Artemova, Mariia

Working Paper

An Order-Invariant Score-Driven Dynamic Factor Model

Tinbergen Institute Discussion Paper, No. TI 2023-067/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Artemova, Mariia (2023) : An Order-Invariant Score-Driven Dynamic Factor Model, Tinbergen Institute Discussion Paper, No. TI 2023-067/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/282880

TI 2023-067/III
Tinbergen Institute Discussion Paper

An Order-Invariant Score-Driven Dynamic Factor Model

Mariia Artemova[1]

1 Vrije Universiteit Amsterdam and Tinbergen Institute

An Order-Invariant Score-Driven Dynamic Factor Model*

Mariia Artemova

Vrije Universiteit Amsterdam and Tinbergen Institute

October 19, 2023

Abstract

This paper introduces a novel score-driven dynamic factor model designed for filtering cross-sectional co-movements in panels of time series. The model is formulated using elliptical distribution for the noise terms, thus allowing the update of the time-varying parameter to be potentially nonlinear and robust to outliers. We derive stochastic properties of the time series generated by the model, such as stationarity and ergodicity, and establish the invertibility of the filter. We prove that the identification of the factors and loadings is achieved by incorporating an orthogonality constraint on the loadings which is invariant to the order of the series in the panel. Given the nonlinearity of the constraint, we propose to exploit a maximum likelihood estimation on the Stiefel manifolds, which ensure that the identification constraint is satisfied numerically, hence allowing a joint estimation of the static and time-varying parameters. Furthermore, the asymptotic properties of the constrained estimator are derived. In a series of Monte Carlo experiments, we find evidence of appropriate finite sample properties of the estimator and resulting score filter for the time-varying parameters. We reveal the empirical usefulness of our factor model for constructing indices of economic activity from a set of macroeconomic and financial variables during the period 1981–2022. An empirical application highlights the importance of the robust update for the time-varying parameters in the presence of V-shaped recessions, such as the COVID-19 recession.

Keywords: Score-driven model; Robust filtering; Factor model; Economic indicators.
JEL Classification: C13, C32, C38.

*Corresponding author email: m.artemova@vu.nl.

1 Introduction

Factor models have become important tools for analyzing and modeling co-movements in panels of economic and financial time series. The central idea is to summarize cross-sectional covariation in a few unobserved factors, which can then be used for variety of purposes. For example, central banks and other policy institutions are often interested in constructing leading and coincident indices for monitoring economic and financial stability. The work on index development was pioneered in the 20th century by the National Bureau of Economic Research (NBER) as a part of the research program on business cycles. After further developments and refinements, factor models have become the dominant approach in this field giving a rise to a wide range of indices and indicators (Stock & Watson, 1989, 2002b; Brave & Kelley, 2017; Lewis et al., 2022).

The COVID-2019 pandemic has created new challenges for macroeconometric analysis (Ng, 2021), particularly in terms of modeling economic and financial time series that experienced large spikes during March 2020 and unanticipated recovery dynamics afterwards. These spikes have a non-negligible effect on the pre-COVID fit of the models and parameter estimates, and also make it substantially more difficult to interpret economic indices. Therefore, the demand for robustness properties in time series econometric models has increased. To address these challenges, we introduce and develop a theory for a novel, order-invariant, score-driven dynamic factor model that allows updates of time-varying parameters to be potentially nonlinear and robust to influential points and outliers. In the empirical application, we reveal the importance and flexibility of our approach in constructing aggregate measures of economic activity in the presence of the COVID-19 recession period.

In the literature, there are two dominant approaches for modeling and estimation of factor models. The first approach assumes that the factors are static, meaning that the dynamics of the factors are not modeled explicitly. This approach typically uses principal component analysis (PCA) and its variations for estimation (Stock & Watson, 2002a; Bai, 2003; Bai & Li, 2012). The second approach models the dynamics of the factors explicitly by casting the model into a state space form. Maximum likelihood-based estimation methods are then available for estimation of the model parameters (Engle & Watson (1981); Watson

& Engle (1983); Quah & Sargent (1993); Doz et al. (2011, 2012)).

Alternatively, the dynamics of the factors can be modeled using an observation-driven modeling approach, in which the dynamics of the factors are driven by past observations (Cox et al., 1981). Observation-driven models are appealing since the likelihood is available in a closed form regardless of the complexity of the distribution. Moreover, recently, Creal et al. (2013) and Harvey (2013) introduced a new class of observation-driven models, known as score-driven models, where the update of the time-varying parameter is based on the scaled score of the predictive likelihood. This new class has given rise to a large strand of observation-driven models since the updating equations based on the score provide a natural updating mechanism while allowing the model to stay flexible and general (Artemova et al., 2022a).

The score-driven modeling approach has been widely used and proved to be suitable in many empirical applications as reviewed by Artemova et al. (2022b). Specifically, Creal et al. (2014) introduced the first score-driven dynamic factor model with an application to mixed-measurement observations. They considered score-driven Gaussian and Student's t factor models for modeling co-movements in the panel of macroeconomic and financial time series. However, in contrast to our paper, Creal et al. (2014) did not discuss the theoretical properties of the model and estimators. Furthermore, to resolve the rotational indeterminacy, a typical problem in factor analysis, the authors adopted an identification condition that is not invariant to the order of the series in the panel. This condition can be restrictive in empirical applications, leading to a lack of model interpretation or even model misspecification. For example, this condition makes it impossible to conduct statistical inference on the restricted loadings, which is a shortcoming in empirical analysis.

In this paper, we introduce an order-invariant score-driven dynamic factor model for capturing co-movements in panels of time series. The model is formulated using a general class of elliptical distributions that covers many empirically relevant distributions, such as Gaussian and Student's t. In the score-driven framework, the choice of the distribution plays a role not only in the modeling of the noise terms, but it also provides an intuitive updating mechanism for the factors. Namely, in the Student's t model, the update of the factors becomes more robust to influential points and outliers with the degrees of freedom parameter adjusting the importance of the robustness. This feature can be important in

empirical applications as evidenced by a growing literature on the development of robust factor models and estimation methods, see, for example, Fan et al. (2021) and He et al. (2023) for robust estimation of factor models with static factors, D'Innocenzo et al. (2023) for robust multivariate location models with score-driven dynamics, and Barigozzi et al. (2023) for robust estimation of factor models for tensor-valued time series, among many others.

We further provide general theoretical results for multivariate score-driven dynamic factor models with elliptically distributed innovations. Specifically, we analyze the stochastic properties of the time series generated by the model such as stationarity and ergodicity and show invertibility of the filter. We also discuss conditions required for the parameter identification and show that for the order-invariant case the estimation can be done using a constrained maximum likelihood estimator which is shown to be consistent and asymptotically normal. The order-invariant identification condition is nonlinear, hence numerically it requires a special treatment. We propose to estimate the model parameters using a constrained maximum likelihood, where the subset of the parameter space is restricted to lie on a Stiefel manifold, (Stiefel, 1935; Edelman et al., 1998). In a series of Monte Carlo experiments, we demonstrate that our estimation procedure is reliable in terms of the good finite sample properties of the maximum likelihood (ML) estimates and the filtered estimates of the factors. In the paper, we also show that the proposed model and estimation procedure can accommodate application-specific parameter restrictions, such as group-factor models with common and group-specific factors.

We apply our model to estimate economic activity indicators from a panel of macroeconomic and financial time series. We find that the estimated factor is closely associated with the US business cycle, where troughs indicate moments of downturns in the US economy. Additionally, our analysis shows the importance of the robust updating mechanism for the time-varying factors, especially when a V-type recession period, such as the COVID-19 recession, is present in the sample. Furthermore, our results emphasize the importance of the order-invariant identification condition, which does not impose zero restrictions on the matrix of loadings. This condition lets the data itself to reveal the structure, and enables us to conduct statistical inference on all the estimated parameters.

The remainder of the paper is organized as follows. Section 2 introduces a score-driven

factor model with elliptically distributed innovations. We establish the stochastic properties of the time series generated by the model and of the filter and discuss in detail model identification. Finally, an order-invariant score-driven factor model is introduced and its link to the group-factor model is established. Section 3 discusses details of the estimation procedure and establishes the asymptotic properties of the estimators. Section 4 presents the results of the Monte Carlo studies where the reliability of our estimation and modeling approaches is revealed. Section 5 demonstrates the results of the empirical application. The proofs of the main theoretical results are contained in Appendix. Further technical details, additional results of the Monte Carlo simulations and details for the empirical application are contained in the Supplementary Appendix.

2 Score-driven factor model

2.1 Model specification

Let $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Nt})^\top$ denote an N-dimensional vector of time series. Assume that the series are subject to a factor structure,

$$\boldsymbol{y}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_t, \boldsymbol{\Sigma}; \boldsymbol{\nu}), \quad t = 1, \ldots, T, \tag{1}$$

where $\boldsymbol{f}_t = (f_{1t}, \ldots, f_{rt})^\top$ is an $r \times 1$ vector of common factors, $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_r]$ is an $N \times r$ matrix of individual specific exposures to the factors with vector $\boldsymbol{\Lambda}_r = (\lambda_{1r}, \ldots, \lambda_{Nr})^\top$ containing the exposures to the common factor f_{rt}, and $\boldsymbol{\varepsilon}_t$ is an independent identically distributed (i.i.d.) $N \times 1$ zero-mean disturbance vector with multivariate density $p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_t, \boldsymbol{\Sigma}; \boldsymbol{\nu})$, diagonal scale matrix $\boldsymbol{\Sigma}$ and other parameters of the distribution collected in a parameter vector $\boldsymbol{\nu}$.

We further assume that $p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_t, \boldsymbol{\Sigma}; \boldsymbol{\nu})$ in (1) is a density from the general class of elliptical distributions with a density generator $g(\cdot)$, that is $\boldsymbol{\varepsilon}_t \sim \mathcal{E}_N(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$. In general, the density of elliptical distribution for some $\boldsymbol{x} \sim \mathcal{E}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ is given by

$$|\boldsymbol{\Sigma}|^{-1/2} g((\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})). \tag{2}$$

The special cases of elliptical distribution are Gaussian with $g(u) = (2\pi)^{-N/2}\exp(-u/2)$ and Student's t with $g(u) = \frac{\Gamma\left(\frac{N+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}(\nu\pi)^{-N/2}\left(1+\frac{u}{\nu}\right)^{-\frac{N+\nu}{2}}$. The choice of the distribution depends on the application at hand. For example, if the data is contaminated by outliers, it can be desirable to consider the Student's t distribution. Therefore, by formulating the model using the general class of elliptical distributions we cover different applications. For more details on the class of elliptical distributions, we refer the reader to Fang et al. (2018).

The goal of this paper is to develop a filter for the vector of dynamic factors $\boldsymbol{f}_t$ that is capable of summarizing co-movements between the series. The factors and innovations are assumed to be mutually uncorrelated, while the factors $\boldsymbol{f}_t$ are allowed to be dynamic. To model the dynamics of the factors, we use the score-driven modeling approach introduced in Creal et al. (2013) and Harvey (2013). Hence, the factors' dynamics are as follows

$$\boldsymbol{f}_{t+1} = \boldsymbol{\omega} + \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{f}_t, \tag{3}$$

with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_r)$, $\boldsymbol{A} = \text{diag}(\alpha_1, \ldots, \alpha_r)$, and $\boldsymbol{B} = \text{diag}(\beta_1, \ldots, \beta_r)$, where $\omega_i, \alpha_i, \beta_i$ for $i = 1, \ldots, N$ are unknown scalar coefficients.

The $r \times 1$ vector $\boldsymbol{s}_t$ is the score of the predictive likelihood ∇_t scaled by a scaling matrix $\boldsymbol{S}_t$, where a common choice for $\boldsymbol{S}_t$ is the inverse of the Fisher information matrix $\boldsymbol{\mathcal{I}}_{t|t-1}$, that is

$$\begin{aligned}
\boldsymbol{s}_t &= \boldsymbol{S}_t \nabla_t, \\
\nabla_t &= \frac{\partial \log p_{\boldsymbol{y}}(\boldsymbol{y}_t|\boldsymbol{f}_t, \mathcal{F}_{t-1}, \boldsymbol{\theta})}{\partial \boldsymbol{f}_t}, \\
\boldsymbol{S}_t &= \boldsymbol{\mathcal{I}}_{t|t-1}^{-1} = \mathbb{E}_{t-1}\left[\nabla_t \nabla_t^\top\right],
\end{aligned}$$

where $p_{\boldsymbol{y}}(\boldsymbol{y}_t|\boldsymbol{f}_t, \mathcal{F}_{t-1}, \boldsymbol{\theta})$ is the predictive conditional density, $\mathcal{F}_{t-1}$ is the filtration representing the set of information available at time $t-1$, and vector $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^q$ collects all the unknown static parameters.

To complete the model specification, in Lemma 1, we derive the score, Fisher information matrix and scaled score expressions for the case of factor model with elliptically distributed innovations. In the proof we exploit that, given model (1), the conditional density

$p_{\boldsymbol{y}}(\boldsymbol{y}_t|\boldsymbol{f}_t, \mathcal{F}_{t-1}, \boldsymbol{\theta})$ is from the class of elliptical distributions, $\boldsymbol{y}_t|\boldsymbol{f}_t, \mathcal{F}_{t-1} \sim \mathcal{E}_N(\boldsymbol{\Lambda}\boldsymbol{f}_t, \boldsymbol{\Sigma}, g)$[1].

Lemma 1. *Let (1) be the observation equation with elliptically distributed innovations, $\boldsymbol{\varepsilon}_t \sim \mathcal{E}_N(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$. Then the score, Fisher information matrix and scaled score take the following form*

$$\begin{aligned}
\nabla_t &= -2\frac{g'(\|\tilde{\boldsymbol{y}}_t\|^2)}{g(\|\tilde{\boldsymbol{y}}_t\|^2)}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{y}}_t, \\
\boldsymbol{\mathcal{I}}_{t|t-1} &= -2C(\|\tilde{\boldsymbol{y}}_t\|, g)\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}, \qquad (4) \\
\boldsymbol{s}_t &= \frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|, g)}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{y}}_t, \qquad (5)
\end{aligned}$$

where $g'(\cdot)$ denotes the derivative of $g(\cdot)$, $C(\|\tilde{\boldsymbol{y}}_t\|, g) := -2\mathbb{E}_{t-1}\left[\|\tilde{\boldsymbol{y}}_t\|^2\left(\frac{g'(\|\tilde{\boldsymbol{y}}_t\|^2)}{g(\|\tilde{\boldsymbol{y}}_t\|^2)}\right)^2\right]$, $\tilde{\boldsymbol{y}}_t := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t)$ and $W(\|\tilde{\boldsymbol{y}}_t\|, g) := \frac{1}{N}C(\|\tilde{\boldsymbol{y}}_t\|, g)\left(\frac{g'(\|\tilde{\boldsymbol{y}}_t\|^2)}{g(\|\tilde{\boldsymbol{y}}_t\|^2)}\right)^{-1}$.

We further follow the literature on factor models and assume that the series are demeaned beforehand such that the unconditional mean of $\boldsymbol{y}_t$ is equal to zero. Hence, in updating equation (3), we can set $\boldsymbol{\omega} = \boldsymbol{0}_r$. Given the score expression in Lemma 1, the updating equation for the factors in case of elliptically distributed innovations is as follows

$$\boldsymbol{f}_{t+1} = \boldsymbol{A}\frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|, g)}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t\right) + \boldsymbol{B}\boldsymbol{f}_t. \qquad (6)$$

Gaussian and Student's t distributions are members of the general class of elliptical distributions. Throughout the paper we consider these models as the main examples. Below, we state the updating equations for the factors in case of the Gaussian and Student's t models. The derivations are presented in the Supplementary Appendix A.

EXAMPLE 1 (Model with Gaussian innovations. Updating equation)**.** *Consider a model with observation equation (1) and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}_N, \boldsymbol{\Sigma})$. The density generator for Gaussian distribution is of the form $g(u) = (2\pi)^{-N/2}\exp(-u/2)$, which given (2) and $\boldsymbol{y}_t|\boldsymbol{f}_t, \mathcal{F}_{t-1} \sim$*

[1]Throughout the paper, we adopt a common notation for norms. Particularly, we use a Euclidean norm for vectors, that is, for any vector $\boldsymbol{x}$, $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top\boldsymbol{x}}$, and a spectral norm for matrices, i.e. for any matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\| = \sqrt{\varrho(\boldsymbol{A}^T\boldsymbol{A})}$, where $\varrho(\boldsymbol{A}^T\boldsymbol{A})$ denotes a spectral radius of matrix $\boldsymbol{A}^T\boldsymbol{A}$.

$\mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{f}_t, \boldsymbol{\Sigma})$, *implies a well-known density function of a multivariate Gaussian distribution,*

$$(2\pi)^{-N/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t)\right).$$

Hence, $C=-N/2$ *and* $W=1$ *and the updating equation for the factors takes the following form*

$$\boldsymbol{f}_{t+1}=\boldsymbol{A}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t\right)+\boldsymbol{B}\boldsymbol{f}_t,$$

or, equivalently,

$$\boldsymbol{f}_{t+1}=\boldsymbol{A}\left(\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t-\boldsymbol{f}_t\right)+\boldsymbol{B}\boldsymbol{f}_t.$$

Clearly, in the case of the Gaussian model, the score update is linear and driven by a scaled prediction error. Intuitively, the prediction error is weighted by the corresponding common factors' exposures, loadings, and downweighted in the case of the large variance of the idiosyncratic components. Therefore, the score update 'automatically' ensures that the series with large loadings contribute more to the update as they contain stronger signal about the common factors, while for the series with the large variance of the error term the effect is limited.

EXAMPLE 2 (Model with Student's t innovations. Updating equation)**.** *Consider a model with observation equation* (1) *and* $\boldsymbol{\varepsilon}_t\sim\boldsymbol{t}_\nu(\boldsymbol{0}_N,\boldsymbol{\Sigma})$ *with* $\nu>1$ *being the degrees of freedom parameter.*

In this case, the density generator is of the form $g(u)=\frac{\Gamma\left(\frac{N+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}(\nu\pi)^{-N/2}\left(1+\frac{u}{\nu}\right)^{-\frac{N+\nu}{2}}$*, hence* $C=-\frac{1}{2}\frac{N(N+\nu)}{(N+\nu+2)}$ *and* $W(\|\tilde{\boldsymbol{y}}_t\|,\nu)=\frac{\nu}{(N+\nu+2)}\left(1+\frac{\|\tilde{\boldsymbol{y}}_t\|^2}{\nu}\right)$ *and the updating equation for the factors takes the following form*

$$\boldsymbol{f}_{t+1}=\boldsymbol{A}\frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|,\nu)}\left(\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t-\boldsymbol{f}_t\right)+\boldsymbol{B}\boldsymbol{f}_t\,.$$

For the Student's t innovations, we obtain a nonlinear updating scheme which due to the presence of a scaling factor $W(\|\tilde{\boldsymbol{y}}_t\|,\nu)$ is robust to influential points and outliers. We

note that when $\nu \to \infty$ the updating equation above simplifies to the Gaussian case.

2.2 Stationarity and invertibility of score-driven factor models

In this section, we state the general conditions for the stationarity and invertibility of the score-driven factor model defined by equations (1) and (6). Note that for a correctly specified model, the updating equation for the factors, can be rewritten in terms of the innovations in the following form

$$\boldsymbol{f}_{t+1} = \boldsymbol{A}\frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|, g)}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t + \boldsymbol{B}\boldsymbol{f}_t. \tag{7}$$

We start with investigating the solutions to equations (1) and (6) given the sequence $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}}$. In other words, we analyze the properties of the sequences generated by the score-driven factor model. First, we state the conditions under which the stationary solutions exist. Second, we show that these solutions are also unique.

Assumption 1. *The matrix $\boldsymbol{\Sigma}$ is diagonal with elements $0 < \underline{c} \leq \sigma_i^2 \leq \bar{c} < \infty$ for every $i = 1, \ldots, N$.*

Assumption 2. $\boldsymbol{P} := \left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}$ *is positive definite with $\|\boldsymbol{P}\| < \infty$.*

Assumption 3. $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}}$ *is an i.i.d. sequence.*

Assumption 4. $\mathbb{E}\log^+\left\|\frac{1}{W(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|, g)}\frac{1}{N}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\right\| < \infty.$

Lemma 2 (Existence and uniqueness of the SE solution)**.** *Let Assumptions 1–4 hold. Then, for all $\boldsymbol{f}_1 \in \mathbb{R}^r$ there exist unique strictly stationary and ergodic causal solutions $\{\boldsymbol{f}_t\}_{t\in\mathbb{Z}}$ and $\{\boldsymbol{y}_t\}_{t\in\mathbb{Z}}$ to equations (1) and (6) if and only if $\|\boldsymbol{B}\| < 1$ for all $\boldsymbol{\theta} \in \Theta$.*

The last condition in Lemma 2 imposes a restriction on the parameter space Θ. The condition is the same as for linear models since, although the filter equation (6) can be nonlinear in $\boldsymbol{f}_t$, the updating equation (7) as a data generating process is always linear. In turn, Assumption 3 restricts the stochastic properties of the innovations and implies that the sequence $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}}$ is strictly stationary and ergodic. Assumptions 2 and 4 imply that the scaled score $\boldsymbol{s}_t$ has a bounded logarithmic moment. We note that Assumption 2 is further

replaced by a stricter condition required for the identification, see Section 2.3. Below, we show that Assumption 4 holds in a number of applications.

EXAMPLE 1 (Ctd., Gaussian model. Assumption 4)**.** *Here, we verify that Assumption 4 is fulfilled in the case of the score-driven factor model with Gaussian innovations with covariance matrix* $\boldsymbol{\Sigma} \succ 0$*, i.e.* $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}_N, \boldsymbol{\Sigma})$*. We verify the existence of the logarithmic moment by showing that there are two bounded moments which, in turn, imply the required condition by Lyapunov's inequality.*

$$\mathbb{E}\left\| \frac{1}{W(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|, g)} \frac{1}{N} \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t \right\|^2 = \frac{1}{N^2}\mathbb{E}\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\right\|^2 = \frac{1}{N} < \infty,$$

where the first equality follows since for all t $W(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|, g) = 1$ *as stated in Section 2.1. Hence,* $\mathbb{E}\log^+ \left\| \frac{1}{W(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|, g)} \frac{1}{N} \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t \right\| < \infty$ *holds.*

EXAMPLE 2 (Ctd., Student's t model. Assumption 4)**.** *Let us verify Assumption 4 for the score-driven factor model with multivariate Student's t distributed innovations with scale matrix* $\boldsymbol{\Sigma} \succ 0$*, i.e.* $\boldsymbol{\varepsilon}_t \sim \boldsymbol{t}_\nu(\mathbf{0}_N, \boldsymbol{\Sigma})$*. Then*

$$\begin{aligned}
&\mathbb{E}\log^+ \left\| \frac{1}{W(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|, g)} \frac{1}{N} \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t \right\| = \mathbb{E}\log^+ \left\| \frac{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|}{NW(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|, g)} \frac{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t}{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|} \right\| \\
&= \log^+ \frac{N+\nu+2}{N\nu} + \mathbb{E}\log^+ \frac{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|}{1+\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|^2/\nu} + \mathbb{E}\log^+ \left\| \frac{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t}{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|} \right\| < \infty,
\end{aligned}$$

as long as $\nu > 0$*.*

The examples above suggest that in many applications it is straightforward to verify that the score has several bounded moments. Hence, by reinforcing Assumptions 2 and 4, we can prove that the data generated by equations (1) and (6) has several moments.

Assumption 3.a. $\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^k < \infty$ *for some* $k > 0$*.*

Assumption 4.a. $\mathbb{E}\left\| \frac{1}{W(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|, g)} \frac{1}{N} \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t \right\|^k < \infty$ *with* k *as defined in Assumption 3.a.*

Lemma 3 (Bounded moments)**.** *Let the conditions of Lemma 2 hold. If, additionally, Assumptions 3.a and 4.a are satisfied, then the solutions to equations (1) and (6) satisfy* $\mathbb{E}\|\boldsymbol{f}_t\|^k < \infty$ *and* $\mathbb{E}\|\boldsymbol{y}_t\|^k < \infty$*.*

EXAMPLE 2 (Ctd., Student's t model. Assumption 4.a)**.** *For the Student's t model, due to the uniform boundedness of the score, we have a stronger result, i.e.* $\sup_t \|\boldsymbol{f}_t\| < \infty$. *Clearly,*

$$\sup_t \left\| \frac{1}{W(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|, g)} \frac{1}{N} \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t \right\| = \frac{N+\nu+2}{N\nu} \sup_t \frac{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|}{1+\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|^2/\nu} < \infty,$$

as long as $\nu > 0$. *In turn, Assumption 3.a holds, hence* $\mathbb{E}\|\boldsymbol{y}_t\|^k < \infty$, *as long as* $\nu > k$.

Next, we analyze the solution to equation (6) given the data $\boldsymbol{y}_t$, not the innovations $\boldsymbol{\varepsilon}_t$ as in the previous case, and also over different $\boldsymbol{\theta} \in \Theta$ which is crucial in deriving the properties of the estimator. Moreover, in practice, the limit sequence $\{\boldsymbol{f}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ is approximated by the filtered sequence $\{\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}, \hat{\boldsymbol{f}}_1)\}_{t\in\mathbb{N}}$ initialized at some value $\hat{\boldsymbol{f}}_1 \in \mathbb{R}^r$. The chosen starting value is fixed, non-random, and is almost surely incorrect. Therefore, for further proofs of consistency and asymptotic normality, it is important to ensure that the choice of the initial value is irrelevant, in other words, we need to show that $\{\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}, \hat{\boldsymbol{f}}_1)\}_{t\in\mathbb{N}}$ is 'asymptotically SE'. The required form of stability is ensured by the filter invertibility (Straumann & Mikosch, 2006; Wintenberger, 2013; Blasques, van Brummelen, Koopman, & Lucas, 2022). We highlight that even under correct model specification the conditions for the filter invertibility and stationarity are not the same (Blasques et al., 2018), hence this notion of stability requires special treatment.

The proposition below establishes that, under certain conditions, the filtered sequence converges exponentially almost surely (e.a.s.) to a stationary and ergodic limit sequence. We note that the filtered sequence $\hat{\boldsymbol{f}}_t$ is defined recursively and to simplify further notation we write the stochastic recurrence equation (SRE) in (6) as $\hat{\boldsymbol{f}}_{t+1}(\boldsymbol{\theta}) := \boldsymbol{\phi}_t(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}), \boldsymbol{\theta})$ where $\boldsymbol{\phi}_t(\cdot) : \mathbb{R}^r \times \Theta \to \mathbb{R}^r$ is a random function $\forall t \in \mathbb{N}$. We further denote as $\boldsymbol{\phi}_t^{(p)}(\,\cdot\,, \boldsymbol{\theta})$ a function that represents a p-fold backward iterate of the dynamic system. For example, $\boldsymbol{\phi}_t^{(3)}(\boldsymbol{f}, \boldsymbol{\theta}) := \boldsymbol{\phi}_t(\boldsymbol{\phi}_{t-1}(\boldsymbol{\phi}_{t-2}(\boldsymbol{f}, \boldsymbol{\theta}), \boldsymbol{\theta}), \boldsymbol{\theta})$.

Assumption 5. $\{\boldsymbol{y}_t\}_{t\in\mathbb{Z}}$ *is strictly stationary and ergodic (SE).*

Proposition 1 (Properties of the filter)**.** *Let Assumption 5 be satisfied. Moreover, let the following conditions hold*

(i) $\mathbb{E}\log^+ \sup_{\boldsymbol{\theta}\in\Theta} \sup_{\boldsymbol{f}\in\mathbb{R}^r} \left\| \frac{\partial \boldsymbol{\phi}_t(\boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{f}} \right\| < \infty$;

(ii) $\mathbb{E}\log^{+}\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\phi}_t(\boldsymbol{f},\boldsymbol{\theta})\| < \infty$ *for some* $\boldsymbol{f}\in\mathbb{R}^r$*;*

(iii) $\mathbb{E}\log\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\frac{\partial\boldsymbol{\phi}_t^{(p)}(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{f}}\right\| < 0$ *for some integer* $p\geq 1$.

Then, the sequence $\{\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ *initialized at some starting value* $\hat{\boldsymbol{f}}_1\in\mathbb{R}^r$ *converges exponentially almost surely (e.a.s.) to a unique strictly stationary and ergodic (SE) solution* $\{\boldsymbol{f}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ *to equation* (6) *uniformly over* Θ*. We have*

$$\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\|\xrightarrow{e.a.s.}0 \quad \text{as } t\to\infty.$$

Proof. Proposition 3.12 in Straumann & Mikosch (2006). ■

Condition (iii) in Proposition 1 is the so-called contraction condition. The contraction condition is usually employed for $p=1$, i.e. $\mathbb{E}\log\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\|\boldsymbol{B}+\boldsymbol{A}\frac{\partial\boldsymbol{s}_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{f}}\|<0$, see, for example, Blasques, van Brummelen, Koopman, & Lucas (2022). However, for multivariate dynamic systems with $r>1$ this contraction condition is restrictive and is rarely satisfied, see, for example, the discussion in Pötscher & Prucha (1997, Chapter 6.4). Therefore, we instead verify that the contraction condition holds for the p-th iterate.

For our main examples, below, we discuss extensively condition (iii) and verify conditions (i) and (ii) of Proposition 1 in the Supplementary Appendix A.

EXAMPLE 1 (Ctd., Gaussian model. Contraction condition)**.** *For the Gaussian model, the contraction condition for* $p=1$ *takes the form* $\log\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{B}-\boldsymbol{A}\|<0$*, which is implied by* $\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{B}-\boldsymbol{A}\|<1$*. Here, due to the fact that matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are diagonal and the updating equation is linear in* $\boldsymbol{f}_t(\boldsymbol{\theta})$*, the contraction condition is easily satisfied for* $p=1$ *as long as the parameter space* Θ *is compact. We also note that the condition required for the filter invertibility turns out to be stronger than the condition required for the existence of the SE solution in Lemma 2. Furthermore, it can be shown, using a similar argument as in the proof of Lemma 2, that for the Gaussian model the contraction condition is both necessary and sufficient for the filter to be invertible since the updating equation is linear in* $\boldsymbol{f}_t(\boldsymbol{\theta})$*.*

EXAMPLE 2 (Ctd., Student's t model. Contraction condition)**.** *For the model with*

Student's t innovations, the contraction condition for $p = 1$ becomes

$$\mathbb{E} \log \sup_{\boldsymbol{\theta} \in \Theta} \sup_{\boldsymbol{f} \in \mathbb{R}^r} \left\| \boldsymbol{B} + \frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|, \nu)} \boldsymbol{A} \times \left(\frac{2}{\nu \left(1 + \frac{\|\tilde{\boldsymbol{y}}_t\|^2}{\nu}\right)} \left(\frac{1}{N} \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \right)^{-1} \right. \right.$$
$$\left. \left. \times \frac{1}{N} \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y}_t - \boldsymbol{\Lambda} \boldsymbol{f} \right) \left(\boldsymbol{y}_t - \boldsymbol{\Lambda} \boldsymbol{f} \right)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} - \boldsymbol{I}_r \right) \right\| < 0.$$

Therefore, as long as Θ is compact, for $\sup_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{B}\| < 1$ and sufficiently small elements of $\boldsymbol{A}$, the contraction condition (iii) might hold. However, for $r > 1$, this parameter region is still very restrictive and is barely satisfied in practice. For $p > 1$ the contraction condition is likely to hold, the system can still be 'stable'.

For $p > 1$, an analytical expression of the contraction condition is cumbersome to analyze. Hence, in practice, as in Blasques, Francq, & Laurent (2022) and Blasques, van Brummelen, Gorgi, & Koopman (2022), for the Student's t model, we suggest verifying the contraction condition for a sufficiently large p using a feasible invertibility condition introduced in Blasques et al. (2018).

We complete this subsection with a lemma that ensures the existence of bounded moments uniformly over Θ of the limit sequence $\boldsymbol{f}_t(\boldsymbol{\theta})$ of the filter that is formulated in terms of the data $\boldsymbol{y}_t$ rather than in terms of the innovations $\boldsymbol{\varepsilon}_t$. The lemma is useful for establishing the theoretical properties of the estimator, such as consistency and asymptotic normality. We note that the result can be obtained under two different sets of conditions depending on whether the score is bounded or not.

Lemma 4. *Let all the assumptions and conditions of Proposition 1 hold. If, furthermore, there exists $k > 0$ such that*

(i.A) $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \sup_{\boldsymbol{f} \in \mathbb{R}^r} \|\boldsymbol{s}_t(\boldsymbol{f}, \boldsymbol{\theta})\|^k < \infty$;

(ii.A) $\sup_{\boldsymbol{\theta} \in \Theta} \sup_{(\boldsymbol{y}, \boldsymbol{f}) \in \mathbb{R}^N \times \mathbb{R}^r} \left\| \frac{\partial \boldsymbol{\phi}_t(\boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{f}} \right\| < 1$;

then, the filter limit sequence as defined in Proposition 1 satisfies $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{f}_t(\boldsymbol{\theta})\|^k < \infty$. If, in addition or alternatively,

(ii.B) there exists a constant $\bar{d} > 0$ such that $\sup_{\boldsymbol{\theta} \in \Theta} \sup_t \|\boldsymbol{s}_t(\boldsymbol{f}_t, \boldsymbol{\theta})\| < \bar{d} < \infty$;

(ii.B) $\sup_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{B}\| < 1$;

then, the filter limit sequence as defined in Proposition 1 satisfies $\sup_t \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}_t(\boldsymbol{\theta})\|^k < \infty$.

The first set of assumptions $(i.A) - (ii.A)$ is a standard and general set of assumptions for establishing bounded moments of the filter in a score-driven framework. However, for multivariate nonlinear models, condition $(ii.A)$ can be very restrictive, which is similar to the issue with the contraction condition discussed above. However, as long as the score is uniformly bounded in t (condition $(i.B)$) and condition $(ii.B)$ holds, the factors possess moments of any order.

EXAMPLE 2 (Ctd., Student's t model. Condition $(ii.B)$)**.** *First, we rewrite the expression for the score* $\boldsymbol{s}_t$ *as follows*

$$\begin{aligned}\boldsymbol{s}_t(\boldsymbol{\theta}) := \boldsymbol{s}(\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta}) &= \frac{N+\nu+2}{\sqrt{\nu}} \left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1} \frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1/2} \\ &\times \frac{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))/\sqrt{\nu}}{1 + (\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))/\nu} \\ &= \frac{N+\nu+2}{\sqrt{\nu}} \boldsymbol{P} \frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1/2} \frac{\boldsymbol{x}_t}{1 + \boldsymbol{x}_t^\top \boldsymbol{x}_t},\end{aligned}$$

where $\boldsymbol{x}_t := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t)/\sqrt{\nu}$. *We notice that as* $\|\boldsymbol{x}_t\| \to 0$ *or as* $\|\boldsymbol{x}_t\| \to \infty$, $\|\boldsymbol{s}_t\| \to 0$ *since* $\Theta \subseteq \mathbb{R}^q$ *and* $\boldsymbol{x}_t \in \mathbb{R}^N$ *as long as* $\|\boldsymbol{P}\| < \infty$, $\boldsymbol{\Sigma} \succ 0$ *and* $0 < \nu < \infty$. *Therefore, the score sequence is uniformly bounded, i.e.* $\sup_t \|\boldsymbol{s}_t(\boldsymbol{\theta})\| < \bar{d}(\boldsymbol{\theta}) < \infty$ *for all* $\boldsymbol{\theta} \in \Theta$. *We further obtain* $\sup_{\boldsymbol{\theta}\in\Theta} \sup_t \|\boldsymbol{s}_t(\boldsymbol{\theta})\| < \bar{d} < \infty$ *as long as the parameter space* Θ *is compact.*

2.3 Identification conditions

In factor models, loadings and factors are not separately identifiable, meaning that they are subject to a rotational indeterminacy problem. There are several ways of imposing restrictions on the loadings and covariance structure of the factors to resolve the rotational indeterminacy problem. Here, we focus on those discussed in Bai & Li (2012), as they are widely used in factor analysis. Furthermore, we discuss the implications of these restrictions on the score-driven updating equations. The conditions are listed in Table 1. We proceed with discussing these conditions in more detail.

IC1. Condition (IC1) requires the upper $r \times r$ block of the matrix of loadings to be equal to the identity matrix while the lower $(N-r) \times r$ block is unrestricted. Essentially,

Table 1: Standard identification restrictions in static factor models.

	Restrictions on $\mathbb{C}ov\boldsymbol{f}_t$	Restrictions on $\boldsymbol{\Lambda}$
IC1	Unrestricted	$\boldsymbol{\Lambda} = (\boldsymbol{I}_r, \boldsymbol{\Lambda}_B^\top)^\top$
IC2	$\mathbb{C}ov\boldsymbol{f}_t$ = diagonal	$\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_A^\top, \boldsymbol{\Lambda}_B^\top)$, where $\boldsymbol{\Lambda}_A$ is an $r \times r$ lower triangular matrix with 1s on the diagonal
IC3	$\mathbb{C}ov\boldsymbol{f}_t = \boldsymbol{I}_r$	$\boldsymbol{\Lambda}^\top = (\boldsymbol{\Lambda}_A^\top, \boldsymbol{\Lambda}_B^\top)^\top$, where $\boldsymbol{\Lambda}_A$ is an $r \times r$ lower triangular matrix with non-zero diagonal elements
IC4	$\mathbb{C}ov\boldsymbol{f}_t$ = diagonal (with distinct elements)	$\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \boldsymbol{I}_r$
IC5	$\mathbb{C}ov\boldsymbol{f}_t = \boldsymbol{I}_r$	$\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$ = diagonal (with distinct elements)

Based on the paper Bai & Li (2012).

this means that the first r factors are measured in terms of the units of the first r series. Additionally, for the upper block, it implies that the first factor is uncorrelated with series $2, \dots r$, the second factor is uncorrelated with series $1, 3, \dots r$, and so on.

The advantage of (IC1) is that no additional restrictions need to be imposed on the factors. For example, factors are allowed to be correlated, oblique factors, making standard estimation methods, such as maximum likelihood, applicable. However, this restriction makes the model not invariant to the ordering of the series. This may lead to a lack of model interpretation and even to an incorrectly specified model. For example, setting $\lambda_{11} = 1$ to meet the restriction may not be appropriate when in the true process $\lambda_{11} = 0$. Furthermore, it can be desirable to conduct inference on the loadings which in this setting would not be possible for the restricted loadings. This problem is in some sense similar to the normalization of the cointegrating vector problem, see Hamilton (1994, Chapter 20). Chan et al. (2018) also discuss the lack of invariance problem in application to factor models in a Bayesian setup.

IC2 & IC3. This type of restriction is adopted by Creal et al. (2014) who introduced the first score-driven factor model for summarizing the co-movements between macroeconomic and financial time series in a few factors. Compared to (IC1), conditions (IC2) and (IC3) impose less restrictions on the matrix of loadings. Intuitively, these conditions allow for more flexibility, as the first factor is allowed to be correlated with all the series, the second factor to be correlated with all but the first series, and so on. However, these identification

conditions are also not order-invariant. Moreover, the relaxation of the restrictions comes at the cost of restricting the covariance matrix of the factors to be diagonal or even the identity matrix. We also note that, although Creal et al. (2014) adopt this restriction, they do not discuss the restrictions on $\mathbb{C}ov(\boldsymbol{f}_t)$ which is required for identification.
Due to the lack of invariance, we do not further proceed with restrictions (IC1)–(IC3).

IC4 & IC5. In turn, conditions (IC4) and (IC5), the so-called orthogonality restrictions, are order-invariant. However, they also impose restrictions both on the factors and loadings. To proceed with the model formulation, we, first, consider the model-implied structure of the covariance matrix $\mathbb{C}ov(\boldsymbol{f}_t)$ under these restrictions assuming the correct model specification.

Lemma 5. *Let all the assumptions and conditions of Lemma 3 hold for $k = 2$. Furthermore, let the parameter space Θ satisfy the identification condition (IC4) on the loadings, i.e. $\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \boldsymbol{I}_r$, and a stochastic process $\{\boldsymbol{f}_t(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ be generated by a score-driven model (1) and (6) with $\boldsymbol{\theta}_0 \in \Theta$. Then,*

$$\mathbb{C}ov(\boldsymbol{f}_t(\boldsymbol{\theta}_0)) = \left(\boldsymbol{I}_r - \boldsymbol{B}^2\right)^{-1} \mathbb{E}\left[\left(\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|}{NW(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right)^2\right] \boldsymbol{A}^2, \tag{8}$$

$$\mathbb{C}ov(\boldsymbol{f}_{t+h}(\boldsymbol{\theta}_0), \boldsymbol{f}_t(\boldsymbol{\theta}_0)) = \boldsymbol{B}^h \mathbb{C}ov(\boldsymbol{f}_t), \tag{9}$$

$$\mathbb{C}ov(\boldsymbol{f}_{t+h}(\boldsymbol{\theta}_0), \boldsymbol{\varepsilon}_t) = \boldsymbol{A}\boldsymbol{B}^{h-1}\boldsymbol{\Lambda}^\top \mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{N^2 W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right], \tag{10}$$

with $\tilde{\boldsymbol{\varepsilon}}_t := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t$.

Given that matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal, the restriction on the loadings and updating equations for the factors imply that $\mathbb{C}ov(f_{mt}, f_{kt}) = 0$ for all $m \neq k$. In other words, the orthogonality restriction on the loadings and the model formulation guarantee that $\mathbb{C}ov(\boldsymbol{f}_t)$ is diagonal. Therefore, if the unconditional variances of the factors differ between the factors, the restriction on the covariance structure of $\boldsymbol{f}_t$ is fulfilled by design, and we only need to ensure that $\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \boldsymbol{I}_r$, which is an order-invariant restriction.

Corollary 1. *Let all the assumptions and conditions of Lemma 3 hold for $k = 2$ and let matrix $\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$ be diagonal (restriction (IC5) on the loadings). If a stochastic process*

$\{\boldsymbol{f}_t(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is generated by a score-driven model (1) and (6) with $\boldsymbol{\theta}_0 \in \Theta$, then, for the restriction on the covariance matrix of the factors in (IC5) to be satisfied, the following condition must hold

$$\left(\boldsymbol{I}_r - \boldsymbol{B}^2\right)^{-1} \mathbb{E}\left[\left(\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|}{NW(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right)^2\right] \left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1} \boldsymbol{A}^2 = \boldsymbol{I}_r.$$

Therefore, by imposing restrictions on the matrix of loadings we guarantee that $\mathbb{C}ov(\boldsymbol{f}_t)$ is diagonal. However, for $\mathbb{C}ov(\boldsymbol{f}_t)$ to be identity we need additional restrictions on the parameter space.

Remark. *Both under (IC4) and (IC5):*

- *$\boldsymbol{\Lambda}$ and $\boldsymbol{f}_t$ are identified up to a column sign change;*
- *in general, the order of the factors and of the columns of matrix $\boldsymbol{\Lambda}$ are identified up to a relabeling. For example, we can always redefine $\tilde{f}_{1t} = f_{2t}$, $\tilde{f}_{2t} = f_{1t}$ and $\tilde{\boldsymbol{\Lambda}}_1 = \boldsymbol{\Lambda}_2$, $\tilde{\boldsymbol{\Lambda}}_2 = \boldsymbol{\Lambda}_1$. That is why, by requiring in (IC4) and (IC5) matrices $\mathbb{C}ov\boldsymbol{f}_t$ and $\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$ to have distinct diagonal elements with decreasing order of magnitude, the ordering issue is resolved.*

To sum up, given the generality, order-invariance and simplicity of the condition (IC4), we conclude that (IC4) is the most suitable condition for an order-invariant score-driven factor model. Then, the score-driven factor model with elliptically distributed innovations takes the following form

$$\boldsymbol{y}_t = \sum_{j=1}^{r} \boldsymbol{\Lambda}_j f_{jt} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{E}_N(\boldsymbol{0}, \boldsymbol{\Sigma}, g), \tag{11}$$

$$\boldsymbol{f}_{t+1} = \boldsymbol{A}\frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|, g)}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t\right) + \boldsymbol{B}\boldsymbol{f}_t, \tag{12}$$

with the static parameters as defined in Section 2.1.

Model (11)–(12) will be the baseline model in our further discussion. We call this model an *order-invariant score-driven dynamic factor model*. Below, we present the updating equation for our main examples under (IC4).

EXAMPLE 3 (Gaussian and Student's t models. Updating equation for common factors)**.** *For the Gaussian and Student's t models, the updating equation for the dynamic factor* $\boldsymbol{f}_t$ *is given by*

$$\boldsymbol{f}_{t+1} = \boldsymbol{A}\frac{1}{W_t}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - \boldsymbol{f}_t\right) + \boldsymbol{B}\boldsymbol{f}_t,$$

where for the Gaussian model $W_t = 1$ *and for the Student's t model* $W_t = \frac{\nu}{(N+\nu+2)}(1 + \frac{(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t)}{\nu})$.

2.4 Model extensions

2.4.1 Score-driven group-factor model

A factor model with group-factor structure is a special case of the standard factor model. In this model, not all the factors are common to all the series but rather there are some common factors as well as group-specific ones. The observation equation, in vector form, remains unchanged, however, the matrix of loadings has some elements set to zero. For example, in the case of $c = 2$ groups, the matrix of loadings has the following structure

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{21} & \boldsymbol{0}_K \\ \boldsymbol{\Lambda}_{12} & \boldsymbol{0}_{N-K} & \boldsymbol{\Lambda}_{32} \end{bmatrix},$$

where K is the number of series in group 1.

Since $\boldsymbol{\Sigma}$ is diagonal, the identifying restriction $\frac{1}{N}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \boldsymbol{I}_r$ can be split into c orthogonality constraints with c being the number of groups. For $c = 2$, we have

$$\frac{1}{K}\begin{bmatrix}\boldsymbol{\Lambda}_{11}\\ \boldsymbol{\Lambda}_{21}\end{bmatrix}\boldsymbol{\Sigma}^{-1}_{K\times K}\begin{bmatrix}\boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{21}\end{bmatrix} = \boldsymbol{I}_{r_c+r_1}, \qquad \frac{1}{N-K}\begin{bmatrix}\boldsymbol{\Lambda}_{12}\\ \boldsymbol{\Lambda}_{32}\end{bmatrix}\boldsymbol{\Sigma}^{-1}_{N-K\times N-K}\begin{bmatrix}\boldsymbol{\Lambda}_{12} & \boldsymbol{\Lambda}_{32}\end{bmatrix} = \boldsymbol{I}_{r_c+r_2},$$

where $\boldsymbol{\Sigma} := \text{diag}(\boldsymbol{\Sigma}_{K\times K}, \boldsymbol{\Sigma}_{(N-K)\times(N-K)})$, r_c is the number of common factors and r_i with $i = \{1, 2\}$ denotes the number of group-specific factors corresponding to group i, hence, the total number of factors is $r = r_c + r_1 + r_2$.

Combining these two restrictions we obtain

$$\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{I}_{r_c} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{K}{N}\boldsymbol{I}_{r_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{N-K}{N}\boldsymbol{I}_{r_2} \end{bmatrix}. \tag{13}$$

Restriction (13) is a modified version of the (IC4) orthogonality restriction. Intuitively, this restriction accounts for the fact that all the series are subject to the common factor, K/N proportion of the series is subject to the first group-specific factor, and $(N-K)/N$ proportion of the series is subject to the second group-specific factor.

Given restriction (13), the updating equation for the common factors remains unchanged. For the group-specific factors, the updating equations are adapted as follows

$$\boldsymbol{f}^j_{t+1} = \boldsymbol{A}_j \frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|, g)} \frac{1}{K_j} \boldsymbol{\Lambda}_j^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t) + \boldsymbol{B}_j \boldsymbol{f}^j_t,$$

where $j = 1, \ldots, c$ and K_j corresponds to the number of series in group j. We notice that the updating equation takes into account the fact that only the proportion of the series is related to the corresponding group-specific factor. Below, we state the updating equations for the group-specific factors for the Student's t model.

EXAMPLE 4 (Gaussian and Student's t models. Updating equations for group-specific factors)**.** *For a group-factor model with Student's t innovations, the group-specific factor j is updated as follows*

$$\boldsymbol{f}^j_{t+1} = \boldsymbol{A}_j \frac{(N+\nu+2)}{\nu} \frac{1}{W_t} \left(\frac{1}{K_j} \boldsymbol{\Lambda}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_t - \boldsymbol{f}_t \right) + \boldsymbol{B}_j \boldsymbol{f}^j_t,$$

with W_t as defined in Example 3.

2.4.2 Temporal dependence in the innovation term

In practice, Assumption 3 of the i.i.d. innovations can be very restrictive. Intuitively, it means that all the dynamic effects in the time series are only due to common factors, but excludes the possibility of individual time series having their own dynamic effects. We relax

this assumption and allow $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}}$ to follow a restricted AR(1) process, i.e.

$$\boldsymbol{\varepsilon}_{t+1} = \rho\boldsymbol{\varepsilon}_t + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim \mathcal{E}_N(\mathbf{0}, \boldsymbol{\Omega}, g), \quad t = 1, \dots, T.$$

Applying the lag operator $(1 - \rho L)$ to the original model equations, this implies that

$$\begin{aligned} \boldsymbol{y}_t &= \rho\boldsymbol{y}_{t-1} + \boldsymbol{\Lambda}\boldsymbol{f}_t^\star + \boldsymbol{u}_t, \\ \boldsymbol{f}_t &= \boldsymbol{f}_t^\star + \rho\boldsymbol{f}_{t-1}, \end{aligned} \tag{14}$$

where the vector of innovations $\boldsymbol{u}_t$ is i.i.d.

Hence, model (11)–(12) can be rewritten in the following form

$$\begin{aligned} \boldsymbol{y}_t &= \rho\boldsymbol{y}_{t-1} + \boldsymbol{\Lambda}\boldsymbol{f}_t^\star + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim \mathcal{E}_N(\mathbf{0}, \boldsymbol{\Omega}, g), \quad t = 1, \dots, T, \\ \boldsymbol{f}_{t+1}^\star &= \boldsymbol{A}\frac{1}{W(\|\tilde{\boldsymbol{y}}_t^\star\|, g)}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Omega}^{-1}\left(\boldsymbol{y}_t^\star - \boldsymbol{\Lambda}\boldsymbol{f}_t^\star\right) + \boldsymbol{B}\boldsymbol{f}_t^\star, \end{aligned} \tag{15}$$

with $\boldsymbol{y}_t^\star \coloneqq \boldsymbol{y}_t - \rho\boldsymbol{y}_{t-1}$ and $\tilde{\boldsymbol{y}}_t^\star \coloneqq \boldsymbol{\Omega}^{-1/2}(\boldsymbol{y}_t^\star - \boldsymbol{\Lambda}\boldsymbol{f}_t^\star)$.

Given the relation between $\boldsymbol{f}_t$ and $\boldsymbol{f}_t^\star$, we can recover $\boldsymbol{f}_t$ itself using (14) or as follows

$$\begin{aligned} \boldsymbol{f}_{t+1} =& \boldsymbol{A}\frac{1}{W(\|\tilde{\boldsymbol{y}}_t^\star\|, g)}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Omega}^{-1}[(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t) - \rho\left(\boldsymbol{y}_{t-1} - \boldsymbol{\Lambda}\boldsymbol{f}_{t-1}\right)] \\ &+ (\boldsymbol{B} + \rho\boldsymbol{I}_r)\boldsymbol{f}_t - \rho\boldsymbol{B}\boldsymbol{f}_{t-1}. \end{aligned}$$

Similar to the discussion in Section 2.2, in practice, the filtered sequence $\{\hat{\boldsymbol{f}}_t^\star(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ is defined recursively using the SRE (15). Clearly, under certain conditions, Proposition 1 ensures that the sequence $\{\hat{\boldsymbol{f}}_t^\star(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ converges e.a.s. and uniformly over Θ to a limit sequence $\{\boldsymbol{f}_t^\star(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ that is strictly stationary and ergodic. Then, as the corollary below states, under additional assumptions on the parameter space Θ, it follows that the filter for the dynamic common factor $\{\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ itself converges e.a.s. and uniformly to an SE limit sequence $\{\boldsymbol{f}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$.

Corollary 2 (Properties of the filter)**.** *Let $\{\hat{\boldsymbol{f}}_t^\star(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ be a solution to SRE (15). Let all the assumptions and conditions of Proposition 1 hold, then the sequence $\{\boldsymbol{f}_t^\star(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ converges e.a.s. and uniformly to a unique SE solution $\{\boldsymbol{f}_t^\star(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ to equation (15). If,*

furthermore, $\sup_{\boldsymbol{\theta}\in\Theta}|\rho|<1$*, then the sequence* $\{\boldsymbol{f}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ *converges e.a.s. and uniformly to a unique SE solution* $\{\boldsymbol{f}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ *to equation* (14).

3 Estimation

Parameters of observation-driven models can be estimated by maximum likelihood which in our case would imply

$$\begin{aligned}\hat{\boldsymbol{\theta}}_T &= \arg\max_{\boldsymbol{\theta}\in\Theta}\frac{1}{T}\sum_{t=2}^{T} l_t(\boldsymbol{\theta}) \\ &:= \arg\max_{\boldsymbol{\theta}\in\Theta}\frac{1}{T}\sum_{t=2}^{T}\left(-\frac{1}{2}\log|\boldsymbol{\Sigma}| + \log g\left((\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t)\right)\right),\end{aligned}$$

where $\boldsymbol{\theta}=((\operatorname{diag}\boldsymbol{\Sigma})^\top,\operatorname{vec}(\boldsymbol{\Lambda})^\top,\operatorname{diag}\boldsymbol{A}^\top,\operatorname{diag}\boldsymbol{B}^\top,\boldsymbol{\nu}^\top)^\top$ and $g(\cdot)$ is a density generator that can correspond to, for example, Gaussian or Student's t distributions.

However, as discussed in Section 2.3, in factor models, special attention should be devoted to parameter identification which is crucial for estimation. Therefore, we start our discussion with formulating the conditions required for the model identification and then turn to the detailed discussion of the estimation procedure.

3.1 Parameter identifiability

We begin this section with the summary of the conditions imposed on the parameter space Θ. In particular, the parameter space Θ is such that

A. matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal with $\|\boldsymbol{B}\|<1$;

B. the scale matrix $\boldsymbol{\Sigma}$ is diagonal with elements $0<\underline{c}\le\sigma_i^2\le\bar{c}<\infty$ for all $i=1,\dots,N$.

As discussed in the previous section, the conditions above ensure the existence of the SE solutions with two bounded moments. Next, we summarize the conditions on the parameter space that guarantee identification. For this, we introduce a restricted parameter set $\tilde{\Theta}$, which is a subset of the original set Θ, i.e. $\tilde{\Theta}\subseteq\Theta$, subject to the following restrictions:

C. $\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}=\boldsymbol{I}_r$;

D. the covariance matrix $\boldsymbol{K} := \mathbb{C}ov(\boldsymbol{f}_t)$ has distinct elements on the diagonal, with $\mathbb{C}ov(\boldsymbol{f}_t)$ as specified in equation (8);

E. if any row of matrix $\boldsymbol{\Lambda}$ is deleted, the remaining matrix can be partitioned into two disjoint submatrices of rank r.

Conditions A–D ensure that (IC4) is satisfied, importance of which is extensively discussed in Section 2.3. Condition E comes from the statistical factor model literature, and it guarantees the identification of the variance of the common and idiosyncratic components (Anderson et al., 1956). The proposition below establishes parameter identification under the assumptions and conditions stated above.

Proposition 2 (Identification for correctly specified models)**.** *Let all the assumptions and conditions of Lemma 3 hold for $k = 2$. Furthermore, let the parameter space $\tilde{\Theta}$ satisfy conditions A–E and let the observed data $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ be a subset of a stochastic process $\{\boldsymbol{y}_t(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ generated by a score-driven model (11)–(12) and with $\boldsymbol{\theta}_0 \in \tilde{\Theta}$. Then, $\boldsymbol{\theta}_0$ is set identifiable.*
If, in addition, parameter space $\tilde{\Theta}$ is such that the sign(λ_{ik}) is known for some $i = 1, \ldots, N$ and for all $k = 1, \ldots, r$, then $\boldsymbol{\theta}_0$ is identifiable meaning that it is not observationally equivalent to any other parameter $\boldsymbol{\theta} \in \tilde{\Theta}$, i.e. $\mathbf{p}(\boldsymbol{y}; \boldsymbol{\theta}) \neq \mathbf{p}(\boldsymbol{y}; \boldsymbol{\theta}_0)$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and some $\boldsymbol{y}$ in a set of non-zero probability.

3.2 Estimation procedure

Given conditions A–E on the parameter space $\tilde{\Theta}$, several constraints are imposed on the parameters. Particularly, condition C imposes a nonstandard nonlinear constraint, making standard constrained optimization inapplicable. Therefore, to ensure that the constraint is fulfilled we resort to optimization on a Stiefel manifold. The Stiefel manifold defines a set of matrices $\boldsymbol{X}$ that satisfy an orthogonality constraint, i.e. $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_r$. For an introduction to manifolds, see Edelman et al. (1998); Boumal (2023). The details on Python implementation can be found in Townsend et al. (2016).

Standard optimization methods, like gradient descent, are not available when carrying out optimization on manifolds since manifold space is nonlinear. Therefore, the existing

optimization methods on manifolds exploit the idea of moving in the direction of the tangent space while remaining on the manifold. In other words, given an initial 'guess', the search continues along the tangent direction and a retraction map ensures that the next search point remains on the manifold.

Hence, to fulfill condition C, we restrict matrix $\boldsymbol{X} := \frac{1}{\sqrt{N}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}$ to lie on the Stiefel manifold. Further restrictions on the loadings, such as those implied by economic theory, can be imposed by using the approach introduced by Liu & Boumal (2020). The remaining parameters diag $\boldsymbol{A}$, diag $\boldsymbol{\Sigma}$ and $\boldsymbol{\nu}$ are reparameterized to ensure the positivity constraints but, in general, they do not require constrained optimization, hence their reparameterized counterparts lie on a Euclidean manifold. Since the Cartesian product of manifolds forms a manifold (Boumal, 2023, Proposition 3.14), in practice, the whole optimization problem is still an optimization on a manifold. In the next subsection, we further discuss the properties of the estimator.

3.3 Asymptotic properties of the constrained estimator

In this section, we establish the consistency and asymptotic normality of the constrained ML estimator. We note that, as typical for score-driven models, the log-likelihood function depends on the filtered sequence $\{\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$. Hence, first, we define the empirical average log-likelihood based on the filtered sequence $\{\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ and on the limit sequence $\{\boldsymbol{f}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$, respectively, as follows

$$\begin{aligned}\hat{\mathcal{L}}_T(\boldsymbol{\theta}) &:= \frac{1}{T}\sum_{t=2}^{T}\hat{l}_t\left(\boldsymbol{\theta}\right) = \frac{1}{T}\sum_{t=2}^{T} l(\boldsymbol{y}_t, \hat{\boldsymbol{f}}_t(\boldsymbol{\theta}), \boldsymbol{\theta}),\\ \mathcal{L}_T(\boldsymbol{\theta}) &:= \frac{1}{T}\sum_{t=2}^{T} l_t\left(\boldsymbol{\theta}\right) = \frac{1}{T}\sum_{t=2}^{T} l\left(\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta}\right).\end{aligned} \tag{16}$$

It is also important to highlight that the loadings and factors are identified up to a sign change, hence the limit criterion function has two global maxima. This implies that one of the conditions, identifiable uniqueness, required for the consistency of an M-estimator to a single point is clearly violated. However, it is possible to establish the consistency of the estimator towards the set of the maximizers of the limit criterion function $L_\infty(\boldsymbol{\theta}) := \mathbb{E}[l_t(\boldsymbol{\theta})]$ (Pötscher & Prucha, 1997). Alternatively, to ensure the consistency towards the point, one

can impose restrictions on the sign of one of the rows of the matrix of loadings as in Proposition 2.

Assumption 6. *The parameter space* Θ *is compact.*

Assumption 7. $\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{l}_t(\boldsymbol{\theta}) - l_t(\boldsymbol{\theta})\right| \xrightarrow{e.a.s.} 0$ *as* $t \to \infty$.

Assumption 8. $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\log g((\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta})))\right| < \infty$.

Theorem 1 (Consistency of the constrained ML under correct model specification)**.** *Let the assumptions and conditions of Propositions 1 and 2 hold. Furthermore, let Assumptions 6–8 be satisfied. Then, the constrained ML estimator* $\hat{\boldsymbol{\theta}}_T$ *is strongly consistent to* $\Theta_0^\star$ *for any filter initialization* $\hat{\boldsymbol{f}}_1 \in \mathbb{R}^r$*,*

$$\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \Theta_0^\star \quad as \quad T \to \infty,$$

where $\Theta_0^\star = \arg\max_{\boldsymbol{\theta}\in\tilde{\Theta}} L_\infty(\boldsymbol{\theta})$.
If, parameter space $\tilde{\Theta}$ *is such that the sign*(λ_{ik}) *is known for some* $i = 1, \ldots, N$ *and for all* $k = 1, \ldots, r$*, then the constrained ML estimator* $\hat{\boldsymbol{\theta}}_T$ *is strongly consistent to* $\boldsymbol{\theta}_0$.

Assumption 7 essentially ensures that the filter initialization has a negligible effect on the empirical likelihood. As we show below, the filter invertibility will be sufficient for this assumption to hold. In turn, Assumption 8 ensures that the log-likelihood function is bounded, hence it allows application of the ergodic law of large numbers in the proof. We verify the assumptions for the Gaussian model in the Supplementary Appendix A.1. Below, we verify that the assumptions of the Theorem 1 hold for the Student's t model.

EXAMPLE 2 (Ctd., Student's t model. Assumptions 7 and 8)**.** *First, we verify that Assumption 8 holds for the score-driven model with Student's t innovations. We have*

$$\begin{aligned}
&\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\log g\left((\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))\right)\right| \\
&\leq \sup_{\boldsymbol{\theta}\in\Theta}\left|\log\Gamma\left(\frac{N+\nu}{2}\right)\right| + \sup_{\boldsymbol{\theta}\in\Theta}\left|\log\Gamma\left(\frac{\nu}{2}\right)\right| + \frac{N}{2}\sup_{\boldsymbol{\theta}\in\Theta}|\log(\nu\pi)| \\
&+ \sup_{\boldsymbol{\theta}\in\Theta}\frac{N+\nu}{2}\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\log\left(1 + \frac{(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))}{\nu}\right)\right| < \infty.
\end{aligned}$$

The first three terms are bounded as long as $0 < \nu < \infty$ *since the parameter space* Θ *is compact (Assumption 6). For the last term we have that*

$$\begin{aligned}
&\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\log\left(1+\frac{(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))}{\nu}\right)\right| \\
&\leq \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))}{\nu}\right| \\
&\leq \sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{\nu}\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta}))\right\|^2 \\
&\leq \sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{\nu}\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Sigma}^{-1}\|\left(c_r\mathbb{E}\|\boldsymbol{y}_t\|^2+c_r\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Lambda}\|^2\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}_t(\boldsymbol{\theta})\|^2\right), \qquad (17)
\end{aligned}$$

where in the last line we used the Loève's c_r *inequality. Given the correct model specification, by Lemma 3* $\mathbb{E}\|\boldsymbol{y}_t\|^2 < \infty$ *as long as* $\boldsymbol{\Sigma} \succ 0$ *and* $\nu > 2$*. By Lemma 4, for the filter limit sequence we also have* $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}_t(\boldsymbol{\theta})\|^k < \infty$ *for any* k*. Hence, given the conditions on the parameter space stated in Section 3.1, the whole expression in (17) is bounded.*

Now, we turn to Assumption 7. By the mean-value theorem, we have

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{l}_t(\boldsymbol{\theta})-l_t(\boldsymbol{\theta})\right| \leq \sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}}\right\|\sup_{\boldsymbol{\theta}\in\Theta}\left\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\right\|.$$

The sequence $\left\{\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}}\right\|\right\}_{t\in\mathbb{Z}}$ *is SE since it is a continuous function of* $\boldsymbol{y}_t$*, which by Lemma 2 given Assumptions 3 and 4 and conditions A-C is SE. Since by Proposition 1 the filter is uniformly invertible,* $\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\| \xrightarrow{e.a.s.} 0$ *as* $t\to\infty$*, by Lemma 2.1 in Straumann & Mikosch (2006),* $\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{l}_t(\boldsymbol{\theta})-l_t(\boldsymbol{\theta})\right| \xrightarrow{e.a.s.} 0$ *as* $t\to\infty$ *as long as the SE sequence* $\left\{\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}}\right\|\right\}_{t\in\mathbb{Z}}$ *has a logarithmic moment. The latter follows since we have*

$$\begin{aligned}
&\mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\frac{\partial\log g((\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}))}{\partial\boldsymbol{f}}\right\| \\
&= \mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|2\frac{N+\nu}{\nu}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\frac{(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})}{1+(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})/\nu}\right\| \\
&\leq \log^+\sup_{\boldsymbol{\theta}\in\Theta}\frac{2(N+\nu)}{\sqrt{\nu}}+\log^+\sqrt{N}+\mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\frac{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})/\sqrt{\nu}}{1+(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})/\nu}\right\|. \qquad (18)
\end{aligned}$$

The first term in the expression above is bounded since the parameter space Θ is compact as long as $0 < \nu < \infty$. Let us now consider the term $\boldsymbol{v}_t(\boldsymbol{\theta}) = \frac{\boldsymbol{x}_t(\boldsymbol{\theta})}{1+\boldsymbol{x}_t(\boldsymbol{\theta})^\top \boldsymbol{x}_t(\boldsymbol{\theta})}$, where $\boldsymbol{x}_t(\boldsymbol{\theta}) := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})/\sqrt{\nu}$. Clearly, for all $\boldsymbol{\theta} \in \Theta$, as $\|\boldsymbol{x}_t(\boldsymbol{\theta})\| \to 0$ or $\|\boldsymbol{x}_t(\boldsymbol{\theta})\| \to \infty$ we have $\|\boldsymbol{v}_t(\boldsymbol{\theta})\| \to 0$. Therefore, the last term in (18) is uniformly bounded in $(\boldsymbol{y}_t, \boldsymbol{f})$. Since parameter space Θ is compact (Assumption 6), we have that $\mathbb{E} \sup_{\boldsymbol{\theta}} \sup_{\boldsymbol{f} \in \mathbb{R}^r} \|\boldsymbol{v}_t(\boldsymbol{\theta})\| < \infty$.

We complete this section with the asymptotic distribution of the ML estimator. First, we state the high-level assumptions required for the proof of the theorem.

Assumption 9. $\boldsymbol{\theta}_0 \in \text{int}(\Theta)$.

Assumption 10. $\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}_T(\hat{\boldsymbol{\theta}}_T) - \mathcal{I}(\boldsymbol{\theta}_0)\| \xrightarrow{P} 0 \quad$ *as* $T \to \infty$, *where* $\mathcal{I}(\boldsymbol{\theta}) := \mathbb{E}\left[\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]$ *and* $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}) := \frac{\partial^2 \mathcal{L}_T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$.

Assumption 11. *For the derivative of the filter we have* $\|\hat{\boldsymbol{f}}_t'(\boldsymbol{\theta}_0) - \boldsymbol{f}_t'(\boldsymbol{\theta}_0)\| \xrightarrow{e.a.s.} 0$ *as* $t \to \infty$, *where* $\{\boldsymbol{f}_t'(\boldsymbol{\theta}_0)\}_{t \in \mathbb{Z}}$ *is a limit sequence which is unique and SE.*

Assumption 12. $\left\| \frac{\partial \hat{l}_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} - \frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\| \xrightarrow{e.a.s.} 0$ *as* $t \to \infty$.

Assumption 13. *The following moment conditions hold:*

(i) $\mathbb{E}\left\| \frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\|^2 < \infty$;

(ii) $\mathbb{E} \log^+ \sup_{\boldsymbol{f} \in \mathbb{R}^r} \left\| \frac{\partial^2 l_t(\boldsymbol{f}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{f} \partial \boldsymbol{f}^\top} \right\| < \infty$;

(iii) $\mathbb{E}\left\| \frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0)}{\partial \boldsymbol{f}} \right\|^\delta < \infty$ *for some* $\delta > 0$ *and* $\mathbb{E}\|\boldsymbol{f}_t'(\boldsymbol{\theta}_0)\|^n < \infty$ *with* $n \geq \frac{2\delta}{2-\delta}$.

Theorem 2 (Asymptotic normality of the constrained ML estimator)**.** *Let the assumptions and conditions of Theorem 1 hold. Furthermore, let Assumptions 9–13 be satisfied. Then, for any filter initialization $\hat{\boldsymbol{f}}_1 \in \mathbb{R}^r$, the constrained ML estimator $\hat{\boldsymbol{\theta}}_T$ satisfies*

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{P}\boldsymbol{V}\boldsymbol{P}) \quad \text{as} \quad T \to \infty,$$

where $\boldsymbol{P} := \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{Q}(\boldsymbol{Q}^\top \boldsymbol{H}^{-1}\boldsymbol{Q})^{-1}\boldsymbol{Q}^\top \boldsymbol{H}^{-1}$, $\boldsymbol{H}(\boldsymbol{\theta}_0) := \mathcal{I}(\boldsymbol{\theta}_0) + \boldsymbol{Q}(\boldsymbol{\theta}_0)\boldsymbol{Q}^\top(\boldsymbol{\theta}_0)$, $\boldsymbol{V}(\boldsymbol{\theta}_0) := \mathbb{E}[l_t'(\boldsymbol{\theta}_0) l_t'(\boldsymbol{\theta}_0)^\top]$, $\mathcal{I}(\boldsymbol{\theta}_0) := \mathbb{E}[l_t''(\boldsymbol{\theta}_0)]$, and $\boldsymbol{Q}(\boldsymbol{\theta}_0) := \nabla_{\boldsymbol{\theta}} \boldsymbol{C}(\boldsymbol{\theta}_0)^\top$ with $\boldsymbol{C}(\boldsymbol{\theta})$ collecting the stacked (IC4) constraints on the loadings.

The proof of this theorem and explicit expressions for all the matrices stated above are provided in the appendices. In appendices, we verify all the conditions of the theorem for our main examples. Specifically, for Gaussian and Student's t models, we verify Assumptions 10 and 11 which results in Lemmas SB.3 and SB.4, respectively. For the examples, Assumption 12 is verified in the Supplementary Appendix A. Finally, Assumption 13 holds given the result in Lemmas SB.5 and SB.6 together with the assumption of two bounded moments required by Proposition 2.

4 Monte Carlo simulations

In this section, we assess the finite sample properties of the constrained maximum likelihood estimator and how well the score-driven filter captures the dynamics of the factors. We consider a simulation design where the data generating process (DGP) is either Gaussian or Student's t score-driven factor model. Additional Monte Carlo simulation setups and results are presented in the Supplementary Appendix D.

4.1 Simulation design

As a DGP, we use Gaussian and Student's t score-driven dynamic factor models, i.e. $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}_N, \boldsymbol{\Sigma})$ or $\boldsymbol{\varepsilon}_t \sim \boldsymbol{t}_\nu(\mathbf{0}_N, \boldsymbol{\Sigma})$, given by equations (11) and (12). In the simulations, we set the values of the static factors' parameters to $\alpha_k = \beta_k = 0.9 - 0.1 \times (k-1)$. That is, in case of $r = 3$ the values of the parameters considered for the simulations are as follows

$$\boldsymbol{A} = \text{diag}(0.9, 0.8, 0.7), \quad \boldsymbol{B} = \text{diag}(0.9, 0.8, 0.7).$$

The choice of the parameter values ensures that matrix $\mathbb{C}ov(\boldsymbol{f}_t)$ has distinct elements on the diagonal with a decreasing order of magnitude, required for the identification. For the idiosyncratic errors, we consider $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ with $\sigma_i^2 \sim U([0.1, 1.1])$ and, in the case of the Student's t factor model, we set $\nu = 5$.

The matrix of loadings is generated from a standard normal distribution. The loadings are further rotated to satisfy condition C, which is $\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \boldsymbol{I}_r$. Due to the rotation, the final values of matrix $\boldsymbol{\Lambda}$ used for generating the time series vary between different

simulation setups with different values of r and/or N. As a result, the experiments with different values of N and r are not directly comparable, but the experiments with different values of T are directly comparable.

In the simulations, we consider different values of N, T, and r, namely, $N = \{10, 20\}$, $T = \{300, 500, 1000\}$, and $r = \{1, 2, 3\}$. Throughout the simulation study, the number of replications is set to 1000. In all the experiments, the number of factors is assumed to be known.

We emphasize that the identification of the factors and loadings is up to a sign and relabeling, as stated in Remark 2.3. Therefore, in the simulations, the signs of the estimated factors and loadings are defined such that the correlation coefficient between the estimated and simulated factors is positive. The order, label, of the factors is defined based on the sample unconditional variance of the factors, so that the covariance matrix of the factors has decreasing elements on the diagonal. This is done merely for presenting the precision metrics for each of the factors separately, instead of using canonical correlation (Frobenius norm) metrics. We do not use canonical correlation metrics because we are specifically interested in assessing the ability of the filter to identify each of the factors separately. We note that the labeling of the factors and loadings is subject to the estimation uncertainty, and as a result, there can be an additional variation in the parameter estimates due to the uncertainty in the labeling.

Below, we present the results for Gaussian models. The results for the Student's t model as well as further details can be found in the Supplementary Appendix D. We note that with an increase in N and/or r the number of the loadings increases substantially. Therefore, a good choice for the starting values of the parameters is important for the convergence of the optimization procedure. We propose to initialize the matrix of loadings using the PCA estimates.

4.2 Simulation results

In Figures 1 and 2 we present the simulation results for Gaussian model for a cross-sectional sizes $N = 10$ and $N = 20$, respectively. The goal is to assess the performance of the score-driven filter for extracting the factors. This is done by demonstrating the kernel density

plots for the root mean squared error (RMSE) of the estimated factors, for different values of r and T. For each factor k, the RMSE is computed as

$$RMSE(\hat{f}_k) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(f_{k,t} - \hat{f}_{k,t}\right)^2}, \quad \text{for } k = 1, \ldots, r,$$

where $f_{k,t}$ is the kth true, simulated, factor, while $\hat{f}_{k,t}$ is the kth estimated factor. Similarly, the kernel density of the RMSE for the loadings estimates is plotted, where RMSEs are computed as follows

$$RMSE(\hat{\boldsymbol{\Lambda}}_k) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\lambda_{i,k} - \hat{\lambda}_{i,k}\right)^2}, \quad \text{for } k = 1, \ldots, r.$$

Based on the experiments, we find that overall our estimation procedure performs well. As expected, both the factor and loading estimates improve with an increase in the size of the sample T. We also find that the RMSEs for the factors estimates decrease with an increase in N, while the loadings estimates are unaffected. Intuitively, as the cross-sectional size N increases, there is more information about the factors, which leads to improvements in the factors' estimates. At the same time, it increases the total number of the loadings parameters substantially which, in turn, increases the estimation uncertainty.

The RMSEs are the smallest for the first factor which is not surprising since for the identification the order of the factors is in the decreasing order of magnitude, and hence the first factor has the largest variation. The RMSEs for the model with $r = 1$ are considerably smaller than for the model with $r = 2$ and $r = 3$ factors. This result can have two potential explanations. The first one is that the identification is more challenging when two or more factors are present. The second explanation is that this result can also be driven by the relabeling issue. However, given that we do not observe several peaks in the distribution of the RMSEs, the latter should not have a significant impact on the results. Further increase in r does not significantly affect the results for the factors and loadings, indicating that the estimation procedure can identify the factors separately even in the presence of several factors.

The results for other static parameters as well as the results for the Student's t model are

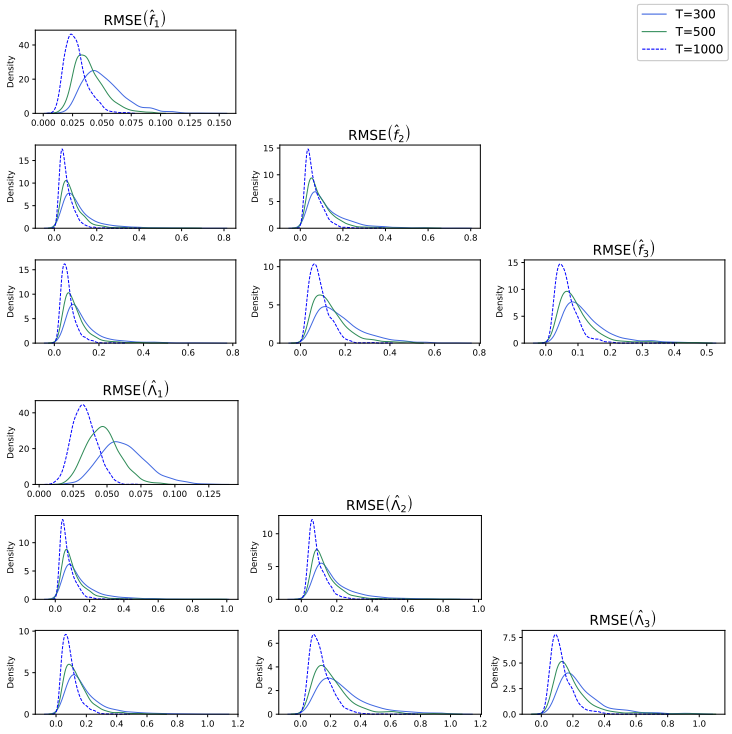


Figure 1: **Kernel density plots of the RMSEs for the factors and loadings.** Monte Carlo simulation results for different values of T and r, $T = \{300, 500, 1000\}$ and $r = \{1, 2, 3\}$. The DGP is a Gaussian score-driven factor model with $N = 10$. The top panel displays the kernel density of the RMSE for the factors, while the bottom panel presents the results for the loadings. The results are based on 1000 Monte Carlo replications.

presented in Supplementary Appendix D. The results in Tables SD.1–SD.4 provide further confirmation that the estimation procedure accurately estimates the static parameters. First, the biases are negligible for all the parameters across all the simulation setups. Moreover, the biases and standard errors decrease with an increase in the sample size T.

5 Empirical application

In this section, we provide an empirical illustration for extracting economic activity indicators from a panel of macroeconomic and financial time series with a specific focus on analyzing the importance of the robustness features in the model. The indicators often have a business and/or financial cycle interpretation and, hence, are of special interest in applied work. In our empirical illustration, we analyze the co-movements between the series and interpret the extracted factors. Given that the model is dynamic, the forecasts and impulse response functions follow straightforwardly from the model. The indicators

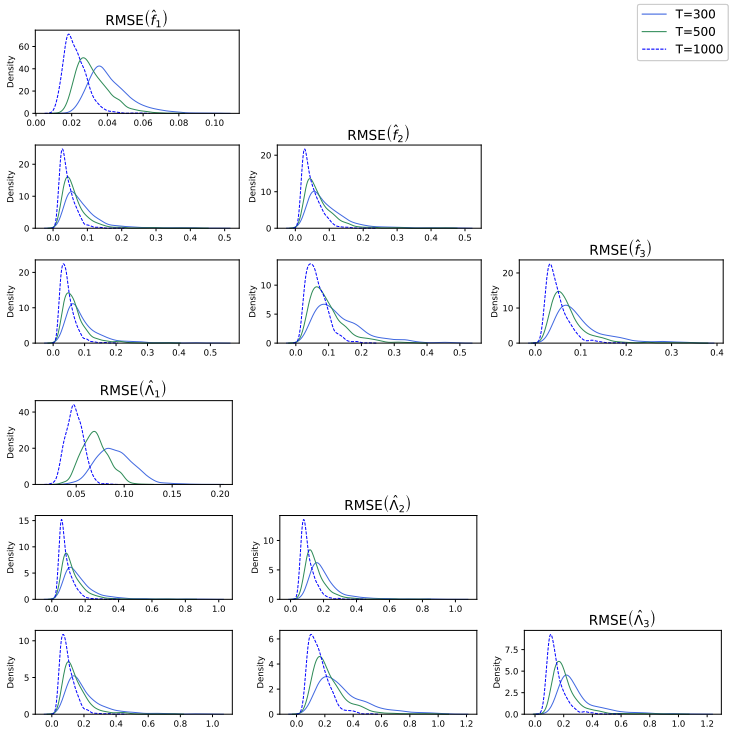


Figure 2: **Kernel density of the RMSEs for the factors and loadings.** Monte Carlo simulation results. The DGP is a Gaussian score-driven factor model with $N = 20$. For further details we refer to Figure 1.

can further be used to construct economic leading indicators or used as control variables for analyzing shocks to economic indicators, see, for example, Loria et al. (2022).

5.1 Data

For our analysis, we use monthly macroeconomic and financial time series for the US starting from January 1981 until February 2022. We follow Creal et al. (2014) and consider the following time series: (i) the annualized industrial production growth rate (INDPRO), (ii) the annual change in the unemployment rate (UNRATE), (iii) the spread between the yield on Baa-rated corporate bonds and the yield on 10-year Treasury bonds (BAATB10Y), (iv) annualized S&P500 returns (S&P500), and (v) stock market volatility (S&P500vol). We further expand the panel by including additional macro-finance variables that are often considered for constructing business and financial cycle indicators, see, for example, Loria et al. (2022). Particularly, given the increasing importance of the trade and service sectors, we append to the dataset the annual change in the log retail sales (RETAIL), $\log RS_t -$

$\log RS_{t-12}$, where RS_t stands for the retail sales at the end of month t. Moreover, to capture the consumer sentiment, we include the annual change in the consumer sentiment index constructed by the University of Michigan (UMSENTx), $CS_t - CS_{t-12}$, where CS_t is the consumer confidence index at the end of month t. Finally, for the finance sector, we consider the annual change in the housing starts index (HOUSING), $HS_t - HS_{t-12}$, as well as the excess bond premium (EBP) index. The EBP index is a novel time series introduced by Gilchrist & Zakrajšek (2012) and updated by Favara et al. (2016). It has been recently recognized as an important time series for constructing economic coincident indicators. The EBP index is constructed as a difference between the average of bond credit spreads and the average of the predicted credit spreads and, intuitively, it captures the risk appetite in the corporate bond market. For further details on the dataset, we refer to the Supplementary Appendix E.

All the series are standardized beforehand such that the mean is equal to zero and the standard deviation to one (Figure 3). Clearly, many time series tend to co-move together with the degree of co-movements becoming stronger during recessions. We also notice that the recession periods are characterized by different shapes. For example, the COVID-19 recession period has a steep decline but a rather quick recovery (V-shaped recession), while the great recession period is characterized by a long period between the decline and recovery (U-shaped recession). Given the difference in the shapes, for the model comparison, we consider two different time periods: the first sample period, January 1981 – December 2011, ends after the great recession, while the second sample period, January 1981 – February 2022, ends after the COVID-19 recession. The latter corresponds to the full sample. The presence of the spikes at the end of the sample indicates that a model equipped with robustness properties might be preferable for the construction of the economic indicators. We investigate this further in the next subsections.

5.2 Model specification and parameter estimates

In our empirical analysis, we estimate parameters of the introduced Gaussian and Student's t score-driven factor models with AR(1) innovations as presented in Section 2.4.2 and with $r = 1, 2, 3$ factors, giving us a total of 6 model specifications. We do not consider

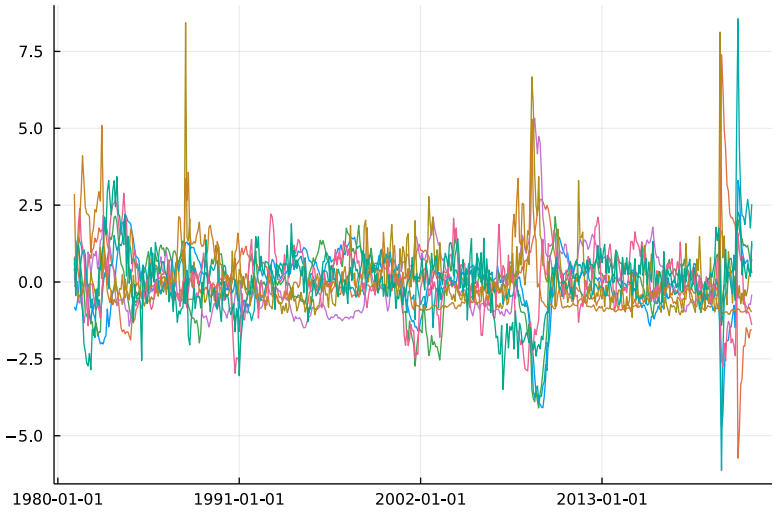


Figure 3: **Time series.**

the specification with i.i.d. errors as it results in significantly worse in-sample model fit. The optimization on manifolds is carried out using the Python package *'Pymanopt'* developed by Townsend et al. (2016). The parameters of the Gaussian models are initialized using PCA estimates, while the parameters of the Student's t models are initialized at the corresponding Gaussian model's parameter estimates. Additionally, to avoid convergence to local maxima, multiple starting values for the parameters are considered.

In Table 2, we report the log-likelihood and Bayesian Information Criterion (BIC) evaluated at the parameter estimates for various model specifications. As mentioned before, we assess the model fit over two different time periods. The results of our analysis in Table 2 suggest that for both sample periods a single factor model with AR(1) innovations provides good fit for the data according to the log-likelihood and BIC. Moreover, we find that regardless of the number of factors included, the BIC is always lower for the Student's t models than for the Gaussian ones. This indicates that the Student's t models provide better in-sample fit to the data compared to the Gaussian models. The differences in the models' fit are particularly pronounced in the full sample period, which includes the COVID-19 recession, where the Student's t models' fit is almost twice better than the fit of the Gaussian models.

Next, we examine the parameter estimates, excluding for now the loadings estimates, which we will turn to later. In Table 3, we present the parameter estimates of the one-factor models selected by the BIC, along with their standard errors. The standard errors are

	$\log L$		BIC	
	$\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$	$t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$	$\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$	$t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$
Sample period 1981-2011				
$r = 1$	-1394.18	-1130.57	2912.59	**2391.29**
$r = 2$	-1357.9	-1110.95	2905.12	2417.14
$r = 3$	-1325.3	-1088.1	2905.0	2436.52
Sample period 1981-2022				
$r = 1$	-2682.74	-1460.73	5495.7	**3057.87**
$r = 2$	-2652.45	-1435.51	5503.32	3075.63
$r = 3$	-2629.9	-1414.76	5526.43	3102.33

Table 2: **Model fit comparison: log-likelihood and BIC values for Gaussian and Student's t score-driven factor models with AR(1) innovations and with 1, 2, and 3 common factors.** $\log L$ denotes the maximized log-likelihood value.

calculated using the asymptotic variance expression given in Theorem 2, assuming correct model specification. From this analysis, we find that parameter ρ estimates of the AR(1) processes are large and significant, indicating that the innovations have a persistent and dynamic structure. Intuitively, a model with only one common factor and i.i.d. innovations is too simplistic to capture all the dynamics present in the time series. By incorporating a more complex structure for the innovations, we are able to better capture the individual-specific dynamics of each series. Moreover, the results show that the estimated degrees of freedom parameter ν appears to be small and lower when the full sample period is considered. This suggests that the improvement in the model's fit documented above is due to the robustness of the Student's t model, especially for the full sample period which includes the COVID-19 recession, where the gains are larger due to the V-shape of the recession.

The results in Table 3 also indicate that the estimates of the static parameters of the Student's t model are relatively stable across different sample periods, while the parameter estimates of the Gaussian model are more sensitive to the changes in the samples. In particular, for the 1981-2011 sample period, the parameter estimates of both models are similar. However, for the 1981-2022 sample period, the estimates of the Gaussian model, especially the factors' static parameters, α_1 and β_1, change dramatically, while those of the Student's t model remain largely unchanged. This confirms the sensitivity of the Gaussian

model to the presence of the spikes in the sample, which is an undesirable feature of the model when V-shaped recession(s) are present in the sample.

	Sample period 1981-2011				*Sample period 1981-2022*			
	$\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$		$t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$		$\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$		$t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$	
	$\hat{\boldsymbol{\theta}}_T$	s.e.	$\hat{\boldsymbol{\theta}}_T$	s.e.	$\hat{\boldsymbol{\theta}}_T$	s.e.	$\hat{\boldsymbol{\theta}}_T$	s.e.
α_1	0.448	0.034	0.399	0.044	0.603	0.039	0.389	0.039
β_1	0.900	0.021	0.870	0.025	0.409	0.042	0.849	0.025
ρ	0.869	0.009	0.890	0.008	0.856	0.008	0.883	0.005
ν			5.432	0.525			4.285	0.250
σ^2_{INDPRO}	0.038	0.002	0.026	0.002	0.090	0.007	0.026	0.002
σ^2_{UNRATE}	0.044	0.003	0.037	0.003	0.213	0.008	0.022	0.002
σ^2_{RETAIL}	0.191	0.010	0.123	0.009	0.282	0.013	0.081	0.005
$\sigma^2_{UMCSENTx}$	0.236	0.015	0.182	0.015	0.263	0.016	0.185	0.014
$\sigma^2_{S\&P500}$	0.094	0.006	0.061	0.005	0.105	0.006	0.063	0.005
$\sigma^2_{S\&P500vol}$	0.659	0.013	0.238	0.019	0.730	0.014	0.249	0.018
$\sigma^2_{BAATB10Y}$	0.065	0.004	0.033	0.003	0.064	0.003	0.032	0.002
$\sigma^2_{HOUSING}$	0.336	0.019	0.269	0.023	0.379	0.020	0.281	0.022
σ^2_{EBP}	0.148	0.008	0.077	0.006	0.162	0.008	0.078	0.005

Table 3: **Parameter estimates and standard errors (s.e.) of the Gaussian and Student's t score-driven factor models with AR(1) innovations and $r = 1$ common factor.**

Next, we examine the estimates, based on the full sample, of the factors and loadings of the one-factor models selected by the information criterion; see Figures 4 and 5. We note that the scale of the factors differs due to the differences in parameter ρ estimates. To facilitate further representation, at the bottom of Figure 4, we also demonstrate the standardized common factors and the filtered factors $\hat{f}_t^\star = \hat{f}_t - \hat{\rho}\hat{f}_{t-1}$.

The results of our analysis indicate that the dynamics of the factors resembles the dynamics of the US business cycle, with the troughs of the common factors corresponding to the US recessions. Overall, the factors of both Gaussian and Student's t models are similar, although differences become more pronounced during recessions. In particular, the factors $\hat{f}_t$ and $\hat{f}_t^\star$ in Figures 4 from the Student's t model heavily downweight the impact of the influential observations during the great recession and even more so during the COVID-19 recession period. We highlight that the model 'automatically' adjusts the 'weight' assigned to extreme observations.

The estimated loadings demonstrated in Figure 5 further reinforce the business cycle

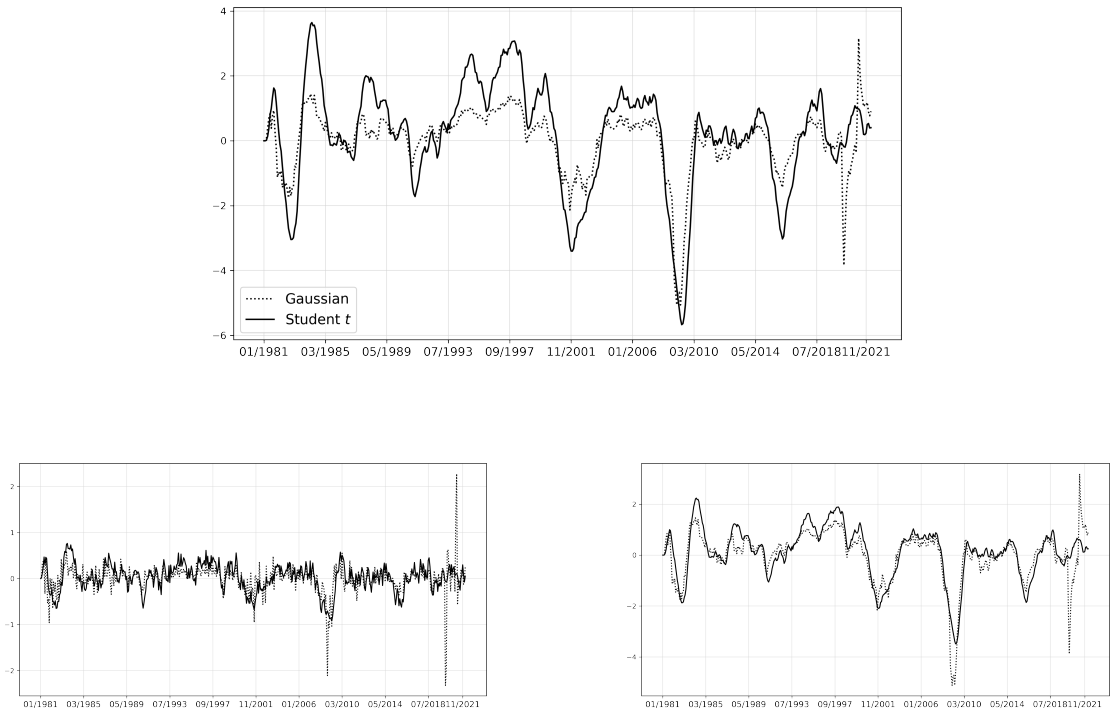


Figure 4: **Estimates of the common factors.** Results for the Gaussian and Student's t models with AR(1) innovations and with $r = 1$ factors. The top panel demonstrates filtered common factors $\hat{f}_t$. The bottom left panel presents $\hat{f}_t^{\star} = \hat{f}_t - \hat{\rho}\hat{f}_{t-1}$. The bottom right panel presents factors $\hat{f}_t$ standardized by the standard deviation.

interpretation for the common factor. Particularly, we find that the loadings of the industrial production, retail sales, S&P500 index, and consumer confidence index are positive, while the loadings of the unemployment rate, excess bond premium, and credit spread are negative. The loadings of the S&P500 volatility and housing starts index appear to be insignificant at the 5% confidence level for both models. Furthermore, we find that the confidence bounds for the loadings for the Student's t model are often narrower than those for the Gaussian model. We highlight that the order-invariant restriction enables us to conduct inference on all the loadings without having to specify the order of the series.

Finally, the results of the standard residual diagnostics are reported in Tables 4 and 5 (see also Figures SE.3, SE.4, and SE.5 in the Supplementary Appendix). We find that our models substantially reduce autocorrelation for most of the time series, indicating the models' ability to capture a considerable number of dynamic features. There are still some traces of the autocorrelation at lag 12 left, which can be attributed to the seasonality

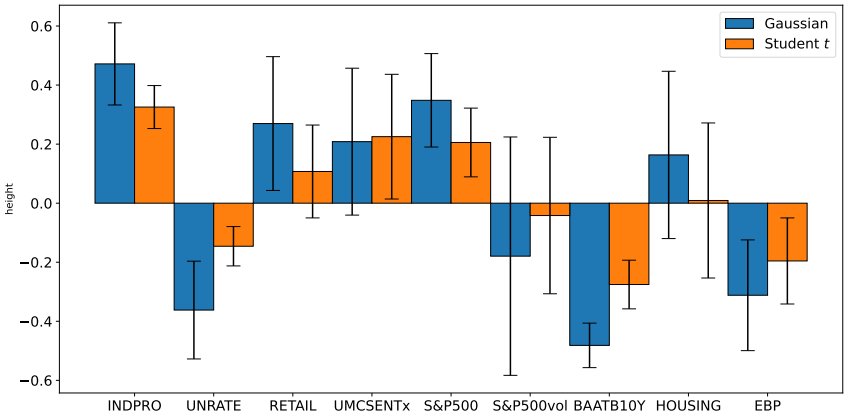


Figure 5: **Estimates of the loadings.** Results for the Gaussian (left bars) and Student's t (right bars) models with AR(1) innovations and $r = 1$ factors.

features unaccounted by the models. Further improvements can be achieved by adding more lags in the dynamic specification of the residuals. The results of the Kolmogorov-Smirnov test in Table 5 further supports the distributional assumptions, namely that for the majority of the series the Student's t distribution provides better fit to the data than the Gaussian distribution. This conclusion is also consistent with the results of the Pearson χ^2 goodness of fit test, as shown in Table SE.5.

	Raw	Residuals N	Residuals t_ν
INDPRO	1959.227	23.870	36.293
UNRATE	1321.451	10.400	12.143
RETAIL	1313.986	26.041	25.244
UMCSENTx	1087.597	14.323	19.291
S&P500	1814.605	32.209	53.957
S&P500vol	445.284	61.883	62.085
BAATB10Y	2328.579	19.778	61.831
HOUSING	1497.350	69.927	67.913
EBP	1848.516	19.765	16.924

Table 4: **The Ljung–Box test for residual serial correlation.** We compare the Ljung–Box test statistics of the standardised raw data to the test statistics of the residuals of the Gaussian and Student's t factor models. We consider the Ljung–Box test for residual autocorrelation up to order 8.

Our results indicate that the Student's t model performs better in the presence of observations from a V-shaped recession, offering more stable parameter estimates, and factor estimates that are less influenced by extreme observations. Unlike the Gaussian model, the Student's t model downweights extreme observations, producing results that

	Residuals N	Residuals t_ν
INDPRO	0.000	0.384
UNRATE	0.000	0.538
RETAIL	0.000	0.323
UMCSENTx	0.130	0.423
S&P500	0.080	0.692
S&P500vol	0.000	0.003
BAATB10Y	0.000	0.034
HOUSING	0.335	0.693
EBP	0.000	0.167

Table 5: **The Kolmogorov-Smirnov test.** We report the p-values of the Kolmogorov-Smirnov test of the equality of the distributions of the residuals with the reference distribution.

are less influenced by spikes. This makes the factors obtained from the model potentially more applicable as coincident economic indicators or as control variables. In the next section, we further examine the out-of-sample performance of the models.

5.3 Forecasting results

In this section, we compare the out-of-sample density forecasting performance during recessions of the Gaussian and Student's t score-driven factor models, both with $r = 1$ factor as suggested by the in-sample BIC. Results for the models with more factors are similar and not included here. Out-of-sample forecasts are generated using a rolling-window estimation with a rolling window size of 312 months (27 years). We produce one-month-ahead density forecasts during the periods of great recession (October 2007 to September 2010) and COVID-19 recession (March 2019 to February 2022). For each window, we re-estimate the models' parameters and produce one-step-ahead density forecasts. In total, for each sample, we have 36 months (3 years) for evaluating models' out-of-sample performance. The density forecasts are further used to compute mean logarithmic scoring rule (LSR) which is a commonly considered loss function in density forecasts evaluation literature.

The results in Table 6 reveal that during the great recession (U-shaped recession), the average LSR is higher for the Gaussian model than for the Student's t model for half of the time series, as indicated by the positive sign of the Diebold-Mariano (DM) test statistics, and vice versa. However, during the COVID-19 recession (V-shaped recession), the Student's t model consistently outperforms the Gaussian model. This suggests that

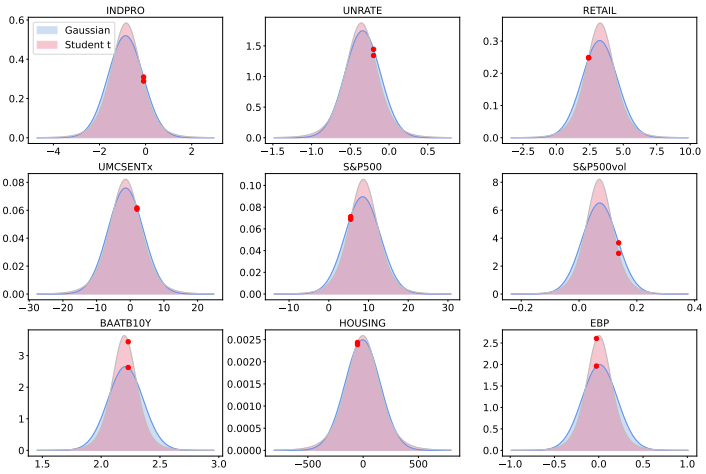
Great recession: October 2007-September 2010					
	INDPRO	UNRATE	RETAIL	UMCSENTx	S&P500
DM	-0.041	1.404	1.948	-0.988	-0.912
p-value	0.967	0.160	0.060	0.330	0.368
	S&P500vol	BAATB10Y	HOUSING	EBP	**Total**
DM	-0.378	-1.300	0.526	-0.878	-1.027
p-value	0.707	0.202	0.602	0.386	0.312
COVID-19 recession: March 2019-February 2022					
	INDPRO	UNRATE	RETAIL	UMCSENTx	S&P500
DM	-1.488	-1.159	-1.729	-1.570	-1.845
p-value	0.137	0.254	0.084	0.116	0.074
	S&P500vol	BAATB10Y	HOUSING	EBP	**Total**
DM	-1.232	-1.748	-0.726	-1.586	-1.573
p-value	0.226	0.089	0.472	0.113	0.125

Table 6: **Diebold-Mariano test.** The test statistics is computed based on the out-of-sample logarithmic scoring rule. A negative value of the statistics corresponds to a lower average logarithmic scoring rule of the Gaussian model. The DM test statistics is computed based on heteroscedasticity robust standard errors.

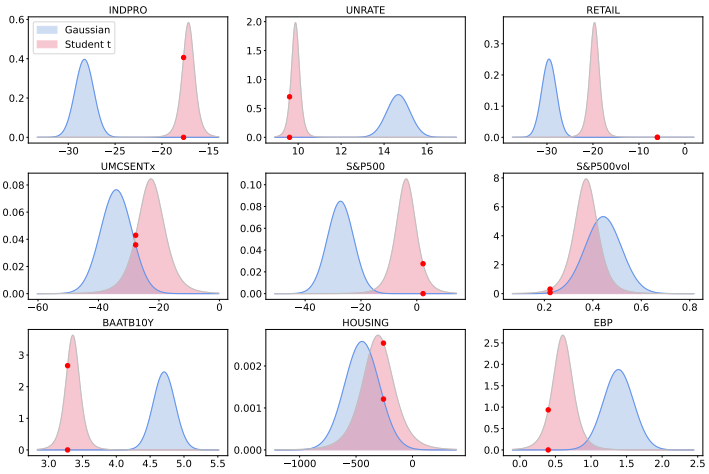
while both models perform similarly during the U-shaped recession, the Gaussian model performs worse during the V-shaped recession. However, the difference between the two models is mostly not statistically significant at a 10% significance level for both out-of-sample periods. Intuitively, both models have similar out-of-sample performance, with the main differences occurring during the short periods of extreme observations. Given that there are only a few extreme observations and a relatively large standard deviation of the loss differential, it is not surprising that the difference is not statistically significant. Nevertheless, for policymakers, it is of high order importance to produce reliable forecasts during turbulent periods like the COVID-19 pandemic, making Gaussian model, which is sensitive to extreme observations, less favorable.

Next, we examine the predictive densities of individual time series during the COVID-19 recession. Our findings reveal that prior to the recession, the model predictions are comparable and well aligned with the actual observations, as shown in Figure 6a. However, after the recession, the performance of the Student's t model is substantially better, as illustrated in Figure 6b. The reason behind this superior performance is the robustness of

the Student's t model to extreme observations, which are a common phenomenon in all time series during the COVID-19 recession.



(a) May 2019.



(b) May 2020

Figure 6: **Density predictions for individual time series.** The densities are constructed based on the parameter estimates obtained using a sample of 312 observations up to April 2019 and April 2020, respectively. The red dot represents the actual observation.

This sensitivity of the Gaussian model to extreme observations during the COVID-19 recession is a major shortcoming, as it can lead to unreliable forecasts. Our analysis of the sensitivity of the loadings estimates used for the out-of-sample forecasts further supports this statement. As shown in Figure 7, the loadings estimates of the Gaussian model are highly impacted by new observations, whereas the Student's t model displays only minor changes. This highlights the advantages of utilizing robust models, such as the Student's t, when dealing with V-shaped recessions like COVID-19 recession.

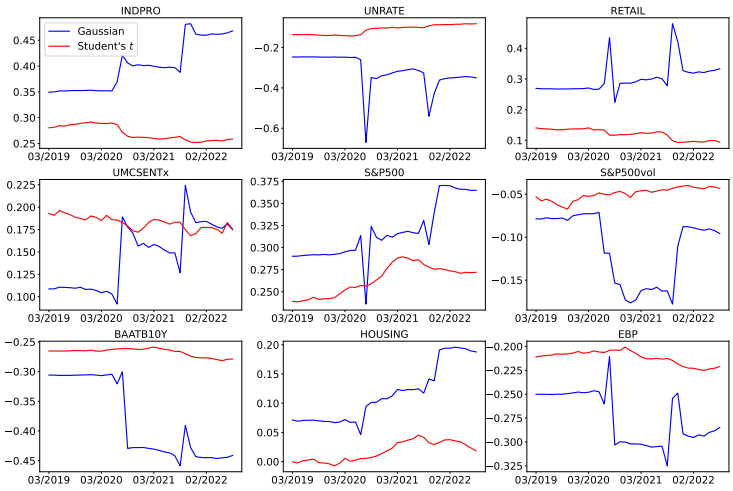


Figure 7: **Rolling-window estimates of loadings for individual time series.** The results are based on the rolling window estimation with a rolling window size equal to 312 months.

6 Conclusion

We have introduced an order-invariant dynamic factor model with elliptically distributed innovations where the dynamics of the factors is driven by the score of the predictive likelihood. The update based on the score allows the dynamics of the factors to be potentially robust to extreme values and outliers. We discuss the model identification and propose a solution to the rotational indeterminacy problem using an order-invariant identification constraint. We also establish theoretical properties of the model and its estimator. A numerical estimation of the model under the order-invariant identification condition is proposed by using optimization methods on the Stiefel manifolds. In an extensive simulation study, we confirm the good finite sample properties of the estimator. The empirical application for constructing coincident economic indicators demonstrates the importance of the robust updating equations in the presence of the COVID-19 recession period in the sample.

Acknowledgments

The author is thankful to Francisco Blasques, Janneke van Brummelen, Paolo Gorgi, Christian Gourieroux, Christian Francq, Siem Jan Koopman, Jean-Michel Zakoian, and conference and seminar participants at the 2022 NBER-NSF Time Series Conference at Boston University, the 2nd International Econometrics PhD Conference at Erasmus University Rot-

terdam, the 33rd EC2 Conference in Paris, the Financial Econometrics Conference at the Tolouse School of Economics, the 2023 Quantitative Finance and Financial Econometrics (QFFE) Conference at Aix-Marseille University, the 2023 Annual SoFiE (Pre)-Conference in Seoul, the 2023 International Association for Applied Econometrics (IAAE) Annual Conference at the BI Norwegian Business School, the Vrije Universiteit Amsterdam, and the Center for Research in Economics and Statistics for useful comments and suggestions. Financial support by the SoFiE and IAAE travel grants is highly acknowledged.

References

Aitchison, J., & Silvey, S. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, *29*(3), 813–828.

Anderson, T. W., Rubin, H., et al. (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (Vol. 5, pp. 111–150).

Artemova, M., Blasques, F., van Brummelen, J., & Koopman, S. J. (2022a). Score-driven models: Methodology and theory. In *Oxford research encyclopedia of economics and finance.*

Artemova, M., Blasques, F., van Brummelen, J., & Koopman, S. J. (2022b). Score-driven models: Methods and applications. In *Oxford research encyclopedia of economics and finance.*

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135–171.

Bai, J., & Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, *40*(1), 436–465.

Barigozzi, M., He, Y., Li, L., & Trapani, L. (2023). Robust estimation of large factor models for tensor-valued time series. *arXiv preprint arXiv:2303.18163*.

Blasques, F., Francq, C., & Laurent, S. (2022). *Autoregressive conditional betas* (Tech. Rep.). Working paper.

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

Blasques, F., Gorgi, P., Koopman, S. J., & 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., & Koopman, S. J. (2022). *Maximum likelihood estimation for non-stationary location models with mixture of normal distributions* (Tech. Rep.). Tinbergen Institute Discussion Paper.

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

Boumal, N. (2023). *An introduction to optimization on smooth manifolds*. Cambridge University Press.

Brave, S. A., & Kelley, D. (2017). Introducing the Chicago FED's new adjusted national financial conditions index. *Chicago Fed Letter*, *386*.

Chan, J., Leon-Gonzalez, R., & Strachan, R. W. (2018). Invariant inference and efficient computation in the static factor model. *Journal of the American Statistical Association*, *113*(522), 819–828.

Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., ... Lauritzen, S. L. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 93–115.

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

Creal, D., Schwaab, B., Koopman, S. J., & 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.

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

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics*, *94*(4), 1014–1024.

D'Innocenzo, E., Luati, A., & Mazzocchi, M. (2023). A robust score-driven filter for multivariate time series. *Econometric Reviews*, 1–30.

Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, *20*(2), 303–353.

Engle, R., & Watson, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, *76*(376), 774–781.

Fan, J., Wang, K., Zhong, Y., & Zhu, Z. (2021). Robust high dimensional factor models with applications to statistical machine learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *36*(2), 303.

Fang, K.-T., Kotz, S., & Ng, K. W. (2018). *Symmetric multivariate and related distributions*. Chapman and Hall/CRC.

Favara, G., Gilchrist, S., Lewis, K. F., & Zakrajšek, E. (2016). Updating the recession risk and the excess bond premium.

Gilchrist, S., & Zakrajšek, E. (2012). Credit spreads and business cycle fluctuations. *American Economic Review*, *102*(4), 1692–1720.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

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

He, Y., Li, L., Liu, D., & Zhou, W.-X. (2023). Huber principal component analysis for large-dimensional factor models. *arXiv preprint arXiv:2303.02817*.

Krengel, U. (1985). *Ergodic theorems* (Vol. 6). Walter de Gruyter.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as statistical method* (Tech. Rep.).

Lewis, D. J., Mertens, K., Stock, J. H., & Trivedi, M. (2022). Measuring real activity using a weekly economic index. *Journal of Applied Econometrics*, *37*(4), 667–687.

Liu, C., & Boumal, N. (2020). Simple algorithms for optimization on Riemannian manifolds with constraints. *Applied Mathematics & Optimization*, *82*(3), 949–981.

Loria, F., Matthes, C., & Zhang, D. (2022). Assessing macroeconomic tail risk. *Available at SSRN 4002665*.

Ng, S. (2021). *Modeling macroeconomic variations after COVID-19* (Tech. Rep.). National Bureau of Economic Research.

Pötscher, B. M., & Prucha, I. (1997). *Dynamic nonlinear econometric models: Asymptotic theory*. Springer Science & Business Media.

Quah, D., & Sargent, T. J. (1993). A dynamic index model for large cross sections. In *Business cycles, indicators, and forecasting* (pp. 285–310). University of Chicago Press.

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

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society*, 577–591.

Silvey, S. D. (1959). The Lagrangian multiplier test. *The Annals of Mathematical Statistics*, *30*(2), 389–407.

Silvey, S. D. (1975). *Statistical inference*. Routledge.

Stiefel, E. (1935). *Richtungsfelder und fernparallelismus in n-dimensionalen mannigfaltigkeiten* (Unpublished doctoral dissertation). ETH Zurich.

Stock, J. H., & Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, *4*, 351–394.

Stock, J. H., & Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, *97*(460), 1167–1179.

Stock, J. H., & Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162.

Straumann, D., & 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.

Townsend, J., Koep, N., & Weichwald, S. (2016). Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *arXiv preprint arXiv:1603.03236*.

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

White, H. (1996). *Estimation, inference and specification analysis* (No. 22). Cambridge University Press.

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

Appendix

A Proofs of the main results

Proof of Lemma 1. Since $\boldsymbol{\varepsilon}_t \sim \mathcal{E}_N(\mathbf{0}, \boldsymbol{\Sigma}, g)$ it implies that $\boldsymbol{y}_t|\boldsymbol{f}_t, \mathcal{F}_{t-1} \sim \mathcal{E}_N(\boldsymbol{\Lambda}\boldsymbol{f}_t, \boldsymbol{\Sigma}, g)$. Hence,

$$p_{\boldsymbol{y}}(\boldsymbol{y}_t|\mathcal{F}_{t-1}) = |\boldsymbol{\Sigma}|^{-1/2} g\left((\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t)\right),$$
$$\log p_{\boldsymbol{y}}(\boldsymbol{y}_t|\mathcal{F}_{t-1}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| + \log g\left((\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t)\right).$$

The result for the score follows immediately by taking the derivative of the log-likelihood with respect to $\boldsymbol{f}_t$. Let us consider the Fisher information matrix

$$\begin{aligned}\boldsymbol{\mathcal{I}}_{t|t-1} &= \mathbb{E}_{t-1}[\nabla_t \nabla_t^\top] = 4\mathbb{E}_{t-1}\left[\left(\frac{g'(\|\tilde{\boldsymbol{y}}_t\|^2)}{g(\|\tilde{\boldsymbol{y}}_t\|^2)}\right)^2 \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{y}}_t\tilde{\boldsymbol{y}}_t^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}\right] \\ &= 4\mathbb{E}_{t-1}\left[\|\tilde{\boldsymbol{y}}_t\|^2\left(\frac{g'(\|\tilde{\boldsymbol{y}}_t\|^2)}{g(\|\tilde{\boldsymbol{y}}_t\|^2)}\right)^2 \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1/2}\frac{\tilde{\boldsymbol{y}}_t}{\|\tilde{\boldsymbol{y}}_t\|}\frac{\tilde{\boldsymbol{y}}_t^\top}{\|\tilde{\boldsymbol{y}}_t\|}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}\right],\end{aligned}$$

where $\tilde{\boldsymbol{y}}_t$ denotes a standardized vector of observations (with mean zero and identity scale matrix), hence it is spherically distributed. Therefore, by Theorem 2.3 in Fang et al. (2018) $\|\tilde{\boldsymbol{y}}_t\|$ and $\tilde{\boldsymbol{y}}_t/\|\tilde{\boldsymbol{y}}_t\|$ are independent. This implies that the expression above can be rewritten as follows

$$\boldsymbol{\mathcal{I}}_{t|t-1} = 4\mathbb{E}_{t-1}\left[\|\tilde{\boldsymbol{y}}_t\|^2\left(\frac{g'(\|\tilde{\boldsymbol{y}}_t\|^2)}{g(\|\tilde{\boldsymbol{y}}_t\|^2)}\right)^2\right]\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1/2}\mathbb{E}_{t-1}\left[\frac{\tilde{\boldsymbol{y}}_t}{\|\tilde{\boldsymbol{y}}_t\|}\frac{\tilde{\boldsymbol{y}}_t^\top}{\|\tilde{\boldsymbol{y}}_t\|}\right]\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}.$$

Exploiting the fact that $\boldsymbol{u}_t := \frac{\tilde{\boldsymbol{y}}_t}{\|\tilde{\boldsymbol{y}}_t\|}$ is the uniform base of the spherical distribution, by Theorem 2.7 in Fang et al. (2018) we have $\mathbb{E}\left[\boldsymbol{u}_t\boldsymbol{u}_t^\top\right] = \frac{1}{N}\boldsymbol{I}_N$ and this implies that

$$\boldsymbol{\mathcal{I}}_{t|t-1} = 4\mathbb{E}_{t-1}\left[\|\tilde{\boldsymbol{y}}_t\|^2\left(\frac{g'(\|\tilde{\boldsymbol{y}}_t\|^2)}{g(\|\tilde{\boldsymbol{y}}_t\|^2)}\right)^2\right]\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}.$$

This completes the proof. ■

Proof of Lemma 2. We first show that there exists a stationary and ergodic causal solution

$\{\boldsymbol{f}_t\}_{t\in\mathbb{Z}}$ to (6). By iterating backwards equation (6), we obtain

$$\boldsymbol{f}_{t+1} = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{f}_t = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\left(\boldsymbol{A}\boldsymbol{s}_{t-1} + \boldsymbol{B}\boldsymbol{f}_{t-1}\right) = \sum_{j=0}^{t-1} \boldsymbol{B}^j \boldsymbol{A}\boldsymbol{s}_{t-j} + \boldsymbol{B}^t \boldsymbol{f}_1, \tag{A.1}$$

where under correct model specification $\boldsymbol{s}_t = \frac{1}{W(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|,g)} \left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1} \frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t$.

The last term in expression (A.1) goes to zero as $t \to \infty$ since $\|\boldsymbol{B}\|^t \to 0$ and $\boldsymbol{f}_1 \in \mathbb{R}^r$. Then, if the limit sequence exists, it is of the form

$$\boldsymbol{f}_{t+1} = \sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A}\boldsymbol{s}_{t-j}. \tag{A.2}$$

We further establish the stochastic properties of the score sequence $\boldsymbol{s}_t$. First, $\boldsymbol{s}_t$ is continuous, hence, measurable function of strictly stationary and ergodic (SE) sequence $\boldsymbol{\varepsilon}_t$. Therefore, by Proposition 4.3 in Krengel (1985) sequence $\{\boldsymbol{s}_t\}_{t\in\mathbb{Z}}$ is also SE. Now we show that the score has a logarithmic moment. By norm submultiplicativity and positive definiteness of matrix $\boldsymbol{P} = \left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}$, we have

$$\begin{aligned}\mathbb{E}\log^+\|\boldsymbol{s}_t\| &\le \log^+\left\|\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{\sqrt{N}}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1/2}\right\| + \mathbb{E}\log^+\left\|\sqrt{N}\tilde{\boldsymbol{s}}_t\right\| \\ &= \frac{1}{2}\underbrace{\log^+\|\boldsymbol{P}\|}_{<\infty,\text{ by Ass. 2}} + \underbrace{\mathbb{E}\log^+\left\|\sqrt{N}\tilde{\boldsymbol{s}}_t\right\|}_{<\infty,\text{ by Ass. 4}} < \infty,\end{aligned}$$

where $\tilde{\boldsymbol{s}}_t := \frac{1}{W(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t\|,g)}\frac{1}{N}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t$.

We further prove that the series in (A.2) converges if and only if $\|\boldsymbol{B}\| < 1$.

Suppose that $\|\boldsymbol{B}\| < 1$. By monotone convergence theorem, norm subadditivity and submultiplicativity, we have

$$\mathbb{E}\sum_{j=0}^{\infty}\left\|\boldsymbol{B}^j\boldsymbol{A}\boldsymbol{s}_{t-j}\right\| \le \sum_{j=0}^{\infty}\|\boldsymbol{B}\|^j\mathbb{E}\left\|\boldsymbol{A}\boldsymbol{s}_{t-j}\right\| \le \underbrace{\|\boldsymbol{A}\|}_{<\infty,\ \Theta\in\mathbb{R}^q}\mathbb{E}\|\boldsymbol{s}_t\|\sum_{j=0}^{\infty}\|\boldsymbol{B}\|^j < \infty, \tag{A.3}$$

where the last claim in (A.3) follows by Lemma 2.1 in Straumann & Mikosch (2006) since $\|\boldsymbol{B}\|^t \to 0$ exponentially as $t \to \infty$ and the sequence $\{\|\boldsymbol{s}_t\|\}_{t\in\mathbb{Z}}$ is SE with $\mathbb{E}\log^+\|\boldsymbol{s}_t\| < \infty$. Therefore, the series in (A.2) converges almost surely and the limit sequence $\{\boldsymbol{f}_t\}_{t\in\mathbb{Z}}$ exists.

If $\|\boldsymbol{B}\| > 1$ then the series in (A.2) diverges at least for one of the vector components which implies that there is a.s. no finite solution to (6) and, consequently, to (1). For $\|\boldsymbol{B}\| = 1$, the series in (A.2) may diverge at least for one of the vector components.

Furthermore, by Proposition 4.3 in Krengel (1985) the limit sequence $\{\boldsymbol{f}_t\}_{t\in\mathbb{Z}}$ is strictly

stationary and ergodic since it is a measurable function of $\{\boldsymbol{s}_t\}_{t\in\mathbb{Z}}$ which is SE. Proposition 4.3 in Krengel (1985) also ensures that $\{\boldsymbol{y}_t\}_{t\in\mathbb{Z}}$ is SE since it is a continuous, hence measurable, function of $\boldsymbol{f}_t$ and $\boldsymbol{\varepsilon}_t$, which are jointly SE.

We now establish the uniqueness of the stationary solutions. Assume that there exists another SE solution $\{\tilde{\boldsymbol{f}}_t\}_{t\in\mathbb{Z}}$ satisfying equation (6), then for $t = t^*$ such that $\boldsymbol{f}_{t^*} \neq \tilde{\boldsymbol{f}}_{t^*}$ and for all $i > 0$, we have

$$0 < \|\boldsymbol{f}_{t^*} - \tilde{\boldsymbol{f}}_{t^*}\| = \|\boldsymbol{B}\|^i \left\|\boldsymbol{f}_{t^*-i} - \tilde{\boldsymbol{f}}_{t^*-i}\right\|.$$

By the conditions of the lemma $\|\boldsymbol{B}\|^i \to 0$ exponentially as $i \to \infty$. Moreover, $\|\boldsymbol{f}_{t^*-i} - \tilde{\boldsymbol{f}}_{t^*-i}\| = O_P(1)$ as the sequences are strictly stationary, hence $\mathbb{P}(\boldsymbol{f}_t = \tilde{\boldsymbol{f}}_t) = 1$ and the uniqueness follows. ■

Proof of Lemma 3. First, we consider the case $k \geq 1$. For this case, we prove that $\mathbb{E}\|\boldsymbol{f}_t\|^k < \infty$ by showing that $\|\boldsymbol{f}_t\|_k \equiv (\mathbb{E}\|\boldsymbol{f}_t\|^k)^{1/k} < \infty$ which clearly implies the desired result.

Given (A.2) and by Minkowski's inequality, we have

$$\|\boldsymbol{f}_t\|_k \leq \sum_{j=0}^{\infty} \|\boldsymbol{B}\|^j \|\boldsymbol{A}\| \|\boldsymbol{s}_{t-j}\|_k \leq \underbrace{\|\boldsymbol{A}\|}_{<\infty, \Theta\subseteq\mathbb{R}^q} \Big(\sum_{j=0}^{\infty} \|\boldsymbol{B}\|^j\Big) \times \underbrace{(\|\boldsymbol{P}\|)^{1/2}}_{<\infty, \text{ by Ass. 2}} \times \underbrace{\left\|\sqrt{N}\tilde{\boldsymbol{s}}_t\right\|_k}_{<\infty, \text{ by Ass. 4.a}},$$

with $\tilde{\boldsymbol{s}}_t$ as defined in the proof of Lemma 2 and positive definite matrix $\boldsymbol{P}$ as defined in Assumption 2. Since, $\|\boldsymbol{B}\| < 1$, we conclude that $(\mathbb{E}\|\boldsymbol{f}_t\|^k)^{1/k} < \infty$ and the result follows.

In the case of $0 < k < 1$, the result immediately follows by application of the Loève's c_r inequality directly to $\mathbb{E}\|\boldsymbol{f}_{t+1}\|^k$.

The proof for $\boldsymbol{y}_t$, i.e. $\mathbb{E}\|\boldsymbol{y}_t\|^k < \infty$, follows by the established above result for the factors, $\mathbb{E}\|\boldsymbol{f}_t\|^k < \infty$, and the Loève's c_r inequality, that is,

$$\mathbb{E}\|\boldsymbol{y}_t\|^k = \mathbb{E}\|\boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t\|^k \leq c_r\|\boldsymbol{\Lambda}\|\mathbb{E}\|\boldsymbol{f}_t\|^k + c_r\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^k < \infty,$$

since $c_r \in \mathbb{R}$, $\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^k < \infty$ by Assumption 3.a and $\Theta \subseteq \mathbb{R}^q$.

■

Proof of Lemma 4. Under the set of conditions (A), the proof is essentially a multivariate extension of Blasques, van Brummelen, Koopman, & Lucas (2022, Proposition 3.1) and, hence, is omitted.

Under the set of conditions (B), we notice that, given that $\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{B}\| < 1$, for large

enough m, we have

$$\sup_t \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}_{t+1}(\boldsymbol{\theta})\| \leq 1 + \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{A}\| \sum_{j=0}^{m} \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{B}^j\| \sup_t \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{s}_{t-j}\|$$
$$\leq 1 + \bar{d} \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{A}\| \sum_{j=0}^{m} \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{B}^j\| \leq 1 + \bar{d} \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{A}\| \frac{1}{1 - \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{B}\|} < \infty,$$

which completes the proof. ■

Proof of Lemma 5. Lemma 3 ensures the existence of the stationary solution with $k = 2$ bounded moments. Then, from (A.2) we obtain

$$\mathbb{C}ov(\boldsymbol{f}_{t+1}) = \mathbb{C}ov\left(\sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A} \boldsymbol{s}_{t-j}\right) = \sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A} \mathbb{C}ov\left(\boldsymbol{s}_{t-j}\right) \boldsymbol{A}\boldsymbol{B}^j = \sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A} \mathbb{C}ov\left(\boldsymbol{s}_{t}\right) \boldsymbol{A}\boldsymbol{B}^j,$$

where we exploited the fact that under correct model specification the score sequence is stationary and white noise.

To simplify further notation, we introduce $\tilde{\boldsymbol{\varepsilon}}_t := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}_t$. Given expression (5) for the score, the covariance matrix for $\boldsymbol{s}_t$ is as follows

$$\begin{aligned}
\mathbb{C}ov\left(\boldsymbol{s}_t\right) &= \left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}\right)^{-1} \frac{1}{N^2} \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1/2} \\
&\times \mathbb{C}ov\left(\frac{1}{(W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g))^2} \tilde{\boldsymbol{\varepsilon}}_t\right) \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Lambda} \left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}\right)^{-1} \\
&= \left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}\right)^{-1} \frac{1}{N^2} \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1/2} \mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{(W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g))^2}\right] \\
&\times \mathbb{E}\left[\frac{\tilde{\boldsymbol{\varepsilon}}_t}{\|\tilde{\boldsymbol{\varepsilon}}_t\|} \frac{\tilde{\boldsymbol{\varepsilon}}_t^\top}{\|\tilde{\boldsymbol{\varepsilon}}_t\|}\right] \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Lambda} \left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}\right)^{-1} \\
&= \mathbb{E}\left[\left(\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|}{NW(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right)^2\right] \left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}\right)^{-1},
\end{aligned}$$

where the second and third equalities follow by application of Theorems 2.3 and 2.7 in Fang et al. (2018) and the fact that $\mathbb{E}[\tilde{\boldsymbol{\varepsilon}}_t] = 0$.
Hence, under restriction (IC4) on the loadings, we obtain

$$\mathbb{C}ov\left(\boldsymbol{s}_t\right) = \mathbb{E}\left[\left(\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|}{NW(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right)^2\right] \boldsymbol{I}_r,$$
$$\mathbb{C}ov(\boldsymbol{f}_{t+1}) = \sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A} \mathbb{C}ov\left(\boldsymbol{s}_t\right) \boldsymbol{A}\boldsymbol{B}^j = \left(\boldsymbol{I}_r - \boldsymbol{B}^2\right)^{-1} \mathbb{C}ov\left(\boldsymbol{s}_t\right) \boldsymbol{A}^2,$$

where we exploited that matrices $\mathbb{C}ov(\boldsymbol{s}_t)$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal and $\|\boldsymbol{B}\| < 1$.
Next, we turn to the autocovariance structure of $\boldsymbol{f}_t$

$$\begin{aligned}
\mathbb{C}ov(\boldsymbol{f}_{t+h+1}, \boldsymbol{f}_{t+1}) &= \mathbb{C}ov\left(\sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A}\boldsymbol{s}_{t+h-j}, \sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A}\boldsymbol{s}_{t-j}\right) \\
&= \mathbb{C}ov\left(\sum_{j=-h}^{\infty} \boldsymbol{B}^{j+h} \boldsymbol{A}\boldsymbol{s}_{t-j}, \sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A}\boldsymbol{s}_{t-j}\right) \\
&= \boldsymbol{B}^h \mathbb{C}ov\left(\sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A}\boldsymbol{s}_{t-j}, \sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A}\boldsymbol{s}_{t-j}\right) = \boldsymbol{B}^h \mathbb{C}ov(\boldsymbol{f}_t).
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{C}ov(\boldsymbol{s}_t, \boldsymbol{\varepsilon}_t) &= \frac{1}{N} \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1/2} \mathbb{C}ov\left(\frac{1}{W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)} \tilde{\boldsymbol{\varepsilon}}_t, \boldsymbol{\Sigma}^{1/2} \tilde{\boldsymbol{\varepsilon}}_t\right) \\
&= \frac{1}{N} \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1/2} \mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)} \frac{\tilde{\boldsymbol{\varepsilon}}_t}{\|\tilde{\boldsymbol{\varepsilon}}_t\|} \frac{\tilde{\boldsymbol{\varepsilon}}_t^\top}{\|\tilde{\boldsymbol{\varepsilon}}_t\|}\right] \boldsymbol{\Sigma}^{1/2} \\
&= \frac{1}{N} \boldsymbol{\Lambda}^\top \mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right] \mathbb{E}\left[\frac{\tilde{\boldsymbol{\varepsilon}}_t}{\|\tilde{\boldsymbol{\varepsilon}}_t\|} \frac{\tilde{\boldsymbol{\varepsilon}}_t^\top}{\|\tilde{\boldsymbol{\varepsilon}}_t\|}\right] = \frac{1}{N^2} \boldsymbol{\Lambda}^\top \mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right], \\
\mathbb{C}ov(\boldsymbol{f}_{t+h+1}, \boldsymbol{\varepsilon}_t) &= \mathbb{C}ov\left(\sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{A}\boldsymbol{s}_{t+h-j}, \boldsymbol{\varepsilon}_t\right) = \boldsymbol{A} \sum_{j=0}^{\infty} \boldsymbol{B}^j \mathbb{C}ov(\boldsymbol{s}_{t+h-j}, \boldsymbol{\varepsilon}_t) \\
&= \boldsymbol{A}\boldsymbol{B}^h \mathbb{C}ov(\boldsymbol{s}_t, \boldsymbol{\varepsilon}_t) = \frac{1}{N^2} \boldsymbol{A}\boldsymbol{B}^h \boldsymbol{\Lambda}^\top \mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right],
\end{aligned}$$

where we applied Theorems 2.3 and 2.7 in Fang et al. (2018) and the fact that $\mathbb{E}[\tilde{\boldsymbol{\varepsilon}}_t] = 0$. Hence, the proof follows. ■

Proof of Proposition 2. Observation equation (11) implies the following covariance structure of the data

$$\mathbb{C}ov(\boldsymbol{y}_t) = \boldsymbol{\Lambda}\boldsymbol{K}\boldsymbol{\Lambda}^\top + \boldsymbol{\Sigma} := \boldsymbol{L}\boldsymbol{L}^\top + \tilde{\boldsymbol{\Sigma}}, \tag{A.4}$$

where, for example, for the Gaussian model $\tilde{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma}$ and for the Student's t model $\tilde{\boldsymbol{\Sigma}} \equiv \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$.

Intuitively, given that $\boldsymbol{\Sigma}$ is diagonal, equation (A.4) implies that all the co-movements between the series are explained by the common components. Identification of $\boldsymbol{L}$ and $\tilde{\boldsymbol{\Sigma}}$ follows immediately from Theorem 5.1 in Anderson et al. (1956) given conditions B and E. Moreover, (IC4) (conditions C–D) guarantees that matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{K}$ are uniquely identified (Lawley & Maxwell, 1971; Bai & Li, 2012). We further note that it suffices to show the identification for $(\boldsymbol{\nu}, \boldsymbol{\Sigma}) = (\boldsymbol{\nu}_0, \boldsymbol{\Sigma}_0)$.

For the remaining parameters, we proceed by contradiction. Assume that there exists

$\tilde{\boldsymbol{\psi}} := (\text{diag}(\tilde{\boldsymbol{A}})^\top, \text{diag}(\tilde{\boldsymbol{B}})^\top)^\top$ such that $\tilde{\boldsymbol{\psi}} \neq \boldsymbol{\psi}$ are observationally equivalent.

We proceed by exploiting the autocovariance structure of $\boldsymbol{y}_t$

$$\begin{aligned}\mathbb{C}ov(\boldsymbol{y}_{t+h}, \boldsymbol{y}_t) = \boldsymbol{\Lambda}\mathbb{C}ov(\boldsymbol{f}_{t+h}, \boldsymbol{f}_t)\boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda}\mathbb{C}ov(\boldsymbol{f}_{t+h}, \boldsymbol{\varepsilon}_t) \\ + \mathbb{C}ov(\boldsymbol{\varepsilon}_{t+h}, \boldsymbol{f}_t)\boldsymbol{\Lambda}^\top + \mathbb{C}ov(\boldsymbol{\varepsilon}_{t+h}, \boldsymbol{\varepsilon}_t), \quad \forall h \geq 1.\end{aligned}$$

From (9) and (10), $\forall h \geq 1$ we have:

$$\begin{aligned}\mathbb{C}ov(\boldsymbol{f}_{t+h}, \boldsymbol{f}_t) &= \boldsymbol{B}^h\boldsymbol{K}, \\ \mathbb{C}ov(\boldsymbol{f}_{t+h}, \boldsymbol{\varepsilon}_t) &= \boldsymbol{A}\boldsymbol{B}^{h-1}\boldsymbol{\Lambda}^\top\mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{N^2W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right].\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{C}ov(\boldsymbol{y}_{t+h}, \boldsymbol{y}_t) &= \boldsymbol{\Lambda}\boldsymbol{B}^h\boldsymbol{K}\boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda}\boldsymbol{B}^{h-1}\boldsymbol{A}\boldsymbol{\Lambda}^\top\mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{N^2W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right] \\ &= \boldsymbol{\Lambda}\boldsymbol{B}^{h-1}\left(\boldsymbol{B}\boldsymbol{K} + \boldsymbol{A}\mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{N^2W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right]\right)\boldsymbol{\Lambda}^\top, \quad \forall h \geq 1.\end{aligned}$$

To shorten further notation we introduce $d := \mathbb{E}\left[\frac{\|\tilde{\boldsymbol{\varepsilon}}_t\|^2}{N^2W(\|\tilde{\boldsymbol{\varepsilon}}_t\|, g)}\right]$. Since there exists an observationally equivalent $\tilde{\boldsymbol{\psi}} \neq \boldsymbol{\psi}$, we have that:

$$\boldsymbol{B}^{h-1}\left(\boldsymbol{B}\boldsymbol{K} + d\boldsymbol{A}\right) = \tilde{\boldsymbol{B}}^{h-1}\left(\tilde{\boldsymbol{B}}\boldsymbol{K} + \tilde{d}\tilde{\boldsymbol{A}}\right), \quad \forall h \geq 1.$$

Hence, for $h = 1$ and $h = 2$, we have

$$\begin{aligned}\left(\boldsymbol{B}\boldsymbol{K} + d\boldsymbol{A}\right) &= \left(\tilde{\boldsymbol{B}}\boldsymbol{K} + \tilde{d}\tilde{\boldsymbol{A}}\right), \\ \boldsymbol{B}\left(\boldsymbol{B}\boldsymbol{K} + d\boldsymbol{A}\right) &= \tilde{\boldsymbol{B}}\left(\tilde{\boldsymbol{B}}\boldsymbol{K} + \tilde{d}\tilde{\boldsymbol{A}}\right).\end{aligned}$$

From this, we can conclude that $\boldsymbol{B} = \tilde{\boldsymbol{B}}$ and $\boldsymbol{A} = \tilde{\boldsymbol{A}}$.

■

Proof of Theorem 1. We note that, although we are dealing with the constrained estimator, the standard consistency proof, for example, White (1996, Theorem 3.5) or Pötscher & Prucha (1997, Lemma 3.1), still applies since the conditions of the theorems do not require $\boldsymbol{\theta}_0$ to belong to the interior of the parameter space $\tilde{\Theta}$. Particularly, the strong consistency of the ML estimator follows from (i) the uniform a.s. convergence of the criterion function; (ii) the regularity of the level sets of the limit criterion function (Pötscher & Prucha, 1997, Definition 4.1).

To prove consistency, we use a similar approach as in the proof of Blasques, van Brum-

melen, Koopman, & Lucas (2022, Theorem 4.6) or Blasques et al. (2023, Theorem 2). We start with showing the uniform convergence. By the triangle inequality

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\tilde{\Theta}}\left|\hat{\mathcal{L}}_T(\boldsymbol{\theta})-\mathcal{L}_\infty(\boldsymbol{\theta})\right| &\leq \sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{\mathcal{L}}_T(\boldsymbol{\theta})-\mathcal{L}_\infty(\boldsymbol{\theta})\right| \\ &\leq \sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{\mathcal{L}}_T(\boldsymbol{\theta})-\mathcal{L}_T(\boldsymbol{\theta})\right| + \sup_{\boldsymbol{\theta}\in\Theta}\left|\mathcal{L}_T(\boldsymbol{\theta})-\mathcal{L}_\infty(\boldsymbol{\theta})\right|. \qquad \text{(A.5)}\end{aligned}$$

We further notice that

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{\mathcal{L}}_T(\boldsymbol{\theta})-\mathcal{L}_T(\boldsymbol{\theta})\right| \leq \frac{1}{T}\sum_{t=2}^{T}\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{l}_t(\boldsymbol{\theta})-l_t(\boldsymbol{\theta})\right|$$

By Lemma 2.1 in Straumann & Mikosch (2006), $\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{\mathcal{L}}_T(\boldsymbol{\theta})-\mathcal{L}_T(\boldsymbol{\theta})\right| \xrightarrow{a.s.} 0$ as $T\to\infty$ since by Assumption 7 $\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{l}_t(\boldsymbol{\theta})-l_t(\boldsymbol{\theta})\right| \xrightarrow{e.a.s.} 0$ as $t\to\infty$.

Now we turn to the second term in (A.5). To show the uniform convergence, we apply the uniform law of large numbers of Rao (1962) to the sequence $\{l_t(\cdot)\}_{t\in\mathbb{Z}}$. The sequence is SE by Krengel (1985, Proposition 4.3) since it is a continuous function on the SE sequence $\{(\boldsymbol{f}_t,\boldsymbol{y}_t)\}_{t\in\mathbb{Z}}$. Moreover, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|l_t(\boldsymbol{\theta})|<\infty$ since

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|l_t(\boldsymbol{\theta})| \leq \frac{1}{2}\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|\log|\boldsymbol{\Sigma}|| + \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\log g\left((\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t)\right)\right|<\infty, \qquad \text{(A.6)}$$

where the first term is bounded by condition B on the parameter space Θ and compactness of Θ. Assumption 8 ensures that the second term in (A.6) is also bounded. Therefore, the conditions of the uniform law of large numbers are satisfied, and we conclude that the second term in (A.5) also goes to 0 almost surely as $T\to\infty$.

The level sets of the limit log-likelihood function are regular since the parameter space Θ is compact (Assumption 6) and the limit criterion function is continuous. Hence, the consistency towards the set of maximizers follows.

If, in addition, $\text{sign}(\lambda_{ik})$ is known for some $i=1,\dots,N$ and for all $k=1,\dots,r$, then parameter $\boldsymbol{\theta}_0$ is point identified; see Proposition 2. Since $\mathbb{E}|l_t(\boldsymbol{\theta})|<\infty$ for all $\boldsymbol{\theta}\in\Theta$, then it immediately follows that $\boldsymbol{\theta}_0$ is the unique maximizer of the limit log-likelihood. This result together with the compactness of Θ and continuity of the limit log-likelihood further imply identifiable uniqueness.

This finishes the proof of the strong consistency.

■

Proof of Theorem 2. The parameter space $\tilde{\Theta}$ is a closed subset of Θ subject to the orthogonality constraint $\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}=\boldsymbol{I}_r$, which introduces $m=\frac{r(r+1)}{2}$ constraints. The

constraints can be summarized as follows

$$C_{\mathrm{d}}(\boldsymbol{\theta}) := \operatorname{diag}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right) - \boldsymbol{\iota}_r = \mathbf{0}_r,$$
$$C_{\mathrm{ndh}}(\boldsymbol{\theta}) := \operatorname{ndiagh}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right) - \mathbf{0}_{r(r-1)/2} = \mathbf{0}_{r(r-1)/2},$$

where $\operatorname{diag}(\cdot)$ and $\operatorname{ndiagh}(\cdot)$ refer to the diagonal and lower diagonal elements of the matrix. To shorten further notation, we introduce $C(\boldsymbol{\theta}) := (C_{\mathrm{d}}(\boldsymbol{\theta}), C_{\mathrm{ndh}}(\boldsymbol{\theta})) = \mathbf{0}_m$, with $C(\boldsymbol{\theta})$ collecting all the constraints.

Given the constraints, we can write the penalized criterion function as

$$\hat{L}_T(\boldsymbol{\zeta}) = T\hat{\mathcal{L}}_T(\boldsymbol{\theta}) + W(\boldsymbol{\zeta}),$$

where $\boldsymbol{\zeta} := (\boldsymbol{\theta}^\top, \boldsymbol{\xi}^\top)^\top$, $\hat{\mathcal{L}}_T(\boldsymbol{\theta})$ denotes the average log-likelihood function based on the filtered time-varying parameter $\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})$ as in Theorem 1, $W(\boldsymbol{\zeta}) := \boldsymbol{\xi}^\top \times C(\boldsymbol{\theta})$ and $\boldsymbol{\xi}$ is an m-dimensional vector of Lagrange multipliers.

Let us further denote the penalized log-likelihood function based on the limit time-varying parameter $\boldsymbol{f}_t(\boldsymbol{\theta})$ as $L_T(\boldsymbol{\zeta}) = T\mathcal{L}_T(\boldsymbol{\theta}) + W(\boldsymbol{\zeta})$. Similar to Blasques et al. (2023), first, we show the asymptotic normality of the estimator $\bar{\boldsymbol{\zeta}}_T$ which maximizes the criterion function $L_T(\boldsymbol{\zeta})$. The mean value theorem around $\boldsymbol{\zeta}_0$ yields

$$\nabla_{\boldsymbol{\zeta}} L_T(\bar{\boldsymbol{\zeta}}_T) = \nabla_{\boldsymbol{\zeta}} L_T(\boldsymbol{\zeta}_0) + \nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}} L_T(\boldsymbol{\zeta}_T^\star)(\bar{\boldsymbol{\zeta}}_T - \boldsymbol{\zeta}_0), \tag{A.7}$$

where $\boldsymbol{\zeta}_T^\star$ lies between $\bar{\boldsymbol{\zeta}}_T$ and $\boldsymbol{\zeta}_0$, and where formally $\boldsymbol{\zeta}_T^\star$ differs between the rows of the Hessian matrix $\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}} L_T$.

Since the estimator $\bar{\boldsymbol{\zeta}}_T$ maximizes $L_T(\boldsymbol{\zeta})$, from the first order condition, we obtain $\nabla_{\boldsymbol{\zeta}} L_T(\bar{\boldsymbol{\zeta}}_T) = 0$. Hence, rearranging the terms

$$\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}} L_T(\boldsymbol{\zeta}_T^\star) \begin{bmatrix} \sqrt{T}(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \\ \sqrt{T}(\bar{\boldsymbol{\xi}}_T - \boldsymbol{\xi}_0) \end{bmatrix} = -\sqrt{T}\nabla_{\boldsymbol{\zeta}} L_T(\boldsymbol{\zeta}_0). \tag{A.8}$$

In our notation, we have

$$\nabla_{\boldsymbol{\zeta}} L_T(\boldsymbol{\zeta}) = \begin{bmatrix} \nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\zeta}) \\ \nabla_{\boldsymbol{\xi}} L_T(\boldsymbol{\zeta}) \end{bmatrix} = \begin{bmatrix} T\nabla_{\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} W(\boldsymbol{\zeta}) \\ \boldsymbol{C}(\boldsymbol{\theta}) \end{bmatrix},$$

$$\nabla_{\zeta\zeta} L_T(\boldsymbol{\zeta}) = \begin{bmatrix} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} L_T(\boldsymbol{\zeta}) & \nabla_{\boldsymbol{\theta}\boldsymbol{\xi}} L_T(\boldsymbol{\zeta}) \\ \nabla_{\boldsymbol{\xi}\boldsymbol{\theta}} L_T(\boldsymbol{\zeta}) & \nabla_{\boldsymbol{\xi}\boldsymbol{\xi}} L_T(\boldsymbol{\zeta}) \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} L_T(\boldsymbol{\zeta}) & \boldsymbol{Q}(\boldsymbol{\theta}) \\ \boldsymbol{Q}(\boldsymbol{\theta})^\top & \mathbf{0}_m \end{bmatrix}$$
$$= \begin{bmatrix} T\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} W(\boldsymbol{\zeta}) & \boldsymbol{Q}(\boldsymbol{\theta}) \\ \boldsymbol{Q}(\boldsymbol{\theta})^\top & \mathbf{0}_m \end{bmatrix},$$

where $\boldsymbol{Q}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \boldsymbol{C}(\boldsymbol{\theta})^\top$. Substituting these expressions into (A.8) and taking into account that $\boldsymbol{C}(\boldsymbol{\theta}_0) = \mathbf{0}_m$, we obtain

$$\begin{bmatrix} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} L_T(\boldsymbol{\zeta}_T^\star) & \boldsymbol{Q}(\boldsymbol{\theta}_T^\star) \\ \boldsymbol{Q}(\boldsymbol{\theta}_T^\star)^\top & \mathbf{0}_m \end{bmatrix} \begin{bmatrix} \sqrt{T}(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \\ \sqrt{T}(\bar{\boldsymbol{\xi}}_T - \boldsymbol{\xi}_0) \end{bmatrix} = \begin{bmatrix} -\sqrt{T}(T\nabla_{\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} W(\boldsymbol{\zeta}_0)) \\ \mathbf{0}_m \end{bmatrix},$$
$$\begin{bmatrix} \frac{1}{T}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} L_T(\boldsymbol{\zeta}_T^\star) & \boldsymbol{Q}(\boldsymbol{\theta}_T^\star) \\ \boldsymbol{Q}(\boldsymbol{\theta}_T^\star)^\top & \mathbf{0}_m \end{bmatrix} \begin{bmatrix} \sqrt{T}(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \\ \frac{1}{\sqrt{T}}(\bar{\boldsymbol{\xi}}_T - \boldsymbol{\xi}_0) \end{bmatrix} = \begin{bmatrix} -\sqrt{T}\nabla_{\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}_0) - \frac{1}{\sqrt{T}}\nabla_{\boldsymbol{\theta}} W(\boldsymbol{\zeta}_0) \\ \mathbf{0}_m \end{bmatrix}.$$

Trivially, $\frac{1}{\sqrt{T}}\nabla_{\boldsymbol{\theta}} W(\boldsymbol{\zeta}_0) \to \mathbf{0}_q$ and $\frac{1}{T}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} W(\boldsymbol{\zeta}_T^\star) \xrightarrow{a.s.} \mathbf{0}_{q\times q}$ as $T \to \infty$. Furthermore, Assumption 10 together with the strong consistency of the estimator $\bar{\boldsymbol{\theta}}_T$ imply $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}_T^\star) \xrightarrow{P} \mathcal{I}(\boldsymbol{\theta}_0)$ as $T \to \infty$. Therefore, by the strong consistency of the estimator $\bar{\boldsymbol{\theta}}_T$, Lemma SB.1 and Slutsky's lemma, it follows that

$$\begin{bmatrix} \mathcal{I}(\boldsymbol{\theta}_0) & \boldsymbol{Q}(\boldsymbol{\theta}_0) \\ \boldsymbol{Q}(\boldsymbol{\theta}_0)^\top & \mathbf{0}_m \end{bmatrix} \begin{bmatrix} \sqrt{T}(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \\ \frac{1}{\sqrt{T}}(\bar{\boldsymbol{\xi}}_T - \boldsymbol{\xi}_0) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \boldsymbol{z} \\ \mathbf{0}_m \end{bmatrix} \quad \text{as} \quad T \to \infty, \tag{A.9}$$

where $\boldsymbol{z}$ is $\mathcal{N}(\mathbf{0}_q, \boldsymbol{V}(\boldsymbol{\theta}_0))$ with $\boldsymbol{V}(\boldsymbol{\theta}_0)$ as defined in Lemma SB.1.

Since $\boldsymbol{\theta}_0$ is not identifiable without restrictions, matrix $\mathcal{I}(\boldsymbol{\theta}_0)$ is singular which leads to a degenerate limiting distribution. Following a similar argument as in Silvey (1959), we notice that (A.9) implies that $\boldsymbol{Q}(\boldsymbol{\theta}_0)^\top \sqrt{T}(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{0}_m$ as $T \to \infty$. Hence, by Slutsky's lemma, the result in (A.9) can be rewritten as

$$\begin{bmatrix} \mathcal{I}(\boldsymbol{\theta}_0) + \boldsymbol{Q}(\boldsymbol{\theta}_0)\boldsymbol{Q}(\boldsymbol{\theta}_0)^\top & \boldsymbol{Q}(\boldsymbol{\theta}_0) \\ \boldsymbol{Q}(\boldsymbol{\theta}_0)^\top & \mathbf{0}_m \end{bmatrix} \begin{bmatrix} \sqrt{T}(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \\ \frac{1}{\sqrt{T}}(\bar{\boldsymbol{\xi}}_T - \boldsymbol{\xi}_0) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \boldsymbol{z} \\ \mathbf{0} \end{bmatrix} \quad \text{as} \quad T \to \infty. \tag{A.10}$$

Since the restriction $\boldsymbol{C}(\boldsymbol{\theta}) = 0$ ensures identifiability of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0$ is a regular point of both $\boldsymbol{Q}(\boldsymbol{\theta}_0)$ and $[\mathcal{I}(\boldsymbol{\theta})^{1/2}, \boldsymbol{Q}(\boldsymbol{\theta})]^\top$, the 'augmented' limit Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta}_0) := \mathcal{I}(\boldsymbol{\theta}_0) + \boldsymbol{Q}(\boldsymbol{\theta}_0)\boldsymbol{Q}(\boldsymbol{\theta}_0)^\top$ is positive definite (Rothenberg, 1971; Silvey, 1975). Therefore, by Aitchison & Silvey (1958, Lemma 3) $\boldsymbol{S}(\boldsymbol{\theta}_0) := \begin{bmatrix} \boldsymbol{H}(\boldsymbol{\theta}_0) & \boldsymbol{Q}(\boldsymbol{\theta}_0) \\ \boldsymbol{Q}(\boldsymbol{\theta}_0)^\top & \mathbf{0}_m \end{bmatrix}$ is non-singular.

Next, let us introduce the following notation for the inverse of matrix $\boldsymbol{S}$, i.e. $\boldsymbol{S}^{-1} := \begin{bmatrix} \boldsymbol{P} & \boldsymbol{D} \\ \boldsymbol{D}^\top & \boldsymbol{R} \end{bmatrix}$, where, to simplify the expressions, we suppress the dependence of the matrices

on $\boldsymbol{\theta}_0$. Then, from the formula for the partitioned inverse, for the block $\boldsymbol{P}$ we have

$$\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{Q}\left(\boldsymbol{Q}^\top \boldsymbol{H}^{-1}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}^\top \boldsymbol{H}^{-1}. \tag{A.11}$$

Therefore, for the limiting distribution of the constrained estimator we have

$$\sqrt{T}\left(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{P}\boldsymbol{V}\boldsymbol{P}\right) \quad \text{as} \quad T \to \infty,$$

with $\boldsymbol{P}$ defined in (A.11).

Now, we turn to the asymptotic properties of the estimator $\hat{\boldsymbol{\theta}}_T$ of the penalized criterion function based on the filtered time-varying parameter. By the mean value theorem

$$\nabla_{\boldsymbol{\zeta}} L_T(\bar{\boldsymbol{\zeta}}_T) = \nabla_{\boldsymbol{\zeta}} L_T(\hat{\boldsymbol{\zeta}}_T) + \nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}} L_T(\boldsymbol{\zeta}_T^\star)(\bar{\boldsymbol{\zeta}}_T - \hat{\boldsymbol{\zeta}}_T), \tag{A.12}$$

where $\boldsymbol{\zeta}_T^\star$ lies (row-wise) between $\bar{\boldsymbol{\zeta}}_T$ and $\hat{\boldsymbol{\zeta}}_T$.

Noticing that $\nabla_{\boldsymbol{\zeta}} L_T(\bar{\boldsymbol{\zeta}}_T) = \nabla_{\boldsymbol{\zeta}} \hat{L}_T(\hat{\boldsymbol{\zeta}}_T) = \mathbf{0}_{q+m}$, from (A.12), we have

$$\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}} L_T(\boldsymbol{\zeta}_T^\star)\sqrt{T}(\bar{\boldsymbol{\zeta}}_T - \hat{\boldsymbol{\zeta}}_T) = \sqrt{T}\left(\nabla_{\boldsymbol{\zeta}} \hat{L}_T(\hat{\boldsymbol{\zeta}}_T) - \nabla_{\boldsymbol{\zeta}} L_T(\hat{\boldsymbol{\zeta}}_T)\right).$$

By Lemma SB.2 and the strong consistency of the estimator $\hat{\boldsymbol{\theta}}_T$, the right hand side of the expression above converges to 0 almost surely as $T \to \infty$. Moreover, by the strong consistency of the estimator $\bar{\boldsymbol{\theta}}_T$ and Assumption 10, we have $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}_T^\star) \xrightarrow{a.s.} \mathcal{I}(\boldsymbol{\theta}_0)$. Then, following similar reasoning as in the first part of the proof, we obtain

$$\begin{bmatrix} \mathcal{I}(\boldsymbol{\theta}_0) + \boldsymbol{Q}(\boldsymbol{\theta}_0)\boldsymbol{Q}(\boldsymbol{\theta}_0)^\top & \boldsymbol{Q}(\boldsymbol{\theta}_0) \\ \boldsymbol{Q}(\boldsymbol{\theta}_0)^\top & \mathbf{0}_m \end{bmatrix} \begin{bmatrix} \sqrt{T}(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \\ \frac{1}{\sqrt{T}}(\bar{\boldsymbol{\xi}}_T - \boldsymbol{\xi}_0) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \text{as} \quad T \to \infty. \tag{A.13}$$

Since matrix $\mathcal{I}(\boldsymbol{\theta}_0) + \boldsymbol{Q}(\boldsymbol{\theta}_0)\boldsymbol{Q}(\boldsymbol{\theta}_0)^\top$ is full rank, it follows that $\sqrt{T}\|\bar{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T\| \xrightarrow{P} 0$ as $T \to \infty$ which implies that the estimator $\hat{\boldsymbol{\theta}}_T$ has the same asymptotic distribution as $\bar{\boldsymbol{\theta}}_T$, thus completing the proof. ■

Supplementary Appendix

An Order-Invariant Score-Driven Dynamic Factor Model

Mariia Artemova

A Specific cases of elliptical distribution

A.1 Gaussian model

A.1.1 Updating equation

For Gaussian distribution the density generator is of the form $g(u) = c\exp(-u/2)$ (Fang et al., 2018, Table 3.1). Therefore, it can be noticed that $g'(u) = -\frac{1}{2}g(u)$, which implies that $-2\frac{g'(u)}{g(u)} = 1$ and $C(\|\tilde{\boldsymbol{y}}_t\|, g) = -\frac{N}{2}$. Therefore, from equation (3) and expression for the score (5), we obtain

$$\boldsymbol{f}_{t+1} = \boldsymbol{\omega} + \boldsymbol{A}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t) + \boldsymbol{B}\boldsymbol{f}_t,$$

or

$$\boldsymbol{f}_{t+1} = \boldsymbol{\omega} + \boldsymbol{A}\left(\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - \boldsymbol{f}_t\right) + \boldsymbol{B}\boldsymbol{f}_t.$$

A.1.2 Condition (i), Proposition 1

For the Gaussian model, condition (i) in Proposition 1 is of the form $\log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{B}-\boldsymbol{A}\| < \infty$, which trivially holds as long as $\Theta \subseteq \mathbb{R}^q$ is compact.

A.1.3 Condition (ii), Proposition 1

Exploiting the expressions derived in Section A.1.1 condition (ii) for the Gaussian model can be rewritten as follows

$$\begin{aligned}
&\mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\boldsymbol{A}\left(\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - \boldsymbol{f}\right) + \boldsymbol{B}\boldsymbol{f}\right\| \\
&\leq 2\log 2 + \mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\boldsymbol{A}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t\right\| + \log^+\|(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{f}\| \\
&\leq 2\log 2 + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{A}\| + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right\| + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{1}{N}\boldsymbol{\Lambda}^\top\right\| \\
&+ \log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\boldsymbol{\Sigma}^{-1}\right\| + \mathbb{E}\log^+\|\boldsymbol{y}_t\| + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{B}-\boldsymbol{A}\| + \log^+\|\boldsymbol{f}\|,
\end{aligned}$$

where we applied Lemma 2.2 from Straumann & Mikosch (2006), norm subadditivity and submultiplicativity. The expression is finite as long as $\mathbb{E}\log^+\|\boldsymbol{y}_t\| < \infty$, $\boldsymbol{\Sigma} \succ 0$ and $\left\|\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right\| < \infty$, $\Theta \subseteq \mathbb{R}^q$ is compact, since $\boldsymbol{f} \in \mathbb{R}^r$. Hence, we conclude that for the Gaussian model condition (ii) holds as long as $\mathbb{E}\log^+\|\boldsymbol{y}_t\| < \infty$, $\boldsymbol{\Sigma} \succ 0$, $\left\|\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right\| < \infty$, and $\Theta \subseteq \mathbb{R}^q$ is compact.

A.1.4 Assumptions 7 and 8.

Assumption 8 for the score-driven model with Gaussian innovations is of the following form

$$\begin{aligned}
&\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\log g\left(\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t\right)^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t\right)\right)\right| \\
\leq& c+0.5\times\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t\right)^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t\right)\right| \\
\leq& c+0.5\sum_{i=1}^{N}\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{(y_{it}-\boldsymbol{\Lambda}_{i\cdot}\boldsymbol{f}_t)^2}{\sigma_i^2}\right| < \infty,
\end{aligned}$$

where c is a constant that does not depend on the parameter $\boldsymbol{\theta}$. The last result follows by condition B and by Lemmas 3 and 4 for $k=2$.

Now we turn to Assumption 7. First, we notice that

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{l}_t(\boldsymbol{\theta})-l_t(\boldsymbol{\theta})\right| = \frac{1}{2}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})\right)^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})\right)\right. \\
&\left.-\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta})\right)^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta})\right)\right| \leq \sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\|\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t\| \\
&+\frac{1}{2}\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})^\top\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})^\top\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{f}_t(\boldsymbol{\theta})\right| \\
&=\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\|\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t\|+\frac{N}{2}\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})^\top\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})^\top\boldsymbol{f}_t(\boldsymbol{\theta})\|, \qquad \text{(SA.1)}
\end{aligned}$$

where in the last line we exploited condition C.

Following a similar argument as in the proof of Lemma TA.14 in Blasques, van Brummelen, Koopman, & Lucas (2022), but generalizing it to a multivariate case, for the second term in (SA.1), by the triangle inequality, we have

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})^\top\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})^\top\boldsymbol{f}_t(\boldsymbol{\theta})\| \leq \sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})^\top\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})^\top\boldsymbol{f}_t(\boldsymbol{\theta})\| \\
&+\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})^\top\boldsymbol{f}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})^\top\boldsymbol{f}_t(\boldsymbol{\theta})\| \leq \sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})\|\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\| \\
&+\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\|\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}_t(\boldsymbol{\theta})\| \leq \sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})+\boldsymbol{f}_t(\boldsymbol{\theta})\|\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\| \\
&+\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\|\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}_t(\boldsymbol{\theta})\| \leq (\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\|)^2 \\
&+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}_t(\boldsymbol{\theta})\|\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\|+\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\|\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}_t(\boldsymbol{\theta})\|.
\end{aligned}$$

By compactness of Θ and Lemma 4 for $k=1$ $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}_t(\boldsymbol{\theta})\| < \infty$. Then, by Lemma 2.1 in Straumann & Mikosch (2006), $\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})^\top\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})^\top\boldsymbol{f}_t(\boldsymbol{\theta})\| \xrightarrow{e.a.s.} 0$ as $t\to\infty$ since by Proposition 1 $\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\| \xrightarrow{e.a.s.} 0$ as $t\to\infty$.

The last result implies that the whole expression in SA.1 converges e.a.s. to 0 as $t\to\infty$ as long as $\{\boldsymbol{y}_t\}_{t\in\mathbb{Z}}$ is SE with $\mathbb{E}\log^+\|\boldsymbol{y}_t\| < \infty$ and the parameter space Θ is compact.

A.2 Student's t model

A.2.1 Updating equation

In the case of the Student's t distributed innovations, the density generator is of the form $g(u) = c\,(1+u/\nu)^{-(N+\nu)/2}$ (Fang et al., 2018, Table 3.1), where ν is the degrees of freedom parameter. Therefore, $g'(u) = -\frac{N+\nu}{2\nu} c\,(1+u/\nu)^{-(N+\nu+2)/2}$ and $-2\frac{g'(u)}{g(u)} = \frac{N+\nu}{\nu}\frac{1}{1+u/\nu}$. Then, we have

$$C(\|\tilde{\boldsymbol{y}}_t\|, g) = -\frac{1}{2}\left(\frac{N+\nu}{\nu}\right)^2 \mathbb{E}_{t-1}\left[\|\tilde{\boldsymbol{y}}_t\|^2 \left(\frac{1}{1+\|\tilde{\boldsymbol{y}}_t\|^2/\nu)}\right)^2\right]. \tag{SA.2}$$

First, we compute the conditional expectation that appears in (SA.2). Given that $\boldsymbol{y}_t|\boldsymbol{f}_t, \mathcal{F}_{t-1}$ is multivariate Student's t distributed we have

$$\begin{aligned}
&\mathbb{E}_{t-1}\left[\|\tilde{\boldsymbol{y}}_t\|^2 \left(\frac{1}{1+\|\tilde{\boldsymbol{y}}_t\|^2/\nu}\right)^2\right] \\
&= \int_{-\infty}^{+\infty} \left(\|\tilde{\boldsymbol{y}}_t\|^2/\nu\right) \frac{\nu\Gamma\left(\frac{N+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{N/2}\pi^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \left(1+\frac{1}{\nu}\|\tilde{\boldsymbol{y}}_t\|^2\right)^{-(N+\nu)/2} \left(1+\frac{1}{\nu}\|\tilde{\boldsymbol{y}}_t\|^2\right)^{-2} \mathrm{dy} \\
&= \frac{\nu\Gamma\left(\frac{N+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{N/2}\pi^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \int_{-\infty}^{+\infty} \left(\|\tilde{\boldsymbol{y}}_t\|^2/\nu\right) \left(1+\frac{1}{\nu}\|\tilde{\boldsymbol{y}}_t\|^2\right)^{-(N+\nu+4)/2} \mathrm{dy} \\
&= \frac{\nu\Gamma\left(\frac{N+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{N/2}\pi^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \int_{-\infty}^{+\infty} \left[\left(1+\frac{1}{\nu}\|\tilde{\boldsymbol{y}}_t\|^2\right)^{-(N+\nu+2)/2} - \left(1+\frac{1}{\nu}\|\tilde{\boldsymbol{y}}_t\|^2\right)^{-(N+\nu+4)/2}\right] \mathrm{dy} \\
&= \frac{\nu\Gamma\left(\frac{N+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{N/2}\pi^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \left[\frac{\Gamma\left(\frac{\nu+2}{2}\right)(\nu+2)^{N/2}\pi^{N/2}\left(\frac{\nu}{\nu+2}\right)^{N/2}|\boldsymbol{\Sigma}|^{1/2}}{\Gamma\left(\frac{N+\nu+2}{2}\right)}\right. \\
&\left. - \frac{\Gamma\left(\frac{\nu+4}{2}\right)(\nu+4)^{N/2}\pi^{N/2}\left(\frac{\nu}{\nu+4}\right)^{N/2}|\boldsymbol{\Sigma}|^{1/2}}{\Gamma\left(\frac{N+\nu+4}{2}\right)}\right] = \frac{\nu\Gamma\left(\frac{N+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left[\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{N+\nu+2}{2}\right)} - \frac{\Gamma\left(\frac{\nu+4}{2}\right)}{\Gamma\left(\frac{N+\nu+4}{2}\right)}\right] \\
&= \frac{\nu\Gamma\left(\frac{N+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{N+\nu+2}{2}\right)}\left[1-\frac{\frac{\nu+2}{2}}{\frac{N+\nu+2}{2}}\right] = \frac{\nu\Gamma\left(\frac{N+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{\Gamma\left(\frac{\nu}{2}\right)\frac{\nu}{2}}{\Gamma\left(\frac{N+\nu}{2}\right)\frac{N+\nu}{2}}\frac{N}{N+\nu+2} \\
&= \frac{\nu^2}{N+\nu}\frac{N}{N+\nu+2}.
\end{aligned}$$

Substituting the last expression into (SA.2) we have

$$C(\|\tilde{\boldsymbol{y}}_t\|, g) = -\frac{1}{2}\left(\frac{N+\nu}{\nu}\right)^2 \frac{\nu^2}{N+\nu}\frac{N}{N+\nu+2} = -\frac{1}{2}\frac{N(N+\nu)}{(N+\nu+2)}.$$

From the last result and equation (5) the scaled score is as follows

$$\boldsymbol{s}_t = \frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|, \nu)}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t),$$

or, alternatively,

$$\boldsymbol{s}_t = \frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|, \nu)}\left(\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - \boldsymbol{f}_t\right),$$

where $W(\|\tilde{\boldsymbol{y}}_t\|, \nu) := \frac{\nu}{N+\nu+2}\left(1 + \frac{(\boldsymbol{y}_t - \boldsymbol{\Lambda} f_t)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda} f_t)}{\nu}\right)$.

A.2.2 Condition (i), Proposition 1

For the Student's t model condition (i) in Proposition 1 takes the form

$$
\begin{aligned}
&\mathbb{E}\log^+ \sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\boldsymbol{B} + \frac{1}{W(\|\tilde{\boldsymbol{y}}_t\|,\nu)}\boldsymbol{A}\frac{N+\nu+2}{\nu} \times \left(\frac{2}{\nu W(\|\tilde{\boldsymbol{y}}_t\|,\nu)}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right.\right.\\
&\left.\left.\times\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - \boldsymbol{f}\right)(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} - \boldsymbol{I}_r\right)\right\|\\
&\le c + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{B}\| + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{2(N+\nu+2)}{\nu}\boldsymbol{A}\right\| + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right\|\\
&+ \mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^q}\left\|\frac{1}{\nu(W(\|\tilde{\boldsymbol{y}}_t\|,\nu))^2}\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - \boldsymbol{f}\right)(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right\|\\
&\le c + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{B}\| + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{2(N+\nu+2)}{\nu}\boldsymbol{A}\right\| + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right\|\\
&+ \mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^q}\left\|\frac{\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - \boldsymbol{f}\right)(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}}{\nu(1 + (\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})/\nu)^2}\right\|,
\end{aligned}
$$

where c is a constant that does not depend on $\boldsymbol{\theta}$ and $\boldsymbol{f}$. The first four terms are bounded as long as $\nu > 0$, $\|\boldsymbol{P}\| < \infty$, and the parameter space $\Theta \subseteq \mathbb{R}^q$ is compact.

For the last term in the expression above we have

$$
\begin{aligned}
\boldsymbol{z}_t &:= \mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^q}\left\|\frac{\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_t - \boldsymbol{f})(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}/\nu}{(1 + (\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})/\nu)^2}\right\|\\
&= \mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\frac{\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}}{(1 + \boldsymbol{x}_t^\top\boldsymbol{x}_t)^2}\right\|\\
&= \log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\| + \mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\frac{\boldsymbol{x}_t\boldsymbol{x}_t^\top}{(1 + \boldsymbol{x}_t^\top\boldsymbol{x}_t)^2}\right\|,
\end{aligned}
$$

where $\boldsymbol{x}_t := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})/\sqrt{\nu}$. As $\|\boldsymbol{x}_t\| \to 0$ or $\|\boldsymbol{x}_t\| \to \infty$, it follows that the term $\|\boldsymbol{z}_t\| \to 0$. Therefore, we conclude that the term $\boldsymbol{z}_t$ is uniformly bounded in $(\boldsymbol{f}, \boldsymbol{y}_t)$. If then the compactness of $\Theta \subseteq \mathbb{R}^q$ holds, condition (i) in Proposition 1 holds.

A.2.3 Condition (ii), Proposition 1

Below, we verify condition (ii) in Proposition 1 for the Student's t model, i.e.

$$
\begin{aligned}
&\mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\boldsymbol{A}\frac{(N+\nu+2)}{\nu}\frac{\left(\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - \boldsymbol{f}\right)}{1 + (\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})/\nu} + \boldsymbol{B}\boldsymbol{f}\right\|\\
&\le 2\log 2 + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{B}\| + \log^+\|\boldsymbol{f}\| + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{A}\| + \log^+\sup_{\boldsymbol{\theta}\in\Theta}\frac{(N+\nu+2)}{\sqrt{N\nu}}\\
&+ \log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\frac{1}{\sqrt{N}}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1/2}\right\|
\end{aligned}
$$

$$+\mathbb{E}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{\boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}\right)/\sqrt{\nu}}{1+\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}\right)/\nu}\right\|\leq c+\log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{B}\|$$
$$+\log^+\|\boldsymbol{f}\|+\log^+\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{A}\|+\log^+\sup_{\boldsymbol{\theta}\in\Theta}\frac{(N+\nu+2)}{\sqrt{N\nu}}+\frac{1}{2}\log^+\sup_{\boldsymbol{\theta}\in\Theta}\left\|\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right\|.$$

where we exploited the fact that the term $\frac{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})/\sqrt{\nu}}{1+(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})/\nu}$ is uniformly bounded. Therefore, if the parameter space Θ is compact, the whole expression above is finite as long as $\boldsymbol{\Sigma}\succ 0$, $\left\|\left(\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right\|<\infty$ and $0<\nu<\infty$.

Assumption 12.

To show that $\left\|\frac{\partial\hat{l}_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}-\frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right\|\xrightarrow{e.a.s.}0$ as $t\to\infty$, we notice that by the mean value theorem

$$\left\|\frac{\partial l_t(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}-\frac{\partial l(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right\|\leq\sup_{\boldsymbol{f}\in\mathbb{R}^r}\left\|\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{f}^\top}\right\|\sup_{\boldsymbol{\theta}\in\Theta}\left\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\right\|\xrightarrow{e.a.s.}0,$$

where the final result follows by Lemma 2.1 in Straumann & Mikosch (2006) since by Proposition 1 $\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})-\boldsymbol{f}_t(\boldsymbol{\theta})\|$ converges to zero e.a.s., while by Lemma SC.6 $\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{f}^\top$ is uniformly bounded in $\boldsymbol{f}$ and $\boldsymbol{\theta}$ for Gaussian and Student's t case and by Krengel (1985, Proposition 4.3) it is also SE sequence.

B Asymptotic normality: Additional lemmas

Lemma SB.1. *Let all the assumptions and conditions of Theorem 2 hold. Then*

$$\sqrt{T}\nabla_{\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}(\mathbf{0},\boldsymbol{V}(\boldsymbol{\theta}_0))\quad\text{as } T\to\infty,$$

where $\boldsymbol{V}(\boldsymbol{\theta}_0)=\mathbb{E}\left[l_t'(\boldsymbol{\theta}_0)l_t'(\boldsymbol{\theta}_0)^\top\right]$ *with* $l_t'(\boldsymbol{\theta}):=\frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ *and* $\nabla_{\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta}):=\frac{\partial\mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$.

Proof. We recall that $\nabla_{\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta}_0)=\frac{1}{T}\sum_{t=2}^T l_t'(\boldsymbol{\theta}_0)$, where

$$l_t'(\boldsymbol{\theta}):=\frac{\partial l(\boldsymbol{y}_t,\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{f}=\boldsymbol{f}_t(\boldsymbol{\theta})}+\left(\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{f}^\top}\Big|_{\boldsymbol{f}=\boldsymbol{f}_t(\boldsymbol{\theta})}\frac{\partial\boldsymbol{f}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top}\right)^\top.\tag{SB.3}$$

To prove the theorem, we apply the Central Limit Theorem (CLT) for stationary and ergodic (SE) martingale difference sequences (mds) of Billingsley (1961) to the sequence $\{l_t'(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$. To apply the theorem, we show that the sequence is an SE mds sequence with two bounded moments.

The stationarity and ergodicity of $\{l_t'(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ immediately follows by application of Krengel (1985, Proposition 4.3) since l_t' is a continuous function on SE sequence $\{(\boldsymbol{y}_t,\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{f}_t'(\boldsymbol{\theta}_0))\}_{t\in\mathbb{Z}}$. The latter sequence is SE by Lemma 2 and Assumption 11 and the fact that $\{\boldsymbol{y}_t\}_{t\in\mathbb{Z}}\equiv\{\boldsymbol{y}_t(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is an SE sequence.

Furthermore, $\{l_t'(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is an mds sequence since, under the correct model specification, we have $\mathbb{E}[l_t(\boldsymbol{\theta}_0)|\mathcal{F}_{t-1}]=0$, hence $\mathbb{E}[l_t'(\boldsymbol{\theta}_0)|\mathcal{F}_{t-1}]=\partial\mathbb{E}[l_t(\boldsymbol{\theta}_0)|\mathcal{F}_{t-1}]/\partial\boldsymbol{\theta}=\mathbf{0}_q$. The interchange of the expectation and derivative is permitted since the likelihood function is continuous and the derivative with respect to $\boldsymbol{\theta}$ is uniformly bounded, which allows the application of the measure theory version of the Leibniz integral rule.

Finally, we show that the second moment of $l_t'(\boldsymbol{\theta}_0)$ is bounded. Specifically, from (SB.3), by the Loève's c_r inequality, we notice that

$$\mathbb{E}\|l_t'(\boldsymbol{\theta}_0)\|^2 \leq c_r\mathbb{E}\left\|\frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right\|^2 + c_r\mathbb{E}\left\|\frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}^\top}\frac{\partial \boldsymbol{f}_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top}\right\|^2. \tag{SB.4}$$

Assumption 13 (i) implies that the first term in (SB.4) is bounded. For the second term, by the generalized Hölder inequality, we have

$$\mathbb{E}\left\|\frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}^\top}\frac{\partial \boldsymbol{f}_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top}\right\|^2 \leq \mathbb{E}\left\|\frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}}\right\|^\delta \mathbb{E}\left\|\boldsymbol{f}_t'(\boldsymbol{\theta}_0)\right\|^n < \infty, \tag{SB.5}$$

where $n = \frac{2\delta}{2-\delta}$ and the last result follows by Assumption 13 (iii). Hence, we conclude that $\mathbb{E}\|l_t'(\boldsymbol{\theta}_0)\| < \infty$ and the desired result follows by the CLT.

■

Lemma SB.2. *Let all the assumptions and conditions of Theorem 2 hold. Then*

$$\sqrt{T}\|\nabla_{\boldsymbol{\theta}}\hat{\mathcal{L}}_T(\boldsymbol{\theta}_0) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta}_0)\| \xrightarrow{a.s.} 0 \quad \text{as } T \to \infty,$$

with $\nabla_{\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta})$ as defined in the Supplementary Appendix C.3.

Proof. First, we show that $\|\hat{l}_t'(\boldsymbol{\theta}_0) - l_t'(\boldsymbol{\theta}_0)\| \xrightarrow{e.a.s.} 0$ as $t \to \infty$. By the norm subadditivity, we have

$$\begin{aligned}\|\hat{l}_t'(\boldsymbol{\theta}_0) - l_t'(\boldsymbol{\theta}_0)\| \leq &\left\|\frac{\partial l_t(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} - \frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right\| \\ &+ \left\|\hat{\boldsymbol{f}}_t'(\boldsymbol{\theta}_0)^\top\frac{\partial l_t(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}} - \boldsymbol{f}_t'(\boldsymbol{\theta}_0)^\top\frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}}\right\|.\end{aligned} \tag{SB.6}$$

The first term on the right hand side of (SB.6) goes to 0 e.a.s. as $t \to \infty$ by Assumption 12.

Now, we turn to the second term on the right hand side of (SB.6). Following a similar argument as in the proof of Lemma TA.14 in Blasques, van Brummelen, Koopman, & Lucas (2022), we have

$$\begin{aligned}&\left\|\hat{\boldsymbol{f}}_t'(\boldsymbol{\theta}_0)^\top\frac{\partial l_t(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}_t} - \boldsymbol{f}_t'(\boldsymbol{\theta}_0)^\top\frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}_t}\right\| \leq (\|\boldsymbol{f}_t'(\boldsymbol{\theta}_0)\| + \|\hat{\boldsymbol{f}}_t'(\boldsymbol{\theta}_0) - \boldsymbol{f}_t'(\boldsymbol{\theta}_0)\|) \\ &\times\left\|\frac{\partial l_t(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}} - \frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}}\right\| + \left\|\frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}}\right\|\|\hat{\boldsymbol{f}}_t'(\boldsymbol{\theta}_0) - \boldsymbol{f}_t'(\boldsymbol{\theta}_0)\|.\end{aligned} \tag{SB.7}$$

Below, we show that both terms on the right hand side of (SB.7) go to 0 exponentially fast almost surely as $t \to \infty$. In particular, for the last term, we have that $\|\hat{\boldsymbol{f}}_t'(\boldsymbol{\theta}_0) - \boldsymbol{f}_t'(\boldsymbol{\theta}_0)\| \xrightarrow{e.a.s.} 0$ as $t \to \infty$ by Assumption 11. Additionally, $\left\|\frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}}\right\|$ is SE with a finite logarithmic moment, where the SE property follows by Krengel (1985, Proposition 4.3) as $\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})/\partial \boldsymbol{f}$ is a continuous function on the SE sequence $\{\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$. The existence of a logarithmic moment follows by Assumption 13 (iii). Then, by Lemma 2.1 in Straumann & Mikosch (2006), the second term on the right hand side of (SB.7) converges e.a.s. to 0 as $t \to \infty$.

For the first term on the right hand side of (SB.7), by the mean value theorem we have

$$\left\| \frac{\partial l_t(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0)}{\partial \boldsymbol{f}} - \frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0)}{\partial \boldsymbol{f}} \right\| \leq \sup_{\boldsymbol{f} \in \mathbb{R}^r} \left\| \frac{\partial^2 l_t(\boldsymbol{f}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{f} \partial \boldsymbol{f}^\top} \right\| \|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}_0) - \boldsymbol{f}_t(\boldsymbol{\theta}_0)\|.$$

By Proposition 1, $\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}) - \boldsymbol{f}_t(\boldsymbol{\theta})\|$ converges to 0 e.a.s. as $t \to \infty$, while by Assumption 13 (ii) $\sup_{\boldsymbol{f}\in\mathbb{R}^r} \|\partial^2 l_t(\boldsymbol{f}, \boldsymbol{\theta}_0)/\partial \boldsymbol{f} \partial \boldsymbol{f}^\top\|$ has a logarithmic bounded moment and by Krengel (1985, Proposition 4.3) it is also an SE sequence. Hence, by Lemma 2.1 in Straumann & Mikosch (2006) we conclude that the whole expression above converges to zero e.a.s. Given that by Assumption 11 $\|\hat{\boldsymbol{f}}'_t(\boldsymbol{\theta}_0) - \boldsymbol{f}'_t(\boldsymbol{\theta}_0)\| \xrightarrow{e.a.s.} 0$ and the sequence $\{\boldsymbol{f}'_t(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is SE with $\mathbb{E}\log^+ \|\boldsymbol{f}'_t(\boldsymbol{\theta}_0)\| < \infty$ (Assumption 13 (iii)), we obtain that $\|\hat{l}'_t(\boldsymbol{\theta}_0) - l'_t(\boldsymbol{\theta}_0)\| \xrightarrow{e.a.s.} 0$ as $t \to \infty$.

Finally, by the norm subadditivity and Lemma 2.1 of Straumann & Mikosch (2006), we have

$$\sqrt{T}\|\nabla_{\boldsymbol{\theta}} \hat{\mathcal{L}}_T(\boldsymbol{\theta}_0) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}_0)\| \leq \frac{1}{\sqrt{T}} \sum_{t=2}^{T} \|\hat{l}'_t(\boldsymbol{\theta}_0) - l'_t(\boldsymbol{\theta}_0)\| \xrightarrow{e.a.s.} 0 \quad \text{as } T \to \infty,$$

which finishes the proof. ■

Lemma SB.3 (Assumption 10 for the Gaussian and Student's t models)**.** *Let all the assumptions and conditions of Theorem 1 hold. Then,*

$$\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}_T(\hat{\boldsymbol{\theta}}_T) - \mathcal{I}(\boldsymbol{\theta}_0)\| \xrightarrow{P} 0 \quad \text{as } T \to \infty, \tag{SB.8}$$

where $\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}[l''_t(\boldsymbol{\theta})]$.

Proof. (i) Gaussian model: First, we notice that the desired result is implied by a stronger result, namely, $\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}) - \mathcal{I}(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$ as $T \to \infty$. We show the latter by application of the ergodic theorem of Rao (1962) to $\{l''_t(\cdot)\}_{t\in\mathbb{Z}}$. Specifically, the uniform convergence follows if 1). the sequence $\{l''_t(\cdot)\}_{t\in\mathbb{Z}}$ is SE and 2). $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|l''_t(\boldsymbol{\theta})\| < \infty$.

The first condition is satisfied by Krengel's theorem since $l''(\cdot)$ is a continuous function on the SE sequence $\{\boldsymbol{y}_t, \boldsymbol{f}_t(\cdot), \boldsymbol{f}'_t(\cdot), \boldsymbol{f}''_t(\cdot)\}_{t\in\mathbb{Z}}$, where the latter is true by Lemmas 2, 4 and SB.4 Next, we turn to the moment bound. We notice that by the norm subadditivity and generalized Hölder inequality

$$\begin{aligned}
\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|l''_t(\boldsymbol{\theta})\| \leq{}& \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{\partial^2 l_t(\boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right\| + 2\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{\partial^2 l_t(\boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{f}^\top} \right\|^2 \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}'_t(\boldsymbol{\theta})\|^2 \\
&+ \sup_t \sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{\partial^2 l_t(\boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})}{\partial \boldsymbol{f} \partial \boldsymbol{f}^\top} \right\| \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}'_t(\boldsymbol{\theta})\|^2 + \sum_{k=1}^{r} \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})}{\partial f_k} \right\|^2 \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|f''_{kt}(\boldsymbol{\theta})\|^2 .
\end{aligned} \tag{SB.9}$$

By Lemma SC.6, the expression above is finite as long as $\mathbb{E}\|\boldsymbol{y}_t\|^2 < \infty$, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}_t(\boldsymbol{\theta})\|^2 < \infty$, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}'_t(\boldsymbol{\theta})\|^2 < \infty$, and $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}''_t(\boldsymbol{\theta})\|^2 < \infty$. $\mathbb{E}\|\boldsymbol{y}_t\|^2 < \infty$ holds by Lemma 3 given Assumption 3 and conditions A and B on the parameter space, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}_t(\boldsymbol{\theta})\|^2$ is bounded by Lemma 4 since $\mathbb{E}\|\boldsymbol{y}_t\|^2 < \infty$ and $\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{B} - \boldsymbol{A}\| < 1$. Finally, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}'_t(\boldsymbol{\theta})\|^2 < \infty$ and $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}''_t(\boldsymbol{\theta})\|^2 < \infty$ follow by Lemma SB.5. This finishes the proof for the Gaussian model.

(ii) Student's t model: Given that by Lemma SB.5 for the Student's t model we only have $\mathbb{E}\|\boldsymbol{f}'_t(\boldsymbol{\theta}_0)\|^2 < \infty$ and $\mathbb{E}\|\boldsymbol{f}''_t(\boldsymbol{\theta}_0)\|^2 < \infty$, and not uniformly over Θ, we take a different approach to prove the convergence of

the hessian. Specifically, we notice that

$$|\nabla_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j}\mathcal{L}_T(\hat{\boldsymbol{\theta}}_T) - \mathcal{I}_{ij}(\boldsymbol{\theta}_0)| \leq |\nabla_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j}\mathcal{L}_T(\boldsymbol{\theta}_0) - \mathcal{I}_{ij}(\boldsymbol{\theta}_0)| + \sup_{\boldsymbol{\theta}\in\Theta} \|\nabla_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta})\| \|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|. \tag{SB.10}$$

By the law of large numbers for SE sequences, $\|\nabla_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j}\mathcal{L}_T(\boldsymbol{\theta}_0) - \mathcal{I}_{ij}(\boldsymbol{\theta}_0)\| \xrightarrow{a.s.} 0$ as $T \to \infty$ since by Krengel's theorem $\{\nabla_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j} l_t(\boldsymbol{\theta}_0)\}$ is SE and $\mathbb{E}|\nabla_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j} l_t(\boldsymbol{\theta}_0)| < \infty$. The latter holds as long as $\mathbb{E}\|\boldsymbol{y}_t\|^2 < \infty$, $\mathbb{E}\|\boldsymbol{f}_t(\boldsymbol{\theta}_0)\|^2 < \infty$, $\mathbb{E}\|\boldsymbol{f}'_t(\boldsymbol{\theta}_0)\|^2 < \infty$, and $\mathbb{E}\|\boldsymbol{f}''_t(\boldsymbol{\theta}_0)\|^2 < \infty$ (see equation SB.9). $\mathbb{E}\|\boldsymbol{y}_t\|^2$ and $\mathbb{E}\|\boldsymbol{f}_t(\boldsymbol{\theta}_0)\|^2$ are bounded by Lemma 3 as long as $\nu > 2$, Assumption 3 and conditions A abd B on the parameter space hold. Furthermore, as long as the conditions of Lemma SB.5 (ii) hold, it ensures that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}'_t(\boldsymbol{\theta}_0)\|^2 < \infty$ and $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}''_t(\boldsymbol{\theta}_0)\|^2 < \infty$.

For the second term in (SB.10), by Theorem 1, we have $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \xrightarrow{a.s.} 0$. Furthermore, given that $\sup_{\boldsymbol{\theta}\in\Theta}\|\nabla_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta})\|$ is SE as it is a continuous function on the SE sequence, it is bounded in probability, i.e. $\sup_{\boldsymbol{\theta}\in\Theta}\|\nabla_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta})\| = O_p(1)$. Therefore, $\sup_{\boldsymbol{\theta}\in\Theta}\|\nabla_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta})\|\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \xrightarrow{P} 0$ as $T \to \infty$. ■

Lemma SB.4 (Derivatives of the filter)**.** *Let all the conditions of Proposition 1 hold. Then, for the Gaussian and Student's t score-driven filters, there exist unique strictly stationary solutions* $\{\boldsymbol{f}'_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$, $\{\boldsymbol{f}''_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$, *and* $\{\boldsymbol{f}'''_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ *to* (SC.11), (SC.12), *and* (SC.13), *respectively, such that*

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}'_t(\boldsymbol{\theta}) - \boldsymbol{f}'_t(\boldsymbol{\theta})\| \xrightarrow{e.a.s.} 0 \quad as \quad t \to \infty,\\
&\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}''_t(\boldsymbol{\theta}) - \boldsymbol{f}''_t(\boldsymbol{\theta})\| \xrightarrow{e.a.s.} 0 \quad as \quad t \to \infty,\\
&\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{\boldsymbol{f}}'''_t(\boldsymbol{\theta}) - \boldsymbol{f}'''_t(\boldsymbol{\theta})\| \xrightarrow{e.a.s.} 0 \quad as \quad t \to \infty.
\end{aligned}$$

Proof. The proof of this lemma is similar to the proof of Proposition 3.4 in Blasques, van Brummelen, Koopman, & Lucas (2022) and of Proposition 2 in Blasques, van Brummelen, Gorgi, & Koopman (2022). As shown in Section C, the SRE for the first derivative of $\boldsymbol{f}_t$ is of the following form

$$\boldsymbol{f}'_{t+1}(\boldsymbol{\theta}) = \boldsymbol{C}_t^{(1)} + \boldsymbol{\Gamma}_t\boldsymbol{f}'_t(\boldsymbol{\theta}),$$

where $\boldsymbol{C}_t^{(1)} = \boldsymbol{C}_t^{(1)}(\boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})$ and $\boldsymbol{\Gamma}_t = \boldsymbol{\Gamma}_t(\boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})$ with explicit expressions for $\boldsymbol{C}_t^{(1)}$ and $\boldsymbol{\Gamma}_t$ in case of the Gaussian and Student's t filters given in Section C . This implies that the filtered sequence $\{\hat{\boldsymbol{f}}'_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ initialized at $\hat{\boldsymbol{f}}'_1$ depends on the filtered sequence $\{\hat{\boldsymbol{f}}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$. In turn, the unperturbed sequence $\{\tilde{\boldsymbol{f}}'_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ initialized at $\hat{\boldsymbol{f}}'_1$ depends on the limit sequence $\{\boldsymbol{f}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$. We denote the limit process as $\{\boldsymbol{f}'_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$.

Hence, to prove this lemma, we use Theorem 2.10 in Straumann & Mikosch (2006) for perturbed stochastic recurrence equations. Condition S.3 of Straumann & Mikosch (2006, Theorem 2.10), the convergence of the perturbed sequence $\{\hat{\boldsymbol{f}}'_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ to the stationary limit sequence $\{\boldsymbol{f}'_t(\boldsymbol{\theta})\}$, corresponds to having

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{C}_t^{(1)}(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}), \boldsymbol{\theta}) - \boldsymbol{C}_t^{(1)}(\boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})\| &\xrightarrow{e.a.s.} 0,\\
\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Gamma}_t(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}), \boldsymbol{\theta}) - \boldsymbol{\Gamma}_t(\boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})\| &\xrightarrow{e.a.s.} 0,
\end{aligned}$$

as $t \to \infty$.

We show this result by using the mean value theorem. Namely, for the first expression, we have

$$\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{C}_t^{(1)}(\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}),\boldsymbol{\theta}) - \boldsymbol{C}_t^{(1)}(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})\| \leq \sup_{\boldsymbol{f}} \sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{\partial \boldsymbol{C}_t^{(1)}}{\partial \boldsymbol{f}} \right\| \sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}) - \boldsymbol{f}_t(\boldsymbol{\theta})\|,$$

and similarly for the second expression. Given that $\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{f}}_t(\boldsymbol{\theta}) - \boldsymbol{f}_t(\boldsymbol{\theta})\| \xrightarrow{e.a.s.} 0$ as $t \to \infty$ and that the derivatives of $\boldsymbol{C}_t^{(1)}$ and $\boldsymbol{\Gamma}_t$ with respect to $\boldsymbol{f}$ are SE, it suffices to show that the derivatives of $\boldsymbol{C}_t^{(1)}$ and $\boldsymbol{\Gamma}_t$ with respect to $\boldsymbol{f}$ are uniformly bounded in $\boldsymbol{f}$ and $\boldsymbol{\theta}$, which is the case given the result in Lemma SC.6. Specifically, given that for both Gaussian and Student's t models $\partial^2 \boldsymbol{s}_t / \partial \boldsymbol{f} \partial \boldsymbol{f}^\top$ is uniformly bounded in both $\boldsymbol{f}$ and $\boldsymbol{\theta}$, it implies the convergence for $\boldsymbol{\Gamma}_t$. Next, the uniform boundedness in $\boldsymbol{f}$ and $\boldsymbol{\theta}$ of $\|\partial s_{k,t}/\partial \boldsymbol{f}^\top\|$ and $\|\partial^2 s_{k,t}/\partial \boldsymbol{\theta}\partial \boldsymbol{f}^\top\|$ for $k = 1, \ldots, r$, implies the convergence for $\boldsymbol{C}_t^{(1)}$. The convergence of the Lipschitz coefficients then follows straightforwardly.

Condition S.1 of Straumann & Mikosch (2006, Theorem 2.10) is fulfilled since $\boldsymbol{\Gamma}_t(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})$ evaluated at the limit sequence $\boldsymbol{f}_t(\boldsymbol{\theta})$ is bounded uniformly over t and $\boldsymbol{\theta}$, and $\boldsymbol{C}_t^{(1)}(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})$ has a bounded logarithmic moment uniformly over Θ since

(i). for the Gaussian model, conditions of Proposition 1 require $\mathbb{E}\log^+\|\boldsymbol{y}_t\| < \infty$ and by Lemma 4 we then also have $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\log^+\|\boldsymbol{f}_t(\boldsymbol{\theta})\| < \infty$, which together with the compactness of the parameter space Θ imply $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\log^+\|\boldsymbol{C}_t^{(1)}(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})\| < \infty$.

(ii). for the Student's t model, $\boldsymbol{C}_t^{(1)}(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})$ is uniformly bounded in t and Θ, hence it has moments of any order uniformly over Θ.

This implies that the unperturbed recurrence equation evaluated at some deterministic point has a bounded logarithmic moment.

The condition S.2 of Straumann & Mikosch (2006, Theorem 2.10) in our case is of the form

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\log^+\|\boldsymbol{\Gamma}_t(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})\| < \infty \quad \text{and} \quad \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\log\|\prod_{j=0}^{p-1}\boldsymbol{\Gamma}_{t-j}(\boldsymbol{f}_{t-j}(\boldsymbol{\theta}),\boldsymbol{\theta})\| < 0.$$

Since $\boldsymbol{\Gamma}_t(\boldsymbol{f},\boldsymbol{\theta})$ is uniformly bounded in $\boldsymbol{f}$ and $\boldsymbol{\theta}$, the first condition trivially holds. Following a similar argument as in the proof of Proposition 2 in Blasques, van Brummelen, Gorgi, & Koopman (2022), we notice that the second condition is implied by $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{f}\in\mathbb{R}^r}\|\partial\phi_t^{(p)}(\boldsymbol{f},\boldsymbol{\theta})/\partial\boldsymbol{f}\| < 0$, which is one of the conditions for the filter invertibility in Proposition 1. The proofs for the second and third order derivatives are similar and are omitted. ■

Lemma SB.5 (Moments of the derivatives of the filter)**.** *Let all the conditions of Lemmas 4 and SB.4 hold. Then, for the limit sequence of the derivative of the filter, we have*

(i). for the Gaussian model, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}_t'(\boldsymbol{\theta})\|^k < \infty$ *and* $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{f}_t''(\boldsymbol{\theta})\|^k < \infty$ *with* k *as defined in Lemmas 4.*

(ii). for the Student's t model, let for some integer $p \geq 1$ *and some* $0 < k < \nu$

$$\mathbb{E}\|\prod_{j=0}^{p-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-j},\boldsymbol{\theta}_0)\|^k < 1$$

with $\boldsymbol{\Gamma}_t$ *as defined in Section C.1. Then,* $\mathbb{E}\|\boldsymbol{f}_t'(\boldsymbol{\theta}_0)\|^k < \infty$ *and* $\mathbb{E}\|\boldsymbol{f}_t''(\boldsymbol{\theta}_0)\|^k < \infty$.

Proof. *(i) Gaussian model:* The stochastic recurrence equation for the derivative process is of the form

$$\boldsymbol{f}'_{t+1}(\boldsymbol{\theta}) = \boldsymbol{C}_t^{(1)} + (\boldsymbol{B} - \boldsymbol{A})\boldsymbol{f}'_t(\boldsymbol{\theta}).$$

Hence, following a similar reasoning as in the proof of Lemma 4, for $k > 1$, we have

$$\|\boldsymbol{f}'_{t+1}(\boldsymbol{\theta})\|_k^{\Theta} \leq 1 + \sum_{j=0}^{m} \|(\boldsymbol{B} - \boldsymbol{A})^j\|^{\Theta} \|\boldsymbol{C}_{t-j}^{(1)}\|_k^{\Theta} \leq 1 + \frac{\|\boldsymbol{C}_t^{(1)}\|_k^{\Theta}}{1 - \|\boldsymbol{B} - \boldsymbol{A}\|^{\Theta}} < \infty,$$

where the second inequality follows since the conditions of Proposition 1 imply that $\sup_{\boldsymbol{\theta}} \|\boldsymbol{B} - \boldsymbol{A}\| < 1$ and since $\|\boldsymbol{C}_{t-j}^{(1)}\|_k^{\Theta} = \|\boldsymbol{C}_t^{(1)}\|_k^{\Theta}$ for every j as it is a function on the SE sequence. The final inequality follows by Lemmas 3 and 4 which imply that $\|\boldsymbol{C}_t^{(1)}\|_k^{\Theta} < \bar{d}_k < \infty$. Hence, we have that $\|\boldsymbol{f}'_{t+1}(\boldsymbol{\theta})\|_k^{\Theta} < \infty$ for $k > 1$. The proof for $0 < k < 1$ follows immediately by the application of the Loève's c_r inequality. The proof for the second order derivative is similar and is omitted.

(ii) Student's t model: The SRE for the derivative of the filter is of the form

$$\boldsymbol{f}'_{t+1}(\boldsymbol{\theta}) = \boldsymbol{C}^{(1)}(\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta}) + \boldsymbol{\Gamma}(\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})\boldsymbol{f}'_t(\boldsymbol{\theta}).$$

And under correct model specification, we have

$$\boldsymbol{f}'_{t+1}(\boldsymbol{\theta}_0) = \boldsymbol{C}^{(1)}(\boldsymbol{\varepsilon}_t, \boldsymbol{f}_t(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0) + \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_t, \boldsymbol{\theta}_0)\boldsymbol{f}'_t(\boldsymbol{\theta}_0),$$

where $\boldsymbol{C}_t^{(1)}(\boldsymbol{\varepsilon}_t, \boldsymbol{f}_t(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0)$ is as defined in Section C.1.

Then, the SRE for the p-th iterate takes the following form

$$\begin{aligned}\boldsymbol{f}'^{(p)}_{t+1}(\boldsymbol{\theta}_0) = &\sum_{j=0}^{p-1} \left(\boldsymbol{C}^{(1)}(\boldsymbol{\varepsilon}_{t-j}, \boldsymbol{f}_{t-j}(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0)\right) \left(\prod_{i=0}^{j-1} \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i}, \boldsymbol{\theta}_0)\right) \\ &+ \left(\prod_{i=0}^{p-1} \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i}, \boldsymbol{\theta}_0)\right) \boldsymbol{f}'^{(p)}_t(\boldsymbol{\theta}_0).\end{aligned}$$

First, by norm subadditivity and submultiplicativity, we have

$$\begin{aligned}\|\boldsymbol{f}'^{(p)}_{t+1}(\boldsymbol{\theta}_0)\| \leq &\sum_{j=0}^{p-1} \left\|\boldsymbol{C}^{(1)}(\boldsymbol{\varepsilon}_{t-j}, \boldsymbol{f}_{t-j}(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0)\right\| \left\|\prod_{i=0}^{j-1} \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i}, \boldsymbol{\theta}_0)\right\| \\ &+ \left\|\prod_{i=0}^{p-1} \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i}, \boldsymbol{\theta}_0)\right\| \left\|\boldsymbol{f}'^{(p)}_t(\boldsymbol{\theta}_0)\right\|.\end{aligned}$$

Iterating backwards the SRE for the pth iterate, we obtain

$$\begin{aligned}\|\boldsymbol{f}'^{(p)}_{t+1}(\boldsymbol{\theta}_0)\| \leq &\left(\sum_{h=0}^{T-1} \left(\sum_{j=0}^{p-1} \|\boldsymbol{C}^{(1)}(\boldsymbol{\varepsilon}_{t-j-ph}, \boldsymbol{f}_{t-j-ph}(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0)\| \left\|\prod_{i=0}^{j-1} \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-ph}, \boldsymbol{\theta}_0)\right\|\right) \right. \\ &\left. \times \prod_{l=0}^{h-1} \left\|\prod_{i=0}^{p-1} \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-pl}, \boldsymbol{\theta}_0)\right\|\right) + \prod_{l=0}^{T-1} \left\|\prod_{i=0}^{p-1} \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-pl}, \boldsymbol{\theta}_0)\right\| \left\|\boldsymbol{f}'^{(p)}_{t-pl}(\boldsymbol{\theta}_0)\right\|.\end{aligned}$$

For any $t \in \mathbb{Z}$, the sequence $\left\{\left\|\prod_{i=0}^{p-1} \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-pl}, \boldsymbol{\theta}_0)\right\|\right\}_{l \in \mathbb{Z}}$ is i.i.d., hence SE, sequence of nonnegative random variables. The conditions of the lemma also imply that $\mathbb{E} \log \|\prod_{j=0}^{p-1} \boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-j}, \boldsymbol{\theta}_0)\| < 1$ (condition

in Proposition 1). Therefore, by Lemma 2.4 in Straumann & Mikosch (2006) we have

$$\prod_{l=0}^{T}\left\|\prod_{i=0}^{p-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-pl},\boldsymbol{\theta}_0)\right\| \xrightarrow{e.a.s.} 0 \quad \text{as } T\to\infty.$$

Since by Lemma SB.4 the sequence $\{\|\boldsymbol{f}_t'^{(p)}(\boldsymbol{\theta}_0)\|\}_{t\in\mathbb{Z}}$ is also SE, there exists large enough l such that

$$\prod_{l=0}^{T-1}\left\|\prod_{i=0}^{p-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-pl},\boldsymbol{\theta}_0)\right\|\left\|\boldsymbol{f}'_{t-pl}{}^{(p)}(\boldsymbol{\theta}_0)\right\|<1, \quad \text{a.s.}.$$

Then, for $k\geq 1$, we have

$$\begin{aligned}
\|\boldsymbol{f}'_{t+1}{}^{(p)}(\boldsymbol{\theta}_0)\|_k \leq & 1+\sum_{h=0}^{T-1}\left\|\left(\sum_{j=0}^{p-1}\left(\|\boldsymbol{C}^{(1)}(\boldsymbol{\varepsilon}_{t-j-ph},\boldsymbol{f}_{t-j-ph}(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)\|\right)\left\|\prod_{i=0}^{j-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-ph},\boldsymbol{\theta}_0)\right\|\right)\right. \\
& \left.\times\prod_{l=0}^{h-1}\left\|\prod_{i=0}^{p-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-pl},\boldsymbol{\theta}_0)\right\|\right\|_k \\
\leq & 1+\sum_{h=0}^{T-1}\left\|\sum_{j=0}^{p-1}\left(\|\boldsymbol{C}^{(1)}(\boldsymbol{\varepsilon}_{t-j-ph},\boldsymbol{f}_{t-j-ph}(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)\|\right)\left\|\prod_{i=0}^{j-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-ph},\boldsymbol{\theta}_0)\right\|\right\|_k \\
& \times\left\|\prod_{l=0}^{h-1}\left\|\prod_{i=0}^{p-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-pl},\boldsymbol{\theta}_0)\right\|\right\|_k \\
\leq & 1+\bar{c}_{k,p}(\boldsymbol{\theta}_0)\sum_{h=0}^{T-1}\left(\left(\mathbb{E}\left\|\prod_{i=0}^{p-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i},\boldsymbol{\theta}_0)\right\|^k\right)^{1/k}\right)^h \\
\leq & 1+\frac{\bar{c}_{k,p}(\boldsymbol{\theta}_0)}{1-\left(\mathbb{E}\left\|\prod_{i=0}^{p-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i},\boldsymbol{\theta}_0)\right\|^k\right)^{1/k}}<\infty,
\end{aligned}$$

where the second inequality follows since $\sum_{j=0}^{p-1}\left(\|\boldsymbol{C}^{(1)}(\boldsymbol{\varepsilon}_{t-j-ph},\boldsymbol{f}_{t-j-ph}(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)\|\right)\left\|\prod_{i=0}^{j-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-ph},\boldsymbol{\theta}_0)\right\|$ and $\prod_{l=0}^{h-1}\left\|\prod_{i=0}^{p-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-pl},\boldsymbol{\theta}_0)\right\|$ are independent; the third inequality follows by Proposition 4.3 in Krengel (1985) since for any $t\in\mathbb{Z}$ $\sum_{j=0}^{p-1}\left(\|\boldsymbol{C}^{(1)}(\boldsymbol{\varepsilon}_{t-j-ph},\boldsymbol{f}_{t-j-ph}(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0)\|\right)\left\|\prod_{i=0}^{j-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-i-ph},\boldsymbol{\theta}_0)\right\|$ is a continuous function on the SE sequence $\{\boldsymbol{f}_t(\boldsymbol{\theta}_0),\ldots,\boldsymbol{f}_{t+1-(h+1)p}(\boldsymbol{\theta}_0),\boldsymbol{\varepsilon}_t,\ldots,\boldsymbol{\varepsilon}_{t+1-(h+1)p}\}_{h\in\mathbb{Z}}$; the forth inequality follows given the conditions of the lemma, i.e. $\mathbb{E}\|\prod_{j=0}^{p-1}\boldsymbol{\Gamma}(\boldsymbol{\varepsilon}_{t-j},\boldsymbol{\theta}_0)\|^k<1$.

For $0<k<1$, the proof follows by Loève's c_k inequality, where the constant $c_k=1$ when $0<k<1$. The proof for the second order derivative is similar and is omitted.

■

C Derivatives

In this section, we provide the expressions for the required derivatives. To simplify further notation, let us further define $\boldsymbol{\sigma}^2=\operatorname{diag}\boldsymbol{\Sigma}$, $\boldsymbol{a}=\operatorname{diag}\boldsymbol{A}$ and $\boldsymbol{b}=\operatorname{diag}\boldsymbol{B}$, so that for the Gaussian model $\boldsymbol{\theta}=(\boldsymbol{\sigma}^{2\top},\operatorname{vec}\boldsymbol{\Lambda}^\top,\boldsymbol{a}^\top,\boldsymbol{b}^\top)^\top$ and for the Student's t model $\boldsymbol{\theta}=(\boldsymbol{\sigma}^{2\top},\operatorname{vec}\boldsymbol{\Lambda}^\top,\boldsymbol{a}^\top,\boldsymbol{b}^\top,\nu)^\top$.

C.1 Derivatives of the time-varying parameter

For the Gaussian and Student's t models we have the following updating equation $\boldsymbol{f}_{t+1} = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{f}_t$, with $\boldsymbol{s}_t = \boldsymbol{s}(\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})$ and $\boldsymbol{f}_t = \boldsymbol{f}_t(\boldsymbol{\theta})$.

As discussed before $\boldsymbol{f}_t$ is an $r \times 1$ vector. Hence, for $k = 1, \ldots, r$

$$f_{k,(t+1)} = \alpha_k s_{k,t} + \beta_k f_{k,t}.$$

Let us introduce the following notation:

$$\begin{aligned}
\boldsymbol{f}'_t(\boldsymbol{\theta}) &:= \begin{bmatrix} f'_{1,t}(\boldsymbol{\theta}) & \ldots & f'_{r,t}(\boldsymbol{\theta}) \end{bmatrix}^\top, \\
\boldsymbol{f}''_t(\boldsymbol{\theta}) &:= \begin{bmatrix} f''_{1,t}(\boldsymbol{\theta}) & \ldots & f''_{r,t}(\boldsymbol{\theta}) \end{bmatrix}^\top,
\end{aligned}$$

where, for $k = 1, \ldots, r$, we have

$$\begin{aligned}
f'_{k,t+1}(\boldsymbol{\theta}) &:= \frac{\partial f_{k,t+1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = C^{(1)}_{k,t} + \boldsymbol{f}'_t(\boldsymbol{\theta})^\top \Gamma_{k,t}, \\
f''_{k,t+1}(\boldsymbol{\theta}) &:= \operatorname{vec}\left(\frac{\partial^2 f_{k,t+1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right) = C^{(2)}_{k,t} + \boldsymbol{f}''_t(\boldsymbol{\theta})^\top \Gamma_{k,t},
\end{aligned}$$

where

$$\begin{aligned}
C^{(1)}_{k,t} &:= C^{(1)}_k(\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta}) = \frac{\partial \alpha_k}{\partial \boldsymbol{\theta}} s_{k,t} + \alpha_k \frac{\partial s_{k,t}}{\partial \boldsymbol{\theta}} + \frac{\partial \beta_k}{\partial \boldsymbol{\theta}} f_{k,t}, \\
\Gamma_{k,t} &:= \Gamma_k(\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta}) = \beta_k \boldsymbol{e}_k + \alpha_k \frac{\partial s_{k,t}}{\partial \boldsymbol{f}}, \\
C^{(2)}_{k,t} &:= C^{(2)}_k(\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta}) \\
&= \operatorname{vec}\left(\frac{\partial C^{(1)}_{k,t}}{\partial \boldsymbol{\theta}^\top} + \frac{\partial C^{(1)}_{k,t}}{\partial \boldsymbol{f}^\top} \boldsymbol{f}'_t(\boldsymbol{\theta}) + \boldsymbol{f}'_t(\boldsymbol{\theta})^\top \left(\frac{\partial \Gamma_{k,t}}{\partial \boldsymbol{\theta}^\top} + \frac{\partial \Gamma_{k,t}}{\partial \boldsymbol{f}^\top} \boldsymbol{f}'_t(\boldsymbol{\theta}) \right) \right) \\
&= \operatorname{vec}\left(\frac{\partial \alpha_k}{\partial \boldsymbol{\theta}} \frac{\partial s_{k,t}}{\partial \boldsymbol{\theta}^\top} + \frac{\partial s_{k,t}}{\partial \boldsymbol{\theta}} \frac{\partial \alpha_k}{\partial \boldsymbol{\theta}^\top} + \alpha_k \frac{\partial^2 s_{k,t}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} + \frac{\partial \beta_k}{\partial \boldsymbol{\theta}} \frac{f_{k,t}}{\partial \boldsymbol{\theta}^\top} \right) \\
&+ \operatorname{vec}\left(\left(\frac{\partial \alpha_k}{\partial \boldsymbol{\theta}} \frac{\partial s_{k,t}}{\partial \boldsymbol{f}^\top} + \alpha_k \frac{\partial^2 s_{k,t}}{\partial \boldsymbol{\theta} \partial \boldsymbol{f}^\top} + \frac{\partial \beta_k}{\partial \boldsymbol{\theta}} \boldsymbol{e}_k^\top \right) \boldsymbol{f}'_t(\boldsymbol{\theta}) \right) \\
&+ \operatorname{vec}\left(\boldsymbol{f}'_t(\boldsymbol{\theta})^\top \left(\alpha_k \frac{\partial^2 s_{k,t}}{\partial \boldsymbol{f} \partial \boldsymbol{f}^\top} \boldsymbol{f}'_t(\boldsymbol{\theta}) + \boldsymbol{e}_k \frac{\partial \beta_k}{\partial \boldsymbol{\theta}^\top} + \frac{\partial s_{k,t}}{\partial \boldsymbol{f}} \frac{\partial \alpha_k}{\partial \boldsymbol{\theta}^\top} + \alpha_k \frac{\partial^2 s_{k,t}}{\partial \boldsymbol{f} \partial \boldsymbol{\theta}^\top} \right) \right),
\end{aligned}$$

where $\boldsymbol{e}_k$ is an $r \times 1$ vector of zeros with the one in the kth position and where we use the notation

$$\begin{aligned}
\frac{\partial s_{k,t}}{\partial \boldsymbol{\theta}} &:= \left. \frac{\partial s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) = (\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})}, \\
\frac{\partial s_{k,t}}{\partial \boldsymbol{f}} &:= \left. \frac{\partial s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{f}} \right|_{(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) = (\boldsymbol{y}_t, \boldsymbol{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta})},
\end{aligned}$$

and similar notations for the other derivatives. Then, the derivatives satisfy the following SREs

$$\boldsymbol{f}'_{t+1}(\boldsymbol{\theta}) = \boldsymbol{C}^{(1)}_t + \boldsymbol{\Gamma}_t \boldsymbol{f}'_t(\boldsymbol{\theta}), \tag{SC.11}$$

$$\boldsymbol{f}''_{t+1}(\boldsymbol{\theta}) = \boldsymbol{C}^{(2)}_t + \boldsymbol{\Gamma}_t \boldsymbol{f}''_t(\boldsymbol{\theta}), \tag{SC.12}$$

where $\boldsymbol{C}^{(i)}_t = (C^{(i)}_{1,t}, \ldots, C^{(i)}_{k,t})^\top$ with $i = \{1, 2\}$. The third derivative can be also represented by a similar

SRE, i.e.

$$\boldsymbol{f}'''_{t+1}(\boldsymbol{\theta}) = \boldsymbol{C}_t^{(3)} + \boldsymbol{\Gamma}_t \boldsymbol{f}'''_t(\boldsymbol{\theta}), \tag{SC.13}$$

with $\boldsymbol{f}'''_t(\boldsymbol{\theta}) := \begin{bmatrix} f'''_{1,t}(\boldsymbol{\theta}) & \dots & f'''_{r,t}(\boldsymbol{\theta}) \end{bmatrix}^\top$, $f'''_{k,t}(\boldsymbol{\theta}) := \operatorname{vec}\left(\frac{\partial f''_{k,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right)$.

C.2 Derivatives of the score

In this section, we provide explicit expressions for the derivatives of the score for the specific cases of Gaussian and Student's t models.

Gaussian model. The score for this model is of the following form

$$\boldsymbol{s}(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) = \frac{1}{N} \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y} - \boldsymbol{f},$$

hence, $s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) = \frac{1}{N} \boldsymbol{\Lambda}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y} - f_k$.

Clearly, $\partial \boldsymbol{s}(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) / \partial \boldsymbol{f}^\top = -\boldsymbol{I}_r$, $\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{f}^\top = \boldsymbol{0}_{q \times r}$ and $\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) / \partial \boldsymbol{f} \partial \boldsymbol{f}^\top = \boldsymbol{0}_{r \times r}$.

Furthermore, we have

$$\frac{\partial s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2} \\ \frac{\partial s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \operatorname{vec} \boldsymbol{\Lambda}} \\ \frac{\partial s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{a}} \\ \frac{\partial s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{N} \operatorname{diag}(\boldsymbol{y}) \boldsymbol{\Sigma}^{-2} \boldsymbol{\Lambda}_k \\ \boldsymbol{e}_k \otimes \left(\frac{1}{N} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}\right) \\ \boldsymbol{0}_r \\ \boldsymbol{0}_r \end{bmatrix},$$

$$\frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \begin{bmatrix} \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2 \partial \boldsymbol{\sigma}^{2\top}} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2 \partial \operatorname{vec} \boldsymbol{\Lambda}^\top} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2 \partial \boldsymbol{a}^\top} & \frac{\partial s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2 \partial \boldsymbol{b}^\top} \\ \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \operatorname{vec} \boldsymbol{\Lambda} \partial \boldsymbol{\sigma}^{2\top}} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \operatorname{vec} \boldsymbol{\Lambda} \partial \operatorname{vec} \boldsymbol{\Lambda}^\top} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \operatorname{vec} \boldsymbol{\Lambda} \partial \boldsymbol{a}^\top} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \operatorname{vec} \boldsymbol{\Lambda} \partial \boldsymbol{b}^\top} \\ \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{a} \partial \boldsymbol{\sigma}^{2\top}} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{a} \partial \operatorname{vec} \boldsymbol{\Lambda}^\top} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\top} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{a} \partial \boldsymbol{b}^\top} \\ \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{\sigma}^{2\top}} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \operatorname{vec} \boldsymbol{\Lambda}^\top} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{a}^\top} & \frac{\partial^2 s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{b}^\top} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2}{N} \operatorname{diag}(\boldsymbol{y} \boldsymbol{\Sigma}^{-3}) \operatorname{diag}(\boldsymbol{\Lambda}_k) & -\left(\boldsymbol{e}_k \otimes \frac{1}{N} \boldsymbol{\Sigma}^{-2} \operatorname{diag}(\boldsymbol{y})\right)^\top & \boldsymbol{0}_{N \times r} & \boldsymbol{0}_{N \times r} \\ -\boldsymbol{e}_k \otimes \frac{1}{N} \boldsymbol{\Sigma}^{-2} \operatorname{diag}(\boldsymbol{y}) & \boldsymbol{0}_{Nr \times Nr} & \boldsymbol{0}_{Nr \times r} & \boldsymbol{0}_{Nr \times r} \\ \boldsymbol{0}_{r \times N} & \boldsymbol{0}_{r \times Nr} & \boldsymbol{0}_{r \times r} & \boldsymbol{0}_{r \times r} \\ \boldsymbol{0}_{r \times N} & \boldsymbol{0}_{r \times Nr} & \boldsymbol{0}_{r \times r} & \boldsymbol{0}_{r \times r} \end{bmatrix},$$

where $\boldsymbol{e}_k$ is an $r \times 1$ vector of zeros with a one in the kth position and operator $\operatorname{diag}(\cdot)$ creates an $N \times N$ diagonal matrix out of an $N \times 1$ vector with the elements of the vector on the diagonal.

Student's t model. The score for the Student's t model is of the form

$$\boldsymbol{s}(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) = \frac{N + \nu + 2}{\nu} \frac{1}{1 + (\boldsymbol{y} - \boldsymbol{\Lambda} \boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\Lambda} \boldsymbol{f}) / \nu} \left(\frac{1}{N} \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y} - \boldsymbol{f}\right),$$

hence, $s_k(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) = \frac{N+\nu+2}{\nu} \frac{1}{K(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})} \left(\frac{1}{N} \boldsymbol{\Lambda}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y} - f_k\right)$ with $K(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) = 1 + (\boldsymbol{y} - \boldsymbol{\Lambda} \boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\Lambda} \boldsymbol{f}) / \nu$.

Then

$$\frac{\partial \boldsymbol{s}(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} = -\frac{N + \nu + 2}{\nu} \frac{1}{K(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})} \boldsymbol{I}_r - \frac{1}{K(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})} \boldsymbol{s}(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta}) \frac{\partial K(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})}{\partial \boldsymbol{f}^\top}.$$

Similar to the case of Gaussian model, we use the following notation

$$\frac{\partial s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\frac{\partial s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial {\boldsymbol{\sigma}^2}^\top}, \frac{\partial s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \operatorname{vec} \boldsymbol{\Lambda}^\top}, \frac{\partial s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{a}^\top}, \frac{\partial s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{b}^\top}, \frac{\partial s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \nu}\right]^\top.$$

We have

$$\frac{\partial s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}))^{-1} \begin{bmatrix} -\frac{N+\nu+2}{\nu}\frac{\operatorname{diag}(\boldsymbol{y})\boldsymbol{\Sigma}^{-2}\boldsymbol{\Lambda}_k}{N} \\ \frac{N+\nu+2}{\nu}\boldsymbol{e}_k \otimes \left(\frac{1}{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right) \\ \mathbf{0}_r \\ \mathbf{0}_r \\ -\frac{N+2}{\nu^2}\left(\frac{1}{N}\boldsymbol{\Lambda}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - f_k\right) \end{bmatrix} - \frac{1}{K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})} s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

where $\boldsymbol{e}_k$ is an $r \times 1$ vector of zeros with a one in the kth position and operator $\operatorname{diag}(\cdot)$ creates an $N \times N$ diagonal matrix out of an $N \times 1$ vector with the elements of the vector on the diagonal.

To simplify further derivations, we introduce

$$\boldsymbol{\pi}_k(\boldsymbol{y}, f_k, \boldsymbol{\theta}) := \left[-\frac{N+\nu+2}{\nu}\frac{\operatorname{diag}(\boldsymbol{y})\boldsymbol{\Sigma}^{-2}\boldsymbol{\Lambda}_k}{N}, \quad \frac{N+\nu+2}{\nu}\boldsymbol{e}_k \otimes \left(\frac{1}{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right), \quad \mathbf{0}_r, \quad \mathbf{0}_r, \quad -\left(\frac{1}{N}\boldsymbol{\Lambda}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - f_k\right)\frac{N+2}{\nu^2}\right]^\top.$$

Then, we have

$$\begin{aligned}\frac{\partial^2 s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{f}^\top} &= -(K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}))^{-2}\boldsymbol{\pi}_k(\boldsymbol{y}, f_k, \boldsymbol{\theta})\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} + \frac{N+2}{\nu^2}(K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}))^{-1}\boldsymbol{v}_q\boldsymbol{e}_k^\top \\ &+ \frac{1}{K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})} s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})\left(\frac{2}{K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} - \frac{\partial^2 K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{f}^\top}\right) \\ &+ \frac{N+\nu+2}{\nu(K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}))^2}\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{e}_k^\top,\end{aligned}$$

where $\boldsymbol{v}_q$ is a $q \times 1$ vector of zeros with a one in the qth position.

$$\begin{aligned}\frac{\partial^2 s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} &= (K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}))^{-1}\boldsymbol{\Pi}_k(\boldsymbol{y}, f_k, \boldsymbol{\theta}) - (K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}))^{-2}\boldsymbol{\pi}_k(\boldsymbol{y}, f_k, \boldsymbol{\theta})\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \\ &+ \frac{1}{(K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}))^2} s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \\ &- \frac{1}{K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})} s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})\frac{\partial^2 K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} - \frac{1}{K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top},\end{aligned}$$

with

$$\boldsymbol{\Pi}_k(\boldsymbol{y}, f_k, \boldsymbol{\theta}) = \begin{bmatrix} \frac{2(N+\nu+2)}{N\nu}\operatorname{diag}(\boldsymbol{y}\boldsymbol{\Sigma}^{-3})\operatorname{diag}(\boldsymbol{\Lambda}_k) & -\frac{N+\nu+2}{N\nu}(\boldsymbol{e}_k \otimes \boldsymbol{\Sigma}^{-2}\operatorname{diag}(\boldsymbol{y}))^\top & \mathbf{0}_{N\times r} & \mathbf{0}_{N\times r} & \frac{N+2}{N\nu^2}\operatorname{diag}(\boldsymbol{y})\boldsymbol{\Sigma}^{-2}\boldsymbol{\Lambda}_k \\ -\frac{N+\nu+2}{\nu}\boldsymbol{e}_k \otimes \left(\frac{1}{N}\boldsymbol{\Sigma}^{-2}\operatorname{diag}(\boldsymbol{y})\right) & \mathbf{0}_{Nr\times Nr} & \mathbf{0}_{Nr\times r} & \mathbf{0}_{Nr\times r} & -\frac{N+2}{\nu^2}\boldsymbol{e}_k \otimes \left(\frac{1}{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right) \\ \mathbf{0}_{r\times N} & \mathbf{0}_{r\times Nr} & \mathbf{0}_{r\times r} & \mathbf{0}_{r\times r} & \mathbf{0}_r \\ \mathbf{0}_{r\times N} & \mathbf{0}_{r\times Nr} & \mathbf{0}_{r\times r} & \mathbf{0}_{r\times r} & \mathbf{0}_r \\ \frac{N+2}{N\nu^2}(\operatorname{diag}(\boldsymbol{y})\boldsymbol{\Sigma}^{-2}\boldsymbol{\Lambda}_k)^\top & -\frac{N+2}{N\nu^2}(\boldsymbol{e}_k \otimes (\frac{1}{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}))^\top & \mathbf{0}_{1\times r} & \mathbf{0}_{1\times r} & \left(\frac{1}{N}\boldsymbol{\Lambda}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - f_k\right)\frac{2(N+2)}{\nu^3} \end{bmatrix}.$$

$$\frac{\partial^2 s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \boldsymbol{f}^\top} = \frac{N+\nu+2}{\nu}\frac{1}{(K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}))^2}\boldsymbol{e}_k\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} - \frac{1}{K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})\frac{\partial^2 K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \boldsymbol{f}^\top}$$
$$+ \frac{1}{(K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}))^2}s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}}\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} - \frac{1}{K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}}\frac{s_k(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}^\top}.$$

Now we turn to the derivatives of $K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})$.

$$\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} = -2(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}/\nu, \quad \frac{\partial^2 K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \boldsymbol{f}^\top} = -\frac{2N}{\nu}\boldsymbol{I}_r,$$
$$\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} -(\operatorname{diag}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})\boldsymbol{\Sigma}^{-1})^2\boldsymbol{\iota}_N/\nu & -2\boldsymbol{f}\otimes\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})/\nu & \boldsymbol{0}_r & \boldsymbol{0}_r & -\frac{(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})}{\nu^2}\end{bmatrix}^\top.$$

$$\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}$$
$$= \begin{bmatrix} 2(\operatorname{diag}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f}))^2\boldsymbol{\Sigma}^{-3}/\nu & (2\boldsymbol{f}\otimes\boldsymbol{\Sigma}^{-2}\operatorname{diag}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})/\nu)^\top & \boldsymbol{0}_{N\times r} & \boldsymbol{0}_{N\times r} & (\operatorname{diag}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})\boldsymbol{\Sigma}^{-1})^2\boldsymbol{\iota}_N/\nu^2 \\ 2\boldsymbol{f}\otimes\boldsymbol{\Sigma}^{-2}\operatorname{diag}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})/\nu & 2(\boldsymbol{f}\boldsymbol{f}^\top)\otimes\operatorname{diag}(\boldsymbol{\Sigma}^{-1})/\nu & \boldsymbol{0}_{Nr\times r} & \boldsymbol{0}_{Nr\times r} & 2\boldsymbol{f}\otimes\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})/\nu^2 \\ \boldsymbol{0}_{r\times N} & \boldsymbol{0}_{r\times Nr} & \boldsymbol{0}_{r\times r} & \boldsymbol{0}_{r\times r} & \boldsymbol{0}_r \\ \boldsymbol{0}_{r\times N} & \boldsymbol{0}_{r\times Nr} & \boldsymbol{0}_{r\times r} & \boldsymbol{0}_{r\times r} & \boldsymbol{0}_r \\ \boldsymbol{\iota}_N^\top(\operatorname{diag}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})\boldsymbol{\Sigma}^{-1})^2/\nu^2 & 2\boldsymbol{f}^\top\otimes\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})/\nu^2 & \boldsymbol{0}_r^\top & \boldsymbol{0}_r^\top & 2\frac{(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})}{\nu^3}\end{bmatrix}.$$

$$\frac{\partial K(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{f}^\top} = \begin{bmatrix} 2\operatorname{diag}(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})\boldsymbol{\Sigma}^{-2}\boldsymbol{\Lambda}/\nu & 2(\operatorname{diag}(\boldsymbol{f})+\boldsymbol{f}\boldsymbol{\iota}_r^\top)\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}/\nu & \boldsymbol{0}_r & \boldsymbol{0}_r & 2\frac{(\boldsymbol{y}-\boldsymbol{\Lambda}\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}}{\nu^2}\end{bmatrix}.$$

C.3 Derivatives of log-likelihood

Here, we provide the expressions for the first and second derivatives of the log-likelihood function $\mathcal{L}_T(\boldsymbol{\theta})$ defined in (16) with respect to $\boldsymbol{\theta} = (\boldsymbol{\sigma}^2, \operatorname{vec}\boldsymbol{\Lambda}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\nu})$, where $\boldsymbol{a} := \operatorname{diag}\boldsymbol{A}$, $\boldsymbol{b} := \operatorname{diag}\boldsymbol{B}$ and $\boldsymbol{\sigma}^2 := \operatorname{diag}\boldsymbol{\Sigma}$.[2]

For the first derivative, we have $\nabla_{\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta}) := \frac{\partial \mathcal{L}_T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{T}\sum_{t=2}^T l_t'(\boldsymbol{\theta})$, where

$$l_t'(\boldsymbol{\theta}) := \frac{\partial l_t(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{f}=\boldsymbol{f}_t(\boldsymbol{\theta})} + \left(\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}^\top}\Big|_{\boldsymbol{f}=\boldsymbol{f}_t(\boldsymbol{\theta})}\frac{\partial \boldsymbol{f}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right)^\top, \tag{SC.14}$$

where

$$\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix}\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2} & \frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \operatorname{vec}\boldsymbol{\Lambda}} & \boldsymbol{0}_r & \boldsymbol{0}_r & \frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \nu}\end{bmatrix}^\top,$$

with

$$\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2} = -\frac{1}{2}\operatorname{diag}\boldsymbol{\Sigma}^{-1} + \frac{\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{2}\boldsymbol{\Sigma}^{-2}(\operatorname{diag}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f}))^2\boldsymbol{\iota}_N,$$
$$\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \operatorname{vec}\boldsymbol{\Lambda}} = \kappa_t(\boldsymbol{f},\boldsymbol{\theta})\boldsymbol{f}\otimes(\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda}\boldsymbol{f})),$$

where for the Gaussian model $\kappa_t = 1$ and for the Student's model $\kappa_t(\boldsymbol{f},\boldsymbol{\theta}) = \frac{N+\nu}{\nu}\frac{1}{K_t(\boldsymbol{f},\boldsymbol{\theta})}$ with $K_t(\boldsymbol{f},\boldsymbol{\theta})$ as defined in Section C.2; $\boldsymbol{\iota}_N$ is an N-dimensional vector of ones.

[2] We define as $\operatorname{diag}(\boldsymbol{A})$ the $r\times 1$ vector holding the diagonal elements of matrix $\boldsymbol{A}$.

In case of the Gaussian model, $\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \nu} = 0$, while for the Student's t model

$$\begin{aligned}\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \nu} &= \frac{1}{2}\psi\left(\frac{\nu+N}{2}\right) - \frac{1}{2}\psi\left(\frac{\nu}{2}\right) \\ &\quad - \frac{N}{2\nu} - \frac{1}{2}\log K_t(\boldsymbol{f},\boldsymbol{\theta}) + \frac{1}{2}\kappa_t(\boldsymbol{f},\boldsymbol{\theta})K_t(\boldsymbol{f},\nu),\end{aligned}$$

where $\psi(\cdot)$ is a digamma function.

The derivatives of $\boldsymbol{f}_t$ are presented in Section C.1, while for $\partial l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial \boldsymbol{f}^\top$ we have

$$\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}^\top} = \kappa_t(\boldsymbol{f},\boldsymbol{\theta})(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}.$$

For the second-order derivatives we have

$$\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}_T(\boldsymbol{\theta}) := \frac{\partial^2 \mathcal{L}_T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} = \frac{1}{T}\sum_{t=2}^{T} l_t''(\boldsymbol{\theta}),$$

where

$$\begin{aligned}l_t''(\boldsymbol{\theta}) &:= \frac{\partial^2 l_t(\boldsymbol{f}_t(\boldsymbol{\theta}),\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} = \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\Big|_{\boldsymbol{f}=\boldsymbol{f}_t(\boldsymbol{\theta})} + \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{f}^\top}\Big|_{\boldsymbol{f}=\boldsymbol{f}_t(\boldsymbol{\theta})}\frac{\partial \boldsymbol{f}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \\ &+ \frac{\partial \boldsymbol{f}_t(\boldsymbol{\theta})^\top}{\partial \boldsymbol{\theta}}\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \boldsymbol{\theta}^\top}\Big|_{\boldsymbol{f}=\boldsymbol{f}_t(\boldsymbol{\theta})} + \frac{\partial \boldsymbol{f}_t(\boldsymbol{\theta})^\top}{\partial \boldsymbol{\theta}}\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \boldsymbol{f}^\top}\Big|_{\boldsymbol{f}=\boldsymbol{f}_t(\boldsymbol{\theta})}\frac{\partial \boldsymbol{f}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \\ &+ \sum_{k=1}^{r}\frac{\partial l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial f_k}\Big|_{\boldsymbol{f}=\boldsymbol{f}_t(\boldsymbol{\theta})}\frac{\partial^2 f_{kt}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}.\end{aligned}$$

$$\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} = \begin{bmatrix} \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2 \partial \boldsymbol{\sigma}^{2\top}} & \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2 \partial \operatorname{vec}\boldsymbol{\Lambda}^\top} & \mathbf{0}_{N,r} & \mathbf{0}_{N,r} & \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2 \partial \nu} \\ \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \operatorname{vec}\boldsymbol{\Lambda} \partial \boldsymbol{\sigma}^{2\top}} & \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \operatorname{vec}\boldsymbol{\Lambda} \partial \operatorname{vec}\boldsymbol{\Lambda}^\top} & \mathbf{0}_{Nr,r} & \mathbf{0}_{Nr,r} & \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \operatorname{vec}\boldsymbol{\Lambda} \partial \nu} \\ \mathbf{0}_{r,N} & \mathbf{0}_{r,Nr} & \mathbf{0}_{r,r} & \mathbf{0}_{r,r} & \mathbf{0}_r \\ \mathbf{0}_{r,N} & \mathbf{0}_{r,Nr} & \mathbf{0}_{r,r} & \mathbf{0}_{r,r} & \mathbf{0}_r \\ \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \nu \partial \boldsymbol{\sigma}^{2\top}} & \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \nu \partial \operatorname{vec}\boldsymbol{\Lambda}^\top} & \mathbf{0}_r^\top & \mathbf{0}_r^\top & \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \nu^2} \end{bmatrix},$$

$$\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \boldsymbol{\theta}^\top} = \begin{bmatrix} \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \boldsymbol{\sigma}^{2\top}} & \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \operatorname{vec}\boldsymbol{\Lambda}^\top} & \mathbf{0}_r & \mathbf{0}_r & \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \nu} \end{bmatrix}.$$

For the Student's t model

$$\begin{aligned}\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{f}\partial \boldsymbol{\theta}^\top} &= \kappa_t(\boldsymbol{f},\boldsymbol{\theta})\begin{bmatrix} -\boldsymbol{\Lambda}^\top \operatorname{diag}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})\boldsymbol{\Sigma}^{-2} & \boldsymbol{\iota}_r^\top \otimes \boldsymbol{y}_t^\top \boldsymbol{\Sigma}^{-1} & \mathbf{0}_{r\times r} & \mathbf{0}_{r\times r} & 0 \end{bmatrix} \\ &\quad - \frac{N+\nu}{\nu}\frac{1}{K_t(\boldsymbol{f},\boldsymbol{\theta})^2}\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t)\frac{\partial K_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2 \partial \boldsymbol{\sigma}^{2\top}} &= \frac{1}{2}\boldsymbol{\Sigma}^{-2} - \kappa_t(\boldsymbol{f},\boldsymbol{\theta})\boldsymbol{\Sigma}^{-3}(\operatorname{diag}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}))^2 \\ &\quad + \frac{1}{2}\frac{\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{\nu K_t(\boldsymbol{f},\boldsymbol{\theta})}\boldsymbol{\Sigma}^{-2}(\operatorname{diag}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}))^2 \boldsymbol{\iota}_N (\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f})^\top \boldsymbol{\Sigma}^{-2}, \\ \frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}^2 \partial \operatorname{vec}\boldsymbol{\Lambda}^\top} &= -\kappa_t(\boldsymbol{f},\boldsymbol{\theta})\boldsymbol{f} \otimes (\boldsymbol{\Sigma}^{-2}\operatorname{diag}(\boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{f}))\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{\nu K_t(\boldsymbol{f},\boldsymbol{\theta})}\boldsymbol{\Sigma}^{-2}(\operatorname{diag}(\boldsymbol{y}_t-\boldsymbol{\Lambda f}))^2\boldsymbol{\iota}_N(\boldsymbol{f}^\top\otimes(\boldsymbol{y}_t-\boldsymbol{\Lambda f})^\top\boldsymbol{\Sigma}^{-1}),\\
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{\sigma}^2\partial\nu} &= \frac{1}{2}\boldsymbol{\Sigma}^{-2}(\operatorname{diag}(\boldsymbol{y}_t-\boldsymbol{\Lambda f}))^2\boldsymbol{\iota}_N\frac{\partial\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\nu},\\
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\operatorname{vec}\boldsymbol{\Lambda}\partial\nu} &= \boldsymbol{f}\otimes(\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda f}))\frac{\partial\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\nu},\\
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\operatorname{vec}\boldsymbol{\Lambda}\partial\operatorname{vec}\boldsymbol{\Lambda}^\top} &= -\kappa_t(\boldsymbol{f},\boldsymbol{\theta})(\boldsymbol{f}\boldsymbol{f}^\top)\otimes\boldsymbol{\Sigma}^{-1}+\frac{\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{\nu K_t(\boldsymbol{f},\boldsymbol{\theta})}\boldsymbol{f}\otimes(\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda f}))(\boldsymbol{f}^\top\otimes(\boldsymbol{y}_t-\boldsymbol{\Lambda f})^\top\boldsymbol{\Sigma}^{-1}),\\
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\nu^2} &= \frac{1}{4}\psi^{(1)}\left(\frac{N+\nu}{2}\right)-\frac{1}{4}\psi^{(1)}\left(\frac{\nu}{2}\right)+\frac{N}{2\nu^2}+\frac{1}{2\nu^2}\frac{1}{K_t(\boldsymbol{f},\boldsymbol{\theta})}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})\\
&\quad-\frac{1}{2\nu^2}\kappa_t(\boldsymbol{f},\boldsymbol{\theta})(\boldsymbol{y}_t-\boldsymbol{\Lambda f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})+\frac{1}{2}K_t(\boldsymbol{f},\boldsymbol{\theta})\frac{\partial\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\nu}\\
&= \frac{1}{4}\psi^{(1)}\left(\frac{N+\nu}{2}\right)-\frac{1}{4}\psi^{(1)}\left(\frac{\nu}{2}\right)+\frac{N}{2\nu^2}-\frac{N}{2\nu^3}\frac{1}{K_t(\boldsymbol{f},\boldsymbol{\theta})}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})\\
&\quad+\frac{1}{2}K_t(\boldsymbol{f},\boldsymbol{\theta})\frac{\partial\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\nu},
\end{aligned}
$$

where $\frac{\partial\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\nu}=\frac{N+\nu}{\nu^3}\frac{1}{K_t(\boldsymbol{f},\boldsymbol{\theta})^2}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})-\frac{N}{\nu^2}\frac{1}{K_t(\boldsymbol{f},\boldsymbol{\theta})}$ and $\psi^{(1)}(\cdot)$ is a polygamma function of order 1.

$$
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{f}\partial\boldsymbol{f}^\top}=-N\kappa_t(\boldsymbol{f},\boldsymbol{\theta})+\frac{2\kappa_t(\boldsymbol{f},\boldsymbol{\theta})}{\nu K_t(\boldsymbol{f},\boldsymbol{\theta})}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})(\boldsymbol{y}_t-\boldsymbol{\Lambda f})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}
$$

For the Gaussian model

$$
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{f}\partial\boldsymbol{\theta}^\top}=\begin{bmatrix}-\boldsymbol{\Lambda}^\top\operatorname{diag}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})\boldsymbol{\Sigma}^{-2} & \boldsymbol{\iota}_r^\top\otimes\boldsymbol{y}_t^\top\boldsymbol{\Sigma}^{-1} & \mathbf{0}_{r\times r} & \mathbf{0}_{r\times r} & 0\end{bmatrix},
$$

$$
\begin{aligned}
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{\sigma}^2\partial{\boldsymbol{\sigma}^2}^\top} &= \frac{1}{2}\boldsymbol{\Sigma}^{-2}-\boldsymbol{\Sigma}^{-3}(\operatorname{diag}(\boldsymbol{y}_t-\boldsymbol{\Lambda f}))^2,\\
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{\sigma}^2\partial\operatorname{vec}\boldsymbol{\Lambda}^\top} &= -\boldsymbol{f}\otimes(\boldsymbol{\Sigma}^{-2}\operatorname{diag}(\boldsymbol{y}_t-\boldsymbol{\Lambda f})),\\
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\operatorname{vec}\boldsymbol{\Lambda}\partial\operatorname{vec}\boldsymbol{\Lambda}^\top} &= -(\boldsymbol{f}\boldsymbol{f}^\top)\otimes\boldsymbol{\Sigma}^{-1},\\
\frac{\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})}{\partial\boldsymbol{f}\partial\boldsymbol{f}^\top} &= -N.
\end{aligned}
$$

The following lemma states the bounds for each of the derivative.

Lemma SC.6. *The derivatives of the log-likelihood and the score can be bounded as follows:*
For the Student's t model:

- $\sup_{t,\boldsymbol{\theta}}\|\partial l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\boldsymbol{\sigma}^2\|\le c_1$;
- $\sup_{t,\boldsymbol{\theta}}\|\partial l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\operatorname{vec}\boldsymbol{\Lambda}\|\le c_1$;
- $\sup_{\boldsymbol{\theta}}\|\partial l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\nu\|\le c_1+c_2\|\boldsymbol{y}_t\|^2+c_3\|\boldsymbol{f}\|^2$;
- $\sup_{t,\boldsymbol{\theta}}\|\partial l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\boldsymbol{f}\|\le c_1$;
- $\sup_{t,\boldsymbol{\theta}}\|\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\boldsymbol{f}\partial\boldsymbol{f}^\top\|\le c_1$;
- $\sup_{t,\boldsymbol{\theta}}\|\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\boldsymbol{f}\partial{\boldsymbol{\sigma}^2}^\top\|\le c_1$;

- $\sup_{t,\boldsymbol{\theta}} \|\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial \boldsymbol{f} \partial \mathrm{vec}\, \boldsymbol{\Lambda}^\top\| \leq c_1$;
- $\sup_{t,\boldsymbol{\theta}} \|\partial \boldsymbol{s}_t/\partial \boldsymbol{f}\| \leq c_1$;
- $\sup_{t,\boldsymbol{\theta}} \|\partial^2 \boldsymbol{s}_t/\partial \boldsymbol{\theta} \partial \boldsymbol{f}^\top\| \leq c_1$;
- $\sup_{t,\boldsymbol{\theta}} \|\partial^2 \boldsymbol{s}_t/\partial \boldsymbol{f} \partial \boldsymbol{f}^\top\| \leq c_1$.

For the Gaussian model:

- $\sup_{t,\boldsymbol{\theta}} \|\partial \boldsymbol{s}_t/\partial \boldsymbol{f}\| \leq c_1$;
- $\sup_{t,\boldsymbol{\theta}} \|\partial^2 \boldsymbol{s}_t/\partial \boldsymbol{\theta} \partial \boldsymbol{f}^\top\| = 0$;
- $\sup_{t,\boldsymbol{\theta}} \|\partial^2 \boldsymbol{s}_t/\partial \boldsymbol{f} \partial \boldsymbol{f}^\top\| = 0$;
- $\sup_{\boldsymbol{\theta}} \|\partial \boldsymbol{s}_t/\partial \boldsymbol{\theta}\| \leq c_1 + c_2\|\boldsymbol{y}_t\|$;
- $\sup_{\boldsymbol{\theta}} \|\partial^2 \boldsymbol{s}_t/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top\| \leq c_1 + c_2\|\boldsymbol{y}_t\|$;
- $\sup_{\boldsymbol{\theta}} \|\partial l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial \boldsymbol{\sigma}^2\| \leq c_1 + c_2\|\boldsymbol{y}_t\|^2 + c_3\|\boldsymbol{f}\|^2$;
- $\sup_{\boldsymbol{\theta}} \|\partial l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\, \mathrm{vec}\, \boldsymbol{\Lambda}\| \leq c_1\|\boldsymbol{y}_t \boldsymbol{f}^\top\| + c_2\|\boldsymbol{f}\|^2$;
- $\sup_{\boldsymbol{\theta}} \|\partial l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial \boldsymbol{f}\| \leq c_1\|\boldsymbol{y}_t\| + c_2\|\boldsymbol{f}\|$;
- $\sup_{\boldsymbol{\theta}} \|\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial \boldsymbol{f} \partial {\boldsymbol{\sigma}^2}^\top\| \leq c_1\|\boldsymbol{y}_t\| + c_2\|\boldsymbol{f}\|$;
- $\sup_{\boldsymbol{f},\boldsymbol{\theta}} \|\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial \boldsymbol{f} \partial \mathrm{vec}\, \boldsymbol{\Lambda}^\top\| \leq c_1\|\boldsymbol{y}_t\|$;
- $\sup_{\boldsymbol{\theta}} \|\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial \boldsymbol{\sigma}^2 \partial {\boldsymbol{\sigma}^2}^\top\| \leq c_1 + c_2\|\boldsymbol{y}_t\|^2 + c_3\|\boldsymbol{f}\|^2$;
- $\sup_{\boldsymbol{\theta}} \|\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\, \mathrm{vec}\, \boldsymbol{\Lambda} \partial {\boldsymbol{\sigma}^2}^\top\| \leq c_1\|\boldsymbol{y}_t \boldsymbol{f}^\top\| + c_2\|\boldsymbol{f}\|^2$;
- $\sup_{\boldsymbol{\theta}} \|\partial^2 l_t(\boldsymbol{f},\boldsymbol{\theta})/\partial\, \mathrm{vec}\, \boldsymbol{\Lambda} \partial \mathrm{vec}\, \boldsymbol{\Lambda}^\top\| \leq c_1\|\boldsymbol{f}\|^2$.

The derivations are straightforward and can be checked using e.g. Mathematica.

D Supplementary Monte Carlo simulation results

In this section, we present additional Monte Carlo results. Specifically, we demonstrate the results for the experiments with $N = 20$ and different values of T and r. For the detailed description of the simulation design we refer the reader to Section 4.1.

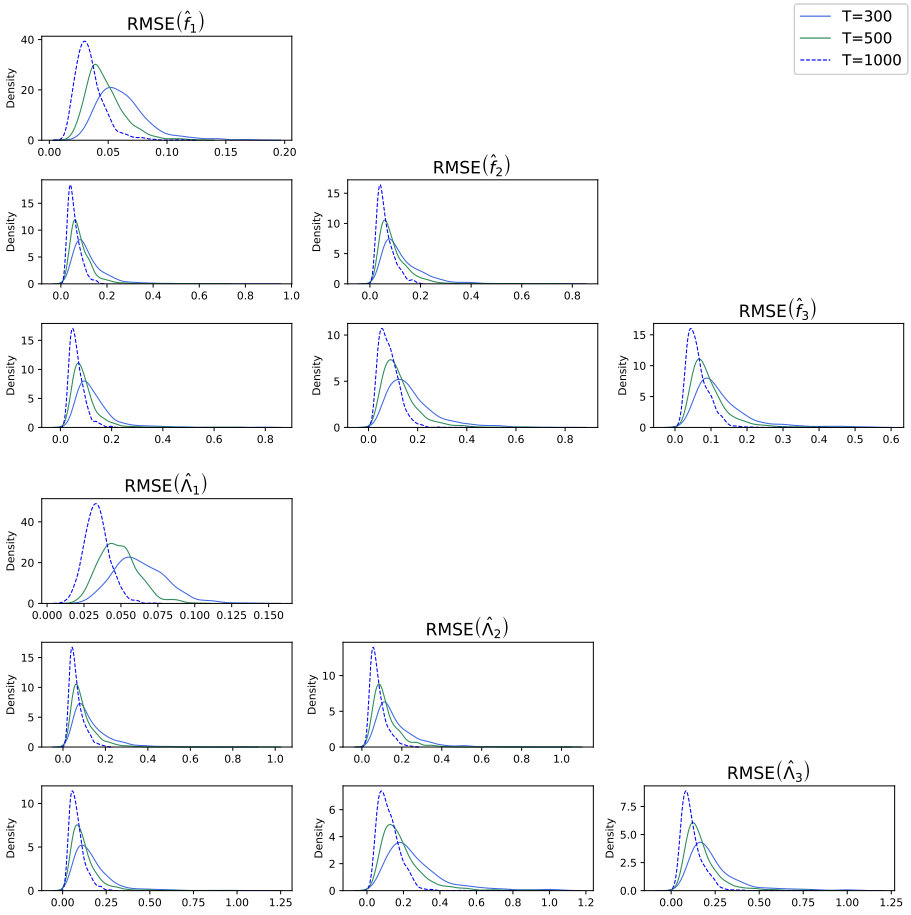


Figure SD.1: **Kernel density of the RMSE for the factors and loadings.** Monte Carlo simulation results. The DGP is a Student's t score-driven factor model with $N = 10$. For further details we refer to Figure 1.

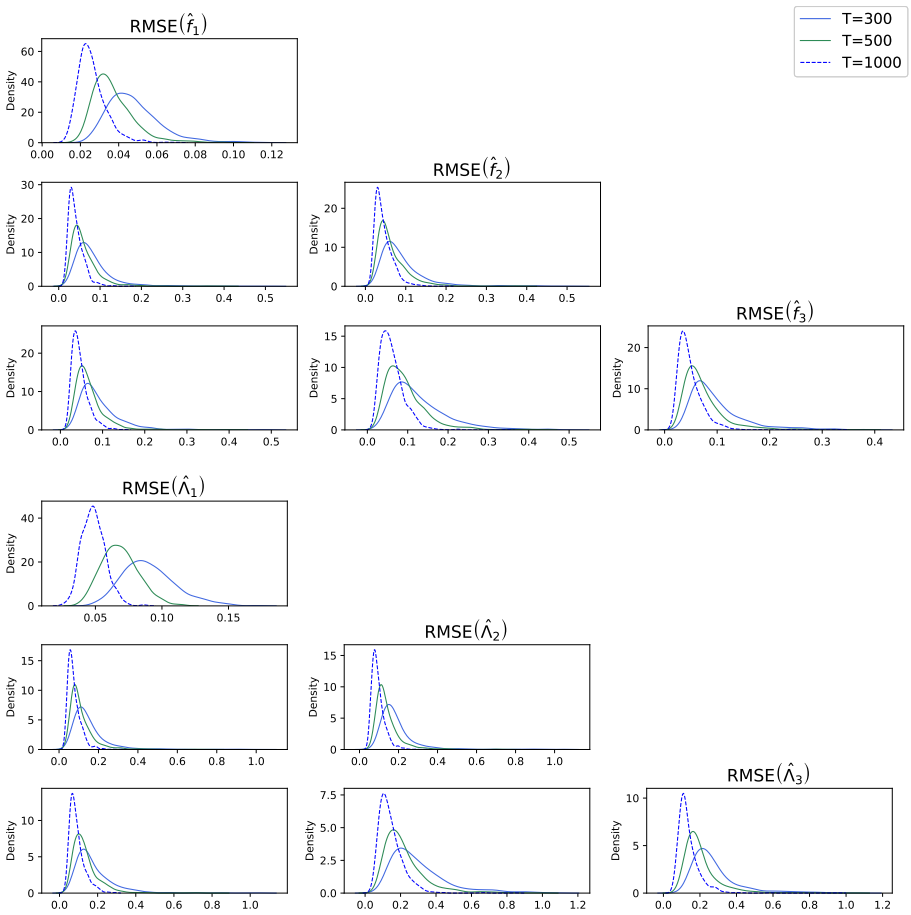


Figure SD.2: **Kernel density of the RMSE for the factors and loadings.** Monte Carlo simulation results. The DGP is a Student's t score-driven factor model with $N = 20$. For further details we refer to Figure 1.

	$T = 300$					$T = 500$					$T = 1000$				
	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.
$r = 2$															
α_1	0.900	0.907	0.007	0.063	0.063	0.900	0.904	0.004	0.048	0.047	0.900	0.902	0.002	0.033	0.033
β_1	0.900	0.895	-0.005	0.040	0.031	0.900	0.897	-0.003	0.027	0.023	0.900	0.898	-0.002	0.020	0.016
$\lambda_{1,1}$	0.953	0.907	-0.046	0.245	0.099	0.953	0.933	-0.020	0.058	0.059	0.953	0.944	-0.009	0.031	0.028
$\lambda_{3,1}$	1.056	1.009	-0.047	0.265	0.145	1.056	1.031	-0.025	0.095	0.094	1.056	1.047	-0.009	0.059	0.059
α_2	0.800	0.802	0.002	0.068	0.063	0.800	0.802	0.002	0.048	0.048	0.800	0.800	-0.000	0.033	0.033
β_2	0.800	0.785	-0.015	0.041	0.048	0.800	0.792	-0.008	0.028	0.038	0.800	0.796	-0.004	0.020	0.025
$\lambda_{1,2}$	-0.073	-0.076	-0.003	0.228	0.266	-0.073	-0.062	0.012	0.176	0.189	-0.073	-0.068	0.005	0.120	0.119
$\lambda_{3,2}$	0.299	0.278	-0.021	0.287	0.306	0.299	0.300	0.001	0.206	0.218	0.299	0.297	-0.002	0.140	0.140
$r = 3$															
α_1	0.900	0.908	0.008	0.061	0.063	0.900	0.903	0.003	0.048	0.047	0.900	0.903	0.003	0.031	0.033
β_1	0.900	0.895	-0.005	0.056	0.030	0.900	0.897	-0.003	0.037	0.022	0.900	0.898	-0.002	0.025	0.016
$\lambda_{1,1}$	0.899	0.848	-0.050	0.112	0.111	0.899	0.873	-0.026	0.072	0.067	0.899	0.888	-0.010	0.043	0.040
$\lambda_{3,1}$	-0.233	-0.214	0.019	0.188	0.176	-0.233	-0.231	0.002	0.138	0.137	-0.233	-0.228	0.005	0.091	0.088
α_2	0.800	0.805	0.005	0.063	0.070	0.800	0.805	0.005	0.047	0.052	0.800	0.802	0.002	0.033	0.034
β_2	0.800	0.791	-0.009	0.054	0.049	0.800	0.795	-0.005	0.036	0.038	0.800	0.797	-0.003	0.023	0.025
$\lambda_{1,2}$	0.027	0.027	-0.000	0.243	0.296	0.027	0.031	0.004	0.198	0.217	0.027	0.027	0.000	0.135	0.137
$\lambda_{3,2}$	-0.201	-0.196	0.004	0.341	0.408	-0.201	-0.172	0.029	0.276	0.346	-0.201	-0.190	0.010	0.199	0.213
α_3	0.700	0.698	-0.002	0.062	0.068	0.700	0.696	-0.004	0.050	0.048	0.700	0.698	-0.002	0.033	0.032
β_3	0.700	0.673	-0.027	0.051	0.068	0.700	0.685	-0.015	0.035	0.051	0.700	0.690	-0.010	0.028	0.032
$\lambda_{1,3}$	-0.480	-0.439	0.042	0.157	0.147	-0.480	-0.462	0.018	0.113	0.116	-0.480	-0.472	0.008	0.073	0.072
$\lambda_{3,3}$	-1.255	-1.145	0.110	0.214	0.216	-1.255	-1.184	0.070	0.151	0.168	-1.255	-1.227	0.028	0.091	0.087

Table SD.1: **Monte Carlo simulation results for 1000 replications for the Gaussian model for different values of T and r ($N = 10$).** The table reports the biases and the standard errors for the static parameters. We report average asymptotic standard deviation (As. std.) as well as the empirical standard deviation (Emp. std.). We present the results for all the elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ as well as for the selected rows of $\boldsymbol{\Lambda}$. We demonstrate the results only for the selected rows due to the large number of the parameters.

	$T = 300$					$T = 500$					$T = 1000$				
	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.
$r = 2$															
α_1	0.900	0.915	0.015	0.060	0.062	0.900	0.909	0.009	0.047	0.045	0.900	0.902	0.002	0.033	0.031
β_1	0.900	0.896	-0.004	0.031	0.026	0.900	0.898	-0.002	0.027	0.021	0.900	0.899	-0.001	0.017	0.013
$\lambda_{1,1}$	0.953	0.917	-0.036	0.083	0.082	0.953	0.935	-0.018	0.050	0.058	0.953	0.945	-0.008	0.031	0.028
$\lambda_{3,1}$	1.056	1.010	-0.046	0.127	0.132	1.056	1.033	-0.022	0.085	0.085	1.056	1.046	-0.010	0.054	0.053
α_2	0.800	0.805	0.005	0.061	0.064	0.800	0.804	0.004	0.047	0.048	0.800	0.801	0.001	0.032	0.031
β_2	0.800	0.788	-0.012	0.035	0.045	0.800	0.793	-0.007	0.027	0.032	0.800	0.797	-0.003	0.018	0.022
$\lambda_{1,2}$	-0.073	-0.068	0.006	0.200	0.232	-0.073	-0.066	0.007	0.155	0.163	-0.073	-0.071	0.002	0.106	0.103
$\lambda_{3,2}$	0.299	0.291	-0.008	0.236	0.271	0.299	0.303	0.004	0.183	0.188	0.299	0.298	-0.001	0.125	0.121
ν	5.000	4.987	-0.013	0.540	0.531	5.000	5.023	0.023	0.406	0.391	5.000	5.015	0.015	0.280	0.280
$r = 3$															
α_1	0.900	0.916	0.016	0.067	0.064	0.900	0.908	0.008	0.049	0.048	0.900	0.904	0.004	0.032	0.033
β_1	0.900	0.896	-0.004	0.044	0.027	0.900	0.898	-0.002	0.033	0.020	0.900	0.900	-0.000	0.023	0.013
$\lambda_{1,1}$	0.899	0.857	-0.042	0.097	0.100	0.899	0.875	-0.024	0.065	0.067	0.899	0.889	-0.009	0.039	0.037
$\lambda_{3,1}$	-0.233	-0.221	0.012	0.172	0.161	-0.233	-0.228	0.005	0.124	0.115	-0.233	-0.230	0.004	0.082	0.078
α_2	0.800	0.808	0.008	0.062	0.068	0.800	0.806	0.006	0.046	0.049	0.800	0.804	0.004	0.032	0.032
β_2	0.800	0.795	-0.005	0.048	0.047	0.800	0.795	-0.005	0.033	0.034	0.800	0.798	-0.002	0.023	0.022
$\lambda_{1,2}$	0.027	0.037	0.010	0.225	0.262	0.027	0.028	0.001	0.173	0.198	0.027	0.021	-0.006	0.119	0.121
$\lambda_{3,2}$	-0.201	-0.188	0.012	0.318	0.398	-0.201	-0.174	0.026	0.258	0.307	-0.201	-0.197	0.004	0.180	0.196
α_3	0.700	0.707	0.007	0.063	0.066	0.700	0.704	0.004	0.047	0.050	0.700	0.700	0.000	0.032	0.032
β_3	0.700	0.679	-0.021	0.049	0.066	0.700	0.683	-0.017	0.042	0.048	0.700	0.692	-0.008	0.022	0.032
$\lambda_{1,3}$	-0.480	-0.440	0.040	0.143	0.147	-0.480	-0.459	0.021	0.103	0.102	-0.480	-0.470	0.010	0.067	0.067
$\lambda_{3,3}$	-1.255	-1.152	0.102	0.204	0.222	-1.255	-1.202	0.052	0.137	0.139	-1.255	-1.232	0.023	0.086	0.086
ν	5.000	4.969	-0.031	0.529	0.525	5.000	4.991	-0.009	0.396	0.386	5.000	5.006	0.006	0.274	0.279

Table SD.2: **Monte Carlo simulation results for 1000 replications for the Student's t model for different values of T and r ($N = 10$).** For further details we refer to Table SD.1

	$T = 300$					$T = 500$					$T = 1000$				
	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.
$r = 2$															
α_1	0.9	0.914	0.014	0.062	0.065	0.9	0.909	0.009	0.043	0.047	0.9	0.904	0.004	0.03	0.032
β_1	0.8	0.817	0.017	0.064	0.066	0.8	0.809	0.009	0.045	0.048	0.8	0.804	0.004	0.03	0.032
$\lambda_{1,1}$	0.9	0.893	-0.007	0.044	0.031	0.9	0.897	-0.003	0.028	0.023	0.9	0.898	-0.002	0.019	0.016
$\lambda_{3,1}$	0.8	0.785	-0.015	0.044	0.049	0.8	0.791	-0.009	0.03	0.037	0.8	0.795	-0.005	0.022	0.025
α_2	0.932	0.881	-0.051	0.1	0.122	0.932	0.905	-0.027	0.07	0.081	0.932	0.923	-0.009	0.045	0.044
β_2	-0.19	-0.197	-0.007	0.206	0.264	-0.19	-0.193	-0.004	0.158	0.197	-0.19	-0.188	0.002	0.11	0.124
$\lambda_{1,2}$	0.845	0.79	-0.054	0.181	0.219	0.845	0.818	-0.026	0.135	0.155	0.845	0.836	-0.009	0.092	0.094
$\lambda_{3,2}$	-0.621	-0.602	0.019	0.239	0.263	-0.621	-0.61	0.011	0.181	0.201	-0.621	-0.613	0.008	0.125	0.132
$r = 3$															
α_1	0.9	0.915	0.015	0.071	0.067	0.9	0.908	0.008	0.049	0.046	0.9	0.902	0.002	0.033	0.033
β_1	0.8	0.822	0.022	0.074	0.072	0.8	0.814	0.014	0.05	0.053	0.8	0.806	0.006	0.031	0.033
$\lambda_{1,1}$	0.7	0.721	0.021	0.07	0.068	0.7	0.71	0.01	0.046	0.049	0.7	0.703	0.003	0.03	0.032
$\lambda_{3,1}$	0.9	0.895	-0.005	0.063	0.031	0.9	0.896	-0.004	0.038	0.023	0.9	0.898	-0.002	0.026	0.016
α_2	0.8	0.789	-0.011	0.051	0.053	0.8	0.793	-0.007	0.037	0.038	0.8	0.795	-0.005	0.025	0.025
β_2	0.7	0.673	-0.027	0.051	0.073	0.7	0.686	-0.014	0.036	0.052	0.7	0.691	-0.009	0.023	0.035
$\lambda_{1,2}$	0.713	0.67	-0.043	0.122	0.126	0.713	0.688	-0.025	0.086	0.092	0.713	0.703	-0.01	-	0.058
$\lambda_{3,2}$	-0.252	-0.217	0.036	0.216	0.292	-0.252	-0.232	0.02	0.164	0.213	-0.252	-0.238	0.015	-	0.145
α_3	0.585	0.519	-0.067	0.19	0.196	0.585	0.549	-0.037	0.133	0.135	0.585	0.573	-0.012	-	0.087
β_3	0.374	0.369	-0.004	0.311	0.418	0.374	0.363	-0.01	0.244	0.306	0.374	0.368	-0.006	-	0.197
$\lambda_{1,3}$	-1.512	-1.303	0.209	0.26	0.341	-1.512	-1.403	0.109	0.176	0.228	-1.512	-1.458	0.054	-	0.138
$\lambda_{3,3}$	0.486	0.428	-0.058	0.366	0.556	0.486	0.476	-0.01	0.269	0.412	0.486	0.499	0.013	-	0.282

Table SD.3: **Monte Carlo simulation results for 1000 replications for the Gaussian model for different values of T and r ($N = 20$).** For further details we refer to the caption of Table SD.1.

	$T = 300$					$T = 500$					$T = 1000$				
	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.	$\boldsymbol{\theta}_0$	Avg. $\hat{\boldsymbol{\theta}}_T$	Bias	As. std.	Emp. std.
$r = 2$															
α_1	0.9	0.925	0.025	0.073	0.066	0.9	0.914	0.014	0.05	0.049	0.9	0.907	0.007	0.036	0.033
β_1	0.8	0.823	0.023	0.063	0.067	0.8	0.814	0.014	0.05	0.049	0.8	0.806	0.006	0.032	0.035
$\lambda_{1,1}$	0.9	0.894	-0.006	0.045	0.027	0.9	0.897	-0.003	0.026	0.021	0.9	0.899	-0.001	0.018	0.013
$\lambda_{3,1}$	0.8	0.789	-0.011	0.038	0.045	0.8	0.789	-0.011	0.029	0.033	0.8	0.797	-0.003	0.017	0.022
α_2	0.932	0.886	-0.046	0.1	0.118	0.932	0.914	-0.018	0.069	0.063	0.932	0.924	-0.008	0.044	0.043
β_2	-0.19	-0.176	0.014	0.221	0.233	-0.19	-0.18	0.009	0.163	0.172	-0.19	-0.188	0.002	0.11	0.11
$\lambda_{1,2}$	0.845	0.809	-0.035	0.184	0.201	0.845	0.826	-0.019	0.131	0.123	0.845	0.835	-0.01	0.089	0.089
$\lambda_{3,2}$	-0.621	-0.592	0.029	0.254	0.257	-0.621	-0.599	0.022	0.184	0.19	-0.621	-0.614	0.007	0.123	0.112
ν	5.0	4.985	-0.015	0.505	0.466	5.0	4.995	-0.005	0.373	0.354	5.0	5.002	0.002	0.255	0.252
$r = 3$															
α_1	0.9	0.925	0.025	0.068	0.068	0.9	0.914	0.014	0.051	0.05	0.9	0.906	0.006	0.033	0.033
β_1	0.8	0.831	0.031	0.08	0.071	0.8	0.817	0.017	0.052	0.051	0.8	0.808	0.008	0.034	0.034
$\lambda_{1,1}$	0.7	0.726	0.026	0.071	0.07	0.7	0.715	0.015	0.053	0.052	0.7	0.708	0.008	0.034	0.034
$\lambda_{3,1}$	0.9	0.897	-0.003	0.054	0.028	0.9	0.897	-0.003	0.036	0.02	0.9	0.898	-0.002	0.023	0.013
α_2	0.8	0.789	-0.011	0.05	0.047	0.8	0.794	-0.006	0.038	0.033	0.8	0.797	-0.003	0.024	0.021
β_2	0.7	0.677	-0.023	0.041	0.068	0.7	0.686	-0.014	0.043	0.048	0.7	0.693	-0.007	0.023	0.032
$\lambda_{1,2}$	0.713	0.675	-0.038	0.128	0.125	0.713	0.691	-0.022	0.091	0.089	0.713	0.704	-0.008	0.058	0.058
$\lambda_{3,2}$	-0.252	-0.213	0.039	0.251	0.275	-0.252	-0.233	0.019	0.185	0.211	-0.252	-0.237	0.015	0.126	0.135
α_3	0.585	0.526	-0.059	0.2	0.184	0.585	0.559	-0.027	0.136	0.132	0.585	0.576	-0.01	0.089	0.089
β_3	0.374	0.356	-0.017	0.334	0.383	0.374	0.357	-0.017	0.255	0.277	0.374	0.384	0.01	0.172	0.173
$\lambda_{1,3}$	-1.512	-1.325	0.186	0.31	0.326	-1.512	-1.413	0.098	0.198	0.237	-1.512	-1.468	0.043	0.124	0.13
$\lambda_{3,3}$	0.486	0.438	-0.048	0.445	0.514	0.486	0.473	-0.014	0.337	0.397	0.486	0.47	-0.016	0.237	0.259
ν	5.0	4.94	-0.06	0.512	0.459	5.0	4.963	-0.037	0.373	0.34	5.0	4.99	-0.01	0.254	0.247

Table SD.4: **Monte Carlo simulation results for 1000 replications for the Student's t model for different values of T and r ($N = 20$).** For further details we refer to the caption of Table SD.1.

E Supplementary empirical application details

E.1 Dataset

Name	Description	Abbreviation
Industrial production	Annual change in log industrial production index, $\log IP_t - \log IP_{t-12}$	INDPRO
Unemployment rate	Annual change in the unemployment rate, $UR_t - UR_{t-12}$	UNRATE
Retail sales	Annual change in log retail sales, $\log RS_t - \log RS_{t-12}$	RETAIL
Consumer sentiment index	Annual change in the survey-based consumer sentiment index constructed by the University of Michigan, $CS_t - CS_{t-12}$	UMCSENTx
S&P500 index	Monthly annual returns on the S&P500 index, $\log SP_t - \log SP_{t-12}$	S&P500
S&P500 volatility	Annualized daily realized volatility computed over the current month	S&P500vol
Credit spread	Difference between the yield on Baa-rated bonds and the yield on 10-year Treasury bonds	BAATB10Y
Housing starts: total new privat. owned	Annual monthly change in the housing starts index $HS_t - HS_{t-12}$	HOUSING
EBP	Excess bond premium	EBP

Table SE.5: **Time series: description, and abbreviation.** All the time series but S&P500 volatility and EBP are retrieved from the FRED MD database McCracken & Ng (2016). The time series for S&P500vol are constructed as discussed in Creal et al. (2014) using the data from the Yahoo finance database.

E.2 Model diagnostics

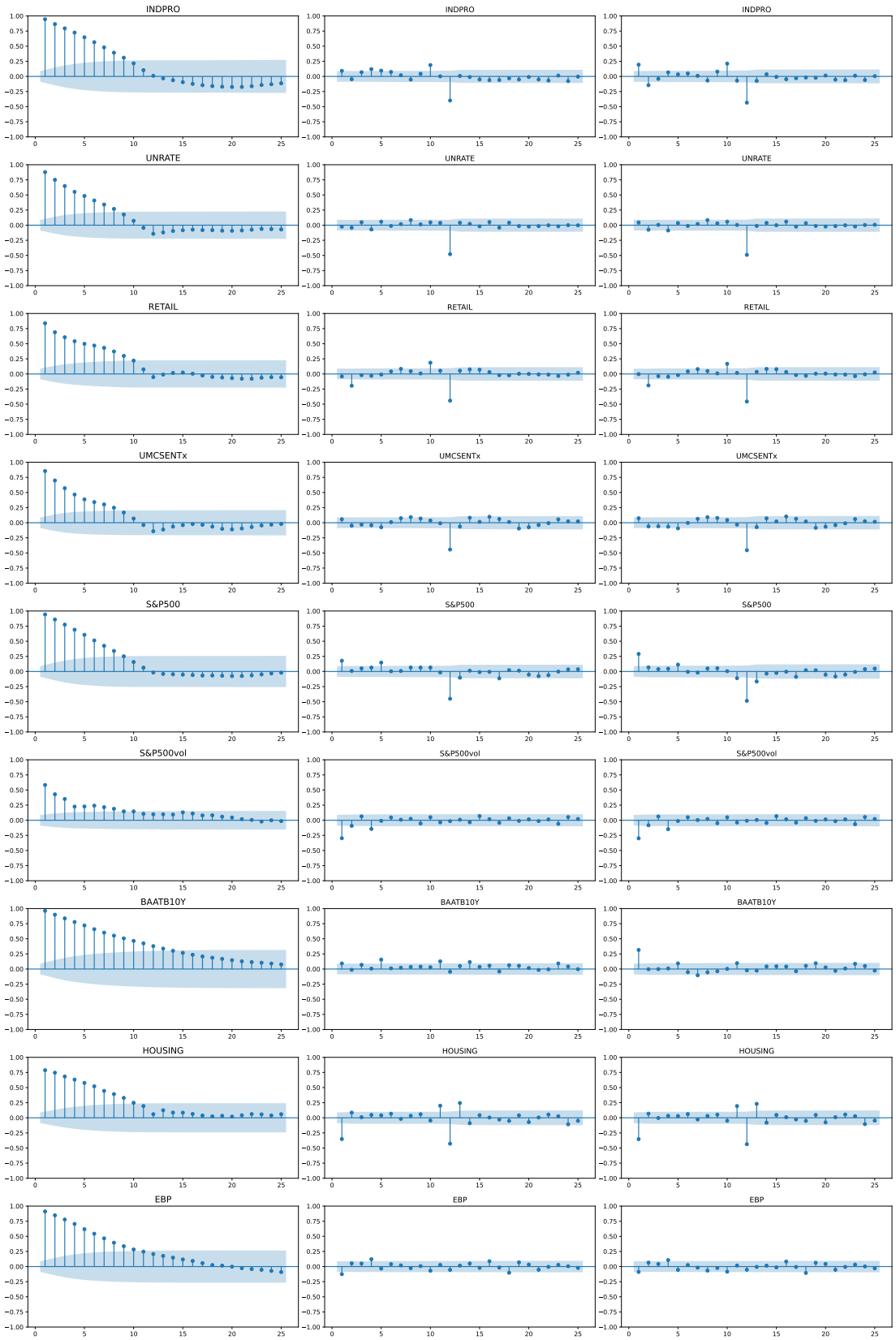


Figure SE.3: **The plots of the autocorrelation functions:** of the data (left column) and the one-step-ahead prediction errors for the individual series of the Gaussian (middle column) and Student's t (right column) models.

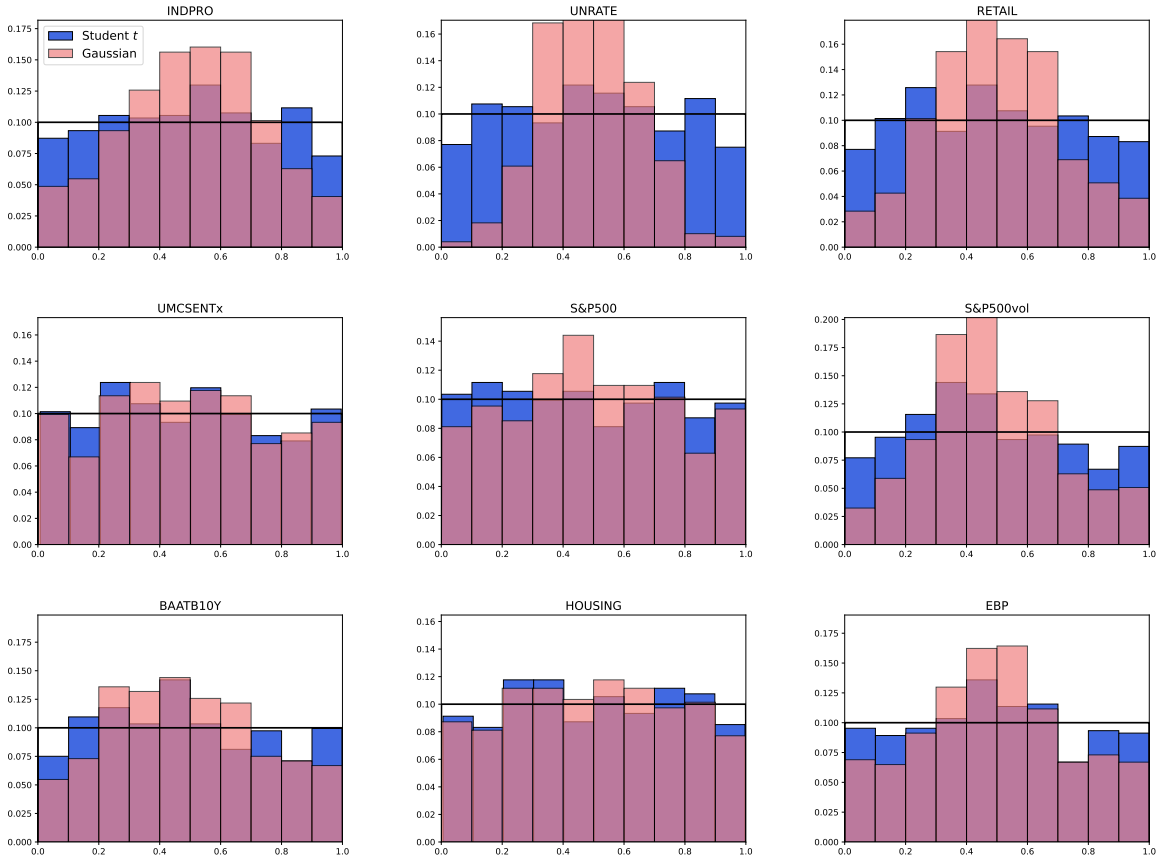


Figure SE.4: **Histograms of the probability integral transforms (PITs) for the Gaussian (light orange) and Student's t (blue) score-driven factor models with $r = 1$ factor.** The PITs were computed using the residuals of the fitted model considered in Section 5.

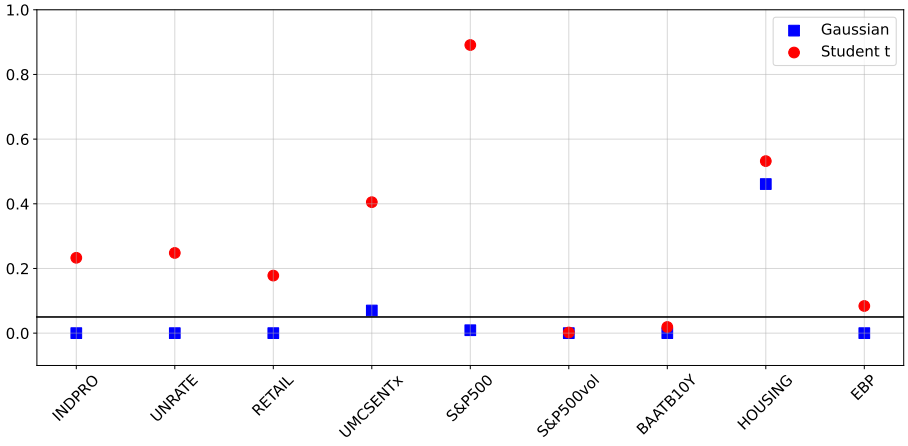


Figure SE.5: **The p-values of the Pearson χ^2 goodness-of-fit test.** The test is applied to the residuals of the score-driven factor models considered in Section 5 with $r = 1$ factors as selected by the BIC. The null hypothesis corresponds to the correct model specification.

References

Billingsley, P. (1961). The Lindeberg-Levy theorem for martingales. *Proceedings of the American Mathematical Society*, *12*(5), 788–792.

Blasques, F., van Brummelen, J., Gorgi, P., & Koopman, S. J. (2022). *Maximum likelihood estimation for non-stationary location models with mixture of normal distributions* (Tech. Rep.). Tinbergen Institute Discussion Paper.

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

Creal, D., Schwaab, B., Koopman, S. J., & 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.

Fang, K.-T., Kotz, S., & Ng, K. W. (2018). *Symmetric multivariate and related distributions*. Chapman and Hall/CRC.

Krengel, U. (1985). *Ergodic theorems* (Vol. 6). Walter de Gruyter.

McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

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

Straumann, D., & 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.