Australian
National
University

Crawford School of Public Policy

CAMA

Centre for Applied Macroeconomic Analysis

# Financial Condition Indices in an Incomplete Data Environment

CAMA Working Paper 42/2023
August 2023

**Miguel Herculano**
University of Glasgow

**Punnoose Jacob**
Reserve Bank of New Zealand
Centre for Applied Macroeconomic Analysis, ANU

## Abstract

We construct a Financial Conditions Index (FCI) for the United States using a dataset that features many missing observations. The novel combination of probabilistic principal component techniques and a Bayesian factor-augmented VAR model resolves the challenges posed by data points being unavailable within a high-frequency dataset. Even with up to 62% of the data missing, the new approach yields a less noisy FCI that tracks the movement of 22 underlying financial variables more accurately both in-sample and out-of-sample.

**Keywords**

Financial Conditions Index, Mixed-Frequency, Bayesian Methods

**JEL Classification**

C11, C32, C52, C53, C66

**Address for correspondence:**

(E) cama.admin@anu.edu.au

**ISSN 2206-0332**

# *Financial condition indices in an incomplete data environment*[*]

Miguel Herculano
University of Glasgow
miguel.herculano@glasgow.ac.uk

Punnoose Jacob
Reserve Bank of New Zealand and Centre for Applied Macroeconomic Analysis - ANU
Punnoose.Jacob@rbnz.govt.nz

August 24, 2023

**Abstract**

We construct a Financial Conditions Index (FCI) for the United States using a dataset that features many missing observations. The novel combination of probabilistic principal component techniques and a Bayesian factor-augmented VAR model resolves the challenges posed by data points being unavailable within a high-frequency dataset. Even with up to 62% of the data missing, the new approach yields a less noisy FCI that tracks the movement of 22 underlying financial variables more accurately both in-sample and out-of-sample.

JEL classification: C11, C32, C52, C53, C66.
Keywords: Financial Conditions Index, Mixed-Frequency, Bayesian Methods

## 1 Introduction

Assessing financial conditions, particularly risks potentially posed by disruptions to credit markets for the business cycle, has become a priority for central banks since the Global Financial Crisis (Bordo 2017). While quantifying risks to financial stability remains a challenge for the profession, a key step in this direction has been the development of financial conditions indices; a financial conditions index (FCI) provides a quantitative summary of common movements in an array of financial variables that describe financial conditions in an economy. Since current financial conditions may reflect relevant information about future risks to the real economy (Adrian et al. 2019 and Giglio et al. 2016), a timely high-frequency FCI is of considerable appeal to the policy-maker. A significant challenge that confronts the econometrician when she sets about the task of constructing a timely FCI is the incompleteness

[*] The views expressed herein are those of the authors, and do not necessarily represent the official views of the Reserve Bank of New Zealand. We are grateful for the advice and guidance of Dimitris Korobilis and Gary Koop who commented on an early draft.

of the dataset of interest. High-frequency financial datasets are often hampered by missing values while macroeconomic time series are usually only reported at low frequencies. This paper offers solutions to problems related to data incompleteness, paving the way for the construction of FCIs that are reliable and timely.

We approach the problem of building a FCI with data featuring many missing observations, by setting up a Factor-Augmented VAR, which we estimate by adapting the algorithm of Koop and Korobilis (2014) to incorporate the strategy of Giannone et al. (2008) that deals with missing data. Following the authors, we employ the two-step estimator of the state-space representation of dynamic factor models proposed by Doz et al. (2011). The estimator is able to deal with missing observations, exploiting the dynamics of the common factors while accounting for common and idiosyncratic heteroskedasticity. In the first step, the parameters of the state-space representation are estimated using probabilistic principal components estimators. This is the key difference between our approach and the one originally proposed by Giannone et al. (2008) who instead take principal components of a 'balanced' version of the data, that only include series without missing values. On the contrary, we initialize the factors on the full incomplete 'unbalanced' panel of data. In the second step, factors are re-estimated with a Kalman Smoother using the entire 'unbalanced' panel of data.

The main purpose of our exercise is to construct an index that tracks the development of financial conditions when many data points are missing. We find that probabilistic solutions to principal component estimation yield better in-sample fit, particularly when working at a weekly frequency. In contrast, Least Squares solutions to principal component estimation deliver very noisy indices, compromising interpretability.

Another key contribution of this paper is to study the sensitivity of the two-step estimator of approximate dynamic factor models proposed by Doz et al. (2011) which is widely used in the literature. We explore several unbalanced panel techniques based on probabilistic principal components as alternatives to deal with noisy and missing data at higher frequencies. We rely on the probabilistic view of a Principal Component Regression (PPCA) proposed by Tipping and Bishop (1999a) which admits missing values and various levels of regularization. This approach addresses the problem that conventional principal component techniques (PCA) overfit the data at higher frequencies. We then integrate PPCA into the two-step estimation procedure of Doz et al. (2011). Although the issue of sensitivity to factor initialization is also relevant in the context of other estimation frameworks[1], in this paper we focus on sensitivity of the two-step Kalman Filter and Smoother (KFS) estimates to factor initialization when working with unbalanced panels of data. Understanding this sensitivity is important because in the first step, the parameters of the model are estimated by treating the principal components as if they were the true common factors. While Doz et al. (2011) prove the consistency of the two-step estimator of approximate DFMs, this result relies on the consistency of principal components as estimators of the span of the common factors. However, in an incomplete data environment, principal components are not available without pre-treatment

[1] For instance Bernanke et al. (2005) conduct inference in a Factor-Augmented VAR, estimating the model with Markov-Chain Monte Carlo where the unobserved factors are pre-estimated with principal components.

or modelling of missing values. One of the key results of our exercise is that inference using a two-step KFS is sensitive to factor initialization.

Our paper relates to a large body of literature on *now-casting*, coincident index construction and mixed-frequency model estimation.[2] Similar to Giannone et al. (2008), our framework can be interpreted as a large bridge model, which uses a large number of variables to bridge higher frequency data releases with the forecast of the relatively lower frequency variables. In line with Koop and Korobilis (2014), our FCI is built employing a very flexible framework which allows for stochastic volatility and time-varying parameters. Recently, Eraslan and Schröder (2022) extend this framework to account for mixed-frequency data by defining the relationship that links quarterly variables to their latent high-frequency counterpart, following Mariano and Murasawa (2003). Similarly, Bańbura and Rünstler (2011) and Bańbura et al. (2013) also extend Giannone et al. (2008) to a mixed-frequency environment by integrating a forecasting equation for quarterly variables which constrain the state-space representation of a dynamic factor model. On the contrary, we follow the original approach by Giannone et al. (2008) and focus instead on the sensitivity of inference to different methods to initialize factors which are relevant if many data points are missing.

While the methodology adopted in this paper inherits all the advantages of the approaches introduced by Koop and Korobilis (2014), it also offers a more flexible way to measure financial conditions that allows for: i) time-varying weights, which define the way each financial variable load into the FCI; ii) structural instability of the relationship between the FCI combining time-series data available at different frequencies. Most FCI models avoid the problem of having to deal with incomplete datasets by restricting the sample size and aggregating higher frequency variables to a lower base frequency. Neglecting the 'unbalanced' nature of the data leads to less informative indicators which may fail to capture potentially relevant movements in financial variables at high frequencies. An exception is the approach of Brave and Butters (2011) that employs the EM algorithm of Stock and Watson (2002) to estimate a Principal Component in a mixed-frequency setting. The EM algorithm allows them to estimate a FCI for the US economy dating back to the early 1970s.

Most FCIs are estimated under the assumption that the underlying parameters and volatilities are time-invariant. However, there are good reasons to believe that, neither the importance of a financial variable in determining broad financial conditions, nor the relation between financial conditions and the macroeconomy, are constant over time, particularly over longer periods. Financial intermediation was subject to important institutional changes in recent decades, characterized by deregulation, financial innovation and globalization. These long-run trends have resulted in a larger size of the financial sector in the economy which, in principle, could increase the relevance of shocks to financial conditions for macroeconomic fluctuations. Indeed, after the Great Financial Crisis of 2007/09, a large literature has established the importance of modelling time-varying parameters, especially when modelling financial time-series (see for instance Stock and Watson 2007 and Koop and Korobilis 2012b).

The remainder of the paper proceeds as follows. Section 2 outlines the econometric problem faced when constructing FCIs with incomplete data, Section 3 explains the data used,

[2] See Stock and Watson (2016) and Foroni and Marcellino (2013) who provide reviews of the literature.

Section 4 defines the model and the estimation methods. Section 5 discusses the results and Section 6 concludes.

## 2 Measuring financial conditions with incomplete data

Missing data arise for several reasons. For instance, the data might have different sampling frequencies. The time series of GDP growth is only available at a quarterly frequency, and will feature many missing observervations if used for the construction of a weekly FCI. Moreover, even when data series are available at compatible frequencies, some series begin later than others, and the timing of the final observations may also not be aligned because data-releases may not be synchronous. In our empirical exercise, macroeconomic variables are available for the full sample period from 1971 to 2020. On the contrary, many financial variables start later. When measuring financial conditions in real time, it is often that many variables have missing observations for the most recent periods. Missing observations can also arise in the middle of the sample, for some time-series, due to data provider issues. The details of how missing data are handled in the literature differ across signal extraction and state-space applications, but it is common to assume that data are missing at random. According to Stock and Watson (2016), the missing-at-random assumption, which rules out any dependency between the missing observation and the latent variables in the model is reasonable in the context of DFMs for macroeconomic applications.

When encountering these problems, many researchers often choose to aggregate the data, or disregard series with missing observations. Missing observations are also often replaced with the historical mean of the corresponding time-series. However, such procedures may lead to relevant information loss that can lead to *'aggregation bias'* due to the neglect of high-frequency dynamics.

## 3 Data

Next, we describe the data used to construct our FCI. FCIs should capture movements in a wide range of variables that contain relevant information about the current state of financial conditions. With this in mind, we select a broad range of variables that extend the data considered by Koop and Korobilis (2014) (see Table 2 for further details). We include variables covering the following categories:

1. We consider variables that measure credit volumes and leverage. These include household debt as a percentage of GDP and households'debt service ratio, the total amount outstanding of asset-backed securities issuance, as a percentage of GDP and total consumer credit owned and securitized as a percentage of GDP. Credit supply, a common determinant of financial distress, is considered by including a survey indicator of credit standards (DRTSCILM).

2. We include one variable that captures exchange rate movements of several important currencies towards the dollar (JPMNEER) and a commodity index which tracks the movements of the price of a number of important commodities (CRY Index).

3. A number of indicators which generally describe the market price of risk are considered. These include, for instance, the TED spread, the spread on 2 year and 10 year

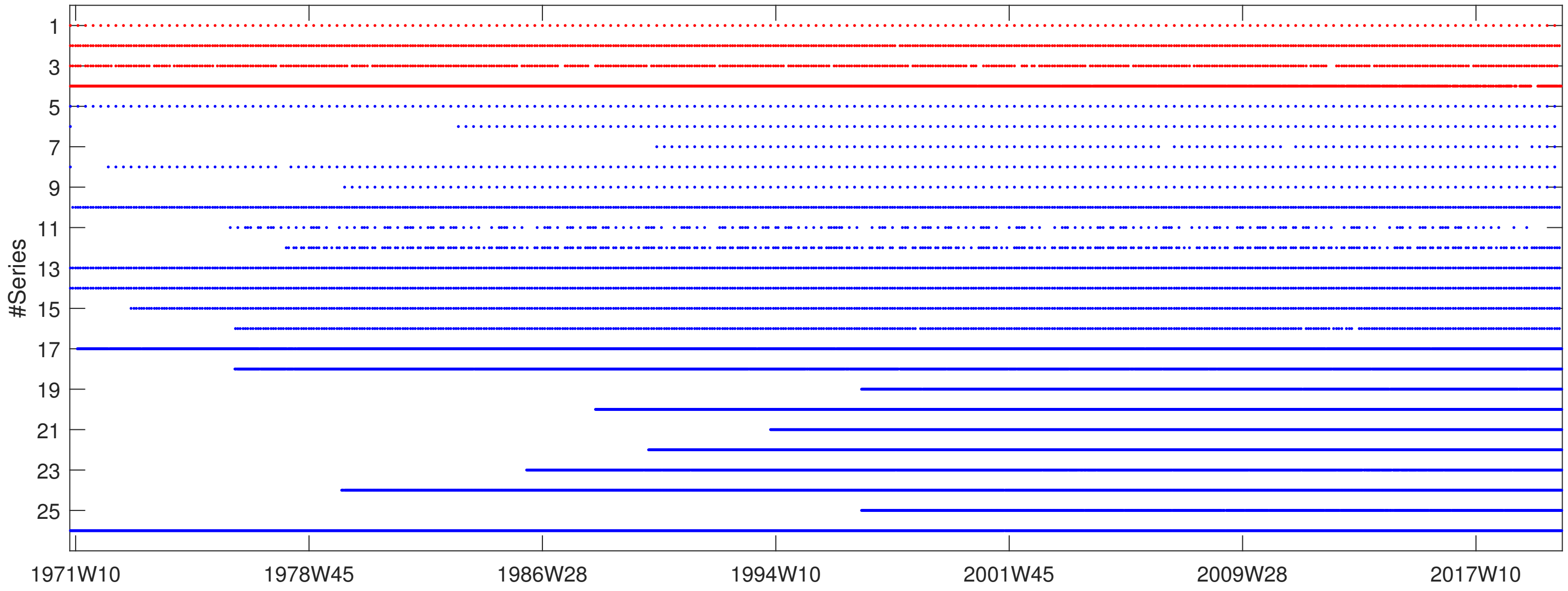
Treasury Bills. We include a number of risk indicators targeting in particular credit markets. These include a Loan Performance index, commercial paper spread, high yield spreads, mortgage spreads, the spread between the bank prime loan rate and the Libor, the finance rate on commercial bank loans. We also include the loan-to-deposit ratio of commercial banks, which measures liquidity risk.

4. Variables measure equity valuations in the economy include the S&P500 price index and the Wilshire 5000 Full Cap Price index.

5. We include variables related to beliefs, expectations and uncertainty. We consider the VIX, which is commonly related to risk-aversion and uncertainty, a survey index of expected changes in financial conditions and a yield curve weighted index of the implied volatility of short-term treasury options.

6. Finally, we also use macroeconomic data to construct the FCI but do not include them in the index. These include real GDP growth, the unemployment rate and inflation measured as the percentage changes in the Consumer Price Index The interest rate used in the observable sector of the model is the effective federal funds rate. The macroeconomic sector of the model allows us to make sure that the FCI is uncontaminated by macroeconomic innovations. Figure 1 gives a visual description of the data used to construct the FCI.

Figure 1: Incomplete data used to build the FCI



**Notes**: The frequency of the data is weekly. The white squares indicate missing values; red/blue squares indicate the availability of macroeconomic/financial variables considered. Macroeconomic variables include 4 time series, while the financial variables include the 22 financial variables used. The y-axis labelled #*Series* refers to the numbering of the time-series according to Table 2.

The rows in red describe the observations in our dataset that relate to macroeconomic variables. With the exception of interest rates, all macroeconomic variables in our model are only available at quarterly and monthly frequency, in contrast to several of the financial variables that are available at the weekly frequency. These are examples of the first reason why some data is missing in large datasets that are used to estimate FCIs - some variables are observed at a lower frequency than others. The rows in blue describe the pattern of sparsity in financial variables. Here observations are missing for three main reasons. First, most financial variables are missing for many periods at the beginning of the sample. Second, there are some missing values in the middle of the sample due to data limitations of the data provider. Lastly, it is possible to observe few variables which are not available for the very recent periods. This is often caused by lags in data availability. In total, the FCI is based on 22 financial variables and 4 macroeconomic series which are explained in detail in Table 2. All variables are transformed so that they are stationary and standardized prior to the estimation of the model.

# 4 Econometric Approach

We briefly describe our model and estimation procedure which is an extension of the time-varying parameter Factor-Augmented VAR (TVP-FAVAR) methodology of Koop and Korobilis (2014), such that factors are initialized with a number of different principal components estimators particularly suited to incomplete data environments.

## 4.1 The TVP-FAVAR

Consider a TVP-FAVAR, written as follows

$$X_t = \lambda_t^F F_t + \lambda_t^y Y_t + u_t, \tag{1}$$

$$\begin{bmatrix} Y_t \\ F_t \end{bmatrix} = \beta_{t,1} \begin{bmatrix} Y_{t-1} \\ F_{t-1} \end{bmatrix} + ... + \beta_{t,p} \begin{bmatrix} Y_{t-p} \\ F_{t-p} \end{bmatrix} + \varepsilon_t, \tag{2}$$

where $X_t$ is an $(N \times T)$ matrix, defining an unbalanced panel of $N$ financial variables that load onto the FCI denoted in this setting by $F_t$. $\lambda_t^F$ are time-dependent loading parameters that define which financial variables $x_{it}$ load onto $F_t$ at each point in time $t$. $\lambda_t^Y$ denotes the coefficients of a set of observable macroeconomic variables organized in $Y_t$, in the measurement equation (1) which are also time-dependent and may contain missing values. This term is key in ensuring that the variation in the financial variables that are attributable to changes in real economic conditions are purged from $F_t$. The state equation (2) describes the relationship between macroeconomic variables and the FCI. The set of parameters that determine this relationship $\{\beta_{t,1}, ..., \beta_{t,p}\}$ are also allowed to vary over time. Lastly, the innovations to the measurement equation $u_t \sim N(0,\ V_t)$ and to the state equation $\varepsilon_t \sim N(0,\ Q_t)$ are Gaussian and their volatilities are allowed to vary with time.

The main objective of the exercise is to estimate a FCI at quarterly, monthly and weekly frequencies. Note that at higher frequencies, the data is very sparse (i.e. there are many missing values). Our strategy to tackle the issue consists of writing the model for the highest available frequency, in which lower frequency variables feature in the model and are bridged with factors calculated at higher frequencies as in Giannone et al. (2008). Following the authors, we apply the two-step estimation procedure of Doz et al. (2011), except that we do not rely solely on standard principal components to initialize the factors. In particular, we use the full unbalanced panel of data $X_t$ to estimate the principal axis, taking into account the missing data. Our preliminary factor estimates are then used as an input to estimate the parameters $(\beta,\ \lambda)$ with a Kalman Filter and the factors $F_t$ with a Kalman Smoother, taking into consideration the full unbalanced data $(X_t,\ Y_t)$. The algorithm is a modified version of that proposed by Koop and Korobilis (2014) where the parameter set is only updated when data is available. If a particular data point is missing, the Kalman Gain in the corresponding updating equation is simply set to zero. The full details of the algorithm are summarized in Appendix A.1.

This methodology combines principal components with the Kalman Filter & Smoother, as proposed by Giannone et al. (2008) and Doz et al. (2011) and is known as "bridging with factors". It should be noted that it is not possible to construct the FCI at higher frequencies by using the standard estimation procedure of Koop and Korobilis (2014).

The model is completed by the last two state equations below,

$$\beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \sim N(0,\ R_t), \tag{3}$$

$$\lambda_t = \lambda_{t-1} + v_t, \quad v_t \sim N(0,\ W_t), \tag{4}$$

which allow the parameters to smoothly change over time. A final remark regarding all variance-covariance matrices $\{V_t,\ Q_t,\ R_t,\ W_t\}$ is in order. While time-variation in $V_t$ and $Q_t$ describes the dynamics of stochastic volatility, $R_t$ and $W_t$ determine the amount of time-variation in the parameters $\beta_t$ and $\lambda_t$. For $V_t$ and $Q_t$, Exponential Weighted Moving Average (EWMA) estimators are used which involves the use of recursive simulation-free variance discounting methods and depend on decay factors $\kappa_1$ and $\kappa_2$[3]. EWMA estimators accurately approximate integrated GARCH processes whereas $R_t$ and $W_t$ are estimated through forgetting factor methods described in Koop and Korobilis (2012a, 2013)and are driven by $\kappa_3$ and $\kappa_4$ [4]. Variance discounting and forgetting factors are commonly used since they offer a computationally feasible way of modelling stochastic volatility and time-variation in the parameters by recursively updating all variance-covariance matrices which are assumed to evolve smoothly over time.

The model defined by equations (1)-(4) differs from that proposed by Koop and Korobilis (2014) in two important dimensions. Firstly, the time series that are included in $Y_t$ and $X_t$ feature many missing values. Koop and Korobilis (2014) estimate their model using data that are available at a quarterly frequency , and set missing values to zero . In contrast, our model is estimated at higher frequencies, where $X_t$ include many missing data points and $(Y_t)$ includes variables such as inflation or GDP growth which are only available at a quarterly frequency. This feature of our model is very useful for policymakers who often need indica-

[3] See Appendix A.1, equations 27-28.

[4] See Appendix A.1, equations 17-18.

tors of financial conditions at higher frequencies. Secondly, we treat the missing values in $X_t$ as latent variables. While Koop and Korobilis (2014) set any missing values in $X_t$ to zero (which is the unconditional mean of all the standardized financial time-series), we estimate any missing values in $X_t$ with probabilistic principal component methods. We will discuss the latter methodology in the next section. The model is estimated at weekly, monthly and quarterly frequencies. While monitoring financial conditions at higher frequencies is indeed appealing, the noise in financial data at higher frequencies may challenge the relevance of higher frequency FCIs. We will later see the value of PPCA in dealing with this problem.

## 4.2 Estimation

The model defined by equations (1)-(4) can be cast in state-space form and is estimated through a Bayesian Kalman filtering and smoothing routine which is based on Koop and Korobilis (2014). The algorithm extends the original framework to an incomplete data environment and involves the following steps:

- **Step 1:** Estimate $F_t$ using principal components methods which accommodate missing values.

- **Step 2:** Conditional on the initial values of the factors, $\tilde{F}_t$, estimate the parameters in the TVP-FAVAR by applying Kalman filtering and smoothing.

- **Step 3:** Conditional on the parameters estimated in Step 2, estimate $\hat{F}_t$, which is used as our FCI, by applying Kalman filtering and smoothing.

The incomplete nature of the data gives rise to a sparse dataset; 'sparsity' implies that at some points across the time dimension, there are very few observations. This is particularly relevant for early periods in the sample for which only a small subset of all variables considered are observed. One of the great advantages of Kalman filtering techniques is that it is straightforward to deal with missing values. The algorithm of Koop and Korobilis (2014) (Steps 2 and 3) is modified in accordance with Giannone et al. (2008) and Koopman and Commandeur (2008) by simply setting the Kalman Gain term to zero in each updating equation, whenever an observation is missing. This guarantees that in the absence of any new information regarding a variable (*i.e.* when it is missing), our best guess for the current state of the respective parameter remains unchanged. In Step 1, prior to the TVP-FAVAR stage, we estimate the factor $F_t$ and reconstruct the data matrix of financial variables. The challenge is to reconstruct the unbalanced panel of data $X_t$ which contains missing observations. The reconstructed matrix is then used as an input to Step 2. In Step 3, a Kalman Filter and Smoother is passed through the pre-estimated factor yielding our final estimate of the FCI. Steps 2 and 3 are similar to the method proposed by Doz et al. (2011).

### 4.2.1 Principal Component Analysis in the presence of missing values

One of the most popular approaches to PCA is based on singular value decomposition (SVD) of the covariance matrix of the data,

$$C = N^{-1} X^T X = UDU^T, \tag{5}$$

where $D$ contains the eigenvalues and $U$ the eigenvectors of the covariance matrix $C$, normalized to have unit-length. This approach is only valid when the panel is balanced, *i.e.* when no observations are missing. In the presence of missing values, Least-Square techniques need to be modified. We explore the following four alternative solutions to finding principal components in incomplete datasets.

i) *Standard PCA with missing values set to zero (PCA)*

ii) *Least Squares Expectation-Maximization PCA (EM-PCA)*

iii) *Probabilistic PCA (PPCA)*

iv) *Variational Bayesian PCA (VBPCA)*

These methods vary in several important ways.[5] The standard PCA removes missing values by setting them to zero. Since the variables in $X$ are standardized, this is equivalent to setting each missing value to the respective variable's unconditional mean. Next, the Expectation-Maximization PCA (EM-PCA) as proposed by Stock and Watson (2002), results in an imputed matrix $X$ where the missing values are reconstructed such that the Least-Squares Likelihood is maximized. In each iteration, any missing points are replaced with their conditional mean. The PCA and EM-PCA belong to the family of Least-Squares methods since they work by minimizing the squared residuals of the PC regression. In contrast, the probabilistic PCA and the Variational Bayes PCA approach the PC regression from a probabilistic point of view.

A probabilistic formulation of PCA offers a number of benefits, including a well-founded treatment of missing values, extendability and regularization. The PPCA is based on the work of Tipping and Bishop (1999b). Estimation in an incomplete data setting is discussed in Ilin and Raiko (2010) who use an EM algorithm treating the latent factors as hidden variables. Although the PPCA performs some regularization (*i.e.* the Gaussian priors set on the factors penalize large values in $F$), this might be insufficient if the data is very sparse. An extension that allows for more strict penalization is proposed by Oba et al. (2003). This approach which is called Variational Bayesian PCA (VBPCA) involves imposing priors over the remaining parameters in the model that penalize more complex explanations of the data.

In a nutshell, estimation of the proposed FCI is based on the TVP-FAVAR which is estimated in three steps. In the first step, a large number of financial data series $X$ are compressed into a common factor by using one of the four methods suggested above. Secondly, taking this factor as given, a Kalman filter and smoother is applied to estimate time-varying parameters and stochastic volatilities. Third, conditional on the parameters obtained, the Kalman Filter and Smoother is applied to the factors resulting in the FCI.

[5] We discuss the main idea and intuition for each method. More details are provided in Appendix A

# 5 Results

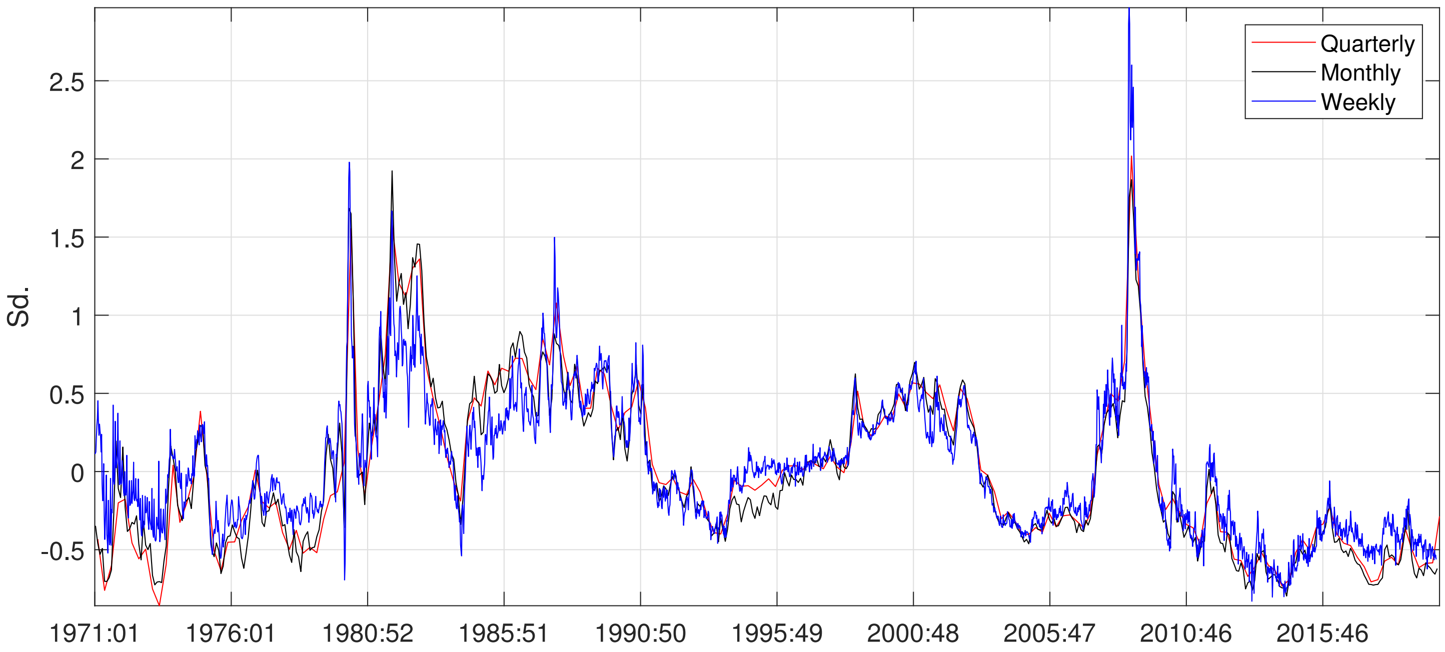
## 5.1 Measuring financial conditions with incomplete data

The main purpose of our exercise is to construct an index that tracks the development of financial conditions when many data points are missing. Figure 2 presents the Financial Conditions Index (FCI) estimated at a weekly, monthly and quarterly frequencies, based on the Variational Bayes approach to PCA, which is the model that yields the best in-sample fit across frequencies. This result is evident in Figure 6 which presents in-sample Mean Squared Errors obtained when estimating a Principal Component (PC) Regression with each of the four alternative methods we employ.Probabilistic solutions to PC estimation yield better in-sample fit. This finding is particularly evident when moving from the quarterly frequency to weekly frequency. With weekly data, the probabilistic PCA noticeably outperforms PCA and EM-PCA.

Figure 5 presents the FCI calculated by running our model where factors are initialized differently, for each of the frequency considered. The key result is that inference in the model is very sensitive to Ff initialization. When data is missing, FCI estimates are very sensitive to the specific method employed to pre-estimate the factor which is then used in the two-step KFS approach. Moreover, Figure 5 also highlights that at higher frequencies, Least Squares solutions to PC estimation (*i.e.* PCA and EM-PCA) deliver very noisy indices compromising interpretability. On the other hand, the probabilistic measures, PPCA and VBPCA deliver interpretable indices.

The indices presented in Figure 2 summarize the movement in the underlying 22 financial variables relevant for real economic developments in the US (see Table 2 in the Appendix for details). By construction, the FCI fluctuates around zero. Times of tight financial conditions such as the period marked by the Global Financial Crisis in 2009, drive the index upwards. In contrast, periods of accommodative financial conditions lead to negative values of the FCI. The figure reveals the size of the ’bias’ that results from neglecting information at higher frequencies. For instance, the FCI built at quarterly frequency tends to underestimate the size of financial distress in 2009, during the Great Recession.

Figure 2: Financial Conditions Indices at different data frequencies

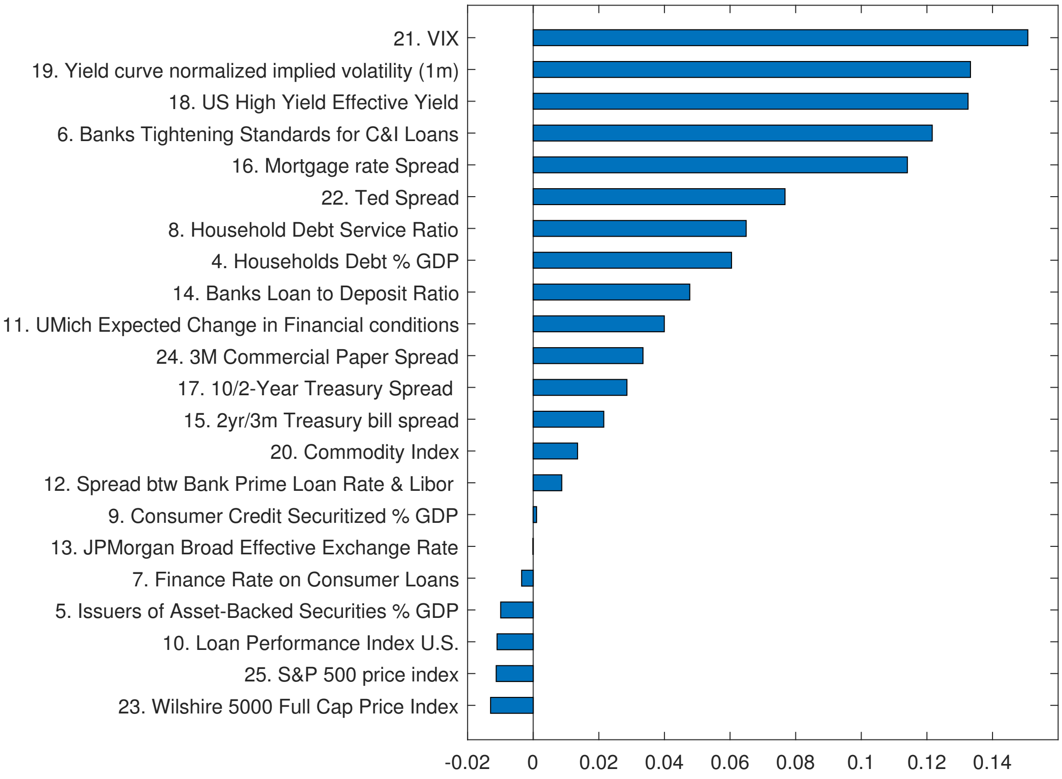


**Notes**: The Financial Conditions Index (FCI) in this figure is estimated for the period 1971-2020 with the Variational Bayes variant of our algorithm for a weekly, monthly and quarterly frequencies. Positive values of the FCI indicate that financial conditions are tighter than average, while negative values indicate financial conditions that are looser than average. Deviations from the mean are measured in standard deviations (Sd.).

## 5.2 Individual contributions to the dynamics of the FCI

We now examine the driving forces and macroeconomic relevance of financial conditions as measured by the best-fitting FCI, the VBPCA. Figure 3 shows the average importance of the underlying financial variables which contribute the most in driving the FCI, throughout the period from 1990-2020.

Figure 3: Average weights of individual financial variables loading into the Financial Conditions Index



**Notes**: The weights represent the average time-varying loadings of each financial indicator into the FCI for the full sample period 1971-2020.

The indicators with the highest positive contribution for the movement in the FCI include the VIX which measures the markets' expectation of near-term volatility and the volatility implied in Treasury options. Credit supply, as measured by the change in credit standards by commercial banks, also features in the most influential variables driving the FCI. Other relevant indicators include two important measures of leverage and liquidity - such as the Household Debt Service Ratio and the Bank Loans to Deposit Ratio; most of the remaining influential variables driving financial conditions relate to credit spreads. These include mortgage spreads, the TED spread and spreads on US government debt.

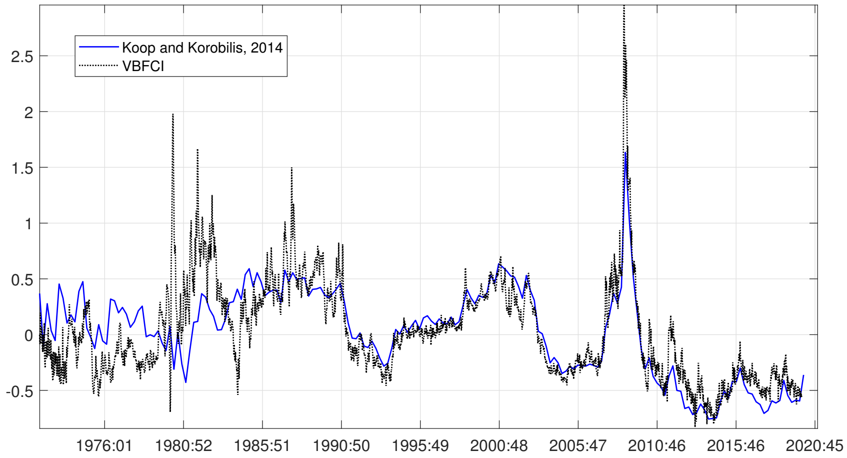
In addition, the financial variables which contribute negatively to the developments of financial conditions are mainly quantity variables (in dollars) which relate to equity valuations and mortgage lending. This result is intuitive, since in times of tight financial conditions, the performance of outstanding loans tends to deteriorate and stock markets tend lose value. Some interest rates and spreads also have a countercyclical contribution to the index. These include the commercial finance rate on consumer loans. This may reflect the fact that the yield curve tends to invert in recessions, when financial conditions also tend to tighten.

## 5.3 Comparison with alternative econometric approaches

The key novelty of the econometric approach used throughout this paper is its treatment of missing data and regularization of noise. The model of Koop and Korobilis (2014) only allows the construction of the FCI at a quarterly frequency. The high-frequency FCI of Brave and Butters (2011) is a Dynamic Factor Model with no time-varying dynamics. Our mixed-frequency model allows us to construct the FCI at a weekly and monthly frequencies and retain all modelling flexibility offered by Koop and Korobilis (2014).

Figure 4 below allows for a comparison between our approach and Koop and Korobilis (2014) estimated on the same dataset at a quarterly frequency. The FCI is naturally more informative than the analogous model estimated at a quarterly frequency. This is because it captures more movements in the underlying data that occur at higher frequencies, but also because it estimates missing data points in financial time series. Overall, the FCI seems to provide more timely signals of financial distress. This is particularly noticeable during the Great Recession.

Figure 4: Comparison between the FCI and alternative methods.



**Notes**: The FCI is estimated with the methods described in section 2, with the Variational Bayes variant of the PCA algorithm presented, at a weekly frequency for the sample period between 1971-2020. The FCI by Koop and Korobilis (2014) is estimated on the same dataset at a quarterly frequency.

## 5.4 The macroeconomic relevance of financial conditions

We now focus on the macroeconomic relevance of financial conditions by studying the out-of-sample performance of our FCI in forecasting GDP growth, inflation, unemployment and interest rates.

We calculate the FCI at a weekly, monthly and quarterly frequencies and compare Mean Square Forecast Error (MSFE) statistics calculated for a given frequency $f$ and forecast horizon $h$ as follows:

$$MSFE_h^f = \frac{\sum_{j \in O_j} (Y_j - \hat{Y}_j)^2}{T - h},$$

where $O_j$ denotes the set of indices $j$ for which $Y_j$ is observed at frequency $f$ and T is the total number of time-series observations. To be clear, $\hat{Y}_j$ is our forecast, h-steps ahead, for all macroeconomic time series at a given frequency $f$. When the data is available at higher frequencies , say monthly, we only observe GDP once every three months. In this case, we calculate MSFE statistics based on the difference $(Y_j - \hat{Y}_j)^2$ series for $j = \{1, 3, 6, 9, ...\}$.

Our goal is to compare the forecasting performance of FCIs where factors are initialized with different signal extraction methods (i.e. PCA, PPCA, EM-PCA and VBPCA) . Although it is tempting to explore how the forecasting performance varies across the data-frequencies considered, this analysis would be spurious because weekly, monthly and quarterly models are estimated with a different number of parameters. We estimate the weekly, monthly and quarterly models with 12, 6 and 4 lags, respectively. The choice of the number of lags to include in the model is a compromise between computational cost, the risks of over-parameterization and neglecting persistence. Moreover, the FCIs estimated for different frequencies also differ, making this comparison even harder. For these reasons, a comparison of forecast performance across frequencies is out of the scope of our current work and is left for future research.

Our out-of-sample forecasting exercise is based on the latest 2020 vintage of data and therefore not performed in real time. The baseline specification is obtained by estimating our model (described in section 4.1) without any latent factor included and therefore boils down to a VAR. The model is ran recursively, on an expanding window of data. We divide our sample into training (from 1971-1981) and testing (from 1981-2020). The first line of Table 5.4 for each macroeconomic variable reports MSFE for the baseline case, without the inclusion of any FCI (i.e. VAR (no FCI)). The lines below report the ratio of the MSFEs obtained from the baseline case and the exact same model augmented with the FCI, computed via one of the four methods discussion thus far (i.e. PCA, EM-PCA, PPCA and VBPCA). A ratio lower than 1 signals that the FCI measured with one of the aforementioned signal extraction methods outperforms the benchmark model. Forecast gains vis-Ã -vis the benchmark are tested with Diebold-Mariano Statistics with automatic lag selection.

The main message of the out-of-sample exercise is that the value added by the econometric methods proposed throughout this paper becomes more obvious when working at higher frequencies with noisy data. Importantly, the value of using FCIs to forecast macroeconomic variables is higher in the very short run, at weekly and monthly frequencies. At a weekly frequency, the FCI computed via VBPCA and PPCA improves forecasting most significantly for unemployment and interest rates. What is particularly striking is the relatively superior performance of the VBPCA and PPCA methods over the EM-PCA used by Brave and Butters (2011) and in many other economic applications. This result emphasizes the importance of regularization in econometric models, in particular when the data is observed at high-frequency and is noisy. At lower frequencies, it is possible to observe some forecasting gains, in particular when forecasting unemployment and inflation, one-month ahead. However, these gains are statistically insignificant at a quarterly frequency. Although the models with the FCI have lower MSFE, this difference does not appear to be statistically significant.

On the other hand, there is little evidence that the alternative signal extraction methods outperform PCA at a quarterly frequency. This finding is consistent with the previously reported result that these methods add value when the data is very noisy (which tends to be the case at higher frequencies).

| **Quarters ahead** | **1q** | **4q** | **8q** | **12q** | **20q** | **Months ahead** | **1m** | **4m** | **6m** | **12m** | **24m** | **Weeks ahead** | **1w** | **4w** | **8w** | **9w** | **12w** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***GDP growth*** | | | | | | ***GDP growth*** | | | | | | ***GDP growth*** | | | | | |
| MF-VAR (no FCI) | 0.68 | 0.78 | 0.73 | 0.57 | 0.54 | MF-VAR (no FCI) | 0.77 | 1.09 | 1.03 | 1.15 | 0.89 | MF-VAR (no FCI) | 0.67 | 0.68 | 0.72 | 0.72 | 0.73 |
| EM-PCA | 1 | 1.02 | 1.11 | 1.08 | 1.03 | EM-PCA | 1.31 | 1.56 | 1.46 | 1.16 | 1.13 | EM-PCA | 0.97 | 0.95** | 0.96* | 0.99 | 1.01 |
| PCA | 0.98 | 1 | 0.99 | 1 | 1.01 | PCA | 1.14 | 1.24 | 1.17 | 1.08 | 1.15 | PCA | 0.97 | 0.99 | 0.94 | 0.96 | 0.98 |
| PPCA | 0.98 | 1.02 | 1 | 0.99 | 1.02 | PPCA | 1.14 | 0.94 | 0.88* | 0.74*** | 0.70*** | PPCA | 0.99 | 1 | 1 | 1.01 | 1 |
| VBPCA | 0.95 | 1.01 | 1 | 0.99 | 1.01 | VBPCA | 1.2 | 1.10 | 1 | 0.95*** | 0.99 | VBPCA | 0.96 | 1.01 | 1.05 | 1.08 | 1.04 |
| ***CPI inflation*** | | | | | | ***CPI inflation*** | | | | | | ***CPI inflation*** | | | | | |
| MF-VAR (no FCI) | 0.59 | 0.61 | 0.71 | 0.68 | 0.66 | MF-VAR (no FCI) | 0.56 | 0.71 | 0.69 | 0.69 | 0.69 | MF-VAR (no FCI) | 0.55 | 0.56 | 0.61 | 0.63 | 0.73 |
| EM-PCA | 0.95** | 0.96* | 0.90 | 0.84 | 0.90 | EM-PCA | 0.94*** | 0.98 | 1.04 | 1.17 | 1.04 | EM-PCA | 0.99 | 1.05 | 1 | 0.97 | 0.86* |
| PCA | 1.04 | 1.07 | 1.12 | 1.03 | 0.95 | PCA | 0.91* | 1.03 | 1.05 | 1.11 | 1.23 | PCA | 1.03 | 1.12 | 1.24 | 1.19 | 1.23 |
| PPCA | 1.06 | 1.12 | 1.21 | 1.13 | 1.01 | PPCA | 0.93*** | 0.94*** | 0.96** | 0.96*** | 0.95*** | PPCA | 0.96 | 0.99 | 1 | 0.98 | 0.92 |
| VBPCA | 1 | 1.06 | 1.12 | 1.06 | 1 | VBPCA | 0.92** | 0.95** | 0.98 | 1.02 | 0.96* | VBPCA | 0.94 | 1 | 1.04 | 1 | 0.86* |
| ***Unemployment*** | | | | | | ***Unemployment*** | | | | | | ***Unemployment*** | | | | | |
| MF-VAR (no FCI) | 0.80 | 0.97 | 0.93 | 0.87 | 0.87 | MF-VAR (no FCI) | 0.81 | 0.88 | 0.95 | 1.19 | 1.13 | MF-VAR (no FCI) | 0.85 | 0.86 | 0.89 | 0.89 | 0.94 |
| EM-PCA | 1 | 1 | 1.02 | 1.04 | 1.01 | EM-PCA | 0.93*** | 0.96* | 0.98 | 0.95* | 0.97 | EM-PCA | 0.97** | 0.98 | 1.01 | 1.02 | 1 |
| PCA | 0.92 | 1 | 1 | 1.01 | 1 | PCA | 1.03 | 1.06 | 1.08 | 1.13 | 1.07 | PCA | 0.93** | 0.95 | 0.94 | 0.94 | 0.92 |
| PPCA | 0.94 | 1.01 | 0.99 | 1.01 | 1.01 | PPCA | 0.97** | 0.96** | 0.95* | 0.89* | 0.84** | PPCA | 0.92*** | 0.93* | 1.02 | 1.03 | 1 |
| VBPCA | 0.98 | 1 | 1.01 | 1.02 | 1 | VBPCA | 0.96** | 0.95*** | 0.92** | 0.90* | 0.96 | VBPCA | 0.94** | 0.94 | 1.03 | 1.05 | 1.01 |
| ***Interest rates*** | | | | | | ***Interest rates*** | | | | | | ***Interest rates*** | | | | | |
| MF-VAR (no FCI) | 0.04 | 0.16 | 0.35 | 0.48 | 0.61 | MF-VAR (no FCI) | 0.02 | 0.21 | 0.32 | 0.71 | 0.89 | MF-VAR (no FCI) | 0.08 | 0.42 | 1.20 | 1.38 | 1.78 |
| EM-PCA | 0.95 | 1.04 | 1.03 | 0.95 | 0.89 | EM-PCA | 1.09 | 0.90 | 0.90 | 0.95 | 0.86 | EM-PCA | 0.73*** | 0.53*** | 0.57*** | 0.61*** | 0.77*** |
| PCA | 0.97* | 1.16 | 1.24 | 1.22 | 1.07 | PCA | 0.79*** | 0.87** | 0.96 | 1.36 | 1.62 | PCA | 0.83*** | 0.81*** | 0.94* | 0.98 | 1.10 |
| PPCA | 1 | 1.12 | 1.17 | 1.15 | 1.04 | PPCA | 1.28 | 1.03 | 0.96 | 0.86** | 0.70*** | PPCA | 0.67*** | 0.58*** | 0.67*** | 0.70*** | 0.85*** |
| VBPCA | 0.90* | 1.09 | 1.15 | 1.13 | 1.04 | VBPCA | 0.97 | 0.81*** | 0.78*** | 0.89*** | 1.04 | VBPCA | 0.58*** | 0.55*** | 0.65*** | 0.68*** | 0.84*** |

Table 1: Out-of-Sample MSFE for different signal extraction methods, data frequencies (i.e. quarterly, monthly and weekly) and horizons for predictions with a Mixed-Frequency VAR (MF-VAR) for GDP growth, inflation, unemployment and interest rates on a training sample 1971-1981 and testing 1981-2020. The methods are (i) standard Principal Component Analysis with missing values set to 0 (PCA) (ii) Least Squares Expectation-Maximization (EM-PCA) (iii) Probabilistic Principal Component Analysis (PPCA) and (iv) Variational Bayesian Probabilistic Principal Component Analysis (VBPCA). The MSFE statistics for the MF-VAR (first line in each table) are obtained by running the model without the inclusion of any FCI. The lines below show the ratio between the MSFE of the model in concern augmented with an FCI and the baseline MF-VAR model without any FCI. Values below 1 in each table for the PCA, EM-PCA, PPCA and VBPCA imply that the relevant model outperforms the baseline model for a given horizon, frequency and macro variable. ***,** and * refer to p-values lower than 0.01, 0.05 and 0.1, respectively, associated to the Diebold-Mariano Statistic with automatic lag-selection.

## 6 Conclusion

We constructed a Financial Conditions Index for the United States with data reported at different frequencies, starting dates and many missing observations. In doing so, we address a challenging aspect of estimating such indices with traditional macroeconometric techniques; the incompleteness of the underlying dataset.

The novel factor-augmented VAR we estimated through a combination of probabilistic PCA methods and Kalman filtering and smoothing routines offers a solution to cope with datasets featuring missing values. In-sample fit as well as out-of-sample forecasting results suggest that the alternative signal-extraction methods that we employ are useful particularly when the data is noisy, which usually tends to be the case at higher frequencies. Our key finding that the two-step Kalman Filter and Smoother approach to estimating approximate Dynamic Factor Models is sensitive to factor initialization when the datasets are incomplete, is also relevant in the context of other estimation frameworks such as Markov-Chain Monte Carlo simulations. However, this related issue is beyond the scope of this paper, and we leave it for future research.

## References

Adrian, T., Boyarchenko, N., and Giannone, D. (2019). Vulnerable Growth. *American Economic Review*, 109(4):1263–1289.

Attias, H. (2000). A variational Bayesian framework for graphical models. *Advances in neural information processing systems*, pages 209–215.

Bańbura, M., D., G., M., M., and L., R. (2013). Chapter 4 - now-casting and the real-time data flow. *Elliott G., Timmermann A. (Eds.), Handbook of economic forecasting, Elsevier*, 2(2).

Bańbura, M. and Rünstler, G. (2011). A Look into the factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting*, 27(2):333–346.

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

Bordo, M. D. (2017). An Historical Perspective on the Quest for Financial Stability and the Monetary Policy Regime. NBER Working Papers 24154, National Bureau of Economic Research, Inc.

Brave, S. and Butters, R. A. (2011). Monitoring financial stability : A financial conditions index approach. *Economic Perspectives*, 35(1):22–43.

Doz, C., Giannone, D., and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164(1):188–205.

Eraslan, S. and Schröder, M. (2022). Nowcasting GDP with a pool of factor models and a fast estimation algorithm. *International Journal of Forecasting*, (39(3)):1460–1476.

Foroni, C. and Marcellino, M. G. (2013). A Survey of Econometric Methods for Mixed-Frequency Data. *SSRN Electronic Journal*.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Giglio, S., Kelly, B., and Pruitt, S. (2016). Systemic risk and the macroeconomy: An empirical evaluation. *Journal of Financial Economics*, 119(3):457–471.

Ilin, A. and Raiko, T. (2010). Practical approaches to Principal Component Analysis in the presence of missing values. *Journal of Machine Learning Research*, 11:1957–2000.

Koop, G. and Korobilis, D. (2012a). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Koop, G. and Korobilis, D. (2012b). Why Has U.S. Inflation Become Harder to Forecast? *International Economic Review*, (53):867â886.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, 177(2):185–198.

Koop, G. and Korobilis, D. (2014). A new index of financial conditions. *European Economic Review*, 71:101–116.

Koopman, S. J. and Commandeur, J. J. (2008). *Introduction to State Space Time Series Analysis*. Oxford University Press.

Mariano, R. S. and Murasawa, Y. (2003). A NEW COINCIDENT INDEX OF BUSINESS CYCLES BASED. 443(October 2002):427–443.

Neal, R. M. and Hinton, G. E. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, page 355â368.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. I., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096.

Raftery, A. E., Kárný, M., and Ettler, P. (2010). Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill. *Technometrics*, 52(1):52–66.

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

Stock, J. H. and Watson, M. W. (2007). Erratum to âWhy Has U.S. In ation Become Harder to Forecast?â. *Banking*, 39(1):3–33.

Stock, J. H. and Watson, M. W. (2016). Factor Models and Structural Vector Autoregressions in Macroeconomics. *Handbook of Macroeconomics*, pages 1–111.

Tipping, M. E. and Bishop, C. M. (1999a). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 61(3):611–622.

Tipping, M. E. and Bishop, C. M. (1999b). Probabilistic Principal Component Analysis. *J. R. Statisit. Soc. B*, 61(3):611–622.

# A Econometric Methods

## A.1 Bayesian Kalman Filter with Incomplete Data

The model defined in (1)-(4) configures a MF-TVP-FAVAR and can be written compactly, in state space form as follows

$$\begin{aligned} X_t &= Z_t\Lambda_t + u_t, \quad u_t \sim N(0, V_t), & (6)\\ Z_t &= Z_{t-1}\beta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, Q_t), & (7)\\ \beta_t &= \beta_{t-1} + \eta_t \quad \eta_t \sim N(0, R_t), & (8)\\ \Lambda_t &= \Lambda_{t-1} + \upsilon_t \quad \upsilon_t \sim N(0, W_t). & (9) \end{aligned}$$

Where $\Lambda_t = [\lambda_t^y, \lambda_t^f]'$, $Z_t = \begin{bmatrix} Y_t \\ F_t \end{bmatrix}$. Note that $Z_t$ depends on the latent factor $F_t$ that is taken as data [6]. Let $\theta_t = \{\Lambda_t, \beta_t\}$ denote the parameter set and $D_t = \{X_t, Z_t\}$ the data for $t = \{1, ..., T\}$. Assuming that we know the posterior of $\theta$ at time $t-1$, Bayesian filtering/smoothing is based on the equations below

$$\begin{aligned} p(\theta_t, \theta_{t-1}|D_{t-1}) &= p(\theta_t|\theta_{t-1}, D_{t-1})p(\theta_{t-1}|D_{t-1}), & (10)\\ p(\theta_t|D_{t-1}) &= \int_\Omega p(\theta_t|\theta_{t-1}, D_{t-1})p(\theta_{t-1}|D_{t-1})d\theta_{t-1}, & (11) \end{aligned}$$

where $\Omega$ is the support of $\theta_{t-1}$. The prediction step is given by the Chapman-Kolmogorov equation A6.

Next, at each iteration $t$, the prior $p(\theta_t|D_{t-1})$ gets updated according to equation A6 and the measurement likelihood $p(D_t|\theta_t)$ is augmented by an additional observation of $D_t$. Hence the posterior distribution is updated according to the Bayes rule

$$p(\theta_t|D_t) = \frac{1}{H_t}p(D_t|\theta_t, D_{t-1})p(\theta_t|D_{t-1}). \qquad (12)$$

Where $H_t = \int p(D_t|\theta_t)p(\theta_t|D_{t-1}))$ is the normalizing constant. Equation A7 is refered to as the updating step. To summarize, the algorithm extends the one derived in Koop and Korobilis (2014) to an incomplete data environment where $D_t$ is allowed to contain missing values. It consists in 2 stFCI, iterating through prediction (A6) and updating (A7) after the system is initialized. These two main stFCI are repeated for $t = \{1, ..., T\}$.

**1. Kalman Filter**

**1.1 Initialization (Priors)** All quantities are initialized according to their priors which are chosen following the diffuse choices of Koop and Korobilis (2014): $f_0 \sim N(0, 10)$, $\Lambda_0 \sim N(0, I)$, $\beta_0 \sim N(0, I)$. The variances of the innovations in (A1)-(A4) can be seen as hyperparameters and are set to $\hat{V}_0 = 0.1 * I$, $\hat{Q}_0 = 0.1 * I$, $\hat{R}_0 = 10^{-5} * I$ and $\hat{W}_0 = 10^{-5} * I$. However,

[6] This quantity is the first principal component calculated in Step 1 by employing one of the four algorithms discussed in the body of the paper and detailed in section A2 to reconstruct the incomplete data matrix $X_t$

in this setting the hyperparameter are allowed to smoothly change over time following a Exponentially Weighted Moving Average (EWMA).

### 1.2 Prediction

$$\Lambda_t \sim N(\Lambda_{t|t-1}, \Sigma^{\Lambda}_{t|t-1}), \tag{13}$$

$$\beta_t \sim N(\beta_{t|t-1}, \Sigma^{\beta}_{t|t-1}). \tag{14}$$

Where $\Lambda_{t|t-1} = \Lambda_{t-1|t-1}$, $\beta_{t|t-1} = \beta_{t-1|t-1}$ and

$$\Sigma^{\beta}_{t|t-1} = \Sigma^{\beta}_{t-1|t-1} + \hat{R}_{t|t-1}, \tag{15}$$

$$\Sigma^{\Lambda}_{t|t-1} = \Sigma^{\Lambda}_{t-1|t-1} + \hat{W}_{t|t-1}. \tag{16}$$

The state covariances in the equations above are estimated by

$$\hat{R}_{t|t-1} = \frac{1}{\kappa_3}\hat{R}_{t-1|t-1}, \tag{17}$$

$$\hat{W}_{t|t-1} = \frac{1}{\kappa_4}\hat{W}_{t-1|t-1}. \tag{18}$$

where $\kappa_3$ and $\kappa_4$ are forgetting factors that define the law of motion of the parameters. We set these quantities to $\kappa_3 = \kappa_4 = 0.99$.[7] The prediction step allows us to compute measurement equation prediction errors that are necessary inputs for the updating step and computed as

$$\hat{u}_t = x_t - \hat{x}_{t|t-1}, \tag{19}$$

$$\hat{\varepsilon}_t = z_t - \hat{z}_{t|t-1}. \tag{20}$$

Where $\hat{x}_{t|t-1} = z_t\Lambda_{t|t-1}$ and $\hat{z}_{t|t-1} = z_{t-1}\beta_{t|t-1}$. With missing data, we simple set to zero the error corresponding to variables missing at a given point in time $t$.

### 1.3 Update

Update each $\Lambda_{it}$ for $i = 1, ..., n$ and $\beta_t$

$$\Lambda_{it} \sim N(\Lambda_{it|t}, \Sigma^{\Lambda}_{ii,t|t}), \tag{21}$$

$$\beta_t \sim N(\beta_{t|t}, \Sigma^{\beta}_{t|t}). \tag{22}$$

The terms in (A16) are calculated as

$$\Lambda_{it|t} = \Lambda_{it|t-1} + \Sigma^{\Lambda}_{ii,t|t-1} z'_t(\hat{V}_t + z_t\Sigma^{\Lambda}_{ii,t|t-1}F'_t)^{-1}\hat{\varepsilon}_t, \tag{23}$$

$$\Sigma^{\Lambda}_{ii,t|t} = \Sigma^{\Lambda}_{ii,t|t-1} - \Sigma^{\Lambda}_{ii,t|t-1} z'_t(\hat{V}_t + z_t\Sigma^{\Lambda}_{ii,t|t-1}z'_t)^{-1} z_t\Sigma^{\Lambda}_{ii,t|t-1}. \tag{24}$$

Where the term $\Sigma^{\Lambda}_{ii,t|t-1} z'_t(\hat{V}_t + z_t\Sigma^{\Lambda}_{ii,t|t-1}F'_t)^{-1}$ is the Kalman Gain for each time period $t$ and is set to zero for the variables missing at each step.

[7] In practice these two equations are approximations of $\hat{R}_{t|t-1} = \hat{R}_{t-1|t-1} + \hat{\eta}_{t-1}\hat{\eta}'_{t-1}$ and $\hat{W}_{t|t-1} = \hat{W}_{t-1|t-1} + \hat{v}_{t-1}\hat{v}'_{t-1}$in the standard Kalman Filter (see Koop and Korobilis (2013) and Raftery et al. (2010))

The terms in (A17) are calculated as

$$\beta_{t|t} = \beta_{t|t-1} + \Sigma^{\beta}_{t|t-1} z'_t (\hat{Q}_t + z_{t-1}\Sigma^{\beta}_{t|t-1} z'_{t-1})^{-1}(z_t - z_{t-1}\hat{\beta}_{t-1}) \tag{25}$$

$$\Sigma^{\beta}_{t|t} = \Sigma^{\beta}_{t|t-1} - \Sigma^{\beta}_{t|t-1} z'_t (\hat{Q}_t + z_t \Sigma^{\beta}_{t|t-1} z'_t)^{-1} z_t \Sigma^{\beta}_{t|t-1} \tag{26}$$

Where the term $\Sigma^{\beta}_{t|t-1} z'_t (\hat{Q}_t + z_{t-1}\Sigma^{\beta}_{t|t-1} z'_{t-1})^{-1}$ is the Kalman Gain for each time period $t$ and is set to zero for the variables missing at each step.

The only outstanding terms that need defining are the measurement equation error covariance matrices that can be obtained using EWMA as follows

$$\hat{V}_t = \kappa_1 \hat{V}_{t-1} + (1-\kappa_1)\hat{u}_t \hat{u}'_t, \tag{27}$$

$$\hat{Q}_t = \kappa_2 \hat{Q}_{t-1} + (1-\kappa_2)\hat{\varepsilon}_t \hat{\varepsilon}'_t. \tag{28}$$

where $\kappa_1$ and $\kappa_2$ are forgetting factors that define the law of motion of the idiosyncratic volatilities in the measurement equation and the volatilities of the observable variables and the factors in the state equation. We set these quantities to $\kappa_1 = \kappa_2 = 0.99$.

**2. Kalman Smoother**

The kalman filter algorithm described in (1.1)-(1.3) above works by forward recursion and outputs estimates of $E(\theta|D^t)$ for all parameters $\theta = [\hat{V}_t, \hat{Q}_t, \hat{R}_t, \hat{W}_t, \hat{\beta}_t, \hat{\Lambda}_t]$ in the model using the data available $D^t$ for $t = 1, ..., t$. However, we are ultimately interested in an estimate of $E(\theta|D^T)$ which yields the parameter states conditional on the entire sample $t = 1, ..., T$. Therefore, the kalman smoother is applied to the output of the kalman filter working by backward recursion. Given that the Kalman filter takes into account missing data, no alteration is necessary to the Kalman smoother.

**3. Kalman Smoother/filter for Factors**

The full three step procedure described in section 4.2 is complete by applying the Kalman filter and smoother algorithm to the factors, which are initialized with a PCA estimate. The algorithm and mixed frequency extensions are analogous to that previously described.

## A.2 PCA algorithms for Incomplete Data

### A.2.1 Least-Squares PCA in the presense of missing data

Several simple ways to deal with missing values in a classical LS PCA framework consist in setting missing values to zero or using an Expectation Maximization PCA (EM PCA). The later technique is used by Stock and Watson (2002) and consists of an iterative procedure that alternates between imputing missing values in $X$ (E-step) and applying standard PCA to the pseudo-balanced panel of data (M-step) until convergence is reached. To summarize the algorithm proceeds as follows:

- **E-step:** Reconstruct $X_t$ by filling in its missing values:

$$X_t^* = \begin{cases} X_t & \text{for observed values} \\ \hat{\Lambda}^k \hat{F_t^k} & \text{for missing values.} \end{cases} \tag{29}$$

- **M-step:** Perform standard PCA by SVD on the infilled matrix $X_t^*$ and obtain new values for $\{\hat{\Lambda^k}, \hat{F_t^k}\}$.

The algorithm alternates between the E-M stFCI until convergence is reached in which case new estimates for the parameters in iteration $k-1$ do not improve the Least-squares minimization problem solved in iteration $k$.

### A.2.2 Variation view of the Expectation Maximization (EM) algorithm for incomplete data environments

Consider the standard PC regression

$$X_t = \Lambda F_t + \xi_t, \quad \xi_t \sim N(0, v_x I). \tag{30}$$

Where $\theta = \{\Lambda, F_t, v_y\}$ are model parameters and a subset $X_{mis}$ of the data matrix is missing and treated as hidden variables. The variational view of the EM algorithm (see Neal and Hinton (1999) and Attias (2000)) consists in minimizing the objective function

$$V(\theta, p(X_{mis})) = \int p(X_{mis}) log \frac{p(X_{mis})}{p(X|\theta)} dX_{mis} = \tag{31}$$

$$\int p(X_{mis}) log \frac{p(X_{mis})}{p(p(X_{mis}|\theta)} dX_{mis} - logp(X_{obs}|\theta), \tag{32}$$

wrt the model parameters $\theta$ and the density over the missing data $p(X_{mis})$. $X_{obs}$ denotes the observed data such that $X = X_{mis} \bigoplus X_{obs}$.

**E-step**. Equation A26 is the Kullback-Leibler divergence between the pdfs over observable and unobservable data. The minimization of this expression wrt $p(X_{mis})$, given $\theta$ is shown to yield

$$p(X_{mis}|\theta) = \prod_{ij \in O} N(\hat{x}(\theta)_{ij}, v_x). \tag{33}$$

where $O$ is the set of indices for which observation $x_{ij}$ is observed, $\hat{x}(\theta)_{ij}$ result from the reconstruction of the incomplete data matrix $X$ from expression A24, for a given $\theta$. This procedure is refered to as the E-step of the algorithm.

**M-step**. Next, the proceedings from the E-step are substituted back into expression (A26). The terms in the resulting expression which depend on $\theta$ are given by

$$-\int p(X_{mis}) logp(X|\theta) dX_{mis}. \tag{34}$$

It can be shown that minimization of (A28) wrt. $\theta$ is equivalent to minimizing the LS objective function in case of no missing data (Neal and Hinton, 1999). Thus, the M-step of the algorithm consists in performing SVD decomposition to the imputed data matrix $X$. The algorithm alternatives between the E-M stFCI until convergence is reached (ie, when the reconstruction error stabilizes).

### A.2.3 Probabilistic PCA (PPCA)

A probabilistic PCA specification has been found to provide a good foundation to handle missing data Ilin and Raiko (2010). The probabilistic PCA set forth by Tipping and Bishop (1999b) can be written as follows

$$X_t = m + \Lambda F_t + \xi_t, \tag{35}$$

where both the principal component and the noise term are assumed normaly distributed as follows

$$p(F_t) \sim N(0, I_K), \tag{36}$$

$$p(\xi_t) \sim N(0, \tau^{-1} I_N), \tag{37}$$

where $\theta = \{m, \Lambda, \tau\}$ are model parameters, $I_N$ and $I_T$ denote identity matrices and $\tau$ is the scalar inverse variance of $\xi$. It can be shown that the Maximum Likelihood (ML) estimation of the PPCA is identical to PCA in the case of non-missing data. The great advantage of the PPCA is that, in case of incomplete data, it allows for regularization that arises naturally from the choice of Gaussian priors. The model can then be estimation with a standard EM algorithm. The necessary extensions to handle missing data are discussed in Ilin and Raiko (2010). Below I summarize their procedure.

**E-step.** Estimate the conditional distribution of the hidden variables $F$ given the data $X$ and model parameters $\theta$,

$$p(F|X, \theta) = \prod_{j=1}^{K} N(\bar{F}_j, \Sigma_{F_j}), \tag{38}$$

based on the following updating rules

$$\Sigma_{F_j} = \tau^{-1}(\tau^{-1} I + \sum_{i \in O_j} \lambda_i \lambda_i^T)^{-1}, \tag{39}$$

$$\bar{F}_j = \tau \Sigma_{F_j} \sum_{i \in O_j} \lambda_i (x_{ij} - m_i), j = 1, ..., K, \tag{40}$$

$$m_i = \frac{1}{|O_i|} \sum_{j \in O_i} (x_{ij} - \lambda_i^T \bar{F}_j). \tag{41}$$

**M-step**. re-estimate the model parameters as

$$\lambda_i = \big( \sum_{j \in O_i} \bar{F}_j \bar{F}_j^T + \Sigma_{F_j} \big)^{-1} \sum_{j \in O_i} \bar{F}_j (x_{ij} - m_i), i = 1, ..., N, \tag{42}$$

$$\tau = \Big[ \frac{1}{N} \sum_{ij \in O} \big[ x_{ij} - \lambda_i^T \bar{x}_j - m_i \big]^2 + \lambda_i^T \Sigma_{F_j} \lambda_i \Big]^{-1}. \tag{43}$$

where $O_i, O_j$ and $O$ denote the set of indices $i, j$ for which $x_{ij}$ is observed.

### A.2.4 Variational Bayesian PCA (VBPCA)

Some studies suggest that the standard PPCA is still vulnerable to over-fitting (see for example Ilin and Raiko (2010)). One possible reason that might lead to overfitted solution might be the nontrivial choice of the number of principal components to include in (A30). Including a large number of common components $F_t$ might cause the model to over-learn the data.

One possible solution to this problem consists in penalizing large values in the matrices $\Lambda$ and $F_t$. The probabilistic pca model is flexible enough to allow for an automatic, data-driven selection of relevant common components by shrinking to zero the solutions $\lambda_j$ that are small relative to the noise variance. This can be achieved through a variational bayesian PCA algorithm as explained below. We follow Oba et al. (2003) in imposing additional regularization to penalize parameter values that yield more complex explanations of the data. Hence, in addition to (A31)-(A32) one can further impose

$$p(m|\tau) \sim N(0, (\gamma_{m0}\tau)^{-1} I_T), \tag{44}$$

$$p(\lambda_j|\tau, \alpha_j) \sim N(0, (\alpha_j \tau)^{-1} I_T), \tag{45}$$

$$p(\tau) \sim G(\tau|\bar{\tau}_0, \gamma_{\tau_0}), \tag{46}$$

where $\psi = \{\gamma_{m0}, \gamma_{\tau_0}, \bar{\tau}_0, \alpha_j\}$ are hyperparameters and $\lambda_j$ are the parameters in column $j$ of the loadings matrix $\Lambda$ that define the importance of each principal component $F_j$, $j = \{1, 2, ..., K\}$. The prior $p(\Lambda|\alpha, \tau)$, which has a hierarchical structure, is called an automatic relevance determination (ARD) prior. This structure plays a key role in guaranteeing parsimony of the model. Its variance $(\alpha_j \tau)^{-1}$ is determined by a hyperparameter $\alpha_j$ that becomes large when the euclidean distance $||\lambda_j||$ is small relative to the noise variance $\tau^{-1}$.

Estimation of the model now requires a variational EM algorithm as proposed by Attias (2000) to cope with the unknown analytical form of the posterior of the parameters $p(\Lambda, F_t, m|X, \psi)$, which invalidates the E-step of the standard EM algorithm. To overcome this difficulty, the author proposes an approximation to this this quantity by a simpler $q(\Lambda, F_t, m)$. Using a variational approach, the E-step is modified such that the objective function approximates $p(\theta|X)$ with a simpler pdf $p(\theta)$, written as follows

$$V(p(\theta), \psi)) = \int p(\theta) log \frac{p(\theta)}{p(X, \theta|\psi)} d\theta = \tag{47}$$

$$\int p(\theta) log \frac{p(\theta)}{p(X, \theta|\psi)} d\theta - log p(X|\psi), \tag{48}$$

**E-step**. Equation (A33) is the Kullback-Leibler divergence between the true posterior and its approximation. In this step the approximation $p(\theta)$ is updated. This corresponds to minimizing this distance wrt $p(\theta)$.

**M-step**. Next, the approximation $p(\theta)$ is used as if it was the actual posterior $p(\theta|X, \psi)$ in order to increase $p(X|\psi)$. This consists in deriving the expression (A33) wrt $\psi$.

The algorithm alternatives between the E-M stFCI until convergence is reached (i.e. when the reconstruction error stabilizes).

## B Additional Tables and Figures

| # | mnemonic | description | frequency | Sample Start | t-code | Source |
|---|---|---|---|---|---|---|
| 1 | GDPC1 | Real Gross Domestic Product SA. Annual Rate | Q | 1971-01-01 | 5 | St. Louis FRED |
| 2 | CPIAUCSL | Consumer Price Index for All Urban Consumers | M | 1971-01-01 | 5 | St. Louis FRED |
| 3 | UNRATE | Unemployment Rate, Percent, SA. | M | 1971-01-01 | 5 | St. Louis FRED |
| 4 | DFF | Effective Federal Funds Rate | D | 1971-01-01 | 1 | St. Louis FRED |
| 5 | CMDEBT | Households and Nonprofit Organizations Debt % GDP | Q | 1971-01-01 | 5 | St. Louis FRED |
| 6 | ABSITCMAHDFS | Issuers of Asset-Backed Securities % GDP | Q | 1983-07-01 | 5 | St. Louis FRED |
| 7 | DRTSCILM | Net Percentage of Domestic Banks Tightening Standards for Commercial and Industrial Loans | Q | 1990-04-01 | 1 | St. Louis FRED |
| 8 | TERMCBAUTO48NS | Finance Rate on Consumer Installment Loans at Commercial Banks, New Autos 48 Month Loan, Percent | Q | 1972-01-01 | 5 | St. Louis FRED |
| 9 | TDSP | Household Debt Service Payments as a Percent of Disposable Personal Income | Q | 1980-01-01 | 1 | St. Louis FRED |
| 10 | TOTALSL | Total Consumer Credit Owned and Securitized % GDP | M | 1971-01-01 | 1 | St. Louis FRED |
| 11 | LOANHPI | Loan Performance Index U.S. | M | 1976-03-01 | 5 | Bloomberg |
| 12 | CONSEXFI | UMich Expected Change in Financial conditions | M | 1978-02-01 | 1 | Uni Michigan |
| 13 | BPLR | Bank Prime Loan Rate / Libor spread | M | 1971-01-01 | 1 | St. Louis FRED |
| 14 | JPMNEER | JPMorgan Broad Nominal Effective Exchange Rate (2010=100) | M | 1971-01-01 | 5 | Bloomberg |
| 15 | LDR | All Commercial Banks Loan to Deposit Ratio | M | 1973-01-01 | 1 | Haver Analytics |
| 16 | 2/3TBS | 2yr/3m Treasury bill spread | M | 1976-06-01 | 1 | St. Louis FRED |
| 17 | MORTGAGE30US | Mortgage rate / 10yr Treasury Bill spread | W | 1971-04-02 | 1 | St. Louis FRED |
| 18 | T10Y2Y | 10-Year Minus 2-Year Treasury Constant Maturity yield, Percent | D | 1976-06-01 | 1 | St. Louis FRED |
| 19 | BAMLH0A0HYM2EY | ICE BofAML US High Yield Master II Effective Yield, Percent | D | 1996-12-31 | 1 | Bloomberg |
| 20 | MOVE Index | Yield curve weighted index of the normalized implied volatility on 1-month Treasury options. | D | 1988-04-04 | 1 | Bloomberg |
| 21 | CRY Index | Thomson Reuters/CoreCommodity CRB Commodity Index | D | 1994-01-03 | 1 | Bloomberg |
| 22 | VXOVIX | Cboe S&P 100/500 Volatility Index | D | 1990-01-02 | 1 | St. Louis FRED |
| 23 | BASPTDSP | Ted Spread | D | 2001-01-02 | 1 | St. Louis FRED |
| 24 | WILL5000PRFC | Wilshire 5000 Full Cap Price Index | D | 1971-01-01 | 5 | St. Louis FRED |
| 25 | CPFF | 3-Month Commercial Paper Minus Federal Funds Rate, Percent, Daily, Not Seasonally Adjusted | D | 1997-01-02 | 1 | St. Louis FRED |
| 26 | SP500 | S&P 500 price index | D | 1971-01-01 | 5 | St. Louis FRED |

**Notes**: Mnemonic refers to the statistical reference with which the time series can be fetched from the source. Frequency is either Q: quarterly, M: monthly, W: weekly or D: Daily. Sample Start date refers to the first observation for a specific time-series in our sample period 1971-2020. t-code refers to transformation applied to each variable. 1:levels; 5: log-differences.

Table 2: Description of the data used to construct the Financial Conditions Index.

Figure 5: FCI calculated with different signal extraction methods within the FAVAR. The four methods include the PCA - simple Principle Components, EM-PCA which stands for Expectation Maximization PCA, PPCA - Probabilistic PCA and VBPCA - Variational Bayes PCA.

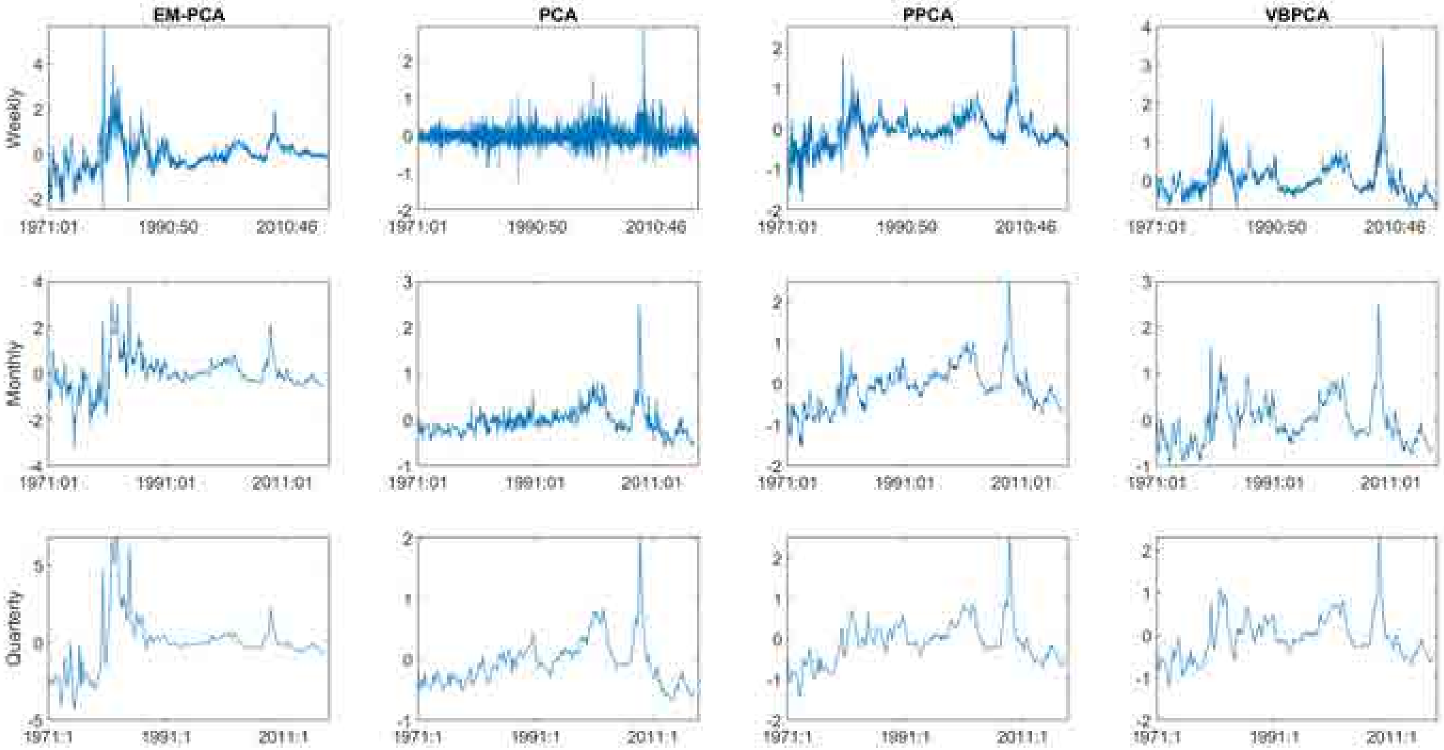
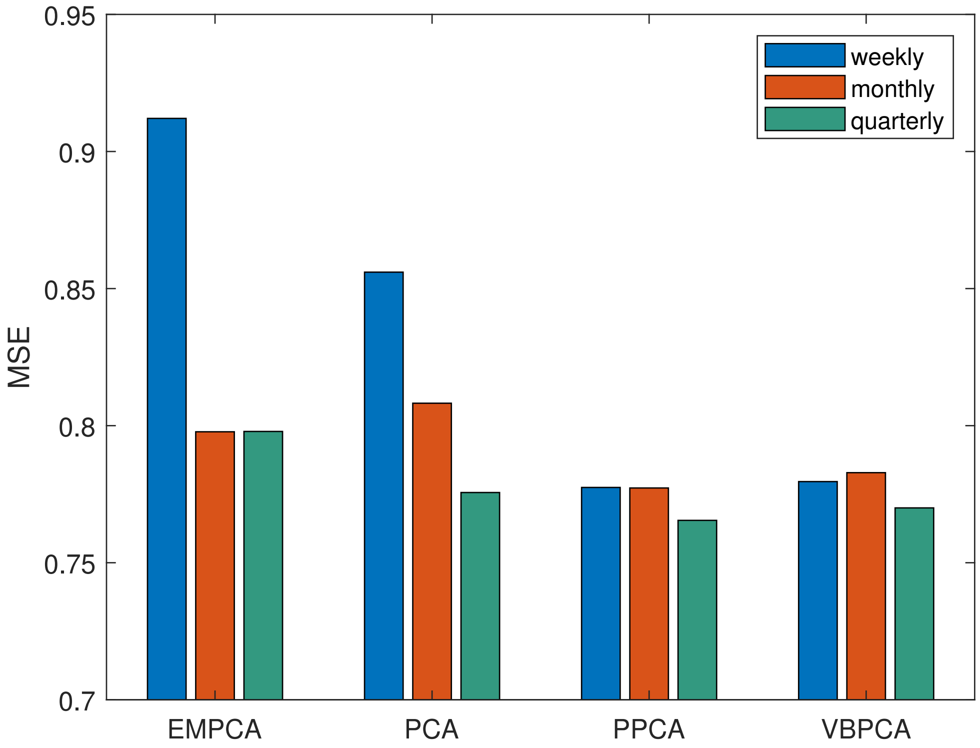


Figure 6: Reconstruction Mean Squared Error across four alternative PCA methods



**Notes**: Mean-squared error statistics are calculated in-sample $MSE = \frac{1}{N}\sum_{ij \in O}(x_{ij} - \hat{x}_{ij})^2$, where $O$ includes all indice for which $x_{ij}$ is observed and $\hat{x}_{ij}$ results from the projection of $X_t$ on the factors $F_t$ in equation 2.1, permitting the reconstruction of the incomplete data set of financial variables that loan onto the FCI. for the observable section of the standardized unbalanced panel of financial indicators across the different signal extraction methods.