGP. For each method, also the averaged differences of the ratios of correctly identified variables between, respectively, the one-layer and the normal prior to the two-layer prior, are shown. The simulated 95% intervals of differences are presented in parentheses.

RMSEs in Panel B. However, the efficiency gain decreases as the degree of sparsity decreases in the DGP. The 95% intervals in Panel C indicate no systematic differences between the two-layer and one-layer prior, whereas the two-layer prior yields a significantly lower RMSE than the normal prior for all s_0.

Tables 2 and 3 contain the results concerning the identification of relevant and irrelevant variables, respectively. The panels in each table contain the results of the three different methods applied to discriminate between relevant and irrelevant variables: Univariate HPD intervals are applied in Panel A, multivariate HPD regions in Panel B, and the zero-row probability in Panel C. In each panel, we show two groups of statistics: The upper part of each panel contains the ratio of correctly identified relevant (Table 2) or irrelevant variables (Table 3), respectively, to the true number of relevant and irrelevant variables in the DGP. The ratio is averaged over replications. A value of one indicates that all 10 variables have been correctly identified in all $R = 65$ datasets. The lower part in the panels contains the average difference of ratios between the one-layer and the two-layer prior, as well as the averaged difference of ratios between the normal prior and the two-layer prior. If the mean is smaller than zero, the

TABLE 3 Simulation results on the identification of irrelevant variables under different priors and methods to assess relevance

	s_0 0.1	0.5	0.9
A. Univariate loadings HPD			
Ratio of correctly identified variables			
Normal prior	0.93	0.91	0.92
One-layer prior	1.00	0.99	0.99
Two-layer prior	1.00	0.99	0.99
Differences between priors (mean, 95% interval)			
Normal minus two-layer	−0.07	−0.08	−0.08
	(−0.20, 0.00)	(−0.30, 0.00)	(−0.20, 0.00)
One-layer minus two-layer	0.00	0.00	0.00
	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
B. Multivariate loadings HPD			
Ratio of correctly identified variables			
Normal prior	0.97	0.96	0.97
One-layer prior	1.00	0.99	0.99
Two-layer prior	1.00	0.99	0.99
Differences between priors (mean, 95% interval)			
Normal minus two-layer	−0.03	−0.04	−0.03
	(−0.10, 0.00)	(−0.20, 0.00)	(−0.10, 0.00)
One-layer minus two-layer	0.00	0.00	−0.00
	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
C. Zero-row probability			
Ratio of correctly identified variables			
Normal prior	0.00	0.00	0.00
One-layer prior	1.00	0.99	0.95
Two-layer prior	1.00	0.99	0.98
Differences between priors (mean, 95% interval)			
Normal minus two-layer	−1.00	−0.99	−0.98
	(−1.00, −1.00)	(−1.00, −0.90)	(−1.00, −0.90)
One-layer minus two-layer	0.00	0.00	−0.02
	(0.00, 0.00)	(0.00, 0.00)	(−0.10, 0.00)

Note. This table provides results on how well irrelevant variables can be identified by the different methods. The methods used to identify relevant variables are the same as in Table 2, and further details on how to interpret the entries can be found in the note to that table.

two-layer prior performs better. We also provide the 95% interval of the differences. If the interval includes zero, the differences are not significant.

The results for the relevant variables in Table 2 can be summarized as follows. From the ratio of correctly identified variables we can see that at least 92% of the relevant variables can be identified correctly. The ratio is slightly smaller for very sparse data with $s_0 = 0.1$ in the DGP compared to the cases $s_0 = 0.5, 0.9$. The two-level and the one-level priors perform equally well, whereas the normal prior performs slightly better in identifying the relevant variables. However, if we look at the lower parts of Panels A, B, and C, we see that the differences between the priors are not significant. Zero is included in the 95% intervals in all cases. Note that the distribution of differences is very skewed with a lot of mass around zero. This leads to 95% intervals with one boundary equal to zero, and a median of zero, too, in the majority of cases (not reported in the table). Thus no systematic difference is present between the priors' ability to identify relevant variables. The differences between the three different methods in Panels A, B, and C to identify relevant variables are also small. Note that under the normal prior the zero-row probability always identifies all relevant variables. What appears as a strength of the normal prior will be a large disadvantage when it comes to identifying irrelevant variables.

TABLE 4 Empirical results on the relevance of variables in international GDP growth data: share of relevant variables identified by univariate loadings HPD intervals

k	1	2	3	4
A. One-layer prior				
All countries	0.47	0.51	0.51	0.51
Africa	0.00 (0 of 6)	0.00 (0 of 6)	0.00 (0 of 6)	0.00 (0 of 6)
Asia 1	0.00 (0 of 6)	0.17 (1 of 6)	0.17 (1 of 6)	0.17 (1 of 6)
Asia 2	0.67 (4 of 6)	0.83 (5 of 6)	0.83 (5 of 6)	0.67 (4 of 6)
Europe	1.00 (18 of 18)	1.00 (18 of 18)	1.00 (18 of 18)	1.00 (18 of 18)
Latin America	0.13 (2 of 16)	0.13 (2 of 16)	0.06 (1 of 16)	0.13 (2 of 16)
North America	1.00 (3 of 3)	1.00 (3 of 3)	1.00 (3 of 3)	1.00 (3 of 3)
Oceania	0.00 (0 of 2)	0.00 (0 of 2)	0.50 (1 of 2)	0.50 (1 of 2)
B. Two-layer prior				
All countries	0.42	0.44	0.49	0.46
Africa	0.00 (0 of 6)	0.00 (0 of 6)	0.00 (0 of 6)	0.00 (0 of 6)
Asia 1	0.00 (0 of 6)	0.00 (0 of 6)	0.17 (1 of 6)	0.17 (1 of 6)
Asia 2	0.33 (2 of 6)	0.33 (2 of 6)	0.67 (4 of 6)	0.67 (4 of 6)
Europe	1.00 (18 of 18)	1.00 (18 of 18)	1.00 (18 of 18)	0.89 (16 of 18)
Latin America	0.06 (1 of 16)	0.06 (1 of 16)	0.06 (1 of 16)	0.13 (2 of 16)
North America	1.00 (3 of 3)	1.00 (3 of 3)	1.00 (3 of 3)	0.67 (2 of 3)
Oceania	0.00 (0 of 2)	0.50 (1 of 2)	0.50 (1 of 2)	0.50 (1 of 2)

Note. The table contains the estimated shares of relevant variables under different priors—the one-layer prior in Panel A and the one-layer prior in Panel B—and different number of factors k, where relevant variables are identified by using univariate loadings HPD intervals as outlined in Section 3.3. In each panel, the second row contains the share of relevant variables in the whole set of variables in the data. The remaining rows contain the shares of relevant variables in different world regions. Absolute numbers of relevant variables and the number of all variables are shown in parentheses.

In Table 3 we provide the results concerning irrelevant variables. The ratio of correctly identified irrelevant variables is always greater than or equal to 93%. The exception is the normal prior together with the zero-row probability, as shown in Panel C. Due to the loss in estimation efficiency under the normal prior, the estimated probability of a row of zero loadings is too low. The corresponding ratios in the table are all equal to zero. Concerning the differences between the sparse priors and the normal prior in Panels A and B, we observe that the sparsity priors perform similarly, whereas the normal prior performs worse in identifying the irrelevant variables. However, the differences are not significant.

We have carried out a number of alternative simulations to check the robustness of the results. In particular, modifying the prior specifications for estimating the factor models, working with an alternative identification scheme by Geweke and Zhou (1996), and using different degrees in persistence of simulated irrelevant variables have been considered. It turns out that the results remain very similar to those just reported.[5] Further checks for robustness include DGPs, where the factor model has a reduced dynamic rank, as discussed in Forni, Hallin, Lippi, & Zaffaroni (2015). Finally, we have also investigated how to determine the number of factors. Results from simulation experiments using different ways to specify the number of factors k are provided in Appendix B. The results show that we can properly choose the right number of factors in the simulations.

To summarize, we find that working with sparse priors generally improves the estimation efficiency of the common component when compared to working with the normal prior. The two-layer prior cannot systematically improve over the one-layer prior; the differences in RMSEs are insignificant. In the vast majority of cases, we can identify more than 90% of the relevant and irrelevant variables regardless of which sparsity prior or which method to discriminate between relevant and irrelevant variables is used. The exception is the normal prior, under which the probability of zero rows cannot identify irrelevant variables.

[5]These results are reported in the working paper version (Kaufmann & Schumacher, 2012).

5 | EMPIRICAL RESULTS

The methods to identify relevant variables in sparse factor models can be regarded as quite general tools of model assessment. To illustrate their scope for applications, we provide two empirical examples:

1. We analyze a multi-country GDP growth dataset as in Francis et al. (2012), who investigate the role of global business cycles in a large factor model.
2. We investigate disaggregated US CPI inflation data taken from Mackowiak et al. (2009) in a sparse factor model setting. Mackowiak et al. (2009) use a factor model to compare alternative theoretical models of price-setting behavior.

The two datasets differ quite a lot in terms of data coverage, and we can expect to find different degrees of sparsity. We estimate a sparse factor model for each of these datasets and identify relevant and irrelevant variables. We provide results for different model specifications concerning the number of factors, the priors, and the methods to identify relevant and irrelevant variables.

5.1 | International GDP growth data

The application follows the literature on international business cycles estimated in large factor models (Kose et al., 2003, 2010; Del Negro & Otrok, 2008). Recently, Francis et al. (2012) estimated global and regional business cycles from a large set of country-specific annual GDP growth series. We follow these authors and choose the Penn World Tables (PWT) to construct the dataset. We use version 7.0 (Heston, Summers, & Aten, 2011) and take all GDP series that are available from 1960 onwards to match the sample period in Francis et al. (2012). We end up with $N = 57$ countries, and compute growth rates by taking first differences of the logarithm of the level series.

To specify the sparse factor model, we use $p = 2$ lags in the factor VAR and $q = 1$ AR lag in the idiosyncratic components. We estimate the model with the one- and the two-layer prior. We specify a prior degree of sparsity of $s_0 = 0.5$ with precision $r_0 = 3$. The other hyperparameters are specified as before in the simulation exercise. We sample 60,000 posterior draws, discard the first 20,000 replications, and retain every fourth one. To determine the number of factors, we apply different methods. Details can be found in Appendix B.3. The results regarding the number of factors differ with respect to the methods chosen. The values for k span a range of between one and four. To account for this uncertainty and check how it affects the identification of relevant and irrelevant variables, we provide results for one to four factors, $k = 1, \dots, 4$. We apply three different methods to separate

TABLE 5 Empirical results on the relevance of variables in international GDP growth data: share of relevant variables identified by multivariate loadings HPD regions

k	1	2	3	4
A. One-layer prior				
All countries	0.53	0.53	0.42	0.39
Africa	0.00 (0 of 6)	0.00 (0 of 6)	0.00 (0 of 6)	0.00 (0 of 6)
Asia 1	0.17 (1 of 6)	0.33 (2 of 6)	0.17 (1 of 6)	0.17 (1 of 6)
Asia 2	0.83 (5 of 6)	1.00 (6 of 6)	0.67 (4 of 6)	0.50 (3 of 6)
Europe	1.00 (18 of 18)	1.00 (18 of 18)	0.89 (16 of 18)	0.83 (15 of 18)
Latin America	0.19 (3 of 16)	0.06 (1 of 16)	0.00 (0 of 16)	0.06 (1 of 16)
North America	1.00 (3 of 3)	1.00 (3 of 3)	0.67 (2 of 3)	0.67 (2 of 3)
Oceania	0.00 (0 of 2)	0.00 (0 of 2)	0.50 (1 of 2)	0.00 (0 of 2)
B. Two-layer prior				
All countries	0.53	0.39	0.40	0.35
Africa	0.00 (0 of 6)	0.00 (0 of 6)	0.00 (0 of 6)	0.00 (0 of 6)
Asia 1	0.17 (1 of 6)	0.00 (0 of 6)	0.17 (1 of 6)	0.00 (0 of 6)
Asia 2	0.83 (5 of 6)	0.33 (2 of 6)	0.50 (3 of 6)	0.50 (3 of 6)
Europe	1.00 (18 of 18)	0.89 (16 of 18)	0.89 (16 of 18)	0.83 (15 of 18)
Latin America	0.19 (3 of 16)	0.06 (1 of 16)	0.00 (0 of 16)	0.00 (0 of 16)
North America	1.00 (3 of 3)	0.67 (2 of 3)	0.67 (2 of 3)	0.67 (2 of 3)
Oceania	0.00 (0 of 2)	0.50 (1 of 2)	0.50 (1 of 2)	0.00 (0 of 2)

Note. This table provides empirical shares of relevant variables, where the relevance of variables is assessed by multivariate loadings HPD regions defined in Equation 16. Table entries can otherwise be interpreted in the same way as in Table 4, and further details can be found in the note to that table.

relevant and irrelevant variables: Results based on univariate HPD intervals are provided in Table 4, results using multivariate HPD regions in Table 5, and results based on the probability of a zero row in Table 6.

In each table, we present the share of identified relevant variables, and the region-specific shares of identified relevant variables. The definition of geographic regions follows Kose et al. (2003) and Francis et al. (2012), who divide the world into six regions: Africa, Asia 1 (less developed), Asia 2 (more developed), Europe, Latin America, North America and Oceania. The detailed country classification can be found in Appendix C.1.

In Table 4 we provide results based on univariate HPD intervals and the one-layer prior in Panel A. For the one-factor model in the first column we find that 47% (27 out of 57) of the countries are identified as relevant. In other words, more than half of the countries in the dataset are not related to the factor, indicating that sparsity seems to matter a lot. If we consider geographic regions, the proportions of relevant variables are as follows: Africa 0% (0 of 6 countries), Asia 1 0% (0 of 6), Asia 2 67% (4 of 6), Europe 100% (18 of 18), Latin America 13% (2 of 16), North America 100% (3 of 3), Oceania 0% (0 of 2). Thus African countries and almost all countries in Latin America are not explained by the single estimated factor, whereas the majority of countries in North America, Europe and Asia 2 are linked to it. The results for more factors ($k = 2, 3, 4$) are similar. In Panel B, the results obtained under the two-layer prior show a slightly lower number of identified relevant variables. The share of relevant variables lies between 42% and 49%, whereas the share obtained under the one-layer prior ranges between 47% and 51%. The results regarding the relevance of variables in the different geographic regions are very similar for the one- and two-layer prior.

The results based on the multivariate HPD regions in Table 5 are overall very similar compared to Table 4. The share of relevant variables is a bit more dispersed: For the one-layer prior, the ratio ranges between 39% and 53%, and for the two-layer prior between 35% and 53%. When the number of factor is increased, the share of relevant variables tends to decrease, but not substantially. The relative results regarding the relevance of variables in the different geographic regions are very similar and comparable to the results in Table 4 based on univariate HPD intervals.

Results based on the probability of a zero row in Table 6 indicate a larger number of relevant variables, in particular if the one-layer prior is chosen (Panel A). The overall share of relevant variables varies between 49% and 77%. Concerning the different geographic regions, however, the results show that still fewer countries in Africa and Latin America are relevant compared to Europe or North America. In this regard, the findings are in line with the other evidence in Tables 4 and 5.

Overall, if we summarize the evidence from the alternative specifications, we find some uncertainty regarding the relevance of variables in the international GDP growth data. However, we can also derive robust findings: From the 24 sets of results in

TABLE 6 Empirical results on the relevance of variables in international GDP growth data: share of relevant variables identified by zero-row probabilities

k	1	2	3	4
A. One-layer prior				
All countries	0.58	0.60	0.70	0.77
Africa	0.17 (1 of 6)	0.17 (1 of 6)	0.17 (1 of 6)	0.17 (1 of 6)
Asia 1	0.17 (1 of 6)	0.33 (2 of 6)	0.33 (2 of 6)	0.33 (2 of 6)
Asia 2	0.83 (5 of 6)	1.00 (6 of 6)	1.00 (6 of 6)	1.00 (6 of 6)
Europe	1.00 (18 of 18)	1.00 (18 of 18)	1.00 (18 of 18)	1.00 (18 of 18)
Latin America	0.31 (5 of 16)	0.25 (4 of 16)	0.50 (8 of 16)	0.75 (12 of 16)
North America	1.00 (3 of 3)	1.00 (3 of 3)	1.00 (3 of 3)	1.00 (3 of 3)
Oceania	0.00 (0 of 2)	0.00 (0 of 2)	1.00 (2 of 2)	1.00 (2 of 2)
B. Two-layer prior				
All countries	0.49	0.51	0.58	0.67
Africa	0.17 (1 of 6)	0.17 (1 of 6)	0.17 (1 of 6)	0.17 (1 of 6)
Asia 1	0.00 (0 of 6)	0.00 (0 of 6)	0.33 (2 of 6)	0.33 (2 of 6)
Asia 2	0.67 (4 of 6)	0.67 (4 of 6)	1.00 (6 of 6)	1.00 (6 of 6)
Europe	1.00 (18 of 18)	1.00 (18 of 18)	1.00 (18 of 18)	1.00 (18 of 18)
Latin America	0.19 (3 of 16)	0.19 (3 of 16)	0.19 (3 of 16)	0.50 (8 of 16)
North America	1.00 (3 of 3)	1.00 (3 of 3)	1.00 (3 of 3)	1.00 (3 of 3)
Oceania	0.00 (0 of 2)	0.50 (1 of 2)	0.50 (1 of 2)	0.50 (1 of 2)

Note. This table provides empirical shares of relevant variables, where the relevance of variables is assessed by zero-row probabilities defined in Equation 17. Table entries can otherwise be interpreted in the same way as in Table 4, and further details can be found in the note to that table.

TABLE 7 Empirical results for disaggregated US CPI inflation: share of relevant variables identified by alternative methods

k Prior	1 One-layer	2 One-layer	1 Two-layer	2 Two-layer
A. Univariate loadings HPD				
All CPI components	0.37	0.41	0.37	0.37
Apparel	0.36 (4 of 11)	0.36 (4 of 11)	0.36 (4 of 11)	0.36 (4 of 11)
Education and communication	0.25 (1 of 4)	0.25 (1 of 4)	0.25 (1 of 4)	0.25 (1 of 4)
Food and beverages	0.24 (7 of 29)	0.34 (10 of 29)	0.24 (7 of 29)	0.24 (7 of 29)
Housing	0.36 (4 of 11)	0.36 (4 of 11)	0.36 (4 of 11)	0.45 (5 of 11)
Medical care	1.00 (5 of 5)	1.00 (5 of 5)	1.00 (5 of 5)	1.00 (5 of 5)
Other goods and services	0.00 (0 of 2)	0.00 (0 of 2)	0.00 (0 of 2)	0.00 (0 of 2)
Recreation	0.71 (5 of 7)	0.71 (5 of 7)	0.71 (5 of 7)	0.71 (5 of 7)
Transportation	0.30 (3 of 10)	0.30 (3 of 10)	0.30 (3 of 10)	0.20 (2 of 10)
B. Multivariate loadings HPD				
All CPI components	0.44	0.42	0.44	0.37
Apparel	0.36 (4 of 11)	0.36 (4 of 11)	0.36 (4 of 11)	0.36 (4 of 11)
Education and communication	0.75 (3 of 4)	0.25 (1 of 4)	0.75 (3 of 4)	0.25 (1 of 4)
Food and beverages	0.28 (8 of 29)	0.34 (10 of 29)	0.28 (8 of 29)	0.21 (6 of 29)
Housing	0.45 (5 of 11)	0.36 (4 of 11)	0.45 (5 of 11)	0.55 (6 of 11)
Medical care	1.00 (5 of 5)	1.00 (5 of 5)	1.00 (5 of 5)	1.00 (5 of 5)
Other goods and services	0.50 (1 of 2)	0.00 (0 of 2)	0.50 (1 of 2)	0.00 (0 of 2)
Recreation	0.71 (5 of 7)	0.71 (5 of 7)	0.71 (5 of 7)	0.71 (5 of 7)
Transportation	0.40 (4 of 10)	0.40 (4 of 10)	0.40 (4 of 10)	0.20 (2 of 10)
C. Zero-row probability				
All CPI components	0.39	0.42	0.38	0.39
Apparel	0.36 (4 of 11)	0.36 (4 of 11)	0.36 (4 of 11)	0.36 (4 of 11)
Education and communication	0.25 (1 of 4)	0.25 (1 of 4)	0.25 (1 of 4)	0.25 (1 of 4)
Food and beverages	0.24 (7 of 29)	0.34 (10 of 29)	0.24 (7 of 29)	0.24 (7 of 29)
Housing	0.45 (5 of 11)	0.36 (4 of 11)	0.36 (4 of 11)	0.45 (5 of 11)
Medical care	1.00 (5 of 5)	1.00 (5 of 5)	1.00 (5 of 5)	1.00 (5 of 5)
Other goods and services	0.00 (0 of 2)	0.00 (0 of 2)	0.00 (0 of 2)	0.00 (0 of 2)
Recreation	0.71 (5 of 7)	0.71 (5 of 7)	0.71 (5 of 7)	0.71 (5 of 7)
Transportation	0.40 (4 of 10)	0.40 (4 of 10)	0.40 (4 of 10)	0.40 (4 of 10)

Note. The table contains the estimated share of relevant variables under different priors and different number of factors *k*. In Panel A, relevant variables are identified by univariate loadings HPD intervals as outlined in Section 3.3. In Panel B, multivariate loadings HPD regions as defined in Equation 16 are used, whereas in Panel C, zero-row probabilities as defined in Equation 17 identify relevant variables. In each panel, the second row contains the share of relevant variables in the whole set of variables in the data. The remaining rows contain the shares of relevant variables in different groups of CPI components. Absolute numbers of relevant variables and the number of all variables are shown in parentheses.

Tables 4, 5, and 6, there are 20 cases where the share of relevant variables is not greater than 60%. Thus at least 40% of variables are not related to the factors in most of the cases, which indicates a relatively high degree of sparsity in the international data. Furthermore, if we compare the share of relevant variables in the different geographic regions, we find that in all specifications at most one country in Africa is identified as relevant, whereas more than 80% of European countries and at least two out of three North American countries are usually identified as relevant.[6]

In relation to Kose et al. (2003, 2010), Del Negro and Otrok (2008), and Francis et al. (2012), we interpret our results in the following way. If variables for a certain country, region, or continent are identified as irrelevant, they are unrelated to the common factors. If these factors represent global or regional common business cycles, this means that countries identified as irrelevant are not linked to common business cycles. In the data used here, this is often the case for nearly all African and most Asia 1 countries.

[6]To check the robustness of the results with respect to Africa further, following a referee's comment we have also considered an extended dataset including more African countries than in Kose et al. (2003). The results remain largely the same also in this case.

5.2 | US sectoral inflation data

Mackowiak et al. (2009) estimate a factor model for disaggregated US CPI inflation data to identify sector-specific and aggregate shocks. By means of impulse response functions for the subindices of the US CPI, the authors discriminate between different theoretical models of price-setting behavior. Shock identification is obtained by assuming that sector-specific shocks are entirely idiosyncratic, whereas aggregate shocks are common. The disaggregated US inflation data of Mackowiak et al. (2009) contain $N = 79$ subindices of the CPI. A detailed sectoral classification can be found in Appendix C.2.

We apply different criteria to determine the number of factors in the model, and details can be found in Appendix B.3. The selection criteria find either one or two factors. Note that Mackowiak et al. (2009) prefer a one-factor model. To consider the uncertainty concerning the specifications, we provide results for both $k = 1, 2$. Furthermore, we specify $p = 4$ lags in the factor VAR and $q = 2$ lags in the idiosyncratic components. The sparse prior is specified with $s_0 = 0.5$ and precision $r_0 = 3$, and the rest of the parameters as in the simulations. The results are reported in Table 7. Results based on univariate HPD intervals are provided in Panel A, results using multivariate HPD regions in Panel B, and results based on the probability of a zero row in Panel C.

In the 12 specifications, we find that between 37% (29 out of 79) and 44% (35 out of 79) of the CPI subcomponents are relevant. Thus a majority of variables are not related to the factor(s). If we follow Nakamura and Steinsson (2008) and form groups of sectoral CPI subindices, we find that almost every group contains at least one series associated with a factor. In particular, if we look at Panel A with $k = 1$ and the one-layer prior, the group-specific shares of relevant variables are (absolute numbers in brackets): apparel 36% (4 of 11 variables), 25% education and communication (1 of 4), food and beverages 24% (7 of 29), housing 36% (4 of 11), medical care 100% (5 of 5), recreation 71% (5 of 7), and transportation 30% (3 of 10). On the other hand, other goods and services (0 of 2) are unrelated to the common factor. The results are quite robust to changes in the model specification regarding the prior and the number of factors.

Note that these results are closely in line with the conclusions drawn by Mackowiak et al. (2009), who find that firms mainly pay attention to sector-specific rather than aggregate conditions when setting prices. Their conclusion relies on the small role of common shocks when compared with the role of idiosyncratic or sector-specific shocks based on variance decompositions. In a sparse factor model, irrelevant variables are not affected by shocks in common factors, as nonzero impulse responses require nonzero factor loadings. Thus our finding of many irrelevant variables in the data matches the finding of the small role of common shocks in the analysis of Mackowiak et al. (2009). Finally, note that the identification of irrelevant variables based on a sparse factor model does not hinge on the identification of structural shocks.

6 | CONCLUSION

In this paper, we propose to use a sparse factor model to discriminate between variables containing relevant information to estimate the unobserved factors and variables which are uninformative. In this context, we define irrelevant variables as those with only zero or insignificant factor loadings. In this case, the variable is entirely driven by idiosyncratic noise and not by the common factors. For various economic questions, it might be helpful to know which variables in the data are relevant and which are irrelevant. For example, any impulse response of a variable to a structural shock in a factor model can only be significant if at least one factor loading is significantly different from zero. We propose to discriminate between relevant and irrelevant variables by means of univariate and multivariate HPD regions and the posterior probability of a row with all loadings equal to zero. To estimate the factor model, we choose a sparse prior distribution for the factor loadings to sharply discriminate between relevant and irrelevant variables. This prior encompasses the normal prior, which is often used to estimate factor models for economic datasets. The inference on relevance or irrelevance is then based on identified posterior draws of the factor loadings.

In Monte Carlo simulations we show that estimation efficiency of the common component improves under a sparse prior specification of the loadings compared with a normal prior specification. In particular, estimates obtained under a sparse prior are quite robust to different degrees of sparsity in the DGP. The higher the degree of sparsity in the data, the better is the estimation performance of the sparse prior over the normal prior. Even in the case of little sparsity in the DGP, the sparse prior generally performs better than the normal prior. The methods to identify relevant and irrelevant variables generally work quite well, as more than 90% of variables are classified correctly.

To illustrate the methods, we identify relevant variables in two economic datasets. We first estimate a sparse factor model for an international GDP growth dataset, which has been used by Francis et al. (2012) and Kose et al. (2003) to analyze the role of global business cycles. We identify between 42% and 77% of the countries as relevant. African countries seem to be—if at all—only weakly related to international business cycles. This is in contrast to the evidence for European and North American countries. An additional application uses disaggregated US CPI inflation data taken from Mackowiak et al. (2009). In these data, we detect even more sparsity, given that only between 37% and 44% of sectoral inflation rates are informative for the factors. Most sectoral price developments seem to be driven by idiosyncratic rather than common determinants. Overall, we find a high degree of sparsity in both datasets. In general, however, results will depend on the data and empirical application chosen. Overall, the proposed methods seem to be adequate tools to determine relevant and irrelevant variables in the presence of heterogeneous information in large economic datasets.

ACKNOWLEDGMENTS

We thank three referees and the Editor, Herman van Dijk, for very helpful comments and suggestions. We are also grateful to Jörg Breitung, Anne Sofie Jore, Michael Owyang, and Markus Pape, as well as participants at the Annual Conference of the German Economic Association Frankfurt 2011, Econometrics Seminar at the University of Kiel 2011, Econometrics Workshop at the University of Frankfurt 2012, Joint Research Workshop of the Norges Bank and Deutsche Bundesbank 2012, DIW Macroeconometric Workshop Berlin 2012, Annual Conference of the German Statistical Association Berlin 2013, ESEM Gothenburg 2013, Applied Time Series Workshop St. Louis Fed 2013, and the Research Seminar at the University of Regensburg 2014, for further discussions and remarks.

REFERENCES

Aguilar, O., & West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, *18*, 338–357.

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., & Veronese, G. (2010). New Eurocoin: Tracking economic growth in real time. *Review of Economics and Statistics*, *92*, 1024–1034.

Anderson, T., & Rubin, H. (1956). Statistical inference in factor analysis. In Neyman, J (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability: Contributions to econometrics, industrial research, and psychometry* (Vol. *5*, pp. 111–150). Berkeley, CA: University of California Press.

Bai, J., & Ng, S. (2013). Principal components estimation and identification of the factors. *Journal of Econometrics*, *176*, 18–29.

Bai, J., & Wang, P. (2014). Identification theory for high dimensional static and dynamic factor models. *Journal of Econometrics*, *178*, 794–804.

Bauwens, L., Lubrano, M., & Richard, J. F. (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford, UK: Oxford University Press.

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

Bhattacharya, A., & Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika*, *98*, 291–306.

Boivin, J., & Ng, S. (2006). Are more data always better for factor analysis?. *Journal of Econometrics*, *132*, 169–194.

Carter, C. K., & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, *84*, 541–553.

Carvalho, C. M. (2006). *Structure and sparsity in high-dimensional multivariate analysis*, (Doctoral dissertation, Institute of Statistics and Decision Sciences, Duke University, Durham, NC).

Carvalho, C. M., Chang, C., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, *103*, 1438–1456.

Conti, G., Frühwirth-Schnatter, S., Heckman, J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics*, *183*, 31–57.

Del Negro, M., & Otrok, C. (2008). Dynamic factor models with time-varying parameters: Measuring changes in international business cycles. (*Staff Report No.326*): New York, NY: Federal Reserve Bank of New York.

Forni, M., Giannone, D., Lippi, M., & Reichlin, L. (2009). Opening the black box: Structural factor models with large cross-sections. *Econometric Theory*, *25*, 1319–1347.

Forni, M., Hallin, M., Lippi, M., & Zaffaroni, P. (2015). Dynamic factor models with infinite-dimensional factor space: One-sided representations. *Journal of Econometrics*, *185*, 359–371.

Francis, N., Owyang, M. T., & Savascin, Ö. (2012). An endogenously clustered factor approach to international business cycles. (*Working Paper 2012-014A*): St.Louis, MO: Federal Reserve Bank of St.Louis.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, *15*, 183–202.

Frühwirth-Schnatter, S. (2011). Dealing with label switching under model uncertainty. In Kerrie, CPR, Mengersen, L, & Titterington, DM (Eds.), *Mixtures: Estimation and Applications* (pp. 213–240). Hoboken, NJ: Wiley.

Frühwirth-Schnatter, S., & Lopes, H. F. (2010). Parsimonious Bayesian factor analysis when the number of factors is unknown. (*Technical report*): Chicago, IL: University of Chicago Booth School of Business.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*, 339–373.

Geweke, J. (1996). Variable selection and model comparison in regression. *Bayesian Statistics*, *5*, 609–620.

Geweke, J., & Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, *9*, 557–587.

Hanson, T., & McMillan, G. T. (2012). Scheffé-style simultaneous credible bands for regression surfaces with application to Ache honey gathering. *Journal of Data Science*, *10*, 175–193.

Held, L. (2004). Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics*, *13*, 20–35.

Heston, A., Summers, R., & Aten, B. (2011). *Penn World Table Version 7.0*: Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In Dodge, Y (Ed.), *Statistical data analysis based on the L_1–norm and related methods* (pp. 405–416). Amsterdam, Netherlands: North-Holland.

Kaufmann, S., & Schumacher, C. (2012). Finding relevant variables in sparse Bayesian factor models: Economic applications and simulation results. (*Discussion Paper No.29/2012*): Frankfurt am Main, Germany: Deutsche Bundesbank.

Kaufmann, S., & Schumacher, C. (2013). Bayesian estimation of sparse dynamic factor models with order-independent identification. (*Working Paper 13.04*): Gerzensee, Switzerland: Study Center Gerzensee.

Knowles, D., & Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modelling. *Annals of Applied Statistics*, *5*(2B), 1534–1552.

Koopman, S. J., & Harvey, A. (2003). Computing observation weights for signal extraction and filtering. *Journal of Economic Dynamics & Control*, *27*, 1317–1333.

Kose, M. A., Otrok, C., & Whiteman, C. H. (2003). International business cycles: World, region and country specific factors. *American Economic Review*, *93*, 1216–1239.

Kose, M. A., Otrok, C., & Whiteman, C. H. (2010). Understanding the evolution of world business cycles. *Journal of International Economics*, *75*, 110–130.

Lawley, J. N., & Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*. London, UK: Butterworths.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions: Five years of experience. *Journal of Business & Economic Statistics*, *4*, 25–38.

Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, *14*, 41–67.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., & West, M. (2006). Sparse statistical modelling in gene expression genomics. In Do, KA, Müller, P, & Vannucci, M (Eds.), *Bayesian inference for gene expression and proteomics* (pp. 155–176). Cambridge, UK: Cambridge University Press.

Mackowiak, B., Mönch, E., & Wiederholt, M. (2009). Sectoral price data and models of price setting. *Journal of Monetary Economics*, *56*, S78–S99.

Nakamura, E., & Steinsson, J. (2008). Five facts about prices: A reevaluation of menu cost models. *Quarterly Journal of Economics*, *123*, 1415–1464.

Stock, J., & Watson, M. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*, 147–162.

West, M. (2003). Bayesian factor regression models in the “large p, small n” paradigm. *Bayesian Statistics*, *7*, 723–732.

How to cite this article: Kaufmann S, Schumacher C. Identifying relevant and irrelevant variables in sparse factor. *J Appl Econ*. 2017;0:1-22. https://doi.org/10.1002/jae.2566

APPENDIX A

POSTERIOR DISTRIBUTIONS OF LOADINGS AND HYPERPARAMETERS

This section provides the conditional posterior distributions used in Steps 3 and 4 of the sampler to update the elements of the factor loading matrix and the hyperparameters. Subsection A.1 covers the two-layer prior, whereas Subsection A.2 covers the one-layer prior and the normal prior, respectively.

A.1 Factor loadings with the two-layer prior

The posterior $\pi\left(\lambda_{ij}|f^T, X^T, \Sigma_\epsilon\right)$ is obtained by first integrating out the variable-specific prior probability of a nonzero loading for each factor j. The prior in Equations 3–5 implies a common base rate of a nonzero factor loading of $E\left(\beta_{ij}\right) = \rho_j b$ across variables. The marginal prior becomes

$$\pi\left(\lambda_{ij}|\rho_j\right) \sim (1-\rho_j b)\delta_0(\lambda_{ij}) + \rho_j b N\left(0, \tau_j\right). \tag{A1}$$

For each factor j, transform the variables to

$$x^*_{it} = x_{it} - \sum_{l=1, l\neq j}^{k} \lambda_{il} f_{lt} = \lambda_{ij} f_{jt} + \epsilon_{it}, \tag{A2}$$

which basically isolates the effect of factor j in variable i. Combine the marginal prior with data information to sample independently across i from

$$\pi\left(\lambda_{ij}|\cdot\right) = \prod_{t=q+1}^{T} \pi(x_{it}^{*}|\cdot)\left\{(1-\rho_j b)\delta_0(\lambda_{ij}) + \rho_j b N\left(0,\tau_j\right)\right\} \tag{A3}$$

$$= P\left(\lambda_{ij}=0|\cdot\right)\delta_0(\lambda_{ij}) + P\left(\lambda_{ij}\neq 0|\cdot\right) N\left(m_{ij}, M_{ij}\right), \tag{A4}$$

with observation density $\pi(x_{it}^{*}|\cdot) = N\left(\lambda_{ij} f_{jt}, \sigma_{\varepsilon,i}^2\right)$ and where

$$M_{ij} = \left(\frac{1}{\sigma_{\varepsilon,i}^2}\sum_{t=1}^{T} f_{jt}^2 + \frac{1}{\tau_j}\right)^{-1}, \qquad m_{ij} = M_{ij}\left(\frac{1}{\sigma_{\varepsilon,i}^2}\sum_{t=1}^{T} f_{jt} x_{it}^{*}\right). \tag{A5}$$

To obtain the posterior odds $P\left(\lambda_{ij}\neq 0|\cdot\right)/P\left(\lambda_{ij}=0|\cdot\right)$ we update the prior odds of nonzero factor loading:

$$\mathrm{PO}_{ij} = \frac{P\left(\lambda_{ij}\neq 0|\cdot\right)}{P\left(\lambda_{ij}=0|\cdot\right)} = \frac{\pi\left(\lambda_{ij}\right)|_{\lambda_{ij}=0}}{\pi\left(\lambda_{ij}|\cdot\right)|_{\lambda_{ij}=0}}\frac{\rho_j b}{1-\rho_j b} = \frac{N\left(0;0,\tau_j\right)}{N\left(0;m_{ij},M_{ij}\right)}\frac{\rho_j b}{1-\rho_j b}. \tag{A6}$$

We choose $\lambda_{ij}\neq 0$ if $U \leqslant \mathrm{PO}_{ij}/(1+\mathrm{PO}_{ij})$, where U is a draw from the uniform distribution over [0, 1]. If the binary decision leads to $\lambda_{ij}\neq 0$, λ_{ij} is drawn from $N\left(m_{ij}, M_{ij}\right)$, otherwise it is set equal to zero.

Conditional on λ_{ij}, we can update the hyperparameters of the two-layer prior: To update the variable specific probabilities β_{ij}, we sample from $\pi(\beta_{ij}|\lambda_{ij},\cdot)$. Conditional on $\lambda_{ij}=0$:

$$\pi(\beta_{ij}|\lambda_{ij}=0,\cdot) \propto (1-\beta_{ij})\left[(1-\rho_j)\delta_0(\beta_{ij}) + \rho_j B(ab, a(1-b))\right], \tag{A7}$$

$$P(\beta_{ij}=0|\lambda_{ij}=0,\cdot) \propto (1-\rho_j), \quad P(\beta_{ij}\neq 0|\lambda_{ij}=0,\cdot) \propto (1-b)\rho_j. \tag{A8}$$

That is, with posterior odds $(1-b)\rho_j/(1-\rho_j)$ we sample from $B(ab, a(1-b)+1)$ and set otherwise β_{ij} equal to zero. Conditional on $\lambda_{ij}\neq 0$, we obtain

$$\pi(\beta_{ij}|\lambda_{ij}\neq 0,\cdot) \propto \beta_{ij} N\left(\lambda_{ij};0,\tau_j\right)\left[(1-\rho_j)\delta_0(\beta_{ij}) + \rho_j B\left(ab, a(1-b)\right)\right], \tag{A9}$$

$$P(\beta_{ij}=0|\lambda_{ij}\neq 0,\cdot) = 0, \quad P(\beta_{ij}\neq 0|\lambda_{ij}\neq 0,\cdot) = 1. \tag{A10}$$

In this case we sample β_{ij} from $B(ab+1, a(1-b))$.

The posterior update of the hyperparameters τ_j and ρ_j is sampled from an inverse Gamma, $\pi\left(\tau_j|\cdot\right) \sim \mathrm{IG}\left(g_j, G_j\right)$, and a Beta distribution, $\pi\left(\rho_j|\cdot\right) \sim B\left(r_{1j}, r_{2j}\right)$, respectively, with

$$g_j = g_0 + \frac{1}{2}\sum_{i=1}^{N} I\{\lambda_{ij}\neq 0\}, \quad G_j = G_0 + \frac{1}{2}\sum_{i=1}^{N}\lambda_{ij}^{2}, \tag{A11}$$

$$r_{1j} = r_0 s_0 + S_j, \quad r_{2j} = r_0(1-s_0) + N - S_j, \text{where} \quad S_j = \sum_{i=1}^{N} I\{\beta_{ij}\neq 0\}. \tag{A12}$$

A.2 One-layer and normal prior

Updating the one-layer prior from Equations 6–7 and the normal prior work in a similar way as in the previous section, as the priors are nested in the two-layer prior.

The one-layer prior consisting of $\pi(\lambda_{ij}) = (1-\rho_j)\delta_0(\lambda_{ij}) + \rho_j N(0, \tau_j)$, and $\pi(\rho_j) = B(r_0 s_0, r_0(1-s_0))$ assumes that there is a common probability across units of a zero loading on factor j. The posterior for λ_{ij} is basically governed by the same moments as derived above for the two-layer prior, but with adjusted posterior odds of zero factor loading:

$$\mathrm{PO}_{ij} = \frac{P(\lambda_{ij} \neq 0|\cdot)}{P(\lambda_{ij} = 0|\cdot)} = \frac{\pi(\lambda_{ij})\,|_{\lambda_{ij}=0}}{\pi(\lambda_{ij}|\cdot)\,|_{\lambda_{ij}=0}} \frac{\rho_j}{1-\rho_j} = \frac{N(0; 0, \tau_j)}{N(0; m_{ij}, M_{ij})} \frac{\rho_j}{1-\rho_j}. \tag{A13}$$

The posterior of the hyperparameter ρ_j is in this case $\pi(\rho_j|\cdot) \sim B(r_{1j}, r_{2j})$ with

$$r_{1j} = r_0 s_0 + S_j, \quad r_{2j} = r_0(1-s_0) + N - S_j, \text{where} \quad S_j = \sum_{i=1}^{N} I\{\lambda_{ij} \neq 0\}. \tag{A14}$$

Under a normal prior $\pi(\lambda_{i\cdot}) \sim N(0, \tau_0)$ with $\tau_0 = \text{diag}\,(\tau_1 \ \dots \ \tau_k)$ and $\pi(\tau_j) \sim \mathrm{IG}(g_0, G_0)$, we can sample independently over i from the posterior:

$$\pi(\lambda_{i\cdot}|\cdot) = N(m_i, M_i), \tag{A15}$$

$$M_i = \left(\frac{1}{\sigma^2_{\varepsilon,i}} F'F + \tau^{-1}\right)^{-1}, \quad m_i = M_i\left(\frac{1}{\sigma^2_{\varepsilon,i}} F' x_i\right), \tag{A16}$$

where F and X_i are, respectively, defined as $F = (f_1', \dots, f_T')'$ and $x_i = (x_{i1}, \dots, x_{iT})'$.

The posterior of the hyperparameter τ_j is $\pi(\tau_j|\cdot) = \mathrm{IG}(g_j, G_j)$ with $g_j = g_0 + 0.5N$ and $G_j = G_0 + 0.5\sum_{i=1}^{N} \lambda_{ij}{}^2$.

APPENDIX B

DETERMINING THE NUMBER OF FACTORS

B.1 Methods

To determine the number of factors, we will vary the number of estimated factors and infer on the number of columns containing more than two nonzero factor loadings (Anderson & Rubin, 1956; Conti, Frühwirth-Schnatter, Heckman & Piatek, 2014, Theorem 1). We set a maximum number $k^{\max}$ of estimated factors, which is chosen to be considerably larger than the expected number of factors in the DGP. We estimate the factor model for $k = 1, 2, \dots, k^{\max}$, and apply three different methods to determine the number of factors:

1. *Implied posterior number of factors*: For each posterior sample g, we obtain a loadings matrix $\lambda^{(g)}$, and the implied number of factors for this draw is equal to the number of columns in $\lambda^{(g)}$ containing more than two nonzero factor loadings:

$$\hat{k}^{\mathrm{imp},(g)}(k) = \sum_{j=1}^{k} I\left\{\left(\sum\nolimits_{i=1}^{N} I\{\lambda_{ij}^{(g)} \neq 0\}\right) > 2\right\}. \tag{B1}$$

   Based on G^{eff} draws, we can investigate the posterior samples for the implicit number of factors $\{\hat{k}^{\mathrm{imp},(g)}(k)\}_{g=1}^{G^{\mathrm{eff}}}$. The mode of the distribution denoted by $\hat{k}^{\mathrm{imp}}(k)$ is a point estimate of the number of factors. Note that this statistic depends on the number k of estimated factors, and we can estimate the implicit number of factors for each $k = 1, 2, \dots, k^{\max}$. Thus we are able to check whether the implicit number of factors changes with the number k of estimated factors.
2. *Permutations in k -medoids*: For each k, we take all the posterior samples of the factors $\{f^{T(g)}\}_{g=1}^{G^{\mathrm{eff}}}$ and remove those draws for which $\hat{k}^{\mathrm{imp},(g)}(k)$ differs from $\hat{k}^{\mathrm{imp}}(k)$, retaining $\tilde{G}^{\mathrm{eff}}$ draws. We apply k-medoids clustering to the $\tilde{G}^{\mathrm{eff}} \times \hat{k}^{\mathrm{imp}}(k)$ draws of factors with $\hat{k}^{\mathrm{imp}}(k)$ clusters. For each draw of the factor series $\{f_{1t}^{(g)}, f_{2t}^{(g)}, \dots, f_{\hat{k}^{\mathrm{imp}}(k),t}^{(g)}\}_{t=1}^{T}$, we obtain a new classification to clusters with index $\tilde{\kappa}^{(g)}$, with elements $\tilde{\kappa}_j^{(g)} \in \{1, \dots, \hat{k}^{\mathrm{imp}}(k)\}$, $j = 1, \dots, \hat{k}^{\mathrm{imp}}(k)$. If this classification equals a permutation of $\{1, \dots, \hat{k}^{\mathrm{imp}}(k)\}$, we keep the draw and reorder factors $\{f_{\tilde{\kappa}_1^{(g)},t}^{(g)}, f_{\tilde{\kappa}_2^{(g)},t}^{(g)}, \dots, f_{\tilde{\kappa}_{\hat{k}^{\mathrm{imp}}(k)}^{(g)},t}^{(g)}\}_{t=1}^{T}$. Otherwise, the draw is discarded (Frühwirth-Schnatter, 2011). We compute the share of retained permutations $G_0(k)$ to the overall number of draws, $G_0(k)/G^{\mathrm{eff}}$, and repeat this for all $k = 1, 2, \dots, k^{\max}$. The number of factors is determined by the largest k for which $G_0(k)/G^{\mathrm{eff}} > 0.95$. This procedure is adopted from the framework of mixture models. Frühwirth-Schnatter (2011) shows that nonpermutations of classifications to clusters are indicators of overfitting with respect to the number of components.

TABLE B1 Simulation results for alternative methods to determine the number of factors under different priors

		k							$\hat{k}$
		1	2	3	4	5	10	15	
A. One-layer prior									
$\hat{k}^{\text{imp}}(k)$	5%	1.000	2.000	2.000	2.000	2.000	3.000	4.000	—
	Mean	1.000	2.000	2.015	2.123	2.308	3.520	4.739	—
	Median	1.000	2.000	2.000	2.000	2.000	3.000	5.000	—
	95%	1.000	2.000	2.000	3.000	3.000	4.250	6.000	—
$G_0(k)/G^{\text{eff}}$	5%	1.000	1.000	0.581	0.326	0.216	0.054	0.014	2.000
	Mean	1.000	1.000	0.746	0.580	0.450	0.116	0.033	2.000
	Median	1.000	1.000	0.778	0.617	0.493	0.123	0.028	2.000
	95%	1.000	1.000	0.831	0.701	0.585	0.191	0.058	2.000
BIC$(k)(\times e^4)$	5%	2.325	2.225	2.252	2.281	2.310	2.472	2.657	2.000
	Mean	2.364	2.288	2.315	2.343	2.372	2.532	2.716	2.000
	Median	2.367	2.295	2.321	2.349	2.378	2.538	2.722	2.000
	95%	2.400	2.352	2.375	2.402	2.431	2.588	2.771	2.000
B. Two-layer prior									
$\hat{k}^{\text{imp}}(k)$	5%	1.000	2.000	2.000	2.000	2.000	3.000	4.000	—
	Mean	1.000	2.000	2.015	2.139	2.242	3.569	4.723	—
	Median	1.000	2.000	2.000	2.000	2.000	4.000	5.000	—
	95%	1.000	2.000	2.000	3.000	3.000	5.000	6.000	—
$G_0(k)/G^{\text{eff}}$	5%	1.000	1.000	0.580	0.291	0.218	0.046	0.013	2.000
	Mean	1.000	1.000	0.744	0.575	0.423	0.111	0.034	2.000
	Median	1.000	1.000	0.779	0.609	0.477	0.098	0.029	2.000
	95%	1.000	1.000	0.832	0.699	0.580	0.187	0.057	2.000
BIC $(k)(\times e^4)$	5%	2.325	2.225	2.252	2.281	2.310	2.472	2.657	2.000
	Mean	2.366	2.288	2.315	2.343	2.372	2.532	2.716	2.000
	Median	2.367	2.295	2.321	2.349	2.378	2.537	2.722	2.000
	95%	2.398	2.352	2.376	2.403	2.430	2.588	2.773	2.000

Note. The table contains results on the choice of the number of factors, when the factor model is estimated on simulated data under the one-layer (Panel A) and the two-layer prior (Panel B). In each panel, the table compares three methods to estimate the number of factors: the implicit number of factors $\hat{k}^{\text{imp}}(k)$ for different k, the relative number of permutations in k-medoids clustering $G_0(k)/G^{\text{eff}}$, and the BIC(k) (multiplied by e^4 for better readability of the entries), as outlined in Appendix B.1. For each method, the mean, the median, as well as the 5th and 95th percentiles of its distribution across $R = 65$ replications from the DGP are shown for different k. The final column for $G_0(k)/G^{\text{eff}}$ and BIC(k) shows the mean and percentiles of the number of factors $\hat{k}$ chosen from the results for different k.

3. *Bayesian information criterion (BIC)*: We estimate the BIC for all $k = 1, 2, \ldots, k^{\max}$ and pick the specification with the lowest BIC (see Lopes & West, 2004). The number of parameters in the factor model used here is equal to $Nk+Nq+N+k^2p$, where q is the number of idiosyncratic lags, whereas p is the number of factor VAR lags. The likelihood is estimated as the average over all MCMC draws.

The first method using the posterior of the implicit number of factors has been used in Frühwirth-Schnatter and Lopes (2010) and Bhattacharya and Dunson (2011). However, this method has not been applied to large economic datasets. The number of permutations has not been used in the context of large factor models before, either. In contrast, the BIC is often used for model selection. Simulations will show whether the methods adapted from the literature on mixture models are helpful in the present context.

B.2 Simulation results

Table B1 contains results based on the simulated data with two factors and $s_0 = 0.5$ in the DGP as outlined in Section 4.1. The sparse factor model is estimated on each simulated dataset conditional on $k = 1, 2, 3, 4, 5, 10, 15$. In Table B1, we provide

TABLE B2 Empirical results for alternative methods to determine the number of factors

		k										
		1	2	3	4	5	6	7	8	9	10	15
A. International GDP growth												
$G_0(k)/G^{\text{eff}}$	One-layer	1.000	0.999	0.984	0.937	0.559	0.160	0.111	0.145	0.047	0.038	0.013
	Two-layer	1.000	0.999	0.951	0.990	0.551	0.165	0.156	0.135	0.048	0.044	0.006
BIC$(k)(\times e^4)$	One-layer	0.662	0.671	0.680	0.697	0.723	0.752	0.782	0.814	0.848	0.883	1.084
	Two-layer	0.662	0.672	0.681	0.696	0.723	0.751	0.782	0.814	0.848	0.883	1.084
B. US CPI inflation												
$G_0(k)/G^{\text{eff}}$	One-layer	1.000	1.000	0.754	0.565	0.445	0.372	0.323	0.312	0.241	0.191	0.090
	Two-layer	1.000	1.000	0.802	0.545	0.495	0.318	0.358	0.298	0.279	0.213	0.109
BIC$(k)(\times e^5)$	One-layer	1.266	1.268	1.271	1.274	1.278	1.283	1.289	1.296	1.304	1.312	1.361
	Two-layer	1.266	1.268	1.271	1.275	1.278	1.283	1.289	1.296	1.304	1.312	1.361

Note. The table contains results on the choice of the number of factors, when the factor model is estimated on international GDP growth data (Panel A) and US CPI inflation data (Panel B) 2. In each panel, the table compares two methods to estimate the number of factors explained in Appendix B.1: the relative number of permutations in k-medoids clustering $G_0(k)/G^{\text{eff}}$ and the BIC(k) multiplied by e^4 in Panel A and multiplied by e^5 in Panel B for better readability of the entries. The entries in the table are the estimated statistics $G_0(k)/G^{\text{eff}}$ and BIC(k) for different k and the one- and the two-layer priors.

results for factor models estimated using the one- and two-layer priors with means $s_0 = 0.5$, $b = 0.8$ and precision $r_0 = a = 0.03$. Compared to the applications in the main text, the precision has been decreased to account for the additional uncertainty regarding the number of factors. We apply the three methods to determine the number of factors above in the following way. We determine the implicit number of factors $\hat{k}^{\text{imp}}(k)$ for each k and report percentiles and the mean obtained from the $R = 1, \ldots, 65$ replications. We hope to find numbers close to two for estimated models with different numbers of factors k. Considering permutations in k-medoids, we provide the statistic for each k and report percentiles and the mean of $G_0(k)/G^{\text{eff}}$ estimated across replications. We pick the largest k for which $G_0(k)/G^{\text{eff}} > 0.95$. This estimate of the number of factors is denoted by $\hat{k}$ and shown in the last column of Table B1, where we provide the percentiles and the mean of $\hat{k}$ across the R simulation samples. With respect to the BIC, we proceed in the same way as for the number of permutations.

The empirical results can be summarized as follows. The implicit number of factors $\hat{k}^{\text{imp}}(k)$ can be used to detect the number of factors, when the estimated number of factors is not strongly overfitting, that is, when $k \leqslant 5$. For $k = 10, 15$ there are signs of overfitting, as the median of the implicit number of factors equals three and five, and the lower 5% bounds of the distribution are greater than the true number of factors. The mean of the distribution is greater than the true number of factors already for $k \geqslant 3$. The reason for the performance is that some column draws of the loadings matrix contain more than two nonzero elements as required in Equation B1. In the relatively large datasets with $N = 60$, we often find cases where up to eight loadings elements are unequal zero. Given the DGP, we can identify these spurious columns: When looking at the posterior samples of the corresponding factors, they typically are not significantly different from zero. In the simulations, the ratio of permutations $G_0(k)/G^{\text{eff}}$ is equal to 1.0 for $k = 1, 2$ and then decreases below 0.95 for all samples. Thus in all cases the true number of factors is determined. The BIC also determines the right number of factors. We find no big differences between the one- and the two-layer priors. Thus, given the number of permutations and the BIC, it is possible to well determine the number of factors. This justifies the choice of fixed $k = 2$ in the simulation experiments.

B.3 Empirical results

Based on the simulation results in the previous section, we determine the number of factors for the two datasets used for the empirical applications. Owing to their relatively good performance, we focus on the relative number of permutations $G_0(k)/G^{\text{eff}}$ and the BIC to determine the number of factors. We estimate models with $k = 1, 2, \ldots, 9, 10, 15$ factors under the one- and two-layer prior specifications as above. The results can be found in Table B2.

For the international GDP growth data, we find that the ratio of permutations is greater than 0.95 for $k \leqslant 4$ when using the two-layer prior and thus indicates $\hat{k} = 4$. Based on the one-layer prior, the number of factors is three according to the ratio of permutations. The BIC indicates a very parsimonious specification with one factor for both priors. For the US CPI inflation data, the ratio of permutations indicates two factors, whereas BIC identifies one factor. The results for the one- and the two-layer priors lead to the same number of factors.

Overall, some uncertainty remains regarding the number of factors in the two empirical datasets. As the main focus of the paper is the identification of relevant and irrelevant variables, we provide a wide range of results for different numbers of factors.

Given the findings in this section, we provide empirical results for $k = 1, \ldots, 4$ when estimating the model on the international GDP growth data, and results for $k = 1, 2$ when estimating the model on US CPI inflation data.

APPENDIX C: DATA

C.1 Country coverage in the international GDP growth dataset

Below, the countries covered in the international GDP growth dataset used in Section 5 above are shown. Each country belongs to one of the geographical regions: Africa, Asia 1 (less developed), Asia 2 (more developed), Europe, Latin America, North America, and Oceania. The definition of geographical regions follows Kose et al. (2003):

1. Africa: Cameroon, Kenya, Morocco, Senegal, South Africa, Zimbabwe
2. Asia 1: Bangladesh, India, Indonesia, Pakistan, Philippines, Sri Lanka
3. Asia 2: Hong Kong, Japan, Korea, Malaysia, Singapore, Thailand
4. Europe: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, UK
5. Latin America: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Honduras, Jamaica, Panama, Paraguay, Peru, Uruguay, Venezuela
6. North America: Canada, Mexico, USA
7. Oceania: Australia, New Zealand

C.2 Sector coverage in the US CPI inflation dataset

Below, the sectors covered in the US CPI inflation dataset used in Section 5 are shown. Each sector belongs to one of the groups defined in Nakamura and Steinsson (2008):

1. *Apparel*: Men's suits, sport coats, and outerwear; Men's furnishings; Men's pants and shorts; Boys' apparel; Women's outerwear; Women's dresses; Girls' apparel; Men's footwear; Boys' and girls' footwear; Women's footwear; Infants' and toddlers' apparel
2. *Education and communication*: Admissions; Educational books and supplies; College tuition and fees; Elementary and high school tuition and fees
3. Food and beverages: Flour and prepared flour mixes; Breakfast cereal; Rice, pasta, cornmeal; Cakes, cupcakes, and cookies; Other bakery products; Uncooked ground beef; Ham; Pork chops; Other meats; Eggs; Cheese and related products; Ice cream and related products; Apples; Bananas; Potatoes; Lettuce; Tomatoes; Other fresh vegetables; Carbonated drinks; Coffee; Sugar and artificial sweeteners; Soups; Frozen and freeze-dried prepared foods; Snacks; Spices, seasonings, condiments, sauces; Beer, ale, and other malt beverages at home; Distilled spirits at home; Wine at home; Alcoholic beverages away from home
4. *Housing*: Rent of primary residence; Housing at school, excluding board; Other lodging away from home including hotels and motels; Owner's equivalent rent of primary residence; Fuel oil; Electricity; Utility (piped) gas service; Water and sewerage maintenance; Garbage and trash collection; Bedroom furniture; Clocks, lamps, and decorator items
5. *Medical care*: Prescription drugs and medical supplies; Internal and respiratory over-the-counter drugs; Nonprescription medical equipment and supplies; Physicians' services; Dental services
6. *Other goods and services*: Postage; Cosmetics, perfume, bath, nail preparations and implements
7. *Recreation*: Televisions; Audio equipment; Sports vehicles including bicycles; Sports equipment; Pets and pet products; Photographic equipment and supplies; Toys
8. *Transportation*: New vehicles; Used cars and trucks; Gasoline (all types); Tires; Motor vehicle body work; Motor vehicle maintenance and servicing; Motor vehicle insurance; Airline fare; Other intercity transportation; Intracity transportation