Specification of FAVAR Models

Robert MacDonald*
University of California, Irvine

Abstract

This paper proposes a novel methodology for determining the specification of factor-augmented vector autoregression (FAVAR) models. Without strong a priori beliefs about the set of possible models, the complexity of the problem renders traditional model selection techniques infeasible. By contrast, my proposed solution only requires the estimation of a single model. This makes the process easy to scale in both the cross-sectional and time series dimensions. An efficient optimization algorithm for model estimation is developed. Monte Carlo studies show the technique to be highly effective in small samples, even in the presence of a low signal-to-noise ratio and missing data. Applications to large datasets of monthly and quarterly U.S. macroeconomic variables identify observed factors not normally considered in the FAVAR literature. The methodology is then used to analyze the asset-pricing model of Fama and French (1993). I find that their constructed factors for firm size and book-to-market equity ratio are likely observed components, but excess market return is not.

JEL classification: C38, C11, C52, C61, G12
Keywords: Factor Models, Variable Selection, Big Data, Expectation-Maximization Algorithm, Asset Pricing

*Department of Economics, University of California-Irvine, 3151 Social Sciences Plaza, Irvine, CA 92697-5100; email: rmacdon1@uci.edu.
I would like to thank my advisor, Ivan Jeliazkov, for introducing me to this project and all of his subsequent help. I would also like to thank Fabio Milani, Eric Swanson, Jonathan Roth, Bruce Hansen, Julian Martinez-Iriarte, Guillaume Rocheteau, Michael Choi, the participants in the UCI Macro Brown Bag, and the participants in the UCI Econometrics Seminar for their helpful comments and advice. An earlier version of this paper was titled, "Identifying Observed Factors in FAVAR Models: a Bayesian Variable Selection Approach."

1 Introduction

Factor-augmented vector autoregressions (FAVARs) are a popular tool in big data time series analysis. The central assumption of any factor model is that much of the variation in a large panel of dependent variables can be explained by a relatively small number of common components. A standard FAVAR with an intercept is written as

$$X_t = \mu_X + \Lambda^f f_t + \Lambda^y y_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma), \tag{1}$$

$$\begin{bmatrix} f_t \\ y_t - \mu_y \end{bmatrix} = \Phi(L) \begin{bmatrix} f_{t-1} \\ y_{t-1} - \mu_y \end{bmatrix} + \eta_t, \quad \eta_t \sim N(0, \Omega). \tag{2}$$

X_t is an $N \times 1$ vector of dependent variables with unit variance, f_t is an $r_f \times 1$ vector of latent factors, y_t is an $r_y \times 1$ vector of observed factors, Λ^f and Λ^y are matrices of loading parameters, μ_X is the intercepts, and ε_t is an $N \times 1$ vector of error terms. The variables are observed in each time period $t = 1, \ldots, T$. The common factors f_t and y_t are assumed to explain all of the covariance in X_t. The idiosyncratic errors ε_t are thus assumed to have diagonal covariance matrix $\Sigma = diag(\sigma_1^2, \ldots, \sigma_N^2)$. The FAVAR reduces to a multivariate regression when $r_f = 0$ and a dynamic factor model (DFM) when $r_y = 0$. This specification allows for parsimonious modeling of high-dimensional data when $r = r_f + r_y << N$, thus offering an alternative to highly-parameterized vector autoregressions (VARs).

The FAVAR was originally developed for structural macroeconomic analysis by Bernanke, Boivin, and Eliasz (2005).[1] The authors assumed the Federal Funds Rate was the only observed factor and did not perform any model comparison with alternative observed factors. Models with observed factors, though not typically cast in terms of an FAVAR, are also common in the asset-pricing literature[2]. The number of possible observed factors to consider has grown quite large. Choosing the best subset from the available "factor zoo" (Cochrane, 2011) is of interest to researchers and investors alike.

There is no existing feasible method for comparing all of the possible FAVAR specifica-

1. See Belviso and Milani (2006); Boivin, Giannoni, and Stevanović (2013); Fernald, Spiegel, and Swanson (2014); Paccagnini (2017) among others for further detail as well as interesting extensions and applications.
2. Among many others, see Chen, Roll, and Ross, 1986; Fama and French, 1993; Fama and French, 2015.

tions. Model selection requires knowing the observed factors, the number of latent factors, and the lag order. Exhaustive model comparison would require estimating millions of models, even with modestly sized datasets. I propose a solution that only requires the estimation of a single model. The procedure exploits the fact that any FAVAR has an equivalent representation as a DFM. I first determine the total number of factors r and the lag order p using existing information criterion methods. The main contribution of this paper is the identification of observed factors. If an observed factor is modeled as a dependent variable in a DFM, then the associated idiosyncratic error term ε_{it} will have true variance $\sigma_i^2 = 0$. I estimate a DFM with all observed variables and place a Bayesian variable selection prior on each σ_i^2. The prior is designed to shrink small variances towards 0 while exercising little influence on larger variances.

Model selection is achieved through maximum a posteriori (MAP) estimation. To facilitate fast model selection, I develop several extensions to the Expectation-Maximization (EM) algorithm for DFMs. The proposed algorithm leverages the rotation and scale invariance of the likelihood to obtain a solution significantly faster than the basic EM algorithm.

The identification of observed factors has been addressed solely by the frequentist literature until this point. The two papers most closely related to this project are Bai and Ng (2006) and Parker and Sul (2016). Bai and Ng (2006) observe that if we can consistently estimate the factor space, then any observed factors will be linear combinations of the estimated factors. Their procedure relies on statistical tests in which some candidate variable is an observed factor under the null hypothesis. This approach is reasonable when the set of possible observed factors is small, but will encounter problems when the set is large. To systematically find the correct observed factors in a dataset with many variables, this requires running dozens or hundreds of independent tests and then performing a correction for multiple hypothesis testing. This method is unlikely to select the true model and may produce incoherent results, such as concluding there are more observed factors than total possible factors. Parker and Sul (2016) build upon the work of Bai and Ng (2006) to develop

a criterion for finding a set of candidate observed factors. When combined with a clustering algorithm, the criterion is effective at finding the set of all possible observed factors. However, this approach is agnostic about choosing between highly correlated candidates. If X_i is an observed factor and $X_j = X_i + \varepsilon_j$, where ε_j has a small variance, the criterion may conclude that either variable could be an observed factor. Both papers assume a balanced panel dataset and do not address the case of missing data. My model selection process does not encounter the same problems.

I apply the procedure to large datasets of monthly and quarterly macroeconomic data, as well as the asset-pricing data of Fama and French (1993). The analysis of quarterly macroeconomic data yields surprising results. Rather than selecting the Federal Funds Rate as an observed factor, the default assumption in the monetary FAVAR literature, the procedure selects the Total Capacity Utilization index. A model of monthly macroeconomic data selects the spread between the 10-Year Treasury Rate and the Federal Funds Rate as the only observed factor. Models restricted to the period following the 2007 Financial Crisis find that employment measures are more likely to be observed factors. I also estimate a model with the same variables as Fama and French (1993). Variables that measure the excess returns attributable to firm size and book-to-market equity ratio are classified as observed factors, while the excess return from a market portfolio is not.

The remainder of the paper proceeds as follows. Section 2 recasts the model selection process as an optimization problem. Section 3 develops an efficient EM algorithm for MAP estimation. Section 4 investigates the performance of the proposed procedure through Monte Carlo studies. Section 5 applies the new approach to large macroeconomic and financial datasets, and section 6 concludes. Mathematical proofs and technical details can be found in the appendix.

2 The Model Selection Procedure

2.1 Rotation Invariant Likelihood

A perennial issue in factor analysis is that the likelihood is invariant under rotations of the factor basis. Consider the case of a DFM with likelihood $f(X|\theta)$. For any square invertible matrix A, $\Lambda f_t = \Lambda A^{-1} A f_t$. Let $F = (f_1, \ldots, f_T)$, $\Lambda^* = \Lambda A^{-1}$, $F^* = AF$, $\Phi_l^* = A\Phi_l A^{-1}$, and $\Omega^* = A\Omega A'$. Assuming the first p instances of the factors come from the stationary distribution, where p is the VAR lag order in (2), we obtain the equality

$$\begin{aligned} f(X|\theta) &= \int f(X|F, \Lambda, \Sigma)\pi(F|\Phi, \Omega)dF \\ &= \int f(X|F^*, \Lambda^*, \Sigma)\pi(F^*|\Phi^*, \Omega^*)dF^* \\ &= f(X|\theta^*). \end{aligned} \tag{3}$$

This means that parameter restrictions are required to identify the likelihood. Unfortunately, a priori restrictions can lead to model misspecification. Identification is usually achieved through restrictions on an $r \times r$ submatrix of Λ. However, the restrictions are only valid if the true submatrix is invertible.

2.2 Rewriting the FAVAR as a DFM

If we stack X_t and y_t in a single vector, we can then rewrite the FAVAR as a special case of a DFM:

$$\begin{bmatrix} X_t \\ y_t \end{bmatrix} = \begin{bmatrix} \mu_X + \Lambda^y \mu_y \\ \mu_y \end{bmatrix} + \begin{bmatrix} \Lambda_f & \Lambda_y \\ 0_{r_y \times r_f} & I_{r_y} \end{bmatrix} \begin{bmatrix} f_t \\ y_t - \mu_y \end{bmatrix} + \varepsilon_t^\dagger, \qquad \varepsilon_t^\dagger \sim N\left(0, \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}\right). \tag{4}$$

Now consider rotating the factors by an arbitrary invertible matrix A: $f_t^* = A[f_t' \;\; y_t' - \mu_y']'$. The FAVAR can then be written as:

$$X_t^* = \mu_{X^*} + \Lambda^* f_t^* + \varepsilon_t^*, \qquad \varepsilon_t^* \sim N\left(0, \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}\right), \tag{5}$$

$$f_t^* = \Phi^*(L) f_{t-1}^* + \eta_t^*, \quad \eta_t^* \sim N(0, \Omega^*), \tag{6}$$

where $X_t^* = [X_t' \;\; y_t']'$, μ_{X^*} is the intercept from (4), and the remaining parameters with asterisks are defined as in section 2.1. We can thus conclude that any DFM in which some idiosyncratic variances are 0 is equivalent to an FAVAR where the corresponding variables are observed factors.

Rather than comparing estimates from different FAVAR specifications. I will estimate a DFM that nests all possible FAVARs with total number of factors r and lag order p. Identification is not an issue if your only aim is to determine the observed factors, r, and p. The elements of Σ do not change when the factor basis is rotated. Since the procedure I propose makes use of a MAP estimate from a Gaussian state-space model, the posterior can be optimized using an EM algorithm, which does not require an identified likelihood to converge to a maximum. This helps us avoid any model misspecification problems that can arise from a priori restrictions. The only normalization I assume is $\Omega = I_r$. This helps to scale identify the factors and facilitates jumping between points of equal probability to accelerate the EM algorithm. Once MAP estimates are obtained, the researcher is free to choose his or her preferred identifying restrictions and rotate the factor basis accordingly.

2.3 Determining the Observed Factors

Let us consider the problem of identifying the observed factors when r and p are known. The theoretically ideal method would be exhaustive Bayes factor comparisons. However, the combinatorial complexity of such a procedure requires prohibitively large computing resources even when r is small. One would have to estimate marginal likelihoods for $\sum_{r_y=0}^{r} \binom{N}{r_y}$ models. This amounts to over 4 million marginal likelihood estimations in the modest case of $N = 100$ and $r = 4$. Even if one used the Bayesian information criterion (BIC) to approximate the marginal likelihood, the model selection process would take months on a standard personal computer.

Since any variables with idiosyncratic variances of 0 must be observed factors, a closely related approach would be to place spike-and-slab priors on the variances. This would take

the form

$$\pi(\sigma_i^2) = (1 - \rho_i)\delta_0(\sigma_i^2) + \rho_i\psi_1(\sigma_i^2). \tag{7}$$

While the spike-and-slab prior recasts the problem in the context of a single model, it does not alleviate the problem of combinatorial complexity. To produce a posterior that is easier to traverse, let us consider a continuous approximation of the point-mass mixture prior. After adding a hierarchical prior on the mixing weight and a latent indicator for the components of the mixture, the prior for σ_i^2 is expressed as

$$\pi(\sigma_i^2|\gamma_i) = \psi_0(\sigma_i^2)^{1-\gamma_i}\psi_1(\sigma_i^2)^{\gamma_i}, \tag{8}$$

$$\psi_q(\sigma_i^2) = \alpha_q e^{-\alpha_q\sigma_i^2}, \tag{9}$$

$$\gamma_i \sim \text{Bernoulli}(\rho_i), \tag{10}$$

$$\rho_i \sim \mathcal{B}(a, a). \tag{11}$$

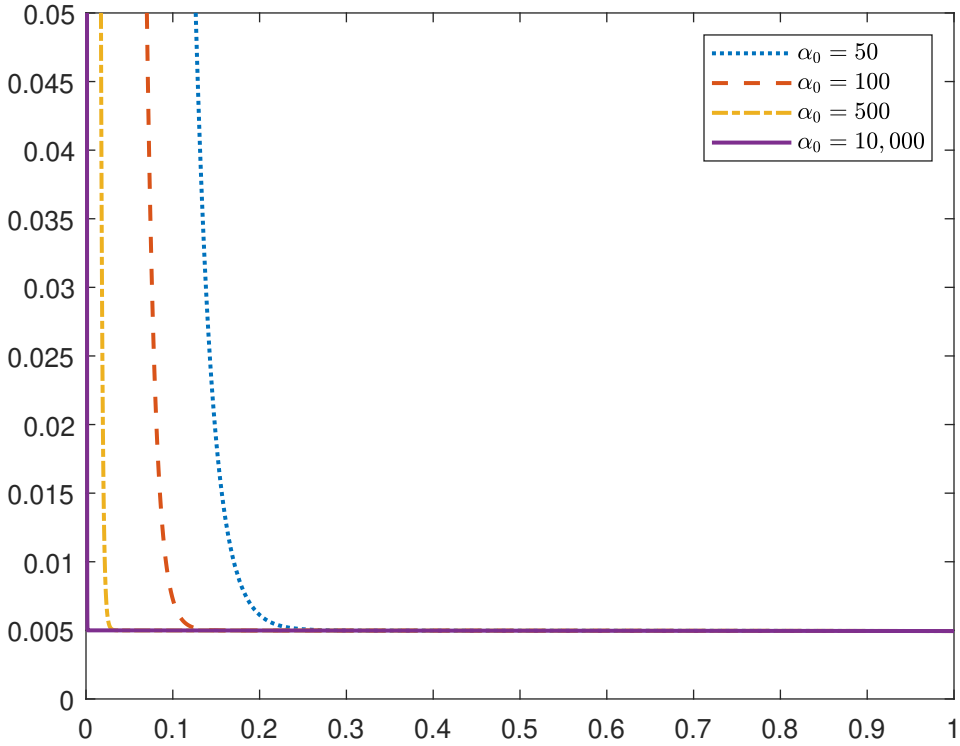
The spike-and-slab densities are exponential distributions. By setting $\alpha_0 >> \alpha_1$ and α_1 close to 0, we can place virtually all of the probability mass of the spike distribution (ψ_0) near 0 while placing nearly all of the mass of the slab distribution (ψ_1) away from 0. Figure 1 shows the mixture density for increasingly large α_0's and $\rho = 0.5$. This is equivalent to the marginal prior after γ_i and ρ_i have been integrated out. We can see that this continuous prior approaches the point-mass mixture prior as $\alpha_0 \to \infty$. I employ parameter expansion by augmenting the prior with the latent indicator variable γ_i. The latent variable formulation is amenable to closed-form updates of variance estimates within an EM algorithm.

Parameter estimates are obtained by solving the optimization problem

$$\begin{aligned} \hat{\theta}_{MAP} &= argmax_\theta \pi(\theta|X) \\ &= argmax_\theta f(X|\theta)\pi(\theta) \\ &= argmax_\theta \ln f(X|\theta) + \ln\pi(\theta). \end{aligned} \tag{12}$$

Care must be taken when optimizing a model for which some $\hat{\sigma}_{i,MAP}^2 = 0$. The EM algorithm requires the output from a Kalman smoother, which gives imprecise estimates when idiosyncratic variances are sufficiently small. I avoid this problem by first estimating a con-

Figure 1: Spike-and-Slab Priors for σ_i^2, $\alpha_1 = 0.01$, $\rho = 0.5$



strained model in which $\sigma_i^2 \geq 10^{-15}$. We can maintain numerical stability with variances of this size by using a square root Kalman smoother that leverages the QR decomposition (Tracy, 2022). After termination of the EM algorithm, I check to see if the posterior density can be increased further by setting any $\hat{\sigma}_{i,MAP}^2 < 10^{-8}$ to 0. The model with exacts 0's was preferred in all estimations.

2.4 Selecting the Number of Factors

Before applying the model selection process, we must first know the number of factors. There has been a great deal of work done with regard to estimating r. Many approaches in the frequentist literature develop information criteria (Bai and Ng, 2002; Hallin and Liška, 2007; Ahn and Horenstein, 2013). Another approach to estimating r is using an overidentified model, with more factors than is likely true, and then forcing factor loadings towards 0.

Frequentist methods accomplish this by applying regularization techniques like the LASSO to factors estimated using Principle Components Analysis (PCA) (Zou, Hastie and Tibshirani, 2006; Witten, Tibshirani and Hastie, 2009). Bayesian solutions typically place hierarchical shrinkage priors on the loading parameters (Carvalho et al., 2008; Frühwirth-Schnatter and Lopes, 2009; Knowles and Ghahramani, 2011; Ročková and George, 2016). These methods pertain to models in which the factors are assumed to be uncorrelated across time. There has been some recent work that extends the approach to restricted DFMs (McAlinn, Ročková, and Saha 2018; Luo and Yu, 2021).

Many Bayesian approaches for selecting r are unfortunately ill-suited to the problem at hand. Methods based on variable selection priors are less effective when the idiosyncratic variances are very small, which will occur for any observed factors as well as any other variables that are particularly well-explained by the common components. When continuous shrinkage priors are used, such as in Ročková and George (2016), very small variances reduce the variable selection penalty to effectively 0. Another issue arises when the factors are highly correlated, a situation that is not precluded by the model under consideration. In fact, highly correlated factors are likely to result when the model is overidentified. Overidentification does not create the same issue in static factor models because the factors are independent a priori. Point mass-density priors may be less susceptible to the problem of small variances, but they are still likely to encounter difficulties with highly correlated factors. Likelihood-based criteria such as Bayes factors and the Deviance Information Criterion unfortunately have a tendency to overfit the number of factors (Beyeler and Kaufmann, 2021). While BIC performed well in simulation studies, it exhibited the same overfitting property in empirical applications, continuing to decrease as the number of factors increased. It is for this reason I instead use the IC_{p2} criterion of Bai and Ng (2002). It can be computed quickly, has very good finite sample properties and tends to give reasonable factor estimates for models calibrated to macroeconomic applications (Stock and Watson, 2016).

2.5 Lag Length Selection

Conditional on the number of factors, I use BIC to select p. Calculation of the BIC is not obvious in this case because I have chosen to maximize the unidentified likelihood, which attains a maximum not at a single point but at a ridge. The derivation of the BIC requires the maximum to be unique for the Laplace approximation to be valid, meaning BIC can only be calculated for an identified model. However, the maximum likelihood value found for the unidentified likelihood will correspond to the maximum likelihood of any model with correct identifying restrictions (assuming there is at least one nonsingular $r \times r$ submatrix in $\hat{\Lambda}_{MLE}$). We can thus use the maximum likelihood from the unidentified model and the penalty from the identified model to compute

$$\text{BIC} = -2\ln f(X|\hat{\theta}_{MLE}) + \ln T(N(r+2) + pr^2 - r(r-1)/2). \tag{13}$$

3 Estimation Algorithm

Our goal is to maximize $\pi(\theta|X) \propto \int f(X|F,\theta)\pi(F|\theta)\pi(\theta)dF$. Maximization of the posterior is achieved using a variant of the EM algorithm. While the posterior can be maximized with a conventional EM algorithm, convergence is considerably slower. I instead use a combination of two EM variants with faster convergence properties: the Expectation/Conditional Maximization (ECME) algorithm (Liu and Rubin, 1994) and the Parameter-Expanded Expectation-Maximization (PX-EM) algorithm (Liu, Rubin, and Wu, 1998). The basic EM algorithm is an iterative process in which one can find parameter updates that monotonically increase the value of an integrated density, such as $\pi(\theta|X)$, by maximizing the posterior expectation of the full data density (Dempster, Laird, and Rubin, 1977). The parameter updates take the form

$$\theta_n = argmax_\theta \mathbb{E}[\ln f(X|F,\theta) + \ln\pi(F|\theta)|X,\theta_{n-1}] + \ln\pi(\theta) = argmax_\theta Q(\theta|\theta_{n-1}). \tag{14}$$

3.1 The PX-EM Algorithm for a DFM

The PX-EM algorithm exploits the rotational invariance of the likelihood. The proposed algorithm proceeds by first maximizing $Q(\theta|\theta_{n-1})$ with respect to $\theta^* = \{\Lambda^*, \Phi^*, \Omega^*, \Sigma\}$. We then take A_n^L to be the lower triangular Cholesky factor of Ω_n^* and set $\Lambda_n = \Lambda_n^* A_n^L$ and $\Phi_{l,n} = A_n^{L^{-1}} \Phi_{l,n}^* A_n^L$ for each lag l. By adopting improper priors for Λ and Φ, we obtain a posterior that is also rotation invariant. The priors for Λ and Φ are $\pi(\Lambda) \propto 1$, and $\pi(\Phi) \propto \mathbb{1}\{\Phi \in \mathcal{A}\}$, where $\mathcal{A}$ is the region of the parameter space for which the roots of the VAR polynomial lie outside the unit circle. The main advantage of improper priors is that they are rotation invariant, which will make the posterior easier to traverse. Improper priors can create convergence issues in Markov Chain Monte Carlo (MCMC) estimation, but they are not a problem in MAP estimation. The sequence of density ordinates generated by the EM updates will still converge to a stationary point (Wu, 1983). Any solution found by optimization is also a solution under an appropriately diffuse, Uniform prior. Diffuse proper priors are unlikely to impact posterior inference for Σ, but they do create difficulties for optimization. One may be inclined to choose diffuse semiconjugate priors such as $\pi(\Lambda_i) = f_N(\Lambda_i|0, \nu_\Lambda I_r)$ and $\pi(\Phi) \propto \mathbb{1}\{\Phi \in \mathcal{A}\} \prod_{i,l} f_N(\Phi_{il}|0, \nu_\Phi I_r)$. Such priors do little to identify the posterior because they are invariant under orthonormal rotations as well as sign and order permutations. However, they are not invariant under oblique rotations such as A^L, so we can no longer use the PX-EM algorithm. Proper priors merely restrict the modes of the posterior to a smaller ridge while making the parameter space more difficult to explore.

The original EM algorithm for maximum likelihood estimation of a DFM can be found in Watson and Engle (1983). I adapt their algorithm to account for the variable selection prior on Σ and the estimation of Ω^*. Details of the PX-EM algorithm are given in Algorithm 1. Define $\hat{f}_t \equiv \mathbb{E}[f_t|X, \theta_n]$, $\hat{F} \equiv (\hat{f}_1, \ldots, \hat{f}_T)'$, $\hat{P} \equiv \sum_t \mathbb{E}[(f_t - \hat{f}_t)(f_t - \hat{f}_t)'|X, \theta_n]$, and $\hat{\gamma}_i \equiv \Pr(\gamma_i = 1|X, \theta_n)$. All conditional moments related to the factors are available directly from the output of a Kalman smoother in which the state vector has been augmented to include an additional lag of f_t. The calculation of $\hat{\gamma}_i$ follows from a straightforward application of

Bayes' formula.

Algorithm 1 PX-EM Algorithm

while $\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}) - (\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}))$ > tolerance level **do**

E Step:

Run a Kalman smoother to obtain $\hat{F}$, $\hat{G}$ $\hat{P}$, $\hat{C}$, and $\hat{P}_g$.

for $1 \leq i \leq N$ **do**

$\hat{\gamma_i} = (1 + \frac{\rho_{i,n-1}}{1-\rho_{i,n-1}} \frac{\alpha_1}{\alpha_0} \exp((\alpha_0 - \alpha_1)\sigma^2_{i,n-1}))^{-1}$.

end for

M Step:

$\Lambda^*_n = X'\hat{F}(\hat{F}'\hat{F} + \hat{P})^{-1}$

for $1 \leq i \leq N$ **do**

$SS_i = \sum_t (X_{it} - \Lambda^*_{i,n}\hat{f}_t)^2 + \Lambda^*_{i,n}\hat{P}\Lambda^{*\prime}_{i,n}$

$\alpha^*_i = (1 - \hat{\gamma_i})\alpha_0 + \hat{\gamma_i}\alpha_1$

$\sigma^2_{i,n} = \frac{SS_i}{\frac{1}{2}(T + \sqrt{T^2 + 2\alpha^*_i SS_i})}$

$\rho_{i,n} = \frac{\hat{\gamma_i}+1-a}{2a-1}$

end for

$\{\Phi^*_n, \Omega^*_n\} = argmax_{\Phi^*_n, \Omega^*_n} Q(\theta|\theta_{n-1})$

Rotation Step:

Calculate A^L_n, the lower Cholesky factor of Ω^*_n.

$\Lambda_n = \Lambda^*_n A^L_n$

for $1 \leq l \leq p$ **do**

$\Phi_{l,n} = A^{L^{-1}}_n \Phi^*_{l,n} A^L_n$

end for

end while

3.2 The PX-ECME Algorithm for a DFM

The algorithm described in the previous section has faster convergence properties than the standard EM algorithm, but still encounters some difficulties maximizing the posterior, especially with respect to Σ, the parameters of greatest interest. To overcome this issue, I occasionally supplement the iterations of the PX-EM algorithm with an iteration from an ECME algorithm. The ECME algorithm works by iteratively maximizing functions of parameter blocks that are conditioned on the values of the remaining parameters from the last iteration. Better convergence properties are obtained by allowing the functions to be either conditional

Q functions, such as $\mathbb{E}[\ln\pi(\theta_1|X, F, \theta_{2,n-1})|X, \theta_{n-1}]$ or conditional log integrated densities, such as $\ln\pi(\theta_1|X, \theta_{2,n-1})$. For the problem at hand, I choose to update ρ using a conditional Q function and update the remaining parameters with conditional posterior densities. All of the conditional maximizations must be unique in order for the sequence of posterior ordinates generated by the ECME algorithm to converge (Liu and Rubin, 1994). The posterior distribution obviously does not have a unique maximum, but the conditional distributions do when the parameters are grouped by observation equation (1) and state equation (2). One option for an ECME iteration would be to first maximize $\ln\pi(\Lambda, \Sigma|X, \Phi_{n-1}, \Omega = I_r)$ with respect to Λ_n and Σ_n, then maximize $\ln\pi(\Phi|X, \Lambda_n, \Sigma_n, \Omega = I_r)$ with respect to Φ_n. I instead modify this step with parameter expansion by first maximizing $\ln\pi(\Lambda^*, \Sigma|X, \Phi_{n-1}, \Omega = I_r)$ with respect to Λ_n^* and Σ_n, then maximizing $\ln\pi(\Phi^*, \Omega^*|X, \Lambda_n^*, \Sigma_n)$ with respect to Φ_n^* and Ω_n^*. The solutions are then rotated back to the scale identified model, as in the PX-EM algorithm. Details are given in Algorithm 2. Except for ρ, all maximizations are done numerically using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Computation time is decreased by using the closed form solution for the gradient that results from the identity $\nabla\ln\pi(\theta_n|X) = \nabla Q(\theta_n|\theta_n)$ (Ruud, 1991).

3.3 Approximations to the Stationary Likelihood

Working with the stationary likelihood is theoretically ideal, but presents several challenges. Maximization of $Q(\theta|\theta_n)$ with respect to the VAR parameters must be done numerically, as no closed form solutions exist. Let us consider the state equation when it is rewritten from a $VAR(p)$ equation to a $VAR(1)$ equation. Let $g_t = (f_t', f_{t-1}', \ldots, f_{t-p+1}')'$.

$$g_t = Bg_{t-1} + \begin{bmatrix} \eta_t \\ 0_{r(p-1)\times 1} \end{bmatrix} \tag{15}$$

Rather than work with the stationary variance of the factor process, I will instead approximate the stationary likelihood by assuming the first p presample instances of the factors follow the distribution

Algorithm 2 PX-ECME Algorithm

while $\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}) - (\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1})) >$ tolerance level **do**
  E Step:
  for $1 \leq i \leq N$ **do**
    $\hat{\gamma}_i = (1 + \frac{\rho_{i,n-1}}{1-\rho_{i,n-1}} \frac{\alpha_1}{\alpha_0} \exp((\alpha_0 - \alpha_1)\sigma^2_{i,n-1}))^{-1}$.
  end for

  M Step:
  for $1 \leq i \leq N$ **do**
    $\rho_{i,n} = \frac{\hat{\gamma}_i + 1 - a}{2a-1}$
  end for
  $\{\Lambda^*_n, \Sigma\} = argmax_{\Lambda^*,\Sigma} \ln\pi(\Lambda, \Sigma|X, \rho_n, \Phi_{n-1}, \Omega = I_r)$
  $\{\Phi^*_n, \Omega^*_n\} = argmax_{\Phi^*_n,\Omega^*_n} \ln\pi(\Phi^*, \Omega^*|X, \rho_n, \Lambda^*_n, \Sigma_n)$

  Rotation Step:
  Calculate A^L_n, the lower Cholesky factor of Ω^*_n.
  $\Lambda_n = \Lambda^*_n A^L_n$
  for $1 \leq l \leq p$ **do**
    $\Phi_{l,n} = A^{L^{-1}}_n \Phi^*_{l,n} A^L_n$
  end for
end while

$$g_0 \sim N(0, \nu_{g_0} I_{pr}). \tag{16}$$

Integration over these presample instances of f_t yields a distribution for the first p instances of f_t of the form

$$g_p \sim N(0, \Omega_g + \nu_{g_0} B^p B^{p\prime} + \sum_{j=1}^{p-1} B^j \Omega_g B^{j\prime}), \;\; \Omega_g = \begin{bmatrix} \Omega & \\ & 0_{r(p-1)} \end{bmatrix}. \tag{17}$$

This functions as an approximation to the stationary distribution. The approximation could be made arbitrarily accurate by making the number of presample factors τ sufficiently large. As $\tau \to \infty$, the marginal covariance matrix of g_p will converge to the stationary covariance matrix P_0. However, such an approach will drastically reduce the efficiency of the EM algorithm. As τ increases, the curvature of $Q(\theta|\theta_n)$ increases, leading to smaller steps being taken in each parameter update. There is likely a more optimal choice of τ. A researcher may run a preliminary algorithm with a small number of iterations, then select his or her preferred τ by choosing a number such that $||P_{0,\tau} - P_0|| < m$, where m is a positive tuning parameter,

$||\cdot||$ is a matrix norm, and $P_{0,\tau}$ is the covariance matrix of g_p for a given τ. Alternatively, one could increase τ until the convergence time of the algorithm begins to suffer. While these approximations greatly increase the efficiency of the algorithm, the likelihood is no longer rotation invariant. The parameter-expanded algorithms I have developed are thus no longer guaranteed to produce updates that monotonically increase the likelihood. One way to make the algorithm monotonic is to only perform rotations if they increase the posterior density, and just perform regular EM updates otherwise. Another option is to only do parameter-expanded steps for a set number of iterations, then switch to a basic EM algorithm. Despite the loss of monotonicity, any fixed points of the parameter-expanded algorithms will also be fixed points of the basic EM algorithm. Non-monotonic updates were not an issue in applications to simulated or real data, while convergence was markedly faster. The PX-EM algorithm with the likelihood approximation is given in Algorithm 3. In addition to the posterior moments defined previously, let $\hat{g}_t \equiv \mathbb{E}[g_t|X,\theta_n]$, $\hat{G} \equiv (\hat{g}_0,\ldots,\hat{g}_{T-1})'$, $\hat{P}_g \equiv \sum_t \mathbb{E}[(g_{t-1}-\hat{g}_{t-1})(g_{t-1}-\hat{g}_{t-1})'|X,\theta_n]$, and $\hat{C} \equiv \sum_t \mathbb{E}[(f_t-\hat{f}_t)(g_{t-1}-\hat{g}_{t-1})'|X,\theta_n]$. The reader will note that all parameters updates are now available in closed-form.

Another option, should one wish to work with the exact stationary likelihood, is to only use ECME steps for updating the parameters of the state equation. Gradient-based methods for this problem require care. Calculation of the numerical gradient requires many runs of either a Kalman filter or a precision-based method for obtaining the integrated likelihood, as well as many high-dimensional matrix inversions to calculate the stationary variance. A precise approximation of the gradient can be computed in significantly less time by augmenting the state vector with many presample factors and using the fact that $\nabla \ln\pi(\theta_n|X) = \nabla Q(\theta_n|\theta_n)$ (Ruud, 1991). Justification for this approach is given by Proposition 1.

Proposition 1

Let $Q_{\tau-p}(\theta|\theta_n) \equiv \mathbb{E}[\ln\pi(X,F,f_0,f_{-1},\ldots,f_{-\tau+p+1},\theta|f_{-\tau+p},\ldots,f_{-\tau+1})|X,\theta_n]$ and assume

θ_n is an interior point of the parameter space.

$$\lim_{\tau \to \infty} \nabla Q_{\tau - p}(\theta_n | \theta_n) = \nabla \ln \pi(\theta_n | X).$$

A proof of this proposition can be found in Appendix A.

Algorithm 3 PX-EM Algorithm with Approximate Likelihood

while $\ln f(X|\theta_{n-1}) + \ln \pi(\theta_{n-1}) - (\ln f(X|\theta_{n-1}) + \ln \pi(\theta_{n-1}))$ > tolerance level **do**

E Step:

Run a Kalman smoother to obtain $\hat{F}$, $\hat{G}$ $\hat{P}$, $\hat{C}$, and $\hat{P}_g$.

for $1 \leq i \leq N$ **do**

$\hat{\gamma}_i = (1 + \frac{\rho_{i,n-1}}{1-\rho_{i,n-1}} \frac{\alpha_1}{\alpha_0} \exp((\alpha_0 - \alpha_1)\sigma^2_{i,n-1}))^{-1}$.

end for

M Step:

$\Lambda^*_n = X' \hat{F} (\hat{F}' \hat{F} + \hat{P})^{-1}$

for $1 \leq i \leq N$ **do**

$SS_i = \sum_t (X_{it} - \Lambda^*_{i,n} \hat{f}_t)^2 + \Lambda^*_{i,n} \hat{P} \Lambda^{*\prime}_{i,n}$

$\alpha^*_i = (1 - \hat{\gamma}_i)\alpha_0 + \hat{\gamma}_i \alpha_1$

$\sigma^2_{i,n} = \frac{SS_i}{\frac{1}{2}(T + \sqrt{T^2 + 2\alpha^*_i SS_i})}$

$\rho_{i,n} = \frac{\hat{\gamma}_i + 1 - a}{2a - 1}$

end for

$\Phi^*_n = (\Phi^*_{1,n}, \ldots, \Phi^*_{p,n}) = (\hat{F}' \hat{G} + \hat{C})(\hat{G}' \hat{G} + \hat{P}_g)^{-1}$

$\Omega^*_n = T^{-1}(\sum_t (\hat{f}_t - \Phi^* \hat{g}_t)(\hat{f}_t - \Phi^* \hat{g}_t)' + \Phi^* \hat{P}_g \Phi^{*\prime} - \Phi^* \hat{C}' - \hat{C} \Phi^{*\prime})$

Rotation Step:

Calculate A^L_n, the lower Cholesky factor of Ω^*_n.

$\Lambda_n = \Lambda^*_n A^L_n$

for $1 \leq l \leq p$ **do**

$\Phi_{l,n} = {A^L_n}^{-1} \Phi^*_{l,n} A^L_n$

end for

end while

The benefit of this approach is that it only requires one matrix inversion and one Kalman smoother run, as opposed to many matrix inversions and Kalman filter runs. One only has to work with the conditional elements of the likelihood, so the gradient is available in closed form. This result is also applicable to stationary VARs and vector autoregressive moving average (VARMA) models. It could be used for maximizing the likelihoods of these models

or for efficient simulation from the posterior distribution using Hamiltonian Monte Carlo, which is an area of active research (Heaps, 2023; Binks et al., 2023).

3.4 Specification of α_q

α_1 should be set so as to have minimal influence on variance estimates. I adopt the convention of $\alpha_1 = 0.01$ in all estimations. Optimal specification of α_0 is not obvious. Values that are too small will not impose sufficient shrinkage on small variances. However, setting α_0 too high means that the EM algorithm is unlikely to assign significant weight to the spike component of the prior, and the resulting estimates will be close to the maximum likelihood estimates. Rather than try to find a single optimal α_0, I adopt the dynamic posterior exploration approach developed by Ročková and George (2016). The authors, drawing on concepts from deterministic annealing, estimate a series of models with increasingly pronounced spike distributions. This is done by using a ladder of increasing spike parameters $\alpha_0 \in I = \{\alpha_0^1, \alpha_0^2, \ldots, \alpha_0^L\}$. α_1 is held constant. Small values of α_0 produce a flatter posterior density that is easier to traverse. As α_0 increases, the posterior becomes spikier. Each estimation is initialized with the MAP estimates from the previous estimation. This "warm start" approach makes the global mode easier to find. The intersection point of the spike and slab densities is given by $\delta(\alpha_1, \alpha_0, \rho) = \frac{1}{\alpha_0 - \alpha_1} \ln\left(\frac{\alpha_0}{\alpha_1}\frac{1-\rho}{\rho}\right)$. The sequence I is defined implicitly by the sequence $\delta(\alpha_1, \alpha_0, \rho = .5) \in I_\delta = \{\delta^1, \delta^2, \ldots, \delta^L\}$. I use the sequence $I_\delta = \{.5, .25, .1, .05, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ for all estimations.

3.5 Missing data

It is rare for a researcher to be blessed with a balanced panel of data. Very often variables are not available for the entire sample period. It may be that they were only recorded after a certain date or were eventually discontinued. It could also be the case that certain entries are intentionally trimmed by the researcher to control for outliers. Missing data represent a major issue in the frequentist literature when factors are estimated with PCA. Missing

values must be imputed with a consistent estimator (Stock and Watson, 2002; Jin, Miao, and Su, 2021). Likelihood-based method do not have the same problem. One only has to restrict the vector of dependent variables in each period to those with non-missing values.

I adjust the variance selection prior to account for differing sample sizes in the presence of missing data. Using the same α_0 and α_1 for every time series would disproportionately penalize the variances of variables with many missing values. Let T_i be the number of time periods for which X_{it} is observed. The variable-specific hyperparameters are then defined as $\alpha_{iq} \equiv \frac{T_i}{T}\alpha_q$. The correction term $\frac{T_i}{T}$ ensures that, conditional on γ_i, the prior has the same influence on each σ_i^2.

Algorithm 4 PX-EM Algorithm with Approximate Likelihood and Missing Data

while $\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}) - (\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}))$ > tolerance level **do**

E Step:

Run a Kalman smoother to obtain $\{\hat{F}_i\}$, $\hat{G}$, $\{\hat{P}_i\}$, $\hat{C}$, and $\hat{P}_g$.

for $1 \leq i \leq N$ **do**

$\hat{\gamma}_i = (1 + \frac{\rho_{i,n-1}}{1-\rho_{i,n-1}}\frac{\alpha_{i1}}{\alpha_{i0}}\exp((\alpha_{i0} - \alpha_{i1})\sigma^2_{i,n-1}))^{-1}$.

end for

M Step:

for $1 \leq i \leq N$ **do**

$\Lambda^*_{i,n} = X_i'\hat{F}_i(\hat{F}_i'\hat{F}_i + \hat{P}_i)^{-1}$

$SS_i = \sum_{t\in O_i}(X_{it} - \Lambda^*_{i,n}\hat{f}_t)^2 + \Lambda^*_{i,n}\hat{P}_i{\Lambda^*}'_{i,n}$

$\alpha^*_i = (1-\hat{\gamma}_i)\alpha_{i0} + \hat{\gamma}_i\alpha_{i1}$

$\sigma^2_{i,n} = \frac{SS_i}{\frac{1}{2}(T_i + \sqrt{T_i^2 + 2\alpha^*_i SS_i})}$

$\rho_{i,n} = \frac{\hat{\gamma}_i + 1 - a}{2a-1}$

end for

$\Phi^*_n = (\Phi^*_{1,n}, \ldots, \Phi^*_{p,n}) = (\hat{F}'\hat{G} + \hat{C})(\hat{G}'\hat{G} + \hat{P}_g)^{-1}$

$\Omega^*_n = T^{-1}(\sum_t(\hat{f}_t - \Phi^*\hat{g}_t)(\hat{f}_t - \Phi^*\hat{g}_t)' + \Phi^*\hat{P}_g\Phi^{*\prime} - \Phi^*\hat{C}' - \hat{C}\Phi^{*\prime})$

Rotation Step:

Calculate A^L_n, the lower Cholesky factor of Ω^*_n.

$\Lambda_n = \Lambda^*_n A^L_n$

for $1 \leq l \leq p$ **do**

$\Phi_{l,n} = A^{L^{-1}}_n \Phi^*_{l,n} A^L_n$

end for

end while

Optimization can proceed with only minor modifications to the algorithms. Posterior

moments are obtained using a Kalman smoother adjusted for missing data. Let the set of time periods O_i be defined as $O_i \equiv \{t : X_{it} \text{ is observed}\}$. Let τ_{im} be the m^{th} entry in O_i. We now define $\hat{F}_i \equiv (\hat{f}_{\tau_{i1}}, \ldots, \hat{f}_{\tau_{iT_i}})'$, $\hat{P}_i \equiv \sum_{t \in O_i} \mathbb{E}[(f_t - \hat{f}_t)(f_t - \hat{f}_t)' | X, \theta_n]$. Algorithm 4 gives the details for a modified PX-EM algorithm.

4 Monte Carlo Studies

This section presents the results of various Monte Carlo studies. In each simulation, the loadings are randomly generated according to $\Lambda_{ij} \sim N(0, 1)$. I consider the case of $p = 1$ lags. Φ is randomly generated using its eigendecomposition $\Phi = VDV^{-1}$. The elements of the eigenvector matrix are distributed $V_{jj'} \sim \mathcal{U}(-1, 1)$ and the eigenvalues are distributed $D_{jj} \sim \mathcal{U}(.4, .6)$. $\Omega = \omega I_r$. ω is chosen such that $\text{Var}(\Lambda_i f_t) = r$.[3]I first examine datasets with every possible combination of N, T, and r for which $N \in \{40, 60, 100, 200\}$, $T \in \{50, 100, 150, 200, 250\}$, and $r \in \{2, 4, 6\}$. In each case, the number of observed and latent factors are equal: $r_f = r_y = r/2$. The factors are simulated by drawing the first p instances from the stationary distribution and then iterating the data generating process forward through time. Each study analyzes 100 simulated datasets.

The first simulation study assumes a balanced panel and $\sigma_i^2 = r$. The results can be seen in Figure 2. The plots give the proportion of datasets for which the procedure correctly identified the true observed factors. The proposed approach is quite good at identifying the observed factors in most cases. The one noticeable limitation is that results suffer when N and T are small and r is large. This is hardly surprising, as we are asking a lot of the model and not providing sufficient data. Thankfully, the success rate is quite high for combinations of N and T that we are likely to encounter in practice.

3. This is done by first calculating the stationary covariance matrix P_0 of a process with transition parameters Φ and covariance matrix $\Omega = I_r$. Let C be the lower Cholesky factor of P_0 such that $P_0 = CC'$. The covariance matrix of innovations is then rescaled to $\Omega = \frac{r}{||C||_2^2} I_r$.

Figure 2: Proportion of Models Correctly Identified, $\sigma_i^2 = r$

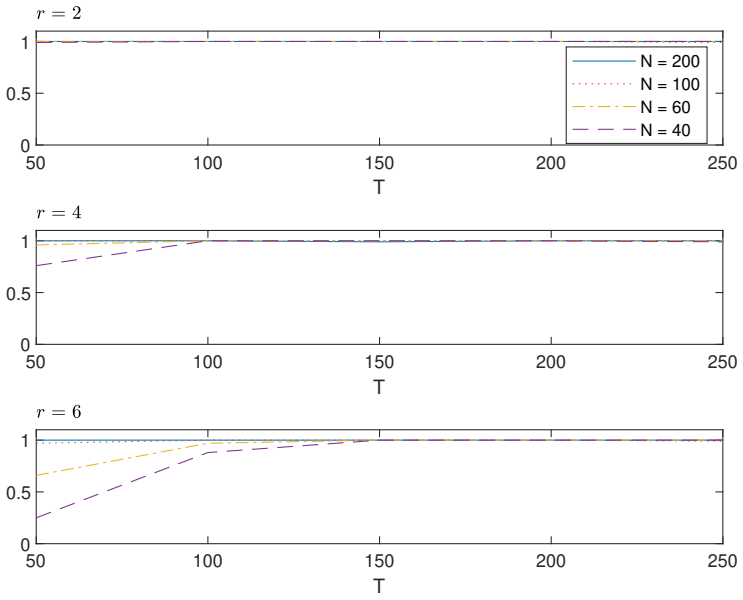


Figure 3: Proportion of Models Correctly Identified, $\sigma_i^2 = 2r$

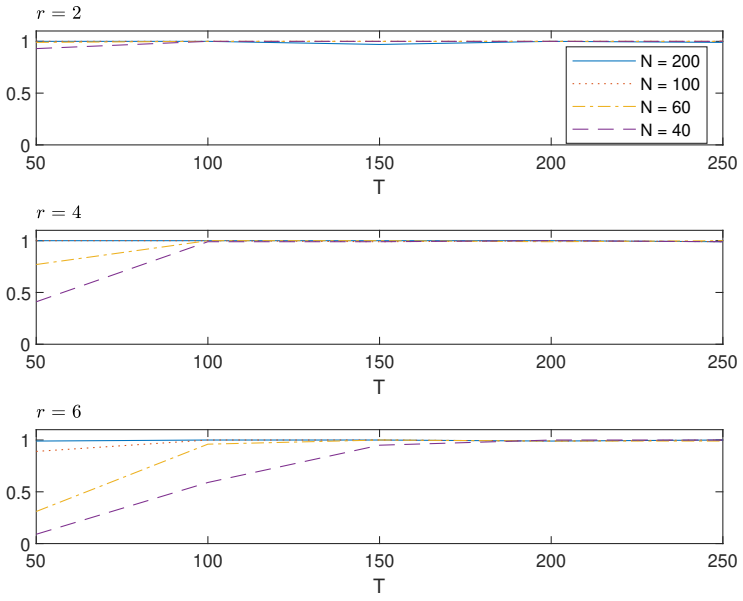


Figure 3 gives the results of a simulation study that uses the exact same parameters and factors as Table Figure 2, only the idiosyncratic variance is now set to $\sigma_i^2 = 2r$. We observe a slight decrease in accuracy for small values of N and T. This is to be expected because

the signal-to-noise ratio has decreased and the factors will not be estimated as precisely. However, we see no noticeable drop in accuracy for $N \geq 60$ and $T \geq 100$.

Figure 4: Proportion of Models Correctly Identified, $\sigma_i^2 = r$, $p_{miss} = 0.05$

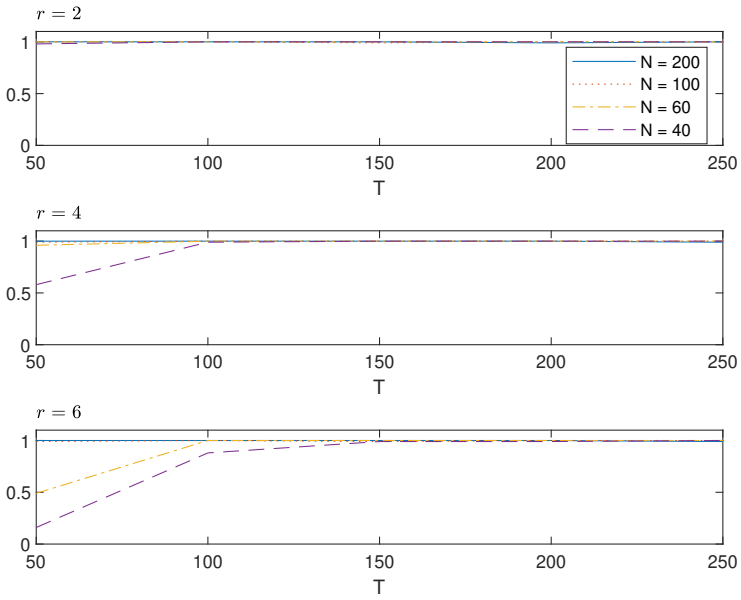
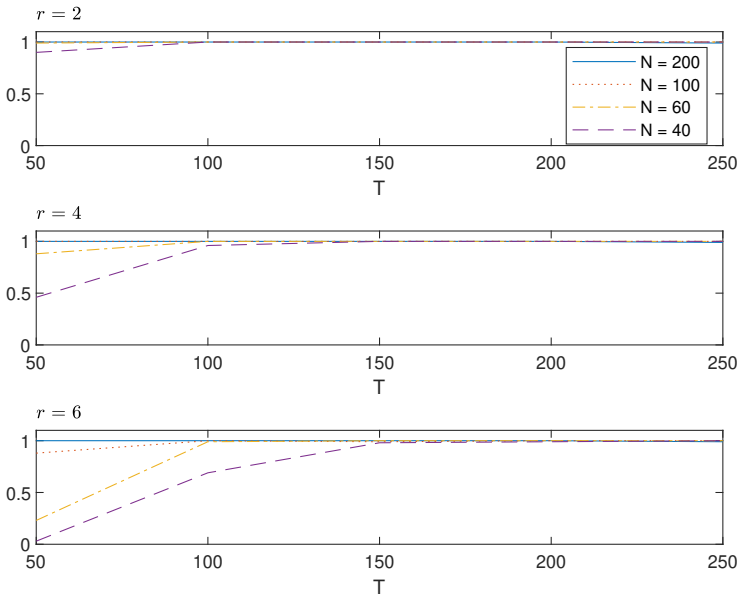


Figure 5: Proportion of Models Correctly Identified, $\sigma_i^2 = r$, $p_{miss} = 0.1$



Figures 4 and 5 give the results of Monte Carlo studies in which $\sigma_i^2 = r$ and a proportion

p_{miss} of the data is missing. The values considered are $p_{miss} = 0.05, 0.1$. There appears to be no substantive difference between the results with missing data and the results with a balanced panel for $N \geq 60$ and $T \geq 100$.

5 Applications

5.1 Quarterly Macroeconomic Data

This section applies the model selection procedure developed above to the FRED-QD dataset (McCracken and Ng, 2020). For the first application, the dataset consists of $N = 246$ macroeconomic variables observed at quarterly intervals. Observations begin in 1959:Q3 and end in 2023:Q1. 38 of the variables were not recorded until midway through the sample period. Variables were transformed to be approximately stationary using the recommended transformation codes of the authors, and then standardized to have unit variance. I performed outlier detection using the same criterion as McCracken and Ng (2020). Any observations that deviated from the sample median by more than ten interquartile ranges were classified as outliers and treated as missing. Initial factor estimates were obtained by replacing missing values with 0 and then using PCA. Jin, Miao, and Su (2021) show that this is a consistent estimator of the true factor space. I analyzed the full sample period as well as the subsamples 1959:Q3 - 2007:Q3 and 2007:Q4 - 2023:Q1. The sample was partitioned to examine any structural changes that may have occurred after the 2007 financial crisis. The outlier classification criterion detected 90 outliers in the full sample, 5 outliers in the pre-financial crisis subsample, and 109 outliers in the post-financial crisis subsample. The IC_{p2} criterion selected 8 factors for the full sample period and 6 factors for each of the subsamples.

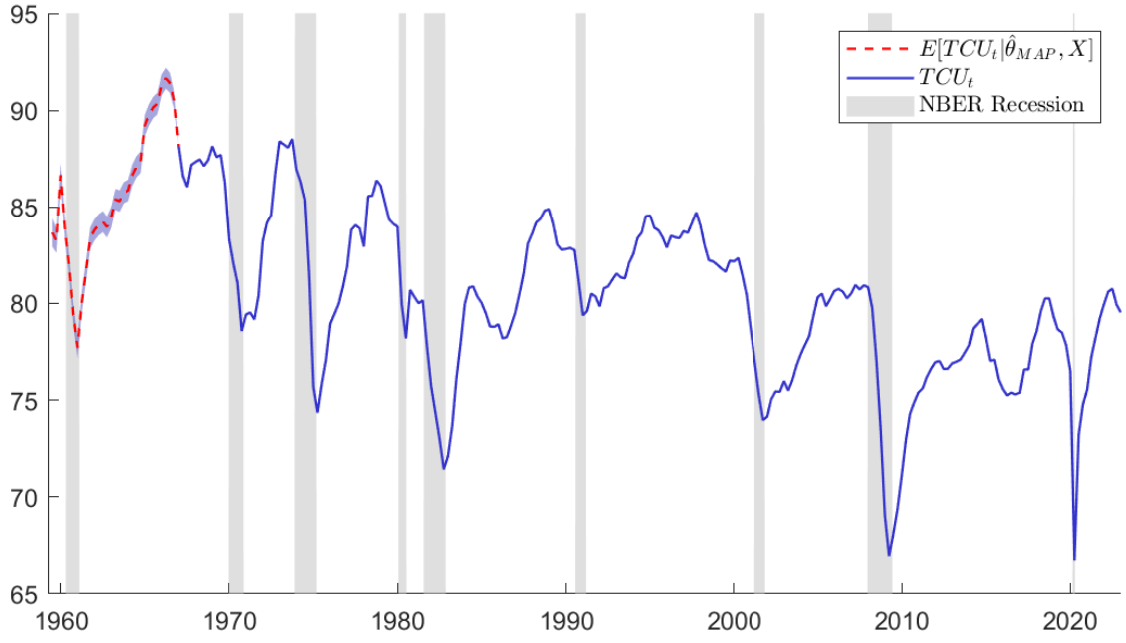
As can be seen from Table 1, Capacity Utilization: Total Industry (TCU) is selected as an observed factor for both the full sample and pre-2007 estimations. TCU is an index that measures the percentage of potential feasible output that is being produced. This is a surprising but not unreasonable finding. Capacity utilization has long been recognized as a

Table 1: Likely Observed Factors in the U.S. Economy, Quarterly Data

Period	N	T	r	y
1959:Q3-2023:Q1	246	255	8	Capacity Utilization: Total Industry
1959:Q3-2007:Q3	246	193	6	Capacity Utilization: Total Industry
2007:Q4-2023:Q1	246	62	6	Business Sector: Real Output All Employees: Service-Providing Industries All Employees: Goods-Producing Industries

leading indicator for inflation and business cycles (Corrado and Mattey, 1997). That TCU was selected demonstrates the necessity of being able to incorporate missing data. TCU was not recorded until 1967:Q1. The existing frequentist methods require a balanced panel dataset, and thus would not have been able to detect this relationship over the periods considered. Another advantage of the Bayesian approach is that unobserved values of observed factors can be imputed naturally using the output from the Kalman smoother.

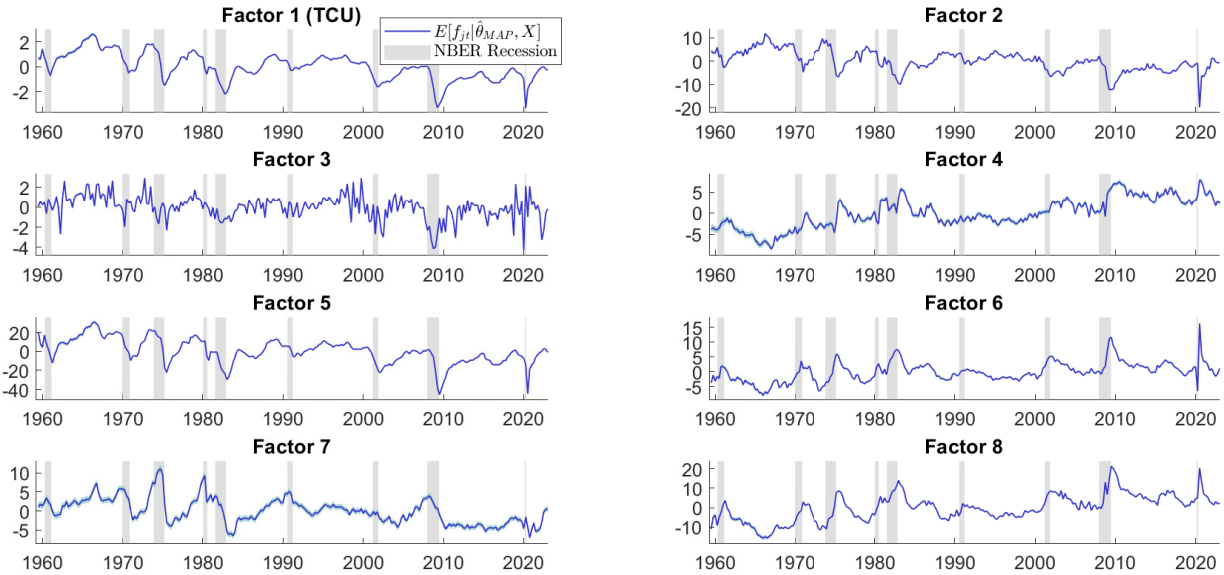
Figure 6: Total Capacity Utilization (Observed and Imputed)



Notes: The shaded region around the imputed values of TCU_t is a 95% credible interval. The variance of TCU_t is available directly from the Kalman smoother.

Figure 6 shows the precise estimates that are obtained for the period 1959:Q3 - 1968:Q4 using this method. TCU is not selected in the post-2007 estimation. With this in mind, the behavior of TCU does seem to be different before and after the financial crisis. Capacity utilization tends to peak in the middle of expansions prior to 2007. The index is already declining prior to the onset of recessions during this period. The inter-recession shape of TCU appears different after 2007. It is approximately level during 2007 and only starts to decline after the 2008 recession has already begun. Estimates of all 8 factors from the full sample estimation are plotted in Figure 7. One can see that estimates of factor 5 are nearly identical to a one period lag of TCU. This suggests that the true Ω might be of reduced rank (Bai and Ng, 2007). It also indicates that TCU is not only an important driver of the economy, but its impact is also persistent.

Figure 7: Factor Estimates from FRED-QD



Notes: The shaded region around the estimated values of f_{jt} is a 95% credible interval. The variance of f_{jt} is available directly from the Kalman smoother.

5.2 Monthly Macroeconomic Data

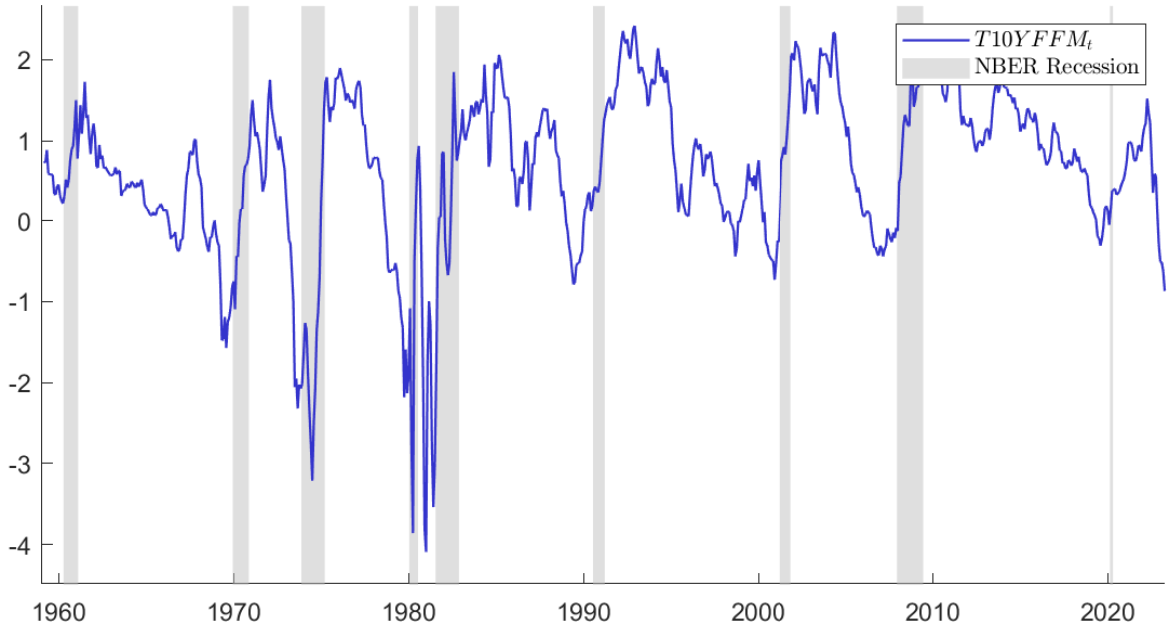
This sections analyzes the FRED-MD dataset (McCracken and Ng, 2016). It consists of $N = 127$ monthly macroeconomic variables over the period 1959:3-2023:6. The variables are transformed using the authors' recommended transformations and standardized to have unit variance. Outliers are identified and removed using the criterion previously discussed. As with the quarterly data, I analyze the full sample, as well as pre- and post-financial crisis subsamples. Results are given in Table 2.

Table 2: Likely Observed Factors in the U.S. Economy, Monthly Data

Period	N	T	r	y
1959:3-2023:6	127	772	7	10-Year Treasury Constant Maturity Minus Federal Funds Rate
1959:3-2007:9	127	583	7	Moody's Seasoned Baa Corporate Bond Minus Federal Funds Rate
2007:10-2023:3	127	189	7	All Employees, Total Nonfarm All Employees: Service-Providing Industries Consumer Price Index for All Urban Consumers: All Items in U.S. City Average S&P 500

The only variable identified as an observed factor in the full sample estimation is 10-Year Treasury Constant Maturity Minus Federal Funds Rate (T10YFFM). This is very similar to measures of the slope of the yield curve, which has been studied for its relationship to business cycles. Figure 8 plots T10YFFM along with NBER recession dates. We can see that T10YFFM often turns negative near the peak of an expansion and then sharply increases during recessions. The pre-financial crisis estimation selects a related variable: Moody's Seasoned Baa Corporate Bond Minus Federal Funds Rate (BAAFFM). The correlation between T10YFFM and BAAFFM during this period is 0.94, so the relationship between BAAFFM and business cycles is nearly identical. The correlation between the two variables drops to 0.89 in the post-2007 subsample. One possible reason that T10YFFM is selected in the full sample estimation is that the Federal Reserve began targeting the yield curve directly after

Figure 8: 10-Year Treasury Constant Maturity Minus Federal Funds Rate



the financial crisis.

While estimations using monthly and quarterly data produce differing results in the pre-financial crisis subsamples, there is some overlap in the selected observed factors for the post-financial crisis subsamples. They both select All Employees: Service-Providing Industries, along with one other employment measure. The importance of service sector employment may stem from its correlation with the impacts of the COVID-19 pandemic.

5.3 Fama-French Portfolio Data

I will now use the model selection process to investigate the asset-pricing model of Fama and French (1993). The authors extend the capital asset pricing model to include factors that measure the excess returns attributable to firm size and book-to-market equity ratio (BE/ME). The Fama-French three-factor model is given by

$$X_{it} = R_{it} - R_t^f = \beta_0 + \beta_{1i}(R_t^m - R_t^f) + \beta_{2i}SMB_t + \beta_{3i}HML_t + \varepsilon_{it}, \tag{18}$$

where R_{it} is the return on portfolio i, R_t^m is the return on a market portfolio, R_t^f is the risk-free return, SMB_t is the firm size factor, and HML_t is the BE/ME factor. I estimate models for a dataset that includes the three Fama-French factors, their measure of the risk-free rate, and the excess returns from 100 portfolios. The portfolios are the intersection of 10 portfolios organized by deciles of firm size and 10 portfolios organized by deciles of BE/ME. The data was collected from Kenneth French's website [4]. Estimating such a model allows us to test whether the 3 factor specification is supported by the data. Factor observations and incomplete portfolio data are available for the period 1926:7-2023:6. I estimated a model for the full sample period as well as a number of subsamples. The subsamples include the time periods considered by Bai and Ng (2006) as well several others. The time periods not previously examined are the interwar period of 1926:7-1945:8, the Bretton Woods period of 1945:9-1972:12, the pre-financial crisis period of 1997:1-2007:9, and the post-financial crisis period of 2007:10-2023:6. Previous studies only examined the validity of the three-factor model after 1960 and did not include all 100 portfolios. Researchers had to delete several portfolios as well as many time periods because their methods required a balanced panel. Results are given in Table 3.

The importance of BE/ME is quite stable. It is selected as an observed factor in the full sample estimation as well as every subsample estimation except for 1973:1-1987:12. Firm size is selected in the full sample estimation, but not in the subsamples for 1973:1-1987:12, 1988:1-1996:12, 1997:1-2007:9, and 1960:1-1996:12. A surprising result is the selection of the portfolio of firms in the tenth deciles of size and BE/ME in 2 subsamples. However, this result should be treated skeptically because there are very few observations of the variable in these subperiods, so there is a good chance of overfitting. The most glaring result is that market excess return is not selected in any estimation. Although the market variable is not selected, we should not interpret this as evidence that market return plays no role in portfolio returns. The estimated variance of the idiosyncratic error for the market variable

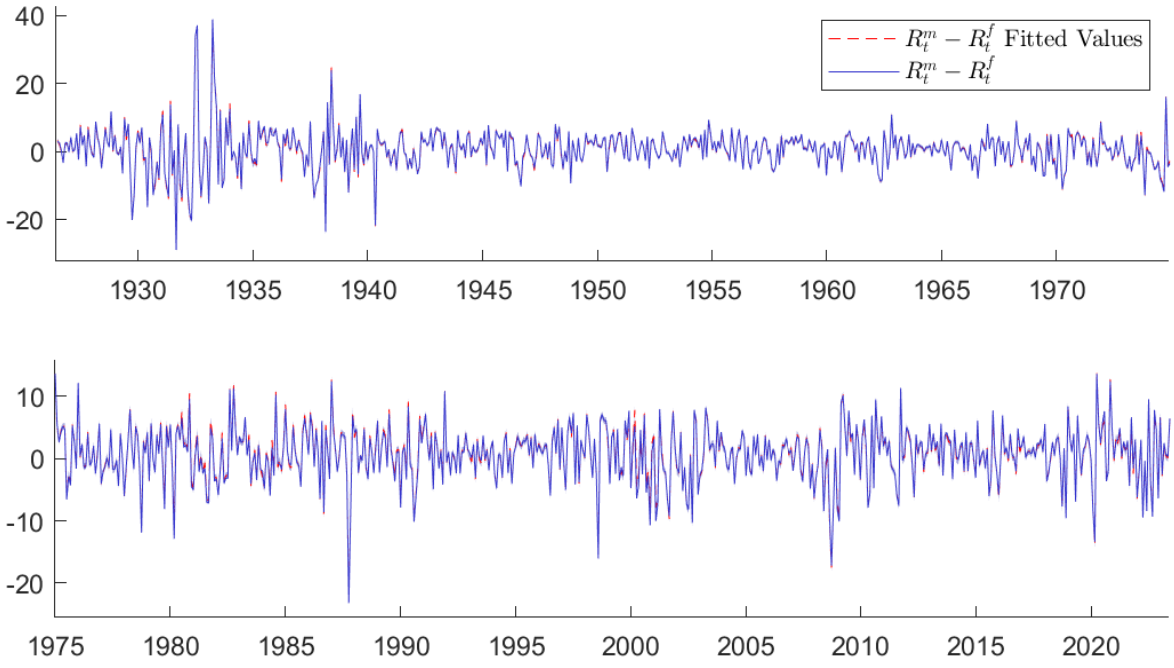
4. See `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html` for further information.

Table 3: Likely Observed Factors in Monthly Fama-French Portfolios

Period	N	T	r	y
1926:7-2023:6	104	1,164	4	Firm Size Book-to-Market Equity Ratio
1926:7-1945:8	104	230	4	Firm Size Book-to-Market Equity Ratio Portfolio of firms in the tenth deciles of size and BE/ME
1945:9-1972:12	104	328	3	Firm Size Book-to-Market Equity Ratio
1973:1-1987:12	104	170	4	Risk-Free Return
1988:1-1996:12	104	108	3	Book-to-Market Equity Ratio
1997:1-2007:9	104	109	5	Book-to-Market Equity Ratio Portfolio of firms in the tenth deciles of size and BE/ME
2007:10-2023:6	104	189	4	Firm Size Book-to-Market Equity Ratio
1960:1-1996:12	104	444	4	Book-to-Market Equity Ratio
1960:1-1982:12	104	276	4	Firm Size Book-to-Market Equity Ratio Risk-Free Return
1982:1-1996:12	104	168	3	Firm Size Book-to-Market Equity Ratio

is less than 0.01 in all but two of the estimations. This suggests that excess market return or some closely related variable is a fundamental factor, but it is not perfectly observed. Figure 9 plots the market variable over the entire sample period along with its fitted values. We can see that there is very little difference between the two. A more surprising result is the occasional selection of the risk-free return as an observed factor. This suggests that excess returns depend on R_t^f in a way that is not simply a function of their dependence on excess market return. It is important to note that the two time periods in which R_t^f is selected include the period in the 1970s and early 1980s when interests rates were extremely volatile.

Figure 9: Actual and Fitted Values of $R_t^m - R_t^f$



Notes: The shaded region around the fitted values of $R_t^m - R_t^f$ is a 95% credible interval. The variances of the fitted values are available directly from the Kalman smoother.

6 Conclusion

I proposed a model selection procedure for FAVARs. Estimation of the total number of factors and the lag length is done using existing methods, although the use of BIC for lag length selection is modified to avoid model misspecification problems. The selection of observed factors is achieved using a Bayesian shrinkage prior. The prior allows us to recast a high-dimensional model selection process as an optimization problem. This enables researchers to differentiate between millions of potential models by estimating just a single model. The procedure has very good small sample properties. Model selection accuracy was virtually 100% in simulated datasets of realistic size.

Several extensions to the EM algorithm for estimating DFMs were proposed. The resulting PX-ECME algorithm exhibited faster convergence properties than the basic EM

algorithm. I also developed an efficient and precise method for calculating the gradient of the log-likelihood of stationary VARMA processes, of which the FAVAR is a special case.

The model selection procedure yielded interesting results when applied to macroeconomic and financial data. The Total Capacity Utilization index was the only observed factor detected in a large dataset of quarterly U.S. macroeconomic variables. The spread between the 10-Year Treasury Constant Maturity Rate and the Federal Funds Rate was the only observed factor detected for monthly data. A specification in which the Federal Funds Rate is the only observed factor, the default assumption in the FAVAR literature, was never selected. Finally, I used the model selection procedure to test the assumptions of the Fama and French (1993) asset-pricing model. The variables constructed for firm size and book-to-market equity ratio were often selected as observed factors, but excess market return was not. That excess market return was not selected is more likely the result of mismeasurement of the variable rather than lack of importance.

There are many avenues for further research. While the approach of this paper seeks to find the most likely model, it may be the case that there are multiple competing models with significant posterior probabilities. MCMC would be the appropriate means of estimation for this end. This model assumes homoskedastic Normal errors, which is unlikely to be realistic in macroeconomic and financial data. One could incorporate errors with stochastic volatility into the state equation as well as the observation equations. Allowing for stochastic volatility in the observation equations permits the possibility of the observed factors changing over time, which is a perfectly reasonable hypothesis.

Appendix A Proof of Proposition 1

A different route to efficient evaluation of the gradient can be seen by noting that the integrated likelihood for a DFM is equivalent to that of a DFM that also includes presample instances of the state variable. If we let $F^{\dagger} = (f_0, f_{-1}, \ldots, f_{-\tau+1})$, the likelihood can then be expressed as

$$f(X|\theta) = \int f(X|F,\theta)\pi(F|\theta)dF = \int\int f(X|F,\theta)\pi(F|F^{\dagger},\theta)\pi(F^{\dagger}|\theta)dFdF^{\dagger}. \tag{19}$$

Since the model with presample factors is also valid, it is also amenable to the construction of an EM algorithm. Define $Q_{\tau}(\theta|\theta_n) \equiv \mathbb{E}[\ln f(X, F, F^{\dagger}|\theta)]$. Using the same result from Ruud (1991), we know that

$$\nabla \ln f(X|\theta_n) = \nabla Q_{\tau}(\theta_n|\theta_n). \tag{20}$$

I will now show that for τ sufficiently large, we can calculate the gradient using only the conditional terms in $Q_{\tau}(\theta_n|\theta_n)$ and omit any terms that involve the stationary distribution of the factors.

Proposition 1

Let $F^{\dagger} = (f_0, f_{-1}, \ldots, f_{-\tau+1})$, $Q_{\tau-p}(\theta|\theta_n) \equiv \mathbb{E}[\ln f(X, F, f_0, f_{-1}, \ldots, f_{-\tau+p+1}|$ $f_{-\tau+p}, \ldots, f_{-\tau+1}, \theta)|X, \theta_n]$ and assume θ_n is an interior point of the parameter space.

$$\lim_{\tau\to\infty} \nabla Q_{\tau-p}(\theta_n|\theta_n) = \nabla \ln f(X|\theta_n).$$

Proof. As $\tau \to \infty$, $\nabla Q_{\tau}(\theta_n|\theta_n)$ becomes an infinite sum. Since $\nabla \ln f(X|\theta_n) = \nabla Q_{\tau}(\theta_n|\theta_n)$, we know that this sum must converge to the desired gradient. All that remains is to show that the terms involving the stationary distribution go to 0. As the Kalman smoother is iterated backwards, the smoothed moments of the factors will converge to the stationary moments: $\mathbb{E}[g_t|X, \theta_n] \to \mathbb{E}[g_t|\theta_n] = 0$, $\mathbb{E}[(g_t - \hat{g}_t)(g_t - \hat{g}_t)'|X, \theta_n] \to P_0$. By Gibb's Inequality, $\mathbb{E}[\nabla \ln \pi(g_t|\theta_n)|\theta_n] = 0$ for any θ_n in the interior of the parameter space. We can thus conclude

that

$$\begin{aligned}\lim_{\tau\to\infty} \nabla Q_\tau(\theta_n|\theta_n) - \nabla Q_{\tau-p}(\theta_n|\theta_n) &= \lim_{\tau\to\infty} \mathbb{E}[\nabla \ln\pi(g_{-\tau+p}|\theta_n)|X,\theta_n] \\ &= \mathbb{E}[\nabla \ln\pi(g_t|\theta_n)|\theta_n] \\ &= 0.\end{aligned}$$

□

For an accurate calculation of the gradient, τ should be chosen so that the smoothed moments converge to the stationary moments. This will obviously depend on the persistence of shocks in the model. For highly persistent models, simulation results suggest that $\tau = 5,000$ is sufficiently large.

References

Ahn, Seung C., and Alex R. Horenstein. 2013. "Eigenvalue Ratio Test for the Number of Factors." *Econometrica* 81 (3): 1203–1227.

Bai, Jushan, and Serena Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica* 70 (1): 191–221.

———. 2006. "Evaluating latent and observed factors in macroeconomics and finance." *Journal of Econometrics* 131 (1): 507–537.

———. 2007. "Determining the Number of Primitive Shocks in Factor Models." *Journal of Business & Economic Statistics* 25 (1): 52–60.

Belviso, Francesco, and Fabio Milani. 2006. "Structural factor-augmented VARs (SFAVARs) and the effects of monetary policy." *Topics in Macroeconomics* 6 (3).

Bernanke, Ben S., Jean Boivin, and Piotr Eliasz. 2005. "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach." *The Quarterly Journal of Economics* 120, no. 1 (February): 387–422.

Beyeler, Simon, and Sylvia Kaufmann. 2021. "Reduced-form factor augmented VAR—Exploiting sparsity to include meaningful factors." *Journal of Applied Econometrics* 36 (7): 989–1012.

Binks, Rachel L., Sarah E. Heaps, Mariella Panagiotopoulou, Yujiang Wang, and Darren J. Wilkinson. 2023. *Bayesian inference on the order of stationary vector autoregressions.* arXiv: 2307.05708 [stat.ME].

Boivin, Jean, Marc P Giannoni, and Dalibor Stevanović. 2013. "Dynamic effects of credit shocks in a data-rich environment." *FRB of New York Staff Report,* no. 615.

Carvalho, Carlos M., Jeffrey Chang, Joseph E. Lucas, Joseph R. Nevins, Quanli Wang, and Mike West. 2008. "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics." *Journal of the American Statistical Association* 103 (484): 1438–1456.

Chen, Nai-Fu, Richard Roll, and Stephen A. Ross. 1986. "Economic Forces and the Stock Market." *The Journal of Business* 59 (3): 383–403.

Cochrane, John H. 2011. "Presidential Address: Discount Rates." *The Journal of Finance* 66 (4): 1047–1108.

Corrado, Carol, and Joe Mattey. 1997. "Capacity Utilization." *Journal of Economic Perspectives* 11, no. 1 (March): 151–167.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data Via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–22.

Fama, Eugene F., and Kenneth R. French. 1993. "Common risk factors in the returns on stocks and bonds." *Journal of financial economics* 33 (1): 3–56.

———. 2015. "A five-factor asset pricing model." *Journal of Financial Economics* 116 (1): 1–22.

Fernald, John G, Mark M Spiegel, and Eric T Swanson. 2014. "Monetary policy effectiveness in China: Evidence from a FAVAR model." *Journal of International Money and Finance* 49:83–103.

Frühwirth-Schnatter, Sylvia, and Hedibert Freitas Lopes. 2009. *Parsimonious Bayesian Factor Analysis When the Number of Factors is Unknown.* Technical report. University of Chicago Booth School of Business.

Hallin, Marc, and Roman Liška. 2007. "Determining the Number of Factors in the General Dynamic Factor Model." *Journal of the American Statistical Association* 102 (478): 603–617.

Heaps, Sarah E. 2023. "Enforcing Stationarity through the Prior in Vector Autoregressions." *Journal of Computational and Graphical Statistics* 32 (1): 74–83.

Jin, Sainan, Ke Miao, and Liangjun Su. 2021. "On factor models with random missing: EM estimation, inference, and cross validation." *Journal of Econometrics* 222 (1, Part C): 745–777.

Knowles, David, and Zoubin Ghahramani. 2011. "Nonparametric Bayesian sparse factor models with application to gene expression modeling." *The Annals of Applied Statistics* 5 (2B): 1534–1552.

Liu, Chuanhai, and Donald B. Rubin. 1994. "The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence." *Biometrika* 81 (4): 633–648. Accessed August 8, 2023.

Liu, Chuanhai, Donald B. Rubin, and Ying Nian Wu. 1998. "Parameter Expansion to Accelerate EM: The PX-EM Algorithm." *Biometrika* 85 (4): 755–770.

Luo, Jiayi, and Cindy Long Yu. 2021. "Determining Number of Factors in Dynamic Factor Models Contributing to GDP Nowcasting." *Mathematics* 9 (22).

McAlinn, Kenichiro, Veronika Ročková, and Enakshi Saha. 2018. *Dynamic Sparse Factor Analysis.* arXiv: 1812.04187.

McCracken, Michael, and Serena Ng. 2016. "FRED-MD: A Monthly Database for Macroeconomic Research." *Journal of Business & Economic Statistics* 34 (4): 574–589.

———. 2020. *FRED-QD: A quarterly database for macroeconomic research.* Technical report. National Bureau of Economic Research.

Paccagnini, Alessia. 2017. *Forecasting with FAVAR: macroeconomic versus financial factors.* NBP Working Papers 256. Narodowy Bank Polski.

Parker, Jason, and Donggyu Sul. 2016. "Identification of Unknown Common Factors: Leaders and Followers." *Journal of Business & Economic Statistics* 34 (2): 227–239.

Ročková, Veronika, and Edward I. George. 2016. "Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity." *Journal of the American Statistical Association* 111 (516): 1608–1622.

Ruud, Paul A. 1991. "Extensions of estimation methods using the EM algorithm." *Journal of Econometrics* 49 (3): 305–341.

Stock, James H, and Mark W Watson. 2002. "Macroeconomic Forecasting Using Diffusion Indexes." *Journal of Business & Economic Statistics* 20 (2): 147–162.

———. 2016. "Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics." Chap. Chapter 8, 2:415–525. Elsevier.

Tracy, Kevin. 2022. *A Square-Root Kalman Filter Using Only QR Decompositions.* arXiv: 2208.06452.

Watson, Mark W., and Robert F. Engle. 1983. "Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models." *Journal of Econometrics* 23 (3): 385–400.

Witten, Daniela M., Robert Tibshirani, and Trevor Hastie. 2009. "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis." *Biostatistics* 10, no. 3 (April): 515–534.

Wu, C. F. Jeff. 1983. "On the Convergence Properties of the EM Algorithm." *The Annals of Statistics* 11 (1): 95–103.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. "Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics* 15 (2): 265–286.