Consistency, distributional convergence, and optimality of score-driven filters

Eric Beutner[1], Yicong Lin[1,2] and Andre Lucas*[1,2]

[1]Vrije Universiteit Amsterdam

[2]Tinbergen Institute

August, 2023

Abstract

We study the in-fill asymptotics of score-driven time series models. For general forms of model mis-specification, we show that score-driven filters are consistent for the Kullback-Leibler (KL) optimal time-varying parameter path, which minimizes the pointwise KL divergence between the statistical model and the unknown dynamic data generating process. This directly implies that for a correctly specified predictive conditional density, score-driven filters consistently estimate the time-varying parameter path even if the model is mis-specified in other respects. We also obtain distributional convergence results for the filtering errors and derive the filter that minimizes the asymptotic filter error variance. Score-driven filters turn out to be optimal under correct specification of the predictive conditional density. The results considerably generalize earlier findings on the continuous-time consistency of volatility filters under mis-specification: they apply to biased filters, use weaker assumptions, allow for more general forms of mis-specification, and consider general time-varying parameters in non-linear time series models beyond the volatility case. Several examples are used to illustrate the theory, including time-varying tail shape models, dynamic copulas, and time-varying regression models.

Keywords: score-driven models, information theoretic optimality, Kullback-Leibler divergence, pseudo true time-varying parameters, in-fill asymptotics.

*Corresponding author: Department of Econometrics and Data Science, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, the Netherlands. E-mail address: a.lucas@vu.nl.

1 Introduction

Since their introduction by Creal et al. (2011, 2013) and Harvey (2013), score-driven models have led to a wide range of new applied flexible non-linear time series models that are successfully applied to describe time variation in economics and finance.[1] These models update the time-varying parameter by the score of the conditional predictive density. The popularity of the score-driven approach lies in its generality, its ease of applicability, and its relative computational simplicity since the models can be estimated by standard maximum likelihood methods. The asymptotic theory of these models has recently been studied for particular cases in, for instance, Harvey (2013), Blasques et al. (2016, 2018), Babii et al. (2019), and Hetland et al. (2023). A more general asymptotic statistical framework for parameter estimation in univariate score-driven models has been formulated in Blasques et al. (2022).

Despite their popularity, the literature has remained remarkably silent on the theoretical cornerstones of score-driven models, such as the consistency of filtered time-varying parameter paths and the accuracy of their asymptotic approximations. This holds true even for the correctly specified model case. If models are (severely) mis-specified, the asymptotic properties of filtered paths of score-driven models in general settings remain largely unknown. It is here that the current paper tries to make its main contribution. Next to the asymptotic framework for parameter estimation mentioned earlier (Blasques et al., 2022), only two other contributions comment on the score-driven framework's general properties. First, Blasques et al. (2015, 2018) show that under specific conditions the Kullback-Leibler (KL) divergence between the true, unknown conditional density and the model density is improved if and only if the time-varying parameter mechanism uses a score-driven update, or an equivalent thereof; see Blasques et al. (2023) for further generalizations of this result, and Creal et al. (2019) for a similar result using other criteria than KL divergence, such as a generalized method of moments criterion.[2] Second, Buccheri et al. (2021) study the in-fill asymptotics for univariate score-driven volatility filters, building on results of Nelson (1990, 1992).

[1] The literature on score-driven models is by now quite extensive, including contributions on modeling time-varying default probabilities and losses-give-default in panel data (see Creal et al., 2014; Babii et al., 2019), univariate time-varying mean and volatility modeling (see for instance Harvey and Luati, 2014; Linton and Wu, 2020), multivariate time-varying volatility modeling, including multivariate realized covariance matrices and multivariate copulas (see for instance Creal et al., 2011; Lucas et al., 2014; Opschoor et al., 2018; Gorgi et al., 2019; Buccheri et al., 2021; Hafner and Wang, 2023; Oh and Patton, 2018, 2023), time-varying regression models (Umlandt, 2023), spatial models with dynamic parameters (Blasques et al., 2016; Gasperoni et al., 2021), time-varying tail shape models, (Massacci, 2017), time-varying cure rate models (Hansen and Schmidtblaicher, 2021), models for dynamic quantiles and tail expectations (Patton et al., 2019; Catania and Luati, 2023), models for circular time-series and wind angles (Harvey et al., 2023), time-varying eigenvalues (Hetland et al., 2023), mixture models for clustering (Lucas et al., 2019; Joao et al., 2023), DSGE models with time-varying structural parameters (Angelini and Gorgi, 2018), state-space models with time-varying parameters (Monache et al., 2021), models for data on bounded intervals (Gorgi, 2020), models for multivariate discrete high-frequency tick-data (Koopman et al., 2018), models for classification trees and forests with time variation (Patton and Simsek, 2023), and much more. For a more complete overview, see for instance the papers section on `http://www.gasmodel.com`.

[2] Further theoretical discrete-time results on KL improvements have recently been derived by van Os (2023) for modified versions of *implicit* score-driven models of Lange et al. (2022) as opposed to the *explicit* score-driven models of Creal et al. (2013).

All of these earlier results, however, remain quite specific. The approach of Blasques et al. (2015) (BKL15 from now on), for instance, suffers from three major drawbacks. First, the framework of BKL15 is in discrete time, limiting its conclusion to the validity of the direction of a (sufficiently small) parameter update. It does not explicitly address whether the updates succeed in minimizing the KL divergence in an asymptotic sense. Second, the results in BKL15 are local rather than global in nature. In particular, they are only applicable in admissible sets, i.e., sets where the true density dominates the model density (see Blasques et al., 2018). Admissibility of this type cannot be checked empirically, which reduces the concept's relevance. Creal et al. (2019) dispense with admissibility by focusing on expected rather than realized KL improvements, but again can still only make statements about the direction of the update. Finally, neither BKL15 nor Creal et al. (2019) establish consistency of the filtered time-varying parameter, nor do they define what consistency could mean in this context or what size of filtering errors to expect in the limit. Similarly, the in-fill asymptotic analysis of Buccheri et al. (2021) restricts itself to the volatility case and robust, score-driven volatility filters. It does not deal with other forms of parameter dynamics such as the wide variety of models mentioned in Footnote 1. It also considers relatively mild forms of mis-specification, where volatility remains the prime source of time-variation, and there is no more fundamental mismatch between the time-varying parameter in the model and in the data generating process (DGP).

In this paper, we address the above issues in a general framework that fills a number of gaps in the literature. First, we show under what conditions score-driven filters are consistent for the true time-varying parameter path, even if the model is dynamically mis-specified otherwise. This includes, for instance, settings where the DGP is of state-space form with a time-varying parameter driven by its own stochastic process, such that the observation-driven score-based filter is obviously mis-specified.

Second, even if the model is more severely mis-specified, whether due to the incorrect choice of the predictive conditional density or the time-varying parameter, we show that score-driven filters are still consistent estimators for the KL-optimal time-varying parameter path. This path minimizes (pointwise) the KL divergence between the possibly mis-specified statistical model and the unknown DGP. The result holds under very general conditions, including cases where the time-varying parameter and the pseudo-true parameter do not coincide, or are even of different dimensions. We do not require a similar notion of admissibility as in Blasques et al. (2018).

Third, we derive the asymptotic behavior of scaled filtering errors, i.e., the scaled difference between the estimated time-varying parameter and its pseudo-true counterpart. Interestingly, using these new asymptotic results, we are able to study the variance of filtering error as a function of the shape of the parameter update. This allows us to construct an optimal update function for observation-driven filters from a minimum variance perspective. Our findings highlight that score-driven filters are the optimal choice if the model's conditional predictive density is correctly specified. If there is mis-specification, we show that the (infeasible) optimal update function is proportional to the true density's score function, even in cases where the model contains fewer time-varying parameters than the DGP.

A number of our tools are taken from the familiar work of Nelson (1990, 1992, 1996) for the continuous time limit of generalized autoregressive conditional heteroskedasticity (GARCH) filters, and as applied to score-driven volatility filters by Buccheri et al. (2021). We also build on the slightly less familiar results of Nelson and Foster (1994). In particular, our results on the asymptotic distribution of filtering errors are inspired by the two-time-scales approach of Nelson and Foster. However, we also considerably generalize their approach. First, Nelson and Foster (1994) establish results for filtering errors around the true time-varying parameter path, which in their case is volatility. In our setting, such a true time-varying parameter path may not exist, and the best we can hope for is to recover a KL-optimal path. The notion of mis-specification of this type, however, invalidates some of the key steps in Nelson and Foster's approach that builds upon the overlap between the true and filtered time-varying parameters, and we show how to overcome these issues. We also show how to establish the asymptotic properties of score-driven filters in mis-specified under-parameterized settings, i.e., settings where the DGP contains more time-varying parameters than the model itself.

Like Nelson (1990, 1992, 1996) and Nelson and Foster (1994), we treat the static parameters of the filtering equations as fixed in this paper and do not consider data-driven (or estimated) choices for these parameters. Rather we concentrate on the properties of the filter itself as an 'estimator' for the unknown KL-optimal parameter path, in line with the above references. We leave the issue of parameter estimation to future work; see for instance Jensen and Lange (2010) for some results in the volatility setting. However, we gather some additional results for a more generic setting than the one considered in the main text, where we allow convergence rates to vary over the different components of the filtering equation. This might provide a relevant stepping-stone for a further analysis of the effect of parameter estimation; see the remarks on convergence rates in Jensen and Lange (2010).

The remainder of this paper is structured as follows. Section 2 offers an intuitive understanding of our asymptotic framework by presenting a simple motivating example. Section 3 then introduces the general formal modeling set-up. Section 4 develops the asymptotic framework for the score-driven filter, including consistency to the true or pseudo-true time-varying parameter path, the asymptotic normality of the filtering errors, and the shape of minimum variance filters. Section 5 contains a number of illustrative applications of the theory. Section 6 concludes. The appendix gathers the proofs and supplementary materials.

We adopt the following notational conventions. Vectors and matrices are in bold, whereas scalars are non-bold. For a vector $\boldsymbol{x} = (x_j) \in \mathbb{R}^n$, its p-norm is denoted by $\|\boldsymbol{x}\|_p = (\sum_{j=1}^n |x_j|^p)^{1/p}$. The induced p-norm for a matrix $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|_p = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \|\boldsymbol{A}\boldsymbol{x}\|_p/\|\boldsymbol{x}\|_p$. The subscripts are omitted whenever $p = 2$. We write $\operatorname{diag}(\boldsymbol{A}, \boldsymbol{B})$ for a block-diagonal matrix with blocks $\boldsymbol{A}$ and $\boldsymbol{B}$. Let $\lfloor x \rfloor$ be the integer part of $x \in \mathbb{R}$. Finally, we use C to denote generic constants that can change from line to line.

2 A motivating example

To set the stage, we first discuss an example that highlights the main aspects of the theory developed in the subsequent sections of this paper. We consider the case of a Pareto distribution with a time-varying tail shape parameter that we attempt to filter using the score-driven approach of Creal et al. (2013) and Harvey (2013). Score-driven filters for time-varying tail shapes have been studied in for instance Massacci (2017) and D'Innocenzo et al. (2023).

Assume a univariate time series $y_h(t_{i,h})$ observed at the discrete time points $t_{i,h} = i \cdot h$ for $i = 1, \ldots, n_h$ with $n_h = \lfloor T_0/h \rfloor$ for some fixed $T_0 > 0$. Note that in order to derive our subsequent results, we thus actually consider a sequence of time series indexed by h for $h \downarrow 0$. We assume that $\Big(y_h(t_{i+1,h}) \,\big|\, \psi_h(t_{i,h}) = \psi\Big)$ has a conditional probability density function (pdf) given by

$$q_h(y; \psi) = \Big(h\,\sigma\,\ell(\psi)\Big)^{-1} \left(\frac{y}{h\,\sigma}\right)^{-1/\ell(\psi)-1}, \qquad y \geq h\,\sigma > 0, \tag{2.1}$$

where $\psi_h(t_{i,h})$ is the time-varying parameter that we try to filter below. The distribution in (2.1) is a Pareto distribution with tail-shape parameter $\ell(\psi)$. Let $\ell(\psi) = 4^{-1}\big(1 + \exp(-\psi/4)\big)^{-1} \in (0, 1/4)$, which ensures that the 4th order conditional moment of $y_h(t_{i+1,h})$ always exists for any $\psi \in \mathbb{R}$. Models such as (2.1) are for instance used to assess the occurrence of extreme risks, the occurrence of which may vary with changing economic and market conditions.

Our main interest lies in the asymptotic properties of observation-driven filters when applied to estimate the time-varying parameter $\psi_h(t_{i,h})$. The time-variation in the true $\psi_h(t_{i,h})$ may itself not be observation-driven. Here and elsewhere in the paper we assume that the dynamics of $\psi_h(t_{i,h})$ are specified by their own discrete-time stochastic process. For this introductory example, we assume

$$\psi_h(t_{i+1,h}) = h\,a + (1 - h\,b)\psi_h(t_{i,h}) + h^{1/2}\,B\,\eta_{i+1}, \tag{2.2}$$

where η_{i+1} are i.i.d. innovations with zero mean and unit variance, and a, b, and B are constants. These underlying dynamics, however, are *unknown* to the statistician. As mentioned, the statistician's goal is to filter the (log) tail shape-parameter from the data. For this, she may use a possibly misspecified model. As an illustration, consider the case where the statistician uses a slightly differently parameterized version of the same Pareto distribution,

$$p_h\big(y;\theta\big) = \big(h\,\sigma\big)^{-1}\,\exp(-\theta)\left(\frac{y}{h\,\sigma}\right)^{-\exp(-\theta)-1}, \qquad y \geq h\sigma > 0, \tag{2.3}$$

where she uses the exponential function to transform a possibly negative θ into the positive tail-shape parameter $\exp(\theta)$. For filtering the tail-shape parameter, the statistician uses a score-driven filter, specified as

$$\theta_h(t_{i+1,h}) = h\,\omega + (1 - h\,\beta)\theta_h(t_{i,h}) + h^{1/2}\,g_h\big(y_h(t_{i+1,h}), \theta_h(t_{i,h})\big), \tag{2.4}$$

with

$$g_h\big(y_h(t_{i+1,h}), \theta_h(t_{i,h})\big) = \alpha\,\frac{\partial \log p_h\big(y_h(t_{i+1,h}); \theta_h(t_{i,h})\big)}{\partial\theta} = \alpha\,\left(\exp\big[-\theta_h(t_{i,h})\big]\ln\left(\frac{y_h(t_{i+1,h})}{h\,\sigma}\right) - 1\right); \tag{2.5}$$

see Creal et al. (2013) for more details, and Massacci (2017) and D'Innocenzo et al. (2023) for applications. We follow Nelson and Foster (1994) and Nelson (1996) by considering fixed values of the parameters α, β and ω in the filtering equation (2.4) and do not consider data-driven (or estimated) parameter choices; more on this can be found in Section 4.4. We thus mainly look at the filter as an estimator for the true, unknown time-varying parameter path $\psi_h(\cdot)$ or its transform $\ell(\psi_h(\cdot))$.

The filter in Eq. (2.4) is clearly mis-specified for the true time-varying tail-shape parameter in at least two ways. First, the true Pareto tail-shape $\ell(\psi_h(t_{i,h}))$ in (2.1) can only take values in the range $(0, 1/4)$, whereas the filtered tail-shape parameter $\exp(\theta_h(t_{i,h}))$ in (2.3) can take values in $\mathbb{R}^+$. Second, the true tail-shape $\psi_h(t_{i+1,h})$ in (2.2) is driven by its own disturbances η_{i+1} and thus has parameter-driven dynamics as defined by Cox (1981), whereas the filter $\theta_h(t_{i+1,h})$ in (2.5) is driven by the values of $y_h(t_{i+1,h})$ and $\theta_h(t_{i,h})$ via the function g_h and thus has observation-driven dynamics. Despite this double mis-specification, this paper shows that $\exp(\theta_h(t_{i,h}))$ consistently estimates the true $\ell(\psi_h(t_{i,h}))$ as $h \downarrow 0$.

The current mis-specification in the model and filter in Eqs. (2.1)–(2.5) is relatively mild. In later examples in Section 5, we also establish consistency for much more severe forms of mis-specification. These include settings where the true density q_h and model density p_h differ, or where the true parameter ψ and the model parameter θ capture different aspects of the distribution or even differ in dimension. In such settings, consistency can in general no longer be towards the true time-varying parameter $\psi_h(t_{i,h})$, but will instead be towards the Kullback-Leibler optimal value.

To develop some intuition for the general results obtained in this paper, consider a new parameter

$\theta_h^\star(t_{i,h})$ defined as

$$\exp(\theta_h^\star(t_{i,h})) = \ell(\psi_h(t_{i,h})) \qquad \Longleftrightarrow \qquad \theta_h^\star(t_{i,h}) = \ln\Big(\ell\big[\psi_h(t_{i,h})\big]\Big). \tag{2.6}$$

This parameter $\theta_h^\star(t_{i,h})$ is the 'pseudo true' value of $\theta_h(t_{i,h})$ in the sense that it gives the same value of the tail-shape parameter $\exp(\theta_h^\star(t_{i,h}))$ as the (transformed) true time-varying parameter $\ell(\psi_h(t_{i,h}))$. Using this $\theta_h^\star(t_{i,h})$, we define the key quantity of this paper, namely the filtering error $z_h(t_{i,h}) = \theta_h(t_{i,h}) - \theta_h^\star(t_{i,h})$. By combining Eqs. (2.2) and (2.4) and defining $\Delta z_h(t_{i+1,h}) = z_h(t_{i+1,h}) - z_h(t_{i,h})$, we obtain

$$\Delta z_h(t_{i+1,h}) = h^{1/2}\, g_h\big(y_h(t_{i+1,h}), \theta_h(t_{i,h})\big) - h^{1/2}\frac{\mathrm{d}\theta_h^\star(t_{i,h})}{\mathrm{d}\psi_h(t_{i,h})}B\eta_{i+1} + O_P(h), \tag{2.7}$$

where we use the notation $O_P(h)$ to denote terms of the form $h \cdot X_h$ for some random sequence (X_h) that is bounded in probability. Using a first order Taylor series approximation of (2.7) around $\theta_h^\star(t_{i,h})$, we get

$$\Delta z_h(t_{i+1,h}) = h^{1/2}\,\frac{\partial g_h\big(y_h(t_{i+1,h}), \theta_h^\star(t_{i,h})\big)}{\partial\theta}\, z_h(t_{i,h}) + h^{1/2}\left[g_h\big(y_h(t_{i+1,h}), \theta_h^\star(t_{i,h})\big) - \frac{\mathrm{d}\theta_h^\star(t_{i,h})}{\mathrm{d}\psi_h(t_{i,h})}B\eta_{i+1}\right] + O_P(h). \tag{2.8}$$

A formal analysis of the validity of all steps is provided in Section 4. Ignoring the $O_P(h)$ term, which converges to zero in probability as $h \downarrow 0$, Eq. (2.8) can be recognized as a first order autoregressive (AR) process for $z_h(t_{i+1,h})$ with random AR coefficient $1 - h^{1/2}\, A_{i+1,h}$ and innovation $h^{1/2}\, \zeta_{i+1,h}$, where

$$A_{i+1,h} = -\,\frac{\partial g_h\big(y_h(t_{i+1,h}), \theta_h^\star(t_{i,h})\big)}{\partial\theta}, \qquad \zeta_{i+1,h} = g_h\big(y_h(t_{i+1,h}), \theta_h^\star(t_{i,h})\big) - \frac{\mathrm{d}\theta_h^\star(t_{i,h})}{\mathrm{d}\psi_h(t_{i,h})}B\eta_{i+1}.$$

Given the definition of $\theta_h^\star(t_{i,h})$ in (2.6) and the fact that $g_h\big(y_h(t_{i+1,h}), \theta_h^\star(t_{i,h})\big)$ from (2.5) is the derivative of the log Pareto density, it follows that the innovations $\zeta_{i+1,h}$ of this autoregressive process have zero mean. After iterating, and neglecting the $O_P(h)$ term, we arrive at

$$\begin{aligned} z_h(t_{i+1,h}) &= z_h(t_{0,h}) \prod_{j=0}^{i}\left[1 - h^{1/2}\, A_{j+1,h}\right] + \sum_{j=0}^{i} h^{1/2}\zeta_{j+1,h} \prod_{k=j+1}^{i}\left[1 - h^{1/2}\, A_{k+1,h}\right] \\ &\approx z_h(t_{0,h}) \exp\left[-\sum_{j=0}^{i} h^{1/2}\, A_{j+1,h}\right] + \sum_{j=0}^{i} h^{1/2}\zeta_{j+1,h} \exp\left[-\sum_{k=j+1}^{i} h^{1/2}\, A_{k+1,h}\right], \end{aligned} \tag{2.9}$$

where the second line follows from the approximation $1 - h^{1/2}\, x \approx \exp(-h^{1/2}\, x)$ for $h \downarrow 0$, and where we use the convention that the product over the empty set equals 1. We re-write the argument of the first exponential in (2.9) as

$$-h^{1/2}\sum_{j=0}^{i}\Big(A_{j+1,h}-\mathbb{E}\big[A_{j+1,h}\mid\psi_h(t_{j,h}),\theta_h^\star(t_{j,h})\big]\Big)-h^{1/2}\sum_{j=0}^{i}\mathbb{E}\big[A_{j+1,h}\mid\psi_h(t_{j,h}),\theta_h^\star(t_{j,h})\big]. \quad (2.10)$$

If i is of order h^{-1}, the first sum in (2.10) converges in distribution and hence is bounded in probability, whereas the second part of (2.10) diverges to $-\infty$ if

$$\mathbb{E}\big[A_{j+1,h}\mid\psi_h(t_{j,h}),\theta_h^\star(t_{j,h})\big]=-\mathbb{E}\left[\frac{\partial g_h\big(y_h(t_{j+1,h}),\theta_h^\star(t_{j,h})\big)}{\partial\theta}\;\middle|\;\psi_h(t_{j,h}),\theta_h^\star(t_{j,h})\right]>0. \quad (2.11)$$

The latter clearly holds in our current example if $\alpha>0$, as we have $-\mathbb{E}[A_{j+1,h}\mid\psi_h(t_{j,h}),\theta_h^\star(t_{j,h})]=\alpha\exp\big(-\theta_h^\star(t_{j,h})\big)\,\mathbb{E}\big[\ln\big(y_h(t_{j+1,h})/(h\sigma)\big)\mid\psi_h(t_{j,h}),\theta_h^\star(t_{j,h})\big]=\alpha\exp\big(-\theta_h^\star(t_{j,h})\big)\,\ell\big(\psi_h(t_{j,h})\big)=\alpha$. Eq. (2.11) also plays a key role in our formal set-up in Section 4.

Combining all elements, the first expression on the right-hand side of (2.9) converges to zero in probability regardless of the starting value $z_h(t_{0,h})$. A similar argument shows that the second expression in (2.9) converges to zero in probability, such that the filtering error $z_h(t_{i+1,h})$ converges to zero in probability and, thus, $\theta_h(t_{i,h})$ is consistent for $\theta_h^\star(t_{i,h})$ as $h\downarrow 0$. For our Pareto example this is illustrated in Figure 1 for two randomly drawn realizations of the data at different frequencies: $h=1/252$ and $h=1/(252\cdot 78)=1/19656$, i.e., a year of daily and a year of 5-minute frequency observations, respectively. As the frequency increases, we clearly see that (i) the filtering errors $z_h(t_{i,h})$ become smaller, such that $\theta_h(t_{i,h})$ and $\theta_h^\star(t_{i,h})$ overlap more and more closely, and (ii) the filter $\theta_h(t_{i,h})$ oscillates faster and faster around its target value $\theta_h^\star(t_{i,h})$.

To understand the latter and to find the right scaling for obtaining an approximating distribution for the filtering error $z_h(t_{i+1,h})$, we take another look at (2.8). By viewing this equation as a discretized stochastic differential equation we see that we must multiply the equation by $h^{-1/4}$ in order for drift and innovation terms to be balanced in the sense that the variance of the innovation terms and the size of the drift terms are of the same order in h. Indeed, defining $\tilde{z}_h(t_{i,h})=h^{-1/4}z_h(t_{i,h})$ and multiplying (2.8) by $h^{-1/4}$, we obtain

$$\begin{aligned}&h^{-1/4}\Delta z_h(t_{i+1,h})=-h^{1/2}\,A_{i+1,h}\;h^{-1/4}z_h(t_{i,h})+h^{1/4}\zeta_{i+1,h}+O_P(h^{3/4})\\ \iff\quad&\Delta\tilde{z}_h(t_{i+1,h})=-h^{1/2}\,A_{i+1,h}\;\tilde{z}_h(t_{i,h})+h^{1/4}\zeta_{i+1,h}+O_P(h^{3/4}). \qquad (2.12)\end{aligned}$$

Even though the drift and the variance terms are now both of order $h^{1/2}$ and thus balanced, their size remains uncommonly large compared to the 'standard' setting where both would be of order h. This is problematic if we wish to approximate the distribution of $\tilde{z}_h(t_{i+1,h})$ by that of a diffusion process with a drift term of bounded variation. To see this, recall that the time increment between

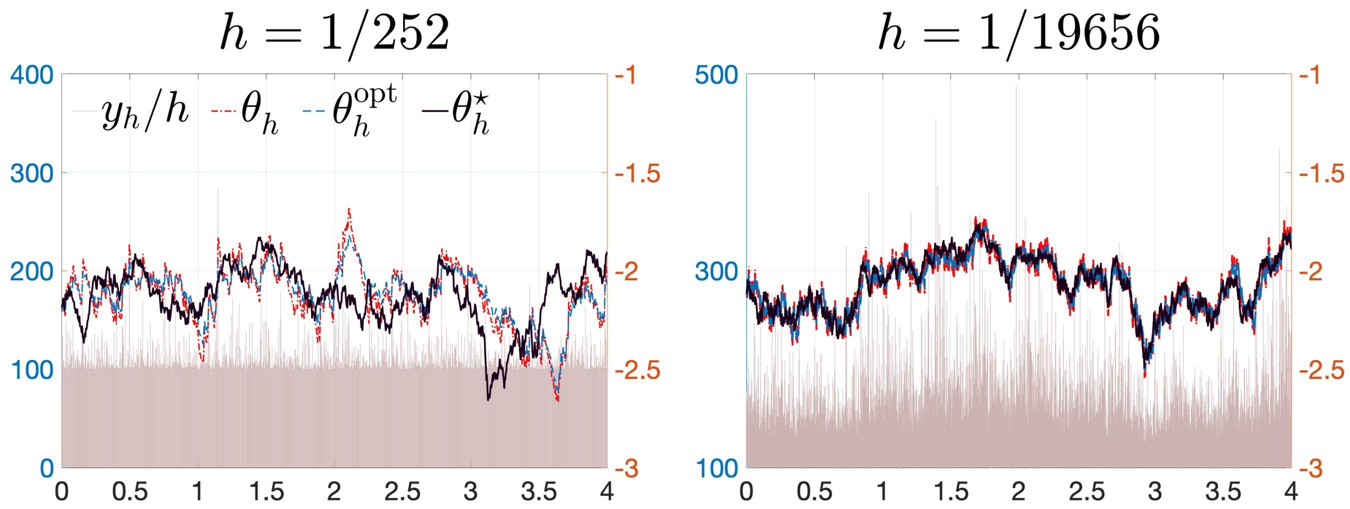

Figure 1: Consistency for the score-driven tail-shape filter for the Pareto distribution
The figure contains the path of a simulated tail-shape model using Eqs. (2.1) and (2.2) over a time span $T_0 = 4$ and h equal to either $1/252$ (daily data) or $1/19656$ (5-minute data). The parameters in the DGP are chosen as $(a, b, B, \sigma) = (1, -3, 3, 100)$ and the innovations η_i follow an i.i.d. standard normal distribution. The resulting target 'true' parameter $\theta_h^\star(t_{i,h})$ from (2.6), with $(\omega, \beta, \alpha) = (0.01, -0.01, 0.6)$, is given in black and uses the right-hand y-axis. The underlying data $y_h(t_{i,h})$ are drawn as gray bars at the bottom and use the left-hand y-axis. The filtered value $\theta_h(t_{i,h})$ using the filter (drawn in red) is given by Eq. (2.4). Drawn in blue, we also provide the filtered $\theta_h(t_{i,h})$ using the optimal filter from Section 4.3.

$\tilde{z}_h(t_{i+1,h})$ and $\tilde{z}_h(t_{i,h})$ is h. Hence, for any bounded $A_{i+1,h}$ the variation of the drift term on a finite time interval would be of order $h^{1/2} \cdot (1/h)$ and thus diverge to infinity for $h \downarrow 0$, i.e., the drift term of the approximating diffusion process would become of unbounded rather than of bounded variation.

A solution is suggested if we again iterate the autoregression for $\tilde{z}_h(t_{i+1,h})$ and use the same approximation as before: $1 - h^{1/2}\,x \approx \exp(-h^{1/2}\,x)$. In contrast to what we did earlier, we now only iterate back to $\tilde{z}_h(t_{m,h})$ rather than $\tilde{z}_h(t_{0,h})$ for some $0 < m < i+1$. Omitting approximation errors, we obtain

$$\tilde{z}_h(t_{i+1,h}) = \exp\left[-\sum_{j=m}^{i} h^{1/2}\,A_{j+1,h}\right]\tilde{z}_h(t_{m,h}) + \sum_{j=m}^{i} h^{1/4}\zeta_{j+1,h}\exp\left[-\sum_{k=j+1}^{i} h^{1/2}\,A_{k+1,h}\right]. \tag{2.13}$$

We can now immediately see that if $i - m$ is of order $h^{-1/2}$, then under appropriate conditions $\sum_{j=m}^{i} h^{1/2}\,A_{j+1,h}$ converges to a constant as it is a sum of $h^{-1/2}$ random terms multiplied by $h^{1/2}$. Similarly, we can expect that under appropriate conditions $\sum_{j=m}^{i} h^{1/4}\zeta_{j+1,h}\exp\left(-\sum_{k=j+1}^{i} h^{1/2}\,A_{k+1,h}\right)$ converges in distribution to a normally distributed random variable, as it is a sum of $h^{-1/2}$ random terms multiplied by $h^{1/4}$. This (heuristically) establishes a distributional convergence result for the scaled (inflated) filtering errors $\tilde{z}_h(t_{i,h})$, which is made precise in Section 4.2.

It is interesting to also highlight here the implication of the above heuristic result for the *time span* between $\tilde{z}_h(t_{i+1,h})$ and $\tilde{z}_h(t_{m,h})$. Because our observations are on an equidistant time grid with grid size h, the time span between $\tilde{z}_h(t_{i+1,h})$ and $\tilde{z}_h(t_{m,h})$ is of order $h^{-1/2}\;h = h^{1/2}$, which converges to zero as $h \downarrow 0$; see again Section 4.2 for the formal results. The formal proofs make use of results on the convergence of Markov chains to diffusion processes, similar to Nelson and Foster (1994). To use these for establishing an approximating distribution for the filtering error, we need a workaround for the time span that shrinks to zero as $h \downarrow 0$. As in Nelson and Foster (1994), the workaround consists of transforming time on a shrinking interval to a new time index on a fixed interval and considering the filtering error process and its convergence properties on this new time scale and fixed interval length.

Figure 2 illustrates the approach. In the two top panels, we plot the unscaled filtering errors $z_h(t_{i,h})$ for two frequencies using the same vertical axis. Looking at times $T = 1/2$ (corresponding to the lower limit, i.e., $t_{m+1,h} \equiv T$) and $T + Mh^{1/2}$ (corresponding to the upper limit, i.e., $t_{i+1,h} \equiv T + Mh^{1/2}$ such that $i - m$ is of order $h^{-1/2}$), each figure has a small (red) box stretching the time span $[T, T + Mh^{1/2}]$. As the frequency $1/h$ increases, this box shrinks in size, both vertically (due to the consistency) and horizontally (due to $T + Mh^{1/2} \to T$ as $h \downarrow 0$). Zooming out the filtering errors, scaling them by $h^{-1/4}$ and considering them on the interval $(0, M]$, we obtain the inserted figures in the top panels or the zoomed-in figures in the middle panels. We see that the scaled-up-and-streched-out filtering errors behave more and more like a non-degenerate stochastic process as $h \downarrow 0$. In Section 4.2 we formally show that this process converges to an Ornstein-Uhlenbeck type process in the limit, which is plausible by comparing the explicit solution of the Ornstein-Uhlenbeck process with Eq. (2.13) and the arguments given below this equation.

Finally, the two bottom panels in Figure 2 show the distributional convergence of the scaled filtering error. Clearly, an asymptotic normality result appears to apply. In particular, we obtain an explicit expression for the asymptotic variance of the filtering errors given the convergence of the scaled filtering errors to the Ornstein-Uhlenbeck process in the transformed time scale. Looking at Eq. (2.8), this variance will be composed of three elements: (i) the variance of the innovation process $B\eta_{i+1}$ of the underlying true time-varying parameter, scaled by $\mathrm{d}\theta_h^\star(t_{i,h})/\mathrm{d}\psi_h(t_{i,h})$, (ii) the variance of the filter innovations $g_h\big(y_h(t_{i+1,h}), \theta_h^\star(t_{i,h})\big)$, and (iii) the filter mean-reversion parameter as defined by the expectation of $A_{i+1,h} \equiv -\partial g_h\big(y_h(t_{i+1,h}), \theta_h^\star(t_{i,h})\big)/\partial\theta$. Given that the variance is a function of the filter's forcing variable $g_h(\,\cdot\,,\,\cdot\,)$, we can therefore ask ourselves what is the best $g_h(\,\cdot\,,\,\cdot\,)$ that minimizes the asymptotic filter error variance. It turns out that this is the score-driven filter of Creal et al. (2013) and Harvey (2013), where the score is scaled with the square root inverse Fisher information matrix in the sense of Creal et al. (2013).

The simulated path of the optimal and a score-driven filter with ad-hoc chosen parameters as shown in Figure 1 reveals that the optimal filter (blue) is much less erratic compared to the score-driven filter with ad-hoc parameters (red). After de-meaning and standardizing the filtering errors for their (different) asymptotic means and variances, the lower panels in Figure 2 show that both the ad-hoc and optimal filter have asymptotically normally distributed filtering errors as $h \downarrow 0$, as shown later in

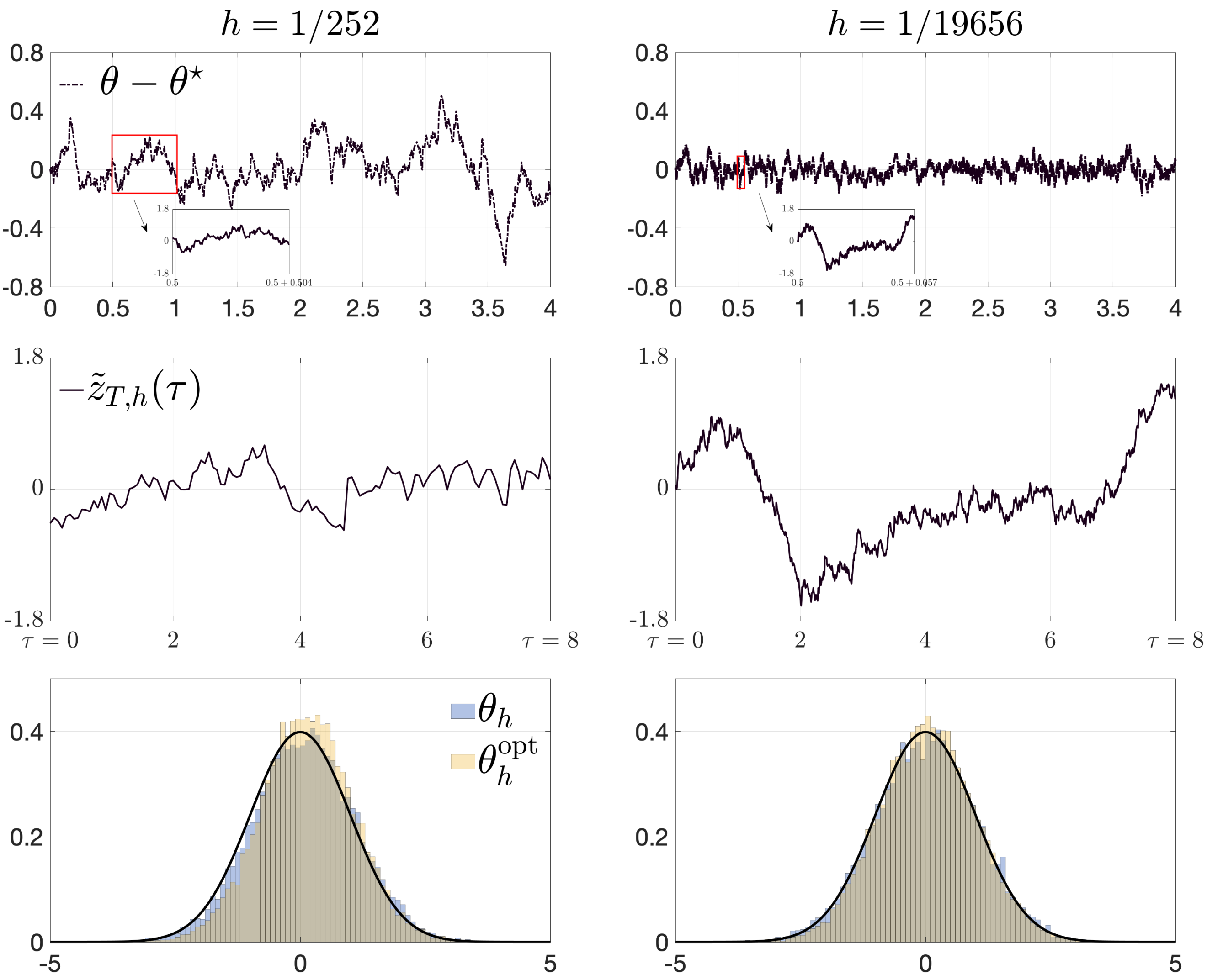

Figure 2: Motivating example: simulated sample paths of filtering errors, scaled and 'stretched' filtering errors, and their distribution

The left and right panels display the results for two different frequencies of h: 1/252 for daily data and 1/19656 for 5-minute data, respectively. The top panels show the filtering errors for a simulated path of the tail-shape model using Eqs. (2.1) and (2.2) over a time span of $T_0 = 4$. The parameters in the DGP are chosen as $(a, b, B, \sigma) = (1, 3, 3, 100)$. The filtered value $\boldsymbol{\theta}_h(t_{i,h})$ is computed using the score-driven filter as given by Eq. (2.4), with $(\omega, \beta, \alpha) = (-0.01, 0.01, 0.6)$. The boxed area shows the range of the fast time scale, $[T, T + Mh^{1/2}]$ for $T = 1/2$, and is zoomed-in inside the top panels as well as in the second row of plots. These plots in the second row scale the filtering errors by $h^{-1/4}$ and 'stretch' the time axis visually to the full width of the figure and range from $\tau \in [0, M]$ using a new time index τ. The bottom panels show the distribution of the scaled and re-centered filtering errors using the expressions for the asymptotic mean and variance in Eq. (4.22). The resulting histogram should become standard normal (benchmark curve) as $h \downarrow 0$. The panels also provide results for the filtered $\theta_h(t_{i,h})$ using the optimal filter (θ_h^{opt}) from Section 4.3.

this paper.

3 General set-up

We now extend Section 2 to the general set-up used in the remainder of this paper. Consider a discretely observed multivariate time series $\boldsymbol{y}_h(t_{i,h})$ of dimension k_y for $t_{i,h} = i \cdot h$, where $i = 1, \ldots, n_h$, $n_h = \lfloor T_0/h \rfloor$, and $T_0 > 0$ as defined in Section 2. Let $\boldsymbol{y}_h(t) = \boldsymbol{y}_h(t_{i,h})$ whenever $t \in [t_{i,h}, t_{i+1,h})$. This defines a sequence of stochastic processes $\{\boldsymbol{y}_h(t)\}_{t \in [0,T_0]}$ indexed by h, where we focus on the in-fill asymptotic setting with $h \downarrow 0$. We assume that $\boldsymbol{y}_h(t_{i+1,h}) \,\big|\, \boldsymbol{\psi}_h(t_{i,h})$ has a conditional probability density function (pdf) denoted as $q_h(\,\cdot\,;\boldsymbol{\psi}_h(t_{i,h}))$, where $\boldsymbol{\psi}_h(t_{i,h}) \in \boldsymbol{\Psi}$ for an open and convex parameter space $\boldsymbol{\Psi} \subset \mathbb{R}^{k_\psi}$, and where $\boldsymbol{\psi}_h(t_{i,h})$ is given through the stochastic recurrence equation

$$\Delta\boldsymbol{\psi}_h(t_{i+1,h}) = \boldsymbol{\psi}_h(t_{i+1,h}) - \boldsymbol{\psi}_h(t_{i,h}) = h\boldsymbol{a}_h\big(\boldsymbol{\psi}_h(t_{i,h})\big) + h^{1/2}\boldsymbol{B}_h\big(\boldsymbol{\psi}_h(t_{i,h})\big)\boldsymbol{\eta}_{i+1}. \tag{3.1}$$

The initial values $\boldsymbol{y}_h(0)$ and $\boldsymbol{\psi}_h(0)$ may be fixed or random. This set-up covers a wide range of data generating processes, including the set-ups for studying volatility of Nelson (1990, 1992) and Buccheri et al. (2021). It is clear, however, that the set-up is not restricted to the volatility case, but can also cover time-varying means, tail shapes, or other time-varying higher order moments.

We assume that neither $q_h(\,\cdot\,;\boldsymbol{\psi}_h(t_{i,h}))$ nor the transition equation (3.1) is known to the researcher. Instead, the researcher (possibly incorrectly) assumes that $\boldsymbol{y}_h(t_{i+1,h}) \,\big|\, \boldsymbol{\theta}_h(t_{i,h})$ has pdf $p_h(\,\cdot\,;\boldsymbol{\theta}_h(t_{i,h}))$, where $\boldsymbol{\theta}_h(t_{i,h}) \in \boldsymbol{\Theta}$ for some open convex parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^{k_\theta}$, and where $\boldsymbol{\theta}_h(t_{i,h})$ is obtained from the filtering equation

$$\Delta\boldsymbol{\theta}_h(t_{i+1,h}) = h\boldsymbol{\omega} + h\boldsymbol{\beta}\boldsymbol{\theta}_h(t_{i,h}) + h^{1/2}\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big), \tag{3.2}$$

where $\boldsymbol{g}_h$ is measurable for any h. To simplify notation, we confine ourselves to the common scenario above. It is worth noting, however, that we can allow the coefficients $\boldsymbol{a}_h$ and $\boldsymbol{B}_h$ in Eq. (3.1) and $\boldsymbol{\omega}$, and $\boldsymbol{\beta}$ in Eq. (3.2) to depend on both $\boldsymbol{\psi}_h(t_{i,h})$ and $\boldsymbol{y}_h(t_{i,h})$, for instance, to allow for asymmetry or leverage type effects. Additionally, we can allow for different shrinking rates of h in Eq. (3.2). We comment on this in Section 4.4.

We call the recursion in (3.2) the filter. In this paper, we particularly focus on score-driven filters. These set the function $\boldsymbol{g}_h(\boldsymbol{y}, \boldsymbol{\theta})$ equal to a scaled version of the 'score' of the predictive conditional model density, i.e., to a scaled version of $\partial \log p_h(\boldsymbol{y};\boldsymbol{\theta})/\partial\boldsymbol{\theta}$; see Creal et al. (2013).

The set-up above covers a range of different statistical models and forms of mis-specification. Next to the score-driven models, for instance, it also covers the case of a (mis-specified) GARCH model for a stochastic volatility DGP by setting $\boldsymbol{g}_h(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})) = h^{-1}\boldsymbol{y}_h(t_{i+1,h})^2 - \boldsymbol{\theta}_h(t_{i,h})$; see Nelson and Foster (1994). However, the set-up above allows for much more general forms of mis-specification. For instance, the DGP may be a skewed Student's t distribution $q_h(\,\cdot\,;\boldsymbol{\psi}_h(t_{i,h}))$ with time-varying skewness parameter $\boldsymbol{\psi}_h(t_{i,h})$, whereas the statistical model is a normal distribution $p_h(\,\cdot\,;\boldsymbol{\theta}_h(t_{i,h}))$ with time-varying mean $\boldsymbol{\theta}_h(t_{i,h})$. This goes considerably beyond the setting studied in Nelson and Foster (1994), which restricts itself to the setting where $q_h(\,\cdot\,;\boldsymbol{\psi}_h(t_{i,h}))$ and $p_h(\,\cdot\,;\boldsymbol{\theta}_h(t_{i,h}))$ may differ, but $\boldsymbol{\psi}_h(t_{i,h})$ and $\boldsymbol{\theta}_h(t_{i,h})$ still have the same interpretation.

Similar to Nelson and Foster (1994), we take the parameters $\boldsymbol{\omega}$ and $\boldsymbol{\beta}$ in (3.2) as given and abstract from the problem of data-driven parameter choice or parameter estimation. Instead, we focus on the problem of estimating the unknown time-varying parameters, referred to as the 'filtering problem' in the literature. The rates in terms of h in (3.2) are chosen in line with Nelson (1990, 1992) and result in a non-degenerate asymptotic statistical theory. It is well known that as $h \downarrow 0$, estimated parameters of GARCH processes tend to the integrated iGARCH case, and this is embedded in the current set-up. Jensen and Lange (2010) study a setting with more general rates of convergence of the different terms in (3.2) when studying parameter estimation. We can extend our current set-up easily to allow for such more general rates of h and we comment on this in Section 4.4. For a clearer exposition in the main text, however, we stick to the simplified set-up as in (3.2).

Given the substantial possible mis-specification between the DGP and the statistical model, our first task is to define a sensible target for the mis-specified filter. For this, we introduce the concept of a pseudo-true parameter and a pseudo-true parameter path. Consider the Kullback-Leibler (KL) divergence from the model density $p_h(\,\cdot\,;\boldsymbol{\theta})$ to the true unknown DGP density $q_h(\,\cdot\,;\boldsymbol{\psi})$, for any $h > 0$,

$$\begin{aligned}\mathrm{KL}\left(q_h(\boldsymbol{\psi}), p_h(\boldsymbol{\theta})\right) &= \int \log\left(\frac{q_h(\boldsymbol{y};\boldsymbol{\psi})}{p_h(\boldsymbol{y};\boldsymbol{\theta})}\right) q_h(\boldsymbol{y};\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{y} \\ &= \int \left(\log q_h(\boldsymbol{y};\boldsymbol{\psi})\right) q_h(\boldsymbol{y};\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{y} - \int \left(\log p_h(\boldsymbol{y};\boldsymbol{\theta})\right) q_h(\boldsymbol{y};\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{y}.\end{aligned} \tag{3.3}$$

We then consider the parameter $\boldsymbol{\theta}$ that minimizes this KL divergence and label it the pseudo-true parameter. The KL divergence has strong information-theoretic roots and provides an adequate target for the filter. It provides the best value of $\boldsymbol{\theta}$ to bring the model density close to the true unknown density, even if we do not know the form of the latter. Note that this definition of the pseudo-true parameter is applicable even in extreme cases of mis-specification, such as the earlier case where $q_h(\,\cdot\,;\boldsymbol{\psi}_h(t_{i,h}))$ is a skewed Student's t with time-varying skewness parameter $\boldsymbol{\psi}_h(t_{i,h})$ and $p_h(\,\cdot\,;\boldsymbol{\theta}_h(t_{i,h}))$ is a normal distribution with time-varying mean $\boldsymbol{\theta}_h(t_{i,h})$. In such cases $\boldsymbol{\psi}_h(t_{i,h})$ and $\boldsymbol{\theta}_h(t_{i,h})$ have very different meanings, and in the extreme could even be defined on very different parameter spaces $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$. Note that if the model is correctly specified, the pseudo-true and true parameter values coincide. However, the correct specification of the distribution family of q_h alone is neither necessary nor sufficient for this.

Throughout the paper, we impose the following assumptions related to KL divergence.

Assumptions:

For any $(\boldsymbol{\theta}, \boldsymbol{\psi}) \in \boldsymbol{\Theta} \times \boldsymbol{\Psi}$ and $h > 0$, let m_j, $j = 0, 1$, be some functions that are independent of $\boldsymbol{\theta}$ and integrable with respect to $Q_h(\,\cdot\,;\boldsymbol{\psi})$, where $Q_h(\,\cdot\,;\boldsymbol{\psi})$ is the cumulative distribution function (cdf) associated with $q_h(\,\cdot\,;\boldsymbol{\psi})$.

KL.1 Existence of KL divergence: $\forall(\boldsymbol{\theta}, \boldsymbol{\psi}) \in \boldsymbol{\Theta}\times\boldsymbol{\Psi}$, $\int \left|\log q_h(\boldsymbol{y};\boldsymbol{\psi})\right| q_h(\boldsymbol{y};\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{y} < \infty$, $\left|\log p_h(\boldsymbol{y},\boldsymbol{\theta})\right| \leq$

$m_0(\boldsymbol{y})$.

KL.2 Identification: $\forall \boldsymbol{\psi} \in \boldsymbol{\Psi}$, $\forall h > 0$, the function $\boldsymbol{\theta} \mapsto \mathrm{KL}\big(q_h(\boldsymbol{\psi}), p_h(\boldsymbol{\theta})\big)$ has a unique minimizer in $\boldsymbol{\Theta}$.

KL.3 Interchangeability of differentiation and integration: $\forall h > 0$, $\boldsymbol{\theta} \mapsto \log p_h(\boldsymbol{y}; \boldsymbol{\theta})$ is differentiable almost surely for $\boldsymbol{y}$; for all $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{k_\theta}) \in \boldsymbol{\Theta}$, $\big|\partial \log p_h(\boldsymbol{y}; \boldsymbol{\theta})/\partial\theta_j\big| \le m_1(\boldsymbol{y})$, $j = 1, \dots, k_\theta$.

KL.4 Existence of global implicit functions: Let $\boldsymbol{f}_h(\boldsymbol{\theta}, \boldsymbol{\psi}) = \int \big(\partial \log p_h(\boldsymbol{y}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}\big) q_h(\boldsymbol{y}; \boldsymbol{\psi}) \,\mathrm{d}\boldsymbol{y}$. For any $(\boldsymbol{\theta}, \boldsymbol{\psi}) \in \boldsymbol{\Theta} \times \boldsymbol{\Psi}$, suppose $\boldsymbol{f}_h$ is a continuous mapping and is continuously differentiable in the first variable $\boldsymbol{\theta}$. Moreover, there exists a unique *mapping $\boldsymbol{\iota}_h : \boldsymbol{\Psi} \to \boldsymbol{\Theta}$ such that $\boldsymbol{f}_h\big(\boldsymbol{\iota}_h(\boldsymbol{\psi}), \boldsymbol{\psi}\big) = \mathbf{0}$.*

The first two assumptions have been previously considered by White (1982, Assumption A3). Assumption KL.1 ensures that the KL divergence is well-defined. Assumption KL.2 allows us to define the pseudo-true parameter at time i, denoted by $\boldsymbol{\theta}^\star_h(t_{i,h})$, as

$$\boldsymbol{\theta}^\star_h(t_{i,h}) = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{KL}\Big(q_h(\boldsymbol{\psi}_h(t_{i,h})), p_h(\boldsymbol{\theta})\Big), \qquad i = 1, \dots, n_h. \tag{3.4}$$

Assumption KL.3 can be found in standard textbooks, e.g., Schilling (2017, Theorem 12.5), Klenke (2020, Theorem 6.28). With $\big|\log p_h(\boldsymbol{y}, \boldsymbol{\theta})\big| \le m_0(\boldsymbol{y})$ in Assumption KL.1, Assumption KL.3 ensures the interchangeability of differentiation (with respect to $\boldsymbol{\theta}$) and the integral sign in the minimization problem (3.4) above. It is possible to impose a weaker condition with more complex notation, see Talvila (2001). By Assumptions KL.1 and KL.3, we obtain the first-order condition (FOC) for $\boldsymbol{\theta}^\star_h(t_{i,h})$:

$$\boldsymbol{f}_h\big(\boldsymbol{\theta}^\star_h(t_{i,h}), \boldsymbol{\psi}_h(t_{i,h})\big) = \int \frac{\partial \log p_h(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta}^\star_h(t_{i,h})} q_h(\boldsymbol{y}; \boldsymbol{\psi}_h(t_{i,h}))\,\mathrm{d}\boldsymbol{y} = \mathbf{0}. \tag{3.5}$$

With the FOC (3.5), Assumption KL.4 implies that $\boldsymbol{\theta}^\star_h(t_{i,h}) = \boldsymbol{\iota}_h\big(\boldsymbol{\psi}_h(t_{i,h})\big)$ for some appropriate function $\boldsymbol{\iota}_h(\,\cdot\,)$. Clearly, $\boldsymbol{\iota}_h(\boldsymbol{\psi}) = \boldsymbol{\psi}$ if the model is correctly specified. One can find a sufficient condition for the existence of a global implicit function in Zhang and Ge (2006). We provide some examples in Section 5.

We call the sequence $\boldsymbol{\theta}^\star_h(t_{i,h})$, for $i = 1, \dots, n_h$, the pseudo-true parameter path and use it as a target for the filter $\boldsymbol{\theta}_h(t_{i,h})$ in (3.2). If $\boldsymbol{\theta}_h(t_{i,h})$ succeeds in recovering $\boldsymbol{\theta}^\star_h(t_{i,h})$ pointwise at every moment in time, then the filter succeeds in adapting the mis-specified dynamic density $p_h(\,\cdot\,; \boldsymbol{\theta}_h(t_{i,h}))$ as best as possible in a KL sense to the unknown true dynamic density $q_h(\,\cdot\,; \boldsymbol{\psi}_h(t_{i,h}))$, despite the density, the time-varying parameter, and the dynamic set-up all being mis-specified. This seems the best one can hope for given the generality of the current set-up. In the next section, we establish the conditions for filter consistency and show that score-driven filters automatically satisfy these conditions. In addition, we establish an asymptotic normality result for the filtering errors and show the conditions under which score-driven filters minimize the asymptotic filter error variance.

Motivating example (continued). For the Pareto tail-shape model from Section 2, Eq. (3.5) boils down to

$$\int \left[\exp\big(-\theta_h^{\star}(t_{i,h})\big)\ \ln\left(\frac{y}{h\,\sigma}\right) - 1 \right] q_h(y; \psi_h(t_{i,h}))\,\mathrm{d}y = \exp\big(-\theta_h^{\star}(t_{i,h})\big)\ \ell\big(\psi_h(t_{i,h})\big) - 1 = 0, \quad (3.6)$$

such that we get the pseudo-true parameter path $\theta_h^{\star}(t_{i,h}) = \ln\big(\ell[\psi_h(t_{i,h})]\big)$, $i = 1, \ldots, n_h$, as in (2.6).

4 Asymptotic theory

In this section, we derive our three main results. First, in Section 4.1 we formulate the conditions under which the filter $\boldsymbol{\theta}_h(t_{i,h})$ in Eq. (3.2) is consistent for $\boldsymbol{\theta}_h^{\star}(t_{i,h})$. We then conclude that the score-driven filter of Creal et al. (2013) satisfies these conditions and is therefore consistent for the true parameter path if the observation density p_h is correctly specified, or for the pseudo-true parameter path if the model is mis-specified. Second, in Section 4.2 we derive the distributional convergence of the filtering errors $\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}_h^{\star}(t_{i,h})$ and obtain the asymptotic filter error variance. As alluded to in Section 2 the distributional convergence of the filtering error is non-standard in two regards: first the scaling sequence is $h^{-1/4}$ rather than the usual $h^{-1/2}$, and second the convergence in distribution takes place on a transformed time axis. In Section 4.3 we then consider the optimal choice of $\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big)$ if the pdf of the DGP $q_h(\boldsymbol{y}_h(t_{i+1,h}); \boldsymbol{\psi}_h(t_{i,h}))$ is known and obtain that score-driven filters are in that case also optimal. We also show that the optimal shape of $\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big)$ may become time-varying if the dimension of $\boldsymbol{\psi}_h(t_{i,h})$ exceeds that of $\boldsymbol{\theta}_h(t_{i,h})$. We conclude with a short discussion on extensions to models with alternative convergence rates.

In the rest of this section, we adopt the following notation. For $i = 0, 1, \ldots$, let

$$\boldsymbol{x}_h(t_{i,h}) = \begin{pmatrix} \boldsymbol{v}_h(t_{i,h}) \\ \boldsymbol{\psi}_h(t_{i,h}) \end{pmatrix}, \qquad \boldsymbol{z}_h(t_{i,h}) = h^{-\kappa}\big(\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}_h^{\star}(t_{i,h})\big), \qquad \kappa \in [0, 1/4], \quad (4.1)$$

such that $\boldsymbol{z}_h(t_{i,h})$ is the scaled filtering error. The new k_v-dimensional process $\{\boldsymbol{v}_h(t_{i,h})\}$ is explained further below and may take different forms, contingent upon the specific application. For instance, in cases where the coefficients $\boldsymbol{a}_h$ and $\boldsymbol{B}_h$ in (3.1) depend on $\boldsymbol{y}_h(t_{i,h})$ or its partial sums, it is appropriate to set $\boldsymbol{v}_h(t_{i,h})$ equal to $\boldsymbol{y}_h(t_{i,h})$, respectively to its partial sums.

The joint process $\big\{\big(\boldsymbol{x}_h(t_{i,h}), \boldsymbol{z}_h(t_{i,h})\big),\ i \geq 0\big\}$ is assumed to be a time-homogeneous Markov chain.

For simplicity and in line with Nelson and Foster (1994), we also assume there is no feedback from $\boldsymbol{z}_h(t_{i,h})$ to $\boldsymbol{x}_h(t_{i,h})$. That is, given $\big(\boldsymbol{x}_h(t_{i,h}), \boldsymbol{z}_h(t_{i,h})\big)$, we assume that $\boldsymbol{x}_h(t_{i+1,h})$ is independent of $\boldsymbol{z}_h(t_{i,h})$; see Eq. (A.5) in Appendix A. Throughout this section, we set $\big(\boldsymbol{\upsilon}_h(t), \boldsymbol{\psi}_h(t), \boldsymbol{z}_h(t), \boldsymbol{\theta}_h(t)\big) = \big(\boldsymbol{\upsilon}_h(t_{i,h}), \boldsymbol{\psi}_h(t_{i,h}), \boldsymbol{z}_h(t_{i,h}), \boldsymbol{\theta}_h(t_{i,h})\big)$ for $t \in [t_{i,h}, t_{i+1,h})$.

4.1 Consistency

We first consider filter consistency, i.e., pointwise convergence of $\boldsymbol{\theta}_h(t)$ to $\boldsymbol{\theta}_h^\star(t)$ for any $t > 0$. The following conditions are imposed to obtain filter consistency (FC).

Assumptions:

FC.1 The dynamics of $\big\{\boldsymbol{\psi}_h(t)\big\}$*:*

(a) $\forall \eta > 0$, $\lim_{h\downarrow 0}\sup_{\|\boldsymbol{\psi}\|\leq\eta}\big\|\boldsymbol{a}_h(\boldsymbol{\psi}) - \boldsymbol{a}(\boldsymbol{\psi})\big\| = 0$, $\lim_{h\downarrow 0}\sup_{\|\boldsymbol{\psi}\|\leq\eta}\big\|\boldsymbol{B}_h(\boldsymbol{\psi}) - \boldsymbol{B}(\boldsymbol{\psi})\big\| = 0$, *where* $\boldsymbol{a}(\cdot)$ *and* $\boldsymbol{B}(\cdot)$ *are continuous;*

(b) $\forall i \in \mathbb{Z}^+$, *the following moment conditions hold almost surely:* $\mathbb{E}\big(\boldsymbol{\eta}_{i+1}\,\big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\big) = \boldsymbol{0}$; $\mathbb{E}\Big(\boldsymbol{\eta}_{i+1}\boldsymbol{\eta}_{i+1}^\top\,\Big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\Big) = \boldsymbol{I}_{k_\psi}$; $\forall \eta > 0$, $\exists K_\eta > 0$ *such that (s.t.)*

$$\sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\leq\eta} \mathbb{E}\Big(\big\|\boldsymbol{\eta}_{i+1}\big\|^4\,\Big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\Big) \leq K_\eta. \tag{4.2}$$

FC.2 The dynamics of $\big\{\boldsymbol{\upsilon}_h(t)\big\}$*: for every* $\eta > 0$*, there exist* $\boldsymbol{\mu}(\cdot) \in \mathbb{R}^{k_\upsilon}$*,* $\boldsymbol{\Omega}_{\upsilon\eta}(\cdot) \in \mathbb{R}^{k_\upsilon\times k_\psi}$*, and* $\boldsymbol{\Omega}_{\upsilon\upsilon}(\cdot) \in \mathbb{R}^{k_\upsilon\times k_\upsilon}$*, s.t.*

$$\lim_{h\downarrow 0}\sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\leq\eta}\left\|h^{-1}\mathbb{E}\Big(\Delta\boldsymbol{\upsilon}_h(t_{i+1,h})\,\big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\Big) - \boldsymbol{\mu}(\boldsymbol{x})\right\| = 0,$$

$$\lim_{h\downarrow 0}\sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\leq\eta}\left\|h^{-1/2}\mathbb{E}\Big(\big(\Delta\boldsymbol{\upsilon}_h(t_{i+1,h})\big)\boldsymbol{\eta}_{i+1}^\top\,\big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\Big) - \boldsymbol{\Omega}_{\upsilon\eta}(\boldsymbol{x})\right\| = 0,$$

$$\lim_{h\downarrow 0}\sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\leq\eta}\left\|h^{-1}\mathbb{E}\Big(\big(\Delta\boldsymbol{\upsilon}_h(t_{i+1,h})\big)\big(\Delta\boldsymbol{\upsilon}_h(t_{i+1,h})\big)^\top\,\Big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\Big) - \boldsymbol{\Omega}_{\upsilon\upsilon}(\boldsymbol{x})\right\| = 0,$$

where $\boldsymbol{\mu}(\cdot)$*,* $\boldsymbol{\Omega}_{\upsilon\eta}(\cdot)$*, and* $\boldsymbol{\Omega}_{\upsilon\upsilon}(\cdot)$*, are continuous, and* $\boldsymbol{\mu}(\cdot)$ *is uniformly bounded. Moreover,*

$$\lim_{h\downarrow 0}\sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\leq\eta} h^{-1}\mathbb{E}\Big(\big\|\Delta\boldsymbol{\upsilon}_h(t_{i+1,h})\big\|^4\,\Big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\Big) = 0, \tag{4.3}$$

FC.3 The initial values $\big(\boldsymbol{x}_h(0), \boldsymbol{z}_h(0)\big)$ *satisfy: (a)* $\limsup_{h\downarrow 0}\big[\mathbb{E}\,\|\boldsymbol{x}_h(0)\|^2 + \mathbb{E}\,\|\boldsymbol{z}_h(0)\|^2\big] < \infty$*; (b) as* $h \downarrow 0$*,* $\boldsymbol{x}_h(0) \Rightarrow \boldsymbol{x}(0)$*, where* $\boldsymbol{x}(0)$ *has probability measure* $\nu_{x,0}$*. Moreover,* $\nu_{x,0}$*,* $\Big(\boldsymbol{\mu}(\boldsymbol{x})^\top, \boldsymbol{a}(\boldsymbol{\psi})^\top\Big)^\top$*, and* $\begin{pmatrix} \boldsymbol{\Omega}_{\upsilon\upsilon}(\boldsymbol{x}) & \boldsymbol{\Omega}_{\upsilon\eta}(\boldsymbol{x})\boldsymbol{B}(\boldsymbol{\psi})^\top \\ \boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{\Omega}_{\upsilon\eta}(\boldsymbol{x})^\top & \boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{B}(\boldsymbol{\psi})^\top \end{pmatrix}$ *uniquely specify the distribution of a diffusion process* $\boldsymbol{x}(t)$

(as the distributional limit of $\boldsymbol{x}_h(t_{i,h})$ as $h \downarrow 0$) with initial distribution $\nu_{x,0}$, drift vector $\left(\boldsymbol{\mu}(\boldsymbol{x})^\top, \boldsymbol{a}(\boldsymbol{\psi})^\top\right)^\top$, and diffusion matrix $\begin{pmatrix} \boldsymbol{\Omega}_{\upsilon\upsilon}(\boldsymbol{x}) & \boldsymbol{\Omega}_{\upsilon\eta}(\boldsymbol{x})\boldsymbol{B}(\boldsymbol{\psi})^\top \\ \boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{\Omega}_{\upsilon\eta}(\boldsymbol{x})^\top & \boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{B}(\boldsymbol{\psi})^\top \end{pmatrix}$.

FC.4 Implicit functions: for $j = 1, 2, \ldots, k_\theta$, let $\boldsymbol{\iota}_{h,j}(\cdot)$ be twice differentiable, where $\boldsymbol{\iota}_{h,j}(\cdot)$ is the j_{th} element of $\boldsymbol{\iota}_h(\cdot)$.

(a) $\forall \eta > 0$, $\exists C_{0,\eta} > 0$ s.t.

$$\limsup_{h\downarrow 0} \sup_{\|\boldsymbol{\psi}\|\le\eta} \left\|\boldsymbol{\iota}_h(\boldsymbol{\psi})\right\| \le C_{0,\eta}. \tag{4.4}$$

Moreover, for any $\kappa \in [0, 1/4)$ in Eq. (4.1),

$$\lim_{h\downarrow 0} \sup_{\|\boldsymbol{\psi}\|\le\eta} h^{1/4-\kappa} \left\|\frac{\partial \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right\| = 0, \qquad \lim_{h\downarrow 0} \sup_{\|\boldsymbol{\psi}\|\le\eta} h^{1/2-\kappa} \left\|\frac{\partial^2 \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\top}\right\| = 0. \tag{4.5}$$

(b) $\forall \boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \in \boldsymbol{\Psi}$, $\forall h > 0$, there exist nondecreasing functions $V_{h,j} : \mathbb{R}^+ \to \mathbb{R}^+$, $j = 1, 2, \ldots, k_\theta$, such that

$$\left\|\frac{\partial^2 \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi}_1)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\top} - \frac{\partial^2 \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi}_2)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\top}\right\| \le V_{h,j}\big(\|\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2\|\big). \tag{4.6}$$

Moreover, $\forall \eta > 0$, $\forall \kappa \in [0, 1/4]$, $j = 1, 2, \ldots, k_\theta$,

$$\lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} h^{-1-2\kappa} \mathbb{E}\Big[V_{h,j}^2\big(\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|\big) \|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|^4 \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\Big] = 0. \tag{4.7}$$

FC.5 Forcing variables: $\forall h > 0$, $\boldsymbol{y} \in \mathbb{R}^{k_y}$, $\boldsymbol{g}_h(\boldsymbol{y}, \cdot)$ is twice differentiable, satisfying

$$\mathbb{E}\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h^\star(t_{i,h})\big) \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\Big] = \boldsymbol{0}. \tag{4.8}$$

Moreover, for every $\eta, \tilde{\eta} > 0$,

$$\lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} h^\kappa\, \mathbb{E}\left[\sup_{\|\boldsymbol{\theta}\|\le\tilde{\eta}} \left\|\frac{\partial^2 \boldsymbol{g}_{h,j}\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}\big)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right\| \,\middle|\, \boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\right] = 0, \tag{4.9}$$

$$\limsup_{h\downarrow 0} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} \mathbb{E}\left[\left\|\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big)\right\|^2 \,\middle|\, \boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\right] \le C_{1,\eta}, \tag{4.10}$$

where $\kappa \in [0, 1/4]$ as in Eq. (4.1), and $\boldsymbol{g}_{h,j}(\boldsymbol{y}, \boldsymbol{\theta})$ is the j_{th} element of $\boldsymbol{g}_h(\boldsymbol{y}, \boldsymbol{\theta})$, $j = 1, \ldots, k_\theta$, *and $C_{1,\eta} > 0$ is some constant. Finally, define*

$$\boldsymbol{A}(\boldsymbol{x}) = -\lim_{h\downarrow 0} \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{g}_h\Big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h^\star(t_{i,h})\Big) \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x} \right]. \tag{4.11}$$

For any $\eta > 0$ and $\|\boldsymbol{x}\| \leq \eta$, all the eigenvalues of $\boldsymbol{A}(\boldsymbol{x})$ have strictly positive real parts.

We briefly discuss the assumptions in turn. Assumption FC.1 restricts the dynamics of the true time-varying parameter $\boldsymbol{\psi}_h(t_{i,h})$. In our current formulation of Assumption FC.1, we consider parameters $\boldsymbol{a}_h$ and $\boldsymbol{B}_h$ that do not depend on $\boldsymbol{y}_h(t_{i,h})$. If necessary, the setup can easily be extended to accommodate such dependence by incorporating $\boldsymbol{y}_h(t_{i,h})$ into $\boldsymbol{v}_h(t_{i,h})$. We do not pursue this here in order not to overburden the notation. Eq. (4.2) excludes too fat-tailed innovation processes for $\boldsymbol{\psi}_h(t_{i,h})$ for consistent filtering results.

Conditions FC.2 and FC.3 are in line with the work on convergence to continuous time volatility filters of Nelson (1990, 1992) and Nelson and Foster (1994) and set the convergence rates of the mean and variance of $\boldsymbol{v}_h(t_{i,h})$ in our in-fill asymptotic experiment. The generality introduced by the new process $\boldsymbol{v}_h(t_{i,h})$ serves two purposes. First, it allows us to consider more general settings than the volatility setting of Nelson (1996) and Nelson and Foster (1994). Second, it also allows us to relax some of the moment conditions required in Nelson (1996), Nelson and Foster (1994), and Buccheri et al. (2021). In particular, note that we do not require fourth-order conditional moments of the data $\boldsymbol{y}_h(t_{i+1,h})$ itself. This is particularly important when considering robust filters as in Creal et al. (2013) or Buccheri et al. (2021). Instead, (4.3) restricts the degree of fat-tailedness of the new $\boldsymbol{v}_h(t_{i+1,h})$ to ensure filter consistency. The condition might possibly be relaxed to the existence of slightly more than second-order moments. We stick to the current stricter formulation to keep in line with Ethier and Nagylaki (1980, 1988).

Condition FC.4 requires sufficient smoothness of the implicit function that links the pseudo-true parameter $\boldsymbol{\theta}_h^\star(t_{i,h})$ to the true time-varying parameter $\boldsymbol{\psi}_h(t_{i,h})$. It is worth noting that Assumption FC.4(b) is automatically satisfied if $\partial \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})/\partial \boldsymbol{\psi}$ is independent of $\boldsymbol{\psi}$, which can be achieved by setting $V_{h,j} \equiv 0$. More generally, the condition holds when $\left\|\partial^2 \boldsymbol{\iota}_{h,j}(\cdot)/\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top\right\| \leq M_{h,j}$, where $M_{h,j} > 0$ is deterministic with $h^{1/2-\kappa} M_{h,j} = o(1)$, for every $j = 1 \ldots, k_\theta$.

Finally, Assumption FC.5 puts conditions on the forcing variable $\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h^\star(t_{i,h})\big)$ of the filter, i.e., on its 'News Impact Curve' (NIC). Eq. (4.8) ensures that the forcing variable has zero conditional expectation when evaluated at the pseudo-true parameter. For instance, if $\boldsymbol{y}_h(t_{i+1,h}) \,\big|\, \boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{v}, \boldsymbol{\psi})$ is independent of the σ-algebra $\sigma\big(\boldsymbol{v}_h(t_{s,h}),\, s \leq i\big)$, then Eq. (4.8) may be equivalently written as

$$\mathbb{E}\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\iota}_h(\boldsymbol{\psi})\big) \,\big|\, \boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{v}, \boldsymbol{\psi})\Big] = \int \boldsymbol{g}_h\big(\boldsymbol{y}, \boldsymbol{\iota}_h(\boldsymbol{\psi})\big) q_h\left(\boldsymbol{y}; \boldsymbol{\psi}\right)\, \mathrm{d}\boldsymbol{y} = \boldsymbol{0}.$$

It is then immediately clear from the FOC (3.5) that a GAS(1,1) score-driven filter as proposed by Creal et al. (2013) with

$$\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\theta}\big) = \boldsymbol{s}_h(\boldsymbol{\theta})\frac{\partial \log p_h\left(\boldsymbol{y}_h(t_{i+1,h});\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}, \tag{4.12}$$

where $\boldsymbol{s}_h(\boldsymbol{\theta})$ is some scaling matrix such as the inverse of the Fisher information matrix, automatically satisfies condition (4.8). Condition (4.8) is also implicitly assumed in Nelson and Foster (1994), but evaluated at $\boldsymbol{\psi}_h(t_{i,h})$ rather than $\boldsymbol{\theta}^\star_h(t_{i,h})$. As we allow for much more general forms of mis-specification, even scenarios where the dimensions of $\boldsymbol{\psi}_h(t_{i,h})$ and $\boldsymbol{\theta}^\star_h(t_{i,h})$ differ, using $\psi_h(t_{i,h})$ in (4.8) would not be viable. Therefore, we make the assumption explicit and also replace the true parameter $\boldsymbol{\psi}_h(t_{i,h})$ with its pseudo-true equivalent $\boldsymbol{\theta}^\star_h(t_{i,h})$. Condition (4.10) is similar to Eq. (3.19) in Nelson and Foster (1994) and is typically easily satisfied by score-driven filters for fat-tailed observations; see also Buccheri et al. (2021).

The final condition in Eq. (4.11), and in particular the positive real part of its eigenvalues, is important for consistency. It coincides with condition Eq. (2.11) in Section 2, where it was used in the intuitive informal derivation to establish independence from initial conditions and a vanishing filtering errors in (2.10). This condition is easily met for many time-varying volatility models as well as for many non-standard examples; see Section 5. Again, the theory can be further generalized to allow for vanishing eigenvalues at appropriate rates at the cost of further notational complexity. For instance, we can allow the right-hand side of (4.11) to vanish at the rate of h^c for some $c \in (0, 1/2)$ and still obtain consistency. For $c = 1/2$, however, consistency is lost and we only obtain mean-reversion of $\boldsymbol{\theta}_h(t_{i,h})$ around $\boldsymbol{\theta}^\star_h(t_{i,h})$. This is an important special case and happens in some settings where we filter for a time-varying mean parameter. We refer to Section 5.3 for an example of this.

With the above set of assumptions in place, we can now formulate the following theorem that establishes the consistency of the filter to the pseudo-true time-varying parameter path.

Theorem 1 (Filter Consistency)
Under Assumptions KL.1 - KL.4, FC.1 - FC.5, we have

$$h^{-\kappa}\big(\boldsymbol{\theta}_h(t) - \boldsymbol{\theta}^\star_h(t)\big) \xrightarrow{p} \mathbf{0}_{k_\theta\times 1}, \qquad \forall\kappa\in[0,1/4), \quad \forall t\in(0,T_0), \tag{4.13}$$

where "$\xrightarrow{p}$" denotes convergence in probability as $h \downarrow 0$.

The strict upper bound 1/4 of κ is reflected in the proof, see Eq. (B.13). It is a bound that is to be expected given the earlier work of Nelson and Foster (1994), who show that the scaled filtering error for $\kappa = 1/4$ has a non-degenerate limiting distribution.

Motivating example (continued). Suppose Assumptions FC.1 and FC.3 hold in the motivating example of Section 2, where η_{i+1} follows a standard normal distribution. The process $\boldsymbol{v}_h$ plays no particular role in the example, so we set it to $\boldsymbol{v}_h \equiv \mathbf{0}$, such that also Assumption FC.3 holds trivially. In this case, conditioning on $\boldsymbol{x}_h(t_{i,h})$ is equivalent to conditioning on $\psi_h(t_{i,h})$ only. Assumption FC.4 is met given the specification of the pseudo-true parameter derived at the end of Section 3 in terms of ψ. Finally, Assumption FC.5 can be verified given the score-driven specification. Note that the function $g_h\big(y_h(t_{i+1,h}), \theta_h^\star(t_{i,h})\big)$ only holds a logarithmic transformation of the data, such that we only require a squared log-moment of $y_h(t_{i+1,h})$ to exist for the moment conditions in the assumption to be satisfied. This holds by construction given $\ell(\psi_h(t_{i,h})) < 1/4$ and thus 4th order conditional moments of $y_h(t_{i+1,h})$ exist. As shown in Section 2 $A(\psi) = \alpha\, \mathbb{E}\Big([\ell(\psi)]^{-1} \ln \big(y_h(t_{i+1,h})/(h\sigma) \big) \mid \psi_h(t_{i,h}) = \psi \Big) = \alpha$, which is positive as long as $\alpha > 0$. As all the assumptions are satisfied, Theorem 1 applies and the score-driven filter in the motivating example is consistent for $\theta_h^\star(t_{i,h})$, even though the filter is mis-specified for the true dynamics as well as for the parameterization. The consistency was visually illustrated in Figure 1. ■

4.2 Weak convergence

In this section, we establish the weak convergence of the filtering error $\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}_h^\star(t_{i,h})$, where again we allow the model to be severely mis-specified. As mentioned at the beginning of Section 4 and illustrated in Section 2, this leads to a non-standard asymptotic setting, similar to that in the volatility context of Nelson and Foster (1994) and Nelson (1996). The non-standard feature is that we do not only need to scale the filtering error by $h^{-1/4}$ to establish a non-degenerate limiting result, but in addition also need to change the time axis and time span over which we consider the filtering error process.

To accommodate the first, let $\widetilde{\boldsymbol{z}}_h(t_{i,h})$ be defined as $\boldsymbol{z}_h(t_{i,h})$, but with $\kappa = 1/4$. This provides exactly the correct inflation factor for the filter error variance. To accommodate the second part, let

$$\widetilde{\boldsymbol{z}}_{T,h}(\tau) = \widetilde{\boldsymbol{z}}_h\big(T + \tau h^{1/2}\big), \tag{4.14}$$

where $T \in [0, T_0)$, τ is a new time index on a 'fast time scale' in the sense of Nelson and Foster (1994) and Nelson (1996). As discussed in the motivating example (Section 2), this new time scale is needed because the filtering error process converges faster and faster (time-wise) to zero as $h \downarrow 0$ compared to a regular time process. As a result, the filter $\boldsymbol{\theta}_h(t_{i,h})$ oscillates faster and faster around the pseudo-true $\boldsymbol{\theta}_h^\star(t_{i,h})$ as $h \downarrow 0$. This may be counter-intuitive at first, as both $\boldsymbol{\theta}_h(t)$ and $\boldsymbol{\theta}_h^\star(t)$ converge on the normal time scale, whereas their difference apparently does not. The reason is that the drift term of the unscaled filtering error process has the same order of magnitude in h as its diffusion part; see the appendix for proof. On a normal time scale, the scaled filtering errors $\widetilde{\boldsymbol{z}}_h(t_{i,h})$ gradually resemble a white noise process when h approaches 0, as shown in the top panel of Figure 2. On the new fast time scale τ, by contrast, the process $\big\{\boldsymbol{x}_{T,h}(\tau) = \boldsymbol{x}_h\big(T + \tau h^{1/2}\big)\big\}$ operates more and more slowly as $h \downarrow 0$ and eventually degenerates into a constant. At the same time $\{\widetilde{\boldsymbol{z}}_{T,h}(\tau)\}$ converges weakly to a diffusion process, yielding a non-degenerate distributional result for the asymptotic scaled filtering errors (the inserted frames in the top panels of Figure 2). We consider this limiting diffusion at a point $T + M\,h^{1/2}$ for some positive real M. This implies that we consider the filtering error process on a shorter and shorter time interval that in the limit collapses to a point. We show in Section 5 that this limiting result serves as a good approximation for a fixed h, even when it is relatively large.

The following assumptions are required to establish the result for the asymptotic distribution (AD). Note that $\Delta\boldsymbol{\psi}_h(t+h) = \boldsymbol{\psi}_h\big((\lfloor t/h\rfloor \cdot h + h\big) - \boldsymbol{\psi}_h\big(\lfloor t/h\rfloor \cdot h\big)$; similar expressions hold for the other processes.

Assumptions:

AD.1 The process $\big\{\big(\boldsymbol{x}_h(t), \tilde{\boldsymbol{z}}_h(t)\big)\big\}$, $t > 0$, satisfies Assumptions FC.1, FC.4, and FC.5, albeit with some modifications as described below.

(a) For $j = 1, \ldots, k_\theta$, if $\partial\boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ does not rely on $\boldsymbol{\psi}$, let $\varphi_1 > 0$ be some constant (can be arbitrarily small). Otherwise, let $\varphi_1 > 2$. Then replace (4.2) in Assumption FC.1(b) by $\mathbb{E}\Big(\big\|\boldsymbol{\eta}_{\lfloor t/h\rfloor+1}\big\|^{2+\varphi_1}\,\Big|\,\boldsymbol{x}_h(t) = \boldsymbol{x},\, \tilde{\boldsymbol{z}}_h(t) = \boldsymbol{z}\Big) \le K_\eta$.

(b) Replace (4.5) and (4.7) in Assumption FC.4, respectively, by

$$\lim_{h\downarrow 0}\ \sup_{\|\boldsymbol{\psi}\|\le\eta}\ h^{1/4}\left\|\frac{\partial\boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right\|^{1+2/\varphi_2} = 0, \qquad \lim_{h\downarrow 0}\ \sup_{\|\boldsymbol{\psi}\|\le\eta}\ h^{1/4}\left\|\frac{\partial^2\boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}\right\| = 0, \tag{4.15}$$

and

$$\lim_{h\downarrow 0}\ \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta}\ h^{-3/2(1+\varphi_2/2)}\mathbb{E}\Big[V_{h,j}^{2+\varphi_2}\big(\|\Delta\boldsymbol{\psi}_h(t+h)\|\big) \\ \|\Delta\boldsymbol{\psi}_h(t+h)\|^{2(2+\varphi_2)}\,\Big|\,\boldsymbol{x}_h(t) = \boldsymbol{x},\, \tilde{\boldsymbol{z}}_h(t) = \boldsymbol{z}\Big] = 0, \tag{4.16}$$

for some $\varphi_2 > 0$.

(c) Replace (4.9) and (4.10) in Assumption FC.5, respectively, by

$$\lim_{h\downarrow 0}\sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} h^{1/2}\,\mathbb{E}\left[\sup_{\|\boldsymbol{\theta}\|\le\tilde{\eta}}\left\|\frac{\partial^2 \boldsymbol{g}_{h,j}\big(\boldsymbol{y}_h(t+h),\boldsymbol{\theta}\big)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right\|^2\,\middle|\,\boldsymbol{x}_h(t)=\boldsymbol{x},\,\widetilde{\boldsymbol{z}}_h(t)=\boldsymbol{z}\right]=0,$$

$$\limsup_{h\downarrow 0}\sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta}\mathbb{E}\left[\left\|\boldsymbol{g}_h\big(\boldsymbol{y}_h(t+h),\boldsymbol{\theta}_h(t)\big)\right\|^{2+\varphi_3}\,\middle|\,\boldsymbol{x}_h(t)=\boldsymbol{x},\,\widetilde{\boldsymbol{z}}_h(t)=\boldsymbol{z}\right]\le C_{1,\eta},$$

where $\varphi_3>0$. Further, $\boldsymbol{A}(\boldsymbol{x})$ in Assumption FC.5 should be read as

$$\boldsymbol{A}(\boldsymbol{x})=-\lim_{h\downarrow 0}\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{g}_h\Big(\boldsymbol{y}_h(t+h),\boldsymbol{\theta}_h^\star(t)\Big)\,\middle|\,\boldsymbol{x}_h(t)=\boldsymbol{x}\right].$$

We require $\boldsymbol{A}(\boldsymbol{x})$ to be twice continuously differentiable in $\boldsymbol{x}$.

AD.2 As $h\downarrow 0$, the following convergence holds uniformly on every bounded $(\boldsymbol{x},\boldsymbol{z})$ set:

$$h^{-1/2}\mathbb{E}\Big(\Delta\boldsymbol{v}_h(t+h)\,\Big|\,\boldsymbol{x}_h(t)=\boldsymbol{x},\,\widetilde{\boldsymbol{z}}_h(t)=\boldsymbol{z}\Big)\to\boldsymbol{0},\tag{4.17}$$

$$h^{-1/2}\,\mathbb{C}\mathrm{ov}\,\Big(\Delta\boldsymbol{v}_h(t+h)\,\Big|\,\boldsymbol{x}_h(t)=\boldsymbol{x},\,\widetilde{\boldsymbol{z}}_h(t)=\boldsymbol{z}\Big)\to\boldsymbol{0}.\tag{4.18}$$

For every $\eta>0$, there exists $C_{2,\eta}>0$ s.t.

$$\lim_{h\downarrow 0}\sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta}\mathbb{E}\Big(\left\|h^{-1/2}\Delta\boldsymbol{v}_h(t+h)\right\|^{2+\varphi_4}\,\Big|\,\boldsymbol{x}_h(t)=\boldsymbol{x},\,\widetilde{\boldsymbol{z}}_h(t)=\boldsymbol{z}\Big)\le C_{2,\eta},\tag{4.19}$$

where $\varphi_4>0$. Moreover,

$$\lim_{h\downarrow 0}\sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta}h^{1/2}\mathbb{E}\left[\left\|\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{g}_h\Big(\boldsymbol{y}_h(t+h),\boldsymbol{\theta}_h^\star(t)\Big)\right\|^2\,\middle|\,\boldsymbol{x}_h(t)=\boldsymbol{x},\,\widetilde{\boldsymbol{z}}_h(t)=\boldsymbol{z}\right]=0.\tag{4.20}$$

Condition (4.16) implies (4.7) for any $\varphi_2>0$. It is easily met in the examples in this paper. Overall, the set of assumptions for weak convergence is stricter than that for filter consistency. However, it is worth noting that the conditional moments specified in Assumption AD.1(a) and (4.19) can potentially be slightly weaker compared to their counterparts for filter consistency. This distinction may arise from the weaker nature of Theorem 2, which solely focuses on the process within an interval that gradually shortens to length zero in the limit.

We now obtain the following limiting distributional approximation for the scaled filtering errors.

Theorem 2 (Asymptotic Distribution)

Let $t_{T,\tau,h}=T+\tau h^{1/2}$ for $T\in[0,T_0)$. Define $\boldsymbol{\zeta}_h(t)=\boldsymbol{g}_h\big(\boldsymbol{y}_h(t),\boldsymbol{\theta}_h^\star(t-h)\big)-\big(\partial\boldsymbol{\iota}_h(\boldsymbol{\psi})/\partial\boldsymbol{\psi}^\top\big)\,\boldsymbol{B}(\boldsymbol{\psi})\,\boldsymbol{\eta}_{\lfloor t/h\rfloor}$, *and*

$$\boldsymbol{\Sigma}(\boldsymbol{x}) = \lim_{h\downarrow 0} \mathbb{E}\Big[\boldsymbol{\zeta}_h\big(t_{T,\tau,h}+h\big)\boldsymbol{\zeta}_h\big(t_{T,\tau,h}+h\big)^\top \,\Big|\, \boldsymbol{x}_h\big(t_{T,\tau,h}\big) = \boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{\psi})\Big]. \tag{4.21}$$

Let $\boldsymbol{\Gamma}_0$ be any bounded open subset of $\mathbb{R}^{k_y+k_\psi+k_\theta}$ on which for some $\varepsilon > 0$, the real parts of all the eigenvalues of $\boldsymbol{A}(\boldsymbol{x})$ are bounded below by ε. Suppose that $\boldsymbol{\Sigma}(\boldsymbol{x})$ is twice continuously differentiable in $\boldsymbol{x}$, and that Assumptions AD.1 - AD.2 hold for $t = t_{T,\tau,h}$. Under Assumptions KL.1 - KL.4, for every $(\boldsymbol{x}, \boldsymbol{z}) \in \boldsymbol{\Gamma}_0$ and $\tau > 0$,

$$\widetilde{\boldsymbol{z}}_{T,h}(\tau) \,\Big|\, \big(\boldsymbol{x}_{T,h}(0), \widetilde{\boldsymbol{z}}_{T,h}(0)\big) = (\boldsymbol{x}, \boldsymbol{z}) \xrightarrow{d} \mathcal{N}\Big(\boldsymbol{b}(\tau, \boldsymbol{x}, \boldsymbol{z}), \boldsymbol{V}(\tau, \boldsymbol{x})\Big), \tag{4.22}$$

where "$\xrightarrow{d}$" denotes convergence in distribution as $h \downarrow 0$, $\boldsymbol{b}(\tau, \boldsymbol{x}, \boldsymbol{z}) = \exp\big[-\tau \boldsymbol{A}(\boldsymbol{x})\big]\boldsymbol{z}$, and

$$\boldsymbol{V}(\tau, \boldsymbol{x}) = \exp\big[-\tau\boldsymbol{A}(\boldsymbol{x})\big]\bigg\{\int_0^\tau \exp\big[s\boldsymbol{A}(\boldsymbol{x})\big]\boldsymbol{\Sigma}(\boldsymbol{x})\exp\big[s\boldsymbol{A}(\boldsymbol{x})^\top\big]\,\mathrm{d}s\bigg\}\exp\big[-\tau\boldsymbol{A}(\boldsymbol{x})^\top\big].$$

As mentioned, Theorem 2 focuses on the process within the interval $[T, T+\tau h^{1/2}]$ which eventually collapses to the point T as $h \downarrow 0$. It is worth noting that Theorem 2 can be extended to the weak convergence of the process $\big\{\widetilde{\boldsymbol{z}}_{T,h}(\tau)\big\}_{\tau\in[0,M]}$, $M < \infty$, i.e., for the time interval $[T, T + Mh^{1/2}]$ on the natural time scale. More specifically, as discussed in Nelson and Foster (1994) and Nelson (1996), it converges weakly to the diffusion

$$\mathrm{d}\boldsymbol{Z}(\tau) = -\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Z}(\tau)\mathrm{d}\tau + \boldsymbol{\Sigma}(\boldsymbol{x})^{1/2}\mathrm{d}\boldsymbol{W}_\tau, \qquad \text{as } h \downarrow 0, \tag{4.23}$$

where $\boldsymbol{W}_\tau$ is a standard Brownian motion.

As opposed to Theorem 1, the weak convergence in Eq (4.22) is conditional on the initial values $\big(\boldsymbol{x}_{T,h}(0), \widetilde{\boldsymbol{z}}_{T,h}(0)\big) = (\boldsymbol{x}, \boldsymbol{z})$. To ensure it is a meaningful condition, one implicitly requires that the sample path of $\big\{\big(\boldsymbol{x}_h(T), \widetilde{\boldsymbol{z}}_h(T)\big)\big\}$ is nonexplosive everywhere for an arbitrary $T \geq 0$. Under this initial condition, some assumptions for filter consistency can be dropped. For instance, FC.3 (a) – (b) are in that case automatically fulfilled. Both theorems require the sample path of the process $\big\{\big(\boldsymbol{x}_h(t), \widetilde{\boldsymbol{z}}_h(t)\big)\big\}$ to have no discrete jumps as guaranteed by (4.2) and (4.3) for filter consistency, and Assumption AD.1(a) and (4.19) for weak convergence. As such, the conditions for obtaining asymptotic distributions and filter consistency are indeed similar.

Since the convergence in distribution in (4.22) holds for every finite $\tau = M$, a variant of Lemma 5.2 of Helland (1982) as given in Appendix F implies it also holds for some $M \equiv M(h) \to \infty$, as $h \downarrow 0$. For this $M(h)$, we may have a simple approximation of the conditional distribution of $\widetilde{\boldsymbol{z}}_h(T + Mh^{1/2})$

given by $\mathcal{N}\big(\mathbf{0}, \boldsymbol{V}^*\big)$, where

$$\boldsymbol{V}^* = \int_0^\infty \exp\big[-s\boldsymbol{A}(\boldsymbol{x})\big]\,\boldsymbol{\Sigma}(\boldsymbol{x})\exp\big[-s\boldsymbol{A}(\boldsymbol{x})^\top\big]\,\mathrm{d}s. \tag{4.24}$$

By Problem 6.6 in Karatzas and Shreve (1998, p. 357), $\boldsymbol{V}^*$ also satisfies the following useful identity

$$\boldsymbol{A}(\boldsymbol{x})\boldsymbol{V}^* + \boldsymbol{V}^*\boldsymbol{A}(\boldsymbol{x})^\top = \boldsymbol{\Sigma}(\boldsymbol{x}). \tag{4.25}$$

It is worth mentioning that neither Lemma 5.2 in Helland (1982) nor its variant in Appendix F specifies the divergence rate of $M(h)$. Therefore, there is no theoretical guidance on determining which sequence $\{M(h)\}$ ensures the validity of (4.22) for $\widetilde{\boldsymbol{z}}_{T,h}(M(h))$. Choosing some rate $M(h)$ with $h^{1/2}M(h) \to 0$, as $h \downarrow 0$, may be used for purposes of approximation, but lacks as yet a formal guarantee that it is the appropriate rate to obtain distributional convergence of $\widetilde{\boldsymbol{z}}_{T,h}(M(h))$ for diverging $M(h)$.

Motivating example (continued). Assumption AD.1 is trivial to check given the functional form of $\theta_h^\star(t_{i,h})$ and the Gaussianity of η_{i+1}. Also part (c) follows easily for $\alpha > 0$ from the existence of 4th order conditional moments of $y_h(t_{i+1,h})$ given $\ell(\psi_h(t_{i,h})) \in (0, 1/4)$ and the simple expression of $A(\boldsymbol{x}) = A(\psi) = \alpha$ as derived at the end of Section 4.1. Finally, Assumption AD.2 is fulfilled given $\boldsymbol{v}_h \equiv \mathbf{0}$ and the finiteness of a squared logarithmic conditional moment of $y_h(t_{i+1,h})$. Therefore, Theorem 2 applies and the filtering errors in the motivating example converge weakly to a normal on the 'fast time scale' τ. This explains the distributional convergence visualized in Figure 2. ■

4.3 Optimality

One might ask whether there is an 'optimal' form of the forcing variable $\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big)$ that defines the filter. In Nelson and Foster (1994) and Nelson (1996), optimality was defined in terms of minimizing the matrix mean-squared error (MSE), i.e., the trace of $\boldsymbol{V}^*$. From the definition of $\boldsymbol{\Sigma}(\boldsymbol{x})$ in Eq. (4.21), we see that the asymptotic filter error variance has three components: (i) the variance of the forcing variable $\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big)$, (ii) the variance of the (re-scaled) innovation $\boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{\eta}_{i+1}$, and (iii) the expectation of their cross-term. The latter can be non-zero if $\boldsymbol{\eta}_{i+1}$ and $\boldsymbol{y}_h(t_{i+1,h})$ are conditionally dependent, which is not the case in our leading example. Each of the above three components plays a role in the expression for the optimal filter under mis-specification.

Further intuition may be obtained as follows. As the asymptotic variance contains both $\boldsymbol{\Sigma}(\boldsymbol{x})$ and $\boldsymbol{A}(\boldsymbol{x})$, the asymptotic variance can be lowered if we can 'minimize' $\boldsymbol{\Sigma}(\boldsymbol{x})$. This can be done by minimizing the expected quadratic distance between $\boldsymbol{g}_h$ and the scaled conditional expectation of $\boldsymbol{\eta}_{i+1}$

given $\boldsymbol{y}_h(t_{i+1,h})$. 'Maximizing' $\boldsymbol{A}(\boldsymbol{x})$, on the other hand, also helps to lower the asymptotic variance and corresponds to maximizing the covariance between $\boldsymbol{g}_h$ and the score of $q_h(\,\cdot\,)$ with respect to $\boldsymbol{\psi}$. This can be seen from the definition of $\boldsymbol{A}(\boldsymbol{x})$, which can be rewritten as

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{x}) &= -\lim_{h\downarrow 0}\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}^{\star\top}}\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\theta}^\star_h(t_{i,h})\big)\,\Big|\,\boldsymbol{x}_h(t_{i,h})=\boldsymbol{x}=(\boldsymbol{v},\boldsymbol{\psi})\right]\\ &= \lim_{h\downarrow 0}\mathbb{E}\left[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\theta}^\star_h(t_{i,h})\big)\frac{\partial\log q_h\big(\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\psi}\big)}{\partial\boldsymbol{\psi}^\top}\,\Big|\,\boldsymbol{x}_h(t_{i,h})=\boldsymbol{x}=\big(\boldsymbol{v},\boldsymbol{\psi}\big)\right]\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)^\top.\end{aligned}$$

This equation is similar to Eq. (4.1) in Nelson and Foster (1994) and Eq. (A.7) in Nelson (1996), except for the additional scaling by $\boldsymbol{Q}_h(\tilde{\boldsymbol{\theta}})$, defined in Eq. (4.31) below, which is due to the fact that our setting allows for much more general forms of mis-specification.

Before continuing, it is worth mentioning that we allow for $k_\theta \leq k_\psi$ rather than restricting $k_\theta = k_\psi$ as in Nelson and Foster (1994) and Nelson (1996). This allows us to also study settings where there are more time-varying parameters in the DGP than there are in the model itself.[3] When discussing filter optimality, this additional flexibility requires some additional notation to establish the appropriate results. To this end, we define an artificial pseudo-true parameter $\boldsymbol{\theta}^\dagger_h(t_{i,h}) = \boldsymbol{\iota}^\dagger_h\big(\boldsymbol{\psi}_h(t_{i,h})\big) \in \mathbb{R}^{k_\psi - k_\theta}$ that summarizes the additional (unknown) parameters in $\boldsymbol{\psi}$. For instance, in the context of our leading example, we could add a time-varying scale parameter to the DGP besides the currently time-varying tail shape. The new parameter $\boldsymbol{\theta}^\dagger_h(t_{i,h})$ would then be a new function of $\boldsymbol{\psi}_h(t_{i,h})$ that would be different from $\boldsymbol{\theta}^\star_h(t_{i,h})$ almost everywhere. We gather $\boldsymbol{\theta}^\star_h(t_{i,h})$ and the new $\boldsymbol{\theta}^\dagger_h(t_{i,h})$ in a new vector and write

$$\tilde{\boldsymbol{\theta}}_h(t_{i,h}) = \tilde{\boldsymbol{\iota}}_h\big(\boldsymbol{\psi}_h(t_{i,h})\big) = \begin{pmatrix}\boldsymbol{\theta}^\star_h(t_{i,h})\\ \boldsymbol{\theta}^\dagger_h(t_{i,h})\end{pmatrix} \in \mathbb{R}^{k_\psi}, \qquad \tilde{\boldsymbol{\iota}}_h(\,\cdot\,) = \begin{pmatrix}\boldsymbol{\iota}_h(\,\cdot\,)\\ \boldsymbol{\iota}^\dagger_h(\,\cdot\,)\end{pmatrix}. \tag{4.26}$$

We assume that this new mapping $\tilde{\boldsymbol{\iota}}(\,\cdot\,)$ is one-to-one.

To establish the optimality result, we also introduce the following notation:

$$\boldsymbol{S}_h\left(\boldsymbol{y},\boldsymbol{\psi}\right) = \frac{\partial\log q_h\left(\boldsymbol{y};\boldsymbol{\psi}\right)}{\partial\boldsymbol{\psi}}, \tag{4.27}$$

$$\boldsymbol{P}_h\left(\boldsymbol{y},\boldsymbol{v},\boldsymbol{\psi}\right) = \left(\frac{\partial\boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^\top}\right)\;\mathbb{E}\Big[\boldsymbol{B}(\boldsymbol{\psi})\,\boldsymbol{\eta}_{i+1}\,\Big|\,(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{v}_h(t_{i,h}),\boldsymbol{\psi}_h(t_{i,h}))=(\boldsymbol{y},\boldsymbol{v},\boldsymbol{\psi})\Big]. \tag{4.28}$$

The score $\boldsymbol{S}_h(\boldsymbol{y},\boldsymbol{\psi})$ is now the score of the DGP density $q_h(\boldsymbol{y};\boldsymbol{\psi})$ rather than of the model density $p_h(\boldsymbol{y};\boldsymbol{\theta})$. We will find later on that this DGP score is key in finding the optimal filter. In fact, if the true DGP density $q_h(\boldsymbol{y};\boldsymbol{\psi})$ is known, a score-driven filter turns out to be optimal. The definition of $\boldsymbol{P}_h(\boldsymbol{y},\boldsymbol{v},\boldsymbol{\psi})$ is needed if the DGP has conditional dependence between $\boldsymbol{y}_h(t_{i+1,h})$ and $\boldsymbol{\eta}_{i+1}$. This

[3] Note that when $k_\theta > k_\psi$, the pseudo-true parameter path is not well-defined, as there may be multiple pseudo-true parameter paths that correspond to the same true path.

can be important for instance in a volatility context with leverage-type effects, i.e., a conditional correlation between financial returns and volatilities, with volatilities increasing more with negative returns than with positive returns $\boldsymbol{y}_h(t_{i+1,h})$.

We now formulate the following assumptions for obtaining filter optimality (FO).

Assumptions:

Let $k_\psi \geq k_\theta$ and let $\widetilde{\boldsymbol{\Theta}} \subset \mathbb{R}^{k_\psi}$ be open and convex.

FO.1 $\forall h > 0$ and $\forall \tilde{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$, $\boldsymbol{\theta} \mapsto \boldsymbol{g}_h\big(\boldsymbol{y}, \tilde{\boldsymbol{\theta}}\big)$ is differentiable almost surely for $\boldsymbol{y}$; $\big|\partial \boldsymbol{g}_{h,i}\big(\boldsymbol{y}, \tilde{\boldsymbol{\theta}}\big)/\partial \theta_j\big|$, $i, j = 1, \ldots, k_\psi$, are dominated by functions that are independent of $\tilde{\boldsymbol{\theta}} = \big(\theta_1, \ldots, \theta_{k_\theta}, \theta_{k_\theta+1}, \ldots, \theta_{k_\psi}\big) \in \widetilde{\boldsymbol{\Theta}}$ and integrable with respect to the distribution $Q_h(\,\cdot\,; \boldsymbol{\psi})$.

FO.2 As in (4.8), we assume

$$\mathbb{E}\Big[\boldsymbol{g}_h\Big(\boldsymbol{y}_h(t_{i+1,h}), \tilde{\boldsymbol{\theta}}_h(t_{i,h})\Big) \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, \widetilde{\boldsymbol{z}}_h(t_{i,h}) = \boldsymbol{z}\Big] = \boldsymbol{0}. \tag{4.29}$$

FO.3 $\forall h > 0$, $q_h\big(\boldsymbol{y}_h(t_{i+1,h})\,; \boldsymbol{\psi}\big)$ is continuously differentiable in $\boldsymbol{\psi}$ almost everywhere, with one-sided partial derivatives everywhere, and for some $\varphi > 0$,

$$\mathbb{E}\bigg(\Big\|\boldsymbol{P}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\upsilon}, \boldsymbol{\psi}\big)\Big\|^{2+\varphi} \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{\upsilon}, \boldsymbol{\psi})\bigg), \qquad \mathbb{E}\bigg(\Big\|\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\psi}\big)\Big\|^{2+\varphi} \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{\upsilon}, \boldsymbol{\psi})\bigg),$$

are bounded uniformly on every bounded $(\boldsymbol{\upsilon}, \boldsymbol{\psi})$ set as $h \downarrow 0$.

FO.4 $\forall \tilde{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$, suppose the inverse map $\tilde{\boldsymbol{\iota}}_h^{-1}(\tilde{\boldsymbol{\theta}})$ exists and is differentiable. There is a unique, positive semidefinite solution $\boldsymbol{W}$ to the matrix Riccati equation:

$$\begin{aligned}
&\mathbb{E}\Big[\boldsymbol{P}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\upsilon}, \boldsymbol{\psi}\big)\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\psi}\big)^\top \boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)^\top \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{\upsilon}, \boldsymbol{\psi})\Big]\boldsymbol{W} \\
&\quad + \boldsymbol{W}\mathbb{E}\Big[\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\psi}\big)\boldsymbol{P}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\upsilon}, \boldsymbol{\psi}\big)^\top \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{\upsilon}, \boldsymbol{\psi})\Big] \\
&\quad + \boldsymbol{W}\mathbb{E}\Big[\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\psi}\big)\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\psi}\big)^\top \boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)^\top \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{\upsilon}, \boldsymbol{\psi})\Big]\boldsymbol{W} \\
&= \mathbb{E}\Bigg\{\bigg[\bigg(\frac{\partial \boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^\top}\bigg)\boldsymbol{B}\boldsymbol{\eta}_{i+1} - \boldsymbol{P}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\upsilon}, \boldsymbol{\psi}\big)\bigg] \\
&\quad \times \bigg[\bigg(\frac{\partial \boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^\top}\bigg)\boldsymbol{B}\boldsymbol{\eta}_{i+1} - \boldsymbol{P}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\upsilon}, \boldsymbol{\psi}\big)\bigg]^\top \,\bigg|\, \boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{\upsilon}, \boldsymbol{\psi})\Bigg\},
\end{aligned} \tag{4.30}$$

where

$$\boldsymbol{Q}_h(\tilde{\boldsymbol{\theta}}) = \left(\frac{\partial \tilde{\boldsymbol{\iota}}_h^{-1}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^{\star\top}}\right)^\top. \tag{4.31}$$

Assumption FO.1 guarantees, similarly to Assumption KL.3, the interchangeability of differentiation (with respect to the parameters) and integration for $\boldsymbol{g}_h$. The moment conditions in Assumption FO.3 ensure that (4.30) and (4.32) are well defined. Finally, Assumption FO.4 ensures that there is a solution to the equation that determines the minimum variance filter. With these assumptions in place, we can formulate the following result.

Theorem 3 (Optimal Filter)
Under Assumptions KL.1 - KL.4, AD.1 - AD.2, FO.1 - FO.4, for every $h > 0$, the trace of covariance matrix is minimized if

$$\boldsymbol{g}_h\big(\,\cdot\,,\boldsymbol{\upsilon},\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big) = \boldsymbol{g}_h\Big(\,\cdot\,,\boldsymbol{\upsilon},\boldsymbol{\iota}_h(\boldsymbol{\psi}),\boldsymbol{\iota}_h^{\dagger}(\boldsymbol{\psi})\Big) = \boldsymbol{P}_h\big(\,\cdot\,,\boldsymbol{\upsilon},\boldsymbol{\psi}\big) + \boldsymbol{W}_h\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)\boldsymbol{S}_h\big(\,\cdot\,,\boldsymbol{\psi}\big), \tag{4.32}$$

where $\boldsymbol{W}_h = \boldsymbol{W}_h(\boldsymbol{\upsilon},\boldsymbol{\psi})$ is the positive semidefinite solution to (4.30), and $\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)$ is defined in Eq. (4.31). Moreover, the minimized covariance matrix is $\boldsymbol{W}_h$.

The result in (4.32) is clearly recognizable as a score-driven model based on the true DGP density $q_h(\boldsymbol{y};\boldsymbol{\psi})$. The score-driven filter transforms the dynamics for $\boldsymbol{\psi}$ into dynamics for $\tilde{\boldsymbol{\theta}}$ via the transformation $\boldsymbol{Q}_h(\tilde{\boldsymbol{\theta}})$. The term involving $\boldsymbol{P}_h(\,\cdot\,,\boldsymbol{\upsilon},\boldsymbol{\psi})$ accounts for the conditional correlation between $\boldsymbol{y}_h(t_{i+1,h})$ and $\boldsymbol{\eta}_{i+1}$, such as leverage effects in the volatility context, and causes the filter to react possibly asymmetrically to new observations $\boldsymbol{y}_h(t_{i+1,h})$.

In general, a closed-form expression of $\boldsymbol{W}_h$ is not available. There are, however, two important special cases of Theorem 3 that are worth mentioning. We summarize them in the following two corollaries.

Corollary 1
If $\boldsymbol{P}_h \equiv \boldsymbol{0}$, then $\boldsymbol{W}_h = \boldsymbol{\mathcal{B}}_h^{-1/2}\big(\boldsymbol{\mathcal{B}}_h^{1/2}\boldsymbol{\mathcal{C}}_h\boldsymbol{\mathcal{B}}_h^{1/2}\big)^{1/2}\boldsymbol{\mathcal{B}}_h^{-1/2}$ is a solution to (4.30) with

$$\boldsymbol{\mathcal{B}}_h = \mathbb{E}\Big[\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\psi}\big)\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\psi}\big)^{\top}\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)^{\top}\Big|\,\boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{\upsilon},\boldsymbol{\psi})\Big],$$

and $\boldsymbol{\mathcal{C}}_h = \left(\dfrac{\partial\boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^{\top}}\right)\boldsymbol{B}\boldsymbol{B}^{\top}\left(\dfrac{\partial\boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^{\top}}\right)^{\top}$.

Corollary 2
Consider $\boldsymbol{\iota}_h(\boldsymbol{\psi}) = \boldsymbol{\psi}$ for any $h > 0$ and $\boldsymbol{\psi} \in \boldsymbol{\Psi}$. In the case where $\boldsymbol{P}_h \equiv \boldsymbol{0}$, and the model density $p_h\big(\,\cdot\,;\boldsymbol{\iota}_h(\boldsymbol{\psi})\big)$ coincides with the DGP density $q_h(\,\cdot\,;\boldsymbol{\psi})$, the optimal filter in (4.32) is given by the score-driven filter as defined by Creal et al. (2013).

Moreover, let's suppose that $\mathbb{E}\Big[\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\psi}\big)\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\psi}\big)^{\top}\,\Big|\,\boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{\upsilon},\boldsymbol{\psi})\Big]$ *and* $\boldsymbol{B}\boldsymbol{B}^{\top}$

commute (Horn and Johnson 2012, Theorem 1.3.12), for instance, they are diagonal, which clearly includes the scalar case $k_\psi = 1$.[4] *In this scenario, the asymptotic covariance matrix $\boldsymbol{W}_h$ simplifies to:*

$$\boldsymbol{W}_h = \left\{ \mathbb{E}\Big[\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\psi}\big) \boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\psi}\big)^\top \,\Big|\, \boldsymbol{x}_h(t_{i,h}) = (\boldsymbol{\upsilon}, \boldsymbol{\psi}) \Big] \right\}^{-1/2} \big(\boldsymbol{B}\boldsymbol{B}^\top\big)^{1/2}. \tag{4.33}$$

The term of conditional expectation in (4.33) represents the Fisher information matrix; a larger Fisher information results in a smaller $\boldsymbol{W}_h$. Furthermore, $\boldsymbol{B}\boldsymbol{B}^\top$ measures the variation of the true parameter path; a larger $\boldsymbol{B}\boldsymbol{B}^\top$ implies less accuracy of the optimal filter. Multiplying the score with $\boldsymbol{W}_h$ in (4.33), we achieve an optimal score-driven filter.

Particularly Corollary 2 is interesting. It highlights that a score-driven filter is not only consistent but also optimal if the true density is known, despite mis-specifications in the stochastic nature of the true process $\{\boldsymbol{\psi}_h(t)\}$. This provides a further theoretical motivation for the use of score-driven models. It also provides guidance for the choice of scaling: for achieving a minimum asymptotic filter error variance, inverse square root information matrix scaling could be preferable to the commonly used inverse information matrix scaling, depending on the value of $\boldsymbol{B}$.

Even in case the true model density is known, we note that (4.32) would have to be re-parameterized before it could be applied: the formulation is in terms of $\boldsymbol{\psi}$ rather than $\boldsymbol{\theta}$. A re-parameterization can easily be obtained by exploiting the one-to-one mapping between $\boldsymbol{\psi}$ and $\tilde{\boldsymbol{\theta}}$. We then obtain

$$\begin{aligned} \boldsymbol{g}_h\Big(\,\cdot\,, \boldsymbol{\upsilon}, \boldsymbol{\iota}_h(\boldsymbol{\psi}), \boldsymbol{\iota}_h^\dagger(\boldsymbol{\psi})\Big) &= \boldsymbol{P}_h\Big(\,\cdot\,, \boldsymbol{\upsilon}, \tilde{\boldsymbol{\iota}}_h^{-1}\Big(\boldsymbol{\iota}_h(\boldsymbol{\psi}), \boldsymbol{\iota}_h^\dagger(\boldsymbol{\psi})\Big)\Big) \\ &\quad + \boldsymbol{W}_h\Big(\boldsymbol{\upsilon}, \tilde{\boldsymbol{\iota}}_h^{-1}\Big(\boldsymbol{\iota}_h(\boldsymbol{\psi}), \boldsymbol{\iota}_h^\dagger(\boldsymbol{\psi})\Big)\Big) \boldsymbol{Q}_h\Big(\boldsymbol{\iota}_h(\boldsymbol{\psi}), \boldsymbol{\iota}_h^\dagger(\boldsymbol{\psi})\Big) \boldsymbol{S}_h\Big(\,\cdot\,, \tilde{\boldsymbol{\iota}}_h^{-1}\Big(\boldsymbol{\iota}_h(\boldsymbol{\psi}), \boldsymbol{\iota}_h^\dagger(\boldsymbol{\psi})\Big)\Big), \end{aligned}$$

and replace $\boldsymbol{\iota}_h\big(\boldsymbol{\psi}_h(t_{i,h})\big)$ by $\boldsymbol{\theta}_h(t_{i,h})$ in the recursion equation (3.2).

4.4 Alternative convergence rates

At this point, it is useful to remark that the consistency and weak convergence results of Sections 4.1 to 4.3 can be generalized to a setting with more general rates of h in the transition equations for $\boldsymbol{\psi}_h(t_{i+1,h})$ and/or $\boldsymbol{\theta}_h(t_{i+1,h})$. More specifically, we are interested in variations of the filter recursion (3.2) of the form

$$\Delta\boldsymbol{\theta}_h(t_{i+1,h}) = h^{\delta_\omega}\boldsymbol{\omega} + h^{\delta_\beta}\boldsymbol{\beta}\boldsymbol{\theta}_h(t_{i,h}) + h^{\delta_\alpha/2}\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big), \qquad \delta_\omega, \delta_\beta \geq \delta_\alpha/2 > 0.$$

[4] Two matrices $\boldsymbol{A}, \boldsymbol{B}$ are said to commute if $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$.

For $\delta_\omega = \delta_\beta = \delta_\alpha = 1$, we recover Eq. (3.2), with $\delta = \delta_\alpha/2$ in the proof of Theorem 1. Theorem 1 and thus consistency therefore continues to hold for $\kappa \in [0, \delta_\alpha/4)$ if one replaces Eq. (4.5) in Assumption FC.4 by

$$\lim_{h\downarrow 0} \sup_{\|\boldsymbol{\psi}\|\leq\eta} h^{1/2-\delta_\alpha/4-\kappa} \left\| \frac{\partial \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \right\| = 0, \qquad \lim_{h\downarrow 0} \sup_{\|\boldsymbol{\psi}\|\leq\eta} h^{1-\delta_\alpha/2-\kappa} \left\| \frac{\partial^2 \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top} \right\| = 0,$$

and (4.7) by

$$\lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\leq\eta} h^{-\delta_\alpha-2\kappa} \mathbb{E}\Big[V_{h,j}^2\big(\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|\big) \|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|^4 \;\Big|\; \boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\ \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z} \Big] = 0.$$

Theorems 2 and 3 continue to hold under the original sets of conditions if $\delta_\alpha = 1$ and $\delta_\omega, \delta_\beta > 1/2$ that remain unspecified. For $\delta_\alpha < 1$, the innovation of the filter is of a larger order of magnitude in h than the DGP's innovation. Such generalizations may seem interesting in the light of simulations shown in Jensen and Lange (2010). These authors study the effect of parameter estimation and conjecture that, depending on the setting, values of δ_α in the range $0.75 - 0.88$ may be relevant in the context of GARCH volatility models with estimated parameters. Parameter estimation is beyond the scope of our current paper and is left for future research. The discussion on general rates of h nevertheless suggests that the results of this paper may still apply in a slightly modified form, albeit with different rates of convergence for the filtering errors. In particular, in that case, the appropriate scaling of the filtering errors becomes $h^{-\delta_\alpha/4}\big(\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}_h^\star(t_{i,h})\big)$ to obtain a non-degenerate limiting result, and the fast time scale has to be altered to $T + \tau h^{1-\delta_\alpha/2}$. A larger impact of the forcing variable in the filtering equation (i.e., lower values of δ_α), thus results in a slower rate of convergence of the filtering error towards zero, and a faster rate of oscillation around zero, i.e., a speeding up of the fast time scale τ. Additionally, as the filter innovation dominates the DGP innovation asymptotically, this leads to an altered asymptotic variance in Theorem 2 with $\boldsymbol{B}(\boldsymbol{\psi})$ in Eq. (4.21) now set to zero. The expression for the optimal filter based on the minimum variance degenerates in this case with both the 'leverage correction' $\boldsymbol{P}_h\big(\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{\psi}\big)$ from Eq. (4.28) and the matrix $\boldsymbol{W}_h$ solving (4.30) collapsing to zero. This degeneracy is easily resolved by replacing the minimum variance criterion by a minimum mean-squared-error criterion, in which case the optimal forcing variable is again directly proportional to the true score function with respect to $\boldsymbol{\psi}$ for any finite M. Such a change of criterion was irrelevant before, as the asymptotic filter error variance up till now could not reach zero given the presence of the DGP innovations $\boldsymbol{\eta}_{i+1}$ in the limiting expressions.

5 Illustrative examples

In this section, we discuss four examples to illustrate different aspects of the theory. In Section 5.1 we illustrate how the variable $\boldsymbol{\upsilon}_h(t_{i,h})$ can play to include exogenous variables as (partial) drivers of time variation in parameters. We do so by extending the motivating example from Section 2.[5] In Section 5.2, we apply the theory to filtering a time-varying copula dependence parameter for the Clayton copula using a score-driven filter. Here, we also show how the theory can be applied to other existing filters in the literature, such as the original copula filter of Patton (2006). We introduce a slight modification of the original filter, which we prove to be consistent and asymptotically normal. Section 5.3 discusses the case of filtering a time-varying regression parameter and shows the consistency result no longer applies when filtering a time-varying drift rate. In that case we obtain mean-reversion of the filtering errors towards zero, but not consistency. Finally, in Section 5.4 we consider a fully mis-specified example where the model and DGP density differ and the true and filtered parameters capture different aspects of the distribution. Here we establish consistency to the Kullback-Leibler optimal value $\boldsymbol{\theta}_h^\star(t_{i,h})$.

5.1 Including an exogenous regressor in the filter

We start by modifying the filter in the motivating example from Section 2 with an exogenous variable $\upsilon_h(t_{i,h})$ as

$$\theta_h(t_{i+1,h}) = h\,\omega + h\,\gamma\,\upsilon_h(t_{i,h}) + (1 - h\,\beta)\theta_h(t_{i,h}) + h^{1/2}\,g_h\big(y_h(t_{i+1,h}), \theta_h(t_{i,h})\big), \tag{5.1}$$

with $g_h\big(y_h(t_{i+1,h}), \theta_h(t_{i,h})\big)$ as in Eq. (2.4), and $(\omega, \gamma, \beta) \in \mathbb{R}^3$. Examples of such models for time-varying tail shapes can be found in for instance D'Innocenzo et al. (2023). We assume $\upsilon_h(t_{i+1,h}) = (1 - c\,h)\upsilon_h(t_{i,h}) + h^{1/2}\,\zeta_{i+1}$ for some i.i.d. innovation sequence $\{\zeta_i\}$ that is independent of $\{\eta_i\}$. The set-up can be further relaxed to a general Markov structure for $\upsilon_h(t_{i,h})$, as we only need the triplet $\big(\upsilon_h(t_{i,h}), \psi_h(t_{i,h}), z_h(t_{i,h})\big)$ to be Markovian for our theory to apply.

The main change to the earlier set-up of the motivating example is that we now have to check Assumptions FC.2 and AD.2. Both of these follow easily if ζ_i has finite fourth order moments. The theoretical results of the paper thus readily extend to time-varying parameter filters that are partly driven by explanatory variables.

[5] Other interesting examples include, for instance, stochastic volatility models with leverage effects, such as Danielsson (1994), Yu (2005), or Li et al. (2019).

5.2 Bivariate copulas

As our second example, consider a filter for a time-varying copula parameter as introduced by Patton (2006), with score-driven applications introduced in Creal et al. (2013). We fully focus on the copula part and abstract from modeling the marginals to simplify the exposition. Consider a random vector $\boldsymbol{y}_h(t_{i+1,h}) = (y_{1h}(t_{i+1,h}), y_{2h}(t_{i+1,h}))$ that follows a Clayton copula distribution $C(\boldsymbol{y}_h(t_{i+1,h}); \tilde{\psi}_h(t_{i,h}))$ given by

$$C(\boldsymbol{y}; \tilde{\psi}) = \Big((y_1)^{-\tilde{\psi}} + (y_2)^{-\tilde{\psi}} - 1\Big)^{-1/\tilde{\psi}}, \qquad \tilde{\psi} := \exp(\psi) > 0, \quad \boldsymbol{y} = (y_1, y_2)^{\top}, \tag{5.2}$$

$$\psi_h(t_{i+1,h}) = \mu h + (1 - ah)\psi_h(t_{i,h}) + \sqrt{h} B \eta_{i+1}, \tag{5.3}$$

where η_{i+1} is i.i.d. standard normal. For convenience, define the short-hand notation $Q(\boldsymbol{y}; \theta) = (y_1)^{-\theta} + (y_2)^{-\theta} - 1$, where $\boldsymbol{y} = (y_1, y_2)$, such that the copula density is given by

$$c(\boldsymbol{y}; \theta) = (1 + \theta)\ (y_1 y_2)^{-(1+\theta)} Q(\boldsymbol{y}; \theta)^{-(2+1/\theta)}. \tag{5.4}$$

We assume the model density is correctly specified, such that p_h and q_h coincide. To derive the (optimal) score-driven filter, we note that

$$\begin{aligned} S(\boldsymbol{y}; \theta) &= \frac{\partial}{\partial \theta} \log c(\boldsymbol{y}; \theta) \\ &= (1 + \theta)^{-1} - \log(y_1 y_2) + \theta^{-2} \log Q(\boldsymbol{y}; \theta) + (\theta^{-1} + 2) \frac{(y_1)^{-\theta} \log y_1 + (y_2)^{-\theta} \log y_2}{Q(\boldsymbol{y}; \theta)}. \end{aligned} \tag{5.5}$$

Using this, we obtain the filtering equation

$$\Delta\theta_h(t_{i+1,h}) = h\omega + h\beta\theta_h(t_{i,h}) + \alpha\sqrt{h}\, S\big(\boldsymbol{y}_h(t_{i+1,h}); \theta_h(t_{i,h})\big). \tag{5.6}$$

Note that the filter will still be mis-specified for the true parameter dynamics in (5.3), as the filter is observation-driven, whereas the DGP is not. Since $\int_{[0,1]^2} (\partial \log c(\boldsymbol{y}; \theta)/\partial\theta)\ c(\boldsymbol{y}; \theta)\, \mathrm{d}\boldsymbol{y} = 0$, we have the global implicit function given by $\theta_h^{\star}(t_{i,h}) = \tilde{\psi}_h(t_{i,h}) = \iota_h(\psi_h(t_{i,h})) = \exp(\psi_h(t_{i,h}))$.

We note that most of the conditions in the assumptions from Section 4 are trivially satisfied in the current model set-up, noting $\boldsymbol{v}_h$ plays no particular role and can thus be set to $\boldsymbol{v}_h \equiv \mathbf{0}$. Assumption FC.5 requires $\alpha > 0$ and the information matrix $\mathcal{I}(\theta)$ of the Clayton copula with respect to θ to exist. For this we can use the explicit expression derived on p. 418 of Oakes (1982),

$$\mathcal{I}(\theta) = \frac{1}{(\theta+1)^2} + \frac{2}{\theta(\theta+1)(2\theta+1)} + \frac{4(\theta+1)}{3\theta+1} - \frac{2(2\theta+1)}{\theta}\rho(\theta), \tag{5.7}$$

where

$$\rho(\theta) = \frac{1}{(3\theta+1)(2\theta+1)}\left\{1+\frac{\theta+1}{2\theta}\left[\Psi^1\left(\frac{1}{2\theta}\right)-\Psi^1\left(\frac{\theta+1}{2\theta}\right)\right]+\frac{1}{2\theta}\left[\Psi^1\left(\frac{\theta+1}{2\theta}\right)-\Psi^1\left(\frac{2\theta+1}{2\theta}\right)\right]\right\},$$

and $\Psi^1(\,\cdot\,)$ denotes the trigamma function. As we have that both $A(\boldsymbol{x})$ and the conditional variance of $g_h\big(\boldsymbol{y}_h(t_{i+1,h}), \theta_h(t_{i,h})\big)$ from Assumption FC.5 are equal to the information matrix, the conditions are satisfied for $\theta > 0$.

To compute the optimal filter, we use the chain rule to obtain the true score function $S_0(\boldsymbol{y};\psi)$ as

$$S_0(\boldsymbol{y};\psi) = \frac{\partial}{\partial\psi}\log\left[c\big(\boldsymbol{y};\tilde{\psi}\big)\right] = \left(\frac{\partial}{\partial\tilde{\psi}}\log\left[c\big(\boldsymbol{y};\tilde{\psi}\big)\right]\right)\frac{\mathrm{d}\tilde{\psi}}{\mathrm{d}\psi} = S\big(\boldsymbol{y};\exp(\psi)\big)\,\exp(\psi),$$

where $S(\,\cdot\,)$ is defined in (5.5). By Corollary 1, we obtain $\mathcal{C}_h = \exp(2\psi)\,B^2$. Moreover, we note that

$$\mathcal{B}_h = \mathbb{E}\left\{\left[S\big(\boldsymbol{y}_h(t_{i+1,h});\exp(\psi)\big)\right]^2\,\middle|\,\psi_h(t_{i,h}) = \psi\right\},$$

which equals the Fisher information of the Clayton copula distribution with respect to $\tilde{\psi}_h(t_{i,h})$, for which we again use the result of Oakes (1982). Therefore,

$$W_h = \left(\frac{\mathcal{C}_h}{\mathcal{B}_h}\right)^{1/2} = \left[\frac{\exp(2\psi)B^2}{\mathcal{I}\big(\exp(\psi)\big)}\right]^{1/2}.$$

Using the reparametrization $\theta_h(t_{i,h}) = \exp(\psi_h(t_{i,h}))$, combining these results gives the optimal filter:

$$\Delta\theta_h^{\mathrm{opt}}(t_{i+1,h}) = h\omega + h\beta\theta_h^{\mathrm{opt}}(t_{i,h}) + \sqrt{h}\,\left[\theta_h^{\mathrm{opt}}(t_{i,h})\,|B|\right]\,\left[\mathcal{I}\big(\theta_h^{\mathrm{opt}}(t_{i,h})\big)\right]^{-1/2} S\big(\boldsymbol{y};\theta_h^{\mathrm{opt}}(t_{i,h})\big), \tag{5.8}$$

which is a score-driven model with square-root inverse information matrix scaling in the sense of Creal et al. (2013), where the coefficient $|B|$ is the standard deviation of the true, unknown innovations to $\psi_h(t_{i+1,h})$ times the Jacobian of the mapping from ψ to θ. Figure 3 confirms the consistency of this optimal filter and also shows that the optimal filter shows a less spiky behavior than the filter in (5.6) with ad-hoc chosen parameters ω, α, and β.

It is also interesting to see how our results can be applied to existing filters for copula parameters. We consider a slightly revised version of the filter from Patton (2006),

$$\xi_h(t_{i+1,h}) = \omega + h\,\beta\,\xi_h(t_{i,h}) + \sqrt{h}\,\alpha\,\big|y_{1h}(t_{i+1,h}) - y_{2h}(t_{i+1,h})\big|, \tag{5.9}$$

where $\xi_h(t_{i,h})$ is the lower-tail-dependence parameter. The original filter has an additional logistic mapping around the right-hand side of (5.9) to ensure that the tail-dependence probability $\xi_h(t_{i,h})$ is

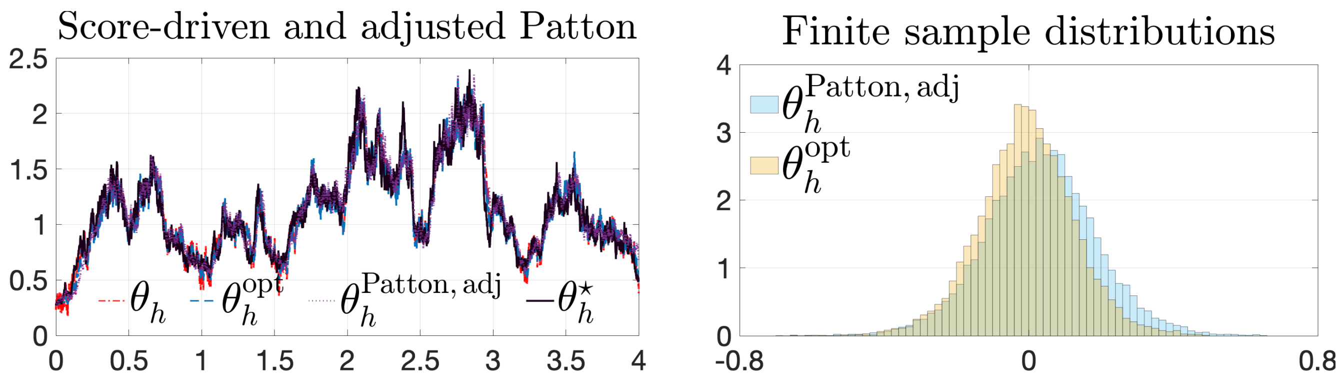

Figure 3: Filter consistency for the Clayton copula

The DGP is a Clayton copula density with true parameter $\exp(\psi_h(t_{i,h}))$ as given in (5.3) and the parameters given by $(\mu, a, B) = (0.5, 5, 1)$. The sampling frequency is $h = 1/19656$. The filter equation is given in (5.6), with the parameters $(\omega, \beta, \alpha) = (0.5, -0.1, 2)$. We take $(\omega, \beta, \alpha) = (0.5, -0.1, -1.4)$ for the adjusted Patton filter.

always between zero and one. The filter as stated in Eq. (5.9) is not well-balanced for an asymptotic analysis, because the forcing variable $\alpha\left|y_{1h}(t_{i+1,h}) - y_{2h}(t_{i+1,h})\right|$ does not have a zero conditional mean. The effect of this is that the filter diverges as $h \downarrow 0$. This defect is easily corrected by subtracting the conditional mean of the forcing variable under the Clayton copula density from the last term in (5.9), leading to the slightly adjusted Patton filter

$$\xi_h(t_{i+1,h}) = \omega + h\,\beta\,\xi_h(t_{i,h}) + \sqrt{h}\,\alpha\Big(\big|y_{1h}(t_{i+1,h}) - y_{2h}(t_{i+1,h})\big| - m(\xi_h(t_{i,h}))\Big), \tag{5.10}$$

$$m(\xi_h(t_{i,h})) = 2^{-1/\theta_h(t_{i,h})}\ {}_2F_1\left(1 + \theta_h(t_{i,h})^{-1},\ 2\theta_h(t_{i,h})^{-1},\ 1 + 2\theta_h(t_{i,h})^{-1},\ 1/2\right) - 1, \tag{5.11}$$

where ${}_2F_1(\,\cdot\,)$ is a confluent hypergeometric function, and $\xi_h(t_{i,h}) = 2^{-1/\theta_h(t_{i,h})}$, i.e., the upper-tail-dependence parameter for the Clayton copula. We refer to Appendix E for a detailed derivation.

Implementing the de-meaned copula filter from Eq. (5.10), we plot $\theta_h^{\text{Patton,adj}} = -1/\log_2(\xi_h(t_{i,h}))$ in Figure 3. The figure clearly shows that this adjusted version of the original filter of Patton (2006) is consistent. In the right-hand panel of the figure we moreover show the simulated distribution of the filtering errors. We see that the optimal filter's mean is closer to zero and its variance is somewhat smaller than that of the adjusted Patton filter, in line with the theory presented. Finally, looking at the histograms of the demeaned and standardized simulations (not shown in figure), we also obtain that both filtering errors are well approximated by the normal distribution with the asymptotic mean and variance as derived in this paper for $h \downarrow 0$.

5.3 Time-varying regression coefficients

Let f be an arbitrary pdf and $\mathcal{G} := \{g_{\mu,\sigma},\ \mu \in \mathbb{R},\ \sigma > 0\}$ be a location-scale family. Assume that for any h the pdf of $\boldsymbol{y}_h(t_{i+1,h}) = \big(y_{1h}(t_{i+1,h}), y_{2h}(t_{i+1,h})\big)$ given $\psi_h(t_{i,h})$ equals

$$q_h\big(\boldsymbol{y}_h(t_{i+1,h}); \psi_h(t_{i,h})\big) = f(y_{1h}(t_{i+1,h}))\, \frac{g_{0,1}\left(\frac{y_{2h}(t_{i+1,h}) - \psi_h(t_{i,h})\, y_{1h}(t_{i+1,h})}{\sigma_\epsilon \sqrt{h}}\right)}{\sigma_\epsilon \sqrt{h}}, \tag{5.12}$$

i.e., we have the following regression model with a time-varying coefficient $\psi_h(t_{i,h})$ and random regressor $y_{1h}(t_{i,h})$:

$$y_{2h}(t_{i+1,h}) = \psi_h(t_{i,h}) y_{1h}(t_{i+1,h}) + \sqrt{h}\sigma_\epsilon \epsilon_{i+1}, \tag{5.13}$$

where ϵ_{i+1} has pdf $g_{0,1} \in \mathcal{G}$. One can interpret $y_{2h}(t_{i+1,h})$ as an excess stock return, while $y_{1h}(t_{i+1,h})$ is the excess return on the market as in the CAPM model with time-varying betas; see for instance Umlandt (2023) for score-driven asset pricing models.

The statistician works with the mis-specified joint pdf

$$p_h\big(\boldsymbol{y}_h(t_{i+1,h}); \theta_h(t_{i,h})\big) = \tilde{f}(y_{1h}(t_{i+1,h})) \frac{1}{\sqrt{2\pi h}} \exp\left(- \frac{[y_{2h}(t_{i+1,h}) - \theta_h(t_{i,h}) y_{1h}(t_{i+1,h})]^2}{2h} \right), \tag{5.14}$$

where $\tilde{f}$ is also an arbitrary pdf. It is clear that the conditional pdf of $y_{2h}(t_{i+1,h}) \,\big|\, \big(y_{1h}(t_{i+1,h}), \theta_h(t_{i,h})\big)$ is specified by the statistician as $\mathcal{N}\big(\theta_h(t_{i,h}) y_{1h}(t_{i+1,h}), h\big)$. The integral in (3.5) here becomes a double integral

$$\begin{aligned}
&\iint y_1 \frac{y_2 - \theta y_1}{h} f(y_1)\, \frac{g_{0,1}\left(\frac{y_2 - \psi\, y_1}{\sigma_\epsilon \sqrt{h}}\right)}{\sigma_\epsilon \sqrt{h}}\, \mathrm{d}y_2\, \mathrm{d}y_1 \\
&\quad = \int y_1 f(y_1) \int \frac{y_2}{h}\, \frac{g_{0,1}\left(\frac{y_2 - \psi\, y_1}{\sigma_\epsilon \sqrt{h}}\right)}{\sigma_\epsilon \sqrt{h}}\, \mathrm{d}y_2\, \mathrm{d}y_1 - \theta \int \frac{y_1^2}{h}\, f(y_1) \int \frac{g_{0,1}\left(\frac{y_2 - \psi\, y_1}{\sigma_\epsilon \sqrt{h}}\right)}{\sigma_\epsilon \sqrt{h}}\, \mathrm{d}y_2\, \mathrm{d}y_1 \\
&\quad = \int y_1 f(y_1)\ \ \psi y_1\, \mathrm{d}y_1 - \theta \int y_1^2 f(y_1)\, \mathrm{d}y_1 = 0.
\end{aligned}$$

Hence, we can readily derive the global implicit function ι_h as $\theta_h^\star(t_{i,h}) = \iota_h\big(\psi_h(t_{i,h})\big) = \psi_h(t_{i,h})$. Clearly, such a relation holds true for any f, $\tilde{f}$, and $g_{0,1}$ with a finite second and first moment, respectively. Therefore, if the score-driven filter is consistent, it remains consistent across a wide range of mis-specified scenarios, reminiscent of the classical consistency results of quasi-maximum likelihood estimation; see White (1982).

Based on (5.14), the score-driven filtering equation used by the statistician is given by

$$\Delta\theta_h(t_{i+1,h}) = h\omega + h\beta\theta_h(t_{i,h}) + \alpha h^{-\zeta}\Big(y_{1h}(t_{i+1,h})\big[y_{2h}(t_{i+1,h}) - \theta_h(t_{i,h})y_{1h}(t_{i+1,h})\big]\Big), \quad \alpha, \zeta > 0, \tag{5.15}$$

where we use a flexible rate $h^{\zeta+1/2}$ for the forcing variable. Without loss of generality, assume that $y_{1h}(i)$ is i.i.d. with a mean $\mu_1 h$ and a variance $\sigma_1^2 h$. Furthermore, let ϵ_i be i.i.d. and follow a standard Student's t distribution with $\nu > 0$ degrees of freedom, denoted as t_ν. The process $\{\boldsymbol{v}_h(t_{i,h})\}$ in Eq. (4.1) plays no role in the current example and is set to $\mathbf{0}$ for all $i \in \mathbb{Z}^+$.

For consistency, only Assumption FC.5 is non-trivial to check. The remaining assumptions are easily verified. For any $\nu > 2$ and using the filter in Eq. (5.15), the expressions for $A(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x})$ as defined in (4.11) and (4.21) become

$$A(\boldsymbol{x}) = \alpha \lim_{h\downarrow 0}\Big\{h^{-(\zeta+1/2)}\big(\mu_1^2h^2 + \sigma_1^2 h\big)\Big\}, \qquad \Sigma(\boldsymbol{x}) = \alpha^2\sigma_\epsilon^2 \lim_{h\downarrow 0}\Big\{h^{-2\zeta}\big(\mu_1^2h^2 + \sigma_1^2 h\big)\Big\}\frac{\nu}{\nu-2} + B^2.$$

It is clear from the above equations that the asymptotic behavior of the filter is determined by the values of σ_1 and ζ. We distinguish two cases.

(a) For $\sigma_1 > 0$, we take $\zeta = 1/2$. Then $A(\boldsymbol{x}) = \alpha\sigma_1^2 > 0$ and $\Sigma(\boldsymbol{x}) = \alpha^2\sigma_1^2\sigma_\epsilon^2\nu/(\nu-2) + B^2$. The (mis-specified) filter in (5.15) is then consistent with asymptotic variance

$$V(\tau, \boldsymbol{x}) = \Big[1 - \exp\big(-2\alpha\sigma_1^2\tau\big)\Big]\frac{1}{2\alpha\sigma_1^2}\bigg(\alpha^2\sigma_1^2\sigma_\epsilon^2\frac{\nu}{\nu-2} + B^2\bigg),$$

which for large τ can be approximately optimized by taking $\alpha = (\sigma_1\sigma_\epsilon)^{-1}|B|\sqrt{(\nu-2)/\nu}$.

(b) For $\sigma_1 = 0$ and $\mu_1 \neq 0$, i.e., in the case of filtering a time-varying intercept, we do not have consistency. More specifically, consider the case $\mu_1 \neq 0$. If we choose $\zeta = 3/2$, we have that $A(\boldsymbol{x}) = \alpha\mu_1^2 > 0$, but $\Sigma(\boldsymbol{x}) = \alpha^2\sigma_\epsilon^2\mu_1^2\nu/(\nu-2)\big\{\lim_{h\downarrow 0} h^{-1}\big\} + B^2 = +\infty$. This implies that the asymptotic variance diverges to infinity under this scaling. If we choose $\zeta = 1$, then $\Sigma(\boldsymbol{x}) < \infty$, but $A(\boldsymbol{x}) = \lim_{h\downarrow 0} O\big(h^{1/2}\big)$, such that Assumption FC.5 fails. In fact, the filter will not be consistent, but instead converge to an Ornstein-Uhlenbeck (OU) process. As a result, the filtering errors become mean-reverting around zero, but do not converge to zero for every time t. This is important for models that filter a time-varying location, such as Harvey and Luati (2014).

The consistency results are visualized in Figure 4. We clearly see that for the non-degenerate regressor case (left-hand plot) the score-driven filter is consistent, whereas for the time-varying intercept case (right-hand plot) it is not.

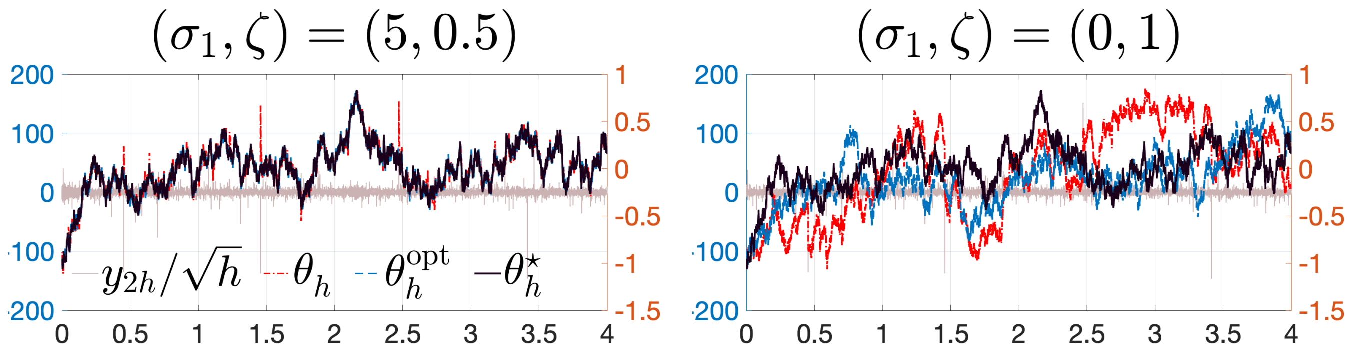

Figure 4: Filter (in)consistency for time-varying regression models

Consider $\sigma_\epsilon = 1$ in Eq. (5.12) and let $\psi_h(t_{i+1,h})$ be generated by $\psi_h(t_{i+1,h}) = (1 - ah)\psi_h(t_{i,h}) + \sqrt{h}B\eta_{i+1}$ with $(a, B) = (10, 1)$. Moreover, ϵ_i is i.i.d. t_ν with $\nu = 2.5$. Fix the level parameter μ_1 of $y_{1h}(t_{i+1,h})$ at $\mu_1 = 5$. The left-hand side gives the path of $\theta_h(t_{i,h})$ (red for the Gaussian, and blue for the optimal filter), where $(\omega, \beta, \alpha) = (0.1, -1, 0.1)$, and $\theta^\star_h(t_{i,h})$ (black) for the case of a non-degenerate regressor $y_{1h}(t_{i+1,h})$. Here the filter is consistent. The right-hand plot gives the same curves for the case of a degenerate regressor (intercept), where the filtering errors are mean-reverting to zero, but the filter is not consistent. Here σ_1 denotes the variance parameter of the regressor $y_{1h}(t_{i+1,h})$, and $h^{-\zeta}$ denotes the scaling of the forcing variable in the filter (5.15). Data of the dependent variable $y_{2h}(t_{i+1,h})$ are in gray bars and use the left-hand y-axis. The filters use the right-hand y-axis.

Now we derive the optimal filter. Note that the score is

$$S_h\big(\boldsymbol{y}_h(t_{i+1,h}), \psi_h(t_{i,h})\big) = \frac{(\nu+1)y_{1h}(t_{i+1,h})\big[y_{2h}(t_{i+1,h}) - \psi_h(t_{i,h})y_{1h}(t_{i+1,h})\big]}{h\nu\sigma_\epsilon^2 + \big[y_{2h}(t_{i+1,h}) - \psi_h(t_{i,h})y_{1h}(t_{i+1,h})\big]^2}.$$

By Corollary 1, we have $\mathcal{B}_h = (\nu+1)(\mu_1^2 h + \sigma_1^2)/\left[(\nu+3)\sigma_\epsilon^2\right]$ and $\mathcal{C}_h = B^2$, leading to

$$W_h = \left(\frac{(\nu+3)\sigma_\epsilon^2}{(\nu+1)(\mu_1^2 h + \sigma_1^2)}\right)^{1/2} |B|.$$

The filtering equation for the optimal filter as stated in Theorem 3 is then given by

$$\begin{aligned}\Delta\theta_h^{\text{opt}}(t_{i+1,h}) = h\omega + h\beta\theta_h^{\text{opt}}(t_{i,h}) + h^{1/2}&\left(\frac{(\nu+3)\sigma_\epsilon^2}{(\nu+1)(\mu_1^2 h + \sigma_1^2)}\right)^{1/2} \\ &\times \frac{(\nu+1)y_{1h}(t_{i+1,h})\big[y_{2h}(t_{i+1,h}) - \psi_h(t_{i,h})y_{1h}(t_{i+1,h})\big]}{h\nu\sigma_\epsilon^2 + \big[y_{2h}(t_{i+1,h}) - \psi_h(t_{i,h})y_{1h}(t_{i+1,h})\big]^2}|B|. \quad (5.16)\end{aligned}$$

The simulated (finite-sample) filtering errors and there de-meaned and scaled counterparts using the asymptotic approximations are shown in Figure 5 for $T = 0.5$ and two different frequencies. In the upper panels, we show the raw filtering errors for different frequencies. Clearly, for lower frequencies the filtering errors are obviously larger. We also see in the top-right graph that the filtering errors based on the optimal filter are more concentrated and have less fat tails. This is also seen from the red peaks in the filter in Figure 4 and may motivate the use of robust filters for time-varying regression parameters as in Umlandt (2023). To assess the adequacy of the asymptotic distribution, the lower panels de-mean and scale the filtering errors by the appropriate expressions for the mean and variance

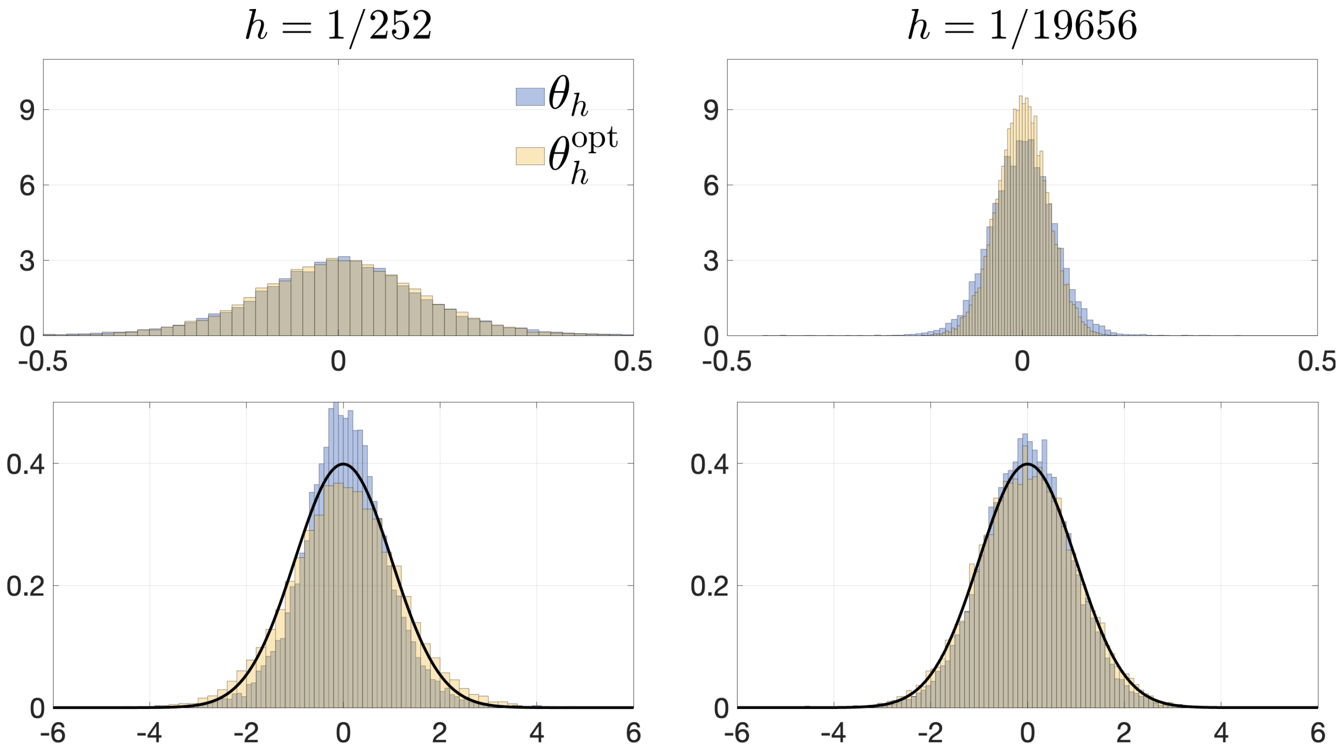

Figure 5: Distributions for filtering errors in time-varying regression models
The top panels give the distribution of the filtering errors for the Gaussian filter in Eq. (5.15) in blue and the optimal filter in yellow. The bottom panels prove the same results, but are now demeaned and scaled using the expressions for the mean and variance of the asymptotic distribution to assess the adequacy of the asymptotic normal approximation, where $(T, M) = (1/2, 4)$. We refer the reader to Figure 4 for the specification of parameters.

of the asymptotic normal distribution. The optimal filter satisfies this asymptotic approximation better for both frequencies considered in the figure.

5.4 Fully mis-specified models

Let the DGP's conditional density be of the generalized Pareto type with a time-varying tail shape,

$$q_h(y_h(t_{i+1,h}); \psi_h(t_{i,h})) = h^{-1}\, \tilde{\psi}_h(t_{i,h}) \cdot \left[1 + h^{-1} y_h(t_{i+1,h})\right]^{-\tilde{\psi}_h(t_{i,h})-1}, \tag{5.17}$$

where $\tilde{\psi}_h(t_{i,h}) = \nu + \exp(\psi_h(t_{i,h}))$ is a scalar time-varying shape parameter with fixed lower bound ν. The conditional expectation of $y_h(t_{i+1,h})$ is given by $h/(\tilde{\psi}_h(t_{i,h}) - 1)$. The dynamics of $\psi_h(t_{i,h})$ are assumed to be given by

$$\Delta\psi_h(t_{i+1,h}) = h\big(a_0 - a_1\, \psi_h(t_{i,h})\big) + h^{1/2} a_2\, \eta_{i+1}, \qquad a_0, a_1, a_2 \geq 0, \tag{5.18}$$

for a scalar i.i.d. standard normal process η_{i+1} that is independent of $y_h(t_{i+1,h})$ conditional on $\psi_h(t_{i,h})$. As the tail shape is parameterized by $\tilde{\psi}_h(t_{i,h}) = \nu + \exp(\psi_h(t_{i,h}))$, we can easily control the number of finite conditional moments of $y_h(t_{i+1,h})$, with ν being a strict upper bound.

As a filter, we employ the Multiplicative Error Model (MEM) of Engle and Gallo (2006). The

MEM model is score-driven for an exponential distribution with a time-varying scale, see (Creal et al., 2013). To ensure the positivity of the scale, we take the log-scale of the MEM as our time-varying parameter that we filter from the data. This gives the conditional model density

$$p_h\big(y_h(t_{i+1,h});\theta_h(t_{i,h})\big) = h^{-1}\exp\Big(-\theta_h(t_{i,h}) - h^{-1}\,e^{-\theta_h(t_{i,h})}\,y_h(t_{i+1,h})\Big), \tag{5.19}$$

with filter dynamics

$$\Delta\theta_h(t_{i+1,h}) = h\big(b_0 - b_1\theta_h(t_{i,h})\big) + h^{1/2}b_2\Big(h^{-1}e^{-\theta_h(t_{i,h})}\,y_h(t_{i+1,h}) - 1\Big). \tag{5.20}$$

To compute the pseudo true parameter, note that Eq. (3.5) collapses to

$$\begin{aligned} &f_h\big(\theta_h^\star(t_{i,h}),\psi_h(t_{i,h})\big) = \int \Big(h^{-1}\,e^{-\theta_h^\star(t_{i,h})}\,y - 1\Big)\,q_h(y;\psi_h(t_{i,h}))\,\mathrm{d}y = 0 \qquad \Longleftrightarrow \\ &\theta_h^\star(t_{i,h}) = \log\int \frac{y}{h}\,q_h(y;\psi_h(t_{i,h}))\,\mathrm{d}y = -\log\big(\tilde{\psi}_h(t_{i,h}) - 1\big) = -\log\big(\nu - 1 + \exp(\psi_h(t_{i,h}))\big), \end{aligned} \tag{5.21}$$

for $\nu + \exp(\psi_h(t_{i,h})) > 1$. The pseudo-true parameter $\theta_h^\star(t_{i,h})$ is thus obviously and in a predictable way mis-specified for the true time-varying parameter $\psi_h(t_{i,h})$. The mis-specification, in this case, is quite severe: the conditional model density (exponential) and true conditional density (generalized Pareto) are quite different. Moreover, the model parameter $\theta_h^\star(t_{i,h})$ even has an entirely different interpretation than $\psi_h(t_{i,h})$, with $\theta_h^\star(t_{i,h})$ measuring scale variation, and $\psi_h(t_{i,h})$ measuring tail shape variation.

Assumptions FC.1 to FC.4 are easily checked by setting $\boldsymbol{v}_h(\,\cdot\,) \equiv \mathbf{0}$, and using the Gaussianity of η_{i+1} and the fact that $\theta_h^\star(t_{i,h}) = -\log\big(\nu - 1 + \exp(\psi_h(t_{i,h}))\big)$ is smooth for $\nu > 1$ and sub-linear: its second derivative with respect to ψ equals $-(\nu-1)e^{-\psi}/(1+(\nu-1)e^{-\psi})^2$, which for $\nu > 1$ has a maximum absolute value of $1/4$. To check the final Assumption FC.5, we use the (score-driven) filter dynamics already stated in Eq. (5.20), which yield

$$\begin{aligned} &\mathbb{E}\Big[g_h\big(y_h(t_{i+1,h}),\theta_h^\star(t_{i,h})\big)\,\big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\; z_h(t_{i,h}) = z\Big] \\ &\qquad\qquad = e^{-\theta_h^\star(t_{i,h})}\,\mathbb{E}\Big[\frac{y_h(t_{i+1,h})}{h}\,\big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\; z_h(t_{i,h}) = z\Big] - 1 = 0, \end{aligned}$$

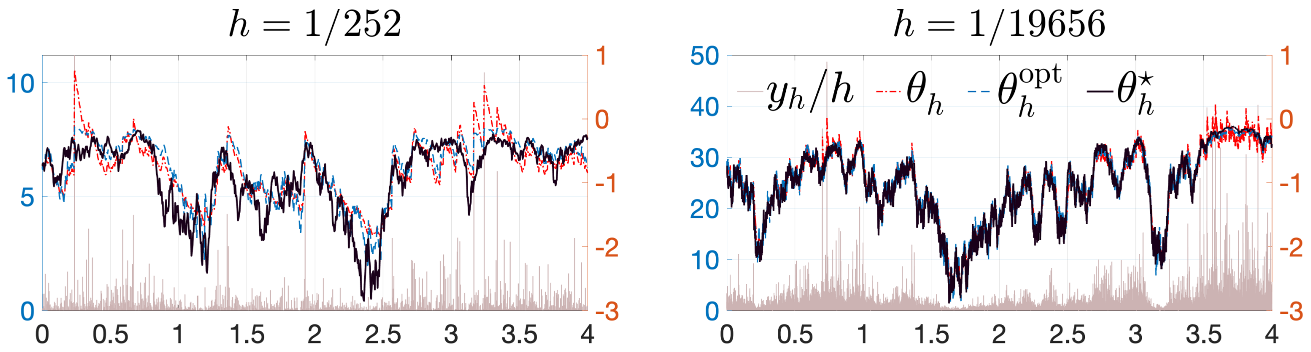

Figure 6: Consistency illustration for the MEM example

The figure contains the path of a simulated tail-shape model using Eqs. (5.17) and (5.18) for $T = 4$ and h equal to either 1/252 (daily data) or 1/19656 (5 minute data). The parameters in the DGP are chosen as $a_0 = 0$, $a_1 = a_2 = 3$, and $\nu = 2.1$. The resulting pseudo-true parameter $\theta_h^\star(t_{i,h})$ from (5.21) is given in black and uses the right-hand y-axis. The underlying data $y_h(t_{i,h})$ are drawn as gray bars at the bottom and use the left-hand y-axis. The filtered value $\theta_h(t_{i,h})$ using the MEM type filter (drawn in red) is given by Eq. (5.20), with $b_0 = \log(\bar{y}/h)$, $b_1 = 3$, $b_2 = 1$, and $\bar{y}$ the sample mean of $y_h(t_{i,h})$. Drawn in blue, we also provide the filtered $\theta_h(t_{i,h})$ using the optimal filter from Section 4.3.

using Eq. (5.21). Also Eq. (4.9) is satisfied for $\nu > 1$, as

$$\lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},z)\|\leq\eta} h^\kappa\, |b_2|\, \mathbb{E}\left[\sup_{\|\theta\|\leq\tilde{\eta}} \left\| \frac{e^{-\theta} y_h(t_{i+1,h})}{h} \right\| \,\middle|\, \boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, z_h(t_{i,h}) = z \right]$$
$$\leq \lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},z)\|\leq\eta} h^\kappa |b_2| \frac{\exp(\tilde{\eta})}{\nu - 1 + \exp(\theta_h^\star(t_{i,h}))} = 0. \quad (5.22)$$

Eq. (4.10) is easily satisfied if we assume $\nu > 2$, because

$$\limsup_{h\downarrow 0} \sup_{\|(\boldsymbol{x},z)\|\leq\eta} \mathbb{E}\left[h^{-2} e^{-2\theta_h(t_{i,h})} y_h(t_{i+1,h})^2 \,\middle|\, \boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x},\, z_h(t_{i,h}) = z \right]$$
$$\leq \limsup_{h\downarrow 0} \sup_{\|(\boldsymbol{x},z)\|\leq\eta} \frac{2e^{-2\theta_h(t_{i,h})}}{(\tilde{\psi}_h(t_{i,h}) - 1)(\tilde{\psi}_h(t_{i,h}) - 2)} \leq \frac{2e^{2\eta}}{(\nu-1)(\nu-2)}. \quad (5.23)$$

Finally, we obtain that $A(\boldsymbol{x}) = b_2 \exp\left(\theta_h^\star(t_{i,h}) - \theta_h^\star(t_{i,h})\right) = b_2$ in (4.11), which is clearly positive if $b_2 > 0$. Therefore, Assumption FC.5 is satisfied.

As a result, for $\nu > 2$, Theorem 1 applies and we obtain consistency of the score-driven filter to the pseudo-true parameter. An example is provided in Figure 6. We show two score-driven filtering results: the MEM filter with ad-hoc parameters, and an optimal filter as derived later on. Both filters satisfy the conditions for consistency, and it is clearly seen that the filtered $\theta_h(t_{i,h})$ in both cases lies close to the pseudo-true $\theta_h^\star(t_{i,h})$, despite severe mis-specification. The fit becomes tighter as $h \downarrow 0$.

For the asymptotic normality of the filtering errors, only Assumption AD.1(c) is less trivial. Note that $g_h\big(y_h(t_{i+1,h}), \theta_h(t_{i,h})\big) = b_2[h^{-1}e^{-\theta_h(t_{i,h})} y_h(t_{i+1,h}) - 1]$, where $y_h(t_{i+1,h})$ has a generalized Pareto conditional distribution as in (5.17). For $\nu > 2$, the first equation in Assumption AD.1(c) is then directly satisfied by a similar argument as in Eq. (5.23). We also obtain

$$\limsup_{h\downarrow 0}\ \sup_{\|(\boldsymbol{x},z)\|\leq\eta}\ e^{-\theta_h(t)}\mathbb{E}\Bigg[\ \Big|h^{-1}y_h(t+h)\Big|^{2+\varphi_3}\ \Bigg|\ \boldsymbol{x}_h(t)=\boldsymbol{x},\ \tilde{z}_h(t)=z\Bigg]\leq C_{1,\eta},$$

as long as $\nu>2$ by choosing $\varphi_3=(\nu-2)/2$. Finally, $A(\boldsymbol{x})=b_2$, establishing that Assumption AD.1(c) holds if $b_2>0$. For Eq. (4.20) we have

$$\lim_{h\downarrow 0}\ \sup_{\|(\boldsymbol{x},z)\|\leq\eta}\ h^{1/2}\mathbb{E}\left[\frac{e^{-2\theta_h^\star(t)}\ y_h(t+h)^2}{h^2}\ \Bigg|\ \boldsymbol{x}_h(t)=\boldsymbol{x},\ \tilde{z}_h(t)=z\right]=\lim_{h\downarrow 0}\ \sup_{\|(\boldsymbol{x},z)\|\leq\eta}\ 2\ h^{1/2}=0.$$

As a result, Theorem 2 applies and the scaled filtering errors are asymptotically normal. As $\exp(-\theta_h^\star(t_{i,h}))y_h(t_{i+1,h})/h$ has a conditional unit exponential distribution, the conditional variance of $g_h\big(y_h(t+h),\theta^\star(t)\big)=h^{-1}\exp\big(-\theta_h^\star(t_{\lfloor t/h\rfloor,h})\big)y_h(t_{\lfloor t/h\rfloor+1,h})-1$ is equal to 1, and thus

$$\Sigma(\boldsymbol{x})=b_2^2\,\frac{\nu+e^\psi}{\nu-2+e^\psi}+a_2^2\frac{e^{2\psi}}{\left(\nu-1+e^\psi\right)^2},\qquad V^*=\int_0^\infty\exp(-2b_2\,s)\Sigma(\boldsymbol{x})\,\mathrm{d}s=(2b_2)^{-1}\Sigma(\boldsymbol{x}).\ (5.24)$$

The asymptotic variance of the filtering error thus clearly increases if the true underlying time-varying parameter $\psi_h(t_{i,h})$ has a larger variance, i.e., if a_2 is larger. Also, the asymptotic variance is non-monotonic in b_2. A larger b_2 means a larger reaction to the score of the model density. This allows the filter to react more timely to changes in the underlying time-varying $\psi_h(t_{i,h})$ and thus reduces V^* via the factor $(2b_2)^{-1}$ in (5.24). At the same time, a larger b_2 may cause over-shooting and thus a larger asymptotic variance, which is the reason why b_2^2 enters $\Sigma(\boldsymbol{x})$. The two effects result in a bias-variance trade-off that can be used to select an optimal value of b_2.

To compute the optimal filter in this example, we first invert the mapping from ψ to $\theta^\star$ and determine its derivatives. We obtain

$$\begin{aligned}\iota_h(\psi)&=-\log(\nu-1+e^\psi) &\iff\qquad \iota_h^{-1}(\theta^\star)&=\psi=\log(1+e^{-\theta^\star}-\nu),\\ \frac{\partial\iota_h(\psi)}{\partial\psi}&=\frac{-e^\psi}{\nu-1+e^\psi}=\frac{-\left(1+e^{-\theta^\star}-\nu\right)}{e^{-\theta^\star}}, & \frac{\partial\iota_h^{-1}(\theta^\star)}{\partial\theta^\star}&=\frac{-e^{-\theta^\star}}{1+e^{-\theta^\star}-\nu}=\frac{-\left(\nu-1+e^\psi\right)}{e^\psi}.\end{aligned}$$

Using the generalized Pareto distribution (5.17), we have

$$\begin{aligned}S_h(y_h(t_{i+1,h}),\psi_h(t_{i,h}))&=\left[\frac{1}{\tilde{\psi}_h(t_{i,h})}-\log\left(1+\frac{y_h(t_{i+1,h})}{h}\right)\right](\tilde{\psi}_h(t_{i,h})-\nu),\\ \mathbb{E}\Big[S_h(y_h(t_{i+1,h}),\psi)^2\,\Big|\,\psi_h(t_{i,h})=\psi\Big]&=\mathbb{E}\Bigg[-\frac{\partial}{\partial\psi}S_h(y_h(t_{i+1,h}),\psi)\,\Bigg|\,\psi_h(t_{i,h})=\psi\Bigg]=\frac{e^{2\psi_h(t_{i,h})}}{\left(\nu+e^{\psi_h(t_{i,h})}\right)^2},\end{aligned}$$

with $\tilde{\psi}_h(t_{i,h})=\nu+e^{\psi_h(t_{i,h})}$. Substituting these results into the expressions for $\mathcal{B}_h$ and W_h using

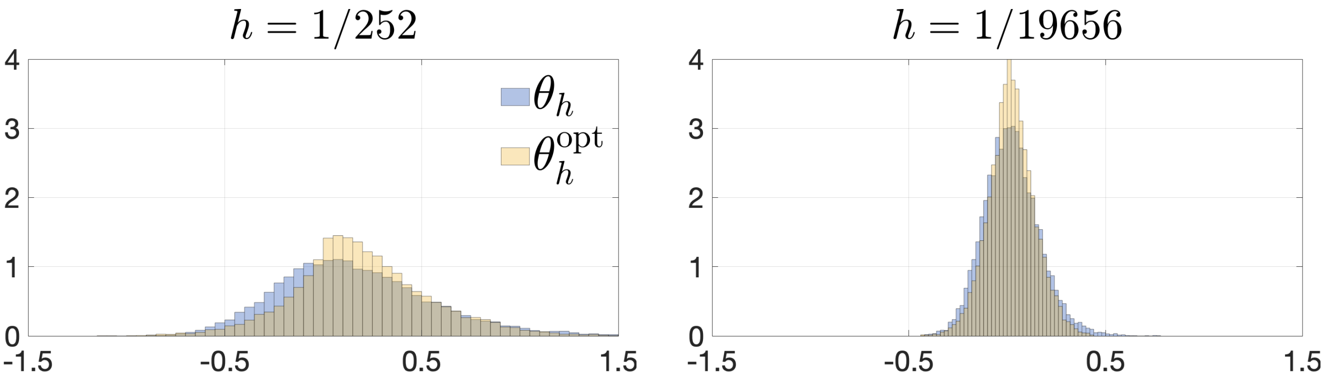

Figure 7: MEM example: finite sample distributions of filtering errors
The figure contains the simulated distributions of filtering errors for the dynamic tail-shape model from Eqs. (5.17) and (5.18) for $T = 1/2$ and $M = 4$ and h equal to either 1/252 (daily data) or 1/19656 (5 minute data). The parameters in the DGP are chosen as $a_0 = 0$, $a_1 = a_2 = 3$, and $\nu = 2.1$. In blue, we draw the filtering errors based on the MEM type filter given by Eq. (5.20) with $b_0 = \log(\bar{y}/h)$, $b_1 = 3$, $b_2 = 1$, and $\bar{y}$ the sample mean of $y_h(t_{i,h})$. In yellow, we provide the distribution of filtering errors using the optimal filter.

Corollary 1, we obtain, conditional on $\psi_h(t_{i,h}) = \psi$,

$$\begin{aligned}
\mathcal{B}_h &= \Big[Q_h\big(\theta_h^\star(t_{i,h})\big)\Big]^2 \frac{e^{2\psi_h(t_{i,h})}}{\big(\nu + e^{\psi_h(t_{i,h})}\big)^2} = \frac{\big(\nu - 1 + e^{\psi}\big)^2}{e^{2\psi}} \frac{e^{2\psi}}{\big(\nu + e^{\psi}\big)^2}, \\
\mathcal{C}_h &= \frac{e^{2\psi}}{\big(\nu - 1 + e^{\psi}\big)^2}\, a_2^2, \\
W_h &= \sqrt{\frac{\mathcal{C}_h}{\mathcal{B}_h}} = \frac{e^{\psi}\ (\nu + e^{\psi})}{\big(\nu - 1 + e^{\psi}\big)^2}\, |a_2| = \big[1 - (\nu - 1)e^{\theta^\star}\big](1 + e^{\theta^\star})\, |a_2|.
\end{aligned}$$

Using the above derivations, we now simulate the distribution of filtering errors using $T = 1/2$ and $M = 4$. The results can be found in Figure 7. The figure shows that the distribution of filtering errors is indeed tighter for the optimal filter than for the MEM type filter, in line with Theorem 3. This holds for both the low and high frequency considered. For low frequencies, the optimal filter shows a slight bias, though its tails are substantially smaller than those of the suboptimal filter. For high frequencies, the bias disappears, and the gain in smaller variance remains and dominates the result.

The intuition for the smaller variance in this case stems from the more robust nature of the score-driven filter for the DGP density. Rather than being linear in $y_h(t_{i+1,h})$, the optimal filter is based on $S_h\big(y_h(t_{i+1,h}), \psi_h(t_{i,h})\big)$ and thus logarithmic in $y_h(t_{i+1,h})$. As such, it reacts much less fiercely on incidental large observations. Such observations may be prevalent if the tail index $\tilde{\psi}_h(t_{i,h}) = \nu + e^{\psi_h(t_{i,h})}$ in the data generating process reaches its lower bound ν, see also the simulated data in Figure 6. As a result, the MEM type filter may behave much more erratically in these settings than the optimal filter, resulting in the more spiky behavior of the filtered path in Figure 6, and the larger tails for the MEM filtering error distribution in Figure 7.

6 Conclusions

In this paper, we studied the theoretical properties of score-driven models in a generic setting using the tools from an in-fill asymptotic experiment. We have established that score-driven filters are consistent for the time-varying parameter if the model's conditional predictive density is correctly specified, or more generally, if the Kullback-Leibler optimal parameter coincides with the true time-varying parameter. If the model is mis-specified, the score-driven filter is still consistent for the Kullback-Leibler optimal parameter path, i.e., for the path of a parameter that at every time point minimizes the Kullback-Leibler divergence between the model density and the unknown DGP. Such a result considerably generalizes earlier results from the literature for volatility models to the more generic non-linear time series context, while at the same time allowing for more general forms of mis-specification. It also generalizes earlier discrete time results as in Blasques et al. (2015) to the continuous time context.

We further derived the asymptotic distribution of filtering errors in this general setting. The asymptotic result was non-standard in the sense that it required both a scaling of filtering errors as well as a 'stretching' of the time axis to obtain a non-degenerate limiting result. The asymptotic normal approximation appeared to work well in settings with different frequencies.

Using the asymptotic normality of the filtering errors, we considered the choice of the optimal filter. The optimal filter turned out to be a score-driven filter based on the true conditional predictive distribution of the DGP. This filter serves as an (infeasible) benchmark, enabling us to assess the quality of other filters. At the same time, it motivates the use of score-driven models in settings where one believes the model to conditional predictive density is correctly specified.

References

Angelini, G. and P. Gorgi (2018). DSGE models with observation-driven time-varying volatility. *Economics Letters 171*, 169–171.

Babii, A., X. Chen, and E. Ghysels (2019). Commercial and residential mortgage defaults: Spatial dependence with frailty. *Journal of Econometrics 212*(1), 47–77.

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

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

Blasques, F., S. J. Koopman, and A. Lucas. (2018). Amendments and corrections: Information theoretic optimality of observation driven time series models for continuous responses. *Biometrika 105*, 753.

Blasques, F., S. J. Koopman, A. Lucas, and J. Schaumburg (2016). Spillover dynamics for systemic risk measurement using spatial financial time series models. *Journal of Econometrics 195*(2), 211–223.

Blasques, F., A. Lucas, and E. Silde (2018). A stochastic recurrence equations approach for score driven correlation models. *Econometric Reviews 37*(2), 166–181.

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

Buccheri, G., G. Bormetti, F. Corsi, and F. Lillo (2021). A score-driven conditional correlation model for noisy and asynchronous data: An application to high-frequency covariance dynamics. *Journal of Business & Economic Statistics 39*(4), 920–936.

Buccheri, G., F. Corsi, F. Flandoli, and G. Livieri (2021). The continuous-time limit of score-driven volatility models. *Journal of Econometrics 221*(2), 655–675.

Catania, L. and A. Luati (2023). Semiparametric modeling of multiple quantiles. *Journal of Econometrics*, (in press).

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 93–115.

Creal, D., S. J. Koopman, and A. Lucas. (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics 29*, 552–563.

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

Creal, D., S. J. Koopman, A. Lucas, and M. Zamojski (2019). Generalized autoregressive method of moments. Discussion paper (revised version), Tinbergen Institute.

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

Danielsson, J. (1994). Stochastic volatility in asset prices estimation with simulated maximum likelihood. *Journal of Econometrics 64*(1), 375–400.

Davidson, J. (1994). *Stochastic Limit Theory.* Oxford: Oxford University Press.

D'Innocenzo, E., A. Lucas, B. Schwaab, and X. Zhang (2023). Modeling extreme events: Time-varying extreme tail shape. *Journal of Business and Economic Statistics*, (in press).

Dudley, R. M. (2002). *Real Analysis and Probability.* Cambridge: Cambridge University Press.

Engle, R. F. and G. M. Gallo (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics 131*(1-2), 3–27.

Ethier, S. N. and T. Nagylaki (1980). Diffusion approximations of Markov chains with two time scales and applications to population genetics. *Advances in Applied Probability 12*(1), 14–49.

Ethier, S. N. and T. Nagylaki (1988). Diffusion approximations of Markov chains with two time scales and applications to population genetics, II. *Advances in Applied Probability 20*(3), 525–545.

Gasperoni, F., A. Luati, L. Paci, and E. D'Innocenzo (2021). Score-driven modeling of spatio-temporal data. *Journal of the American Statistical Association*, 1–12.

Gorgi, P. (2020). Beta–negative binomial auto-regressions for modelling integer-valued time series with extreme observations. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(5), 1325–1347.

Gorgi, P., P. R. Hansen, P. Janus, and S. J. Koopman (2019). Realized Wishart-GARCH: A score-driven multi-asset volatility model. *Journal of Financial Econometrics 17*(1), 1–32.

Hafner, C. M. and L. Wang (2023). A dynamic conditional score model for the log correlation matrix. *Journal of Econometrics*, (in press).

Hansen, P. R. and M. Schmidtblaicher (2021). A dynamic model of vaccine compliance: How fake news undermined the Danish HPV vaccine program. *Journal of Business & Economic Statistics 39*(1), 259–271.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series.* Econometric Series Monographs. Cambridge: Cambridge University Press.

Harvey, A. C., S. Hurn, D. Palumbo, and S. Thiele (2023). Modeling circular time series. *Journal of Econometrics* (in press).

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

Helland, I. S. (1982). Central limit theorems for martingales with discrete or continuous time. *Scandinavian Journal of Statistics 9*(2), 79–94.

Hetland, S., R. S. Pedersen, and A. Rahbek (2023). Dynamic conditional eigenvalue GARCH. *Journal of Econometrics*, (in press).

Horn, R. A. and C. R. Johnson (2012). *Matrix Analysis* (2nd ed.). Cambridge: Cambridge University Press.

Jensen, A. T. and T. Lange (2010). On convergence of the QMLE for misspecified GARCH models. *Journal of Time Series Econometrics 2*(1).

Joao, I. C., A. Lucas, J. Schaumburg, and B. Schwaab (2023). Dynamic clustering of multivariate panel data. *Journal of Econometrics*.

Karatzas, I. and S. E. Shreve (1998). *Brownian Motion and Stochastic Calculus*. NY: Springer New York.

Klenke, A. (2020). *Probability Theory: A Comprehensive Course* (3rd ed.). Universitext. Cham: Springer.

Koopman, S. J., R. Lit, A. Lucas, and A. Opschoor (2018). Dynamic discrete copula models for high-frequency stock price changes. *Journal of Applied Econometrics 33*(7), 966–985.

Lancaster, P. and M. Tismenetsky (1985). *The Theory of Matrices* (2nd ed.). San Deigo: Academic Press.

Lange, R.-J., B. van Os, and D. J. van Dijk (2022). Robust observation-driven models using proximal-parameter updates. *TI Discussion Papers 2022-066/III*.

Li, H., K. Yang, and D. Wang (2019). A threshold stochastic volatility model with explanatory variables. *Statistica Neerlandica 73*(1), 118–138.

Linton, O. and J. Wu (2020). A coupled component DCS-EGARCH model for intraday and overnight volatility. *Journal of Econometrics 217*(1), 176–201.

Lucas, A., J. Schaumburg, and B. Schwaab (2019). Bank business models at zero interest rates. *Journal of Business & Economic Statistics 37*(3), 542–555.

Lucas, A., B. Schwaab, and X. Zhang (2014). Conditional euro area sovereign default risk. *Journal of Business & Economic Statistics 32*(2), 271–284.

Massacci, D. (2017). Tail risk dynamics in stock returns: Links to the macroeconomy and global markets connectedness. *Management Science 63*(9), 3072–3089.

Monache, D. D., I. Petrella, and F. Venditti (2021). Price dividend ratio and long-run stock returns: A score-driven state space model. *Journal of Business & Economic Statistics 39*(4), 1054–1065.

Nelson, D. B. (1990). ARCH models as diffusion approximations. *Journal of Econometrics 45*(1-2), 7–38.

Nelson, D. B. (1992). Filtering and forecasting with misspecified ARCH models I: Getting the right variance with the wrong model. *Journal of Econometrics 52*(1), 61–90.

Nelson, D. B. (1996). Asymptotic filtering theory for multivariate ARCH models. *Journal of Econometrics 71*(1), 1–47.

Nelson, D. B. and D. P. Foster (1994). Asymptotic filtering theory for univariate ARCH models. *Econometrica 62*(1), 1–41.

Norman, M. F. (1975). Diffusion approximation of non-Markovian processes. *The Annals of Probability 3*(2), 358–364.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(3), 414–422.

Oh, D. H. and A. J. Patton (2018). Time-varying systemic risk: Evidence from a dynamic copula model of CDS spreads. *Journal of Business & Economic Statistics 36*(2), 181–195.

Oh, D. H. and A. J. Patton (2023). Dynamic factor copula models with estimated cluster assignments. *Journal of Econometrics*, 105374.

Opschoor, A., P. Janus, A. Lucas, and D. Van Dijk (2018). New HEAVY models for fat-tailed realized covariances and returns. *Journal of Business & Economic Statistics 36*(4), 643–657.

Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review 47*(2), 527–556.

Patton, A. J. and Y. Simsek (2023). Generalized autoregressive score trees and forests. *arXiv preprint arXiv:2305.18991*.

Patton, A. J., J. F. Ziegel, and R. Chen (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics 211*(2), 388–413.

Revuz, D. and M. Yor (1999). *Continuous Martingales and Brownian Motion* (3rd ed.). Number 293 in Grundlehren der mathematischen Wissenschaften. Berlin: Springer.

Schilling, R. L. (2017). *Measures, Integrals and Martingales* (2nd ed.). Cambridge: Cambridge University Press.

Talvila, E. (2001). Necessary and sufficient conditions for differentiating under the integral sign. *American Mathematical Monthly 108*(6), 544–548.

Umlandt, D. (2023). Score-driven asset pricing: Predicting time-varying risk premia based on cross-sectional model performance. *Journal of Econometrics*, 105470.

van Os, B. (2023). Information-theoretic time-varying density modeling. *TI Discussion Papers 2023-037/III*.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Yu, J. (2005). On leverage in a stochastic volatility model. *Journal of Econometrics 127*(2), 165–178.

Zhang, W. and S. S. Ge (2006). A global implicit function theorem without initial point and its applications to control of non-affine systems of high dimensions. *Journal of Mathematical Analysis and Applications 313*(1), 251–261.

Appendix to: Consistency, distributional convergence and optimality of score-driven filters

Eric Beutner[1], *Yicong Lin*[1,2] *and Andre Lucas*[1,2]

[1] *Vrije Universiteit Amsterdam*

[2] *Tinbergen Institute*

Contents

A General asymptotic results

In this section, we provide an overview of the fundamental conditions necessary for establishing the asymptotic properties of observation-driven filters under mis-specification. Let $\{\boldsymbol{V}_h(t_{i,h}),\, i=0,1,\ldots,\}$ be a generic time-homogeneous Markov chain in a metric space $\boldsymbol{E}_h$ corresponding to a one-step transition function. In our context, $\boldsymbol{V}_h(t_{i,h})$ encompasses various components, typically including the observed data or transformations thereof, the dynamic parameters, and the (scaled) filtering error, for example,

$$\boldsymbol{V}_h(t_{i,h}) = \Big(\boldsymbol{y}_h(t_{i,h}), \boldsymbol{\psi}_h(t_{i,h}), h^{-\kappa}\big(\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}^{\star}_h(t_{i,h})\big)\Big), \tag{A.1}$$

for some $\kappa \in [0, 1/4]$. Note that for a time-inhomogeneous process $\{\boldsymbol{V}(t)\}$, Exercise 1.10 in Revuz and Yor (1999, Chapter III) implies that the time-space process $\{(t, \boldsymbol{V}(t))\}$ is time-homogeneous. Thus, it is possible to make the transition probability depend on time by making the time index $t_{i,h}$ an element of $\boldsymbol{V}_h(t_{i,h})$. Let's assume that both $\Phi_h : \boldsymbol{E}_h \to \mathbb{R}^n$ and $\Psi_h : \boldsymbol{E}_h \to \mathbb{R}^m$ are Borel measurable. We can then define $\boldsymbol{X}_h(t_{i,h}) = \Phi_h\big(\boldsymbol{V}_h(t_{i,h})\big)$ and $\boldsymbol{Z}_h(t_{i,h}) = \Psi_h\big(\boldsymbol{V}_h(t_{i,h})\big)$. In most of the examples we are considering, the expression $\boldsymbol{X}_h(t_{i,h})$ can take the form of $\big(\boldsymbol{y}_h(t_{i,h}), \boldsymbol{\psi}_h(t_{i,h})\big)$, and $\boldsymbol{Z}_h(t_{i,h})$ would be represented as $h^{-\kappa}\big(\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}^{\star}_h(t_{i,h})\big)$. Here, $\boldsymbol{V}_h(t_{i,h})$ corresponds to the expression given in (A.1).

We assume $\big\{\big(\boldsymbol{X}_h(t_{i,h}), \boldsymbol{Z}_h(t_{i,h})\big),\, i = 0, 1, \ldots\big\}$ forms a time-homogeneous Markov chain in $\mathbb{R}^n \times \mathbb{R}^m$. However, if the process is non-Markovian, we refer to the discussions in Norman (1975) and Ethier and Nagylaki (1988, p. 527) for further insights. More specifically, let ν_h be a probability measure on $(\mathbb{R}^n \times \mathbb{R}^m, \mathcal{B}(\mathbb{R}^n \times \mathbb{R}^m))$, where $\mathcal{B}(\mathbb{R}^n \times \mathbb{R}^m)$ are the Borel sets on $\mathbb{R}^n \times \mathbb{R}^m$. We use $\mathcal{D}([0,\infty), \mathbb{R}^n \times \mathbb{R}^m)$ to denote the space of functions, endowed with the Skorohod metric, from $[0,\infty)$ into $\mathbb{R}^n \times \mathbb{R}^m$ that are right continuous with finite left limits. For $h > 0$, let $\Pi_h(\boldsymbol{x}, \boldsymbol{z}, \cdot)$ be a transition function on $\mathbb{R}^n \times \mathbb{R}^m$ and $\mathbb{P}_h$ be a (fixed) probability measure on $\mathcal{D}([0,\infty), \mathbb{R}^n \times \mathbb{R}^m)$ such that

$$\mathbb{P}_h\big[\big(\boldsymbol{X}_h(t_{i,h}), \boldsymbol{Z}_h(t_{i,h})\big) \in \boldsymbol{\Gamma}\big] = \nu_h(\boldsymbol{\Gamma}), \qquad \forall \boldsymbol{\Gamma} \in \mathcal{B}(\mathbb{R}^n \times \mathbb{R}^m), \tag{A.2}$$

$$\mathbb{P}_h\big[\big(\boldsymbol{X}_h(t), \boldsymbol{Z}_h(t)\big) = \big(\boldsymbol{X}_h(t_{i,h}), \boldsymbol{Z}_h(t_{i,h})\big),\; t_{i,h} \le t < t_{i+1,h}\big] = 1, \tag{A.3}$$

$$\mathbb{P}_h\big[\big(\boldsymbol{X}_h(t_{i+1,h}), \boldsymbol{Z}_h(t_{i+1,h})\big) \in \boldsymbol{\Gamma} \,\big|\, \mathcal{F}_h(t_{i,h})\big] = \Pi_h\big(\boldsymbol{X}_h(t_{i,h}), \boldsymbol{Z}_h(t_{i,h}), \boldsymbol{\Gamma}\big), \tag{A.4}$$

where $\mathcal{F}_h(t_{i,h})$ is a σ-algebra generated by $\boldsymbol{X}_h(0), \boldsymbol{X}_h(t_{1,h}), \ldots, \boldsymbol{X}_h(t_{i,h})$ and $\boldsymbol{Z}_h(0), \boldsymbol{Z}_h(t_{1,h}), \ldots, \boldsymbol{Z}_h(t_{i,h})$, and moreover, (A.4) holds almost surely under $\mathbb{P}_h$ for all $i \ge 0$ and all $\boldsymbol{\Gamma} \in \mathcal{B}(\mathbb{R}^n \times \mathbb{R}^m)$. For convenience, we adopt the following condition that rules out feedback from $\{\boldsymbol{Z}_h(t_{i,h})\}$ to $\{\boldsymbol{X}_h(t_{i,h})\}$: for every Borel subset $\boldsymbol{\Gamma}_x$ of $\mathbb{R}^n$ and for all $h > 0$,

$$\Pi_h(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\Gamma}_x \times \mathbb{R}^m) \quad \text{is independent of } \boldsymbol{z}. \tag{A.5}$$

This assumption can also be omitted by making some modifications to the assumptions presented below; see also the discussion following the Assumptions below.

The following assumptions are adapted from Ethier and Nagylaki (1988) and Nelson (1992). The first four assumptions imply the weak convergence of the process $\{\boldsymbol{X}_h(t)\}$ to $\{\boldsymbol{X}(t)\}$, as $h \downarrow 0$, whose distribution can be uniquely determined (Assumption A4).

Assumptions:

A1 $\boldsymbol{X}_h(0) \Rightarrow \boldsymbol{X}(0)$ *as* $h \downarrow 0$*, where* $\boldsymbol{X}(0)$ *has probability measure* $\tilde{\nu}_0$*.*

A2 For every $\eta > 0$*,*

$$\lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} \left\| h^{-1}\mathbb{E}\left[\Delta \boldsymbol{X}_h(t_{i+1,h}) \,\middle|\, \boldsymbol{X}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{Z}_h(t_{i,h}) = \boldsymbol{z}\right] - \widetilde{\boldsymbol{\mu}}(\boldsymbol{x}) \right\| = 0,$$

$$\lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} \left\| h^{-1}\mathbb{E}\left[\left(\Delta \boldsymbol{X}_h(t_{i+1,h})\right)\left(\Delta \boldsymbol{X}_h(t_{i+1,h})\right)^{\top} \,\middle|\, \boldsymbol{X}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{Z}_h(t_{i,h}) = \boldsymbol{z}\right] - \widetilde{\boldsymbol{\Omega}}(\boldsymbol{x}) \right\| = 0,$$

where $\widetilde{\boldsymbol{\mu}}(\,\cdot\,)$ *and* $\widetilde{\boldsymbol{\Omega}}(\,\cdot\,)$ *are continuous, and the expectations above are taken under* $\mathbb{P}_h$*.*

A3 For every $\eta > 0$ *and all* $j = 1, \ldots, n$*,*

$$\lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} \left\| h^{-1}\mathbb{E}\left[\left(\Delta \boldsymbol{X}_h(t_{i+1,h})\right)_j^4 \,\middle|\, \boldsymbol{X}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{Z}_h(t_{i,h}) = \boldsymbol{z}\right] \right\| = 0, \tag{A.6}$$

where $(\boldsymbol{a})_j$ *denotes the* j_{th} *element of a vector* $\boldsymbol{a}$*.*

A4 $\tilde{\nu}_0$*,* $\widetilde{\boldsymbol{\mu}}(\boldsymbol{x})$*, and* $\widetilde{\boldsymbol{\Omega}}(\boldsymbol{x})$ *uniquely specify the distribution of a diffusion process* $\boldsymbol{X}(t)$ *with initial distribution* $\tilde{\nu}_0$*, drift vector* $\widetilde{\boldsymbol{\mu}}(\boldsymbol{x})$*, and diffusion matrix* $\widetilde{\boldsymbol{\Omega}}(\boldsymbol{x})$*.*

A5 For some δ*,* $0 < \delta < 1$*, and for every* $\eta > 0$*,*

$$\lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} \left\| h^{-\delta}\mathbb{E}\left[\Delta \boldsymbol{Z}_h(t_{i+1,h}) \,\middle|\, \boldsymbol{X}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{Z}_h(t_{i,h}) = \boldsymbol{z}\right] - \boldsymbol{c}(\boldsymbol{x},\boldsymbol{z}) \right\| = 0, \tag{A.7}$$

where for all $\boldsymbol{x} \in \mathbb{R}^n$*,* $\boldsymbol{c}(\boldsymbol{x},\boldsymbol{0}) = \boldsymbol{0}$*, and*

$$\lim_{h\downarrow 0} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} \left\| h^{-\delta}\mathbb{E}\left[\left(\Delta \boldsymbol{Z}_h(t_{i+1,h})\right)\left(\Delta \boldsymbol{Z}_h(t_{i+1,h})\right)^{\top} \,\middle|\, \boldsymbol{X}_h(t_{i,h}) = \boldsymbol{x},\, \boldsymbol{Z}_h(t_{i,h}) = \boldsymbol{z}\right] \right\| = 0. \tag{A.8}$$

A6 For each $\boldsymbol{x} \in \mathbb{R}^n$*,* $\boldsymbol{z} \in \mathbb{R}^m$*, define the differential equation with an initial condition*

$$\frac{\mathrm{d}\boldsymbol{Z}(t,\boldsymbol{x},\boldsymbol{z})}{\mathrm{d}t} = \boldsymbol{c}\left(\boldsymbol{x}, \boldsymbol{Z}\left(t,\boldsymbol{x},\boldsymbol{z}\right)\right), \qquad \boldsymbol{Z}\left(0,\boldsymbol{x},\boldsymbol{z}\right) = \boldsymbol{z}. \tag{A.9}$$

We require $\boldsymbol{0}_{m\times 1}$ *is a global asymptotically stable solution of* (A.9) *for bounded values of* $\boldsymbol{x}$ *and*

$\boldsymbol{z}$, i.e., for every $\eta > 0$,

$$\lim_{t\to\infty} \sup_{\|(\boldsymbol{x},\boldsymbol{z})\|\le\eta} \|\boldsymbol{Z}(t,\boldsymbol{x},\boldsymbol{z})\| = 0. \tag{A.10}$$

A7 There exist Borel functions $\rho_h : \boldsymbol{E}_h \to [0,\infty)$, and constants $\lambda(\eta,h) > 0$ such that

$$\lim_{\eta\to\infty} \liminf_{h\downarrow 0} \inf_{\boldsymbol{v}\notin\boldsymbol{\mathcal{G}}_{\eta,h}} \rho_h(\boldsymbol{v}) = \infty, \qquad \boldsymbol{\mathcal{G}}_{\eta,h} = \Big\{\boldsymbol{v}\in\boldsymbol{E}_h : \big\|(\Phi_h(\boldsymbol{v}),\Psi_h(\boldsymbol{v}))\big\| \le \eta\Big\}, \tag{A.11}$$

$$\limsup_{\eta\to\infty} \limsup_{h\downarrow 0} \lambda(\eta,h) < \infty, \tag{A.12}$$

$$\limsup_{h\downarrow 0} \mathbb{E}\left[\rho_h\big(\boldsymbol{V}_h(0)\big)\right] < \infty, \tag{A.13}$$

and for every $\eta > 0$ and $h > 0$,

$$\sup_{\boldsymbol{v}\in\boldsymbol{\mathcal{G}}_{\eta,h}} \Big\{ h^{-1}\mathbb{E}\Big[\rho_h\big(\boldsymbol{V}_h(t_{i+1,h})\big) - \rho_h(\boldsymbol{v}) \,\Big|\, \boldsymbol{V}_h(t_{i,h}) = \boldsymbol{v}\Big] - \lambda(\eta,h)\rho_h(\boldsymbol{v})\Big\} \le 0. \tag{A.14}$$

Note that the moment conditions in Assumption A2 are independent of $\boldsymbol{z}$ under $\mathbb{P}_h$ by Eq. (A.5). Similar conditions have been imposed in Ethier and Nagylaki (1988), see their Eqs. (1.10), (1.15), and (1.16). One can consider allowing $\widetilde{\boldsymbol{\mu}}(\,\cdot\,)$ and $\widetilde{\boldsymbol{\Omega}}(\,\cdot\,)$ to depend on $\boldsymbol{z}$, thereby modifying Assumption A4 accordingly. In this case, condition (A.5) can be dropped. Appendix A in Nelson (1990) summarizes a set of conditions that imply the distributional uniqueness as required in Assumption A4. Further intuition may be given as follows. Assumptions A2, A3, and A5, may imply that, for some $\delta \in (0,1)$,

$$h^{-1}\mathbb{E}\left[\begin{pmatrix}\Delta\boldsymbol{X}_h(t_{i+1,h})\\ \Delta\boldsymbol{Z}_h(t_{i+1,h})\end{pmatrix} \,\middle|\, \boldsymbol{X}_h(t_{i,h}) = \boldsymbol{x},\ \boldsymbol{Z}_h(t_{i,h}) = \boldsymbol{z}\right] \approx \begin{pmatrix}\widetilde{\boldsymbol{\mu}}(\boldsymbol{x})\\ h^{\delta-1}\boldsymbol{c}(\boldsymbol{x},\boldsymbol{z})\end{pmatrix},$$

as $h \downarrow 0$. Note that $h^{\delta-1} \to \infty$ as $h \downarrow 0$. If the process $\big\{\big(\boldsymbol{X}_h(t),\boldsymbol{Z}_h(t)\big)\big\}$ does not explode in finite time, which is guaranteed by Assumptions A3, A4, and A7, then $\boldsymbol{Z}_h(t)$ must converge to $\boldsymbol{0}$, requiring the differential equation in Assumption A6 has a stable solution of $\boldsymbol{0}$. Therefore, we have the following theorem that serves as a valuable result for demonstrating the filter consistency, see Theorem 2.1 of Ethier and Nagylaki (1988) as well as Theorem 2.2 of Nelson (1992).

Theorem A.1 (Ethier and Nagylaki, 1988)
Let Assumptions A1 - A7 hold. Then, for each $t > 0$,

$$\boldsymbol{Z}_h(t) \overset{p}{\to} \boldsymbol{0}_{m\times 1}, \qquad \text{as } h \downarrow 0. \tag{A.15}$$

We now present the asymptotic results that establish the weak convergence of filtering errors.

To achieve this, we define $\boldsymbol{V}_h(t_{i,h})$ as given in Eq. (4.1) with $\kappa = 1/4$. Furthermore, we define $\widetilde{\boldsymbol{X}}_h(t_{i,h}) = \Phi_h\big(\boldsymbol{V}_h(t_{i,h})\big) = \big(\boldsymbol{\upsilon}_h(t_{i,h}), \boldsymbol{\psi}_h(t_{i,h})\big)$ and $\widetilde{\boldsymbol{Z}}_h(t_{i,h}) = \Psi_h\big(\boldsymbol{V}_h(t_{i,h})\big) = h^{-1/4}\big(\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}^{\star}_h(t_{i,h})\big)$. For $\tau \geq 0$, let $\widetilde{\boldsymbol{X}}_{T,h}(\tau) = \widetilde{\boldsymbol{X}}_h\big(T + \tau h^{1/2}\big)$ and $\widetilde{\boldsymbol{Z}}_{T,h}(\tau) = \widetilde{\boldsymbol{Z}}_h\big(T + \tau h^{1/2}\big)$. We begin by presenting the assumptions.

Assumptions:

Let t be a function of (T, τ, h) given by $t \equiv T + \tau h^{1/2}$.

A8 The following functions are well-defined and twice-differentiable in $\boldsymbol{x} = (\boldsymbol{\upsilon}, \boldsymbol{\psi})$:

$$\boldsymbol{A}(\boldsymbol{x}) \equiv - \lim_{h \downarrow 0} \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}^{\top}} \boldsymbol{g}_h\Big(\boldsymbol{y}_h(t+h), \boldsymbol{\theta}^{\star}_h(t)\Big) \,\middle|\, \widetilde{\boldsymbol{X}}_h(t) = \boldsymbol{x} \right], \tag{A.16}$$

$$\begin{aligned} \boldsymbol{\Sigma}(\boldsymbol{x}) \equiv \lim_{h \downarrow 0} \mathbb{E}\Bigg\{ & \left[\boldsymbol{g}_h\Big(\boldsymbol{y}_h(t+h), \boldsymbol{\theta}^{\star}_h(t)\Big) - \left(\frac{\partial \boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^{\top}}\right) \boldsymbol{B} \boldsymbol{\eta}_{\lfloor t/h \rfloor + 1} \right] \\ & \times \left[\boldsymbol{g}_h\Big(\boldsymbol{y}_h(t+h), \boldsymbol{\theta}^{\star}_h(t)\Big) - \left(\frac{\partial \boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^{\top}}\right) \boldsymbol{B} \boldsymbol{\eta}_{\lfloor t/h \rfloor + 1} \right]^{\top} \,\Bigg|\, \widetilde{\boldsymbol{X}}_h(t) = \boldsymbol{x} = (\boldsymbol{\upsilon}, \boldsymbol{\psi}) \Bigg\}. \end{aligned} \tag{A.17}$$

Further,

$$h^{-1/2}\mathbb{E}\Big[\Delta \widetilde{\boldsymbol{X}}_h(t+h) \,\big|\, \widetilde{\boldsymbol{X}}_h(t) = \boldsymbol{x},\, \widetilde{\boldsymbol{Z}}_h(t) = \boldsymbol{z}\Big] \to \boldsymbol{0}, \tag{A.18}$$

$$h^{-1/2}\,\mathbb{C}\mathrm{ov}\Big[\Delta \widetilde{\boldsymbol{X}}_h(t+h) \,\big|\, \widetilde{\boldsymbol{X}}_h(t) = \boldsymbol{x},\, \widetilde{\boldsymbol{Z}}_h(t) = \boldsymbol{z}\Big] \to \boldsymbol{0}, \tag{A.19}$$

and

$$h^{-1/2}\mathbb{E}\Big[\Delta \widetilde{\boldsymbol{Z}}_h(t+h) \,\big|\, \widetilde{\boldsymbol{X}}_h(t) = \boldsymbol{x},\, \widetilde{\boldsymbol{Z}}_h(t) = \boldsymbol{z}\Big] \to -\boldsymbol{A}(\boldsymbol{x})\boldsymbol{z}, \tag{A.20}$$

$$h^{-1/2}\,\mathbb{C}\mathrm{ov}\Big[\Delta \widetilde{\boldsymbol{Z}}_h(t+h) \,\big|\, \widetilde{\boldsymbol{X}}_h(t) = \boldsymbol{x},\, \widetilde{\boldsymbol{Z}}_h(t) = \boldsymbol{z}\Big] \to \boldsymbol{\Sigma}(\boldsymbol{x}), \tag{A.21}$$

as $h \downarrow 0$ uniformly on every bounded $(\boldsymbol{x}, \boldsymbol{z})$ set.

A9 For some $\varphi > 0$, the following conditional expectations

$$\mathbb{E}\Big[\big\|h^{-1/2}\Delta \widetilde{\boldsymbol{X}}_h(t+h)\big\|^{2+\varphi} \,\Big|\, \widetilde{\boldsymbol{X}}_h(t) = \boldsymbol{x},\, \widetilde{\boldsymbol{Z}}_h(t) = \boldsymbol{z}\Big], \tag{A.22}$$

are bounded as $h \downarrow 0$, uniformly on every bounded $(\boldsymbol{x}, \boldsymbol{z})$ set. Moreover,

$$h^{-1/2}\mathbb{E}\Big[\big\|\Delta \widetilde{\boldsymbol{Z}}_h(t+h)\big\|^{2+\varphi} \,\Big|\, \widetilde{\boldsymbol{X}}_h(t) = \boldsymbol{x},\, \widetilde{\boldsymbol{Z}}_h(t) = \boldsymbol{z}\Big] \to 0, \tag{A.23}$$

uniformly on every bounded $(\boldsymbol{x}, \boldsymbol{z})$ set.

The set of Assumptions A8 - A9 is high-level and appears to be isolated from Assumptions A1 - A7. However, as seen in Theorems 1 - 2, the verification of these two sets of assumptions leads to similar conditions.

The following theorem is an adaption of Theorem 2.1 in Nelson (1996). The univariate counterpart is established in Nelson and Foster (1994, Theorem 3.1). It states that, for $M > 0$, the process $\big\{\widetilde{\boldsymbol{Z}}_{T,h}(\tau)\big\}_{\tau\in[0,M]}$, defined on the fast time scale, converges weakly to a diffusion process conditional on $\big(\widetilde{\boldsymbol{X}}_{T,h}(0), \widetilde{\boldsymbol{Z}}_{T,h}(0)\big) = (\boldsymbol{x}, \boldsymbol{z})$.

Theorem A.2 (Nelson, 1996)

Let Assumptions A8 - A9 be satisfied. Let $\boldsymbol{\Gamma}_0$ be a bounded, open subset of $\mathbb{R}^{k_y+k_\psi+k_\theta}$ on which for some $\varepsilon > 0$, the real parts of all the eigenvalues of $\boldsymbol{A}(\boldsymbol{x})$ are bounded below by ε. Then for every $(\boldsymbol{x}, \boldsymbol{z}) \in \boldsymbol{\Gamma}_0$, conditional on $\big(\widetilde{\boldsymbol{X}}_{T,h}(0), \widetilde{\boldsymbol{Z}}_{T,h}(0)\big) = (\boldsymbol{x}, \boldsymbol{z})$, $\big\{\widetilde{\boldsymbol{Z}}_{T,h}(\tau)\big\}_{\tau\in[0,M]}$ converges weakly to the diffusion

$$\mathrm{d}\boldsymbol{Z}(\tau) = -\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Z}(\tau)\mathrm{d}\tau + \boldsymbol{\Sigma}(\boldsymbol{x})^{1/2}\mathrm{d}\boldsymbol{W}_\tau, \qquad \textit{as } h \downarrow 0, \tag{A.24}$$

where $\boldsymbol{W}_\tau$ is a standard Brownian motion. This convergence is uniform on $\boldsymbol{\Gamma}_0$. Further, for every $(\boldsymbol{x}, \boldsymbol{z}) \in \boldsymbol{\Gamma}_0$,

$$\widetilde{\boldsymbol{Z}}_{T,h}(\tau) \,\Big|\, \big(\widetilde{\boldsymbol{X}}_{T,h}(0), \widetilde{\boldsymbol{Z}}_{T,h}(0)\big) = (\boldsymbol{x}, \boldsymbol{z}) \xrightarrow{d} \mathcal{N}\Big(\boldsymbol{b}(\tau, \boldsymbol{x}, \boldsymbol{z}), \boldsymbol{V}(\tau, \boldsymbol{x})\Big), \qquad \textit{as } h \downarrow 0, \tag{A.25}$$

where $\boldsymbol{b}(\tau, \boldsymbol{x}, \boldsymbol{z}) = \exp\big[-\tau\boldsymbol{A}(\boldsymbol{x})\big]\boldsymbol{z}$ and

$$\boldsymbol{V}(\tau, \boldsymbol{x}) = \exp\big[-\tau\boldsymbol{A}(\boldsymbol{x})\big]\left\{\int_0^\tau \exp\big[s\boldsymbol{A}(\boldsymbol{x})\big]\boldsymbol{\Sigma}(\boldsymbol{x})\exp\big[s\boldsymbol{A}(\boldsymbol{x})^\top\big]\,\mathrm{d}s\right\}\exp\big[-\tau\boldsymbol{A}(\boldsymbol{x})^\top\big].$$

It is important to note that Eqs. (A.18) - (A.19) are inherently satisfied within the DGP in Nelson (1996). Therefore, they are only implicitly imposed by Nelson (1996). Given the additional mis-specification in our setting, we explicate these assumptions. Additionally, instead of enforcing (A.23), Nelson (1996) requires that

$$\mathbb{E}\left[\left\|\boldsymbol{g}_h\Big(\boldsymbol{y}_h(t+h), \boldsymbol{\theta}_h(t)\Big)\right\|^{2+\varphi} \,\middle|\, \widetilde{\boldsymbol{X}}_h(t) = \boldsymbol{x}, , \widetilde{\boldsymbol{Z}}_h(t) = \boldsymbol{z}\right]$$

be bounded as $h \downarrow 0$, uniformly on every bounded $(\boldsymbol{x}, \boldsymbol{z})$ set. While this condition implies (A.23) within Nelson's framework, it may not necessarily hold true within our own framework. We therefore again make this assumption explicit.

The key to proving Theorem A.2 involves three steps.

(i) By Eqs. (A.18) - (A.22), one argues that the first two conditional moments

$$h^{-1/2}\mathbb{E}\left[\begin{pmatrix}\Delta\widetilde{\boldsymbol{X}}_h(t+h)\\ \Delta\widetilde{\boldsymbol{Z}}_h(t+h)\end{pmatrix}\,\middle|\,\widetilde{\boldsymbol{X}}_h(t)=\boldsymbol{x},\,\widetilde{\boldsymbol{Z}}_h(t)=\boldsymbol{z}\right]\to\begin{pmatrix}\boldsymbol{0}\\ -\boldsymbol{A}(\boldsymbol{x})\boldsymbol{z}\end{pmatrix},$$

as $h\downarrow 0$, uniformly on every bounded $(\boldsymbol{x},\boldsymbol{z})$ set, and similarly,

$$h^{-1/2}\,\mathbb{C}\mathrm{ov}\left[\begin{pmatrix}\Delta\widetilde{\boldsymbol{X}}_h(t+h)\\ \Delta\widetilde{\boldsymbol{Z}}_h(t+h)\end{pmatrix}\,\middle|\,\widetilde{\boldsymbol{X}}_h(t)=\boldsymbol{x},\,\widetilde{\boldsymbol{Z}}_h(t)=\boldsymbol{z}\right]\to\mathrm{diag}\Big(\boldsymbol{0},\boldsymbol{\Sigma}(\boldsymbol{x})\Big).$$

(ii) By Eqs. (A.22) and (A.23),

$$h^{-1/2}\mathbb{E}\left[\left\|\begin{pmatrix}\Delta\widetilde{\boldsymbol{X}}_h(t+h)\\ \Delta\widetilde{\boldsymbol{Z}}_h(t+h)\end{pmatrix}\right\|^{2+\varphi}\,\middle|\,\widetilde{\boldsymbol{X}}_h(t)=\boldsymbol{x},\,\widetilde{\boldsymbol{Z}}_h(t)=\boldsymbol{z}\right]\to 0,$$

uniformly on every bounded $(\boldsymbol{x},\boldsymbol{z})$ set.

(iii) Let $\Delta=1/2$ in Theorem 2.1 of Nelson and Foster (1994). This leads to the joint weak convergence of the process $\big\{\big(\widetilde{\boldsymbol{X}}_h(t),\widetilde{\boldsymbol{Z}}_h(t)\big)=\big(\widetilde{\boldsymbol{X}}_{T,h}(\tau),\widetilde{\boldsymbol{Z}}_{T,h}(\tau)\big)\big\}_{\tau\in[0,M]}$ for $t=T+\tau h^{1/2}$. More importantly, since the drift and covariance of $\widetilde{\boldsymbol{X}}_{T,h}(\tau)$ are asymptotically zeros, the process $\big\{\widetilde{\boldsymbol{X}}_{T,h}(\tau)\big\}_{\tau\in[0,M]}$ degenerates to the initial value $\widetilde{\boldsymbol{X}}_{T,h}(0)=\widetilde{\boldsymbol{X}}_h(T)$ in the limit, while $\big\{\widetilde{\boldsymbol{Z}}_{T,h}(\tau)\big\}_{\tau\in[0,M]}$ becomes a diffusion.

In contrast to Theorem A.1, the weak convergence described in Eqs. (A.24) and (A.25) is conditional on the initial values $\big(\widetilde{\boldsymbol{X}}_{T,h}(0),\widetilde{\boldsymbol{Z}}_{T,h}(0)\big)=(\boldsymbol{x},\boldsymbol{z})$. Since $T\geq 0$ can take any value, this condition fundamentally requires that the sample path of $\big\{\big(\widetilde{\boldsymbol{X}}_h(t),\widetilde{\boldsymbol{Z}}_h(t)\big)\big\}$ is non-explosive everywhere, as specified in Assumption A7. Under this (strong) condition, it is intuitively possible to relax or drop certain assumptions for filter consistency. For instance, Assumption A1 is automatically satisfied in this case. However, both theorems still require the sample path of the process $\big\{\widetilde{\boldsymbol{X}}_h(t)\big\}$ to be finite and, asymptotically, free from discrete jumps. This explains why checking the two sets of assumptions, namely Assumptions A1 - A7 and Assumptions A8 - A9, ultimately leads to similar constraints, despite their apparent differences.

B Filter consistency

For any $\eta>0$, define two compact sets $\boldsymbol{\mathcal{N}}_1(\eta)=\{\boldsymbol{x}=(\boldsymbol{\upsilon},\boldsymbol{\psi}):\|(\boldsymbol{\upsilon},\boldsymbol{\psi})\|\leq\eta\}\subset\mathbb{R}^{k_\upsilon}\times\boldsymbol{\Psi}$ and $\boldsymbol{\mathcal{N}}_2(\eta)=\{(\boldsymbol{x},\boldsymbol{z})=(\boldsymbol{\upsilon},\boldsymbol{\psi},\boldsymbol{z}):\|(\boldsymbol{\upsilon},\boldsymbol{\psi},\boldsymbol{z})\|\leq\eta\}\subset\mathbb{R}^{k_\upsilon}\times\boldsymbol{\Psi}\times\mathbb{R}^{k_\theta}$.

Proof of Theorem 1 Let $\boldsymbol{V}_h(t_{i,h})$ in Appendix A be specified as $\boldsymbol{V}_h(t_{i,h})=\Big(\boldsymbol{\upsilon}_h(t_{i,h}),\boldsymbol{\psi}_h(t_{i,h}),h^{-\kappa}\big(\boldsymbol{\theta}_h(t_{i,h})-$

$\boldsymbol{\theta}_h^\star(t_{i,h})\big)\Big)$. Define Φ_h and Ψ_h such that $\boldsymbol{x}_h(t_{i,h}) = \Phi_h\big(\boldsymbol{V}_h(t_{i,h})\big)$ and $\boldsymbol{z}_h(t_{i,h}) = \Psi_h\big(\boldsymbol{V}_h(t_{i,h})\big)$, where $\boldsymbol{x}_h(t_{i,h})$ and $\boldsymbol{z}_h(t_{i,h})$ are shown in Eq. (4.1). To apply Theorem A.1, we shall verify Assumptions A1 - A7 for $\big\{\big(\boldsymbol{x}_h(t_{i,h}), \boldsymbol{z}_h(t_{i,h})\big)\big\}$. Let $\mathbb{E}_{i,h}\left[\,\cdot\,\right] = \mathbb{E}\big[\,\cdot\,\big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x} = (\boldsymbol{\upsilon}, \boldsymbol{\psi}),\, \boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}\big]$. We separate the proof into three main parts. Note that Parts I and III do not impose conditions on the value of κ as given in Eq. (4.1).

I. Verify Assumptions A1 - A4

We first verify Assumptions A2 - A3. Let $\widetilde{\boldsymbol{\mu}}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{\mu}(\boldsymbol{x}) \\ \boldsymbol{a}(\boldsymbol{\psi}) \end{pmatrix}$ and $\widetilde{\boldsymbol{\Omega}}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{\Omega}_{\upsilon\upsilon}(\boldsymbol{x}) & \boldsymbol{\Omega}_{\upsilon\eta}(\boldsymbol{x})\boldsymbol{B}(\boldsymbol{\psi})^\top \\ \boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{\Omega}_{\upsilon\eta}(\boldsymbol{x})^\top & \boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{B}(\boldsymbol{\psi})^\top \end{pmatrix}$. By Assumption FC.2, we have

$$\begin{aligned}\lim_{h\downarrow 0}\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} &\left\|h^{-1}\mathbb{E}_{i,h}\left[\Delta\boldsymbol{x}_h(t_{i+1,h})\right] - \widetilde{\boldsymbol{\mu}}(\boldsymbol{x})\right\| \\ &= \lim_{h\downarrow 0}\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\|\begin{pmatrix} h^{-1}\mathbb{E}_{i,h}\left[\Delta\boldsymbol{\upsilon}_h(t_{i+1,h})\right] - \boldsymbol{\mu}(\boldsymbol{x}) \\ \boldsymbol{a}_h(\boldsymbol{\psi}) - \boldsymbol{a}(\boldsymbol{\psi})\end{pmatrix}\right\| = 0, \end{aligned} \tag{B.1}$$

and similarly,

$$\lim_{h\downarrow 0}\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\|h^{-1}\mathbb{E}_{i,h}\left[\left(\Delta\boldsymbol{x}_h(t_{i+1,h})\right)\left(\Delta\boldsymbol{x}_h(t_{i+1,h})\right)^\top\right] - \widetilde{\boldsymbol{\Omega}}(\boldsymbol{x})\right\| = 0.$$

Then Assumption A2 is fulfilled. Using the conditions of the fourth-order moments given in Assumptions FC.1 - FC.2, and the c_r-inequality, one can immediately verify Assumption A3. Finally, Assumption FC.3 contains Assumptions A1 and A4.

II. Verify Assumptions A5 - A6

We first verify Eq. (A.7) in Assumption A5. It is worth noting that the results presented below, including Eq. (B.9), hold true for any $\kappa \in [0, 1/4]$. Take $\delta = 1/2$. For every $\eta > 0$,

$$\begin{aligned} &h^{-\delta}\mathbb{E}_{i,h}\left[\Delta\boldsymbol{z}_h(t_{i+1,h})\right] \\ &= h^{-1/2-\kappa}\mathbb{E}_{i,h}\left[\Delta\boldsymbol{\theta}_h(t_{i+1,h}) - \Delta\boldsymbol{\theta}_h^\star(t_{i+1,h})\right] \\ &= h^{1/2-\kappa}\boldsymbol{\beta}\mathbb{E}_{i,h}\left[\boldsymbol{\theta}_h(t_{i,h})\right] + h^{-\kappa}\mathbb{E}_{i,h}\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big)\Big] - h^{-1/2-\kappa}\mathbb{E}_{i,h}\left[\Delta\boldsymbol{\theta}_h^\star(t_{i+1,h})\right] + h^{1/2-\kappa}\boldsymbol{\omega}. \end{aligned} \tag{B.2}$$

Conditional on $\big(\boldsymbol{x}_h(t_{i,h}), \boldsymbol{z}_h(t_{i,h})\big) = (\boldsymbol{\upsilon}, \boldsymbol{\psi}, \boldsymbol{z})$, we have by definition $\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}_h^\star(t_{i,h}) = \boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\iota}_h(\boldsymbol{\psi}) = h^\kappa\boldsymbol{z}$. Using Assumption FC.4, we have the first term in (B.2)

$$\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\|h^{1/2-\kappa}\boldsymbol{\beta}\mathbb{E}_{i,h}\left[\boldsymbol{\theta}_h(t_{i,h})\right]\right\| \le C\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left[h^{1/2}\|\boldsymbol{z}\| + h^{1/2-\kappa}\|\boldsymbol{\iota}_h(\boldsymbol{\psi})\|\right] = o(1).$$

For the second term in (B.2), we expand element-wise $\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big)$ in a Taylor series around $\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h^\star(t_{i,h})\big)$. Using the identity $\boldsymbol{a}^\top \boldsymbol{H}\boldsymbol{a} = \operatorname{tr}\left(\boldsymbol{a}^\top \boldsymbol{H}\boldsymbol{a}\right) = \left[\operatorname{vec}\left(\boldsymbol{H}^\top\right)\right]^\top \operatorname{vec}(\boldsymbol{a}\boldsymbol{a}^\top)$ for any $\boldsymbol{a} \in \mathbb{R}^K$, $\boldsymbol{H} \in \mathbb{R}^{K\times K}$, $K \in \mathbb{Z}^+$, by Eq. (4.8) in Assumption FC.5, we arrive at

$$h^{-\kappa}\mathbb{E}_{i,h}\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big)\Big] = \mathbb{E}_{i,h}\left[\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h^\star(t_{i,h})\big)\right]\boldsymbol{z} + h^{-\kappa}\mathbb{E}_{i,h}\big[R_{i,h}\big(\boldsymbol{g}_h, \bar{\boldsymbol{\theta}}_h^\star(t_{i,h})\big)\big], \tag{B.3}$$

with

$$R_{i,h}\big(\boldsymbol{g}_h, \bar{\boldsymbol{\theta}}_h^\star(t_{i,h})\big) = \frac{1}{2}\begin{pmatrix} \operatorname{vec}\left[\dfrac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\boldsymbol{g}_{h,1}\Big(\boldsymbol{y}_h(t_{i+1,h}), \bar{\boldsymbol{\theta}}_{h,1}^\star(t_{i,h})\Big)\right]^\top \\ \vdots \\ \operatorname{vec}\left[\dfrac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\boldsymbol{g}_{h,k_\theta}\Big(\boldsymbol{y}_h(t_{i+1,h}), \bar{\boldsymbol{\theta}}_{h,k_\theta}^\star(t_{i,h})\Big)\right]^\top \end{pmatrix} \operatorname{vec}\left[\big(\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}_h^\star(t_{i,h})\big)\big(\boldsymbol{\theta}_h(t_{i,h}) - \boldsymbol{\theta}_h^\star(t_{i,h}\right.$$

where $\boldsymbol{g}_{h,j}\big(\boldsymbol{y}_h(t_{i+1,h}), \cdot\,\big)$ is the j_{th} element of the vector $\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \cdot\,\big)$, $j = 1, \dots, k_\theta$, and $\bar{\boldsymbol{\theta}}_{h,j}^\star(t_{i,h})$ lie on the segment joining $\boldsymbol{\theta}_h(t_{i,h})$ and $\boldsymbol{\theta}_h^\star(t_{i,h})$. Note that $\big\|\bar{\boldsymbol{\theta}}_{h,j}^\star(t_{i,h}) - \boldsymbol{\theta}_h^\star(t_{i,h})\big\| \leq h^\kappa\|\boldsymbol{z}\| \leq h^\kappa\eta$ conditional on $\boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}$, $j = 1, \dots, k_\theta$. For every $\eta > 0$ and every $j = 1, \dots, k_\theta$, there exists a constant $\delta_1 = \delta_1(\eta) > 0$ such that $\big\|\bar{\boldsymbol{\theta}}_{h,j}^\star(t_{i,h})\big\| \leq \big\|\boldsymbol{\theta}_h^\star(t_{i,h})\big\| + \big\|\bar{\boldsymbol{\theta}}_{h,j}^\star(t_{i,h}) - \boldsymbol{\theta}_h^\star(t_{i,h})\big\| \leq \delta_1$ by Eq. (4.4) in Assumption FC.4, as $h \downarrow 0$. By Eq. (4.9), we obtain

$$\Big\|h^{-\kappa}\mathbb{E}_{i,h}\big[R_{i,h}\big(\boldsymbol{g}_h, \bar{\boldsymbol{\theta}}_h^\star(t_{i,h})\big)\big]\Big\| \leq C\|\boldsymbol{z}\|^2\sum_{j=1}^{k_\theta} h^\kappa\mathbb{E}_{i,h}\left\{\sup_{\|\boldsymbol{\theta}\|\leq\delta_1}\left\|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\boldsymbol{g}_{h,j}\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}\big)\right\|\right\} = o(1), \tag{B.4}$$

uniformly on the bounded set $\boldsymbol{\mathcal{N}}_2(\eta)$. Therefore, for every $\eta > 0$,

$$\lim_{h\downarrow 0} h^{-\kappa}\mathbb{E}_{i,h}\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\theta}_h(t_{i,h})\big)\Big] = -\boldsymbol{A}(\boldsymbol{x})\boldsymbol{z}, \qquad \text{uniformly on } \boldsymbol{\mathcal{N}}_2(\eta), \tag{B.5}$$

where $\boldsymbol{A}(\boldsymbol{x})$ is defined in Assumption FC.5.

For the third term in (B.2), note that $\Delta\boldsymbol{\theta}_h^\star(t_{i+1,h}) = \boldsymbol{\iota}_h\big(\boldsymbol{\psi}_h(t_{i+1,h})\big) - \boldsymbol{\iota}_h\big(\boldsymbol{\psi}_h(t_{i,h})\big)$. Recall $\boldsymbol{\iota}_{h,j}(\cdot)$ is the j_{th} element of $\boldsymbol{\iota}_h(\cdot)$. For $j = 1, \dots, k_\theta$, using a Taylor expansion of $\boldsymbol{\iota}_{h,j}\big(\boldsymbol{\psi}_h(t_{i+1,h})\big)$ around $\boldsymbol{\iota}_{h,j}\big(\boldsymbol{\psi}_h(t_{i,h})\big)$, the j_{th} element of $\mathbb{E}_{i,h}\big[\Delta\boldsymbol{\theta}_h^\star(t_{i+1,h})\big]$ can be written as

$$\mathbb{E}_{i,h}\left[\left(\frac{\partial\boldsymbol{\iota}_{h,j}\big(\boldsymbol{\psi}_h(t_{i,h})\big)}{\partial\boldsymbol{\psi}^\top}\right)\Delta\boldsymbol{\psi}_h(t_{i+1,h})\right] + \frac{1}{2}\mathbb{E}_{i,h}\left[\big(\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big)^\top\left(\frac{\partial^2\boldsymbol{\iota}_{h,j}\big(\bar{\boldsymbol{\psi}}_h(t_{i,h})\big)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}\right)\big(\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big)\right],$$

where $\bar{\boldsymbol{\psi}}_h(t_{i,h})$ lies on the segment joining $\boldsymbol{\psi}_h(t_{i+1,h})$ and $\boldsymbol{\psi}_h(t_{i,h})$. Note that, by Assumptions FC.1

and FC.4,

$$\left\| h^{-1/2-\kappa}\mathbb{E}_{i,h}\left[\left(\frac{\partial \boldsymbol{\iota}_{h,j}\big(\boldsymbol{\psi}_h(t_{i,h})\big)}{\partial \boldsymbol{\psi}^\top}\right)\Delta\boldsymbol{\psi}_h(t_{i+1,h})\right]\right\| = \left\| h^{1/2-\kappa}\left(\frac{\partial \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^\top}\right)\boldsymbol{a}_h(\boldsymbol{\psi})\right\| = o(1), \tag{B.6}$$

uniformly on $\boldsymbol{\mathcal{N}}_2(\eta)$. Recall $\|\boldsymbol{A}^\top\boldsymbol{B}\boldsymbol{A}\| \le \|\boldsymbol{B}\|\,\|\boldsymbol{A}^\top\boldsymbol{A}\|$ for any symmetric matrix $\boldsymbol{B}$. Since $\|\bar{\boldsymbol{\psi}}_h(t_{i,h})-\boldsymbol{\psi}_h(t_{i,h})\| \le \|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|$, $\forall \eta > 0$,

$$\begin{aligned}
&\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\| h^{-1/2-\kappa}\mathbb{E}_{i,h}\left[\big(\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big)^\top\left(\frac{\partial^2 \boldsymbol{\iota}_{h,j}\big(\bar{\boldsymbol{\psi}}_h(t_{i,h})\big)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\top}\right)\big(\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big)\right]\right\| \\
&\le \sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} h^{-1/2-\kappa}\mathbb{E}_{i,h}\left\{\left\|\frac{\partial^2 \boldsymbol{\iota}_{h,j}\big(\bar{\boldsymbol{\psi}}_h(t_{i,h})\big)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\top}\right\|\big\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big\|^2\right\} \\
&\le \sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} h^{-1/2-\kappa}\mathbb{E}_{i,h}\left\{\left\|\frac{\partial^2 \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\top}\right\|\big\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big\|^2 + V_{h,j}\big(\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|\big)\big\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big\|^2\right\} \\
&\le \sup_{\|\boldsymbol{\psi}\|\le\eta} h^{1/2-\kappa}\left\|\frac{\partial^2 \boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\top}\right\| + \sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} h^{-1/2-\kappa}\mathbb{E}_{i,h}\left[V_{h,j}\big(\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|\big)\big\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big\|^2\right] \\
&= o(1) + \sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} h^{-1/2-\kappa}\mathbb{E}_{i,h}\left[V_{h,j}\big(\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|\big)\big\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big\|^2\right],
\end{aligned}$$

where the second inequality and the last step follow from (4.6) and (4.5) in Assumption FC.4, respectively. Moreover, for any random variable X, we note that $\big(\mathbb{E}_{i,h}|X|^r\big)^{1/r} \le \big(\mathbb{E}_{i,h}|X|^s\big)^{1/s}$, $s > r > 0$, by the conditional Hölder inequality. Therefore, by (4.7) in Assumption FC.4,

$$\begin{aligned}
&\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} h^{-1/2-\kappa}\mathbb{E}_{i,h}\left[V_{h,j}\big(\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|\big)\big\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big\|^2\right] \\
&\qquad\le \left\{\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} h^{-1-2\kappa}\mathbb{E}_{i,h}\left[V_{h,j}^2\big(\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|\big)\big\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big\|^4\right]\right\}^{1/2} = o(1).
\end{aligned} \tag{B.7}$$

Combining the results above, we obtain

$$\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\| h^{-1/2-\kappa}\mathbb{E}_{i,h}\left[\big(\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big)^\top\left(\frac{\partial^2 \boldsymbol{\iota}_{h,j}\big(\bar{\boldsymbol{\psi}}_h(t_{i,h})\big)}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\top}\right)\big(\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big)\right]\right\| = o(1). \tag{B.8}$$

By (B.6) - (B.8), we have $h^{-1/2-\kappa}\mathbb{E}_{i,h}\big[\Delta\boldsymbol{\theta}_h^\star(t_{i+1,h})\big] = o(1)$ uniformly on $\boldsymbol{\mathcal{N}}_2(\eta)$ for every $\eta > 0$.

Combining (B.2) - (B.8), for every $\eta > 0$, we obtain

$$\lim_{h\downarrow 0} h^{-\delta}\mathbb{E}_{i,h}\big[\Delta\boldsymbol{z}_h(t_{i+1,h})\big] = -\boldsymbol{A}(\boldsymbol{x})\boldsymbol{z}, \qquad \text{uniformly on } \boldsymbol{\mathcal{N}}_2(\eta). \tag{B.9}$$

Let $\boldsymbol{c}(\boldsymbol{x},\boldsymbol{z}) = -\boldsymbol{A}(\boldsymbol{x})\boldsymbol{z}$. Clearly, $\boldsymbol{c}(\boldsymbol{x},\boldsymbol{0}) = \boldsymbol{0}$ for all $\boldsymbol{x} \in \mathbb{R}^n$, and thus Eq. (A.7) follows.

We move on to checking Eq. (A.8) in Assumption A5. Assume $\kappa < 1/4$ now. It suffices to show:

$$\lim_{h\downarrow 0} \sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} \left\| h^{-1/2-2\kappa}\mathbb{E}_{i,h}\left[\left(\Delta\boldsymbol{\theta}_h(t_{i+1,h})\right)\left(\Delta\boldsymbol{\theta}_h(t_{i+1,h})\right)^\top\right]\right\| = \mathbf{0}, \tag{B.10}$$

$$\lim_{h\downarrow 0} \sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} \left\| h^{-1/2-2\kappa}\mathbb{E}_{i,h}\left[\left(\Delta\boldsymbol{\theta}^\star_h(t_{i+1,h})\right)\left(\Delta\boldsymbol{\theta}^\star_h(t_{i+1,h})\right)^\top\right]\right\| = \mathbf{0}, \tag{B.11}$$

$$\lim_{h\downarrow 0} \sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} \left\| h^{-1/2-2\kappa}\mathbb{E}_{i,h}\left[\left(\Delta\boldsymbol{\theta}_h(t_{i+1,h})\right)\left(\Delta\boldsymbol{\theta}^\star_h(t_{i+1,h})\right)^\top\right]\right\| = \mathbf{0}, \tag{B.12}$$

for every $\eta > 0$. Note that by the c_r-inequality and Assumption FC.4,

$$\begin{aligned}\left\| h^{-1/2-2\kappa}\mathbb{E}_{i,h}\left[\left(\Delta\boldsymbol{\theta}_h(t_{i+1,h})\right)\left(\Delta\boldsymbol{\theta}_h(t_{i+1,h})\right)^\top\right]\right\| &\leq h^{-1/2-2\kappa}\mathbb{E}_{i,h}\left[\left\|\Delta\boldsymbol{\theta}_h(t_{i+1,h})\right\|^2\right] \\ &\leq Ch^{1/2-2\kappa}\mathbb{E}_{i,h}\left[\left\|\boldsymbol{g}_h\left(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\theta}_h(t_{i,h})\right)\right\|^2\right] + o\left(h^{1/2}\right),\end{aligned} \tag{B.13}$$

where the $o\left(h^{1/2}\right)$-term is uniform on $\boldsymbol{\mathcal{N}}_2(\eta)$. Then Eq. (B.10) is immediate by Eqs. (B.13) and (4.10) in Assumption FC.5. Similarly, by the second-order Taylor expansion of $\boldsymbol{\iota}_h\left(\boldsymbol{\psi}_h(t_{i+1,h})\right)$ around $\boldsymbol{\iota}_h\left(\boldsymbol{\psi}_h(t_{i,h})\right)$, the c_r-inequality, and using similar arguments for (B.8), $\forall\eta > 0$, we have

$$\begin{aligned}\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} \left\| h^{-1/2-2\kappa}\mathbb{E}_{i,h}\left[\left(\Delta\boldsymbol{\theta}^\star_h(t_{i+1,h})\right)\left(\Delta\boldsymbol{\theta}^\star_h(t_{i+1,h})\right)^\top\right]\right\| &\leq C\sum_{j=1}^{k_\theta}\left\{\sup_{\|\boldsymbol{\psi}\|\leq\eta} h^{1/2-2\kappa}\left\|\frac{\partial\iota_{h,j}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^\top}\right\|^2\right. \\ \left. + \sup_{\|\boldsymbol{\psi}\|\leq\eta} h^{3/2-2\kappa}\left\|\frac{\partial^2\iota_{h,j}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}\right\|^2 + \sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} h^{-1/2-2\kappa}\mathbb{E}_{i,h}\left[V^2_{h,j}\left(\left\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\right\|\right)\left\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\right\|^4\right]\right\} &= o(1).\end{aligned} \tag{B.14}$$

We obtain (B.11). As a result, Eq. (B.12) immediately follows from (B.13) - (B.14) and the Cauchy-Schwarz inequality.

Finally, for every $\boldsymbol{x} \in \mathbb{R}^{k_\upsilon} \times \mathbb{R}^{k_\psi}$, $\boldsymbol{z} \in \mathbb{R}^{k_\theta}$, define the following ordinary differential equation:

$$\frac{d\boldsymbol{Z}(t,\boldsymbol{x},\boldsymbol{z})}{dt} = -\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Z}(t,\boldsymbol{x},\boldsymbol{z}), \qquad \boldsymbol{Z}(0,\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{z}. \tag{B.15}$$

Since all the eigenvalues of $\boldsymbol{A}(\boldsymbol{x})$ have strictly positive real parts in every bounded subset of $\mathbb{R}^{k_\upsilon} \times \mathbb{R}^{k_\psi}$ (Assumption FC.5), Assumption A6 is fulfilled, see e.g., Eq. (3.15) in Nelson (1992), Ethier and Nagylaki (1980, Remarks, p. 20).

III. Verify Assumption A7

We follow Nelson (1992, Proof of Theorem 3.1) and define $\phi(\boldsymbol{z}) = \boldsymbol{z}^\top\boldsymbol{z}\left[1 - \exp\left(-\boldsymbol{z}^\top\boldsymbol{z}\right)\right]$, $\omega(\boldsymbol{x}) =$

$\left(\boldsymbol{x}^\top \boldsymbol{x}\right)^{1/2}\left[1-\exp\left(-\boldsymbol{x}^\top \boldsymbol{x}\right)\right]$,

$$\rho_h(\boldsymbol{x},\boldsymbol{z})=\rho(\boldsymbol{x},\boldsymbol{z})=2+\phi(\boldsymbol{z})+\omega(\boldsymbol{x}). \tag{B.16}$$

Then Eq. (A.11) is immediate. Moreover, Part (a) in Assumption FC.3 directly implies Eq. (A.13). It suffices to show that there is a $\lambda(\eta,h)$ satisfying (A.12) such that for every $\eta>0$ and $h>0$,

$$\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\{h^{-1}\mathbb{E}_{i,h}\left[\omega\big(\boldsymbol{x}_h(t_{i+1,h})\big)-\omega(\boldsymbol{x})\right]-\lambda(\eta,h)\left[1+\omega(\boldsymbol{x})\right]\right\}\le 0. \tag{B.17}$$

$$\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\{h^{-1}\mathbb{E}_{i,h}\left[\phi\big(\boldsymbol{z}_h(t_{i+1,h})\big)-\phi(\boldsymbol{z})\right]-\lambda(\eta,h)\left[1+\phi(\boldsymbol{z})\right]\right\}\le 0, \tag{B.18}$$

Note that (B.17) and (B.18) would follow if for every $\eta>0$ and $h>0$, there exist $K_1(\eta,h), K_2(\eta,h)>0$ such that

$$\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\{h^{-1}\mathbb{E}_{i,h}\left[\omega\big(\boldsymbol{x}_h(t_{i+1,h})\big)-\omega(\boldsymbol{x})\right]-K_1(\eta,h)\right\}\le 0, \tag{B.19}$$

$$\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\{h^{-1}\mathbb{E}_{i,h}\left[\phi\big(\boldsymbol{z}_h(t_{i+1,h})\big)-\phi(\boldsymbol{z})\right]-K_2(\eta,h)\right\}\le 0. \tag{B.20}$$

Indeed, as $\phi(\cdot),\omega(\cdot)\ge 0$, (B.17) and (B.18) follow from (B.19) and (B.20) by letting $\lambda(\eta,h)=K_1(\eta,h)\vee K_2(\eta,h)$. If $\lambda(\eta,h)$ satisfies (A.12), then we obtain Assumption A7.

For (B.19), simple algebra implies

$$\sup_{\boldsymbol{x}\in\mathbb{R}^{k_y}\times\mathbb{R}^{k_\psi}}\left\|\frac{\partial\omega(\boldsymbol{x})}{\partial\boldsymbol{x}}\right\|=\sup_{\boldsymbol{x}\in\mathbb{R}^{k_y}\times\mathbb{R}^{k_\psi}}\left\|\frac{\boldsymbol{x}}{\sqrt{\boldsymbol{x}^\top\boldsymbol{x}}}\left[1-\exp\big(-\boldsymbol{x}^\top\boldsymbol{x}\big)\right]+2\boldsymbol{x}\sqrt{\boldsymbol{x}^\top\boldsymbol{x}}\exp\big(-\boldsymbol{x}^\top\boldsymbol{x}\big)\right\|\le C.$$

By a Taylor series and Eq. (B.1), we have

$$\begin{aligned}\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\|h^{-1}\mathbb{E}_{i,h}\left[\omega\big(\boldsymbol{x}_h(t_{i+1,h})\big)-\omega(\boldsymbol{x})\right]\right\|&=\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\|\frac{\partial\omega(\bar{\boldsymbol{x}}_h(t_{i,h}))}{\partial\boldsymbol{x}^\top}h^{-1}\mathbb{E}_{i,h}\left[\Delta\boldsymbol{x}_h(t_{i+1,h})\right]\right\|\\&\le C\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\|h^{-1}\mathbb{E}_{i,h}\left[\Delta\boldsymbol{x}_h(t_{i+1,h})\right]\right\|\\&\le C\left(1+\sup_{\boldsymbol{x}\in\boldsymbol{\mathcal{N}}_1(\eta)}\left\|\boldsymbol{\mu}(\boldsymbol{x})\right\|\right),\end{aligned}$$

as $h\downarrow 0$, where $\bar{\boldsymbol{x}}_h(t_{i,h})$ is on the line segment between $\boldsymbol{x}_h(t_{i+1,h})$ and $\boldsymbol{x}_h(t_{i,h})=\boldsymbol{x}$. Since $\boldsymbol{\mu}(\cdot)$ is uniformly bounded (Assumption FC.2), one can find a sufficiently large $K_1>0$, which is independent of η and h, such that (B.19) holds with $K_1(\eta,h)\equiv K_1$.

For (B.20), let $\bar{\boldsymbol{z}}_h(t_{i,h})$ lie on the segment joining $\boldsymbol{z}_h(t_{i,h}) = \boldsymbol{z}$ and $\boldsymbol{z}_h(t_{i+1,h})$, then

$$\begin{aligned}
&h^{-1}\mathbb{E}_{i,h}\big[\phi\big(\boldsymbol{z}_h(t_{i+1,h})\big) - \phi(\boldsymbol{z})\big] \\
&= h^{-1}\frac{\partial\phi(\boldsymbol{z})}{\partial\boldsymbol{z}^\top}\mathbb{E}_{i,h}\big[\Delta\boldsymbol{z}_h(t_{i+1,h})\big] + \frac{1}{2}h^{-1}\mathbb{E}_{i,h}\Big[R_{i,h}\big(\phi, \bar{\boldsymbol{z}}_h(t_{i,h})\big)\Big] \\
&= h^{\delta-1}\bigg\{\frac{\partial\phi(\boldsymbol{z})}{\partial\boldsymbol{z}^\top}h^{-\delta}\mathbb{E}_{i,h}\big[\Delta\boldsymbol{z}_h(t_{i+1,h})\big] + \frac{1}{2}h^{-\delta}\mathbb{E}_{i,h}\Big[R_{i,h}\big(\phi, \bar{\boldsymbol{z}}_h(t_{i,h})\big)\Big]\bigg\},
\end{aligned} \tag{B.21}$$

where $R_{i,h}\big(\phi, \bar{\boldsymbol{z}}_h(t_{i,h})\big) = \big[\Delta\boldsymbol{z}_h(t_{i+1,h})\big]^\top \dfrac{\partial^2\phi(\bar{\boldsymbol{z}}_h(t_{i,h}))}{\partial\boldsymbol{z}\partial\boldsymbol{z}^\top}\big[\Delta\boldsymbol{z}_h(t_{i+1,h})\big]$,

$$\begin{aligned}
\frac{\partial\phi(\boldsymbol{z})}{\partial\boldsymbol{z}} &= 2\boldsymbol{z}\left[1 + \big(\boldsymbol{z}^\top\boldsymbol{z} - 1\big)\exp\big(-\boldsymbol{z}^\top\boldsymbol{z}\big)\right], \\
\frac{\partial^2\phi(\boldsymbol{z})}{\partial\boldsymbol{z}\partial\boldsymbol{z}^\top} &= 2\boldsymbol{I}_{k_\theta}\left[1 + \big(\boldsymbol{z}^\top\boldsymbol{z} - 1\big)\exp\big(-\boldsymbol{z}^\top\boldsymbol{z}\big)\right] - 4\boldsymbol{z}\boldsymbol{z}^\top\left[\big(\boldsymbol{z}^\top\boldsymbol{z} - 2\big)\exp\big(-\boldsymbol{z}^\top\boldsymbol{z}\big)\right].
\end{aligned}$$

Before continuing, we point out two properties of the partial derivatives above. First, using

$$0 < 1 + \big(\boldsymbol{z}^\top\boldsymbol{z} - 1\big)\exp\big(-\boldsymbol{z}^\top\boldsymbol{z}\big) = \big[1 - \exp\big(-\boldsymbol{z}^\top\boldsymbol{z}\big)\big] + \boldsymbol{z}^\top\boldsymbol{z}\exp\big(-\boldsymbol{z}^\top\boldsymbol{z}\big) \leq C, \tag{B.22}$$

$\big\|\boldsymbol{z}\boldsymbol{z}^\top\big[(\boldsymbol{z}^\top\boldsymbol{z} - 2)\exp\big(-\boldsymbol{z}^\top\boldsymbol{z}\big)\big]\big\| \leq C$, it is not hard to obtain $\sup_{\boldsymbol{z}\in\mathbb{R}^{k_\theta}}\Big(\|\partial\phi(\boldsymbol{z})/\partial\boldsymbol{z}\|/\|\boldsymbol{z}\|\Big) \leq C$ and $\sup_{\boldsymbol{z}\in\mathbb{R}^{k_\theta}}\big\|\partial^2\phi(\boldsymbol{z})/\partial\boldsymbol{z}\partial\boldsymbol{z}^\top\big\| \leq C$. Second, note that, for any $x \geq 0$, we have $|1 - \exp(-x)| \leq |x|$, $\exp(-x) \leq 1$, and $|(x-2)\exp(-x)| \leq 2$. If $\|\boldsymbol{z}\| \to 0$, then

$$\begin{aligned}
\left\|\frac{\partial\phi(\boldsymbol{z})}{\partial\boldsymbol{z}}\right\| &\leq 2\|\boldsymbol{z}\|\Big(\Big|\boldsymbol{z}^\top\boldsymbol{z}\exp\big(-\boldsymbol{z}^\top\boldsymbol{z}\big)\Big| + \Big|1 - \exp\big(-\boldsymbol{z}^\top\boldsymbol{z}\big)\Big|\Big) \leq 4\|\boldsymbol{z}\|^3, \\
\left\|\frac{\partial^2\phi(\boldsymbol{z})}{\partial\boldsymbol{z}\partial\boldsymbol{z}^\top}\right\| &\leq 4\|\boldsymbol{z}\|^2 + 8\|\boldsymbol{z}\|^2 \leq 12\|\boldsymbol{z}\|^2.
\end{aligned} \tag{B.23}$$

Moreover, by similar steps for showing Eq. (A.8), for every $\eta > 0$, we have

$$\limsup_{h\downarrow 0}\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)}\left\|h^{-\delta}\mathbb{E}_{i,h}\Big\{\big[\Delta\boldsymbol{z}_h(t_{i+1,h})\big]^\top\big[\Delta\boldsymbol{z}_h(t_{i+1,h})\big]\Big\}\right\| = 0. \tag{B.24}$$

With a slight abuse of notation and utilizing the uniformly bounded property of $\partial^2\phi(\boldsymbol{z})/\partial\boldsymbol{z}\partial\boldsymbol{z}^\top$, Eqs. (B.9), (B.21), and (B.24), we arrive at

$$h^{-1}\mathbb{E}_{i,h}\big[\phi\big(\boldsymbol{z}_h(t_{i+1,h})\big) - \phi(\boldsymbol{z})\big] = h^{\delta-1}\left\{-\frac{\partial\phi(\boldsymbol{z})}{\partial\boldsymbol{z}^\top}\boldsymbol{A}(\boldsymbol{x})\boldsymbol{z} + o\left(\left\|\frac{\partial\phi(\boldsymbol{z})}{\partial\boldsymbol{z}}\right\|\right) + o\left(\left\|\frac{\partial^2\phi(\boldsymbol{z})}{\partial\boldsymbol{z}\partial\boldsymbol{z}^\top}\right\|\right)\right\}, \tag{B.25}$$

where the $o(\cdot)$-terms are uniform on $\boldsymbol{\mathcal{N}}_2(\eta)$. Clearly, (B.20) and (A.12) are immediate if $\|\boldsymbol{z}\| = 0$ (because $\phi(\boldsymbol{z}) \equiv 0$). We assume $\|\boldsymbol{z}\| \neq 0$ next. By (B.25) above, we observe two main cases for $h^{-1}\mathbb{E}_{i,h}\big[\phi\big(\boldsymbol{z}_h(t_{i+1,h})\big) - \phi(\boldsymbol{z})\big]$.

(i) For any $K > 0$ (possibly depending on η) and $\Delta > 0$, in neighborhoods of the form $0 < \|\boldsymbol{z}\| \leq Kh^{\Delta}$, the right-hand side (RHS) of Eq. (B.25) can be written as

$$h^{\delta-1}\Big\{-2\left[1+\left(\boldsymbol{z}^\top\boldsymbol{z}-1\right)\exp\left(-\boldsymbol{z}^\top\boldsymbol{z}\right)\right]\boldsymbol{z}^\top\boldsymbol{A}(\boldsymbol{x})\boldsymbol{z}+o(\|\boldsymbol{z}\|^3)+o(\|\boldsymbol{z}\|^2)\Big\}$$
$$= h^{\delta-1}\Big\{O(\|\boldsymbol{z}\|^2)+o(\|\boldsymbol{z}\|^3)+o(\|\boldsymbol{z}\|^2)\Big\}, \quad \text{(B.26)}$$

using the construction of $\partial\phi(\boldsymbol{z})/\partial\boldsymbol{z}$ and (B.23), where the $O(\|\boldsymbol{z}\|^2)$ term is strictly negative because $1+\left(\boldsymbol{z}^\top\boldsymbol{z}-1\right)\exp\left(-\boldsymbol{z}^\top\boldsymbol{z}\right) > 0$ as mentioned in (B.22). Clearly, the $O(\|\boldsymbol{z}\|^2)$ term is in the strict sense, i.e., it is not $o(\|\boldsymbol{z}\|^2)$. As such, it dominates the other terms asymptotically. Therefore, $h^{-1}\mathbb{E}_{i,h}\left[\phi\left(\boldsymbol{z}_h(t_{i+1,h})\right)-\phi(\boldsymbol{z})\right]$ is asymptotically negative (may be bounded or unbounded). In this case, there exists a sufficiently large $K_2 > 0$, which is independent of η and h, such that (B.20) holds with $K_2(\eta,h) \equiv K_2$. Clearly, $\lambda(\eta,h) = K_1(\eta,h) \vee K_2(\eta,h) \equiv K_1 \vee K_2$ fulfills (A.12).

(ii) Otherwise, the first term in the curly brackets on the RHS of (B.25) dominates the remaining terms, and as a result, $h^{-1}\mathbb{E}_{i,h}\left[\phi\left(\boldsymbol{z}_h(t_{i+1,h})\right)-\phi(\boldsymbol{z})\right]$ diverges to $-\infty$. The arguments in Part (i) follow immediately.

The verification of Assumptions A1 - A7 allows one to use Theorem A.1, yielding Theorem 1. ■

C Weak convergence

Proof of Theorem 2 Let $t \equiv T + \tau h^{1/2}$, without the dependence on (T, τ, h) for brevity in notation. Recall that $\widetilde{\boldsymbol{z}}_h(t) = h^{-\kappa}\left(\boldsymbol{\theta}_h(t)-\boldsymbol{\theta}_h^\star(t)\right)$ with $\kappa = 1/4$. Without confusion, let $\mathbb{E}_{t,h}$ ($\mathbb{C}\text{ov}_{t,h}$) be the expectation (covariance) conditional on $\left(\boldsymbol{x}_h(t), \widetilde{\boldsymbol{z}}_h(t)\right) = (\boldsymbol{x}, \boldsymbol{z})$. We shall apply Theorem A.2. As such, we verify Assumptions A8 - A9. The equations (A.18) - (A.19) in Assumption A8 are directly derived from the dynamics of by the dynamics of $\left\{\Delta\boldsymbol{\psi}_h(t+h)\right\}$ and (4.17) - (4.18). Furthermore, Eq. (B.9) continues to hold for $\widetilde{\boldsymbol{z}}_h(t)$ with minor modifications under Assumption AD.1 (subsequently, Assumptions FC.1, FC.4, FC.5). Hence, we have (A.20), and moreover,

$$\begin{aligned}
&h^{-1/2}\,\mathbb{C}\text{ov}_{t,h}\left(\Delta\widetilde{\boldsymbol{z}}_h(t+h)\right)\\
&= h^{-1/2}\mathbb{E}_{t,h}\Big[\left(\Delta\widetilde{\boldsymbol{z}}_h(t+h)\right)\left(\Delta\widetilde{\boldsymbol{z}}_h(t+h)\right)^\top\Big] - h^{1/2}\Big[h^{-1/2}\mathbb{E}_{t,h}\left(\Delta\widetilde{\boldsymbol{z}}_h(t+h)\right)\Big]\Big[h^{-1/2}\mathbb{E}_{t,h}\left(\Delta\widetilde{\boldsymbol{z}}_h(t+h)\right)^\top\Big]\\
&= h^{-1}\mathbb{E}_{t,h}\Big[\left(\Delta\boldsymbol{\theta}_h(t+h)-\Delta\boldsymbol{\theta}_h^\star(t+h)\right)\left(\Delta\boldsymbol{\theta}_h(t+h)-\Delta\boldsymbol{\theta}_h^\star(t+h)\right)^\top\Big] + O\left(h^{1/2}\right),
\end{aligned}$$

where the $O(\cdot)$-term is uniform on $\boldsymbol{\mathcal{N}}_2(\eta)$, $\eta > 0$. Next, we use the second-order Taylor expansions repeatedly to find the terms dominating the asymptotic order of $\Delta\boldsymbol{\theta}_h(t+h) - \Delta\boldsymbol{\theta}_h^\star(t+h)$. Note that, conditional on $\big(\boldsymbol{x}_h(t), \widetilde{\boldsymbol{z}}_h(t)\big) = (\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{\upsilon}, \boldsymbol{\psi}, \boldsymbol{z}) \in \boldsymbol{\mathcal{N}}_2(\eta)$,

$$h^{-1/2}\big(\Delta\boldsymbol{\theta}_h(t+h) - \Delta\boldsymbol{\theta}_h^\star(t+h)\big) =: \sum_{\ell=1}^{3} \boldsymbol{\varXi}_{h,\ell} + \widetilde{R}_{t,h}\big(\boldsymbol{g}_h, \boldsymbol{x}, \boldsymbol{z}\big) + \breve{R}_{t,h}\big(\boldsymbol{\iota}_h, \boldsymbol{x}, \boldsymbol{z}\big) + O\big(h^{1/2}\big), \quad \text{(C.1)}$$

where the $O(\cdot)$-term is uniform on $\boldsymbol{\mathcal{N}}_2(\eta)$,

$$\boldsymbol{\varXi}_{h,1} = \boldsymbol{g}_h\Big(\boldsymbol{y}_h(t+h), \boldsymbol{\theta}_h^\star(t)\Big) - \left(\frac{\partial \boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^\top}\right)\boldsymbol{B}_h(\boldsymbol{\psi})\boldsymbol{\eta}_{\lfloor t/h\rfloor+1},$$
$$\boldsymbol{\varXi}_{h,2} = h^{1/4}\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{g}_h\Big(\boldsymbol{y}_h(t+h), \boldsymbol{\theta}_h^\star(t)\Big)\boldsymbol{z}, \qquad \boldsymbol{\varXi}_{h,3} = -h^{1/2}\frac{\partial \boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^\top}\boldsymbol{a}_h(\boldsymbol{\psi}).$$

By Eq. (4.20) and Assumption FC.4(a), respectively, we have $\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} \mathbb{E}_{t,h}\big\|\boldsymbol{\varXi}_{h,2}\big\|^2 = o(1)$ and $\sup_{\|\boldsymbol{\psi}\|\le\eta}\big\|\boldsymbol{\varXi}_{h,3}\big\|^2 = o(h^{1/2})$. Moreover, the first remainder term is bounded as, for some $K_{1,\eta}, K_{2,\eta} > 0$,

$$\Big\|\widetilde{R}_{t,h}\big(\boldsymbol{g}_h, \boldsymbol{x}, \boldsymbol{z}\big)\Big\| \le K_{1,\eta}\sum_{j=1}^{k_\theta}\left\{h^{1/2}\sup_{\|\boldsymbol{\theta}\|\le K_{2,\eta}}\left\|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\boldsymbol{g}_{h,j}\big(\boldsymbol{y}_h(t+h), \boldsymbol{\theta}\big)\right\|\right\}.$$

Then AD.1(c) implies $\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} \mathbb{E}_{t,h}\Big\|\widetilde{R}_{t,h}\big(\boldsymbol{g}_h, \boldsymbol{x}, \boldsymbol{z}\big)\Big\|^2 = o(h^{1/2})$.

For the second remainder term, there are two cases. If $\partial\boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}^\top$ does not depend on $\boldsymbol{\psi}$, and thus $\partial^2\boldsymbol{\iota}_{h,j}(\cdot)/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top \equiv 0$, we have $\big\|\breve{R}_{t,h}\big(\boldsymbol{\iota}_h, \boldsymbol{x}, \boldsymbol{z}\big)\big\| \equiv 0$, and thus, $\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} \mathbb{E}_{t,h}\big\|\breve{R}_{t,h}\big(\boldsymbol{\iota}_h, \boldsymbol{x}, \boldsymbol{z}\big)\big\|^2 \equiv 0$ for any $\varphi_1 > 0$, where φ_1 is given in Assumption AD.1(a). Otherwise, it is bounded as

$$\Big\|\breve{R}_{t,h}\big(\boldsymbol{\iota}_h, \boldsymbol{x}, \boldsymbol{z}\big)\Big\| \le C\sum_{j=1}^{k_\theta}\left\{h^{1/2}\left\|\frac{\partial^2\boldsymbol{\iota}_{h,j}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}\right\|\big\|\boldsymbol{\eta}_{\lfloor t/h\rfloor+1}\big\|^2 + h^{-1/2}V_{h,j}\big(\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\|\big)\big\|\Delta\boldsymbol{\psi}_h(t_{i+1,h})\big\|^2\right\}.$$

In this case, we require $\varphi_1 \ge 2$ to ensure the existence of $\mathbb{E}_{t,h}\big\|\boldsymbol{\eta}_{\lfloor t/h\rfloor+1}\big\|^4$. For $\varphi_1 \ge 2$, by Assumption AD.1(b) , we obtain $\sup_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{\mathcal{N}}_2(\eta)} \mathbb{E}_{t,h}\big\|\breve{R}_{t,h}\big(\boldsymbol{\iota}_h, \boldsymbol{x}, \boldsymbol{z}\big)\big\|^2 = o(h^{1/2}) + o(h^{1/2})$. Combining the results above into (C.1) and using conditional Cauchy-Schwarz inequality, we obtain $\lim_{h\downarrow 0} h^{-1/2}\,\mathbb{C}\mathrm{ov}_{t,h}\big(\Delta\widetilde{\boldsymbol{z}}_h(t+h)\big) = \boldsymbol{\varSigma}(\boldsymbol{x})$ uniformly on $\boldsymbol{\mathcal{N}}_2(\eta)$. Therefore, Eq. (A.21) in Assumption A8 is satisfied.

Now we check Assumption A9. Recall φ_j, $j = 2, 3, 4$, from Assumptions AD.1(c), AD.2, respectively, and let $\varphi = \min\{\varphi_1/2 - 1, \varphi_2, \varphi_3, \varphi_4\}$. By Assumptions AD.1(a) and (4.19), (A.22) is immediate. Moreover, by the c_r-inequality and Assumption AD.1(c), (A.23) can be written as

$$h^{-1/2}\mathbb{E}_{t,h}\Big[\big\|\Delta\widetilde{\boldsymbol{z}}_h(t+h)\big\|^{2+\varphi}\Big] = h^{-(1+\varphi/4)}\mathbb{E}_{t,h}\Big[\big\|\Delta\boldsymbol{\theta}_h^\star(t+h)\big\|^{2+\varphi}\Big] + O(h^{\varphi/4}), \quad \text{(C.2)}$$

where the $O(\cdot)$-term is uniform on $\boldsymbol{\mathcal{N}}_2(\eta)$. By Assumption AD.1(b) and employing similar reasoning as discussed earlier, we can conclude that the first component in (C.2) is $o(1)$, which holds uniformly on $\boldsymbol{\mathcal{N}}_2(\eta)$. After verifying (A.23), we proceed to apply Theorem A.2, which yields the pointwise limiting distribution given by (4.22). ■

D Optimality

Proof of Theorem 3 As in Nelson (1996, Proof of Theorem 2.2), we first guess a solution and then verify its global optimality. We use $\mathbb{E}_{i,h}[\,\cdot\,] = \mathbb{E}\big[\,\cdot\,\big|\,\boldsymbol{x}_h(t_{i,h}) = \boldsymbol{x} = (\boldsymbol{v},\boldsymbol{\psi})\big]$ in the current proof with a slight abuse of notation. Recall $\tilde{\boldsymbol{\iota}}_h(\cdot) = \big(\boldsymbol{\iota}_h(\cdot), \boldsymbol{\iota}_h^{\dagger}(\cdot)\big)^{\top}$ in Eq. (4.26) and let $\tilde{\boldsymbol{\theta}} = \big(\boldsymbol{\theta}^{\star}, \boldsymbol{\theta}^{\dagger}\big)^{\top} = \tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})$. Before continuing, the following two identities, similar to Lemma A.1 and Eq. (A.7) in Nelson (1996), are useful:

$$\mathbb{E}_{i,h}\left[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)^{\top}\left(\frac{\partial\boldsymbol{\iota}_h(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^{\top}}\right)\boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{\eta}_{i+1}\right] = \mathbb{E}_{i,h}\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)^{\top}\boldsymbol{P}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{v},\boldsymbol{\psi}\big)\Big], \tag{D.1}$$

and

$$-\mathbb{E}_{i,h}\left[\frac{\partial}{\partial\boldsymbol{\theta}^{\star\top}}\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)\right] = \mathbb{E}_{i,h}\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\psi}\big)^{\top}\Big]\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)^{\top}, \tag{D.2}$$

where $\partial\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)/\partial\boldsymbol{\theta}^{\star\top} = \partial\boldsymbol{g}_h\Big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\iota}_h(\boldsymbol{\psi}),\boldsymbol{\iota}_h^{\dagger}(\boldsymbol{\psi})\Big)/\partial\boldsymbol{\theta}^{\star\top}$ denotes the partial derivative w.r.t. the second vector component $\boldsymbol{\iota}_h(\cdot)$. Moreover, $\boldsymbol{P}_h(\cdot)$, $\boldsymbol{S}_h(\cdot)$, and $\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)$, are defined in Eqs. (4.28), (4.27), and (4.31), respectively. The first identity (D.1) is a straightforward result from the law of iterated expectation for nested subfields (Davidson, 1994, Theorem 10.26, p. 155). To see the second identity (D.2), we note that by the condition (4.29) and $\boldsymbol{\psi} = \tilde{\boldsymbol{\iota}}_h^{-1}\big(\tilde{\boldsymbol{\theta}}\big)$ (Assumption FO.4),

$$\begin{aligned}
\boldsymbol{0} &= \frac{\partial}{\partial\boldsymbol{\theta}^{\star\top}}\mathbb{E}_{i,h}\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)\Big] \\
&= \frac{\partial}{\partial\boldsymbol{\theta}^{\star\top}}\int \boldsymbol{g}_h\big(\boldsymbol{y},\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)q_h(\boldsymbol{y};\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{y} \\
&= \int\left[\frac{\partial\boldsymbol{g}_h\big(\boldsymbol{y},\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)}{\partial\boldsymbol{\theta}^{\star\top}}q_h(\boldsymbol{y};\boldsymbol{\psi}) + \boldsymbol{g}_h\big(\boldsymbol{y},\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)\frac{\partial q_h(\boldsymbol{y};\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^{\top}}\frac{\partial\tilde{\boldsymbol{\iota}}_h^{-1}\big(\tilde{\boldsymbol{\theta}}\big)}{\partial\boldsymbol{\theta}^{\star\top}}\right]\mathrm{d}\boldsymbol{y} \\
&= \left[\int\frac{\partial\boldsymbol{g}_h\big(\boldsymbol{y},\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)}{\partial\boldsymbol{\theta}^{\star\top}}q_h(\boldsymbol{y};\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{y}\right] + \left[\int\boldsymbol{g}_h\big(\boldsymbol{y},\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)\boldsymbol{S}_h(\boldsymbol{y},\boldsymbol{\psi})^{\top}q_h(\boldsymbol{y};\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{y}\right]\frac{\partial\tilde{\boldsymbol{\iota}}_h^{-1}\big(\tilde{\boldsymbol{\theta}}\big)}{\partial\boldsymbol{\theta}^{\star\top}} \\
&= \mathbb{E}_{i,h}\left[\frac{\partial}{\partial\boldsymbol{\theta}^{\star\top}}\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)\right] + \mathbb{E}_{i,h}\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}),\boldsymbol{\psi}\big)^{\top}\Big]\frac{\partial\tilde{\boldsymbol{\iota}}_h^{-1}\big(\tilde{\boldsymbol{\theta}}\big)}{\partial\boldsymbol{\theta}^{\star\top}}.
\end{aligned}$$

Recall the definition of $\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)$ in Eq. (4.31). We obtain Eq. (D.2) by subtracting the first term from both sides.

I. The guess step

Let $\boldsymbol{L}_{\boldsymbol{\psi},h} = \partial \boldsymbol{\iota}_h(\boldsymbol{\psi})/\partial \boldsymbol{\psi}^\top$. Without taking $\lim_{h\downarrow 0}$, we similarly define

$$\begin{aligned}
\boldsymbol{A}_h(\boldsymbol{g}_h) &= -\mathbb{E}_{i,h}\left[\frac{\partial}{\partial \boldsymbol{\theta}^{\star\top}} \boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)\right],\\
\boldsymbol{\Sigma}_h(\boldsymbol{g}_h) &= \mathbb{E}_{i,h}\left\{\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big) - \boldsymbol{L}_{\boldsymbol{\psi},h}\boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{\eta}_{i+1}\Big]\Big[\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big) - \boldsymbol{L}_{\boldsymbol{\psi},h}\boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{\eta}_{i+1}\Big]^\top\right\},
\end{aligned}$$

with explicit dependence on $\boldsymbol{g}_h$. For any arbitrarily small $h > 0$, the optimal $\boldsymbol{g}_h$ must fulfill the matrix Riccati equation as in (4.25):

$$\boldsymbol{A}_h(\boldsymbol{g}_h)\boldsymbol{V}_h(\boldsymbol{g}_h) + \boldsymbol{V}_h(\boldsymbol{g}_h)\boldsymbol{A}_h(\boldsymbol{g}_h)^\top = \boldsymbol{\Sigma}_h(\boldsymbol{g}_h), \tag{D.3}$$

where $\boldsymbol{V}_h(\boldsymbol{g}_h)$ is the resulting solution (for $\boldsymbol{x} = (\boldsymbol{\upsilon}, \boldsymbol{\psi})$ given). Taking the trace on both sides of the matrix Riccati equation, and interchanging it with expectations, yields

$$\begin{aligned}
0 &= \operatorname{tr}\big(\boldsymbol{\Sigma}_h(\boldsymbol{g}_h)\big) - 2\operatorname{tr}\big(\boldsymbol{A}_h(\boldsymbol{g}_h)\boldsymbol{V}_h(\boldsymbol{g}_h)\big)\\
&= \mathbb{E}_{i,h}\Big\{\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)^\top \boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big) - 2\boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)^\top \boldsymbol{P}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\upsilon}, \boldsymbol{\psi}\big)\\
&\qquad + \boldsymbol{\eta}_{i+1}^\top \boldsymbol{B}(\boldsymbol{\psi})^\top \boldsymbol{L}_{\boldsymbol{\psi},h}^\top \boldsymbol{L}_{\boldsymbol{\psi},h}\boldsymbol{B}(\boldsymbol{\psi})\boldsymbol{\eta}_{i+1} - 2\Big[\boldsymbol{V}_h(\boldsymbol{g}_h)\boldsymbol{Q}_h(\tilde{\boldsymbol{\theta}})\boldsymbol{S}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \boldsymbol{\psi}\big)\Big]^\top \boldsymbol{g}_h\big(\boldsymbol{y}_h(t_{i+1,h}), \tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big)\Big\},
\end{aligned}$$

using the identities (D.1) - (D.2). Then we differentiate both sides w.r.t. $\boldsymbol{g}_h(\,\cdot\,,\,\cdot\,)$, treating $\boldsymbol{g}_h(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$ as a separate choice variable for each $(\boldsymbol{y}, \tilde{\boldsymbol{\theta}})$. It implies that the optimal $\boldsymbol{g}_h$, for $\boldsymbol{x} = (\boldsymbol{\upsilon}, \boldsymbol{\psi})$ given, takes the following form:

$$\boldsymbol{g}_h\big(\,\cdot\,, \boldsymbol{\upsilon}, \tilde{\boldsymbol{\iota}}_h(\boldsymbol{\psi})\big) = \boldsymbol{P}_h\big(\,\cdot\,, \boldsymbol{\upsilon}, \boldsymbol{\psi}\big) + \boldsymbol{V}_h(\boldsymbol{\upsilon}, \boldsymbol{\psi})\boldsymbol{Q}_h\big(\tilde{\boldsymbol{\theta}}\big)\boldsymbol{S}_h\big(\,\cdot\,, \boldsymbol{\psi}\big), \tag{D.4}$$

where, for every h, $\boldsymbol{V}_h$ solely depends on $(\boldsymbol{\upsilon}, \boldsymbol{\psi})$ by reparametrization. Substituting $\boldsymbol{g}_h$ in (D.4) back to the matrix Riccati equation (D.3) implies $\boldsymbol{V}_h(\boldsymbol{\upsilon}, \boldsymbol{\psi})$ is the solution to Eq. (4.30) by routine algebra and using a similar argument for (D.1).

II. Global optimality

We drop all function arguments when no confusion is caused. We consider another choice of $\boldsymbol{g}_h$ that satisfies the assumptions in Theorem 2, say $\tilde{\boldsymbol{g}}_h = \boldsymbol{P}_h + \boldsymbol{W}_h\boldsymbol{Q}_h\boldsymbol{S}_h + \boldsymbol{H}_h$, where $\boldsymbol{H}_h$ is a vector-valued function with the same arguments of $\boldsymbol{P}_h$ and $\boldsymbol{S}_h$. As required in Assumption FC.5, we assume that all the eigenvalues of $\boldsymbol{A}_h(\tilde{\boldsymbol{g}}_h)$ have strictly positive real parts so that there is a bounded $\boldsymbol{V}_h(\tilde{\boldsymbol{g}}_h)$ that is the asymptotic covariance matrix of the filter errors (Lancaster and Tismenetsky, 1985, Chapter 12.3, Theorem 3). By the matrix Riccati equation and the definition of $\boldsymbol{\Sigma}_h(\cdot)$, we have

$$\boldsymbol{A}_h(\tilde{\boldsymbol{g}}_h)\boldsymbol{V}_h(\tilde{\boldsymbol{g}}_h) + \boldsymbol{V}_h(\tilde{\boldsymbol{g}}_h)\boldsymbol{A}_h(\tilde{\boldsymbol{g}}_h)^\top = \boldsymbol{\Sigma}_h(\tilde{\boldsymbol{g}}_h), \tag{D.5}$$

where

$$\boldsymbol{\Sigma}_h(\tilde{\boldsymbol{g}}_h) = \mathbb{E}_{i,h}\Big[\big(\boldsymbol{P}_h + \boldsymbol{W}_h\boldsymbol{Q}_h\boldsymbol{S}_h + \boldsymbol{H}_h - \boldsymbol{L}_{\psi,h}\boldsymbol{B}\boldsymbol{\eta}_{i+1}\big)\big(\boldsymbol{P}_h + \boldsymbol{W}_h\boldsymbol{Q}_h\boldsymbol{S}_h + \boldsymbol{H}_h - \boldsymbol{L}_{\psi,h}\boldsymbol{B}\boldsymbol{\eta}_{i+1}\big)^\top\Big].$$

Similarly, the construction in (D.4) gives (D.3) with

$$\boldsymbol{\Sigma}_h(\boldsymbol{g}_h) = \mathbb{E}_{i,h}\Big[\big(\boldsymbol{P}_h + \boldsymbol{W}_h\boldsymbol{Q}_h\boldsymbol{S}_h - \boldsymbol{L}_{\psi,h}\boldsymbol{B}\boldsymbol{\eta}_{i+1}\big)\big(\boldsymbol{P}_h + \boldsymbol{W}_h\boldsymbol{Q}_h\boldsymbol{S}_h - \boldsymbol{L}_{\psi,h}\boldsymbol{B}\boldsymbol{\eta}_{i+1}\big)^\top\Big].$$

Note that by the law of iterated expectations, we have

$$\mathbb{E}_{i,h}\Big[\big(\boldsymbol{P}_h - \boldsymbol{L}_{\psi,h}\boldsymbol{B}\boldsymbol{\eta}_{i+1}\big)\boldsymbol{H}_h^\top\Big] = \boldsymbol{0}. \tag{D.6}$$

With (D.6), subtracting (D.3) from (D.5) and simplifying the expressions leads to

$$\boldsymbol{A}_h(\tilde{\boldsymbol{g}}_h)\big(\boldsymbol{V}_h(\tilde{\boldsymbol{g}}_h) - \boldsymbol{V}_h(\boldsymbol{g}_h)\big) + \big(\boldsymbol{V}_h(\tilde{\boldsymbol{g}}_h) - \boldsymbol{V}_h(\boldsymbol{g}_h)\big)\boldsymbol{A}_h(\tilde{\boldsymbol{g}}_h)^\top = \mathbb{E}_{i,h}\big(\boldsymbol{H}_h\boldsymbol{H}_h^\top\big). \tag{D.7}$$

The same arguments below (A.10) in Nelson (1996) implies that $\boldsymbol{V}_h(\tilde{\boldsymbol{g}}_h) - \boldsymbol{V}_h(\boldsymbol{g}_h)$ is positive semidefinite, and $\boldsymbol{V}_h(\tilde{\boldsymbol{g}}_h) = \boldsymbol{V}_h(\boldsymbol{g}_h)$ if and only if $\mathbb{E}_{i,h}\big(\boldsymbol{H}_h\boldsymbol{H}_h^\top\big) = \boldsymbol{0}$. That is, $\boldsymbol{H}_h \equiv \boldsymbol{0}$ and thus $\tilde{\boldsymbol{g}}_h = \boldsymbol{g}_h$. ■

E Patton filter correction for Clayton copula

Recall $c(\boldsymbol{y};\theta)$ in Eq. (5.4), where $\boldsymbol{y} = (y_1, y_2)$. To demean the Patton filter, we need to compute the conditional expectation of $|y_{1h}(t_{i+1,h}) - y_{2h}(t_{i+1,h})|$. For this, we first realize that due to the exchangeability of the Clayton copula we have that

$$\begin{aligned}\int_0^1\int_0^1 |y_1 - y_2|\ c(\boldsymbol{y};\theta)\,\mathrm{d}y_1\,\mathrm{d}y_2 &= \int_0^1\int_0^{y_1} y_1\ c(\boldsymbol{y};\theta)\,\mathrm{d}y_2\,\mathrm{d}y_1 - \int_0^1\int_{y_2}^1 y_2\ c(\boldsymbol{y};\theta)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &\quad - \int_0^1\int_{y_1}^1 y_1\ c(\boldsymbol{y};\theta)\,\mathrm{d}y_2\,\mathrm{d}y_1 + \int_0^1\int_0^{y_2} y_2\ c(\boldsymbol{y};\theta)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &= 2\int_0^1\int_0^{y_2} y_2\ c(\boldsymbol{y};\theta)\,\mathrm{d}y_1\,\mathrm{d}y_2 - 2\int_0^1\int_{y_2}^1 y_2\ c(\boldsymbol{y};\theta)\,\mathrm{d}y_1\,\mathrm{d}y_2.\end{aligned}$$

Note that

$$\int c(\boldsymbol{y};\theta)\mathrm{d}y_1 = (y_2)^{-\theta-1}\Big((y_1)^{-\theta}+(y_2)^{-\theta}-1\Big)^{-1-1/\theta} = \left(\frac{y_1}{y_2}\right)^{\theta+1}\Big(1+(y_1)^{\theta}\left((y_2)^{-\theta}-1\right)\Big)^{-1-1/\theta}.$$

We then get $\int_0^{y_2} y_2\; c(\boldsymbol{y};\theta)\,\mathrm{d}y_1 = y_2\,\left(2-(y_2)^{\theta}\right)^{-1-1/\theta}$ and $\int_{y_2}^1 y_2\; c(\boldsymbol{y};\theta)\,\mathrm{d}y_1 = y_2 - y_2\,\left(2-(y_2)^{\theta}\right)^{-1-1/\theta}$. Taking these results together, we arrive at

$$\int_0^1\int_0^1 |y_1-y_2|\; c(\boldsymbol{y};\theta)\,\mathrm{d}y_1\,\mathrm{d}y_2 = 4\int_0^1 y_2\,\Big(2-(y_2)^{\theta}\Big)^{-1-1/\theta}\mathrm{d}y_2 - 1.$$

Finally, note that

$$\begin{aligned}
\int_0^1 y_2\,\Big(2-(y_2)^{\theta}\Big)^{-1-1/\theta}\mathrm{d}y_2 \quad &= \quad \int_0^1 y_2\;2^{-1-1/\theta}\Big(1-2^{-1}(y_2)^{\theta}\Big)^{-1-1/\theta}\mathrm{d}y_2\\
&= \int_0^1 y_2\;2^{-1-1/\theta}\sum_{k=0}^{\infty}\left(\theta^{-1}+1\right)_k\frac{2^{-k}(y_2)^{k\theta}}{k!}\,\mathrm{d}y_2\\
&= 2^{-1-1/\theta}\sum_{k=0}^{\infty}\left(\theta^{-1}+1\right)_k\frac{2^{-k}}{k!}\int_0^1 (y_2)^{k\theta+1}\,\mathrm{d}y_2\\
&= 2^{-1-1/\theta}\sum_{k=0}^{\infty}\frac{\left(\theta^{-1}+1\right)_k}{\left(k+2\theta^{-1}\right)}\;\theta^{-1}\;\frac{2^{-k}}{k!}\\
&= \frac{2^{-1-1/\theta}}{\theta}\sum_{k=0}^{\infty}\frac{\left(\theta^{-1}+1\right)_k\;\left(2\theta^{-1}\right)_k}{2\theta^{-1}\;\left(1+2\theta^{-1}\right)_k}\;\frac{2^{-k}}{k!}\\
&= 2^{-2-1/\theta}\;{}_2F_1\Big(1+\theta^{-1}\;,\;2\theta^{-1}\;;\;1+2\theta^{-1}\;;\;2^{-1}\Big),
\end{aligned}$$

where $(a)_k = a(a+1)\cdots(a+k-1)$ denotes the Pochhammer symbol, and where we used $(1+z)^{-a} = \sum_{k=0}^{\infty}(a)_k z^k/k!$ for $|z|<1$ and $a>0$. ■

F Helland's result for weak convergence

We first give the part a) of Lemma 5.2 in Helland (1982).

Lemma F.1 (Helland, 1982)

Let Y, $(Y_n)_{n\geq1}$ and $(Y_{n,k})_{n\geq1,k\geq1}$ be random variables. Suppose that

(i) $Y_{n,k}\xrightarrow{p}Y_n$ as $k\to\infty$ for each n;

(ii) $Y_n\xrightarrow{p}Y$ as $n\to\infty$.

Then for any increasing sequence (k_n) such that $k_n\to\infty$ fast enough we have $Y_{n,k_n}\xrightarrow{p}Y$.

Clearly, for the claim made in the main text we need Helland's result with convergence in probability replaced by convergence in distribution in the assumptions as well as in the conclusion. Because we need one additional assumption we state the result in Helland (1982) with convergence in probability replaced by weak convergence explicitly.

Lemma F.2

Let Y, $(Y_n)_{n\geq 1}$ and $(Y_{n,k})_{n\geq 1,k\geq 1}$ be random variables. Suppose that

(i) $Y_{n,k} \overset{d}{\to} Y_n$ *as* $k \to \infty$ *for each* n*;*

(ii) $Y_n \overset{d}{\to} Y$ *as* $n \to \infty$*;*

(iii) $Y_1, Y_2, \ldots$ *and* Y *have a continuous distribution.*

Then there exists an increasing sequence (k_n) *such that* $k_n \to \infty$ *and we have* $Y_{n,k_n} \overset{d}{\to} Y$*.*

Proof We first show that there exists a sequence $(\breve{k}(n))$ such that the sequence of random variables $Y_{n,\breve{k}(n)}$ is tight. Note first that for any $\epsilon > 0$ there exists by assumption (i) and Prohorov's theorem an $M(\epsilon)$ such that

$$\mathbb{P}\big(||Y_n|| \leq M(\epsilon)\big) > 1 - \frac{\epsilon}{2}, \qquad \forall n \in \mathbb{N}.$$

For $n = 1$ there exists by the Portmanteau lemma and the fact that Y_1 has a continuous distribution according to assumption (iii) a $\tilde{k}(1)$ such

$$\big|\mathbb{P}\big(||Y_{1,k}|| \leq M(\epsilon)\big) - \mathbb{P}\big(||Y_1|| \leq M(\epsilon)\big)\big| \leq \frac{\epsilon}{2}, \qquad \forall k \geq \tilde{k}(1).$$

Therefore,

$$\mathbb{P}\big(||Y_{1,k}|| \leq M(\epsilon)\big) > 1 - \epsilon, \qquad \forall k \geq \tilde{k}(1).$$

By the same arguments there exists a $\tilde{k}(2)$ with

$$\mathbb{P}\big(||Y_{2,k}|| \leq M(\epsilon)\big) > 1 - \epsilon, \qquad \forall k \geq \tilde{k}(2).$$

Continuing like this there is a sequence $(\tilde{k}(n))$ such that

$$\mathbb{P}\big(||Y_{n,k}|| \leq M(\epsilon)\big) > 1 - \epsilon, \qquad \forall k \geq \tilde{k}(n).$$

Now define the sequence $(\breve{k}(n))$ by

$$\begin{array}{rcl} \breve{k}(1) & = & \tilde{k}(1); \\ \breve{k}(2) & = & \max\{\breve{k}(1), \tilde{k}(2)\} + 1; \\ \breve{k}(3) & = & \max\{\breve{k}(2), \tilde{k}(3)\} + 1; \\ \vdots & \vdots & \vdots \end{array}$$

Clearly $(\breve{k}(n))$ is an increasing sequence and by construction we have

$$\mathbb{P}\big(||Y_{n,\breve{k}(n)}|| \leq M(\epsilon)\big) > 1 - \epsilon, \qquad \forall n.$$

I.e. the sequence $(Y_{n,\breve{k}(n)})$ is tight and therefore has by Prohorov's theorem a subsequence that converges weakly. Denote the distributional limit of this convergent subsequence by X. We are going to show that X is distributed as Y which will finish the proof. For this let d be any metric on the space of probability measures that metricises convergence in distribution; see, for example, Dudley (2002, Chapter 11). Assume now that X is not distributed as Y. Then there exists a $\delta > 0$ such $d(X, Y) > \delta$. Let $\epsilon > 0$ be arbitrary. Then

a) By assumption (ii) there exists an n_1 such that

$$d(Y_n, Y) \leq \frac{\epsilon}{3}, \qquad \forall n \geq n_1;$$

b) Because $Y_{n,\breve{k}(n)}$ converges weakly to X there is an n_2 such that

$$d(Y_{n,\breve{k}(n)}, X) \leq \frac{\epsilon}{3}, \qquad \forall n \geq n_2;$$

c) From the above we know that for $n_3 := \max\{n_1, n_2\}$ there is a $\bar{k}(n_3)$ such that

$$d(Y_{n_3,k}, Y_{n_3}) \leq \frac{\epsilon}{3}, \qquad \forall k \geq \bar{k}(n_3).$$

Choosing $\epsilon = \delta$ in a)-c) they imply that

$$d(Y, X) \leq \delta,$$

which is a contradiction. Hence, X is distributed as Y and this finishes the proof. ■