Blasques, Francisco; van Brummelen, Janneke; Gorgi, Paolo; Koopman, Siem Jan

Working Paper

Maximum likelihood estimation for non-stationary location models with mixture of normal distributions

Tinbergen Institute Discussion Paper, No. TI 2022-001/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Blasques, Francisco; van Brummelen, Janneke; Gorgi, Paolo; Koopman, Siem Jan (2022) : Maximum likelihood estimation for non-stationary location models with mixture of normal distributions, Tinbergen Institute Discussion Paper, No. TI 2022-001/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/248792

TI 2022-001/III
Tinbergen Institute Discussion Paper

Maximum Likelihood Estimation for Non-Stationary Location Models with Mixture of Normal Distributions

Francisco Blasques[1,2]
Janneke van Brummelen[1,2]
Paolo Gorgi[1,2]
Siem Jan Koopman[1,2,3]

[1] Vrije Universiteit Amsterdam
[2] Tinbergen Institute
[3] CREATES, Aarhus University

Maximum Likelihood Estimation for Non-Stationary Location Models with Mixture of Normal Distributions [1]

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

[(a)] Vrije Universiteit Amsterdam and Tinbergen Institute
[(b)] CREATES, Aarhus University

January, 2022

Abstract

We consider an observation-driven location model where the unobserved location variable is modeled as a random walk process and where the error variable is from a mixture of normal distributions. The mixed normal distribution can approximate many continuous error distributions accurately. We obtain a flexible modeling framework which is particularly designed for robust filtering and forecasting. We provide sufficient conditions for the strong consistency and asymptotic normality of the maximum likelihood estimator of the parameter vector in the specified model. The asymptotic properties are valid under correct model specification and can be generalized to allow for potential misspecification of the model. A simulation study is carried out to monitor the forecast accuracy improvements when extra mixture components are added to the model. In an empirical study we show that our approach is able to outperform alternative observation-driven location models in forecast accuracy for a time-series of electricity spot prices.

Key words: time-varying parameters, asymmetric and heavy-tailed distributions, robust filter, invertibility, consistency, asymptotic normality.

JEL classification: C13, C22.

[1]Blasques thanks the Dutch Science Foundation (NWO; grant VI.Vidi.195.099) for financial support. Koopman acknowledges support from CREATES, Aarhus University, Denmark, funded by the Danish National Research Foundation, (DNRF78). Corresponding author: S.J. Koopman, Vrije Universiteit Amsterdam, School of Business and Economics, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands. Phone: +31205986019, Fax: +31205986020, Email: `s.j.koopman@vu.nl`.

1. Introduction

In the analysis and forecasting of economic and financial time-series it remains of key importance to have an accurate estimate of the mean or location. The forecasting of a time-series can only be successful when an accurate estimate of the conditional mean is available. The conditional mean at a particular time-period is typically a function of *past* time-series observations. We introduce a non-stationary time-varying location model for the mixture of normal density. The mean or location is modeled as a random walk process with drift and with innovations specified as a function of past observations. Hence, our framework belongs to the class of *observation-driven models*; see the discussion in Cox (1981). In particular, the innovation for the random walk process is defined as the score of the mixed normal density, conditional on past observations (predicted), with respect to the mean variable. Hence, this formulation of a time-varying location model belongs to the class of *score-driven models* which is introduced by Creal et al. (2013) and Harvey (2013).

This time-varying location model relies on two particular novelties within score-driven models: the choice for the mixed normal density and the non-stationary specification for the location variable. The mixed normal density is known to be a very flexible framework as it can approximate almost all continuous density functions, including those with heavy tails, multiple modes, excess kurtosis and asymmetry. Any smooth density can be approximated for any arbitrary amount of error by a finite mixture of normals, when enough components are added to the mixture; see the discussions in McLachlan and Peel (2004), Goodfellow et al. (2016, Section 3.9.6) and Nielsen (2021). The mixed normal distribution is typically designed for settings where observations mostly fluctuate in a moderate fashion but occasionally are contaminated by spikes and large "outliers". The observations may appear to be generated by a fat-tailed distribution, possibly with excess kurtosis. In such a case, we can consider a mixed normal where some components have a moderate variance and a high weight while some components have a large variance and a low weight. A similar construction is used for the so-called 'contaminated' normal distribution which is discussed by Tukey (1960) and Huber (1964), and which forms a fundamental part of the literature on *robust statistics*. The mixed normal model is also used in the context of time-series modeling. An example is the mixed autoregressive model of Wong and Li (2000) where a mixture of stationary and non-stationary autoregressive components is considered for the treatment of possible multimodality in the predictive density.

Our second novelty is the non-stationary specification for the time-varying location. Almost all score-driven models that have been studied in the literature so far have been

analyzed in a stationary context. However, the score framework can also be applied in settings where the observations are non-stationary without further adjustments. We propose a filter for estimating the non-stationary time-varying parameter and a maximum likelihood estimator for the unknown parameter vector. The theoretical properties of the filter and the maximum likelihood estimator have been established for the stationary case; see Blasques et al. (2015) and Blasques et al. (2021). However, in a non-stationary setting the development of theoretical results is more challenging and non-standard because the filtered path of the time-varying location will not be asymptotically stationary. Fortunately, in the current setting, we can show that under fairly general conditions, the difference between the observations and the filtered conditional mean converges to a unique stationary and ergodic limit sequence. This result implies that consistency and asymptotic normality of the MLE can be derived using a similar approach as one would adopt in stationary settings.

Earlier time-varying location models within the family of score-driven models have been proposed. Harvey and Luati (2014) and Caivano and Harvey (2014) consider a similar non-stationary setting but for the Student's t distribution and the exponential generalized beta (EGB) distribution (of the second kind), respectively. More involved dynamic specifications for the location are considered by Caivano et al. (2016). In these contributions, the starting value of the time-varying process for the location are assumed known or as an unknown parameter that needs to be estimated. We will argue that under certain conditions, the starting value is irrelevant in the limit and therefore it does not need to be estimated. In this way, we can provide a complete theoretical framework for the non-stationary location process under a mixed normal error density. Catania (2021) introduced a general dynamic mixture model which also falls within the class of score-driven models. It allows for time-varying mixture components and a time-varying composition, which indeed lead to a flexible and general modeling framework. However, our model is inherently different as we consider a basic signal-plus-noise model where the noise comes from a *time-invariant* mixed normal distribution. It is within this targeted framework that we are able to provide a complete theoretical foundation.

The remainder of the paper is organized as follows. In Section 2 we provide the model specification in detail. In Section 3 we discuss how to filter the time-varying parameter and how to estimate the static parameter vector by means of maximum likelihood. We further formulate conditions for the invertibility of the filter and establish consistency and asymptotic normality of the maximum likelihood estimator of the parameter vector. We argue that these results are also valid under misspecification of the model, when certain conditions on the observations are valid. In Section 4 we carry out two Monte Carlo experiments. The first experiment is to measure the accuracy of the filter in tracking the

time-varying parameter in a finite sample. The second experiment is to verify whether the correct number of components in the mixed normal can be determined within our estimation framework. In Section 5 we analyze a daily time-series of electricity spot prices and we show that our modeling approach can outperform other models, both in-sample as well as out-of-sample. In Section 6 we provide concluding remarks. The proofs of the theorems are given in the Appendix. The technical details and discussions are collected in a Technical Appendix.

2. A non-stationary location model with mixed errors

We aim to develop a filter for the conditional expectation μ_t of univariate stochastic sequences $\{y_t\}_{t\in\mathbb{Z}}$ with stationary and ergodic increments $\{\Delta y_t\}_{t\in\mathbb{Z}}$. Given a sample of observed data $y_1, \ldots, y_T$, our object of interest is thus the sequence $\mu_1, \ldots, \mu_T$ with elements

$$\mu_t = \mathbb{E}(y_t|\mathcal{F}_{t-1}) \ , \quad t = 1, \ldots, T \ ,$$

where $\{\mathcal{F}_t\}_{t\in\mathbb{Z}}$ is the filtration composed of sigma algebras $\mathcal{F}_t = \sigma(y_t, y_{t-1}, \ldots)$. We propose a simple yet flexible way of filtering μ_t which ensures an analytically tractable log-likelihood even in the presence of nonlinear dynamics and non-Gaussian innovations. In particular, we employ a filtering model with an observation equation given by,

$$y_t = \mu_t + \varepsilon_t \, , \tag{1}$$

where we assume that $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a zero mean, independent and identically distributed (iid) sequence of mixed normal random variables. In particular, the innovations are assumed to be drawn from a mixture of J normal distributions such that the probability density function of the innovation ε_t is given by

$$p_\varepsilon(x) = \sum_{j=1}^{J} \frac{w_j}{\sigma_j} \phi\left(\frac{x - c_j}{\sigma_j}\right) \ ,$$

where $\phi(\cdot)$ is the standard normal density function, with fixed weights $w_j \geq 0$ which are subject to $\sum_{j=1}^{J} w_j = 1$, means $c_j \in \mathbb{R}$ which are subject to $\sum_{j=1}^{J} c_j w_j = 0$, and standard deviations $\sigma_j > 0$, for $j = 1, \ldots, J$. The restriction on the means ensures that the innovations have mean zero. The observation equation implies that the density of the observations is given by $p_y(y_t|\mu_t) = p_\varepsilon(y_t - \mu_t)$. The time-varying location μ_t is specified as in the score-driven models of Creal et al. (2013) and Harvey (2013). Given the observation equation (1), the location updating equation is

$$\mu_{t+1} = \omega + \mu_t + \alpha s(y_t - \mu_t) \, , \tag{2}$$

where $\omega \in \mathbb{R}$ is an unknown fixed drift coefficient, $\alpha \in \mathbb{R}$ is an unknown fixed coefficient, and function $s(\cdot) : \mathbb{R} \to \mathbb{R}$ is the score of the innovation density multiplied by the derivative of the inverse link function with the link function denoted by $f(\cdot)$ such that $y_t = f(\mu_t, \varepsilon_t) = \mu_t + \varepsilon_t$. We obtain

$$s(x) = \frac{\partial \log\left(p_\varepsilon(x)\right)}{\partial x} \cdot \frac{\partial f^{-1}(\mu_t, y_t)}{\partial \mu} = \frac{\sum_{j=1}^{J} \sigma_j^{-2} w_j \phi'(z_j)}{\sum_{j=1}^{J} \sigma_j^{-1} w_j \phi(z_j)} \cdot (-1) = \frac{\sum_{j=1}^{J} \sigma_j^{-1} h_j(x)\, z_j}{\sum_{j=1}^{J} h_j(x)}, \quad (3)$$

with standardized variable $z_j = \sigma_j^{-1}(x - c_j)$ and the j-th scaling factor $h_j(x) = \sigma_j^{-1} w_j \phi(z_j)$, for $j = 1, \ldots, J$, where we notice that $\phi'(z_j) := \partial\phi(z_j)/\partial z_j = -z_j\phi(z_j)$. It follows that the score $s(y_t - \mu_t)$ is equal to the score of the conditional observation density $p_y(y_t|\mu_t)$ with respect to μ_t. We assume that the process $\{\mu_t\}_{t=1}^{T}$ is initialized at some value μ_1 that is either a fixed but unknown constant, or a random variable that takes values in $\mathbb{R}$. Given the model equations (1)–(3), it is implied that the process $\{\mu_t\}_{t\in\mathbb{N}}$ is a random walk with drift parameter ω and iid innovation $s(\varepsilon_t)$. The coefficient α can be interpreted as the *signal-to-noise* parameter as it determines how much μ_{t+1} changes in relation to μ_t, given the the innovation $s(\varepsilon_t)$. In case $J = 1$ we have a linear filtering equation, since then $s(x) = x/\sigma_1^2$. This result is immediate because for $J = 1$ the innovations are normally distributed with mean zero and variance σ_1^2. Also, the case of $J = 1$ can be regarded as an observation-driven analogue of the local level model made popular by Harvey (1989) and Durbin and Koopman (2012, Chapter 2).

Our proposed filter (1)–(3) provides considerable flexibility. In practice, the mixture density allows us to capture complex innovation densities which may possibly lead to asymmetric updating mechanisms for (3). This feature can be readily exploited by score-driven models where the updating mechanism reflects the innovation density directly. It is widely established that the mixture density is able to approximate arbitrarily accurately any heavy-tailed (or fat-tailed) density. The filter can subsequently achieve any level of robustness to outliers. Hence, the score update can downweight innovations on any compact set, and hence downweight any observation in a given sample. This ability applies to both symmetric and asymmetric heavy-tailed distribution; see Titterington et al. (1985) and McLachlan and Peel (2004). The use of a mixture density provides robustness and flexibility, but it also allows us to establish the filtering and estimation theory for this score-driven location model. We conclude that our specific use of nonlinear score-driven filtering methods can be regarded as advantageous from both theoretical and practical perspectives.

When we view the set of equations (1)–(3) as a statistical dynamic model, or as a data generating process (DGP), we stress that the dynamic model produces a time-varying location parameter $\{\mu_t\}_{t\in\mathbb{Z}}$ which exhibits non-stationary random walk dynamics. It means that the model generates non-stationary data $\{y_t\}_{t\in\mathbb{Z}}$ with random walk dynamics

and nonlinear moving average innovations. In the special case of $\omega = 0$ and Gaussian innovations ($J = 1$), the model is able to generate linear random-walk data $\{y_t\}_{t\in\mathbb{Z}}$ with Gaussian iid innovations, see Remarks 1 and 2. The conditions for stationary score-driven models are well established and the asymptotic theory has been developed recently in the literature, see Blasques et al. (2021). In Section 3 we consider the theoretical aspects of inference for *non-stationary* data as generated by the score-driven model (1)–(3).

Remark 1. *As a data generating process, the model stated in (1)–(3) generates a sequence* $\{\mu_t\}_{t\in\mathbb{Z}}$ *that is a random-walk with drift* ω *featuring iid innovations* $v_t = \alpha s(\varepsilon_{t-1})$*,*

$$\mu_{t+1} = \omega + \mu_t + v_t.$$

Naturally, if $\omega = 0$ *and* $J = 1$ *then* $\{\mu_t\}_{t\in\mathbb{Z}}$ *follows a random walk with iid Gaussian innovations,*

$$\mu_{t+1} = \mu_t + \frac{\alpha}{\sigma_1^2}\varepsilon_t.$$

Remark 2. *Let* $\omega = 0$*. Then, as a data generating process, the model stated in (1)–(3) generates a random-walk sequence* $\{y_t\}_{t\in\mathbb{Z}}$ *with nonlinear moving average innovations,*

$$y_t = y_{t-1} + \varepsilon_t + \phi(\varepsilon_{t-1}) \ ,$$

where $\phi(\varepsilon_{t-1}) = \alpha s(\varepsilon_{t-1}) - \varepsilon_{t-1}$ *is the nonlinear component. Furthermore, if* $J = 1$ *and* $\alpha = \sigma_1^2$ *we have that* $\phi(\varepsilon_{t-1}) = 0$*, and hence,*

$$y_t = y_{t-1} + \varepsilon_t.$$

3. Parameter estimation and asymptotic properties

The parameter vector for the model (1)–(3) is given by

$$\boldsymbol{\theta} = (\omega, \alpha, \boldsymbol{\psi}')', \text{ where } \boldsymbol{\psi} = \left(c_1, \ldots, c_{J-1}, \sigma_1^2, \ldots, \sigma_J^2, w_1, \ldots, w_{J-1}\right)'.$$

The parameters w_J and c_J are not included in $\boldsymbol{\theta}$ as they are set as functions of the other parameters and are given by

$$w_J = 1 - \sum_{j=1}^{J-1} w_j \, , \quad \text{and} \quad c_J = -\frac{\sum_{j=1}^{J-1} w_j c_j}{w_J} \, ,$$

where the expression for w_J enforces the weights to sum to unity while the one for c_J follows from the restriction that the mean of ε_t is set to zero. Suppose that we have a sample of T observations $\{y_t\}_{t=1}^T$, which is a subset of a sequence $\{y_t\}_{t\in\mathbb{N}}$ generated by the

model equations (1) and (2) under some true parameter $\boldsymbol{\theta}_0$. We do not observe $\{\mu_t\}_{t=1}^T$. Hence, to construct the log-likelihood, we consider the filtered sequence $\{\hat{\mu}_t(\boldsymbol{\theta})\}_{t=1}^T$ as given by

$$\hat{\mu}_{t+1}(\boldsymbol{\theta}) = \omega + \hat{\mu}_t(\boldsymbol{\theta}) + \alpha s(y_t - \hat{\mu}_t(\boldsymbol{\theta}); \boldsymbol{\psi}), \tag{4}$$

where the filtered sequence can for example be initialized using the first observation in the sample $\hat{\mu}_1(\boldsymbol{\theta}) = y_1$ and where the $\boldsymbol{\psi}$ in $s(\cdot\,; \boldsymbol{\psi})$ indexes the mixed normal parameters used in the score function.

The true and unknown parameter $\boldsymbol{\theta}_0$ can be estimated using the method of maximum likelihood (ML). The average log-likelihood function is given by

$$\hat{L}_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=2}^{T} \ell(y_t, \hat{\mu}_t(\boldsymbol{\theta}); \boldsymbol{\theta}), \tag{5}$$

where

$$\ell(y_t, \hat{\mu}_t(\boldsymbol{\theta}); \boldsymbol{\theta}) := \log \left[\sum_{j=1}^{J} h_j \left(y_t - \hat{\mu}_t(\boldsymbol{\theta}) \right) \right],$$

with function $h_j(\cdot)$ given below equation (3). The ML estimator is then defined as

$$\hat{\boldsymbol{\theta}}_T = \arg\max_{\boldsymbol{\theta} \in \Theta} \hat{L}_T(\boldsymbol{\theta}), \tag{6}$$

where Θ is assumed to be a compact parameter set with elements satisfying restrictions for the parameters of the Gaussian mixture as given below.

Assumption PS: $\Theta \subset \mathbb{R}^{3J}$ is a compact set such that for some $\kappa > 0$, for each $\boldsymbol{\theta} \in \Theta$:

(i) $\sigma_1^2 > \ldots > \sigma_J^2 \geq \kappa$ with $\min_{\boldsymbol{\theta} \in \Theta} \sigma_j^2 - \sigma_{j+1}^2 \geq \kappa$ for every $j = 1, 2, \ldots, J-1$,

(ii) $w_j \geq \kappa$ for $j = 1, \ldots, J-1$ and $\sum_{j=1}^{J-1} w_j < 1 - \kappa$.

The restrictions in Assumption PS are identification conditions. Condition (i) may be relaxed to allow for components with equal variances. However, this would require some ordering condition for the means and this would complicate the developments below.

In Section 3.1 we study the invertibility of the filter in this unit-root framework. In Section 3.2 we show that the ML estimator defined in (6) is consistent and asymptotically normal. In Section 3.3 we discuss how the proposed non-stationary score-driven model is a reliable filter for general misspecified I(1) processes.

3.1. Invertibility of the filter and bounded moments of its derivatives

A key ingredient to derive the asymptotic properties of the ML estimator is to ensure the invertibility of the filter $\hat{\mu}_t(\boldsymbol{\theta})$. Even when we assume that the true parameter $\boldsymbol{\theta}_0$ is

known, we still have generally $\hat{\mu}_t(\boldsymbol{\theta}_0) \neq \mu_t$ because the true starting value μ_1 is unknown, meaning that the filter has to be initialized by some other value, for example the first available observation. Invertibility, as defined by Straumann and Mikosch (2006), entails that the filter $\hat{\mu}_t(\boldsymbol{\theta}_0)$ converges to the true conditional expectation μ_t as t goes to infinity. However, these invertibility results only apply for cases where the data generating process is stationary and ergodic. In particular, under stationarity, the filter can be shown to converge uniformly over the parameter space to a stationary and ergodic sequence, which is also known as continuous invertibility; see the discussions in Wintenberger (2013). Continuous invertibility can be used to study the limit properties of the likelihood function and derive the consistency and asymptotic normality of the ML estimator. In our case, the filter $\hat{\mu}_t(\boldsymbol{\theta})$ does not converge to a stationary and ergodic sequence because $\{y_t\}$ is a unit root process. We consider a different approach to ensure an appropriate form of convergence of the log-likelihood function. The likelihood function in (5) depends on the process $\{y_t\}$ only through the prediction error $\hat{g}_t(\boldsymbol{\theta}) \equiv y_t - \hat{\mu}_t(\boldsymbol{\theta})$. In the following, we show that the prediction error $\hat{g}_t(\boldsymbol{\theta})$ converges to a stationary and ergodic sequence uniformly over Θ. This will imply that the terms of the log-likelihood are asymptotically stationary and ergodic. Therefore, the limit properties of the log-likelihood function can be derived using standard arguments.

The prediction errors sequence $\{\hat{g}_t(\boldsymbol{\theta})\}_{t=1}^{T}$ can be expressed through the stochastic recurrence equation (SRE) as given by

$$\hat{g}_{t+1}(\boldsymbol{\theta}) = \hat{g}_t(\boldsymbol{\theta}) - \omega - \alpha s(\hat{g}_t(\boldsymbol{\theta}); \boldsymbol{\psi}) + \Delta y_{t+1}, \tag{7}$$

where we choose $\hat{g}_1(\boldsymbol{\theta}) = 0$, because we choose $\hat{\mu}_1(\boldsymbol{\theta}) = {y_1}^2$. It follows that the prediction error $\hat{g}_t(\boldsymbol{\theta})$ is a SRE with stationary and ergodic innovations since Δy_t is a nonlinear moving average process as shown in Remark 2.

The proposition below shows that, under a contraction condition, the prediction error converges exponentially fast almost surely (e.a.s.) to a stationary and ergodic limit process with a bounded log moment uniformly over Θ. Furthermore, it shows that the filter evaluated at the true parameter value converges to the true conditional expectation of the data generating process. For notational convenience, we write the SRE in (7) as $\hat{g}_{t+1}(\boldsymbol{\theta}) = \phi_t(\hat{g}_t(\boldsymbol{\theta}), \boldsymbol{\theta})$, where ϕ_t is a random function from $\mathbb{R} \times \Theta$ to $\mathbb{R}$, defined by $\phi_t(g, \boldsymbol{\theta}) = g - \omega - \alpha s(g; \boldsymbol{\psi}) + \Delta y_{t+1}$. We define the r-fold convolution of the function ϕ_t as $\phi_t^{(r)}(\cdot, \boldsymbol{\theta}) = \phi_t(\cdot, \boldsymbol{\theta}) \circ \ldots \circ \phi_{t-r+1}(\cdot, \boldsymbol{\theta})$ and we define the derivative of $\phi_t^{(r)}$ as $\dot{\phi}_t^{(r)}(g, \boldsymbol{\theta}) = \partial \phi_t^{(r)}(g, \boldsymbol{\theta}) / \partial g$.

[2]The setting $\hat{g}_1(\boldsymbol{\theta}) = 0$ is not strictly needed for the results presented in the remainder of this section as they hold irrespective of the initialization $\hat{g}_1(\boldsymbol{\theta})$.

Proposition 1. *Let $\{y_t\}_{t\in\mathbb{Z}}$ satisfy the model's equations (1)-(3) for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with $\boldsymbol{\theta}_0 \in \Theta$. Furthermore, assume that Θ satisfies Assumption PS and that the following condition is satisfied for some integer $r \geq 1$:*

$$\mathbb{E} \log \sup_{\boldsymbol{\theta}\in\Theta} \sup_{g\in\mathbb{R}} |\dot{\phi}_t^{(r)}(g, \boldsymbol{\theta})| < 0. \tag{8}$$

Then, the following results hold true:

(i) The prediction error $\hat{g}_t$ converges e.a.s. to a unique stationary and ergodic sequence $\{g_t\}_{t\in\mathbb{Z}}$ uniformly over Θ, with a finite $\log^+$-moment, i.e.

$$\sup_{\boldsymbol{\theta}\in\Theta} |\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})| \overset{e.a.s.}{\to} 0, \quad as \quad t \to \infty,$$

where $\mathbb{E} \log^+ \sup_{\boldsymbol{\theta}\in\Theta} |g_t(\boldsymbol{\theta})| < \infty$.

(ii) The filter $\mu_t(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ converges e.a.s. to the true conditional expectation μ_t,

$$|\hat{\mu}_t(\boldsymbol{\theta}_0) - \mu_t| \overset{e.a.s.}{\to} 0, \quad as \quad t \to \infty.$$

The proof relies on Straumann and Mikosch (2006, Theorem 2.8) and Bougerol (1993, Theorem 3.1), which both provide sufficient conditions for the stability of SREs with stationary and ergodic innovations. Proposition 1 entails that the limit of the prediction error $g_t(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta}_0$, is equal to the error term of the data generating process, i.e. $g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ almost surely.

The contraction condition in (8) requires the r-th iterate of the function ϕ_t to be contracting on average, for some integer r. For $r = 1$, the condition simplifies to a uniform contraction, i.e. that $|1 - \alpha s'(g; \boldsymbol{\psi})| < 1$ uniformly over $g \in \mathbb{R}$ and $\boldsymbol{\theta} \in \Theta$, but in practice this is typically too strict and we need to consider $r > 1$. An analytical closed form expression for the condition is not available in that case. For the simple case $r = 1$ and $J = 2$, it may be shown that the parameter set satisfying (8) is not degenerate. In practice, this condition for $r > 1$ can be checked based on the observed data sample following the approach of Blasques et al. (2018) for feasible invertibility conditions, where empirical boundaries of the parameter space are constructed using the observations. In the current setting, the verification requires a numerical maximization of the derivative of the r-fold iterate over g for every observation, which is somewhat computationally intensive, but not infeasible. In case of our empirical study in Section 5, the feasible version of the invertibility condition holds for large values of r.

In order to establish asymptotic normality of the ML estimator, we also need to know whether the first and second order derivative of the empirical prediction errors $\hat{g}_t(\boldsymbol{\theta})$ converge to a limit stationary and ergodic process e.a.s. Furthermore, we need certain

moments of $g_t(\boldsymbol{\theta})$ and its first two derivatives to exist. The following proposition is similar to Proposition of 3.4 of Blasques et al. (2021).

Proposition 2. *Let $\{y_t\}_{t\in\mathbb{Z}}$ satisfy the model's equations (1)-(3) for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with $\boldsymbol{\theta}_0 \in \Theta$. Furthermore, assume that Θ satisfies Assumption PS. Let ϕ_t and its r-th iterate $\phi_t^{(r)}$ be defined as in Proposition 1. Let for some integer $r \geq 1$ and some $n > 0$,*

$$\mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} \sup_{g\in\mathbb{R}} \left| \dot{\phi}_t^{(r)}(g, \boldsymbol{\theta}) \right|^n < 1 \,. \tag{9}$$

Then the following results hold:

(i) $\{\hat{g}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ *converges e.a.s. uniformly over Θ to a unique stationary and ergodic sequence $\{g_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ where $g_t(\boldsymbol{\theta})$ has n bounded moments uniformly over Θ, i.e. $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |g_t(\boldsymbol{\theta})|^n < \infty$.*

(ii) $\{\partial \hat{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}_{t\in\mathbb{N}}$ *converges e.a.s. uniformly over Θ to a unique stationary and ergodic sequence $\{\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}_{t\in\mathbb{Z}}$ where $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\|^n < \infty$.*

(iii) $\{\partial^2 \hat{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\}_{t\in\mathbb{N}}$ *converges e.a.s. uniformly over Θ to a unique stationary and ergodic sequence $\{\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\}_{t\in\mathbb{Z}}$ where $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\|^{n/2} < \infty$.*

Notice that the conditions of Proposition 2 are stronger than those of Proposition 1. In case condition (8) of Proposition 1 holds for $r = 1$, there is a uniform contraction, implying that the contraction condition of Proposition 2 holds trivially. In case there is no uniform contraction, a feasible version of the condition in (9) can be verified in the same way as is suggested for the condition of Proposition 1. As we shall see below, the conditions of Proposition 1 are sufficient for the consistency of the ML estimator, instead, asymptotic normality relies on the conditions of Proposition 2 with $n \geq 4$.

3.2. Asymptotic properties of the maximum likelihood estimator

The next result delivers the consistency of the ML estimator under the assumptions of Proposition 1.

Theorem 1 (Consistency)**.** *Assume that the assumptions of Proposition 1 hold. Then the ML estimator $\hat{\boldsymbol{\theta}}_T$ satisfies $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $T \to \infty$.*

If additionally the assumptions of Proposition 2 hold, then the following theorem provides the asymptotic normality of the ML estimator. In this theorem, $\ell_t(\boldsymbol{\theta})$ denotes the t-th log-likelihood contribution evaluated in the limit filter, i.e. $\ell_t(\boldsymbol{\theta}) \equiv \ell(y_t, g_t(\boldsymbol{\theta}) - y_t; \boldsymbol{\theta})$, where $\ell(\cdot, \cdot; \cdot)$ is defined below (5).

Theorem 2 (Asymptotic Normality)**.** *Let the assumptions of Proposition 2 hold for $n \geq 4$. Then, if $\boldsymbol{\theta}_0$ lies in the interior of Θ, the ML estimator $\hat{\boldsymbol{\theta}}_T$ satisfies*

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}) \quad as \quad T \to \infty\,,$$

where $\mathcal{I} := -\mathbb{E}[\partial^2 \ell_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']$ is the Fischer information matrix with its expression given in Technical Appendix C.4.

We do not need the restriction $\alpha_0 \neq 0$, to enforce identification of ω and β and to rule out that μ_t is deterministic, since $\alpha_0 = 0$ is already ruled out by the contraction conditions of Proposition 2. This notion follows immediately as these contraction conditions can never hold if $\alpha = 0$. There is no explicit expression of the Fischer information matrix in terms of the parameters, because expectations such as $\mathbb{E}[s(\varepsilon_t; \boldsymbol{\psi}_0)^2]$ cannot be evaluated analytically for a general $\boldsymbol{\psi}_0$. Therefore, we can only provide an expression of the Fischer information matrix in terms of expectations and parameters, see Technical Appendix C.4.

3.3. Filtering and estimation under misspecifcation

In the proof of Proposition 1, the correct specification assumption is only used to make sure that $\{\Delta y_t\}$ has certain properties. Hence, it can be proved straightforwardly that the following corollary holds.

Corollary 1. *Let $\{y_t\}_{t\in\mathbb{Z}}$ be a sequence with first differences $\{\Delta y_t\}_{t\in\mathbb{Z}}$ that are stationary and ergodic with $n > 0$ bounded moments. Then if Θ satisfies Assumption PS and condition* (8) *holds for Θ and for some integer $r \geq 1$, the prediction error $\hat{g}_t$ converges e.a.s. to a unique stationary and ergodic sequence $\{g_t\}_{t\in\mathbb{Z}}$ uniformly over Θ, i.e.*

$$\sup_{\boldsymbol{\theta}\in\Theta} |\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})| \overset{e.a.s.}{\to} 0, \quad as \quad t \to \infty.$$

If also Δy_t is independent of Δy_{t-s} for all integers $s \geq 2$, then $\mathbb{E}\log^+ \sup_{\boldsymbol{\theta}\in\Theta} |g_t(\boldsymbol{\theta})| < \infty$.

It follows that we can prove the existence of filter invertibility without the need to assume the correct specification of the model. This is an important result because correct model specification is a rather strong assumption in many practical settings. Corollary 1 implies that even if the observed data is not generated by the model under consideration, the filter still forgets its initialization in the limit and the prediction errors will converge to an $I(0)$ sequence in the limit e.a.s. Therefore, filtering and estimation based on this model can be justified, even when the correct specification assumption seems to be unrealistic. Notice that the condition on the independence of Δy_t and Δy_{t-s} for $s \geq 2$, that is needed for a bounded $\log^+$-moment of the limit prediction error, can be weakened in case there is a stronger contraction condition. In particular, if there is a uniform contraction, in

other words, if condition (8) holds for $r = 1$, then this independence assumption is not needed. It can be verified that the correct specification assumption of Proposition 2 can be replaced by the same conditions on $\{y_t\}_{t\in\mathbb{Z}}$ as in Corollary 1.

The consistency and asymptotic normality results can also be generalized to a model misspecification setting, in a similar way as shown by Blasques et al. (2021). In particular, the strong consistency of Theorem 1 can be attained under potential misspecification if the conditions of Corollary 1 hold, and if we assume there is a unique maximizer $\boldsymbol{\theta}_0 \in \Theta$ of the limit log-likelihood function $\mathbb{E}[l_t(\boldsymbol{\theta})]$. In this case, $\boldsymbol{\theta}_0$ is the so-called pseudo-true parameter which minimizes the Kullback-Leibler divergence between the probability measure of the sample and the one implied by the model. A thorough discussion of statistical inference based on misspecified models is provided in the book of White (1994), for example. To obtain asymptotic normality under misspecification, Theorem 2 should be altered to impose that the data $\{y_t\}_{t\in\mathbb{Z}}$ is such that the results of Proposition 2 hold for $n \geq 4$ under misspecification; see the discussion above. Furthermore, we require that there is a central limit theorem (CLT) that can be applied to the first derivative of the log-likelihood function. For instance, a CLT for near epoch dependent sequences on a mixing sequence of some size as in Blasques et al. (2021); see also Pötscher and Prucha (1997, Chapter 6). Finally, the invertibility of the limit Hessian, evaluated at $\boldsymbol{\theta}_0$, has to be assumed as well.

4. Monte Carlo study

To assess whether our modeling framework is able to filter the conditional expectation μ_t from a time-series accurately and whether it is able to select the correct J adequately in a finite sample, we conduct a Monte Carlo study consisting of two experiments.

4.1. Design of first Monte Carlo experiment

We simulate observations $\{y_t\}_{t=1}^T$ from the data generation process (DGP) given by

$$y_t = \mu_t + \varepsilon_t\,, \qquad \{\varepsilon_t\}_{t=1}^T \sim \text{iid Student's Skew-}t(\tau = 0, \sigma^2 = 2, \nu = 3.5, \gamma = 0.8)\,, \quad (10)$$

where the innovation ε_t is assumed to come from a (standardized) Student's Skew-t distribution, with its skewness enforced as in Fernández and Steel (1998), and with mean $\tau = 0$, variance $\sigma^2 = 2$, degrees of freedom $\nu = 3.5$ and skewness parameter $\gamma = 0.8$. Hence, the distribution of the innovation ε_t is negatively skewed. We consider four different paths, with sample length T, for the time-varying location parameter μ_t in the DGP model (10):

1. Random walk:

$$\mu_{t+1} = \mu_t + v_t\,, \qquad \{v_t\}_{t=1}^{T} \sim \text{iid } \mathcal{N}(0, 0.25)\,,$$

see Figure 1 for the specific path that is generated.

2. Random walk with drift:

$$\mu_{t+1} = 0.2 + \mu_t + v_t\,, \qquad \{v_t\}_{t=1}^{T} \sim \text{iid } \mathcal{N}(0, 0.25)\,,$$

where we use the same random draws for v_t as in 1.

3. Linear trend with cycle: $\mu_t = 0.04 + 5 \times \sin(5 \cdot 2\pi t \,/\, T)$.

4. Linear trend with break: $\mu_t = 0.01\,t + 5 \times I\{t > T/2\}$, where $I\{\cdot\}$ is the indicator function returning 1 when argument is true, and 0 otherwise.

We have selected these particular paths because they have a unit root and/or a deterministic trend. Such dynamic features in the data can also be generated by our proposed model (1)–(3) in Section 2. However, for other selection of paths, also those without such trends, we have found similar results as for those reported below[3].

The Monte Carlo results are based on experiments where we have generated, for each μ_t specification as given above, 1000 corresponding time-series y_t from (10), with time-series length $T = 1000$. For each simulated time-series, we consider the mixed-normal score-driven location model (1)–(3), with $J = 1, 2, 3$ components, and estimate its parameters by the method of maximum likelihood as discussed in Section 3. We should emphasize that, as the number of components increases, the optimization becomes more challenging because the log-likelihood surface may be somewhat ill-behaved, which results in the optimization process having to overcome local optima. Notice that adopting an expectation-maximization approach, as is common for mixture models, is no solution to this, because the distributional parameters $\boldsymbol{\psi}$ occur both in the log-likelihood explicitly and in the updating function of μ_t. To facilitate these estimation challenges for the model with $J = 3$, we reduce the dimension of the parameter vector by restricting the component variances as $\sigma_3^2 = \sigma^2, \sigma_2^2 = k\sigma^2, \sigma_1^2 = k^2\sigma^2$ for the 'new' parameters $\sigma^2 > 0$ and $k > 1$. This restriction is more or less arbitrary, but it is effective and the estimation results are similar to those obtained from unrestricted estimation. In a few cases, the restriction even leads to a higher maximized log-likelihood value. In these cases, the unrestricted estimation process may have ended in a local optimum. When parameters need to be estimated on a case-by-case basis in an empirical setting, it is advised to start the (unrestricted) estimation process with different starting values of the parameters for

[3] Additional Monte Carlo results can be made available upon request.

the optimization. In our first simulation study, for motivations of feasibility, we have adopted the restriction as given above.

4.2. Filtering precision

For all DGPs, the model (1)–(3), for any choice of J, is clearly mis-specified, but we anticipate that the mixture of normals can accommodate the thick tails and skewness of the skewed Student's t distribution accurately. Furthermore, our score-driven updating mechanism provides a robust filtering method for the underlying location parameter μ_t. Our Monte Carlo results support these propositions. The median filtered paths and their 2.5 and 97.5 percentiles, for the models with $J = 1$ and $J = 3$, are presented in Figure 1, together with the true paths. The filtered path for $J = 2$ is omitted, because it is very similar to the one for $J = 3$, only slightly less accurate. Our score-driven model (1)–(3) is able to filter the true path accurately. For the DGP with a linear trend with break, it takes some time for both filters to move to the right level after the break.

Furthermore, we can learn from the Monte Carlo results in Figure 1 that the $J = 1$ percentiles of the filtered estimates are further apart when compared to the $J = 3$ percentiles. This finding is expected since our filter with $J = 3$ is robust to large observations which occur regularly due to the thick-tailed error distribution used in the DGP model (10). The filter with $J = 1$ does not have such robustness features as it is based on a linear function of past observations which are assumed to come from a normal distribution. Another finding of interest from the results for the filters with $J = 1$ and $J = 3$ is the relatively large differences between their 2.5 percentiles compared to the 97.5 percentiles. The filter with $J = 3$ is able to account for the negative skewness of the innovation distribution, in contrast to the linear and symmetric filter with $J = 1$. Overall we can conclude that the nonlinear $J = 3$ filter is more accurate in the filtering of the time-varying location μ_t compared to the linear $J = 1$ filter. Hence, the consideration of a nonlinear filter can be more effective in estimating time-varying location variables.

Our Monte Carlo results are also summarized in Table 1. Given the filtered paths $\hat{\mu}_t$ for the 1000 simulated time-series, it presents the averages of the maximized log-likelihood values and the mean squared error $T^{-1}\sum_{t=1}^{T}(\mu_t - \hat{\mu}_t)^2$ values. The filtered locations have smaller mean squared errors when they are based on models with multiple mixture components, when compared to the linear filter ($J = 1$), for all four DGPs. The only exception is the DGP of the linear trend with break. In this case, the linear filter of $J = 1$ is able to respond relatively quick to the break in the trend, which makes it as competitive as the $J = 2$ and $J = 3$ filters. Finally, the fit in terms of maximized log-likelihood values confirms that filters based on the model (1)–(3), with $J > 1$, considerably outperform the linear filter ($J = 1$) for time-series coming from one of the considered DGPs. It is not

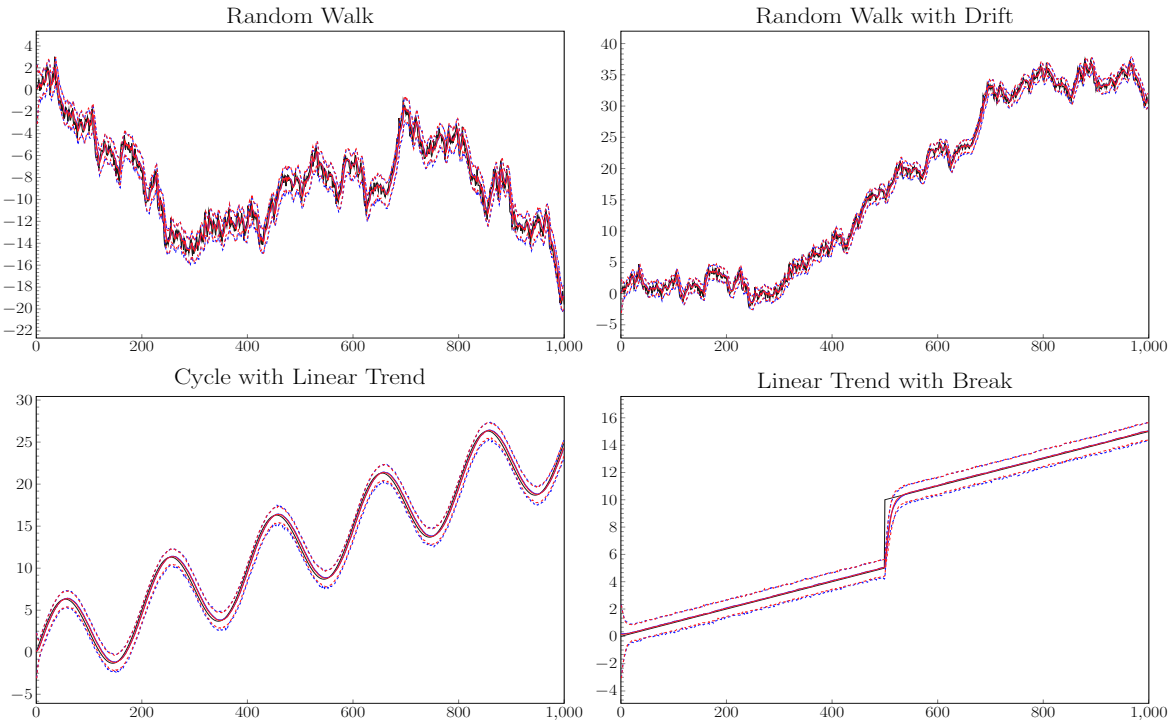

Figure 1. Median of the 1000 filtered paths $\hat{\mu}_t$ for $J = 1$ (blue solid) and $J = 3$ (red solid), the corresponding 2.5 and 97.5 percentiles (dashed), and the true simulated path μ_t (black solid)

Table 1. Monte Carlo results for filtering precisions.

	Log-likelihood value			Mean squared error		
	$J = 1$	$J = 2$	$J = 3$	$J = 1$	$J = 2$	$J = 3$
Random Walk	-1.935	-1.851	-1.846	0.853	0.749	0.744
Random Walk w/Drift	-1.935	-1.851	-1.845	0.848	0.745	0.740
Fixed Trend w/Cycle	-1.865	-1.754	-1.746	0.480	0.371	0.364
Fixed Trend w/Break	-1.814	-1.710	-1.699	0.245	0.261	0.247

We report the averages of the maximized log-likelihood values and the mean squared error $T^{-1}\sum_{t=1}^{T}(\mu_t - \hat{\mu}_t)^2$ values, for 1000 simulated time-series, with $T = 1000$, from the data generation process (DGP) model (10) with a particular path μ_t as depicted in Figure 1.

surprising that the average maximized log-likelihood value increases with J because the models with $J > 1$ components nest the linear $J = 1$ component model, when ignoring the identification restrictions.

4.3. Design of second Monte Carlo experiment

The second part of our Monte Carlo study focuses on the selection of the number of components J. In the asymptotic theory, as developed in Section 3, we have assumed that the number of components J has been selected beforehand, *a-priori*. In practice,

we typically rely on likelihood-ratio statistics to determine a value of J. However, such tests suffer from identification problems on the boundary of the admissible parameter set, under the null hypothesis. Hence, the resulting test statistics do not have standard limit distributions. To avoid such inference issues, we consider the use of Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC). In this Monte Carlo simulation study, we assess how accurate these information criteria are in selecting the correct number of components. For this purpose, we revisit the Monte Carlo design as above but with the correct model in (1)–(3) as DGP with $J = 2$. We estimate the parameters of the models with $J = 1$, $J = 2$ and $J = 3$ components for different sample sizes ($T = 250, 500, 1000$) and obtain the AIC and BIC. Table 2 reports the number of times each model is selected over 1000 Monte Carlo replications.

4.4. Accuracy in the selection of J

We can summarize the results reported in Table 2 as follows. The BIC selects the correct number of components ($J = 2$) much more often when compared to the AIC. Further, the AIC tends to produce an overestimation of the number of components J as it is selecting $J = 3$ more often than $J = 2$. Both AIC and BIC improve as the sample size increases. The BIC performs particularly well in small sample sizes, attaining 98% of correct selections for $T = 1000$. For computational reasons we have considered $J = 3$ as the maximum number of components of the model for this simulation study. The likelihood becomes more flat when the model is over-parameterized which makes numerical optimization more challenging. Therefore, a limitation of this simulation study is that the number of times $J = 2$ is selected may be overestimated, especially for AIC. However, the reported results provide a good indication of the relative performance of AIC and BIC; they highlight how selection improves as the sample size increases. Overall, we can conclude that the BIC should be preferred to AIC as a practical way to select the number of components of the mixture, especially for small T.

Table 2. Accuracy in the selection of J.

	$T = 250$			$T = 500$			$T = 1000$		
	$J = 1$	$J = 2$	$J = 3$	$J = 1$	$J = 2$	$J = 3$	$J = 1$	$J = 2$	$J = 3$
AIC	0	126	874	0	242	758	0	373	627
BIC	5	773	222	0	917	83	0	980	20

Number of times each number of components $J = 1, 2, 3$ is selected by AIC and BIC for different sample sizes. The results are based on 1000 Monte Carlo replications. The DGP is the model in (1)–(3) with $J = 2$ and parameter vector $(\alpha, \omega, \sigma_1^2, \sigma_2^2, c_1, w_1) = (2, 0, 27.5, 3.5, 3.4, 0.1)$.

5. Empirical study

We provide an empirical illustration where a time-series of daily electricity spot prices is analyzed. Since the worldwide deregulation of wholesale electricity markets in the 1990s, electricity spot prices have been studied to investigate the extremely volatile behaviour in more detail. This volatility makes the prediction of spot prices more challenging. Time-series of electricity prices can be characterised by random walk dynamics but contaminated by unexpected occurences or spikes of much higher prices during a short time period. Such spikes are usually not encountered in daily asset prices traded at regulated financial markets. According to Escribano et al. (2011), this erratic behaviour is primarily caused by the non-storability of electricity, which implies that demand and supply must always be balanced. Therefore, shocks in either supply or demand will induce large price movements. It also plays a role that the demand of electricity is fairly price inelastic, since the demand will barely react to price changes.

We consider a time-series of daily electricity spot prices of the PJM electricity market, which serves 13 states in the United States, from June 1, 2007 to April 30, 2021. We take logs of the prices and multiply them by 10; see Figure 2 for a plot of this time-series. The final three years of the sample are used for out-of-sample forecast evaluations. Hence, the implied in-sample time-series is of length $T = 3987$ (from June 1, 2007 to April 30, 2018). The time-series plot confirms the large (mainly positive) spikes in the data. After a spike, the price typically quickly reverts back to its pre-spike level. For this reason, if we are interested in filtering the conditional mean for this time-series, the case for a robust filter is compelling. The filter should namely not react too strongly to occasional large observations (or spikes), because these price shifts are usually temporary in nature. Our score-driven mixed-normal location model with $J > 1$ components can deliver this needed robustness, because it can take into account fat tails and asymmetry in the density of the errors. This empirical study provides an illustration of our methodology. We do not claim that our suggested model captures all the dynamic features in the data. For example, we ignore the possible importance of conditional volatility in this time-series, as well as the possible seasonal effects such as day-in-week, holidays, and quarterly variations.

5.1. Model specifications and parameter estimation

In our econometric analysis, we estimate the parameters for the the score-driven mixed normal model (1)–(3), with $J = 1, 2, 3$. We also consider the score-driven model based on the Student's t distribution with ν degrees of freedom. This model is defined in equation (18) of Harvey and Luati (2014). The filter implied by the Student's t model is also robust against outliers because the score contribution converges to zero for large observations.

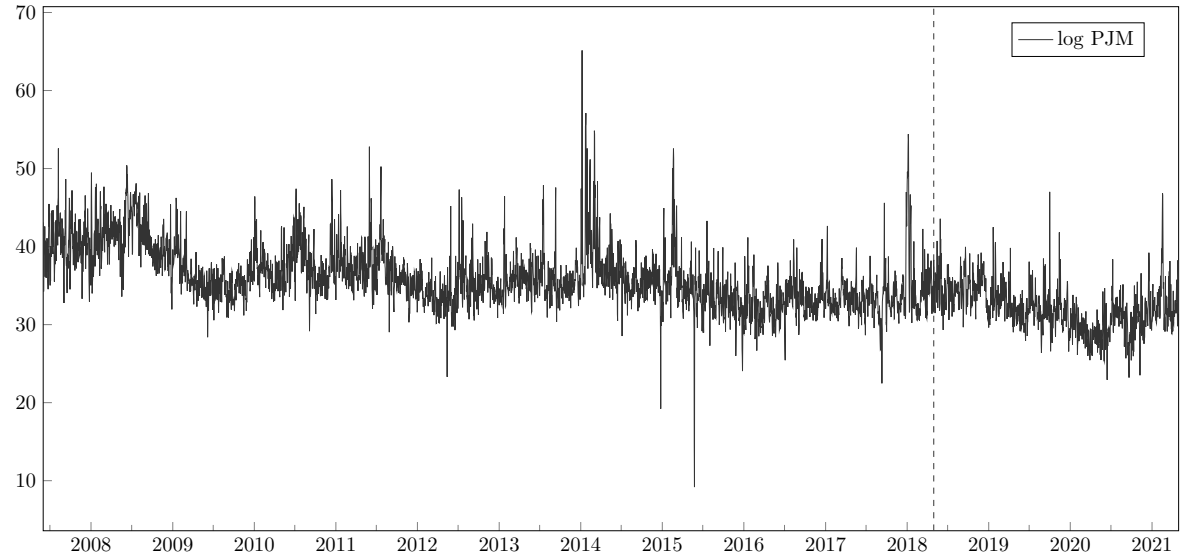

Figure 2. The log daily electricity spot prices of PJM from 1-6-2007 to 30-4-2021. The Dashed line at 30-04-2018 indicates the end of the 'in-sample data' used for the estimates in Table 3.

This leads to a parametric form of trimming on the observations. It is this property of the Student's t filter which induces a theoretical problem in the current unit root setting, as implied by the updating equation (2). Since the derivative of the score function also converges to zero in the limit, we cannot formally establish filter invertibility for the Student's t model. However, we still consider the model of Harvey and Luati (2014) in our empirical study because it is a natural candidate to compare it with our mixture models.

We present in Table 3 the parameter estimates, the maximized log-likelihood values and their corresponding information criteria AIC and BIC, for the mixed normal models with $J = 1$, 2 and 3 components and for the Student's t model. The model specifications are without intercept, that is $\omega = 0$ in equation (2), because the AIC and BIC values are higher in all cases with $\omega \neq 0$. The reported standard errors in Table 3 are obtained by using the asymptotic variance matrix expression in Theorem 2, under the assumption of correct model specification. Within the table the increase of J leads to lower values for AIC and BIC. The model with $J = 4$ is not reported in Table 3 because it shows an increase of both of these values. Recall that estimating the parameters in our model (1)–(3) for higher values of J, including $J = 4$, is challenging. To ensure that the estimation process has not ended in a local optimum, we have repeated the estimation for all models many times with different starting values for the parameters. However, the estimation process overall has not caused too many challenges; this is mainly due to the time-series length which is sufficiently high.

For the model with $J = 1$ components, the contraction condition of Propositions 1 and

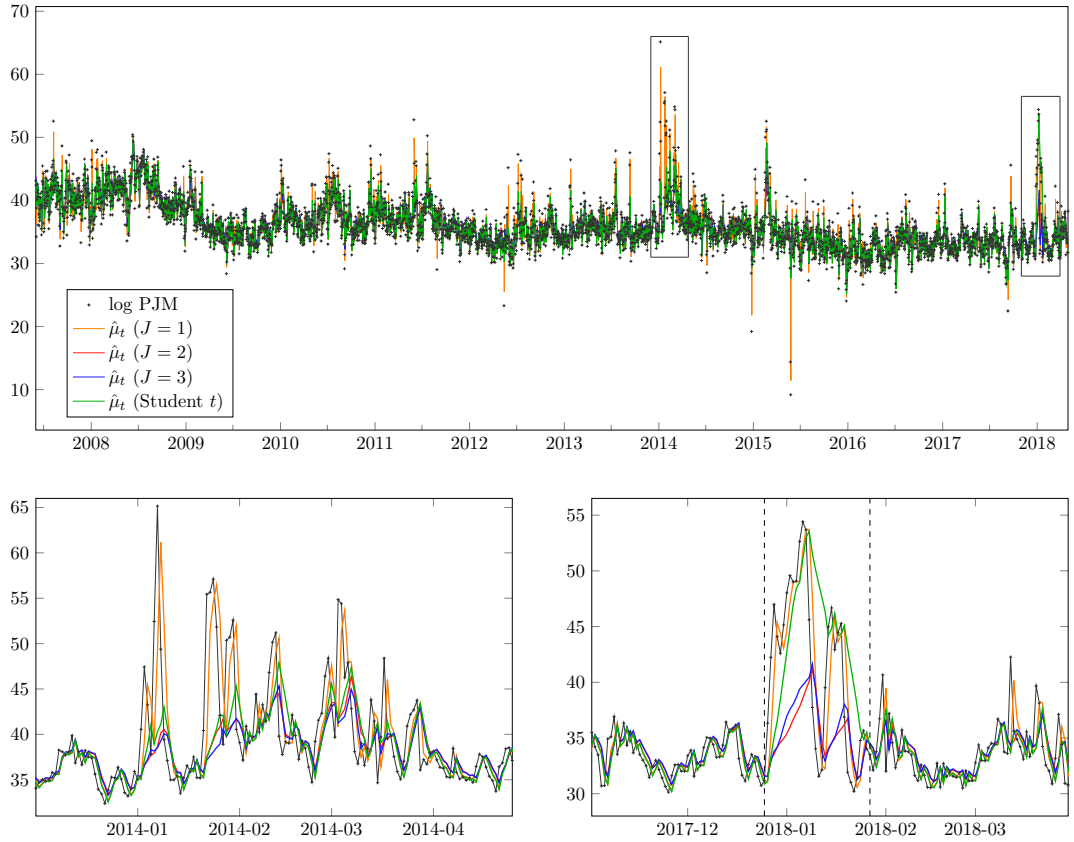

Figure 3. In the top panel we display the log daily electricity spot prices of PJM from June 1, 2007 to April 30, 2018, together with the filtered locations based on the fitted mixture models and the Student's t model. The two bottom panels present the same contents but for two sub-periods which are indicated by the squares in the top panel.

2 hold for the maximum likelihood estimates (MLEs), since we have $\alpha/\sigma_1^2 = 0.758 < 1$. For $J = 2$ and $J = 3$, a feasible version of the contraction condition of Proposition 1 holds at the MLEs for $r = 50$ and $r = 100$, respectively, which indicates that the filters of these fitted models are invertible. This is confirmed by the fact that the MLEs barely change when different starting values $\hat{\mu}_1$ are used. Also, for the models with $J = 2$ and $J = 3$ components, a feasible version of the contraction condition in Proposition 2 for $n = 4$ holds for $r = 90$ and $r = 225$, respectively. These results indicate that $g_t(\boldsymbol{\theta})$ has the required number of bounded moments needed for asymptotic normality as stated in Theorem 2. Hence, the reported standard errors in Table 3 are reliable, under the assumption of correct specification.

The MLEs for the model with $J = 2$ components show that the highest weight is given to the second component with a mean close to zero and a moderately sized variance while the first component has a low weight, a large mean and a very large variance, all in relative terms. These estimation results are typical when the mixture model is used to analyze data with many patches of spikes; see Figure 2. Similar interpretations can be given to

the MLEs for the model with $J = 3$ components. In case of the Student's t model, the MLE of the degrees-of-freedom parameter ν is equal to 6.2 which is sufficiently low to imply a heavy-tailed distribution for the innovations.

5.2. In-sample filtering

The filtered locations $\{\hat{\mu}_t\}_{t=1}^T$ for the four different models in our study are presented in Figure 3. For illustrative purposes, we have magnified two sub-periods where the time-series is more volatile and where the four score-driven updating functions behave distinctively different from each other. In the overall more tranquile periods, the filtered paths of all four fitted models are rather similar. However, in the first selected sub-period the linear filter, our model with $J = 1$, responds more heavily on temporary spikes in the prices, when compared to the other three models. The filtered path of the Student's t model is between the paths from the mixed normal with $J \geq 2$ and the path of the linear filter. The second sub-period also shows more volatile data. However, the filtered paths from the linear filter and the Student's t model respond strongly to the spikes in a similar way, while those from the mixed normal with $J \geq 2$ respond in a less pronounced way. The second sub-periods show spikes during a pro-longed time period and hence the former set of filters appear to represent the underlying location more precisely in this sub-period. On the other hand, in this sub-period there are also a substantive amount of lower values (second half of January 2018). The linear filter reacts to these lower values when they occur while the Student's t filter does not react. As we learn from Table 3, the maximized log-likelihood value is the highest for our model with $J = 3$. However, the linear filter clearly provides the highest log-likelihood contribution in the second sub-period: the log-likelihood values between the dashed lines in Figure 3 are -116.1, -159.9, -164.5, and -160.1, for the mixed models with $J = 1, 2, 3$ and the Student's t model, respectively. Although, the filter of the Student's t behaves somewhat more in tune with the linear filter, its contribution is similar to the robust mixed model filters with $J = 2, 3$. Hence, while the filters behave somewhat exceptional in the second sub-period, the mixed model with $J = 3$ provides our preferred filter.

The estimated innovation densities and corresponding scaled scores $\alpha s(x)$ are presented in Figure 4. The mixed normal densities (for $J = 2, 3$) have a different appearance from the normal and Student's t densities, but their shapes are not excessively different. From a closer look, we learn that the densities from the mixed normal models are skewed to the right while those of the other models are symmetric. These differences become more apparent from the corresponding score functions as presented in the right panel of Figure 4. The score plots show that especially moderately large positive values of $y_t - \hat{\mu}_t$ are downweighted more heavily by the mixed normal filters ($J = 2, 3$) compared to the

Table 3. Parameter estimates with their standard errors (s.e) for four different models, using daily log PJM electricity spot prices from June 1, 2007 to April 30, 2018.

$\boldsymbol{\theta}$	Normal ($J=1$)		Mixed N ($J=2$)		Mixed N ($J=3$)		Student's t	
	$\widehat{\boldsymbol{\theta}}_T$	s.e.	$\widehat{\boldsymbol{\theta}}_T$	s.e.	$\widehat{\boldsymbol{\theta}}_T$	s.e.	$\widehat{\boldsymbol{\theta}}_T$	s.e.
α	4.527	*0.180*	2.155	*0.117*	2.336	*0.113*	0.866	*0.037*
σ_1^2	6.022	*0.135*	32.585	*2.398*	66.731	*5.593*	3.988	*0.123*
σ_2^2			3.783	*0.110*	9.604	*1.315*		
σ_3^2					3.490	*0.137*		
c_1			3.487	*0.398*	5.169	*0.859*		
c_2			-0.291	-	2.145	*0.666*		
c_3					-0.415	-		
w_1			0.077	*0.008*	0.024	*0.004*		
w_2			0.923	-	0.111	*0.032*		
w_3					0.866	-		
ν							6.196	*0.429*
$\hat{L}_T(\boldsymbol{\theta})$	-9234.38		-9031.84		-8998.55		-9078.95	
AIC	18472.77		18073.68		18013.10		18163.9	
BIC	18485.35		18105.13		18063.42		18182.77	

Parameter estimates are reported for the mixed normal model (1)–(3) and for the Student's t model (4). The maximized log-likelihood values $\hat{L}_T(\boldsymbol{\theta})$ in equation (5), together with their corresponding information criteria AIC and BIC, are also reported.

linear filter ($J = 1$). The score functions of the Student's t model and the mixed normal models look somewhat similar for negative values. However, on the positive side of the real line, the Student's t filter has a higher response for values around 5. We notice that the score of the Student's t distribution converges to zero in the limit, while the score of the mixed normal models goes to infinity, eventually. Nevertheless, on the interval that we consider, the response to positive values is typically smaller for the mixture models than for the Student's t model.

5.3. Out-of-sample forecasting results

We complete the empirical study with a comparison in the accuracies of the out-of-sample density forecasts from the four models. The out-of-sample forecasts are obtained from a rolling window. We construct one-step-ahead density forecasts for the observations from May 1, 2018 to April 30, 2021 (3987 forecasts). The point forecast of y_{T+1}, based on the observations until T, is simply equal to $\hat{\mu}_{T+1}(\hat{\boldsymbol{\theta}}_T)$. We emphasize that for each

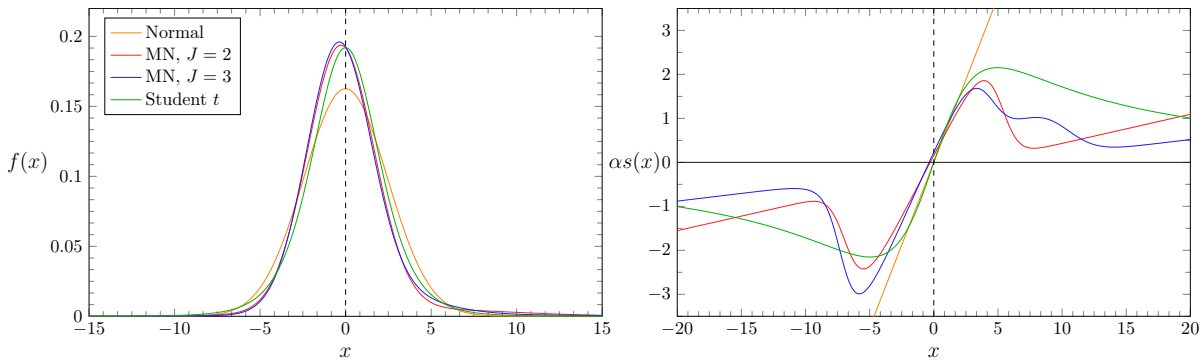

Figure 4. The estimated probability densities (left) and corresponding scaled score $\alpha s(x)$ (right) for the filters of the four different fitted models corresponding to estimates in Table 3.

forecast, the parameter vector is re-estimated using the rolling window of observations. The density forecasts are used to compute the *mean logarithmic scoring rule* (MLSR) as a measure of forecast accuracy; see Geweke and Amisano (2011). The MLSR is the average of the 3987 realized one-step ahead forecast log-densities. We notice that the difference of MLSRs between models can be regarded an approximation of the difference in Kullback-Leibler divergence between true and conditional densities of models; see, for example, Gneiting and Raftery (2007, Section 4.1). The MLSR values for each model are provided in Table 4. We also report the Diebold-Mariano (DM) test based on heteroscedasticity robust standard errors. The DM test verifies whether the MSLR values are significantly different from each other. We can conclude from Table 4 that the $J = 3$ model has a better MLSR than the $J = 2$ model. Also, both mixed normal models have a better MLSR than the Student's t model. All differences in MSLR are significantly different from zero at a 5% or even 1% level. The DM test applied to the logarithmic scoring rules of two models reduces to a likelihood-ratio test for the out-of-sample observations. Hence, the mixed normal models with $J = 2$ and $J = 3$ components have a significantly better out-of-sample log-likelihood value than the Student's t model. It is reassuring that this superiority remains to hold out-of-sample because it possibly counters the argument of overfitting in the discussion of the in-sample results.

6. Conclusion

We have introduced a novel score-driven mixed-normal location model for time-series observations. We treat effectively the time-varying mean of the mixed normal density as a non-stationary random walk process with a (potentially) non-zero intercept and with the score of the predicted mixed-normal density treated as the innovations. With respect to the general framework of score-driven models, our treatment of a mixture of normal model with a non-stationary process for location is innovative. Also, the

Table 4. Mean logarithmic scoring rule for forecasts of four different models using daily log PJM prices from May 1,2018 to April 30, 2021 and the Diebold-Mariano statistics.

		Normal ($J=1$)	Mixed N ($J=2$)	Mixed N ($J=3$)	Student's t
MLSR		-2.1900	-2.1202	-2.1054	-2.1387
DM	$J=2$	-3.233**			
	$J=3$	-4.132**	-3.585**		
	Student's t	-2.523*	2.158*	3.370**	

A positive/negative value for the Diebold-Mariano (DM) statistic corresponds to the model (in rows) having a lower/higher mean logarithmic scoring rule (MLSR) value than another model (in columns). The asterisks indicate significance at 5% (*) and 1% (**) levels.

theoretical developments in establishing consistency and asymptotic normality for the maximum likelihood estimator of the parameter vector are to some extent novel. The score-driven location model for the Student's t density can be viewed as an obvious competitive alternative to the mixed-normal model. However, in the context of a non-stationary location, the theoretical developments are not valid for the Student's t model due to its lack of an invertibility condition. We have shown in a Monte Carlo study that the mixed-normal location model is able to filter the true path of the time-varying mean very accurately in finite samples The selection of the correct number of mixture components can also be done accurately using likelihood-based information criteria, in case the different components are sufficiently identifiable. We further have shown the empirical relevance of our time-varying location model for an illustration of daily time-series of electricity prices. The empirical results are convincing from both in-sample and out-of-sample perspectives. In particular, the mixed-normal model outperforms both the linear filter and the Student's t model in terms of fit.

Appendix

A. Proofs of main results

Proof Proposition 1. (i) To prove the uniform convergence result $\sup_{\boldsymbol{\theta}\in\Theta} |\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})| \xrightarrow{e.a.s.} 0$, we follow the approach of (Straumann and Mikosch, 2006, Proposition 3.12). Note that we can write

$$\hat{g}_{t+1} = \tilde{\phi}_t(\hat{g}_t)\,,$$

initialized at some $\hat{g}_1(\boldsymbol{\theta}) = \hat{g}_1 \in \mathbb{R}$ for all $\boldsymbol{\theta} \in \Theta$, where $\tilde{\phi}_t$ is a random map defined by $\tilde{\phi}_t(g) = \phi_t(g(\cdot), \cdot)$ where $g \in \mathbb{C}(\Theta, \mathbb{R})$, for a compact set Θ. So we can view $\{\hat{g}_t\}_{t\in\mathbb{N}}$ initialized at $\hat{g}_1$ as a

sequence of random functions that lies in the separable Banach space $\mathbb{C}(\Theta, \mathbb{R})$ that is equipped with the norm $\|\cdot\|^{\Theta}$, where $\|g_t(\boldsymbol{\theta})\|^{\Theta} \equiv \sup_{\boldsymbol{\theta}\in\Theta} |g_t(\boldsymbol{\theta})|$ and $\|g_t(\boldsymbol{\theta})\|_n^{\Theta} \equiv (\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |g_t(\boldsymbol{\theta})|^n)^{1/n}$.

Under the assumption that the innovations $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ are i.i.d. together with the fact that ϕ_t is a continuous function of $\Delta y_{t+1} = \omega_0 + \alpha_0 s(\varepsilon_t; \boldsymbol{\psi}_0) + \varepsilon_{t+1} - \varepsilon_t$ (using the correct specification assumption) for every $(g, \boldsymbol{\theta}) \in \mathbb{R} \times \Theta$, it follows from Krengel (1985, Proposition 4.3) that $\{\tilde{\phi}_t(\cdot)\}_{t\in\mathbb{Z}}$ is stationary and ergodic. So we can apply Straumann and Mikosch (2006, Proposition 3.12). See also the proof of Proposition TA.4 in Blasques et al. (2021) for a more formal verification of this.

Proceeding in a similar manner as the proof in Blasques et al. (2018, Proposition 3.1), it follows from the mean value theorem that for any integer $r \geq 1$:

$$\sup_{g_1,g_2\in\mathbb{R},g_1\neq g_2} \frac{\left|\phi_t^{(r)}(g_1, \boldsymbol{\theta}) - \phi_t^{(r)}(g_2, \boldsymbol{\theta})\right|}{|g_1 - g_2|} \leq \sup_{g\in\mathbb{R}} \left|\dot{\phi}_t^{(r)}(g, \boldsymbol{\theta})\right| ,$$

which implies that the following conditions are sufficient to be able to apply Bougerol (1993, Theorem 3.1) and obtain the convergence to zero of $\|\hat{g}_t - g_t\|^{\Theta}$:

(a) $\mathbb{E}\log^+ \|\phi_t(\bar{g}, \cdot)\|^{\Theta} < \infty$ for some $\bar{g}$,

(b) $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\sup_{g\in\mathbb{R}}\log^+ |\dot{\phi}_t(g, \boldsymbol{\theta})| < \infty$,

(c) $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\sup_{g\in\mathbb{R}}\log |\dot{\phi}_t^{(r)}(g, \boldsymbol{\theta})| < 0$.

Condition (a) holds as it is implied by Lemma 1. Condition (b) holds as it is implied by Lemma 2. Finally, condition (c) holds by assumption as it is the same as condition (8) in the statement of the proposition. Therefore, it follows from (a), (b) and (c) that (Straumann and Mikosch, 2006, Proposition 3.12) can be applied, so $\|\hat{g}_t - g_t\|^{\Theta} \overset{e.a.s.}{\to} 0$ where $\{g_t\}_{t\in\mathbb{Z}}$ is a unique stationary and ergodic sequence.

Next, we show that $g_t(\boldsymbol{\theta})$ has a finite uniform log moment: $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\log^+ |g_t(\boldsymbol{\theta})| < \infty$. For convenience, we define $\rho_{r,t} = \sup_{\boldsymbol{\theta}\in\Theta}\sup_{g\in\mathbb{R}} |\dot{\phi}_t^{(r)}(g, \boldsymbol{\theta})|$. We notice that Lemma 2 implies that, for any given r, $\rho_{r,t}$ is bounded by the constant K^r. Furthermore, given a decreasing sequence of positive numbers $\{n_i\}_{\in\mathbb{N}}$ such that $n_i \to 0$ and $n_1 = 1$, we have that $\{K^r - (\rho_{r,t}^{n_i} - 1)/n_i\}_{i\in\mathbb{N}}$ is an increasing sequence of positive-valued random variables that converges a.s. to $K^r - \log \rho_{r,t}$. Therefore, an application of the monotone convergence theorem entails that $\lim_{n\to 0} \mathbb{E}(\rho_{r,t}^n - 1)/n = \mathbb{E}\log \rho_{r,t}$. This implies that there exists an $n > 0$ such that $\mathbb{E}\rho_{r,t}^n < 1$ since $\mathbb{E}\log \rho_{r,t} < 0$ by assumption. As a result, we obtain that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |g_t(\boldsymbol{\theta})|^n < \infty$ for some $n > 0$ by an application of Lemma 3. We therefore conclude that $\mathbb{E}\log^+ \sup_{\boldsymbol{\theta}\in\Theta} |g_t(\boldsymbol{\theta})| < \infty$.

(ii) The second result of the proposition follows straightforwardly from result *(i)*. We namely have that $\hat{g}_t(\boldsymbol{\theta}) = y_t - \hat{\mu}_t(\boldsymbol{\theta})$, so

$$\begin{aligned} |\hat{\mu}_t(\boldsymbol{\theta}_0) - \mu_t| &= |\hat{\mu}_t(\boldsymbol{\theta}_0) - y_t + y_t - \mu_t| \\ &= |-\hat{g}_t(\boldsymbol{\theta}_0) + \varepsilon_t| \\ &= |g_t(\boldsymbol{\theta}_0) - \hat{g}_t(\boldsymbol{\theta}_0) + \varepsilon_t - g_t(\boldsymbol{\theta}_0)| \end{aligned}$$

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

as $t \to \infty$, where the first term goes to zero e.a.s. by result *(i)* and where the second term goes to zero because $g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ with probability one. The reason for the latter fact is that the limit sequence $\{g_t(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is determined by the SRE in (7). Therefore, using the model equations (1) and (2), we have:

$$\begin{aligned} g_{t+1}(\boldsymbol{\theta}_0) &= g_t(\boldsymbol{\theta}_0) - \omega_0 - \alpha_0 s(g_t(\boldsymbol{\theta}_0); \boldsymbol{\psi}_0) + \Delta y_{t+1} \\ &= g_t(\boldsymbol{\theta}_0) + \alpha_0[s(\varepsilon_t; \boldsymbol{\psi}_0) - s(g_t(\boldsymbol{\theta}_0); \boldsymbol{\psi}_0)] + \Delta\varepsilon_{t+1}\,. \end{aligned}$$

Clearly, a solution to this SRE is $g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ for every t and because result *(i)* of this proposition states that the limit sequence $\{g_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ is unique, it must be that $g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ with probability one. □

Proof Proposition 2. (i) The e.a.s. convergence result follows directly from Proposition 1 *(i)*, because condition (8) holds whenever the contraction condition of this proposition holds. Namely, by Jensen's inequality, for any random variable x: $\mathbb{E}|x|^k < 1$ for some small $k > 0$ implies $\mathbb{E}\log|x| < 0$. The bounded moment result follows from Lemma 3.

(ii) and *(iii)*: The proof of *(ii)* and *(iii)* follows the same approach as the proof of Proposition 3.4 of Blasques et al. (2021), which is a similar proposition but then for a more general score-driven model. The only important difference is that here we do not have a uniform contraction, i.e. that $\partial\phi_t/\partial g$ is uniformly bounded between -1 and 1 in all its arguments, but we only have a contraction in expectation of the r-th iterates of the stochastic mapping functions.

Step 1: convergence. The convergence result can be shown by applying Theorem 2.10 of Straumann and Mikosch (2006) in virtually the same way as the proof of Proposition 3.4 in Blasques et al. (2021). The theorem considers a perturbed stochastic recurrence equation (SRE) $x_{t+1} = \hat{\phi}_t^*(x_t)$, where the sequence of maps $\{\hat{\phi}_t^*\}_{t\in\mathbb{N}}$ converges to a stationary limit $\{\phi_t^*\}_{t\in\mathbb{Z}}$. In the current setting, the perturbed SRE of the first derivative process corresponds to $\{\partial\hat{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}_{t\in\mathbb{N}}$, which is initialized at zero and depends on the initialized sequence $\{\hat{g}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$. The perturbed SRE of the second derivative process corresponds to $\{\partial^2\hat{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\}_{t\in\mathbb{N}}$, and is initialized at zero and depends on the initialized sequences $\{\hat{g}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ and $\{\partial\hat{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}_{t\in\mathbb{N}}$. The unperturbed SREs instead depend on the limit processes $\{g_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ and $\{\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}_{t\in\mathbb{Z}}$. The derivative processes are practically the same as those in Blasques et al. (2021) and can be described by the following equations:

$$\frac{\partial g_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = A_t^{(1)} + \frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} B_t\,, \qquad \frac{\partial^2 g_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = A_t^{(2)} + \frac{\partial^2 g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} B_t\,,$$

where $A_t^{(1)} = A^{(1)}(\boldsymbol{\theta}; g_t(\boldsymbol{\theta}))$ is a vector, $A_t^{(2)} = A^{(2)}(\boldsymbol{\theta}; g_t(\boldsymbol{\theta}), \partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$ is a matrix, and $B_t = B(\boldsymbol{\theta}; g_t(\boldsymbol{\theta}))$ is a scalar, which are defined in Technical Appendix C.2.

The conditions of Theorem 2.10 of Straumann and Mikosch (2006) regarding the convergence of the perturbed $\{\hat{\phi}_t^*\}_{t\in\mathbb{N}}$ to the stationary limit $\{\phi_t^*\}_{t\in\mathbb{Z}}$ can be verified by showing that

$$\sup_{\boldsymbol{\theta}\in\Theta} |B(\boldsymbol{\theta}; \hat{g}_t(\boldsymbol{\theta})) - B(\boldsymbol{\theta}; g_t(\boldsymbol{\theta}))| \overset{e.a.s.}{\to} 0\,,$$

$$\sup_{\boldsymbol{\theta}\in\Theta} |A_j^{(1)}(\boldsymbol{\theta}; \hat{g}_t(\boldsymbol{\theta})) - A_j^{(1)}(\boldsymbol{\theta}; g_t(\boldsymbol{\theta}))| \overset{e.a.s.}{\to} 0\,, \quad \text{and}$$

$$\sup_{\boldsymbol{\theta}\in\Theta} |A_{i,j}^{(2)}(\boldsymbol{\theta}; \hat{g}_t(\boldsymbol{\theta}), \partial\hat{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}) - A_{i,j}^{(2)}(\boldsymbol{\theta}; \hat{g}_t(\boldsymbol{\theta}), \partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta})| \overset{e.a.s.}{\to} 0\,,$$

where $A_j^{(1)}$ denotes the j-th element of the vector $A^{(1)}$ and $A_{i,j}^{(2)}$ denotes the i,j-th element of the matrix $A^{(2)}$. The convergence result can be shown in the same way as in the proof of Blasques et al. (2021, Proposition 3.4). The convergence of B_t and $A_t^{(1)}$ is shown using the mean value theorem. Because we know from *(i)* that $\sup_{\boldsymbol{\theta}\in\Theta} |\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})| \overset{e.a.s.}{\to} 0$, for *(ii)* it suffices to show that the derivatives of B_t and $A_t^{(1)}$ with respect to their second argument g is uniformly bounded in g and $\boldsymbol{\theta}$. In Lemma TA.1, it is shown that this is the case. More specifically, $\sup_{g\in\mathbb{R},\boldsymbol{\theta}\in\Theta} |\partial^2 s(g;\boldsymbol{\psi})/\partial g^2|$ can be bounded by a finite constant, which can be used to show the convergence of B_t. Also, $\sup_{g\in\mathbb{R},\boldsymbol{\theta}\in\Theta} |\partial s(g;\boldsymbol{\psi})/\partial g|$ and $\sup_{g\in\mathbb{R},\boldsymbol{\theta}\in\Theta} |\partial^2 s(g;\boldsymbol{\psi})/\partial\boldsymbol{\psi}_i\partial g|$ can be bounded by a finite constant, which is used to show the convergence of $A_t^{(1)}$. Finally, to show the convergence of $A_t^{(2)}$, we can use exactly the same approach as the proof of Lemma TA.17 of Blasques et al. (2021), using the asymptotic stationarity results of *(i)* and *(ii)* and that certain derivatives can be uniformly bounded. In Lemma TA.1, it is namely argued that $\sup_{g\in\mathbb{R},\boldsymbol{\theta}\in\Theta} |\partial^3 s(g;\boldsymbol{\psi})/\partial\boldsymbol{\psi}_i\partial\boldsymbol{\psi}_j'\partial g|$, $\sup_{g\in\mathbb{R},\boldsymbol{\theta}\in\Theta} |\partial^3 s(g;\boldsymbol{\psi})/\partial\boldsymbol{\psi}\partial g^2|$ and $\sup_{g\in\mathbb{R},\boldsymbol{\theta}\in\Theta} |\partial^2 s(g;\boldsymbol{\psi})/\partial g^3|$ can be bounded by a finite constant.

Furthermore, condition S.1 of Straumann and Mikosch (2006, Theorem 2.10) says that we need a bounded $\log^+$ moment for the unperturbed recurrence ϕ_t^* evaluated at some deterministic point. In other words we must show that $A_t^{(i)}$ and B_t evaluated at the stationary limit sequences $\{g_t(\boldsymbol{\theta})\}$ and $\{\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}$ have a finite $\log^+$ moment uniformly on Θ. This follows automatically from the fact that $A_t^{(1)}$ and B_t have $n > 0$ bounded moments, and $A_t^{(2)}$ has $n/2 > 0$ bounded moments, uniformly over Θ, which will be shown in Step 2 of the proof.

Finally, condition S.2 of Straumann and Mikosch (2006, Theorem 2.10) must hold. It was shown in the proof of Proposition 1, that it suffices to show that

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\sup_{g\in\mathbb{R}} \log^+ \|\dot{\phi}_t^*(g,\boldsymbol{\theta})\| < \infty \quad \text{and} \quad \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\sup_{g\in\mathbb{R}} \log \|\dot{\phi}_t^{*(r)}(g,\boldsymbol{\theta})\| < 0\,.$$

The first condition holds trivially, because the value of each of the elements of $\dot{\phi}_t^*$ is equal to $B(\boldsymbol{\theta}, g_t(\boldsymbol{\theta})) = 1 - \alpha s'(g_t(\boldsymbol{\theta}), \boldsymbol{\theta})$ for both derivative processes, and B_t is uniformly bounded. The second condition follows from the contraction in (9). Namely, it is not hard to see that for both the first and the second derivative process, the derivative of the r-th convolution of the corresponding stochastic mapping ϕ_t^* with respect to each single element is equal to

$$\prod_{i=0}^{r-1} B(\boldsymbol{\theta}; g_{t-i}(\boldsymbol{\theta}))\,,$$

which does not depend on $\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, but only on $g_t(\boldsymbol{\theta})$. Hence, the supremum over g can be dropped and the contraction condition simplifies to

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \log \left| \prod_{i=0}^{r-1} B(\boldsymbol{\theta}; g_{t-i}(\boldsymbol{\theta})) \right| < 0\,.$$

Upon closer inspection, it can be seen that this condition is weaker than condition (8) of Proposition 1. Namely, $\phi_t^{(r)}(g, \boldsymbol{\theta})$ has the same form, but then in the place of $g_{t-i}(\boldsymbol{\theta})$, there is $\phi_{t-i-1}^{(r-i)}(g, \boldsymbol{\theta})$. Hence, we obtain

$$\begin{aligned}
\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \log \left| \prod_{i=0}^{r-1} B(\boldsymbol{\theta}; g_{t-i}(\boldsymbol{\theta})) \right| &= \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \log \left| \prod_{i=0}^{r-1} B(\boldsymbol{\theta}; \phi_{t-i-1}^{(r-i)}(g_{t-r}(\boldsymbol{\theta}), \boldsymbol{\theta})) \right| \\
&\leq \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \sup_{g \in \mathbb{R}} \log \left| \prod_{i=0}^{r-1} B(\boldsymbol{\theta}; \phi_{t-i-1}^{(r-i)}(g, \boldsymbol{\theta})) \right| \\
&= \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \sup_{g \in \mathbb{R}} \log \left| \phi_t^{(r)}(g, \boldsymbol{\theta}) \right| < 0 \,,
\end{aligned}$$

where it was argued in *(i)* that the final inequality holds because of (9). This finishes the proof of the convergence result in *(ii)* and *(iii)*.

Step 2: bounded moments. We can almost directly apply the proof of Lemma 3 to the unperturbed systems of the derivative processes. If we denote by ϕ_t^* the stochastic mapping defining the SREs of the first or second derivative process, then we can use that by the discussion above we have that $|\dot{\phi}_t^{*(r)}(\boldsymbol{\theta})| \leq \sup_{g \in \mathbb{R}} |\dot{\phi}_t^{(r)}(g, \boldsymbol{\theta})| \equiv \boldsymbol{\rho}_t^{(r)}$. This is convenient, because by assumption we have $(\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} (\boldsymbol{\rho}_0^{(r)}(\boldsymbol{\theta}))^n)^{1/n} < 1$. This means we can straightforwardly apply the proof of Lemma 3, using this bounding argument, as long as we can show that

$$\|\phi_t^{*(r)}(\bar{g}, \boldsymbol{\theta})\|_{n^*}^{\Theta} < \infty \,,$$

which proves that the derivative process has n^* bounded moments.

We will proceed by showing that this moment condition holds for $n^* = n$ the first derivative process and for $n^* = n/2$ for the second derivative process, which will prove the bounded moment result in *(ii)* and *(iii)* respectively. Notice that

$$\phi_t^{*(r)}(\bar{g}, \boldsymbol{\theta}) = \sum_{j=0}^{r-1} \left(\prod_{k=0}^{j-1} B(\boldsymbol{\theta}; g_{t-k}(\boldsymbol{\theta})) \right) A_{t-j}^{(i)} + \bar{g} \prod_{j=0}^{r-1} B(\boldsymbol{\theta}; g_{t-j}(\boldsymbol{\theta})) \,,$$

if we let ϕ_t^* denote the mapping function of the first derivative process $i = 1$ or the second derivative process $i = 2$. By inspecting the expression of B_t in Technical Appendix C.2 it follows that B_t has bounded moments of any order, since $\partial s(g; \boldsymbol{\psi})/\partial g$ is uniformly bounded in g and $\boldsymbol{\theta}$. Hence, using the C_r-inequality of Loève (1977) in a similar way as in the proof of Proposition 3.4 of Blasques et al. (2021), it follows that to show that the moment condition holds, it suffices to show that the elements of $A_t^{(1)}$ and $A_t^{(2)}$ have n and $n/2$ bounded moments respectively.

For $A_{j,t}^{(1)}$, the j-th element of $A_t^{(1)}$, it is clear that the number of moments is equal to n, because in Lemma TA.1 it is given that $s(g; \boldsymbol{\psi})$ and the elements of $\partial s(g; \boldsymbol{\psi})/\partial \boldsymbol{\psi}$ can be bounded by $d_1 + d_2|g|$ for positive constants d_1 and d_2, and from *(i)* we know that $g_t(\boldsymbol{\theta})$ has n bounded moments. Finally, the i, j-th element of $A_t^{(2)}$, $A_{i,j,t}^{(2)}$ has $n/2$ bounded moments, because $\partial s(g; \boldsymbol{\psi})/\partial \boldsymbol{\theta}$ and $\partial^2 s(g; \boldsymbol{\psi})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$ have n bounded moments, $\partial s(g; \boldsymbol{\psi})/\partial g$, $\partial^2 s(g; \boldsymbol{\psi})/\partial \boldsymbol{\theta} \partial g$ and

$\partial^2 s(g;\boldsymbol{\psi})/\partial g^2$ are uniformly bounded. Therefore, using that we have that $\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ has n bounded moments by the result in *(ii)*, we obtain that

$$\frac{\partial s(g_t(\boldsymbol{\theta});\boldsymbol{\psi})}{\partial g}\frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\,, \quad \frac{\partial^2 s(g_t(\boldsymbol{\theta});\boldsymbol{\psi})}{\partial \boldsymbol{\theta}\partial g}\frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\,, \quad \frac{\partial^2 s(g_t(\boldsymbol{\theta});\boldsymbol{\psi})}{\partial g^2}\frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\,,$$

have n, n and $n/2$ bounded moments respectively. This finishes the proof. □

Proof of Theorem 1. First consider some notation. It is convenient to use the following average log-likelihood representation

$$\hat{L}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=2}^{T}\hat{\ell}_t(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=2}^{T}\ell(\hat{g}_t(\boldsymbol{\theta}),\boldsymbol{\psi}) \tag{A.11}$$

where $\hat{\ell}_t(\boldsymbol{\theta}) \equiv \ell(\hat{g}_t(\boldsymbol{\theta}),\boldsymbol{\psi}) \equiv \log p_y(y_t|\hat{\mu}_t(\boldsymbol{\theta});\boldsymbol{\psi})$ is the log-likelihood contribution of the t-th observation:

$$\hat{\ell}_t(\boldsymbol{\theta}) = \log p_y(y_t|\hat{\mu}_t(\boldsymbol{\theta});\boldsymbol{\psi}) = \log p_\varepsilon(\hat{g}_t(\boldsymbol{\theta});\boldsymbol{\psi}) = \log\left(\sum_{j=1}^{J}\frac{w_j}{\sigma_j}\exp\left(-\frac{(\hat{g}_t(\boldsymbol{\theta})-c_j)^2}{2\sigma_j^2}\right)\right)\,, \tag{A.12}$$

because $y_t - \hat{\mu}_t(\boldsymbol{\theta}) = \hat{g}_t(\boldsymbol{\theta})$. We ignore the constant $1/\sqrt{2\pi}$, because it is irrelevant for the maximization of the log-likelihood over $\boldsymbol{\theta}$. Define $L_T(\boldsymbol{\theta})$ as the average log-likelihood with $\hat{g}_t(\boldsymbol{\theta})$ replaced by $g_t(\boldsymbol{\theta})$ of Proposition 1 for every t:

$$L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=2}^{T}\ell_t(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=2}^{T}\ell(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})\,, \tag{A.13}$$

where $\ell_t(\boldsymbol{\theta}) \equiv \ell(g_t(\boldsymbol{\theta}),\boldsymbol{\psi}) \equiv \log p_\varepsilon(g_t(\boldsymbol{\theta});\boldsymbol{\psi})$. Also define $L(\boldsymbol{\theta}) \equiv \mathbb{E}\ \ell_1(\boldsymbol{\theta})$, where we note that $\{\ell_t(\boldsymbol{\theta})\}$ is stationary and ergodic by Proposition 4.3 in Krengel (1985) because $\{g_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ is a stationary and ergodic sequence by Proposition 1 and $\ell(g,\boldsymbol{\psi})$ is continuous in g for every $\boldsymbol{\theta}\in\Theta$.

Now turn to the real consistency proof. We use the same approach as in the proof of Blasques et al. (2018, Theorem 4.1), which is similar to that of Straumann and Mikosch (2006, Theorem 4.1). Following the proof of Theorem 4.1 of Blasques et al. (2018), the following conditions are sufficient for the strong consistency of the MLE $\hat{\boldsymbol{\theta}}_T$ to the true parameter value $\boldsymbol{\theta}_0$:

(A1) The function $\hat{L}_T$ converges almost surely to L_T uniformly over Θ:

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{L}_T(\boldsymbol{\theta}) - L_T(\boldsymbol{\theta})\right| \overset{a.s.}{\to} 0 \quad \text{as } T\to\infty\,.$$

(A2) $\sup_{(g,\boldsymbol{\theta})\in\mathbb{R}\times\Theta}\ell(g,\boldsymbol{\psi}) < \infty$ and $\mathbb{E}|\ell_1(\boldsymbol{\theta}_0)| < \infty$

(A3) The model is identifiable, so the parameter $\boldsymbol{\theta}_0$ is the unique maximizer of the limit log-likelihood $L(\boldsymbol{\theta})$, i.e. $L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}_0)$ for any $\boldsymbol{\theta}\in\Theta$, $\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$.

Below we complete the proof by showing that (A1)-(A3) hold.

(A1) Is satisfied by Lemma 4.

(A2) The first claim of this condition, namely that the log-likelihood is uniformly bounded from above over $\mathbb{R} \times \Theta$, is not hard to verify when looking at $\ell(g, \boldsymbol{\psi})$ which is defined in (A.12). We namely have that Θ is compact and that for every j, σ_j is bounded away from zero by *(iii)* of Assumption PS. Furthermore, for any $g \in \mathbb{R}$ and $\boldsymbol{\theta} \in \Theta$, it is clear that $\exp(-(g-c_j)^2/2\sigma_j^2)$ is bounded from above by 1. Hence, $\ell(g, \boldsymbol{\psi})$ is uniformly bounded from above by some finite value which depends on the bounds of Θ. Note that this uniform upper bound on $\ell(g, \boldsymbol{\psi})$ implies that $\mathbb{E}\, \ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) = L(\boldsymbol{\theta})$ exists for any $\boldsymbol{\theta} \in \Theta$ and that $L(\boldsymbol{\theta}) \in \{-\infty\} \cup \mathbb{R}$.
The second claim is that $\mathbb{E}|\ell(g_1(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)| < \infty$. Recall from the proof of Proposition 1 *(ii)* that $g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ with probability one for every t. Therefore, using the form of $\ell(g, \boldsymbol{\psi})$ given in (B.16):

$$\begin{aligned}\mathbb{E}\,|\ell(g_1(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)| &\leq \mathbb{E}\left|\frac{(g_1(\boldsymbol{\theta}_0) - c_{1,0})^2}{2\sigma_{1,0}^2}\right| + \mathbb{E}\left|\log\left(\frac{w_{1,0}}{\sigma_{1,0}} + B(g_1(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)\right)\right| \\ &\leq \mathbb{E}\left|\frac{(\varepsilon_1 - c_{1,0})^2}{2\sigma_{1,0}^2}\right| + C < \infty\end{aligned}$$

where we use that $|\log(w_{10}/\sigma_{10} + B(g, \boldsymbol{\psi}_0)|$ is uniformly bounded by some constant C by arguments given in the proof of Lemma 4. Furthermore, we use that ε_t has bounded moments of any order because it is mixed normally distributed and that Θ is compact and condition (i) of Assumption PS implies that $\sigma_{1,0}^2 > 0$.

(A3) By (A2) we know that $\ell_1(\boldsymbol{\theta})$ is uniformly bounded from above and therefore it is integrable, so in other words $L(\boldsymbol{\theta}) = \mathbb{E}\ell_1(\boldsymbol{\theta})$ exists for any $\boldsymbol{\theta} \in \Theta$. Furthermore, we know by (A2) that $|L(\boldsymbol{\theta}_0)| < \infty$ and that $L(\boldsymbol{\theta}) \in \{-\infty\} \cup \mathbb{R}$ for any other $\boldsymbol{\theta} \in \Theta$. For any $\boldsymbol{\theta} \in \Theta$ for which $L(\boldsymbol{\theta}) = -\infty$, it is therefore clear that $L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}_0)$. So we can from now on consider values of $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ for which $L(\boldsymbol{\theta}) \in \mathbb{R}$ and show that for those values $L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}_0)$ also holds.

A condition that implies uniqueness of $\boldsymbol{\theta}_0$ as a maximizer of $L(\boldsymbol{\theta})$ is that $\ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) = \ell(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)$ almost surely if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Lemma 5 says that this condition holds under the current assumptions. This condition implies identification by a standard argument (see for example the proof of Theorem 4.1 of Blasques et al. (2018)) which we will repeat here: The well-known inequality $\log(x) \leq x - 1$, which holds for any $x > 0$ and is only an equality for $x = 1$, implies that

$$\ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) - \ell(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0) \leq \frac{p_\varepsilon(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})}{p_\varepsilon(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)} - 1\,.$$

Clearly, if we can show that $\ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) = \ell(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)$ almost surely if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then the inequality above will be strict with a probability greater than zero for any $\boldsymbol{\theta} \in \Theta$ with $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Hence, if this condition holds, then for any $\boldsymbol{\theta} \in \Theta$ other than $\boldsymbol{\theta}_0$,

$$\mathbb{E}[\ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) - \ell(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)] < \mathbb{E}\left[\frac{p_\varepsilon(g_t(\boldsymbol{\theta}); \boldsymbol{\psi})}{p_\varepsilon(g_t(\boldsymbol{\theta}_0); \boldsymbol{\psi})}\right] - 1 = \mathbb{E}\left[\mathbb{E}\left[\frac{p_\varepsilon(g_t(\boldsymbol{\theta}); \boldsymbol{\psi})}{p_\varepsilon(\varepsilon_t; \boldsymbol{\psi}_0)}\Big|\mathcal{F}_{t-1}\right]\right] - 1 = 0\,,$$

using the law of total expectation, the fact that $g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ with probability one and that the conditional expectation has value one. The true conditional density function is namely given by $p_\varepsilon(\varepsilon_t; \boldsymbol{\psi}_0)$, which implies that

$$\begin{aligned}\mathbb{E}\left[\frac{p_\varepsilon(g_t(\boldsymbol{\theta}); \boldsymbol{\psi})}{p_\varepsilon(\varepsilon_t; \boldsymbol{\psi}_0)}\Big|\mathcal{F}_{t-1}\right] &= \int \frac{p_\varepsilon(h+\varepsilon_t; \boldsymbol{\psi})}{p_\varepsilon(\varepsilon_t; \boldsymbol{\psi}_0)} p_\varepsilon(\varepsilon_t; \boldsymbol{\psi}_0)\,\mathrm{d}\varepsilon_t\Big|_{h=g_t(\boldsymbol{\theta})-\varepsilon_t} \\ &= \int p_\varepsilon(h+\varepsilon_t; \boldsymbol{\psi})\,\mathrm{d}\varepsilon_t\Big|_{h=g_t(\boldsymbol{\theta})-\varepsilon_t} = 1\,,\end{aligned}$$

where we use that $g_t(\boldsymbol{\theta}) - \varepsilon_t$ is $\mathcal{F}_{t-1}$-measurable, because $\hat{g}_t(\boldsymbol{\theta}) - \varepsilon_t = \mu_t(\boldsymbol{\theta}_0) - \hat{\mu}_t(\boldsymbol{\theta})$ is clearly $\mathcal{F}_{t-1}$-measurable and will converge e.a.s. to $g_t(\boldsymbol{\theta}) - \varepsilon_t$ by the same arguments as used in the proof of Proposition 1 for the convergence of $\hat{g}_t(\boldsymbol{\theta})$ to $g_t(\boldsymbol{\theta})$. The final integral is equal to one because the integrand is a probability density function of a mixed normally distributed random variable with j-th component mean $c_j^* = c_j - h$. Therefore, we have that

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_0) = \mathbb{E}[\ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) - \ell(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)] < 0\,,$$

in case $\ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) = \ell(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)$ almost surely if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Hence, (A3) must hold, because Lemma 5 shows that this condition is satisfied here. □

Proof Theorem 2. This proof follows the same approach as the asymptotic normality result of Theorem 3.1 of Gorgi and Koopman (2021), which is based on the asymptotic normality proof in Section 7 of Straumann and Mikosch (2006). This approach starts by deriving the asymptotic distribution of the ML estimator $\tilde{\boldsymbol{\theta}}_T$, defined by

$$\tilde{\boldsymbol{\theta}}_T = \arg\max_{\boldsymbol{\theta}\in\Theta} L_T(\boldsymbol{\theta})\,,$$

where L_T is the limit log-likelihood, see (A.13). The final result is then derived by showing that $\tilde{\boldsymbol{\theta}}_T$ and $\hat{\boldsymbol{\theta}}_T$ have the same asymptotic distribution.

L_T is twice continuously differentiable in Θ, see Technical Appendix C.3 for the expressions of the first and second derivatives L_T' and L_T'', which are based on the limit processes $\{g_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$, $\{\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}_{t\in\mathbb{Z}}$ and $\{\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\}_{t\in\mathbb{Z}}$, which are stationary and ergodic by Proposition 2. It follows from Krengel (1985, Proposition 4.3) that L_T' and L_T'' are stationary and ergodic. Looking at the proof of Theorem 1, it is clear that $\tilde{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ lies in the interior of Θ by assumption. In the proof of Gorgi and Koopman (2021), it is argued that $\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \Omega)$ as $T \to \infty$ with $\Omega = -\mathbb{E}[\ell_t''(\boldsymbol{\theta}_0)]^{-1}$, if the following conditions hold:

(A) $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\ell_t''\| < \infty$,

(B) $-\mathbb{E}[\ell_t''(\boldsymbol{\theta}_0)]$ is positive definite,

(C) $\sqrt{T}L_T'(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \Omega^{-1})$.

It is shown that these three conditions hold in Lemmas 6, 7 and 8, respectively. Now it is argued by Gorgi and Koopman (2021) that given condition (A), the following condition is sufficient to show that the asymptotic distribution of $\tilde{\boldsymbol{\theta}}_T$ and $\hat{\boldsymbol{\theta}}_T$ are equal:

$$\sqrt{T}\sup_{\boldsymbol{\theta}\in\Theta}\|L_T'(\boldsymbol{\theta})-\hat{L}_T'(\boldsymbol{\theta})\| \overset{a.s.}{\to} 0\,,$$

which is proved to hold in Lemma 9. This finishes the proof. □

B. Lemmas

Lemma 1. *Let the assumptions of Proposition 1 hold. Then,* $\|\phi_t(\bar{g},\cdot)\|_n^{\Theta} < \infty$ *for any* $n > 0$ *and for any* $\bar{g} \in \mathbb{R}$.

Proof. For any $n \geq 1$, we can use the sub-additivity of the norm $\|\cdot\|_n^{\Theta}$. It follows immediately that the condition holds for $0 < n < 1$. We can take $n \geq 1$ and have

$$\begin{aligned}\|\phi_t(\bar{g},\cdot)\|_n^{\Theta} &\leq |\omega| + \sup_{\boldsymbol{\theta}\in\Theta}|\alpha|\cdot\|s(\bar{g};\boldsymbol{\psi})\|_n^{\Theta} + \|\Delta y_{t+1}\|_n \\ &\leq |\omega| + \sup_{\boldsymbol{\theta}\in\Theta}|\alpha|\cdot\|s(\bar{g};\boldsymbol{\psi})\|_n^{\Theta} + |\omega_0| + (\mathbb{E}|\varepsilon_{t+1}|^n)^{1/n} + (\mathbb{E}|\varepsilon_t|^n)^{1/n} + \\ &\quad |\alpha_0|\cdot\|s(\varepsilon_t;\boldsymbol{\psi}_0)\|_n\,.\end{aligned}$$

Since $\Theta \ni \boldsymbol{\theta}_0$ is compact and ε_t has bounded moments of any order, it suffices to show that $\|s(\varepsilon_t;\boldsymbol{\psi})\|_n^{\Theta}$ is finite (which implies $\|s(\bar{g};\boldsymbol{\psi}_0)\|_n$ is finite for some $\bar{g} \in \mathbb{R}$ too). Using the notation $f_j(x;\boldsymbol{\psi}) \equiv \exp(-(x-c_j)^2/2\sigma_j^2)w_j/\sigma_j$, we have

$$s(x;\boldsymbol{\psi}) = \frac{\sum_{j=1}^J \frac{x-c_j}{\sigma_j} f_j(x;\boldsymbol{\psi})}{\sum_{j=1}^J f_j(x;\boldsymbol{\psi})} = \frac{\frac{x-c_1}{\sigma_1} + \sum_{j=2}^J \frac{x-c_j}{\sigma_j^2}\frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}}{1+\sum_{j=2}^J \frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}} \leq \frac{x-c_1}{\sigma_1^2} + \sum_{j=2}^J \frac{x-c_j}{\sigma_j^2}\frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\,,$$

where we divide the numerator and denominator by $f_1(x;\boldsymbol{\psi})$, which is strictly positive for every $(x,\boldsymbol{\theta}) \in (\mathbb{R},\Theta)$, since Θ is compact and $w_1 \geq \kappa > 0$ for every $\boldsymbol{\theta} \in \Theta$. The inequality follows from the fact that the denominator $1+\sum_{j=2}^J f_j(x;\boldsymbol{\psi})/f_1(x;\boldsymbol{\psi})$ only attains values on the interval $[1,\infty)$ for any $(x,\boldsymbol{\theta}) \in (\mathbb{R},\Theta)$. Now we will argue that the sum from $j=2$ to J in the final expression can be uniformly bounded by a finite value over all $(x,\boldsymbol{\theta}) \in \mathbb{R}\times\Theta$. Namely, because $\sigma_1^2 - \sigma_j^2 \geq \kappa > 0$ and $w_j \geq \kappa > 0$ for any $\boldsymbol{\theta}$ in the compact set Θ, it is clear that for any $j = 2,\ldots,J$:

$$\lim_{|x|\to\infty}\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{x-c_j}{\sigma_j^2}\frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\right| = \lim_{|x|\to\infty}\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{w_j\sigma_1}{w_1\sigma_j}\frac{x-c_j}{\sigma_j^2}\exp\left(-\frac{1}{2}\left[\frac{(x-c_j)^2}{\sigma_j^2}-\frac{(x-c_1)^2}{\sigma_1^2}\right]\right)\right|,$$

equals zero by l'Hôpital's rule, because the $\exp(\cdot)$-term will go to zero at an exponential rate. By the compactness of Θ and because for all j, $\sigma_j^2 \geq \kappa > 0$, it follows that there exists a finite number C such that

$$\sup_{x\in\mathbb{R}}\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{j=2}^J \frac{x-c_j}{\sigma_j^2}\frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\right| \leq C < \infty\,.$$

Therefore, we have that:

$$\|s(\varepsilon_t;\boldsymbol{\psi})\|_n^{\Theta} = \left\|\frac{\varepsilon_t - c_1}{\sigma_1^2}\right\|_n^{\Theta} + C \leq \frac{1}{\kappa}(\|\varepsilon_t\|_n^{\Theta} + \sup_{\boldsymbol{\theta}\in\Theta}|c_1|) + C < \infty\,,$$

because ε_t has bounded moments of any order. Hence, the desired result holds. □

Lemma 2. *Let the assumptions of Proposition 1 hold. Then, there is a positive constant K such that* $\sup_{\boldsymbol{\theta}\in\Theta}\sup_{g\in\mathbb{R}}|\dot{\phi}_t(g,\boldsymbol{\theta})| \leq K < \infty$.

Proof. We have that $\dot{\phi}_t(g,\boldsymbol{\theta}) = 1 - \alpha\cdot\partial s(g,\boldsymbol{\psi})/\partial g$, so by the compactness of Θ, it will suffice to show that

$$\sup_{\boldsymbol{\theta}\in\Theta}\sup_{x\in\mathbb{R}}\left|\frac{\partial s(x,\boldsymbol{\psi})}{\partial x}\right| < \infty\,.$$

When again using the notation $f_j(x;\boldsymbol{\psi}) \equiv \exp(-(x-c_j)^2/2\sigma_j^2)w_j/\sigma_j$, we have

$$\begin{aligned}\frac{\partial s(x,\boldsymbol{\psi})}{\partial x} &= \left(\sum_{j=1}^{J} f_j(x;\boldsymbol{\psi})\right)^{-2}\left(\sum_{j=1}^{J} f_j(x;\boldsymbol{\psi})\right)\left(\sum_{j=1}^{J}\frac{1}{\sigma_j^2} f_j(x;\boldsymbol{\psi})\right)\\ &\quad - \left(\sum_{j=1}^{J} f_j(x;\boldsymbol{\psi})\right)^{-2}\sum_{j=1}^{J}\sum_{i=j+1}^{J}\left(\frac{x-c_j}{\sigma_j^2} - \frac{x-c_i}{\sigma_i^2}\right)^2 f_j(x;\boldsymbol{\psi}) f_i(x;\boldsymbol{\psi})\,.\end{aligned}$$

Dividing the numerator and the denominator by $f_1(x;\boldsymbol{\psi})$, so the factor of the component with the largest variance, leads to

$$\begin{aligned}\frac{\partial s(x,\boldsymbol{\psi})}{\partial x} &= \left(1+\sum_{j=2}^{J}\frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\right)^{-2}\left(1+\sum_{j=2}^{J}\frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\right)\left(\frac{1}{\sigma_1^2}+\sum_{j=2}^{J}\frac{1}{\sigma_j^2}\frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\right)\\ &\quad - \left(1+\sum_{j=2}^{J}\frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\right)^{-2}\sum_{j=1}^{J}\sum_{i=j+1}^{J}\left(\frac{x-c_j}{\sigma_j^2} - \frac{x-c_i}{\sigma_i^2}\right)^2 \frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\frac{f_i(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\,.\end{aligned}$$

We will argue that this expression can be bounded uniformly. The denominator is bounded from below by 1, because $f_j(x;\boldsymbol{\psi})/f_1(x;\boldsymbol{\psi}) \geq 0$ for every $x\in\mathbb{R}$ and $\boldsymbol{\theta}\in\Theta$. All the factors in the numerator can also be uniformly bounded, because as $|x|\to\infty$ none of the sums diverge and it can be seen that for any value $x\in\mathbb{R}$ all sums are finite uniformly over Θ, given the restrictions on Θ that are in place. For example, for any $j = 1,\ldots,J$ and $i > j$:

$$\begin{aligned}&\lim_{|x|\to\infty}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left(\frac{x-c_j}{\sigma_j^2} - \frac{x-c_i}{\sigma_i^2}\right)^2 \frac{f_j(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\frac{f_i(x;\boldsymbol{\psi})}{f_1(x;\boldsymbol{\psi})}\right|\\ &= \lim_{|x|\to\infty}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left(\left(\frac{1}{\sigma_j^2}-\frac{1}{\sigma_i^2}\right)x - \left(\frac{c_j}{\sigma_j^2}-\frac{c_i}{\sigma_i^2}\right)\right)^2 \frac{w_j w_i \sigma_1^2}{\sigma_j\sigma_i w_1^2}\right.\\ &\left.\times\exp\left(-\frac{1}{2}\left[\frac{(x-c_j)^2}{\sigma_j^2}+\frac{(x-c_i)^2}{\sigma_i^2}-2\frac{(x-c_1)^2}{\sigma_1^2}\right]\right)\right|\\ &= 0\,,\end{aligned}$$

because of the restrictions on the compact set Θ and in particular the fact that for every $\boldsymbol{\theta} \in \Theta$, $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_J^2$. More specifically, the $\exp(\cdot)$-factor will converge to zero at an exponential rate, while the term containing x^2 will only diverge at a quadratic rate. The convergence to zero can be shown formally by bounding the expression on Θ and applying l'Hôpital's rule. $\square$

Lemma 3. *Let the assumptions of Proposition 1 hold. Furthermore, assume that for some integer $r \geq 1$*

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \sup_{g \in \mathbb{R}} |\dot{\phi}_t^{(r)}(g, \boldsymbol{\theta})|^n < 1, \quad n > 0.$$

Then, the following uniform moment condition is satisfied

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |g_t(\boldsymbol{\theta})|^n < \infty.$$

Proof. This proof is similar to the proof of Proposition 3.1 of Blasques et al. (2021), but this proof is more general because we look at the r-th iterate and do not impose i.i.d. 'innovations' Δy_t. Consider the stationary limit sequence $\{g_t(\boldsymbol{\theta})\}_{t \in \mathbb{Z}}$ that exists by Proposition 1. We can bound $\|g_t(\boldsymbol{\theta})\|^\Theta$ as follows, for any $\bar{g} \in \mathbb{R}$:

$$\begin{aligned}
\|g_t(\boldsymbol{\theta})\|^\Theta &= \|\phi_{t-1}^{(r)}(g_{t-r}(\boldsymbol{\theta}), \boldsymbol{\theta})\|^\Theta \\
&\leq \|\phi_{t-1}^{(r)}(g_{t-r}(\boldsymbol{\theta}), \boldsymbol{\theta}) - \phi_{t-1}^{(r)}(\bar{g}, \boldsymbol{\theta})\|^\Theta + \|\phi_{t-1}^{(r)}(\bar{g}, \boldsymbol{\theta})\|^\Theta \\
&= \sup_{\boldsymbol{\theta} \in \Theta} \left(|g_{t-r}(\boldsymbol{\theta}) - \bar{g}| \times \frac{|\phi_{t-1}^{(r)}(g_{t-r}(\boldsymbol{\theta}), \boldsymbol{\theta}) - \phi_{t-1}^{(r)}(\bar{g}, \boldsymbol{\theta})|}{|g_{t-r}(\boldsymbol{\theta}) - \bar{g}|} \right) + \|\phi_{t-1}^{(r)}(\bar{g}, \boldsymbol{\theta})\|^\Theta \\
&\leq \|g_{t-r}(\boldsymbol{\theta}) - \bar{g}\|^\Theta \times \sup_{\boldsymbol{\theta} \in \Theta} \sup_{g_1, g_2 \in \mathbb{R}, g_1 \neq g_2} \frac{|\phi_{t-1}^{(r)}(g_1, \boldsymbol{\theta}) - \phi_{t-1}^{(r)}(g_2, \boldsymbol{\theta})|}{|g_1 - g_2|} + \|\phi_{t-1}^{(r)}(\bar{g}, \boldsymbol{\theta})\|^\Theta \\
&\leq \|g_{t-r}(\boldsymbol{\theta}) - \bar{g}\|^\Theta \times \sup_{\boldsymbol{\theta} \in \Theta} \sup_{g \in \mathbb{R}} |\dot{\phi}_t^{(r)}(g, \boldsymbol{\theta})| + \|\phi_{t-1}^{(r)}(\bar{g}, \boldsymbol{\theta})\|^\Theta \\
&\leq \sup_{\boldsymbol{\theta} \in \Theta} \boldsymbol{\rho}_{t-1}^{(r)}(\boldsymbol{\theta}) \cdot \|g_{t-r}(\boldsymbol{\theta})\|^\Theta + \sup_{\boldsymbol{\theta} \in \Theta} \boldsymbol{\rho}_{t-1}^{(r)}(\boldsymbol{\theta}) \cdot |\bar{g}| + \|\phi_{t-1}^{(r)}(\bar{g}, \boldsymbol{\theta})\|^\Theta,
\end{aligned}$$

where we use the subadditivity of the $\|\cdot\|^\Theta$-norm and the mean-value theorem. Also, we use the notation $\boldsymbol{\rho}_t^{(r)}(\boldsymbol{\theta}) \equiv \sup_{g \in \mathbb{R}} |\dot{\phi}_t^{(r)}(g, \boldsymbol{\theta})|$. Now unfold this recursion k steps backwards:

$$\begin{aligned}
\|g_t(\boldsymbol{\theta})\|^\Theta \leq &\left(\prod_{i=0}^{k-1} \sup_{\boldsymbol{\theta} \in \Theta} \boldsymbol{\rho}_{t-ri-1}^{(r)}(\boldsymbol{\theta}) \right) \cdot \|g_{t-rk}(\boldsymbol{\theta})\|^\Theta \\
&+ \sum_{j=0}^{k-1} \left(\prod_{i=0}^{j-1} \sup_{\boldsymbol{\theta} \in \Theta} \boldsymbol{\rho}_{t-ri-1}^{(r)}(\boldsymbol{\theta}) \right) \left(\sup_{\boldsymbol{\theta} \in \Theta} \boldsymbol{\rho}_{t-rj-1}^{(r)}(\boldsymbol{\theta}) \cdot |\bar{g}| + \|\phi_{t-rj-1}^{(r)}(\bar{g}, \boldsymbol{\theta})\|^\Theta \right).
\end{aligned} \tag{B.14}$$

Because $\phi_t(g, \boldsymbol{\theta}) = g - \omega - \alpha s(g; \boldsymbol{\psi}) + \Delta y_{t+1}$, it is clear that $\dot{\phi}_t(g, \boldsymbol{\theta})$ does not depend on Δy_t. It follows that $\boldsymbol{\rho}_t^{(r)}(\boldsymbol{\theta})$ does not depend on Δy_{t+1}, but only on $\Delta y_t, \ldots, \Delta y_{t-r+2}$. As was argued in the proof of Proposition 1, $\Delta y_t \perp \Delta y_{t-s}$ for $s \geq 2$, because the innovations sequence $\{\varepsilon_t\}_{t \in \mathbb{Z}}$ is i.i.d. It follows that for any $s \in \mathbb{Z}$, $\{\sup_{\boldsymbol{\theta} \in \Theta} \boldsymbol{\rho}_{s+rt}^{(r)}(\boldsymbol{\theta})\}_{t \in \mathbb{Z}}$ is an i.i.d. and therefore stationary and ergodic sequence. Furthermore, its elements are nonnegative random variables

with $\mathbb{E}\log\sup_{\boldsymbol{\theta}\in\Theta}\boldsymbol{\rho}_0^{(r)}(\boldsymbol{\theta}) < 0$ by the conditions of Proposition 1. Therefore, it follows from (Straumann and Mikosch, 2006, Lemma 2.4) that for every t we have that:

$$\prod_{i=0}^{k}\sup_{\boldsymbol{\theta}\in\Theta}\boldsymbol{\rho}_{t-ri-1}^{(r)}(\boldsymbol{\theta}) \overset{e.a.s.}{\to} 0\,, \qquad \text{as } k\to\infty\,.$$

Together with the fact that $\{\|g_t(\boldsymbol{\theta})\|^{\Theta}\}_{t\in\mathbb{Z}}$ is stationary and ergodic by the first part of this proposition, this implies that there exists some large $k\in\mathbb{N}$, such that

$$\left(\prod_{i=0}^{k-1}\sup_{\boldsymbol{\theta}\in\Theta}\boldsymbol{\rho}_{t-ri-1}^{(r)}(\boldsymbol{\theta})\right)\cdot\|g_{t-rk}(\boldsymbol{\theta})\|^{\Theta} < 1\,, \qquad \text{a.s.} \tag{B.15}$$

First consider the case where $n\geq 1$, so then $\|\cdot\|_n^{\Theta}$ is sub-additive. Clearly, showing that $\mathbb{E}\|g_t(\boldsymbol{\theta})\|_n^{\Theta}<\infty$ implies the result $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|g_t(\boldsymbol{\theta})|^n<\infty$. Now taking a large enough k such that (B.15) holds, we have by (B.14) and the sub-additivity of $\|\cdot\|_n^{\Theta}$, that

$$\begin{aligned}
\|g_t(\boldsymbol{\theta})\|_n^{\Theta} &\leq 1+\sum_{j=0}^{k-1}\left\|\left(\prod_{i=0}^{j-1}\sup_{\boldsymbol{\theta}\in\Theta}\boldsymbol{\rho}_{t-ri-1}^{(r)}(\boldsymbol{\theta})\right)\left(\sup_{\boldsymbol{\theta}\in\Theta}\boldsymbol{\rho}_{t-rj-1}^{(r)}(\boldsymbol{\theta})\cdot|\bar{g}|+\|\phi_{t-rj-1}^{(r)}(\bar{g},\boldsymbol{\theta})\|^{\Theta}\right)\right\|_n \\
&\leq 1+|\bar{g}|\cdot\sum_{j=0}^{k-1}\left(\prod_{i=0}^{j-1}\left(\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left(\boldsymbol{\rho}_{t-ri-1}^{(r)}(\boldsymbol{\theta})\right)^n\right)^{1/n}\right)\left(\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left(\boldsymbol{\rho}_{t-rj-1}^{(r)}(\boldsymbol{\theta})\right)^n\right)^{1/n} \\
&\quad+\sum_{j=0}^{k-1}\left(\prod_{i=0}^{j-2}\left(\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left(\boldsymbol{\rho}_{t-ri-1}^{(r)}(\boldsymbol{\theta})\right)^n\right)^{1/n}\right)\cdot\left\|\sup_{\boldsymbol{\theta}\in\Theta}\boldsymbol{\rho}_{t-r(j-1)-1}^{(r)}(\boldsymbol{\theta})\cdot\|\phi_{t-rj-1}^{(r)}(\bar{g},\boldsymbol{\theta})\|^{\Theta}\right\|_n \\
&\leq 1+((\bar{\rho}_n^{(r)})^2\ |\bar{g}|+K^r\|\phi_{t-rj-1}^{(r)}(\bar{g},\boldsymbol{\theta}))\|_n^{\Theta})\sum_{j=0}^{k-1}\left(\bar{\rho}_n^{(r)}\right)^{j-1} \\
&\leq 1+\frac{(\bar{\rho}_n^{(r)})^2\ |\bar{g}|+K^r\|\phi_{t-rj-1}^{(r)}(\bar{g},\boldsymbol{\theta}))\|_n^{\Theta}}{1-\bar{\rho}_n^{(r)}}<\infty\,,
\end{aligned}$$

where $\|\cdot\|_n\equiv(\mathbb{E}[\cdot]^n)^{1/n}$, $\bar{\rho}_n^{(r)}\equiv(\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}(\boldsymbol{\rho}_0^{(r)}(\boldsymbol{\theta}))^n)^{1/n}<1$, which holds by assumption and where K is the uniform bound in Lemma 2. Also, we use that $\{\sup_{\boldsymbol{\theta}\in\Theta}\boldsymbol{\rho}_{s+rt}^{(r)}(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence for any $s\in\mathbb{Z}$ and that $\phi_{s+rt}^{(r)}$ is independent of $\boldsymbol{\rho}_{s+r(t+i)}^{(r)}$ for all $i\in\mathbb{N}\setminus\{0,1\}$, by the independence of the innovations $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ [4]. The third inequality holds because $\sup_{\boldsymbol{\theta}\in\Theta}\boldsymbol{\rho}_{t-r(j-1)-1}^{(r)}(\boldsymbol{\theta})$ can be uniformly bounded by K^r because of Lemma 2. Lastly, Lemma 1 shows that $\|\phi_t(\bar{g},\boldsymbol{\theta})\|_n^{\Theta}<\infty$ for any $n>0$. Using the same approach (so using that $\sup_{g,\boldsymbol{\theta}}s(g,\boldsymbol{\psi})\leq d_1+d_2|g|$ for finite constants d_1 and d_2), it can be shown that for any integer $r\geq 1$ also $\|\phi_t^{(r)}(\bar{g},\boldsymbol{\theta})\|_n^{\Theta}<\infty$. This finishes the proof of $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|g_t(\boldsymbol{\theta})|^n<\infty$ for $n\geq 1$.

For $0<n<1$, $\|\cdot\|_n^{\Theta}$ is no longer sub-additive, but then the result can be proved by simply directly using $(\|\cdot\|_n^{\Theta})^n=\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|\cdot|^n$, which is sub-additive in that case. □

[4] More specifically, it can be verified that $\boldsymbol{\rho}_{s+r(t+i)}^{(r)}$ only depends directly on $\Delta y_{s+r(t+i)},\ldots,\Delta y_{s+r(t+i-1)+2}$, and that $\phi_{s+rt}^{(r)}$ only depends directly on $\Delta y_{s+rt+1},\ldots,\Delta y_{s+r(t-1)+2}$. So $\boldsymbol{\rho}_{s+r(t+i)}^{(r)}\perp\phi_{s+rt}^{(r)}$ for $i\geq 2$ because $\Delta y_t\perp\Delta y_{t-q}$ for any integer $q\geq 2$.

Lemma 4. *Let the assumptions of Theorem 1 hold. Then the function* $\hat{L}_T$ *defined in* (A.11) *converges almost surely to* L_T *defined in* (A.13) *uniformly over* Θ*:*

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{L}_T(\boldsymbol{\theta}) - L_T(\boldsymbol{\theta})\right| \overset{a.s.}{\to} 0 \quad \text{as } T\to\infty\,.$$

Proof. First is convenient to rewrite the log-likelihood contribution $\ell(g,\boldsymbol{\psi})$ defined implicitly in (A.12) in the following manner:

$$\begin{aligned}\ell(g,\boldsymbol{\psi}) &= \log\left(\exp\left(-\frac{(g-c_1)^2}{2\sigma_1^2}\right)\left[\frac{w_1}{\sigma_1}+\sum_{j=2}^{J}\frac{w_j}{\sigma_j}\exp\left(-\left[\frac{(g-c_j)^2}{2\sigma_j^2}-\frac{(g-c_1)^2}{2\sigma_1^2}\right]\right)\right]\right)\\ &= -\frac{(g-c_1)^2}{2\sigma_1^2}+\log\left(\frac{w_1}{\sigma_1}+\underbrace{\sum_{j=2}^{J}\frac{w_j}{\sigma_j}\exp\left(-\left[\frac{(g-c_j)^2}{2\sigma_j^2}-\frac{(g-c_1)^2}{2\sigma_1^2}\right]\right)}_{\equiv B(g,\boldsymbol{\psi})}\right). \qquad \text{(B.16)}\end{aligned}$$

By part (*iii*) of Assumption PS and because Θ is compact, it follows that $w_1/\sigma_1 + B(g,\boldsymbol{\psi})$ can be uniformly bounded from below and above over $\mathbb{R}\times\Theta$. The uniform upper bound follows from the mentioned assumption which says that σ_1^2 is strictly greater than all other σ_j^2's and that all σ_j^2's are bounded away from zero. So, the arguments in the $\exp(\cdot)$'s go to $-\infty$ as $|g|\to\infty$ uniformly over Θ, because $\sigma_1^2 \geq \sigma_j^2+\kappa$ for $j=1,2,\ldots$ and $\kappa>0$. The arguments in the $\exp(\cdot)$'s can be positive for intermediate values of $g+\varepsilon_t$ if $c_j\neq c_1$, but this value can be uniformly bounded by a finite constant because of the compactness of Θ. The uniform lower bound follows from part (*iv*) of Assumption PS, which says that for any $\boldsymbol{\theta}\in\Theta$ we have $w_j\geq\kappa>0$ for all j. Together with the fact that $B(g,\boldsymbol{\psi})$ only attains non-negative values, this assumption implies that the argument in the log, so $w_1/\sigma_1+B(g,\boldsymbol{\psi})$, can be uniformly bounded from below by $\kappa/\overline{\sigma}_1^2>0$, over $\mathbb{R}\times\Theta$. Here $\overline{\sigma}_1^2<\infty$ is the maximum value that σ_1^2 attains in the compact set Θ. In conclusion, the second term of $\ell(g,\boldsymbol{\psi})$ in (B.16) can be uniformly bounded between two finite values.

Now we can turn to showing that $\hat{L}_t$ converges to L_t almost surely uniformly over Θ. Using the expression for $\ell(g,\boldsymbol{\psi})$ in (B.16) and that $\hat{\ell}_t=\ell(\hat{g}_t,\boldsymbol{\psi})$ and $\ell_t=\ell(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})$, we have that

$$\begin{aligned}\hat{L}_T(\boldsymbol{\theta})-L_T(\boldsymbol{\theta}) &= \frac{1}{T}\sum_{t=2}^{T}-\frac{1}{2\sigma_1^2}\left((\hat{g}_t(\boldsymbol{\theta})-c_1)^2-(g_t(\boldsymbol{\theta})-c_1)^2\right)\\ &\quad+\frac{1}{T}\sum_{t=2}^{T}\left[\log\left(\frac{w_1}{\sigma_1}+B(\hat{g}_t(\boldsymbol{\theta}),\boldsymbol{\psi})\right)-\log\left(\frac{w_1}{\sigma_1}+B(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})\right)\right].\end{aligned}$$

Both these averages converges to zero uniformly almost surely over Θ. For the first term this holds true, because we can show that its terms converge to zero exponentially fast almost surely uniformly over Θ. It follows from (Straumann and Mikosch, 2006, Lemma 2.1) that the *sum* of these terms from $t=1$ to ∞ then converges almost surely, and therefore the *average* of these terms converges to zero almost surely. The terms converge to zero e.a.s. because

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{2\sigma_1^2}\left((\hat{g}_t(\boldsymbol{\theta})-c_1)^2-(g_t(\boldsymbol{\theta})-c_1)^2\right)\right| \leq \frac{1}{2\kappa^2}\sup_{\boldsymbol{\theta}\in\Theta}\left|(\hat{g}_t(\boldsymbol{\theta})-c_1)^2-(g_t(\boldsymbol{\theta})-c_1)^2\right|,$$

where we use that by part (*iii*) of Assumption PS, $\sigma_1^2 \geq \kappa > 0$ for any $\boldsymbol{\theta} \in \Theta$. It follows from Proposition 1 that

$$\sup_{\boldsymbol{\theta}\in\Theta} |(\hat{g}_t(\boldsymbol{\theta}) - c_1) - (g_t(\boldsymbol{\theta}) - c_1)| = \sup_{\boldsymbol{\theta}\in\Theta} |\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})| \overset{e.a.s.}{\to} 0\,,$$

as $t \to \infty$, where g_t has a uniform $\log^+$ moment, i.e. $\mathbb{E}\log^+ \sup_{\boldsymbol{\theta}\in\Theta} |g_t(\boldsymbol{\theta})| < \infty$. Therefore, it follows that the difference of the squares of $\hat{g}_t(\boldsymbol{\theta}) - c_1$ and $g_t(\boldsymbol{\theta}) - c_1$ also converges to zero uniformly over Θ, see Lemma TA.17 of Blasques et al. (2021).

We finalize the proof by arguing that the second average in (B.16) also converges to zero almost surely. Again we do this by showing that the terms of the average converge to zero e.a.s. We will do so by applying the mean value theorem twice. First apply the mean value theorem to $\log(w_1/\sigma_1^2 + x)$ to get

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta} &\left|\log\left(\frac{w_1}{\sigma_1} + B(\hat{g}_t(\boldsymbol{\theta}), \boldsymbol{\psi}\right) - \log\left(\frac{w_1}{\sigma_1} + B(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})\right)\right| \\ &\leq \sup_{\boldsymbol{\theta}\in\Theta, g\in\mathbb{R}} \left|\frac{1}{\frac{w_1}{\sigma_1} + B(g, \boldsymbol{\psi})}\right| \cdot \sup_{\boldsymbol{\theta}\in\Theta} |B(\hat{g}_t(\boldsymbol{\theta}), \boldsymbol{\psi}) - B(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})|\,,\end{aligned}$$

where $\sup_{\boldsymbol{\theta}\in\Theta} |1/(w_1/\sigma_1 + B(g, \boldsymbol{\psi}))| < \infty$ because in the discussion above we argued that $w_1/\sigma_1 + B(g, \boldsymbol{\psi})$ is bounded from below on $\mathbb{R} \times \Theta$ by a strictly positive value. Now apply the mean value theorem again to the continuously differentiable function $B(g, \boldsymbol{\psi})$ for fixed $\boldsymbol{\psi}$ to get:

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta} |B(\hat{g}_t(\boldsymbol{\theta}), \boldsymbol{\psi}) - B(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})| &\leq \sup_{\boldsymbol{\theta}\in\Theta, g\in\mathbb{R}} \left|\frac{\partial B(g, \boldsymbol{\psi})}{\partial g}\right| \cdot \sup_{\boldsymbol{\theta}\in\Theta} |\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})| \\ &\leq \bar{B}' \cdot \sup_{\boldsymbol{\theta}\in\Theta} |\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})| \overset{e.a.s.}{\to} 0\,,\end{aligned}$$

as $t \to \infty$. The convergence result follows from the e.a.s. convergence of $\hat{g}_t$ to g_t uniformly over Θ that follows from Proposition 1. Furthermore, we use that the derivative $\partial B(g, \boldsymbol{\psi})/\partial g$ is uniformly bounded over g and $\boldsymbol{\psi}$ by a finite constant $\bar{B}'$, because:

$$\left|\frac{\partial B(g, \boldsymbol{\psi})}{\partial g}\right| = \left|\sum_{j=2}^{J} -\frac{w_j}{\sigma_j}\left(\frac{g - c_j}{\sigma_j^2} - \frac{g - c_1}{\sigma_1^2}\right) \cdot \exp\left(-\left[\frac{(g - c_j)^2}{2\sigma_j^2} - \frac{(g - c_1)^2}{2\sigma_1^2}\right]\right)\right| < \bar{B}' < \infty\,,$$

where the bounding constant $\bar{B}'$ exists by similar arguments that we used to argue that the term $B(g, \boldsymbol{\psi})$ itself is bounded. More specifically, we have by part (*iii*) of Assumption PS that $\sigma_1^2 - \sigma_j^2 \geq \kappa > 0$, for every $j > 1$, which ensures that the factor $\exp(\cdot)$ converges to zero at an exponential rate as $|g| \to \infty$. The term in front of the $\exp(\cdot)$-function diverges as $|g| \to \infty$, but not at an exponential rate. Therefore, for every j the product converges to zero as $|g| \to \infty$, uniformly over Θ (this can be shown more formally using l'Hôpital's rule). Because of the compactness of Θ and the fact that the variances are bounded away from zero, it is now straightforward to see that the derivative can be uniformly bounded by some finite value $\bar{B}'$. This finishes the proof. □

Lemma 5. *Let the assumptions of Theorem 1 hold. Then*

$$\ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) = \ell(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0) \quad \textit{almost surely if and only if} \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0 \,.$$

Proof. We proceed in a way similar to the approach used, for example, by Blasques et al. (2021). Recall from the proof of Proposition 1(*ii*) that $g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ almost surely for every t. Our first step is to show that $\ell(h + \varepsilon_t, \boldsymbol{\psi}) = \ell(\varepsilon_t, \boldsymbol{\psi}_0)$ can only hold with probability one if $h = 0$ and $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. Because ε_t is mixed normally distributed with parameter $\boldsymbol{\psi}_0$ from Θ, ε_t has a positive density function on the entire real line $\mathbb{R}$. This implies that it is sufficient to show that for any $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \in \Theta$, $\ell(h + x, \boldsymbol{\psi}_1) = \ell(x, \boldsymbol{\psi}_2)$ can only hold for every $x \in \mathbb{R}$ if $h = 0$ and $\boldsymbol{\psi}_1 = \boldsymbol{\psi}_2$. Look at the expression of $\ell(g, \boldsymbol{\psi})$ given in (A.12). For any $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ from Θ and any $h \in \mathbb{R}$, it is clear that to have

$$\sum_{j=1}^{J} \frac{w_{j1}}{\sigma_{j1}} \exp\left(-\frac{(h + x - c_{j1})^2}{2\sigma_{j1}^2}\right) = \sum_{j=1}^{J} \frac{w_{j2}}{\sigma_{j2}} \exp\left(-\frac{(x - c_{j2})^2}{2\sigma_{j2}^2}\right),$$

for any $x \in \mathbb{R}$, we must have $\boldsymbol{\psi}_1 = \boldsymbol{\psi}_2$ and $h = 0$. By Assumption PS we namely have that the component variances $\sigma_j^2 > 0$ are such that $\sigma_1^2 > \cdots > \sigma_J^2$, that all weights are non-zero and that the J-th component mean c_J is such that $\sum_{j=1}^{J} w_j c_j = 0$. Hence, it follows from this discussion that for some t, $\ell(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0) = \ell(\varepsilon_t, \boldsymbol{\psi}_0) = \ell(h + \varepsilon_t, \boldsymbol{\psi})$ can only hold almost surely if $h = 0$ and $\boldsymbol{\psi} = \boldsymbol{\psi}_0$.

Therefore, to show that $\ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) = \ell(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)$ a.s. holds if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, it only remains to be shown that given that $\boldsymbol{\theta} = (\omega, \alpha, \boldsymbol{\psi}) \in \Theta$ is such that $\boldsymbol{\psi} = \boldsymbol{\psi}_0$, then $g_t(\boldsymbol{\theta}) = g_t(\boldsymbol{\theta}_0)$ almost surely if and only if $(\omega, \alpha) = (\omega_0, \alpha_0)$. Recall that $\{g_t(\boldsymbol{\theta})\}_{t \in \mathbb{Z}}$ is stationary and ergodic. Hence, we can use that if $g_t(\boldsymbol{\theta}) = g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ almost surely, for some t, then it holds for every $t \in \mathbb{Z}$. Now, given that $g_t(\boldsymbol{\theta}) = g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ almost surely, then $g_{t+1}(\boldsymbol{\theta})$ will satisfy

$$\begin{aligned} g_{t+1}(\boldsymbol{\theta}) &= \varepsilon_t - \omega - \alpha s(\varepsilon_t; \boldsymbol{\psi}_0) + \Delta y_{t+1} \\ &= \omega_0 - \omega + (\alpha_0 - \alpha) s(\varepsilon_t; \boldsymbol{\psi}_0) + \varepsilon_{t+1}, \end{aligned}$$

using the model equations (1) and (2) to work out Δy_{t+1}. First, we can argue that if $g_{t+1}(\boldsymbol{\theta}) = \varepsilon_{t+1}$ almost surely, then we must have $\omega = \omega_0$. Namely, if $\omega \neq \omega_0$, then we must have $(\alpha_0 - \alpha) s(\varepsilon_t; \boldsymbol{\psi}_0) = \omega - \omega_0 \neq 0$ almost surely. This implies that we must have that $\alpha_0 - \alpha$ and $s(\varepsilon_t; \boldsymbol{\psi}_0)$ are non-zero constants. However, $s(\varepsilon_t; \boldsymbol{\psi}_0)$ is not degenerate because clearly $\partial s(x; \boldsymbol{\psi}_0)/\partial x\big|_{x=\varepsilon_t} \neq 0$ for almost every ε_t. So we must have $\omega = \omega_0$. Because $\omega = \omega_0$ and $s(\varepsilon_t; \boldsymbol{\psi}_0)$ is non-zero with probability one, we can only have $g_{t+1}(\boldsymbol{\theta}) = \varepsilon_{t+1}$ almost surely if $\alpha = \alpha_0$. This finishes the proof. □

Lemma 6. *Let the assumptions of Theorem 2 hold. Then:*

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\ell_t''(\boldsymbol{\theta})\| < \infty \,,$$

where $\ell_t''(\boldsymbol{\theta}) = \partial^2 \ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$, as defined in Technical Appendix C.3 and $\|\cdot\|$ denotes the operator norm induced by the L_1-norm.

Proof. For notational convenience, we define the sup-norm as $\|\cdot\|^{\Theta} \equiv \sup_{\boldsymbol{\theta}\in\Theta}\|\cdot\|$ and we define the following functions: $\ell^{a}(g,\boldsymbol{\psi}) = \partial\ell(g,\boldsymbol{\psi})/\partial a$ and $\ell^{ab}(g,\boldsymbol{\psi}) = \partial^2\ell(g,\boldsymbol{\psi})/\partial a\partial b'$ with $a,b\in\{\boldsymbol{\theta},g\}$. Using the sub-additivity of the sup-norm

$$\begin{aligned}\mathbb{E}\left\|\ell_t''(\boldsymbol{\theta})\right\|^{\Theta} \leq \mathbb{E}\left\|\ell^{\boldsymbol{\theta}\boldsymbol{\theta}}(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})\right\|^{\Theta} + \mathbb{E}\left\|\ell^{\boldsymbol{\theta}g}(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right\|^{\Theta} + \mathbb{E}\left\|\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\ell^{g\boldsymbol{\theta}}(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})\right\|^{\Theta} \\ + \mathbb{E}\left\|\ell^{gg}(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right\|^{\Theta} + \mathbb{E}\left\|\ell^{g}(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})\frac{\partial^2 g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\|^{\Theta},\end{aligned}$$

so it suffices to show that each of these terms is bounded. By the assumptions in Theorem 2, Proposition 2 implies that

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|g_t(\boldsymbol{\theta})|^n < \infty, \quad \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\|^n < \infty, \quad \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\|^{n/2} < \infty,$$

for some $n\geq 4$. Furthermore, from the bounds of the log-likelihood derivatives given in Lemma TA.1, it follows that the first term is finite as $\ell_t''(\boldsymbol{\theta})$ has $n/2>1$ uniform bounded moments. To show the boundedness of the next terms, use the generalized Hölder's inequality, which says that if $\|\cdot\|_p = (\mathbb{E}\|\cdot\|^p)^{1/p}$, for random variables or vectors x and y, $\|x\cdot y\|_1 \leq \|x\|_p\|y\|_q$, where $p,q>0$ are such that $1 = p*q/(p+q)$. For the second and third term use the generalized Hölder inequality and the submultiplicativity of the sup-norm to see that:

$$\mathbb{E}\left\|\ell^{\boldsymbol{\theta}g}(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right\|^{\Theta} \leq \left\|\ell^{\boldsymbol{\theta}g}(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})\right\|_2^{\Theta}\left\|\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right\|_2^{\Theta} < \infty,$$

This expression is finite because both $\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\ell^{\boldsymbol{\theta}g}(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})$ have $n\geq 4$ bounded moments. The fourth term is also finite, because $\ell^{gg}(g,\boldsymbol{\psi})$ is uniformly bounded in g and $\boldsymbol{\theta}$, and $\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ has $n\geq 4$ bounded moments. Finally, for the last term apply the generalized Hölder's inequality, and use that $\ell^{g}(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})$ has $n\geq 4$ bounded moments and $\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ has $n/2\geq 2$ bounded moments. □

Lemma 7. *Let the assumptions of Theorem 2 hold. Then the Fischer information matrix*

$$\mathcal{I} = -\mathbb{E}[\ell_t''(\boldsymbol{\theta}_0)] = \mathbb{E}[\ell_t'(\boldsymbol{\theta}_0)\ell_t'(\boldsymbol{\theta}_0)^\top]$$

is positive definite.

Proof. We use a proof similar to that of Lemma A.5 of Gorgi and Koopman (2021). First of all, $-\mathbb{E}[\ell_t''(\boldsymbol{\theta}_0)] = \mathbb{E}[\ell_t'(\boldsymbol{\theta}_0)\ell_t'(\boldsymbol{\theta}_0)^\top]$ because of the Fischer information matrix equality. We namely assume that the model is correctly specified, so $\ell_t(\boldsymbol{\theta}_0)$ is the true log density evaluated at y_t and the second derivative of the log-likelihood function has a finite moment by Lemma 6. Therefore, we can obtain the result above via a standard argument.

Now turning to the positive definiteness result: $\mathbb{E}[\ell_t'(\boldsymbol{\theta}_0)\ell_t'(\boldsymbol{\theta}_0)^\top]$ is positive semi-definite by construction, so it only remains to be shown that it is invertible. This can be done by proving that

$$v^\top\ell_t'(\boldsymbol{\theta}_0) = v^\top\left[\begin{pmatrix}0\\0\\\frac{\partial\ell(\varepsilon_t,\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}}\end{pmatrix} + \begin{pmatrix}\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial\alpha}\\\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial\omega}\\\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\psi}}\end{pmatrix}\ell^{g}(\varepsilon_t,\boldsymbol{\psi}_0)\right] = 0,$$

almost surely, only if $v = 0$, where $v \in \mathbb{R}^{3J}$ and ℓ^g is defined in the proof of Lemma 6. We can split up the vector $v = (v_1', v_{-1}')'$ with $v_1 \in \mathbb{R}^2$ and $v_{-1} \in \mathbb{R}^{3J-2}$, such that v_1 corresponds to the derivative with respect to α and ω and v_{-1} corresponds to the derivative with respect to $\boldsymbol{\psi}$. Recall from the proof of Lemma 5 that $g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ almost surely. Furthermore, notice that $\ell^g(g, \boldsymbol{\psi}) = -s(g; \boldsymbol{\psi})$, so it is equal to the score function.

Similarly to the proof of Gorgi and Koopman (2021), we can now argue that $v^\top \ell_t'(\boldsymbol{\theta}_0) = 0$ almost surely, can only hold for $v \neq 0$, if we are in one of the following cases: (i) $v_1 \neq 0$ and $v_{-1} = 0$, (ii) $v_1 = 0$ and $v_{-1} \neq 0$, and (iii) $v_1 \neq 0$ and $v_{-1} \neq 0$. For (i) to hold, we must have

$$s(\varepsilon_t; \boldsymbol{\psi}_0) v_1^\top \begin{pmatrix} \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \alpha} \\ \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \omega} \end{pmatrix} = 0\,.$$

Looking at the score function, it is clear that $s(\varepsilon_t; \boldsymbol{\psi}_0) \neq 0$ with probability one. So for the equation to hold, we must have $v_{1,1}\partial g_t(\boldsymbol{\theta}_0)/\partial\alpha + v_{1,2}\partial g_t(\boldsymbol{\theta}_0)/\partial\omega = 0$ a.s. Note that the derivative processes of $\partial g_t(\boldsymbol{\theta}_0)/\partial\alpha$ and $\partial g_t(\boldsymbol{\theta}_0)/\partial\omega$ are virtually the same, but the former has 'innovation' $-s(\varepsilon_t; \boldsymbol{\psi})$ and the latter has 'innovation' 1, see Technical Appendix C.2. It is clear that both processes are not degenerate and since $s(\varepsilon_t; \boldsymbol{\psi}) \neq -1$ with probability one, the derivative processes are linearly independent. Hence, option (i) is ruled out.

For case (ii), we would need to have

$$v_{-1}^\top \left[\frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} (s(\varepsilon_t; \boldsymbol{\psi}_0))^{-1} - \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}} \right] = 0\,,$$

this is ruled out because the elements of the vector in brackets are linearly independent. To see this, first notice that $\partial g_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\psi}$ is $\mathcal{F}_{t-1}$-measurable, and $\partial\ell(\varepsilon_t, \boldsymbol{\psi}_0)/\partial\boldsymbol{\psi}$ and $s(\varepsilon_t; \boldsymbol{\psi}_0)$ are not. Next, look at the expression of the vector $\partial\ell(\varepsilon_t, \boldsymbol{\psi})/\partial\boldsymbol{\psi}$, which can be constructed using the building blocks in Technical Appendix C.1:

$$i = 1, \ldots, J-1: \qquad \frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi})}{\partial w_i} = \frac{\frac{1}{\sigma_i} \exp\left(-\frac{(\varepsilon_t - c_i)^2}{2\sigma_i^2}\right) - \left[1 - \frac{\varepsilon_t - c_J}{\sigma_J^2}(c_J - c_i)\right] \frac{1}{\sigma_J} \exp\left(-\frac{(\varepsilon_t - c_J)^2}{2\sigma_J^2}\right)}{\sum_{j=1}^J \frac{w_j}{\sigma_j} \exp\left(-\frac{(\varepsilon_t - c_j)^2}{2\sigma_j^2}\right)}\,,$$

$$i = 1, \ldots, J-1: \qquad \frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi})}{\partial c_i} = \frac{\frac{\varepsilon_t - c_i}{\sigma_i^2} \frac{w_i}{\sigma_i} \exp\left(-\frac{(\varepsilon_t - c_i)^2}{2\sigma_i^2}\right) - \frac{\varepsilon_t - c_J}{\sigma_J^2} \frac{w_i}{\sigma_J} \exp\left(-\frac{(\varepsilon_t - c_J)^2}{2\sigma_J^2}\right)}{\sum_{j=1}^J \frac{w_j}{\sigma_j} \exp\left(-\frac{(\varepsilon_t - c_j)^2}{2\sigma_j^2}\right)}\,,$$

$$i = 1, \ldots, J: \qquad \frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi})}{\partial \sigma_i^2} = \frac{\left[\frac{(\varepsilon_t - c_i)^2}{\sigma_i^2} - 1\right] \frac{w_i}{2\sigma_i^3} \exp\left(-\frac{(\varepsilon_t - c_i)^2}{2\sigma_i^2}\right)}{\sum_{j=1}^J \frac{w_j}{\sigma_j} \exp\left(-\frac{(\varepsilon_t - c_j)^2}{2\sigma_j^2}\right)}\,.$$

It is immediately clear that all the elements of $\partial\ell(\varepsilon_t, \boldsymbol{\psi}_0)/\partial\boldsymbol{\psi}$ are linearly independent given the identification conditions in Assumption PS, and they are also non-degenerate. Hence, there is no nonzero vector v_{-1} such that $v_{-1}^\top \partial\ell(\varepsilon_t, \boldsymbol{\psi}_0)/\partial\boldsymbol{\psi} = 0$ almost surely, and in effect, the same counts for $(s(\varepsilon_t; \boldsymbol{\psi}_0))^{-1} v_{-1}^\top \partial\ell(\varepsilon_t, \boldsymbol{\psi}_0)/\partial\boldsymbol{\psi} = 0$. This rules out case (ii), because $\partial g_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\psi}$ is independent of $\partial\ell(\varepsilon_t, \boldsymbol{\psi}_0)/\partial\boldsymbol{\psi}(s(\varepsilon_t; \boldsymbol{\psi}_0))^{-1}$.

Lastly, for case (iii), we would need to have

$$v_1^\top \begin{pmatrix} \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \alpha} \\ \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \omega} \end{pmatrix} + v_{-1}^\top \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}} = (s(\varepsilon_t; \boldsymbol{\psi}_0))^{-1} v_{-1}^\top \frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} ,$$

but this condition will not hold for a similar reason. The left hand side is namely $\mathcal{F}_{t-1}$-measurable and the right hand side is not, and the derivatives on both sides are not degenerate. This namely implies that for this condition to hold, we must have that both sides of the equation are equal to zero, and we just argued that the elements of $\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)/\partial \boldsymbol{\psi}$ are linearly independent, so there exists no $v_{-1} \in \mathbb{R}^{3J-2}$ for which the right hand side is equal to zero almost surely. This means case (iii) is also ruled out, which finishes the proof. □

Lemma 8. *Let the assumptions of Theorem 2 hold. Then:*

$$\sqrt{T} L_T'(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, K) \qquad \text{as } T \to \infty ,$$

with $K = \mathbb{E}[\ell_t'(\boldsymbol{\theta}_0)\ell_t'(\boldsymbol{\theta}_0)']$ *and where* L_T' *and* $\ell_t'(\boldsymbol{\theta}) = \partial \ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})/\partial \boldsymbol{\theta}$ *are defined in Technical Appendix C.3.*

Proof. We obtain this result by applying the Central Limit theorem for stationary and ergodic martingale difference sequences of Billingsley (1999). In order to be able to apply this theorem, we have to argue that $\{\ell_t'(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is a stationary and ergodic martingale difference sequence with a finite second moment.

It follows from (Krengel, 1985, Proposition 4.3) that $\{\ell_t'(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is stationary and ergodic, because ℓ_t' is a continuous function of $g_t(\boldsymbol{\theta})$ and $\partial g_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$, while Proposition 2 states both are elements of stationary and ergodic sequences. Furthermore, $\{\ell_t'(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is a martingale difference sequence since the model is assumed to be correctly specified and $\ell_t'(\boldsymbol{\theta}_0)$ is the conditional score, defined as the derivative of the conditional log-likelihood with respect to $\boldsymbol{\theta}$, evaluated at the true parameter value $\boldsymbol{\theta}_0$. The weak regularity conditions needed for this result are met here, because the observational density function is continuously differentiable and its derivative with respect to $\boldsymbol{\theta}$ can be uniformly bounded in all its arguments by some finite constant, which implies Leibniz integral rule can be applied.

Finally, we show that the second moment of $\ell_t'(\boldsymbol{\theta}_0)$ is finite, i.e. $\mathbb{E}\|\ell_t'(\boldsymbol{\theta}_0)\|^2 < \infty$, where $\|\cdot\|$ denotes the L_1-norm. This is equivalent to showing that $\|\ell_t'(\boldsymbol{\theta}_0)\|_2 < \infty$, where $\|\cdot\|_n \equiv (\mathbb{E}\|\cdot\|^n)^{1/n}$, which is sub-additive if $n \geq 1$. Hence:

$$\|\ell_t'(\boldsymbol{\theta}_0)\|_2 \leq \left\| \ell^{\boldsymbol{\theta}}(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0) \right\|_2 + \left\| \ell^{g}(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0) \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\|_2 ,$$

where $\ell^{\boldsymbol{\theta}}$ and ℓ^{g} are defined in the proof of Lemma 6. By Lemma TA.1 we know that the first derivative has $n/2$ bounded moments, where $n \geq 4$ are the number of bounded moments of $g_t(\boldsymbol{\theta})$ by the assumptions in Theorem 2 and the results in Proposition 2. Hence, the expectation is finite. For the second term, we can use the generalized Hölder's inequality, which says that for

random variables x and y, $\|x \cdot y\|_1 \leq \|x\|_p \|y\|_q$, where $p, q > 0$ are such that $n = p * q/(p+q)$. It follows that $\|x \cdot y\|_n \leq \|x\|_{np} \|y\|_{nq}$. Hence, we can bound the absolute second moment of the separate elements of the term as follows:

$$\left\| \ell^g(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0) \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_i} \right\|_2 \leq \|\ell^g(g_t(\boldsymbol{\theta}_0), \boldsymbol{\psi}_0)\|_4 \left\| \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_i} \right\|_4 .$$

Both terms on the right hand side are finite, because $\partial g_t(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$ has $n \geq 4$ bounded moments by the assumptions in Theorem 2 and the results in Proposition 2, and the other term also has n bounded moments, as is argued in Lemma TA.1. Alternatively, we could have used that $g_t(\boldsymbol{\theta}_0) = \varepsilon_t$ almost surely, and that $\ell^g(\varepsilon_t, \boldsymbol{\psi}_0)$ and $\partial g_t(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$ are independent. This finishes the proof. □

Lemma 9. *Let the assumptions of Theorem 2 hold. Then:*

$$\sqrt{T} \sup_{\boldsymbol{\theta} \in \Theta} \left\| \hat{L}'_T(\boldsymbol{\theta}) - L'_T(\boldsymbol{\theta}) \right\| \xrightarrow{a.s.} 0 \quad \text{as } T \to \infty ,$$

see Technical Appendix C.3 for the expression of the log-likelihood derivative. $\|\cdot\|$ *is the* L_1*-norm.*

Proof. For notational convenience, define the sup-norm as $\| \cdot \|^\Theta \equiv \sup_{\boldsymbol{\theta} \in \Theta} \| \cdot \|$. Similar to Lemma A.6 of Gorgi and Koopman (2021), we will show that the convergence result holds, by showing that

$$\|\hat{\ell}'_t(\boldsymbol{\theta}) - \ell'_t(\boldsymbol{\theta})\|^\Theta \xrightarrow{e.a.s.} 0 , \tag{B.17}$$

as $t \to \infty$, as this implies that

$$T \left\| \hat{L}'_T(\boldsymbol{\theta}) - L'_T(\boldsymbol{\theta}) \right\|^\Theta \leq \sum_{t=2}^{T} \sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\ell}'_t(\boldsymbol{\theta}) - \ell'_t(\boldsymbol{\theta})\|^\Theta < \infty ,$$

almost surely, by the subadditivity of the sup-norm and Lemma 2.1 of Straumann and Mikosch (2006). This in turn implies the final result $\sqrt{T}\|\hat{L}'_T(\boldsymbol{\theta}) - L'_T(\boldsymbol{\theta})\|^\Theta \xrightarrow{a.s} 0$.

To show the result in (B.17), consider the expression given in Technical Appendix C.3 and using the subadditivity of the sup-norm bound it as follows:

$$\|\hat{\ell}'_t(\boldsymbol{\theta}) - \ell'_t(\boldsymbol{\theta})\|^\Theta \leq \left\| \ell^{\boldsymbol{\theta}}(\hat{g}_t(\boldsymbol{\theta}), \boldsymbol{\psi}) - \ell^{\boldsymbol{\theta}}(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) \right\|^\Theta \tag{B.18}$$

$$+ \left\| \ell^g(\hat{g}_t(\boldsymbol{\theta}), \boldsymbol{\psi}) \frac{\partial \hat{g}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \ell^g(g_t(\boldsymbol{\theta}), \boldsymbol{\psi}) \frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\|^\Theta , \tag{B.19}$$

where ℓ^g and $\ell^{\boldsymbol{\theta}}$ are defined in the proof of Lemma 6. Hence, it suffices to show that both terms on the right hand side vanish e.a.s. as $t \to \infty$. For the second term, we use Corollary TA.16 of Blasques et al. (2021), which says it suffices to show that

$$\|\ell^g(\hat{g}_t(\boldsymbol{\theta}), \boldsymbol{\psi}) - \ell^g(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})\|^\Theta \xrightarrow{e.a.s.} 0 , \qquad \text{and} \qquad \left\| \frac{\partial \hat{g}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\|^\Theta \xrightarrow{e.a.s.} 0 ,$$

as long as $\ell^g(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})$ and $\partial g_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ are stationary and ergodic and have a finite $\log^+ \|\cdot\|$ moment, which is the case here. In particular, $\partial g_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is stationary and ergodic and has $n \geq 4$ finite moments by the assumptions in Theorem 2 and the results in Proposition 2. Furthermore, $\ell^g(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})$ is also stationary and ergodic by (Krengel, 1985, Proposition 4.3) and also has $n \geq 4$ moments by the bounding result given in Lemma TA.1. The convergence of the first derivative $\partial \hat{g}_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ to the stationary limit $\partial g_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ also follows directly from Proposition 2. Finally, use the mean value theorem in order to see that:

$$\|\ell^g(\hat{g}_t(\boldsymbol{\theta}), \boldsymbol{\psi}) - \ell^g(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})\|^{\Theta} \leq \sup_{\boldsymbol{\theta} \in \Theta} \sup_{g \in \mathbb{R}} |\ell^{gg}(g, \boldsymbol{\psi})| \cdot \|\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})\|^{\Theta} \overset{e.a.s.}{\to} 0\,,$$

where the convergence result follows from the fact that $\ell^{gg}(g, \boldsymbol{\psi})$ is uniformly bounded in g and $\boldsymbol{\theta}$, see Lemma TA.1, and $\|\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})\|^{\Theta}$ converges to zero e.a.s.

The first term of the right hand side of B.18 needs a more intricate treatment, because not all terms of the derivative vector $\ell^{\boldsymbol{\theta} g}(g, \boldsymbol{\psi})$ are uniformly bounded. To be more precise, all the terms are bounded except for the term corresponding to σ_1^2. For the terms that are uniformly bounded, we can again apply the mean value theorem in the same way. So it is not hard to see, that it now suffices to show that also $\partial \ell(\hat{g}_t(\boldsymbol{\theta}), \boldsymbol{\psi})/\partial \sigma_1^2$ converges to $\partial \ell(g_t(\boldsymbol{\theta}), \boldsymbol{\psi})/\partial \sigma_1^2$ e.a.s. uniformly over Θ. The expression of $\partial \ell(g, \boldsymbol{\psi})/\partial \sigma_1^2$ is given by

$$\frac{\partial \ell(g, \boldsymbol{\psi})}{\partial \sigma_1^2} = \frac{\left[\frac{(g-c_1)^2}{\sigma_1^2} - 1\right] \frac{w_1}{2\sigma_1^3}}{1 + \sum_{j=2}^{J} \frac{w_j}{\sigma_j} \exp\left(-\left(\frac{(g-c_j)^2}{2\sigma_j^2} - \frac{(g-c_1)^2}{2\sigma_1^2}\right)\right)}\,.$$

Notice that the denominator cannot be smaller than one given Assumption PS. Denote the numerator by $N(g; \boldsymbol{\psi})$ and the denominator by $D(g; \boldsymbol{\psi})$. We can apply Corollary TA.16 of Blasques et al. (2021) to show that $N(\hat{g}_t(\boldsymbol{\theta}); \boldsymbol{\psi})$ converges e.a.s. to $N(g_t(\boldsymbol{\theta}); \boldsymbol{\psi})$ uniformly over Θ and the same for the inverse of $D(g; \boldsymbol{\psi})$. We namely know that both factors are stationary and ergodic if evaluated at $g_t(\boldsymbol{\theta})$ by (Krengel, 1985, Proposition 4.3) and they both have a finite moment, because $g_t(\boldsymbol{\theta})$ has $n \geq 4$ bounded moments by Proposition 2, Θ is compact by Assumption PS and $D(g; \boldsymbol{\psi})$ is never smaller than 1 given Assumption PS. For the convergence of the numerator:

$$\begin{aligned}
\|N(\hat{g}_t(\boldsymbol{\theta}); \boldsymbol{\psi}) - N(g_t(\boldsymbol{\theta}); \boldsymbol{\psi})\|^{\Theta} &= \left\| \left[\frac{(\hat{g}_t(\boldsymbol{\theta}) - c_1)^2}{\sigma_1^2} - 1\right] \frac{w_1}{2\sigma_1^3} - \left[\frac{(g_t(\boldsymbol{\theta}) - c_1)^2}{\sigma_1^2} - 1\right] \frac{w_1}{2\sigma_1^3} \right\|^{\Theta} \\
&= C \left\| (\hat{g}_t(\boldsymbol{\theta}) - c_1)^2 - (g_t(\boldsymbol{\theta}) - c_1)^2 \right\|^{\Theta} \overset{e.a.s.}{\to} 0\,,
\end{aligned}$$

where C is some finite constant and where the convergence result follows from another application of Corollary TA.16 of Blasques et al. (2021), since $\|\hat{g}_t(\boldsymbol{\theta}) - c_1 - (g_t(\boldsymbol{\theta}) - c_1)\|^{\Theta} = \|\hat{g}_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta})\|^{\Theta} \overset{e.a.s.}{\to} 0$. Finally, to show that the inverse of $D(\hat{g}_t(\boldsymbol{\theta}); \boldsymbol{\psi})$ converges e.a.s. to the inverse of $D(g_t(\boldsymbol{\theta}); \boldsymbol{\psi})$, the mean value theorem approach from above can be applied, since it is straightforward to see that $\partial(1/D(g; \boldsymbol{\psi}))/\partial g$ is uniformly bounded over $\boldsymbol{\theta} \in \Theta$ and $g \in \mathbb{R}$, since:

$$\frac{\partial}{\partial g}\left(\frac{1}{D(g; \boldsymbol{\psi})}\right) = \frac{\partial}{\partial g} \frac{1}{1 + \sum_{j=2}^{J} \frac{w_j}{\sigma_j} \exp\left(-\left(\frac{(g-c_j)^2}{2\sigma_j^2} - \frac{(g-c_1)^2}{2\sigma_1^2}\right)\right)}$$

$$= \frac{\sum_{j=2}^{J} \frac{w_j}{\sigma_j} \left[\frac{g-c_j}{\sigma_j^2} - \frac{g-c_1}{\sigma_1^2} \right] \exp\left(- \left(\frac{(g-c_j)^2}{2\sigma_j^2} - \frac{(g-c_1)^2}{2\sigma_1^2} \right) \right)}{\left[1 + \sum_{j=2}^{J} \frac{w_j}{\sigma_j} \exp\left(- \left(\frac{(g-c_j)^2}{2\sigma_j^2} - \frac{(g-c_1)^2}{2\sigma_1^2} \right) \right) \right]^2},$$

which can be uniformly bounded given the parameter restrictions in Assumption PS, in particular because $\sigma_1^2 < \sigma_j^2$ for all $j \neq 1$. This finishes the proof. □

Technical Appendix

C. Derivatives

C.1. Derivatives of model expressions

In this section we will give the derivatives of $s(g;\boldsymbol{\psi})$ and $\ell(g,\boldsymbol{\psi})$ with respect to $\boldsymbol{\psi}$ and g and argue how they can be bounded. Let

$$F(g;\boldsymbol{\psi}) = \sum_{j=1}^{J} \frac{w_j}{\sigma_j} \exp\left(-\frac{(g-c_j)^2}{2\sigma_j^2}\right), \tag{C.20}$$

and let

$$G(g;\boldsymbol{\psi}) = -\frac{\partial F(g;\boldsymbol{\psi})}{\partial g} = \sum_{j=1}^{J} \frac{g-c_j}{\sigma_j^2}\frac{w_j}{\sigma_j} \exp\left(-\frac{(g-c_j)^2}{2\sigma_j^2}\right),$$

such that $\ell(g,\boldsymbol{\psi}) = \log F(g,\boldsymbol{\psi})$ and $s(g;\boldsymbol{\psi}) = -\partial\ell(g,\boldsymbol{\psi})/\partial g = G(g;\boldsymbol{\psi})/F(g;\boldsymbol{\psi})$. Let $G_k(g;\boldsymbol{\psi}) := \partial G(g;\boldsymbol{\psi})/\partial k$ where k can be g or the vector $\boldsymbol{\psi}$, and use similar notation for the function F and for the second and third derivatives. Then we have:

$$\begin{aligned}
\frac{\partial s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} &= \frac{G_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})} - \frac{G(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})^2},\\
\frac{\partial^2 s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}'} &= \frac{G_{\boldsymbol{\psi}\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})} - \frac{F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})G_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} - \frac{G_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} - \frac{G(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2}\\
&\quad + 2\frac{G(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^3},\\
\frac{\partial s(g;\boldsymbol{\psi})}{\partial g} &= \frac{G_g(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})} + \frac{(G(g;\boldsymbol{\psi}))^2}{(F(g;\boldsymbol{\psi}))^2},\\
\frac{\partial^2 s(g;\boldsymbol{\psi})}{\partial g^2} &= \frac{G_{gg}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})} + \frac{3G(g;\boldsymbol{\psi})G_g(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} + \frac{2(G(g;\boldsymbol{\psi}))^3}{(F(g;\boldsymbol{\psi}))^3},\\
\frac{\partial^3 s(g;\boldsymbol{\psi})}{\partial g^3} &= \frac{G_{ggg}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})} + \frac{4G(g;\boldsymbol{\psi})G_{gg}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} + \frac{3(G_g(g;\boldsymbol{\psi}))^2}{(F(g;\boldsymbol{\psi}))^2} + \frac{12(G(g;\boldsymbol{\psi}))^2G_g(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^3} + \frac{6(G(g;\boldsymbol{\psi}))^4}{(F(g;\boldsymbol{\psi}))^4},\\
\frac{\partial^2 s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial g} &= \frac{G_{\boldsymbol{\psi}g}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})} + \frac{2G(g;\boldsymbol{\psi})G_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} - \frac{F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})G_g(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} - \frac{2(G(g;\boldsymbol{\psi}))^2F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^3},\\
\frac{\partial^3 s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial g^2} &= \frac{G_{\boldsymbol{\psi}gg}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})} + \frac{3G(g;\boldsymbol{\psi})G_{\boldsymbol{\psi}g}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} - \frac{F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})G_{gg}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} + \frac{3G_g(g;\boldsymbol{\psi})G_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2}\\
&\quad + \frac{6(G(g;\boldsymbol{\psi}))^2G_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^3} - \frac{6G(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})G_g(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^3} - \frac{6(G(g;\boldsymbol{\psi}))^3F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^4},\\
\frac{\partial^3 s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}'\partial g} &= \frac{G_{\boldsymbol{\psi}\boldsymbol{\psi}'g}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})} + \frac{2G(g;\boldsymbol{\psi})G_{\boldsymbol{\psi}\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} - \frac{F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})G_{g\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} - \frac{G_{g\boldsymbol{\psi}}(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2}\\
&\quad - \frac{G_g(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} + \frac{2G_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})G_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2} - \frac{2(G(g;\boldsymbol{\psi}))^2F_{\boldsymbol{\psi}\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^3}\\
&\quad - \frac{4G(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})G_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^3} - \frac{4G(g;\boldsymbol{\psi})G_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^3}\\
&\quad + \frac{2F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})G_g(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^3} + \frac{6(G(g;\boldsymbol{\psi}))^2F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})F_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^4}.
\end{aligned}$$

The derivatives of the log likelihood function are given by

$$\frac{\partial \ell(g,\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \frac{F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})},$$

$$\frac{\partial^2 \ell(g,\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} = \frac{F_{\boldsymbol{\psi}\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{F(g;\boldsymbol{\psi})} - \frac{F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi}) F_{\boldsymbol{\psi}'}(g;\boldsymbol{\psi})}{(F(g;\boldsymbol{\psi}))^2}\,,$$

and the other derivatives of interest of the log likelihood function $\ell(g,\boldsymbol{\psi})$ are already known given the derivatives of $s(g;\boldsymbol{\psi})$ above:

$$\frac{\partial \ell(g,\boldsymbol{\psi})}{\partial g} = -s(g;\boldsymbol{\psi})\,, \qquad \frac{\partial^2 \ell(g,\boldsymbol{\psi})}{\partial g^2} = -\frac{\partial}{\partial g} s(g;\boldsymbol{\psi})\,, \qquad \frac{\partial^3 \ell(g,\boldsymbol{\psi})}{\partial g^3} = -\frac{\partial^2}{\partial g^2} s(g;\boldsymbol{\psi})\,,$$
$$\frac{\partial^2 \ell(g,\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial g} = -\frac{\partial}{\partial \boldsymbol{\psi}} s(g;\boldsymbol{\psi})\,, \qquad \frac{\partial^3 \ell(g,\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial g^2} = -\frac{\partial^2}{\partial \boldsymbol{\psi} \partial g} s(g;\boldsymbol{\psi})\,, \qquad \frac{\partial^3 \ell(g,\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}' \partial g} = -\frac{\partial^2}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} s(g;\boldsymbol{\psi})\,.$$

We will now list all the derivatives of F and G that occur above. We have to take into account that w_J and c_J are determined by the other weights and means in the derivatives. Remember that by Assumption PS we have that $w_J = 1 - \sum_{j=1}^{J-1} w_j$ and that $c_J = -\frac{\sum_{j=1}^{J-1} w_j c_j}{w_J}$, which implies that for every integer $1 \le i \le J-1$:

$$\frac{\partial w_J}{\partial w_i} = -1\,, \qquad \frac{\partial c_J}{\partial w_i} = \frac{1}{w_J}(c_J - c_i)\,, \qquad \frac{\partial c_J}{\partial c_i} = -\frac{w_i}{w_J}\,.$$

To keep the expressions readable, we introduce the notation $f_i(g) := \exp\left(-(g-c_i)^2/(2\sigma_i^2)\right)$. Note that $\partial f_i(g)/\partial c_i = (g-c_i) f_i(g)/\sigma_i^2$, $\partial f_i(g)/\partial \sigma_i^2 = (g-c_i)^2 f_i(g)/(2\sigma_i^4)$ and $\partial f_i(g)/\partial g = -(g-c_i) f_i(g)/\sigma_i^2$. Hence, the elements of $F_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})$ are given by:

$$w_i : i \le J-1, \qquad F_{w_i}(g;\boldsymbol{\psi}) = \frac{1}{\sigma_i} f_i(g) - \left(1 - \frac{g-c_J}{\sigma_J^2}(c_J - c_i)\right) \frac{1}{\sigma_J} f_J(g)\,,$$
$$c_i : i \le J-1, \qquad F_{c_i}(g;\boldsymbol{\psi}) = \frac{g-c_i}{\sigma_i^2}\frac{w_i}{\sigma_i} f_i(g) - \frac{g-c_J}{\sigma_J^2}\frac{w_i}{\sigma_J} f_J(g)\,,$$
$$\sigma_i^2 : i \le J, \qquad F_{\sigma_i^2}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 1\right) \frac{w_i}{2\sigma_i^3} f_i(g)\,.$$

The elements of $F_{\boldsymbol{\psi}\boldsymbol{\psi}'}(g;\boldsymbol{\psi})$ are given by:

$$w_i, w_j : i,j \le J-1, \qquad F_{w_i w_j}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_J)^2}{\sigma_J^2} - 1\right) \frac{(c_J - c_i)(c_J - c_j)}{w_J \sigma_J^3} f_J(g)\,,$$
$$c_i, c_i : i \le J-1, \qquad F_{c_i c_i}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 1\right) \frac{w_i}{\sigma_i^3} f_i(g) + \left(\frac{(g-c_J)^2}{\sigma_J^2} - 1\right) \frac{w_i^2}{\sigma_J^3 w_J} f_J(g)\,,$$
$$c_i, c_j : i,j \le J-1, j \ne i, \qquad F_{c_i c_j}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_J)^2}{\sigma_J^2} - 1\right) \frac{w_i w_j}{\sigma_J^3 w_J} f_J(g)\,,$$
$$\sigma_i^2, \sigma_i^2 : i \le J, \qquad F_{\sigma_i^2 \sigma_i^2}(g;\boldsymbol{\psi}) = \left(3 - 6\frac{(g-c_i)^2}{\sigma_i^2} + \frac{(g-c_i)^4}{\sigma_i^4}\right) \frac{w_i}{4\sigma_i^5} f_i(g)\,,$$
$$\sigma_i^2, \sigma_j^2 : i,j \le J, j \ne i, \qquad F_{\sigma_i^2 \sigma_j^2}(g;\boldsymbol{\psi}) = 0\,,$$
$$w_i, c_i : i \le J-1, \qquad F_{w_i c_i}(g;\boldsymbol{\psi}) = \frac{g-c_i}{\sigma_i^3} f_i(g) + \left(1 - \frac{(g-c_J)^2}{\sigma_J^2}\right) \frac{w_i(c_J - c_i)}{w_J \sigma_J^3} f_J(g) - \frac{g-c_J}{\sigma_J^3} f_J(g)\,,$$
$$w_i, c_j : i,j \le J-1, j \ne i, \qquad F_{w_i c_j}(g;\boldsymbol{\psi}) = \left(1 - \frac{(g-c_J)^2}{\sigma_J^2}\right) \frac{(c_J - c_i) w_j}{w_J \sigma_J^3} f_J(g)\,,$$
$$w_i, \sigma_i^2 : i \le J-1, \qquad F_{w_i \sigma_i^2}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 1\right) \frac{1}{2\sigma_i^3} f_i(g)$$
$$w_i, \sigma_j^2 : i,j \le J-1, j \ne i, \qquad F_{w_i \sigma_j^2}(g;\boldsymbol{\psi}) = 0\,,$$
$$w_i, \sigma_J^2 : i \le J-1, \qquad F_{w_i \sigma_J^2}(g;\boldsymbol{\psi}) = \left(1 - 3\frac{(g-c_J)(c_J - c_i)}{\sigma_J^2} - \left(1 - \frac{(g-c_J)(c_J - c_i)}{\sigma_J^2}\right) \frac{(g-c_J)^2}{\sigma_J^2}\right) \frac{1}{2\sigma_J^3} f_J(g)\,,$$
$$c_i, \sigma_i^2 : i \le J-1, \qquad F_{c_i \sigma_i^2}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 3\right) \frac{(g-c_i) w_i}{2\sigma_i^5} f_i(g)\,,$$
$$c_i, \sigma_j^2 : i,j \le J-1, j \ne i, \qquad F_{c_i \sigma_j^2}(g;\boldsymbol{\psi}) = 0\,,$$

$$c_i, \sigma_J^2 : i \leq J-1, \qquad F_{c_i\sigma_J^2}(g;\boldsymbol{\psi}) = \left(3 - \frac{(g-c_J)^2}{\sigma_J^2}\right)\frac{(g-c_J)w_i}{2\sigma_J^5} f_J(g)\,.$$

The elements of $G_{\boldsymbol{\psi}}(g;\boldsymbol{\psi})$ are given by:

$$\begin{aligned}
w_i : i \leq J-1, \qquad & G_{w_i}(g;\boldsymbol{\psi}) = \frac{g-c_i}{\sigma_i^3} f_i(g) - \left(g - c_i - \frac{(g-c_J)^2}{\sigma_J^2}(c_J - c_i)\right)\frac{1}{\sigma_J^3} f_J(g)\,,\\
c_i : i \leq J-1, \qquad & G_{c_i}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 1\right)\frac{w_i}{\sigma_i^3} f_i(g) - \left(\frac{(g-c_J)^2}{\sigma_J^2} - 1\right)\frac{w_i}{\sigma_J^3} f_J(g)\,,\\
\sigma_i^2 : i \leq J, \qquad & G_{\sigma_i^2}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 3\right)\frac{(g-c_i)w_i}{2\sigma_i^5} f_i(g)\,.
\end{aligned}$$

The elements of $G_{\boldsymbol{\psi}\boldsymbol{\psi}'}(g;\boldsymbol{\psi})$ are given by:

$$\begin{aligned}
w_i, w_j : i,j \leq J-1, \qquad & G_{w_iw_j}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_J)^2}{\sigma_J^2} - 3\right)\frac{(c_J-c_i)(c_J-c_j)(g-c_J)}{\sigma_J^5 w_J} f_J(g)\,,\\
c_i, c_i : i \leq J-1, \qquad & G_{c_ic_i}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 3\right)\frac{(g-c_i)w_i}{\sigma_i^5} f_i(g) + \left(\frac{(g-c_J)^2}{\sigma_J^2} - 3\right)\frac{(g-c_J)w_i^2}{\sigma_J^5 w_J} f_J(g)\,,\\
c_i, c_j : i,j \leq J-1, j \neq i, \qquad & G_{c_ic_j}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_J)^2}{\sigma_J^2} - 3\right)\frac{(g-c_J)w_iw_j}{\sigma_J^5 w_J} f_J(g)\,,\\
\sigma_i^2, \sigma_i^2 : i \leq J, \qquad & G_{\sigma_i^2\sigma_i^2}(g;\boldsymbol{\psi}) = \left(15 - 10\frac{(g-c_i)^2}{\sigma_i^2} + \frac{(g-c_i)^4}{\sigma_i^4}\right)\frac{(g-c_i)w_i}{4\sigma_i^7} f_i(g)\,,\\
\sigma_i^2, \sigma_j^2 : i,j \leq J, j \neq i, \qquad & G_{\sigma_i^2\sigma_j^2}(g;\boldsymbol{\psi}) = 0\,,\\
w_i, c_i : i \leq J-1, \qquad & G_{w_ic_i}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 1\right)\frac{1}{\sigma_i^3} f_i(g) - \left(\frac{(g-c_J)^2}{\sigma_J^2} - 1\right)\frac{1}{\sigma_J^3} f_J(g)\\
& \qquad + \left(3 - \frac{(g-c_J)^2}{\sigma_J^2}\right)\frac{(g-c_J)(c_J-c_i)w_i}{w_J\sigma_J^5} f_J(g)\,,\\
w_i, c_j : i,j \leq J-1, j \neq i, \qquad & G_{w_ic_j}(g;\boldsymbol{\psi}) = \left(3 - \frac{(g-c_J)^2}{\sigma_J^2}\right)\frac{(g-c_J)(c_J-c_i)w_j}{w_J\sigma_J^5} f_J(g)\,,\\
w_i, \sigma_i^2 : i \leq J-1, \qquad & G_{w_i\sigma_i^2}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 3\right)\frac{g-c_i}{2\sigma_i^5} f_i(g)\,,\\
w_i, \sigma_j^2 : i,j \leq J-1, j \neq i, \qquad & G_{w_i\sigma_j^2}(g;\boldsymbol{\psi}) = 0\,,\\
w_i, \sigma_J^2 : i \leq J-1, \qquad & G_{w_i\sigma_J^2}(g;\boldsymbol{\psi}) = \left(3(g-c_i) - 6\frac{(g-c_J)^2(c_J-c_i)}{\sigma_J^2}\right)\frac{1}{2\sigma_J^5} f_J(g)\\
& \qquad - \left(1 - \frac{(g-c_J)(c_J-c_i)}{\sigma_J^2}\right)\frac{(g-c_J)^3}{2\sigma_J^7} f_J(g)\,,\\
c_i, \sigma_i^2 : i \leq J-1, \qquad & G_{c_i\sigma_i^2}(g;\boldsymbol{\psi}) = \left(3 - 6\frac{(g-c_i)^2}{\sigma_i^2} + \frac{(g-c_i)^4}{\sigma_i^4}\right)\frac{w_i}{2\sigma_i^5} f_i(g)\,,\\
c_i, \sigma_j^2 : i,j \leq J-1, j \neq i, \qquad & G_{c_i\sigma_j^2}(g;\boldsymbol{\psi}) = 0\,,\\
c_i, \sigma_J^2 : i \leq J-1, \qquad & G_{c_i\sigma_J^2}(g;\boldsymbol{\psi}) = -\left(3 - 6\frac{(g-c_J)^2}{\sigma_J^2} + \frac{(g-c_J)^4}{\sigma_J^4}\right)\frac{w_i}{2\sigma_J^5} f_J(g)\,.
\end{aligned}$$

Next, $G_g(g;\boldsymbol{\psi})$, $G_{gg}(g;\boldsymbol{\psi})$ and $G_{ggg}(g;\boldsymbol{\psi})$ are given by:

$$\begin{aligned}
G_g(G;\boldsymbol{\psi}) &= \sum_{j=1}^{J}\left(1 - \frac{(g-c_j)^2}{\sigma_j^2}\right)\frac{w_j}{\sigma_j^3} f_j(g)\,,\\
G_{gg}(G;\boldsymbol{\psi}) &= \sum_{j=1}^{J}\left(\frac{(g-c_j)^2}{\sigma_j^2} - 3\right)\frac{(g-c_j)w_j}{\sigma_j^5} f_j(g)\,,\\
G_{ggg}(G;\boldsymbol{\psi}) &= \sum_{j=1}^{J}\left(-3 + 6\frac{(g-c_j)^2}{\sigma_j^2} - \frac{(g-c_j)^4}{\sigma_j^4}\right)\frac{w_j}{\sigma_j^5} f_j(g)\,.
\end{aligned}$$

The elements of $G_{\boldsymbol{\psi} g}(g;\boldsymbol{\psi})$ are given by

$$w_i, g : i \leq J-1, \qquad G_{w_i g}(g;\boldsymbol{\psi}) = \left(1 - \frac{(g-c_i)^2}{\sigma_i^2}\right)\frac{1}{\sigma_i^3} f_i(g) - \frac{1}{\sigma_J^3} f_J(g) + \left(g - c_J + \left(3 - \frac{(g-c_J)^2}{\sigma_J^2}\right)(c_J - c_i)\right)\frac{g-c_J}{\sigma_J^5} f_J(g)\,,$$

$$c_i, g : i \leq J-1, \qquad G_{c_i g}(g;\boldsymbol{\psi}) = \left(3 - \frac{(g-c_i)^2}{\sigma_i^2}\right)\frac{(g-c_i)w_i}{\sigma_i^5} f_i(g) - \left(3 - \frac{(g-c_J)^2}{\sigma_J^2}\right)\frac{(g-c_J)w_i}{\sigma_J^5} f_J(g)\,,$$

$$\sigma_i^2, g : i \leq J, \qquad G_{\sigma_i^2 g}(g;\boldsymbol{\psi}) = \left(-3 + 6\frac{(g-c_i)^2}{\sigma_i^2} - \frac{(g-c_i)^4}{\sigma_i^4}\right)\frac{w_i}{2\sigma_i^5} f_i(g)\,.$$

The elements of $G_{\boldsymbol{\psi} gg}(g;\boldsymbol{\psi})$ are given by

$$w_i, g, g : i \leq J-1, \qquad G_{w_i gg}(g;\boldsymbol{\psi}) = \left(\frac{(g-c_i)^2}{\sigma_i^2} - 3\right)\frac{g-c_i}{\sigma_i^5} f_i(g) + 3\frac{c_J - c_i}{\sigma_J^5} f_J(g) + \left(3 - 6\frac{(g-c_J)(c_J-c_i)}{\sigma_J^2} - \left(1 - \frac{(g-c_J)(c_J-c_i)}{\sigma_J^2}\right)\frac{(g-c_J)^2}{\sigma_J^2}\right)\frac{g-c_J}{\sigma_J^5} f_J(g)\,.$$

$$c_i, g, g : i \leq J-1, \qquad G_{c_i gg}(g;\boldsymbol{\psi}) = 3\frac{w_i}{\sigma_i^5} f_i(g) - \left(6 - \frac{(g-c_i)^2}{\sigma_i^2}\right)\frac{(g-c_i)^2 w_i}{\sigma_i^7} f_i(g) - 3\frac{w_i}{\sigma_J^5} f_J(g) - \left(6 - \frac{(g-c_J)^2}{\sigma_J^2}\right)\frac{(g-c_J)^2 w_i}{\sigma_J^7} f_J(g)\,,$$

$$\sigma_i^2, g, g : i \leq J, \qquad G_{\sigma_i^2 gg}(g;\boldsymbol{\psi}) = \left(15 - 10\frac{(g-c_i)^2}{\sigma_i^2} + \frac{(g-c_i)^4}{\sigma_i^4}\right)\frac{(g-c_i)w_i}{2\sigma_i^7} f_i(g)\,.$$

Finally, the elements of $G_{\boldsymbol{\psi}\boldsymbol{\psi}' g}(g;\boldsymbol{\psi})$ are given by

$$w_i, w_j, g : i, j \leq J-1, \qquad G_{w_i w_j g}(g;\boldsymbol{\psi}) = \left(-3 + 6\frac{(g-c_J)^2}{\sigma^2} - \frac{(g-c_J)^4}{\sigma^4}\right)\frac{(c_J-c_i)(c_J-c_j)}{\sigma_J^5 w_J} f_J(g)\,,$$

$$c_i, c_i, g : i \leq J-1, \qquad G_{c_i c_i g}(g;\boldsymbol{\psi}) = \left(-3 + 6\frac{(g-c_i)^2}{\sigma_i^2} - \frac{(g-c_i)^4}{\sigma_i^4}\right)\frac{w_i}{\sigma_i^5} f_i(g) + \left(-3 + 6\frac{(g-c_J)^2}{\sigma_J^2} - \frac{(g-c_J)^4}{\sigma_J^4}\right)\frac{w_i^2}{\sigma_J^5 w_J} f_J(g)\,,$$

$$c_i, c_j, g : i, j \leq J-1, j \neq i, \qquad G_{c_i c_j g}(g;\boldsymbol{\psi}) = \left(-3 + 6\frac{(g-c_J)^2}{\sigma_J^2} - \frac{(g-c_J)^4}{\sigma_J^4}\right)\frac{w_i w_j}{\sigma_J^5 w_J} f_J(g)$$

$$\sigma_i^2, \sigma_i^2, g : i \leq J, \qquad G_{\sigma_i^2\sigma_i^2 g}(g;\boldsymbol{\psi}) = \left(15 - 45\frac{(g-c_i)^2}{\sigma_i^2} + 15\frac{(g-c_i)^4}{\sigma_i^4} - \frac{(g-c_i)^6}{\sigma_i^6}\right)\frac{w_i}{4\sigma_i^7} f_i(g)$$

$$\sigma_i^2, \sigma_j^2, g : i, j \leq J, j \neq i, \qquad G_{\sigma_i^2\sigma_j^2 g}(g;\boldsymbol{\psi}) = 0\,,$$

$$w_i, c_i, g : i \leq J-1, \qquad G_{w_i c_i g}(g;\boldsymbol{\psi}) = \left(3 - \frac{(g-c_i)^2}{\sigma_i^2}\right)\frac{g-c_i}{\sigma_i^5} f_i(g) - \left(3 - \frac{(g-c_J)^2}{\sigma_J^2}\right)\frac{g-c_J}{\sigma_J^5} f_J(g) + \left(3 - 6\frac{(g-c_J)^2}{\sigma_J^2} + \frac{(g-c_J)^4}{\sigma_J^4}\right)\frac{(c_J-c_i)w_i}{w_J\sigma_J^5} f_J(g)\,,$$

$$w_i, c_j, g : i, j \leq J-1, j \neq i, \qquad G_{w_i c_j g}(g;\boldsymbol{\psi}) = \left(3 - 6\frac{(g-c_J)^2}{\sigma_J^2} + \frac{(g-c_J)^4}{\sigma_J^4}\right)\frac{(c_J-c_i)w_j}{w_J\sigma_J^5} f_J(g)\,,$$

$$w_i, \sigma_i^2, g : i \leq J-1, \qquad G_{w_i\sigma_i^2 g}(g;\boldsymbol{\psi}) = \left(-3 + 6\frac{(g-c_i)^2}{\sigma_i^2} - \frac{(g-c_i)^4}{\sigma_i^4}\right)\frac{1}{2\sigma_i^5} f_i(g)$$

$$w_i, \sigma_j^2, g : i, j \leq J-1, j \neq i, \qquad G_{w_i\sigma_j^2 g}(g;\boldsymbol{\psi}) = 0\,,$$

$$w_i, \sigma_J^2, g : i \leq J-1, \qquad G_{w_i\sigma_J^2 g}(g;\boldsymbol{\psi}) = \left(1 - \frac{(g-c_J)}{\sigma_J^2}\left(2(g-c_i) + 3(c_J-c_i)\right)\right)\frac{3}{2\sigma_J^5} f_J(g) + \left((g-c_i) + (c_J-c_i)\left(9 - \frac{(g-c_J)^2}{\sigma_J^2}\right)\right)\frac{(g-c_J)^3}{2\sigma_J^9} f_J(g)\,,$$

$$
\begin{aligned}
&c_i, \sigma_i^2, g : i \leq J-1, && G_{c_i \sigma_i^2 g}(g; \boldsymbol{\psi}) = \left(-15 + 10 \frac{(g-c_i)^2}{\sigma_i^2} - \frac{(g-c_i)^4}{\sigma_i^4} \right) \frac{(g-c_i) w_i}{2\sigma_i^7} f_i(g) \,, \\
&c_i, \sigma_j^2, g : i, j \leq J-1, j \neq i, && G_{c_i \sigma_j^2 g}(g; \boldsymbol{\psi}) = 0 \,, \\
&c_i, \sigma_J^2, g : i \leq J-1, && G_{c_i \sigma_J^2 g}(g; \boldsymbol{\psi}) = - \left(-15 + 10 \frac{(g-c_J)^2}{\sigma_J^2} - \frac{(g-c_J)^4}{\sigma_J^4} \right) \frac{(g-c_J) w_i}{2\sigma_J^7} f_J(g)
\end{aligned}
$$

The resulting expressions of the derivatives of s and l typically do not simplify to convenient forms, so we will refrain from filling in the derivatives of F and G to obtain them. However, it is straightforward to check whether the expressions can be bounded by a constant or by $a|g|^k + b$ for some k as follows. By Assumption PS we namely have that σ_1^2 is the largest component variance. So if we write the derivatives of s and l as one big fraction and we divide and multiply the resulting expression by $f_1(g)$ to the power of F in the denominator, say k, then the resulting denominator is bounded away from zero. It is clear that the only terms of the resulting numerator that can cause one of the derivatives to *not* be uniformly bounded, are the terms containing $f_1(g)$ k times. We namely know that $g^p f_j(g)/f_1(g)$ will converge to zero as $|g| \to 0$ for any $p \in \mathbb{R}$ and any $j > 1$, because $f_j(g)/f_1(g)$ will go to zero at an exponential rate. Hence, all of the derivatives with respect to w_i, c_i and σ_i^2 for $i > 1$ can be disregarded, because they can be trivially bounded in this way, as they do not have $f_1(g)$ in their numerator k times. For the derivatives with respect to w_1, c_1 and/or σ_1^2 that *do* contain $f_1(g)$ k times in the numerator, it turns out that the terms containing $f_1(g)$ are often canceled out in the derivatives of s and l.

The following lemma states how each of the derivatives can be bounded uniformly over Θ. We will not show the derivations of this, because it is straightforward, yet tedious, but it can be checked using e.g. Mathematica.

Lemma TA.1. *The derivatives of the log likelihood $\ell(g, \boldsymbol{\psi})$ and the score function $s(g; \boldsymbol{\psi})$ can be bounded as follows:*

- $\sup_{\boldsymbol{\psi}} |s(g; \boldsymbol{\psi})| = \sup_{\boldsymbol{\psi}} |\partial \ell(g, \boldsymbol{\psi})/\partial g| \leq d_1 + d_2|g|$ *for some finite constants d_1 and d_2. See the proof of Proposition 1 for a derivation.*
- $\sup_{g, \boldsymbol{\psi}} |\partial s(g; \boldsymbol{\psi})/\partial g| = \sup_{g, \boldsymbol{\psi}} |\partial^2 \ell(g, \boldsymbol{\psi})/\partial g^2| \leq d_1$ *, see the proof of Proposition 1 for a derivation.*
- $\sup_{g, \boldsymbol{\psi}} |\partial^2 s(g; \boldsymbol{\psi})/\partial g^2| = \sup_{g, \boldsymbol{\psi}} |\partial^3 \ell(g, \boldsymbol{\psi})/\partial g^3| \leq d_1$,
- $\sup_{g, \boldsymbol{\psi}} |\partial^3 s(g; \boldsymbol{\psi})/\partial g^3| \leq d_1$,
- *The elements of $\partial \ell(g, \boldsymbol{\psi})/\partial \boldsymbol{\psi}$ can be uniformly bounded by d_1, except for:*
  - $\sup_{\boldsymbol{\psi}} |\partial \ell(g, \boldsymbol{\psi})/\partial c_1| \leq d_1 + d_2|g|$,
  - $\sup_{\boldsymbol{\psi}} |\partial \ell(g, \boldsymbol{\psi})/\partial \sigma_1^2| \leq d_1 + d_2|g|^2$,
- *The elements of $\partial^2 \ell(g, \boldsymbol{\psi})/\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'$ can be uniformly bounded by d_1, except for:*
  - $\sup_{\boldsymbol{\psi}} |\partial^2 \ell(g, \boldsymbol{\psi})/\partial (\sigma_1^2)^2| \leq d_1 + d_2|g|^2$,
  - $\sup_{\boldsymbol{\psi}} |\partial^2 \ell(g, \boldsymbol{\psi})/\partial c_1 \partial \sigma_1^2| \leq d_1 + d_2|g|$,
- *The elements of $\partial s(g; \boldsymbol{\psi})/\partial \boldsymbol{\psi}$ and thus of $\partial^2 \ell(g, \boldsymbol{\psi})/\partial \boldsymbol{\psi} \partial g$ can be uniformly bounded by d_1, except for:*
  - $\sup_{\boldsymbol{\psi}} |\partial s(g; \boldsymbol{\psi})/\partial \sigma_1^2| = \sup_{\boldsymbol{\psi}} |\partial^2 \ell(g, \boldsymbol{\psi})/\partial \sigma_1^2 \partial g| \leq d_1 + d_2|g|$,
- *The elements of $\partial^2 s(g; \boldsymbol{\psi})/\partial \boldsymbol{\psi} \partial g$ and thus of $\partial^3 \ell(g, \boldsymbol{\psi})/\partial \boldsymbol{\psi} \partial g^2$ can all be uniformly bounded by d_1.*

- *The elements of* $\partial^3 s(g;\boldsymbol{\psi})/\partial\boldsymbol{\psi}\partial g^2$ *can all be uniformly bounded by* d_1*.*
- *The elements of* $\partial^2 s(g;\boldsymbol{\psi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'$ *and thus of* $\partial^3 \ell(g,\boldsymbol{\psi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'\partial g$ *can be uniformly bounded by* d_1*, except for:*
  - $\sup_{\boldsymbol{\psi}} |\partial^2 s(g;\boldsymbol{\psi})/\partial(\sigma_1^2)^2| = \sup_{\boldsymbol{\psi}} |\partial^3 \ell(g,\boldsymbol{\psi})/\partial(\sigma_1^2)^2\partial g| \leq d_1 + d_2|g|$
- *The elements of* $\partial^3 s(g;\boldsymbol{\psi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'\partial g$ *can all be uniformly bounded by* d_1*.*

If evaluated at $g_t(\boldsymbol{\theta})$, where $g_t(\boldsymbol{\theta})$ has n bounded moments, the derivatives that can be uniformly bounded by a constant have bounded moments of any order. The derivatives that can be bounded by $d_1 + d_2|g|^p$ where $p = 1$ or 2, have bounded moments of order n and $\frac{1}{2}n$ respectively.

C.2. Derivatives of prediction error process

In this section we give the first and second derivative of $g_t(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We use the same notation as in Technical Appendix D.2 of Blasques et al. (2021). Considering the updating equation of $g_t(\boldsymbol{\theta})$ in (7), it follows that:

$$\frac{\partial g_{t+1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = A_t^{(1)} + \frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} B_t$$

where

$$A_t^{(1)} = A^{(1)}(\boldsymbol{\theta}; g_t(\boldsymbol{\theta})) = \frac{\partial \omega}{\partial \boldsymbol{\theta}} - \frac{\partial \alpha}{\partial \boldsymbol{\theta}} s(g_t(\boldsymbol{\theta});\boldsymbol{\psi}) - \alpha \frac{\partial s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\theta}}\bigg|_{g=g_t(\boldsymbol{\theta})},$$
$$B_t = B(\boldsymbol{\theta}; g_t(\boldsymbol{\theta})) = 1 - \alpha s'(g_t(\boldsymbol{\theta});\boldsymbol{\psi}).$$

For the second derivative process we get:

$$\frac{\partial^2 g_{t+1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = A_t^{(2)} + \frac{\partial^2 g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} B_t,$$

where

$$\begin{aligned}
A_t^{(2)} &= A^{(2)}(\boldsymbol{\theta}; g_t(\boldsymbol{\theta}), \partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}) = \frac{\partial A_t^{(1)}}{\partial \boldsymbol{\theta}'} + \frac{\partial A_t^{(1)}}{\partial g_t}\frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} + \frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial B_t}{\partial \boldsymbol{\theta}'} + \frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial B_t}{\partial g_t}\frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \\
&= -\left(\frac{\partial \alpha}{\partial \boldsymbol{\theta}}\frac{\partial s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\theta}'}\bigg|_{g=g_t(\boldsymbol{\theta})} + \frac{\partial s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\theta}}\bigg|_{g=g_t(\boldsymbol{\theta})}\frac{\partial \alpha}{\partial \boldsymbol{\theta}'} + \alpha\frac{\partial^2 s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\bigg|_{g=g_t(\boldsymbol{\theta})}\right) \\
&\quad - \left(\frac{\partial \alpha}{\partial \boldsymbol{\theta}}\frac{\partial s(g;\boldsymbol{\psi})}{\partial g}\bigg|_{g=g_t(\boldsymbol{\theta})} + \alpha\frac{\partial^2 s(g;\boldsymbol{\psi})}{\partial \boldsymbol{\theta}\partial g}\bigg|_{g=g_t(\boldsymbol{\theta})}\right)\frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \\
&\quad - \frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\frac{\partial s(g;\boldsymbol{\psi})}{\partial g}\bigg|_{g=g_t(\boldsymbol{\theta})}\frac{\partial \alpha}{\partial \boldsymbol{\theta}'} + \alpha\frac{\partial^2 s(g;\boldsymbol{\psi})}{\partial g\partial \boldsymbol{\theta}'}\bigg|_{g=g_t(\boldsymbol{\theta})}\right) - \alpha\frac{\partial^2 s(g;\boldsymbol{\psi})}{\partial g^2}\bigg|_{g=g_t(\boldsymbol{\theta})}\frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}.
\end{aligned}$$

For the expressions of the derivatives of the score function $s(g;\boldsymbol{\psi})$, we refer to Section C.1.

C.3. Derivatives of log likelihood

Here we give the first and second derivative of the log likelihood function with respect to $\boldsymbol{\theta} = (\omega, \alpha, \boldsymbol{\psi}')'$. Using the notation

$$L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=2}^{T} \ell(g_t(\boldsymbol{\theta});\boldsymbol{\psi}), \qquad \text{where} \quad \ell(g,\boldsymbol{\psi}) = \log F(g;\boldsymbol{\psi}),$$

and where F is defined in (C.20). The first derivative is $\partial L_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \frac{1}{T}\sum_{t=2}^{T}\partial\ell(g_t(\boldsymbol{\theta});\boldsymbol{\psi})/\partial\boldsymbol{\theta}$, where

$$\frac{\partial\ell(g_t(\boldsymbol{\theta}),\boldsymbol{\psi})}{\partial\boldsymbol{\theta}} = \frac{\partial\ell(g,\boldsymbol{\psi})}{\partial\boldsymbol{\theta}}\Big|_{g=g_t(\boldsymbol{\theta})} + \frac{\partial\ell(g,\boldsymbol{\psi})}{\partial g}\Big|_{g=g_t(\boldsymbol{\theta})}\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\,,$$

where we know that $\partial\ell(g,\boldsymbol{\psi})/\partial g = -s(g;\boldsymbol{\psi})$ and $\partial\ell(g,\boldsymbol{\psi})/\partial\boldsymbol{\theta}$ is equal to

$$\frac{\partial\ell(g,\boldsymbol{\psi})}{\partial\boldsymbol{\theta}} = \begin{pmatrix} 0 \\ 0 \\ \frac{\partial\ell(g,\boldsymbol{\psi})}{\partial\boldsymbol{\psi}} \end{pmatrix}.$$

See section C.1 for the derivatives of the $\ell(g,\boldsymbol{\psi})$ function and see section C.2 for the derivative process of $g_t(\boldsymbol{\theta})$.

Now for the second derivative we have $\partial^2 L_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}' = \frac{1}{T}\sum_{t=2}^{T}\partial^2\ell(g_t(\boldsymbol{\theta});\boldsymbol{\psi})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, where

$$\begin{aligned}\frac{\partial^2\ell(g_t(\boldsymbol{\theta});\boldsymbol{\psi})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} &= \frac{\partial^2\ell(g,\boldsymbol{\psi})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Big|_{g=g_t(\boldsymbol{\theta})} + \frac{\partial^2\ell(g,\boldsymbol{\psi})}{\partial\boldsymbol{\theta}\partial g}\Big|_{g=g_t(\boldsymbol{\theta})}\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} + \frac{\partial^2\ell(g,\boldsymbol{\psi})}{\partial g\partial g}\Big|_{g=g_t(\boldsymbol{\theta})}\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} \\ &\quad + \frac{\partial g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial^2\ell(g,\boldsymbol{\psi})}{\partial g\partial\boldsymbol{\theta}'}\Big|_{g=g_t(\boldsymbol{\theta})} + \frac{\partial\ell(g,\boldsymbol{\psi})}{\partial g}\Big|_{g=g_t(\boldsymbol{\theta})}\frac{\partial^2 g_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\,,\end{aligned}$$

where $\partial^2\ell(g,\boldsymbol{\psi})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ is equal to

$$\frac{\partial^2\ell(g,\boldsymbol{\psi})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \frac{\partial^2\ell(g,\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'} \\ 0 & 0 & & \end{pmatrix}.$$

C.4. Fischer Information Matrix expression

Lemma TA.2. *Under the assumptions of Theorem 2 the Fischer Information matrix is given by*

$$\mathcal{I} = \mathbb{E}[\ell_t'(\boldsymbol{\theta}_0)\ell_t'(\boldsymbol{\theta}_0)^\top] = \begin{pmatrix} A & C \\ C^\top & B \end{pmatrix}$$

where

$$\begin{aligned} A &= \frac{1}{1-b}\begin{pmatrix} c^2 & -d \\ -d & \frac{1+a}{\alpha_0} \end{pmatrix}, \\ B &= D + \frac{\alpha_0}{1-b}\left[c\alpha_0\ E + \frac{2(\alpha_0 e - c)}{c}FF^\top - \alpha_0(HF^\top + FH^\top)\right], \\ C &= \frac{\alpha_0}{1-b}\begin{pmatrix} c\ G^\top - d\ F^\top \\ \frac{e}{c}\ F^\top +\ H^\top \end{pmatrix}, \end{aligned}$$

where a, b, c, d and e are the scalar-valued:

$$\begin{aligned} a &= 1 - \alpha_0\mathbb{E}[s'(\varepsilon_t;\boldsymbol{\psi}_0)] = 1 - \alpha_0 c\,, \\ b &= 1 - 2\alpha_0\mathbb{E}[s'(\varepsilon_t;\boldsymbol{\psi}_0)] + \alpha_0^2\mathbb{E}[s'(\varepsilon_t;\boldsymbol{\psi}_0)^2] = 1 - 2\alpha_0 c + \alpha_0^2 e\,, \\ c &= \mathbb{E}[s'(\varepsilon_t;\boldsymbol{\psi}_0)] = \mathbb{E}[s(\varepsilon_t;\boldsymbol{\psi}_0)^2]\,, \\ d &= \mathbb{E}[s'(\varepsilon_t,\boldsymbol{\psi}_0)s(\varepsilon_t,\boldsymbol{\psi}_0)]\,, \\ e &= \mathbb{E}[s'(\varepsilon_t,\boldsymbol{\psi}_0)^2]\,, \end{aligned}$$

and where D, E, F, G and H are the matrix and vector-valued:

$$
\begin{aligned}
D &= \mathbb{E}\left[\frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} \frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}'}\right], \\
E &= \mathbb{E}\left[\frac{s(\varepsilon_t; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} \frac{s(\varepsilon_t; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^\top}\right] \\
F &= \mathbb{E}\left[\frac{\partial s(\varepsilon_t; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right] = -\mathbb{E}\left[s(\varepsilon_t; \boldsymbol{\psi}_0) \frac{\ell(\varepsilon_t; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right], \\
G &= \mathbb{E}\left[s(\varepsilon_t; \boldsymbol{\psi}_0) \frac{\partial s(\varepsilon_t; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right], \\
H &= \mathbb{E}\left[s'(\varepsilon_t; \boldsymbol{\psi}_0) \frac{\partial s(\varepsilon_t; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right].
\end{aligned}
$$

Proof. By the information equality, we know that $\mathcal{I} = \mathbb{E}[\ell_t'(\boldsymbol{\theta}_0)\ell_t'(\boldsymbol{\theta}_0)^\top]$, so it follows from the expression of the derivative of the log likelihood function in section C.3, that

$$
\mathcal{I} = \begin{pmatrix} A & C \\ C^\top & B \end{pmatrix}
$$

with

$$
\begin{aligned}
A &= \mathbb{E}[s(\varepsilon_t, \boldsymbol{\psi}_0)^2]\, \mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_1} \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_1^\top}\right], \\
B &= \mathbb{E}\left[\frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} \frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^\top}\right] + \mathbb{E}[s(\varepsilon_t, \boldsymbol{\psi}_0)^2]\, \mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}} \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}^\top}\right] \\
&\quad - \mathbb{E}\left[s(\varepsilon_t, \boldsymbol{\psi}_0) \frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right] \mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}^\top}\right] - \mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}\right]\, \mathbb{E}\left[s(\varepsilon_t, \boldsymbol{\psi}_0) \frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^\top}\right] \\
C &= \mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_1}\right] \mathbb{E}\left[\frac{\partial \ell(\varepsilon_t, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^\top} s(\varepsilon_t, \boldsymbol{\psi}_0)\right] + \mathbb{E}[s(\varepsilon_t, \boldsymbol{\psi}_0)^2]\, \mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_1} \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}^\top}\right],
\end{aligned}
$$

for $\boldsymbol{\theta}_1 = (\alpha, \omega)^\top$, where we use that $g_t(\boldsymbol{\theta}_0)$ is equal to ε_t almost surely. Notice that we use that the derivative of $\ell(g, \boldsymbol{\psi})$ with respect to g is equal to $-s(g; \boldsymbol{\psi})$. Proving the equalities that are claimed to hold in the definitions of the constants c and F, can be done by taking the derivative of $\mathbb{E}[s(\varepsilon; \boldsymbol{\psi}_0)]$ (which is zero by construction) with respect to ε_t and $\boldsymbol{\psi}$ respectively and by interchanging the integral and the derivative, which can be done by a standard argument. For the expectations of the derivatives of $\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ and its square, we can use the derived expressions of section C.2 and the fact that by the assumptions of Theorem 2 and Proposition 2 we know that these expectations exist and are finite and that the derivatives are SE. Using the notation a as defined in the Lemma, for the expectation of $\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ we get:

$$
\begin{aligned}
\mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \alpha}\right] &= \frac{\mathbb{E}[-s(\varepsilon_t; \boldsymbol{\psi}_0)]}{1-a} = 0 \\
\mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \omega}\right] &= \frac{1}{1-a} \\
\mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}\right] &= -\frac{\alpha_0}{1-a} \mathbb{E}\left[\frac{\partial s(\varepsilon_t; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right] = -\frac{\alpha_0\, F}{1-a}
\end{aligned}
$$

Similarly, for the expectation of the square of $\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$, using the notation defined in the Lemma, we get:

$$
\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \alpha}\right)^2\right] &= \frac{\mathbb{E}[s(\varepsilon_t; \boldsymbol{\psi}_0)^2]}{1-b} = \frac{c}{1-b} \\
\mathbb{E}\left[\left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \omega}\right)^2\right] &= \frac{1+a}{(1-a)(1-b)}
\end{aligned}
$$

$$\mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}^\top}\right] = \frac{\alpha_0^2}{1-b}\Bigg[\mathbb{E}\left[\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^\top}\right] + \frac{1}{1-a}\Bigg[2\ \mathbb{E}\left[\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right]\mathbb{E}\left[\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^\top}\right]$$
$$- \alpha_0\ \mathbb{E}\left[s'(\varepsilon_t;\boldsymbol{\psi}_0)\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right]\mathbb{E}\left[\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^\top}\right] - \alpha_0\ \mathbb{E}\left[\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^\top}\right]\mathbb{E}\left[s'(\varepsilon_t;\boldsymbol{\psi}_0)\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right]\Bigg]\Bigg]$$
$$= \frac{\alpha_0^2}{1-b}\left[E + \frac{1}{1-a}\left[2FF^\top - \alpha_0 HF^\top - \alpha_0 FH^\top\right]\right]$$
$$\mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \alpha}\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \omega}\right] = -\frac{\alpha_0}{(1-a)(1-b)}\mathbb{E}[s'(\varepsilon_t;\boldsymbol{\psi}_0)s(\varepsilon_t;\boldsymbol{\psi}_0)]$$
$$= -\frac{\alpha_0\ d}{(1-a)(1-b)}$$
$$\mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \alpha}\right] = \frac{\alpha_0}{1-b}\left[\mathbb{E}\left[\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}s(\varepsilon_t;\boldsymbol{\psi}_0)\right] - \frac{\alpha_0}{1-a}\mathbb{E}\left[\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right]\mathbb{E}[s'(\varepsilon_t,\boldsymbol{\psi}_0)s(\varepsilon_t,\boldsymbol{\psi}_0)]\right]$$
$$= \frac{\alpha_0}{1-b}\left[G - \frac{\alpha_0\ d}{1-a}F\right]$$
$$\mathbb{E}\left[\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \omega}\right] = \frac{\alpha_0}{(1-a)(1-b)}\left[-2\ \mathbb{E}\left[\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right] + \alpha_0\mathbb{E}\left[\frac{\partial s(\varepsilon_t;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}s(\varepsilon_t;\boldsymbol{\psi}_0)\right]\right]$$
$$= \frac{\alpha_0}{(1-a)(1-b)}\left[-2\ F + \alpha_0 G\right]$$

Plugging these expectations into the expressions of A, B and C, and simlpifying the resulting expressions leads to the final forms in the lemma. □

D. Supplementary Monte Carlo simulation results

To supplement the asymptotic results on the MLE, we consider a Monte Carlo simulation study to investigate the small sample results of the ML estimator for the model under consideration. Consider the model with $J = 2$ and $J = 3$ components. The chosen parameter values are close to those that were estimated in the application of Section 5, see Table 5, and just as in the application we set $\omega = 0$. We simulate 1000 times for sample sizes $T = 500$, $T = 1000$ and $T = 5000$.

The results of the simulation study are reported in Table 5 and Figure 5. Figure 5 displays the estimated kernel density function for each of the estimated parameters. As the sample size increases, the average estimates move towards the true values and the standard deviations become smaller. This is also visible in the kernel density plots, as for most of the parameters the density is symmetric around the true value and as the sample size increases, the estimates move closer around the true value. Furthermore, the empirical standard deviation is generally larger than the average asymptotic standard deviation calculated based on the asymptotic variance matrix of Theorem 2. As the sample size grows, the two standard deviations move closer together. We also see in Table 5 that the parameters σ_i^2 and c_i corresponding to components with a small weight w_i are estimated less accurately, taking into account the true value of these parameters might be larger for the component with the smaller weight.

The results for $J = 3$ components are notably worse than for $J = 2$; the average estimates are further away from their true values, especially for the smaller sample sizes. This is also visible in the plotted kernel densities of Figure 5. What stands out in particular, is that for $J = 3$ it seems to be difficult to identify the second component. Namely, for the smaller sample sizes, a part of the estimates of σ_2^2 is not around its true value 10, but around 3, which is the true value of σ_3^2. The second component weight w_2 also has a peak around 0.8, while the true value is 0.15. Hence, for $J = 3$ or more components, it seems like the ML estimator struggles to distinguish the different components, which has to do with the log likelihood being ill-behaved. For the sample size $T = 5000$, this problem is less pronounced, as the weight

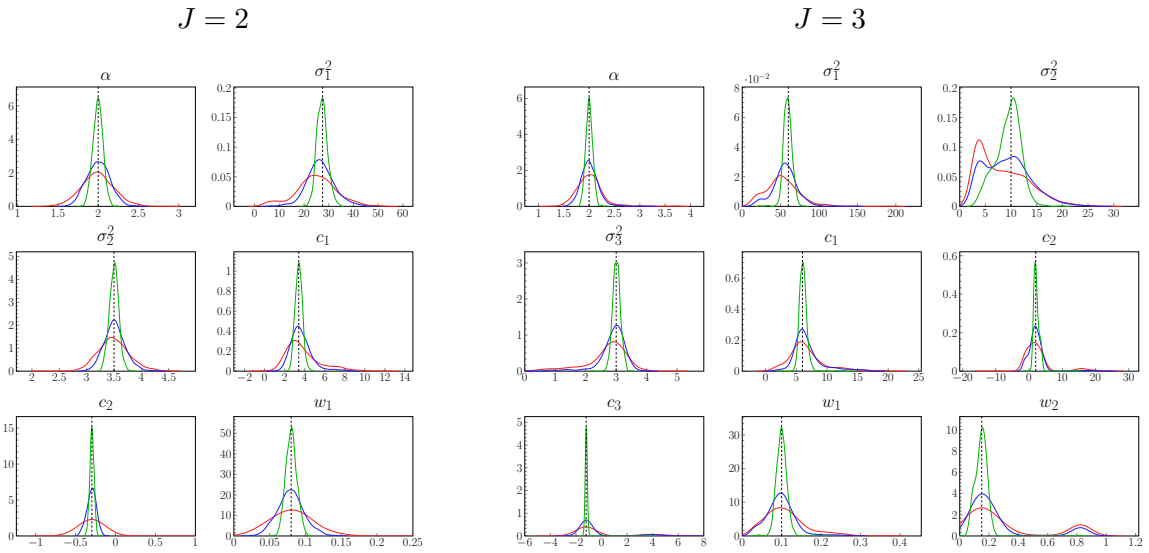

Figure 5. Kernel density estimates of estimated parameters of model with $J = 2$ (left) and $J = 3$ (right) components for sample sizes $T = 500$ (red), $T = 1000$ (blue) and $T = 5000$ (green). The results are based on 1000 Monte Carlo replications. Black dashed line represents true parameter value.

w_2 is estimated rather accurately, but still the estimates of σ_2^2 are not very precise and there is more mass left of the true value. In practice, when the number of components increases, the ML estimator may not be very stable, especially if the sample size is small.

Table 5. Monte Carlo simulation results for 1000 replications

		$T = 500$				$T = 1000$				$T = 5000$			
	$\boldsymbol{\theta}_0$	Avg $\hat{\boldsymbol{\theta}}_T$	Bias	As. Std.	Emp. Std.	Avg $\hat{\boldsymbol{\theta}}_T$	Bias	As. Std.	Emp. Std.	Avg $\hat{\boldsymbol{\theta}}_T$	Bias	As. Std.	Emp. Std.
$J = 2$													
α	2	1.991	-0.009	0.198	0.202	2.001	0.001	0.139	0.141	1.997	-0.003	0.062	0.061
σ_1^2	27.5	24.868	-2.632	7.079	8.271	26.675	-0.825	5.428	5.505	27.315	-0.185	2.276	2.151
σ_2^2	3.5	3.494	-0.006	0.276	0.284	3.513	0.013	0.193	0.193	3.501	0.001	0.085	0.085
c_1	3.4	3.820	0.420	1.240	1.763	3.646	0.246	0.918	1.075	3.407	0.007	0.365	0.358
c_2	-0.296	-0.262	0.033	-	0.590	-0.300	-0.005	-	0.059	-0.296	0.0001	-	0.026
w_1	0.08	0.084	0.004	0.025	0.063	0.079	-0.001	0.018	0.018	0.080	0.0003	0.008	0.008
w_2	0.92	0.916	-0.004	-	0.063	0.921	0.001	-	0.018	0.920	-0.0003	-	0.008
$J = 3$													
α	2	2.014	0.014	0.209	0.271	2.012	0.012	0.147	0.176	2.002	0.002	0.065	0.068
σ_1^2	60	52.064	-7.936	11.554	22.104	55.832	-4.168	10.098	16.311	59.138	-0.862	4.719	5.658
σ_2^2	10	8.338	-1.662	3.693	4.807	9.223	-0.777	3.332	4.402	9.408	-0.592	1.926	2.404
σ_3^2	3	2.684	-0.316	0.511	0.694	2.882	-0.118	0.368	0.466	3.011	0.011	0.120	0.129
c_1	6	6.658	0.658	1.776	3.032	6.827	0.827	1.332	2.431	6.121	0.121	0.519	0.725
c_2	2	2.671	0.671	1.164	4.663	2.257	0.257	1.015	3.107	2.205	0.205	0.637	0.872
c_3	-1.2	0.019	1.219	-	3.072	-0.561	0.639	-	1.916	-1.179	0.021	-	0.237
w_1	0.1	0.110	0.010	0.0307	0.055	0.098	-0.002	0.022	0.038	0.100	-0.0004	0.010	0.013
w_2	0.15	0.300	0.150	0.069	0.278	0.239	0.089	0.055	0.223	0.147	-0.003	0.031	0.048
w_3	0.75	0.591	-0.159	-	0.285	0.664	-0.086	-	0.232	0.753	0.003	-	0.045

Avg $\hat{\boldsymbol{\theta}}_T$ reports the average estimate, Bias stands for the difference between $\boldsymbol{\theta}_0$ and the average estimate, As. Std. is the average asymptotic standard deviation and Emp. Std. is the empirical standard deviation of the estimated parameters.

References

Billingsley, P. (1999). *Convergence of Probability Measures*. John Wiley & Sons, 2nd edition.

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

Blasques, F., Koopman, S. J., and Lucas, A. (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika*, 102(2):325–343.

Blasques, F., van Brummelen, J., Koopman, S. J., and Lucas, A. (2021). Maximum likelihood estimation for score-driven models. *Journal of Econometrics*. forthcoming.

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

Caivano, M. and Harvey, A. C. (2014). Time-series models with an egb2 conditional distribution. *Journal of Time Series Analysis*, 35(6):558–571.

Caivano, M., Harvey, A. C., and Luati, A. (2016). Robust time series models with trend and seasonal components. *SERIEs*, 7(1):99–120.

Catania, L. (2021). Dynamic adaptive mixture models with an application to volatility and risk. *Journal of Financial Econometrics*, 19(4):531–564.

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

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

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

Escribano, A., Ignacio Peña, J., and Villaplana, P. (2011). Modelling electricity prices: International evidence. *Oxford Bulletin of Economics and Statistics*, 73(5):622–650.

Fernández, C. and Steel, M. F. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164(1):130–141.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gorgi, P. and Koopman, S. J. (2021). Beta observation-driven models with exogenous regressors: a joint analysis of realized correlation and leverage effects. *Journal of Econometrics*. forthcoming.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

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

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

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

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

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

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley.

Nielsen, F. (2021). Fast approximations of the Jeffreys divergence between univariate Gaussian mixtures via mixture conversions to exponential-polynomial distributions. *Entropy*, 23(11):forthcoming.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic nonlinear econometric models: Asymptotic theory*. Springer Verlag.

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

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Econometric Society Monographs. 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.

Wong, C. S. and Li, W. K. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):95–115.