# Essays on asynchronous time series and related multidimensional data

Filippo Pellegrino

Department of Statistics

London School of Economics and Political Science

A thesis submitted for the degree of *Doctor of Philosophy*

May 2022

# Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

# Statement of co-authored work

I confirm that Chapter 3 was jointly co-authored with Professor Matteo Barigozzi and I have contributed 80% of this work.

# Abstract

This thesis focusses on asynchronous time series and related multidimensional data: time-dependent measurements with varying publication delays. This class of data exists in a broad range of fields. In social sciences, most official time series and repeated surveys are indeed asynchronous in nature since statistical offices need time to collect and aggregate raw data. In STEM, statistical offices are generally less relevant and most publication delays are caused by more exotic factors. For instance, with series derived from technological networks, they are usually generated by a direct reference (digital or textual) of the past (e.g., publishing pictures of a trip done a week ago that was also photographed and posted in real time by a friend). As a result, the study of data releases is key for developing accurate real-time models and finds applications in forecasting, policy and risk management.

# Contents

# Preface

This thesis is a collection of three articles focussing on asynchronous time series and related multidimensional data: time-dependent measurements with varying publication delays. This class of data exists in a broad range of fields. In social sciences, most official time series and repeated surveys are indeed asynchronous in nature since statistical offices need time to collect and aggregate raw data. In STEM, statistical offices are generally less relevant and most publication delays are caused by more exotic factors. For instance, with series derived from technological networks, they are usually generated by a direct reference (digital or textual) of the past (e.g., publishing pictures of a trip done a week ago that was also photographed and posted in real time by a friend). As a result, the study of data releases is key for developing accurate real-time models and finds applications in forecasting, policy and risk management.

The first article approaches the *fil rouge* of this dissertation by focussing on hyperparameter selection for forecasting models based on potentially incomplete time series. Indeed, hyperparameter selection precedes any analysis and should be properly done to control, for instance, the tendency of over-fitting in-sample, but performing poorly out-of-sample. Even though there are methods for selecting hyperparameters for dependent data problems, they are usually limited in scope by the underlying subsampling methods. I have proposed to overcome the problem by employing a generalisation of the delete-$d$ jackknife (Wu, 1986; Shao and Wu, 1989) in which the data removal step is replaced with a fictitious deletion that consists in imposing (artificial) patterns of missing observations on the data. This allows to have plain compatibility with time-series problems while retaining the efficiency of subsampling methods for independent data.

The second article covers a different, but equally fundamental topic: bridging the gap between traditional forecasting models and machine learning in order to exploit non-linear dynamics. In particular, it proposes to extend the information set of time-series regression trees with latent stationary factors extracted via state-space methods. In doing so, this approach generalises time-series regression trees on two dimensions. First, it allows to handle predictors that exhibit measurement error, non-stationary trends, seasonality and/or irregularities such as missing observations. Second, it gives a transparent way for using domain-specific theory to inform time-series regression trees. As a byproduct,

this technique sets the foundations for structuring powerful ensembles. Their real-world applicability is studied in an empirical application: a set of key macroeconomic indicators is modelled via a state-space representation informed by economic theory and a latent stationary factor interpretable as the business cycle is then used for predicting equity volatility. Results show that these factor-augmented tree ensembles outperform all benchmark methods in most cases of interest.

Finally, the third article focusses on more complex data. In spatio-temporal filtering and smoothing, the space-related information allows computing predictions that are specific to each region included in the dataset, for each point in time. For instance, with a meteorological dataset, it is possible to compute predictions for different cities, conditioning on the surrounding environment, winds and likelihood of a storm moving from one city to another. This can be done since the space-dimension gives precise ordering on a map. However, with higher dimensional problems this is not as straightforward. This paper proposes a solution that involves extracting interpretable dynamic factors over multiple dimensions and time. Results are specialised to model microeconomic data on US households jointly with macroeconomic aggregates. This approach allows to generate localised predictions, counterfactuals and impulse response functions for individual households, accounting for traditional time-series complexities. The model is also compatible with the growing focus of policymakers for real-time economic analysis as it is able to process observations online, while handling missing values and asynchronous releases.

# Notation

This brief description provides a concise reference on the mathematical notation I have used throughout the thesis. More specialised notation is described directly in the text when deemed necessary.

ASYMPTOTIC THEORY, STATISTICS AND PROBABILITY. The expected value and probabilities are indicated with the symbols $\mathbb{E}$ and Pr. $M_1, M_2, \ldots$ denote generic positive and finite constants (unless otherwise stated).

MATRIX NOTATION. Matrices, vectors and vector-valued functions are written using bold symbols (or, bold notation). The $(i, j)$-th element of a generic matrix $\mathbf{A}$ is denoted as $A_{i,j}$. The transpose of $\mathbf{A}$ is indicated as $\mathbf{A}'$. The notation $\mathbf{A}_{1:i,1:j}$ is used for referring to the matrix built by taking the first $i$ rows and $j$ columns of $\mathbf{A}$. The $(k, k)$ entry-wise matrix norm of $\mathbf{A}$ is denoted as $\|\mathbf{A}\|_{k,k} = (\sum_i \sum_j |A_{i,j}|^k)^{1/k}$, while the Frobenius norm $\|\mathbf{A}\|_{\mathrm{F}} \equiv \|\mathbf{A}\|_{2,2}$. The Euclidean norm of a vector is denoted with standard notation. The vectorised and half-vectorised versions of $\mathbf{A}$ are indicated with $\mathrm{vec}(\mathbf{A})$ and $\mathrm{vech}(\mathbf{A})$. $\mathrm{vec}(\mathbf{A})$ and $\mathrm{vech}(\mathbf{A})$ are column vectors. Finally, the symbol $\odot$ is used for denoting the Hadamard (or element-wise) product.

SET NOTATION. The calligraphic alphabet is used for denoting sets only. Standard notation is used for number sets, intervals and operations.

SPECIAL SYMBOLS. $\mathbf{0}_{j\times k}$ and $\boldsymbol{\iota}_{j\times k}$ denote $j \times k$ matrices of zeros and ones. $\mathbf{I}_k$ indicates a $k \times k$ identity matrix. $\mathbb{I}$ denotes an indicator function equal to one when the condition in its subscript is verified (zero otherwise).

USE OF BRACKETS. For grouping, the preferred sequence of brackets in this thesis is $\{[()]\}$. For sets and intervals, I have used $\{\}$ and $[()]$. For composite functions I have denoted in bold the outer parenthesis. For instance, $g\boldsymbol{(}f(...)\boldsymbol{)}$ for some generic functions $f$ and $g$.

# Acknowledgements

First, I would like to thank Professor Matteo Barigozzi and Dr Kostas Kalogeropoulos. They have been the very best PhD supervisors one can hope to have. Their sustained encouragement and suggestions have been instrumental to develop the ideas contained in this thesis.

I am also grateful to the London School of Economics and Political Science for making the last few years so pleasant, interesting and rewarding. Special thanks go to staff, colleagues and friends including Professor Rita Astuti, Dr Yining Chen, Dr Huang Feng, Gianluca Giudice, Penny Montague, Dr Alice Pignatelli di Cerchiara, Dr Xinghao Qiao, Dr Yan Qu, Ragvir Sabharwal, Dr Tianlin Xu, Professor Qiwei Yao and Dr Xiaolin Zhu. Moreover, I am thankful to external academics, co-authors and researchers including Dr Paolo Andreini, Thomas Hasenzagl, Dr Cosimo Izzo, Dr Chiara Perricone, Dr Giovanni Ricco and Professor Lucrezia Reichlin for their helpful comments on the first chapter.

Finally, I would like to thank my family and Serena Lariccia to whom I am deeply indebted for their constant patience, understanding and everlasting support.

# 1 Selecting time-series hyperparameters with the artificial jackknife

*This article proposes a generalisation of the delete-d jackknife to solve hyperparameter selection problems for time series. I call it artificial delete-d jackknife to stress that this approach substitutes the classic removal step with a fictitious deletion, wherein observed datapoints are replaced with artificial missing values. This procedure keeps the data order intact and allows plain compatibility with time series. This manuscript justifies the use of this approach asymptotically and shows its finite-sample advantages through simulation studies. Besides, this article describes its real-world advantages by regulating high-dimensional forecasting models for foreign exchange rates.*

## 1.1. Introduction

Using large datasets with standard predictive models is not straightforward. There is often a proliferation of parameters, high estimation uncertainty and the tendency of over-fitting in-sample, but performing poorly out-of-sample. This so-called curse of dimensionality is often handled regularising statistical models with a collection of tuning parameters. Since the latter are often determined before the estimation process takes place, they are denoted as hyperparameters. This paper proposes a systematic approach for selecting them in the case of time-series data.

There is a large number of techniques for high-dimensional prediction problems. Classical methods include ridge (Hoerl and Kennard, 1970), LASSO (Tibshirani, 1996) and elastic-net (Zou and Hastie, 2005) regressions. They make the estimation feasible for linear regressions by penalising the magnitude of the coefficients to downweight the variables that do not help in predicting. The strength of the penalties is tuned with a vector of hyperparameters. Regression trees (Morgan and Sonquist, 1963; Breiman et al., 1984; Quinlan, 1986) are a classical example from the machine learning literature for exploring high-dimensional datasets. These techniques can handle non-linearities and complex data generating processes. However, they must be regulated via a range of penalties and stopping rules to perform well out-of-sample. This is again achieved using hyperparameters.

Large datasets are also commonly handled with Bayesian methods. In this literature, hyperparameters are often necessary to define prior distributions (Gelman et al., 2014) and obtain parsimonious models with shrinkage techniques similar or equivalent to ridge and LASSO (Giannone et al., 2017), and the elastic-net (Li and Lin, 2010). Hyperparameters are also crucial for low-dimensional problems. For example, hyperparameters such as the number of lags for autoregressive models are fundamental for structuring forecasting exercises.

Cross-validation (Stone, 1974) is among the most well-known approaches for selecting hyperparameters in independent data settings. It is a statistical method to estimate the expected accuracy of a model on unseen data. Its basic formulation is straightforward: data is split into complementary partitions, and the resulting subsamples are used for estimating and validating a predictive method. The performance within the validation samples is used as an estimate of the prediction error on unseen datapoints and the hyperparameters are generally selected to minimise this measure.

Cross-validation is challenging for time series since data is ordered and autocorrelated. Several authors have proposed generalisations to handle these complexities. One of the first contributions came from Snijders (1988). The latter used insights from Brown et al. (1975) and Ljung and Söderström (1983) to propose a cross-validatory method based on realised pseudo out-of-sample errors. Indeed, it suggested to split the observed data into complementary partitions and then use the first as an estimation sample, and the remaining observations to measure the realised pseudo out-of-sample error. The hyperparameters are selected to minimise this error measure.

While this approach is very intuitive and consistent with the structure of the data, it is not necessarily robust, since it uses only a single estimation and validation set. Kunst (2008) proposed overcoming this downside by applying standard pseudo out-of-sample evaluations to random subsamples. However, the results are relatively difficult to interpret since the algorithm used for generating these partitions is initialised with in-sample regression parameters.

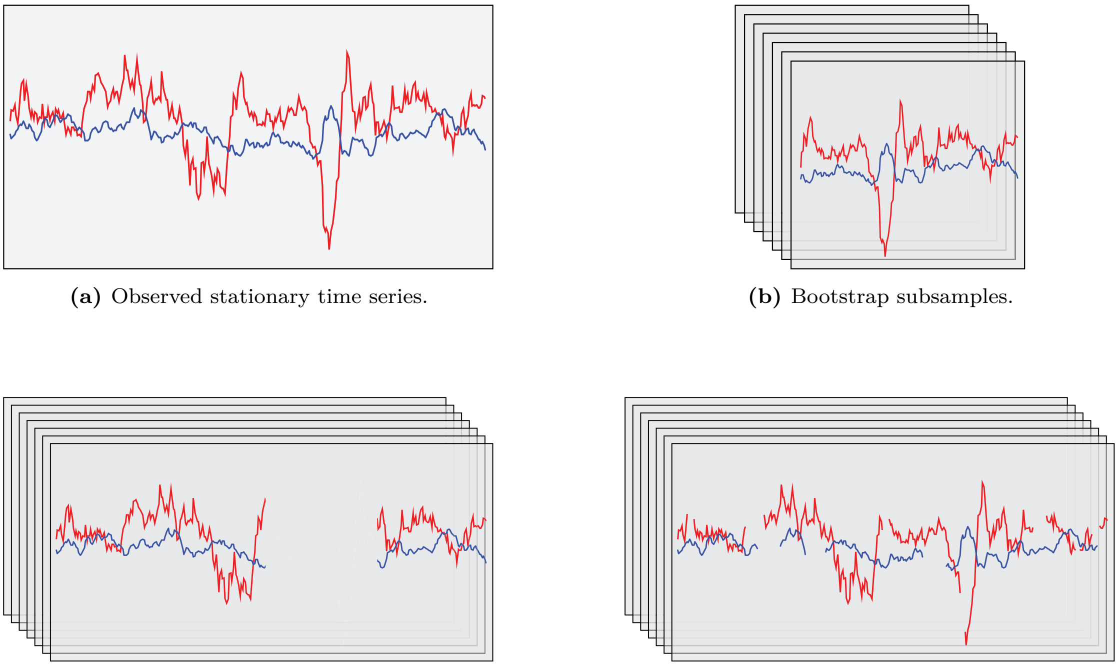
Burman et al. (1994) introduced a different way to address this problem: the so-called $h$-block cross-validation. This methodology, based on Györfi et al. (1989) and Burman and Nolan (1992), uses blocking techniques to generate validation samples independent from the data used for estimation. Indeed, Burman et al. (1994) proposed creating a set of estimation samples by removing, in turn, each block of dimension $2h + 1$ (for a given $h$) from the data. $h$-block cross-validation then uses the median item of this block as a one-dimensional validation sample. Even though this approach has interesting properties, keeping a fixed distance between the partitions is costly, given that a large share of observations is lost in the process. This is especially severe when there are not so many observations, because the number of validation samples available is small.

Most recently, Bergmeir et al. (2018) proposed cross-validating autoregressive models with uncorrelated errors with techniques for i.i.d. data. This approach makes good use of all available observations, but its properties do not hold for models with correlated errors and it disregards the order in the data.

Jackknife (Quenouille, 1956; Tukey, 1958) and bootstrap (Efron, 1979a,b, 1981) can be used as alternative approaches to estimate the prediction error on unseen datapoints and thus select hyperparameters. These techniques are typically more efficient than cross-validation (Efron, 1979a; Efron and Gong, 1983) since they measure the accuracy of a model on the average prediction error committed over a large range of data subsamples. Bootstrap builds these partitions sampling with replacement from the data. Instead, jackknife constructs subsamples by removing sets of observations from the observables. In particular, the delete-$d$ jackknife (Wu, 1986; Shao and Wu, 1989) generates a sequence of partitions by removing, in turn, all the combinations of $d > 0$ observations from the data.

Jackknife and bootstrap require modifications to be compatible with time series, since the subsampling schemes do not take the data order into account. Kunsch (1989) extended these methodologies to stationary series. Indeed, building on Carlstein (1986), Kunsch (1989) proposed developing block-wise subsampling schemes. Let $c$ be an integer lower or equal to the total number of observed time periods. The block jackknife generates the partitions by removing or down-weighting, in turn, all the $c$-dimensional blocks of consecutive observations from the data. Instead, the block bootstrap draws with replacement a fixed number of $c$-dimensional blocks of observations from the data. Politis and Romano (1992, 1994) developed this technique further proposing the so-called stationary bootstrap. This approach wraps the data “in a circle”, so that the first observation follows the last, and generates the bootstrap samples drawing and merging blocks of random length. Differently than the block bootstrap and its variations, the block jackknife does not impact the data order when constructing subsamples.

This paper introduces a version of the standard delete-$d$ jackknife compatible with time series. In this version of the jackknife, the data removal step is replaced with a fictitious deletion that consists in imposing (artificial) patterns of missing observations on the data. I call this new approach artificial delete-$d$ jackknife (or artificial jackknife) to emphasise that $d$ observations are artificially removed from the original data to generate each subsample. This article proposes using this new methodology to compute a robust measure of the forecast error (or, artificial jackknife error) as a means for selecting hyperparameters. The advantages of this approach depend on the finite-sample properties of the artificial jackknife. In fact, all errors based on pseudo out-of-sample evaluations converge in probability to the true error with the same rate (as shown in section 1.B). However, the artificial jackknife error has a smaller finite-sample variance than the pseudo



**(a)** Observed stationary time series.

**(b)** Bootstrap subsamples.

**(c)** Block jackknife subsamples.

**(d)** Artificial delete-$d$ jackknife subsamples.

**Figure 1.1:** Subsampling schemes for dependent data. (b) Blocks of random (stationary bootstrap) or fixed (block bootstrap) length are drawn with replacement from the data. (c) Subsamples are constructed down-weighting, in turn, all the $c$-dimensional blocks of consecutive observations from the data. As in section 1.2.2, the down-weighting scheme is operated by turning blocks of consecutive observations into missing values. (d) Subsamples are constructed imposing (artificial) patterns of missing data to the original sample. This is a generalisation of the delete-$d$ jackknife.

out-of-sample error and the block jackknife (for most configurations of $c$ and $d$). This is crucial for stability and to select hyperparameters when the number of observations (i.e., time periods) is limited.

The artificial delete-$d$ jackknife is compatible with forecasting models able to handle missing observations. Within the scope of this paper, this is not a strong restriction. Most predictive problems with missing observations in the measurements can be written in state-space form and estimated via a large number of methods, as surveyed in Shumway and Stoffer (2011, ch. 6) and Särkkä (2013, ch. 12).

As an illustration, this article employs the artificial jackknife for tuning vector autoregressive moving average (VARMA) models regulated via an elastic-net penalty (Zou and Hastie, 2005). These models are estimated on a high-dimensional dataset of weekly exchange rate returns. In order to provide full compatibility with the artificial jackknife, this article proposes to estimate the VARMAs with an Expectation-Conditional Maximisation (ECM) algorithm (Meng and Rubin, 1993) able to handle incomplete time series. This estimation method is a secondary contribution of the paper given that, to my best knowledge, the literature has not proposed a way for handling missing observations in the measurements with similar settings.[1]

[1] The replication code for this empirical application is available on GitHub.

## 1.2. Methodology

One of the main objectives of time series is to predict the future. This article aims to select optimal vectors of hyperparameters consistently with this maxim and thus in a way that minimises the expected forecast error.

### 1.2.1. Foundations

This subsection sets out the foundations for the hyperparameter selection process and delimits the scope of the article to a broad family of forecasting methods that encompasses common techniques such as ARMA, ADL and VARMA models.[2]

**Assumption 1** (Data). Let $n, T \in \mathbb{N}$ and $n_Z \in \mathbb{N}_0$. Assume that $Y_{i,t}$ and $Z_{j,t}$ are finite realisations of some real-valued stochastic processes observed at time periods in the sets $\mathscr{T}_i, \mathscr{T}_j \subseteq \{t : t \in \mathbb{Z},\, 1 \leq t \leq T\}$ for $i = 1, \dots, n$ and $j = n+1, \dots, n+n_Z$.

**Assumption 2** (Lags). Define $q, r \in \mathbb{N}_0$ to be such that $p := \max(q, r)$ and $0 < p \ll T-1$.

**Assumption 3** (Predictors). Let $\mathbf{X}_t := (\mathbf{Y}'_t \; \dots \mathbf{Y}'_{t-q+1} \; \mathbf{Z}'_t \; \dots \mathbf{Z}'_{t-r+1})'$ be $m \times 1$ and defined at any point in time $t \in \mathbb{Z}$.

**Assumption 4** (Model structure). Finally, assume that

$$\mathbf{Y}_{t+1} = \mathbf{f}(\mathbf{X}_t, \mathbf{\Psi}) + \mathbf{V}_{t+1}, \tag{1.1}$$

where $\mathbf{f}$ is a finite function, $\mathbf{\Psi}$ is a matrix of finite coefficients, $\mathbf{V}_{t+1} \overset{i.i.d.}{\sim} (\mathbf{0}_{n\times 1}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ being a positive definite matrix[3] and $\mathbb{E}(\mathbf{V}_{t+1}|\mathbf{X}_t) = 0$, for any integer $t$.[4]

**Remark.** The dependence on the sample size is highlighted in the notation only when strictly necessary, in order to ease the reading experience. Also, this article uses the same symbols to indicate the realisations at some integer point in time $t$ and their general value in the underlying process. This is again for simplifying the notation and it should be clear from the context whether the manuscript is referring to the first or second category.

Knowing the data generating process in assumption 4, one could use it for obtaining the most accurate prediction (true forecast) for $\mathbf{Y}_{t+1}$ given $\mathbf{X}_t$ at any point in time $t$.

[2] Please note that this subsection and section 1.2 in its entirety do not limit the manuscript by looking at a specific forecasting model. Hence, the theoretical results are widely applicable.

[3] This part of assumption 4 could be relaxed following an approach similar to the one employed in Barigozzi and Luciani (2020). However, this is outside the scope of the paper.

[4] Under assumptions 1–3, $\mathbf{X}_t$ is allowed to include $\mathbf{Y}_t, \dots, \mathbf{Y}_{t-q+1}$ and $\mathbf{V}_t, \dots, \mathbf{V}_{t-r+1}$ (for some $0 \leq q \leq p$ and $0 \leq r \leq p$), and more explanatory variables referring up to time $t-p+1$.

**Definition 1** (True error). Under a weighted square loss, the expected error associated with the true forecast is

$$err := \sum_{i=1}^{n} w_i \, \mathbb{E}\left[ |Y_{i,t+1} - f_i\left(\mathbf{X}_t, \mathbf{\Psi}\right)|^2 \right] = \sum_{i=1}^{n} w_i \, \mathbb{E}(V_{i,t+1}^2) = \sum_{i=1}^{n} w_i \, \mathbf{\Sigma}_{i,i},$$

with $w_i \geq 0$ for $1 \leq i \leq n$. This article refers to *err* as the true error.

In most practical applications, the data generating process is unknown and forecasters' objective can be then reduced to approximating the true forecast.

**Assumption 5** (Information set). Formally, at any time period $p \leq s \leq T$, forecasters have an information set $\mathscr{I}(s)$ containing the data observed up to that point and their expectation for $\mathbf{Y}_{t+1}$ conditional on $\mathscr{I}(s)$ is

$$\hat{\mathbf{Y}}_{t+1|s}(\boldsymbol{\gamma}) := \mathbb{E}\left[ \mathbf{Y}_{t+1} | \mathbf{X}_t, \hat{\boldsymbol{\theta}}_s(\boldsymbol{\gamma}) \right] = \mathbf{g}(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_s(\boldsymbol{\gamma})),$$

where $\mathbf{g}(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_s(\boldsymbol{\gamma}))$ is a finite function whose coefficients $\hat{\boldsymbol{\theta}}_s(\boldsymbol{\gamma})$ are also finite and estimated on the basis of the data in $\mathscr{I}(s)$, and given a vector of hyperparameters $\boldsymbol{\gamma}$.

**Remark.** The forecast function is further specified in section 1.2.3 with assumption 8. Note that forecasters' may consider different predictors than those in the true forecast function when constructing their approximation. The article does not explicitly consider this case to simplify notation, but allowing for it would not change the results.

The predictions generated as in assumption 5 are clearly less accurate than the corresponding true forecasts. However, as shown in the following subsections, empirical error estimators converge in probability to the true error for a wide class of forecast functions. Therefore, the use of these approximations can be justified asymptotically.

### 1.2.2. Error estimators

This subsection expands on these empirical error estimators. It starts by describing the most well-known, broadens the discussion with the block jackknife and introduces the artificial delete-$d$ jackknife error.

Before getting into details, I need to define a loss function for measuring the forecast error at each point in time.

**Definition 2** (Loss). Consistently with definition 1, this paper uses

$$L(\mathbf{Y}_{t+1}, \hat{\mathbf{Y}}_{t+1|s}(\boldsymbol{\gamma})) := \sum_{i \in \mathscr{D}(t+1)} w_i \left[ Y_{i,t+1} - \hat{Y}_{i,t+1|s}(\boldsymbol{\gamma}) \right]^2,$$

where $\mathscr{D}(t+1) := \{i : 1 \leq i \leq n \text{ and } Y_{i,t+1} \neq \text{NA}\}$, NA denotes a generic missing value, for any $p \leq t \leq T-1$ and $p \leq s \leq T$.

The next thing to consider is the conceptual relation between the forecast and its conditioning set. As a general point, the difficulty in obtaining an accurate $\hat{\mathbf{Y}}_{t+1|s}(\boldsymbol{\gamma})$ changes depending on whether $\mathscr{I}(s)$ includes information about the future. This is what leads to the distinction between the two most common categories of forecast error estimators: in-sample and pseudo out-of-sample.

**Definition 3** (In-sample error). The in-sample error

$$\overline{err}(\boldsymbol{\gamma}) := \frac{1}{T-p} \sum_{t=p+1}^{T} L(\mathbf{Y}_t, \hat{\mathbf{Y}}_{t|T}(\boldsymbol{\gamma}))$$

is a measure of the average loss between the data and predictions generated conditioning on the full information set.

Estimating the coefficients once and on the full information set is beneficial for very short time-series problems, as there may not be enough observations to compute more sophisticated estimators. However, this approach tends to overstate the forecast accuracy since the information set is (at least partially) aware of the future. Indeed, in a realistic environment, forecasters would only have information about the past when computing their predictions.

**Definition 4** (Pseudo out-of-sample error). The pseudo out-of-sample error

$$\widehat{err}(\boldsymbol{\gamma}) := \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} L(\mathbf{Y}_{t+1}, \hat{\mathbf{Y}}_{t+1|t}(\boldsymbol{\gamma})),$$

overcomes this limitation by using forecasts generated on the basis of an expanding and backward-looking information set, starting from $p \leq t_0 \leq T-1$.

**Remark.** The pseudo out-of-sample error can be extended to forecast horizons larger than one, but this is not further explored in the manuscript.[5]

Unfortunately, the pseudo out-of-sample error can be either over or under confident depending on the time periods used for estimating and validating the model. This can be overcome using estimators based on the average of pseudo out-of-sample errors computed on a series of data subsamples. The article generates these partitions using time-series generalisations of the jackknife (Quenouille, 1956; Tukey, 1958). The generic jackknife

[5] When long run predictions are calculated iteratively from the one step ahead forecast, it is not necessary to generalise definition 4 to handle longer horizons. The latter would need to be modified only in the case of direct forecast. It is important to stress that when the model is correctly specified, producing iterative forecasts is more efficient than computing horizon-specific ones. However, the latter are more robust to misspecification (Marcellino et al., 2006). For simplicity, this paper focusses only on the one-step ahead forecast, wherein iterative and direct forecasts are identical. Implicitly, this approach is also consistent with iterative forecast methods targetting longer horizons.

error in definition 5 is an estimator that can accommodate for different jackknife partitioning algorithms.

**Definition 5** (Generic jackknife error). Let $\mathscr{J}$ be an indexed family of sets such that each element contains ordered pairs $(i, t)$ with $1 \leq i \leq n$ and $0 \leq t \leq T$. The generic jackknife pseudo out-of-sample error

$$\widetilde{err}(\mathscr{J}, \boldsymbol{\gamma}) := \frac{1}{|\mathscr{J}| \cdot (T - t_0)} \sum_{j=1}^{|\mathscr{J}|} \sum_{t=t_0}^{T-1} L(\mathbf{Y}_{t+1}^{-j}, \hat{\mathbf{Y}}_{t+1|t}^{-j}(\boldsymbol{\gamma})),$$

where $\mathbf{Y}^{-j}$ is the $n \times T$ matrix such that

$$Y_{i,t}^{-j} := \begin{cases} Y_{i,t}, & \text{if } (i,t) \notin \mathscr{J}_j, \\ \text{NA}, & \text{if } (i,t) \in \mathscr{J}_j, \end{cases}$$

$\hat{\mathbf{Y}}_{t+1|t}^{-j}(\boldsymbol{\gamma})$ is analogous to $\hat{\mathbf{Y}}_{t+1|t}(\boldsymbol{\gamma})$, but the autoregressive data component of $\mathbf{X}_t$ is now based on $\hat{\mathbf{Y}}^{-j}$. As for definition 4, $p \leq t_0 \leq T - 1$.

**Remark.** Allowing the ordered pairs $(i, t)$ to have $t = 0$, permits to write the pseudo out-of-sample error as a banal case of jackknife error in which $\mathscr{J}$ contains only one element external to the sample. For instance via $\mathscr{J} = \{(1, 0)\}$. Indeed, the actual data has observations referring to the points in time between 1 and $T$ (included). Therefore, any $t = 0$ is to be considered external.

The most well-known approach to generate jackknife subsamples for dependent data is the block jackknife (Kunsch, 1989). This technique partitions the data into block jackknife samples by removing or down-weighting, in turn, all the unique non-interrupted blocks of $1 \leq c \leq T$ observations.

**Definition 6** (Block jackknife error). This paper denotes the block jackknife error as

$$\widetilde{err}^{BJK}(c, \boldsymbol{\gamma}) \equiv \widetilde{err}(\mathscr{B}(c), \boldsymbol{\gamma}), \tag{1.2}$$

where $\mathscr{B}(c)$ is the family of sets

$$\mathscr{B}(c) := \{\mathscr{B}(1, c), \ldots, \mathscr{B}(T - c + 1, c)\}$$

and

$$\mathscr{B}(j, c) := \{(i, t) : 1 \leq i \leq n \text{ and } j \leq t \leq j + c - 1\}.$$

**Remark.** In other words, this article constructs the individual blocks by replacing, in turn, all the unique non-interrupted blocks of $c$ observations with missing values. This is compatible with Kunsch (1989) since imposing blocks of NAs can be interpreted as fully down-weighting groups of observations. Furthermore, it simplifies the use of the block jackknife to estimate hyperparameters in forecasting settings. In fact, by processing the data via filtering and smoothing techniques compatible with missing observations, it is easier to estimate forecasting models without pre-processing the measurements to remove breaks introduced in the subsampling process.

The main issue with this estimator is that the number of partitions that can be generated from the data is generally small. Thus, the overall improvement over the standard pseudo out-of-sample error is somewhat limited. Also, for those partitions wherein a huge chunk of observations are removed after $t_0$, dividing for a factor of $T - t_0$ may produce inaccurate estimates of the expected error. This is especially true in small-sample problems where $c$ is large relative to $T - t_0$. A simple way for reducing this issue in a finite-sample problem consists in adjusting the $\widetilde{err}^{BJK}(c, \gamma)$ multiplying it by

$$\frac{T - t_0}{|\mathscr{B}(c)|} \sum_{j=1}^{|\mathscr{B}(c)|} \frac{1}{|\{(i, t) \in \mathscr{B}(j, c) : t > t_0\}|}.$$

However, this is difficult to justify asymptotically.

This paper proposes to surpass these problems using an error estimator based on a generalisation of delete-$d$ jackknife (Wu, 1986; Shao and Wu, 1989) compatible with time-series problems: the artificial delete-$d$ jackknife. The classical delete-$d$ jackknife for i.i.d. data (Wu, 1986; Shao and Wu, 1989) generates subsamples by removing, in turn, all the combinations of $d > 0$ observations from the data. This is clearly incompatible with dependent data, since the autocorrelation structure would break during the subsampling process. The artificial jackknife overcomes this complexity by generating the partitions replacing, in turn, all the combinations of $d$ observations with (artificial) missing values. This allows to handle dependent data, as the resulting partitions keep the original ordering and the autocorrelation structure is not altered. Moreover, this approach permits to generate a much larger number of subsamples than block jackknife.[6]

**Definition 7** (Artificial delete-$d$ jackknife error). Let

$$\mathscr{P} := \{i \in \mathbb{Z} : 1 \leq i \leq n\} \times \{t \in \mathbb{Z} : 1 \leq t \leq T\}$$

be the set of all data pairs. Hence, define $\mathscr{A}(d)$ as a family of sets with cardinality

$$|\mathscr{A}(d)| = \frac{(nT)!}{d!\,(nT - d)!}$$

[6] It is interesting to notice that the block jackknife in equation 1.2 is a special case of the artificial delete-$d$ jackknife, in which blocks of consecutive datapoints are replaced with missing values.

such that each element is a $d$-dimensional combination of $\mathscr{P}$. Next, let

$$\widetilde{err}^{AJK}(d, \boldsymbol{\gamma}) \equiv \widetilde{err}(\mathscr{A}(d), \boldsymbol{\gamma}). \tag{1.3}$$

This is the artificial delete-$d$ jackknife error.

The higher reliability of this error estimator is given by the large number of partitions that the artificial delete-$d$ jackknife is able to generate and their heterogeneity. This can be formalised in terms of efficiency as follows.

**Assumption 6** (Finite-sample variance). Assume, for simplicity of notation, that the constituent pseudo out-of-sample errors in definition 5 follow a common finite-sample distribution with variance $\sigma^2(T - t_0, \boldsymbol{\gamma})$.

**Proposition 1.** *Under assumption 6, it follows that, in finite-sample problems,*

$$\begin{aligned} var\left[\widetilde{err}^{AJK}(d, \boldsymbol{\gamma})\right] &\leq var(\widehat{err}(\boldsymbol{\gamma})), \\ var\left[\widetilde{err}^{BJK}(c, \boldsymbol{\gamma})\right] &\leq var(\widehat{err}(\boldsymbol{\gamma})). \end{aligned}$$

*Proof.* The proof is reported in section 1.A.1. □

**Remark.** Assumption 6 can be released without impacting the structure of the proof. However, the notation becomes quite convoluted and hard to read.

Section 1.A.2 compares the variance of the block and artificial jackknife errors through a simulation exercise. This exercise shows that the artificial jackknife outperforms the block jackknife especially in small-sample problems.

When $nT$ is large and $\sqrt{nT} < d < nT$, the cardinality $|\mathscr{A}(d)|$ can be large and it might not be computationally feasible to calculate equation 1.3 evaluating all combinations. Following common practice (Efron and Tibshirani, 1994, p. 149), this computational issue is handled with an approximation. Define $\widetilde{\mathscr{A}}(d) \subset \mathscr{A}(d)$ as a family of sets constructed by drawing at random, without replacement, for a sufficiently large number of times from $\mathscr{A}(d)$. Hence, use this newly defined subset to compute $\widetilde{err}(\widetilde{\mathscr{A}}(d), \boldsymbol{\gamma})$, an approximation of the artificial delete-$d$ jackknife error. Clearly, the accuracy of the approximation

$$\widetilde{err}^{AJK}(d, \boldsymbol{\gamma}) \approx \widetilde{err}(\widetilde{\mathscr{A}}(d), \boldsymbol{\gamma}) \tag{1.4}$$

depends on how close $|\widetilde{\mathscr{A}}(d)|$ is to $|\mathscr{A}(d)|$.

In most empirical problems, the artificial jackknife will likely be truncated. Thus, this article proposes a simple heuristics for selecting its number of artificial missing observations. As detailed in section 1.A and with the simplified notation in assumption 6, the artificial jackknife error variance depends on two factors: $\sigma^2(T - t_0, \boldsymbol{\gamma})$ and the heterogeneity across jackknife subsamples. The latter is controlled by $d$ and, ceteribus paribus,

$\mathrm{var}(\widetilde{err}^{AJK}(d, \boldsymbol{\gamma}))$ is at its minimum when the subsamples are the most diverse. The exact functional form of this variance is unknown and, in the case of the truncated artificial jackknife, one would need to choose a value for $d$ that guarantees a large pool of combinations. Besides, it would be ideal to exclude from $\widetilde{\mathscr{A}}(d)$ the combinations that are the most similar to the block jackknife. In other words, those where all series are missing for one or more periods.

**Conjecture** (Rule of thumb for selecting d). As a result, this paper proposes selecting $d$ for the truncated artificial jackknife error to be

$$\hat{d} = \arg\max_{\underline{d}} \binom{nT}{\underline{d}} - \mathbb{I}_{\underline{d} \geq n} \binom{nT-n}{\underline{d}-n} T - \sum_{i=2}^{\lfloor \underline{d}/n \rfloor} (-1)^{i-1} \binom{T}{i} \binom{nT-in}{\underline{d}-in}, \tag{1.5}$$

where

$$\mathbb{I}_{d \geq n} \binom{nT-n}{d-n} T + \sum_{i=2}^{\lfloor d/n \rfloor} (-1)^{i-1} \binom{T}{i} \binom{nT-in}{d-in},$$

is the amount of subsamples with points in time where all series are artificially missing.[7]

**Remark.** The maximisation is trivial since the objective function is particularly fast to compute for each admissible $\underline{d}$, that is every integer $\underline{d} \in [1, nT]$.

### 1.2.3. Asymptotic properties for estimators based on pseudo out-of-sample evaluations

This subsection provides the asymptotic justification needed for using the approximation in assumption 5 to forecast the target data. It starts by describing the underlying assumptions and continues by proving that pseudo out-of-sample evaluations are consistent, even in the presence of missing observations. The proofs are reported in section 1.B.

**Assumption 7** (Absolute summability). For any finite $n > 0$,

$$\begin{aligned}
&\sum_{i=1}^{n} w_i \leq M_1, \\
&\sum_{i=1}^{n} \sum_{j=1}^{n} |\operatorname{cov}(V_{i,t}, V_{j,t})| \leq M_2, \\
&\sum_{i=1}^{n} \sum_{j=1}^{n} |\operatorname{cov}(V_{i,t}^2, V_{j,t}^2)| \leq M_3,
\end{aligned}$$

where $M_1, M_2, M_3 \in (0, \infty)$ are non-negative finite constants.

[7] To further reduce the effect of those combinations where all series are missing in one or more points in time, the simulation algorithm employed in this article is structured to exclude them from $\widetilde{\mathscr{A}}(d)$.

**Remark.** Recall that the elements of $\mathbf{w}$ are non-negative. Thus, for any finite $n > 0$,

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} |w_i|$$

by definition.

**Assumption 8** (Mean squared error of the forecast). For any $t > 0$,

$$\mathbb{E}\left(\|\mathbf{f}(\mathbf{X}_t, \boldsymbol{\Psi}) - \mathbf{g}(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))\|_2^2\right) \leq M_4/t,$$

where $M_4 \in (0, \infty)$ is a positive finite constant.

**Remark.** Note that

$$\sup_t \|\mathbf{f}(\mathbf{X}_t, \boldsymbol{\Psi}) - \mathbf{g}(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))\|_2^2 \leq \sup_t \||\mathbf{f}(\mathbf{X}_t, \boldsymbol{\Psi})| + |\mathbf{g}(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))|\|_2^2.$$

Since under assumptions 4–5 both the true forecast and its approximation are always finite, the assumption holds within the context of this paper. However, this bound can be loose as it is a function on the problem at hand and it depends on the true forecast and all modelling choices.

**Assumption 9** (Limiting size of the presample). Assume that

$$\lim_{T \to \infty} t_0/T = 0.$$

**Assumption 10** (Limiting number of missing observations). Denote with $0 \leq t_{NA} < T - t_0$ the number of periods between $t_0 + 1$ and $T$ (included) where the data contains missing observations, and assume that

$$\lim_{T \to \infty} t_{NA}/T = 0.$$

**Remark.** Note that assumption 9 serves a crucial purpose: making sure that as $T$ approaches infinity, the pseudo out-of-sample period increases. Similarly, assumption 10 limits the number of periods with missing observations, as $T$ approaches infinity. This implies that as $T$ increases the information set expands, because the number of observed datapoints increases. Without assumption 10, the total number of missing values could become predominant, relative to the amount of observed datapoints.

**Proposition 2.** *Denote with $\widehat{err}_T(\boldsymbol{\gamma})$ the pseudo out-of-sample error for a dataset with $T$ periods. Under assumptions 1–5 and assumptions 7–9, and with complete data it holds* *that*

$$\lim_{T\to\infty} \frac{T}{\ln T}\, \mathbb{E}\left[\left|\widehat{err}_T(\boldsymbol{\gamma}) - err\right|\right] \leq M_1\, M_4.$$

This proposition shows that with complete data, pseudo out-of-sample errors are consistent estimators of the true error. This is a first stepping stone to prove convergence in probability for the generic jackknife errors. Proposition 3 bridges further the gap by extending these results to estimators based on potentially incomplete data.

**Proposition 3.** *Under assumptions 1–5 and assumptions 7–10, and with potentially incomplete data it holds that*

$$\lim_{T\to\infty} \frac{T}{\ln T}\, \mathbb{E}\left[\left|\widehat{err}_T(\boldsymbol{\gamma}) - err\right|\right] \leq M_1\, M_4.$$

**Remark.** Under assumption 10 the rate of convergence in proposition 2 is preserved with potentially incomplete data.

The following corollary of proposition 3 extends its conclusions to the generic jackknife pseudo-out-of-sample error estimators described in section 1.2.2. Clearly, this includes the artificial delete-$d$ jackknife error.

**Corollary 3.1.** *Let $\widetilde{err}_T(\mathscr{J}, \boldsymbol{\gamma})$ be a generic jackknife pseudo out-of-sample error based on a dataset with $T$ time periods. Under the assumptions of proposition 3, it holds that*

$$\lim_{T\to\infty} \frac{T}{\ln T}\, \mathbb{E}\left[\left|\widetilde{err}_T(\mathscr{J}, \boldsymbol{\gamma}) - err\right|\right] \leq M_1\, M_4.$$

### 1.2.4. Hyperparameter selection

Having justified asymptotically the use of the forecasters' approximation in assumption 5, this subsection shows how to optimise its accuracy through hyperparameter selection. It does so by exploring a grid of candidate hyperparameters to find the minimiser for a pseudo out-of-sample error estimator of choice between those reported in section 1.2.2.[8]

Prior to entering into details, let me formalise the hyperparameter selection problem in general terms.

**Definition 8** (Search region and optimal hyperparameters). Let $\mathscr{H}$ be a compact set of ordered tuples that defines the region of existence of the vector of hyperparameters of

[8] This is in line with classical empirical risk minimisation (see, for instance, Elliott and Timmermann, 2016, ch. 3 for a complete survey).

interest. Hence, the optimal hyperparameters are

$$\begin{cases} \hat{\underline{\gamma}}(\mathscr{H}) := \arg\min_{\underline{\gamma}\in\mathscr{H}} \widehat{err}(\underline{\gamma}), & \text{when using the estimator in definition 4,} \\ \tilde{\underline{\gamma}}^{BJK}(c, \mathscr{H}) := \arg\min_{\underline{\gamma}\in\mathscr{H}} \widetilde{err}^{BJK}(c, \underline{\gamma}), & \text{when using the estimator in definition 6,} \\ \tilde{\underline{\gamma}}^{AJK}(d, \mathscr{H}) := \arg\min_{\underline{\gamma}\in\mathscr{H}} \widetilde{err}^{AJK}(d, \underline{\gamma}), & \text{when using the estimator in definition 7.} \end{cases}$$

The simplest way to explore a region of interest is via a grid search.

**Definition 9** (Grid search). Let $\mathscr{H}^{GS} \subseteq \mathscr{H}$ be a finite set of candidate vectors of hyperparameters. Grid search considers every candidate in $\mathscr{H}^{GS}$ and computes

$$\begin{cases} \hat{\underline{\gamma}}(\mathscr{H}^{GS}), & \text{when using the estimator in definition 4,} \\ \tilde{\underline{\gamma}}^{BJK}(c, \mathscr{H}^{GS}), & \text{when using the estimator in definition 6,} \\ \tilde{\underline{\gamma}}^{AJK}(d, \mathscr{H}^{GS}), & \text{when using the estimator in definition 7,} \end{cases}$$

via a naive brute-force optimisation.

This approach explores a small to medium finite grid of candidates and evaluates the relevant error estimator for each one of them. It then returns the candidate vector of hyperparameters associated to the smaller error. The set $\mathscr{H}^{GS}$ is generally constructed to include combinations of hyperparameters within some predetermined ranges. This can be done agnostically (e.g., specifying a rule to take candidates lying in some broad range) or via user expertise (e.g., selecting a few candidates of interest according to a judgmental component).[9] In both cases, there is a strong risk of excluding valid candidates, since it is unfeasible to explore large search regions by using a brute force approach. A simple solution for this problem is given by a random search.

**Definition 10** (Random search). Define $\mathscr{H}^{RS} \subseteq \mathscr{H}$ as a set of candidate vectors of hyperparameters constructed via means of independent and uniform draws without replacement from $\mathscr{H}$. A random search considers every candidate in $\mathscr{H}^{RS}$ and computes

$$\begin{cases} \hat{\underline{\gamma}}(\mathscr{H}^{RS}), & \text{when using the estimator in definition 4,} \\ \tilde{\underline{\gamma}}^{BJK}(c, \mathscr{H}^{RS}), & \text{when using the estimator in definition 6,} \\ \tilde{\underline{\gamma}}^{AJK}(d, \mathscr{H}^{RS}), & \text{when using the estimator in definition 7,} \end{cases}$$

with the same approach employed for grid search.

In its most naive implementation, it is a grid search based on a region of interest constructed by taking random candidates from $\mathscr{H}$. This operation allows to keep the computational advantages of grid search, while exploring a more heterogeneous section of

[9] The latter is also called manual search (Bergstra and Bengio, 2012).

$\mathscr{H}$. This is especially relevant if $\boldsymbol{\gamma}$ is high-dimensional, since it is difficult to generate a proper set of candidates on the basis of some deterministic or subjective rule.

This formulation for the random search is rather naive. Nonetheless, Bergstra and Bengio (2012) showed that it is (at least) as good as more advanced versions of random search. Further details on these algorithms can be found in Solis and Wets (1981) and Andradóttir (2015). Random search tends to be less effective for cases where the number of hyperparameters to tune is very large. For these cases, alternative and more powerful techniques (e.g., simulated annealing, particle swarm optimization) surveyed in Weise (2009) could help. However, since they would inevitably increase the computational burden and the complexity of the hyperparameter optimisation, they are left for future research.

A final point that should be taken into account is that while definition 8 is intuitive, it is also prone to errors in some circumstances. Indeed, when the expected error surface is flat, it is hard to pick one candidate in particular. In these cases, it is often more sensible to evaluate the whole grid of interest and use the threshold where the surface starts flattening as optimal hyperparameters.

## 1.3. Empirical application

This section illustrates the functionality of the artificial delete-$d$ jackknife by tuning the hyperparameters of penalised VARMAs on weekly exchange rate returns.

The exchange rates complexities serve as a good empirical example to benchmark different techniques for selecting hyperparameters.[10] Starting with the contribution of Meese and Rogoff (1983), a large body of empirical economic research has found that forecasting models for exchange rates based on macroeconomic data or informed by economic theory are often outperformed by simple univariate techniques and parsimonious multivariate methods usually difficult to tune.

### 1.3.1. Penalised VARMA

This subsection describes the case in which forecasters form their predictions using high-dimensional elastic-net VARMA($q$, $r$) models.[11] Clearly, the following assumptions and definitions affect only the empirical example in section 1.3.

[10] It is important to remark that this manuscript does not intend to find the best model (among a class of techniques) for predicting exchange rates, but rather it aims to show that the artificial jackknife is a valid approach for tuning the models in this example.

[11] It is important to stress that the VARMA model encompasses common univariate (i.e., AR, MA, ARIMA) and multivariate (i.e., VAR, VMA, VARIMA) forecasting methods.

**Assumption 11** (VARMA model). Within section 1.3, forecasters form their expectations assuming that

$$\mathbf{Y}_{t+1} = \mathbf{\Pi}_1 \mathbf{Y}_t + \ldots + \mathbf{\Pi}_q \mathbf{Y}_{t-q+1} + \mathbf{\Xi}_1 \mathbf{V}_t + \ldots + \mathbf{\Xi}_r \mathbf{V}_{t-r+1} + \mathbf{V}_{t+1}, \tag{1.6}$$

where $\mathbf{V}_{t+1} \overset{w.n.}{\sim} N\left(\mathbf{0}_{n\times 1}, \mathbf{\Sigma}\right)$ with $\mathbf{\Sigma}$ being positive definite, $t \in \mathbb{Z}$.[12] The autoregressive and moving average coefficients are $n \times n$ matrices for which the VARMA is causal and invertible (Brockwell et al., 1991, pp. 418-420).

For simplicity of notation, let

$$\begin{aligned} \mathbf{\Pi} &:= \begin{pmatrix} \mathbf{\Pi}_1 & \ldots & \mathbf{\Pi}_q \end{pmatrix}, \\ \mathbf{\Xi} &:= \begin{pmatrix} \mathbf{\Xi}_1 & \ldots & \mathbf{\Xi}_r \end{pmatrix}. \end{aligned}$$

Moreover, consider only parametrisations where $\min(q, r) = 0$.[13]

**Definition 11** (Penalised maximum likelihood estimation). Forecasters use penalised maximum likelihood estimation to estimate the estimated VARMA coefficients. With complete data, this implies

$$\hat{\boldsymbol{\theta}}_s(\boldsymbol{\gamma}) := \underset{\underline{\boldsymbol{\theta}} \in \mathscr{R}}{\arg\max}\ \mathcal{L}(\underline{\boldsymbol{\theta}} \,|\, \mathbf{Y}_{1:s}) - \mathcal{P}(\underline{\boldsymbol{\theta}}, \boldsymbol{\gamma}),$$

where $\mathscr{R}$ is the region of interest for the parameters implicitly defined in assumption 11,

$$\mathcal{L}(\underline{\boldsymbol{\theta}} \,|\, \mathbf{Y}_{1:s}) \simeq -\frac{s}{2} \ln |\underline{\mathbf{\Sigma}}| - \frac{1}{2} \operatorname{Tr}\left[\sum_{t=1}^{s} \underline{\mathbf{\Sigma}}^{-1} \mathbf{V}_t(\underline{\boldsymbol{\theta}}) \mathbf{V}_t(\underline{\boldsymbol{\theta}})'\right]$$

denotes the log-likelihood of the VARMA model (Lütkepohl, 2005, ch. 11) and $\mathcal{P}(\underline{\boldsymbol{\theta}}, \boldsymbol{\gamma})$ is a penalty function, for $\max(q, r) \leq s \leq T$.[14] By extension, $\underline{\mathbf{\Pi}}$, $\underline{\mathbf{\Xi}}$ and $\underline{\mathbf{\Sigma}}$ are the VARMA coefficients built from $\underline{\boldsymbol{\theta}}$.

Performing penalised maximum likelihood with incomplete data is non-trivial. In order to overcome the related complexities, this article uses an Expectation-Conditional Maximisation (ECM) algorithm (Meng and Rubin, 1993). The details of this iterative estimation procedure are described in section 1.C.

The penalty function of interest for this empirical application builds on the elastic-net literature (Zou and Hastie, 2005; Zou and Zhang, 2009).

[12] The data is assumed to have zero mean and unit standard deviation for simplicity of notation.

[13] Under assumption 2, $\max(q, r) > 0$. Therefore, letting $\min(q, r) = 0$ does not exclude the white noise case, since the model could be parametrised to have autoregressive and moving average coefficients equal to zero. This point is purely to simplify the notation in section 1.C.

[14] The innovations $\mathbf{V}_t(\underline{\boldsymbol{\theta}}) \equiv \mathbf{V}_t$ in equation 1.6. This notation is used for stressing its dependence from the coefficients in $\underline{\boldsymbol{\theta}}$ and obtain a compact formula.

**Definition 12** (Generalised elastic-net penalty). For any $p \in \mathbb{N}$, let

$$\mathbf{\Gamma}(\boldsymbol{\gamma}, p) := \lambda \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times n} & \dots & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times n} & \beta \cdot \mathbf{I}_n & \dots & \mathbf{0}_{n\times n} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0}_{n\times n} & \dots & \dots & \beta^{p-1} \cdot \mathbf{I}_n \end{pmatrix}$$

where $\boldsymbol{\gamma} := (q \;\; r \;\; \lambda \;\; \alpha \;\; \beta)'$ is a given vector of hyperparameters with $\lambda \geq 0$, $0 \leq \alpha \leq 1$ and $\beta \geq 1$. Building on that, this manuscript uses the penalty

$$\mathcal{P}(\underline{\boldsymbol{\theta}}, \boldsymbol{\gamma}) := \begin{cases} \frac{1-\alpha}{2} \left\| \underline{\mathbf{\Pi}}\, \mathbf{\Gamma}(\boldsymbol{\gamma}, q)^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 + \frac{\alpha}{2} \left\| \underline{\mathbf{\Pi}}\, \mathbf{\Gamma}(\boldsymbol{\gamma}, q) \right\|_{1,1} & \text{if } q > 0 \text{ and } r = 0, \\ \frac{1-\alpha}{2} \left\| \underline{\mathbf{\Xi}}\, \mathbf{\Gamma}(\boldsymbol{\gamma}, r)^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 + \frac{\alpha}{2} \left\| \underline{\mathbf{\Xi}}\, \mathbf{\Gamma}(\boldsymbol{\gamma}, r) \right\|_{1,1} & \text{if } q = 0 \text{ and } r > 0. \end{cases} \tag{1.7}$$

**Remark.** Note that when the penalty is active (i.e., $\lambda > 0$), $\mathbf{\Gamma}(\boldsymbol{\gamma}, q)$ and $\mathbf{\Gamma}(\boldsymbol{\gamma}, r)$ are diagonal and positive definite matrices, and thus

$$\mathcal{P}(\underline{\boldsymbol{\theta}}, \boldsymbol{\gamma}) = \begin{cases} \sum_{i=1}^{n} \sum_{j=1}^{nq} \frac{1-\alpha}{2}\, \underline{\Pi}_{i,j}^2 \; [\mathbf{\Gamma}(\boldsymbol{\gamma}, q)]_{j,j} + \frac{\alpha}{2} \, |\underline{\Pi}_{i,j}| \; [\mathbf{\Gamma}(\boldsymbol{\gamma}, q)]_{j,j} & \text{if } q > 0 \text{ and } r = 0, \\ \sum_{i=1}^{n} \sum_{j=1}^{nr} \frac{1-\alpha}{2}\, \underline{\Xi}_{i,j}^2 \; [\mathbf{\Gamma}(\boldsymbol{\gamma}, r)]_{j,j} + \frac{\alpha}{2} \, |\underline{\Xi}_{i,j}| \; [\mathbf{\Gamma}(\boldsymbol{\gamma}, r)]_{j,j} & \text{if } q = 0 \text{ and } r > 0. \end{cases}$$

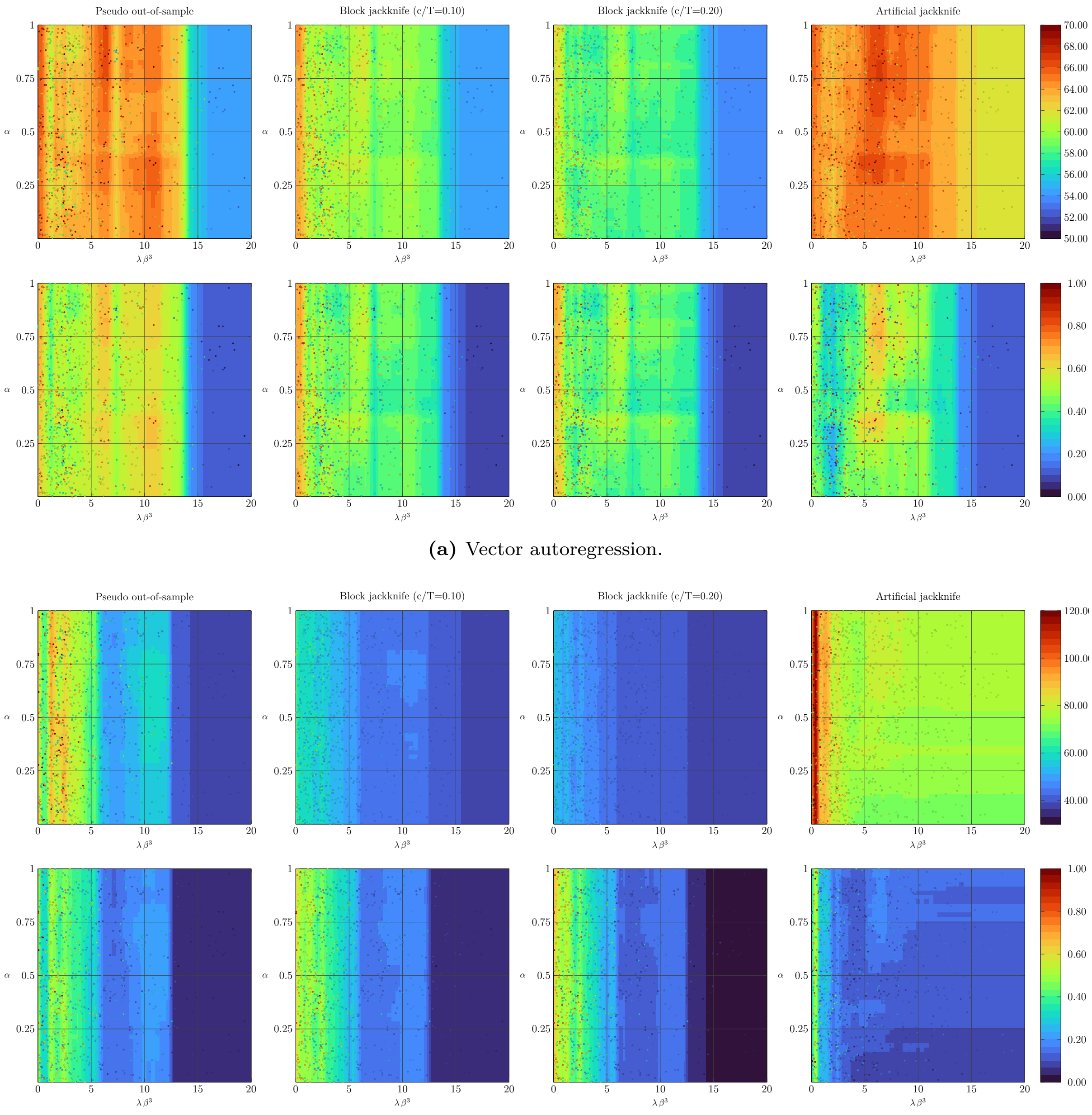
The penalty $\mathcal{P}(\underline{\boldsymbol{\theta}}, \boldsymbol{\gamma})$ is a generalisation of the elastic-net that allows to penalise more autoregressive and moving average coefficients referring to distant points in time. As for its standard implementation, when $\alpha = 1$ and $\alpha = 0$ the function is equivalent to the LASSO (Tibshirani, 1996) and ridge (Hoerl and Kennard, 1970) penalties. These penalties perform differently depending on the empirical setting in which they are employed, as extensively described in Zou and Hastie (2005).[15] For $0 < \alpha < 1$ the model allows for a sparse model and benefits from the co-movement of correlated predictors. With respect to the standard elastic-net, the penalty function in equation 1.7 includes $\beta$, an additional hyperparameter. If $\beta > 1$, then $\mathcal{P}(\underline{\boldsymbol{\theta}}, \boldsymbol{\gamma})$ penalises more coefficients referring to distant points in time.[16]

### 1.3.2. Results

The time series for the exchange rates are collected from the Federal Reserve Board H.10 and include regular weekly (Friday, EOP) observations from January 1999 to the end

[15] LASSO gives a sparse representation of the model and thus a simple regression in few predictors. Ridge does not select subsets of regressors, but it shrinks all of them jointly. The ability of LASSO in selecting the same covariates over time is rather poor when some of them are highly correlated.

[16] This idea is commonly used in time series and a simple parallel can be made by looking at Bayesian VARs with Minnesota priors (Doan et al., 1984; Litterman, 1986). Indeed, in stationary settings, this set of priors shrinks the vector autoregression toward a white noise (i.e., it shrinks the coefficients to zero) and penalises more distant lags. The penalty in equation 1.7 is similar in spirit, but it allows for a sparse representation of the model and for the use of moving average coefficients.



**(a)** Vector autoregression.

**(b)** Vector moving average.

**Figure 1.2:** Expected error for the candidate hyperparameters in $\mathscr{H}$.
**Notes**: For each model, the first row describes the expected error in absolute terms, while the second one shows it in relative terms (per subsampling method). The scalar $\lambda\,\beta^3$ denotes the shrinkage associated to the farthest lag. The block jackknife output is adjusted to reduce the finite-sample methodological defects as described in section 1.2.2.

of December 2020, for a set of major economies reported in table 1.D.1. This dataset contains a total of 1,148 weeks and 21,812 observations. These are all the exchange rates in the Federal Reserve Board H.10 that did not have a fixed or pegged rate with the dollar in the sample. Moreover, these exchange rates are not taken in levels, but they are transformed in weekly log-returns instead.

The sample is divided into three blocks: a presample (January 1999 to December 1999), a selection sample (January 2000 to December 2001) and a test sample (January

2002 to December 2020). The presample is used for computing $\mathbf{w}$ and then discarded. Each entry in this vector is equal to 1 over the variance of the corresponding series in the presample. This is done to equally weight each series, regardless of its volatility. Next, the grid of candidate hyperparameters $\mathscr{H} = \mathscr{H}_p \times \mathscr{H}_\lambda \times \mathscr{H}_\alpha \times \mathscr{H}_\beta$ is explored on the selection sample following section 1.2.4 to compute the associated expected errors. This is done settings $\mathscr{H}_p := \{4\}$, $\mathscr{H}_\lambda := [10^{-2}, 2.5]$, $\mathscr{H}_\alpha := [0, 1]$ and $\mathscr{H}_\beta := [1, 2]$ and letting the methods based on pseudo out-of-sample criteria defining $t_0$ to be such that the part of the selection sample used for estimation purposes ends in December 2000. The set $\mathscr{H}_p$ fixes the number of lags to 4: a value considered large enough to forecast the weekly financial returns.[17] The sets referring to the remaining hyperparameters allow to control the overall shrinkage level and kill superfluous lags, if needed. Finally, the expected error associated to each hyperparameter is assessed by computing the realised pseudo out-of-sample error over the test sample, having estimated the relevant model (i.e., vector autoregression or vector moving average) once on the full selection sample.

Figure 1.2 describes the random search output obtained with the error estimators described in section 1.2.2.[18] The vector autoregression results show a series of important features. First, it is evident that the expected error decreases when the shrinkage level increases, no matter the error estimator. Second, the area with the lowest expected error is where $\lambda\beta^3 \geq 15$. This location seems independent from $\alpha$ and it is again found regardless of the error estimator of choice. Third, there are strong differences in the scale of the expected error obtained via different estimators. Indeed, the artificial jackknife estimates are the most conservative, since the expected error is higher in scale across all candidate hyperparameters. The pseudo out-of-sample gives similar expected errors for any configuration with $\lambda\beta^3 < 15$, but a more pronounced fall before the common flattening. The block jackknife measures are the least conservative. The picture observed through the lenses of the vector moving average results is quite different. Indeed, while the expected error still decreases when the shrinkage level increases, it does so at a different rate for the case of the artificial jackknife. Indeed, the artificial jackknife expected error decreases sharply already at $\lambda\beta^3 \approx 1$, whereas it takes about a fivefold figure to start decreasing when estimated with the pseudo out-of-sample and block jackknife. Furthermore, the artificial jackknife expected error starts flattening at a much smaller shrinkage level compared to the benchmarks.

These expected errors are then compared with the pseudo out-of-sample error realised in test sample. Table 1.1 summarises the quality of each selection method for both vector autoregressions and moving averages. These results show that the artificial jackknife

[17] Since the vector moving averages in this manuscript are constrained to be invertible, they account for a higher persistence than the vector autoregressions of the same order. Indeed, any invertible VMA can be equivalently thought as a VAR($\infty$).

[18] Figure 1.D.1 shows the same output using an alternative graphical representation.

| Error estimator | Selection RMSE | |
|---|---|---|
| | Vector autoregression | Vector moving average |
| In-sample error | 7.28 | 1.50 |
| Pseudo out-of-sample error | 1.00 | 1.00 |
| Block jackknife, $c/T = 0.1$ | 1.68 | 1.20 |
| Block jackknife, $c/T = 0.2$ | 2.47 | 1.30 |
| Block jackknife, $c/T = 0.1$ (adjusted) | 1.17 | 1.14 |
| Block jackknife, $c/T = 0.2$ (adjusted) | 1.28 | 1.20 |
| Artificial delete-$\hat{d}$ jackknife | 0.93 | 0.89 |

**Table 1.1:** Selection relative mean squared error.
**Notes**: The selection MSE is computed by averaging the squared error between the realised and expected error associated to each candidate hyperparameter. The realised error is the pseudo out-of-sample error computed in the test sample. The selection RMSE is rescaled so that values lower than one indicate a better performance compared to the pseudo-out-of-sample selection.

gives the best estimate of the expected error across all candidate hyperparameters. This is evident both for vector autoregressions and vector moving averages. A further interesting result is that the pseudo out-of-sample is better than the block jackknife (both raw and adjusted). This is likely due to the small number of partitions that the block jackknife is able to generate in this empirical application. Finally, it follows from table 1.1 and figure 1.2 that the best configuration for vector moving averages requires a much smaller shrinkage level compared to vector autoregressions.

## 1.4. Concluding comments

This article proposes a new approach for selecting hyperparameters in time series denoted as artificial delete-$d$ jackknife: a generalisation of the delete-$d$ jackknife.

By contrast with existing approaches, the artificial delete-$d$ jackknife can partition dependent data into a large set of unique partitions, even when $T$ is relatively small. These partitions are used for constructing a robust forecast error estimator, based on pseudo out-of-sample evaluations. The artificial delete-$d$ jackknife has strong finite-sample advantages and converges in probability to the true error. Empirical results on weekly exchange rate returns are also promising.

While the theory developed in this paper is based on a weighted mean square loss, the artificial jackknife error could be extended to other loss functions for prediction and classification problems. Also, it could be expanded to compute the uncertainty around sample statistics in time series. These and a few other points are not fully developed in this article and they are left for future research.

# Appendix

## 1.A. Finite sample results

### 1.A.1. Proposition 1

PROOF OF PROPOSITION 1. Let

$$\widehat{err}^{-j}(\boldsymbol{\gamma}) := \frac{1}{T-t_0}\sum_{t=t_0}^{T-1} L(\mathbf{Y}_{t+1}^{-j}, \hat{\mathbf{Y}}_{t+1|t}^{-j}(\boldsymbol{\gamma})),$$

for any $j = 1, \ldots, |\mathscr{J}|$. Hence,

$$\widetilde{err}(\mathscr{J}, \boldsymbol{\gamma}) = \frac{1}{|\mathscr{J}| \cdot (T-t_0)} \sum_{j=1}^{|\mathscr{J}|}\sum_{t=t_0}^{T-1} L(\mathbf{Y}_{t+1}^{-j}, \hat{\mathbf{Y}}_{t+1|t}^{-j}(\boldsymbol{\gamma})) = \frac{1}{|\mathscr{J}|}\sum_{j=1}^{|\mathscr{J}|} \widehat{err}^{-j}(\boldsymbol{\gamma}).$$

It follows that

$$\begin{aligned}
&\operatorname{var}\big(\widetilde{err}(\mathscr{J}, \boldsymbol{\gamma})\big) \\
&\quad = \frac{1}{|\mathscr{J}|^2} \operatorname{var}\left[\sum_{j=1}^{|\mathscr{J}|} \widehat{err}^{-j}(\boldsymbol{\gamma})\right] \\
&\quad = \frac{1}{|\mathscr{J}|^2} \sum_{i=1}^{|\mathscr{J}|}\sum_{j=1}^{|\mathscr{J}|} \operatorname{cov}\left[\widehat{err}^{-i}(\boldsymbol{\gamma}),\, \widehat{err}^{-j}(\boldsymbol{\gamma})\right] \\
&\quad = \frac{1}{|\mathscr{J}|^2} \left\{ \sum_{i=1}^{|\mathscr{J}|} \operatorname{var}\left[\widehat{err}^{-i}(\boldsymbol{\gamma})\right] + \sum_{i=1}^{|\mathscr{J}|}\sum_{\substack{j=1\\ j\neq i}}^{|\mathscr{J}|} \operatorname{cov}\left[\widehat{err}^{-i}(\boldsymbol{\gamma}),\, \widehat{err}^{-j}(\boldsymbol{\gamma})\right] \right\}.
\end{aligned}$$

Under assumption 6,

$$\operatorname{var}\big(\widetilde{err}(\mathscr{J}, \boldsymbol{\gamma})\big) = \frac{\sigma^2(T-t_0, \boldsymbol{\gamma})}{|\mathscr{J}|} + \frac{1}{|\mathscr{J}|^2} \sum_{i=1}^{|\mathscr{J}|}\sum_{\substack{j=1\\ j\neq i}}^{|\mathscr{J}|} \operatorname{cov}\left[\widehat{err}^{-i}(\boldsymbol{\gamma}),\, \widehat{err}^{-j}(\boldsymbol{\gamma})\right].$$

For any $\mathscr{J}$ with cardinality equal to one, including the pseudo out-of-sample error (cf.

definition 5 remark),

$$\operatorname{var}\big(\widetilde{err}(\mathscr{J},\boldsymbol{\gamma})\big) = \sigma^2(T - t_0, \boldsymbol{\gamma}).$$

Instead, for $\mathscr{J}$ with a larger cardinality and heterogeneous partitions, the $\operatorname{var}\big(\widetilde{err}(\mathscr{J},\boldsymbol{\gamma})\big)$ is lower or equal to $\sigma^2(T-t_0,\boldsymbol{\gamma})$. Among others, this is the case of the block and artificial delete-$d$ jackknife errors. Hence,

$$\begin{aligned}\operatorname{var}\left[\widetilde{err}^{AJK}(d,\boldsymbol{\gamma})\right] &\leq \operatorname{var}\big(\widehat{err}(\boldsymbol{\gamma})\big),\\ \operatorname{var}\left[\widetilde{err}^{BJK}(c,\boldsymbol{\gamma})\right] &\leq \operatorname{var}\big(\widehat{err}(\boldsymbol{\gamma})\big).\end{aligned}$$

□

### 1.A.2. Simulation results

Determining which one between the block and artificial jackknife error estimators has the smaller variance is a little more complicated, since it requires to study the covariances.

Note that

$$\frac{1}{|\mathscr{J}|^2}\sum_{i=1}^{|\mathscr{J}|}\sum_{\substack{j=1\\ j\neq i}}^{|\mathscr{J}|}\operatorname{cov}\left[\widehat{err}^{-i}(\boldsymbol{\gamma}),\, \widehat{err}^{-j}(\boldsymbol{\gamma})\right]$$

is inversely proportional to the heterogeneity across subsamples. One way to measure it is through the expected value of the Jaccard similarity coefficient (denoted as sim) of a pair of subsamples $\mathscr{S}_1$ and $\mathscr{S}_2$ selected at random among those in the family of sets $\mathscr{J}$. The lower it is, the more diverse the partitions. Formally, the expected value of this Jaccard index is denoted as $\mathbb{E}[\operatorname{sim}(\mathscr{S}_1,\mathscr{S}_2)]$. Besides, for all partitioning methods in which the elements of $\mathscr{J}$ have the same cardinality,

$$\mathbb{E}\Big[\operatorname{sim}(\mathscr{S}_1,\mathscr{S}_2)\Big] = \mathbb{E}\left[\frac{|\mathscr{S}_1\cap\mathscr{S}_2|}{|\mathscr{S}_1\cup\mathscr{S}_2|}\right] = \sum_{k=0}^{|\mathscr{S}_1|}\frac{k}{2|\mathscr{S}_1|-k}\operatorname{Pr}(|\mathscr{S}_1\cap\mathscr{S}_2|=k).$$

In the case of the artificial delete-$d$ jackknife, the random subsamples are selected from those in $\mathscr{A}(d)$, they are dependent on $d$ and such that $|\mathscr{S}_1(d)| = |\mathscr{S}_2(d)| = d$. Therefore,

$$\mathbb{E}\Big[\operatorname{sim}\big(\mathscr{S}_1(d),\mathscr{S}_2(d)\big)\Big] = \sum_{k=1}^{d-1}\frac{k}{2d-k}\,\frac{\binom{d}{k}\binom{nT-d}{d-k}}{\binom{nT}{d}}.$$

With the block jackknife, the random subsamples are selected from those in $\mathscr{B}(c)$, they are dependent on $c$ and such that $|\mathscr{S}_1(c)| = |\mathscr{S}_2(c)| = c$. In this case, the partitions are non-interrupted blocks of consecutive observations. Considering each $i$-th subsample as a different case and employing the law of total probability allows to determine that

$$\Pr(|\mathcal{S}_1(c) \cap \mathcal{S}_2(c)| = k) = \frac{1}{(T-c+1)^2} \sum_{i=1}^{T-c+1} \left( \mathbb{I}_{T-c+1-i \geq c-k} + \mathbb{I}_{i-1 \geq c-k} \right),$$

for $1 \leq k < c$. Hence, the expected Jaccard index for the block jackknife is

$$\mathbb{E}\Big[\text{sim}\big(\mathcal{S}_1(c), \mathcal{S}_2(c)\big)\Big] = \frac{1}{(T-c+1)^2} \sum_{k=1}^{c-1} \sum_{i=1}^{T-c+1} \frac{k}{2c-k} \left( \mathbb{I}_{T-c+1-i \geq c-k} + \mathbb{I}_{i-1 \geq c-k} \right).$$

These expectations are compared within deterministic computer simulations to understand whether the artificial jackknife produces more heterogeneous partitions than the block jackknife. In order to put $d$ and $c$ onto the same scale, the former is set to be equal to $nc$ in this simulation exercise.[19]

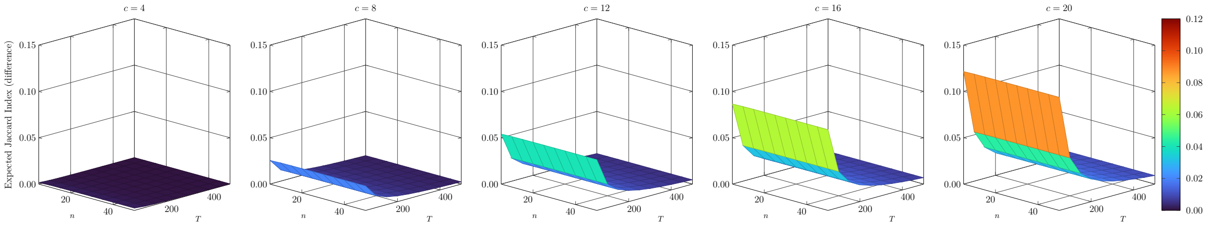


**Figure 1.A.1:** Difference between the expected Jaccard similarity associated to the block and artificial jackknifes for a broad set of configurations.
**Notes**: A positive difference indicates a configuration in which the artificial jackknife outperforms the block version. The $n$-axis ranges from 5 to 50 with a step size of 5. The $T$-axis ranges from 50 to 500 with a step size of 50. Hence, the surface comprises 100 points.

Results in figure 1.A.1 show that the artificial jackknife significantly outperforms the block jackknife especially in small samples. However, its relative advantage decreases as the sample size increases. This is consistent for different $c$ and compatible with the asymptotic results in section 1.B.

[19] Indeed, these versions of the jackknife place the same number of artificial missing observations when $d = nc$, since $c$ refers to a number of time periods.

# 1.B. Asymptotic results

## 1.B.1. Proposition 2

Recall that with complete data

$$\begin{aligned}\widehat{err}_T(\boldsymbol{\gamma}) = &+ \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \sum_{i=1}^{n} w_i \, V_{i,t+1}^2 \\ &+ \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \sum_{i=1}^{n} w_i \Big[ f_i(\mathbf{X}_t, \boldsymbol{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma})) \Big]^2 \\ &+ \frac{2}{T-t_0} \sum_{t=t_0}^{T-1} \sum_{i=1}^{n} w_i \, V_{i,t+1} \Big[ f_i(\mathbf{X}_t, \boldsymbol{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma})) \Big],\end{aligned} \tag{1.8}$$

where $f_i(\mathbf{X}_t, \boldsymbol{\Psi}) \equiv [\mathbf{f}(\mathbf{X}_t, \boldsymbol{\Psi})]_i$, $g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma})) \equiv \Big[\mathbf{g}(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))\Big]_i$ for $i = 1, \ldots, n$ and any integer point in time $t$. The proof of proposition 2 relies on the following three lemmas, and it is reported hereinafter. Each term on the RHS of equation 1.8 is linked to one of these lemmas.

**Lemma 1.** *Under assumptions 1–5, assumption 7, assumption 9 and with complete data, it holds that*

$$\lim_{T\to\infty} T \, \mathbb{E} \left( \left| \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \sum_{i=1}^{n} w_i \, V_{i,t+1}^2 - err \right|^2 \right) \leq M_1^2 \, M_3.$$

PROOF. Recall that the vector of weights $\mathbf{w}$ is made of $n$ given non-negative finite scalars, $\mathbf{V}_{t+1} \overset{i.i.d.}{\sim} (\mathbf{0}_{n\times 1}, \boldsymbol{\Sigma})$, the model parameters are finite, $\boldsymbol{\Sigma}$ is a positive definite matrix and $err = \sum_{i=1}^{n} w_i \, \boldsymbol{\Sigma}_{i,i}$. Using the bias-variance decomposition

$$\mathbb{E} \left( \left| \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \sum_{i=1}^{n} w_i \, V_{i,t+1}^2 - \sum_{i=1}^{n} w_i \, \boldsymbol{\Sigma}_{i,i} \right|^2 \right)$$

can be written in the equivalent form

$$\left[ \text{bias} \left( \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \sum_{i=1}^{n} w_i \, V_{i,t+1}^2 \right) \right]^2 + \text{var} \left( \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \sum_{i=1}^{n} w_i \, V_{i,t+1}^2 \right).$$

The bias is equal to zero since

$$\text{bias} \left( \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \sum_{i=1}^{n} w_i \, V_{i,t+1}^2 \right) = \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \sum_{i=1}^{n} \mathbb{E} \left( w_i \, V_{i,t+1}^2 \right) - \sum_{i=1}^{n} w_i \, \boldsymbol{\Sigma}_{i,i} = 0.$$

The variance is

$$\begin{aligned}
&\operatorname{var}\left(\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\, V_{i,t+1}^2\right)\\
&= \frac{1}{(T-t_0)^2}\sum_{t=t_0}^{T-1}\sum_{s=t_0}^{T-1}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i\, w_j\, \left[\mathbb{E}\left(V_{i,t+1}^2\, V_{j,s+1}^2\right) - \mathbb{E}\left(V_{i,t+1}^2\right)\mathbb{E}\left(V_{j,s+1}^2\right)\right]\\
&= \frac{1}{(T-t_0)^2}\sum_{t=t_0}^{T-1}\sum_{s=t_0}^{T-1}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i\, w_j\, \operatorname{cov}\left(V_{i,t+1}^2, V_{j,s+1}^2\right).
\end{aligned}$$

Given that $\mathbf{V}_{t+1}$ is i.i.d.

$$\operatorname{var}\left(\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\, V_{i,t+1}^2\right) = \frac{1}{(T-t_0)^2}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i\, w_j\, \operatorname{cov}\left(V_{i,t+1}^2, V_{j,t+1}^2\right).$$

Since all weights are non-negative and under assumption 7,

$$\begin{aligned}
&\mathbb{E}\left(\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\, V_{i,t+1}^2 - \sum_{i=1}^{n} w_i\, \mathbf{\Sigma}_{i,i}\right|^2\right)\\
&\leq \frac{M_1^2}{(T-t_0)^2}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}\left(V_{i,t+1}^2, V_{j,t+1}^2\right)\\
&\leq \frac{M_1^2\, M_3}{T(1-t_0/T)}.
\end{aligned}$$

Hence, under assumption 9,

$$\lim_{T\to\infty} T\,\mathbb{E}\left(\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\, V_{i,t+1}^2 - err\right|^2\right) \leq M_1^2\, M_3.$$

□

**Lemma 2.** *Under assumptions 1–5, assumptions 7–9 and with complete data*

$$\lim_{T\to\infty}\frac{T}{\ln T}\,\mathbb{E}\left\{\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\Big[f_i(\mathbf{X}_t, \mathbf{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\gamma))\Big]^2\right|\right\} \leq M_1\, M_4.$$

PROOF. Note that, since all weights are non-negative,

$$\begin{aligned}
&\mathbb{E}\left\{\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\Big[f_i(\mathbf{X}_t, \mathbf{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\gamma))\Big]^2\right|\right\}\\
&= \frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,\mathbb{E}\left[|f_i(\mathbf{X}_t, \mathbf{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\gamma))|^2\right].
\end{aligned}$$

Under assumptions 7–8,

$$\begin{aligned}
&\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,\mathbb{E}\left[|f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))|^2\right]\\
&\quad\le \frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,\mathbb{E}\left[\|\mathbf{f}(\mathbf{X}_t,\boldsymbol{\Psi})-\mathbf{g}(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\|_2^2\right]\\
&\quad\le \frac{M_4}{T-t_0}\sum_{t=t_0}^{T-1}\frac{1}{t}\sum_{i=1}^{n} w_i\\
&\quad\le \frac{M_1\,M_4}{T-t_0}\sum_{t=t_0}^{T-1}\frac{1}{t}.
\end{aligned}$$

Since

$$\sum_{t=t_0}^{T-1}\frac{1}{t}=\sum_{s=1}^{T-t_0}\frac{1}{s+t_0-1}\le\sum_{s=1}^{T}\frac{1}{s+t_0-1}\le\sum_{s=1}^{T}\frac{1}{s}\le\int_1^T\frac{1}{s}\,ds\le\ln T+1,$$

it follows that

$$\mathbb{E}\left\{\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\Big[f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\Big]^2\right|\right\}\le\frac{(\ln T+1)(M_1\,M_4)}{T(1-t_0/T)}.$$

Therefore, under assumption 9, it holds that

$$\lim_{T\to\infty}\frac{T}{\ln T}\,\mathbb{E}\left\{\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\Big[f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\Big]^2\right|\right\}\le M_1\,M_4.$$

□

**Lemma 3.** *Under assumptions 1–5, assumptions 7–9 and with complete data*

$$\lim_{T\to\infty}\sqrt{T}\,\mathbb{E}\left\{\left|\frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,V_{i,t+1}\left[f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\right]\right|\right\}\le 4M_1\sqrt{M_2}\sqrt{M_4}.$$

PROOF. Since all weights are non-negative, it follows from the Cauchy-Schwarz inequality that

$$\begin{aligned}
&\mathbb{E}\left\{\left|\frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,V_{i,t+1}\left[f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\right]\right|\right\}\\
&\quad\le\frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,\mathbb{E}\Big[|\,V_{i,t+1}\,|\,|f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))|\Big]\\
&\quad\le\frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,\sqrt{\mathbb{E}\Big[|\,V_{i,t+1}\,|^2\Big]}\,\sqrt{\mathbb{E}\Big[|f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))|^2\Big]}.
\end{aligned}$$

Under assumptions 7–8,

$$\begin{aligned}
&\frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i \sqrt{\mathbb{E}\Big[|\,V_{i,t+1}\,|^2\Big]}\,\sqrt{\mathbb{E}\Big[|f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))|^2\Big]}\\
&\le \frac{2\sqrt{M_2}}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i \sqrt{\mathbb{E}\Big[|f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))|^2\Big]}\\
&\le \frac{2\sqrt{M_2}}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i \sqrt{\mathbb{E}\Big[\|\mathbf{f}(\mathbf{X}_t,\boldsymbol{\Psi})-\mathbf{g}(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\|_2^2\Big]}\\
&\le \frac{2\sqrt{M_2}\sqrt{M_4}}{T-t_0}\sum_{t=t_0}^{T-1}\frac{1}{\sqrt{t}}\sum_{i=1}^{n} w_i\\
&\le \frac{2M_1\sqrt{M_2}\sqrt{M_4}}{T-t_0}\sum_{t=t_0}^{T-1}\frac{1}{\sqrt{t}}.
\end{aligned}$$

Note that

$$\sum_{t=t_0}^{T-1}\frac{1}{\sqrt{t}}=\sum_{s=1}^{T-t_0}\frac{1}{\sqrt{s+t_0-1}}\le\sum_{s=1}^{T}\frac{1}{\sqrt{s+t_0-1}}\le\sum_{s=1}^{T}\frac{1}{\sqrt{s}}\le\int_1^T\frac{1}{\sqrt{s}}\,ds\le 2\sqrt{T}-1.$$

Thus,

$$\mathbb{E}\left\{\left|\frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,V_{i,t+1}\left[f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\right]\right|\right\}\le\frac{2M_1\sqrt{M_2}\sqrt{M_4}\,(2\sqrt{T}-1)}{T(1-t_0/T)}$$

and, under assumption 9,

$$\lim_{T\to\infty}\sqrt{T}\,\mathbb{E}\left\{\left|\frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,V_{i,t+1}\left[f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\right]\right|\right\}\le 4M_1\sqrt{M_2}\sqrt{M_4}.$$

□

PROOF OF PROPOSITION 2. From equation 1.8 it follows that

$$\begin{aligned}
\mathbb{E}\Big[\Big|\widehat{err}_T(\gamma)-err\Big|\Big]\le{}&+\mathbb{E}\left(\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,V_{i,t+1}^2-\sum_{i=1}^{n} w_i\,\boldsymbol{\Sigma}_{i,i}\right|\right)\\
&+\mathbb{E}\left\{\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\Big[f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\Big]^2\right|\right\}\\
&+\mathbb{E}\left\{\left|\frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\,V_{i,t+1}\left[f_i(\mathbf{X}_t,\boldsymbol{\Psi})-g_i(\mathbf{X}_t,\hat{\boldsymbol{\theta}}_t(\gamma))\right]\right|\right\}.
\end{aligned}$$

By lemmas 1–3 the second term has the slowest rate of convergence.

Therefore, under assumption 9, it holds that

$$\lim_{T\to\infty} \frac{T}{\ln T}\, \mathbb{E}\left[\left|\widehat{err}_T(\boldsymbol{\gamma}) - err\right|\right] \leq M_1\, M_4.$$

□

## 1.B.2. Proposition 3

Recall that with potentially incomplete data

$$\begin{aligned}\widehat{err}_T(\boldsymbol{\gamma}) = &+ \frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2 \\ &+ \frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\Big[f_i(\mathbf{X}_t, \boldsymbol{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))\Big]^2 \\ &+ \frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}\Big[f_i(\mathbf{X}_t, \boldsymbol{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))\Big].\end{aligned} \tag{1.9}$$

Lemma 4 analyses the first term of equation 1.9. This is then used for structuring the proof of proposition 3 (reported hereinafter).

**Lemma 4.** *Under assumptions 1–5, assumption 7, assumptions 9–10 and with potentially incomplete data, it holds that*

$$\lim_{T\to\infty} T\, \mathbb{E}\left(\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2 - err\right|^2\right) \leq M_1^2\, M_3.$$

PROOF. Similarly to lemma 1, using the bias-variance decomposition,

$$\mathbb{E}\left(\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2 - \sum_{i=1}^{n} w_i\, \boldsymbol{\Sigma}_{i,i}\right|^2\right)$$

can be re-written in the equivalent form

$$\left[\text{bias}\left(\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2\right)\right]^2 + \text{var}\left(\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2\right).$$

Consider that

$$\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2 = \frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\left(\mathbb{I}_{|\mathscr{D}(t+1)|=n}\sum_{i=1}^{n} w_i\, V_{i,t+1}^2 + \mathbb{I}_{|\mathscr{D}(t+1)|\neq n}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2\right).$$

It follows that the bias is

$$\begin{aligned}
&\operatorname{bias}\left(\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2\right)\\
&\quad= \frac{T-t_0-t_{NA}}{T-t_0}\sum_{i=1}^{n} w_i\,\boldsymbol{\Sigma}_{i,i} + \frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\mathbb{I}_{|\mathscr{D}(t+1)|\neq n}\sum_{i\in\mathscr{D}(t+1)} w_i\,\boldsymbol{\Sigma}_{i,i} - \sum_{i=1}^{n} w_i\,\boldsymbol{\Sigma}_{i,i}\\
&\quad\leq \frac{1-t_0/T-t_{NA}/T}{1-t_0/T}\sum_{i=1}^{n} w_i\,\boldsymbol{\Sigma}_{i,i} + \frac{t_{NA}/T}{1-t_0/T}M_5 - \sum_{i=1}^{n} w_i\,\boldsymbol{\Sigma}_{i,i},
\end{aligned}$$

where

$$M_5 := \sup_t\left(\mathbb{I}_{|\mathscr{D}(t+1)|\neq n}\sum_{i\in\mathscr{D}(t+1)} w_i\,\boldsymbol{\Sigma}_{i,i}\right)$$

and thus $0 \leq M_5 < \sum_{i=1}^{n} w_i\,\boldsymbol{\Sigma}_{i,i}$. Furthermore, following an approach analogous to lemma 1 (but allowing for potentially incomplete data), it holds that

$$\begin{aligned}
&\operatorname{var}\left(\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2\right) =\\
&\quad+\frac{1}{(T-t_0)^2}\sum_{t=t_0}^{T-1}\mathbb{I}_{|\mathscr{D}(t+1)|=n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i\, w_j\ \operatorname{cov}\left(V_{i,t+1}^2, V_{j,t+1}^2\right)\\
&\quad+\frac{1}{(T-t_0)^2}\sum_{t=t_0}^{T-1}\mathbb{I}_{|\mathscr{D}(t+1)|\neq n}\sum_{i\in\mathscr{D}(t+1)}\sum_{j\in\mathscr{D}(t+1)} w_i\, w_j\ \operatorname{cov}\left(V_{i,t+1}^2, V_{j,t+1}^2\right).
\end{aligned}$$

Under assumption 7,

$$\begin{aligned}
&\operatorname{var}\left(\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2\right)\\
&\quad\leq \frac{(T-t_0-t_{NA})\,(M_1^2\,M_3)}{(T-t_0)^2} + \frac{t_{NA}\,M_1^2\,M_6}{(T-t_0)^2}\\
&\quad\leq \frac{(1-t_0/T-t_{NA}/T)\,(M_1^2\,M_3)}{T(1-t_0/T)^2} + \frac{(t_{NA}/T)\,(M_1^2\,M_6)}{T(1-t_0/T)^2},
\end{aligned}$$

where

$$M_6 := \sup_t\left[\mathbb{I}_{|\mathscr{D}(t+1)|\neq n}\sum_{i\in\mathscr{D}(t+1)}\sum_{j\in\mathscr{D}(t+1)} |\operatorname{cov}(V_{i,t+1}^2, V_{j,t+1}^2)|\right]$$

and thus $0 \leq M_6 < M_3$.

Under assumptions 9–10,

$$\lim_{T\to\infty} \text{bias}\left(\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2\right) = 0,$$

$$\lim_{T\to\infty} T\,\text{var}\left(\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2\right) \leq M_1^2\, M_3.$$

As a result,

$$\lim_{T\to\infty} T\,\mathbb{E}\left(\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2 - err\right|^2\right) \leq M_1^2\, M_3.$$

□

**Remark** (Upper bounds in lemma 4). Defining $M_5$ and $M_6$ to be equal to $\sum_{i=1}^n w_i\, \mathbf{\Sigma}_{i,i}$ and $M_3$ would give loose upper bounds for the bias and variance. This would mask the relevancy of assumption 10, and it would not be ideal especially for the case in which $t_{NA}$ is large and all series are always jointly missing.

PROOF OF PROPOSITION 3. In the presence of potentially incomplete data,

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{err}_T(\boldsymbol{\gamma}) - err\right|\right] \leq &+ \mathbb{E}\left(\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2 - \sum_{i=1}^n w_i\, \mathbf{\Sigma}_{i,i}\right|\right)\\ &+ \mathbb{E}\left\{\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\Big[f_i(\mathbf{X}_t, \mathbf{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))\Big]^2\right|\right\}\\ &+ \mathbb{E}\left\{\left|\frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}\Big[f_i(\mathbf{X}_t, \mathbf{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))\Big]\right|\right\}.\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{err}_T(\boldsymbol{\gamma}) - err\right|\right] \leq &+ \mathbb{E}\left(\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i\in\mathscr{D}(t+1)} w_i\, V_{i,t+1}^2 - \sum_{i=1}^n w_i\, \mathbf{\Sigma}_{i,i}\right|\right)\\ &+ \mathbb{E}\left\{\left|\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\Big[f_i(\mathbf{X}_t, \mathbf{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))\Big]^2\right|\right\}\\ &+ \frac{2}{T-t_0}\sum_{t=t_0}^{T-1}\sum_{i=1}^{n} w_i\, \mathbb{E}\Big[|\,V_{i,t+1}\,|\,|f_i(\mathbf{X}_t, \mathbf{\Psi}) - g_i(\mathbf{X}_t, \hat{\boldsymbol{\theta}}_t(\boldsymbol{\gamma}))|\Big].\end{aligned}$$

By lemmas 2–4 the second term has the slowest rate of convergence. Thus, under assumptions 9–10,

$$\lim_{T\to\infty} \frac{T}{\ln T}\, \mathbb{E}\left[\left|\widetilde{err}_T(\boldsymbol{\gamma}) - err\right|\right] \leq M_1\, M_4.$$

even with potentially incomplete data. □

PROOF OF COROLLARY 3.1. It follows from the argument in section 1.A that $\widetilde{err}_T(\mathscr{J}, \boldsymbol{\gamma})$ is the average of a number of i.d. pseudo out-of-sample errors. Trivially, this implies that proposition 3 is also valid for this case. □

# 1.C. Estimation of penalised VARMA models for incomplete data

Traditional estimation methods for VARMA$(q, r)$ models are unable to handle incomplete time series. This is especially problematic for the scope of the manuscript since the artificial jackknife introduces missing values into the data. This appendix proposes overcoming the issue with an ECM algorithm (Meng and Rubin, 1993).

## 1.C.1. State-space representation

The ECM algorithm developed in this appendix is structured similarly to the EM algorithm in Shumway and Stoffer (1982) and Watson and Engle (1983), and thus starting from a model representation in state-space form.

**Definition 13** (State-space). Recall that $\min(q, r) = 0$ and $\max(q, r) \geq 1$, and let $m := nq + nr + n\,\mathbb{I}_{q=0}$. The representation chosen for the VARMA($q$, $r$) is such that, for any integer $t$,

$$\mathbf{Y}_t = \mathbf{B}\mathbf{X}_t + \boldsymbol{\epsilon}_t, \tag{1.10}$$

$$\mathbf{X}_t = \mathbf{C}\mathbf{X}_{t-1} + \mathbf{D}\tilde{\mathbf{V}}_t, \tag{1.11}$$

where $\mathbf{X}_t$ denotes a vector of $m$ latent states, $\boldsymbol{\epsilon}_t \overset{w.n.}{\sim} N\left(\mathbf{0}_{n\times 1}, \mathbf{R}\right)$, $\tilde{\mathbf{V}}_t \overset{w.n.}{\sim} N\left(\mathbf{0}_{n\times 1}, \tilde{\boldsymbol{\Sigma}}\right)$,

$$\mathbf{B} := \left( \mathbf{I}_n \,\middle|\, \mathbf{B}_* \right),$$

$$\mathbf{R} := \varepsilon \cdot \mathbf{I}_n,$$

$$\mathbf{D} := \left( \frac{\mathbf{I}_n}{\mathbf{0}_{m-n\times n}} \right),$$

$\varepsilon$ is a small positive real number and $\tilde{\mathbf{\Sigma}}$ is a $n \times n$ positive definite covariance matrix.[20]

The structure of $\mathbf{B}_*$ and $\mathbf{C}$ is described in definitions 14–15, differentiating between the VAR and VMA cases. It follows from these definitions that $\mathbf{V}_t \approx \tilde{\mathbf{V}}_t$ and $\mathbf{\Sigma} \approx \tilde{\mathbf{\Sigma}}$. The precision of these approximations is inversely proportional to $\varepsilon$.

**Definition 14** (Direct state-space representation: VAR). If $q > 0$ and $r = 0$,

$$\begin{aligned}
\mathbf{B}_* &:= \mathbf{0}_{n \times m-n} \\
\mathbf{C} &:= \left(\begin{array}{ccccc}
\tilde{\mathbf{\Pi}}_1 & \tilde{\mathbf{\Pi}}_2 & \dots & \tilde{\mathbf{\Pi}}_{q-1} & \tilde{\mathbf{\Pi}}_q \\
\hline
\mathbf{I}_n & \mathbf{0}_{n \times n} & \dots & \dots & \mathbf{0}_{n \times n} \\
\mathbf{0}_{n \times n} & \ddots & \ddots & \vdots & \vdots \\
\vdots & \ddots & \ddots & \vdots & \vdots \\
\mathbf{0}_{n \times n} & \dots & \dots & \mathbf{I}_n & \mathbf{0}_{n \times n}
\end{array}\right),
\end{aligned}$$

where $\tilde{\mathbf{\Pi}}_i$ denotes a $n \times n$ matrix, for any integer $1 \leq i \leq q$. With this representation, the vectors of latent states can be partitioned as

$$\mathbf{X}_t = \left(\tilde{\mathbf{Y}}_t' \quad \dots \quad \tilde{\mathbf{Y}}_{t-q+1}'\right)',$$

where $\mathbf{Y}_t \approx \tilde{\mathbf{Y}}_t$. The precision of this approximation is inversely proportional to $\varepsilon$.

**Definition 15** (Direct state-space representation: VMA). If $q = 0$ and $r > 0$,

$$\begin{aligned}
\mathbf{B}_* &:= \left( \tilde{\mathbf{\Xi}}_1 \quad \dots \quad \tilde{\mathbf{\Xi}}_r \right) \\
\mathbf{C} &:= \left(\begin{array}{ccccc}
\mathbf{0}_{n \times n} & \mathbf{0}_{n \times n} & \dots & \mathbf{0}_{n \times n} & \mathbf{0}_{n \times n} \\
\hline
\mathbf{I}_n & \mathbf{0}_{n \times n} & \dots & \dots & \mathbf{0}_{n \times n} \\
\mathbf{0}_{n \times n} & \ddots & \ddots & \vdots & \vdots \\
\vdots & \ddots & \ddots & \vdots & \vdots \\
\mathbf{0}_{n \times n} & \dots & \dots & \mathbf{I}_n & \mathbf{0}_{n \times n}
\end{array}\right),
\end{aligned}$$

where $\tilde{\mathbf{\Xi}}_i$ denotes a $n \times n$ matrix, for any integer $1 \leq i \leq r$. With this representation, the vectors of latent states can be partitioned as

$$\mathbf{X}_t = \left(\tilde{\mathbf{V}}_t' \quad \dots \quad \tilde{\mathbf{V}}_{t-r}'\right)'.$$

**Assumption 12** (Initial conditions). Consistently with assumption 1, it is assumed that the first observation for the measurements refers to $t = 1$. Therefore, the state-space

[20] In the empirical implementation reported in this manuscript $\varepsilon = 10^{-4}$.

representation is initialised such that $\mathbf{X}_0 \overset{w.n.}{\sim} N(\boldsymbol{\mu}_0, \boldsymbol{\Omega}_0)$, where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Omega}_0$ denote a generic $m \times 1$ vector and a $m \times m$ positive definite covariance matrix.

**Assumption 13** (Causality and invertibility of the state-space model). In all cases, the state-space parameters are assumed to be such that the model is causal and invertible. The region in which the parameters must lie is indicated with $\mathscr{R}$ (as in section 1.3.1).

**Example 1** (Vector autoregression of order 2). For this example, let $q = 2$ and $r = 0$. Under definition 14, the state-space representation in equations 1.10–1.11 becomes

$$\mathbf{Y}_t = \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{Y}}_t \\ \tilde{\mathbf{Y}}_{t-1} \end{pmatrix} + \boldsymbol{\epsilon}_t,$$

$$\begin{pmatrix} \tilde{\mathbf{Y}}_t \\ \tilde{\mathbf{Y}}_{t-1} \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{\Pi}}_1 & \tilde{\boldsymbol{\Pi}}_2 \\ \mathbf{I}_n & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{Y}}_{t-1} \\ \tilde{\mathbf{Y}}_{t-2} \end{pmatrix} + \begin{pmatrix} \mathbf{I}_n \\ \mathbf{0}_{n\times n} \end{pmatrix} \tilde{\mathbf{V}}_t.$$

**Example 2** (Vector moving average of order 2). For this example, let $q = 0$ and $r = 2$. Under definition 15, the state-space representation in equations 1.10–1.11 becomes

$$\mathbf{Y}_t = \begin{pmatrix} \mathbf{I}_n & \tilde{\boldsymbol{\Xi}}_1 & \tilde{\boldsymbol{\Xi}}_2 \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{V}}_t \\ \tilde{\mathbf{V}}_{t-1} \\ \tilde{\mathbf{V}}_{t-2} \end{pmatrix} + \boldsymbol{\epsilon}_t,$$

$$\begin{pmatrix} \tilde{\mathbf{V}}_t \\ \tilde{\mathbf{V}}_{t-1} \\ \tilde{\mathbf{V}}_{t-2} \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{I}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times n} & \mathbf{I}_n & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{V}}_{t-1} \\ \tilde{\mathbf{V}}_{t-2} \\ \tilde{\mathbf{V}}_{t-3} \end{pmatrix} + \begin{pmatrix} \mathbf{I}_n \\ \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times n} \end{pmatrix} \tilde{\mathbf{V}}_t.$$

## 1.C.2. The Expectation-Conditional Maximisation algorithm

The ECM algorithm builds on the penalised maximum likelihood problem described in definition 11 and proposes an iterative estimation method compatible with missing observations.

Let

$$\boldsymbol{\vartheta} := \begin{pmatrix} \operatorname{vec}(\mathbf{B})' & \operatorname{vech}(\mathbf{R})' & \operatorname{vec}(\mathbf{C})' & \operatorname{vech}(\tilde{\boldsymbol{\Sigma}})' & \operatorname{vec}(\boldsymbol{\mu}_0)' & \operatorname{vech}(\boldsymbol{\Omega}_0)' \end{pmatrix}'.$$

**Assumption 14.** Assume that the ECM algorithm is initialised as in section 1.C.3 and denote with $\hat{\boldsymbol{\vartheta}}_s^0(\boldsymbol{\gamma})$ the corresponding initial vector of VARMA coefficients.

The ECM algorithm proceeds from the initial value assigned under assumption 14 and repeats the process described in definition 16 until it converges.

**Assumption 15** (Convergence). The ECM algorithm is said to be converged when the conditions described in algorithm 1 are reached.

**Definition 16** (Estimation routine). At any iteration $k+1>0$, the ECM algorithm computes the vector of coefficients

$$\hat{\boldsymbol{\vartheta}}_s^{k+1}(\boldsymbol{\gamma}) := \underset{\underline{\boldsymbol{\vartheta}} \in \mathscr{R}}{\arg\max}\ \mathbb{E}\left[\mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Y}_{1:s}, \mathbf{X}_{1:s}) \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^{k}(\boldsymbol{\gamma})\right] - \mathbb{E}\left[\mathcal{P}(\underline{\boldsymbol{\vartheta}}, \boldsymbol{\gamma}) \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^{k}(\boldsymbol{\gamma})\right],$$

where $\mathscr{Y}(s)$ is the information set available at time $t=s$,

$$\begin{aligned}
\mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Y}_{1:s}, \mathbf{X}_{1:s}) \simeq & -\frac{1}{2}\ln|\underline{\boldsymbol{\Omega}_0}| - \frac{1}{2}\operatorname{Tr}\left[\underline{\boldsymbol{\Omega}_0}^{-1}(\mathbf{X}_0 - \underline{\boldsymbol{\mu}_0})(\mathbf{X}_0 - \underline{\boldsymbol{\mu}_0})'\right] \\
& -\frac{s}{2}\ln|\underline{\tilde{\boldsymbol{\Sigma}}}| - \frac{1}{2}\operatorname{Tr}\left[\sum_{t=1}^{s}\underline{\tilde{\boldsymbol{\Sigma}}}^{-1}(\mathbf{X}_{1:n,t} - \underline{\mathbf{C}}_*\mathbf{X}_{t-1})(\mathbf{X}_{1:n,t} - \underline{\mathbf{C}}_*\mathbf{X}_{t-1})'\right] \\
& -\frac{s}{2}\ln|\underline{\mathbf{R}}| - \frac{1}{2}\operatorname{Tr}\left[\sum_{t=1}^{s}\underline{\mathbf{R}}^{-1}(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)'\right],
\end{aligned}$$

$\underline{\mathbf{C}}_* \equiv \underline{\mathbf{C}}_{1:n,1:m}$ and the underlined coefficients denote the state-space parameters within $\underline{\boldsymbol{\vartheta}}$. The function $\mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Y}_{1:s}, \mathbf{X}_{1:s})$ is known as complete-data (i.e., as if the states were known and the data was fully observed) log-likelihood.[21] Finally,

$$\mathcal{P}(\underline{\boldsymbol{\vartheta}}, \boldsymbol{\gamma}) := \begin{cases} \frac{1-\alpha}{2}\left\|\underline{\mathbf{C}}_*\boldsymbol{\Gamma}(\boldsymbol{\gamma}, q)^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\left\|\underline{\mathbf{C}}_*\boldsymbol{\Gamma}(\boldsymbol{\gamma}, q)\right\|_{1,1} & \text{if } q>0 \text{ and } r=0, \\ \frac{1-\alpha}{2}\left\|\underline{\mathbf{B}}_*\boldsymbol{\Gamma}(\boldsymbol{\gamma}, r)^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\left\|\underline{\mathbf{B}}_*\boldsymbol{\Gamma}(\boldsymbol{\gamma}, r)\right\|_{1,1} & \text{if } q=0 \text{ and } r>0. \end{cases}$$

This function is a compact version of the elastic-net penalty in equation 1.7 for the state-space representations illustrated in this appendix.

Every $\hat{\boldsymbol{\vartheta}}_s^{k+1}(\boldsymbol{\gamma})$ is estimated via the so-called E-step and CM-step. The E-step corresponds to the operation of computing the complete-data penalised log-likelihood expectation, conditional on the parameters estimated at the $k$-th iteration and $\mathscr{Y}(s)$. The CM-step estimates $\hat{\boldsymbol{\vartheta}}_s^{k+1}(\boldsymbol{\gamma})$ to conditionally maximise the resulting expected penalised log-likelihood.

In order to formalise the E-step for the complete-data log-likelihood, it is convenient to clarify which measurements are observed at each point in time.

**Definition 17** (Observed measurements). Let

$$\begin{aligned}
\mathscr{T} &:= \bigcup_{i=1}^{n}\mathscr{T}_i, \\
\mathscr{T}(s) &:= \{t : t \in \mathscr{T},\ 1 \leq t \leq s\},
\end{aligned}$$

describe two sets representing the points in time (either over the full sample or up to

[21] Direct maximisation of the complete-data log-likelihood is not feasible since there are missing observations in the measurements.

time $s$) in which at least one measurement is observed, for $1 \leq s \leq T$. Thus, let

$$\begin{aligned}\mathbf{Y}_t^{obs} &:= \Big(Y_{i,t}\Big)_{i \in \mathscr{D}(t)} \\ \mathbf{B}_t^{obs} &:= \mathbf{A}_t \mathbf{B}\end{aligned}$$

be the vector of observed measurements at time $t$ and the corresponding $|\mathscr{D}(t)| \times m$ matrix of coefficients, for $t \in \mathscr{T}$. Every $\mathbf{A}_t$ is indeed a selection matrix constituted by ones and zeros that permits to retrieve the appropriate rows of $\mathbf{B}$ for every $t \in \mathscr{T}$.

Building on Shumway and Stoffer (1982) and Watson and Engle (1983), it is also handy to formalise the E-step by using the Kalman smoother output described in definition 18. Indeed, lemmas 5–6 build on that to compute a series of conditional expectations. Finally, proposition 4 uses these results to design the E-step for the complete-data log-likelihood.

**Definition 18** (Kalman smoother output). The Kalman smoother output used for formalising the E-step is

$$\begin{aligned}\hat{\mathbf{X}}_t &:= \mathbb{E}\left[\mathbf{X}_t \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right], \\ \hat{\mathbf{P}}_{t,t-j} &:= \operatorname{Cov}\left[\mathbf{X}_t, \mathbf{X}_{t-j} \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right],\end{aligned}$$

for any $k \geq 0$, $0 \leq j \leq t$ and $t \geq 0$. Let also $\hat{\mathbf{P}}_t \equiv \hat{\mathbf{P}}_{t,t}$.

**Remark.** These estimates can be computed using a range of different recursions. This article follows the approach in Durbin and Koopman (2012) for $\hat{\mathbf{X}}_t$ and $\hat{\mathbf{P}}_t$, and the one in Watson and Engle (1983) for $\hat{\mathbf{P}}_{t,t-1}$.

**Lemma 5.** *Building on definition 18, it follows that*

$$\mathbb{E}\left[\mathbf{X}_t\mathbf{X}_{t-j}' \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] = \hat{\mathbf{X}}_t\hat{\mathbf{X}}_{t-j}' + \hat{\mathbf{P}}_{t,t-j},$$

*for any $k \geq 0$, $0 \leq j \leq t$ and $t \geq 0$.*

PROOF. Note that $\mathbf{X}_t = \hat{\mathbf{X}}_t + (\mathbf{X}_t - \hat{\mathbf{X}}_t)$. Thus,

$$\begin{aligned}&\mathbb{E}\left[\mathbf{X}_t\mathbf{X}_{t-j}' \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] \\ &\quad = \mathbb{E}\left\{\mathbf{X}_t\Big[\hat{\mathbf{X}}_{t-j} + (\mathbf{X}_{t-j} - \hat{\mathbf{X}}_{t-j})\Big]' \,\Big|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right\} \\ &\quad = \mathbb{E}\left[\mathbf{X}_t\hat{\mathbf{X}}_{t-j}' \,\Big|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] + \mathbb{E}\left[\mathbf{X}_t\mathbf{X}_{t-j}' \,\Big|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] - \mathbb{E}\left[\mathbf{X}_t\hat{\mathbf{X}}_{t-j}' \,\Big|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right].\end{aligned}$$

Since

$$\mathbb{E}\left[\mathbf{X}_t\hat{\mathbf{X}}_{t-j}' \,\Big|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] = \mathbb{E}\left[\mathbf{X}_t \,\Big|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right]\mathbb{E}\left[\mathbf{X}_{t-j} \,\Big|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right]' = \hat{\mathbf{X}}_t\hat{\mathbf{X}}_{t-j}',$$

it holds that

$$\mathbb{E}\left[\mathbf{X}_t\mathbf{X}'_{t-j}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right] = \hat{\mathbf{X}}_t\hat{\mathbf{X}}'_{t-j} + \hat{\mathbf{P}}_{t,t-j}.$$

□

**Lemma 6.** *Building on definition 13 and definition 18, it follows that*

$$\mathbb{E}\left[\mathbf{X}_{i_1:i_2,t}\mathbf{X}'_{i_3:i_4,t-j}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right] = \left[\hat{\mathbf{X}}_t\hat{\mathbf{X}}'_{t-j} + \hat{\mathbf{P}}_{t,t-j}\right]_{i_1:i_2,i_3:i_4},$$

*for any* $k \geq 0$, $1 \leq i_1 \leq i_2 \leq m$, $1 \leq i_3 \leq i_4 \leq m$ *and* $0 \leq j \leq t$.

PROOF. Note that

$$\mathbf{X}_{i_1:i_2,t}\mathbf{X}'_{i_3:i_4,t-j} = \left[\mathbf{X}_t\mathbf{X}'_{t-j}\right]_{i_1:i_2,i_3:i_4}.$$

It then follows from the same logic employed in the proof of lemma 5 that

$$\mathbb{E}\left[\mathbf{X}_{i_1:i_2,t}\mathbf{X}'_{i_3:i_4,t-j}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right] = \left[\hat{\mathbf{X}}_t\hat{\mathbf{X}}'_{t-j} + \hat{\mathbf{P}}_{t,t-j}\right]_{i_1:i_2,i_3:i_4}.$$

□

**Proposition 4.** *Let*

$$\mathcal{L}_e\left[\underline{\boldsymbol{\vartheta}}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right] \equiv \mathbb{E}\left[\mathcal{L}(\underline{\boldsymbol{\vartheta}}\,|\,\mathbf{Y}_{1:s},\mathbf{X}_{1:s})\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right].$$

*Building on definition 13 and definition 18, it follows that*

$$\begin{aligned}\mathcal{L}_e\left[\underline{\boldsymbol{\vartheta}}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right] \simeq &-\frac{1}{2}\ln|\underline{\boldsymbol{\Omega}_0}| - \frac{1}{2}\operatorname{Tr}\left[\underline{\boldsymbol{\Omega}_0}^{-1}(\hat{\mathbf{E}} - \hat{\mathbf{X}}_0\underline{\boldsymbol{\mu}_0}' - \underline{\boldsymbol{\mu}_0}\hat{\mathbf{X}}_0' + \underline{\boldsymbol{\mu}_0}\,\underline{\boldsymbol{\mu}_0}')\right] \\ &-\frac{s}{2}\ln|\underline{\tilde{\boldsymbol{\Sigma}}}| - \frac{1}{2}\operatorname{Tr}\left[\underline{\tilde{\boldsymbol{\Sigma}}}^{-1}(\hat{\mathbf{F}}_s - \hat{\mathbf{G}}_s\underline{\mathbf{C}}'_* - \underline{\mathbf{C}}_*\hat{\mathbf{G}}_s' + \underline{\mathbf{C}}_*\hat{\mathbf{H}}_s\underline{\mathbf{C}}'_*)\right] \\ &-\frac{1}{2\varepsilon}\operatorname{Tr}\left\{\sum_{t\in\mathscr{T}(s)}\left[\left(\mathbf{Y}^{obs}_t - \underline{\mathbf{B}}^{obs}_t\hat{\mathbf{X}}_t\right)\left(\mathbf{Y}^{obs}_t - \underline{\mathbf{B}}^{obs}_t\hat{\mathbf{X}}_t\right)' + \underline{\mathbf{B}}^{obs}_t\hat{\mathbf{P}}_t\underline{\mathbf{B}}^{obs\prime}_t\right]\right\},\end{aligned}$$

*where*

$$\begin{aligned}\hat{\mathbf{E}} &:= \mathbb{E}\left[\mathbf{X}_0\mathbf{X}'_0\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right] = \hat{\mathbf{X}}_0\hat{\mathbf{X}}'_0 + \hat{\mathbf{P}}_0, \\ \hat{\mathbf{F}}_s &:= \sum_{t=1}^{s}\mathbb{E}\left[\mathbf{X}_{1:n,t}\mathbf{X}'_{1:n,t}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right] = \sum_{t=1}^{s}\left(\hat{\mathbf{X}}_t\hat{\mathbf{X}}'_t + \hat{\mathbf{P}}_t\right)_{1:n,1:n}, \\ \hat{\mathbf{G}}_s &:= \sum_{t=1}^{s}\mathbb{E}\left[\mathbf{X}_{1:n,t}\mathbf{X}'_{t-1}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right] = \sum_{t=1}^{s}\left(\hat{\mathbf{X}}_t\hat{\mathbf{X}}'_{t-1} + \hat{\mathbf{P}}_{t,t-1}\right)_{1:n,1:m}, \\ \hat{\mathbf{H}}_s &:= \sum_{t=1}^{s}\mathbb{E}\left[\mathbf{X}_{t-1}\mathbf{X}'_{t-1}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}^k_s(\boldsymbol{\gamma})\right] = \sum_{t=1}^{s}\left(\hat{\mathbf{X}}_{t-1}\hat{\mathbf{X}}'_{t-1} + \hat{\mathbf{P}}_{t-1}\right).\end{aligned}$$

PROOF. Note that

$$-\frac{s}{2}\ln|\underline{\mathbf{R}}| - \frac{1}{2}\operatorname{Tr}\left[\sum_{t=1}^{s}\underline{\mathbf{R}}^{-1}(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)'\right] \simeq -\frac{1}{2\varepsilon}\operatorname{Tr}\left[\sum_{t=1}^{s}(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)'\right],$$

since the covariance matrix $\underline{\mathbf{R}} = \varepsilon \cdot \mathbf{I}_n$. Thus, the complete-data log-likelihood

$$\begin{aligned}\mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Y}_{1:s}, \mathbf{X}_{1:s}) \simeq &-\frac{1}{2}\ln|\underline{\boldsymbol{\Omega}_0}| - \frac{1}{2}\operatorname{Tr}\left[\underline{\boldsymbol{\Omega}_0}^{-1}(\mathbf{X}_0\mathbf{X}_0' - \mathbf{X}_0\underline{\boldsymbol{\mu}_0}' - \underline{\boldsymbol{\mu}_0}\mathbf{X}_0' + \underline{\boldsymbol{\mu}_0}\,\underline{\boldsymbol{\mu}_0}')\right] \\ &-\frac{s}{2}\ln|\underline{\tilde{\boldsymbol{\Sigma}}}| - \frac{1}{2}\operatorname{Tr}\left[\sum_{t=1}^{s}\underline{\tilde{\boldsymbol{\Sigma}}}^{-1}(\mathbf{X}_{1:n,t}\mathbf{X}_{1:n,t}' - \mathbf{X}_{1:n,t}\mathbf{X}_{t-1}'\underline{\mathbf{C}}_*')\right] \\ &-\frac{1}{2}\operatorname{Tr}\left[\sum_{t=1}^{s}\underline{\tilde{\boldsymbol{\Sigma}}}^{-1}(-\underline{\mathbf{C}}_*\mathbf{X}_{t-1}\mathbf{X}_{1:n,t}' + \underline{\mathbf{C}}_*\mathbf{X}_{t-1}\mathbf{X}_{t-1}'\underline{\mathbf{C}}_*')\right] \\ &-\frac{1}{2\varepsilon}\operatorname{Tr}\left[\sum_{t=1}^{s}(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)'\right].\end{aligned} \tag{1.12}$$

It follows from definition 18 and lemma 5 that the expectation of the terms in the first row of equation 1.12, conditional on the information set $\mathscr{Y}(s)$ and $\hat{\boldsymbol{\vartheta}}_s^k(\gamma)$, is

$$-\frac{1}{2}\ln|\underline{\boldsymbol{\Omega}_0}| - \frac{1}{2}\operatorname{Tr}\left[\underline{\boldsymbol{\Omega}_0}^{-1}(\hat{\mathbf{E}} - \hat{\mathbf{X}}_0\underline{\boldsymbol{\mu}_0}' - \underline{\boldsymbol{\mu}_0}\hat{\mathbf{X}}_0' + \underline{\boldsymbol{\mu}_0}\,\underline{\boldsymbol{\mu}_0}')\right].$$

The following terms are a bit harder to handle. It follows from lemma 6 that

$$\begin{aligned}\mathbb{E}\Big[\mathbf{X}_{1:n,t}\mathbf{X}_{1:n,t}' \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\gamma)\Big] &= \Big(\hat{\mathbf{X}}_t\hat{\mathbf{X}}_t' + \hat{\mathbf{P}}_t\Big)_{1:n,1:n}, \\ \mathbb{E}\Big[\mathbf{X}_{1:n,t}\mathbf{X}_{t-1}' \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\gamma)\Big] &= \Big(\hat{\mathbf{X}}_t\hat{\mathbf{X}}_{t-1}' + \hat{\mathbf{P}}_{t,t-1}\Big)_{1:n,1:m}.\end{aligned}$$

Building on that, it holds that the expectation of the terms in the second and third row of equation 1.12, conditional on the information set $\mathscr{Y}(s)$ and $\hat{\boldsymbol{\vartheta}}_s^k(\gamma)$, is

$$-\frac{s}{2}\ln|\underline{\tilde{\boldsymbol{\Sigma}}}| - \frac{1}{2}\operatorname{Tr}\left[\underline{\tilde{\boldsymbol{\Sigma}}}^{-1}(\hat{\mathbf{F}}_s - \hat{\mathbf{G}}_s\underline{\mathbf{C}}_*' - \underline{\mathbf{C}}_*\hat{\mathbf{G}}_s' + \underline{\mathbf{C}}_*\hat{\mathbf{H}}_s\underline{\mathbf{C}}_*')\right].$$

Finally, it follows directly from Shumway and Stoffer (1982, Section 3) that

$$\begin{aligned}&-\frac{1}{2\varepsilon}\operatorname{Tr}\left\{\mathbb{E}\Big[\sum_{t=1}^{s}(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)(\mathbf{Y}_t - \underline{\mathbf{B}}\mathbf{X}_t)' \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\gamma)\Big]\right\} \\ &\simeq -\frac{1}{2\varepsilon}\operatorname{Tr}\left\{\sum_{t\in\mathscr{T}(s)}\left[\Big(\mathbf{Y}_t^{obs} - \underline{\mathbf{B}}_t^{obs}\,\hat{\mathbf{X}}_t\Big)\Big(\mathbf{Y}_t^{obs} - \underline{\mathbf{B}}_t^{obs}\,\hat{\mathbf{X}}_t\Big)' + \underline{\mathbf{B}}_t^{obs}\,\hat{\mathbf{P}}_t\,\underline{\mathbf{B}}_t^{obs'}\right]\right\}.\end{aligned}$$

□

**Lemma 7.** *Building on definition 13 and definitions 16–18, it follows that*

$$\mathbb{E}\left[\mathcal{P}(\underline{\boldsymbol{\vartheta}}, \gamma) \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\gamma)\right] = \mathcal{P}(\underline{\boldsymbol{\vartheta}}, \gamma).$$

PROOF. A formal proof is not reported since it is immediate. Indeed, the penalty function in this ECM algorithm depends on the current coefficients (i.e., $\underline{\mathbf{C}}_*$ or $\underline{\mathbf{B}}_*$) and hyperparameters only. □

Proposition 4 and lemma 7 give the structure of the expected penalised log-likelihood

$$\mathcal{M}_e\left[\underline{\boldsymbol{\vartheta}}, \gamma \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\gamma)\right] := \mathcal{L}_e\left[\underline{\boldsymbol{\vartheta}} \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\gamma)\right] - \mathcal{P}(\underline{\boldsymbol{\vartheta}}, \gamma). \tag{1.13}$$

The CM-step conditionally maximises $\mathcal{M}_e\left[\underline{\boldsymbol{\vartheta}}, \gamma \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\gamma)\right]$ to estimate the state-space coefficients. Lemmas 8–11 detail the estimation procedure. For internal consistency, the estimated coefficients are denoted with the same naming used in definitions 14–15, a “‘hat”’ symbol on top, an $s$ in the subscript to highlight the sample size and a superscript denoting the reference to the ECM iteration.

**Lemma 8.** *Building on definition 13 and definitions 16–18, it follows that the ECM estimators at a generic iteration* $k+1>0$ *for* $\boldsymbol{\mu}_0$ *and* $\boldsymbol{\Omega}_0$ *are*

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{0,s}^{k+1}(\gamma) &= \hat{\mathbf{X}}_0,\\ \hat{\boldsymbol{\Omega}}_{0,s}^{k+1}(\gamma) &= \hat{\mathbf{P}}_0.\end{aligned}$$

PROOF. The derivative of equation 1.13 with respect to $\underline{\boldsymbol{\mu}_0}$ is

$$\frac{\partial \mathcal{M}_e\left[\underline{\boldsymbol{\vartheta}}, \gamma \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\gamma)\right]}{\partial \underline{\boldsymbol{\mu}_0}} = -\frac{1}{2}\underline{\boldsymbol{\Omega}_0}^{-1}\left(-2\hat{\mathbf{X}}_0 + 2\underline{\boldsymbol{\mu}_0}\right).$$

It follows that the maximiser for the expected penalised log-likelihood is

$$\hat{\boldsymbol{\mu}}_{0,s}^{k+1}(\gamma) = \hat{\mathbf{X}}_0.$$

The derivative of equation 1.13 with respect to $\underline{\boldsymbol{\Omega}_0}$ and fixing $\underline{\boldsymbol{\mu}_0} = \hat{\boldsymbol{\mu}}_{0,s}^{k+1}(\gamma)$ is

$$-\frac{1}{2}\underline{\boldsymbol{\Omega}_0}^{-1} + \frac{1}{2}\underline{\boldsymbol{\Omega}_0}^{-1}\left[\hat{\mathbf{E}} - \hat{\mathbf{X}}_0\hat{\boldsymbol{\mu}}_{0,s}^{k+1'}(\gamma) - \hat{\boldsymbol{\mu}}_{0,s}^{k+1}(\gamma)\hat{\mathbf{X}}_0' + \hat{\boldsymbol{\mu}}_{0,s}^{k+1}(\gamma)\,\hat{\boldsymbol{\mu}}_{0,s}^{k+1'}(\gamma)\right]\underline{\boldsymbol{\Omega}_0}^{-1},$$

or,

$$-\frac{1}{2}\underline{\boldsymbol{\Omega}_0}^{-1} + \frac{1}{2}\underline{\boldsymbol{\Omega}_0}^{-1}\left[\hat{\mathbf{E}} - \hat{\mathbf{X}}_0\hat{\mathbf{X}}_0' - \hat{\mathbf{X}}_0\hat{\mathbf{X}}_0' + \hat{\mathbf{X}}_0\,\hat{\mathbf{X}}_0'\right]\underline{\boldsymbol{\Omega}_0}^{-1}.$$

Thus, due to the structure of $\hat{\mathbf{E}}$,

$$\hat{\boldsymbol{\Omega}}_{0,s}^{k+1}(\gamma) = \hat{\mathbf{P}}_0.$$

□

**Lemma 9.** *Partition the output of the following Cartesian as*

$$\{1, 2, \ldots, n\} \times \{1, 2, \ldots, m\} = \Big\{\mathscr{E}(i, j), \, (i, j), \, \mathscr{E}''(i, j)\Big\},$$

*and let*

$$\mathscr{F}(i, j) := \Big\{\mathscr{E}(i, j), \, \mathscr{E}''(i, j)\Big\},$$

*for any integer $1 \leq i \leq n$ and $1 \leq j \leq m$. Let also $\mathcal{S}(a, b) := \mathit{sign}(a) \max(|a| - b, 0)$ be the soft-thresholding operator, for any $a, b \in \mathbb{R}$. Building on definition 13 and definitions 16–18, it follows that, if $q > 0$ and $r = 0$, the ECM estimators at a generic iteration $k+1 > 0$ for $\mathbf{C}$ is such that*

$$\hat{\mathbf{C}}^{k+1}_{*,s}(\boldsymbol{\gamma}) = \begin{pmatrix} \hat{C}^{k+1}_{1,1,s}(\boldsymbol{\gamma}) & \ldots & \hat{C}^{k+1}_{1,m,s}(\boldsymbol{\gamma}) \\ \vdots & \ddots & \vdots \\ \hat{C}^{k+1}_{n,1,s}(\boldsymbol{\gamma}) & \ldots & \hat{C}^{k+1}_{n,m,s}(\boldsymbol{\gamma}), \end{pmatrix}$$

*where*

$$\hat{C}^{k+1}_{i,j,s}(\boldsymbol{\gamma}) = \frac{\mathcal{S}\left[\sum_{l_1=1}^{n} \hat{\tilde{\Sigma}}^{k-1}_{i,l_1,s}(\boldsymbol{\gamma}) \, \hat{G}_{l_1,j,s} - \sum_{(l_1,l_2)\in\mathscr{F}(i,j)} \hat{\tilde{\Sigma}}^{k-1}_{i,l_1,s}(\boldsymbol{\gamma}) \, \hat{C}^{k+\mathbb{I}_{(l_1,l_2)\in\mathscr{E}(i,j)}}_{l_1,l_2,s}(\boldsymbol{\gamma}) \, \hat{H}_{l_2,j,s}, \ \frac{\alpha}{2} \, \Gamma_{j,j}\left(\boldsymbol{\gamma}, q\right)\right]}{\hat{\tilde{\Sigma}}^{k-1}_{i,i,s}(\boldsymbol{\gamma}) \, \hat{H}_{j,j,s} + (1-\alpha) \, \Gamma_{j,j}\left(\boldsymbol{\gamma}, q\right)}$$

*for any integer $1 \leq i \leq n$ and $1 \leq j \leq m$, and the remaining entries are constant and specified according to the prescriptions in definition 14. If $q = 0$ and $r > 0$, it follows from definition 15 that $\hat{\mathbf{C}}^{k+1}_{*,s}(\boldsymbol{\gamma}) = \mathbf{0}_{n\times m}$.*

PROOF. This proof starts with the case in which $q > 0$ and $r = 0$. Given that the absolute value function in the penalty is not differentiable at zero, this part of the ECM algorithm estimates the free entries of $\mathbf{C}$ one-by-one starting from the $C_{1,1}$, in a column-major order and conditioning on a series of coefficients estimated in previous rounds of the same algorithm. Namely, it estimates every $C_{i,j}$ by fixing $\underline{\tilde{\boldsymbol{\Sigma}}} = \hat{\tilde{\boldsymbol{\Sigma}}}^{k}_{s}(\boldsymbol{\gamma})$ and any other free entry of $\mathbf{C}$ to the latest estimate available, with $1 \leq i \leq n$ and $1 \leq j \leq m$.[22] In other words, the derivative of equation 1.13 with respect to $C_{i,j}$ is taken having fixed the parameters as described in the last sentence. If $\underline{C}_{i,j} \neq 0$, this is

$$\begin{aligned} &+ \sum_{l_1=1}^{n} \hat{\tilde{\Sigma}}^{k-1}_{i,l_1,s}(\boldsymbol{\gamma}) \, \hat{G}_{l_1,j,s} - \hat{\tilde{\Sigma}}^{k-1}_{i,i,s}(\boldsymbol{\gamma}) \, \underline{C}_{i,j} \hat{H}_{j,j,s} - \sum_{(l_1,l_2)\in\mathscr{F}(i,j)} \hat{\tilde{\Sigma}}^{k-1}_{i,l_1,s}(\boldsymbol{\gamma}) \, \hat{C}^{k+\mathbb{I}_{(l_1,l_2)\in\mathscr{E}(i,j)}}_{l_1,l_2,s}(\boldsymbol{\gamma}) \, \hat{H}_{l_2,j,s} \\ &\qquad - (1-\alpha)\underline{C}_{i,j} \, \Gamma_{j,j}\left(\boldsymbol{\gamma}, q\right) - \frac{\alpha}{2} \, \Gamma_{j,j}\left(\boldsymbol{\gamma}, q\right) \mathrm{sign}(\underline{C}_{i,j}). \end{aligned}$$

[22] This approach is similar, in spirit, to Friedman et al. (2010).

It follows that

$$\hat{C}^{k+1}_{i,j,s}(\boldsymbol{\gamma}) = \frac{\mathcal{S}\left[\sum_{l_1=1}^{n} \hat{\tilde{\Sigma}}^{k-1}_{i,l_1,s}(\boldsymbol{\gamma})\, \hat{G}_{l_1,j,s} - \sum_{(l_1,l_2)\in\mathscr{F}(i,j)} \hat{\tilde{\Sigma}}^{k-1}_{i,l_1,s}(\boldsymbol{\gamma})\, \hat{C}^{k+\mathbb{I}_{(l_1,l_2)\in\mathscr{E}(i,j)}}_{l_1,l_2,s}(\boldsymbol{\gamma})\, \hat{H}_{l_2,j,s},\ \frac{\alpha}{2}\,\Gamma_{j,j}\left(\boldsymbol{\gamma},q\right)\right]}{\hat{\tilde{\Sigma}}^{k-1}_{i,i,s}(\boldsymbol{\gamma})\, \hat{H}_{j,j,s} + (1-\alpha)\,\Gamma_{j,j}\left(\boldsymbol{\gamma},q\right)}.$$

When $q = 0$ and $r > 0$ the coefficients of interest for this proof are not free parameters and fixed to zero as described in definition 15. □

**Lemma 10.** *Building on definition 13 and definitions 16–18, it follows that the ECM estimators at a generic iteration $k+1 > 0$ for $\tilde{\boldsymbol{\Sigma}}$ is*

$$\hat{\tilde{\boldsymbol{\Sigma}}}^{k+1}_{s}(\boldsymbol{\gamma}) = \frac{1}{s}\left[\hat{\mathbf{F}}_s - \hat{\mathbf{G}}_s \hat{\mathbf{C}}^{k+1'}_{*,s}(\boldsymbol{\gamma}) - \hat{\mathbf{C}}^{k+1}_{*,s}(\boldsymbol{\gamma})\,\hat{\mathbf{G}}'_s + \hat{\mathbf{C}}^{k+1}_{*,s}(\boldsymbol{\gamma})\,\hat{\mathbf{H}}_s\,\hat{\mathbf{C}}^{k+1'}_{*,s}(\boldsymbol{\gamma})\right].$$

PROOF. The derivative of equation 1.13 with respect to $\underline{\tilde{\boldsymbol{\Sigma}}}$ and fixing $\underline{\mathbf{C}}_* = \hat{\mathbf{C}}^{k+1}_{*,s}$ is

$$-\frac{s}{2}\,\underline{\tilde{\boldsymbol{\Sigma}}}^{-1} + \frac{1}{2}\,\underline{\tilde{\boldsymbol{\Sigma}}}^{-1}\left[\hat{\mathbf{F}}_s - \hat{\mathbf{G}}_s \hat{\mathbf{C}}^{k+1'}_{*,s}(\boldsymbol{\gamma}) - \hat{\mathbf{C}}^{k+1}_{*,s}(\boldsymbol{\gamma})\,\hat{\mathbf{G}}'_s + \hat{\mathbf{C}}^{k+1}_{*,s}(\boldsymbol{\gamma})\,\hat{\mathbf{H}}_s\,\hat{\mathbf{C}}^{k+1'}_{*,s}(\boldsymbol{\gamma})\right]\underline{\tilde{\boldsymbol{\Sigma}}}^{-1}.$$

It follows that

$$\hat{\tilde{\boldsymbol{\Sigma}}}^{k+1}_{s}(\boldsymbol{\gamma}) = \frac{1}{s}\left[\hat{\mathbf{F}}_s - \hat{\mathbf{G}}_s \hat{\mathbf{C}}^{k+1'}_{*,s}(\boldsymbol{\gamma}) - \hat{\mathbf{C}}^{k+1}_{*,s}(\boldsymbol{\gamma})\,\hat{\mathbf{G}}'_s + \hat{\mathbf{C}}^{k+1}_{*,s}(\boldsymbol{\gamma})\,\hat{\mathbf{H}}_s\,\hat{\mathbf{C}}^{k+1'}_{*,s}(\boldsymbol{\gamma})\right].$$

□

**Remark** (Vector moving average case). When $q = 0$ and $r > 0$, $\hat{\mathbf{C}}^{k+1}_{*,s}(\boldsymbol{\gamma}) = \mathbf{0}_{n\times m}$. Thus, it follows that $\hat{\tilde{\boldsymbol{\Sigma}}}^{k+1}_{s}(\boldsymbol{\gamma}) = \frac{1}{s}\,\hat{\mathbf{F}}_s$.

**Lemma 11.** *Let*

$$\begin{aligned} \hat{\mathbf{M}}_s &:= \sum_{t\in\mathscr{T}(s)} \mathbf{A}'_t \mathbf{Y}^{obs}_t\,\hat{\mathbf{X}}'_t, \\ \hat{\mathbf{N}}_t &:= \mathbf{A}'_t \mathbf{A}_t, \\ \hat{\mathbf{O}}_t &:= \hat{\mathbf{X}}_t \hat{\mathbf{X}}'_t + \hat{\mathbf{P}}_t. \end{aligned}$$

*Building on definition 13 and definitions 16–18, it follows that, if $q = 0$ and $r > 0$, the ECM estimators at a generic iteration $k+1 > 0$ for $\mathbf{B}$ is such that*

$$\hat{\mathbf{B}}^{k+1}_{*,s}(\boldsymbol{\gamma}) = \begin{pmatrix} \hat{B}^{k+1}_{1,1,s}(\boldsymbol{\gamma}) & \dots & \hat{B}^{k+1}_{1,nr,s}(\boldsymbol{\gamma}) \\ \vdots & \ddots & \vdots \\ \hat{B}^{k+1}_{n,1,s}(\boldsymbol{\gamma}) & \dots & \hat{B}^{k+1}_{n,nr,s}(\boldsymbol{\gamma}), \end{pmatrix}$$

*where*

$$\hat{B}_{i,j,s}^{k+1}(\boldsymbol{\gamma}) = \frac{\mathcal{S}\left[\hat{M}_{i,j,s} - \sum_{t\in\mathscr{T}(s)} \sum_{(l_1,l_2)\in\mathscr{F}(i,j)} \hat{N}_{i,l_1,t}\, \hat{B}_{l_1,l_2,s}^{k+\mathbb{I}_{(l_1,l_2)\in\mathscr{E}(i,j)}}(\boldsymbol{\gamma})\, \hat{O}_{l_2,j,t},\ \frac{\alpha}{2}\,\varepsilon\,\Gamma_{j-n,j-n}(\boldsymbol{\gamma}, r)\right]}{\sum_{t\in\mathscr{T}(s)} \hat{N}_{i,i,t}\, \hat{O}_{j,j,t} + (1-\alpha)\,\varepsilon\,\Gamma_{j-n,j-n}(\boldsymbol{\gamma}, r)}$$

*for any integer $1 \leq i \leq n$ and $n+1 \leq j \leq m$, and the remaining entries are constant and specified according to the prescriptions in definition 15. If $q > 0$ and $r = 0$, it follows from definition 14 that $\hat{\mathbf{B}}_{*,s}^{k+1}(\boldsymbol{\gamma}) = \mathbf{0}_{n\times m-n}$.*

PROOF. This proof starts with the case in which $q = 0$ and $r > 0$. Note that

$$\begin{aligned}
&\sum_{t\in\mathscr{T}(s)} \left[\left(\mathbf{Y}_t^{obs} - \underline{\mathbf{B}}_t^{obs}\, \hat{\mathbf{X}}_t\right)\left(\mathbf{Y}_t^{obs} - \underline{\mathbf{B}}_t^{obs}\, \hat{\mathbf{X}}_t\right)' + \underline{\mathbf{B}}_t^{obs}\, \hat{\mathbf{P}}_t\, \underline{\mathbf{B}}_t^{obs'}\right] \\
&\quad = \sum_{t\in\mathscr{T}(s)} \left[\left(\mathbf{Y}_t^{obs} - \mathbf{A}_t\underline{\mathbf{B}}\hat{\mathbf{X}}_t\right)\left(\mathbf{Y}_t^{obs} - \mathbf{A}_t\underline{\mathbf{B}}\hat{\mathbf{X}}_t\right)' + \mathbf{A}_t\underline{\mathbf{B}}\hat{\mathbf{P}}_t\underline{\mathbf{B}}'\mathbf{A}_t'\right] \\
&\quad = \sum_{t\in\mathscr{T}(s)} \left[\mathbf{Y}_t^{obs}\, \mathbf{Y}_t^{obs'} - \mathbf{Y}_t^{obs}\, \hat{\mathbf{X}}_t'\underline{\mathbf{B}}'\mathbf{A}_t' - \mathbf{A}_t\underline{\mathbf{B}}\hat{\mathbf{X}}_t\mathbf{Y}_t^{obs'} + \mathbf{A}_t\underline{\mathbf{B}}\left(\hat{\mathbf{X}}_t\hat{\mathbf{X}}_t' + \hat{\mathbf{P}}_t\right)\underline{\mathbf{B}}'\mathbf{A}_t'\right].
\end{aligned}$$

Since the absolute value function in the penalty is not differentiable at zero, this part of the ECM algorithm estimates the free entries of $\mathbf{B}$ one-by-one starting from the $B_{1,n+1}$, in a column-major order and conditioning on a series of coefficients estimated in previous rounds of the same algorithm. Indeed, as in lemma 9, the derivative of equation 1.13 with respect to $B_{i,j}$ is taken having fixed any other free entry of $\mathbf{B}$ to the latest estimate available, for $1 \leq i \leq n$ and $n+1 \leq j \leq m$. If $\underline{B}_{i,j} \neq 0$, this is

$$\begin{aligned}
&+\varepsilon^{-1}\hat{M}_{i,j,s} - \underline{B}_{i,j} \sum_{t\in\mathscr{T}(s)} \varepsilon^{-1}\hat{N}_{i,i,t}\, \hat{O}_{j,j,t} - \sum_{t\in\mathscr{T}(s)} \sum_{(l_1,l_2)\in\mathscr{F}(i,j)} \varepsilon^{-1}\hat{N}_{i,l_1,t}\, \hat{B}_{l_1,l_2,s}^{k+\mathbb{I}_{(l_1,l_2)\in\mathscr{E}(i,j)}}(\boldsymbol{\gamma})\, \hat{O}_{l_2,j,t} \\
&\quad - (1-\alpha)\underline{B}_{i,j}\, \Gamma_{j-n,j-n}(\boldsymbol{\gamma}, r) - \frac{\alpha}{2}\, \Gamma_{j-n,j-n}(\boldsymbol{\gamma}, r)\, \mathrm{sign}(\underline{B}_{i,j}).
\end{aligned}$$

It follows that

$$\hat{B}_{i,j,s}^{k+1}(\boldsymbol{\gamma}) = \frac{\mathcal{S}\left[\hat{M}_{i,j,s} - \sum_{t\in\mathscr{T}(s)} \sum_{(l_1,l_2)\in\mathscr{F}(i,j)} \hat{N}_{i,l_1,t}\, \hat{B}_{l_1,l_2,s}^{k+\mathbb{I}_{(l_1,l_2)\in\mathscr{E}(i,j)}}(\boldsymbol{\gamma})\, \hat{O}_{l_2,j,t},\ \frac{\alpha}{2}\,\varepsilon\,\Gamma_{j-n,j-n}(\boldsymbol{\gamma}, r)\right]}{\sum_{t\in\mathscr{T}(s)} \hat{N}_{i,i,t}\, \hat{O}_{j,j,t} + (1-\alpha)\,\varepsilon\,\Gamma_{j-n,j-n}(\boldsymbol{\gamma}, r)}.$$

When $q > 0$ and $r = 0$ the coefficients of interest for this proof are not free parameters and fixed to zero as described in definition 14. □

### 1.C.3. Initialisation of the ECM algorithm

In small-data settings, the ECM algorithm is initialised interpolating missing observations (if any) with sample average of the observed datapoints. If $q > 0$ and $r = 0$, the model is initialised via OLS (in small data settings) or ridge (in high-dimensional settings). If

$q = 0$ and $r > 0$, an estimate of the VMA innovations is computed by taking the sample residuals of a VAR with $\lfloor \sqrt{T} \rfloor$ lags. Indeed, these residuals can be interpreted as those of a truncated VAR($\infty$) resulting from an invertible VMA. The VMA coefficients are then initialised regressing the data on the estimated residuals (either with OLS or ridge, depending on the problem dimensionality).

In both cases, the approach in section 1.C.4 is used for making sure that the estimated coefficients are within the feasible region $\mathscr{R}$.

## 1.C.4. Enforcing causality and invertibility

The ECM algorithm makes sure that the autoregressive and moving average coefficients are causal and invertible. If, at any iteration $k + 1 > 1$, the matrix of autoregressive coefficients $\hat{\hat{\boldsymbol{\Pi}}}_s^{k+1}(\boldsymbol{\gamma})$ needs to be adjusted, it is replaced by the causal

$$\eta^{k+1} \, \hat{\hat{\boldsymbol{\Pi}}}_s^{k+1}(\boldsymbol{\gamma}) + (1 - \eta^{k+1}) \, \hat{\hat{\boldsymbol{\Pi}}}_s^{k}(\boldsymbol{\gamma})$$

associated to the largest feasible $\eta^{k+1} \in \{0, 0.1, 0.2, \ldots 0.9\}$. An analogous procedure is followed to adjust $\hat{\hat{\boldsymbol{\Xi}}}_s^{k+1}(\boldsymbol{\gamma})$ when necessary. This approach can be thought as a slowing mechanism that restricted the CM-step to the feasible region $\mathscr{R}$.

## 1.C.5. Estimation algorithm summary

**Algorithm 1:** VARMA with elastic-net penalty

Initialization

The ECM algorithm is initialised as described in section 1.C.3.

Estimation

```
for $k \leftarrow 1$ to $max\_iter$ do
    for $j \leftarrow 1$ to $m$ do
        Run the Kalman filter and smoother using $\hat{\boldsymbol{\vartheta}}_s^{k-1}(\boldsymbol{\gamma})$;
        if converged then
            Store the parameters and stop the loop.
        end
        Estimate $\hat{\boldsymbol{\mu}}_{s,0}^k(\boldsymbol{\gamma})$ and $\hat{\boldsymbol{\Omega}}_{s,0}^k(\boldsymbol{\gamma})$ as in lemma 8;
        Estimate $\hat{\mathbf{C}}_s^k(\boldsymbol{\gamma})$, $\hat{\hat{\boldsymbol{\Sigma}}}_s^k(\boldsymbol{\gamma})$ and $\hat{\mathbf{B}}_s^k(\boldsymbol{\gamma})$ as in lemmas 9–11;
        Build $\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})$;
    end
end
```

**Notes**

- The results are computed fixing *max_iter* to 1000. This is a conservative number, since the algorithm generally requires substantially less iterations to converge.
- The ECM algorithm is considered to be converged when the estimated coefficients (all relevant parameters in lemmas 9–11) do not significantly change in two subsequent iterations. This is done by computing the absolute relative change per parameters and comparing at the same time the median and $95^{th}$ quantile respectively with a fixed tolerance of $10^{-3}$ and $10^{-2}$. Intuitively, when the coefficients do not change much, the expected log-likelihood and the parameters in lemma 8 should also be stable.
- The scalar $\varepsilon$ is summed to the denominator of each relative change in order to ensure numerical stability.

The replication code for this paper is available on GitHub.

## 1.D. Tables and charts

| Mnemonic | Description | Transformation |
|---|---|---|
| australia | Foreign exchange rate: Australia / USA | Week-on-week (log-returns) |
| brazil | Foreign exchange rate: Brazil / USA | Week-on-week (log-returns) |
| canada | Foreign exchange rate: Canada / USA | Week-on-week (log-returns) |
| denmark | Foreign exchange rate: Denmark / USA | Week-on-week (log-returns) |
| ea | Foreign exchange rate: EA / USA | Week-on-week (log-returns) |
| india | Foreign exchange rate: India / USA | Week-on-week (log-returns) |
| japan | Foreign exchange rate: Japan / USA | Week-on-week (log-returns) |
| mexico | Foreign exchange rate: Mexico / USA | Week-on-week (log-returns) |
| new_zealand | Foreign exchange rate: New Zealand / USA | Week-on-week (log-returns) |
| norway | Foreign exchange rate: Norway / USA | Week-on-week (log-returns) |
| singapore | Foreign exchange rate: Singapore / USA | Week-on-week (log-returns) |
| south_africa | Foreign exchange rate: South Africa / USA | Week-on-week (log-returns) |
| south_korea | Foreign exchange rate: South Korea / USA | Week-on-week (log-returns) |
| sweden | Foreign exchange rate: Sweden / USA | Week-on-week (log-returns) |
| switzerland | Foreign exchange rate: Switzerland / USA | Week-on-week (log-returns) |
| taiwan | Foreign exchange rate: Taiwan / USA | Week-on-week (log-returns) |
| thailand | Foreign exchange rate: Thailand / USA | Week-on-week (log-returns) |
| uk | Foreign exchange rate: UK / USA | Week-on-week (log-returns) |

**Table 1.D.1:** Foreign exchange rates used for the empirical application in section 1.3.
**Notes**: The time series are collected from the Federal Reserve Board H.10 and include regular weekly (Friday, EOP) observations from January 1999 to the end of December 2020. This dataset contains a total of 1,148 weeks and 21,812 observations.
**Source**: Federal Reserve Board.

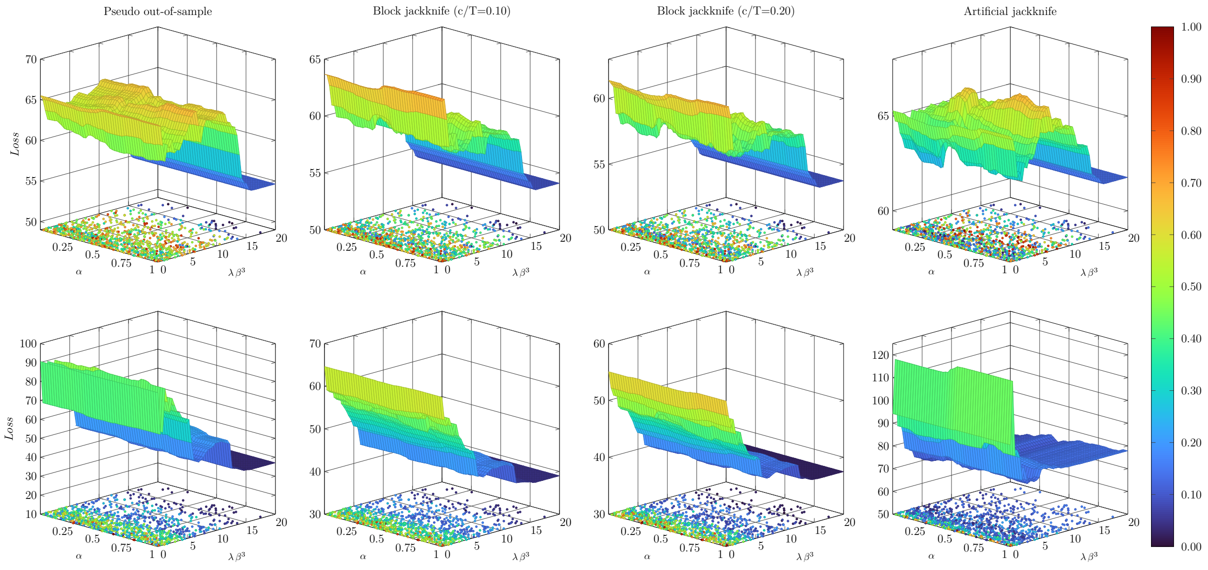


**Figure 1.D.1:** Expected error for the candidate hyperparameters in $\mathscr{H}$.
**Notes**: Alternative graphical representation for the VAR (first row) and VMA (second row) selection. The colormap highlights the relative scale of the expected error for each subsampling method. The scalar $\lambda\,\beta^3$ denotes the shrinkage associated to the farthest lag. The block jackknife output is adjusted to reduce the finite-sample methodological defects as described in section 1.2.2.

# 2 Factor-augmented tree ensembles

*This manuscript proposes to extend the information set of time-series regression trees with latent stationary factors extracted via state-space methods. In doing so, this approach generalises time-series regression trees on two dimensions. First, it allows to handle predictors that exhibit measurement error, non-stationary trends, seasonality and/or irregularities such as missing observations. Second, it gives a transparent way for using domain-specific theory to inform time-series regression trees. As a byproduct, this technique sets the foundations for structuring powerful ensembles. Their real-world applicability is studied under the lenses of empirical macro-finance.*

## 2.1. Introduction

In time series, the simplicity of regression trees (Morgan and Sonquist, 1963; Breiman et al., 1984; Quinlan, 1986) comes at a cost: irregularities, complicated periodic patterns and non-stationary trends cannot be explicitly modelled, and this is unfortunate given that many real-world examples are subject to them.

Following, in spirit, Harvey et al. (1998), this paper proposes to pre-process problematic predictors using state-space representations general enough to deal with all these complexities at once. This operation can be thought as an automated feature engineering process that extracts stationary patterns hidden across multiple predictors, while handling problematic data characteristics. Besides, when the state-space representation is compatible with domain-specific theory, this becomes a transparent way for extracting signals with structural interpretation. The stationary common components recovered from the data, referred hereinbelow as stationary dynamic factors, are then employed as regular predictors for standard time-series regression trees. This manuscript calls them factor-augmented regression trees to stress their dependence on latent components.

For this article, I have built on a broad body of theoretical research on time series. Indeed, factor models originated in psychometrics (Lawley and Maxwell, 1962) as a dimensionality reduction technique. They were later generalised to take into account the autocorrelation structure of time series with the work of Geweke (1977) on dynamic factor models. Over time, these methodologies have been further developed within the state-space literature pioneered by Harvey (1985) to be compatible with data exhibiting peculiar patterns (e.g., non-stationary trends, seasonality, missing observations). Relevant developments include Forni et al. (2000, 2005, 2009), Forni and Lippi (2001), Bernanke et al. (2005), Doz et al. (2012), Barigozzi and Luciani (2020), Li et al. (2020).

Factor-augmented regression trees are also strongly motivated by empirical results in economics and finance. Recent literature on semi-structural models, including Hasenzagl et al. (2022a,b), proposed to enrich statistical trend-cycle decompositions by using a minimal set of economic-driven restrictions. The main advantage of these semi-structural models is that they are able to extract unobserved cyclical and persistent components with economic interpretation, while allowing the data to speak. However, it is often hard to determine reasonable restrictions for most high-dimensional problems. Indeed, in macroeconomics and finance, theory is unclear on the exact dynamics of classical aggregate variables (e.g., stock market indices) and disaggregated indicators (e.g., single shares). Also, the literature is not mature enough to understand the precise drivers of new data (e.g., Google searches). Factor-augmented regression trees can be seen as a bridge between the output of small-dimensional semi-structural models (i.e., interpretable cyclical unobserved components) and time series that are not entirely understood from the theoretical standpoint and/or exhibit non-linear dynamics.

As for standard regression trees, the factor-augmented version can suffer from overfitting. Tree ensembles are an efficient way to reduce it without having to use complex vectors of hyperparameters. In order to do that, these methods generally fit a series of regression trees on a range of data subsamples and return aggregate forecasts. This article constructs the ensembles following Breiman (1996). These factor-augmented ensembles are similar to the rotation forest proposed in Rodriguez et al. (2006) and Pardo et al. (2013), but they take into account the autocorrelation structure in the data when estimating the factors and have the higher flexibility embedded in state-space modelling.

Their real-world applicability is studied under the lenses of empirical macro-finance. In particular, this article extracts a measure of the business cycle similar to the one employed in Hasenzagl et al. (2022a,b) and structure factor-augmented ensembles to target US equity volatility. Empirical results are encouraging and show that the forecasting accuracy of factor-augmented ensembles is notably higher compared to naive benchmarks and regular tree ensembles.

## 2.2. Methodology

### 2.2.1. Regression trees

This subsection describes the population model implied by standard regression trees and their most common estimation method (Breiman et al., 1984).

**Assumption 16** (Data). Let $T \in \mathbb{N}$ and $n \in \mathbb{N}_0$. Assume that $Y_t$ and $Z_{j,t}$ are finite realisations of some real-valued mean-stationary stochastic processes observed at the time periods $t = 1, \dots, T$ and with $j = 1, \dots, n$.

**Assumption 17** (Predictors of standard regression trees). Let $\mathbf{X}_t := (Y_t \; Z_{1,t} \; \dots \; Z_{n,t})'$ be $m \times 1$ dimensional and defined for any point in time $t \in \mathbb{Z}$.

**Remark.** Throughout the manuscript, the dependency on $n$ and $T$ is highlighted only when strictly necessary to ease the reading experience. Furthermore, specific realisations at some integer point in time $t$ and their general value in the underlying stochastic processes are denoted with the same symbols. However, it should be clear from the context whether the manuscript is referring to the first or second category.

This article describes the regression trees as non-linear forecasting models for $Y_t$ based on the information included in $\mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-p}$, for some number of lags $p \in \mathbb{N}$.[1]

**Assumption 18** (Lags). Let $0 < p \ll T - 1$.

**Assumption 19** (Regression tree model). In a regression tree setting,

$$Y_t = \sum_{i=1}^{|\mathscr{F}|} b_i \mathbb{I} \left\{ (\mathbf{X}_{t-1} \; \dots \; \mathbf{X}_{t-p}) \in \mathscr{F}_i \right\} + \epsilon_t,$$

whereas $\mathscr{F}$ is an indexed family of disjoint sets of matrices, every $b_i$ is a finite constant, $\epsilon_t \overset{i.i.d.}{\sim} (0, \sigma^2)$ with $\sigma > 0$ and finite, for any integer $t$ and $i = 1, \dots, |\mathscr{F}|$. Besides, regression trees assume that

$$\mathbb{E}(Y_t | \mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-p}, \mathbf{b}, \sigma, \mathscr{F}) = \sum_{i=1}^{|\mathscr{F}|} b_i \mathbb{I} \left\{ (\mathbf{X}_{t-1} \; \dots \; \mathbf{X}_{t-p}) \in \mathscr{F}_i \right\},$$

for any integer $t$.

Regression trees estimate $\mathbf{b}$, $\sigma$ and $\mathscr{F}$ recursively partitioning the predictor space to find the best fit. There are several modelling choices to take when performing this operation. This article follows common practice by focussing on binary partitions and using the CART algorithm (Breiman et al., 1984). At its first iteration, this estimation method looks for the best possible way to split the predictor space into two regions. This assessment is performed fitting a constant model in each region and minimising the mean square forecast error. Moreover, the splits are computed by inspecting, in turn, each covariate separately. The algorithm iteratively repeats the same operation for each of the resulting regions until some stopping criteria is reached. This manuscript uses DecisionTree.jl to implement it and refers to the estimated parameters with $\hat{\boldsymbol{\theta}}(\boldsymbol{\gamma})$ where $\boldsymbol{\gamma}$ is a vector of hyperparameters.

[1] Without loss of generality, this manuscript focusses on one-step ahead forecasts. Long-run predictions can be generated by computing direct forecasts in the same spirit of Marcellino et al. (2006).

### 2.2.2. Factor-augmented regression trees

This subsection introduces the factor-augmented regression trees: a version of the model in section 2.2.1 able to handle predictors with irregularities such as structural breaks and missing observations, intricate periodic patterns and non-stationary trends. In order to deal with these complexities, this subsection introduces a series of changes to the model and estimation algorithm.

Factor-augmented regression trees allow for these complexities in the data redefining $\mathbf{Z}_t$ and $\mathbf{X}_t$ as detailed in assumptions 20–22.

**Assumption 20** (State-space representation: data). Assume that $Z_{i,t}$ is finite realisations of some real-valued stochastic process observed at the time periods in the set $\mathscr{T}_i \subseteq \{t : t \in \mathbb{Z},\ 1 \leq t \leq T\}$ for every $i = 1, \ldots, n$.

**Assumption 21** (State-space representation: structure). Let $\mathbf{Z}_t$ be a $n \times 1$ real random vector that allows the state-space representation

$$\begin{aligned} Z_{i,t} &= g_{i,t}(\mathbf{\Phi}_t, \xi_{i,t}), \\ \mathbf{\Phi}_t &= \mathbf{h}_t(\mathbf{\Phi}_{t-1}, \boldsymbol{\zeta}_t), \end{aligned}$$

where $g_{i,t}$ and $\mathbf{h}_t$ are continuous and differentiable functions, $\mathbf{\Phi}_t$ denotes a vector of $q > 0$ latent states, $\boldsymbol{\xi}_t \overset{i.i.d.}{\sim} (\mathbf{0}_{n\times 1}, \mathbf{R}_t)$ and $\boldsymbol{\zeta}_t \overset{i.i.d.}{\sim} (\mathbf{0}_{q\times 1}, \mathbf{Q}_t)$, for any integer $t$ and $i = 1, \ldots, n$. Also, it is assumed that every $\mathbf{\Phi}_t$ includes a $\bar{q} \times 1$ vector of stationary common factors $\boldsymbol{\phi}_t$, with $0 < \bar{q} \ll n$ and $q \geq \bar{q}$. Since the observations start from the time period $t = 1$, it is further assumed that $\mathbf{\Phi}_1 = \mathbf{h}_0(\boldsymbol{\zeta}_0)$. This allows the evaluation of the state-space representation with the observed data.

**Remark** (Non-stationarity). Differently than with assumptions 16–17, assumption 20 does not assume that the underlying stochastic process is stationary. As a result, assumption 21 is compatible with non-stationary trends and co-integrated relationships.

**Assumption 22** (Predictors of factor-augmented trees). Factor-augmented regression trees include the stationary common factors in the predictors (as is, transformed in a way that does not alter data ordering and preserves stationarity, or both). Formally, this is achieved including these common components in the predictor matrix $\mathbf{X}_t$ jointly with $Y_t$ and updating $m$ accordingly.

**Remark** (Information set). Factor-augmented regression trees extend the information set of a tree autoregression for $Y_t$ with stationary common factors, while discarding idiosyncratic noise in the predictors and non-stationary trends, and handling data irregularities. The simplest case is when the predictor matrix is extended to include these stationary factors as they come out from the state-space. Formally, this is achieved by letting $\mathbf{X}_t := (Y_t \ \ \boldsymbol{\phi}_t)$ be a $m \times 1$ vector of time series, with $m := 1 + \bar{q}$.

The structure of the model is exactly as described in section 2.2.1, but uses the newly defined predictor matrix and value for $m$. However, the estimation process is different and structured as a two-step method. In the first step, the state space in assumption 21 is estimated with any algorithm compatible with the data complexities described above, including, but not limited to, the EM (Dempster et al., 1977; Rubin and Thayer, 1982; Shumway and Stoffer, 1982; Watson and Engle, 1983; Bańbura and Modugno, 2014; Barigozzi and Luciani, 2020), ECM (Meng and Rubin, 1993; Pellegrino, 2023a) and ECME algorithms (Liu and Rubin, 1994).[2] In the second and final step, the predictor matrix is formed on the basis of the estimated states and the regression tree is trained with CART.

It is worth stressing that the main difference between factor-augmented regression trees and the individual base learners of rotation forests lies in the technique used for reducing the dimensionality of the data. Instead of using Principal Component Analysis, factor-augmented regression tree models use a state-space. In doing so, this approach explicitly models the temporal factors dynamics[3], permits to pinpoint specific unobserved components and allows for data that exhibits peculiar patterns such as non-stationary trends.[4] Example 3 describes a correlated empirical problem, in which the state-space is used for extracting a factor compatible with structural economic interpretation. This helps stressing further the empirical motivation underlying these techniques.

**Example 3** (Financial returns and the business cycle). Finding empirical relationships between financial and macroeconomic data is difficult given their complex dynamics. Academic insights indicate that financial returns are linked to macroeconomic fundamentals in an undefined non-linear fashion (e.g., in periods of high economic uncertainty, they react differently to new information with respect to normal times).

Theoretically, a simple way to exploit this behaviour in forecasting would be running a non-linear predictive regression using the lagged business cycle as predictor and some function of a financial return of interest as a response. However, this is easier said than done since the business cycle itself is an unobserved variable that reflects the cyclical co-movement between a series of non-stationary economic indicators (e.g., output, employment and inflation). Also, the non-linear links have an unclear form, and thus it is hard to model them in a parametric way.

Factor-augmented regression trees represent a simple approach to the problem, com-

[2] Bayesian techniques surveyed, for instance, in Särkkä (2013) can also be used. In that case, $\boldsymbol{\phi}_t$ would be a point estimate (e.g., mean or median) of the stationary dynamic factors distribution at time $t$.

[3] This is fundamentally the same difference between traditional and dynamic factor models (see, for example, Barigozzi and Luciani, 2020, for a comparison between these approaches).

[4] Factor-augmented regression trees could be extended to use selected idiosyncratic periodic patterns as additional predictors. This could be done by redefining $\boldsymbol{\phi}$ into a vector of "selected cycles", both common and idiosyncratic. However, this would increase the computational burden and, without limitations, the risk of generating spurious splits.

patible with its complexities. The state-space in assumption 21 can be thought as a way for extracting the business cycle from a set of predictors, and the regression tree as a model that does not require an a-priori parametrisation of the non-linear link between macroeconomic and financial data.

**Remark.** Example 3 is discussed in section 2.3 with greater detail. While the emphasis in this article is given to economic and financial data, similar time-series models are applicable in other fields including geography, meteorology and engineering. Examples can be found in Harvey (1990).

### 2.2.3. Tree ensembles

Tree ensembles are methods that combine multiple regression trees, in order to produce more efficient predictions than the individual base learners (i.e., the trees themselves).

For simplicity of illustration, this article focusses on ensemble averaging and, in particular, on bootstrap aggregating or bagging (Breiman, 1996).[5] This method obtains the increase in efficiency estimating a large number of regression trees on random data subsamples and combining their predictions taking a sample average. Intuitively, this reduces over-fitting since the base learners are not trained on the original data, but on random subsamples generated from it. The more heterogeneous and numerous the subsamples the better in terms of efficiency.[6]

This article follows common practice and uses the bootstrap version proposed in Efron and Gong (1983, section 7) to generate the subsamples. This approach considers each covariate-response pair as a single datapoint and constructs data subsamples via independent bootstrap (Efron, 1979a,b, 1981). In other words, it resamples covariate-response pairs from the original data to generate the subsamples. In particular, in the case of the factor-augmented trees this is done focussing on the factor-response pairs.

## 2.3. Results

### 2.3.1. Data

This section develops further the narrative in example 3 and illustrates how factor-augmented tree ensembles are an effective technique for empirical macro-finance.

The problem at hand consists in forecasting US equity volatility[7] for the financial indices in table 2.3.1 as a function of its own past and a dynamic factor identifying the

[5] That being said, factor-augmented trees could be used for structuring more complex tree ensembles.
[6] This can be formally established following an approach equivalent to Hastie et al. (2009, section 15.2).
[7] Measured in terms of squared returns.

| Mnemonic | Description | Transformation | Source |
|---|---|---|---|
| TCU | Capacity utilization: total index | Levels | FRB |
| INDPRO | Industrial production: total index | Levels | FRB |
| RPCE | Real personal consumption expendit. | Levels | BEA |
| PAYEMS | Total nonfarm employment | Levels | BLS |
| EMRATIO | Employment-population ratio | Levels | BLS |
| UNRATE | Unemployment rate | Levels | BLS |
| WTISPLC | Spot crude oil price (WTI) | YoY returns | FRBSL |
| CPIAUCNS | CPI: all items | YoY returns | BLS |
| CPILFENS | CPI: all items excl. food and energy | YoY returns | BLS |
| WILL5000IND | Wilshire 5000 TMI | MoM returns (squared) | WA |
| WILLLRGCAP | Wilshire US Large-Cap TMI | MoM returns (squared) | WA |
| WILLLRGCAPVAL | Wilshire US Large-Cap Value TMI | MoM returns (squared) | WA |
| WILLLRGCAPGR | Wilshire US Large-Cap Growth TMI | MoM returns (squared) | WA |
| WILLMIDCAP | Wilshire US Mid-Cap TMI | MoM returns (squared) | WA |
| WILLMIDCAPVAL | Wilshire US Mid-Cap Value TMI | MoM returns (squared) | WA |
| WILLMIDCAPGR | Wilshire US Mid-Cap Growth TMI | MoM returns (squared) | WA |
| WILLSMLCAP | Wilshire US Small-Cap TMI | MoM returns (squared) | WA |
| WILLSMLCAPVAL | Wilshire US Small-Cap Value TMI | MoM returns (squared) | WA |
| WILLSMLCAPGR | Wilshire US Small-Cap Growth TMI | MoM returns (squared) | WA |

**Table 2.3.1:** Monthly macro-financial indicators. The macroeconomic data is sampled from January 1984 to December 2020 and downloaded in a real-time fashion from the Archival Federal Reserve Economic Data (ALFRED) database. The financial indicators are sampled from January 1984 to January 2021 and downloaded from the Federal Reserve Economic Data (FRED) database.
**Notes**: Table 2.B.1 provides a glossary for the acronyms.

state of the economy, in a real-time fashion. In order to estimate the state of the economy, this section uses a state-space representation similar, in spirit, to the one proposed in Hasenzagl et al. (2022a,b). This modelling choice implies that each macroeconomic indicator in table 2.3.1 is considered as the sum of non-stationary trends and causal cycles, one of which can be interpreted as the US business cycle. The trends account for the persistence in the data and provide a view on a series of structural components such as the natural rate of unemployment and trend inflation. By linking together key variables such as the real personal consumption expenditures, unemployment rate and inflation through the business cycle, the model is compatible with economic relationships such as the Phillips curve and the Okun's law (interpreting consumption as a proxy for GDP). A complex lag structure in the coefficients associated with the business cycle allows to take into account frictions in the economy (for instance, in the labour market). Finally, the idiosyncratic cycles account for autocorrelation in the error terms (if any). This is all formalised in assumption 23.

**Assumption 23** (State-space representation: trend-cycle model). For any integer $t$, let $\mathbf{Z}_t$ represent the macroeconomic indicators in table 2.3.1 (first block of series reported in the table, in the same order) referring to time $t$. Let also the data in $\mathbf{Z}_t$ be standardised such that each $i$-th series is divided for a given scaling factor $\eta_i$, for $i = 1, \dots, n$ with

$n = 9$. Hence, assume that

$$\begin{pmatrix} Z_{1,t} \\ Z_{2,t} \\ Z_{3,t} \\ Z_{4,t} \\ Z_{5,t} \\ Z_{6,t} \\ Z_{7,t} \\ Z_{8,t} \\ Z_{9,t} \end{pmatrix} = \begin{pmatrix} \tau_{1,t} \\ \tau_{2,t} \\ \tau_{3,t} \\ \tau_{4,t} \\ \tau_{5,t} \\ \tau_{6,t} \\ \tau_{7,t} \\ \frac{\tau_{8,t}}{\eta_8} \\ \frac{\tau_{8,t}}{\eta_9} \end{pmatrix} + \begin{pmatrix} 1 \\ \Upsilon_{1,1} + \Upsilon_{1,2} L + \ldots + \Upsilon_{1,p} L^{p-1} \\ \Upsilon_{2,1} + \Upsilon_{2,2} L + \ldots + \Upsilon_{2,p} L^{p-1} \\ \Upsilon_{3,1} + \Upsilon_{3,2} L + \ldots + \Upsilon_{3,p} L^{p-1} \\ \Upsilon_{4,1} + \Upsilon_{4,2} L + \ldots + \Upsilon_{4,p} L^{p-1} \\ \Upsilon_{5,1} + \Upsilon_{5,2} L + \ldots + \Upsilon_{5,p} L^{p-1} \\ \Upsilon_{6,1} + \Upsilon_{6,2} L + \ldots + \Upsilon_{6,p} L^{p-1} \\ \Upsilon_{7,1} + \Upsilon_{7,2} L + \ldots + \Upsilon_{7,p} L^{p-1} \\ \Upsilon_{8,1} + \Upsilon_{8,2} L + \ldots + \Upsilon_{8,p} L^{p-1} \end{pmatrix} \psi_{1,t} + \begin{pmatrix} \psi_{2,t} \\ \psi_{3,t} \\ \psi_{4,t} \\ \psi_{5,t} \\ \psi_{6,t} \\ \psi_{7,t} \\ \psi_{8,t} \\ \psi_{9,t} \\ \psi_{10,t} \end{pmatrix} + \boldsymbol{\xi}_t$$

where $\psi_{1,t}$ is a causal AR($p$) cycle; $\tau_{1,t}, \ldots, \tau_{8,t}$ are second-order smooth trends (Kitagawa and Gersch, 1996, section 8.1); $\psi_{2,t}, \ldots, \psi_{10,t}$ are causal AR(1) idiosyncratic noises; $\boldsymbol{\xi}_t \overset{w.n.}{\sim} N\left(\mathbf{0}_{9\times1}, \varepsilon \cdot \mathbf{I}_9\right)$ for a small positive $\varepsilon$, similarly to Bańbura and Modugno (2014).[8] Hereinafter, the number of lags $p$ is assumed being equal to 12 (months).

**Remark** (Business cycle). $\psi_{1,t}$ represents the business cycle at time $t$.

**Remark** (Trend inflation). Headline and core inflation share a common trend.

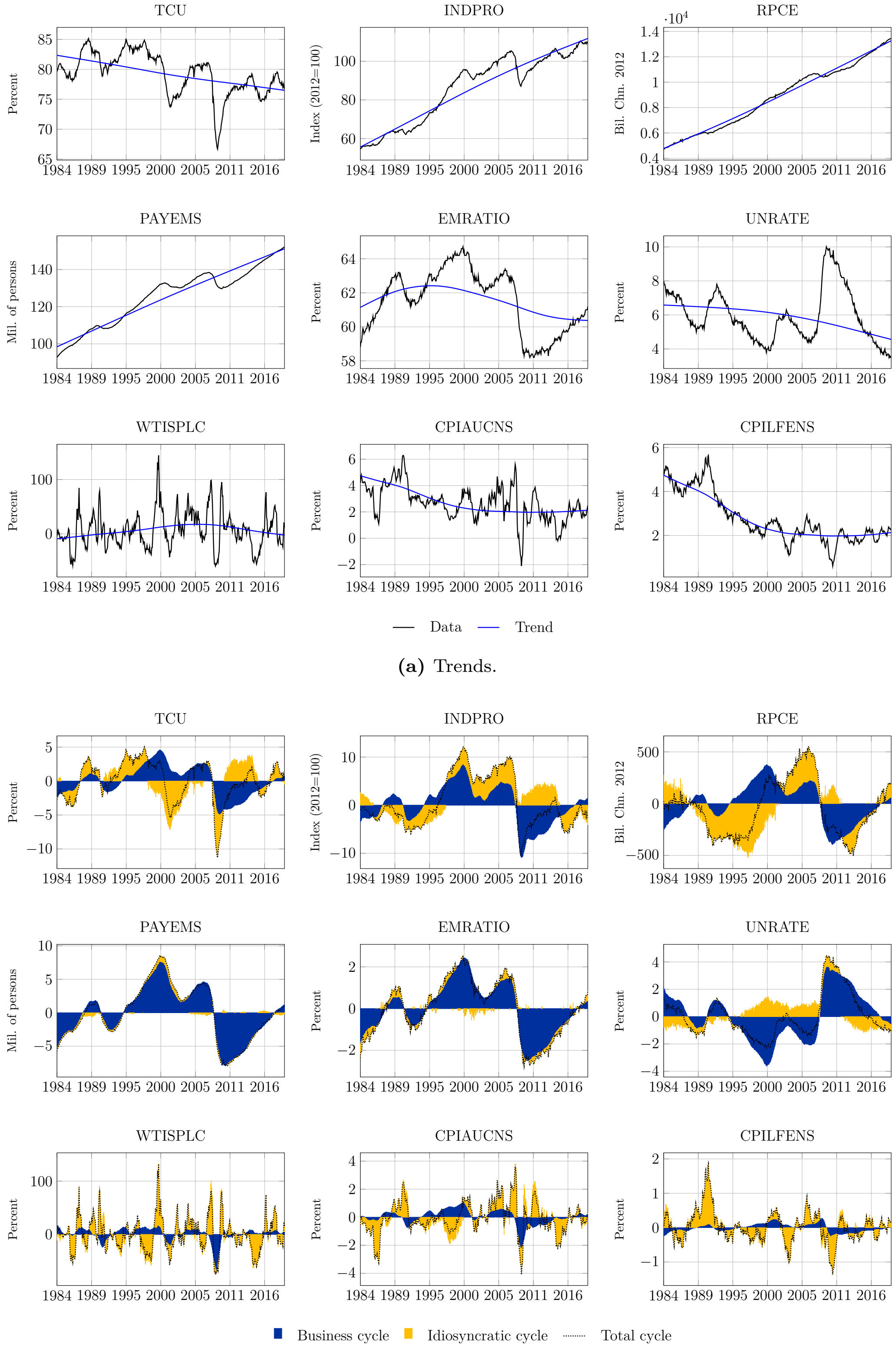
The dynamics for the latent states and the estimation method for this trend-cycle model are further detailed in section 2.A. The estimation process uses an elastic-net penalty analogous to the one in Pellegrino (2023a). Post estimation, the standardisation is removed to attribute the original scaling. In doing so, the scaling factors associated with trend inflation are also removed. Hence, headline and core inflation have the exact same trend once the standardisation is lifted.

Figure 2.3.1 shows an in-sample snapshot of the economy captured by this trend-cycle decomposition. This is obtained by estimating the model with monthly data from January 1984 to January 2020. The top subplot compares the macroeconomic data in levels with the estimated trends, while the bottom subplot decomposes the cycles into common and idiosyncratic fluctuations. These results indicate a strong heterogeneity in the data. Nonetheless, they also show that the business cycle is synchronised with the NBER's recession dates and able to explain most of the cyclical fluctuations. This is in line with the results in Hasenzagl et al. (2022a,b).

## 2.3.2. Empirical settings for the tree ensembles

The factor-augmented ensembles considered for this empirical problem extend the information set of traditional autoregression trees by including an estimate of the US business

[8] In this empirical example, $\varepsilon = 10^{-4}$.



**(a)** Trends.

**(b)** Historical decomposition of the cycles.

**Figure 2.3.1:** In-sample output of the trend-cycle decomposition.
**Notes**: The model is estimated with monthly data from January 1984 to December 2020 (full dataset as at 28th February 2020).

cycle. The latter is used both in levels and via a selected range of transformations. Formally, in order to compute a prediction referring to a generic time $t+1$, the factor-augmented ensembles use a vector of predictors containing the target values referring to time $t, \ldots, t-11$ and the augmentation

$$\left(\begin{array}{c}
\hat{\psi}_{1,t+11|t} \\
\vdots \\
\hat{\psi}_{1,t-11|t} \\
\hline
\hat{\psi}_{1,t+11|t} - \hat{\psi}_{1,t+10|t} \\
\vdots \\
\hat{\psi}_{1,t-10|t} - \hat{\psi}_{1,t-11|t} \\
\hline
\hat{\psi}_{1,t+11|t} - \hat{\psi}_{1,t|t} \\
\vdots \\
\hat{\psi}_{1,t+2|t} - \hat{\psi}_{1,t|t} \\
\hline
\hat{\psi}_{1,t|t} - \hat{\psi}_{1,t-2|t} \\
\vdots \\
\hat{\psi}_{1,t|t} - \hat{\psi}_{1,t-11|t}
\end{array}\right),$$

where $\hat{\psi}_{1,t+j|t}$ denotes the estimate of the business cycle for a generic period $t+j$ computed with the information set available at time $t$. While the first block in the factor augmentation gives a direct view on the business cycle levels, the following ones are useful for computing splits directly on its turning points and making a better use of the data.

Each ensemble is regulated via a vector of hyperparameters that includes those specifics to the elastic-net penalty of the state-space model (section 2.A) and the minimum number of observations per leaf. These tuning parameters are determined on a sample going from January 1984 to the end of January 2005. The ALFRED data vintage used for structuring the macroeconomic selection sample includes the information was available right before the end of January 2005. Since this article uses a two-step method, hyperparameters are selected first for the trend-cycle model and then for the factor-augmented ensembles. The trend-cycle model is tuned as illustrated in section 2.A.5. Next, the minimum number of observations per leaf of each ensemble is determined with a pseudo out-of-sample criterion and a grid search on the equally spaced $\mathscr{H}_{RT} := \{0.01, 10, 15, \ldots, 0.5\}$ with $|\mathscr{H}_{RT}| = 25$.[9] The minimum number of observations per leaf is expressed in percentage terms with respect to the number of time periods available. Both steps use the first half of the selection sample for the estimation and the second half to validate the results.

[9] This difference in the selection method is determined by the higher computational complexity required to estimate and forecast with factor-augmented tree ensembles.

### 2.3.3. Model evaluation

Having selected the hyperparameters, these factor-augmented tree ensembles are then estimated, in turn, for each target on the full selection sample. Next, they are tested in pseudo out-of-sample on the remaining observations. This operation is performed within an online framework in which the macroeconomic data is downloaded in the form of real-time vintages from the Archival Federal Reserve Economic Data (ALFRED) database.[10] This ensures that models do not "cheat" by looking forward in time. The models are re-estimated every time a new ALFRED vintage is released.

| Target | Pre COVID-19 | | Post COVID-19 | |
|---|---|---|---|---|
| | Autoregressive | Augmented | Autoregressive | Augmented |
| WILL5000IND | 0.760 | **0.739** | 0.754 | **0.739** |
| WILLLRGCAP | 0.768 | **0.745** | 0.756 | **0.738** |
| WILLLRGCAPVAL | 0.770 | **0.765** | 0.779 | **0.773** |
| WILLLRGCAPGR | 0.765 | **0.702** | 0.729 | **0.689** |
| WILLMIDCAP | 0.783 | **0.758** | 0.815 | **0.803** |
| WILLMIDCAPVAL | 0.784 | **0.763** | 0.863 | **0.859** |
| WILLMIDCAPGR | 0.835 | **0.720** | 0.799 | **0.732** |
| WILLSMLCAP | 0.750 | **0.704** | 0.788 | **0.767** |
| WILLSMLCAPVAL | 0.757 | **0.753** | **0.816** | 0.822 |
| WILLSMLCAPGR | 1.132 | **0.683** | 1.007 | **0.713** |

**Table 2.3.2:** Mean squared error relative to a forecast constant at zero. Values lower than 1 denote cases where this naive benchmark was outperformed by alternative forecasting models.
**Notes**: The mean squared errors are computed using a one-month ahead forecast horizon, in real-time, over the target observations spanning from February 2005 to January 2021. The columns marked as "Autoregressive" refer to ensembles whose predictors are the lags of the target variable. The columns marked as "Augmented" refer to the factor-augmented ensembles in section 2.3. The Pre COVID-19 period uses the ALFRED vintages up to the 28th February 2020 release (included) and the corresponding Wilshire data.

Table 2.3.2 summarises the pseudo out-of-sample results in the form of mean squared error relative to a forecast constant at zero, a simple naive benchmark. The baseline ensembles do not use the factor-augmentation described above. The output shows that the business cycle is helpful in forecasting even in the post COVID-19 period. In order to better understand the predictability drivers, this section uses figure 2.3.2 and the additional figures in section 2.B. These charts compare the factor-augmented ensembles estimated pre and post COVID-19 (i.e., estimating it first with data as at 28th February 2020 and then with the latest vintage available) by looking at the bagging importance weights: the number of times, in percentage points, that a predictor is selected to create a split in the underlying factor-augmented regression trees. Essentially, an internal ranking of the predictors. Figure 2.3.2 reports the total importance of the lagged target and factor augmentation. Mid to small cap shares are usually more vulnerable than blue chips to changes in economic conditions. Indeed, a broad range of papers such as Gertler

[10] The macroeconomic test sample includes 861 vintages and a minimum of 2277 observations per vintage.

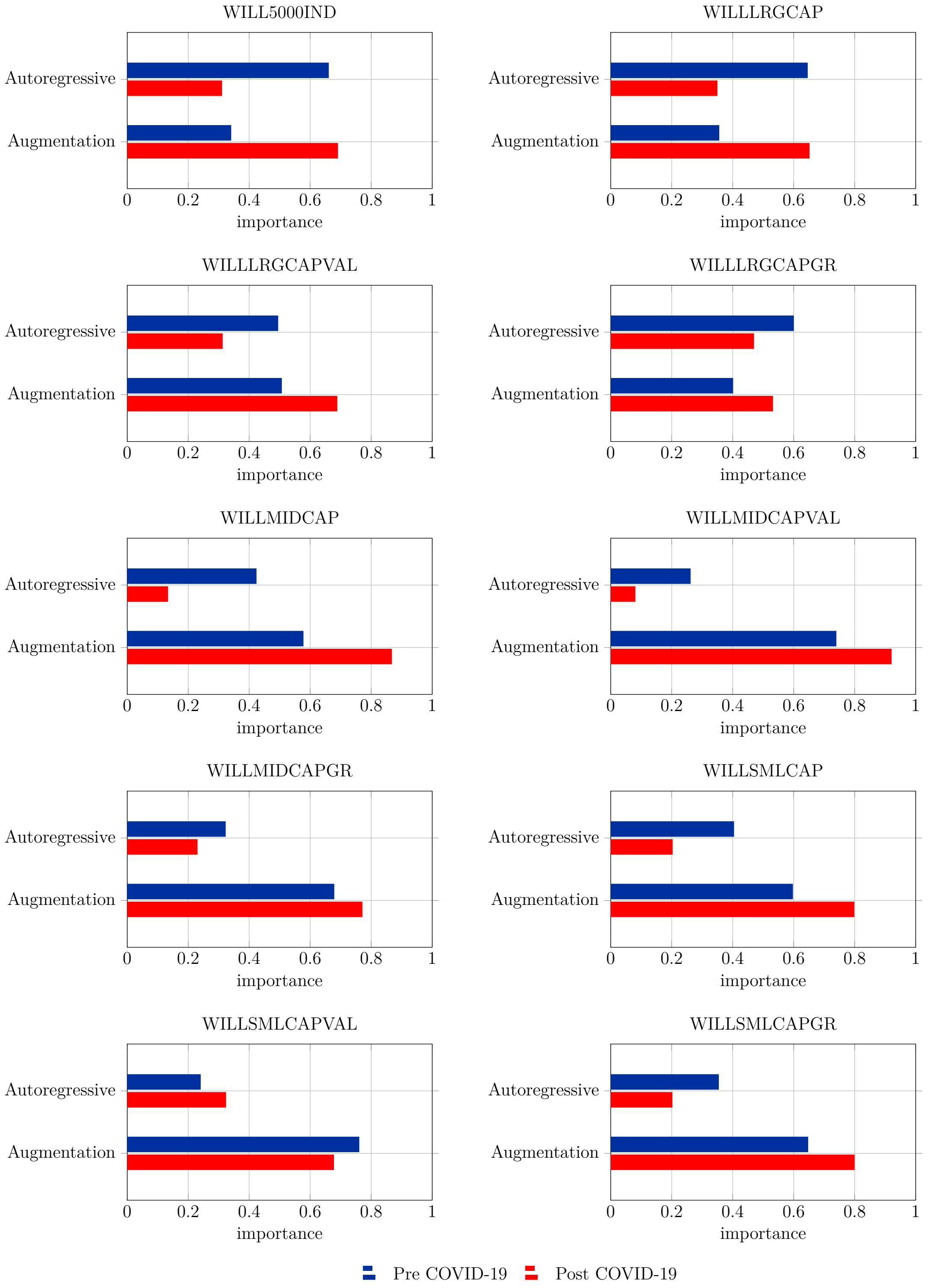


**Figure 2.3.2:** Importance weights pre and post COVID-19.
**Notes**: The "Autoregressive" and "Augmentation" bars reflect the cumulative weights for the lagged target and transformed business cycle. Pre COVID-19 weights are computed using the macroeconomic series available on the 28th February 2020 on ALFRED and the corresponding Wilshire data.

and Gilchrist (1994) and Bernanke et al. (1996) argue that small firms do not have a broad range of financing options and mostly use intermediaries to access credit. This leaves them more at risk during a downturn when banks become more selective with respect to credit extensions. Therefore, it is not surprising that the factor augmentation is especially crucial for the Wilshire indices referring to these market capitalisations. In addition, figure 2.3.2 highlights how the factor augmentation is even more relevant post COVID-19, a period of unprecedented high volatility and uncertainty. The additional charts in section 2.B highlight a high heterogeneity across targets.

## 2.4. Concluding remarks

This manuscript proposes a two-step method for handling predictors that exhibit measurement error, non-stationary trends, seasonality and/or irregularities such as missing observations within standard time-series regression trees. This approach can be intuitively thought as an automated feature engineering process that extracts a series of stationary and common patterns hidden in the predictors, while discarding troublesome characteristics. Given that this technique builds on a state-space model, the process can be easily enriched with domain-specific theory.

Section 2.3 shows promising results for empirical macro-finance problems based on bootstrap aggregating. Indeed, it proposes to use these ensembles for studying unclear non-linear links between the US business cycle and equity volatility within a forecasting setting. These factor-augmented ensembles outperform both naive benchmarks and standard bagging for all targets, both pre and post COVID-19. The models are further studied under the lenses of importance weights: an automated and internal ranking of the predictors. This shows that the augmentation is crucial and especially relevant for predicting volatility of mid to small cap equity indices. This is consistent with the literature that considers smaller companies as particularly vulnerable to negative changes in the business cycle due to their limited financing options.

Factor-augmented trees can be easily used for building more sophisticated ensembles or to study other problems: for instance to model the yield curve with a framework compatible with unspecified non-linearities.

# Appendix

## 2.A. Business cycle estimation

### 2.A.1. Trend-cycle model

Re-write the model in assumption 23 in the state-space form

$$\begin{aligned}\mathbf{Z}_t &= \mathbf{B}\boldsymbol{\Phi}_t + \boldsymbol{\xi}_t,\\ \boldsymbol{\Phi}_t &= \mathbf{C}\boldsymbol{\Phi}_{t-1} + \mathbf{D}\boldsymbol{\zeta}_t.\end{aligned}$$

The measurement coefficient matrix is sparse and the non-zero entries are such that

$$\mathbf{B} := \left( \underbrace{\begin{matrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{1}{\eta_8} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{1}{\eta_9} \end{matrix}}_{9\times 8} \middle| \underbrace{\begin{matrix} \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \end{matrix}}_{9\times 8} \middle| \underbrace{\begin{matrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{matrix}}_{9\times 9} \middle| \underbrace{\begin{matrix} 1 & \cdot & \dots & \cdot \\ \tilde{\Upsilon}_{1,1} & \tilde{\Upsilon}_{1,2} & \dots & \tilde{\Upsilon}_{1,p} \\ \tilde{\Upsilon}_{2,1} & \tilde{\Upsilon}_{2,2} & \dots & \tilde{\Upsilon}_{2,p} \\ \tilde{\Upsilon}_{3,1} & \tilde{\Upsilon}_{3,2} & \dots & \tilde{\Upsilon}_{3,p} \\ \tilde{\Upsilon}_{4,1} & \tilde{\Upsilon}_{4,2} & \dots & \tilde{\Upsilon}_{4,p} \\ \tilde{\Upsilon}_{5,1} & \tilde{\Upsilon}_{5,2} & \dots & \tilde{\Upsilon}_{5,p} \\ \tilde{\Upsilon}_{6,1} & \tilde{\Upsilon}_{6,2} & \dots & \tilde{\Upsilon}_{6,p} \\ \tilde{\Upsilon}_{7,1} & \tilde{\Upsilon}_{7,2} & \dots & \tilde{\Upsilon}_{7,p} \\ \tilde{\Upsilon}_{8,1} & \tilde{\Upsilon}_{8,2} & \dots & \tilde{\Upsilon}_{8,p} \end{matrix}}_{9\times p} \right),$$

where $\tilde{\boldsymbol{\Upsilon}}$ is a $8 \times p$ matrix of finite real parameters and $\tilde{\Upsilon}_{i,j} \approx \Upsilon_{i,j}$, for any $i = 1, \dots, 8$ and $j = 1, \dots, p$.[11] The transition matrices are also sparse and the non-zero entries are

[11] This numerical approximation is governed by $\varepsilon$ in assumption 23.

such that

$$\mathbf{C} := \begin{array}{c}\left(\begin{array}{ccc|ccc|ccc|ccccc}
1 & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \ddots & \cdot & \cdot & \ddots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \ddots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \pi_1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \ddots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \pi_n & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \pi_{n+1} & \pi_{n+2} & \dots & \pi_{n+p-1} & \pi_{n+p} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \dots & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \ddots & \vdots & \vdots \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \vdots & \ddots & \ddots & \vdots & \vdots \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \dots & 1 & \cdot
\end{array}\right) \\
\underbrace{\qquad\quad}_{q\times 8}\ \underbrace{\qquad\quad}_{q\times 8}\ \underbrace{\qquad\quad}_{q\times 9}\ \underbrace{\qquad\qquad\qquad\qquad}_{q\times p}
\end{array},$$

$$\mathbf{D} := \begin{array}{c}\left(\begin{array}{ccc|ccc|c}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \ddots & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \ddots & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right) \\
\underbrace{\qquad\quad}_{q\times 8}\ \underbrace{\qquad\quad}_{q\times 9}\ \underbrace{\ }_{q\times 1}
\end{array},$$

where $\boldsymbol{\pi}$ is a $n + p \times 1$ vector of finite real parameters and $q = 25 + p$. The innovation in the transition equation $\boldsymbol{\zeta}_t \overset{w.n}{\sim} N\left(\mathbf{0}_{r\times 1}, \boldsymbol{\Sigma}\right)$ with $\boldsymbol{\Sigma}$ being a $r \times r$ positive definite real diagonal matrix and $r = 18$. This representation implies that

$$\boldsymbol{\Phi}_t := \Big( \underbrace{\tau_{1,t} \quad \dots \quad \tau_{8,t}}_{8\times 1} \Big| \underbrace{\delta_{1,t} \quad \dots \quad \delta_{8,t}}_{8\times 1} \Big| \underbrace{\psi_{2,t} \quad \psi_{3,t} \quad \dots \quad \psi_{10,t}}_{9\times 1} \Big| \underbrace{\psi_{1,t} \quad \psi_{1,t-1} \quad \dots \quad \psi_{1,t-p+1}}_{p\times 1} \Big)'.$$

The initial conditions for the states are such that $\mathbf{\Phi}_0 \overset{w.n.}{\sim} N(\boldsymbol{\mu}_0, \mathbf{\Omega}_0)$, where $\boldsymbol{\mu}_0$ and $\mathbf{\Omega}_0$ denote a $q \times 1$ real vector and a $q \times q$ positive definite real covariance matrix. The matrix $\mathbf{\Omega}_0$ is sparse and the entries that can differ from zero are those with coordinates $(i, j) \in \{(i, j) : i = j \text{ and } 1 \leq i \leq 25\} \cup \{(i, j) : 25 < i \leq q \text{ and } 25 < j \leq q\}$.

**Remark.** The empirical application assumes that $\mathbf{\Sigma}$ is diagonal. This implies an *exact* factor model (i.e., no cross-sectional dependence in the idiosyncratic components) and, in doing so, it simplifies the narrative. However, this assumption may be too restrictive for more general problems. For similar problems, the assumptions could be relaxed and the ECM algorithm described in this appendix could be presented as a penalised quasi maximum likelihood estimation method building on the theoretical results in Barigozzi and Luciani (2020).

## 2.A.2. The Expectation-Conditional Maximisation algorithm

Denote the model free parameters with

$$\boldsymbol{\vartheta} := \begin{pmatrix} \boldsymbol{\mu}_0' & \text{vech}(\mathbf{\Omega}_0)' & \text{vec}(\tilde{\mathbf{\Upsilon}})' & \boldsymbol{\pi}' & \Sigma_{1,1} & \Sigma_{2,2} & \dots & \Sigma_{r,r} \end{pmatrix}'.$$

The ECM algorithm estimates these coefficients by repeating the optimisation process illustrated in definition 19 until convergence.

**Definition 19** (ECM estimation routine). At any $k+1 > 1$ iteration, the ECM algorithm computes the vector of coefficients

$$\hat{\boldsymbol{\vartheta}}_s^{k+1}(\gamma) := \underset{\underline{\boldsymbol{\vartheta}} \in \mathscr{R}}{\arg\max}\ \mathbb{E}\left[\mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Z}_{1:s}, \mathbf{\Phi}_{1:s}) \,|\, \mathscr{X}(s),\, \hat{\boldsymbol{\vartheta}}_s^{k}(\gamma)\right] - \mathbb{E}\left[\mathcal{P}(\underline{\boldsymbol{\vartheta}}, \gamma) \,|\, \mathscr{X}(s),\, \hat{\boldsymbol{\vartheta}}_s^{k}(\gamma)\right],$$

where $\mathscr{R}$ denotes the region in which the AR cycles (common and idiosyncratic) are causal, $\mathscr{X}(s)$ is the information set available at time $s$,

$$\begin{aligned} \mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Z}_{1:s}, \mathbf{\Phi}_{1:s}) \simeq & -\frac{1}{2} \ln |\underline{\mathbf{\Omega}_0}| - \frac{1}{2} \operatorname{Tr}\left[\underline{\mathbf{\Omega}_0}^{-1}(\mathbf{\Phi}_0 - \underline{\boldsymbol{\mu}_0})(\mathbf{\Phi}_0 - \underline{\boldsymbol{\mu}_0})'\right] \\ & -\frac{s}{2} \ln |\underline{\mathbf{\Sigma}}| - \frac{1}{2} \operatorname{Tr}\left[\sum_{t=1}^{s} \underline{\mathbf{\Sigma}}^{-1}(\mathbf{\Phi}_{*,t} - \underline{\mathbf{C}}_*\mathbf{\Phi}_{t-1})(\mathbf{\Phi}_{*,t} - \underline{\mathbf{C}}_*\mathbf{\Phi}_{t-1})'\right] \\ & -\frac{s}{2} \ln |\underline{\mathbf{R}}| - \frac{1}{2} \operatorname{Tr}\left[\sum_{t=1}^{s} \underline{\mathbf{R}}^{-1}(\mathbf{Z}_t - \underline{\mathbf{B}}\mathbf{\Phi}_t)(\mathbf{Z}_t - \underline{\mathbf{B}}\mathbf{\Phi}_t)'\right], \end{aligned} \tag{2.1}$$

$\mathbf{\Phi}_{*,t} \equiv \underline{\mathbf{D}}'\mathbf{\Phi}_t$, $\underline{\mathbf{C}}_* \equiv \underline{\mathbf{D}}'\underline{\mathbf{C}}$, and the underlined matrices denote the state-space coefficients implied by $\underline{\boldsymbol{\vartheta}}$. Besides,

$$\begin{aligned}\mathcal{P}(\underline{\boldsymbol{\vartheta}}, \gamma) := &+\frac{1-\alpha}{2}\left(\left\|\underline{\boldsymbol{\pi}}_{1:n}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},1)^{\frac{1}{2}}\right\|_{\mathrm{F}}^{2} + \left\|\underline{\boldsymbol{\pi}}'_{n+1:n+p}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},1)^{\frac{1}{2}}\right\|_{\mathrm{F}}^{2} + \left\|\underline{\tilde{\boldsymbol{\Upsilon}}}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},p)^{\frac{1}{2}}\right\|_{\mathrm{F}}^{2}\right)\\ &+\frac{\alpha}{2}\left(\left\|\underline{\boldsymbol{\pi}}_{1:n}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},1)\right\|_{1,1} + \left\|\underline{\boldsymbol{\pi}}'_{n+1:n+p}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},1)\right\|_{1,1} + \left\|\underline{\tilde{\boldsymbol{\Upsilon}}}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},p)\right\|_{1,1}\right)\end{aligned}$$

where, for any $l \in \mathbb{N}$,

$$\boldsymbol{\Gamma}(\boldsymbol{\gamma}, l) := \lambda \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \beta & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \beta^{l-1} \end{pmatrix},$$

$\lambda \geq 0$, $0 \leq \alpha \leq 1$ and $\beta \geq 1$ are hyperparameters included in $\boldsymbol{\gamma}$. The state-space coefficients for the first iteration are initialised as in section 2.A.3.

**Remark** (Objective functions). The function in equation 2.1 is the so-called complete-data (i.e., fully observed data and known latent states) log-likelihood, while $\mathcal{P}(\underline{\boldsymbol{\vartheta}}, \boldsymbol{\gamma})$ represents the generalised elastic-net penalty used in Pellegrino (2023a).

**Remark** (Underlined coefficients). Some of the underlined state-space coefficients are partially or fully fixed in accordance with the structure in section 2.A.1. For instance, $\mathbf{D} = \underline{\mathbf{D}}$ since $\mathbf{D}$ does not contain free parameters.

**Assumption 24** (Convergence). The ECM algorithm is said to be converged when the criteria in section 2.A.6 are met.

The optimisation in definition 19 is performed in two steps. The first one (E-step) involves the computation of the expectations in equation 2.1. The second step (CM-step) conditionally maximises the resulting expected penalised log-likelihood with respect to the free parameters.

It is convenient to write down the E-step on the basis of the output of a Kalman smoother compatible with incomplete time series, as in Shumway and Stoffer (1982) and Watson and Engle (1983). The required output is introduced in definition 20 and used in proposition 5 to compute the expected log-likelihood.

**Definition 20** (Kalman smoother output). The hereinbefore mentioned Kalman smoother output is

$$\begin{aligned}\hat{\boldsymbol{\Phi}}_t &:= \mathbb{E}\left[\boldsymbol{\Phi}_t \,|\, \mathscr{X}(s), \hat{\boldsymbol{\vartheta}}^{k}_{s}(\boldsymbol{\gamma})\right],\\ \hat{\mathbf{P}}_{t,t-j} &:= \mathrm{Cov}\left[\boldsymbol{\Phi}_t, \boldsymbol{\Phi}_{t-j} \,|\, \mathscr{X}(s), \hat{\boldsymbol{\vartheta}}^{k}_{s}(\boldsymbol{\gamma})\right],\end{aligned}$$

for any $k \geq 0$, $0 \leq j \leq t$ and $t \geq 0$. Let also $\hat{\mathbf{P}}_t \equiv \hat{\mathbf{P}}_{t,t}$.

**Remark.** These estimates are computed as in Pellegrino (2023a).

Furthermore, definition 21 is useful to formalise which measurements are observed at every single point in time.

**Definition 21** (Observed measurements). Let

$$\begin{aligned} \mathscr{T} &:= \bigcup_{i=1}^{n} \mathscr{T}_i, \\ \mathscr{T}(s) &:= \{t : t \in \mathscr{T},\, 1 \leq t \leq s\}, \end{aligned}$$

describe two sets representing the points in time (either over the full sample or up to time $s$) in which we observe at least one measurement, for $1 \leq s \leq T$. Let also

$$\mathscr{V}_t := \{i : t \in \mathscr{T}_i,\, 1 \leq i \leq n\},$$

for $1 \leq t \leq T$. Thus, let

$$\begin{aligned} \mathbf{Z}_t^{obs} &:= \left(Z_{i,t}\right)_{i \in \mathscr{V}_t} \\ \mathbf{B}_t^{obs} &:= \mathbf{A}_t \mathbf{B} \end{aligned}$$

be the vector of observed measurements at time $t$ and the corresponding $|\mathscr{V}_t| \times q$ matrix of coefficients, for any $t \in \mathscr{T}$. Every $\mathbf{A}_t$ is indeed a selection matrix constituted by ones and zeros that permits to retrieve the appropriate rows of $\mathbf{B}$ for every $t \in \mathscr{T}$.

**Proposition 5.** *Let*

$$\mathcal{L}_e \left[ \underline{\boldsymbol{\vartheta}} \,|\, \mathscr{Z}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma}) \right] \equiv \mathbb{E} \left[ \mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Z}_{1:s}, \boldsymbol{\Phi}_{1:s}) \,|\, \mathscr{Z}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma}) \right].$$

*Building on definition 20, it follows that*

$$\begin{aligned} \mathcal{L}_e \left[ \underline{\boldsymbol{\vartheta}} \,|\, \mathscr{Z}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma}) \right] \simeq &- \frac{1}{2} \ln |\underline{\boldsymbol{\Omega}_0}| - \frac{1}{2} \operatorname{Tr} \left[ \underline{\boldsymbol{\Omega}_0}^{-1} (\hat{\mathbf{E}} - \hat{\boldsymbol{\Phi}}_0 \underline{\boldsymbol{\mu}_0}' - \underline{\boldsymbol{\mu}_0} \hat{\boldsymbol{\Phi}}_0' + \underline{\boldsymbol{\mu}_0}\, \underline{\boldsymbol{\mu}_0}') \right] \\ &- \frac{s}{2} \ln |\underline{\boldsymbol{\Sigma}}| - \frac{1}{2} \operatorname{Tr} \left[ \underline{\boldsymbol{\Sigma}}^{-1} (\hat{\mathbf{F}}_s - \hat{\mathbf{G}}_s \underline{\mathbf{C}}'_* - \underline{\mathbf{C}}_* \hat{\mathbf{G}}_s' + \underline{\mathbf{C}}_* \hat{\mathbf{H}}_s \underline{\mathbf{C}}'_*) \right] \\ &- \frac{1}{2\varepsilon} \operatorname{Tr} \left\{ \sum_{t \in \mathscr{T}(s)} \left[ \left( \mathbf{Z}_t^{obs} - \underline{\mathbf{B}}_t^{obs} \hat{\boldsymbol{\Phi}}_t \right) \left( \mathbf{Z}_t^{obs} - \underline{\mathbf{B}}_t^{obs} \hat{\boldsymbol{\Phi}}_t \right)' + \underline{\mathbf{B}}_t^{obs} \hat{\mathbf{P}}_t \underline{\mathbf{B}}_t^{obs'} \right] \right\}, \end{aligned}$$

*where*

$$\begin{aligned} \hat{\mathbf{E}} &:= \mathbb{E} \Big[ \boldsymbol{\Phi}_0 \boldsymbol{\Phi}_0' \,|\, \mathscr{Z}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma}) \Big] = \hat{\boldsymbol{\Phi}}_0 \hat{\boldsymbol{\Phi}}_0' + \hat{\mathbf{P}}_0, \\ \hat{\mathbf{F}}_s &:= \sum_{t=1}^{s} \mathbb{E} \Big[ \boldsymbol{\Phi}_{*,t} \boldsymbol{\Phi}'_{*,t} \,|\, \mathscr{Z}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma}) \Big] = \sum_{t=1}^{s} \underline{\mathbf{D}}' \left( \hat{\boldsymbol{\Phi}}_t \hat{\boldsymbol{\Phi}}_t' + \hat{\mathbf{P}}_t \right) \underline{\mathbf{D}}, \end{aligned}$$

$$\begin{aligned}\hat{\mathbf{G}}_s &:= \sum_{t=1}^{s} \mathbb{E}\Big[\mathbf{\Phi}_{*,t}\mathbf{\Phi}'_{t-1} \,|\, \mathscr{X}(s), \hat{\boldsymbol{\vartheta}}^k_s(\gamma)\Big] = \sum_{t=1}^{s} \underline{\mathbf{D}}' \left(\hat{\mathbf{\Phi}}_t\hat{\mathbf{\Phi}}'_{t-1} + \hat{\mathbf{P}}_{t,t-1}\right), \\ \hat{\mathbf{H}}_s &:= \sum_{t=1}^{s} \mathbb{E}\Big[\mathbf{\Phi}_{t-1}\mathbf{\Phi}'_{t-1} \,|\, \mathscr{X}(s), \hat{\boldsymbol{\vartheta}}^k_s(\gamma)\Big] = \sum_{t=1}^{s} \left(\hat{\mathbf{\Phi}}_{t-1}\hat{\mathbf{\Phi}}'_{t-1} + \hat{\mathbf{P}}_{t-1}\right).\end{aligned}$$

PROOF. The proof is analogous to the one in Pellegrino (2023a, Proposition 4). □

**Remark.** $\mathbf{D} = \underline{\mathbf{D}}$ plays the role of a selection matrix. In particular premultiplying by $\underline{\mathbf{D}}'$ allows to select rows and postmultiplying by $\underline{\mathbf{D}}$ columns.

**Lemma 12.** *The conditional expectation for the penalty in definition 19 is*

$$\mathbb{E}\left[\mathcal{P}(\underline{\boldsymbol{\vartheta}}, \gamma) \,|\, \mathscr{X}(s), \hat{\boldsymbol{\vartheta}}^k_s(\gamma)\right] = \mathcal{P}(\underline{\boldsymbol{\vartheta}}, \gamma).$$

PROOF. A formal proof is not reported since it is immediate. Indeed, the penalty function in this ECM algorithm depends only on the current vector of coefficients and hyperparameters. □

The CM-step conditionally maximises the expected penalised log-likelihood

$$\mathcal{M}_e\left[\underline{\boldsymbol{\vartheta}}, \gamma \,|\, \mathscr{X}(s), \hat{\boldsymbol{\vartheta}}^k_s(\gamma)\right] := \mathcal{L}_e\left[\underline{\boldsymbol{\vartheta}} \,|\, \mathscr{X}(s), \hat{\boldsymbol{\vartheta}}^k_s(\gamma)\right] - \mathcal{P}(\underline{\boldsymbol{\vartheta}}, \gamma) \tag{2.2}$$

to estimate the state-space parameters. The estimated coefficients are denoted with an "hat" symbol. Besides, an $s$ subscript is used for highlighting the sample size and a superscript for denoting the ECM iteration.

**Lemma 13.** *The ECM estimator at a generic iteration $k+1 > 0$ for $\boldsymbol{\mu}_0$ is*

$$\hat{\boldsymbol{\mu}}^{k+1}_{0,s}(\gamma) = \hat{\mathbf{\Phi}}_0$$

*and the estimator for $\mathbf{\Omega}_0$ is a sparse covariance matrix whose non-zero entries are*

$$\left[\hat{\mathbf{\Omega}}^{k+1}_{0,s}(\gamma)\right]_{i,j} = \left[\hat{\mathbf{P}}_0\right]_{i,j},$$

*for $(i,j) \in \{(i,j) : i = j$ and $1 \leq i \leq 25\} \cup \{(i,j) : 25 < i \leq q$ and $25 < j \leq q\}$.*

PROOF. The derivative of equation 2.2 with respect to $\underline{\boldsymbol{\mu}_0}$ is

$$\frac{\partial \mathcal{M}_e\left[\underline{\boldsymbol{\vartheta}}, \gamma \,|\, \mathscr{X}(s), \hat{\boldsymbol{\vartheta}}^k_s(\gamma)\right]}{\partial \underline{\boldsymbol{\mu}_0}} = -\frac{1}{2}\underline{\mathbf{\Omega}_0}^{-1}\left(-2\hat{\mathbf{\Phi}}_0 + 2\underline{\boldsymbol{\mu}_0}\right).$$

It follows that the maximiser for the expected penalised log-likelihood is

$$\hat{\boldsymbol{\mu}}^{k+1}_{0,s}(\gamma) = \hat{\mathbf{\Phi}}_0.$$

The derivative of equation 2.2 with respect to $\underline{\mathbf{\Omega}_0}$ and fixing $\underline{\boldsymbol{\mu}_0} = \hat{\boldsymbol{\mu}}_{0,s}^{k+1}(\boldsymbol{\gamma})$ is

$$-\frac{1}{2}\underline{\mathbf{\Omega}_0}^{-1} + \frac{1}{2}\underline{\mathbf{\Omega}_0}^{-1}\left[\hat{\mathbf{E}} - \hat{\mathbf{\Phi}}_0\hat{\mathbf{\Phi}}_0'\right]\underline{\mathbf{\Omega}_0}^{-1} = -\frac{1}{2}\underline{\mathbf{\Omega}_0}^{-1} + \frac{1}{2}\underline{\mathbf{\Omega}_0}^{-1}\hat{\mathbf{P}}_0\,\underline{\mathbf{\Omega}_0}^{-1}.$$

as also shown in Pellegrino (2023a). Given the assumptions in section 2.A.1 on the structure of $\mathbf{\Omega}_0$, it follows that $\hat{\mathbf{\Omega}}_{0,s}^{k+1}(\boldsymbol{\gamma})$ is a sparse matrix whose non-zero entries are

$$\left[\hat{\mathbf{\Omega}}_{0,s}^{k+1}(\boldsymbol{\gamma})\right]_{i,j} = \left[\hat{\mathbf{P}}_0\right]_{i,j},$$

for $(i,j) \in \{(i,j) : i = j \text{ and } 1 \leq i \leq 25\} \cup \{(i,j) : 25 < i \leq q \text{ and } 25 < j \leq q\}$. □

**Definition 22.** Let $\tilde{\mathbf{\Gamma}}(\boldsymbol{\gamma})$ be a diagonal $q \times q$ matrix whose non-zero entries are such that

$$\tilde{\mathbf{\Gamma}}(\boldsymbol{\gamma}) := \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \lambda\,\mathbf{I}_9 & \cdot \\ \cdot & \cdot & \mathbf{\Gamma}(\boldsymbol{\gamma}, p) \end{pmatrix}.$$

**Lemma 14.** *The ECM estimator at a generic iteration $k+1 > 0$ for $\mathbf{C}$ is such that*

$$\hat{C}_{i,j,s}^{k+1}(\boldsymbol{\gamma}) = \frac{\mathcal{S}\left[\hat{\Sigma}_{i,i,s}^{k-1}(\boldsymbol{\gamma})\Big(\hat{G}_{i,j,s} - \sum_{l=1,\,l\neq j}^{q}\hat{C}_{i,l,s}^{k+\mathbb{I}_{l<j}}(\boldsymbol{\gamma})\,\hat{H}_{l,j,s}\Big),\ \frac{\alpha}{2}\,\tilde{\Gamma}_{j,j}\,(\boldsymbol{\gamma})\right]}{\hat{\Sigma}_{i,i,s}^{k-1}(\boldsymbol{\gamma})\,\hat{H}_{j,j,s} + (1-\alpha)\,\tilde{\Gamma}_{j,j}\,(\boldsymbol{\gamma})},$$

*for any $(i,j) \in \{(i,j) : i = j$ and $17 \leq i \leq 25\} \cup \{(i,j) : i = 26$ and $26 \leq j \leq q\}$, and constant to the values in section 2.A.1 for the remaining entries.*

PROOF. Given that the absolute value function in the penalty is not differentiable at zero, this part of the ECM algorithm estimates, in turn, the free entries of $\mathbf{C}$ (i.e., $\pi_1, \ldots, \pi_{n+p}$) while fixing $\underline{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_s^k(\boldsymbol{\gamma})$ and any other free entry of $\mathbf{C}$ to their latest estimate. For any $\underline{C}_{i,j} \neq 0$ corresponding to a free parameter, the derivative of equation 2.2 with respect to $\underline{C}_{i,j}$ having fixed the coefficients as described in the previous sentence is

$$+\,\hat{\Sigma}_{i,i,s}^{k-1}(\boldsymbol{\gamma})\left(\hat{G}_{i,j,s} - \underline{C}_{i,j}\hat{H}_{j,j,s} - \sum_{\substack{l=1\\ l\neq j}}^{q}\hat{C}_{i,l,s}^{k+\mathbb{I}_{l<j}}\,(\boldsymbol{\gamma})\,\hat{H}_{l,j,s}\right) - (1-\alpha)\,\tilde{\Gamma}_{j,j}\,(\boldsymbol{\gamma})\,\underline{C}_{i,j} - \frac{\alpha}{2}\,\tilde{\Gamma}_{j,j}\,(\boldsymbol{\gamma})\,\mathrm{sign}(\underline{C}_{i,j}),$$

since $\hat{\mathbf{\Sigma}}_s^k(\boldsymbol{\gamma})$ is diagonal. It follows that

$$\hat{C}_{i,j,s}^{k+1}(\boldsymbol{\gamma}) = \frac{\mathcal{S}\left[\hat{\Sigma}_{i,i,s}^{k-1}(\boldsymbol{\gamma})\Big(\hat{G}_{i,j,s} - \sum_{l=1,\,l\neq j}^{q}\hat{C}_{i,l,s}^{k+\mathbb{I}_{l<j}}(\boldsymbol{\gamma})\,\hat{H}_{l,j,s}\Big),\ \frac{\alpha}{2}\,\tilde{\Gamma}_{j,j}\,(\boldsymbol{\gamma})\right]}{\hat{\Sigma}_{i,i,s}^{k-1}(\boldsymbol{\gamma})\,\hat{H}_{j,j,s} + (1-\alpha)\,\tilde{\Gamma}_{j,j}\,(\boldsymbol{\gamma})},$$

for any $(i,j) \in \{(i,j) : i = j \text{ and } 17 \leq i \leq 25\} \cup \{(i,j) : i = 26 \text{ and } 26 \leq j \leq q\}$, and constant to the values in section 2.A.1 for the remaining entries. □

**Lemma 15.** *The ECM estimator at a generic iteration $k+1>0$ for $\boldsymbol{\Sigma}$ is such that*

$$\hat{\Sigma}^{k+1}_{i,i,s}(\boldsymbol{\gamma}) = \frac{1}{s}\left[\hat{\mathbf{F}}_s - \hat{\mathbf{G}}_s\hat{\mathbf{C}}^{k+1'}_s(\boldsymbol{\gamma}) - \hat{\mathbf{C}}^{k+1}_s(\boldsymbol{\gamma})\,\hat{\mathbf{G}}'_s + \hat{\mathbf{C}}^{k+1}_s(\boldsymbol{\gamma})\,\hat{\mathbf{H}}_s\hat{\mathbf{C}}^{k+1'}_s(\boldsymbol{\gamma})\right]_{i,i}$$

*for $i=1,\dots,r$ and zero for the remaining entries.*

PROOF. The proof is equivalent to the one reported in Pellegrino (2023a, Lemma 10). However, in this manuscript, $\hat{\boldsymbol{\Sigma}}^{k+1}_s(\boldsymbol{\gamma})$ is diagonal as indicated in section 2.A.1. □

**Lemma 16.** *Let*

$$\begin{aligned}
\hat{\mathbf{M}}_s &:= \sum_{t\in\mathscr{T}(s)} \mathbf{A}'_t\mathbf{Z}^{obs}_t\,\hat{\boldsymbol{\Phi}}'_t, \\
\hat{\mathbf{N}}_t &:= \mathbf{A}'_t\mathbf{A}_t, \\
\hat{\mathbf{O}}_t &:= \hat{\boldsymbol{\Phi}}_t\hat{\boldsymbol{\Phi}}'_t + \hat{\mathbf{P}}_t.
\end{aligned}$$

*The ECM estimator at a generic iteration $k+1>0$ for $\mathbf{B}$ is such that*

$$\hat{B}^{k+1}_{i,j,s}(\boldsymbol{\gamma}) = \frac{\mathcal{S}\left[\hat{M}_{i,j,s} - \sum_{t\in\mathscr{T}(s)}\hat{N}_{i,i,t}\sum_{l=1,l\neq j}^{q}\hat{B}^{k+\mathbb{I}_{l<j}}_{i,l,s}(\boldsymbol{\gamma})\,\hat{O}_{l,j,t},\ \frac{\alpha}{2}\,\varepsilon\,\tilde{\Gamma}_{j,j}\left(\boldsymbol{\gamma}\right)\right]}{\sum_{t\in\mathscr{T}(s)}\hat{N}_{i,i,t}\hat{O}_{j,j,t} + (1-\alpha)\,\varepsilon\,\tilde{\Gamma}_{j,j}\left(\boldsymbol{\gamma}\right)},$$

*for any $(i,j)\in\{(i,j) : 2\leq i\leq n \text{ and } 26\leq j\leq q\}$, and constant to the values in section 2.A.1 for the remaining entries.*

PROOF. Note that

$$\begin{aligned}
&\sum_{t\in\mathscr{T}(s)}\left[\left(\mathbf{Z}^{obs}_t - \underline{\mathbf{B}}^{obs}_t\,\hat{\boldsymbol{\Phi}}_t\right)\left(\mathbf{Z}^{obs}_t - \underline{\mathbf{B}}^{obs}_t\,\hat{\boldsymbol{\Phi}}_t\right)' + \underline{\mathbf{B}}^{obs}_t\,\hat{\mathbf{P}}_t\,\underline{\mathbf{B}}^{obs'}_t\right] \\
&\quad= \sum_{t\in\mathscr{T}(s)}\left[\left(\mathbf{Z}^{obs}_t - \mathbf{A}_t\underline{\mathbf{B}}\hat{\boldsymbol{\Phi}}_t\right)\left(\mathbf{Z}^{obs}_t - \mathbf{A}_t\underline{\mathbf{B}}\hat{\boldsymbol{\Phi}}_t\right)' + \mathbf{A}_t\underline{\mathbf{B}}\hat{\mathbf{P}}_t\underline{\mathbf{B}}'\mathbf{A}'_t\right] \\
&\quad= \sum_{t\in\mathscr{T}(s)}\left[\mathbf{Z}^{obs}_t\,\mathbf{Z}^{obs'}_t - \mathbf{Z}^{obs}_t\,\hat{\boldsymbol{\Phi}}'_t\underline{\mathbf{B}}'\mathbf{A}'_t - \mathbf{A}_t\underline{\mathbf{B}}\hat{\boldsymbol{\Phi}}_t\mathbf{Z}^{obs'}_t + \mathbf{A}_t\underline{\mathbf{B}}\left(\hat{\boldsymbol{\Phi}}_t\hat{\boldsymbol{\Phi}}'_t + \hat{\mathbf{P}}_t\right)\underline{\mathbf{B}}'\mathbf{A}'_t\right].
\end{aligned}$$

Note also that all $\hat{\mathbf{N}}_t$ are diagonal. Indeed, at any point in time $t$ when all series are observed $\mathbf{A}_t = \hat{\mathbf{N}}_t = \mathbf{I}_n$. Besides, at any other $t\in\mathscr{T}(s)$,

$$\hat{N}_{i,i,t} = \begin{cases} 1 & \text{if the } i\text{-th series is observed at time } t, \\ 0 & \text{otherwise,} \end{cases}$$

for $i=1,\dots,n$. Given that the absolute value function in the penalty is not differentiable at zero, this part of the ECM algorithm estimates, in turn, the free entries of $\mathbf{B}$ (i.e., $\tilde{\Upsilon}_{1,1},\dots,\tilde{\Upsilon}_{1,p},\dots,\tilde{\Upsilon}_{8,p}$) while fixing any other free entry of $\mathbf{B}$ to their latest estimate.

For any $\underline{B}_{i,j} \neq 0$ corresponding to a free parameter, the derivative of equation 2.2 with respect to $\underline{B}_{i,j}$ having fixed the coefficients as described in the previous sentence is

$$+\,\varepsilon^{-1}\left(\hat{M}_{i,j,s} - \sum_{t\in\mathscr{T}(s)} \hat{N}_{i,i,t} \sum_{\substack{l=1\\ l\neq j}}^{q} \hat{B}_{i,l,s}^{k+\mathbb{I}_{l<j}}(\boldsymbol{\gamma})\,\hat{O}_{l,j,t}\right) - \underline{B}_{i,j}\left(\varepsilon^{-1}\sum_{t\in\mathscr{T}(s)} \hat{N}_{i,i,t}\hat{O}_{j,j,t} + (1-\alpha)\,\tilde{\Gamma}_{j,j}(\boldsymbol{\gamma})\right)$$
$$-\,\frac{\alpha}{2}\,\tilde{\Gamma}_{j,j}(\boldsymbol{\gamma})\,\mathrm{sign}(\underline{B}_{i,j}),$$

since all $\hat{\mathbf{N}}_t$ are diagonal. It follows that

$$\hat{B}_{i,j,s}^{k+1}(\boldsymbol{\gamma}) = \frac{\mathcal{S}\left[\hat{M}_{i,j,s} - \sum_{t\in\mathscr{T}(s)} \hat{N}_{i,i,t} \sum_{l=1,l\neq j}^{q} \hat{B}_{i,l,s}^{k+\mathbb{I}_{l<j}}(\boldsymbol{\gamma})\,\hat{O}_{l,j,t},\ \frac{\alpha}{2}\,\varepsilon\,\tilde{\Gamma}_{j,j}\left(\boldsymbol{\gamma}\right)\right]}{\sum_{t\in\mathscr{T}(s)} \hat{N}_{i,i,t}\hat{O}_{j,j,t} + (1-\alpha)\,\varepsilon\,\tilde{\Gamma}_{j,j}\left(\boldsymbol{\gamma}\right)},$$

for any $(i,j) \in \{(i,j) : 2 \leq i \leq n \text{ and } 26 \leq j \leq q\}$, and constant to the values in section 2.A.1 for the remaining entries. □

## 2.A.3. Initialisation of the Expectation-Maximisation algorithm

The first step in the initialisation involves computing a first approximation for the trends. This is achieved via univariate trend-cycle decompositions. In the case of headline and core inflation, the initialisation of the trend involves a further operation. Trend inflation is initialised by taking the mean between the persistent components estimated for headline and core inflation, appropriately rescaled by $\eta_8$ and $\eta_9$. The variances of the innovations are calculated on the double differenced initial trends.

The second step involves the initialisation of the cycles, which is performed on the de-trended data. The business cycle is approximated by the first principal component of the de-trended data and a series of ridge regressions is used for computing the coefficients of the cycles. The restrictions described in section 2.A.1 are enforced on each regression. The variances of the innovations are computed on the sample residuals.

## 2.A.4. Enforcing causality during the estimation

The ECM algorithm used in this manuscript ensures that the AR states (i.e., the common cycle and idiosyncratic noise components) are causal at every iteration. This is achieved with the approach proposed in Pellegrino (2023a, Section C.4) for vector autoregressions.

## 2.A.5. Hyperparameter selection

The hyperparameters are selected using the artificial jackknife selection method proposed in Pellegrino (2023a). In the empirical application in section 2.3, the grid of candidate hyperparameters $\mathscr{H} = \mathscr{H}_p \times \mathscr{H}_\lambda \times \mathscr{H}_\alpha \times \mathscr{H}_\beta$ is such that $\mathscr{H}_p := \{12\}$, $\mathscr{H}_\lambda := [10^{-2}, 2.5]$, $\mathscr{H}_\alpha := [0, 1]$ and $\mathscr{H}_\beta := [1, 1.2]$. The selection process returns the specification with the lowest expected forecast error for headline inflation, following a rational similar to Jarocinski and Lenza (2015).[12]

## 2.A.6. Estimation algorithm

**Algorithm 2:** ECM algorithm for the trend-cycle decomposition

Initialization

The ECM algorithm is initialised as described in section 2.A.3.

Estimation

```
for $k \leftarrow 1$ *to max_iter* do
    for $j \leftarrow 1$ *to m* do
        Run the Kalman filter and smoother using $\hat{\boldsymbol{\vartheta}}_s^{k-1}(\boldsymbol{\gamma})$;
        if *converged* then
            Store the parameters and stop the loop.
        end
        Estimate $\hat{\boldsymbol{\mu}}_{s,0}^k(\boldsymbol{\gamma})$ and $\hat{\boldsymbol{\Omega}}_{s,0}^k(\boldsymbol{\gamma})$ as in lemma 13;
        Estimate $\hat{\mathbf{C}}_s^k(\boldsymbol{\gamma})$, $\hat{\boldsymbol{\Sigma}}_s^k(\boldsymbol{\gamma})$ and $\hat{\mathbf{B}}_s^k(\boldsymbol{\gamma})$ as in lemmas 14–16;
        Build $\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})$;
    end
end
```

**Notes**

- The results are computed fixing *max_iter* to 1000. This is a conservative number, since the algorithm generally requires substantially less iterations to converge.
- The ECM algorithm is considered to be converged when the estimated coefficients (all relevant parameters in lemmas 14–16) do not significantly change in two subsequent iterations. This is done by computing the absolute relative change per parameters and comparing at the same time the median and $95^{th}$ quantile respectively with a fixed tolerance of $10^{-3}$ and $10^{-2}$. Intuitively, when the coefficients do not change much, the expected log-likelihood and the parameters in lemma 13 should also be stable.
- The scalar $\varepsilon$ is summed to the denominator of each relative change in order to ensure numerical stability.

The replication code for this paper is available on GitHub.

[12] The weights in Pellegrino (2023a) are set to be equal to zero for all variables with the exception of headline inflation which has weight equal to one.

## 2.B. Additional charts and tables

| Acronym | Description |
|---|---|
| BEA | Bureau of Economic Analysis |
| BLS | Bureau of Labor Statistics |
| CPI | Consumer Price Index |
| FRB | Federal Reserve Board |
| FRBSL | Federal Reserve Bank of St. Louis |
| TMI | Total Market Index |
| WA | Wilshire Associates |

**Table 2.B.1:** Glossary for the acronyms in table 2.3.1.

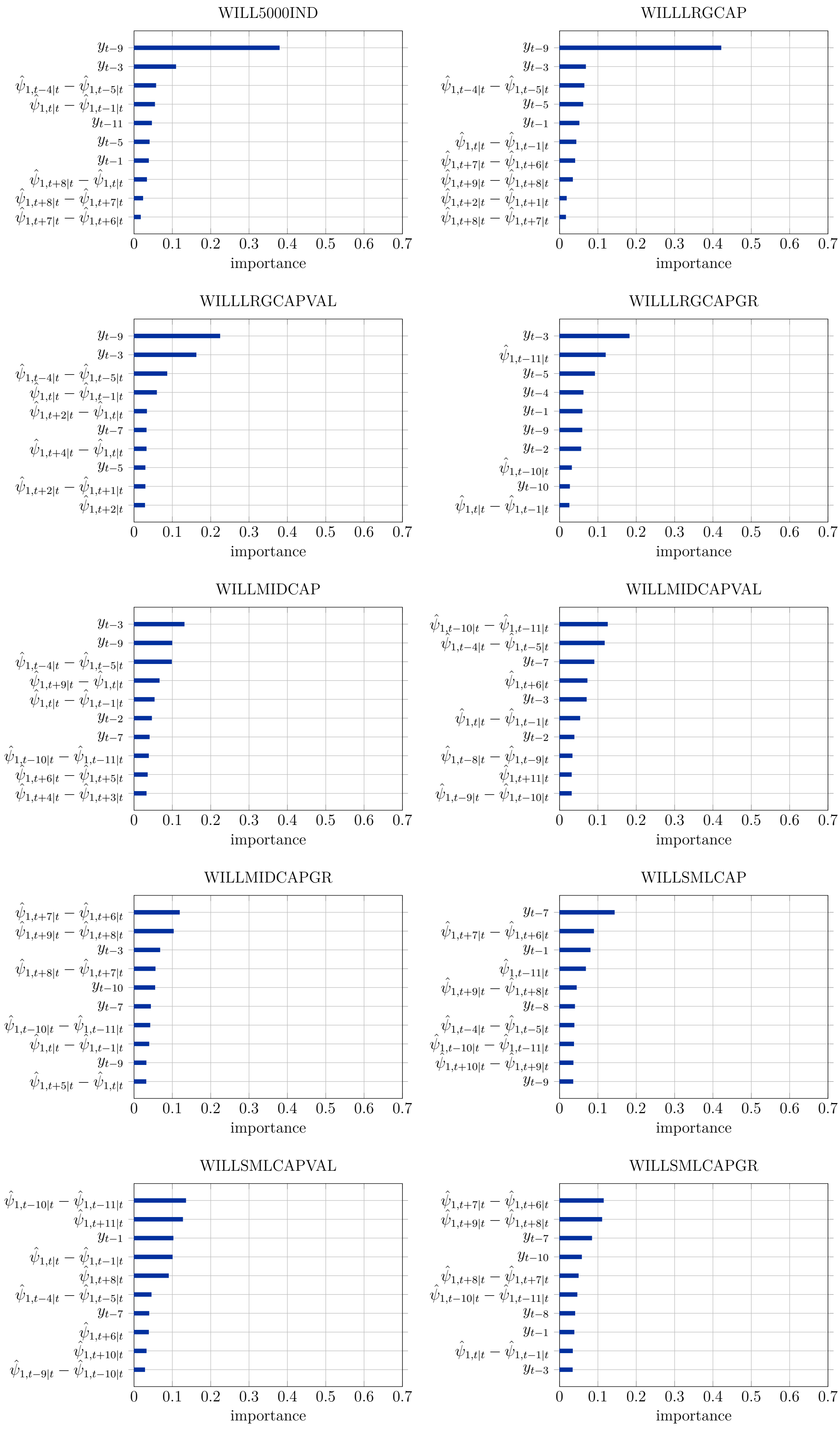


**Figure 2.B.1:** Importance weights pre COVID-19: top 10 predictors.
**Notes**: Pre COVID-19 weights are computed using the macroeconomic series available on the 28th February 2020 on ALFRED and the corresponding Wilshire data.

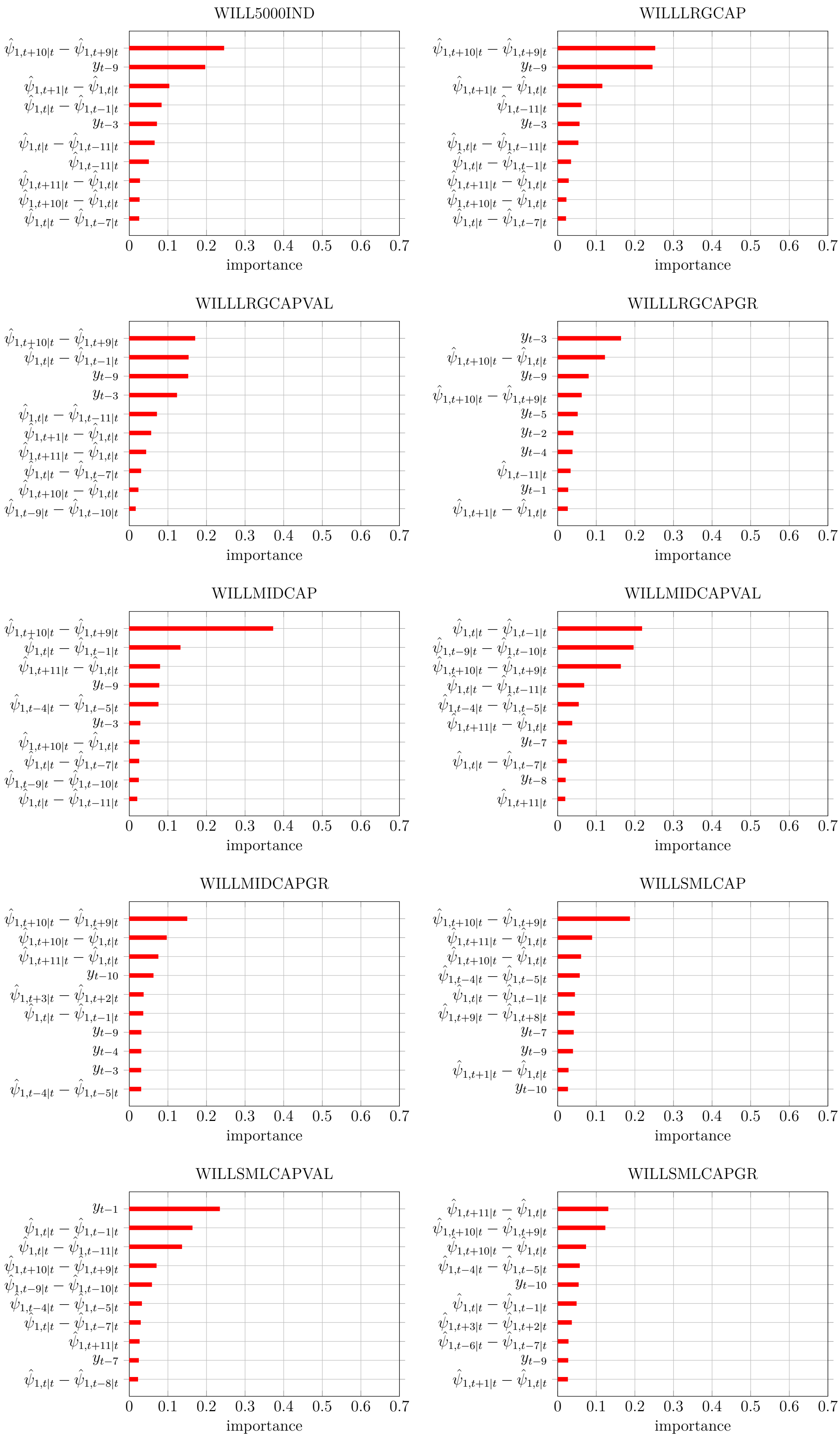


**Figure 2.B.2:** Importance weights post COVID-19: top 10 predictors.

# 3 Multidimensional dynamic factor models

*This paper generalises dynamic factor models for multidimensional dependent data. In doing so, it develops an interpretable technique to study complex information sources ranging from repeated surveys with a varying number of respondents to panels of satellite images. We specialise our results to model microeconomic data on US households jointly with macroeconomic aggregates. This results in a powerful tool able to generate localised predictions, counterfactuals and impulse response functions for individual households, accounting for traditional time-series complexities depicted in the state-space literature. The model is also compatible with the growing focus of policymakers for real-time economic analysis as it is able to process observations online, while handling missing values and asynchronous data releases.*

## 3.1. Introduction

Nowadays, it is easy to find datasets with millions of observations and measurements taken over a broad range of time periods. However, complexity increases with the number of dimensions considered per period and, thus, not all datasets are created equal.

Tabular datasets are often easier to model than more abstract cases, including time series of satellite images and texts. This reduced complexity inspired the development of interpretable models with straightforward policy applications. For instance, tabular multivariate time series are commonly studied via impulse response functions and conditional forecasts to determine appropriate fiscal and monetary policy actions. Unstructured datasets have been mostly studied through less explainable models and, as a result, they are not as used for policy. This is a pity, given that they could be handy for a broad range of applications, including studying poverty through satellite images and predicting volatility from market-risk reports, as surveyed in Mullainathan and Spiess (2017). Agreeing with similar considerations, we propose a framework compatible with multidimensional dependent data that retains the explainability of traditional statistical models.

Our approach suggests to reshape multidimensional data into a tabular multivariate time series with a peculiar vectorisation that accounts for temporal variations in the sample size and composition. Once the transformation is completed, we propose to model the resulting data on the basis of state-space methods (e.g., Harvey, 1990) and reduce the magnitude of the problem by extracting unobserved common components across multiple dimensions and time. In doing so, we manage to obtain an explainable technique flexible enough for handling complex datasets and capable of being linked with domain-specific concepts through identification schemes such as those in Bai and Wang (2015). This dimensionality reduction technique can be interpreted as a generalisation of the one employed by dynamic factor models. As such, the origins of our methodology are rooted in psychometrics (Lawley and Maxwell, 1962) and time-series econometrics (Geweke, 1977; Forni et al., 2000, 2005, 2009; Forni and Lippi, 2001; Bernanke et al., 2005; Doz et al., 2012; Barigozzi and Luciani, 2020). In light of that, we call it multidimensional dynamic factor model.

We specialise our manuscript for analysing microeconomic data on households and macroeconomic time series jointly. This problem is indeed multidimensional since, at each point in time $t$, we observe a survey containing $N_t$ households with $K$ characteristics of interest. Our modelling choice is compatible with economic theory and flexible enough to describe the characteristics of different groups of households, thus helping measuring income inequality.

Our use of repeated microeconomic surveys is different from traditional approaches: we neither pretreat the time-series cross-sectional data by transforming it into aggregate indices, nor model it via cross-sectional regressions with a linear trend predictor. Instead, it shares similarities with the approach in Liu and Plagborg-Møller (2021). We both use macroeconomic aggregates, households data and state-space modelling. However, we model everything in one step and within a single state space, while they use a two-step method in which the latent components are extracted from macroeconomic aggregates only. This allows us to have a system able to handle microeconomic complexities such as the temporal dynamics of each household. As a result, we make a better use of the data and model serial correlation in household income across groups of demographics. Besides, using cyclical and non-stationary latent components we distinguish between transitory and persistent determinants of household income. We are not aware of other papers handling similar complexities at once and refer to the introduction of Liu and Plagborg-Møller (2021) for an in-depth survey of correlated articles.

Our empirical analysis is based on a large dataset containing macroeconomic aggregates from the Archival Federal Reserve Economic Data (ALFRED) and households information collected in the Consumer Expenditure (CE) Public Use Microdata (PUMD). Our empirical results highlight differences in household income among distinct demographics. In particular, we find that our MDFM is able to capture persistent parts of income linked with education and ethnicity. Besides, we show that our model is able to track the demographics surveyed in the CE PUMD before its official publication date, thus extending the findings in Giannone et al. (2008) and the scope of nowcasting to microeconomic problems.

# 3.2. Methodology

## 3.2.1. Data processing

This subsection illustrates our approach to process multidimensional multivariate dependent data.

**Assumption 25** (Data). Let $\mathbf{H}_t \in \mathbb{R}^{N_t \times K}$ be a data matrix with $N_t > 0$ and $K > 0$, and denote with $\mathbf{h}_t$ the $N_t K \times 1$ vectorisation of $\mathbf{H}_t$, for every point in time $t$. Besides, assume that $\mathbf{H}_t$ is a finite realisation of some stochastic process observed at any point in time $t \in \mathscr{T} \subseteq \{1, \dots, T\}$ where $T \geq 1$.

**Remark.** In our notation, $N_t$ denotes the number of subjects at each point in time. It is important to stress that we talk about "subjects" figuratively. Indeed, our definition is not restricted to individuals, but extends to any abstract thing with a data structure compatible with assumption 25.

**Example 4** (Time-series cross sections). In the case of time-series cross-sectional data

$$\mathbf{H}_t = \begin{pmatrix} H_{1,1,t} & \dots & H_{1,K,t} \\ \vdots & \ddots & \vdots \\ H_{N_t,1,t} & \dots & H_{N_t,K,t} \end{pmatrix}$$

represents a cross section referring to time $t$ and

$$\mathbf{h}_t = \begin{pmatrix} H_{1,1,t} & \dots & H_{1,K,t} & \dots & H_{N_t,1,t} & \dots & H_{N_t,K,t} \end{pmatrix}',$$

where $N_t > 0$ is the number of cross-sectional observations for time $t$ and $K$ is the total number of covariates. In social sciences, similar datasets generally represent complex surveys with a varying number of respondents. However, $\mathbf{H}_t$ could also represent more exotic data. For instance, the RGB representation of a satellite image taken at time $t$ with $N_t$ pixels.

**Example 5** (Time series). The banal case in which $\mathbf{H}_t \in \mathbb{R}^{N_t}$ gives a time series dataset. Indeed, $\mathbf{H}_t = (H_{1,t} \;\; \dots, H_{N_t,t})'$ for any $t \in \mathscr{T}$. The value taken by $N_t$ over all $t \in \mathscr{T}$ controls whether this dataset represents a univariate or a multivariate time series, and if it is fully observed.

With empirical problems involving this data structure, new (old) subjects can be added (removed) over time. This implies that the same subject can be observed in our dataset at different positions across multiple points in time. We account for this complexity by associating a subject-characteristic identifier to each entry of $\mathbf{h}_t$, for every point in time $t \in \mathscr{T}$ in which at least one characteristic is observed.

**Definition 23** (Identifiers). In order to allow for this complexity, we let

$$\mathscr{S}_t := \{f(i,t) : 1 \leq i \leq N_t K\},$$

for any $t \in \mathscr{T}$. The function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ is a convenient way for categorising different subjects. Indeed, we structure it to be equal to one when evaluated at $(1,1)$ and to have an incremental value for any pair referring to a new feature of the same subject, or to a new subject. This implies that $f(i_1,t_1) = f(i_2,t_2)$ if and only if $(i_1,t_1)$ and $(i_2,t_2)$ refer to the same subject-characteristic pair. Hence, we let

$$\mathscr{S} := \bigcup_{t \in \mathscr{T}} \mathscr{S}_t$$

be the set of all (observed) subject-characteristic pairs. For simplicity, we let $N := \frac{|\mathscr{S}|}{K}$ be the number of unique subjects.

**Remark.** Under assumption 25, a minimum of one subject is observed across the whole sample and thus $N > 0$. Besides, note that $N$ is a natural number by construction.

Finally, we build on definition 23 and reshape the data into a multivariate time series.

**Definition 24.** Indeed, we let $\mathbf{Y}_t := \mathbf{W}_t \, \mathbf{h}_t$ to be a $NK \times 1$ vector of time series where $\mathbf{W}_t$ is a $NK \times N_t K$ matrix such that

$$W_{i,j,t} = \begin{cases} 1, & \text{if } f(j,t) = i \text{ and } i \in \mathscr{S}_t, \\ 0, & \text{otherwise,} \end{cases}$$

for any $t \in \mathscr{T}$, $1 \leq i \leq NK$ and $1 \leq j \leq N_t K$. In order to have a record of the non-missing entries of each $i$-th subject-characteristic pair, we also let $\mathscr{T}_i \subseteq \mathscr{T}$ to be the set of points in time in which it is observed. Interestingly, $\mathscr{T} = \bigcup_{i=1}^{N} \mathscr{T}_i$.

**Remark.** Note that by taking track of the observed datapoints via the $\mathscr{T}_i$ sets, we can distinguish between real zeros and missing values. Indeed, we do so in the estimation method described below. Moreover, due to assumption 25, all $\mathscr{T}_i$ referring to a single subject are identical (i.e., when we observe a subject we measure all of its characteristics).

### 3.2.2. Multidimensional dynamic factor models

This subsection formalises our approach for modelling generic multidimensional multivariate data via dynamic factor models. This methodology is then specialised in section 3.2.3 focussing on economics.

A multidimensional dynamic factor model (MDFM) is a decomposition of the data class described in section 3.2.1 into mutually orthogonal common and idiosyncratic components at all leads and lags.

**Assumption 26** (Generic MDFM). Going forward, we assume that the model for any $\mathbf{Y}_t$ is the multidimensional dynamic factor model

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{B}(L)\mathbf{\Phi}_t + \mathbf{e}_t, & \mathbf{e}_t &\overset{w.n.}{\sim} N(\mathbf{0}_{NK \times 1}, \mathbf{R}), \\ \mathbf{\Phi}_t &= \mathbf{C}(L)\mathbf{\Phi}_{t-1} + \mathbf{D}\mathbf{u}_t, & \mathbf{u}_t &\overset{w.n.}{\sim} N(\mathbf{0}_{r \times 1}, \mathbf{\Sigma}), \end{aligned}$$

where $\mathbf{\Phi}_t$ denotes a vector of latent components (stationary and/or non-stationary) and for some positive definite covariance matrices $\mathbf{R}$ and $\mathbf{\Sigma}$. The vector $\mathbf{\Phi}_t$ is $q$-dimensional with $1 \leq q \ll NK$, linked to the measurements via the matrix $\mathbf{B}(L)$ and with dynamics determined by $\mathbf{C}(L)$. Besides, $\mathbf{D}$ is $q \times r$ with $1 \leq r \leq q$.

**Remark.** In this article, the common components are not restricted to be stationary.

This model is extremely general and requires a set of restrictions in the parameters to be uniquely identified. This problem is analogous to the one observed with generic dynamic factor models (which are a particular case of MDFM) and it can be handled with the approaches proposed in Bai and Ng (2013) and Bai and Wang (2015). However, if the empirical problem at hand requires the extraction of multiple factors from datasets with subjects observed once or very few times over the full sample, it becomes hard to do. In those cases, it is often convenient to construct the dataset using both time series and multidimensional data. In doing so, a minimal number of subjects (i.e., the number of time series) is observed for most/all points in time and can be used for defining more solid identifying restrictions. Section 3.2.3 follows this approach for proposing a specialised model for economic data.

As for the case of standard dynamic factor models, the MDFM can be estimated with an EM (Dempster et al., 1977; Rubin and Thayer, 1982; Shumway and Stoffer, 1982; Watson and Engle, 1983; Bańbura and Modugno, 2014; Barigozzi and Luciani, 2020), ECM (Meng and Rubin, 1993; Pellegrino, 2023a,b) or ECME algorithm (Liu and Rubin, 1994), as well as with Bayesian methods (Särkkä, 2013). These techniques allow to have missing observations in the measurements, which are often found in the class of multidimensional data described in this manuscript.

### 3.2.3. A microfounded dynamic factor model

We specialise our approach to model microeconomic data jointly with macroeconomic aggregates. This subsection introduces our empirical research question and the economics-informed restrictions we employ for identifying the MDFM.

We propose to use a MDFM for understanding the effect of expansions and recessions on individual US households. In particular, we aim to do so studying the sensitivity of their real income per head to changes in the business cycle (BC), while taking into account the differences that exist across demographic groups (both temporary and persistent). This is an important question for politicians and central bankers. Indeed, an accurate answer would allow to systematically target fiscal and monetary policies for addressing the needs of specific demographic groups.

We collect data on US households from the Consumer Expenditure (CE) Public Use Microdata (PUMD). This is a vast dataset containing information on consumers and their household, including demographic characteristics, income and expenditure figures. The data is collected by the Census Bureau for the Bureau of Labor Statistics in the Interview Survey and Diary Survey. We focus on the first – which is the one describing major and/or recurring items – to gather information on quarterly income before tax and descriptive characteristics at the household level.[1]

In particular, we use the FMLI and ITBI files published from 1990 to 2020 for constructing a quarterly dataset containing demographic and nominal income data.[2] We exclude the subset of households that has not provided enough information to be categorised under one or more of the demographic characteristics in table 3.2.1, those whose attributes changed over time and the consumer units that have not provided any information on their income at all.[3] Moreover, we focus on prime working age urban consumer units (i.e., 25 to 54 years). The resulting dataset comprises a total of approximately 87,000 households.

**Definition 25** (Groups). Define $\mathscr{G}$ as the Cartesian product of the household attributes on education and ethnicity in table 3.2.1: a set with cardinality four such that each member is a unique combination of characteristics that identifies a specific demographic group. For simplicity, we refer to these groups in the order: (0, 0), (0, 1), (1, 0), (1, 1) whereas zero and one refers to the values taken by the binary variables EDUC_HH and WHITE_HH. Finally, we also let $\boldsymbol{\omega} \equiv \boldsymbol{\omega}(\mathscr{G})$ be the vector of integers denoting the number of households per group observed across all periods.

[1] Note that the BLS refers to households as consumer units (CUs). We use them as synonyms.

[2] We have decided to start from the 1990 file since the ITBI data was not available from 1981 to 1989. Note that the 1990 file also includes data referring to 1989 (from October).

[3] We do not exclude households whose income changed over time or with an incomplete income record, as long as we have at least one observation.

| Description | Mnemonic | Categorical | File |
|---|---|---|---|
| Census region | REGION | Y | FMLI |
| College educated household | EDUC_HH | Y | FMLI |
| Family size | FAM_SIZE | N | FMLI |
| Family type | FAM_TYPE | Y | FMLI |
| Prime working age | PRIME_AGE | Y | FMLI |
| Real household income per head (before tax) | INCOME | N | FMLI and ITBI |
| Urban consumers | BLS_URBN | Y | FMLI |
| White household | WHITE_HH | Y | FMLI |

**Table 3.2.1:** Consumer Expenditure (CE) Public Use Microdata (PUMD) selection. The data is extracted from the FMLI and ITBI files published from 1990 to 2020. We deflate the nominal income per head data in the ITBI using the PCE price index in table 3.2.2. Further details on the data construction are reported in section 3.A.
**Source:** Census Bureau for the Bureau of Labor Statistics, Bureau of Economic Analysis.

In addition to the microeconomic data, we also use macroeconomic aggregates. We first transform the nominal income figures obtained from the CE PUMD into real terms deflating them with the headline PCE price index – keeping them at household level. Next, we merge the resulting real income figures with the macro dataset in table 3.2.2. In order to perform these operations correctly, we download the macroeconomic series from the Archival Federal Reserve Economic Data (ALFRED) database and use the vintage released right after the 2020 CE PUMD Interview Survey's publication date.

**Definition 26** (Empirical data). We then arrange the data to match the structure in definition 24 and let

$$\mathbf{Y}_t = \left(\mathbf{X}_t' \; \mathbf{Z}_{1,t}' \; \ldots \; \mathbf{Z}_{4,t}'\right)',$$

where $\mathbf{X}_t$ denotes the vector of macroeconomic aggregates and each $\mathbf{Z}_{i,t}$ represents the vector of real income per head for all households in group $1 \leq i \leq 4$.

**Remark.** Note that every $\mathbf{Z}_{i,t}$ is $\omega_i \times 1$ dimensional. Since the CE PUMD is structured to survey the same household for a maximum of 4 quarters, the $\mathbf{Z}_{i,t}$ vectors are sparse. Besides, the missing observations in $\mathbf{Y}_t$ are handled as in definition 24. Finally, recall that this application focusses on the four demographic groups indicated in definition 25.

For simplicity, the macroeconomic aggregates are used in the same order reported in table 3.2.2. Any within-group ordering for the households is equivalent for our MDFM. We collect them in ascending order, on the basis of the official NEWID identifier available in Consumer Expenditure Public Use Microdata.[4]

Having shaped the data in the form prescribed in section 3.2.1 we are now ready to specialise the MDFM for this household problem. Similarly to recent work on semi-

[4] Note that the last digit of the NEWID refers to the interview number and the previous ones identify the consumer units. As a result, we have not considered the last digit of NEWID to identify the households and determine the within-group ordering.

| Description | Mnemonic | Source |
|---|---|---|
| Real gross domestic product | GDPC1 | BEA |
| Real personal consumption expenditures | PCECC96 | BEA |
| Real gross private domestic investment | GPDIC1 | BEA |
| Total nonfarm employment | PAYEMS | BLS |
| Employment-population ratio | EMRATIO | BLS |
| Unemployment rate | UNRATE | BLS |
| Spot crude oil price (WTI) | WTISPLC | FRBSL |
| Headline PCE | PCEPI | BEA |

**Table 3.2.2:** Macroeconomic aggregates. The dataset is quarterly and includes all observations available in the vintage released right after the 2020 CE PUMD Interview Survey's publication date, starting from October 1989 (to be aligned with the 1990 ITBI). All series are downloaded and used in levels, except for the prices which are transformed in quarterly year-on-year percentage changes. **Source:** Archival Federal Reserve Economic Data (ALFRED) database.

structural models including Hasenzagl et al. (2022a,b) and the empirical application in Pellegrino (2023b), we identify the model via economics-informed restrictions in order to extract interpretable unobserved components.

**Assumption 27.** Formally, we let

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \\ \vdots \\ X_{8,t} \\ \mathbf{Z}_{1,t} \\ \mathbf{Z}_{2,t} \\ \mathbf{Z}_{3,t} \\ \mathbf{Z}_{4,t} \end{pmatrix} = \begin{pmatrix} \tau_{1,t} \\ \tau_{2,t} \\ \vdots \\ \tau_{8,t} \\ \tau_{9,t}\,\boldsymbol{\iota}_{\omega_1} \\ \tau_{10,t}\,\boldsymbol{\iota}_{\omega_2} \\ (\tau_{9,t}+\tau_{11,t})\,\boldsymbol{\iota}_{\omega_3} \\ (\tau_{10,t}+\tau_{11,t})\,\boldsymbol{\iota}_{\omega_4} \end{pmatrix} + \begin{pmatrix} 1 \\ \sum_{i=1}^{p} \Lambda_{1,i} L^{i-1} \\ \vdots \\ \sum_{i=1}^{p} \Lambda_{7,i} L^{i-1} \\ \sum_{i=1}^{p} \Lambda_{8,i}\,\boldsymbol{\iota}_{\omega_1} L^{i-1} \\ \sum_{i=1}^{p} \Lambda_{9,i}\,\boldsymbol{\iota}_{\omega_2} L^{i-1} \\ \sum_{i=1}^{p} \Lambda_{10,i}\,\boldsymbol{\iota}_{\omega_3} L^{i-1} \\ \sum_{i=1}^{p} \Lambda_{11,i}\,\boldsymbol{\iota}_{\omega_4} L^{i-1} \end{pmatrix} \psi_t + \begin{pmatrix} \xi_{1,t} \\ \xi_{2,t} \\ \vdots \\ \xi_{8,t} \\ \xi_{9,t} \\ \xi_{10,t} \\ \xi_{11,t} \\ \xi_{12,t} \end{pmatrix} + \mathbf{e}_t$$

where $\psi_t$ is a causal AR($p$) cycle denoting the business cycle; the $\tau$ denote smooth trends of order two modelled as in Kitagawa and Gersch (1996, ch. 8); $\xi_{1,t}, \ldots, \xi_{8+|\mathscr{G}|,t}$ are causal AR(1) latent components representing idiosyncratic noise; $\boldsymbol{\iota}$ denotes a vector of ones with length indicated in the subscript. Hereinafter, the number of lags $p$ is assumed being equal to 4 (quarters).

**Remark** (Trends). Recall that a generic smooth trend $\underline{\tau}$ modelled as in Kitagawa and Gersch (1996, ch. 8) is of order $k$ if $(1-L)^k\,\underline{\tau}$ is a white noise. Besides, note that the income figures share common trends. In particular: $\boldsymbol{\tau}_9$ models the persistent part of income for not college educated, not white households; $\boldsymbol{\tau}_{10}$ models the persistent part of income for not college educated, white households; $\boldsymbol{\tau}_{11}$ models the persistent offset of college educated households.

**Remark** (CE PUMD data). Assumption 27 implies that each household is modelled as a function of its own group and the dedicated parameters. In other words, all members of the $i$-th group are modelled via the same set of coefficients and latent factors, for every $1 \leq i \leq 4$. While the generic structure proposed in assumption 26 could allow for a more disaggregate model, we do not have enough observations in the CE PUMD to do it. That being said, the model in assumption 27 has quite a few advantages compared to these granular theoretical alternatives. Most importantly, it is less subject to idiosyncratic noise and due to the dimensionality reduction into group factors it is easier to interpret.

The dynamics for the latent factors and the estimation method proposed for this model are illustrated in section 3.B. The estimation is based on penalised quasi maximum likelihood estimation (PQMLE) and built on an ECM algorithm similar to the one employed in Pellegrino (2023b).

## 3.3. Empirical results

### 3.3.1. Pre COVID-19 output

We start analysing the results focussing on the pre COVID-19 period (1989 to 2019) and using a model estimated with the same cutoff.

Figure 3.3.1 reports the macroeconomic aggregates and their trends. The model uses them for describing the slow-moving and persistent component typical of economic time series. The difference between data and trend is the cycle. In the case of real GDP and unemployment rate, their trends are unobserved quantities of economic interest: the so-called potential output and non-accelerating inflation rate of unemployment (NAIRU). The Congressional Budget Office (CBO) publishes their own estimates for these objects which we use to benchmark ours. It is evident from figure 3.3.1 that there are strong differences only in the case of potential output. Indeed, our calculations imply a causal cycle with mean zero, whereas the CBO estimates a negative cycle for most periods. This is consistent with trend-cycle decompositions based purely on macroeconomic aggregates. Economic implications of this difference in view on potential output are discussed in Hasenzagl et al. (2022a,b).

Figure 3.3.2 shows similar results for the income figures extracted from the CE PUMD. The main difference is that each subplot represents a group of households, not a single aggregate indicator. The demographic information is presented graphically through the following summary statistics: average, 25% and 75% quantiles. The trends do not refer to any specific household, but rather on the whole group. From a distributional standpoint, figure 3.3.2 shows four important points: most households have below-average income; a few individuals have disproportionate high revenues compared to the rest of their own demographic; white households are usually higher earners; college education increases the average income level. Our trend structure, further remarked after assumption 27, is

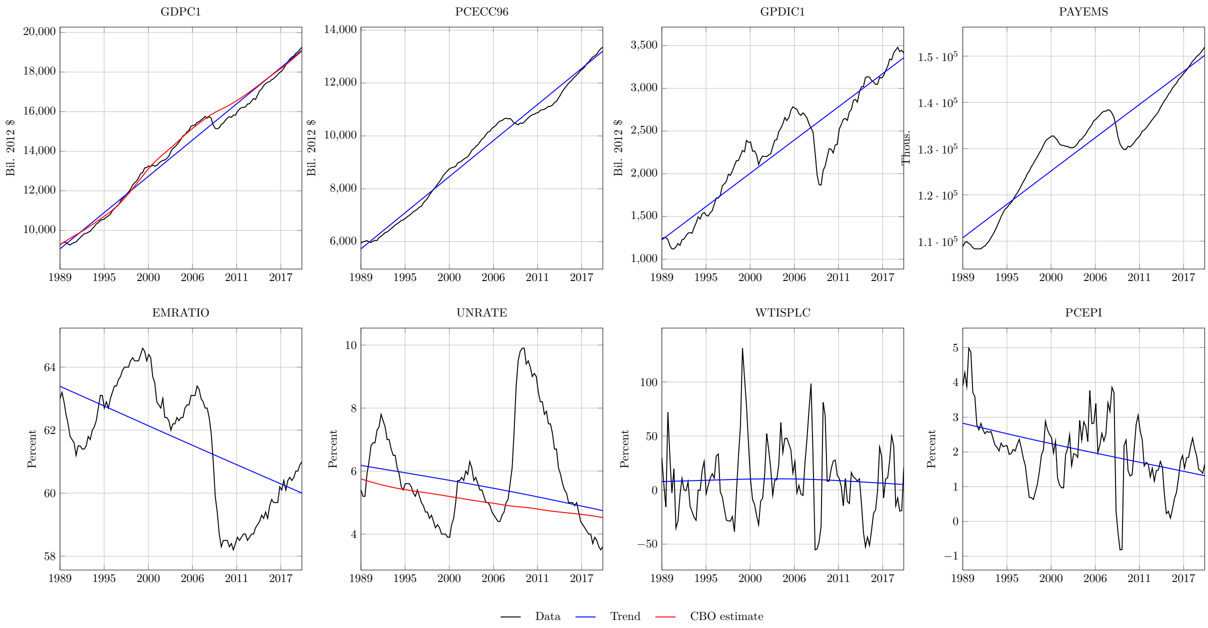


**Figure 3.3.1:** Macroeconomic data and trends.
**Notes**: The model is estimated with quarterly data from October 1989 to December 2019. The congressional budget office estimates are aligned with the ALFRED vintage in table 3.2.2.

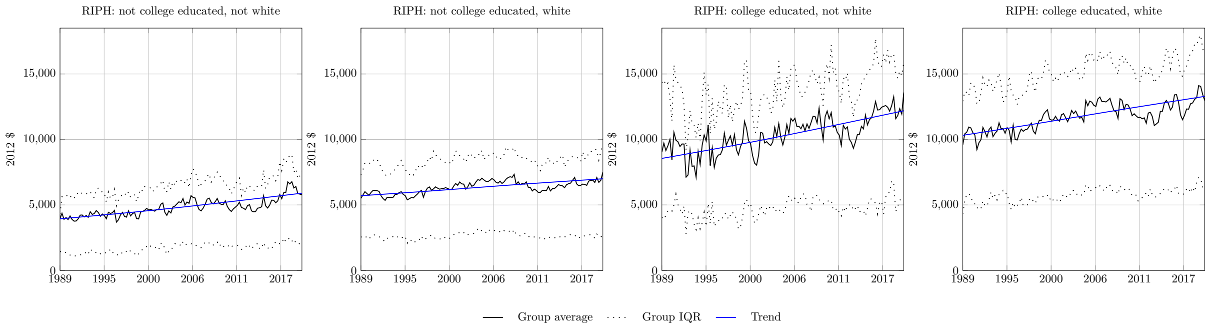


**Figure 3.3.2:** Microeconomic data and trends.
**Notes**: The model is estimated with quarterly data from October 1989 to December 2019.

flexible enough to accommodate for these features. Indeed, figure 3.3.2 show that white and college educated households have persistently higher trends.

Figure 3.3.3 breaks down the cycles to highlight commonalities and idiosyncrasies. The former are modelled through the business cycle and explain most of the cyclical fluctuations across macroeconomic aggregates and demographic groups. Idiosyncratic fluctuations, on the other hand, depict unique movements in specific macroeconomic indicators or groups. These are most prevalent for microeconomic data. Indeed, while the effect of the business cycle is comparable across demographics, each group exhibits distinct idiosyncratic patterns. Figure 3.C.1 builds on this further reporting the core drivers of the demographic groups: the sum between their trends and business cycles. Stripping out the idiosyncratic cycle helps visualising the crucial parts of real household income. Indeed, the resulting series is less impacted by outliers and noise.

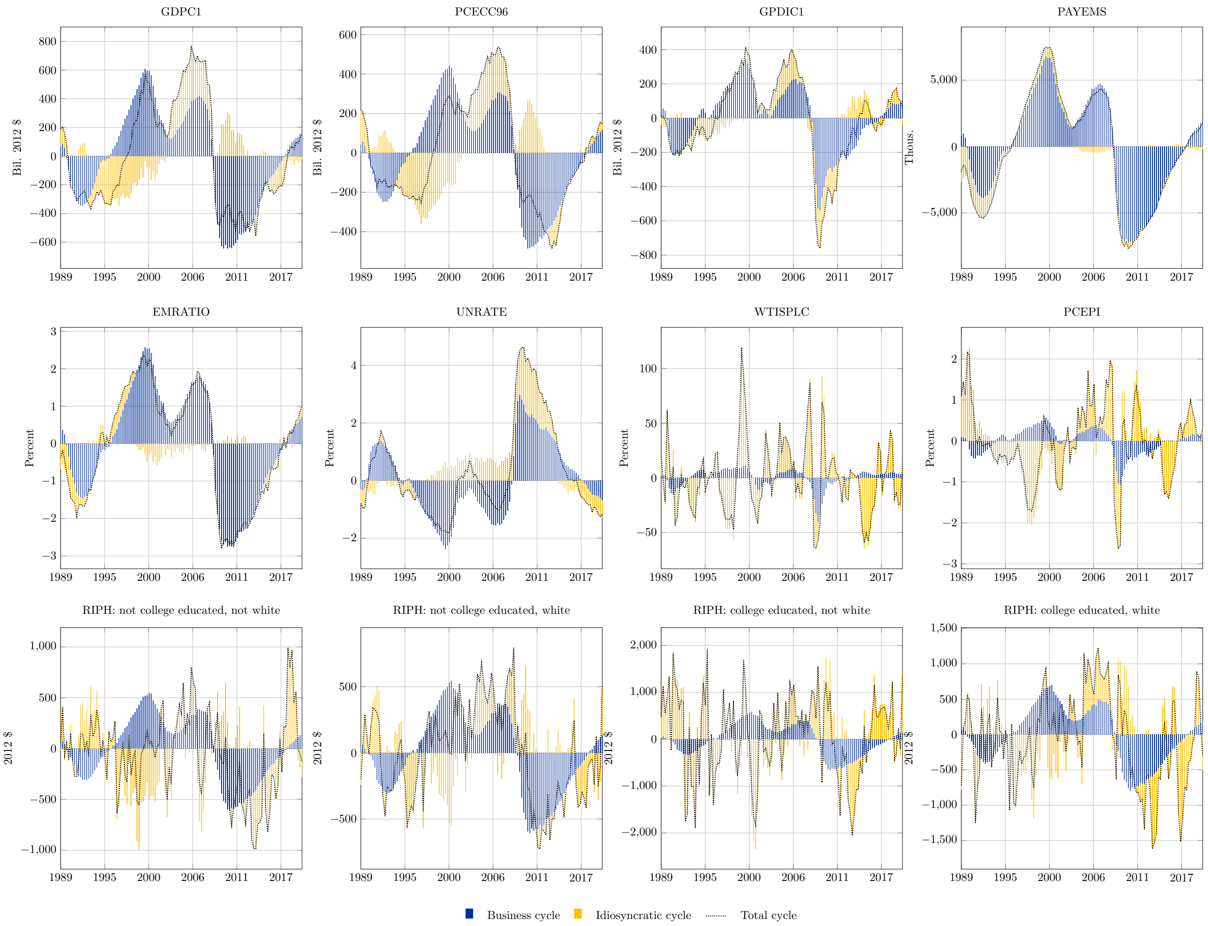


**Figure 3.3.3:** Historical decomposition of the cycles.
**Notes**: The model is estimated with quarterly data from October 1989 to December 2019.

## 3.3.2. COVID-19 dataflow

We now focus on the dataflow from January 2020 to March 2021 for studying the impact of COVID-19 on our estimates for the demographic groups. Throughout this subsection, we keep using the coefficients estimated with data from 1989 to 2019 to avoid altering the business cycle periodicity with a non traditional recession.

Before getting into the results, it is important to mention that the CE PUMD files are released in one block for the whole year and with a large delay from their reference period. Indeed, it is usually possible to access data in the Interview Survey only after 3 months from the end of its reference year. For instance, the 2020 CE PUMD was released at the end of March 2021. However, extending our information set with more timely macroeconomic data we can compute early estimates.

We process the hereinbefore mentioned dataflow in a pseudo real-time fashion and generate early estimates for the microeconomic data at each release. In particular, we produce backcasts and nowcasts: forecasts referring to the previous and current reference quarters (Giannone et al., 2008). Given that the CE PUMD data is released one go, we keep backcasting previous quarters until the publication date. Figure 3.3.4 reports

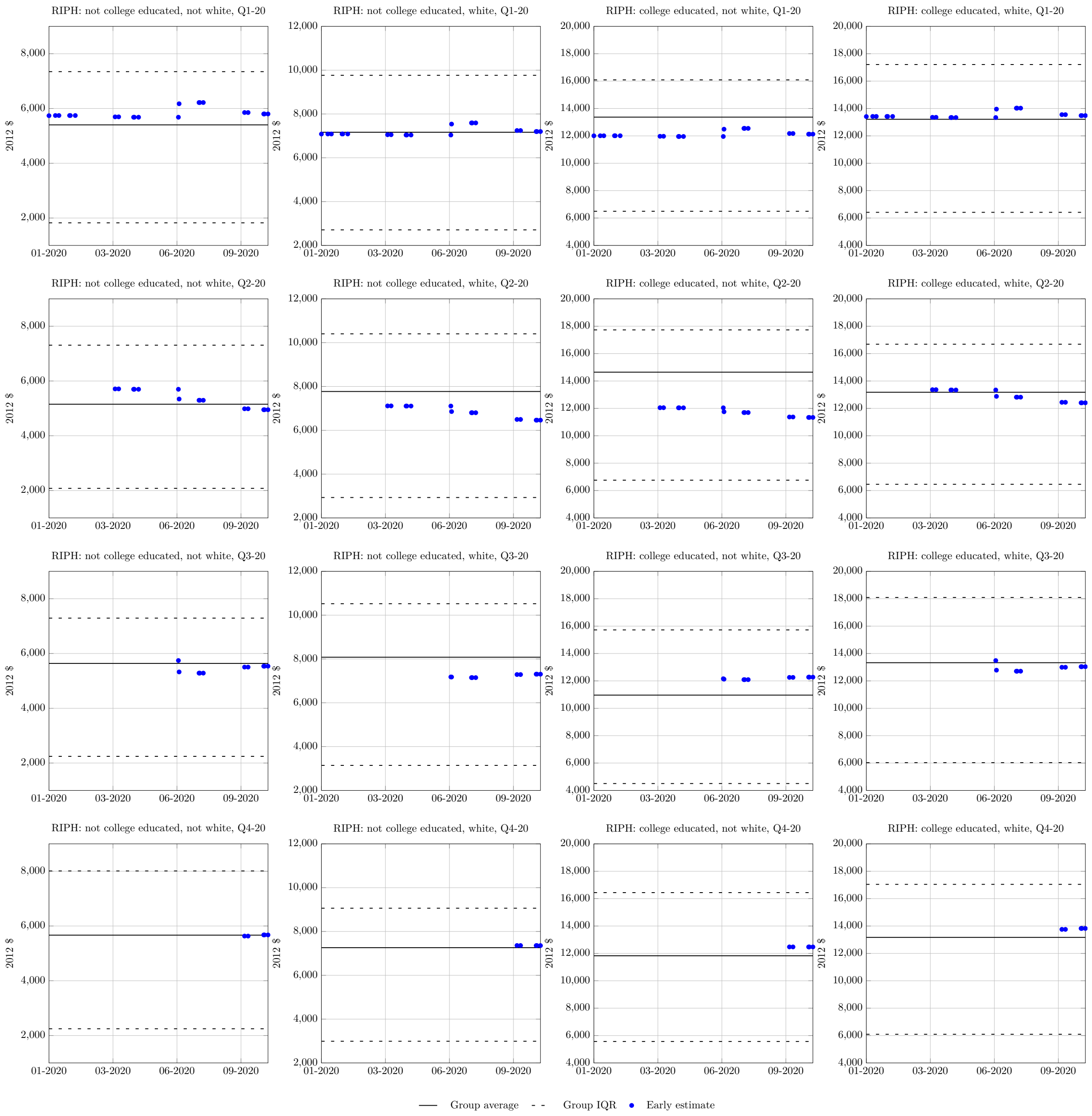


**Figure 3.3.4:** COVID-19 period dataflow.
**Notes**: The model is estimated with quarterly data from October 1989 to December 2019. The dataflow contains macroeconomic releases from January 2020 to March 2021. The x-axis reports release dates.

the results for each demographic of interest and, for simplicity, denotes backcasts and nowcasts as “early estimates”. Overall, these predictions fluctuate closely to the ex-post group averages. This happens almost immediately and, thus, the expanding macroeconomic information set does not have a strong impact. The early estimates for the second quarter are the most distant from the ex-post group averages. This is not surprising since the strongest effect of COVID-19 on economic data was measured in that quarter. We can also see that sentiment (and expectations) became increasingly negative since after March, when the World Health Organization (WHO) declared COVID-19 a pandemic.

## 3.4. Concluding remarks

This article proposes to generalise Dynamic Factor Models to multidimensional data. The resulting framework is flexible enough to accommodate complex datasets ranging from surveys with varying number of respondents to time series of satellite images. However, it retains the interpretability typical of traditional factor models.

We specialise our approach to model macroeconomic aggregates jointly with microeconomic data on household income. In this analysis, we study the effect of college education and ethnicity on the household income levels. In doing so, we find that our model is capable of recognising differences among demographics consistent with well-known stylised facts. Indeed, it finds that college education has a positive and persistent effect on household income and that white consumer units usually have higher earnings. We also explore the cyclical fluctuations in the data and highlight the heterogeneity among demographics.

Finally, realising that CE PUMD files are released with a large delay from their reference period, we show how to track them in real time focussing on the macroeconomic dataflow between January 2020 and March 2021. This is in line with the nowcasting literature (Giannone et al., 2008) and, to the best of our knowledge, the first attempt to perform a similar exercise on microeconomic data.

# Appendix

## 3.A. CE PUMD

The demographic characteristics in table 3.2.1 are constructed at the household level. The following paragraphs give further details on each variable.

- *Census region*: Census Bureau classification for US regions (1 Northeast, 2 Midwest, 3 South, 4 West). We use it for excluding CUs that moved across the US during the sampling period.
- *College educated households*: describes the highest level of education of the reference person and spouse (if any). It is a dummy variable equal to 1 for CUs in which the highest education level is, at least, at an undergraduate level and 0 otherwise.
- *Family size*: Number of family members. We use it for computing real household income per head (before tax).
- *Family type*: Family categorisation. We use it for determine whether we there is a spouse to consider when constructing the other variables in this appendix.
- *Prime working age*: dummy variable equal to 1 for CUs with average age between 25 and 55 years (excluded) and 0 otherwise. The average age is computed by taking the sample mean between the age of the reference person and spouse (if any). We use it for excluding non prime working age households.
- *Urban consumers*: dummy variable equal to 1 for urban CUs and 0 otherwise. We use it for excluding rural CUs.
- *White household*: dummy variable equal to 1 for white CUs and 0 otherwise.

The nominal income per head (before tax) is computed by constructing total nominal income from the ITBI files and dividing it for the number of CUs members in the FMLI files. The identifiers or Universal Classification Code (UCC) for each single income component used for computing this total are summarised in table 3.A.1.

| Releases | Universal Classification Codes (UCCs) |
|---|---|
| 1990 to 2003 | 900000, 900010, 900020, 900030, 900040, 900080, 900050, 900060, 900070, 900100, 900110, 900090, 900120, 900150, 900131, 900132, 800700I, 800710I, 900140 |
| 2004 to 2012 | 900000, 900010, 900020, 900030, 900040, 900080, 900050, 900060, 900070, 900100, 900110, 900090, 900120, 900150, 900131, 900132, 800700, 800710, 900140 |
| 2013 to 2020 | 900000, 900160, 900030, 900170, 900180, 900190, 900200, 900090, 900120, 900150, 900210, 800700, 800710, 900140 |

**Table 3.A.1:** Universal Classification Codes (UCCs) used for computing nominal income.
**Source:** Census Bureau for the Bureau of Labor Statistics.

# 3.B. ECM algorithm

This appendix develops an ECM algorithm (Meng and Rubin, 1993) to estimate the MDFM in section 3.2.3. The design builds on Pellegrino (2023a, Appendix C) and Pellegrino (2023b, Appendix A). This manuscript uses the “hat” symbol to denote the estimated coefficients, an $s$ subscript to indicate the sample size and a $k$ superscript for the ECM iteration.

## 3.B.1. State-space representation

Recall that

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{B}(L)\mathbf{\Phi}_t + \mathbf{e}_t, & \mathbf{e}_t &\overset{w.n.}{\sim} N(\mathbf{0}_{NK\times 1}, \mathbf{R}), \\ \mathbf{\Phi}_t &= \mathbf{C}(L)\mathbf{\Phi}_{t-1} + \mathbf{D}\mathbf{u}_t, & \mathbf{u}_t &\overset{w.n.}{\sim} N(\mathbf{0}_{r\times 1}, \mathbf{\Sigma}). \end{aligned}$$

The matrices $\mathbf{C}(L)$ and $\mathbf{D}$ are sparse and their non-zero entries are such that

$$\mathbf{C}(L) = \begin{array}{c} \left(\begin{array}{c|ccc|ccccc|c}
2\mathbf{I}_{7+|\mathscr{G}|} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\mathbf{I}_{7+|\mathscr{G}|} \\ \hline
\cdot & \pi_1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \ddots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \pi_{8+|\mathscr{G}|} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \hline
\cdot & \cdot & \cdot & \cdot & \pi_{8+|\mathscr{G}|+1} & \pi_{8+|\mathscr{G}|+2} & \cdots & \pi_{8+|\mathscr{G}|+p-1} & \pi_{8+|\mathscr{G}|+p} & \cdot \\
\cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \ddots & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \hline
\mathbf{I}_{7+|\mathscr{G}|} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right) \\
\begin{array}{cccc} \underbrace{\qquad}_{q\times 7+|\mathscr{G}|} & \underbrace{\qquad\qquad}_{q\times 8+|\mathscr{G}|} & \underbrace{\qquad\qquad\qquad\qquad\qquad}_{q\times p} & \underbrace{\qquad}_{q\times 7+|\mathscr{G}|} \end{array} \end{array}$$

and

$$\mathbf{D} = \begin{array}{c} \left(\begin{array}{c|c|c}
\mathbf{I}_{7+|\mathscr{G}|} & \cdot & \cdot \\ \hline
\cdot & \mathbf{I}_{8+|\mathscr{G}|} & \cdot \\ \hline
\cdot & \cdot & 1 \\ \hline
\cdot & \cdot & \cdot
\end{array}\right) \\
\begin{array}{ccc} \underbrace{\qquad}_{q\times 7+|\mathscr{G}|} & \underbrace{\qquad}_{q\times 8+|\mathscr{G}|} & \underbrace{\quad}_{q\times 1} \end{array} \end{array}$$

where $\boldsymbol{\pi}$ is a $8+|\mathscr{G}|+p\times 1$ vector of finite real parameters which ensures that the cyclical components are causal. Due to the structure of $\mathbf{C}$ and $\mathbf{D}$, it follows that $r = 16 + 2|\mathscr{G}|$ and $q = 22 + 3|\mathscr{G}| + p$. The measurement coefficient matrix $\mathbf{B}(L)$ is also sparse and its non-zero entries are

$$\left(\underbrace{\begin{matrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \ddots & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{matrix}}_{NK\times 7+|\mathscr{G}|} \middle| \underbrace{\begin{matrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \ddots & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{matrix}}_{NK\times 8+|\mathscr{G}|} \middle| \underbrace{\begin{matrix} 1 & \cdot & \dots & \cdot \\ \Lambda_{1,1} & \Lambda_{1,2} & \dots & \Lambda_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ \Lambda_{7,1} & \Lambda_{7,2} & \dots & \Lambda_{7,p} \\ \Lambda_{8,1}\,\boldsymbol{\iota}_{\omega_1} & \Lambda_{8,2}\,\boldsymbol{\iota}_{\omega_1} & \dots & \Lambda_{8,p}\,\boldsymbol{\iota}_{\omega_1} \\ \vdots & \vdots & \vdots & \vdots \\ \Lambda_{7+|\mathscr{G}|,1}\,\boldsymbol{\iota}_{\omega_{|\mathscr{G}|}} & \Lambda_{7+|\mathscr{G}|,2}\,\boldsymbol{\iota}_{\omega_{|\mathscr{G}|}} & \dots & \Lambda_{7+|\mathscr{G}|,p}\,\boldsymbol{\iota}_{\omega_{|\mathscr{G}|}} \end{matrix}}_{NK\times p} \middle| \underbrace{\begin{matrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{matrix}}_{NK\times 7+|\mathscr{G}|} \right).$$

As a result,

$$\boldsymbol{\Phi}_t := \Big( \underbrace{\tau_{1,t} \quad \dots \quad \tau_{7+|\mathscr{G}|,t}}_{7+|\mathscr{G}|\times 1} \Big| \underbrace{\xi_{1,t} \quad \dots \quad \xi_{8+|\mathscr{G}|,t}}_{8+|\mathscr{G}|\times 1} \Big| \underbrace{\psi_t \quad \psi_{t-1} \quad \dots \quad \psi_{t-p+1}}_{p\times 1} \Big| \underbrace{\tau_{1,t-1} \quad \dots \quad \tau_{7+|\mathscr{G}|,t-1}}_{7+|\mathscr{G}|\times 1} \Big)'.$$

**Assumption 28** (Initial conditions). Given that we observe data at time $t = 1$, we further assume that $\boldsymbol{\Phi}_0 \overset{w.n.}{\sim} N(\boldsymbol{\mu}_0, \boldsymbol{\Omega}_0)$ for some finite vector of real parameters $\boldsymbol{\mu}_0$ and a positive definite covariance matrix $\boldsymbol{\Omega}_0$. The latter is assumed to be sparse and such that the only entries allowed to differ from zero are those denoting the initial auto-covariances of each state.

**Remark** (Non-zero entries of $\boldsymbol{\Omega}_0$). In other words, the entries of $\boldsymbol{\Omega}_0$ that are allowed to differ from zero are those with coordinates $(i, j)$ in the union of the following sets:

- $\{(i, j) : i = j \text{ and } 1 \leq i < r\}$;
- $\{(i, j) : r \leq i < r + p \text{ and } r \leq j < r + p\}$.

## 3.B.2. Estimation

This manuscript builds on the theoretical results in Barigozzi and Luciani (2020) and estimates the model via quasi penalised maximum likelihood estimation (PQMLE) by considering $\boldsymbol{\Sigma}$ as a diagonal matrix and $\mathbf{R} = \varepsilon\,\mathbf{I}_{NK}$ for a small positive $\varepsilon$.[5] Formally, this implies that the free parameters to estimate are

$$\boldsymbol{\vartheta} := \Big( \boldsymbol{\mu}_0' \quad \operatorname{vech}(\boldsymbol{\Omega}_0)' \quad \operatorname{vec}(\boldsymbol{\Lambda})' \quad \boldsymbol{\pi}' \quad \Sigma_{1,1} \quad \Sigma_{2,2} \quad \dots \quad \Sigma_{r,r} \Big)'.$$

The estimation is performed with an ECM algorithm: an optimisation method that repeats the operations in definition 27 until it reaches convergence.

[5] We set $\varepsilon = 10^{-2}$.

**Definition 27** (ECM estimation routine). At any $k+1 > 1$ iteration, the ECM algorithm computes the vector of coefficients

$$\hat{\boldsymbol{\vartheta}}_s^{k+1}(\boldsymbol{\gamma}) := \underset{\underline{\boldsymbol{\vartheta}} \in \mathscr{R}}{\arg\max}\; \mathbb{E}\left[\mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Y}_{1:s}, \boldsymbol{\Phi}_{1:s}) \,|\, \mathscr{Y}(s),\, \hat{\boldsymbol{\vartheta}}_s^{k}(\boldsymbol{\gamma})\right] - \mathbb{E}\left[\mathcal{P}(\underline{\boldsymbol{\vartheta}}, \boldsymbol{\gamma}) \,|\, \mathscr{Y}(s),\, \hat{\boldsymbol{\vartheta}}_s^{k}(\boldsymbol{\gamma})\right],$$

where $\mathscr{R}$ denotes the region in which the AR cycles (common and idiosyncratic) are causal, $\mathscr{Y}(s)$ is the information set available at time $s$,

$$\begin{aligned}\mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Y}_{1:s}, \boldsymbol{\Phi}_{1:s}) \simeq & -\frac{1}{2}\ln|\underline{\boldsymbol{\Omega}_0}| - \frac{1}{2}\operatorname{Tr}\left[\underline{\boldsymbol{\Omega}_0}^{-1}(\boldsymbol{\Phi}_0 - \underline{\boldsymbol{\mu}_0})(\boldsymbol{\Phi}_0 - \underline{\boldsymbol{\mu}_0})'\right] \\ & -\frac{s}{2}\ln|\underline{\boldsymbol{\Sigma}}| - \frac{1}{2}\operatorname{Tr}\left[\sum_{t=1}^{s}\underline{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\Phi}_{1:r,t} - \underline{\mathbf{C}}_*\boldsymbol{\Phi}_{t-1})(\boldsymbol{\Phi}_{1:r,t} - \underline{\mathbf{C}}_*\boldsymbol{\Phi}_{t-1})'\right] \\ & -\frac{s}{2}\ln|\underline{\mathbf{R}}| - \frac{1}{2}\operatorname{Tr}\left[\sum_{t=1}^{s}\underline{\mathbf{R}}^{-1}(\mathbf{Y}_t - \underline{\mathbf{B}}\boldsymbol{\Phi}_t)(\mathbf{Y}_t - \underline{\mathbf{B}}\boldsymbol{\Phi}_t)'\right],\end{aligned} \tag{3.1}$$

$\underline{\mathbf{C}}_* \equiv \underline{\mathbf{C}}_{1:r,1:q}$ and the underlined coefficients denote the parameters implied by $\underline{\boldsymbol{\vartheta}}$. The function in equation 3.1 is the so-called complete-data (i.e., fully observed data and known latent states) log-likelihood. Besides,

$$\begin{aligned}\mathcal{P}(\underline{\boldsymbol{\vartheta}}, \boldsymbol{\gamma}) := & +\frac{1-\alpha}{2}\left(\left\|\underline{\boldsymbol{\pi}}_{1:8+|\mathscr{G}|}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},1)^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 + \left\|\underline{\boldsymbol{\pi}}'_{8+|\mathscr{G}|+1:8+|\mathscr{G}|+p}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},p)^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 + \left\|\underline{\boldsymbol{\Lambda}}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},p)^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right) \\ & +\frac{\alpha}{2}\left(\left\|\underline{\boldsymbol{\pi}}_{1:8+|\mathscr{G}|}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},1)\right\|_{1,1} + \left\|\underline{\boldsymbol{\pi}}'_{8+|\mathscr{G}|+1:8+|\mathscr{G}|+p}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},p)\right\|_{1,1} + \left\|\underline{\boldsymbol{\Lambda}}\,\boldsymbol{\Gamma}(\boldsymbol{\gamma},p)\right\|_{1,1}\right)\end{aligned}$$

is a version of the elastic-net penalty in Pellegrino (2023a) in which, for any $l \in \mathbb{N}$,

$$\boldsymbol{\Gamma}(\boldsymbol{\gamma}, l) := \rho \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \beta & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \beta^{l-1} \end{pmatrix},$$

$\rho \geq 0$, $0 \leq \alpha \leq 1$ and $\beta \geq 1$ are hyperparameters included in $\boldsymbol{\gamma}$. The state-space coefficients for the first iteration are initialised as in section 3.B.3.

**Assumption 29** (Convergence). The ECM algorithm is considered to be converged when the estimated coefficients do not significantly change in two subsequent iterations. This is done by computing the absolute relative change per parameters and comparing at the same time the median and $95^{th}$ quantile with a fixed tolerance of $10^{-3}$ and $10^{-2}$ respectively.

The operation in definition 27 is performed in two steps: the so-called E-step and CM-step. The E-step computes the expectations in definition 27, while the CM-step conditionally maximises them with respect to the free parameters.

We write down the E-step on the basis of the output of a Kalman smoother compatible with incomplete data, as originally proposed in Shumway and Stoffer (1982) and Watson and Engle (1983). For that, we use the notation in definition 28.

**Definition 28** (Kalman smoother output). The hereinbefore mentioned Kalman smoother output is

$$\begin{aligned}\hat{\mathbf{\Phi}}_t &:= \mathbb{E}\left[\mathbf{\Phi}_t \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right],\\ \hat{\mathbf{P}}_{t,t-j} &:= \operatorname{Cov}\left[\mathbf{\Phi}_t, \mathbf{\Phi}_{t-j} \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right],\end{aligned}$$

for any $k \geq 0$, $0 \leq j \leq t$ and $t \geq 0$. Let also $\hat{\mathbf{P}}_t \equiv \hat{\mathbf{P}}_{t,t}$.

**Remark.** These estimates are computed as in Pellegrino (2023a).

We also use the notation in definition 29 to further deal with missing observations.

**Definition 29** (Observed measurements). Recall that $\mathscr{T} = \bigcup_{i=1}^N \mathscr{T}_i$ and let

$$\mathscr{T}(s) := \{t : t \in \mathscr{T},\, 1 \leq t \leq s\},$$

for $1 \leq s \leq T$. Let also

$$\mathscr{D}_t := \{i : t \in \mathscr{T}_i,\, 1 \leq i \leq NK\},$$

for $1 \leq t \leq T$. Finally, let

$$\begin{aligned}\mathbf{Y}_t^{obs} &:= \left(Y_{i,t}\right)_{i \in \mathscr{D}_t}\\ \mathbf{B}_t^{obs} &:= \mathbf{A}_t \mathbf{B}\end{aligned}$$

be the $|\mathscr{D}_t| \times 1$ vector of observed measurements at time $t$ and the corresponding $|\mathscr{D}_t| \times q$ matrix of coefficients, for any $t \in \mathscr{T}$. Every $\mathbf{A}_t$ is indeed a selection matrix constituted by ones and zeros that permits to retrieve the appropriate rows of $\mathbf{B}$ for every $t \in \mathscr{T}$.

Moreover, in order to simplify the notation, we let

$$\mathcal{L}_e\left[\underline{\boldsymbol{\vartheta}} \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] \equiv \mathbb{E}\left[\mathcal{L}(\underline{\boldsymbol{\vartheta}} \,|\, \mathbf{Y}_{1:s}, \mathbf{\Phi}_{1:s}) \,|\, \mathscr{Y}(s), \hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right]$$

for $1 \leq s \leq T$.

It then follows directly from Pellegrino (2023b, Proposition 1) that

$$\begin{aligned}\mathcal{L}_e\left[\underline{\boldsymbol{\vartheta}}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] \simeq & -\frac{1}{2}\ln|\underline{\boldsymbol{\Omega}_0}| - \frac{1}{2}\operatorname{Tr}\left[\underline{\boldsymbol{\Omega}_0}^{-1}\Big(\hat{\boldsymbol{\Phi}}_0\hat{\boldsymbol{\Phi}}_0' + \hat{\mathbf{P}}_0 - \hat{\boldsymbol{\Phi}}_0\underline{\boldsymbol{\mu}_0}' - \underline{\boldsymbol{\mu}_0}\hat{\boldsymbol{\Phi}}_0' + \underline{\boldsymbol{\mu}_0}\,\underline{\boldsymbol{\mu}_0}'\Big)\right] \\ & - \frac{s}{2}\ln|\underline{\boldsymbol{\Sigma}}| - \frac{1}{2}\operatorname{Tr}\left[\underline{\boldsymbol{\Sigma}}^{-1}\Big(\hat{\mathbf{E}}_s^{(1)} - \hat{\mathbf{E}}_s^{(2)}\underline{\mathbf{C}}_*' - \underline{\mathbf{C}}_*\hat{\mathbf{E}}_s^{(2)'} + \underline{\mathbf{C}}_*\hat{\mathbf{E}}_s^{(3)}\underline{\mathbf{C}}_*'\Big)\right] \\ & - \frac{1}{2\varepsilon}\operatorname{Tr}\left\{\sum_{t\in\mathscr{T}(s)}\left[\Big(\mathbf{Y}_t^{obs} - \underline{\mathbf{B}}_t^{obs}\,\hat{\boldsymbol{\Phi}}_t\Big)\Big(\mathbf{Y}_t^{obs} - \underline{\mathbf{B}}_t^{obs}\,\hat{\boldsymbol{\Phi}}_t\Big)' + \underline{\mathbf{B}}_t^{obs}\,\hat{\mathbf{P}}_t\,\underline{\mathbf{B}}_t^{obs'}\right]\right\},\end{aligned}$$

where

$$\begin{aligned}\hat{\mathbf{E}}_s^{(1)} &:= \sum_{t=1}^{s}\mathbb{E}\Big[\boldsymbol{\Phi}_{1:r,t}\boldsymbol{\Phi}_{1:r,t}'\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\Big] = \sum_{t=1}^{s}\Big(\hat{\boldsymbol{\Phi}}_t\hat{\boldsymbol{\Phi}}_t' + \hat{\mathbf{P}}_t\Big)_{1:r,1:r}, \\ \hat{\mathbf{E}}_s^{(2)} &:= \sum_{t=1}^{s}\mathbb{E}\Big[\boldsymbol{\Phi}_{1:r,t}\boldsymbol{\Phi}_{t-1}'\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\Big] = \sum_{t=1}^{s}\Big(\hat{\boldsymbol{\Phi}}_t\hat{\boldsymbol{\Phi}}_{t-1}' + \hat{\mathbf{P}}_{t,t-1}\Big)_{1:r,1:q}, \\ \hat{\mathbf{E}}_s^{(3)} &:= \sum_{t=1}^{s}\mathbb{E}\Big[\boldsymbol{\Phi}_{t-1}\boldsymbol{\Phi}_{t-1}'\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\Big] = \sum_{t=1}^{s}\Big(\hat{\boldsymbol{\Phi}}_{t-1}\hat{\boldsymbol{\Phi}}_{t-1}' + \hat{\mathbf{P}}_{t-1}\Big),\end{aligned}$$

for $1 \leq s \leq T$. Furthermore, it follows from Pellegrino (2023b, Lemma 1) that

$$\mathbb{E}\left[\mathcal{P}(\underline{\boldsymbol{\vartheta}},\boldsymbol{\gamma})\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] = \mathcal{P}(\underline{\boldsymbol{\vartheta}},\boldsymbol{\gamma}).$$

The CM-step conditionally maximises the expected penalised log-likelihood

$$\mathcal{M}_e\left[\underline{\boldsymbol{\vartheta}},\boldsymbol{\gamma}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] := \mathcal{L}_e\left[\underline{\boldsymbol{\vartheta}}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right] - \mathcal{P}(\underline{\boldsymbol{\vartheta}},\boldsymbol{\gamma}) \tag{3.2}$$

with respect to the free parameters.

The CM-steps for all free parameters of the transition equation are reported in Pellegrino (2023b, Lemmas 2−4). For clarity, we recall just the lemmas' statements in lemmas 17–19 with the adequate minimal notational changes.

**Lemma 17.** *The ECM estimator at a generic iteration $k+1>0$ for $\boldsymbol{\mu}_0$ is*

$$\hat{\boldsymbol{\mu}}_{0,s}^{k+1}(\boldsymbol{\gamma}) = \hat{\boldsymbol{\Phi}}_0$$

*and the estimator for $\boldsymbol{\Omega}_0$, denoted with $\hat{\boldsymbol{\Omega}}_{0,s}^{k+1}(\boldsymbol{\gamma})$, is a sparse covariance matrix whose entries allowed to differ from zero are*

$$\left[\hat{\boldsymbol{\Omega}}_{0,s}^{k+1}(\boldsymbol{\gamma})\right]_{i,j} = \left[\hat{\mathbf{P}}_0\right]_{i,j},$$

*and the coordinates $(i,j)$ are those described in assumption 28.*

**Lemma 18.** *Note that $\boldsymbol{\Gamma}(\boldsymbol{\gamma},1)=\rho$ and let $\tilde{\boldsymbol{\Gamma}}(\boldsymbol{\gamma})$ be a $q\times q$ sparse matrix whose non-zero* *elements are*

$$\tilde{\mathbf{\Gamma}}(\boldsymbol{\gamma}) := \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \rho\,\mathbf{I}_{8+|\mathscr{G}|} & \cdot & \cdot \\ \cdot & \cdot & \mathbf{\Gamma}(\boldsymbol{\gamma}, p) & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

*Moreover, let*

$$\begin{aligned} &\mathscr{U}_C := \{1, \ldots, r\} \times \{1, \ldots, q\}, \\ &\mathscr{U}_\pi := \{(i,j) : i = j \text{ and } 7 + |\mathscr{G}| < i < r\} \cup \{(i,j) : i = r \text{ and } r \leq j < r + p\}, \end{aligned}$$

*wherein the latter can be partitioned as $\{\mathscr{C}(i,j), (i,j), \mathscr{C}''(i,j)\}$ for any $(i,j) \in \mathscr{U}_\pi$. Hence, the ECM estimator at a generic iteration $k+1 > 0$ for $\mathbf{C}$ is such that*

$$\hat{C}^{k+1}_{i,j,s}(\boldsymbol{\gamma}) = \frac{\mathcal{S}\left[\hat{\Sigma}^{k-1}_{i,i,s}(\boldsymbol{\gamma})\,\hat{E}^{(2)}_{i,j,s} - \sum_{\substack{(l_1,l_2)\in\mathscr{U}_C \\ (l_1,l_2)\neq(i,j)}} \hat{\Sigma}^{k-1}_{i,l_1,s}(\boldsymbol{\gamma})\,\hat{C}^{k+\mathbb{I}_{(l_1,l_2)\in\mathscr{C}(i,j)}}_{l_1,l_2,s}(\boldsymbol{\gamma})\,\hat{E}^{(3)}_{l_2,j,s},\ \frac{\alpha}{2}\,\tilde{\Gamma}_{j,j}(\boldsymbol{\gamma})\right]}{\hat{\Sigma}^{k-1}_{i,i,s}(\boldsymbol{\gamma})\,\hat{E}^{(3)}_{j,j,s} + (1-\alpha)\,\tilde{\Gamma}_{j,j}(\boldsymbol{\gamma})},$$

*for any $(i,j) \in \mathscr{U}_\pi$ and constant to the values in section 3.B.1 for the remaining entries, and with $\mathcal{S}$ being the soft-thresholding operator.*

**Lemma 19.** *The ECM estimator at a generic iteration $k+1 > 0$ for $\mathbf{\Sigma}$ is such that*

$$\hat{\Sigma}^{k+1}_{i,i,s}(\boldsymbol{\gamma}) = \frac{1}{s}\left[\hat{\mathbf{E}}^{(1)}_s - \hat{\mathbf{E}}^{(2)}_s \hat{\mathbf{C}}^{k+1'}_s(\boldsymbol{\gamma}) - \hat{\mathbf{C}}^{k+1}_s(\boldsymbol{\gamma})\,\hat{\mathbf{E}}^{(2)'}_s + \hat{\mathbf{C}}^{k+1}_s(\boldsymbol{\gamma})\,\hat{\mathbf{E}}^{(3)}_s \hat{\mathbf{C}}^{k+1'}_s(\boldsymbol{\gamma})\right]_{i,i}$$

*for $i = 1, \ldots, r$ and zero for the remaining entries.*

The CM-step for the free parameters of the measurement equation requires an ad-hoc approach due to the implicit equality constraints for the households described in section 3.2.3 – i.e., all households within a given group have the same factor loadings. While linear constraints have been handled before in EM-like algorithms for time-series models (for instance, in order to apply mixed frequency aggregation constrains in now-casting problems as in Bańbura and Modugno, 2014), our problem is a bit more complicated. Indeed, it is not advised to estimate an unconstrained version of the model and apply the restrictions ex-post, since each consumer unit is observed for very short periods of time. Hence, we handle this CM-step as the constrained optimisation problem in proposition 6.[6]

[6] We do not need to use Lagrangian multipliers, given that the constraints can be implemented by directly plugging them into the expected log-likelihood via $\mathbf{B}$ as described in section 3.B.1.

**Proposition 6.** *Let*

$$\begin{aligned}\ddot{\mathbf{A}}_t &:= \mathbf{A}_t'\mathbf{A}_t,\\ \hat{\mathbf{F}}_t &:= \hat{\boldsymbol{\Phi}}_t\hat{\boldsymbol{\Phi}}_t' + \hat{\mathbf{P}}_t,\\ \hat{\mathbf{G}}_s &:= \sum_{t\in\mathscr{T}(s)} \mathbf{A}_t'\mathbf{Y}_t^{obs}\,\hat{\boldsymbol{\Phi}}_t',\end{aligned}$$

*and*

$$\begin{aligned}\hat{\text{Б}}_{i,j,s} &:= \sum_{k=\omega_i^\dagger}^{\omega_i^\ddagger} \hat{G}_{k,j,s},\\ \text{Л}_{i,j,t} &:= \sum_{k=\omega_i^\dagger}^{\omega_i^\ddagger} \ddot{A}_{k,j,t},\\ \text{Ш}_{i,j,t} &:= \sum_{k=\omega_i^\dagger}^{\omega_i^\ddagger}\sum_{l=\omega_j^\dagger}^{\omega_j^\ddagger} \ddot{A}_{k,l,t},\end{aligned}$$

*where* $\omega_i^\dagger := 9 + \sum_{1\leq j<i-7}\omega_j$ *and* $\omega_i^\ddagger := 8 + \sum_{1\leq j\leq i-7}\omega_j$, *for* $i = 8,\ldots,7+|\mathscr{G}|$. *Let also*

$$\mathscr{U}_\Lambda := \{1,\ldots,7+|\mathscr{G}|\}\times\{0,\ldots,p-1\}.$$

*It follows that, when the penalty is not active and at a generic* $k+1$ *iteration of the ECM algorithm, the factor loadings*

$$\begin{aligned}\hat{\Lambda}_{i,j+1}^{QMLE,\,k+1} = \frac{1}{\sum_{t\in\mathscr{T}(s)}\hat{F}_{j+r,j+r,t}\,\ddot{A}_{i+1,i+1,t}}\Bigg\{&\hat{G}_{i+1,j+r,s} - \sum_{t\in\mathscr{T}(s)}\Bigg[\sum_{l_1=1}^{NK}\sum_{l_2=1}^{r-1}\hat{F}_{j+r,l_2,t}\,B_{l_1,l_2}\,\ddot{A}_{i+1,l_1,t}\\ &+\hat{F}_{j+r,r,t}\,\ddot{A}_{1,i+1,t} + \sum_{\substack{(l_1,l_2)\in\mathscr{U}_\Lambda\\(l_1,l_2)\neq(i,j)}}\hat{F}_{j+r,l_2+r,t}\,\hat{\Lambda}^{\diamond}_{l_1,l_2+1}\left(\mathbb{I}_{l_1\leq 7}\,\ddot{A}_{l_1+1,i+1,t} + \mathbb{I}_{l_1>7}\,\text{Л}_{l_1,i+1,t}\right)\Bigg]\Bigg\}\end{aligned}$$

*for* $i = 1,\ldots,7$, *and*

$$\begin{aligned}\hat{\Lambda}_{i,j+1}^{QMLE,\,k+1} = \frac{1}{\sum_{t\in\mathscr{T}(s)}\hat{F}_{j+r,j+r,t}\,\text{Ш}_{i,i,t}}\Bigg\{&\hat{\text{Б}}_{i,j+r,s} - \sum_{t\in\mathscr{T}(s)}\Bigg[\sum_{l_1=1}^{NK}\sum_{l_2=1}^{r-1}\hat{F}_{j+r,l_2,t}\,B_{l_1,l_2}\,\text{Л}_{i,l_1,t}\\ &+\hat{F}_{r,j+r,t}\,\text{Л}_{i,1,t} + \sum_{\substack{(l_1,l_2)\in\mathscr{U}_\Lambda\\(l_1,l_2)\neq(i,j)}}\hat{F}_{j+r,l_2+r,t}\,\hat{\Lambda}^{\diamond}_{l_1,l_2+1}\left(\mathbb{I}_{l_1\leq 7}\,\text{Л}_{i,l_1+1,t} + \mathbb{I}_{l_1>7}\,\text{Ш}_{l_1,i,t}\right)\Bigg]\Bigg\}\end{aligned}$$

*for* $i = 8,\ldots,7+|\mathscr{G}|$ *where* $\hat{\Lambda}^{\diamond}_{l_1,l_2+1}$ *is the most up-to-date estimate for* $\Lambda_{l_1,l_2+1}$ *available when computing* $\hat{\Lambda}^{k+1}_{i,j+1}$, *for* $j = 0,\ldots,p-1$. *More generally, it follows that, at a generic*

$k+1$ *iteration of the ECM algorithm, the factor loadings*

$$\hat{\Lambda}^{k+1}_{i,j+1}(\boldsymbol{\gamma}) = \begin{cases} \dfrac{\mathcal{S}\left[\hat{\Lambda}^{QMLE,\,k+1}_{i,j+1} \sum_{t\in\mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \ddot{A}_{i+1,i+1,t}\,,\ \frac{\varepsilon\alpha}{2}\Gamma_{j+1,j+1}(\boldsymbol{\gamma},p)\right]}{\varepsilon(1-\alpha)\Gamma_{j+1,j+1}(\boldsymbol{\gamma},p) + \sum_{t\in\mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \ddot{A}_{i+1,i+1,t}}, & 1 \leq i \leq 7, \\ \dfrac{\mathcal{S}\left[\hat{\Lambda}^{QMLE,\,k+1}_{i,j+1} \sum_{t\in\mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \text{Ш}_{i,i,t}\,,\ \frac{\varepsilon\alpha}{2}\Gamma_{j+1,j+1}(\boldsymbol{\gamma},p)\right]}{\varepsilon(1-\alpha)\Gamma_{j+1,j+1}(\boldsymbol{\gamma},p) + \sum_{t\in\mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \text{Ш}_{i,i,t}}, & 8 \leq i \leq 7+|\mathscr{G}|, \end{cases}$$

*for* $j = 0, \ldots, p-1$.

PROOF. We develop the proof in three steps. Step (i) derives the part of the expected log-likelihood that depends on the factor loadings. Step (ii) solves the maximisation problem assuming that the penalty is not active (i.e., $\rho = 0$). This leads to a CM-step similar to the M-step usually employed in non-regularised EM-algorithms for dynamic factor models such as Bańbura and Modugno (2014). Step (iii) builds on that to write down the formula for the final estimator.

(i) Note that, for the linearity of the trace,

$$\begin{aligned} &\mathrm{Tr}\left\{\sum_{t\in\mathscr{T}(s)} \left[\left(\mathbf{Y}^{obs}_t - \underline{\mathbf{B}}^{obs}_t\, \hat{\boldsymbol{\Phi}}_t\right)\left(\mathbf{Y}^{obs}_t - \underline{\mathbf{B}}^{obs}_t\, \hat{\boldsymbol{\Phi}}_t\right)' + \underline{\mathbf{B}}^{obs}_t\, \hat{\mathbf{P}}_t\, \underline{\mathbf{B}}^{obs'}_t\right]\right\} \\ &\quad = \sum_{t\in\mathscr{T}(s)} \mathrm{Tr}\left[\left(\mathbf{Y}^{obs}_t - \underline{\mathbf{B}}^{obs}_t\, \hat{\boldsymbol{\Phi}}_t\right)\left(\mathbf{Y}^{obs}_t - \underline{\mathbf{B}}^{obs}_t\, \hat{\boldsymbol{\Phi}}_t\right)' + \underline{\mathbf{B}}^{obs}_t\, \hat{\mathbf{P}}_t\, \underline{\mathbf{B}}^{obs'}_t\right]. \end{aligned}$$

The part of this trace that depends on the measurement coefficients is

$$\begin{aligned} &\sum_{t\in\mathscr{T}(s)} \mathrm{Tr}\left(\underline{\mathbf{B}}^{obs}_t\, \hat{\mathbf{F}}_t\, \underline{\mathbf{B}}^{obs'}_t - \underline{\mathbf{B}}^{obs}_t\, \hat{\boldsymbol{\Phi}}_t\, \mathbf{Y}^{obs'}_t - \mathbf{Y}^{obs}_t\, \hat{\boldsymbol{\Phi}}'_t\, \underline{\mathbf{B}}^{obs'}_t\right) \\ &\quad = \sum_{t\in\mathscr{T}(s)} \mathrm{Tr}\left(\mathbf{A}_t\underline{\mathbf{B}}\, \hat{\mathbf{F}}_t\, \underline{\mathbf{B}}'\mathbf{A}'_t - \mathbf{A}_t\underline{\mathbf{B}}\, \hat{\boldsymbol{\Phi}}_t\, \mathbf{Y}^{obs'}_t - \mathbf{Y}^{obs}_t\, \hat{\boldsymbol{\Phi}}'_t\, \underline{\mathbf{B}}'\mathbf{A}'_t\right) \\ &\quad = \sum_{t\in\mathscr{T}(s)} \mathrm{Tr}\left(\mathbf{A}_t\underline{\mathbf{B}}\, \hat{\mathbf{F}}_t\, \underline{\mathbf{B}}'\mathbf{A}'_t - \underline{\mathbf{B}}\, \hat{\boldsymbol{\Phi}}_t\, \mathbf{Y}^{obs'}_t\, \mathbf{A}_t - \mathbf{A}'_t\, \mathbf{Y}^{obs}_t\, \hat{\boldsymbol{\Phi}}'_t\, \underline{\mathbf{B}}'\right) \\ &\quad = \sum_{t\in\mathscr{T}(s)} \mathrm{Tr}\left(\underline{\mathbf{B}}\, \hat{\mathbf{F}}_t\, \underline{\mathbf{B}}'\, \ddot{\mathbf{A}}_t\right) - \mathrm{Tr}\left(\underline{\mathbf{B}}\, \hat{\mathbf{G}}'_s\right) - \mathrm{Tr}\left(\hat{\mathbf{G}}_s\, \underline{\mathbf{B}}'\right) \\ &\quad = \sum_{t\in\mathscr{T}(s)} \mathrm{Tr}\left(\underline{\mathbf{B}}\, \hat{\mathbf{F}}_t\, \underline{\mathbf{B}}'\, \ddot{\mathbf{A}}_t\right) - 2\,\mathrm{Tr}\left(\underline{\mathbf{B}}\, \hat{\mathbf{G}}'_s\right). \end{aligned}$$

In order to write down the CM-step for the factor loadings, we develop these traces as functions of $\underline{\boldsymbol{\Lambda}}$.[7] We start with the simpler trace. Under the identification restrictions

[7] Indeed, these are the only free parameters in $\underline{\mathbf{B}}$.

and constraints in section 3.2.3,

$$
\begin{aligned}
&\mathrm{Tr}\left(\underline{\mathbf{B}}\,\hat{\mathbf{G}}_s'\right)\\
&\quad=\sum_{i=1}^{NK}\sum_{j=1}^{q}\underline{B}_{i,j}\,\hat{G}_{i,j,s}\\
&\quad=\sum_{i=1}^{NK}\sum_{j=1}^{r-1}\underline{B}_{i,j}\,\hat{G}_{i,j,s}+\sum_{i=1}^{NK}\sum_{j=r}^{q}\underline{B}_{i,j}\,\hat{G}_{i,j,s}\\
&\quad=\sum_{i=1}^{NK}\sum_{j=1}^{r-1}\underline{B}_{i,j}\,\hat{G}_{i,j,s}+\sum_{i=1}^{8}\sum_{j=r}^{q}\underline{B}_{i,j}\,\hat{G}_{i,j,s}+\sum_{i=9}^{NK}\sum_{j=r}^{q}\underline{B}_{i,j}\,\hat{G}_{i,j,s}\\
&\quad=\sum_{i=1}^{NK}\sum_{j=1}^{r-1}\underline{B}_{i,j}\,\hat{G}_{i,j,s}+\sum_{i=1}^{8}\sum_{j=r}^{r+p-1}\underline{B}_{i,j}\,\hat{G}_{i,j,s}+\sum_{i=9}^{NK}\sum_{j=r}^{r+p-1}\underline{B}_{i,j}\,\hat{G}_{i,j,s}\\
&\quad=\sum_{i=1}^{NK}\sum_{j=1}^{r-1}\underline{B}_{i,j}\,\hat{G}_{i,j,s}+\sum_{i=1}^{8}\sum_{j=0}^{p-1}\underline{B}_{i,j+r}\,\hat{G}_{i,j+r,s}+\sum_{i=9}^{NK}\sum_{j=0}^{p-1}\underline{B}_{i,j+r}\,\hat{G}_{i,j+r,s}\\
&\quad=\sum_{i=1}^{NK}\sum_{j=1}^{r-1}\underline{B}_{i,j}\,\hat{G}_{i,j,s}\;+\;\hat{G}_{1,r,s}+\sum_{i=1}^{7}\sum_{j=0}^{p-1}\underline{\Lambda}_{i,j+1}\,\hat{G}_{i+1,j+r,s}+\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\underline{\Lambda}_{i,j+1}\hat{\text{Б}}_{i,j+r,s}.
\end{aligned}
$$

Thus,

$$
\mathrm{Tr}\left(\underline{\mathbf{B}}\,\hat{\mathbf{G}}_s'\right)\propto\sum_{i=1}^{7}\sum_{j=0}^{p-1}\underline{\Lambda}_{i,j+1}\,\hat{G}_{i+1,j+r,s}+\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\underline{\Lambda}_{i,j+1}\hat{\text{Б}}_{i,j+r,s}. \tag{3.3}
$$

Next, we focus on the most complicated component of the expected log-likelihood. Since $\ddot{\mathbf{A}}_t$ and $\hat{\mathbf{F}}_t$ are symmetric,

$$
\begin{aligned}
&\mathrm{Tr}\left(\underline{\mathbf{B}}\,\hat{\mathbf{F}}_t\,\underline{\mathbf{B}}'\,\ddot{\mathbf{A}}_t\right)\\
&\quad=\sum_{i=1}^{NK}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=1}^{NK}\sum_{j=1}^{r-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}+\sum_{i=1}^{NK}\sum_{j=r}^{q}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\qquad+\sum_{i=1}^{NK}\sum_{j=1}^{r-1}\sum_{k=r}^{q}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}+\sum_{i=1}^{NK}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=1}^{NK}\sum_{j=1}^{r-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}+2\sum_{i=1}^{NK}\sum_{j=r}^{q}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\qquad+\sum_{i=1}^{NK}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}.
\end{aligned}
$$

Thus, the only part of the latter trace that depends on the factor loadings is

$$\sum_{i=1}^{NK}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}+2\sum_{i=1}^{NK}\sum_{j=r}^{q}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}. \tag{3.4}$$

The first term of equation 3.4 is

$$\begin{aligned}
&\sum_{i=1}^{NK}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=1}^{8}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=1}^{8}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}+\sum_{i=9}^{NK}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=9}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\qquad+\sum_{i=1}^{8}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=9}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}+\sum_{i=9}^{NK}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=1}^{8}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=1}^{8}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=1}^{8}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}+\sum_{i=9}^{NK}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=9}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\qquad+2\sum_{i=1}^{8}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=9}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}.
\end{aligned}$$

Under the identification restrictions and constraints in section 3.2.3,

$$\begin{aligned}
&\sum_{i=1}^{8}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=1}^{8}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\hat{F}_{r,r,t}\,\ddot{A}_{1,1,t}+2\sum_{i=2}^{8}\sum_{j=r}^{q}\underline{B}_{i,j}\,\hat{F}_{j,r,t}\,\ddot{A}_{1,i,t}+\sum_{i=2}^{8}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=2}^{8}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\hat{F}_{r,r,t}\,\ddot{A}_{1,1,t}+2\sum_{i=2}^{8}\sum_{j=r}^{r+p-1}\underline{B}_{i,j}\,\hat{F}_{j,r,t}\,\ddot{A}_{1,i,t}+\sum_{i=2}^{8}\sum_{j=r}^{r+p-1}\sum_{k=r}^{r+p-1}\sum_{l=2}^{8}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\hat{F}_{r,r,t}\,\ddot{A}_{1,1,t}+2\sum_{i=2}^{8}\sum_{j=0}^{p-1}\underline{B}_{i,j+r}\,\hat{F}_{j+r,r,t}\,\ddot{A}_{1,i,t}+\sum_{i=2}^{8}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=2}^{8}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k+r,t}\,\underline{B}_{l,k+r}\,\ddot{A}_{l,i,t}\\
&\quad=\hat{F}_{r,r,t}\,\ddot{A}_{1,1,t}+2\sum_{i=1}^{7}\sum_{j=0}^{p-1}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,r,t}\,\ddot{A}_{1,i+1,t}+\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=1}^{7}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\ddot{A}_{l+1,i+1,t}.
\end{aligned}$$

Also,

$$\begin{aligned}
&\sum_{i=9}^{NK}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=9}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=9}^{NK}\sum_{j=r}^{r+p-1}\sum_{k=r}^{r+p-1}\sum_{l=9}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=9}^{NK}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=9}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k+r,t}\,\underline{B}_{l,k+r}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=8}^{7+|\mathscr{G}|}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\text{Ш}_{l,i,t}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\sum_{i=1}^{8}\sum_{j=r}^{q}\sum_{k=r}^{q}\sum_{l=9}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=1}^{8}\sum_{j=r}^{r+p-1}\sum_{k=r}^{r+p-1}\sum_{l=9}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=1}^{8}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=9}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k+r,t}\,\underline{B}_{l,k+r}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{k=0}^{p-1}\sum_{l=9}^{NK}\hat{F}_{r,k+r,t}\,\underline{B}_{l,k+r}\,\ddot{A}_{l,1,t}+\sum_{i=2}^{8}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=9}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k+r,t}\,\underline{B}_{l,k+r}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=9}^{NK}\sum_{j=0}^{p-1}\hat{F}_{r,j+r,t}\,\underline{B}_{i,j+r}\,\ddot{A}_{i,1,t}+\sum_{i=2}^{8}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=9}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k+r,t}\,\underline{B}_{l,k+r}\,\ddot{A}_{l,i,t}\\
&\quad=\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\hat{F}_{r,j+r,t}\,\underline{\Lambda}_{i,j+1}\,\text{Л}_{i,1,t}+\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=8}^{7+|\mathscr{G}|}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\text{Л}_{l,i+1,t}.
\end{aligned}$$

Thus, the first term of equation 3.4 is proportional to

$$\begin{aligned}
&2\sum_{i=1}^{7}\sum_{j=0}^{p-1}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,r,t}\,\ddot{A}_{1,i+1,t}+\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=1}^{7}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\ddot{A}_{l+1,i+1,t}\\
&\quad+\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=8}^{7+|\mathscr{G}|}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\text{Ш}_{l,i,t}+2\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\hat{F}_{r,j+r,t}\,\underline{\Lambda}_{i,j+1}\,\text{Л}_{i,1,t}\\
&\qquad+2\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=8}^{7+|\mathscr{G}|}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\text{Л}_{l,i+1,t}.
\end{aligned}\tag{3.5}$$

Finally, the second term of equation 3.4 is

$$
\begin{aligned}
&2\sum_{i=1}^{NK}\sum_{j=r}^{q}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\qquad=2\sum_{i=1}^{NK}\sum_{j=r}^{r+p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j}\,\hat{F}_{j,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\qquad=2\sum_{i=1}^{NK}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{l,i,t}\\
&\qquad=2\sum_{i=1}^{NK}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i,l,t}\\
&\qquad=2\sum_{i=1}^{8}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i,l,t}+2\sum_{i=9}^{NK}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i,l,t},
\end{aligned}
$$

where

$$
\begin{aligned}
&\sum_{i=1}^{8}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i,l,t}\\
&\qquad=\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\hat{F}_{r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{1,l,t}+\sum_{i=2}^{8}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i,l,t}\\
&\qquad=\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\hat{F}_{r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{1,l,t}+\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i+1,l,t}
\end{aligned}
$$

and

$$
\sum_{i=9}^{NK}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{B}_{i,j+r}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i,l,t}=\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\text{Л}_{i,l,t}.
$$

Hence, the second term of equation 3.4 is proportional to

$$
2\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i+1,l,t}+2\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\text{Л}_{i,l,t}.
\tag{3.6}
$$

Combining equations 3.4–3.6, it follows that

$$
\begin{aligned}
&\operatorname{Tr}\left(\underline{\mathbf{B}}\,\hat{\mathbf{F}}_t\,\underline{\mathbf{B}}'\,\ddot{\mathbf{A}}_t\right) \qquad (3.7)\\
&\quad\propto 2\sum_{i=1}^{7}\sum_{j=0}^{p-1}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,r,t}\,\ddot{A}_{1,i+1,t}+\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=1}^{7}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\ddot{A}_{l+1,i+1,t}\\
&\qquad+2\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\hat{F}_{r,j+r,t}\,\underline{\Lambda}_{i,j+1}\,\text{Л}_{i,1,t}+\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=8}^{7+|\mathscr{G}|}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\text{Ш}_{l,i,t}\\
&\qquad\quad+2\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i+1,l,t}+2\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\text{Л}_{i,l,t}\\
&\qquad\qquad+2\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=8}^{7+|\mathscr{G}|}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\text{Л}_{l,i+1,t}.
\end{aligned}
$$

Finally, it follows from equation 3.3 and equation 3.7 that

$$
\begin{aligned}
&\sum_{t\in\mathscr{T}(s)}\operatorname{Tr}\left(\underline{\mathbf{B}}\,\hat{\mathbf{F}}_t\,\underline{\mathbf{B}}'\,\ddot{\mathbf{A}}_t\right)-2\operatorname{Tr}\left(\underline{\mathbf{B}}\,\hat{\mathbf{G}}_s'\right) \qquad (3.8)\\
&\quad\propto\sum_{t\in\mathscr{T}(s)}\sum_{i=1}^{7}\sum_{j=0}^{p-1}\Bigg(2\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,r,t}\,\ddot{A}_{1,i+1,t}+\sum_{k=0}^{p-1}\sum_{l=1}^{7}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\ddot{A}_{l+1,i+1,t}\\
&\qquad+2\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\ddot{A}_{i+1,l,t}\Bigg)-2\sum_{i=1}^{7}\sum_{j=0}^{p-1}\underline{\Lambda}_{i,j+1}\,\hat{G}_{i+1,j+r,s}\\
&\qquad\quad+\sum_{t\in\mathscr{T}(s)}\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\Bigg(2\hat{F}_{r,j+r,t}\,\underline{\Lambda}_{i,j+1}\,\text{Л}_{i,1,t}+\sum_{k=0}^{p-1}\sum_{l=8}^{7+|\mathscr{G}|}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\text{Ш}_{l,i,t}\\
&\qquad\qquad+2\sum_{k=1}^{r-1}\sum_{l=1}^{NK}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k,t}\,\underline{B}_{l,k}\,\text{Л}_{i,l,t}\Bigg)-2\sum_{i=8}^{7+|\mathscr{G}|}\sum_{j=0}^{p-1}\underline{\Lambda}_{i,j+1}\hat{\text{Б}}_{i,j+r,s}\\
&\qquad\qquad\quad+2\sum_{t\in\mathscr{T}(s)}\sum_{i=1}^{7}\sum_{j=0}^{p-1}\sum_{k=0}^{p-1}\sum_{l=8}^{7+|\mathscr{G}|}\underline{\Lambda}_{i,j+1}\,\hat{F}_{j+r,k+r,t}\,\underline{\Lambda}_{l,k+1}\,\text{Л}_{l,i+1,t}.
\end{aligned}
$$

We have rearranged the terms in equation 3.8 so that the first two rows refer to the factor loadings of the macroeconomic indicators, the third and fourth row refer to the ones of the households and the last row to both of them.

(ii) When the penalty is not active, the CM-step is computed from equation 3.8 since

$$
\frac{\partial\mathcal{M}_e\left[\underline{\boldsymbol{\vartheta}},\boldsymbol{\gamma}\,|\,\mathscr{Y}(s),\hat{\boldsymbol{\vartheta}}_s^k(\boldsymbol{\gamma})\right]}{\partial\underline{\boldsymbol{\Lambda}}}=-\frac{1}{2\varepsilon}\frac{\partial\left[\sum_{t\in\mathscr{T}(s)}\operatorname{Tr}\left(\underline{\mathbf{B}}\,\hat{\mathbf{F}}_t\,\underline{\mathbf{B}}'\,\ddot{\mathbf{A}}_t\right)-2\operatorname{Tr}\left(\underline{\mathbf{B}}\,\hat{\mathbf{G}}_s'\right)\right]}{\partial\underline{\boldsymbol{\Lambda}}}.
$$

We structure the CM-step by following analogous steps to those in Pellegrino (2023b). Indeed, we estimate $\boldsymbol{\Lambda}$ one entry at the time, starting from the $\Lambda_{1,1}$ and in column-major order. In other words, the derivative of equation 3.8 with respect to $\Lambda_{i,j+1}$ is taken having fixed the other factors loadings to their latest estimate, for any $i=1,\ldots,7+|\mathscr{G}|$ and

$j = 0, \dots, p-1$. Formally, at a generic $k+1$ iteration of the ECM algorithm, this derivative is equal to

$$\begin{aligned}
&\frac{\hat{G}_{i+1,j+r,s}}{\varepsilon} - \frac{1}{\varepsilon} \sum_{t \in \mathscr{T}(s)} \Bigg( \sum_{l_1=1}^{NK} \sum_{l_2=1}^{r-1} \hat{F}_{j+r,l_2,t}\, B_{l_1,l_2}\, \ddot{A}_{i+1,l_1,t} + \hat{F}_{j+r,r,t}\, \ddot{A}_{1,i+1,t} + \underline{\Lambda}_{i,j+1} \hat{F}_{j+r,j+r,t}\, \ddot{A}_{i+1,i+1,t} \\
&\qquad + \sum_{\substack{(l_1,l_2) \in \mathscr{U}_\Lambda \\ (l_1,l_2) \neq (i,j) \\ l_1 \leq 7}} \hat{F}_{j+r,l_2+r,t}\, \hat{\Lambda}^{\diamond}_{l_1,l_2+1}\, \ddot{A}_{l_1+1,i+1,t} + \sum_{\substack{(l_1,l_2) \in \mathscr{U}_\Lambda \\ (l_1,l_2) \neq (i,j) \\ l_1 > 7}} \hat{F}_{j+r,l_2+r,t}\, \hat{\Lambda}^{\diamond}_{l_1,l_2+1}\, \text{Л}_{l_1,i+1,t} \Bigg)
\end{aligned}$$

when computed with respect to any factor loading associated to the macroeconomic aggregates, and

$$\begin{aligned}
&\frac{\hat{\text{Б}}_{i,j+r,s}}{\varepsilon} - \frac{1}{\varepsilon} \sum_{t \in \mathscr{T}(s)} \Bigg( \sum_{l_1=1}^{NK} \sum_{l_2=1}^{r-1} \hat{F}_{j+r,l_2,t}\, B_{l_1,l_2}\, \text{Л}_{i,l_1,t} + \hat{F}_{r,j+r,t}\, \text{Л}_{i,1,t} + \underline{\Lambda}_{i,j+1}\, \hat{F}_{j+r,j+r,t}\, \text{Ш}_{i,i,t} \\
&\qquad + \sum_{\substack{(l_1,l_2) \in \mathscr{U}_\Lambda \\ (l_1,l_2) \neq (i,j) \\ l_1 \leq 7}} \hat{\Lambda}^{\diamond}_{l_1,l_2+1}\, \hat{F}_{l_2+r,j+r,t}\, \text{Л}_{i,l_1+1,t} + \sum_{\substack{(l_1,l_2) \in \mathscr{U}_\Lambda \\ (l_1,l_2) \neq (i,j) \\ l_1 > 7}} \hat{F}_{j+r,l_2+r,t}\, \hat{\Lambda}^{\diamond}_{l_1,l_2+1}\, \text{Ш}_{l_1,i,t} \Bigg)
\end{aligned}$$

when computed with respect to any factor loading associated to the households data. These derivatives can be equivalently written in the compact forms

$$\begin{aligned}
&\frac{\hat{G}_{i+1,j+r,s}}{\varepsilon} - \frac{1}{\varepsilon} \sum_{t \in \mathscr{T}(s)} \Bigg[ \sum_{l_1=1}^{NK} \sum_{l_2=1}^{r-1} \hat{F}_{j+r,l_2,t}\, B_{l_1,l_2}\, \ddot{A}_{i+1,l_1,t} + \hat{F}_{j+r,r,t}\, \ddot{A}_{1,i+1,t} + \underline{\Lambda}_{i,j+1} \hat{F}_{j+r,j+r,t}\, \ddot{A}_{i+1,i+1,t} \\
&\qquad + \sum_{\substack{(l_1,l_2) \in \mathscr{U}_\Lambda \\ (l_1,l_2) \neq (i,j)}} \hat{F}_{j+r,l_2+r,t}\, \hat{\Lambda}^{\diamond}_{l_1,l_2+1} \left( \mathbb{I}_{l_1 \leq 7}\, \ddot{A}_{l_1+1,i+1,t} + \mathbb{I}_{l_1 > 7}\, \text{Л}_{l_1,i+1,t} \right) \Bigg]
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\hat{\text{Б}}_{i,j+r,s}}{\varepsilon} - \frac{1}{\varepsilon} \sum_{t \in \mathscr{T}(s)} \Bigg[ \sum_{l_1=1}^{NK} \sum_{l_2=1}^{r-1} \hat{F}_{j+r,l_2,t}\, B_{l_1,l_2}\, \text{Л}_{i,l_1,t} + \hat{F}_{r,j+r,t}\, \text{Л}_{i,1,t} + \underline{\Lambda}_{i,j+1}\, \hat{F}_{j+r,j+r,t}\, \text{Ш}_{i,i,t} \\
&\qquad + \sum_{\substack{(l_1,l_2) \in \mathscr{U}_\Lambda \\ (l_1,l_2) \neq (i,j)}} \hat{F}_{j+r,l_2+r,t}\, \hat{\Lambda}^{\diamond}_{l_1,l_2+1} \left( \mathbb{I}_{l_1 \leq 7}\, \text{Л}_{i,l_1+1,t} + \mathbb{I}_{l_1 > 7}\, \text{Ш}_{l_1,i,t} \right) \Bigg]
\end{aligned}$$

respectively. It follows that, when the penalty is not active and at a generic $k+1$ iteration of the ECM algorithm,

$$\begin{aligned}
\hat{\Lambda}^{QMLE,\,k+1}_{i,j+1} &= \frac{1}{\sum_{t \in \mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \ddot{A}_{i+1,i+1,t}} \Bigg\{ \hat{G}_{i+1,j+r,s} - \sum_{t \in \mathscr{T}(s)} \Bigg[ \sum_{l_1=1}^{NK} \sum_{l_2=1}^{r-1} \hat{F}_{j+r,l_2,t}\, B_{l_1,l_2}\, \ddot{A}_{i+1,l_1,t} \\
&\qquad + \hat{F}_{j+r,r,t}\, \ddot{A}_{1,i+1,t} + \sum_{\substack{(l_1,l_2) \in \mathscr{U}_\Lambda \\ (l_1,l_2) \neq (i,j)}} \hat{F}_{j+r,l_2+r,t}\, \hat{\Lambda}^{\diamond}_{l_1,l_2+1} \left( \mathbb{I}_{l_1 \leq 7}\, \ddot{A}_{l_1+1,i+1,t} + \mathbb{I}_{l_1 > 7}\, \text{Л}_{l_1,i+1,t} \right) \Bigg] \Bigg\}
\end{aligned}$$

for $i = 1, \ldots, 7$, and

$$\hat{\Lambda}_{i,j+1}^{QMLE,\,k+1} = \frac{1}{\sum_{t \in \mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \text{Ш}_{i,i,t}} \Bigg\{ \hat{\text{Б}}_{i,j+r,s} - \sum_{t \in \mathscr{T}(s)} \Bigg[ \sum_{l_1=1}^{NK} \sum_{l_2=1}^{r-1} \hat{F}_{j+r,l_2,t}\, B_{l_1,l_2}\, \text{Л}_{i,l_1,t} \\ + \hat{F}_{r,j+r,t}\, \text{Л}_{i,1,t} + \sum_{\substack{(l_1,l_2) \in \mathscr{U}_\Lambda \\ (l_1,l_2) \neq (i,j)}} \hat{F}_{j+r,l_2+r,t}\, \hat{\Lambda}^{\diamond}_{l_1,l_2+1} \left( \mathbb{I}_{l_1 \leq 7}\, \text{Л}_{i,l_1+1,t} + \mathbb{I}_{l_1 > 7}\, \text{Ш}_{l_1,i,t} \right) \Bigg] \Bigg\}$$

for $i = 8, \ldots, 7 + |\mathscr{G}|$.

(iii) It follows directly from the results in step (i) and step (ii), and the proof of Pellegrino (2023b, Lemma 5) that, at a generic $k+1$ iteration of the ECM algorithm,

$$\hat{\Lambda}_{i,j+1}^{k+1}(\boldsymbol{\gamma}) = \begin{cases} \dfrac{\mathcal{S}\left[\hat{\Lambda}_{i,j+1}^{QMLE,\,k+1} \sum_{t \in \mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \ddot{A}_{i+1,i+1,t}\,,\ \frac{\varepsilon\alpha}{2} \Gamma_{j+1,j+1}(\boldsymbol{\gamma},p)\right]}{\varepsilon(1-\alpha)\Gamma_{j+1,j+1}(\boldsymbol{\gamma},p) + \sum_{t \in \mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \ddot{A}_{i+1,i+1,t}}, & \text{if } 1 \leq i \leq 7, \\ \dfrac{\mathcal{S}\left[\hat{\Lambda}_{i,j+1}^{QMLE,\,k+1} \sum_{t \in \mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \text{Ш}_{i,i,t}\,,\ \frac{\varepsilon\alpha}{2} \Gamma_{j+1,j+1}(\boldsymbol{\gamma},p)\right]}{\varepsilon(1-\alpha)\Gamma_{j+1,j+1}(\boldsymbol{\gamma},p) + \sum_{t \in \mathscr{T}(s)} \hat{F}_{j+r,j+r,t}\, \text{Ш}_{i,i,t}}, & \text{if } 8 \leq i \leq 7 + |\mathscr{G}|, \end{cases}$$

for $j = 0, \ldots, p-1$. □

## 3.B.3. Initialisation

First, we compute group averages for the microeconomic data. Then we apply the procedure described in Pellegrino (2023b, section A.3.) on both the macroeconomic indices and group averages. For simplicity, we set $\lambda = 2.573$, $\alpha = 0.667$ and $\beta = 1.326$. These are the optimal values in Pellegrino (2023b) converted for quarterly frequency data.

## 3.B.4. Enforcing causality during the estimation

We enforce causality during the estimation following Pellegrino (2023b, section A.4.).

## 3.C. Additional charts

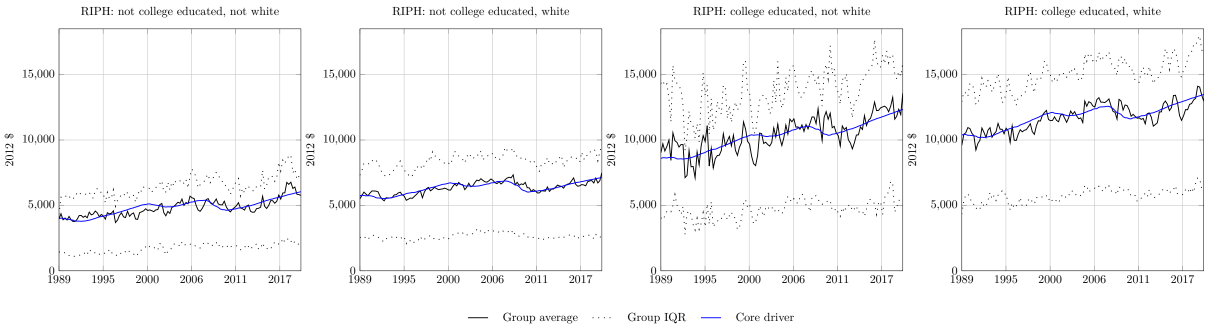


**Figure 3.C.1:** Core driver of RIPH computed as the sum of trend and business cycle.
**Notes**: The model is estimated with quarterly data from October 1989 to December 2019.

# Bibliography

S. Andradóttir. A review of random search methods. In *Handbook of Simulation Optimization*, pages 277–292. Springer, 2015.

J. Bai and S. Ng. Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29, 2013.

J. Bai and P. Wang. Identification and bayesian estimation of dynamic factor models. *Journal of Business & Economic Statistics*, 33(2):221–240, 2015.

M. Bańbura and M. Modugno. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, 29 (1):133–160, 2014.

M. Barigozzi and M. Luciani. Quasi maximum likelihood estimation and inference of large approximate dynamic factor models via the em algorithm. *arXiv preprint arXiv:1910.03821*, 2020.

C. Bergmeir, R. J. Hyndman, and B. Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

B. Bernanke, M. Gertler, and S. Gilchrist. The financial accelerator and the flight to quality. *Review of Economics and Statistics*, 78(1):1–15, 1996.

B. S. Bernanke, J. Boivin, and P. Eliasz. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly journal of economics*, 120(1):387–422, 2005.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees.* CRC press, 1984.

P. J. Brockwell, R. A. Davis, and S. E. Fienberg. *Time series: theory and methods: theory and methods*. Springer Science & Business Media, 1991.

R. L. Brown, J. Durbin, and J. M. Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 149–192, 1975.

P. Burman and D. Nolan. Data-dependent estimation of prediction functions. *Journal of Time Series Analysis*, 13(3):189–207, 1992.

P. Burman, E. Chow, and D. Nolan. A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358, 1994.

E. Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, pages 1171–1179, 1986.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

T. Doan, R. Litterman, and C. Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100, 1984.

C. Doz, D. Giannone, and L. Reichlin. A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of economics and statistics*, 94(4):1014–1024, 2012.

J. Durbin and S. J. Koopman. *Time series analysis by state space methods*. Oxford university press, 2012.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979a.

B. Efron. Computers and the theory of statistics: thinking the unthinkable. *SIAM review*, 21(4):460–480, 1979b.

B. Efron. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981.

B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

G. Elliott and A. Timmermann. *Economic forecasting*. Princeton university press, 2016.

M. Forni and M. Lippi. The generalized dynamic factor model: representation theory. *Econometric theory*, pages 1113–1141, 2001.

M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, 82(4):540–554, 2000.

M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, 100(471):830–840, 2005.

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

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

M. Gertler and S. Gilchrist. Monetary policy, business cycles, and the behavior of small manufacturing firms. *The Quarterly Journal of Economics*, 109(2):309–340, 1994.

J. F. Geweke. The dynamic factor analysis of economic time series model. *Latent variables in socio-economic models*, pages 365 – 383, 1977.

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

D. Giannone, M. Lenza, and G. E. Primiceri. Economic predictions with big data: The illusion of sparsity. *CEPR Discussion Paper No. DP12256*, 2017.

L. Györfi, W. Härdle, P. Sarda, and P. Vieu. *Nonparametric curve estimation from time series*. Springer, 1989.

A. Harvey, S. J. Koopman, and J. Penzer. Messy time series: a unified approach. *Advances in econometrics*, 13:103–144, 1998.

A. C. Harvey. Trends and cycles in macroeconomic time series. *Journal of Business & Economic Statistics*, 3(3):216–227, 1985.

A. C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge university press, 1990.

T. Hasenzagl, F. Pellegrino, L. Reichlin, and G. Ricco. A model of the fed's view on inflation. *The Review of Economics and Statistics*, 104(4):686–704, 2022a.

T. Hasenzagl, F. Pellegrino, L. Reichlin, and G. Ricco. Monitoring the economy in real time: Trends and gaps in real activity and prices. *arXiv preprint arXiv:2201.05556*, 2022b.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

M. Jarocinski and M. Lenza. Output gap and inflation forecasts in a bayesian dynamic factor model of the euro area. *manuscript, European Central Bank*, 2015.

G. Kitagawa and W. Gersch. *Smoothness priors analysis of time series*, volume 116. Springer Science & Business Media, 1996.

H. R. Kunsch. The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, pages 1217–1241, 1989.

R. M. Kunst. Cross validation of prediction models for seasonal time series by parametric bootstrapping. *Austrian journal of Statistics*, 37(3/4):271–284, 2008.

D. N. Lawley and A. E. Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229, 1962.

D. Li, J. Tosasukul, and W. Zhang. Nonlinear factor-augmented predictive regression models with functional coefficients. *Journal of Time Series Analysis*, 41(3):367–386, 2020.

Q. Li and N. Lin. The bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010.

R. B. Litterman. Forecasting with bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38, 1986.

C. Liu and D. B. Rubin. The ecme algorithm: a simple extension of em and ecm with faster monotone convergence. *Biometrika*, 81(4):633–648, 1994.

L. Liu and M. Plagborg-Møller. Full-information estimation of heterogeneous agent models using macro and micro data. *arXiv preprint arXiv:2101.04771*, 2021.

L. Ljung and T. Söderström. *Theory and practice of recursive identification.* MIT press, 1983.

H. Lütkepohl. *New introduction to multiple time series analysis.* Springer Science & Business Media, 2005.

M. Marcellino, J. H. Stock, and M. W. Watson. A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of econometrics*, 135(1-2):499–526, 2006.

R. A. Meese and K. Rogoff. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of international economics*, 14(1-2):3–24, 1983.

X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302):415–434, 1963.

S. Mullainathan and J. Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

C. Pardo, J. F. Diez-Pastor, C. García-Osorio, and J. J. Rodríguez. Rotation forests for regression. *Applied Mathematics and Computation*, 219(19):9914–9924, 2013.

F. Pellegrino. Selecting time-series hyperparameters with the artificial jackknife. *arXiv preprint arXiv:2002.04697*, 2023a.

F. Pellegrino. Factor-augmented tree ensembles. *arXiv preprint arXiv:2111.14000*, 2023b.

D. N. Politis and J. P. Romano. A circular block-resampling procedure for stationary data. *Exploring the limits of bootstrap*, pages 263–270, 1992.

D. N. Politis and J. P. Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.

M. H. Quenouille. Notes on bias in estimation. *Biometrika*, 1956.

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28 (10):1619–1630, 2006.

D. B. Rubin and D. T. Thayer. Em algorithms for ml factor analysis. *Psychometrika*, 47 (1):69–76, 1982.

S. Särkkä. *Bayesian filtering and smoothing.* Cambridge University Press, 2013.

J. Shao and C. J. Wu. A general theory for jackknife variance estimation. *The Annals of Statistics*, pages 1176–1197, 1989.

R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis*, 3(4):253–264, 1982.

R. H. Shumway and D. S. Stoffer. *Time series regression and exploratory data analysis.* Springer, 2011.

T. A. Snijders. On cross-validation for predictor evaluation in time series. *On Model Uncertainty and its Statistical Implications*, 307:56–69, 1988.

F. J. Solis and R. J.-B. Wets. Minimization by random search techniques. *Mathematics of operations research*, 6(1):19–30, 1981.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pages 111–147, 1974.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

J. W. Tukey. Bias and confidence in not-quite large samples. *The Annals of Mathematical Statistics*, 1958.

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

T. Weise. *Global optimization algorithms-theory and application.* Self-Published, 2009.

C.-F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, pages 1261–1295, 1986.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.