TI 2024-062/III
Tinbergen Institute Discussion Paper

Robust Multivariate Observation-Driven Filtering for a Common Stochastic Trend: Theory and Application

Francisco Blasques[1]
Janneke van Brummelen[1]
Paolo Gorgi[1]
Siem Jan Koopman[1]

1 Vrije Universiteit Amsterdam and Tinbergen Institute

Robust Multivariate Observation-Driven Filtering for a Common Stochastic Trend: Theory and Application[*]

Francisco Blasques[(a)] Janneke van Brummelen[(a)]
Paolo Gorgi[(a)] Siem Jan Koopman[(a)]

[(a)] Vrije Universiteit Amsterdam and Tinbergen Institute

September, 2024

Abstract

We introduce a nonlinear semi-parametric model that allows for the robust filtering of a common stochastic trend in a multivariate system of cointegrated time series. The observation-driven stochastic trend can be specified using flexible updating mechanisms. The model provides a general approach to obtain an outlier-robust trend-cycle decomposition in a cointegrated multivariate process. A simple two-stage procedure for the estimation of the parameters of the model is proposed. In the first stage, the loadings of the common trend are estimated via ordinary least squares. In the second stage, the other parameters are estimated via Gaussian quasi-maximum likelihood. We formally derive the theory for the consistency of the estimators in both stages and show that the observation-driven stochastic trend can also be consistently estimated. A simulation study illustrates how such robust methodology can enhance the filtering accuracy of the trend compared to a linear approach as considered in previous literature. The practical relevance of the method is shown by means of an application to spot prices of oil-related commodities.

Key words: consistency; cycle; non-stationary time series; two-step estimation; vector autoregression.

[*] Francisco Blasques acknowledges the financial support of the Dutch Science Foundation (NWO), The Netherlands under grant Vidi.195.099. Corresponding author: J. van Brummelen. E-mail: `j.n.van.brummelen@vu.nl`

1. Introduction

We introduce a semi-parametric model that can be used for outlier-robust filtering of the common stochastic trend of multiple time series variables from a cointegrated system. The most widely used method for modeling cointegrated time series is the vector error correction model (Johansen, 1995). In this framework, the unobserved common trend or trends that are driving the cointegrated time series can be constructed as a linear combination of past residuals. In case one or more time series under consideration contain outliers, the constructed trend will be severely impacted. This is undesirable, as the trend represents the long-term expectation of the time series. The trend should therefore not be impacted by temporary spikes. In a vector error correction model, the underlying common trends are not explicitly modeled and, hence, robustifying the trends against outliers is difficult; see also the discussions in Lucas (1997) and Franses and Lucas (1998).

We propose a nonlinear multivariate observation-driven common trend model which explicitly contains a dynamic specification for the common trend while it still can be formulated as a vector error correction model. The common trend specification enables us to straightforwardly robustify the trend filter. In particular, we propose to use a trend filter that allows for new information to impact the trend in a nonlinear manner, for instance, by means of quasi-score or flexible cubic spline functions that transform the residuals. We consider a semi-parametric specification, in the sense that we do not have to assume a particular distribution for the innovations of the model. To keep the asymptotic theory tractable and contained, we consider the setting where a single common trend drives k unit root time series, such that the cointegrating rank is $k-1$. The methodology can however be extended to allow for more common trends driving the time series.

Our proposed observation-driven common trend model also relates to other multiple time series models, including the common level model (Durbin and Koopman, 2012, Chapter 3.3.2) or the more general multivariate unobserved component models (Harvey, 1989, Chapter 8). Such models are not observation-driven but parameter-driven in the classification of Cox (1981), as the time-varying components are driven by their own independent innovations. A special case of the former type of model is the single latent common trend model introduced by Chang et al. (2009). This model can be shown to be equivalent to the linear Gaussian version of our model. Chang et al. (2009) demonstrate that their model, and therefore also our model, can be rewritten as an infinite-order vector error correction model. In the same way as for the more general state space models mentioned above, parameter estimation and filtering for this model rely on Kalman filter methods. However, in the case of non-Gaussian innovations or nonlinear updates, the Kalman filter is no longer applicable and thus it cannot deliver the log likelihood function. There exist different solutions for this, but they all require modifications which are rather involved, both conceptually and computationally. These results can therefore not be straightforwardly generalized in nonlinear settings. The proposed model in this paper is observation-driven, since the trend component is driven by past observations. A convenient consequence is that the (quasi) log likelihood can be constructed in closed-form and the filtering of the trend is simple, even if the model is nonlinear.

Our model enables the decomposition of a multivariate time series into a common trend and a transitory vector process with the aim to distinguish between long-term or permanent movements (trend) and temporary fluctuations around the trend (cycle). Such trend-cycle decompositions are, for example, used to judge whether the current value of a time series is above or below its forecasted long-term growth path. Common ways of extracting trends and cycles are via the canonical decomposition of autoregressive moving average (ARMA) models of Beveridge and Nelson (1981) or directly via unobserved components models (Harvey, 1985; Clark, 1987). Originally, such decompositions were mainly applied to univariate macro-economic time series variables at yearly or quarterly frequencies, such as gross domestic product. However, when applied to data observed at a higher frequency with a display of occasional erratic behaviour, it can be beneficial to use a robust filter for the long-term trend (Blasques et al., 2024b). Multivariate extensions of trend-cycle decompositions have earlier been considered (Ariño and Newbold, 1998; Murasawa, 2015) but not in the context where a single common trend is imposed or with an outlier-robust trend.

We consider a two-stage procedure for parameter estimation, similar to that of Engle and Granger (1987). In the first stage, the loadings of the common trend are estimated using ordinary least squares. In the second stage, the remaining parameters are estimated using Gaussian quasi-maximum likelihood (QML) with the first-stage estimator "plugged into" the quasi log likelihood. The first-stage estimator is shown to be super-consistent. The theoretical properties of the second-stage estimator rely on the invertibility of the nonlinear trend filter, for which we find a sufficient contraction condition, which can be feasibly verified in many settings of practical interest. Once filter invertibility is established, we show consistency of the second-stage QML estimator and we prove that the trajectory of the stochastic trend can be consistently extracted. The filtering ability of the model is explored in a Monte Carlo simulation study. This experiment demonstrates that in the presence of outliers, our robust methodology is able to filter an underlying trend more accurately than a linear trend filter. Finally, we present the results of an empirical study where a set of spot prices of oil commodities are analysed. This multivariate analysis of weekly energy prices includes the extraction of the common long-term trend which is of key importance for economic policy makers and investors alike.

The contributions of this paper are relative to the following earlier work. When the innovations of our common trend model are drawn from a certain distribution and we use the score corresponding to this distribution for the updating of the common trend, then our model resembles the score-driven model as advocated by Creal et al. (2013) and Harvey (2013). In particular, our model can be regarded as a multivariate generalisation of the univariate stationary score-driven conditional location model of Harvey and Luati (2014) and its non-stationary counterpart studied by Blasques et al. (2024a). Blazsek et al. (2021) also consider a related non-stationary score-driven model, but they do not thoroughly develop the asymptotic theory for maximum likelihood estimation. Łasak and Lont (2020) consider a fractional vector error correction model with a score-driven cointegration vector, variance and cointegration degree parameter, while the location (or trend) is not score-driven. Other models that bear resemblance to ours are score-driven dynamic factor models, such as those proposed by Creal et al. (2014) and Artemova (2023), but they do not allow for non-stationarity.

In light of the earlier work, our current study provides two key novelties. First, we extend the literature on non-stationary observation-driven conditional location models to a multivariate setting. The asymptotic theory of the (quasi) maximum likelihood estimator in the current context is more challenging than the univariate setting in Blasques et al. (2024a). In the univariate case, the (quasi) log likelihood contributions are stationary in the limit. This property does not hold for to the multivariate case, unless the (quasi) log likelihood is evaluated at the true loading vector. Hence, we cannot straightforwardly apply standard theorems to obtain consistency and asymptotic normality. Second, we provide a complete theoretical treatment of consistency for a nonlinear observation-driven conditional location model with both long-run and short-run dynamics. For instance, Blasques et al. (2024a) discuss a model with both long-run and short-run dynamics, but only present a complete theoretical treatment for the model with just long-run dynamics. In the model under consideration, short-run dynamics are introduced by allowing the innovations to be generated through a general vector autoregressive (VAR) process. This modeling framework enables us to derive theoretical results for a given set of conditions which can often be easily verified in empirical work.

The outline of the paper is as follows. In Section 2 we formally introduce the model and discuss its basic properties. In Section 3 we present the two-step estimation procedure, discuss filter invertibility, and give the asymptotic properties of the proposed estimators. In Section 4 we discuss the results of a Monte Carlo simulation study on the filtering performance of our model in various settings. In Section 5 we present the results of the empirical application. Section 6 concludes. The proofs of the propositions, corollaries and theorems can be found in the Supplementary Appendix.

2. Model specification

We consider the observable variable $\boldsymbol{y}_t = (y_{1,t}, \dots, y_{k,t})^\top$ to be a k-dimensional time series process given by

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{\beta} f_t + \boldsymbol{\varepsilon}_t\,, \qquad t \in \mathbb{N}, \tag{1}$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^\top$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^\top$ are k-dimensional vectors of unknown parameters, f_t is a non-stationary univariate stochastic trend and $\boldsymbol{\varepsilon}_t$ is the stationary component of the process. We consider a vector autoregressive specification of order p, VAR(p), for $\boldsymbol{\varepsilon}_t$ as given by

$$\boldsymbol{\varepsilon}_t = \boldsymbol{A}_1 \boldsymbol{\varepsilon}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{\varepsilon}_{t-p} + \boldsymbol{u}_t\,, \qquad \{\boldsymbol{u}_t\}_{t\in\mathbb{Z}} \sim (0, \boldsymbol{\Sigma})\,, \tag{2}$$

where $\boldsymbol{A}_j$, $j = 1, \dots, p$, are $k \times k$ autoregressive matrices with unknown coefficients and $\boldsymbol{u}_t$ is a k-dimensional error term, which is assumed to be a strictly stationary and ergodic (SE) martingale difference sequence (mds) with covariance matrix $\boldsymbol{\Sigma}$. For ease of notation, we also represent the VAR(p) process as $\boldsymbol{\varepsilon}_t = \boldsymbol{A}(L)^{-1}\,\boldsymbol{u}_t$, where $\boldsymbol{A}(L)$ denotes the lag polynomial $\boldsymbol{A}(L) = I - \boldsymbol{A}_1 L - \ldots - \boldsymbol{A}_p L^p$. We consider the following observation-driven specification for the stochastic trend f_t

$$f_{t+1} = \omega + f_t + \boldsymbol{\alpha}^\top s\left(\boldsymbol{A}(L)\left[\boldsymbol{y}_t - \boldsymbol{\mu} - \boldsymbol{\beta} f_t\right]; \boldsymbol{\psi}\right)\,, \;\; t \in \mathbb{N}, \tag{3}$$

where ω is a scalar drift, $\boldsymbol{\alpha}$ is k-dimensional coefficient vector that determines the impact of the innovation on the underlying trend, and $s(\cdot\,; \boldsymbol{\psi}) : \mathbb{R}^k \to \mathbb{R}^k$ is a known parametric function that is indexed by a q-dimensional parameter vector $\boldsymbol{\psi} \in \boldsymbol{\Psi} \subset \mathbb{R}^q$, which may contain elements of the parameter matrix $\boldsymbol{\Sigma}$. We assume that $\mathbb{E}[s(\boldsymbol{u}_t; \boldsymbol{\psi})|\mathcal{F}_{t-1}] = 0$, where $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ is the filtration of sigma algebras $\mathcal{F}_t = \sigma(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \dots, \boldsymbol{y}_1)$. The initial value of the process $\{f_t\}$ is an unknown value f_1, which can be treated as a real value.

From (1)-(3), it follows that the stochastic trend f_t is a random walk with drift ω and SE mds innovations $\boldsymbol{\alpha}^\top s(\boldsymbol{u}_t; \boldsymbol{\psi})$. Therefore, the elements of the observable process $\boldsymbol{y}_t$ are integrated of order one, where the non-stationarity arises from the single common factor f_t. If the function $s(\cdot\,;\cdot)$ is linear in its first argument, then it can be shown that $\{\Delta \boldsymbol{y}_t\}$ is a restricted vector ARMA$(p, p+1)$ process. The one-step-ahead forecast function is given by $\mathbb{E}(\boldsymbol{y}_t|\mathcal{F}_{t-1}) = \boldsymbol{A}(1)\boldsymbol{\mu} + \boldsymbol{A}(L)\boldsymbol{\beta} f_t + \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p}$. The stationarity of the VAR process $\{\boldsymbol{\varepsilon}_t\}$ entails that the long-term expectation of $\boldsymbol{y}_t$ at time $t-1$ is

$$\lim_{T\to\infty} \mathbb{E}(\boldsymbol{y}_{t+T}|\mathcal{F}_{t-1}) - T\omega\boldsymbol{\beta} = \boldsymbol{\mu} + \boldsymbol{\beta} f_t\,.$$

A convenient aspect of the model is this immediate expression, $\boldsymbol{\mu} + \boldsymbol{\beta} f_t$, of the permanent component in a Beveridge-Nelson decomposition (Beveridge and Nelson, 1981). This trend-cycle representation is valid in a univariate setting as discussed in Blasques et al. (2024b), but can straightforwardly be extended to the current multivariate setting.

The stochastic trend f_t and the parameter vectors $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ are not uniquely identified without further restrictions. For example, by taking a shift and scale transformation of f_t and a corresponding transformation of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$, we obtain an observationally equivalent process $\boldsymbol{y}_t$. In order to ensure identification, we normalize the first elements of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ to $\mu_1 = 0$ and $\beta_1 = 1$. This choice of the identification restriction entails that the long-term expectation of the first element of $\boldsymbol{y}_t$ is f_t. Alternative identification restrictions may also be considered. For instance, one may restrict the elements of the vector $\boldsymbol{\mu}$ to sum to 0 and those of $\boldsymbol{\beta}$ to sum to k, which would lead to a different interpretation of the stochastic trend f_t. From the model specification, it follows that if $\boldsymbol{\beta}$ does not contain zeros, $\{\boldsymbol{y}_t\}$ is a cointegrated process with cointegrating rank $k-1$. Under the current identification restriction, the cointegration vectors are $\tilde{\boldsymbol{\beta}}_i = (-\beta_{i+1}, e_{i,k-1}^\top)^\top$ for $i = 1, \dots, k-1$, where $e_{i,k}$ denotes a k-dimensional unit vector with a one at index i, because it is clear that $\tilde{\boldsymbol{\beta}}_i^\top \boldsymbol{\beta} = 0$ and therefore $\tilde{\boldsymbol{\beta}}_i^\top \boldsymbol{y}_t = \tilde{\boldsymbol{\beta}}_i^\top \boldsymbol{\mu} + \tilde{\boldsymbol{\beta}}_i^\top \boldsymbol{\varepsilon}_t$ is integrated of order zero.

The parametric function $s(\cdot\,; \boldsymbol{\psi})$ in (3) determines the updating mechanism of f_t, and can be chosen freely by the researcher under some restrictions that shall be discussed in the next section. The most straightforward choice is a linear function $s(\boldsymbol{u}; \boldsymbol{\psi}) = \boldsymbol{u}$. In many settings, a nonlinear function is more suitable. For instance, when a set of multivariate observations displays outliers after which the time series reverts back to a value close to the pre-outlier level, indicating that the underlying long-term conditional expectation f_t has not changed much. Such a situation can be described by the model in (1)-(3) and having a fat-tailed distribution for $\boldsymbol{u}_t$ and a nonlinear updating function $s(\cdot\,; \boldsymbol{\psi})$ that gives relatively more weight to moderate innovation values than large innovation values. In other words, a function $s(\boldsymbol{u}; \boldsymbol{\psi})$ that has a higher slope for small values than for large values of $\boldsymbol{u}$. This gives updates of f_t that are robust against outliers compared to linear updates. Within our model, a large value of $\boldsymbol{u}_t$ can still have a moderately large impact on the observations for some time, depending on how persistent the VAR(p) process $\{\boldsymbol{\varepsilon}_t\}$ is. However, in the limit the effect on $\boldsymbol{\varepsilon}_t$ will die out. On the other hand, for the long-term component f_t, an outlier in $\boldsymbol{\alpha}^\top s(\boldsymbol{u}_t; \boldsymbol{\psi})$ will have an everlasting effect. Therefore, from a theoretical perspective, it may be sensible to choose a nonlinear function $s(\cdot\,; \boldsymbol{\psi})$ that limits the effect of large values of $\boldsymbol{u}_t$.

A practical way to select $s(\cdot\,; \boldsymbol{\psi})$ is to consider a flexible parametric function. A convenient choice is a piecewise polynomial specification such as a natural cubic spline, which will also be considered in the Monte Carlo study and the empirical application. An alternative choice is to employ a quasi-score function, i.e. to consider the score function of a flexible probability distribution in the spirit of the score-driven framework of Creal et al. (2013) and Harvey (2013). For instance, a flexible specification that can handle extreme observations is the score of a Gaussian mixture as considered in Blasques et al. (2024a).

This option is also considered in the Monte Carlo study and empirical application. We note that the resulting model is not a score-driven model as there is no distributional assumption made on the error $\boldsymbol{u}_t$. Instead, it follows the quasi score-driven approach of Blasques et al. (2023) where the score is used to specify the dynamics of the model in a semi-parametric setting.

3. Estimation of the model parameters

We partition the parameter vector of the model into two sub-vectors $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\xi}^\top)^\top$, which will be separately estimated in two stages. The vector $\boldsymbol{\gamma} \in \Gamma$ contains the first-stage parameters

$$\boldsymbol{\gamma} = (\boldsymbol{b}^\top, \boldsymbol{m}^\top)^\top,$$

where $\boldsymbol{b} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_k)^\top$ and $\boldsymbol{m} = (\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k)^\top$; and the vector $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ contains the second-stage parameters

$$\boldsymbol{\xi} = (\omega\ , \boldsymbol{\alpha}^\top, \boldsymbol{\psi}_{-1}^\top, \mathrm{vec}(\boldsymbol{A}_1)^\top, \ldots, \mathrm{vec}(\boldsymbol{A}_p)^\top, \mathrm{vech}(\boldsymbol{\Sigma})^\top)^\top,$$

where $\boldsymbol{\psi}_{-1} \in \mathbb{R}^{q^*}$ contains the elements of $\boldsymbol{\psi}$ that are not in $\boldsymbol{\Sigma}$. The vector $\boldsymbol{\theta}$ takes values in some parameter space $\Theta = \Gamma \times \boldsymbol{\Xi} \subset \mathbb{R}^{v_1} \times \mathbb{R}^{v_2}$ with $v_1 = 2(k-1)$ and $v_2 = 1 + k + q^* + k^2 p + k(k+1)/2$. In particular cases, it suffices to consider a subset of this vector. For instance, one may fix $\omega = 0$ or $\boldsymbol{m} = 0$ and remove them from the parameter vector when it is assumed that there is no drift in the stochastic trend or that there are no shift transformations. Furthermore, in empirical work, the VAR dynamics in $\{\boldsymbol{\varepsilon}_t\}$ can often be assumed idiosyncratic. Then, all the matrices $\boldsymbol{A}_j$, for $j = 1, \ldots, p$, may be restricted to be diagonal or scalar. The discussions and results presented below remain applicable for all such restricted cases. However, the specific restrictions $\omega = 0$ and $\omega \neq 0$ need to be treated separately as they affect the asymptotic properties of estimators.

Assume that a k-variate sequence of observations $\{\boldsymbol{y}_t\}_{t\in\mathbb{N}}$ is generated by the model in (1)-(3) based on some true parameter value $\boldsymbol{\theta}_0$, and suppose that we observe T consecutive observations from this sequence; i.e. $\{\boldsymbol{y}_t\}_{t=1}^T$. We are interested in estimating the true parameter $\boldsymbol{\theta}_0$. We propose to estimate the parameter vector in two stages in the spirit of the two-step procedure of Engle and Granger (1987). Our approach relies on a regression step and a quasi-likelihood step that delivers a simple estimation approach with tractable asymptotic properties.

In the first stage, we estimate $\boldsymbol{b}$ and $\boldsymbol{m}$ by regressing $\tilde{\boldsymbol{y}}_t = (y_{2t}, \ldots, y_{kt})^\top$ on y_{1t}. In particular, we can write

$$\tilde{\boldsymbol{y}}_t = \boldsymbol{m}_0 + \boldsymbol{b}_0\ y_{1t} + \boldsymbol{v}_t, \tag{4}$$

where $\boldsymbol{v}_t = \tilde{\boldsymbol{\varepsilon}}_t - \boldsymbol{b}_0 \varepsilon_{1t}$, $\tilde{\boldsymbol{\varepsilon}}_t = (\varepsilon_{2t}, \ldots, \varepsilon_{kt})^\top$, is a stationary process with mean zero, whereas y_{1t} has a unit root. This regression will therefore lead to a super-consistent estimator $\widehat{\boldsymbol{b}}_T$

and a consistent estimator $\widehat{\boldsymbol{m}}_T$. In the second stage, we estimate the parameter vector $\boldsymbol{\xi}$ by Gaussian quasi-maximum likelihood (QML) with $\widehat{\boldsymbol{\beta}}_T = (1, \widehat{\boldsymbol{b}}_T^\top)^\top$ and $\widehat{\boldsymbol{\mu}}_T = (0, \widehat{\boldsymbol{m}}_T^\top)^\top$ plugged into the equation of the quasi log likelihood. Since the sequence $\{f_t\}_{t=1}^T$ is not directly observable, the construction of the quasi log likelihood relies on the filtered sequence $\{\hat{f}_t(\widehat{\boldsymbol{\gamma}}_T, \boldsymbol{\xi})\}_{t=1}^T$, with $\widehat{\boldsymbol{\gamma}}_T = (\widehat{\boldsymbol{m}}_T^\top, \widehat{\boldsymbol{b}}_T^\top)^\top$, where for a general $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$, $\hat{f}_t(\boldsymbol{\gamma}, \boldsymbol{\xi})$ can be calculated using the updating equation

$$\hat{f}_{t+1}(\boldsymbol{\gamma}, \boldsymbol{\xi}) = \omega + \hat{f}_t(\boldsymbol{\gamma}, \boldsymbol{\xi}) + \boldsymbol{\alpha}^\top s(\boldsymbol{A}(L)[\boldsymbol{y}_t - \boldsymbol{\mu} - \boldsymbol{\beta}\hat{f}_t(\boldsymbol{\gamma}, \boldsymbol{\xi})]; \boldsymbol{\psi}),$$

where we set $(\boldsymbol{y}_0 - \boldsymbol{\mu} - \boldsymbol{\beta}\hat{f}_0), \ldots, (\boldsymbol{y}_{-p+1} - \boldsymbol{\mu} - \boldsymbol{\beta}\hat{f}_{-p+1})$ to zero and select a particular starting value for $\hat{f}_1$, such as the value of the first observation, i.e. $\hat{f}_1 = y_{1,1}$. Whenever it is convenient, we use the notation $\hat{f}_t(\boldsymbol{\theta}) \equiv \hat{f}_t(\boldsymbol{\gamma}, \boldsymbol{\xi})$. The QML estimator $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$ is defined as

$$\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T) = \arg\max_{\boldsymbol{\xi} \in \boldsymbol{\Xi}} \widehat{L}_T(\widehat{\boldsymbol{\gamma}}_T, \boldsymbol{\xi}),$$

where $\widehat{L}_T(\widehat{\boldsymbol{\gamma}}_T, \boldsymbol{\xi}) = \frac{1}{T}\sum_{t=p+1}^T \hat{\ell}_t(\widehat{\boldsymbol{\gamma}}_T, \boldsymbol{\xi})$ and

$$\hat{\ell}_t(\widehat{\boldsymbol{\gamma}}_T, \boldsymbol{\xi}) = -[\boldsymbol{A}(L)(\boldsymbol{y}_t - \widehat{\boldsymbol{\mu}}_T - \widehat{\boldsymbol{\beta}}_T\hat{f}_t(\widehat{\boldsymbol{\gamma}}_T, \boldsymbol{\xi}))]^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{A}(L)(\boldsymbol{y}_t - \widehat{\boldsymbol{\mu}}_T - \widehat{\boldsymbol{\beta}}_T\hat{f}_t(\widehat{\boldsymbol{\gamma}}_T, \boldsymbol{\xi}))] - \log|\boldsymbol{\Sigma}|.$$

Below we formally discuss the asymptotic properties of the first-stage estimator of the long-run parameter $\boldsymbol{\gamma}$ in Section 3.1 and of the second-stage estimator of the parameter $\boldsymbol{\xi}$ in Section 3.2.

3.1. Asymptotic properties of first stage estimates

We impose the following regularity conditions to establish the consistency and rate of convergence of the first-stage estimator.

A1 $\{\boldsymbol{y}_t\}_{t=1}^T$ is generated by the model in (1)-(3) with $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \in \Theta$ and $\{\boldsymbol{u}_t\}_{t\in\mathbb{Z}}$ is an SE mds with a finite and positive definite covariance matrix.

A2 All solutions λ of the characteristic equation $|\boldsymbol{A}_0(\lambda)| = 0$ are outside the unit circle.

A3 $\{s(\boldsymbol{u}_t; \boldsymbol{\psi}_0)\}$ is an mds with finite second moment, i.e. $\mathbb{E}\|s(\boldsymbol{u}_t; \boldsymbol{\psi}_0)\|^2 < \infty$.

Condition **A1** ensures that the model is correctly specified. The innovations $\{\boldsymbol{u}_t\}_{t\in\mathbb{Z}}$ are assumed to be an mds. Therefore, they must be uncorrelated but are not required to be independent, allowing for instance GARCH dynamics. Condition **A2** ensures that the process $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}}$ is SE. Condition **A3** is imposed to apply a (functional) central limit theorem to f_t that is used to establish the consistency of the first-stage estimator. A sufficient condition for **A3** is that $\boldsymbol{u}_t$ conditional on $\mathcal{F}_{t-1}$ is symmetrically distributed, and $s(\boldsymbol{x}; \boldsymbol{\psi}_0)$ is an odd function in $\boldsymbol{x}$. Another sufficient condition is that $\{\boldsymbol{u}_t\}$ is an independent sequence of random variables with a zero unconditional mean for $s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)$. We also note that **A3** can be relaxed to $\mathbb{E}[s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)|\mathcal{F}_{t-1}]$ being a constant, but in that case ω is no longer identifiable as the drift. In such a case, subtracting $\mathbb{E}[s(\boldsymbol{u}_t;\boldsymbol{\psi})]$ from $s(\boldsymbol{u}_t;\boldsymbol{\psi})$ in the updating equation in (3), leads to this condition being satisfied.

The next theorem provides the rate of convergence of the ordinary least squares (OLS) estimators $\widehat{\boldsymbol{b}}_T$ and $\widehat{\boldsymbol{m}}_T$ of $\boldsymbol{b}_0$ and $\boldsymbol{m}_0$, respectively, in regression model (4).

Theorem 1. *Let **A1**-**A3** hold. Then the first-stage OLS estimators $\widehat{\boldsymbol{b}}_T$ and $\widehat{\boldsymbol{m}}_T$ satisfy the following properties as $T \to \infty$:*

(i) If $\omega_0 = 0$, then $T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) = O_p(1)$ and $T^{1/2}(\widehat{\boldsymbol{m}}_T - \boldsymbol{m}_0) = O_p(1)$.

(ii) If $\omega_0 \neq 0$, then $T^{3/2}(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) = O_p(1)$ and $T^{1/2}(\widehat{\boldsymbol{m}}_T - \boldsymbol{m}_0) = O_p(1)$.

These results follow from standard theory of regressions of integrated processes. The estimator of $\boldsymbol{b}_0$ is super-consistent and its rate of convergence depends on the drift parameter ω_0. We refer the reader to Section D of the Supplementary Appendix for the asymptotic distribution of this estimator. The theory of the second-stage estimator discussed in the next section only relies on the rate of consistency of the first-stage estimators of $\boldsymbol{b}_0$ and $\boldsymbol{m}_0$. Therefore, alternative estimators that meet this consistency requirement can also be considered as an alternative first-stage estimator. In general, for instance, it can be beneficial to take into account the serial dependence in the innovations $\boldsymbol{v}_{-1,t}$ of regression model (4) to obtain a more efficient estimator. Furthermore, even if $\boldsymbol{v}_{-1,t}$ is uncorrelated over time (i.e. if $p = 0$), $\boldsymbol{v}_t$ is correlated with y_{1t}. Due to this endogeneity, the OLS estimator will be inefficient. In Section D of the Supplementary Appendix we propose a modified OLS estimator as an efficient alternative. As additive outliers could distort the OLS estimates of $\boldsymbol{b}_0$ in finite samples, another appealing option could be to use a robust estimator for $\boldsymbol{b}_0$ instead of (modified) OLS. A possible alternative is for instance the fully modified least absolute deviations (FM-LAD) estimator, as suggested by Phillips (1995). Other relevant references in the context of cointegration analysis in the presence of outliers are for instance Lucas (1997) and Franses and Lucas (1998).

3.2. Asymptotic properties of second stage estimates

3.2.1. Filter invertibility

We turn to the estimation of the short-run parameters $\boldsymbol{\xi}$. As pointed out above, we need to use the sequence of filtered values $\hat{f}_t(\boldsymbol{\theta})$ in the construction of the quasi log likelihood, as the true f_t is unobserved. Since we do not know the true starting value f_1, it is of key importance that the filter is invertible, which means that the filtered sequence $\{\hat{f}_t(\boldsymbol{\theta})\}_{t=1}^T$ initialized at some $\hat{f}_1 \in \mathbb{R}$ "forgets" its starting value in the limit. Typically, invertibility is discussed in a stationary context, where the invertibility of the filter also entails the convergence of the filtered sequence to an SE sequence, e.g. in Wintenberger (2013), Blasques et al. (2018), and Blasques et al. (2022). In the current setting, this will not be the case as $\{\boldsymbol{y}_t\}$ is integrated of order 1. However, for $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, i.e. in $\boldsymbol{b} = \boldsymbol{b}_0$ and $\boldsymbol{m} = \boldsymbol{m}_0$, we will show that the differences between the true value of f_t and the filtered value $\hat{f}_t(\boldsymbol{\gamma}, \boldsymbol{\xi})$ converge to an SE sequence under a contraction condition. This will be used to show consistency of the plug-in estimator $\widehat{\boldsymbol{\xi}}_T(\boldsymbol{\gamma}_0)$. In fact, this result holds whenever $\boldsymbol{b} = \boldsymbol{b}_0$, even if $\boldsymbol{m} \neq \boldsymbol{m}_0$, but it is more convenient to immediately consider $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. Also, for $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}_0$ we can still formulate a condition for invertibility of the filter, which is also crucial in our derivation of consistency of the plug-in second-stage QMLE $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$.

We start by considering the filtering error in $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. Define the notation $\hat{g}_t(\boldsymbol{\xi}) \equiv f_t - \hat{f}_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi})$. The sequence $\{\hat{g}_t(\boldsymbol{\xi})\}_{t=1}^T$ follows the stochastic recurrence equation (SRE):

$$\hat{g}_{t+1}(\boldsymbol{\xi}) = \omega_0 - \omega + \hat{g}_t(\boldsymbol{\xi}) + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0) - \boldsymbol{\alpha}^\top s(\boldsymbol{A}(L)[\boldsymbol{\beta}_0 \hat{g}_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t]; \boldsymbol{\psi}),$$

initialized at some value $\hat{g}_1 = f_1 - \hat{f}_1$. Hence, the filtering errors for $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ follow an SRE with SE innovations, as $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}}$ is SE by **A2**. Therefore, we can show that $\{\hat{g}_t(\boldsymbol{\xi})\}_{t\in\mathbb{N}}$ converges exponentially fast almost surely[†] (e.a.s.) to an SE limit process under a contraction condition. We rewrite the SRE in vector form to obtain a first order dynamical system. We do so by defining $\hat{\boldsymbol{g}}_t(\boldsymbol{\xi}) = (\hat{g}_t(\boldsymbol{\xi}), \hat{g}_{t-1}(\boldsymbol{\xi}), \dots, \hat{g}_{t-p}(\boldsymbol{\xi}))^\top$, which follows the SRE:

$$\hat{\boldsymbol{g}}_{t+1}(\boldsymbol{\xi}) = \phi_t(\hat{\boldsymbol{g}}_t(\boldsymbol{\xi}), \boldsymbol{\xi}),$$

initialized at $\hat{\boldsymbol{g}}_1 = (f_1 - \hat{f}_1, 0, \dots, 0)^\top$ and where ϕ_t is a random function $\phi_t : \mathbb{R}^p \times \boldsymbol{\Xi} \to \mathbb{R}^p$, defined by

$$\phi_t(\boldsymbol{g}, \boldsymbol{\xi}) = \begin{pmatrix} \phi_{1t}(\boldsymbol{g}, \boldsymbol{\xi}) \\ g_1 \\ \vdots \\ g_p \end{pmatrix},$$

with $\boldsymbol{g} = (g_1, \dots, g_{p+1})^\top$ and

$$\begin{aligned} \phi_{1t}(\boldsymbol{g}, \boldsymbol{\xi}) = \omega_0 - \omega + g_1 + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0) \\ - \boldsymbol{\alpha}^\top s(\boldsymbol{A}(L)\boldsymbol{\varepsilon}_t + \boldsymbol{\beta}_0 g_1 - \boldsymbol{A}_1 \boldsymbol{\beta}_0 g_2 - \ldots - \boldsymbol{A}_p \boldsymbol{\beta}_0 g_{p+1}; \boldsymbol{\psi}). \end{aligned} \tag{5}$$

Let $\phi_t^{(r)}(\cdot, \boldsymbol{\xi})$ denote the r-th convolution of ϕ_t, i.e. $\phi_t^{(r)}(\cdot, \boldsymbol{\xi}) = \phi_t(\cdot, \boldsymbol{\xi}) \circ \ldots \circ \phi_{t-r+1}(\cdot, \boldsymbol{\xi})$. Under the following regularity conditions, and an additional contraction condition, we

[†] We say some sequence of random variables $\{\hat{x}_t\}$ converges e.a.s. to another sequence $\{x_t\}$ if there exists a constant $\rho > 1$ such that $\rho^t |\hat{x}_t - x_t| \overset{a.s.}{\to} 0$ as $t \to \infty$.

can establish invertibility of the filter. We denote by $\|\cdot\|$ the L^p-norm, $\|\boldsymbol{x}\| = (|x_1|^p + \ldots + |x_n|^p)^{1/p}$, for some $p \geq 1$ when applied to some n-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, and the operator norm induced by the L^p-norm when applied to a matrix.

IN1 Conditions A1-A2 are satisfied.

IN2 The function $s(\boldsymbol{x};\boldsymbol{\psi})$ satisfies the following conditions:
(i) $\boldsymbol{\psi} \mapsto s(\boldsymbol{x}\ ;\boldsymbol{\psi})$ is continuous for any $\boldsymbol{x} \in \mathbb{R}^k$.
(ii) $\boldsymbol{x} \mapsto s(\boldsymbol{x}\ ;\boldsymbol{\psi})$ is differentiable for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}$.
(iii) $\boldsymbol{x} \mapsto s(\boldsymbol{x};\boldsymbol{\psi})$ is Lipschitz continuous uniformly over $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, i.e. there is a $K < \infty$ such that $\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \|s(\boldsymbol{x};\boldsymbol{\psi}) - s(\boldsymbol{x}^*;\boldsymbol{\psi})\| \leq K\|\boldsymbol{x} - \boldsymbol{x}^*\|$ for any $\boldsymbol{x}, \boldsymbol{x}^* \in \mathbb{R}^k$.

IN3 The parameter set Θ is compact and $\boldsymbol{\Sigma}$ is positive definite for any $\boldsymbol{\theta} \in \Theta$.

Condition IN2 contains regularity conditions on the function s which will be used to derive the properties of the $\{\hat{g}_t(\boldsymbol{\xi})\}$ process, where in particular the Lipschitz condition is used to derive bounded moments of the elements of the limit process. Condition IN3 is a standard condition that is helpful in the derivation of uniform results.

The proposition below establishes convergence of $\{\hat{g}_t(\boldsymbol{\xi})\}_{t\in\mathbb{N}}$ to an SE sequence with two bounded moments under conditions IN1-IN3 and a contraction condition. This in turn implies invertibility of the filtered location $\hat{f}_t(\boldsymbol{\theta})$ evaluated in $\boldsymbol{\beta}_0$. The proof of the asymptotic stationarity result uses Straumann and Mikosch (2006, Theorem 2.8), which is based on Bougerol (1993, Theorem 3.1). This proposition is a more general version of Proposition 3.2 of Blasques et al. (2022), because here we consider a filter with higher-order dependence.

Proposition 1. *Let conditions IN1-IN3 be satisfied. Then if for some $\bar{\boldsymbol{g}} \in \mathbb{R}^{p+1}$ there is an integer $r \geq 1$ such that:*

$$\sup_{\substack{\boldsymbol{\xi}\in\Xi, \boldsymbol{g}\in\mathbb{R}^{p+1},\\ \boldsymbol{\varepsilon}_t,\ldots,\boldsymbol{\varepsilon}_{t-p-r+1}\in\mathbb{R}^k}} \left\| \frac{\partial \phi_t^{(r)}(\boldsymbol{g},\boldsymbol{\xi})}{\partial \boldsymbol{g}^\top} \right\| = \sup_{\substack{\boldsymbol{\xi}\in\Xi,\\ \boldsymbol{z}_1,\ldots,\boldsymbol{z}_r\in\mathbb{R}^k}} \left\| \prod_{i=1}^{r} \Phi(\boldsymbol{z}_i, \boldsymbol{\gamma}_0, \boldsymbol{\xi}) \right\| < 1\,, \tag{6}$$

where

$$\Phi(\boldsymbol{z},\boldsymbol{\gamma},\boldsymbol{\xi}) = \begin{pmatrix} 1 - \boldsymbol{\alpha}^\top s'(\boldsymbol{z};\boldsymbol{\psi})\boldsymbol{\beta} & \boldsymbol{\alpha}^\top s'(\boldsymbol{z};\boldsymbol{\psi})\boldsymbol{A}_1\boldsymbol{\beta} & \ldots & \boldsymbol{\alpha}^\top s'(\boldsymbol{z};\boldsymbol{\psi})\boldsymbol{A}_p\boldsymbol{\beta} \\ & & & 0 \\ & I_p & & \vdots \\ & & & 0 \end{pmatrix}, \tag{7}$$

with $s'(\boldsymbol{z};\boldsymbol{\psi}) = \partial s(\boldsymbol{z};\boldsymbol{\psi})/\partial \boldsymbol{z}$, then

(i) $\hat{g}_t(\boldsymbol{\xi})$ converges e.a.s. to a unique SE sequence of random variables $\{g_t(\boldsymbol{\xi})\}_{t\in\mathbb{Z}}$ uniformly over Ξ.

(ii) The filter evaluated in $\boldsymbol{\theta}_0$ converges e.a.s. to the true f_t: $|\hat{f}_t(\boldsymbol{\theta}_0) - f_t| \overset{e.a.s.}{\to} 0$ as $t \to \infty$.

(iii) $g_t(\boldsymbol{\xi})$ has two bounded moments uniformly over Ξ: $\mathbb{E}\sup_{\boldsymbol{\xi}\in\Xi} |g_t(\boldsymbol{\xi})|^2 < \infty$.

Given the companion form of the matrix $\Phi(\boldsymbol{z}, \boldsymbol{\gamma}_0, \boldsymbol{\xi})$, at least $r = p + 1$ iterations are needed for these contraction conditions to hold. When $s(\cdot\, ; \boldsymbol{\psi})$ is linear, $\Phi(\boldsymbol{z}, \boldsymbol{\gamma}_0, \boldsymbol{\xi})$ is a deterministic matrix, so then the contraction condition will hold for some r, if and only if the spectral radius of this matrix is smaller than 1 uniformly over the parameters. When $s(\cdot\, ; \boldsymbol{\psi})$ is nonlinear and $p \geq 1$, then it is less straightforward to verify whether the contraction condition of this proposition holds for a given r. We will discuss some simple sufficient conditions for (6) below Proposition 2. Notice that if $p = 0$, we have the simple contraction condition $\sup_{\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\psi}} |1 - \boldsymbol{\alpha}^\top s'(\boldsymbol{z}, \boldsymbol{\psi})\boldsymbol{\beta}_0| < 1$, which is easy to verify.

For $\boldsymbol{b} \neq \boldsymbol{b}_0$, the filtering errors $\{f_t - \hat{f}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ will not converge to an SE limit sequence, because then the unit root does not cancel out in the filtering updates. However, we can still establish invertibility of the filter under a stricter contraction condition. The proof of the following proposition does not rely on the correct specification of the model, in particular it holds regardless of the properties of $\{\boldsymbol{y}_t\}_{t\in\mathbb{N}}$.

Proposition 2. *Let conditions* ***IN1**-**IN3*** *be satisfied. Also, let the following condition be satisfied for some integer $r \geq 1$:*

$$\sup_{\substack{\boldsymbol{\theta}\in\Theta,\\ \boldsymbol{z}_1,\ldots,\boldsymbol{z}_r\in\mathbb{R}^k}} \left\| \prod_{i=1}^{r} \Phi(\boldsymbol{z}_i, \boldsymbol{\gamma}, \boldsymbol{\xi}) \right\| < 1\,, \tag{8}$$

where the function Φ is defined in (7), then the filter $\hat{f}_t$ is uniformly invertible over Θ, meaning that if $\{\tilde{f}_t\}_{t\in\mathbb{N}}$ is a filtered sequence initialized at some alternative starting value $\tilde{f}_1 \in \mathbb{R}$, then $\sup_{\boldsymbol{\theta}\in\Theta} |\hat{f}_t(\boldsymbol{\theta}) - \tilde{f}_t(\boldsymbol{\theta})| \overset{e.a.s.}{\to} 0$ as $t \to \infty$, for any $\hat{f}_1, \tilde{f}_1 \in \mathbb{R}$.

Clearly, Condition (8) is simply a stronger version of Condition (6) in Proposition 1, as we take the supremum over $\boldsymbol{\theta} \in \Theta$, instead of fixing $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ and just taking the supremum over $\boldsymbol{\xi} \in \Xi$. It can be verified that we must have $\inf_{\boldsymbol{z}\in\mathbb{R}^k,\boldsymbol{\theta}\in\Theta} \boldsymbol{\alpha}^\top s'(\boldsymbol{z}; \boldsymbol{\psi})\boldsymbol{A}(1)\boldsymbol{\beta} > 0$ for condition (8) to hold. So if $s'(\boldsymbol{z}; \boldsymbol{\psi})$ is diagonal and $p = 0$, then it follows that the condition can only hold if at least one sub-function s_i is monotonically increasing or decreasing in $\boldsymbol{z}$. As we will typically use the same function for each index $i = 1, \ldots, k$, it follows that this function must be monotone for the proposition to apply.

The corollary below gives a simple sufficient condition for the contraction condition of Proposition 2 to hold.

Corollary 1. *A sufficient condition for condition (8) of Proposition 2 to be satisfied for $r = p+1$ for the matrix norm induced by the L^∞-norm is:*

$$\sup_{\boldsymbol{z}\in\mathbb{R}^k,\boldsymbol{\theta}\in\Theta} \left\{|1-\boldsymbol{\alpha}^\top s'(\boldsymbol{z};\boldsymbol{\psi})\boldsymbol{\beta}| + |\boldsymbol{\alpha}^\top s'(\boldsymbol{z};\boldsymbol{\psi})\boldsymbol{A}_1\boldsymbol{\beta}| + \cdots + |\boldsymbol{\alpha}^\top s'(\boldsymbol{z};\boldsymbol{\psi})\boldsymbol{A}_p\boldsymbol{\beta}|\right\} < 1\,, \tag{9}$$

where $s'(\boldsymbol{z};\boldsymbol{\psi}) = \partial s(\boldsymbol{z};\boldsymbol{\psi})/\partial\boldsymbol{z}^\top$.

An even simpler sufficient condition to verify is the condition in the corollary below.

Corollary 2. *For each $\boldsymbol{\theta}\in\Theta$, let $s'(\boldsymbol{z};\boldsymbol{\psi})$ be a diagonal matrix and let the coefficient matrices $\boldsymbol{A}_i$ be diagonal, with $\sum_{i=1}^p |\boldsymbol{A}_{i,jj}| < 1$ for $j = 1,\ldots,k$, where $\boldsymbol{A}_{i,jj}$ denotes the j-th diagonal element of $\boldsymbol{A}_i$. Then, for condition (9) to hold it is sufficient that*

$$\inf_{\boldsymbol{z}\in\mathbb{R}^k,\boldsymbol{\theta}\in\Theta} \boldsymbol{\alpha}^\top s'(\boldsymbol{z};\boldsymbol{\psi})\boldsymbol{\beta} > 0\,, \qquad \textit{and} \qquad \sup_{\boldsymbol{z}\in\mathbb{R}^k,\boldsymbol{\theta}\in\Theta} \boldsymbol{\alpha}^\top s'(\boldsymbol{z};\boldsymbol{\psi})\boldsymbol{\beta} < 1\,. \tag{10}$$

This corollary shows that there will be a non-degenerate set of parameters that lead to an invertible filter whenever the function $s(\cdot\,;\cdot)$ is such that the diagonal elements of its derivative are bounded from below by a positive constant and from above by a finite constant. In other words, if $s(\boldsymbol{x};\boldsymbol{\psi})$ is monotonically increasing and Lipschitz continuous in $\boldsymbol{x}$. The latter condition is already assumed in **IN2**. We discuss some choices of $s(\cdot\,;\cdot)$ that can lead to an invertible filter in the next section.

We discussed in Section 2 how the choice of $s(\cdot\,;\cdot)$ has an impact on the properties of the model. Naturally, this choice also impacts the filter. In case $s(\boldsymbol{z};\boldsymbol{\psi}) = \boldsymbol{z}$ the marginal effect of an increase in the magnitude of the prediction error on $\hat{f}_t$ is the same everywhere. In the presence of outliers, however, it is typically preferred to let the effect of large prediction errors be relatively smaller. In other words, in such cases it is beneficial to choose a function $s(\cdot\,;\boldsymbol{\psi})$ which is relatively steep around zero and relatively flat for large values of $\boldsymbol{x}$. However, not all choices of $s(\cdot\,;\boldsymbol{\psi})$ will lead to a filter for which invertibility can be theoretically established. We already noted that $s(\cdot\,;\boldsymbol{\psi})$ must be monotone in most cases for the contraction condition in Proposition 2 to hold. Just as in the univariate case (Blasques et al., 2024a), there is an additional restriction on $s(\cdot\,;\cdot)$ which is stated in the following corollary.

Corollary 3. *If the limits of $s(\boldsymbol{z};\boldsymbol{\psi})$ and $\partial s(\boldsymbol{z};\boldsymbol{\psi})/\partial\boldsymbol{z}$ as $\|\boldsymbol{z}\|\to\infty$ exist and*

$$\lim_{\|\boldsymbol{z}\|\to\infty} \frac{\partial s(\boldsymbol{z};\boldsymbol{\psi})}{\partial\boldsymbol{z}^\top} = \mathbf{0}_{k\times k} \quad \textit{for some} \quad \boldsymbol{\psi}\in\boldsymbol{\Psi}\,,$$

then conditions (6) and thus (8) fail to hold, as this implies that there is a $\boldsymbol{\xi}\in\boldsymbol{\Xi}$ such that $\sup_{\boldsymbol{g}\in\mathbb{R}^{p+1}} \|\partial\phi_t^{(r)}(\boldsymbol{g},\boldsymbol{\xi})/\partial\boldsymbol{g}^\top\| \geq 1$ almost surely for any integer $r\geq 1$.

A proof is omitted, as this can be shown straightforwardly by the same argument as for Corollary 1 of Blasques et al. (2024a). This result implies that the contraction conditions of the previous section fail if $s(\boldsymbol{x};\boldsymbol{\psi})$ converges to a constant vector as $\boldsymbol{x}$ grows large. More precisely, it implies that we only get an invertible filter if at least one of the elements of $s(\cdot\,;\boldsymbol{\psi})$ diverges linearly in the limit. This means that our filter will not be robust in the classical sense, see e.g. Calvet et al. (2015) for a formal definition of robust filters. We can however still have some form of robustness, by having a function with a higher slope around zero than far away from zero.

We will give a few examples of functions s for which invertibility can potentially be shown using Proposition 2. For instance, consider the function $s(\cdot\,;\boldsymbol{\psi})$ where each element $s_i(\cdot\,;\boldsymbol{\psi})$ for $i = 1,\ldots,k$ is the score function of the location of a finite mixture of normals; see Blasques et al. (2024a). This quasi-score function diverges linearly in the limit and is flexible, as the mixture of normals itself is very flexible, for instance allowing for asymmetry. The score function that one would use in a score-driven location model with for instance Student's t or exponential generalized beta distribution of the second kind (EGB2) innovations, does not diverge linearly and therefore invertibility of a filter based on either of these functions cannot be established based on Proposition 2. As mentioned before, another flexible option is a natural cubic spline function for a given set of knots, where the function is linearly extrapolated beyond the outer knots, such that invertibility can be established.

3.2.2. Consistency

Now we turn to the consistency of the estimator of $\boldsymbol{\xi}$. We first consider the consistency of the estimator $\widehat{\boldsymbol{\xi}}_T(\boldsymbol{\gamma}_0)$ and then we extend this consistency result to $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$. To prove these results, we need (subsets of) the following conditions:

C1 Define $s(\cdot\,;\boldsymbol{\Sigma},\boldsymbol{\psi}_{-1}) \equiv s(\cdot\,;\boldsymbol{\psi})$. Let s and $\boldsymbol{\theta}_0 \in \Theta$ be such that $\omega+\boldsymbol{\alpha}^\top s(\boldsymbol{u}_t;\boldsymbol{\Sigma}_0,\boldsymbol{\psi}_{-1}) = \omega_0 + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)$ a.s., if and only if $(\omega,\boldsymbol{\alpha},\boldsymbol{\psi}_{-1}) = (\omega_0,\boldsymbol{\alpha}_0,\boldsymbol{\psi}_{-1,0})$.

C2 The conditions of Proposition 1 are satisfied.

C3 Condition **A3** and the conditions of Proposition 2 are satisfied.

Condition **C1** is needed for identification of $\boldsymbol{\theta}_0$. Under condition **C2**, we have invertibility of the filter in $\boldsymbol{\gamma}_0$ uniformly over Ξ by Proposition 1, and the sequence of filtering errors will be SE in the limit. If the stronger contraction condition in **C3** holds, then we have uniform invertibility of the filter uniformly over Θ by Proposition 2. Assumption **A3** is needed to be able to distinguish between the case with and without drift in the consistency proof of $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$.

Under the conditions above, the terms of the quasi log likelihood in $\boldsymbol{\gamma}_0$ will be asymptotically stationary, such that we can prove strong consistency using standard techniques, in a similar way as in Blasques et al. (2024a).

Theorem 2 (Consistency of $\widehat{\boldsymbol{\xi}}_T(\boldsymbol{\gamma}_0)$)**.** *Let* ***C1*** *and* ***C2*** *hold. Then* $\widehat{\boldsymbol{\xi}}_T(\boldsymbol{\gamma}_0)$ *satisfies* $\widehat{\boldsymbol{\xi}}_T(\boldsymbol{\gamma}_0) \overset{a.s.}{\to} \boldsymbol{\xi}_0$ *as* $T \to \infty$*.*

Under some additional conditions, we can show that the quasi log likelihood evaluated in $\widehat{\boldsymbol{\gamma}}_T$ converges in probabilty to the quasi log likelihood evaluated in $\boldsymbol{\gamma}_0$ as $T \to \infty$, uniformly over Ξ. It then follows that $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$ converges in probability to $\widehat{\boldsymbol{\xi}}_T(\boldsymbol{\gamma}_0)$ as $T \to \infty$, in other words that $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$ is consistent.

Theorem 3 (Consistency of $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$)**.** *Let* ***C1*** *and* ***C3*** *hold. Then if either*

(i) $\omega_0 = 0$ *and* $\|\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0\| = o_p(T^{-1/2})$*,*

(ii) or $\omega_0 \neq 0$ *and* $\|\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0\| = o_p(T^{-1})$ *,*

and $\|\widehat{\boldsymbol{m}}_T - \boldsymbol{m}_0\| = o_p(1)$*, then* $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T) \overset{p}{\to} \boldsymbol{\xi}_0$ *as* $T \to \infty$*.*

It follows that all first-stage estimators $\widehat{\boldsymbol{\gamma}}_T$ that are consistent at an appropriate rate, lead to a consistent estimator $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$. Thus, the first-stage estimator does not need to be efficient, although efficiency could lead to a more preferable asymptotic distribution of the plug-in estimator. Furthermore, this consistency result is not contingent on the choice of identification scheme for $\boldsymbol{\beta}$ (and $\boldsymbol{\mu}$), although we present the result for the particular identification scheme under consideration. It is also applicable under any other exact identification scheme for $\boldsymbol{\beta}$, as long as the first-stage estimator $\widehat{\boldsymbol{\beta}}_T$ is consistent at an appropriate rate. Notice that we do not obtain strong consistency here, because the first-step estimator is only assumed to be weakly consistent.

It also follows from the consistency of $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$, together with the super-consistency of $\widehat{\boldsymbol{b}}_T$, and the invertibility of the filter, that we can recover the true path of the time-varying common location f_t in the limit.

Proposition 3. *Let the conditions of Theorem 3 be satisfied. Then,* $|\hat{f}_{T+1}(\widehat{\boldsymbol{\gamma}}_T, \widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)) - f_{T+1}| \overset{p}{\to} 0$ *as* $T \to \infty$ *, for any initialization* $\hat{f}_1 \in \mathbb{R}$*.*

4. Monte Carlo study: filtering ability

To investigate the filtering ability of our modeling framework, we carry out a Monte Carlo simulation study. We simulate samples $\{\boldsymbol{y}_t\}_{t=1}^{T}$, where $\boldsymbol{y}_t$ is three-dimensional, from the following data generating process (DGP):

$$\boldsymbol{y}_t = \boldsymbol{d}\mu_t + \boldsymbol{c}_t\,, \qquad \text{where} \qquad \begin{array}{ll} \mu_{t+1} = \mu_t + \eta_t\,, & \text{with } \{\eta_t\}_{t=1}^T \sim \text{iid } \mathcal{N}(0,\sigma_\eta^2)\,, \quad \text{and} \\ \boldsymbol{c}_{t+1} = \boldsymbol{\Phi}\boldsymbol{c}_t + \boldsymbol{\zeta}_t\,, & \text{with } \{\boldsymbol{\zeta}_t\}_{t=1}^T \sim \text{iid Student's } t(0,\boldsymbol{\Omega},\nu)\,, \end{array}$$

where $\boldsymbol{d}$ is a three-dimensional vector of fixed parameters, μ_t is a scalar Gaussian random walk initialized at zero, where the innovations η_t have variance σ_η^2, and $\boldsymbol{c}_t$ is a VAR(1) process, with autoregressive coefficient matrix $\boldsymbol{\Phi}$ and multivariate Student's t innovations $\boldsymbol{\zeta}_t$ with mean zero, scale $\boldsymbol{\Omega}$ and degrees of freedom ν. Hence, this is a parameter-driven version of the model given in (1)-(3). We will evaluate how well our model can filter the values of μ_t. We set $\boldsymbol{d} = (1,1,1)^\top$ and $\sigma_\eta^2 = 0.2$. We consider two settings for $\boldsymbol{\Phi}$: in Setting 1 we set $\boldsymbol{\Phi} = 0$ (no short-run dynamics) and in Setting 2 we set $\boldsymbol{\Phi} = 0.6\boldsymbol{I}_3$, where $\boldsymbol{I}_3$ denotes the 3×3 identity matrix. Let $\boldsymbol{C} = 0.2\ \iota_3\iota_3^\top + 0.1\ \boldsymbol{I}_3$, where ι_3 is a column vector of ones with length 3. For the degrees of freedom parameter ν and the scaling matrix $\boldsymbol{\Omega}$ we consider three combinations: (1) $\nu = 5$ and $\boldsymbol{\Omega} = \boldsymbol{C}$ (such that $\mathbb{V}\text{ar}(\zeta_{it}) = 0.5$), (2) $\nu = 5$ and $\boldsymbol{\Omega} = 1.8\ \boldsymbol{C}$ (such that $\mathbb{V}\text{ar}(\zeta_{it}) = 0.9$) and (3) $\nu = 3$ and $\boldsymbol{\Omega} = \boldsymbol{C}$ (such that $\mathbb{V}\text{ar}(\zeta_{it}) = 0.9$).

We consider 1000 replications for the sample sizes $T = 500, 1000$ and 2000, plus 20 time points which we use as a burn-in period for the filter. The filtered sequences are initialized at the first observation of the sample $y_{1,1}$. For simplicity, we set $\omega = 0$ and $\boldsymbol{\mu} = 0$. Furthermore, for Setting 1, we let $p = 0$ and for Setting 2, we let $p = 1$ with a diagonal VAR specification. Furthermore, we use the following three specifications for the function s, where $\boldsymbol{\Sigma}_{ii}$ denotes the i-th diagonal element of the matrix $\boldsymbol{\Sigma}$:

(i) a linear specification: $s_i(\boldsymbol{x};\boldsymbol{\psi}) = x_i/\boldsymbol{\Sigma}_{ii}$;

(ii) a nonlinear specification using splines: $s_i(\boldsymbol{x};\boldsymbol{\psi}) = g(x_i/\sqrt{\boldsymbol{\Sigma}_{ii}};\tau)/\sqrt{\boldsymbol{\Sigma}_{ii}}$, where $g(\cdot;\tau)$ is a symmetric natural cubic spline function with four knots and the parameter $\tau > 0$ is an element of $\boldsymbol{\psi}$. The knots $k_1 < k_2 < k_3 < k_4$ are located at the 2%, 25%, 75% and 98% quantiles of the standard normal distribution. To enforce symmetry and identification, we set $g(k_2;\tau) = k_2$, $g(k_3;\tau) = k_3$, $g(k_1;\tau) = -\tau$ and $g(k_4;\tau) = \tau$;

(iii) a nonlinear specification using the location score of a mixture of normal distributions derived in Blasques et al. (2024a):

$$s_i(\boldsymbol{x};\boldsymbol{\psi}) = g(x_i/\sqrt{\boldsymbol{\Sigma}_{ii}};\boldsymbol{\psi})/\sqrt{\boldsymbol{\Sigma}_{ii}}\,, \quad \text{where}$$

$$g(z;\boldsymbol{\psi}) = z\,\frac{\frac{1}{\sigma_1^2}wf(z;\sigma_1) + \frac{1}{\sigma_2^2}(1-w)f(z;\sigma_2)}{wf(z;\sigma_1) + (1-w)f(z;\sigma_2)}\,, \quad \text{where} \quad f(z;\sigma) = \frac{1}{\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)\,,$$

where $w \in (0,1)$ and $\sigma_1 > \sigma_2 > 0$. To have parsimony we impose that $w\sigma_1^2 + (1-w)\sigma_2^2 = 1$, and we use the following parametrization: $\boldsymbol{\psi} = (\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}, \boldsymbol{\Sigma}_{33}, \tau, w)$, where $\tau \equiv \sigma_1^2/\sigma_2^2 > 1$, such that $\sigma_2^2 = 1/(w\tau + 1 - w)$ and $\sigma_1^2 = \sigma_2^2\tau$. In this simulation study, we fix w at 0.5 and only estimate τ.

We refer to Poirier (1973) for details on how to construct a cubic spline as a system of linear equations. Natural cubic spline functions are such that the second order derivative is equal to zero at the outer knots, and we choose to linearly extrapolate the function beyond the outer knots. The results are not sensitive to small changes in the knot values. For option (ii), if τ is large enough, the function g is monotonically increasing. If additionaly τ is smaller than k_4, the function has a higher slope around zero than when it is far away from zero, which induces filter robustness. When $\tau = k_4$, then cases (i) and (ii) are equivalent. Hence, the function in (i) is a special case of the function in (ii).

The mixed normal quasi-score function in (iii) also tends to the linear functions as $\tau \to 1$. This function also has a steeper slope around zero than for larger values by construction, which leads to robustness of the filter compared to the linear case.

To have parsimony, we impose that $\boldsymbol{\alpha} = \alpha \cdot (1, 1, 1)^\top$, so that the elements of $\boldsymbol{\alpha}$ are equal to each other. This is the reason why we divide by $\boldsymbol{\Sigma}_{ii}$ in case (i), as it ensures that the variability of the innovations of the different time series is accounted for in the filter updates. For case (ii), we scale the input of the spline function by the standard deviation of the corresponding innovation, after which the output of the spline function is again divided by this standard deviation. The first scaling justifies the placement of the knots at certain quantiles of the standard normal distribution. Similarly, for case (iii), the first scaling justifies the restriction $w\sigma_1^2 + (1 - w)\sigma_2^2 = 1$. The second scaling in (ii) and (iii) ensures that the linear function s of (i) is a special case of these nonlinear functions.

For each of the generated samples, we estimate the parameters of the three models described above using our two-step Gaussian QML approach, where we use the modified OLS method discussed in Supplementary Appendix D for the first step. We evaluate the filtering performance of our model, by comparing the filtered values $\widehat{\boldsymbol{\beta}}_T \hat{f}_{t+1}(\widehat{\boldsymbol{\theta}}_T)$ with the 'true' value of the long-term component of $\boldsymbol{y}_t$, which is $\boldsymbol{d}\mu_t$. Notice that μ_t is strictly speaking not the long-term expectation of $\boldsymbol{y}_t$, and that ideally we would compare $\widehat{\boldsymbol{\beta}}_T \hat{f}_{t+1}(\widehat{\boldsymbol{\theta}}_T)$ to the estimated long-term trend $\mathbb{E}[\mu_t|\mathcal{F}_t]$, which could be approximated using a particle filter, but we choose to keep this simulation study simple. Our model is misspecified, as the DGP is parameter-driven and non-Gaussian. Yet, we expect our model based on the nonlinear functions described in points (ii) and (iii) above, to filter the value of μ_t more accurately than for the linear function described in point (i), due to their ability to be less sensitive to outliers.

Filtering results

We present the average root mean squared filtering errors (RMSE) and mean absolute filtering errors (MAE) in Table 1. As expected, the filtering performance of the nonlinear models is better than that of the linear model, and this difference is more pronounced for $\nu = 3$ than for $\nu = 5$, so if the tails of the noise distribution are fatter. Allowing for a nonlinear update of the location process f_t improves the filtering ability of our model in case of fat-tailed disturbances. This effect is clearly visible for the case with and without short-run dynamics. Between the two nonlinear models, the spline model has a smaller average RMSE and MAE than the model that uses the mixed normal score function.

For Setting 2, the RMSE and MAE are considerably higher than for Setting 1, which is not surprising, because the short-run dynamics complicate the tracking of the underlying long-term location component. Notice that for $\mathbb{V}\text{ar}(\zeta_{it}) = 0.9$, the RMSE and MAE are lower for $\nu = 3$ than for $\nu = 5$, which seems counter-intuitive, but is caused by the distribution for $\nu = 5$ having to be 'streched out' more, to obtain the same variance as for $\nu = 3$. Under identical scaling matrices, the RMSE for $\nu = 5$ is considerably lower than that for $\nu = 3$. This is caused by the variance of the noise terms $\boldsymbol{\zeta}_t$ being proportional to $\nu/(\nu - 2)$, implying that for $\nu = 5$ the variance is lower, which makes it easier to track μ_t based on the observations. As the sample size increases, the filtering accuracy tends to slightly decrease. As T increases, the parameters should be closer to their theoretically optimal values, which should lead to a better filter, but on the other hand, the path $\{\boldsymbol{d}\mu_t\}$ that the filtered sequence $\{\widehat{\boldsymbol{\beta}}_T \hat{f}_{t+1}(\widehat{\boldsymbol{\theta}}_T)\}$ tries to trace is longer for larger T. Here, the latter effect apparently dominates in most cases.

5. Empirical study

To demonstrate the empirical relevance of our model, we analyse three time series of prices of commodities related to oil traded in the United States. Specifically, we consider spot prices of West Texas Intermediate crude oil, and heating oil and gasoline traded in the New York Harbor market[‡]. In order not to have day-of-the-week effects, we take the spot price in dollars per gallon of every Friday (the end of each trading week), delivering a weekly trivariate time series from June 6, 1986 to March 27, 2024, with length $T = 1974$. We take logs of all time series and multiply by 10 for numerical purposes; the resulting data are presented in Figure 1. This plot clearly shows non-stationarity and co-movement. This co-movement is not surprising, because heating oil and gasoline are obtained by refining crude oil. The level of the crude oil log prices is slightly lower than that of the other two liquids, because unlike the others, crude oil still needs to be refined. Intuitively, the long-term expectation of these time series should be roughly equal, taking into account their different levels, which is why our model could be suitable for filtering the common long-term trend of these time series. Occasionally, there are large outliers in the data. The use of a robust filter for the long-term trend seems therefore most appropriate.

[‡] Data were retrieved from the website of the U.S. Energy Information Administration: www.eia.gov.

Table 1. Monte Carlo simulation filtering results*

		Setting 1 ($\boldsymbol{\Phi} = 0$)					
		RMSE			MAE		
		$\nu = 5$	$\nu = 5$	$\nu = 3$	$\nu = 5$	$\nu = 5$	$\nu = 3$
T		$\mathbb{V}(\zeta_{it}) = 0.5$	$\mathbb{V}(\zeta_{it}) = 0.9$	$\mathbb{V}(\zeta_{it}) = 0.9$	$\mathbb{V}(\zeta_{it}) = 0.5$	$\mathbb{V}(\zeta_{it}) = 0.9$	$\mathbb{V}(\zeta_{it}) = 0.9$
500	linear	0.4430	0.5357	0.5262	0.3449	0.4197	0.3979
		(0.0224)	*(0.0284)*	*(0.0587)*	*(0.0162)*	*(0.0212)*	*(0.0283)*
	spline	**0.4319**	**0.5178**	**0.4680**	**0.3379**	**0.4076**	**0.3627**
		(0.0198)	*(0.0253)*	*(0.0249)*	*(0.0150)*	*(0.0195)*	*(0.0182)*
	MN	0.4330	0.5203	0.4907	0.3383	0.4090	0.3731
		(0.0203)	*(0.0257)*	*(0.0631)*	*(0.0152)*	*(0.0197)*	*(0.0277)*
1000	linear	0.4428	0.5353	0.5273	0.3444	0.4190	0.3975
		(0.0157)	*(0.0199)*	*(0.0409)*	*(0.0111)*	*(0.0146)*	*(0.0194)*
	spline	**0.4315**	**0.5171**	**0.4671**	**0.3375**	**0.4069**	**0.3616**
		(0.0136)	*(0.0175)*	*(0.0175)*	*(0.0104)*	*(0.0136)*	*(0.0126)*
	MN	0.4325	0.5198	0.4936	0.3377	0.4082	0.3739
		(0.0140)	*(0.0182)*	*(0.0473)*	*(0.0104)*	*(0.0136)*	*(0.0212)*
2000	linear	0.4431	0.5358	0.5301	0.3446	0.4194	0.3984
		(0.0112)	*(0.0142)*	*(0.0306)*	*(0.0078)*	*(0.0102)*	*(0.0145)*
	spline	**0.4319**	**0.5177**	**0.4674**	**0.3377**	**0.4073**	**0.3616**
		(0.0097)	*(0.0125)*	*(0.0122)*	*(0.0073)*	*(0.0095)*	*(0.0089)*
	MN	0.4329	0.5203	0.4993	0.3379	0.4086	0.3764
		(0.0104)	*(0.0136)*	*(0.0402)*	*(0.0074)*	*(0.0098)*	*(0.0189)*

		Setting 2 ($\boldsymbol{\Phi} = 0.6\boldsymbol{I}_3$)					
		RMSE			MAE		
		$\nu = 5$	$\nu = 5$	$\nu = 3$	$\nu = 5$	$\nu = 5$	$\nu = 3$
T		$\mathbb{V}(\zeta_{it}) = 0.5$	$\mathbb{V}(\zeta_{it}) = 0.9$	$\mathbb{V}(\zeta_{it}) = 0.9$	$\mathbb{V}(\zeta_{it}) = 0.5$	$\mathbb{V}(\zeta_{it}) = 0.9$	$\mathbb{V}(\zeta_{it}) = 0.9$
500	linear	0.7519	0.8738	0.8606	0.6014	0.6991	0.6839
		(0.0618)	*(0.0747)*	*(0.0987)*	*(0.0504)*	*(0.0612)*	*(0.0728)*
	spline	**0.7338**	**0.8437**	**0.7758**	**0.5855**	**0.6741**	**0.6172**
		(0.0596)	*(0.0708)*	*(0.0668)*	*(0.0486)*	*(0.0579)*	*(0.0546)*
	MN	0.7363	0.8485	0.8023	0.5872	0.6776	0.6344
		(0.0600)	*(0.0714)*	*(0.0993)*	*(0.0487)*	*(0.0583)*	*(0.0711)*
1000	linear	0.7535	0.8753	0.8641	0.6014	0.6987	0.6849
		(0.0442)	*(0.0536)*	*(0.0708)*	*(0.0361)*	*(0.0439)*	*(0.0524)*
	spline	**0.7354**	**0.8446**	**0.7772**	**0.5853**	**0.6731**	**0.6166**
		(0.0428)	*(0.0511)*	*(0.0489)*	*(0.0348)*	*(0.0417)*	*(0.0389)*
	MN	0.7372	0.8490	0.8087	0.5868	0.6764	0.6371
		(0.0430)	*(0.0515)*	*(0.0764)*	*(0.0350)*	*(0.0418)*	*(0.0545)*
2000	linear	0.7566	0.8788	0.8700	0.6035	0.7009	0.6885
		(0.0323)	*(0.0393)*	*(0.0518)*	*(0.0264)*	*(0.0321)*	*(0.0377)*
	spline	**0.7392**	**0.8487**	**0.7803**	**0.5877**	**0.6757**	**0.6185**
		(0.0307)	*(0.0369)*	*(0.0338)*	*(0.0252)*	*(0.0303)*	*(0.0273)*
	MN	0.7408	0.8527	0.8192	0.5892	0.6788	0.6440
		(0.0316)	*(0.0380)*	*(0.0635)*	*(0.0257)*	*(0.0310)*	*(0.0452)*

*Average root mean squared filtering error (RMSE) and mean absolute filtering error (MAE) over 1000 replications, with standard deviations in brackets. For each replication, RMSE and MAE are calculated for $t = 1, \ldots, T$, based on filtering errors $\widehat{\boldsymbol{\beta}}_T \hat{f}_{t+1} - \boldsymbol{d}\mu_t$. MN stands for the mixture of normals quasi-score. Lowest errors per DGP and T are highlighted in bold.

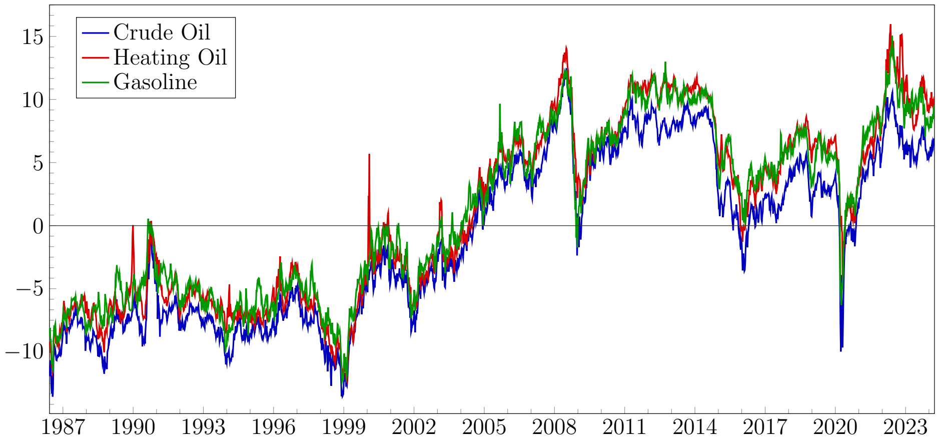

Figure 1. Spot prices per gallon every Friday from June 6, 1986 to March 27, 2024, in logs and multiplied by 10.

We proceed to estimate the parameters of the common trend model in (1)-(3), in order to eventually filter the long-term trend of the spot prices of the three oil-associated products. For the function s we choose the linear, cubic spline and mixed normal score function introduced in Section 4, where for the latter we no longer restrict $w = 0.5$.

We consider a scalar VAR specifications for $\boldsymbol{\varepsilon}_t$ to enforce parsimony and because a diagonal VAR specification leads to worse information criterion values. This simplifying restriction does not imply that we do not allow the different time series to impact each other, as they all affect the value of the filtered long-term trend. No drift is added, i.e. ω is not included in the model, because including ω does not substantially improve the fit.

Empirical results

The parameters in $\boldsymbol{\gamma}$ are estimated using the modified OLS method discussed in Supplementary Appendix D, see Table 2 for the results. The estimates of β_2 and β_3 are close to one and μ_2 and μ_3 are both positive, as expected. These estimates are plugged into the quasi log likelihood, which we maximize to obtain an estimate of the other parameters; see Table 3. We burn the first 20 observations to reduce the impact of the initialization of the filter. We consider $p = 0$ and $p = 2$ VAR lags. Adding more VAR lags does not lead to better information criterion values. For each model, we display the optimized quasi log-likelihood value, AIC, BIC and Takeuchi Information Criterion (TIC); see Takeuchi (1976). The latter information criterion is included, because unlike the AIC and BIC it is a valid information criterion under QML estimation.

For the linear and spline models with $p = 0$ and $p = 2$, it can be shown that the conditions in Corollary 2 are satisfied, such that we have filter invertibility. Regarding the models that use the mixed normal quasi-score, for $p = 0$ the invertibility of the filter can be shown using Corollary 2, but for $p = 2$ the function s is non-montone, so we cannot show invertibility using Propositions 1 or 2. This does not mean the latter filter is not invertible, as the conditions are sufficient, not necessary.

Allowing for an autoregressive specification for $\boldsymbol{\varepsilon}_t$, that is $p \geq 1$, leads to a large improvement of the maximized quasi log-likelihood value. In other words, imposing that $\boldsymbol{\varepsilon}_t$ is serially uncorrelated is too restrictive for these data, which is also indicated by the high autocorrelation that is present in the residuals of the models with $p = 0$ VAR lags. The estimated VAR dynamics are rather persistent, with autoregressive coefficients of around 0.8 for the first lag and around 0.13 for the second lag.

For comparison, we also give the results of the parameter-driven common trend model of Chang et al. (2009) in Table 3. The model is reparametrized such that it has a similar identification scheme as our model, and to be able to compare the results, we also employ the two-step procedure for this model, using the modified OLS estimates of Table 2. Estimating the parameters jointly leads to very similar estimates and almost the same log likelihood value. The log likelihood of the two-step estimator of the model of Chang et al. (2009) is slightly better than that of the linear model with $p = 0$ lags, but the difference is small. It can be shown that if we use the update $f_{t+1} = f_t + \alpha \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{u}_t$, then these two models are exactly equivalent to each other.

Furthermore, we observe that for lag order $p = 2$, the nonlinear models have better information criterion values than the linear models, while for lag order $p = 0$ the opposite is true. For the case $p = 0$, the estimate of τ for the model that uses the spline function is relatively high, leading to a relatively less robust filter than for $p = 2$. For $p = 0$, the estimate of τ is not significantly below $k_4 \approx 2.053$ at a 1% significance level, according to a t-test, while for $p = 2$ it is. This is evidence that indeed the spline specification is superior to the linear specification for $p = 2$ and not for $p = 0$. For the model with $p = 0$ that uses the mixed normal quasi-score, w is relatively high and $\tau = \sigma_1^2/\sigma_2^2$ is relatively close to 1, which also leads to a filter that approaches the linear filter. The overall preferred model is the spline model with $p = 2$, as it has the best information criterion values.

The updating functions $\hat{\alpha}_T s_1(\boldsymbol{x}; \widehat{\boldsymbol{\psi}}_T)$ of the fitted models with $p = 2$ are plotted in Figure 2. The robustness of the nonlinear filters is clearly visible, as the slope of the

Table 2. Parameter estimates $\widehat{\boldsymbol{\gamma}}_T$ obtained from modified ordinary least squares estimation, for log price series shown in Figure 1. Standard errors are reported in brackets.

β_2	β_3	μ_2	μ_3
1.063	1.006	2.031	2.044
(0.008)	*(0.008)*	*(0.055)*	*(0.056)*

Table 3. Parameter estimates $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$, with $\widehat{\boldsymbol{\gamma}}_T$ from Table 2.

		$p = 0$			$p = 2$		
$\widehat{\boldsymbol{\xi}}_T$	SSM	linear	spline	MN	linear	spline	MN
$\alpha(\sigma^2_{\text{state}})$	0.218	0.192	0.207	0.195	0.028	0.039	0.024
	(0.008)	*(0.011)*	*(0.015)*	*(0.010)*	*(0.006)*	*(0.008)*	*(0.007)*
τ			1.873	1.701		1.107	4.203
			(0.184)	*(0.759)*		*(0.302)*	*(1.222)*
w				0.826			0.524
				(0.390)			*(0.230)*
$\boldsymbol{A}_1$					0.794	0.806	0.805
					(0.063)	*(0.063)*	*(0.063)*
$\boldsymbol{A}_2$					0.137	0.128	0.128
					(0.043)	*(0.045)*	*(0.044)*
σ^2_1	0.331	0.538	0.536	0.535	0.312	0.310	0.310
	(0.015)	*(0.063)*	*(0.063)*	*(0.063)*	*(0.025)*	*(0.026)*	*(0.026)*
σ^2_{12}	-0.190	0.045	0.044	0.045	0.217	0.214	0.214
	(0.009)	*(0.025)*	*(0.023)*	*(0.023)*	*(0.013)*	*(0.013)*	*(0.013)*
σ^2_2	0.343	0.605	0.605	0.608	0.334	0.330	0.331
	(0.011)	*(0.097)*	*(0.094)*	*(0.095)*	*(0.035)*	*(0.034)*	*(0.035)*
σ^2_{13}	-0.081	0.132	0.132	0.131	0.217	0.215	0.215
	(0.012)	*(0.050)*	*(0.049)*	*(0.049)*	*(0.017)*	*(0.018)*	*(0.018)*
σ^2_{23}	-0.248	0.007	0.008	0.008	0.208	0.206	0.206
	(0.011)	*(0.025)*	*(0.025)*	*(0.025)*	*(0.012)*	*(0.012)*	*(0.012)*
σ^2_3	0.477	0.711	0.713	0.711	0.340	0.339	0.338
	(0.024)	*(0.065)*	*(0.062)*	*(0.062)*	*(0.019)*	*(0.020)*	*(0.020)*
$\widehat{L}_T$	-6836.96	-6837.31	-6835.40	-6835.72	-3798.43	-3788.70	-3788.43
AIC	13687.92	13688.62	13686.79	13689.44	7614.86	7597.40	7598.87
TIC	13704.37	13722.51	13727.19	13730.59	7765.45	7753.86	7752.28
BIC	13726.96	13727.67	13731.41	13739.64	7665.06	7653.17	7660.22

*Standard errors are reported in brackets. MN stands for mixture of normals. $\widehat{L}_T$ denotes maximized quasi log likelihood. SSM corresponds to state space model of Chang et al. (2009) reparametrized to match the current identification scheme (so β_1 is fixed at 1 and σ^2_{state} is variance of state innovation).

nonlinear functions is higher around zero than for the linear function, while the opposite is true for large values. So even though all three functions diverge linearly in the limit, the level of the nonlinear functions is considerably lower for large values.

Figure 3 shows the filtered long-term trends for the models with $p = 2$ autoregressive lags. The plot also shows a 'standardized' version of the data by subtracting the estimated values of μ_i and dividing by β_i. The filtered trend is more smooth than the observation

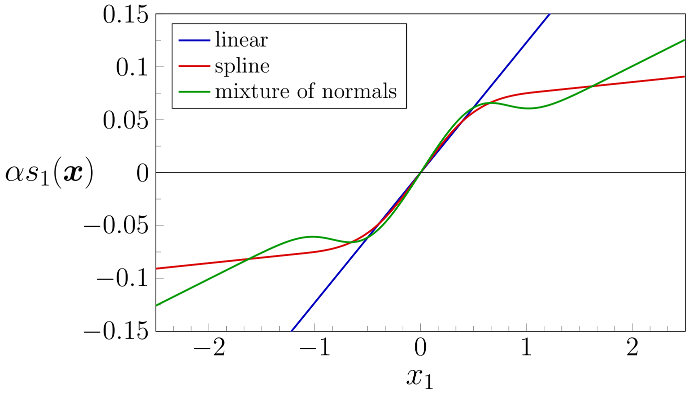

Figure 2. Plots of $\hat{\alpha}_T s_1(\boldsymbol{x}; \widehat{\boldsymbol{\psi}}_T)$ for the estimates of the models with $p = 2$, see Table 3.

series and whenever there is a shift in the level of the observations, the long-term trend is usually rather conservative. Whenever the prices are increasing, the filtered trend is typically below the observations, and vice versa in case of a prolonged decrease in the prices. In this way, the filtered trend behaves like one would expect a long-term component to behave. The path of the linear and nonlinear filters differ occasionally, while the two nonlinear filters lead to virtually identical paths. The most notable differences occur after the 2008 financial crisis and the COVID pandemic in 2020. Here, the linear filter has a more pronounced reaction to the sudden drop in the prices than the nonlinear filters. During such a period the robustness of the nonlinear filter makes a difference.

The filtered long-term trends can be used to perform a multivariate Beveridge-Nelson trend-cycle decomposition on the observed series. The model $\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{\beta} f_{t+1} + \boldsymbol{c}_t$ decomposes the observations in the trend component $\boldsymbol{\mu} + \boldsymbol{\beta} f_{t+1}$, and the cycle component $\boldsymbol{c}_t = \boldsymbol{y}_t - \boldsymbol{\mu} - \boldsymbol{\beta} f_{t+1}$. The cycle components for the models with $p = 2$ lags are presented in Figure 4. The filtered cycle is persistent and can deviate from zero for a longer time,

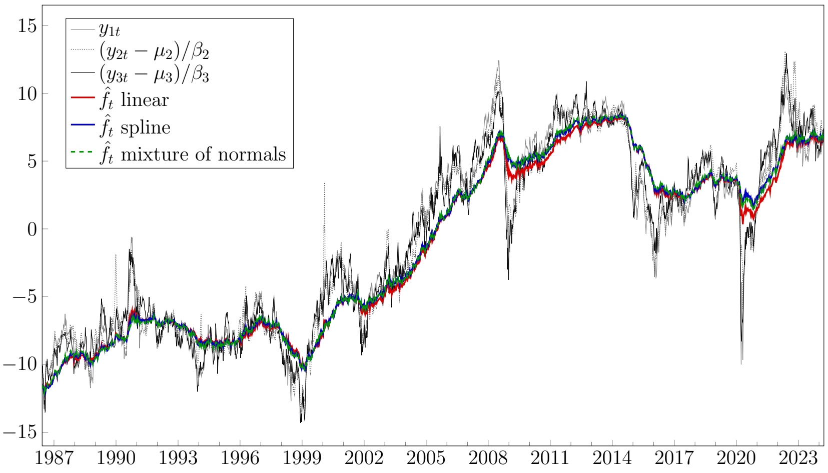

Figure 3. Filtered long-term trend using model with $p = 2$, see Table 3, where y_{1t} corresponds to crude oil, y_{2t} to heating oil and y_{3t} to gasoline

for instance around the financial crisis in 2008, the COVID pandemic in 2020 and the start of the war in Ukraine in 2022, but overall they tend to fluctuate around zero.

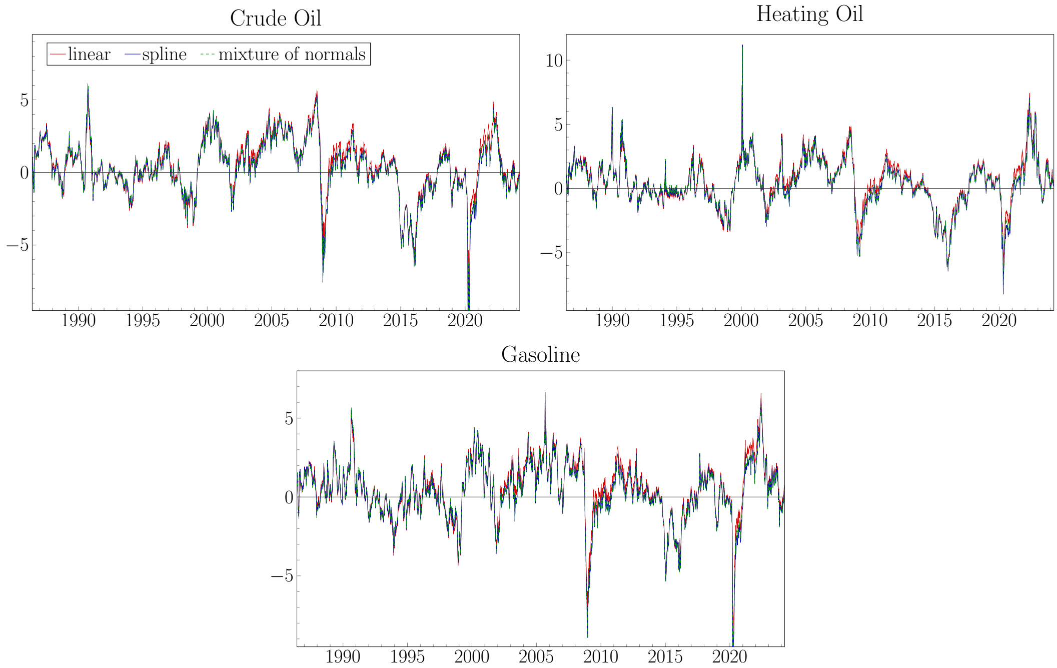

Figure 4. Cycle components $\boldsymbol{y}_t - \widehat{\boldsymbol{\mu}}_T - \widehat{\boldsymbol{\beta}}_T \hat{f}_{t+1}(\widehat{\boldsymbol{\theta}}_T)$ corresponding to the long-term trend plotted in Figure 3 for the three different time series.

6. Conclusion

We have introduced a novel robust multivariate conditional location model with a single common stochastic trend and possible vector autoregressive dynamics for the innovations. Our model offers a convenient way for the robust filtering of the long-term expectation of groups of time series that are driven by a common trend. We introduce a simple two-step estimation procedure, where in the first step we estimate the long-run parameters $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ via an OLS regression and in the second step we estimate the other parameters via Gaussian QML estimation where the first-step estimates are plugged into the quasi log likelihood. The asymptotic distribution of the first-stage estimator follows from standard regression theory for integrated processes. We have developed sufficient conditions for the invertibility of the filter and for the consistency of the second-stage estimator.

A natural extension is to allow for multiple stochastic trends driving the observations. From a methodological standpoint, this is a straightforward extension but deriving the theoretical results will be tedious. Another extension is to consider robust estimation in the first and second stages which might benefit the empirical performance. We discuss such possible robust alternatives to OLS in Supplementary Appendix D. Further, for the second estimation stage, we can postulate a distribution for the innovations and use maximum likelihood estimation instead of Gaussian QML. Finally, instead of modeling $\boldsymbol{\varepsilon}_t$ as a VAR process, we can extend the model with idiosyncratic stationary unobserved variables for each time series variable, similar to the univariate model of Blasques et al. (2024a), allowing for a robust update of both the long-term and short-term components. Such an extension will further complicate the theory and we expect that relatively more stringent conditions need to be put in place, in order to obtain such theoretical results.

References

Andrews, D. W. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60(4):953–966.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(4):817–858.

Ariño, M. A. and Newbold, P. (1998). Computation of the Beveridge–Nelson decomposition for multivariate economic time series. *Economics Letters*, 61(1):37–42.

Artemova, M. (2023). An order-invariant score-driven dynamic factor model. Technical Report TI 2023-067/III, Tinbergen Institute Discussion Paper.

Beveridge, S. and Nelson, C. R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics*, 7(2):151–174.

Blasques, F., Francq, C., and Laurent, S. (2023). Quasi score-driven models. *Journal of Econometrics*, 234(1):251–275.

Blasques, F., Gorgi, P., Koopman, S. J., and Wintenberger, O. (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics*, 12(1):1019–1052.

Blasques, F., van Brummelen, J., Gorgi, P., and Koopman, S. J. (2024a). Maximum likelihood estimation for non-stationary location models with mixture of normal distributions. *Journal of Econometrics*, 238(1):105575.

Blasques, F., van Brummelen, J., Gorgi, P., and Koopman, S. J. (2024b). A robust Beveridge-Nelson decomposition using a score-driven approach with an application. *Economics Letters*, 236:111588.

Blasques, F., van Brummelen, J., Koopman, S. J., and Lucas, A. (2022). Maximum likelihood estimation for score-driven models. *Journal of Econometrics*, 227(2):325–346.

Blazsek, S., Escribano, A., and Licht, A. (2021). Co-integration with score-driven models: an application to US real GDP growth, US inflation rate, and effective Federal Funds rate. *Macroeconomic Dynamics*, pages 1–21.

Bougerol, P. (1993). Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization*, 31(4):942–959.

Calvet, L. E., Czellar, V., and Ronchetti, E. (2015). Robust filtering. *Journal of the American Statistical Association*, 110(512):1591–1606.

Chang, Y., Miller, J. I., and Park, J. Y. (2009). Extracting a common stochastic trend: Theory with some applications. *Journal of Econometrics*, 150(2):231–247.

Clark, P. K. (1987). The cyclical component of U.S. economic activity. *Quarterly Journal of Economics*, 102(4):797–814.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, 8(2):93–115.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Creal, D., Schwaab, B., Koopman, S. J., and Lucas, A. (2014). Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics*, 96(5):898–915.

Davidson, J. (2000). *Econometric Theory*. Wiley.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, 55(2):251–276.

Franses, P. H. and Lucas, A. (1998). Outlier detection in cointegration analysis. *Journal of Business & Economic Statistics*, 16(4):459–468.

Hansen, B. E. (1992). Consistent covariance matrix estimation for dependent heterogeneous processes. *Econometrica*, 60(4):967–972.

Harvey, A. C. (1985). Trends and cycles in macroeconomic time series. *Journal of Business & Economic Statistics*, 3(3):216–227.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge university press.

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, volume 52. Cambridge University Press.

Harvey, A. C. and Luati, A. (2014). Filtering with heavy tails. *Journal of the American Statistical Association*, 109(507):1112–1122.

Johansen, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models.* Oxford University Press.

Krengel, U. (1985). *Ergodic theorems*, volume 6 of *De Gruyter Studies in Mathematics.* de Gruyter, Berlin.

Łasak, K. and Lont, J. (2020). Observation driven long run equilibria. *Computational Economics*, 55(2):551–575.

Loève, M. (1977). *Probability theory.* Springer-Verlag, New York.

Lucas, A. (1997). Cointegration testing using pseudolikelihood ratio tests. *Econometric Theory*, 13(2):149–169.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis.* Springer.

Murasawa, Y. (2015). The multivariate Beveridge–Nelson decomposition with I(1) and I(2) series. *Economics Letters*, 137:157–162.

Newey, W. K. and West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653.

Park, J. Y. and Phillips, P. C. (1988). Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory*, 4(3):468–497.

Phillips, P. C. (1995). Robust nonstationary regression. *Econometric Theory*, 11(5):912–951.

Phillips, P. C. and Durlauf, S. N. (1986). Multiple time series regression with integrated processes. *The Review of Economic Studies*, 53(4):473–495.

Phillips, P. C. and Hansen, B. E. (1990). Statistical inference in instrumental variables regression with I(1) processes. *The Review of Economic Studies*, 57(1):99–125.

Poirier, D. J. (1973). Piecewise regression using cubic splines. *Journal of the American Statistical Association*, 68(343):515–524.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 33(5):659–680.

Saikkonen, P. (1991). Asymptotically efficient estimation of cointegration regressions. *Econometric Theory*, 7(1):1–21.

Saikkonen, P. (1993). Estimation of cointegration vectors with linear restrictions. *Econometric Theory*, 9(1):19–35.

Shin, Y. (1994). A residual-based test of the null of cointegration against the alternative of no cointegration. *Econometric Theory*, 10(1):91–115.

Straumann, D. and Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics*, 34(5):2449–2495.

Takeuchi, K. (1976). The distribution of information statistics and the criterion of goodness of fit of models. *Mathematical Science*, 153:12–18.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.

Wintenberger, O. (2013). Continuous invertibility and stable QML estimation of the EGARCH(1, 1) model. *Scandinavian Journal of Statistics*, 40(4):846–867.

Supplementary Appendix of
Robust Multivariate Observation-Driven Filtering for a Common Stochastic Trend: Theory and Application

A. Proofs of main results

Proof of Theorem 1. This result follows directly from Theorem D.1 which contains the asymptotic distribution results for this estimator. For that Theorem we have an additional assumption on the positive definiteness of the long-run covariance matrix of the vector $(\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0), \boldsymbol{\varepsilon}_{-1,t}^\top - \varepsilon_{1t}\boldsymbol{\beta}_{-1,0}^\top)^\top$, but this assumption is not necessary for showing the rate of consistency of the estimator. □

Proof of Proposition 1. This proof is follows the same approach as Proposition 3.2 of Blasques et al. (2022), but here we consider a slightly more general case, as we have an updating function with higher-order dependence. We start by noting that the equality in equation (6) holds by Lemma B.1.

Part (i), SE: We can follow the same steps as the proof of Proposition TA.3 of Blasques et al. (2022), which uses Bougerol (1993, Theorem 3.1), but then for this specific setting. Let $\mathbb{C}(\boldsymbol{\Xi}, \mathbb{R}^{p+1})$ denote the space of continuous $\mathbb{R}^{p+1}$-valued functions, equipped with the sup-norm $\|\boldsymbol{h}(\boldsymbol{\xi})\|^{\boldsymbol{\Xi}} = \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\boldsymbol{h}(\boldsymbol{\xi})\|$. Notice that $\mathbb{C}(\boldsymbol{\Xi}, \mathbb{R}^{p+1})$ is a separable Banach space. Then let us define the random map $\tilde{\phi}_t(\cdot,\cdot) : \mathbb{C}(\boldsymbol{\Xi}, \mathbb{R}^{p+1}) \times \boldsymbol{\Xi} \to \mathbb{C}(\boldsymbol{\Xi}, \mathbb{R}^{p+1})$, which is the same as ϕ_t with the important difference that this mapping $\phi_t(\cdot,\cdot) : \mathbb{R}^{p+1} \times \boldsymbol{\Xi} \to \mathbb{R}^{p+1}$. So $\hat{\boldsymbol{g}}_{t+1}(\cdot) = \tilde{\phi}_t(\hat{\boldsymbol{g}}_t(\cdot), \cdot)$, implying that we can view $\{\hat{\boldsymbol{g}}_t\}_{t\in\mathbb{N}}$ initialized at some $\hat{\boldsymbol{g}}_1$ as a sequence of random functions taking values in $\mathbb{C}(\boldsymbol{\Xi}, \mathbb{R}^{p+1})$. Define $\tilde{\phi}_t^{(r)}$ in the same way as $\phi_t^{(r)}$.

We will now apply Theorem 3.1 of Bougerol (1993). We have that $\{\boldsymbol{u}_t\}$ is an SE sequence and under the maintained assumptions $\{\boldsymbol{\varepsilon}_t\}$ is SE, as for instance was shown in the proof of D.1. Because due to **IN2** the updating function ϕ_t is a continuous function of $(\boldsymbol{u}_t, \boldsymbol{\varepsilon}_t, \ldots, \boldsymbol{\varepsilon}_{t-p})$ for every $(\boldsymbol{g}, \boldsymbol{\xi}) \in \mathbb{R}^{p+1} \times \boldsymbol{\Xi}$, it follows from Krengel (1985, Proposition 4.3) that the random sequence $\{\tilde{\phi}_t\}_{t\in\mathbb{Z}}$ is SE. Condition (C1) of Bougerol (1993, Theorem 3.1) is $\mathbb{E}\log^+ \|\tilde{\phi}_t(\boldsymbol{h}(\cdot), \cdot)\|^{\boldsymbol{\Xi}} < \infty$ for some $\boldsymbol{h} \in \mathbb{C}(\Theta^2, \mathbb{R}^{p+1})$, which is implied by $\mathbb{E}\log^+ \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\phi_t(\bar{\boldsymbol{g}}, \boldsymbol{\xi})\| < \infty$ for some $\bar{\boldsymbol{g}} \in \mathbb{R}^{p+1}$, as we can then take the function $\boldsymbol{h}(\boldsymbol{\xi}) = \bar{\boldsymbol{g}}$ for any $\boldsymbol{\xi} \in \boldsymbol{\Xi}$. It is clear that $\mathbb{E}\log^+ \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\phi_t(\bar{\boldsymbol{g}}, \boldsymbol{\xi})\| \leq \mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\phi_t(\bar{\boldsymbol{g}}, \boldsymbol{\xi})\| < \infty$, where the expectation is finite by Lemma B.2. We can apply this lemma for $n = 1$ because $\mathbb{E}\|\boldsymbol{u}_t\| < \infty$ as $\boldsymbol{u}_t$ has a finite covariance matrix by assumption. Lastly, condition (C2) of Bougerol (1993, Theorem 3.1) can be shown to hold using the same steps as in the proof of Proposition TA.3 of Blasques et al. (2022), as for any integer $s \geq 1$

$$\begin{aligned}
&\sup_{\boldsymbol{g}_1,\boldsymbol{g}_2 \in \mathbb{C}(\boldsymbol{\Xi},\mathbb{R}^{p+1}), \|\boldsymbol{g}_1-\boldsymbol{g}_2\|^{\boldsymbol{\Xi}}>0} \frac{\|\tilde{\phi}_t^{(s)}(\boldsymbol{g}_1(\cdot),\cdot) - \tilde{\phi}_t^{(s)}(\boldsymbol{g}_2(\cdot),\cdot)\|^{\boldsymbol{\Xi}}}{\|\boldsymbol{g}_1(\cdot)-\boldsymbol{g}_2(\cdot)\|^{\boldsymbol{\Xi}}} \\
&\quad = \sup_{\boldsymbol{g}_1,\boldsymbol{g}_2 \in \mathbb{C}(\boldsymbol{\Xi},\mathbb{R}^{p+1}), \|\boldsymbol{g}_1-\boldsymbol{g}_2\|^{\boldsymbol{\Xi}}>0} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \frac{\|\phi_t^{(s)}(\boldsymbol{g}_1(\boldsymbol{\xi}),\boldsymbol{\xi}) - \phi_t^{(s)}(\boldsymbol{g}_2(\boldsymbol{\xi}),\boldsymbol{\xi})\|}{\|\boldsymbol{g}_1(\cdot)-\boldsymbol{g}_2(\cdot)\|^{\boldsymbol{\Xi}}} \\
&\quad \leq \sup_{\boldsymbol{g}_1,\boldsymbol{g}_2 \in \mathbb{C}(\boldsymbol{\Xi},\mathbb{R}^{p+1}), \|\boldsymbol{g}_1-\boldsymbol{g}_2\|^{\boldsymbol{\Xi}}>0} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}|\boldsymbol{g}_1(\boldsymbol{\xi})\neq\boldsymbol{g}_2(\boldsymbol{\xi})} \frac{\|\phi_t^{(s)}(\boldsymbol{g}_1(\boldsymbol{\xi}),\boldsymbol{\xi}) - \phi_t^{(s)}(\boldsymbol{g}_2(\boldsymbol{\xi}),\boldsymbol{\xi})\|}{\|\boldsymbol{g}_1(\boldsymbol{\xi})-\boldsymbol{g}_2(\boldsymbol{\xi})\|} \\
&\qquad \times \left(\sup_{\boldsymbol{g}_1,\boldsymbol{g}_2 \in \mathbb{C}(\boldsymbol{\Xi},\mathbb{R}^{p+1}), \|\boldsymbol{g}_1-\boldsymbol{g}_2\|^{\boldsymbol{\Xi}}>0} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \frac{\|\boldsymbol{g}_1(\boldsymbol{\xi})-\boldsymbol{g}_2(\boldsymbol{\xi})\|}{\|\boldsymbol{g}_1(\cdot)-\boldsymbol{g}_2(\cdot)\|^{\boldsymbol{\Xi}}} \right) \\
&\quad \leq \sup_{\bar{\boldsymbol{g}}_1,\bar{\boldsymbol{g}}_2\in\mathbb{R}^{p+1}, \boldsymbol{g}_1\neq\boldsymbol{g}_2} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \frac{\|\phi_t^{(s)}(\bar{\boldsymbol{g}}_1,\boldsymbol{\xi}) - \phi_t^{(s)}(\bar{\boldsymbol{g}}_2,\boldsymbol{\xi})\|}{\|\bar{\boldsymbol{g}}_1-\bar{\boldsymbol{g}}_2\|} \\
&\quad \leq \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \sup_{\bar{\boldsymbol{g}}^*\in\mathbb{R}^{p+1}} \left\| \frac{\partial \phi_t^{(s)}(\bar{\boldsymbol{g}}^*,\boldsymbol{\xi})}{\partial \boldsymbol{g}} \right\| \equiv \Lambda_t^{(s)} \, ,
\end{aligned}$$

where the final inequality follows from the mean value theorem. Note that we have from equation (6), that we must have $\Lambda_t^{(r)} < 1$, from which it follows that $\mathbb{E}\log\Lambda_t^{(r)} < 0$. Also, it is clear from the compactness of $\boldsymbol{\Xi}$, the Lipschitz continuity of $s(\cdot\ ;\boldsymbol{\psi})$ assumed in **IN2** and the form of $\partial\phi_t(\boldsymbol{g},\boldsymbol{\xi})/\partial\boldsymbol{g}$, see (C.1), that we must have $\mathbb{E}\log^+\Lambda_t^{(1)} < \infty$.

It now follows from Bougerol (1993, Theorem 3.1) that we have $\|\hat{\boldsymbol{g}}_t - \boldsymbol{g}_t\|^{\boldsymbol{\Xi}} \overset{a.s.}{\to} 0$ as $t \to \infty$. The uniqueness of the SE limit sequence $\{\boldsymbol{g}_t\}_{t\in\mathbb{Z}}$ and the exponentially fast convergence follows from Straumann and Mikosch (2006, Theorem 2.8).

Part (ii), filter invertibility: This follows straightforwardly, as $\hat{g}_t(\boldsymbol{\theta}) \equiv f_t - \hat{f}_t(\boldsymbol{\gamma}_0,\boldsymbol{\xi})$, so

$$|\hat{f}_t(\boldsymbol{\theta}_0) - f_t| = |\hat{g}_t(\boldsymbol{\theta}_0) - g_t(\boldsymbol{\theta}_0)| + |g_t(\boldsymbol{\theta}_0)| \overset{e.a.s.}{\to} 0 \, ,$$

as $t \to \infty$, where the first term vanishes e.a.s. by the first result of the proposition, and where the second term is equal to zero, because $g_t(\boldsymbol{\theta}_0) = 0$ almost surely, because we have

$$\begin{aligned}
g_{t+1}(\boldsymbol{\theta}_0) &= g_t(\boldsymbol{\theta}_0) \\
&\quad + \boldsymbol{\alpha}_0^\top [s(\boldsymbol{u}_t;\boldsymbol{\psi}_0) - s(\boldsymbol{\beta}_0 g_t(\boldsymbol{\theta}_0) - \boldsymbol{A}_1\boldsymbol{\beta}_0 g_{t-1}(\boldsymbol{\theta}_0) - \cdots - \boldsymbol{A}_p\boldsymbol{\beta}_0 g_{t-p}(\boldsymbol{\theta}) + \boldsymbol{u}_t;\boldsymbol{\psi}_0)] \, ,
\end{aligned}$$

so it is clear that $g_t(\boldsymbol{\theta}_0) = 0$ a.s. for all t, is a solution to this stochastic recurrence equation, and by the result above this solution is unique.

Part (iii), bounded moments: By the contraction condition in (6), there exists a constant $0 \leq \bar{c} < 1$ such that for all $(\bar{\boldsymbol{g}}_1,\bar{\boldsymbol{g}}_2)$ and any realisation of $(\boldsymbol{\varepsilon}_t,\ldots,\boldsymbol{\varepsilon}_{t-p-r+1})$:

$$\begin{aligned}
\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\phi_t^{(r)}(\bar{\boldsymbol{g}}_1,\boldsymbol{\xi}) - \phi_t^{(r)}(\bar{\boldsymbol{g}}_2,\boldsymbol{\xi})\| &\leq \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \sup_{\bar{\boldsymbol{g}}^*\in\mathbb{R}^{p+1}} \left\| \frac{\partial\phi_t^{(r)}(\bar{\boldsymbol{g}}^*,\cdot)}{\partial\boldsymbol{g}^\top} (\bar{\boldsymbol{g}}_1 - \bar{\boldsymbol{g}}_2) \right\| \\
&\leq \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \sup_{\bar{\boldsymbol{g}}^*\in\mathbb{R}^{p+1}} \left\| \frac{\partial\phi_t^{(r)}(\bar{\boldsymbol{g}}^*,\cdot)}{\partial\boldsymbol{g}^\top} \right\| \times \|\bar{\boldsymbol{g}}_1 - \bar{\boldsymbol{g}}_2\|
\end{aligned}$$

$$\leq \bar{c} \ \|\bar{\boldsymbol{g}}_1 - \bar{\boldsymbol{g}}_2\| \ ,$$

where the first inequality follows from the mean value theorem. The second inequality follows from the fact that for vectors $\|\cdot\|$ is an L^p-norm and for matrices it is the operator norm induced by this L^p-norm. Now let $\bar{\boldsymbol{g}} \in \mathbb{R}^{p+1}$, then we can bound $\|\boldsymbol{g}_t(\cdot)\|^{\boldsymbol{\Xi}}$ as follows

$$\begin{aligned}
\|\boldsymbol{g}_t(\cdot)\|^{\boldsymbol{\Xi}} &= \|\tilde{\phi}^{(r)}_{t-1}(\boldsymbol{g}_{t-r}(\cdot),\cdot)\|^{\boldsymbol{\Xi}} \\
&\leq \|\tilde{\phi}^{(r)}_{t-1}(\boldsymbol{g}_{t-r}(\cdot),\cdot) - \phi^{(r)}_{t-1}(\bar{\boldsymbol{g}},\cdot)\|^{\boldsymbol{\Xi}} + \|\phi^{(r)}_{t-1}(\bar{\boldsymbol{g}},\cdot)\|^{\boldsymbol{\Xi}} \\
&= \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\phi^{(r)}_{t-1}(\boldsymbol{g}_{t-r}(\boldsymbol{\xi}),\boldsymbol{\xi}) - \phi^{(r)}_{t-1}(\bar{\boldsymbol{g}},\boldsymbol{\xi})\| + \|\phi^{(r)}_{t-1}(\bar{\boldsymbol{g}},\cdot)\|^{\boldsymbol{\Xi}} \\
&\leq \bar{c}\cdot \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\boldsymbol{g}_{t-r}(\boldsymbol{\xi}) - \bar{\boldsymbol{g}}\| + \|\phi^{(r)}_{t-1}(\bar{\boldsymbol{g}},\cdot)\|^{\boldsymbol{\Xi}} \\
&\leq \bar{c}\cdot \|\boldsymbol{g}_{t-r}(\cdot)\|^{\boldsymbol{\Xi}} + \bar{c}\ \|\bar{\boldsymbol{g}}\| + \|\phi^{(r)}_{t-1}(\bar{\boldsymbol{g}},\cdot)\|^{\boldsymbol{\Xi}} .
\end{aligned}$$

So if we unfold this recursion l steps backwards, we obtain

$$\begin{aligned}
\|\boldsymbol{g}_t(\cdot)\|^{\boldsymbol{\Xi}} &\leq (\bar{c})^l \ \|\boldsymbol{g}_{t-lr}(\cdot)\|^{\boldsymbol{\Xi}} + \sum_{i=1}^{l} \bar{c}^{i-1} \ ((\bar{c}) \ \|\bar{\boldsymbol{g}}\| + \|\phi^{(r)}_{t-1-ir}(\bar{\boldsymbol{g}},\cdot)\|^{\boldsymbol{\Xi}}) \\
&\leq 1 + \sum_{i=1}^{l}(\bar{c})^{i-1} \ (\bar{c}\ \|\bar{\boldsymbol{g}}\| + \|\phi^{(r)}_{t-1-ir}(\bar{\boldsymbol{g}},\cdot)\|^{\boldsymbol{\Xi}}) ,
\end{aligned}$$

where the second inequality holds almost surely for large enough l, because $(\bar{c})^l$ goes to zero at an exponential rate and $\|\boldsymbol{g}_{t-lr}(\cdot)\|^{\boldsymbol{\Xi}}$ is SE by the first part of the proposition and Proposition 4.3 of Krengel (1985). Now let $\|\cdot\|_n \equiv (\mathbb{E}(\cdot)^n)^{1/n}$ and let $\|\cdot\|_n^{\boldsymbol{\Xi}} \equiv (\mathbb{E}(\|\cdot\|^{\boldsymbol{\Xi}})^n)^{1/n}$. Note that for $n \geq 1$, $\|\cdot\|_n$ is a norm and therefore it is sub-additive. For l large enough such that the above inequality holds with probability one, take this norm $\|\cdot\|_n$ for $n = 2$ on both sides:

$$\begin{aligned}
\|\boldsymbol{g}_t(\cdot)\|_2^{\boldsymbol{\Xi}} &\leq 1 + \sum_{i=1}^{l}(\bar{c})^{i-1} \ (\bar{c}\ \|\bar{\boldsymbol{g}}\| + \|\phi^{(r)}_{t-1-ir}(\bar{\boldsymbol{g}},\cdot)\|_2^{\boldsymbol{\Xi}}) \\
&\leq 1 + \frac{\bar{c}\ \|\bar{\boldsymbol{g}}\| + \|\phi^{(r)}_0(\bar{\boldsymbol{g}},\cdot)\|_2^{\boldsymbol{\Xi}}}{1-\bar{c}} < \infty ,
\end{aligned}$$

using that $\bar{c} < 1$ and $\|\phi^{(r)}_t(\bar{\boldsymbol{g}},\cdot)\|_2^{\boldsymbol{\Xi}}$ is constant over t by the stationarity of $\{\boldsymbol{\varepsilon}_t\}$ and it is finite by Lemma B.2. Hence, $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}(\|\boldsymbol{g}_t(\boldsymbol{\xi})\|)^2 < \infty$, which clearly also implies that $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|g_t(\boldsymbol{\xi})|^2 < \infty$.

□

Proof of Proposition 2. To show the result, we rewrite the process $\{\hat{f}_t\}$ in vector form such that the updating mechanism is a first-order stochastic recurrence equation. We write $\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}) = (\hat{f}_t(\boldsymbol{\theta}), \hat{f}_{t-1}(\boldsymbol{\theta}), \ldots, \hat{f}_{t-p}(\boldsymbol{\theta}))^\top$, which then follows the stochastic recurrence equation:

$$\hat{\boldsymbol{f}}_{t+1}(\boldsymbol{\theta}) = \breve{\phi}_t(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}),\boldsymbol{\theta}) ,$$

for $t = 1, 2, \ldots$, initialized at $\hat{\boldsymbol{f}}_1 = (\hat{f}_1, 0, \ldots, 0)^\top$ for some $\hat{f}_1 \in \mathbb{R}$ and where $\breve{\phi}_t : \mathbb{R}^p \times \Theta \to \mathbb{R}^p$ is a random function, defined by

$$\begin{aligned}
&\breve{\phi}_t(\boldsymbol{f}, \boldsymbol{\theta}) \\
&\quad = \begin{pmatrix} \omega + f_1 + \boldsymbol{\alpha}^\top s(\boldsymbol{y}_t - \boldsymbol{\mu} - \boldsymbol{\beta} f_1 - \boldsymbol{A}_1(\boldsymbol{y}_{t-1} - \boldsymbol{\mu} - \boldsymbol{\beta} f_2) - \ldots - \boldsymbol{A}_p(\boldsymbol{y}_{t-p} - \boldsymbol{\mu} - \boldsymbol{\beta} f_{p+1})) \\ f_1 \\ \vdots \\ f_p \end{pmatrix},
\end{aligned}$$

with $\boldsymbol{f} = (f_1, \ldots, f_{p+1})^\top$ and where we set $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{1-p}$ equal to zero. Furthermore, we define $\tilde{\boldsymbol{f}}_t(\boldsymbol{\theta})$ which is also defined as $\tilde{\boldsymbol{f}}_{t+1}(\boldsymbol{\theta}) = \breve{\phi}_t(\tilde{\boldsymbol{f}}_t(\boldsymbol{\theta}), \boldsymbol{\theta})$, but then initialized at $\tilde{\boldsymbol{f}}_1 = (\tilde{f}_1, 0, \ldots, 0)^\top$ for some $\tilde{f}_1 \in \mathbb{R}$. Let $\breve{\phi}_t^{(r)}$ denote the r-th convolution of $\breve{\phi}_t(\boldsymbol{f}, \boldsymbol{\theta})$, so $\breve{\phi}_t^{(r)}(\cdot, \boldsymbol{\theta}) = \breve{\phi}_t(\cdot, \boldsymbol{\theta}) \circ \cdots \circ \breve{\phi}_{t-r+1}(\cdot, \boldsymbol{\theta})$. Then it follows from the mean value theorem and the sub-multiplicativity of the operator norm that

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}) - \tilde{\boldsymbol{f}}_t(\boldsymbol{\theta})\| \\
&\qquad = \sup_{\boldsymbol{\theta} \in \Theta} \left\| \breve{\phi}_{t-1}^{(r)}(\hat{\boldsymbol{f}}_{t-r}(\boldsymbol{\theta}), \boldsymbol{\theta}) - \breve{\phi}_{t-1}^{(r)}(\tilde{\boldsymbol{f}}_{t-r}(\boldsymbol{\theta}), \boldsymbol{\theta}) \right\| \\
&\qquad \leq \sup_{\boldsymbol{\theta} \in \Theta} \sup_{\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^{p+1}, \boldsymbol{f}_1 \neq \boldsymbol{f}_2} \frac{\left\| \breve{\phi}_{t-1}^{(r)}(\boldsymbol{f}_1, \boldsymbol{\theta}) - \breve{\phi}_{t-1}^{(r)}(\boldsymbol{f}_2, \boldsymbol{\theta}) \right\|}{\|\boldsymbol{f}_1 - \boldsymbol{f}_2\|} \|\hat{\boldsymbol{f}}_{t-r}(\boldsymbol{\theta}) - \tilde{\boldsymbol{f}}_{t-r}(\boldsymbol{\theta})\| \\
&\qquad \leq \sup_{\boldsymbol{\theta} \in \Theta, \boldsymbol{f}^* \in \mathbb{R}^{p+1}} \left\| \frac{\partial \breve{\phi}_{t-1}^{(r)}(\boldsymbol{f}^*, \boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} \right\| \cdot \sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{f}}_{t-r}(\boldsymbol{\theta}) - \tilde{\boldsymbol{f}}_{t-r}(\boldsymbol{\theta})\| \\
&\qquad \leq b \cdot \sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{f}}_{t-r}(\boldsymbol{\theta}) - \tilde{\boldsymbol{f}}_{t-r}(\boldsymbol{\theta})\|,
\end{aligned}$$

almost surely, where b is some constant $0 \leq b < 1$, as we have by the chain rule that:

$$\left\| \frac{\partial \breve{\phi}_t^{(r)}(\boldsymbol{f}^*, \boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} \right\| = \left\| \left. \frac{\partial \breve{\phi}_t(\boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} \right|_{\boldsymbol{f} = \breve{\phi}_{t-1}^{(p)}(\boldsymbol{f}^*, \boldsymbol{\theta})} \cdots \frac{\partial \breve{\phi}_{t-p}(\boldsymbol{f}^*, \boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} \right\|,$$

where

$$\frac{\partial \breve{\phi}_t(\boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} = \begin{pmatrix} 1 - \boldsymbol{\alpha}^\top s'(\boldsymbol{z}; \boldsymbol{\psi})\boldsymbol{\beta} & \boldsymbol{\alpha}^\top s'(\boldsymbol{z}; \boldsymbol{\psi})\boldsymbol{A}_1\boldsymbol{\beta} & \ldots & \boldsymbol{\alpha}^\top s'(\boldsymbol{z}; \boldsymbol{\psi})\boldsymbol{A}_p\boldsymbol{\beta} \\ & & & 0 \\ & I_p & & \vdots \\ & & & 0 \end{pmatrix} \equiv \breve{\Phi}(\boldsymbol{z}, \boldsymbol{\theta}),$$

and where $\boldsymbol{z} = \boldsymbol{y}_t - \boldsymbol{\mu} - \boldsymbol{\beta} f_1 - \boldsymbol{A}_1(\boldsymbol{y}_{t-1} - \boldsymbol{\mu} - \boldsymbol{\beta} f_2) - \cdots - \boldsymbol{A}_p(\boldsymbol{y}_{t-p} - \boldsymbol{\mu} - \boldsymbol{\beta} f_{p+1})$. So it is clear that taking the supremum over $\boldsymbol{f}$ and $\boldsymbol{y}_t, \ldots \boldsymbol{y}_{t-p-r+1}$, is equivalent to taking the supremum over $\boldsymbol{z} \in \mathbb{R}^k$ for each factor of the matrix product, by the same reasoning as in Lemma B.1. In other words for any $t \geq r$:

$$\sup_{\boldsymbol{\theta} \in \Theta, \boldsymbol{f}^* \in \mathbb{R}^{p+1}, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t-p-r-1} \in \mathbb{R}^k} \left\| \frac{\partial \breve{\phi}_t^{(r)}(\boldsymbol{f}^*, \boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} \right\| = \sup_{\boldsymbol{\theta} \in \Theta, (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_r) \in \mathbb{R}^{kr}} \left\| \prod_{i=1}^{r} \breve{\Phi}(\boldsymbol{z}_i, \boldsymbol{\theta}) \right\| \leq b < 1,$$

where this number b exists because the condition in (8) holds by assumption. It is also clear that (8) can only hold if

$$\sup_{\boldsymbol{\theta}\in\Theta,\boldsymbol{z}\in\mathbb{R}^k} \left\|\breve{\Phi}(\boldsymbol{z},\boldsymbol{\theta})\right\| = a < \infty\,, \tag{A.1}$$

for some real number $1 \leq a < \infty$, because if this condition is violated, it means that not all elements of the matrix are uniformly bounded. It follows that for any $t \geq r+1$ we can unfold backwards the inequality we found above $\lfloor (t-1)/r \rfloor$ times, which leads to the inequality

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}) - \tilde{\boldsymbol{f}}_t(\boldsymbol{\theta})\| &\leq b^{\lfloor (t-1)/r \rfloor} \sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{f}}_{t-r\lfloor (t-1)/r \rfloor}(\boldsymbol{\theta}) - \tilde{\boldsymbol{f}}_{t-r\lfloor (t-1)/r \rfloor}(\boldsymbol{\theta})\|\,,\\ &\leq b^{\lfloor (t-1)/r \rfloor} a^{r-1} \|\hat{\boldsymbol{f}}_1 - \tilde{\boldsymbol{f}}_1\|\,,\end{aligned}$$

almost surely, where for the second inequality we unfold $\hat{\boldsymbol{f}}_{t-r\lfloor (t-1)/r \rfloor}(\boldsymbol{\theta}) - \tilde{\boldsymbol{f}}_{t-r\lfloor (t-1)/r \rfloor}(\boldsymbol{\theta})$ backwards further one step at a time, applying the mean value theorem again for each step and using the equality in (A.1). Lastly, we use that the 'rest term' $t - r(\lfloor (t-1)/r \rfloor) - 1 = (t-1) \bmod r$ is at most $r-1$. We can further bound this final expression as follows:

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}) - \tilde{\boldsymbol{f}}_t(\boldsymbol{\theta})\| &\leq b^{(t-r-1)/r} a^{r-1} \|\hat{\boldsymbol{f}}_1 - \tilde{\boldsymbol{f}}_1\|\\ &= (b^{1/r})^t\, b^{(1-r)/r} a^{r-1} \|\hat{\boldsymbol{f}}_1 - \tilde{\boldsymbol{f}}_1\|\,,\end{aligned}$$

where the first inequality holds because $\lfloor (t-1)/r \rfloor \geq (t-r-1)/r$, so $b^{\lfloor (t-1)/r \rfloor} \leq b^{(t-r-1)/r}$ as $0 \leq b < 1$. Notice that $b^{1/r} < 1$ because $b < 1$ and $r \geq 1$, so it follows that for any real number $1 < \rho < b^{-1/r}$:

$$\rho^t \sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}) - \tilde{\boldsymbol{f}}_t(\boldsymbol{\theta})\| \leq \rho^t (b^{1/r})^t\, b^{-2} a^{r-1} \|\hat{\boldsymbol{f}}_1 - \tilde{\boldsymbol{f}}_1\| \overset{a.s.}{\to} 0\,,$$

as $t \to \infty$, which follows from $\rho{\cdot}b^{1/r} < 1$ and the fact that the other terms are finite constants. In effect, also the first elements of the vector $\hat{\boldsymbol{f}}_t$ and $\tilde{\boldsymbol{f}}_t$ must converge to each other exponentially fast almost surely. □

Proof of Corollary 1. Note that for a matrix $\boldsymbol{A} = (a_{ij})$, $\|\boldsymbol{A}\|_\infty$ is the maximum absolute row sum of $\boldsymbol{A}$: $\|\boldsymbol{A}\|_\infty = \max_i \sum_j |a_{ij}|$. We introduce the notation:

$$\begin{pmatrix} 1-\boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t)\boldsymbol{\beta} & \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t)\boldsymbol{A}_1\boldsymbol{\beta} & \dots & \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t)\boldsymbol{A}_p\boldsymbol{\beta}\\ & & & 0\\ & I_p & & \vdots\\ & & & 0 \end{pmatrix} \equiv \begin{pmatrix} a_1(\boldsymbol{z}_t) & a_2(\boldsymbol{z}_t) & \dots & a_{p+1}(\boldsymbol{z}_t)\\ & & & 0\\ & I_p & & \vdots\\ & & & 0 \end{pmatrix},$$

where we supress the dependence of $a_i(\boldsymbol{z})$ on $\boldsymbol{\theta}$ for notational convenience. We want to show that under the assumptions of the corollary we have

$$\sup_{\boldsymbol{\theta}\in\Theta,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{p+1}\in\mathbb{R}^k}\left\|\begin{pmatrix} a_1(\boldsymbol{z}_1) & \ldots & a_{p+1}(\boldsymbol{z}_1) \\ & & 0 \\ & I_p & \vdots \\ & & 0 \end{pmatrix}\cdots\begin{pmatrix} a_1(\boldsymbol{z}_{p+1}) & \ldots & a_{p+1}(\boldsymbol{z}_{p+1}) \\ & & 0 \\ & I_p & \vdots \\ & & 0 \end{pmatrix}\right\|_\infty < 1\,.$$

We now argue that the inequality holds because each row of the resulting matrix can be shown to have an absolute sum smaller than 1, uniformly over $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{p+1}$. It follows from Condition (9) that $\sup_{\boldsymbol{z}\in\mathbb{R}^k,\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{p+1}|a_i(\boldsymbol{z})|<1$. Then, the i-th row of the product of the first i matrices in the product is equal to $(a_1(\boldsymbol{z}_i) \;\ldots\; a_{p+1}(\boldsymbol{z}_i))$, which clearly has an absolute sum smaller than 1 uniformly over $\boldsymbol{z}_i$ and $\boldsymbol{\theta}$, for each $i=1,\ldots,p+1$. Also, it follow that for any $(p+1)$-dimensional row vector $(b_1(\boldsymbol{x}),\ldots,b_{p+1}(\boldsymbol{x}))$ with an absolute sum smaller than 1 uniformly over $\boldsymbol{x}$ and $\boldsymbol{\theta}$, with $\boldsymbol{x}\in\mathbb{R}^{mk}$ for some $m\in\mathbb{N}$,

$$\begin{aligned}&\Big(b_1(\boldsymbol{x}) \quad \ldots \quad b_{p+1}(\boldsymbol{x})\Big)\begin{pmatrix} a_1(\boldsymbol{z}) & \ldots & a_p(\boldsymbol{z}) & a_{p+1}(\boldsymbol{z}) \\ & & & 0 \\ & I_p & & \vdots \\ & & & 0 \end{pmatrix}\\ &= \Big(b_1(\boldsymbol{x})a_1(\boldsymbol{z})+b_2(\boldsymbol{x}) \quad \ldots \quad b_1(\boldsymbol{x})a_p(\boldsymbol{z})+b_{p+1}(\boldsymbol{x}) \quad b_1(\boldsymbol{x})a_{p+1}(\boldsymbol{z})\Big)\,,\end{aligned}$$

again is absolutely summable as

$$\begin{aligned}&\sup_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta}}\{|b_1(\boldsymbol{x})a_1(\boldsymbol{z})+b_2(\boldsymbol{x})|+|b_1(\boldsymbol{x})a_2(\boldsymbol{z})+b_3(\boldsymbol{x})|+\ldots\\ &\qquad\qquad +|b_1(\boldsymbol{x})a_p(\boldsymbol{z})+b_{p+1}(\boldsymbol{x})|+|b_1(\boldsymbol{x})a_{p+1}(\boldsymbol{z})|\}\\ &\le\sup_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta}}\left\{|b_1(\boldsymbol{x})|\sum_{i=1}^{p+1}|a_i(\boldsymbol{z})|+\sum_{i=2}^{p+1}|b_i(\boldsymbol{x})|\right\}\\ &\le\sup_{\boldsymbol{x},\boldsymbol{\theta}}\sum_{i=1}^{p+1}|b_i(\boldsymbol{x})|<1\,.\end{aligned}$$

The claim above is true since, as we have argued above, the i-th row of the product of the first i matrices will have an absolute sum smaller than 1, so after further multiplications these rows will remain having an absolute sum smaller than one. This leads to the product of the $p+1$ matrices having a maximum absolute row sum smaller than 1 uniformly, which completes the proof. □

Proof of Corollary 2. Under the conditions of the corollary, letting $s_i'(z;\boldsymbol{\psi})$ denote the i-th diagonal element of the matrix $s'(z;\boldsymbol{\psi})$, there exist real numbers $0<\underline{d}<\bar{d}<1$ such that

$\underline{d} \leq \boldsymbol{\alpha}^\top s'(z;\boldsymbol{\psi})\boldsymbol{\beta} = \sum_{i=1}^k \alpha_i\beta_i s_i'(z;\boldsymbol{\psi}) \leq \bar{d}$ for any $\boldsymbol{\theta} \in \Theta$ and $z_i \in \mathbb{R}$. Furthermore, there must be a real number $0 < \bar{c} < 1$ such that $\sum_{i=1}^p |\boldsymbol{A}_{i,jj}| \leq \bar{c}$ for each $j = 1, \ldots, k$ and each $\boldsymbol{\theta} \in \Theta$. Hence

$$\begin{aligned}
&\sup_{z\in\mathbb{R}^k, \boldsymbol{\theta}\in\Theta} \Big\{ |1 - \boldsymbol{\alpha}^\top s'(z;\boldsymbol{\psi})\boldsymbol{\beta}| + |\boldsymbol{\alpha}^\top s'(z;\boldsymbol{\psi})\boldsymbol{A}_1\boldsymbol{\beta}| + \cdots + |\boldsymbol{\alpha}^\top s'(z;\boldsymbol{\psi})\boldsymbol{A}_p\boldsymbol{\beta}| \Big\} \\
&\leq \sup_{z\in\mathbb{R}^k, \boldsymbol{\theta}\in\Theta} \left\{ 1 - \sum_{i=1}^k \alpha_i\beta_i s_i'(z;\boldsymbol{\psi}) + \sum_{i=1}^k \alpha_i\beta_i s_i'(z;\boldsymbol{\psi})|\boldsymbol{A}_{1,ii}| + \cdots + \sum_{i=1}^k \alpha_i\beta_i s_i'(z;\boldsymbol{\psi})|\boldsymbol{A}_{p,ii}| \right\} \\
&= 1 - \inf_{z\in\mathbb{R}^k, \boldsymbol{\theta}\in\Theta} \sum_{i=1}^k \alpha_i\beta_i s_i'(z;\boldsymbol{\psi})(1 - |\boldsymbol{A}_{1,ii}| - \cdots - |\boldsymbol{A}_{p,ii}|) \\
&\leq 1 - (1-\bar{c}) \inf_{z\in\mathbb{R}^k, \boldsymbol{\theta}\in\Theta} \sum_{i=1}^k \alpha_i\beta_i s_i'(z;\boldsymbol{\psi}) \\
&\leq 1 - (1-\bar{c})\ \underline{d} < 1\,,
\end{aligned}$$

as $0 < 1 - \bar{c} < 1$ and $0 < \underline{d} < 1$. The result now follows from an application of Corollary 1. □

Proof of Theorem 2. Because the parameter space $\boldsymbol{\Xi}$ is compact, it follows from standard arguments, see e.g. Wald (1949), that it is sufficient to show that

(a) The criterion function converges to some continuous deterministic limit function uniformly over $\boldsymbol{\Xi}$: $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} |\widehat{L}_T(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - L(\boldsymbol{\gamma}_0, \boldsymbol{\xi})| \overset{a.s.}{\to} 0$ as $T \to \infty$.

(b) The true parameter $\boldsymbol{\xi}_0$ is the unique maximizer of the limit criterion: $L(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) < L(\boldsymbol{\gamma}_0, \boldsymbol{\xi}_0)$ for any $\boldsymbol{\xi} \in \boldsymbol{\Xi}$, $\boldsymbol{\xi} \neq \boldsymbol{\xi}_0$.

To show condition (a) holds, we first notice that by the triangle inequality:

$$|\widehat{L}_T(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - L(\boldsymbol{\gamma}_0, \boldsymbol{\xi})| \leq |\widehat{L}_T(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - L_T(\boldsymbol{\gamma}_0, \boldsymbol{\xi})| + |L_T(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - L(\boldsymbol{\gamma}_0, \boldsymbol{\xi})|\,,$$

where L_T denotes the quasi log likelihood evaluated using the limit prediction errors $g_t(\boldsymbol{\xi})$, i.e. based on $\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi})$, which is defined as $\hat{\ell}_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi})$ but then with $\boldsymbol{y}_t - \boldsymbol{\mu}_0 - \boldsymbol{\beta}_0 \hat{f}_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) = \boldsymbol{y}_t - \boldsymbol{\mu}_0 + \boldsymbol{\beta}_0(\hat{g}_t(\boldsymbol{\xi}) - f_t) = \boldsymbol{\beta}_0 \hat{g}_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t$ replaced by $\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t$. Here $L(\boldsymbol{\gamma}_0, \boldsymbol{\xi})$ will be the expected value of $\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}_0)$. It is shown in Lemmas B.3 and B.4, respectively, that the two terms on the right-hand side of this inequality converge to zero almost surely uniformly over $\boldsymbol{\Xi}$.

Lastly, condition (b) holds by Lemma B.5. □

Proof of Theorem 3. For ease of exposition, we start by proving the result for the model without the parameter $\boldsymbol{\mu}$ (or equivalently, for the case where $\boldsymbol{\mu}_0$ is known), such that $\boldsymbol{\gamma}$ only consists of $\boldsymbol{b}$.

(Result 1: $\gamma = \boldsymbol{b}$) For this part of the proof we will use some abuse of notation, by filling in $\boldsymbol{b}$ in the place of $\boldsymbol{\gamma}$ in functions such as $\widehat{L}_T(\boldsymbol{\gamma}, \boldsymbol{\xi})$ and $\hat{f}_t(\boldsymbol{\gamma}, \boldsymbol{\xi})$, such that it is clear that we do not consider $\boldsymbol{\mu}$.

As in the proof of Theorem 2, because the parameter space Ξ is compact, it is sufficient to show that

(a) The criterion function converges to some continuous deterministic limit function uniformly over Ξ: $\sup_{\boldsymbol{\xi}\in\Xi} |\widehat{L}_T(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}) - L(\boldsymbol{b}_0, \boldsymbol{\xi})| \xrightarrow{p} 0$ as $T \to \infty$.

(b) The true parameter $\boldsymbol{\xi}_0$ is the unique maximizer of the limit criterion: $L(\boldsymbol{b}_0, \boldsymbol{\xi}) < L(\boldsymbol{b}_0, \boldsymbol{\xi}_0)$ for any $\boldsymbol{\xi} \in \Xi$, $\boldsymbol{\xi} \neq \boldsymbol{\xi}_0$.

For condition (a), we can use that by the triangle inequality:

$$|\widehat{L}_T(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}) - L(\boldsymbol{b}_0, \boldsymbol{\xi})| \leq |\widehat{L}_T(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}) - \widehat{L}_T(\boldsymbol{b}_0, \boldsymbol{\xi})| + |\widehat{L}_T(\boldsymbol{b}_0, \boldsymbol{\xi}) - L(\boldsymbol{b}_0, \boldsymbol{\xi})| \,.$$

We have shown in the proof of Theorem 2 that the second term vanishes almost surely uniformly over Ξ. By Lemma B.6, the first term also converges to zero in probability uniformly over Ξ. Condition (b) again holds by Lemma B.5. This completes the proof for the case where $\boldsymbol{\mu}$ is excluded from the model.

(Result 2: $\boldsymbol{\gamma} = (\boldsymbol{b}^\top, \boldsymbol{\mu}^\top)^\top$) We just give a sketch of the proof, because the same approach as for the first result can be used. The same steps can be used, such that it just remains to be shown that

$$\sup_{\boldsymbol{\xi}\in\Xi} \left| \widehat{L}_T(\widehat{\boldsymbol{\gamma}}_T, \boldsymbol{\xi}) - \widehat{L}_T(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) \right| \xrightarrow{p} 0 \,, \tag{A.2}$$

as $T \to \infty$. To show this, Lemma B.6 can be extended. In particular, using the same approach as in the proof of Lemma B.8 we can write

$$\begin{aligned}
&\hat{f}_{t+1}(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - \hat{f}_{t+1}(\boldsymbol{\gamma}, \boldsymbol{\xi}) \\
&\quad = \left(1 - \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*; \boldsymbol{\psi})\boldsymbol{\beta}\right) (\hat{f}_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{\gamma}, \boldsymbol{\xi})) \\
&\qquad + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*; \boldsymbol{\psi}) \Big(\boldsymbol{A}_1 \boldsymbol{\beta} (\hat{f}_{t-1}(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - \hat{f}_{t-1}(\boldsymbol{\gamma}, \boldsymbol{\xi})) + \dots \\
&\qquad + \boldsymbol{A}_p \boldsymbol{\beta} (\hat{f}_{t-p}(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - \hat{f}_{t-p}(\boldsymbol{\gamma}, \boldsymbol{\xi})) \Big) \\
&\qquad + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*; \boldsymbol{\psi}) \boldsymbol{A}(L) \left[(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \hat{f}_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) + \boldsymbol{\mu} - \boldsymbol{\mu}_0 \right] .
\end{aligned}$$

We are again interested in bounding the expectation of the square of $\hat{f}_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{\gamma}, \boldsymbol{\xi})$, so if we treat $\boldsymbol{z}_t^*$ as deterministic values, it follows from the linearity of the updating equation above, that we can decompose it as follows: $\hat{f}_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - f_t(\boldsymbol{\gamma}, \boldsymbol{\xi}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \hat{\boldsymbol{h}}_{1t}(\boldsymbol{\theta}) + (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \hat{\boldsymbol{h}}_{2t}(\boldsymbol{\theta})$, similarly as in the proof of Lemma B.8. More specifically, the definition of $\hat{\boldsymbol{h}}_{1t}(\boldsymbol{\theta})$ will be virtually the same as that of $\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})$ of this lemma, although the values of $\boldsymbol{z}_t^*$ will be different here and the innovation term is based on $\hat{f}_t(\gamma_0,\boldsymbol{\xi})$ instead of $\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})$. The process $\{\hat{\boldsymbol{h}}_{2t}(\boldsymbol{\theta})\}$ will be initialized at $\hat{\boldsymbol{h}}_{2,1}=0$ and is updated by

$$\begin{aligned}\hat{\boldsymbol{h}}_{2,t+1}(\boldsymbol{\theta}) = (1-\boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{\beta})\hat{\boldsymbol{h}}_{2t}(\boldsymbol{\theta}) + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}_1\boldsymbol{\beta}\hat{\boldsymbol{h}}_{2,t-1}(\boldsymbol{\theta}) + \dots \\ + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}_p\boldsymbol{\beta}\hat{\boldsymbol{h}}_{2,t-p}(\boldsymbol{\theta}) + \boldsymbol{A}(1)s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})^\top\boldsymbol{\alpha}\,.\end{aligned}$$

It then follows automatically that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_{1t}(\boldsymbol{\theta})\|^2$ can be bounded in the same way as $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_{t}(\boldsymbol{\theta})\|^2$. It is also not hard to show that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_{2t}(\boldsymbol{\theta})\|$ can be bounded by some constant c, using a similar approach as in Lemma B.8 and using that $\boldsymbol{A}(1)s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})^\top\boldsymbol{\alpha}$ is uniformly bounded by a constant as $s(\cdot,\boldsymbol{\psi})$ is Lipschitz continuous uniformly over $\boldsymbol{\psi}\in\boldsymbol{\Psi}$ and the parameter space Θ is compact. Using these results, it can then be shown that (A.2) holds in the same way as in the proof of Lemma B.4, using the assumptions on the rate of convergence of $\|\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_0\|$ and $\|\widehat{\boldsymbol{\mu}}_T-\boldsymbol{\mu}_0\|$. □

Proof of Proposition 3. It follows from the triangle inequality, and $\hat{g}_t(\boldsymbol{\xi})\equiv f_t-\hat{f}_t(\gamma_0,\boldsymbol{\xi})$ that

$$\begin{aligned}|\hat{f}_{T+1}(\widehat{\gamma}_T,\widehat{\boldsymbol{\xi}}_T(\widehat{\gamma}_T))-f_{T+1}| \le{}& \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_{T+1}(\widehat{\gamma}_T,\boldsymbol{\xi})-\hat{f}_{T+1}(\gamma_0,\boldsymbol{\xi})| \\ &+\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{g}_{T+1}(\boldsymbol{\xi})-g_{T+1}(\boldsymbol{\xi})| \\ &+|g_{T+1}(\widehat{\boldsymbol{\xi}}_T(\widehat{\gamma}_T))-g_{T+1}(\boldsymbol{\xi}_0)|\,,\end{aligned}\tag{A.3}$$

where $g_t(\boldsymbol{\xi}_0)=0$ almost surely, which follows directly from Proposition 1. We can show that every term on the right-hand side of this inequality converges to zero in probability as $T\to\infty$. For the first term, we can use the extended version if Lemma B.8 that we discuss in the proof of Theorem 3, by which we know that:

$$\begin{aligned}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_{T+1}(\widehat{\gamma}_T,\boldsymbol{\xi})-\hat{f}_{T+1}(\gamma_0,\boldsymbol{\xi})| &\le \sup_{\boldsymbol{\theta}\in\Theta}|(\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_0)^\top\hat{\boldsymbol{h}}_{1,t}(\boldsymbol{\theta})| + \sup_{\boldsymbol{\theta}\in\Theta}|(\widehat{\boldsymbol{\mu}}_T-\boldsymbol{\mu}_0)^\top\hat{\boldsymbol{h}}_{2,t}(\boldsymbol{\theta})| \\ &\le \|\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_0\|\cdot\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_{1,t}(\boldsymbol{\theta})\| + \|\widehat{\boldsymbol{\mu}}_T-\boldsymbol{\mu}_0\|\cdot\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_{2,t}(\boldsymbol{\theta})\|\,,\end{aligned}$$

where the first term is the product of either an $o_p(T^{-1/2})$ and an $O_p(\sqrt{T})$ term (in case $\omega_0=0$), or an $o_p(T^{-1})$ and an $O_p(T)$ term (in case $\omega_0\neq 0$). This implies that the product converges to zero in probability in both cases. The second term is the product of an $o_p(1)$ term and an $O_p(1)$ term. That $\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_{1,t}(\boldsymbol{\theta})\|$ is either $O_p(\sqrt{T})$ or $O_p(T)$, follows from Lemma B.8, an application of Jensen's inequality and an application of Markov's theorem, using a similar approach as in the proof of Lemma B.6. That $\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_{2,t}(\boldsymbol{\theta})\|$ is $O_p(1)$ can be argued by using that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_{2t}(\boldsymbol{\theta})\|$ is finite, as was argued in the proof of Theorem 3, and invoking Markov's theorem.

The second term on the right-hand side of (A.3) converges to zero exponentially fast almost surely as $T\to\infty$ by the result of Proposition 1.

Finally, for the last term of the right-hand side of (A.3), we use the definition of convergence in probability and the continuity of $g_t(\boldsymbol{\xi})$ in its argument. Define the notation $B_\varepsilon(\boldsymbol{\xi}_0) = \{\boldsymbol{\xi} \in \Xi; \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| < \varepsilon\}$ for any $\varepsilon > 0$. By the convergence in probability of $\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)$ to $\boldsymbol{\xi}_0$ as $T \to \infty$, we know that for any $\tilde{\varepsilon} > 0$ and $\tilde{\delta} > 0$, there exists an integer $N_{\tilde{\varepsilon},\tilde{\delta}}$ such that for any $T > N_{\tilde{\varepsilon},\tilde{\delta}}$, $\mathbb{P}(\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T) \notin B_{\tilde{\varepsilon}}(\boldsymbol{\xi}_0)) < \tilde{\delta}$. For any $\varepsilon > 0$, $\tilde{\varepsilon} > 0$, $\tilde{\delta} > 0$ and $T \geq N_{\tilde{\varepsilon},\tilde{\delta}}$ we then have that

$$\begin{aligned} \sup_{s \in \mathbb{N}} \mathbb{P}\left(|g_s(\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)) - g_s(\boldsymbol{\xi}_0)| > \varepsilon\right) &\leq \sup_{s \in \mathbb{N}} \mathbb{P}\left(\sup_{\boldsymbol{\xi} \in B_{\tilde{\varepsilon}}(\boldsymbol{\xi}_0)} |g_s(\boldsymbol{\xi}) - g_s(\boldsymbol{\xi}_0)| > \varepsilon\right) \\ &\quad + \mathbb{P}\left(\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T) \notin B_{\tilde{\varepsilon}}(\boldsymbol{\xi}_0)\right) \\ &< \mathbb{P}\left(\sup_{\boldsymbol{\xi} \in B_{\tilde{\varepsilon}}(\boldsymbol{\xi}_0)} |g_s(\boldsymbol{\xi}) - g_s(\boldsymbol{\xi}_0)| > \varepsilon\right) + \tilde{\delta}\,, \end{aligned}$$

where the supremum over s after the second inequality can be dropped because of the stationarity of $\{g_t\}$. Clearly, as $\tilde{\varepsilon}$ approaches zero, the set $B_{\tilde{\varepsilon}}(\boldsymbol{\xi}_0)$ converges to the set only containing $\boldsymbol{\xi}_0$. So due to the continuity of $g_t(\boldsymbol{\xi})$ in its argument, the probability on the right-hand side will go towards zero as $\tilde{\varepsilon}$ approaches zero. It is therefore clear that for any choice of $\varepsilon > 0$ and $\delta > 0$, there will exist some $\tilde{\varepsilon} > 0$ and $\tilde{\delta} > 0$ small enough such that $\mathbb{P}\left(\sup_{\boldsymbol{\xi} \in B_{\tilde{\varepsilon}}(\boldsymbol{\xi}_0)} |g_s(\boldsymbol{\xi}) - g_s(\boldsymbol{\xi}_0)| > \varepsilon\right) + \tilde{\delta} < \delta$. Combining this with the inequality derived above, it then follows that for any $T \geq N_{\tilde{\varepsilon},\tilde{\delta}}$, $\sup_{s \in \mathbb{N}} \mathbb{P}\left(|g_s(\widehat{\boldsymbol{\xi}}_T(\widehat{\boldsymbol{\gamma}}_T)) - g_s(\boldsymbol{\xi}_0)| > \varepsilon\right) < \delta$, from which it follows that the final term on the right-hand side of (A.3) converges to zero in probability.

□

B. Lemmas

The proofs of these lemmas can be found in Section C of this Supplementary Appendix.

Lemma B.1. *The equality in equation (6) in Proposition 1 holds.*

Lemma B.2. *Let the assumptions of Proposition 1 hold. Then, for any integer $h \geq 1$ and for any $\bar{\boldsymbol{g}} \in \mathbb{R}^{p+1}$, there exists a finite constant Z_h such that $\mathbb{E} \sup_{\boldsymbol{\xi} \in \Xi} \|\phi_t^{(h)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi})\|^2 \leq Z_h < \infty$.*

Lemma B.3. *Let the assumptions of Theorem 2 hold. Then*

$$\sup_{\boldsymbol{\xi} \in \Xi} |\widehat{L}_T(\gamma_0, \boldsymbol{\xi}) - L_T(\gamma_0, \boldsymbol{\xi})| \overset{a.s.}{\to} 0 \qquad \text{as } T \to \infty\,.$$

Lemma B.4. *Let the assumptions of Theorem 2 hold. Then*

$$\sup_{\boldsymbol{\xi} \in \Xi} |L_T(\gamma_0, \boldsymbol{\xi}) - \mathbb{E}\ell_t(\gamma_0, \boldsymbol{\xi})| \overset{a.s.}{\to} 0 \qquad \text{as } T \to \infty\,.$$

Lemma B.5. *Let the assumptions of Theorem 1 hold. Then for any $\boldsymbol{\xi} \in \Xi$ such that $\boldsymbol{\xi} \neq \boldsymbol{\xi}_0$, we have*

$$\mathbb{E}\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) < \mathbb{E}\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}_0)\,.$$

Lemma B.6. *Let the assumptions of Theorem 3 hold and suppose that $\boldsymbol{\mu}$ is excluded from the model. Then*

$$\sup_{\boldsymbol{\xi}\in\Xi} |\widehat{L}_T(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}) - \widehat{L}_T(\boldsymbol{b}_0, \boldsymbol{\xi})| \xrightarrow{p} 0 \qquad \text{as } T \to \infty\,.$$

Lemma B.7. *Let the assumptions of Theorem 3 be satisfied and let $\boldsymbol{\mu}$ be excluded from the model. Then there exists a finite constant b such that for any t*

$$\mathbb{E}\sup_{\boldsymbol{\xi}\in\Xi} |\hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi})|^2 \leq \begin{cases} bt & \text{if } \omega_0 = 0\,, \\ bt^2 & \text{if } \omega_0 \neq 0\,. \end{cases} \tag{B.1}$$

Lemma B.8. *Let the assumptions of Theorem 3 be satisfied and suppose that $\boldsymbol{\mu}$ is excluded from the model. Then there exists a k-variate process $\{\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$, such that $\hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b}, \boldsymbol{\xi}) \equiv (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \hat{\boldsymbol{h}}_t(\boldsymbol{\theta})$, and there exists a finite constant a, such that for any t*

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|^2 \leq \begin{cases} at & \text{if } \omega_0 = 0\,, \\ at^2 & \text{if } \omega_0 \neq 0\,. \end{cases}$$

C. Proofs of lemmas

Proof of Lemma B.1. The derivative of ϕ_t with respect to $\boldsymbol{g}$ takes the form:

$$\begin{aligned} \frac{\partial \phi_t(\boldsymbol{g}, \boldsymbol{\xi})}{\partial \boldsymbol{g}^\top} &= \left.\begin{pmatrix} 1 - \boldsymbol{\alpha}^\top s'(\boldsymbol{u}; \boldsymbol{\psi})\boldsymbol{\beta}_0 & \boldsymbol{\alpha}^\top s'(\boldsymbol{u}; \boldsymbol{\psi})\boldsymbol{A}_1\boldsymbol{\beta}_0 & \dots & \boldsymbol{\alpha}^\top s'(\boldsymbol{u}; \boldsymbol{\psi})\boldsymbol{A}_p\boldsymbol{\beta}_0 \\ & & & 0 \\ & I_p & & \vdots \\ & & & 0 \end{pmatrix}\right|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_t(\boldsymbol{g},\boldsymbol{\xi})} \\ &\equiv \Phi(\hat{\boldsymbol{u}}_t(\boldsymbol{g}, \boldsymbol{\xi}), \boldsymbol{\gamma}_0, \boldsymbol{\xi})\,, \end{aligned} \tag{C.1}$$

where $\hat{\boldsymbol{u}}_t : \mathbb{R}^{p+1} \times \Xi \to \mathbb{R}^k$ is a random function specified as $\hat{\boldsymbol{u}}_t(\boldsymbol{g}, \boldsymbol{\xi}) = \boldsymbol{\beta}_0 g_1 + \boldsymbol{\varepsilon}_t - \boldsymbol{A}_1(\boldsymbol{\beta}_0 g_2 + \boldsymbol{\varepsilon}_{t-1}) - \ldots - \boldsymbol{A}_p(\boldsymbol{\beta}_0 g_{p+1} + \boldsymbol{\varepsilon}_{t-p})$. So, by the chain rule we have that

$$\left\| \frac{\partial \phi_t^{(r)}(\boldsymbol{g}, \boldsymbol{\xi})}{\partial \boldsymbol{g}^\top} \right\| = \left\| \left.\frac{\partial \phi_t(\boldsymbol{g}^*, \boldsymbol{\xi})}{\partial \boldsymbol{g}^\top}\right|_{\boldsymbol{g}^*=\phi_{t-1}^{(r-1)}(\boldsymbol{g},\boldsymbol{\xi})} \cdots \frac{\partial \phi_{t-r+1}(\boldsymbol{g}, \boldsymbol{\xi})}{\partial \boldsymbol{g}^\top} \right\|$$

$$= \left\| \prod_{i=1}^{r} \Phi(\hat{\boldsymbol{u}}_{t-i+1}(\phi_{t-i}^{(r-i)}(\boldsymbol{g},\boldsymbol{\xi}),\boldsymbol{\xi}),\gamma_0,\boldsymbol{\xi}) \right\| .$$

Therefore, it is easy to verify that taking the supremum over $\boldsymbol{\varepsilon}_t,\dots,\boldsymbol{\varepsilon}_{t-p-r+1} \in \mathbb{R}^k$ of this quantity boils down to taking the supremum over $\hat{\boldsymbol{u}}_t,\dots,\hat{\boldsymbol{u}}_{t-r+1} \in \mathbb{R}^k$, or equivalently

$$\sup_{\boldsymbol{\varepsilon}_t,\dots,\boldsymbol{\varepsilon}_{t-p-r+1}\in\mathbb{R}^k} \left\| \frac{\partial \phi_t^{(r)}(\boldsymbol{g},\boldsymbol{\xi})}{\partial \boldsymbol{g}^\top} \right\| = \sup_{\boldsymbol{z}_1,\dots,\boldsymbol{z}_r\in\mathbb{R}^k} \left\| \prod_{i=1}^{r} \Phi(\boldsymbol{z}_i,\gamma_0,\boldsymbol{\xi}) \right\| .$$

□

Proof of Lemma B.2. We give a proof by induction on h.

Step 1. (Base case) We first show the claim holds for $h=1$. For any $\bar{\boldsymbol{g}} = (\bar{g}_1,\dots,\bar{g}_{p+1})^\top \in \mathbb{R}^{p+1}$ we have that

$$\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\phi_t(\bar{\boldsymbol{g}},\boldsymbol{\xi})\|^2 = \mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \left\| \begin{pmatrix} \phi_{1t}(\bar{\boldsymbol{g}},\boldsymbol{\xi}) \\ \bar{g}_1 \\ \vdots \\ \bar{g}_p \end{pmatrix} \right\|^2 ,$$

so it is clear that it suffices to show that $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} |\phi_{1t}(\bar{\boldsymbol{g}},\boldsymbol{\xi})|^2$ can be bounded by a finite constant. By the definition of ϕ_{1t} in (5) and the triangle inequality, we have the bound

$$\begin{aligned} |\phi_{1t}(\bar{\boldsymbol{g}},\boldsymbol{\xi})| \le |\bar{g}_1| + |\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)| + \Big| \boldsymbol{\alpha}^\top s(\boldsymbol{\beta}_0\bar{g}_1 + \boldsymbol{\varepsilon}_t - \boldsymbol{A}_1(\boldsymbol{\beta}_0\bar{g}_2 + \boldsymbol{\varepsilon}_{t-1}) - \\ \dots - \boldsymbol{A}_p(\boldsymbol{\beta}_0\bar{g}_{p+1} + \boldsymbol{\varepsilon}_{t-p});\boldsymbol{\psi}) \Big| + |\omega_0 - \omega| . \end{aligned} \tag{C.2}$$

Now we bound the expectation of the terms on the right-hand side uniformly over $\boldsymbol{\Xi}$. The first term is a finite constant and the last term is also finite uniformly over $\boldsymbol{\Xi}$ by the compactness of the parameter set. For the remaining two terms, we first bound the expectation of $\|s(\boldsymbol{z}_t;\boldsymbol{\psi})\|$ for a general vector $\boldsymbol{z}_t$. From **IN2** we have that the function $s(\cdot\,;\boldsymbol{\psi})$ is Lipschitz continuous uniformly on $\boldsymbol{\Psi}$. Hence, for any $\boldsymbol{z}_t, \bar{\boldsymbol{z}} \in \mathbb{R}^{p+1}$, we have

$$\begin{aligned} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|s(\boldsymbol{z}_t;\boldsymbol{\psi})\| &\le \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|s(\boldsymbol{z}_t;\boldsymbol{\psi}) - s(\bar{\boldsymbol{z}};\boldsymbol{\psi})\| + \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|s(\bar{\boldsymbol{z}};\boldsymbol{\psi})\| \\ &\le K\ \|\boldsymbol{z}_t - \bar{\boldsymbol{z}}\| + \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|s(\bar{\boldsymbol{z}};\boldsymbol{\psi})\| \\ &\le K\ \|\boldsymbol{z}_t\| + K\ \|\bar{\boldsymbol{z}}\| + \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|s(\bar{\boldsymbol{z}};\boldsymbol{\psi})\| . \end{aligned} \tag{C.3}$$

Starting with the simple term $\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)$, it follows that there is a finite constant W_1 such that

$$\begin{aligned} \mathbb{E}|\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)|^2 &\le C^2 \bar{\boldsymbol{\alpha}}_0^n\ \mathbb{E}\|s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)\|^2 \\ &\le C^2 \bar{\boldsymbol{\alpha}}_0^2\ (DK^n\ \mathbb{E}\|\boldsymbol{u}_t\|^2 + DK^n\|\bar{\boldsymbol{u}}\|^2 + D\|s(\bar{\boldsymbol{u}};\boldsymbol{\psi}_0)\|^2) \le W_1 < \infty , \end{aligned} \tag{C.4}$$

where the first inequality holds because

$$|\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)| \le \sum_{i=1}^{k} |\boldsymbol{\alpha}_{i0}| \cdot |s_i(\boldsymbol{u}_t;\boldsymbol{\psi}_0)| \le \bar{\boldsymbol{\alpha}}_0 \sum_{i=1}^{k} |s_i(\boldsymbol{u}_t;\boldsymbol{\psi}_0)|\,,$$

where $\bar{\boldsymbol{\alpha}}_0$ denotes the largest element of $\boldsymbol{\alpha}_0$ and where we use that by norm equivalence there exists a constant C such that $\sum_{i=1}^{k} |s_i(\boldsymbol{u}_t;\boldsymbol{\psi}_0)| \le C\|s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)\|$. The second inequality uses inequality (C.3) and the C_n inequality of Loève (1977), which says that for any $n \ge 0$, there exists a constant D such that $(a+b+c)^n \le D(a^n + b^n + c^n)$ for any $a, b, c \in [0,\infty)$. The final expression is finite, because $\|\bar{\boldsymbol{u}}\|$ and $\|s(\bar{\boldsymbol{u}};\boldsymbol{\psi}_0)\|$ are finite constants and $\mathbb{E}\|\boldsymbol{u}_t\|^2 < \infty$ by Assumption **A1**, as it is assumed that the covariance matrix $\boldsymbol{\Sigma}_0$ of $\boldsymbol{u}_t$ is finite.

Using very similar arguments, if we define $\boldsymbol{z}_t(\boldsymbol{\xi}) \equiv \boldsymbol{\beta}_0\bar{g}_1 + \boldsymbol{\varepsilon}_t - \boldsymbol{A}_1(\boldsymbol{\beta}_0\bar{g}_2 + \boldsymbol{\varepsilon}_{t-1}) - \ldots - \boldsymbol{A}_p(\boldsymbol{\beta}_0\bar{g}_{p+1} + \boldsymbol{\varepsilon}_{t-p})$, we have, for any vector $\bar{\boldsymbol{z}} \in \mathbb{R}^k$,

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} |\boldsymbol{\alpha}^\top s(\boldsymbol{z}_t(\boldsymbol{\xi});\boldsymbol{\psi})|^2 &\le C^2\bar{\boldsymbol{\alpha}}^2\ \mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|s(\boldsymbol{z}_t;\boldsymbol{\psi})\|^2 \\ &\le C^2\bar{\boldsymbol{\alpha}}^2\ (DK^2\ \mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\boldsymbol{z}_t(\boldsymbol{\xi})\|^2 + DK^2\|\bar{\boldsymbol{z}}\|^2 \\ &\qquad + D\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|s(\bar{\boldsymbol{z}};\boldsymbol{\psi})\|^2)\,,\end{aligned}$$

where $\bar{\boldsymbol{\alpha}} = \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \max_i |\boldsymbol{\alpha}_i|$ denotes the supremum over $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ of the largest element of the vector $\boldsymbol{\alpha}$ and C, K and D are the same as in (C.4). The term $\|\bar{\boldsymbol{z}}\|^2$ is clearly finite and $\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \|s(\bar{\boldsymbol{z}};\boldsymbol{\psi})\|^2$ is finite by the continuity and uniform Lipschitz contintuity over the compact set $\boldsymbol{\Psi}$ of the function $s(\cdot;\boldsymbol{\psi})$, from Assumption **IN2**. Finally, we there must exist a finite constant W_2 such that

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\boldsymbol{z}_t(\boldsymbol{\xi})\|^2 &= \mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\boldsymbol{\beta}_0\bar{g}_1 + \boldsymbol{\varepsilon}_t - \boldsymbol{A}_1(\boldsymbol{\beta}_0\bar{g}_2 + \boldsymbol{\varepsilon}_{t-1}) - \ldots - \boldsymbol{A}_p(\boldsymbol{\beta}_0\bar{g}_{p+1} + \boldsymbol{\varepsilon}_{t-p})\|^2 \\ &\le \tilde{D}R^2 + \tilde{D}(\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^2 + \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\boldsymbol{A}_1\|^2\ \mathbb{E}\|\boldsymbol{\varepsilon}_{t-1}\|^2 + \\ &\qquad \ldots + \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\boldsymbol{A}_p\|^2\ \mathbb{E}\|\boldsymbol{\varepsilon}_{t-p}\|^2) \le W_2 < \infty\,,\end{aligned}$$

where $\tilde{D}$ is a finite constant that exists by the C_n inequality of Loève (1977), and where the constant R contains the supremum over $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ of the deterministic part of $\boldsymbol{z}_t(\boldsymbol{\xi})$, which is finite because $\boldsymbol{\Xi}$ is compact and $\bar{\boldsymbol{g}} \in \mathbb{R}^{p+1}$. Finally $\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^2 < \infty$ because $\mathbb{E}\|\boldsymbol{u}_t\|^2 < \infty$ and $\boldsymbol{A}_0(L)$ is an invertible polynomial by Assumption **A2**.

By combining (C.2) with these results, it follows from another application of the C_n inequality that there must exist a finite constant Z_1 such that $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\phi_t(\bar{\boldsymbol{g}},\boldsymbol{\xi})\|^2 \le Z_1 < \infty$.

Step 2. (Induction Step) Say that for some integer $h \ge 1$, there exists a constant Z_h such that $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \|\phi_t^{(h)}(\bar{\boldsymbol{g}},\boldsymbol{\xi})\|^2 \le Z_h < \infty$, then we can show that such a bounding constant also exists for $h+1$. In particular, we have that $\phi_t^{(h+1)}(\bar{\boldsymbol{g}},\boldsymbol{\xi}) = \phi_t(\phi_{t-1}^{(h)}(\bar{\boldsymbol{g}},\boldsymbol{\xi}),\boldsymbol{\xi})$. By an application of the mean value theorem to the function $\phi_t(\cdot, \boldsymbol{\xi})$, we get that for any $\boldsymbol{g}^\dagger \in \mathbb{R}^{p+1}$

$$\begin{aligned} &\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \left\| \phi_t(\phi_{t-1}^{(h)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi}), \boldsymbol{\xi}) - \phi_t(\boldsymbol{g}^\dagger, \boldsymbol{\xi}) \right\| \\ &\qquad \leq \sup_{\boldsymbol{g}^* \in \mathbb{R}^{p+1}, \boldsymbol{\xi}\in\boldsymbol{\Xi}} \left\| \frac{\partial \phi_t(\boldsymbol{g}^*, \boldsymbol{\xi})}{\partial \boldsymbol{g}^\top} \right\| \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_{t-1}^{(h)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi}) - \boldsymbol{g}^\dagger \| \\ &\qquad \leq \bar{K} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_{t-1}^{(h)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi}) - \boldsymbol{g}^\dagger \| \,, \end{aligned} \tag{C.5}$$

where $\bar{K}$ is a finite constant. Looking at the expression of $\partial \phi_t(\boldsymbol{g}, \boldsymbol{\xi})/\partial \boldsymbol{g}^\top$, it is clear that this $\bar{K}$ exists because $\boldsymbol{\Xi}$ is compact and because $s(\cdot\,; \boldsymbol{\psi})$ is uniformly Lipschitz continuous by **IN2**. More specifically, by the definition of a derivative, it is not hard to prove that for continuous and differentiable functions $s(\cdot\,; \boldsymbol{\psi})$, the Lipschitz continuity condition implies that the norm of the derivative of $s(\cdot\,; \boldsymbol{\psi})$ is uniformly bounded by a finite constant.

It follows from multiple applications of the triangle inequality to (C.5) that for any $\bar{\boldsymbol{g}}, \boldsymbol{g}^\dagger \in \mathbb{R}^{p+1}$ that:

$$\mathbb{E} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_t^{(h+1)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi}) \| \leq \mathbb{E} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_t(\boldsymbol{g}^\dagger, \boldsymbol{\xi}) \| + \bar{K} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_{t-1}^{(h)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi}) \| + \bar{K} \; \| \boldsymbol{g}^\dagger \| \,.$$

So by raising both sides of the inequality to the power 2 and applying the C_n inequality, there must exist a finite constant $\tilde{W}$ such that

$$\begin{aligned} \mathbb{E} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_t^{(h+1)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi}) \|^2 &\leq D\bar{K}^2 \; \mathbb{E} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_{t-1}^{(h)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi}) \|^2 + D\bar{K}^2 \; \| \boldsymbol{g}^\dagger \|^2 \\ &\qquad + D \; \mathbb{E} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_t(\boldsymbol{g}^\dagger, \boldsymbol{\xi}) \|^2 \leq \tilde{W} < \infty \,, \end{aligned}$$

where we applied the C_n inequality to conclude that there must exist a finite constant D such that the inequality holds. The constant $\tilde{W}$ exists, because

$$\mathbb{E} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_{t-1}^{(h)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi}) \|^2 = \mathbb{E} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_t^{(h)}(\bar{\boldsymbol{g}}, \boldsymbol{\xi}) \|^2 \leq Z_h < \infty \,,$$

which holds by the stationarity of $\{\phi_t^{(h)}\}$ and the induction hypothesis, $\|\boldsymbol{g}^\dagger\|$ is some finite constant and finally

$$\mathbb{E} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \| \phi_t(\boldsymbol{g}^\dagger, \boldsymbol{\xi}) \|^2 \leq Z_1 < \infty \,,$$

which is the base case that was shown to be true in Step 1 of the proof. □

Proof of Lemma B.3. It follows from Straumann and Mikosch (2006, Lemma 2.1) that $x_t \overset{e.a.s.}{\to} 0$ as $t \to \infty$, implies that $T^{-1} \sum_{t=1}^T x_t \overset{a.s.}{\to} 0$ as $T \to \infty$. Hence, by the triangle inequality it suffices to show that $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} |\hat{\ell}_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) - \ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi})| \overset{e.a.s.}{\to} 0$ as $t \to \infty$. Under the maintained assumptions we can apply Proposition 1, which tells us that $|\hat{g}_t(\boldsymbol{\xi}) - g_t(\boldsymbol{\xi})| \overset{e.a.s.}{\to} 0$ as $t \to \infty$, where $\{g_t(\boldsymbol{\xi})\}_{t\in\mathbb{Z}}$

is a unique SE sequence and $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|g_t(\boldsymbol{\xi})|^2<\infty$. Define $\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})\equiv\boldsymbol{A}(L)(\boldsymbol{\beta}_0\hat{g}_t(\boldsymbol{\xi})+\boldsymbol{\varepsilon}_t)$ and let $\boldsymbol{x}_t(\boldsymbol{\xi})$ be the same, but then with $\hat{g}_t(\boldsymbol{\xi})$ replaced by $g_t(\boldsymbol{\xi})$. Then we have that

$$\begin{aligned}|\hat{\ell}_t(\gamma_0,\boldsymbol{\xi})-\ell_t(\gamma_0,\boldsymbol{\xi})| &= \left|\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})^\top\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})-\boldsymbol{x}_t(\boldsymbol{\xi})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_t(\boldsymbol{\xi})\right|\\ &\leq \left|(\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})-\boldsymbol{x}_t(\boldsymbol{\xi}))^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_t(\boldsymbol{\xi})\right|\\ &\quad+\left|(\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})-\boldsymbol{x}_t(\boldsymbol{\xi}))^\top\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})-\boldsymbol{x}_t(\boldsymbol{\xi}))\right|\\ &\quad+\left|\boldsymbol{x}_t(\boldsymbol{\xi})^\top\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})-\boldsymbol{x}_t(\boldsymbol{\xi}))\right|\,,\end{aligned}\tag{C.6}$$

where the inequality follows from adding and subtracting $x_t(\boldsymbol{\xi})$ on both sides of $\boldsymbol{\Sigma}^{-1}$ in the first term and then invoking the triangle inequality. We have that

$$\begin{aligned}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})-\boldsymbol{x}_t(\boldsymbol{\xi})\| &= \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{A}(L)\boldsymbol{\beta}_0(\hat{g}_t(\boldsymbol{\xi})-g_t(\boldsymbol{\xi}))\|\\ &\leq\|\boldsymbol{\beta}_0\|\ \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{g}_t(\boldsymbol{\xi})-g_t(\boldsymbol{\xi})|\\ &\quad+\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{A}_1\boldsymbol{\beta}_0\|\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{g}_{t-1}(\boldsymbol{\xi})-g_{t-1}(\boldsymbol{\xi})|\\ &\quad+\ldots+\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{A}_p\boldsymbol{\beta}_0\|\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{g}_{t-p}(\boldsymbol{\xi})-g_{t-p}(\boldsymbol{\xi})|\ \overset{e.a.s}{\to}0\,,\end{aligned}$$

as $t\to\infty$, where we use the uniform convergence result of Proposition 1, the sub-additivity of $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\cdot\|$ and the compactness of $\boldsymbol{\Xi}$. Now we can use that for any square symmetric matrix B and vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of appropriate dimensions: $|\boldsymbol{x}^\top B\boldsymbol{y}|\leq\lambda_{\max}(B)(\boldsymbol{x}^\top\boldsymbol{x})^{1/2}(\boldsymbol{y}^\top\boldsymbol{y})^{1/2}$, where $\lambda_{\max}(B)$ is the largest eigenvalue of B. Here $(\boldsymbol{x}^\top\boldsymbol{x})^{1/2}$ is the Euclidean norm, so by norm equivalence, there exists a finite constant C such that $(\boldsymbol{x}^\top\boldsymbol{x})^{1/2}\leq C\|\boldsymbol{x}\|$ for any $\boldsymbol{x}$. Hence, we can bound the supremum of the right-hand side of (C.6) as follows:

$$\begin{aligned}&\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{\ell}_t(\gamma,\boldsymbol{\xi})-\ell_t(\gamma,\boldsymbol{\xi})|\\ &\leq C^2\left(\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left|\frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right|\right)\Bigg(2\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})-\boldsymbol{x}_t(\boldsymbol{\xi})\|\cdot\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{x}_t(\boldsymbol{\xi})\|\\ &\qquad\qquad+\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\hat{\boldsymbol{x}}_t(\boldsymbol{\xi})-\boldsymbol{x}_t(\boldsymbol{\xi})\|^2\Bigg)\ \overset{e.a.s}{\to}\ 0\qquad\text{as } t\to\infty\,,\end{aligned}$$

where we use that $\inf_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\lambda_{\min}(\boldsymbol{\Sigma})>0$ because $\boldsymbol{\Xi}$ is compact and for any $\boldsymbol{\xi}\in\boldsymbol{\Xi}$, $\boldsymbol{\Sigma}$ is positive definite by **IN3**. Furthermore, the first term in brackets goes to zero e.a.s. by Lemma 2.1 of Straumann and Mikosch (2006), as it is a product of a variable that converges to zero e.a.s. and $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{x}_t(\boldsymbol{\xi})\|$, which can be shown to have a bounded moment and be SE straightforwardly. For the bounded moment, see for instance the proof of Lemma B.4, and $\{\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{x}_t(\boldsymbol{\xi})\|\}$ is SE by Krengel (1985, Proposition 4.3). Finally, it is immediately clear that the second term in brackets goes to zero e.a.s. by the continuous mapping theorem. This completes the proof.

□

Proof of Lemma B.4. Here $L_T(\boldsymbol{\gamma}_0, \boldsymbol{\xi})$ denotes the average of $\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi})$ from $t = 1$ to T, where

$$\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) = -(\boldsymbol{A}(L)(\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{A}(L)(\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t)) - \log|\boldsymbol{\Sigma}|\,,$$

with $g_t(\boldsymbol{\xi})$ from the SE limit sequence $\{g_t(\boldsymbol{\xi})\}_{t\in\mathbb{Z}}$ that exists by Proposition 1. Hence, the sequence $\{\ell_t(\boldsymbol{\gamma}_0, \cdot)\}$ has elements that take values in the space of continuous functions $\mathbb{C}(\boldsymbol{\Xi}, \mathbb{R})$. It is also an SE sequence by Proposition 4.3 in Krengel (1985), because for each t it is a continuous function of $\{(g_t, \dots, g_{t-p}, \boldsymbol{\varepsilon}_t, \dots, \boldsymbol{\varepsilon}_{t-p})\}$ which is an SE sequence. As in the proof of Lemma TA.6 of Blasques et al. (2022) we can apply the ergodic theorem for separable Banach spaces of Rao (1962) to $\{\ell_t(\boldsymbol{\gamma}_0, \cdot)\}$ if we can show $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} |\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi})| < \infty$. Using the notation $\boldsymbol{x}_t(\boldsymbol{\xi}) \equiv \boldsymbol{A}(L)(\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t)$ and invoking the triangle inequality, we have that:

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} |\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi})| &\le \mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \left|\boldsymbol{x}_t(\boldsymbol{\xi})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t(\boldsymbol{\xi})\right| + \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} |\log|\boldsymbol{\Sigma}|\,| \\ &\le \sup_{\boldsymbol{\xi}\in\Theta}\left[\frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\right] \mathbb{E}\sup_{\boldsymbol{\xi}\in\Theta}\left[\sum_{i=1}^{k} (\boldsymbol{x}_{it}(\boldsymbol{\xi}))^2\right] + \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} |\log|\boldsymbol{\Sigma}|\,|\,, \\ &\le C_1 \sum_{i=1}^{k} \mathbb{E}\sup_{\boldsymbol{\xi}\in\Theta}\left[(\boldsymbol{x}_{it}(\boldsymbol{\xi}))^2\right] + C_2 < \infty\,,\end{aligned}$$

where $\lambda_{\min}(\boldsymbol{\Sigma})$ denotes the smallest eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{x}_{it}(\boldsymbol{\xi})$ denotes the i-th element of the vector $\boldsymbol{x}_t(\boldsymbol{\xi})$. For the second inequality we use the well-known fact that for any square symmetric matrix B and vector $\boldsymbol{x}$ of appropriate dimensions: $|\boldsymbol{x}^\top B\boldsymbol{x}| \le \lambda_{\max}(B)\boldsymbol{x}^\top\boldsymbol{x}$. The finite constants C_1 and C_2 exist, because by **IN3** we have that $\boldsymbol{\Sigma}$ is positive definite for each $\boldsymbol{\xi} \in \boldsymbol{\Xi}$, and $\boldsymbol{\Xi}$ is compact. Finally, the expectations are finite because by the C_n-inequality of Loève (1977), we can bound the expectation of the square of $\boldsymbol{x}_{it}$ by a constant times the sum of the squares of its terms. The square of each term has a finite expectation, because (i) $\boldsymbol{\Xi}$ is a compact set, (ii) $\boldsymbol{\varepsilon}_t$ has two bounded moments under the current assumptions, and (iii) g_t has two bounded moments uniformly over $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ by Proposition 1. Hence, we can apply a uniform law of large numbers and the result follows.

□

Proof of Lemma B.5. We can write

$$\mathbb{E}\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) = -\mathbb{E}\left[(\boldsymbol{A}(L)(\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{A}(L)(\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t))\right] - \log|\boldsymbol{\Sigma}|\,.$$

Notice that substituting $\boldsymbol{\varepsilon}_t = \boldsymbol{A}_{1,0}\boldsymbol{\varepsilon}_{t-1} + \ldots + \boldsymbol{A}_{p,0}\boldsymbol{\varepsilon}_{t-p} + \boldsymbol{u}_t$, and using that $\boldsymbol{A}(L) = I - \boldsymbol{A}_1 L - \cdots - \boldsymbol{A}_p L^p$, gives

$$\boldsymbol{A}(L)(\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t) = \boldsymbol{A}(L)\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + (\boldsymbol{A}(L) - \boldsymbol{A}_0(L))\boldsymbol{\varepsilon}_t + \boldsymbol{u}_t\,,$$

where $\boldsymbol{A}_0(L)$ denotes the lag polynomial $\boldsymbol{A}(L)$ for $\boldsymbol{A}_i = \boldsymbol{A}_{i,0}$ for $i = 1, \dots, p$. Therefore, we can rewrite

$$\mathbb{E}\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi}) = -\mathbb{E}[\boldsymbol{u}_t^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{u}_t] - \mathbb{E}\Big[(\boldsymbol{A}(L)\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + (\boldsymbol{A}(L) - \boldsymbol{A}_0(L))\boldsymbol{\varepsilon}_t)^\top \boldsymbol{\Sigma}^{-1} \\ (\boldsymbol{A}(L)\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + (\boldsymbol{A}(L) - \boldsymbol{A}_0(L))\boldsymbol{\varepsilon}_t)\Big] - \log|\boldsymbol{\Sigma}|\,, \tag{C.7}$$

which follows from

$$\mathbb{E}\Big[\boldsymbol{u}_t^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{A}(L)\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + (\boldsymbol{A}(L) - \boldsymbol{A}_0(L))\boldsymbol{\varepsilon}_t)\Big] = 0\,,$$

which holds by the law of iterated expectations, because $\{\boldsymbol{u}_t\}_{t\in\mathbb{Z}}$ is mds, and $\boldsymbol{A}(L)\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + (\boldsymbol{A}(L) - \boldsymbol{A}_0(L))\boldsymbol{\varepsilon}_t$ only depends on $\boldsymbol{\varepsilon}_{t-1}, \boldsymbol{\varepsilon}_{t-2}, \dots$, and the second factor clearly has a finite expectation under the current assumptions.

The parameters ω, $\boldsymbol{\alpha}$, $\boldsymbol{\psi}_{-1}$, and $\boldsymbol{A}_1$, $\dots$, $\boldsymbol{A}_p$, only occur in the second term of (C.7). We start by investigating for which values of $\boldsymbol{\xi}$ other than $\boldsymbol{\xi}_0$ this term can potentially be minimized. Because $\boldsymbol{\Sigma}$ is a positive definite matrix by assumption, this expectation can only be non-negative. Thus any parameter vector $\boldsymbol{\xi}$ that is such that

$$\boldsymbol{A}(L)\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + (\boldsymbol{A}(L) - \boldsymbol{A}_0(L))\boldsymbol{\varepsilon}_t = 0 \qquad \text{a.s.}\,, \tag{C.8}$$

will be a minimizer of this expectation. Recall that we have

$$g_{t+1}(\boldsymbol{\xi}) = \omega_0 - \omega + g_t(\boldsymbol{\xi}) + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0) - \boldsymbol{\alpha}^\top s(\boldsymbol{A}(L)(\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t); \boldsymbol{\psi}) \\ = \omega_0 - \omega + g_t(\boldsymbol{\xi}) + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0) - \boldsymbol{\alpha}^\top s(\boldsymbol{A}(L)\boldsymbol{\beta}_0 g_t(\boldsymbol{\xi}) + (\boldsymbol{A}(L) - \boldsymbol{A}_0(L))\boldsymbol{\varepsilon}_t + \boldsymbol{u}_t; \boldsymbol{\psi})\,.$$

So if (C.8) holds, then it follows that we have almost surely for every t:

$$g_{t+1}(\boldsymbol{\xi}) = \omega_0 - \omega + g_t(\boldsymbol{\xi}) + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0) - \boldsymbol{\alpha}^\top s(\boldsymbol{u}_t; \boldsymbol{\psi})\,.$$

Hence, unless $\omega_0 + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0) = \omega + \boldsymbol{\alpha}^\top s(\boldsymbol{u}_t; \boldsymbol{\psi})$ almost surely, this would mean that the prediction error process has random walk dynamics and/or a deterministic drift, which is ruled out by the result of Proposition 1, which says $\{g_t\}_{t\in\mathbb{Z}}$ is a unique SE process, uniformly over $\boldsymbol{\xi} \in \boldsymbol{\Xi}$. Thus, we must have $\omega_0 + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0) = \omega + \boldsymbol{\alpha}^\top s(\boldsymbol{u}_t; \boldsymbol{\psi})$ almost surely if (C.8) holds, in which case $\boldsymbol{\xi}$ is such that $g_t(\boldsymbol{\xi}) = 0$ almost surely.

It follows that whenever $\boldsymbol{\xi}$ is such that (C.8) holds, we must have $(\boldsymbol{A}(L) - \boldsymbol{A}_0(L))\boldsymbol{\varepsilon}_t$ almost surely. Notice that $(\boldsymbol{A}(L) - \boldsymbol{A}_0(L))\boldsymbol{\varepsilon}_t = (\boldsymbol{A}_{1,0} - \boldsymbol{A}_1)\boldsymbol{\varepsilon}_{t-1} + \ldots + (\boldsymbol{A}_{p,0} - \boldsymbol{A}_p)\boldsymbol{\varepsilon}_{t-p}$, where the term $(\boldsymbol{A}_{1,0} - \boldsymbol{A}_1)\boldsymbol{\varepsilon}_{t-1} = (\boldsymbol{A}_{1,0} - \boldsymbol{A}_1)(\boldsymbol{A}_{1,0}\boldsymbol{\varepsilon}_{t-2} + \cdots + \boldsymbol{A}_{p,0}\boldsymbol{\varepsilon}_{t-p-1} + \boldsymbol{u}_{t-1})$ contains $\boldsymbol{u}_{t-1}$, which is uncorrelated with all the other (remaining) terms in (C.8) due to the mds property of $\{\boldsymbol{u}_t\}_{t\in\mathbb{Z}}$. Because the elements of $\boldsymbol{u}_t$ have positive variance and are not perfectly correlated by Assumption **IN3**, it follows that for (C.8) to be true, this $\boldsymbol{u}_{t-1}$ must be multiplied by zero, meaning that we must have $\boldsymbol{A}_1 = \boldsymbol{A}_{10}$. The same argument can be used for the remanining $\boldsymbol{A}_i$'s, from which it follows that we must have $\boldsymbol{A}(L) = \boldsymbol{A}_0(L)$.

Now we have shown that the second term of (C.7) is only maximized if $\omega + \boldsymbol{\alpha}^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}) = \omega_0 + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0)$ almost surely and $\boldsymbol{A}_i = \boldsymbol{A}_{i,0}$ for $i = 1, \ldots, p$, such that the term is equal to zero. The remaining two terms of the log likelihood in (C.7) only depend on $\boldsymbol{\Sigma}$. It is a standard result that $\boldsymbol{\Sigma}_0$ is the unique maximizer of the remaining quantity:

$$-\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}[\boldsymbol{u}_t\boldsymbol{u}_t^\top]\right) - \log|\boldsymbol{\Sigma}| = -\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\right) - \log|\boldsymbol{\Sigma}|.$$

Combining this with the derivations above, it follows that for $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ to be a maximizer of $\mathbb{E}\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi})$, we must have $\omega_0 + \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0) = \omega + \boldsymbol{\alpha}^\top s(\boldsymbol{u}_t; \boldsymbol{\Sigma}_0, \boldsymbol{\psi}_{-1})$ almost surely, which by Assumption **C1** holds if and only if $(\omega, \boldsymbol{\alpha}, \boldsymbol{\psi}_{-1}) = (\omega_0, \boldsymbol{\alpha}_0, \boldsymbol{\psi}_{-1,0})$. Hence, we must have $\boldsymbol{\xi} = \boldsymbol{\xi}_0$, which indeed leads to $g_t(\boldsymbol{\xi}) = 0$ almost surely by Proposition 1. In conclusion, $\boldsymbol{\xi}_0$ is the unique maximizer of $\mathbb{E}\ell_t(\boldsymbol{\gamma}_0, \boldsymbol{\xi})$. □

Proof of Lemma B.6. To analyse $\widehat{L}_T(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}) - \widehat{L}_T(\boldsymbol{b}_0, \boldsymbol{\xi})$, it is convenient to first study the difference $\hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi}) - \hat{f}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi})$. By Lemma B.8 there is a k-variate process $\{\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$, such that $\hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b}, \boldsymbol{\xi}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \hat{\boldsymbol{h}}_t(\boldsymbol{\theta})$ for each t, and $\{\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ is such that there exists a constant a such that for any t, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|^2 \leq at$ if $\omega_0 = 0$, and $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|^2 \leq at^2$ if $\omega_0 \neq 0$. The proof of this lemma is based on Lemma B.7 from which it follows that there exists a constant b such that for any t, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |\hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi})|^2 \leq bt$ and $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} |\hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi})|^2 \leq bt^2$, for $\omega_0 = 0$ and $\omega_0 \neq 0$ respectively. Next, we turn to the expression under consideration:

$$\begin{aligned}
&|\widehat{L}_T(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}) - \widehat{L}_T(\boldsymbol{b}, \boldsymbol{\xi})| \\
&= \Bigg| \frac{1}{T}\sum_{t=1}^T (\boldsymbol{A}(L)(\boldsymbol{y}_t - \widehat{\boldsymbol{\beta}}_T \hat{f}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi})))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{A}(L)(\boldsymbol{y}_t - \widehat{\boldsymbol{\beta}}_T \hat{f}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}))) \\
&\quad - \frac{1}{T}\sum_{t=1}^T (\boldsymbol{A}(L)(\boldsymbol{y}_t - \boldsymbol{\beta}_0 \hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi})))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{A}(L)(\boldsymbol{y}_t - \boldsymbol{\beta}_0 \hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi}))) \Bigg| \\
&\leq \left| \frac{1}{T}\sum_{t=1}^T (\boldsymbol{A}(L)(\boldsymbol{\beta}_0 \hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi}) - \widehat{\boldsymbol{\beta}}_T \hat{f}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi})))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{A}(L)(\boldsymbol{\beta}_0 \hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi}) - \widehat{\boldsymbol{\beta}}_T \hat{f}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}))) \right| \\
&\quad + 2 \left| \frac{1}{T}\sum_{t=1}^T (\boldsymbol{A}(L)(\boldsymbol{\beta}_0 \hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi}) - \widehat{\boldsymbol{\beta}}_T \hat{f}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi})))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{A}(L)(\boldsymbol{\beta}_0 \hat{g}_t(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_t)) \right|. \qquad \text{(C.9)}
\end{aligned}$$

We will show that the supremum over $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ of the two terms on the right-hand side will converge to zero in probability. We can use that for any square symmetric matrix B and vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of appropriate dimensions: $|\boldsymbol{x}^\top B \boldsymbol{y}| \leq \lambda_{\max}(B)(\boldsymbol{x}^\top\boldsymbol{x})^{1/2}(\boldsymbol{y}^\top\boldsymbol{y})^{1/2}$, where $\lambda_{\max}(B)$ is the largest eigenvalue of B. Here $(\boldsymbol{x}^\top\boldsymbol{x})^{1/2}$ is the Euclidean norm, so by norm equivalence, for any vector norm $\|\cdot\|$, there exists a finite constant C such that $(\boldsymbol{x}^\top\boldsymbol{x})^{1/2} \leq C\|\boldsymbol{x}\|$ for any $\boldsymbol{x}$. For ease of notation define $\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}) \equiv \boldsymbol{\beta}_0 \hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi}) - \widehat{\boldsymbol{\beta}}_T \hat{f}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi})$. For the first term we now have

$$\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \left| \frac{1}{T}\sum_{t=1}^T (\boldsymbol{A}(L)\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi}))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{A}(L)\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T, \boldsymbol{\xi})) \right|$$

$$\leq \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \frac{C^2}{\lambda_{\min}(\boldsymbol{\Sigma})} \frac{1}{T}\sum_{t=1}^{T} \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \left\| \boldsymbol{A}(L)\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\right\|^2$$

$$\leq C^2KD\,\frac{1}{T}\sum_{t=1}^{T}\left(\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left\|\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\right\|^2 + \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{A}_1\|^2 \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left\|\hat{\boldsymbol{x}}_{t-1}(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\right\|^2 \right.$$

$$\left. + \ldots + \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{A}_p\|^2 \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left\|\hat{\boldsymbol{x}}_{t-p}(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\right\|^2\right),$$

where the second inequality uses that there must be a finite number $K \equiv \sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \lambda_{\min}(\boldsymbol{\Sigma})$ as $\boldsymbol{\Xi}$ is compact and $\boldsymbol{\Sigma}$ is positive definite for any $\boldsymbol{\xi}\in\boldsymbol{\Xi}$ by **IN3**. Furthermore, it uses the sub-additivity and sub-multiplicativity of the matrix norm $\|\cdot\|$ and the fact that there exists a constant D such that $(a_1+\cdots+a_{p+1})^2 \leq D(a_1^2+\cdots+a_p^2)$ for any $a_1,\ldots,a_{p+1}$ numbers, by the C_n inequality of Loève (1977). It thus suffices to show that $\frac{1}{T}\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\|^2 = o_p(1)$. Notice that using the result of Lemma B.8 we can rewrite

$$\begin{aligned}\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi}) &= \boldsymbol{\beta}_0\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \widehat{\boldsymbol{\beta}}_T\hat{f}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\\ &= (\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}_T)\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) + \widehat{\boldsymbol{\beta}}_T(\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi}))\\ &= (\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}_T)\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) + \widehat{\boldsymbol{\beta}}_T(\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_0)^\top\hat{\boldsymbol{h}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi}),\end{aligned} \tag{C.10}$$

where with a slight abuse of notation we let $\hat{\boldsymbol{h}}_t(\boldsymbol{b},\boldsymbol{\xi}) \equiv \hat{\boldsymbol{h}}_t(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{b}^\top,\boldsymbol{\xi}^\top)^\top$. Using this decomposition, we can bound

$$\frac{1}{T}\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\|^2 \leq D\|\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_0\|^2\,\frac{1}{T}\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|^2$$

$$+ D\|\widehat{\boldsymbol{\beta}}_T\|^2\|\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_0\|^2\,\frac{1}{T}\sum_{t=1}^{T}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|^2.$$

This follows from the sub-additivity of the matrix norm $\|\cdot\|$ and the C_n inequality. Furthermore, for the second term we use that by the Cauchy-Schwarz inequality $|\boldsymbol{x}^\top\boldsymbol{y}| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$ for any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of appropriate dimension, which implies that also $|\boldsymbol{x}\top\boldsymbol{y}|^2 \leq \|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2$. By the continuous mapping theorem we furthermore have that $\|\widehat{\boldsymbol{\beta}}_T\|^2 = O_p(1)$, because under the current assumptions $\widehat{\boldsymbol{\beta}}_T \xrightarrow{p} \boldsymbol{\beta}_0$ as $T\to\infty$, which is a finite and real-valued limit. Hence, to show that the expression is $o_p(1)$, it suffices to show that

$$\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|^2 = \begin{cases} O_p(T^2) & \text{if } \omega_0 = 0\,,\\ O_p(T^3) & \text{if } \omega_0 \neq 0\,,\end{cases} \qquad \text{and}$$

$$\sum_{t=1}^{T}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|^2 = \begin{cases} O_p(T^2) & \text{if } \omega_0 = 0\,,\\ O_p(T^3) & \text{if } \omega_0 \neq 0\,,\end{cases}$$

as an assumption of the theorem is that $\|\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_0\| = o_p(T^{-1/2})$ if $\omega_0 = 0$ and $o_p(T^{-1})$ otherwise. Using Markov's inequality we can straightforwardly show that indeed the sums above are $O_p(T^2)$

if $\omega_0 = 0$ and $O_p(T^3)$ if $\omega_0 \neq 0$, because a random variable X_T is $O_p(1)$ if for every $\varepsilon > 0$, there is a k_ε and a T_ε, such that $\mathbb{P}(|X_T| > k_\varepsilon) < \varepsilon$, for every $T > T_\varepsilon$. By Markov's inequality we have $\mathbb{P}(|X_T| > k) \leq \mathbb{E}|X_T|/k$, so if $\mathbb{E}|X_T| \leq \bar{C} < \infty$ for some constant $\bar{C}$ that does not depend on T, it is clear that X_T must be $O_p(1)$. It follows from the results of Lemma B.7 that for example:

$$\frac{1}{T^2}\mathbb{E}\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|^2 \leq \frac{1}{T^2}\sum_{t=1}^{T} bt \leq b\,, \qquad \text{if } \omega_0 = 0\,, \qquad \text{and}$$

$$\frac{1}{T^3}\mathbb{E}\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|^2 \leq \frac{1}{T^3}\sum_{t=1}^{T} bt^2 \leq b\,, \qquad \text{if } \omega_0 \neq 0\,,$$

and the same can be done for $\mathbb{E}\sum_{t=1}^{T}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|^2$ using Lemma B.8 for the bounding constant a. This implies that the first term of (C.9) is $o_p(1)$.

To show that the second term of (C.9) is $o_p(1)$, we can take the virtually the same route to get the following sufficient condition:

$$\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}\|\,\frac{1}{T}\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left\|\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\right\| + \frac{1}{T}\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left\|\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\right\|\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\beta}_0\hat{g}_{t-j}(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_{t-j}\| = o_p(1)\,, \tag{C.11}$$

for any $j = 0, 1, \ldots, p$. The first term of this expression is $o_p(1)$ by the compactness of $\boldsymbol{\Xi}$ and because $\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left\|\boldsymbol{x}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\right\|$ can be shown to be $o_p(1)$. This can be shown using the same approach as we used to show $\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left\|\boldsymbol{x}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\right\|^2 = o_p(1)$, but this time using that by Jensen's inequality $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})| \leq (\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|^2)^{1/2} \leq (bt)^{1/2}$ or $\leq (bt^2)^{1/2}$ depending on the value of ω_0, and the same for $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|$.

The second term of the left-hand side of (C.11) can be bounded as follows, using the decomposition of $\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})$ in (C.10)

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left\|\hat{\boldsymbol{x}}_t(\widehat{\boldsymbol{b}}_T,\boldsymbol{\xi})\right\|\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\beta}_0\hat{g}_{t-j}(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_{t-j}\| \\
&\leq \left\|\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\right\|\cdot \\
&\quad \frac{1}{T}\sum_{t=1}^{T}\left(\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})| + \|\widehat{\boldsymbol{\beta}}_T\|\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|\right)\left(\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\beta}_0\hat{g}_{t-j}(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_{t-j}\|\right).
\end{aligned}$$

This expression can be shown to be $o_p(1)$, by again invoking Markov's inequality, because from Hölder's inequality we have

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\beta}_0\hat{g}_{t-j}(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_{t-j}\|\right) \\
&\qquad \leq \left(\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|^2\;\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\beta}_0\hat{g}_{t-j}(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}_{t-j}\|^2\right)^{1/2},
\end{aligned}$$

and

$$\mathbb{E}\left(\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\beta}_0\hat{g}_{t-j}(\boldsymbol{\xi})+\boldsymbol{\varepsilon}_{t-j}\|\right)$$
$$\leq\left(\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|^2\,\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\beta}_0\hat{g}_{t-j}(\boldsymbol{\xi})+\boldsymbol{\varepsilon}_{t-j}\|^2\right)^{1/2}.$$

Under the maintained assumptions, $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\beta}_0\hat{g}_{t-j}(\boldsymbol{\xi})+\boldsymbol{\varepsilon}_{t-j}\|^2$ can be bounded by a finite constant for any t and j, by arguments we have adopted before repeatedly. Furthermore, $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|^2$ and $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|^2$ can again be bounded using Lemmas B.7 and B.8, respectively. Hence, using the same approach as for the other term, by Markov's inequality we have that

$$\sum_{t=1}^{T}\left(\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|+\|\widehat{\boldsymbol{\beta}}_T\|\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|\right)\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{\beta}_0\hat{g}_{t-j}(\boldsymbol{\xi})+\boldsymbol{\varepsilon}_{t-j}\|\,,$$

is $O_p(T^{3/2})$ if $\omega_0=0$ and $O_p(T^2)$ if $\omega_0\neq 0$, which means that if the sum is multiplied by $\left\|\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_0\right\|/T$, it converges to zero in probability under the assumptions of the theorem. This completes the proof.

□

Proof of Lemma B.7. Notice that for each t we can write $\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})=f_t-\hat{g}_t(\boldsymbol{\xi})$, by the definition of $\hat{g}_t(\boldsymbol{\xi})$. It follows from Proposition 1 that $\{\hat{g}_t(\boldsymbol{\xi})\}$ converges to the unique SE $\{g_t(\boldsymbol{\xi})\}$ e.a.s. uniformly over $\boldsymbol{\Xi}$ and that $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|g_t(\boldsymbol{\xi})|^2<\infty$. Under the current assumptions, using the same unfolding approach as in the proof of Proposition 1, it can be shown that there is a finite constant C such that $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{g}_t(\boldsymbol{\xi})|^2<C$ for any t. Furthermore, we have

$$\begin{aligned}\mathbb{E}(f_t^2)&=\mathbb{E}\left(f_1+(t-1)\omega_0+\sum_{i=1}^{t-1}\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_i;\boldsymbol{\psi}_0)\right)^2\\&=f_1^2+(t-1)^2\omega_0^2+\sum_{i=1}^{t-1}\mathbb{E}(\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_i;\boldsymbol{\psi}_0))^2+\sum_{i=1}^{t-1}\sum_{j\neq i}\mathbb{E}(\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_i;\boldsymbol{\psi}_0)\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_j;\boldsymbol{\psi}_0))\\&\quad+(t-1)\omega_0\;f_1+(t-1)\omega_0\sum_{i=1}^{t-1}\mathbb{E}(\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_i;\boldsymbol{\psi}_0))+f_1\sum_{i=1}^{t-1}\mathbb{E}(\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_i;\boldsymbol{\psi}_0))\\&=f_1^2+(t-1)^2\omega_0^2+\sum_{i=1}^{t-1}\mathbb{E}|\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_i;\boldsymbol{\psi}_0)|^2+0+(t-1)\omega_0f_1+0+0\\&=f_1^2+\left[\mathbb{E}(\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0))^2+\omega_0\;f_1\right](t-1)+\omega_0^2(t-1)^2\,,\end{aligned}$$

where we use that f_1 is a constant and where the third equality uses that $\{\boldsymbol{u}_t\}$ is mds and Assumption **A3** that ensures that $s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)$ is an mds. The values in front of $(t-1)^k$ for $k=0,1,2$ in the final equality are finite, as $\mathbb{E}(\boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0))^2$ is finite under the current assumptions; see the proof of Lemma B.2.

Because $\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) = f_t - \hat{g}_t(\boldsymbol{\xi})$, it is not hard to see that $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})|^2 \leq \mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|f_t|^2 + \mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{g}_t(\boldsymbol{\xi})|^2$. We know from the proof of Proposition 1 that $\mathbb{E}\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}|\hat{g}_t(\boldsymbol{\xi})|^2$ can be bounded by a finite constant. Thus, it follows from the C_n inequality of Loève (1977), and the results above that there exists a finite constant b, such that (B.1) is satisfied. □

Proof of Lemma B.8. By the mean value theorem we have

$$\begin{aligned}
\hat{f}_{t+1}(\boldsymbol{b}_0,\boldsymbol{\xi}) - &\hat{f}_{t+1}(\boldsymbol{b},\boldsymbol{\xi}) \\
&= \hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b},\boldsymbol{\xi}) + \boldsymbol{\alpha}^\top\Big(s(\boldsymbol{A}(L)(\boldsymbol{y}_t - \boldsymbol{\beta}_0\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}));\boldsymbol{\psi}) \\
&\quad - s(\boldsymbol{A}(L)(\boldsymbol{y}_t - \boldsymbol{\beta}\hat{f}_t(\boldsymbol{b},\boldsymbol{\xi}));\boldsymbol{\psi})\Big) \\
&= \hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b},\boldsymbol{\xi}) \\
&\quad + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})[\boldsymbol{A}(L)(\boldsymbol{\beta}\hat{f}_t(\boldsymbol{b},\boldsymbol{\xi}) - \boldsymbol{\beta}_0\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}))]\,,
\end{aligned}$$

where $s'(\boldsymbol{z};\boldsymbol{\psi}) = \partial s(\boldsymbol{z};\boldsymbol{\psi})/\partial\boldsymbol{z}$ and where $\boldsymbol{z}_t^*$ is some k-dimensional vector on the line segment between $\boldsymbol{A}(L)(\boldsymbol{y}_t - \boldsymbol{\beta}_0\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}))$ and $\boldsymbol{A}(L)(\boldsymbol{y}_t - \boldsymbol{\beta}\hat{f}_t(\boldsymbol{b},\boldsymbol{\xi}))$. For notational convenience, we supress the dependence of $\boldsymbol{z}_t^*$ on the parameters and on the observations. Now we can use that

$$\boldsymbol{\beta}\hat{f}_t(\boldsymbol{b},\boldsymbol{\xi}) - \boldsymbol{\beta}_0\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \boldsymbol{\beta}(\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b},\boldsymbol{\xi}))\,,$$

from which it follows that

$$\begin{aligned}
\hat{f}_{t+1}(\boldsymbol{b}_0,\boldsymbol{\xi}) - &\hat{f}_{t+1}(\boldsymbol{b},\boldsymbol{\xi}) \\
&= \Big(1 - \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{\beta}\Big)\left(\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b},\boldsymbol{\xi})\right) \\
&\quad + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\Big(\boldsymbol{A}_1\boldsymbol{\beta}(\hat{f}_{t-1}(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_{t-1}(\boldsymbol{b},\boldsymbol{\xi})) \\
&\qquad\qquad + \ldots + \boldsymbol{A}_p\boldsymbol{\beta}(\hat{f}_{t-p}(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_{t-p}(\boldsymbol{b},\boldsymbol{\xi}))\Big) \\
&\quad + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}(L)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})\,,
\end{aligned}$$

where the lag operators in $\boldsymbol{A}(L)$ only work on $\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})$ and not on $s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})$. Throughout this proof we ignore that in this update, $\boldsymbol{z}_t^*$ indirectly depends on past values of $\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})$ and $\hat{f}_t(\boldsymbol{b},\boldsymbol{\xi})$. This would be problematic if we were to establish the dynamic properties of the process $\{\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b},\boldsymbol{\xi})\}$, but here we are just interested in bounding its expectation, so we can safely ignore this dependence between $\boldsymbol{z}_t^*$ and the elements of $\{\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b},\boldsymbol{\xi})\}$ and treat $\boldsymbol{z}_t^*$ as deterministic (until we start taking expectations). Hence, $\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b},\boldsymbol{\xi})$ is updated using a linear AR($p+1$)-type update, with 'innovation term' $\boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}(L)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top\boldsymbol{A}(L)^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})^\top\boldsymbol{\alpha}\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})$. It follows that we can define the k-variate process $\{\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\}$, such that $\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) - \hat{f}_t(\boldsymbol{b},\boldsymbol{\xi}) \equiv (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})$. In other words, define this process such that $\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})$ is initialized at $\hat{\boldsymbol{h}}_t(\boldsymbol{\theta}) = 0$ and for $t = 1, 2, \ldots, T$ is updated by

$$\hat{\boldsymbol{h}}_{t+1}(\boldsymbol{\theta}) = (1 - \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{\beta})\hat{\boldsymbol{h}}_t(\boldsymbol{\theta}) + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}_1\boldsymbol{\beta}\hat{\boldsymbol{h}}_{t-1}(\boldsymbol{\theta})$$

$$+ \ldots + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}_p\boldsymbol{\beta}\hat{\boldsymbol{h}}_{t-p}(\boldsymbol{\theta}) + \boldsymbol{A}(L)^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})^\top \boldsymbol{\alpha}\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi}) ,$$

where the lag operators in $\boldsymbol{A}(L)^\top$ only work on $\hat{f}_t(\boldsymbol{b}_0,\boldsymbol{\xi})$, and not on $s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})^\top$. So, ignoring the dependence of $\boldsymbol{z}_t^*$ on past values of $\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})$, $\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})$ is updated using a VAR$(p+1)$ type scheme, where all autoregressive coefficients are scalar.

Let $\hat{\boldsymbol{h}}_{it}(\boldsymbol{\theta})$ denote the i-th element of $\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})$. By norm equivalence, there exists a constant $\bar{c}$, such that

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{h}}_t(\boldsymbol{\theta})\|^2 \leq \bar{c}\;\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left(\sum_{i=1}^k |\hat{\boldsymbol{h}}_{it}(\boldsymbol{\theta})|\right)^2 \leq \bar{c}\;\bar{d}\;\sum_{i=1}^k \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|\hat{\boldsymbol{h}}_{it}(\boldsymbol{\theta})|^2 ,$$

where the second inequality holds for some $\bar{d}$ by the C_n inequality of Loève (1977) and the sub-additivity of the sup-norm. Hence, it suffices to show that there exists a finite constant $\bar{a}$ such that for any t, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|\hat{\boldsymbol{h}}_{it}(\boldsymbol{\theta})|^2 \leq \bar{a}t^l$, with $l=1$ if $\omega_0 = 0$ and $l=2$ if $\omega_0 \neq 0$.

Because in the update of $\hat{\boldsymbol{h}}_t$ the 'autoregressive' lags of $\hat{\boldsymbol{h}}_t$ have a scalar coefficient, it is clear that we can study the elements of $\hat{\boldsymbol{h}}_t$ separately. Start by defining the vector $\hat{\boldsymbol{h}}_{it}^\dagger(\boldsymbol{\theta}) = (\hat{\boldsymbol{h}}_{it}(\boldsymbol{\theta}),\ldots,\hat{\boldsymbol{h}}_{i,t-p}(\boldsymbol{\theta}))^\top$, such that we can write the updating scheme of $\hat{\boldsymbol{h}}_{it}(\boldsymbol{\theta})$ as a first-order SRE. Let $\bar{\phi}_{it} : \mathbb{R}^{p+1}\times\Theta \to \mathbb{R}^{p+1}$ denote a random function that is such that $\hat{\boldsymbol{h}}_{i,t+1}^\dagger(\boldsymbol{\theta}) = \bar{\phi}_{it}(\hat{\boldsymbol{h}}_{it}^\dagger(\boldsymbol{\theta}),\boldsymbol{\theta})$. More specifically, we can write

$$\bar{\phi}_{it}(\boldsymbol{h}^\dagger,\boldsymbol{\theta}) = \begin{pmatrix} \bar{\phi}_{1,it}(\boldsymbol{h}^\dagger,\boldsymbol{\theta}) \\ h_1^\dagger \\ \vdots \\ h_p^\dagger \end{pmatrix} ,$$

with $\boldsymbol{h}^\dagger = (h_1^\dagger,\ldots,h_{p+1}^\dagger)^\top$ and

$$\begin{aligned}\bar{\phi}_{1,it}(\boldsymbol{h}^\dagger,\boldsymbol{\theta}) = {} & (1-\boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{\beta})h_1^\dagger + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}_1\boldsymbol{\beta}h_2^\dagger + \ldots + \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}_p\boldsymbol{\beta}h_{p+1}^\dagger \\ & + e_i^\top \Big(s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})^\top\boldsymbol{\alpha}\hat{f}_t(\boldsymbol{\gamma},\boldsymbol{\xi}) - \boldsymbol{A}_1^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})^\top\boldsymbol{\alpha}\hat{f}_{t-1}(\boldsymbol{\gamma},\boldsymbol{\xi}) \\ & - \ldots - \boldsymbol{A}_p^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})^\top\boldsymbol{\alpha}\hat{f}_{t-p}(\boldsymbol{\gamma},\boldsymbol{\xi})\Big) ,\end{aligned}$$

where e_i denotes the i-th standard basis vector of $\mathbb{R}^k$.

Notice that for any $\bar{\boldsymbol{h}}_1, \bar{\boldsymbol{h}}_2 \in \mathbb{R}^{p+1}$ have

$$\begin{aligned}\bar{\phi}_{it}(\bar{\boldsymbol{h}}_1,\boldsymbol{\theta}) - \bar{\phi}_{it}(\bar{\boldsymbol{h}}_2,\boldsymbol{\theta}) & \\ & = \begin{pmatrix} 1-\boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{\beta} & \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}_1\boldsymbol{\beta} & \ldots & \boldsymbol{\alpha}^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})\boldsymbol{A}_p\boldsymbol{\beta} \\ & & & 0 \\ & I_p & & \vdots \\ & & & 0 \end{pmatrix} (\bar{\boldsymbol{h}}_1 - \bar{\boldsymbol{h}}_2) \\ & \equiv \bar{\Phi}(\boldsymbol{z}_t^*,\boldsymbol{\theta})(\bar{\boldsymbol{h}}_1 - \bar{\boldsymbol{h}}_2) ,\end{aligned}$$

and that for the r-fold convolutions of $\bar{\phi}_t$, we have

$$\bar{\phi}_{it}^{(r)}(\bar{\boldsymbol{h}}_1, \boldsymbol{\theta}) - \bar{\phi}_{it}^{(r)}(\bar{\boldsymbol{h}}_2, \boldsymbol{\theta}) = \left(\prod_{j=1}^{r} \bar{\Phi}(\boldsymbol{z}_{t-j+1}^*, \boldsymbol{\theta}) \right) (\bar{\boldsymbol{h}}_1 - \bar{\boldsymbol{h}}_2) \,.$$

By condition **C3**, for some $r \geq 1$, a constant $\kappa < 1$ exists such that

$$\sup_{\boldsymbol{\theta} \in \Theta, (\boldsymbol{z}_1, \dots, \boldsymbol{z}_r) \in \mathbb{R}^{kr}} \left\| \prod_{j=1}^{r} \bar{\Phi}(\boldsymbol{z}_j, \boldsymbol{\theta}) \right\| = \kappa < 1 \,. \tag{C.12}$$

For this r, we unfold the recursion of $\hat{\boldsymbol{h}}_{it}^{\dagger}(\boldsymbol{\theta})$ backwards r times, such that we have for any $\bar{\boldsymbol{h}} \in \mathbb{R}^{p+1}$:

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{h}}_{it}^{\dagger}(\boldsymbol{\theta})\| &\leq \sup_{\boldsymbol{\theta} \in \Theta} \|\bar{\phi}_{i,t-1}^{(r)}(\hat{\boldsymbol{h}}_{i,t-r}^{\dagger}(\boldsymbol{\theta}), \boldsymbol{\theta}) - \bar{\phi}_{i,t-1}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})\| + \sup_{\boldsymbol{\theta} \in \Theta} \|\bar{\phi}_{i,t-1}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})\| \\
&\leq \sup_{\boldsymbol{\theta} \in \Theta} \left\| \prod_{j=1}^{r} \bar{\Phi}(\boldsymbol{z}_{t-j}^*, \boldsymbol{\theta}) \right\| \sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{h}}_{i,t-r}^{\dagger}(\boldsymbol{\theta}) - \bar{\boldsymbol{h}}\| + \sup_{\boldsymbol{\theta} \in \Theta} \|\bar{\phi}_{i,t-1}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})\| \\
&\leq \kappa \sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{h}}_{i,t-r}^{\dagger}(\boldsymbol{\theta})\| + \kappa \, \|\bar{\boldsymbol{h}}\| + \sup_{\boldsymbol{\theta} \in \Theta} \|\bar{\phi}_{i,t-1}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})\| \,,
\end{aligned}$$

where we use the triangle inequality multiple times. In the second inequality we use (C.12) and that for any $q \times q$ matrix A and q-vector v, the L^p-norm and the corresponding operator norm have the property $\|Av\| \leq \|A\| \, \|v\|$. Say $t \geq r+1$, then unfolding this recursion backwards $\lfloor (t-1)/r \rfloor$ times, gives us

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{h}}_{it}^{\dagger}(\boldsymbol{\theta})\| \leq \kappa^{\lfloor (t-1)/r \rfloor} &\sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{h}}_{i,t-r\lfloor (t-1)/r \rfloor}^{\dagger}(\boldsymbol{\theta})\| \\
&+ \sum_{j=0}^{\lfloor (t-1)/r \rfloor - 1} \kappa^j (\kappa \|\bar{\boldsymbol{h}}\| + \sup_{\boldsymbol{\theta} \in \Theta} \|\bar{\phi}_{i,t-1-rj}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})\|) \,.
\end{aligned}$$

We now take $(\mathbb{E}|\cdot|^2)^{1/2}$, on both sides of the inequality, which is sub-additive as it is an L^2-norm in the vector space of real-valued random variables. Then we get:

$$\begin{aligned}
\left(\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{h}}_{it}^{\dagger}(\boldsymbol{\theta})\|^2 \right)^{1/2} \leq \kappa^{\lfloor (t-1)/r \rfloor} &\left(\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\boldsymbol{h}}_{i,t-r\lfloor (t-1)/r \rfloor}^{\dagger}(\boldsymbol{\theta})\|^2 \right)^{1/2} \\
&+ \sum_{j=0}^{\lfloor (t-1)/r \rfloor - 1} \kappa^j \left(\kappa \|\bar{\boldsymbol{h}}\| + \left(\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\bar{\phi}_{i,t-1-rj}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})\|^2 \right)^{1/2} \right) ,
\end{aligned} \tag{C.13}$$

Notice that it follows from the definition of $\bar{\phi}_{i,t}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})$ that we can bound $\sup_{\boldsymbol{\theta} \in \Theta} \|\bar{\phi}_{i,t}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})\|$ as follows,

$$\sup_{\boldsymbol{\theta} \in \Theta} \|\bar{\phi}_{i,t}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})\| \leq \bar{K}^r \|\bar{\boldsymbol{h}}\| + \sum_{i=1}^{r} \bar{K}^{i-1} \sup_{\boldsymbol{\theta} \in \Theta} |\hat{\boldsymbol{x}}_{t-i+1}(\boldsymbol{\theta})| \,,$$

where $\bar{K} \equiv \sup_{\boldsymbol{\theta} \in \Theta, \boldsymbol{z} \in \mathbb{R}^k} \|\bar{\Phi}(\boldsymbol{z}, \boldsymbol{\theta})\|$, which is finite by (C.12) and where

$$\hat{\boldsymbol{x}}_t(\boldsymbol{\theta}) \equiv e_i^{\top} \Big(s'(\boldsymbol{z}_t^*; \boldsymbol{\psi})^{\top} \boldsymbol{\alpha} \hat{f}_t(\boldsymbol{b}_0, \boldsymbol{\xi}) - \boldsymbol{A}_1^{\top} s'(\boldsymbol{z}_t^*; \boldsymbol{\psi})^{\top} \boldsymbol{\alpha} \hat{f}_{t-1}(\boldsymbol{b}_0, \boldsymbol{\xi})$$

$$- \ldots - \boldsymbol{A}_p^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\psi})^\top \boldsymbol{\alpha} \hat{f}_{t-p}(\boldsymbol{b}_0, \boldsymbol{\xi})\Big) \,.$$

Because clearly $\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{A}_j^\top s'(\boldsymbol{z}_t^*;\boldsymbol{\theta})^\top \boldsymbol{\alpha}\| < \infty$ for any j under the maintained assumptions and by the result of Lemma B.7, an inspection of the expression above implies that if we take an expectation, then for any fixed r and $\bar{\boldsymbol{h}}$ there must be a constant C such that

$$\mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} \|\bar{\phi}_{i,t}^{(r)}(\bar{\boldsymbol{h}}, \boldsymbol{\theta})\|^2 \leq \begin{cases} Ct\,, & \text{if } \omega_0 = 0\,, \text{ and} \\ Ct^2\,, & \text{if } \omega_0 \neq 0\,. \end{cases} \tag{C.14}$$

Now going back to (C.13), it follows that if $\omega_0 = 0$:

$$\begin{aligned}
\left(\mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{h}}_{it}^\dagger(\boldsymbol{\theta})\|^2\right)^{1/2} &\leq \kappa^{\lfloor (t-1)/r \rfloor}\, B + \sum_{j=0}^{\lfloor (t-1)/r \rfloor - 1} \kappa^j \left(\kappa\|\bar{\boldsymbol{h}}\| + C^{1/2}(t-1-rj)^{1/2}\right) \\
&\leq B + \frac{\kappa\|\bar{\boldsymbol{h}}\|}{1-\kappa} + C^{1/2} t^{1/2} \sum_{j=0}^{\lfloor (t-1)/r \rfloor - 1} \kappa^j \left(\frac{(t-1-rj)}{t}\right)^{1/2} \\
&\leq B + \frac{\kappa\|\bar{\boldsymbol{h}}\|}{1-\kappa} + C^{1/2} t^{1/2} \sum_{j=0}^{\lfloor (t-1)/r \rfloor - 1} \kappa^j \leq (\bar{A}t)^{1/2}\,,
\end{aligned}$$

where we use that there must be a finite constant B, such that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{h}}_{i,t-r\lfloor (t-1)/r \rfloor}^\dagger(\boldsymbol{\theta})\|^2 \leq B$. This can be shown by unfolding $\hat{\boldsymbol{h}}_{i,t-r\lfloor (t-1)/r \rfloor}^\dagger(\boldsymbol{\theta})$ backwards one step at a time until $\hat{\boldsymbol{h}}_{i,1} = 0$ is reached, which will take at most $r-1$ steps, and bounding the resulting expression using the same results we used above. For the second inequality we use that $0 \leq \kappa < 1$. For the third inequality we use that $t - 1 - rj < t$ for any index j of the sum. It follows from $\kappa < 1$ that there exists a finite constant $\bar{A}$ that does not depend on t, for which the final inequality holds. If $\omega_0 \neq 0$, then based on (C.14) the same strategy can be used to show that for any t, the expectation can be bounded by $\bar{A}t$, for some finite constant $\bar{A}$. It follows that there must also exist a finite constant $\bar{a}$ such that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{h}}_{it}(\boldsymbol{\theta})\|^2$ can be bounded by $\bar{a}t$ or $\bar{a}t^2$ for $\omega_0 = 0$ and $\omega_0 \neq 0$, respectively. This completes the proof. □

D. Estimation of the long-run parameters

To obtain an estimate of $\boldsymbol{b}$ and $\boldsymbol{m}$, we suggest to regress $\boldsymbol{y}_{-1,t} = (y_{2t}, \ldots, y_{kt})^\top$ on y_{1t}, as we can write

$$\begin{aligned}
&\boldsymbol{y}_{-1,t} = \boldsymbol{m}_0 + \boldsymbol{b}_0 y_{1t} + \boldsymbol{v}_{-1,t}\,, \quad \text{where} \\
&\boldsymbol{v}_{-1,t} = \boldsymbol{\varepsilon}_{-1,t} - \boldsymbol{b}_0 \varepsilon_{1t}\,, \quad \text{and} \quad \Delta y_{1t} = \omega_0 + v_{1,t-1} + \Delta\varepsilon_{1t}\,,
\end{aligned}$$

and where $\boldsymbol{\varepsilon}_{-1,t} = (\varepsilon_{2t}, \ldots, \varepsilon_{kt})^\top$ and $v_{1t} = \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t; \boldsymbol{\psi}_0)$. We can find the limiting distribution of the OLS estimator for $\boldsymbol{m}$ and $\boldsymbol{b}$ in the regression above based on standard theory for regressions with integrated processes. Define

$$\boldsymbol{v}_t \equiv \begin{pmatrix} v_{1t} \\ \boldsymbol{v}_{-1,t} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0) \\ \boldsymbol{\varepsilon}_{-1,t} - \boldsymbol{b}_0\varepsilon_{1t} \end{pmatrix} .$$

It is clear that the expectation of $\boldsymbol{v}_t$ is a vector of zeros. Let $\boldsymbol{\Omega}$ denote the long-run covariance matrix of $\boldsymbol{v}_t$:

$$\boldsymbol{\Omega} = \lim_{T\to\infty} \frac{1}{T}\mathbb{V}\mathrm{ar}\left(\sum_{t=1}^{T} \boldsymbol{v}_t\right) = \lim_{T\to\infty} \sum_{s=-T}^{T} \mathbb{E}[\boldsymbol{v}_t\boldsymbol{v}_{t-s}^\top] = \begin{pmatrix} \omega_{11} & \boldsymbol{\Omega}_{21}^\top \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} ,$$

where we use that $\{\boldsymbol{v}_t\}_{t\in\mathbb{Z}}$ is SE. The following proposition is useful for deriving the limiting distributions of the OLS estimators for $\boldsymbol{b}$ and $\boldsymbol{m}$.

Proposition D.1. *If A1-A3 are satisfied and $\boldsymbol{\Omega}$ is positive definite, then for every $r \in [0,1]$*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} \boldsymbol{v}_t \xrightarrow{d} \boldsymbol{B}(r) = \begin{pmatrix} B_1(r) \\ \boldsymbol{B}_2(r) \end{pmatrix} , \qquad \textit{as } T \to \infty ,$$

where we partition $\boldsymbol{B}(r) = (B_1(r), \boldsymbol{B}_2(r)^\top)^\top$ conformably with $\boldsymbol{v}_t$ and where $\boldsymbol{B}(r)$ is a k-dimensional Brownian motion with covariance matrix $\boldsymbol{\Omega}$. Also, for $S_t = \sum_{i=1}^{t-1} v_{1t}$

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{v}_{-1,t}S_t \xrightarrow{d} \int_0^1 B_1(r)\,\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21}, \qquad \textit{as } T \to \infty ,$$

where $\Delta_{21} = \sum_{s=1}^{\infty} \mathbb{E}[\boldsymbol{v}_{-1,t}v_{1,t-s}] = \boldsymbol{C}(\boldsymbol{A}_0(1)^{-1} - I_k)\mathbb{E}[\boldsymbol{u}_t\alpha_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)]$ with $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{b}_0 & I_{k-1} \end{pmatrix}$, where I_n denotes an $n \times n$ identity matrix.

In Theorem D.1, we consider the limiting distribution of the OLS estimator for the parameters in the regression above in different scenarios. The results of this theorem are standard results from the literature on regressions with integrated processes; see, for example, Park and Phillips (1988).

Theorem D.1. *Let A1-A3 be satisfied and assume $\boldsymbol{\Omega}$ is positive definite. Then:*

(i) Let $\omega_0 = 0$ and $\boldsymbol{m}_0 = 0$: then if $\widehat{\boldsymbol{b}}_T$ denotes the OLS estimator of $\boldsymbol{b}_0$ in the regression $\boldsymbol{y}_{-1,t} = \boldsymbol{b}_0 y_{1t} + \boldsymbol{v}_{-1,t}$,

$$T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) \xrightarrow{d} \left(\int_0^1 B_1(r)^2\,\mathrm{d}r\right)^{-1}\left(\int_0^1 B_1(r)\,\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21} + \Gamma_\varepsilon\right) ,$$

as $T \to \infty$, with Δ_{21} as defined in Proposition D.1 and $\Gamma_\varepsilon = \mathbb{E}[\boldsymbol{v}_{-1,t}\varepsilon_{1t}]$.

(ii) Let $\omega_0 = 0$: then if $\widehat{\boldsymbol{b}}_T$ and $\widehat{\boldsymbol{m}}_T$ denote the OLS estimators of $\boldsymbol{b}_0$ and $\boldsymbol{m}_0$ in the regression $\boldsymbol{y}_{-1,t} = \boldsymbol{m}_0 + \boldsymbol{b}_0 y_{1t} + \boldsymbol{v}_{-1,t}$,

$$T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) \xrightarrow{d} \left(\int_0^1 \bar{B}_1(r)^2\,\mathrm{d}r\right)^{-1}\left(\int_0^1 \bar{B}_1(r)\,\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21} + \Gamma_\varepsilon\right) , \textit{ and,}$$

$$T^{1/2}(\widehat{\boldsymbol{m}}_T - \boldsymbol{m}_0) \xrightarrow{d}$$
$$\boldsymbol{B}_2(1) - \left(\int_0^1 B_1(r)\,\mathrm{d}r\right)\left(\int_0^1 \bar{B}_1(r)^2\,\mathrm{d}r\right)^{-1}\left(\int_0^1 \bar{B}_1(r)\,\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21} + \Gamma_\varepsilon\right),$$

as $T \to \infty$, with Δ_{21} and Γ_ε the same as in (i) and $\bar{B}_1(r) = B_1(r) - \int_0^1 B_1(r)\,\mathrm{d}r$ is a demeaned Brownian motion.

(iii) Let $\omega_0 \neq 0$ and $\boldsymbol{m}_0 = 0$: then if $\widehat{\boldsymbol{b}}_T$ and $\widehat{\boldsymbol{m}}_T$ denote the OLS estimators of $\boldsymbol{b}_0$ and $\boldsymbol{m}_0$ in the regression $\boldsymbol{y}_{-1,t} = \boldsymbol{b}_0 y_{1t} + \boldsymbol{v}_{-1,t}$,

$$T^{3/2}(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) \xrightarrow{d} \frac{3}{\omega_0}\left[\boldsymbol{B}_2(1) - \int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r\right] = \mathcal{N}\left(0, \frac{3}{\omega_0^2}\boldsymbol{\Omega}_{22}\right),$$

as $T \to \infty$.

(iv) Let $\omega_0 \neq 0$: then if $\widehat{\boldsymbol{b}}_T$ and $\widehat{\boldsymbol{m}}_T$ denote the OLS estimators of $\boldsymbol{b}_0$ and $\boldsymbol{m}_0$ in the regression $\boldsymbol{y}_{-1,t} = \boldsymbol{m}_0 + \boldsymbol{b}_0 y_{1t} + \boldsymbol{v}_{-1,t}$,

$$T^{3/2}(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) \xrightarrow{d} \frac{12}{\omega_0}\left[\frac{1}{2}\boldsymbol{B}_2(1) - \int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r\right] = \mathcal{N}\left(0, \frac{12}{\omega_0^2}\boldsymbol{\Omega}_{22}\right), \quad \textit{and,}$$
$$T^{1/2}(\widehat{\boldsymbol{m}}_T - \boldsymbol{m}_0) \xrightarrow{d} 2\left[3\int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r - \boldsymbol{B}_2(1)\right] = \mathcal{N}(0, 4\boldsymbol{\Omega}_{22}),$$

as $T \to \infty$.

These OLS estimators can be calculated equation by equation, as each regression equation has a single regressor y_{1t} (plus possibly an intercept). In case the regression models in *(i)* or *(iii)* are used, while in fact $\boldsymbol{m}_0 \neq 0$, then the resulting estimator of $\boldsymbol{b}$ is still consistent, but at a lower rate. On the other hand, including $\boldsymbol{m}$ in the model leads to an estimator with a larger asymptotic variance. Therefore, some care must be taken in selecting which deterministic components are to be added to the model.

When $\omega_0 = 0$, the OLS estimator of $\boldsymbol{b}$ will be T-consistent, but inefficient due to the endogeneity of y_{1t} in the regression model, since we have the terms $\Delta_{21} + \Gamma_\varepsilon$ appear in the limiting distribution, and since $\boldsymbol{B}_2$ is correlated with B_1, which causes $\int_0^1 B_1(r)\,\mathrm{d}\boldsymbol{B}_2(r)$ to be skewed (Shin, 1994). Even when $\boldsymbol{\varepsilon}_t = \boldsymbol{u}_t$, that is when $\Delta_{21} = 0$ and $\boldsymbol{B}_2$ and B_1 are independent, the term Γ_ε remains in the limiting distribution. In Theorem D.2 we consider a modified efficient estimator for this case. When $\omega_0 \neq 0$, then the OLS estimator will be $T^{3/2}$-consistent and asymptotically normal; see points *(iii)* and *(iv)* of the theorem.

Notice that for case *(iii)* in Theorem D.1, we can construct the standard error of the i-th element of $\widehat{\boldsymbol{b}}_T$ by taking the square root of

$$\hat{s}_i^2 = T^{-3}3\widehat{\boldsymbol{\Omega}}_{22,ii}/\hat{\omega}_T^2,$$

where $\widehat{\boldsymbol{\Omega}}_{22,ii}$ is the i-th diagonal element of $\widehat{\boldsymbol{\Omega}}_{22}$, which in turn is some consistent estimator of the long-run variance $\boldsymbol{\Omega}_{22}$ based on the regression residuals. For instance, $\widehat{\boldsymbol{\Omega}}_{22}$ can be a kernel estimator with an appropriate bandwidth; see Andrews (1991), Newey and West (1994) and Hansen (1992). However, these kernel estimators can be inaccurate if there are VAR dynamics in $\boldsymbol{\varepsilon}_t$, so it can be preferable to do a pre-whitening procedure before estimating the long-run variance (Andrews and Monahan, 1992). With respect to $\hat{\omega}_T$ in the expression of $\hat{s}_i^2$, some consistent estimator of ω_0 can be used; for example, the estimator $\hat{\omega}_T$ that is obtained in the second step of the estimation procedure. It follows from the above discussion and the result of Theorem D.1, that the t-test statistic $(\widehat{\boldsymbol{b}}_{i,T} - \boldsymbol{b}_{i,0})/\hat{s}_i$ will be asymptotically standard normal. For the case of *(iv)*, the standard errors can be calculated using the same approach.

To obtain an efficient estimator, we consider the modified dynamic OLS estimator of Saikkonen (1991); see also Shin (1994). This modified estimator is obtained by adding leads and/or lags of Δy_{1t} to the regression model, to account for the serial correlation in and between $v_{1t}+\Delta\varepsilon_t$ and $\boldsymbol{v}_{-1,t}$. The theorem below, which follows from an application of Theorem 4.1 of Saikkonen (1991), gives the limiting distribution of the OLS estimator of $\boldsymbol{b}$ based on the modified regression model.

Theorem D.2. *Let **A1**-**A3** be satisfied, assume $\boldsymbol{\Omega}$ is positive definite, and let $\omega_0 = 0$. Define $\tilde{v}_{1t} = v_{1t} + \Delta\varepsilon_{1,t+1}$ and accordingly $\tilde{\boldsymbol{v}}_t = (\tilde{v}_{1t}, \boldsymbol{v}_{-1,t})$. Assume the autocovariances of $\tilde{\boldsymbol{v}}_t$ are absolutely summable. Let the spectral density matrix of $\tilde{\boldsymbol{v}}_t$ be continuous and bounded away from zero ($f_{\boldsymbol{vv}}(\lambda) \geq aI$ for $a > 0$) and let condition (17) of Saikkonen (1991) on the absolute summability of the fourth order cumulants be satisfied. Then consider the regression model $\boldsymbol{y}_{-1,t} = \boldsymbol{b}_0 y_{1t} + \sum_{j=-K}^{K} \boldsymbol{\pi}_j \Delta y_{1t-j} + \boldsymbol{v}^*_{-1,t}$, with K diverging at a rate such that $K^3/T \to 0$ as $K, T \to \infty$ and $T^{1/2} \sum_{|j|>K}^{\infty} \|\boldsymbol{\pi}_j\| \to 0$.*

(i) *Let $\boldsymbol{m}_0 = 0$. Then if $\widehat{\boldsymbol{b}}_T$ denotes the OLS estimator of $\boldsymbol{b}_0$ in the modified regression model defined above,*

$$T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) \overset{d}{\to} \left(\int_0^1 B_1(r)^2 \,\mathrm{d}r\right)^{-1} \int_0^1 B_1(r) \,\mathrm{d}\boldsymbol{B}_{2\cdot 1}(r)\,, \qquad \text{as} \quad T \to \infty\,,$$

where $\boldsymbol{B}_{2\cdot 1} \equiv \boldsymbol{B}_2 - \boldsymbol{\Omega}_{21}\omega_{11}^{-1} B_1$, which is independent of B_1.

(ii) *If $\widehat{\boldsymbol{b}}_T$ and $\widehat{\boldsymbol{m}}_T$ denote the OLS estimator of $\boldsymbol{b}_0$ and $\boldsymbol{m}_0$ in the regression model above with an intercept $\boldsymbol{m}_0$ included, then*

$$T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) \overset{d}{\to} \left(\int_0^1 \bar{B}_1(r)^2 \,\mathrm{d}r\right)^{-1} \int_0^1 \bar{B}_1(r) \,\mathrm{d}\boldsymbol{B}_{2\cdot 1}(r)\,, \qquad \text{and,}$$

$$T^{1/2}(\widehat{\boldsymbol{m}}_T - \boldsymbol{m}_0) \overset{d}{\to}$$

$$\boldsymbol{B}_{2\cdot 1}(1) - \left(\int_0^1 B_1(r)\,\mathrm{d}r\right)\left(\int_0^1 \bar{B}_1(r)^2 \,\mathrm{d}r\right)^{-1}\left(\int_0^1 \bar{B}_1(r)\,\mathrm{d}\boldsymbol{B}_{2\cdot 1}(r)\right)\,,$$

as $T \to \infty$, where $\bar{B}_1(r) = B_1(r) - \int_0^1 B_1(r)\,\mathrm{d}r$ is a demeaned Brownian motion and where $\boldsymbol{B}_{2\cdot 1}$ is the same as in (i).

In practice, the lag length K can, for instance, be chosen using an information criterion such as Akaike's information criterion (AIC) or Bayesian information criterion (BIC). We notice that the same limiting results can be established for other modified estimators such as the semi-parametric fully modified OLS estimator of Phillips and Hansen (1990). The appeal of the modified estimator we consider here, is that it only requires a single estimation, unlike the fully modified estimator. Finally, we notice that a cointegration test in the spirit of Shin (1994) can be carried out based on the residuals of the regression models of Theorem D.1*(iii)*-*(iv)* and Theorem D.2.

The limiting distributions of the efficient estimators in Theorem D.2 can be used for inference. For case *(i)*, the standard error of the i-th element of $\widehat{\boldsymbol{b}}_T$ can be constructed from the square root of

$$s_i^2 = \widehat{\boldsymbol{\Omega}}_{2\cdot 1,ii}\left(\sum_{t=1}^{T}(y_{1t})^2\right)^{-1},$$

where $\widehat{\boldsymbol{\Omega}}_{2\cdot 1,ii}$ is an estimate of the i-th diagonal element of $\boldsymbol{\Omega}_{2\cdot 1} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\omega_{11}^{-1}\boldsymbol{\Omega}_{21}^{\top}$, which can be estimated based on the residuals of the modified regression model using a kernel estimator. If the estimator of $\boldsymbol{\Omega}_{2\cdot 1}$ is consistent, then this leads to a t-test statistic $(\widehat{\boldsymbol{b}}_{i,T} - \boldsymbol{b}_{i,0})/s_i$ that is asymptotically standard normally distributed. For case *(ii)*, the same strategy can be used to construct the standard errors of $\widehat{\boldsymbol{b}}_{i,T}$, but then based on $\sum_{t=1}^{T}(y_{1t})^2 - \frac{1}{T}(\sum_{t=1}^{T} y_{1t})^2$. Similarly, it is easy to verify that the standard errors of $\widehat{\boldsymbol{m}}_{i,T}$ can be estimated by the square root of

$$s_i^2 = \frac{1}{T}\widehat{\boldsymbol{\Omega}}_{2\cdot 1,ii}\left(1 + \frac{1}{T}\left(\sum_{t=1}^{T} y_{1t}\right)^2\left(\sum_{t=1}^{T}(y_{1t})^2 - \frac{1}{T}\left(\sum_{t=1}^{T} y_{1t}\right)^2\right)^{-1}\right).$$

E. Proof of Theorems of Section D

Proof of Proposition D.1. Using the notation $\boldsymbol{C} := \begin{pmatrix}\boldsymbol{b}_0 & I_{k-1}\end{pmatrix}$, we can write $\boldsymbol{v}_{-1,t} = \boldsymbol{\varepsilon}_{-1,t} - \boldsymbol{b}_0\varepsilon_{1t} = \boldsymbol{C}\boldsymbol{\varepsilon}_t$. Consider the so-called Beveridge-Nelson decomposition of $\boldsymbol{\varepsilon}_t$, see for instance Saikkonen (1993), or Lütkepohl (2005) for the multivariate version. By Assumption **A2**, the lag polynomial $\boldsymbol{A}_0(L)$ is invertible, so we know that $\boldsymbol{\varepsilon}_t$ can be represented by a vector moving average process of infinite order: $\boldsymbol{\varepsilon}_t = \boldsymbol{A}_0(L)^{-1}\boldsymbol{u}_t = \sum_{j=0}^{\infty}\Xi_j\boldsymbol{u}_{t-j}$, for some coefficient matrices Ξ_j, with $\Xi_0 = I_k$ and $\sum_{j=0}^{\infty} j\|\Xi_j\| < \infty$. It can then be verified straightforwardly that the following decomposition is valid:

$$\boldsymbol{\varepsilon}_t = \boldsymbol{A}_0(L)^{-1}\boldsymbol{u}_t = \boldsymbol{A}_0(1)^{-1}\boldsymbol{u}_t + \zeta_t - \zeta_{t-1},$$

where

$$\zeta_t = \sum_{j=0}^{\infty}\boldsymbol{D}_j\boldsymbol{u}_{t-j}, \qquad \text{with} \quad \boldsymbol{D}_j = -\sum_{i=j+1}^{\infty}\Xi_i.$$

The process $\{\zeta_t\}_{t\in\mathbb{Z}}$ is well-defined under the invertibility of $\boldsymbol{A}_0(L)$ (see Saikkonen, 1993, p. 167). Together with the strict stationarity of $\{\boldsymbol{u}_t\}_{t\in\mathbb{Z}}$, it follows from Proposition 4.3 in Krengel (1985), that $\{\zeta_t\}_{t\in\mathbb{Z}}$ is strictly stationary and ergodic. This decomposition allows us to write the partial sums of $\boldsymbol{\varepsilon}_t$ as a random walk with mds innovations plus a stationary term.

Using the new notation introduced above, we can write

$$\boldsymbol{v}_t = \begin{pmatrix} \alpha_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0) \\ \boldsymbol{C}[\boldsymbol{A}_0(1)^{-1}\boldsymbol{u}_t + \zeta_t - \zeta_{t-1}] \end{pmatrix}.$$

It follows that the first result of the proposition holds, by considering the following expression for any $r \in [0,1]$:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} \boldsymbol{v}_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} \begin{pmatrix} \alpha_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0) \\ \boldsymbol{C}\boldsymbol{A}_0(1)^{-1}\boldsymbol{u}_t \end{pmatrix} + \frac{1}{\sqrt{T}}\begin{pmatrix} 0 \\ \boldsymbol{C}[\zeta_{[Tr]} - \zeta_0] \end{pmatrix}.$$

Clearly, the second term will vanish in probability as $T \to \infty$, because it is a strictly stationary term divided by $T^{1/2}$. Hence, we can focus on the first term. We notice that the elements of the vector $(\alpha_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0), (\boldsymbol{C}\boldsymbol{A}_0(1)^{-1}\boldsymbol{u}_t)^\top)^\top$ are strictly stationary under the current assumptions, and furthermore they are mds by Assumptions **A1** and **A3**. Also because **A3** and because $\boldsymbol{\Sigma}_0$ is assumed to be finite, the elements of the vector have a bounded variance. Due to the elements of the vector being uncorrelated over time, it also follows that the long-run covariance matrix of $\boldsymbol{v}_t$, denoted by Ω, is in fact equal to the covariance matrix of this vector, which we assume is positive definite and which is clearly finite. Hence, we can apply Theorem 15.2.1 of Davidson (2000), which gives us the first result of the proposition.

For the second result of the proposition, let $S_t = \sum_{i=1}^{t-1} v_{1t}$ and use the Beveridge-Nelson decomposition of $\boldsymbol{\varepsilon}_t$ given above to rewrite:

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{v}_{-1,t} S_t = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{C}\boldsymbol{A}_0(1)^{-1}\boldsymbol{u}_t S_t + \boldsymbol{C}\,\frac{1}{T}\sum_{t=1}^{T}(\zeta_t - \zeta_{t-1})S_t\,,$$

Recall that we just argued that for the vector $(v_{1t}, (\boldsymbol{C}\boldsymbol{A}_0(1)^{-1}\boldsymbol{u}_t)^\top)^\top$, the conditions of Theorem 15.2.1 of Davidson (2000) are satisfied. It follows automatically, that we can also apply Theorem 15.2.3 of Davidson (2000) to this vector, from which it follows that

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{C}\boldsymbol{A}_0(1)^{-1}\boldsymbol{u}_t S_t \xrightarrow{d} \int_0^1 B_1(r)\,\mathrm{d}\boldsymbol{B}_2(r)\,,$$

where we again use that the elements of $\{\boldsymbol{u}_t\}_{t\in\mathbb{Z}}$ are uncorrelated over time. We rewrite the second term as follows:

$$\begin{aligned}\boldsymbol{C}\,\frac{1}{T}\sum_{t=1}^{T}(\zeta_t - \zeta_{t-1})S_t &= \boldsymbol{C}\left(\frac{1}{T}\sum_{t=1}^{T}\zeta_t S_t - \frac{1}{T}\sum_{t=1}^{T}\zeta_{t-1}S_{t-1} - \frac{1}{T}\sum_{t=1}^{T}\zeta_{t-1}v_{1,t-1}\right) \\ &= \boldsymbol{C}\left(\frac{1}{T}\zeta_T S_T - \frac{1}{T}\sum_{t=0}^{T-1}\zeta_t v_{1,t}\right).\end{aligned}$$

The first term converges to zero in probability as $T \to \infty$, because $\zeta_T S_T$ is $O_P(T^{1/2})$, as ζ_T is strictly stationary over time. The last term converges in probability to $\mathbb{E}[\zeta_t v_{1,t}]$ by the law of large numbers for SE sequences, as it is straightforward to show that $\{\zeta_t v_{1,t}\}_{t\in\mathbb{Z}}$ is SE with a bounded moment under the current assumptions. Due to $\{\boldsymbol{u}_t\}_{t\in\mathbb{Z}}$ being mds, it follows that

$$\begin{aligned}\mathbb{E}[\zeta_t v_{1,t}] &= \mathbb{E}\left[\sum_{j=0}^{\infty} \boldsymbol{D}_j \boldsymbol{u}_{t-j} \alpha_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)\right] = D_0 \mathbb{E}[\boldsymbol{u}_t \alpha_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)] \\ &= -(A_0(1)^{-1} - I_k)\mathbb{E}[\boldsymbol{u}_t \alpha_0^\top s(\boldsymbol{u}_t;\boldsymbol{\psi}_0)] =: \Delta_{21}\,.\end{aligned}$$

It is not hard to see that indeed $\Delta_{21} = \sum_{s=1}^{\infty} \mathbb{E}[\boldsymbol{v}_{-1,t} v_{1,t-s}]$.

□

Proof of Theorem D.1. (i) The OLS estimator takes the form

$$\widehat{\boldsymbol{b}}_T = \frac{\sum_{t=1}^T \boldsymbol{y}_{-1,t} y_{1t}}{\sum_{t=1}^T y_{1t}^2} = \boldsymbol{b}_0 + \frac{\sum_{t=1}^T \boldsymbol{v}_{-1,t} y_{1t}}{\sum_{t=1}^T y_{1t}^2}\,,$$

where $y_{1t} = f_1 + S_t + \varepsilon_{1t}$, with $S_t = \sum_{s=1}^t v_{1s}$. It follows that:

$$\begin{aligned}T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) &= \frac{T^{-1}\sum_{t=1}^T \boldsymbol{v}_{-1,t}(f_1 + S_t + \varepsilon_{1t})}{T^{-2}\sum_{t=1}^T (f_1 + S_t + \varepsilon_{1t})^2} \\ &\xrightarrow{d} \left(\int_0^1 B_1(r)^2 \,\mathrm{d}r\right)^{-1} \left(\int_0^1 B_1(r)\,\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21} + \Gamma_\varepsilon\right)\,,\end{aligned}$$

as $T \to \infty$. As it follows from the result of Proposition D.1 and the continuous mapping theorem for functionals that the well-known standard results of for instance Lemma 3.1 Phillips and Durlauf (1986) and Proposition C.18 of Lütkepohl (2005) hold. From this it follows that

$$\begin{aligned}\frac{1}{T^2}\sum_{t=1}^T S_t^2 &\xrightarrow{d} \int_0^1 B_1^2(r)\,\mathrm{d}r\,, \\ \frac{1}{T}\sum_{t=1}^T \boldsymbol{v}_{-1,t} S_t &\xrightarrow{d} \int_0^1 B_1(r)\,\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21}\,, \\ \frac{1}{T}\sum_{t=1}^T \boldsymbol{v}_{-1,t}\varepsilon_{1t} &\xrightarrow{p} \mathbb{E}[\boldsymbol{v}_{-1,t}\varepsilon_{1t}] \equiv \Gamma_\varepsilon\,,\end{aligned}$$

as $T \to \infty$. The last result follows from the law of large numbers for SE sequences. The other terms in the numerator and denominator all converge to zero in probability by applications of standard law of large numbers and (F)CLTs.

(ii) Here the OLS estimator takes the form

$$\begin{aligned}\widehat{\boldsymbol{b}}_T &= \frac{T\sum_{t=1}^T \boldsymbol{y}_{-1,t} y_{1t} - \sum_{t=1}^T \boldsymbol{y}_{-1,t}\sum_{t=1}^T y_{1t}}{T\sum_{t=1}^T y_{1t}^2 - (\sum_{t=1}^T y_{1t})^2} \\ &= \boldsymbol{b}_0 + \frac{T\sum_{t=1}^T \boldsymbol{v}_{-1,t} y_{1t} - \sum_{t=1}^T \boldsymbol{v}_{-1,t}\sum_{t=1}^T y_{1t}}{T\sum_{t=1}^T y_{1t}^2 - (\sum_{t=1}^T y_{1t})^2}\,.\end{aligned}$$

By again using $y_{1t} = f_1 + S_t + \varepsilon_{1t}$, with $S_t = \sum_{s=1}^{t} v_{1s}$, we have that

$$\begin{aligned}
&T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) \\
&\quad = \frac{T^{-1}\sum_{t=1}^{T} \boldsymbol{v}_{-1,t}(f_1 + S_t + \varepsilon_{1t}) - (T^{-1/2}\sum_{t=1}^{T} \boldsymbol{v}_{-1,t})(T^{-3/2}\sum_{t=1}^{T}(f_1 + S_t + \varepsilon_{1t}))}{T^{-2}\sum_{t=1}^{T}(f_1 + S_t + \varepsilon_{1t})^2 - (T^{-3/2}\sum_{t=1}^{T}(f_1 + S_t + \varepsilon_{1t}))^2} \\
&\quad \xrightarrow{d} \left(\int_0^1 \bar{B}_1(r)^2 \,\mathrm{d}r\right)^{-1} \left(\int_0^1 \bar{B}_1(r)\,\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21} + \Gamma_\varepsilon\right),
\end{aligned}$$

as $T \to \infty$. This follows from Proposition D.1, because together with the functional continuous mapping theorem, it gives us

$$\begin{aligned}
\frac{1}{T^2}\sum_{t=1}^{T} S_t^2 - \left(T^{-3/2}\sum_{t=1}^{T} S_t\right)^2 &\xrightarrow{d} \int_0^1 B_1(r)^2 \,\mathrm{d}r - \left(\int_0^1 B_1(r)\,\mathrm{d}r\right)^2 \\
&= \int_0^1 \left(B_1(r) - \int_0^1 B_1(s)\,\mathrm{d}s\right)^2 \mathrm{d}r\,,
\end{aligned}$$

as $T \to \infty$, and

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{v}_{-1,t}(S_t + \varepsilon_{1t}) - &\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T} \boldsymbol{v}_{-1,t}\right)\left(\frac{1}{T^{3/2}}\sum_{t=1}^{T} S_t\right) \\
&\xrightarrow{d} \int_0^1 B_1(r)\,\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21} + \mathbb{E}[\boldsymbol{v}_{-1,t}\varepsilon_{1t}] - \boldsymbol{B}_2(1)\int_0^1 B_1(r)\,\mathrm{d}r \\
&= \int_0^1 \left(B_1(r) - \int_0^1 B_1(s)\,\mathrm{d}s\right)\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21} + \Gamma_\varepsilon\,,
\end{aligned}$$

as $T \to \infty$. The other terms in the numerator and denominator all converge to zero in probability.

For $\widehat{\boldsymbol{m}}_T$ we have that

$$\widehat{\boldsymbol{m}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_{-1,t} - \widehat{\boldsymbol{b}}_T \frac{1}{T}\sum_{t=1}^{T} y_{1t} = \boldsymbol{m}_0 - (\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0)\frac{1}{T}\sum_{t=1}^{T} y_{1t} + \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{v}_{-1,t}\,,$$

from which it follows that:

$$\begin{aligned}
T^{1/2}(\widehat{\boldsymbol{m}}_T - \boldsymbol{m}_0) &= \frac{1}{T^{1/2}}\sum_{t=1}^{T} \boldsymbol{v}_{-1,t} - \frac{1}{T^{3/2}}\sum_{t=1}^{T} y_{1t} \cdot T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) \\
&\xrightarrow{d} \boldsymbol{B}_2(1) - \int_0^1 \bar{B}_1(r)\,\mathrm{d}r \left(\int_0^1 \bar{B}_1(r)^2\,\mathrm{d}r\right)^{-1}\left(\int_0^1 \bar{B}_1(r)\,\mathrm{d}\boldsymbol{B}_2(r) + \Delta_{21} + \Gamma_\varepsilon\right),
\end{aligned}$$

as $T \to \infty$, where we use the limiting distribution of $T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0)$ derived above and standard results such as $T^{-3/2}\sum_{t=1}^{T} S_t \xrightarrow{d} \int_0^1 B_1(r)\,\mathrm{d}r$ as $T \to \infty$.

(iii) The OLS estimator of this regression has the form:

$$\widehat{\boldsymbol{b}}_T = \frac{\sum_{t=1}^{T} \boldsymbol{y}_{-1,t} y_{1t}}{\sum_{t=1}^{T} y_{1t}^2} = \boldsymbol{b}_0 + \frac{\sum_{t=1}^{T} \boldsymbol{v}_{-1,t} y_{1t}}{\sum_{t=1}^{T} y_{1t}^2}\,.$$

So using that $y_{1t} = t \cdot \omega_0 + f_1 + S_t + \varepsilon_{1t}$, with $S_t = \sum_{s=1}^{t} v_{1s}$, we have that

$$T^{3/2}(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) = \frac{T^{-3/2}\sum_{t=1}^{T} \boldsymbol{v}_{-1,t}(t \cdot \omega_0 + f_1 + S_t + \varepsilon_{1t})}{T^{-3}\sum_{t=1}^{T}(t \cdot \omega_0 + f_1 + S_t + \varepsilon_{1t})^2}$$

$$\xrightarrow{d} \frac{3}{\omega_0}\left[\boldsymbol{B}_2(1) - \int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r\right],$$

as $T \to \infty$, by Proposition D.1), because have that in the denominator $T^{-3}\sum_{t=1} t^2 \to 3^{-1}$ and in the numerator $T^{-3/2}\sum_{t=1}^T t\boldsymbol{v}_{-1,t} \xrightarrow{d} \boldsymbol{B}_2(1) - \int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r$ as $T \to \infty$, and the other terms in the numerator and denominator all converge to zero in probability, which follows from standard results. Finally, it is not hard to show that the limit distribution is Gaussian with mean zero and its variance can also be deduced straightforwardly.

If this regression model is used while in fact $\boldsymbol{\mu}_0 \neq 0$, then the resulting estimator $\widehat{\boldsymbol{b}}_T$ will no longer be $T^{3/2}$-consistent, but it will only be T-consistent with limiting distribution:

$$T(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) = \frac{T^{-2}\sum_{t=1}^T(\boldsymbol{m}_0 + \boldsymbol{v}_{-1,t})(t\cdot\omega_0 + f_1 + S_t + \varepsilon_{1t})}{T^{-3}\sum_{t=1}^T(t\cdot\omega_0 + f_1 + S_t + \varepsilon_{1t})^2} \xrightarrow{p} \boldsymbol{m}_0\frac{3}{2\omega_0},$$

as $T \to \infty$. So having a regression model with intercept leads to a higher asymptotic variance, but not taking into account the intercept, leads to a lower rate of consistency.

(iv) The OLS estimator of $\boldsymbol{b}_0$ of this regression has the form:

$$\begin{aligned}\widehat{\boldsymbol{b}}_T &= \frac{T\sum_{t=1}^T \boldsymbol{y}_{-1,t}y_{1t} - \sum_{t=1}^T \boldsymbol{y}_{-1,t}\sum_{t=1}^T y_{1t}}{T\sum_{t=1}^T y_{1t}^2 - (\sum_{t=1}^T y_{1t})^2}\\ &= \boldsymbol{b}_0 + \frac{T\sum_{t=1}^T \boldsymbol{v}_{-1,t}y_{1t} - \sum_{t=1}^T \boldsymbol{v}_{-1,t}\sum_{t=1}^T y_{1t}}{T\sum_{t=1}^T y_{1t}^2 - (\sum_{t=1}^T y_{1t})^2},\end{aligned}$$

by again using $y_{1t} = t\cdot\omega_0 + f_1 + S_t + \varepsilon_{1t}$, with $S_t = \sum_{s=1}^t v_{1s}$, we have

$$\begin{aligned}T^{3/2}(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0) &= \frac{T^{-3/2}\sum_{t=1}^T \boldsymbol{v}_{-1,t}y_{1t} - (T^{-1/2}\sum_{t=1}^T \boldsymbol{v}_{-1,t})(T^{-2}\sum_{t=1}^T y_{1t})}{T^{-3}\sum_{t=1}^T y_{1t}^2 - (T^{-2}\sum_{t=1}^T y_{1t})^2}\\ &\xrightarrow{d} \frac{12}{\omega_0}\left[\boldsymbol{B}_2(1) - \int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r - \frac{1}{2}\boldsymbol{B}_2(1)\right]\\ &= \frac{12}{\omega_0}\left[\frac{1}{2}\boldsymbol{B}_2(1) - \int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r\right],\end{aligned}$$

as $T \to \infty$, again by Proposition D.1, as we have that $T^{-(i+1)}\sum_{t=1}^T t^i \to (i+1)^{-1}$, so in de denominator $T^{-3}\sum_{t=1}^T t^2 - (T^{-2}\sum_{t=1}^T t)^2 \to 3^{-1} - 4^{-1} = 12^{-1}$. Also, in the numerator

$$T^{-3/2}\sum_{t=1}^T t\boldsymbol{v}_{-1,t} \xrightarrow{d} \boldsymbol{B}_2(1) - \int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r, \quad\text{and}\quad \left(T^{-1/2}\sum_{t=1}^T \boldsymbol{v}_{-1,t}\right)\left(T^{-2}\sum_{t=1}^T t\right) \xrightarrow{d} 2^{-1}\boldsymbol{B}_2(1),$$

as $T \to \infty$.

The OLS estimator of $\boldsymbol{m}_0$ has the form:

$$\widehat{\boldsymbol{m}}_T = \frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_{-1,t} - \widehat{\boldsymbol{\beta}}_T\frac{1}{T}\sum_{t=1}^T y_{1t} = \boldsymbol{m}_0 - (\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0)\frac{1}{T}\sum_{t=1}^T y_{1t} + \frac{1}{T}\sum_{t=1}^T \boldsymbol{v}_{-1,t}.$$

So we have that:

$$T^{1/2}(\widehat{\boldsymbol{m}}_T - \boldsymbol{m}_0) = -T^{3/2}(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0)\frac{1}{T^2}\sum_{t=1}^T y_{1t} + \frac{1}{T^{1/2}}\sum_{t=1}^T \boldsymbol{v}_{-1,t}$$

$$
\begin{aligned}
&\xrightarrow{d} -\frac{12}{\omega_0}\left[\frac{1}{2}\boldsymbol{B}_2(1) - \int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r\right]\cdot\frac{\omega_0}{2} + \boldsymbol{B}_2(1) \\
&= 6\left[\int_0^1 \boldsymbol{B}_2(r)\,\mathrm{d}r - \frac{1}{3}\boldsymbol{B}_2(1)\right],
\end{aligned}
$$

as $T \to \infty$, using the limiting distribution of $T^{3/2}(\widehat{\boldsymbol{b}}_T - \boldsymbol{b}_0)$ and results that we also used earlier in the proof.

Finally, it is clear that for both $\widehat{\boldsymbol{b}}_T$ and $\widehat{\boldsymbol{m}}_T$ the limiting distribution is Gaussian with mean zero, and the variance can be deduced straightforwardly.

□

Proof of Theorem D.2. (i) This result follows directly from Theorem 4.1 of Saikkonen (1991). Similar results as in Proposition D.1 can straightforwardly also be derived for $\tilde{\boldsymbol{v}}_t$. It is assumed that the autocovariances of the process $\tilde{\boldsymbol{v}}_t$ are absolutely summable, whcich implies that condition (16) in Saikkonen (1991) is satisfied. Furthermore,

$$
\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]}\tilde{\boldsymbol{v}}_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]}\boldsymbol{v}_t + \frac{1}{\sqrt{T}}\left[\begin{pmatrix}\varepsilon_{1,[Tr]} - \varepsilon_{1,0} \\ \mathbf{0}\end{pmatrix}\right] \xrightarrow{d} \boldsymbol{B}(r),
$$

as $T \to \infty$, so the limiting distribution of the partial sums of $\tilde{\boldsymbol{v}}_t$ is the same as that of $\boldsymbol{v}_t$, and the long-run covariance matrix is still equal to $\boldsymbol{\Omega}$. Hence, it follows that the result holds.

(ii) Although Theorem 4.1 of Saikkonen (1991) does not apply to models with deterministic components, the theorem can clearly be extended to allow for this case, which will lead to the given limiting distribution; see also Lemma 1 of Shin (1994) which considers this case but for a single equation model. □