Vallarino, Pierluigi

Working Paper

Dynamic kernel models

Tinbergen Institute Discussion Paper, No. TI 2024-082/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Vallarino, Pierluigi (2024) : Dynamic kernel models, Tinbergen Institute Discussion Paper, No. TI 2024-082/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/311620

TI 2024-082/III
Tinbergen Institute Discussion Paper

Dynamic kernel models

Pierluigi Vallarino[1]

1 Erasmus University Rotterdam, Tinbergen Institute

Dynamic kernel models

Pierluigi Vallarino*
Econometric Institute, Erasmus University Rotterdam

December 29, 2024

Abstract

This paper introduces the family of Dynamic Kernel models. These models approximate the predictive density function of a time series through a weighted average of kernel densities possessing a dynamic bandwidth. A general specification is presented and several particular models are studied in detail. We propose an M-estimator for model parameters and derive its asymptotic properties under a misspecified setting. A consistent density estimator is also introduced. Monte Carlo results show that the new models effectively track the time-varying distribution of several data generating processes. Dynamic Kernel models outperform extant kernel-based approaches in tracking the predictive distribution of GDP growth.

Keywords: Time-varying density function; time-varying parameter models; M estimation; density forecasts.

*I am grateful to Leopoldo Catania, Valentina Corradi, Christian Francq, Andrew Harvey, Oliver Linton, Rutger-Jan Lange, Jordi Llorens Terrazas, Alessandra Luati, Paolo Santucci de Magistris, Lorenzo Trapani, Dick van Dijk and Jean-Michel Zakoïan for their precious remarks and suggestions.

1 Introduction

Random phenomena can be fully characterized through their probability distribution. As a result, density forecasts are widely employed in natural and social sciences such as climatology (Campbell and Diebold, 2005), epidemiology (Liu et al., 2021), and econometrics (Elliott and Timmermann, 2016). For instance, fiscal and monetary authorities increasingly rely on density forecasts of GDP growth (Adrian et al., 2019), inflation (Lopez-Salido and Loria, 2024), and the unemployment rate (Kiley, 2022).

Motivated by this growing interest in density forecasts, we introduce a method to track the predictive distribution of the real-valued time series $Y = \{Y_t; t \in \mathbb{Z}\}$. In particular, we use an approach inspired by Kernel Density Estimation (KDE) to model the density function of $Y_{t+1}|\mathcal{F}_t$, for $\mathcal{F}_t$ a sigma-field containing all past information on Y. In KDE, the unconditional density of Y is modeled as an average or kernel functions, which are centred around past values of Y and possess a share a scale component called bandwidth (see Rosenblatt, 1956 for a seminal reference). Since we focus on predictive rather than unconditional densities, modifications are required to handle the conditioning information set $\mathcal{F}_t$. Harvey and Oryshchenko (2012) and Jeon and Taylor (2012) include past information in KDE by weighting the kernel functions through exponentially decaying weights and treat the bandwidth as a static parameter. In this way, the two methods approximate the predictive density with a mixture of kernel densities sharing a time-invariant scale term.[1] As is common in time series analysis, mixture weights assign more importance to densities centred around more recent values of Y. Unfortunately, neither approach can handle random variables whose distributional properties change abruptly over time.

As a first contribution, this paper develops a family of kernel-based models to track and

[1] Throughout the paper, the term *approximate* refers to studying a density of interest without assuming correct specification of the model-based density.

predict the time-varying density of $Y_{t+1}|\,\mathcal{F}_t$. These models extend the idea of Harvey and Oryshchenko (2012) and Jeon and Taylor (2012) by introducing more general weighting schemes and by treating the bandwidth as a time-varying parameter. Using more flexible weights makes model-based densities more responsive to new information on Y, thereby allowing our models to handle more rapidly changing processes. This is also facilitated by the use of a dynamic bandwidth, which allows our density estimator to quickly adapt to changes in the shape of the target density. In contrast, a static bandwidth implies a constant smoothness of the density estimator, which can be overly restrictive when the shape of the true density changes rapidly. As far as model-based distributional properties are concerned, we show that a fixed bandwidth entails similar dynamics for all predictive moments of Y. This could be problematic when, for instance, the predictive mean changes smoothly over time while the conditional variance is rapidly varying. The use of a dynamic bandwidth solves this issue, and also allows researchers to incorporate stylized facts of the data in kernel-based models (this is done by specifying an appropriate dynamics for the bandwidth, as discussed with examples in Section 2). Because the new models are designed for time-varying conditional distributions, we call the resulting family: *Dynamic Kernel models*. Statistical properties are studied in detail and we derive closed-form expressions for predictive moments. One-step ahead quantiles are easily obtained by inverting model-based distribution functions.

The second contribution concerns estimation and inference within Dynamic Kernel models. Model parameters are estimated by maximizing an objective function based on model-implied densities. To allow for potential misspecification of these densities, we derive asymptotic properties of the resulting M-estimator relative to the minimizer of the Kullback-Leibler divergence between the true and the model-based probability measures.

Additionally, we introduce a consistent estimator of the density process that is closest, in Kullback-Leibler terms, to the true ones. These contributions also apply to the approach of Harvey and Oryshchenko (2012), which is a particular case of our class of models and for which no asymptotic theory has been provided yet.[2]

Dynamic Kernel models are related to but different from other kernel-based approaches for predictive densities. Indeed, the semi-parametric GARCH models of Drost and Klaassen (1997) and Sun and Stengos (2006), and the semi-parametric score-driven model of Blasques et al. (2016) use a kernel density estimator for the density of the innovation terms in an observation-driven location-scale model.[3] Hao et al. (2018) consider a similar method for stochastic volatility models. All these approaches rely on the time-varying location and scale, and obtain the distribution of interest as a by-product of the location-scale structure. Conversely, we start from a kernel structure on the density of interest and obtain dynamic distributional properties as a consequence of this approximating functional form. Because standard kernel density estimators are a mixture of densities, our family of models is also related to dynamic mixture models such as the Mixture Autoregressive (MAR) model of Wong and Li (2000), its heteroskedastic counterpart by Wong and Li (2001), and its Student's t version by Wong et al. (2009).

Simulations show that Dynamic Kernel models outperform the approach of Harvey and Oryshchenko (2012) when tracking predictive distributions under several data generating processes. An empirical illustration involving US real GDP growth validates these numerical results. In particular, coverage tests from Kupiec (1995) and Christoffersen (1998) show that Dynamic Kernel models improve the goodness-of-fit of the estimated time-varying

[2] Wang et al. (2018) and Garcin et al. (2023) propose alternative M-estimators for the methodology of Harvey and Oryshchenko (2012). While empirically relevant, our estimator is consistent for the minimizer of the Kullback-Leibler divergence, thus being optimal in an information-theoretic sense.

[3] Engle and Gonzalez-Rivera (1991) considers a semi-parametric ARCH model based on a non-parametric estimator of the innovations' density. While they focus on the discrete maximum penalized likelihood estimator of Scott et al. (1980), a kernel density estimator could be also used here.

quantiles compared with the approach of Harvey and Oryshchenko (2012). Similar results hold for the predictive mean and variance, as implied by residuals analysis. Density forecast results are equally good and remark the importance of a time-varying bandwidth.

The rest of the paper is structured as follows: Section 2 introduces Dynamic Kernel models; Section 3 presents an M-estimator for model parameters and a consistent density estimator; Sections 4 and 5 concern an application to tracking and predicting the distribution of US GDP growth; Section 6 concludes the paper. Assumptions for the asymptotic analysis, derivations and useful mathematical expressions are in Appendices A, B, C and F. Appendix D details a Monte Carlo analysis, while Appendix E reports further empirical results. All these appendices are in the supplementary material.

2 Dynamic Kernel models

We consider a real-valued, strictly stationary and ergodic (SE) process $Y \coloneqq \{Y_t;\, t \in \mathbb{Z}\}$ defined over the probability space $(\Omega, \mathcal{F}, P)$, and that generates the filtration $\mathbb{F} = \{\mathcal{F}_t; t \in \mathbb{Z}\}$ for $\mathcal{F}_t \coloneqq \sigma\left(Y_{t-s},\, s \geq 0\right)$. Let $F^0_{t+1|t}\left(y\right) \coloneqq P\left(Y_{t+1} \leq y | \, \mathcal{F}_t\right)$, with $F^0_{t+1|t} \in \mathcal{C}^0\left(\mathbb{R}\right)$, and $f^0_{t+1|t}\left(y\right) \coloneqq \frac{d}{dy} F^0_{t+1|t}\left(y\right)$ denote the true distribution and density functions of $Y_{t+1} | \, \mathcal{F}_t$, respectively. For any $y \in \mathbb{R}$, we propose to approximate $f^0_{t+1|t}(y)$ with the function

$$f_{t+1|t}\left(y\right) = \frac{1}{h_{t+1}} \sum_{i=0}^{\infty} \omega_i \mathcal{K}\left(\frac{y - y_{t-i}}{h_{t+1}}\right), \tag{1}$$

where y_t denotes a realization of Y_t, $\frac{1}{h_{t+1}}\mathcal{K}\left(\frac{y-y_{t-i}}{h_{t+1}}\right)$ is a kernel density centred in y_{t-i}, the positive weights $\{\omega_i;\, i \in \mathbb{N}\}$ sum to one, and $\{h_t;\, t \in \mathbb{Z}\}$ is the almost surely (a.s.) positive bandwidth process. As in traditional kernel density estimation, h_{t+1} controls the smoothness of $f_{t+1|t}$: higher values of h_{t+1} imply greater dispersion of the kernel densities $\mathcal{K}\left(\cdot\right)$ around past values of Y or, equivalently, a smoother $f_{t+1|t}$. Because the density

of interest, i.e. $f^0_{t+1|t}$, is time-varying even when Y is stationary, the required degree of smoothness may change over time. This is particularly true when the shape of $f^0_{t+1|t}$ changes frequently, in which case the bandwidth must adapt quickly. Sections 2.3 and 4.3 further discuss the benefits of a time-varying bandwidth when studying predictive moments implied by (1). Predictive densities are consistently estimated without requiring $h_{t+1} \to 0$ (either a.s. or in probability) as $t \to \infty$. This difference with standard KDE is due to our focus on predictive rather than unconditional densities, and is further stressed in Section 3.2, which introduces a consistent density estimator based on (1).

2.1 The choice of the weighting scheme

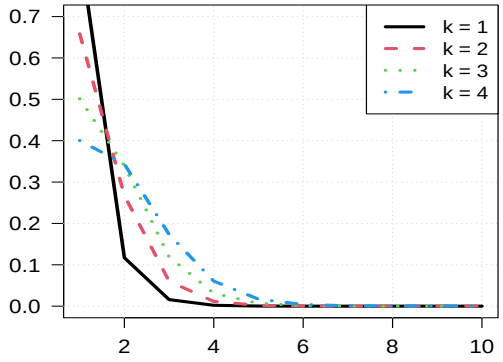
This paper considers the four weighting schemes reported in Table 1. As is common in time series analysis, we choose weights that assign more importance to kernels centered around more recent values of Y. EWMA weights are those employed by Harvey and Oryshchenko (2012); larger values of the parameter θ imply a slower, yet always exponential, convergence to zero. Gamma weights are based on the standard Gamma function $\Gamma(\cdot)$ and on the upper incomplete Gamma function with truncation parameter λi, i.e. $\Gamma(\cdot; \lambda i)$ (see Abramowitz and Stegun, 1964 for a discussion of these functions). Because $\Gamma(k; \lambda\, i)/\Gamma(k) = O\left(\exp\{-\lambda\, i\}\, i^{k-1}\right)$, these weights eventually decay exponentially fast but allow for different behaviors when i is relatively small. When $k = 1$ and $\lambda = -\log(\theta)$ for $\theta \in (0, 1)$, the i-th Gamma and EWMA weights coincide. Hence, Gamma weights generalize EWMA ones to allow for more flexibility with respect to i.[4] Figure 1a shows these weights for different values of k when $\lambda = 2$. If $k = 1$ (black solid line), we observe the exponential decay of EWMA weights. As k gets larger: (i) more importance is attached

[4]In the contexts of volatility and quantiles modeling, Li and Zhu (2020) and Zhu (2023) provide other generalizations of EWMA weights. Future research may employ these generalizations to our framework.

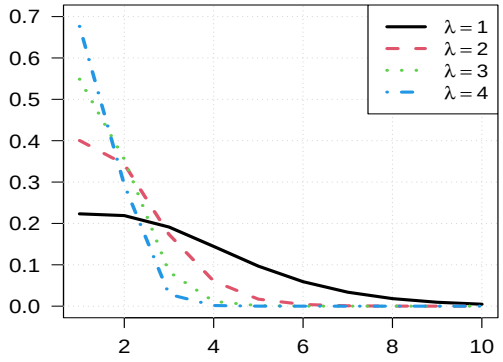
Table 1: Weighting schemes considered in this paper.

Name	ω_i	Parameters	Decay rate
EWMA	$(1-\theta)\,\theta^i$	$\theta \in (0,1)$	Exponential $\forall\, i \geq 0$
Gamma	$a_1 \Gamma(k;\lambda i)/\Gamma(k)$	$\lambda > 0,\ k > 0$	Exponential as $i \to \infty$
Hyperbolic	$a_2\,(1+i)^{-\theta}$	$\theta > 1$	Hyperbolic $\forall\, i \geq 0$
Flexible hyperbolic	$a_3\,(1+\lambda i)^{-\theta}$	$\theta > 1,\ \lambda > 0$	Hyperbolic $\forall\, i \geq 0$

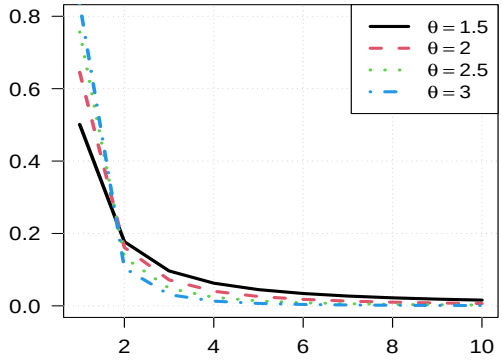
Note: a_1, a_2, and a_3 are strictly positive constants ensuring that $\sum_{i=0}^{\infty} \omega_i = 1$, e.g. $a_2 = \left(\sum_{i=0}^{\infty} (1+i)^{-\theta}\right)^{-1}$. $\Gamma(\cdot)$ is the standard Gamma function and $\Gamma(\cdot;\lambda i)$ the upper incomplete Gamma function with truncation parameter λi.



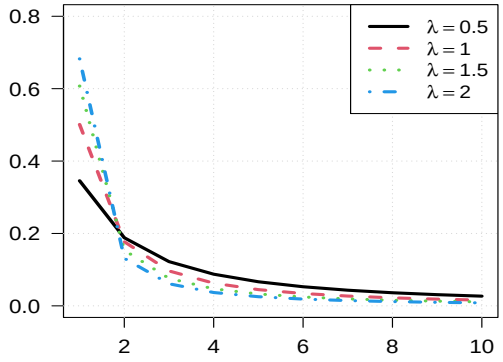
(a) Gamma weights, $k \in \{1,2,3,4\}$; $\lambda = 2$.



(b) Gamma weights, $\lambda \in \{1,2,3,4\}$; $k = 4$.



(c) Hyperbolic weights, $\theta \in \{1.5, 2, 2.5, 3\}$.



(d) Flexible hyperbolic weights, $\lambda \in \{0.5, 1, 1.5, 2\}$ and $\theta = 1.5$.

Figure 1: Upper panels: Gamma weights for different values of k (Panel a) and λ (Panel b). Lower panels: hyperbolic (Panel c) and flexible hyperbolic (Panel d) weights.

to older information, as implied by the black solid line being eventually dominated by all the others; (ii) the decay is not always exponential. Figure 1b shows Gamma weights for different values of λ when $k = 4$; larger values of λ entail a faster decay. Hyperbolic

weights exhibit a power-law relation with i, thus resembling coefficients of long memory filters (see Baillie, 1996, among others). Higher values of θ entail a faster convergence to zero, as shown in Figure 1c. Flexible hyperbolic weights possess the same power-law behavior but offer more flexibility through the parameter λ. Figure 1d shows that smaller values of λ imply a slower, yet always hyperbolic, decay.

2.2 The choice of the bandwidth process

We model the time-varying bandwidth through an observation-driven approach, i.e. h_{t+1} is measurable with respect to $\mathcal{F}_t$. Moreover, we assume that h_t is supported on the convex set $\mathcal{H} \subset \mathbb{R}^+$ and write $h_{t+1} = \phi(h_t, \varepsilon_t)$, where $\varepsilon_t \coloneqq \left(Y_t - \hat{\mu}_{t|t-1}\right)$ is the one-step ahead predictive error based on the predictive mean $\hat{\mu}_{t|t-1}$ (discussed in Section 2.3). The function $\phi : \mathcal{H} \times \mathbb{R} \to \mathcal{H}$ is time-invariant; its choice is discretionary and allows researchers to incorporate empirical features of the data in the model.

Because h_{t+1} is a scale component for the kernel densities, we specify $\phi(\cdot)$ starting from conditional heteroskedasticity models. An example is the GARCH-like recursion:

$$h_{t+1}^2 = \bar{h} + \beta h_t^2 + \alpha \varepsilon_t^2, \tag{2}$$

for $\bar{h} > 0$, $\alpha > 0$ and $\beta \geq 0$. While similar, equation (2) differs from the GARCH(1,1) model of Bollerslev (1986) in that the innovation term ε_t^2 is not a function of h_t^2 under (2), while $\varepsilon_t^2 = h_t^2 z_t^2$ for $z_t \overset{\text{i.i.d.}}{\sim} N(0,1)$ in GARCH(1,1). Under (2), differently signed predictive errors have the same impact on h_{t+1}^2 whenever their magnitudes coincide. To allow for an asymmetric response of h_{t+1}^2, we consider the recursion:

$$h_{t+1}^2 = \bar{h} + \beta h_t^2 + \left[\alpha + \gamma \mathbb{1}\left(\varepsilon_t < 0\right)\right] \varepsilon_t^2, \tag{3}$$

for $\bar{h} > 0$, $\beta \geq 0$, $\alpha \geq 0$ and $\gamma \geq 0$ with $\alpha + \gamma > 0$. This dynamics is based on the GJR model of Glosten et al. (1993) and we call it GJR-like. The additional term $\gamma \mathbb{1}(\varepsilon_t < 0) \varepsilon_t^2$ is beneficial whenever the (true) predictive distribution of Y responds differently to differently signed predictive errors. For instance, negative values of ε_t may have a larger impact on the variance of $Y_{t+1} | \mathcal{F}_t$ than positive ones (as in the leverage effect of Black, 1976 for stock returns). In this case, we expect a significantly positive estimate of γ, so that h_{t+1}^2 is more sensitive to negative than to positive values of ε_t (recall that a larger bandwidth implies more disperse kernel densities and, *ceteris paribus*, a higher predictive variance).

Under (2) and (3), h_{t+1}^2 is an unbounded function of ε_t. Hence, the bandwidth can become arbitrarily large after a sizeable predictive error. Among others, Creal et al. (2013) and Harvey (2013) showed that this is undesirable when tracking time-varying parameters. Thus, we also consider a specification where the log-bandwidth $\left\{\bar{h}_t := \log(h_t); t \in \mathbb{Z}\right\}$ evolves as:

$$\bar{h}_{t+1} = \bar{h} + \beta \bar{h}_t + \alpha u_t + \gamma \operatorname{sgn}(-\varepsilon_t)(u_t + 1), \tag{4}$$

with $\operatorname{sgn}(-\varepsilon_t) = 2\mathbb{1}(\varepsilon_t < 0) - 1$ and $u_t := \frac{(\nu+1)\varepsilon_t^2}{\nu+\varepsilon_t^2} - 1$, so that $-1 \leq u_t < \nu$ for $\nu > 0$. Because u_t has bounded support, $\bar{h}_{t+1}$ is a bounded function of ε_t, making the log-bandwidth process robust to outliers in the predictive error. Moreover equation (4) allows for a leverage behavior similar to (3). [5] Note that u_t mimics the innovation term of the Beta-t-EGARCH model of Harvey and Chakravarty (2008) and Harvey (2013), which is a Dynamic Conditional Score (DCS) model whose innovation term is based on the derivative of a Student's t log-density with respect to its scale parameter.[6] This is why we refer to

[5] While alternative methods for incorporating asymmetric responses in DCS scale models are discussed by Harvey and Chakravarty (2008) and Harvey and Lange (2017), our approach is closer to that of Harvey and Sucarrat (2014). Despite their different specifications, all these models successfully incorporate asymmetric responses in DCS models for the scale.

[6] Clearly, u_t is not related to the derivative of $\log(f_{t|t-1})$ with respect to h_t in our case. This decoupling between the score-based innovation term and the model-based predictive density is also present in the quasi-score driven approach by Blasques et al. (2023).

the dynamic in (4) as DCS-EGARCH.

In practice, we choose the optimal combination of weights, bandwidth and kernel density based on empirical performances of the model. As illustrated in Sections 4 and 5, we consider information criteria, model diagnostics, and density forecasts results. Because h_t does not (and need not, see also Remark 3.1) shrink to zero under (2) - (4), the choice of $\mathcal{K}(\cdot)$ is likely to impact the empirical performances of Dynamic Kernel models (this is confirmed by the Monte Carlo analysis of Appendix D).

Proposition 2.1 shows that the stochastic recurrence equations in (2) to (4) admit an a.s. unique SE solution under the next assumption. Its proof is reported in Appendix B and relies on results due to Brandt (1986). Notably, inspection of the proof shows that no moment condition is required for the DCS-EGARCH specification.

Assumption 1. *$\{Y_t;\, t \in \mathbb{Z}\}$ is SE with $\mathbb{E}\left[|Y_t|^{\delta}\right] < \infty$ for a $\delta > 0$ such that $\sum_{i=0}^{\infty} \omega_i^{\delta} < \infty$.*

Exponential weights always satisfy the summability condition while $\delta > 1$ is required for hyperbolic ones.

Proposition 2.1. *Under Assumption 1: i) If $0 < \beta < 1$, the a.s. unique SE solution to equation (2) is:*

$$h_{t+1}^2 = \frac{\bar{h}}{1-\beta} + \alpha \sum_{s=0}^{\infty} \beta^s \varepsilon_{t-s}^2;$$

ii) If $0 < \beta < 1$, the a.s. unique SE solution to equation (3) is:

$$h_{t+1}^2 = \frac{\bar{h}}{1-\beta} + \sum_{s=0}^{\infty} \beta^s \left[\alpha + \gamma \mathbb{1}\left(\varepsilon_{t-s} < 0\right)\right] \varepsilon_{t-s}^2;$$

iii) If $|\beta| < 1$, the a.s. unique SE solution to equation (4) is:

$$\bar{h}_{t+1} = \frac{\bar{h}}{1-\beta} + \sum_{s=0}^{\infty} \beta^s \left[\alpha u_{t-s} + \gamma \operatorname{sgn}\left(-\varepsilon_{t-s}\right)\left(u_{t-s} + 1\right)\right].$$

Before moving on with the exposition, it is important to remark that the presence of indicators makes $\phi(\cdot)$ not everywhere $\mathcal{C}^2$ in ε_t under (3) and (4). Since ε_t depends on parameters of the weighting scheme, models based on (3) and (4) are not $\mathcal{C}^2$ over the parameter space. To simplify estimation and inference, we replace the function $\mathbb{1}(x < 0)$ with the smooth analog $G(x) = \left(1 + \exp\left(-\frac{x}{c}\right)\right)^{-1}$, for $c > 0$ a smoothing parameter selected *a priori*, such that $G(x) \to \mathbb{1}(x < 0)$ almost everywhere as $c \to 0$. Note that the same problem and solution are discussed by Zakoian (1994) for threshold heteroskedastic models, and by van Dijk et al. (2002) for smooth transition autoregressive models. Finally, Proposition 2.1 still holds when $\mathbb{1}(\varepsilon_t < 0)$ is replaced by $G(\varepsilon_t)$ in (3) and (4), and in the corresponding solutions.

2.3 Predictive moments

The mixture structure in (1) allows us to derive closed form expressions for the predictive moments of Y. To do it, we parametrize $\mathcal{K}(\cdot)$ so that $\frac{1}{h_{t+1}}\mathcal{K}\left(\frac{y-y_{t-i}}{h_{t+1}}\right)$ is the density of a random variable with mean y_{t-i} and variance h_{t+1}^2. This reparametrization is unnecessary if one is not interested in predictive moments and may impose unnecessary constraints on the parameter space. As an example, equation (5) shows $f_{t+1|t}(y)$ when $\mathcal{K}(\cdot)$ is either a Gaussian ($f_{t+1|t,g}$) or a Student's t kernel ($f_{t+1|t,\mathcal{T}}$) with $\nu > 2$ degrees of freedom:

$$\begin{aligned} f_{t+1|t,g}(y) &= \frac{1}{h_{t+1}\sqrt{2\pi}} \sum_{i=0}^{\infty} \omega_i \exp\left\{-\frac{(y-y_{t-i})^2}{2h_{t+1}^2}\right\}, \\ f_{t+1|t,\mathcal{T}}(y) &= \frac{1}{h_{t+1}\sqrt{\pi(\nu-2)}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \sum_{i=0}^{\infty} \omega_i \left(1 + \frac{(y-y_{t-i})^2}{h_{t+1}^2(\nu-2)}\right)^{-\frac{\nu+1}{2}}. \end{aligned} \tag{5}$$

While the densities in (5) already imply a skewed predictive distribution, skewed kernels can also be considered, e.g. the Skew-Normal density of Azzalini (1985) with suitably adjusted

parameters. Power-law tails can be accommodated through the Student's t kernel or the Skewed Student's t one of Azzalini and Capitanio (2003). Proposition 2.2 reports formulas for the one-step ahead predictive mean and variance (while expressions for higher order moments can be derived , they entail a more cumbersome notation and depend on the chosen kernel). In its statement, $\mathbb{E}_{\widehat{P}}$ and $\mathrm{Var}_{\widehat{P}}$ denote the expected value and variance taken under model-based probability measure $\left(\widehat{P}\right)$, which is not assumed to coincide with the true probability measure P. The proposition is a particular case of Proposition C.1, which contains expressions for the k-step ahead mean and variance, and that we report in Appendix C.1 along with its proof.

Proposition 2.2. *If* $\frac{1}{h_{t+1}}\int_{\mathbb{R}} y\mathcal{K}\left(\frac{y-y_{t-i}}{h_{t+1}}\right)dy = y_{t-i}$ *and* $\frac{1}{h_{t+1}}\int_{\mathbb{R}}(y-y_{t-i})^2\mathcal{K}\left(\frac{y-y_{t-i}}{h_{t+1}}\right)dy = h_{t+1}^2$*:*

$$\hat{\mu}_{t+1|t} \coloneqq \mathbb{E}_{\widehat{P}}\left[Y_{t+1}\middle|\,\mathcal{F}_t\right] = \sum_{i=0}^{\infty}\omega_i y_{t-i}; \tag{6}$$

$$\hat{\sigma}^2_{t+1|t} \coloneqq \mathrm{Var}_{\widehat{P}}\left[Y_{t+1}\middle|\,\mathcal{F}_t\right] = \hat{h}^2_{t+1|t} + \sum_{i=0}^{\infty}\omega_i y^2_{t-i} - \hat{\mu}^2_{t+1|t}, \tag{7}$$

where $\hat{h}^2_{t+1|t} \coloneqq \mathbb{E}_{\widehat{P}}\left[h^2_{t+1}\middle|\,\mathcal{F}_t\right]$.

$\hat{\mu}_{t+1|t}$ is a convex linear combination of all past values of Y, which is similar to the result of Koopman and Harvey (2003) for unobserved component models. The predictive variance has a two-component structure: the first one is the one-step ahead squared bandwidth $\hat{h}^2_{t+1|t}$ (available in closed form for all specifications of Section 2.2), while the second one is not a function of the bandwidth as it is given by $\left\{\sum_{i=0}^{\infty}\omega_i y^2_{t-i} - \hat{\mu}^2_{t+1|t}\right\}$.[7] In practice, the second component controls the time-varying level of the variance, while the first one drives short-

[7]Equation (27) resembles the one-step ahead predictive variance of the heteroskedastic Mixture Autoregressive (MAR) model of Wong and Li (2001): this is expected given the mixture structure in Equation (1). Note also the analogy with the original MAR model of Wong and Li (2000) when h_{t+1} is fixed.

term fluctuations around this level (see the discussion in Section 4.3). When the bandwidth is fixed, i.e. $h_{t+1} \equiv h > 0$ for any $t \in \mathbb{Z}$, parameters of the weighting scheme control how past observations impact all predictive moments, i.e. all time-varying distributional properties. This could be overly restrictive, especially in the case of substantial differences in the dynamics of different moments, e.g. the (true) predictive mean changes smoothly over time while the variance is more rapidly varying. Moreover, we can use bandwidth dynamics as those of Section 2.2 to impose a structure on the relation between predictive errors and predictive moments, e.g. using leverage terms to impose asymmetric responses, which can then be tested empirically. Finally, the fixed bandwidth h is also the lower bound of $\hat{\sigma}^2_{t+1|t}$ (note that $\sum_{i=0}^{\infty} \omega_i y^2_{t-i} - \hat{\mu}^2_{t+1|t} > 0$ a.s. at any point in time). This could result in further modeling challenges, as h needs to provide the optimal degree of smoothing at any point in time, while also allowing $\hat{\sigma}^2_{t+1|t}$ to become sufficiently small, if required.

3 Estimation

We collect model parameters into the vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$ and partition it as: $\boldsymbol{\theta} = (\boldsymbol{\theta}'_{\mathcal{K}}, \boldsymbol{\theta}'_{\omega}, \boldsymbol{\theta}'_{h})'$, for $\boldsymbol{\theta}_{\mathcal{K}} \in \boldsymbol{\Theta}_{\mathcal{K}} \subset \mathbb{R}^{d_{\mathcal{K}}}$, $\boldsymbol{\theta}_{\omega} \in \boldsymbol{\Theta}_{\omega} \subset \mathbb{R}^{d_{\omega}}$ and $\boldsymbol{\theta}_{h} \in \boldsymbol{\Theta}_{h} \subset \mathbb{R}^{d_{h}}$ with $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\mathrm{k}} \times \boldsymbol{\Theta}_{\omega} \times \boldsymbol{\Theta}_{\omega}$. $\boldsymbol{\theta}_{\mathcal{K}}$ collects the parameters of $\mathcal{K}$, e.g. $\boldsymbol{\theta}_{\mathcal{K}} = \nu$ for a Student's t kernel; parameters of the weights are grouped into $\boldsymbol{\theta}_{\omega}$, e.g. $\boldsymbol{\theta}_{\omega} = (k, \lambda)$ for Gamma weights, while $\boldsymbol{\theta}_{h}$ contains those of the bandwidth process, e.g. $\boldsymbol{\theta}_{h} = \left(\bar{h}, \alpha, \beta\right)$ for the GARCH case.

Let us consider the generic element of the SE sequence of model-based densities:

$$f_{t+1|t}\left(y; \boldsymbol{\theta}\right) = \frac{1}{h_{t+1}\left(\boldsymbol{\theta}\right)} \sum_{i=0}^{\infty} \mathcal{K}\left(\frac{y - Y_{t-i}}{h_{t+1}\left(\boldsymbol{\theta}\right)}; \boldsymbol{\theta}_{\mathcal{K}}\right) \omega_i\left(\boldsymbol{\theta}_{\omega}\right),$$

when evaluated at $y \in \mathbb{R}$. The sequence $\{h_t\left(\boldsymbol{\theta}\right); t \in \mathbb{Z}\}$ is the SE bandwidth process based on the parameter hypothesis $\boldsymbol{\theta}$. In particular, it depends on $\boldsymbol{\theta}_{h}$ through the function $\phi\left(\cdot\right)$

and on $\boldsymbol{\theta}_\omega$ through the predictive error $\varepsilon_t(\boldsymbol{\theta}_\omega) = (Y_t - \sum_{i=0}^{\infty} \omega_i(\boldsymbol{\theta}_\omega) Y_{t-1-i})$.

Because we assume that Dynamic Kernel models only approximate the density of interest, we carry out inference under a misspecified setting, thereby establishing asymptotic properties with respect to a pseudo-true parameter value $\boldsymbol{\theta}^*$. As customary in M-estimation of time-varying parameter models (see Blasques et al., 2018, 2023, among others), we define $\boldsymbol{\theta}^*$ as the minimizer of the expected Kullback-Leibler divergence between the true and the model based probability measure, viz.

$$\boldsymbol{\theta}^* \coloneqq \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min}\ \mathbb{E}\left[KL\left(f^0_{t+1|t}(Y_{t+1}), f_{t+1|t}(Y_{t+1};\boldsymbol{\theta})\right)\right], \tag{8}$$

for $f^0_{t+1|t}(Y_{t+1})$ the true predictive density function evaluated at Y_{t+1}. Assumption 1 and the next two conditions ensure existence of $\boldsymbol{\theta}^*$ (we denote by $\mathcal{K}(x;\cdot)$ the function $\mathcal{K}$ evaluated at $x \in \mathbb{R}$ for any value of $\boldsymbol{\theta}_\mathcal{K}$).

Assumption 2. *The parameter space $\boldsymbol{\Theta}$ is compact.*

Assumption 3. $\mathcal{K}(\cdot;\boldsymbol{\theta}_\mathcal{K})$, $h_t(\boldsymbol{\theta})$, $\omega(\boldsymbol{\theta}_\omega) \in \mathcal{C}^2(\boldsymbol{\Theta})$; $\mathcal{K}(x;\cdot) \in \mathcal{C}^2(\mathbb{R}), \mathcal{K}(x;\cdot) > 0$, $\forall x \in \mathbb{R}$.

After observing a sample $y_{0:T} = \{y_0, \ldots, y_T\}$ from Y, we construct a sequence of pseudo-densities $\left\{\hat{f}_{1|0}, \ldots, \hat{f}_{T|T-1}\right\}$ and use them to define the M-estimator:

$$\widehat{\boldsymbol{\theta}}_T \coloneqq \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \sum_{t=m}^{T-1} \hat{\varphi}_{t+1}(\boldsymbol{\theta}), \tag{9}$$

for

$$\hat{\varphi}_{t+1}(\boldsymbol{\theta}) \coloneqq \log\left\{\hat{f}_{t+1|t}(Y_{t+1};\boldsymbol{\theta})\right\} = \log\left\{\frac{1}{\hat{h}_{t+1}(\boldsymbol{\theta})}\sum_{i=0}^{t}\mathcal{K}\left(\frac{Y_{t+1}-y_{t-i}}{\hat{h}_{t+1}(\boldsymbol{\theta})};\boldsymbol{\theta}_\mathcal{K}\right)\omega_i(\boldsymbol{\theta}_\omega)\right\}, \tag{10}$$

where $\left\{\hat{h}_t(\boldsymbol{\theta});\ t = 1, \ldots, T\right\}$ is the bandwidth process recovered from $y_{0:T}$ and initialized

at the arbitrary value $\hat{h}_1 \in \mathcal{H}$. This initial value makes the sequence non SE. While $\hat{f}_{t+1|t}(y;\boldsymbol{\theta}) > 0$ for any $y \in \mathbb{R}$, the fact that $\sum_{i=0}^{t} \omega_i < 1$ implies $\int_{\mathbb{R}} \hat{f}_{t+1|t}(y;\boldsymbol{\theta})\, dy < 1$. This is why we call these functions pseudo-densities. Section 3.2 describes a (consistent) proper density estimator based on the sample $y_{0:T}$. Finally, the sum in (9) starts at $t = m$ because we discard m observations to start the procedure. m can be a function of the sample size, i.e. $m = m(T)$, in which case asymptotic properties of $\widehat{\boldsymbol{\theta}}_T$ (see the next section) hold as long as $\sqrt{T - m(T)} = O\left(T^{1/2}\right)$, e.g. $m(T) = \lfloor T^{\alpha} \rfloor$ or $m(T) = \lfloor \alpha T \rfloor$ for $\alpha \in (0,1)$.[8]

3.1 Asymptotic properties of the M-estimator

The objective function in (9) is not an SE sequence. To study the asymptotic properties of $\widehat{\boldsymbol{\theta}}_T$, we introduce the estimator

$$\boldsymbol{\theta}_T \coloneqq \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{t=m}^{T-1} \varphi_{t+1}(\boldsymbol{\theta}),$$

where

$$\varphi_{t+1}(\boldsymbol{\theta}) \coloneqq \log\left\{ \frac{1}{h_{t+1}(\boldsymbol{\theta})} \sum_{i=0}^{\infty} \mathcal{K}\left(\frac{y_{t+1} - y_{t-i}}{h_{t+1}(\boldsymbol{\theta})}; \boldsymbol{\theta}_{\mathcal{K}} \right) \omega_i(\boldsymbol{\theta}_{\omega}) \right\}, \tag{11}$$

so that $\boldsymbol{\theta}_T$ is based on the SE counterpart of the objective function in (9). While empirically infeasible, we derive the asymptotic properties of $\boldsymbol{\theta}_T$ and then establish conditions such that they hold for $\widehat{\boldsymbol{\theta}}_T$. Theorem 1 states the strong consistency of $\widehat{\boldsymbol{\theta}}_T$ for $\boldsymbol{\theta}^*$, while its asymptotic distribution is presented in Theorem 2. Both theorems are proved in Appendix C, while Appendix A reports the assumptions. These are presented at a high-level, i.e. involving model components such as the weights and the bandwidth, and can be verified under more

[8] The objective function in (9) can be modified to prevent spikes in the gradient when $\hat{f}_{t+1|t}(Y_{t+1};\boldsymbol{\theta})$ is infinitesimal. For instance, small values can be trimmed as in Fermanian and Salanie (2004). However, selecting which observations to trim is non trivial, as an excessive trimming hampers the quality of parameter estimates (see Monte Carlo results in Fermanian and Salanie, 2004). Because of these difficulties, we defer the study of trimmed estimators to future work.

primitive conditions, e.g. restrictions on $\boldsymbol{\Theta}$, once a model has been specified. We provide these more primitive conditions for all bandwidth processes of Section 2 in Appendix B. As detailed in Appendices A and C, primitive conditions are milder, and some assumptions can also be discarded, under exponentially decaying weights. Moment conditions can be substantially relaxed when h_t is either constant or a.s. bounded. The latter is the case for the DCS-EGARCH specification in (4), as well as under other DCS-based dynamics.

Theorem 1. *Under Assumptions 1 to 6,* $\widehat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}^*$ *as* $T \to \infty$.

Theorem 2. *Under Assumptions 1 to 11 and as* $T \to \infty$: $\sqrt{T}\left(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ *for* $\boldsymbol{\Sigma} \coloneqq \mathbf{Q}^{-1}\mathbf{V}\mathbf{Q}^{-1}$, *where:* $\mathbf{Q} = \mathbb{E}\left[\nabla^2 \varphi_t(\boldsymbol{\theta}^*)\right]$ *and*

$$\mathbf{V} = \lim_{T\to\infty} \frac{1}{T-m} \mathbb{E}\left[\left(\sum_{t=m}^{T-1} \nabla \varphi_t(\boldsymbol{\theta}^*)\right)\left(\sum_{t=m}^{T-1} \nabla \varphi_t(\boldsymbol{\theta}^*)\right)'\right].$$

3.2 A consistent density estimator

To construct a proper estimator of the density of interest, we introduce the sequence $\left\{\tilde{f}_{1|0}, \ldots, \tilde{f}_{T|T-1}\right\}$ with generic element:

$$\tilde{f}_{t+1|t}(y; \boldsymbol{\theta}) \coloneqq \frac{1}{\tilde{h}_{t+1}(\boldsymbol{\theta})} \sum_{i=0}^{t} \mathcal{K}\left(\frac{y_{t+1} - y_{t-i}}{\tilde{h}_{t+1}(\boldsymbol{\theta})}; \boldsymbol{\theta}_{\mathcal{K}}\right) \tilde{\omega}_{i,t}(\boldsymbol{\theta}_\omega), \tag{12}$$

when evaluated at $y \in \mathbb{R}$. The collection of weights $\tilde{\omega}_t = \{\tilde{\omega}_{0,t}, \ldots, \tilde{\omega}_{t,t}\}$ is (deterministically) time-varying and such that $\sum_{i=0}^{t} \tilde{\omega}_{i,t}(\boldsymbol{\theta}_\omega) = 1$ at any point in time. We define these weights so that $\tilde{\omega}_{i,t} \to \omega_i$ for any $i \in \mathbb{N}$ as $t \to \infty$, e.g. $\tilde{\omega}_{i,t} = \frac{1-\theta}{1-\theta^{t+1}}\theta^i$ when $\omega_i = (1-\theta)\theta^i$.

The predictive error based on $\tilde{\omega}_t$ and $y_{0:t}$ is $\tilde{\varepsilon}_{t+1}(\boldsymbol{\theta}_\omega) \coloneqq Y_{t+1} - \sum_{i=0}^{t} \tilde{\omega}_{i,t}(\boldsymbol{\theta}_\omega) y_{t-i}$, and we use it to construct the sequence $\left\{\tilde{h}_t(\boldsymbol{\theta}); t \in \mathbb{N}\right\}$ where $\tilde{h}_{t+1}(\boldsymbol{\theta}) = \phi\left(\tilde{h}_t(\boldsymbol{\theta}), \tilde{\varepsilon}_t(\boldsymbol{\theta}_\omega)\right)$, for $\phi(\cdot)$ as in Section 2.2, and that is initialized at the arbitrary value $\tilde{h}_1 \in \mathcal{H}$. Proposition 3.1

establishes that if $\widehat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}^*$ then we can use the sequence $\left\{\tilde{f}_{1|0}, \ldots, \tilde{f}_{T|T-1}\right\}$ to consistently estimate the predictive densities that are closest, in Kullback-Leibler terms, to the true ones. Its proof is reported in Appendix C and relies on assumptions presented in Appendix A.

Proposition 3.1. *Under Assumptions 1 to 6 and 12,* $\left|\tilde{f}_{T+1|T}\left(y; \widehat{\boldsymbol{\theta}}_T\right) - f_{T+1|T}\left(y; \boldsymbol{\theta}^*\right)\right| \xrightarrow{p} 0$ *as* $T \to \infty$ *and for any* $y \in \mathbb{R}$.

Remark 3.1. *Differently from standard kernel density estimation, Proposition 3.1 does not require that* $\tilde{h}_{t+1}(\boldsymbol{\theta}) \to 0$ *(either a.s. or in probability) as* $t \to \infty$. *On the other hand, we need that* $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left|\tilde{h}_t(\boldsymbol{\theta}) - h_t(\boldsymbol{\theta})\right| \xrightarrow{p} 0$ *as* $t \to \infty$ *and at a certain rate discussed in the proof. Similar conditions are pervasive in the asymptotic analysis of time-varying parameter models (see Francq and Zakoian, 2019 among others).*

4 Empirical illustration

4.1 Data, model specifications and preliminary results

This section studies how well different Dynamic Kernel models can track the predictive density of US real GDP growth. In particular, we consider quarter-on-quarter growth rate of US real GDP between Q1:1947 and Q4:2019. Figure 2 shows this time series.

Figure 2: Quarter-on-quarter growth rate of US real GDP between Q1:1947 and Q4:2019.

As a starting point, we estimate thirty-two Dynamic Kernel models based on the four weighting schemes of Section 2.1, the three bandwidth processes of Section 2.2 and a

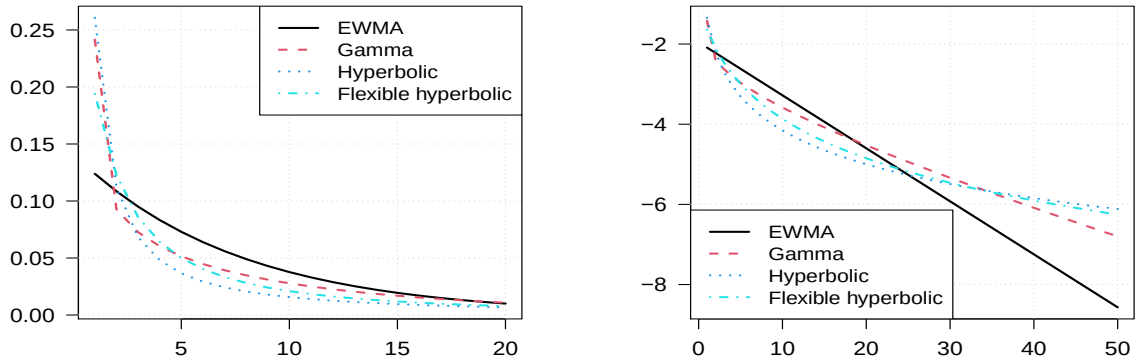
Table 2: Estimation results, values of objective function and BIC.

	Gaussian – Objective function				Student's t – Objective function			
	GARCH	GJR	DCS	Fixed	GARCH	GJR	DCS	Fixed
EWMA	326.23	319.71	318.33	335.82	324.84	319.61	324.69	331.81
Gamma	321.75	316.22	315.29	332.65	321.57	316.24	322.78	329.44
Hyperbolic	322.37	319.53	318.47	333.32	322.36	319.57	325.20	332.15
F-Hyperbolic	321.97	318.09	315.33	332.01	321.70	318.01	323.03	330.12
	Gaussian – BIC				Student's t – BIC			
	GARCH	GJR	DCS	Fixed	GARCH	GJR	DCS	Fixed
EWMA	337.44	333.73	335.14	341.42	338.85	336.43	344.31	340.22
Gamma	335.77	**333.04**	334.91	341.06	338.39	335.86	345.21	340.66
Hyperbolic	333.58	333.55	335.29	338.92	336.37	336.38	344.82	340.56
F-Hyperbolic	335.98	334.91	334.95	340.42	338.52	337.64	345.45	341.33

Note: The optimal model according to the BIC is in bold font. Column headers refer to which bandwidth is being specified, with: GARCH refers to equation (2), GJR is for equation (3), DCS for equation (4), and *Fixed* for a fixed bandwidth.

constant bandwidth as in Harvey and Oryshchenko (2012). We consider both Gaussian and Student's t kernels, and initialize the estimation procedure with $m = 20$ observations.

The upper (lower) panels of Table 2 report values of the objective function (Bayesian Information Criterion, BIC) for all specifications. In all but two cases, the BIC is higher when a Student's t kernel is adopted. Hence, Gaussian models are preferred for this time series. Under Gaussian kernels: i) EWMA weights always provide a worse fit to the data; ii) using a fixed bandwidth is the worst choice irrespectively of the weighting scheme. Thus, departing from the specification of Harvey and Oryshchenko (2012) is desirable in this analysis. A combination of Gamma weights and GJR bandwidth returns the optimal model according to the BIC (Machado, 1993 shows that the BIC consistently selects the pseudo-optimal model in M-estimation settings like ours).



(a) Weights. (b) Logarithmic weights.

Figure 3: Left panel: estimated weights for the twenty most recent observations. Right panel: estimated logarithmic weights for the fifty most recent observations.

4.2 The impact of different weighting schemes

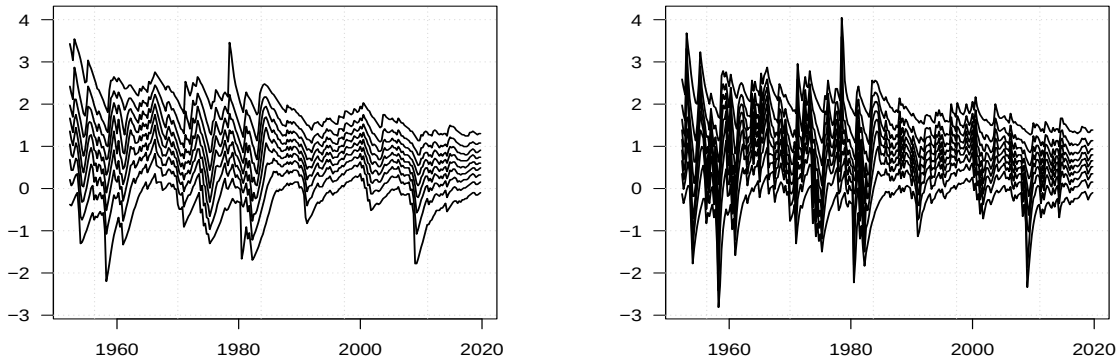
To understand the impact of different weights, we consider Gaussian Dynamic Kernel models based on a fixed bandwidth and on the weights of Table 1. The left panel of Figure 3 shows estimates of the weights for the twenty most recent observations.[9] The first estimated weight is much smaller under EWMA weights (black solid line) than under the other ones. Because this is the weight attached to the most recent information on Y, models based on EWMA weights will be less responsive to new information on GDP growth. The right panel contains the logarithm of the first fifty estimated weights. Gamma weights (red dashed line) first decrease at a slower than exponential rate and eventually become an exponentially decreasing function of the lags, thus showing the flexibility of these weights. Figure 4 plots the one-step ahead mean under these four specifications. Attention is restricted to the time period Q1:2000 to Q4:2019 to better appreciate the impact of the Dot-com bubble and of the Great Financial Crisis. Results are qualitatively identical when considering the whole dataset as well as other bandwidths. Large changes in $\hat{\mu}_{t+1|t}$ take more time to reabsorb under EWMA weights than under other weighting

[9]Precisely, Figure 3a reports the first twenty elements of $\{\tilde{\omega}_{0,T}, \ldots, \tilde{\omega}_{T,T}\}$, i.e. the time-varying weights of Section 3.2 for the last observation of the sample.

Figure 4: One-step ahead predictive mean under different weighting schemes and a fixed bandwidth. All time series refer to the quarter-on-quarter growth rate of US real GDP between Q1:2000 and Q4:2019.

schemes. This is a consequence of failing to timely incorporate new information on Y into the predictive mean. Such drawback hinders the fit of EWMA-based specifications, as unexpectedly negative (positive) values of Y lead EWMA-based models to underestimate (overestimate) subsequent realizations of Y for a substantial amount of time.

Figure 5 further studies these results by plotting the time-varying deciles of the one-step ahead distribution of GDP growth. That is, we consider quantiles for probability levels $\tau = 10\%,...,90\%$ that we extract by numerically solving the equation $\widehat{F}_{t+1|t}(x) = \tau$, for $\tau \in (0,1)$ and $\widehat{F}_{t+1|t}(\cdot)$ the estimated conditional c.d.f. of Y. We carry out this exercise using a fixed bandwidth and either EWMA (left panel) or hyperbolic (right panel) weights. Deciles based on EWMA weights behave more smoothly than those based on hyperbolic ones. As in Figure 4, spikes are quicker to reabsorb when more importance is attached to more recent information. Hence, moving from EWMA to hyperbolic weights makes the entire distribution more sensitive to new information on Y.



(a) EWMA weights. (b) Hyperbolic weights.

Figure 5: Time-varying deciles of the conditional distribution of US real GDP growth using EWMA (left panel) or hyperbolic (right panel) weights and a fixed bandwidth. The sample runs from Q1:1952 to Q4:2019.

4.3 The effect of different bandwidth processes

To study how different bandwidth dynamics influence Dynamic Kernel models, we use EWMA weights and consider the four bandwidth specifications of Table 2. Similar results were obtained using the other weighting schemes.

Figure 6 shows the one-step ahead predictive variance implied by these four models. As for the predictive mean, we focus on the sample Q1:2000 to Q4:2019 for the sake of illustration. The fixed-bandwidth specification (black solid line) returns a much smoother variance process than the others. In particular, shocks do not have a great impact on the process and their effect reabsorbs slowly. This suggests that a time-varying bandwidth is necessary to obtain a more responsive variance process.[10] The red solid (green dashed) vertical line denotes the observation date for the smallest (largest) value of ε_t in this subsample. These lines are instrumental to understand the role of the leverage parameter γ in (3) and (4). Indeed, the GARCH-based variance (red dashed line) always increases

[10]Unreported results showed that the sample auto-correlation function of h_t^2 decays much faster than that of $\sum_{i=0}^{\infty} \omega_i Y_{t-i}^2 - \hat{\mu}_{t+1|t}^2$. This result and the discussion of Figure 6 suggest that the latter is a level component for the predictive variance process while the former captures short term fluctuations.

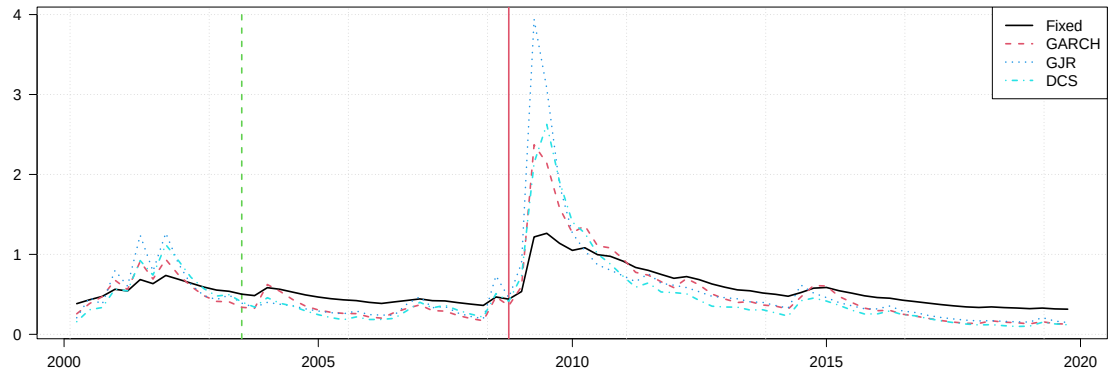
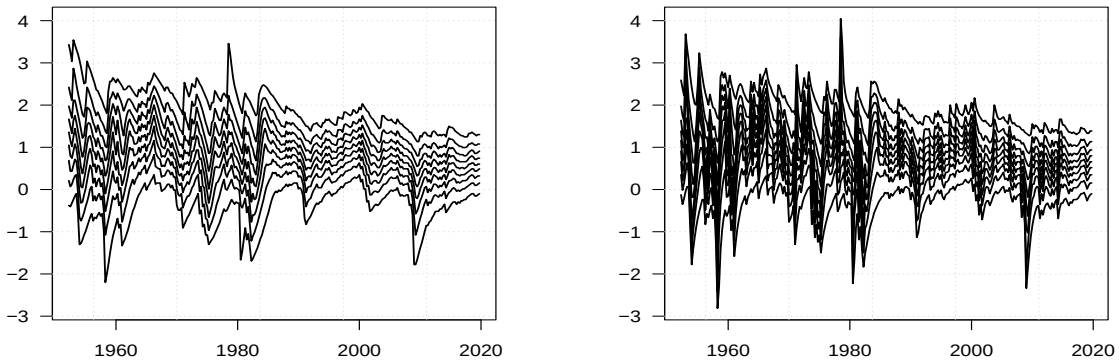


Figure 6: One-step ahead predictive variance under EWMA weights and different bandwidth processes. All time series refer to quarter-on-quarter growth rate of US real GDP between Q1:2000 and Q4:2019.



(a) Fixed bandwidth.

(b) GARCH bandwidth.

Figure 7: Time-varying deciles of the conditional distribution of US real GDP growth using either a fixed (left panel) or a GARCH (right panel) bandwidth and EWMA weights. The sample runs from Q1:1952 to Q4:2019.

after a large value of ε_t^2. Conversely, large spikes cluster after particularly negative values of ε_t under the other dynamics. Finally, the presence of ν tames the spike under the DCS specification (light-blue dash-dotted line) during the Great Financial Crisis. This is not the case for the one based on the GJR bandwidth (blue dotted line), which increase much more than the DCS one.

Figure 7 repeats the same analysis of Figure 5, this time considering EWMA weights with either a fixed (left panel) or a GARCH bandwidth (right panel). Time-varying deciles vary more rapidly when the bandwidth is dynamic, thus confirming that a dynamic band-

Table 3: Empirical rejection frequencies for unconditional and conditional coverage tests.

	Unconditional coverage				Conditional coverage			
	GARCH	GJR	DCS	Fixed	GARCH	GJR	DCS	Fixed
EWMA	0.152	0.101	0.051	0.263	0.323	0.323	0.141	0.283
Gamma	0.222	0.182	0.121	0.414	0.202	0.152	0.121	0.263
Hyperbolic	0.212	0.232	0.283	0.384	0.202	0.192	0.131	0.253
F-Hyperbolic	0.212	0.202	0.121	0.313	0.182	0.121	0.101	0.263

Note: Tests are run for $J = 99$ one-step ahead quantiles at probability levels: $1\%, 2\%, \ldots, 99\%$.

width makes the whole distribution more sensitive to new information on Y.

4.4 Model diagnostics

We carry out model diagnostics through the unconditional coverage (UC) and conditional coverage (CC) tests of Kupiec (1995) and Christoffersen (1998), respectively. Further results based on residual analysis are in Appendix E.

Given $\tau_j \in (0,1)$, both tests consider the sequence of quantile violations $\left\{z_{j,t|t-1}; t \in \mathbb{Z}\right\}$ where $z_{j,t|t-1} = \mathbb{1}\left(Y_t \leq q_{j,t|t-1}\right)$ for $q_{j,t|t-1}$ the model-based τ_j-quantile of $Y_t|\,\mathcal{F}_{t-1}$. If $q_{j,t|t-1}$ is correctly specified, violations form an i.i.d. sequence of Bernoulli random variables with success probability τ_j. The UC test considers the null hypothesis of correct coverage, i.e. $\mathbb{E}\left[z_{j,t}\right] = \tau_j$, while the CC test looks at a composite null of correct coverage and independence of the quantile violations. We test the two null hypotheses for $J = 99$ equally spaced probability levels between 1% and 99% at significance level 5%. For both tests, we report empirical rejection frequencies across probability levels in Table 3, i.e. $\sum_{j=1}^{J} \mathbb{1}\left(p_j \leq 0.05\right)/J$ for p_j the p-value of one of the tests for the τ_j- quantile.

Given a bandwidth process, EWMA weights always return the highest rejection frequencies for the CC test and the lowest ones for the UC test. Hence, problems with the

CC test arise from a lack of independence in quantile violations. This is likely due to the slowly-varying behavior of EWMA-based quantiles. Given a weighting scheme, models with a dynamic bandwidth provide better coverage than those based on a static one. Thus, departing from the approach of Harvey and Oryshchenko (2012) implies more accurate coverages for most of the cases. Figure 8 reports p-values for unconditional (upper panels) and conditional (lower panels) coverage tests. Left panels consider EWMA weights and different bandwidths, while right ones fix the bandwidth to the DCS one and compare p-values across different weighting schemes. Rejections of the two null hypotheses cluster around probability levels between sixty and ninety percent and are more frequent with a static bandwidth.

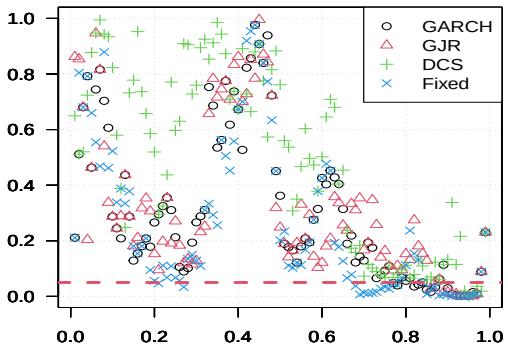
5 Density predictions

We now employ Dynamic Kernel models to predict the density of quarter-on-quarter growth rate of US real GDP. We split the sample of Section 4 into two parts and proceed with an expanding window estimation-forecasting approach. The first part runs from Q1:1947 to Q1:1980 while the second one covers the remaining portion of the sample.

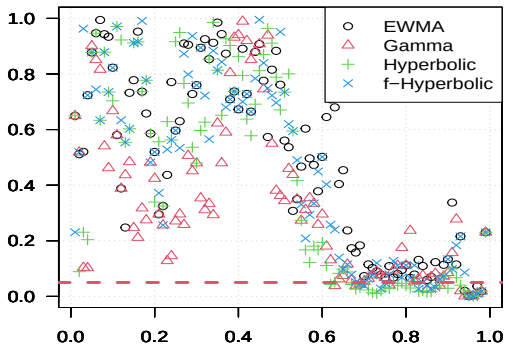
The focus is on density predictions that we asses with the weighted continuous ranked probability score (wCRPS) of Gneiting and Ranjan (2011). At time $t+1$ this error metric is given by:

$$\text{wCRPS}_{t+1} \coloneqq \int_{\mathbb{R}} w\left(u\right) \left(\widehat{F}_{t+1|t}\left(u\right) - \mathbb{1}\left\{y_{t+1} \leq u\right\}\right)^2 du,$$

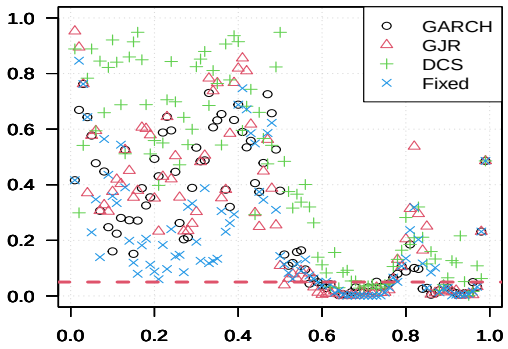
for $w : \mathbb{R} \to \mathbb{R}^+$ a weighting function which assigns more importance to certain regions of the support of the distribution. Table 4 reports the weighting functions proposed by Gneiting and Ranjan (2011) and used in our analysis. Column headers display the emphasized region of the support. We consider a Gaussian kernel and all weights and bandwidth



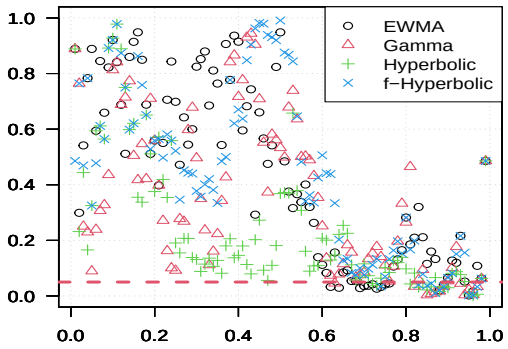
(a) Unconditional coverage, EWMA weights



(b) Unconditional coverage, DCS bandwidth



(c) Conditional coverage, EWMA weights



(d) Conditional coverage, DCS bandwidth

Figure 8: p-values for unconditional (upper panels) and conditional (lower panels) coverage tests based on $J = 99$ one-step ahead quantiles for at probability levels: $1\%, 2\%, \ldots, 99\%$ (the horizontal axis). Left panels consider EWMA weights and compare p-values across the four bandwidth specifications (see legends). Right panels use the DCS-EGARCH bandwidth and compare p-values across different weights (see legends). Red dashed lines denote the 5% level of significance.

processes of Section 4. These models are compared with the heteroskedastic MAR of Wong and Li (2001) that we specify based on $K = 2$ components, each of them being an autoregressive process of order two with ARCH(1) conditional variance.[11] Figure 9 plots the average (over time) wCRPS of each Dynamic Kernel model as a fraction of the same metric for the MAR. Values below one (the horizontal line) indicate that a Dynamic Kernel model is outperforming the benchmark. For all regions, there is at least one kernel-based model that improves upon the MAR. Results are particularly promising when one emphasizes

[11] This is the MAR that delivers the best density forecasts across nine possible specifications. In particular, we consider mixtures of $K = 2$ $AR(p)$ processes for $p = 0, 1, 2$. For each of these processes, we consider conditional variances that are either constant (the original MAR by Wong and Li, 2000), or ARCH(q) with $q = 1, 2$ (as in Wong and Li, 2001). Results on these other models are available upon request.

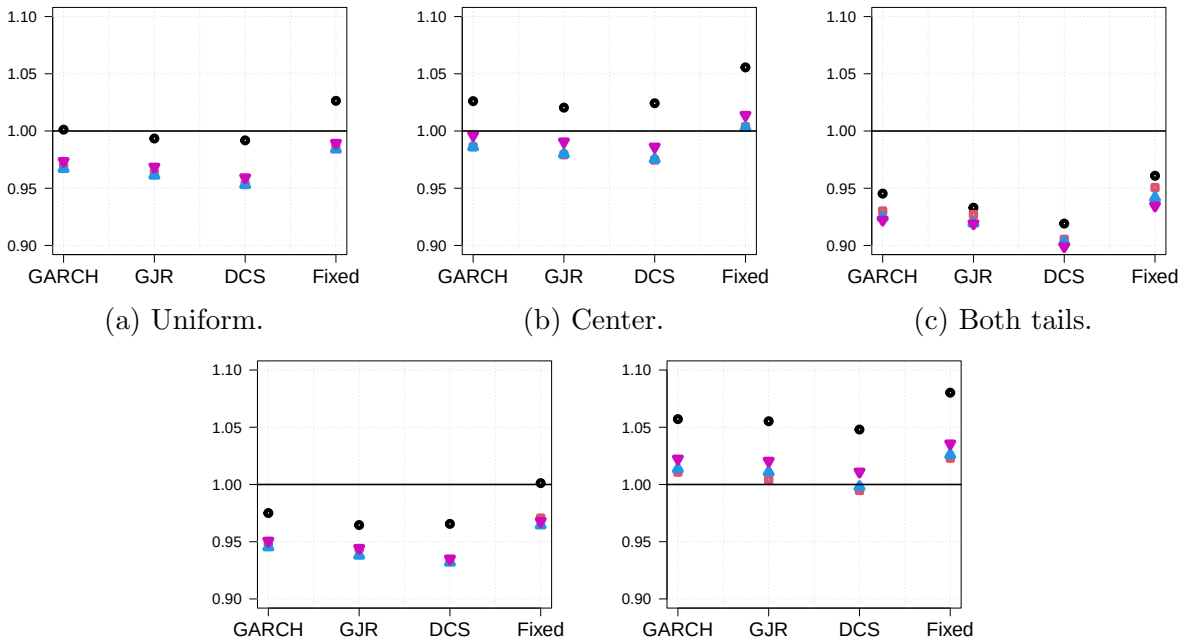
Table 4: Weighting functions for the wCRPS.

	Uniform	Center	Tails	Right tail	Left tail
$w(u)$	1	$\phi(u)$	$1-\phi(z)/\phi(0)$	$\Phi(u)$	$1-\Phi(u)$

Note: $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and cumulative distribution functions of a standard Gaussian, respectively.

either the right tail or both tails at once. Regardless of the bandwidth process, EWMA weights (black dots) yield the less accurate forecasts for any region of interest. Gamma (purple down-pointing triangles), hyperbolic (red squares) and flexible hyperbolic (blue triangles) weights consistently return very similar performances. In particular, the optimal model outperforms those based on Gamma weights by at most 1.4% for the left tail and by no more than 0.1% for the other regions. Hence, we can find Dynamic Kernel models that provide reliable density forecasts and for which the asymptotic analysis holds under less restrictive conditions . Table 5 reports the previous results along with p-values for the test of equal forecasts accuracy of Diebold and Mariano (1995) in parentheses. These are based on testing the null hypothesis of equal accuracy between forecasts implied by a Dynamic Kernel model and by the benchmark (similar results hold when running the model confidence set procedure of Hansen et al., 2011 at the 90% confidence level). The table also presents results for a non-parametric approach where standard kernel density estimation is applied to the (expanding) estimation sample, and the resulting estimator is used as one-step ahead predictive density (this is also one of the benchmarks considered by Jeon and Taylor, 2012; see also Aıt-Sahalia and Lo, 2000 for a financial application of this approach).[12] Predictions based on standard kernel density estimation are always signifi-

[12]We avoid reporting comparisons with Jeon and Taylor (2012) as their approach is more suitable when the conditioning information set is spanned by past and present information on some covariates, and not by past information on the variable of interest (a similar argument holds for other conditional KDE approaches, e.g. that of Izbicki and Lee, 2016). Nevertheless, we tried implementing their approach with



(a) Uniform. (b) Center. (c) Both tails.

(d) Right tail. (e) Left tail.

Figure 9: wCRPS for sixteen Dynamic Kernel models. For each bandwidth process, black dots refer to EWMA weights, purple down-pointing triangles to Gamma weights, red squares to hyperbolic weights and blue triangles to flexible hyperbolic ones. Results are reported as a fraction of the same error metric for the MAR benchmark.

cantly less accurate than those implied by Dynamic Kernel models. The MAR benchmark is consistently outperformed when the emphasis is either on both tails or just on the right one, especially when we move away from EWMA weights; this holds at least at the 5% level of significance. Rejections of the null of equal forecast accuracy are also possible when no particular region is emphasized at levels of significance 10% and 5%. Finally, using a dynamic rather than a fixed bandwidth always improves forecasting results (unreported results showed that these gains are significant for all weighting schemes and emphasized regions).

the latest value of GDP growth as covariate; results are available upon request. We had to use only one covariate due to the high computational cost: the procedure requires estimating $N+2$ parameters by cross-validation, for N the number of covariates (doing it within our expanding window analysis and considering only fifteen values for each parameter already required three hours on a cluster of sixteen cores).

Table 5: Density prediction results.

	Uniform				Center				Tails			
	GARCH	GJR	DCS	Fixed	GARCH	GJR	DCS	Fixed	GARCH	GJR	DCS	Fixed
EWMA	1.001	0.993	0.992	1.026	1.026	1.020	1.024	1.056	0.945	0.933	0.919	0.961
	(0.969)	(0.823)	(0.803)	(0.372)	(0.413)	(0.502)	(0.481)	(0.078)	(0.091)	(0.044)	(0.025)	(0.199)
Gamma	0.973	0.969	0.959	0.989	0.996	0.991	0.986	1.014	0.922	0.919	0.899	0.935
	(0.208)	(0.163)	(0.095)	(0.582)	(0.870)	(0.683)	(0.590)	(0.510)	(<0.001)	(0.002)	(<0.001)	(0.001)
Hyperbolic	0.969	0.963	0.953	0.987	0.986	0.979	0.975	1.004	0.930	0.927	0.905	0.951
	(0.128)	(0.081)	(0.025)	(0.492)	(0.521)	(0.339)	(0.244)	(0.851)	(0.001)	(0.002)	(<0.001)	(0.008)
F-Hyperbolic	0.967	0.961	0.953	0.984	0.986	0.980	0.975	1.003	0.925	0.919	0.903	0.942
	(0.103)	(0.067)	(0.045)	(0.386)	(0.517)	(0.363)	(0.317)	(0.881)	(<0.001)	(0.001)	(<0.001)	(0.001)
Kernel		1.105				1.098				1.120		
		(<0.001)				(0.001)				(<0.001)		

	Right tail				Left tail			
	GARCH	GJR	DCS	Fixed	GARCH	GJR	DCS	Fixed
EWMA	0.975	0.965	0.966	1.001	1.057	1.055	1.048	1.080
	(0.427)	(0.256)	(0.323)	(0.969)	(0.094)	(0.085)	(0.156)	(0.013)
Gamma	0.951	0.944	0.935	0.968	1.022	1.020	1.011	1.036
	(0.025)	(0.021)	(0.013)	(0.102)	(0.335)	(0.368)	(0.667)	(0.101)
Hyperbolic	0.949	0.944	0.934	0.971	1.011	1.004	0.995	1.023
	(0.017)	(0.012)	(0.003)	(0.122)	(0.632)	(0.863)	(0.814)	(0.263)
F-Hyperbolic	0.945	0.938	0.932	0.965	1.014	1.011	0.998	1.026
	(0.008)	(0.005)	(0.007)	(0.055)	(0.568)	(0.663)	(0.937)	(0.235)
Kernel		1.110				1.094		
		(<0.001)				(0.016)		

Note: Results are reported as the average wCRPS with respect to the same quantity for the benchmark MAR. The type of wCRPS is reported as title of each panel. p-values for the Diebold and Mariano test of equal forecast accuracy against the benchmark are in parentheses.

6 Conclusions and further lines of research

This paper introduces the family of Dynamic Kernel models. These models generalise and improve upon extant approaches to kernel density estimation for time series data. An M-estimator for model parameters is proposed and its asymptotic properties are derived under a misspecified setting. The optimal, in Kullback-Leibler terms, sequence of model-based densities can be consistently estimated from the data. An empirical illustration shows that the new models reliably track the predictive distribution of US real GDP growth.

The paper can be extended in several ways. First, multiple-component exponential models may be devised. As suggested by Granger (1980), these models mimic hyperbolic ones while simplifying estimation and inference on model parameters. Covariates can be included in Dynamic Kernel models either through the bandwidth process or through the weighting scheme. The fit in the tails can be improved by letting the bandwidth vary across different regions of the support as in variable bandwidth kernel density estimation (see the review by Markovich 2008). Finally, the paper can be extended to multivariate distributions, with bivariate ones already sufficing for predictive systemic risk measures such as the CoVaR of Adrian and Brunnermeier (2016) and the SRISK of Brownlees and Engle (2017).

References

Abramowitz, M. and I. A. Stegun (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Volume 55. US Government printing office.

Adrian, T., N. Boyarchenko, and D. Giannone (2019). Vulnerable growth. *American Economic Review 109*(4), 1263–89.

Adrian, T. and M. K. Brunnermeier (2016). CoVaR. *The American Economic Review 106*(7),

1705.

Aıt-Sahalia, Y. and A. W. Lo (2000). Nonparametric risk management and implied risk aversion. *Journal of econometrics 94*(1-2), 9–51.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics 12*(2), 171–178.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology 65*(2), 367–389.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of econometrics 73*(1), 5–59.

Black, F. (1976). Studies of stock market volatility changes. In *Meeting of the American Statistical Association Business and Economic Statistics Section*, Washington, DC.

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

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

Blasques, F., J. Ji, and A. Lucas (2016). Semiparametric score driven volatility models. *Computational Statistics & Data Analysis 100*, 58–69.

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.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*(3), 307–327.

Brandt, A. (1986). The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability 18*(1), 211–220.

Brownlees, C. and R. F. Engle (2017). SRISK: a conditional capital shortfall measure of systemic

risk. *The Review of Financial Studies 30*(1), 48–79.

Campbell, S. D. and F. X. Diebold (2005). Weather forecasting for weather derivatives. *Journal of the American Statistical Association 100*(469), 6–16.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review 39*(4), 841–862.

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

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 20*(1), 134–144.

Drost, F. C. and C. A. Klaassen (1997). Efficient estimation in semiparametric garch models. *Journal of Econometrics 81*(1), 193–221.

Elliott, G. and A. Timmermann (2016). Forecasting in economics and finance. *Annual Review of Economics 8*, 81–110.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics 20*(3), 339–350.

Engle, R. F. and G. Gonzalez-Rivera (1991). Semiparametric arch models. *Journal of Business & Economic Statistics 9*(4), 345–359.

Fermanian, J.-D. and B. Salanie (2004). A nonparametric simulated maximum likelihood estimation method. *Econometric Theory 20*(4), 701–734.

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

Francq, C. and J.-M. Zakoïan (2009). Bartlett's formula for a general class of nonlinear processes. *Journal of Time Series Analysis 30*(4), 449–465.

Francq, C. and J.-M. Zakoian (2019). *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons.

Garcin, M., J. Klein, and S. Laaribi (2023). Estimation of time-varying kernel densities and chronology of the impact of COVID-19 on financial markets. *Journal of Applied Statistics 51*(11), 2157–2177.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance 48*(5), 1779–1801.

Gneiting, T. and R. Ranjan (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics 29*(3), 411–422.

Granger, C. W. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics 14*(2), 227–238.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Hao, H.-X., J.-G. Lin, X.-F. Huang, H.-X. Wang, and Y.-Y. Zhao (2018). Estimation and application of semiparametric stochastic volatility models based on kernel density estimation and hidden markov models. *Applied Stochastic Models in Business and Industry 34*(3), 355–375.

Harvey, A. and R.-J. Lange (2017). Volatility modeling with a generalized t distribution. *Journal of Time Series Analysis 38*(2), 175–190.

Harvey, A. and G. Sucarrat (2014). Egarch models with fat tails, skewness and leverage. *Computational Statistics & Data Analysis 76*, 320–338.

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

Harvey, A. C. and T. Chakravarty (2008). Beta-t-(E)GARCH. Technical report, Faculty of Economics, University of Cambridge.

Harvey, A. C. and V. Oryshchenko (2012). Kernel density estimation for time series data. *International Journal of Forecasting 28*(1), 3–14.

Izbicki, R. and A. B. Lee (2016). Nonparametric conditional density estimation in a high-

dimensional regression setting. *Journal of Computational and Graphical Statistics 25*(4), 1297–1316.

Jeon, J. and J. W. Taylor (2012). Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association 107*(497), 66–79.

Kiley, M. T. (2022). Unemployment risk. *Journal of Money, Credit and Banking 54*(5), 1407–1424.

Koopman, S. J. and A. C. Harvey (2003). Computing observation weights for signal extraction and filtering. *Journal of Economic Dynamics and Control 27*(7), 1317–1333.

Koopman, S. J., A. Lucas, and M. Scharth (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *Review of Economics and Statistics 98*(1), 97–110.

Krengel, U. (1985). *Ergodic theorems*. Berlin: De Gruyter studies in Mathematics.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives 3*, 73–84.

Li, D. and K. Zhu (2020). Inference for asymmetric exponentially weighted moving average models. *Journal of Time Series Analysis 41*(1), 154–162.

Linero, A. and A. Rosalsky (2013). On the Toeplitz lemma, convergence in probability, and mean convergence. *Stochastic Analysis and Applications 31*(4), 684–694.

Liu, L., H. R. Moon, and F. Schorfheide (2021). Panel forecasts of country-level covid-19 infections. *Journal of Econometrics 220*(1), 2–22.

Lopez-Salido, D. and F. Loria (2024). Inflation at risk. *Journal of Monetary Economics*, 103570.

Machado, J. A. (1993). Robust model selection and M-estimation. *Econometric Theory 9*(3), 478–493.

Markovich, N. (2008). *Nonparametric analysis of univariate heavy-tailed data: research and practice*. John Wiley & Sons.

Pötscher, B. M. and I. Prucha (1997). *Dynamic nonlinear econometric models: asymptotic theory*.

Springer Science & Business Media.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics 33*(2), 659–680.

Robinson, P. M. and P. Zaffaroni (2006). Pseudo-maximum likelihood estimation of ARCH (∞) models. *The Annals of Statistics 34*(3), 1049–1074.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics 27*(3), 832–837.

Scott, D. W., R. A. Tapia, and J. R. Thompson (1980). Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. *The annals of statistics 8*(4), 820–832.

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

Sun, Y. and T. Stengos (2006). Semiparametric efficient adaptive estimation of asymmetric garch models. *Journal of Econometrics 133*(1), 373–386.

van Dijk, D., T. Teräsvirta, and P. H. Franses (2002). Smooth transition autoregressive models—a survey of recent developments. *Econometric reviews 21*(1), 1–47.

Wang, X., C. P. Tsokos, and A. Saghafi (2018). Improved parameter estimation of time dependent kernel density by using artificial neural networks. *The Journal of Finance and Data Science 4*(3), 172–182.

Wong, C. S., W.-S. Chan, and P. Kam (2009). A student t-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika 96*(3), 751–760.

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

Wong, C. S. and W. K. Li (2001). On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association 96*(455), 982–995.

Zakoian, J.-M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control 18*(5), 931–955.

Zhu, K. (2023). A new generalized exponentially weighted moving average quantile model and its statistical inference. *Journal of Econometrics 237*(1), 105510.

SUPPLEMENTARY MATERIAL

A Notation and assumptions

This appendix reports and discusses the assumptions required for the asymptotic analysis of Section 3. The following notation is employed throughout the whole supplementary material: $\|x\|_{\boldsymbol{\Theta}} \coloneqq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|x(\boldsymbol{\theta})\|$ for any function $x(\boldsymbol{\theta})$ and norm $\|\cdot\|$; $\|\mathbf{v}\| \coloneqq \max_{i=1,\ldots,n} |\mathbf{v}_i|$ for any $\mathbf{v} \in \mathbb{R}^n$; $\|\mathbf{A}\| \coloneqq \max_{i=1,\ldots,m} \sum_{j=1}^{n} |A_{i,j}|$ for any $\mathbf{A} \in \mathbb{R}^{m\times n}$; $\mathcal{K}_{i,t}(\boldsymbol{\theta}) \coloneqq \mathcal{K}\left(\frac{Y_t - Y_{t-1-i}}{h_t(\boldsymbol{\theta})}; \boldsymbol{\theta}_{\mathcal{K}}\right)$, $d_x \coloneqq \frac{d}{dx}$ and $d_x \nabla_{\mathcal{K}} \mathcal{K}(x; \boldsymbol{\theta}_{\mathcal{K}})$ is a vector of $\mathbb{R}^{d_{\mathcal{K}}}$ whose i-th entry is the derivative with respect to x of the i-th entry of $\nabla_{\mathcal{K}} \mathcal{K}(x; \boldsymbol{\theta}_{\mathcal{K}})$.[13] Positive finite scalars are denoted by $c_0, c_1, c_2, \ldots$ and their values may change from line to line. We use the notations $X_t \xrightarrow{a.s.} 0$ and $X_t = o_{a.s.}(1)$ interchangeably, and similarly for $X_t = o_p(1)$ and $X_t = o(1)$.

Assumption 4. *There exists $D > 0$ such that $\left|h_t^{-1}\right|_{\boldsymbol{\Theta}} \leq D < \infty$, a.s. $\forall t \in \mathbb{Z}$.*

Assumption 5. *i) $\mathbb{E}[|Y_t|] < \infty$; ii) There exists $\gamma > 1$ such that $t^{\gamma} \left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ as $t \to \infty$; iii) $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{E}\left[\mathcal{K}\left(\frac{Y_{t+1} - Y_t}{h_{t+1}(\boldsymbol{\theta})}; \boldsymbol{\theta}_{\mathcal{K}}\right)^{-1}\right] < \infty$.*

Assumption 6. $\mathbb{E}[\varphi_t(\boldsymbol{\theta}^*)] > \mathbb{E}[\varphi_t(\boldsymbol{\theta})]$ *for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$.*

We show that the bandwidth processes of Section 2 satisfy Assumptions 4 and 5 (ii) in Appendix B. Assumption 5 (iii) can be relaxed when using exponentially decaying weights. Indeed, existence of a positive logarithmic moment for $\mathcal{K}\left(\frac{Y_t - Y_{t-1}}{h_{t+1}(\boldsymbol{\theta})}; \boldsymbol{\theta}_{\mathcal{K}}\right)^{-1}$ suffices in this case (see Remark C.3 in Appendix C). The latter condition is implied by $\mathbb{E}[Y_t^2] < \infty$ when $\mathcal{K}(\cdot)$ is Gaussian and by $\mathbb{E}\left[|Y_t|^{\delta}\right] < \infty$ for some $\delta > 0$ in the Student's t case. Assumption 6 ensures that the objective function has an identifiable unique maximizer. As discussed in Pötscher and Prucha (1997) this condition may be violated in a misspecified context.

[13] For instance, if $\boldsymbol{\theta}_{\mathcal{K}} = (\xi, \nu)'$ for some real-valued parameters ξ and ν we have that $d_x \nabla_{\mathcal{K}} \mathcal{K}(x; \boldsymbol{\theta}_{\mathcal{K}}) = \left(d_x \partial_{\xi} \mathcal{K}(x; \boldsymbol{\theta}_{\mathcal{K}}), d_x \partial_{\nu} \mathcal{K}(x; \boldsymbol{\theta}_{\mathcal{K}})\right)'$.

When this condition is not satisfied, consistency of $\widehat{\boldsymbol{\theta}}_T$ can be established with respect to the set of minimizers of the Kullback-Leibler divergence in (8). This result follows from Lemma 4.2 in Pötscher and Prucha (1997) after noting that $\varphi_t(\boldsymbol{\theta})$ has regular level sets under Assumption 3 (Definition 4.1 in Pötscher and Prucha, 1997).

To study the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_T$, we introduce the gradient and the Hessian matrix of $\varphi_t(\boldsymbol{\theta})$:

$$\nabla\varphi_t(\boldsymbol{\theta}) = \begin{pmatrix} \nabla_{\mathcal{K}}\varphi_t(\boldsymbol{\theta}) \\ \nabla'_{\omega}\varphi_t(\boldsymbol{\theta}) \\ \nabla_h\varphi_t(\boldsymbol{\theta}) \end{pmatrix}, \qquad \nabla^2\varphi_t(\boldsymbol{\theta}) = \begin{bmatrix} \nabla^2_{\mathcal{K}\mathcal{K}}\varphi_t(\boldsymbol{\theta}) & \nabla^2_{\mathcal{K}\omega}\varphi_t(\boldsymbol{\theta}) & \nabla^2_{\mathcal{K}h}\varphi_t(\boldsymbol{\theta}) \\ \nabla^2_{\omega\mathcal{K}}\varphi_t(\boldsymbol{\theta}) & \nabla^2_{\omega\omega}\varphi_t(\boldsymbol{\theta}) & \nabla^2_{\omega h}\varphi_t(\boldsymbol{\theta}) \\ \nabla^2_{h\mathcal{K}}\varphi_t(\boldsymbol{\theta}) & \nabla^2_{h\omega}\varphi_t(\boldsymbol{\theta}) & \nabla^2_{hh}\varphi_t(\boldsymbol{\theta}) \end{bmatrix},$$

where $\nabla_{\mathcal{K}}$ and $\nabla^2_{\mathcal{K}\mathcal{K}}$ denote the gradient and the Hessian of $\varphi_t(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}_{\mathcal{K}}$, respectively. A similar notation holds for the remaining entries of $\nabla\varphi_t(\boldsymbol{\theta})$ and $\nabla^2\varphi_t(\boldsymbol{\theta})$. The next assumptions allow us to derive the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_T$.

Assumption 7. *$\boldsymbol{\theta}^*$ belongs to the interior of $\boldsymbol{\Theta}$.*

Assumption 8. *(i) There exists $\delta > 0$ such that $\mathbb{E}\left[\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta}^*)\right\}\right\|^{2+\delta}\right] < \infty$, $\mathbb{E}\left[\left|d_x\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta}^*)\right\}\right|^{8+\delta}\right] < \infty$, $\mathbb{E}\left[\left\|\nabla h_t(\boldsymbol{\theta}^*)\right\|^{4+\delta}\right] < \infty$, $\mathbb{E}\left[\mathcal{K}\left(\frac{Y_{t+1}-Y_t}{h_{t+1}(\boldsymbol{\theta}^*)};\boldsymbol{\theta}^*_{\mathcal{K}}\right)^{-(2+\delta)}\right] < \infty$ and $\mathbb{E}\left[|Y_t|^{8+\delta}\right] < \infty$;*

(ii) $\forall\,\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $\sum_{i=0}^{\infty}\left\|\nabla_{\omega}\omega_i(\boldsymbol{\theta}_{\omega})\right\| = S < \infty$.

We use Assumption 8 to show that $\mathbb{E}\left[\left\|\nabla\varphi_t(\boldsymbol{\theta}^*)\right\|^{2+\delta}\right] < \infty$ for some $\delta > 0$ (Lemma C.2 in Appendix C.3); this condition is required apply a central limit theorem (CLT) for near epoch dependent (n.e.d.) sequences, e.g. Theorem 10.2 in Pötscher and Prucha (1997). The moment condition on $\left\|\nabla h_t(\boldsymbol{\theta}^*)\right\|$ is implied by that on Y_t for all bandwidth processes of Section 2. The same holds for the moment restrictions on the derivatives of $\mathcal{K}_{i,t}(\boldsymbol{\theta})$ when

the latter is either Gaussian or Student's t. All weighting schemes of Section 2.1 satisfy Assumption 8 (ii).

Assumption 9. *(i)* $\forall\,(x,\boldsymbol{\theta}) \in \mathbb{R}\times\boldsymbol{\Theta}$, *there exist* $\widetilde{U}_{\mathcal{K}} > 0$ *and* $\widehat{U}_{\mathcal{K}} > 0$ *such that:* $\left\|\nabla^2_{\mathcal{K}\mathcal{K}}\mathcal{K}\left(x;\boldsymbol{\theta}_{\mathcal{K}}\right)\right\| < \widetilde{U}_{\mathcal{K}}$ *and* $\left|\frac{d^2}{dx^2}\mathcal{K}\left(x;\boldsymbol{\theta}_{\mathcal{K}}\right)\right| < \widehat{U}_{\mathcal{K}}$;

(ii) $\forall\boldsymbol{\theta} \in \boldsymbol{\Theta}$: $\mathbb{E}\left[\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\|^2\right] < \infty$, $\mathbb{E}\left[\left|d_x\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right|^8\right] < \infty$, $\mathbb{E}\left[\mathcal{K}\left(\frac{Y_{t+1}-Y_t}{h_{t+1}(\boldsymbol{\theta})};\boldsymbol{\theta}_{\mathcal{K}}\right)^{-2}\right] < \infty$ *and* $\mathbb{E}\left[\left\|\nabla h_t\left(\boldsymbol{\theta}\right)\right\|^4\right] < \infty$;

(iii) $\forall\,\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $\sum_{i=0}^{\infty}\left\|\nabla^2_{\omega\omega}\omega_i\left(\boldsymbol{\theta}_{\omega}\right)\right\| = S_2 < \infty$;

(iv) $\forall\,\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $\mathbb{E}\left[\left\|\nabla^2 h_t\left(\boldsymbol{\theta}\right)\right\|^2\right] < \infty$;

(v) $\mathbb{E}\left[\nabla^2\varphi_t\left(\boldsymbol{\theta}^*\right)\right]$ *is negative definite.*

Points (i) - (iv) give conditions for $\mathbb{E}\left[\left\|\nabla^2\varphi_t\right\|_{\boldsymbol{\Theta}}\right] < \infty$ (Lemma C.3 in Appendix C.3), so that the strong uniform law of large numbers of Rao (1962) applies to the Hessian process. Point (i) holds for various kernel densities such as the Student's t, and the skewed Gaussian and Student's t of Azzalini (1985) and Azzalini and Capitanio (2003), respectively; considerations made for Assumption 8 hold also for points (ii) - (iv) of Assumption 9.

Assumption 10. *(i) There exists* $\gamma > 2$ *such that:* $t^{\gamma}\left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ *as* $t\to\infty$;

(ii) For any $(x,\boldsymbol{\theta})\in\mathbb{R}\times\boldsymbol{\Theta}$, *there exist* $U_{\mathcal{K}} > 0$ *and* $\breve{U}_{\mathcal{K}} > 0$ *such that:* $\left\|\nabla_{\mathcal{K}}\mathcal{K}\left(x;\boldsymbol{\theta}_{\mathcal{K}}\right)\right\| < U_{\mathcal{K}}$ *and* $\left\|d_x\nabla_{\mathcal{K}}\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\| < \breve{U}_{\mathcal{K}}$. *Moreover,* $\mathbb{E}\left[h_t^2\left(\boldsymbol{\theta}\right)\right] < \infty$ *for any* $\boldsymbol{\theta}\in\boldsymbol{\Theta}$;

(iii) There exists $\pi > 1/2$ *such that* $t^{\pi}\sum_{i=t}^{\infty}\left|\omega_i\right|_{\boldsymbol{\Theta}} \to 0$ *and* $t^{\pi}\sum_{i=t}^{\infty}\left\|\nabla_{\omega}\omega_i\right\|_{\boldsymbol{\Theta}} \to 0$, *as* $t\to\infty$.

(iv) There exist $\tilde{\gamma} > 1/2$ *such that* $t^{\tilde{\gamma}}\left\|\nabla\hat{h}_t - \nabla h_t\right\|^2_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ *as* $t\to\infty$

Assumption 10 allows us to show that $\widehat{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}_T$ are asymptotically equivalent in probability. Hence, they share the same asymptotic distribution. Point (i) strengthens Assumption 5 (ii) and we discuss its validity for the bandwidth processes of of Section 2 in Appendix B. Considerations made for Assumption 9 (i) hold also for point (ii), while all the bandwidths of Section 2 have finite second moment under Assumption 8 (i). All weighting schemes of Section 2 satisfy point (iii) under the same conditions required for consistency (see Assumption 13 in Appendix B). The same appendix shows that all bandwidth processes satisfy point (iv). When considering the bandwidths of Section 2 along with exponentially decaying weights, points (i), (ii) and (iv) become redundant, and we can show that the two estimators are asymptotically equivalent almost surely (see Remark C.4 in Appendix C).

Assumption 11. *The sequence* $\{\nabla\varphi_t(\boldsymbol{\theta}^*)\,;\, t\in\mathbb{Z}\}$ *is near epoch dependent of size* -1 *with respect to a* ϕ*-mixing process of size* $r/(r-1)$ *for some* $r>2$.

Assumption 11 is a condition on the memory properties of $\nabla\varphi_t(\boldsymbol{\theta}^*)$ and it allows us to use a CLT for n.e.d. sequences. This kind of CLT is needed since $\nabla\varphi_t(\boldsymbol{\theta}^*)$ may not to be a martingale difference sequence under the measure P due to misspecification of the model.

The next assumption is instrumental to prove Proposition 3.1. We establish its validity for the bandwidths of Section 2 in Appendix B.

Assumption 12. *There exists* $\tilde{\gamma}>0$ *such that* $t^{\tilde{\gamma}}\left|\tilde{h}_t-h_t\right|_{\boldsymbol{\Theta}}\xrightarrow{a.s.}0$ *as* $t\to\infty$.

B Properties of the bandwidth processes

This appendix studies the properties of the bandwidth processes of Section 2. All properties are established over the probability space $(\Omega,\mathcal{F},P)$ and we omit mentioning it in what follows. The appendix also contains two additional assumptions with respect to the main body: Assumption 13 is a condition on the decay rate of weights $\{\omega_i;\in\mathbb{Z}\}$ that is required

for convergence towards the a.s. unique SE solution of the bandwidth processes in Section 2; Assumption 15 is a similar condition for the time-varying weights of Section 3.2. Both assumptions are not necessary for the general asymptotic theory of Section 3 and are only required to show that the bandwidths of Section 2 have the desired asymptotic behavior.

Lemma B.1. *Under Assumption 1, the sequence* $\varepsilon = \{\varepsilon_t;\, t \in \mathbb{Z}\}$ *is SE.*

Proof. The t-th element of ε is given by $\varepsilon_t = Y_t - \sum_{i=0}^{\infty} \omega_i Y_{t-1-i} = \sum_{l=0}^{\infty} \chi_l Y_{t-l}$ for $\chi_l = \mathbb{1}(l=0) - \mathbb{1}(l>0)\,\omega_{l-1}$, whence $\sum_{l=0}^{\infty} \chi_l < \infty$. Thus, ε_t is a measurable map of contemporaneous and past values of the SE sequence Y, so that ε is itself SE by Proposition 4.3 in Krengel (1985). □

B.1 Proof of Proposition 2.1

We start from the GARCH-like case. Lemma B.1 and Proposition 4.3 in Krengel (1985) imply that $\left\{\bar{h} + \alpha\varepsilon_t^2;\, t \in \mathbb{Z}\right\}$ is an SE sequence. Because $0 < \beta < 1$, Theorem 1 in Brandt (1986) gives us that

$$h_{t+1}^2 = \frac{\bar{h}}{1-\beta} + \alpha \sum_{s=0}^{\infty} \beta^s \varepsilon_{t-s}^2$$

is the a.s. unique SE solution of $h_{t+1}^2 = \bar{h} + \alpha\varepsilon_t^2 + \beta h_t^2$ as long as $\mathbb{E}\left[\log^+\left(\bar{h} + \alpha\varepsilon_t^2\right)\right] < \infty$ for $x^+ = \max(x, 0)$. Since $\mathbb{E}\left[\left(\bar{h} + \alpha\varepsilon_t^2\right)^\rho\right] < \infty$ for some $\rho > 0$ implies $\mathbb{E}\left[\log^+\left(\bar{h} + \alpha\varepsilon_t^2\right)\right] < \infty$ studying finiteness of this power moment suffices. When $\rho \in (0, 1)$, the c_r-inequality implies that $E\left[\left(\bar{h} + \alpha\varepsilon_t^2\right)^\rho\right] \leq \bar{h}^\rho + \alpha^\rho E\left[\varepsilon_t^{2\rho}\right]$, where $\varepsilon_t = \sum_{l=0}^{\infty} \chi_l Y_{t-l}$ as in Lemma B.1. Hence,

$$\begin{aligned}
\mathbb{E}\left[\left(|\varepsilon_t|^\rho\right)^2\right] &\leq \mathbb{E}\left[\left\{\left(\sum_{l=0}^{\infty} |\chi_l|\,|Y_{t-l}|\right)^\rho\right\}^2\right] \\
&\leq \mathbb{E}\left[\left(\sum_{l=0}^{\infty} |\chi_l|^\rho\,|Y_{t-l}|^\rho\right)^2\right]
\end{aligned}$$

$$= \sum_{l=0}^{\infty} |\chi_l|^{2\rho} \mathbb{E}\left[|Y_{t-l}|^{2\rho}\right] + 2\sum_{l=1}^{\infty}\sum_{k=0}^{\infty} |\chi_k|^{\rho} |\chi_{k+l}|^{\rho} \mathbb{E}\left[|Y_{t-k}|^{\rho} |Y_{t-k-l}|^{\rho}\right],$$
$$= \sum_{l=0}^{\infty} |\chi_l|^{\delta} \mathbb{E}\left[|Y_{t-l}|^{\delta}\right] + 2\sum_{l=1}^{\infty}\sum_{k=0}^{\infty} |\chi_k|^{\delta/2} |\chi_{k+l}|^{\delta/2} \mathbb{E}\left[|Y_{t-k}|^{\delta/2} |Y_{t-k-l}|^{\delta/2}\right],$$

where we defined $\delta \coloneqq 2\rho$ and the upper bound is finite under Assumption 1.

The proof for point (ii) is equivalent after noting that $\bar{h} + \varepsilon_t^2 [\alpha + \gamma \mathbb{1}(\varepsilon_t < 0)]$ is the generic element of an SE sequence (Lemma B.1 and Proposition 4.3 in Krengel, 1985) and $E\left[\left(\bar{h} + \varepsilon_t^2 [\alpha + \gamma \mathbb{1}(\varepsilon_t < 0)]\right)^{\rho}\right] \leq \bar{h}^{\rho} + (\alpha^{\rho} + \gamma^{\rho}) E\left[\varepsilon_t^{2\rho}\right]$. For point (iii), Lemma B.1 and Proposition 4.3 in Krengel (1985) imply that $\left\{\bar{h} + \alpha u_t + \gamma \text{sgn}(-\varepsilon_t)(u_t + 1); t \in \mathbb{Z}\right\}$ is an SE sequence. Moreover,

$$\begin{aligned}\left|\bar{h} + \alpha u_t + \gamma \text{sgn}(-\varepsilon_t)(u_t + 1)\right| &\leq \left|\bar{h}\right| + |\alpha| |u_t| + |\gamma| |\text{sgn}(-\varepsilon_t)| |(u_t + 1)| \\ &\leq \left|\bar{h}\right| + |\alpha| \max(1, \nu) + |\gamma| (\max(1, \nu) + 1),\end{aligned}$$

a.s. at any point in time (from the main body, $-1 < u_t < \nu$ so that $|u_t| < \max(1, \nu)$ a.s. and for any $t \in \mathbb{Z}$.). Thus, the random variable at the left hand side has finite positive logarithmic moment so that

$$\bar{h}_{t+1} = \frac{\bar{h}}{1 - \beta} + \sum_{s=0}^{\infty} \beta^s [\alpha u_{t-s} + \gamma \,\text{sgn}(-\varepsilon_{t-s})(u_{t-s} + 1)]$$

is the a.s. unique SE solution to (4) as long as $|\beta| < 1$.

B.2 Validity of Assumptions 4 and 5 (ii)

Assumptions 4 and 5 (ii) are verified for the smooth processes:

$$h_{t+1}^2(\boldsymbol{\theta}) = \bar{h} + \beta h_t^2(\boldsymbol{\theta}) + \alpha \varepsilon_t^2(\boldsymbol{\theta}_\omega); \tag{13}$$

$$h_{t+1}^2(\boldsymbol{\theta}) = \bar{h} + \beta h_t^2(\boldsymbol{\theta}) + \{\alpha + \gamma G(\varepsilon_t(\boldsymbol{\theta}_\omega))\} \varepsilon_t^2(\boldsymbol{\theta}_\omega); \tag{14}$$

$$\bar{h}_{t+1}(\boldsymbol{\theta}) = \bar{h} + \beta \bar{h}_t(\boldsymbol{\theta}) + \alpha u_t(\boldsymbol{\theta}) + \gamma \{2G(\varepsilon_t(\boldsymbol{\theta}_\omega)) - 1\}(u_t(\boldsymbol{\theta}) + 1), \tag{15}$$

where $G(x) = \left(1 + \exp\left\{-\frac{x}{c}\right\}\right)^{-1}$.

Proposition B.1. *Under Assumptions 1, 2 and 3 there exists $D > 0$ such that $\left|h_t^{-1}\right|_{\boldsymbol{\Theta}} \leq D < \infty$ a.s. for any $h_t(\boldsymbol{\theta})$ based on (13) to (15).*

Proof. The a.s. unique SE solutions of (13) and (14) imply that:

$$h_{t+1}(\boldsymbol{\theta}) \geq \sqrt{\frac{\bar{h}}{1-\beta}},$$

almost surely. Taking the suprema of the reciprocals, which are finite under Assumption 2, concludes the proof.

The a.s. unique SE solution to (15) entails:

$$\begin{aligned}
|\log(h_{t+1}(\boldsymbol{\theta}))| &\leq \left|\frac{\bar{h}}{1-\beta}\right| + \sum_{s=0}^{\infty} |\beta|^s |\alpha u_{t-s} + \gamma \{2G(\varepsilon_{t-s}) - 1\}(u_{t-s} + 1)| \\
&\leq \left|\frac{\bar{h}}{1-\beta}\right| + |\alpha| \sum_{s=0}^{\infty} |\beta|^s |u_{t-s}| + |\gamma| \sum_{s=0}^{\infty} |\beta|^s |u_{t-s} + 1| \\
&< \left|\frac{\bar{h}}{1-\beta}\right| + \frac{|\alpha|}{1-|\beta|}\max(1,\nu) + \frac{|\gamma|}{1-|\beta|}(\max(1,\nu) + 1) \\
&\leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \left|\frac{\bar{h}}{1-\beta}\right| + \frac{|\alpha|}{1-|\beta|}\max(1,\nu) + \frac{|\gamma|}{1-|\beta|}(\max(1,\nu) + 1) \right\},
\end{aligned}$$

a.s. and where the supremum is finite thanks to Assumption 2. Thus, the bandwidth

process is a.s. bounded and the statement holds. □

To verify Assumption 5 (ii), we consider the smooth bandwidth processes recovered from the sample $y_{0:T}$ as in Section 3:

$$\hat{h}_{t+1}^2(\boldsymbol{\theta}) = \bar{h} + \beta \hat{h}_t^2(\boldsymbol{\theta}) + \alpha \hat{\varepsilon}_t^2(\boldsymbol{\theta}_\omega); \tag{16}$$

$$\hat{h}_{t+1}^2(\boldsymbol{\theta}) = \bar{h} + \beta \hat{h}_t^2(\boldsymbol{\theta}) + \{\alpha + \gamma G(\hat{\varepsilon}_t(\boldsymbol{\theta}_\omega))\} \hat{\varepsilon}_t^2(\boldsymbol{\theta}_\omega); \tag{17}$$

$$\hat{\bar{h}}_{t+1}(\boldsymbol{\theta}) = \bar{h} + \beta \hat{\bar{h}}_t(\boldsymbol{\theta}) + \alpha \hat{u}_t(\boldsymbol{\theta}) + \gamma \{2G(\hat{\varepsilon}_t(\boldsymbol{\theta}_\omega)) - 1\}(\hat{u}_t(\boldsymbol{\theta}) + 1), \tag{18}$$

where $\hat{\varepsilon}_t(\boldsymbol{\theta}_\omega) \coloneqq \left(Y_t - \sum_{i=0}^{t-1} \omega_i(\boldsymbol{\theta}_\omega) Y_{t-1-i}\right)$ and $\hat{u}_t(\boldsymbol{\theta}) = \frac{(\nu+1)\hat{\varepsilon}_t^2(\boldsymbol{\theta}_\omega)}{\nu+\hat{\varepsilon}_t^2(\boldsymbol{\theta}_\omega)} - 1$. Showing that these processes satisfy Assumption 5 (ii) requires an assumption on the weighting scheme as well as a result on the limit behaviour of $|\hat{\varepsilon}_t - \varepsilon_t|_{\boldsymbol{\Theta}}$. The latter is established under the next assumption. Note that this condition is always satisfied by exponentially decaying weights, while $\theta > 3$ is required in the hyperbolic case.

Assumption 13. *There exists $\pi > 2$ such that $t^\pi \sum_{i=t}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} \to 0$ as $t \to \infty$.*

Lemma B.2. *Under Assumptions 1, 2, 3 and 13, there exists $\delta > 2$ such that $t^\delta |\hat{\varepsilon}_t - \varepsilon_t|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$, as $t \to \infty$.*

Proof. Observe that:

$$|\hat{\varepsilon}_t - \varepsilon_t|_{\boldsymbol{\Theta}} \leq \sum_{i=t}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} |Y_{t-1-i}|,$$

where the upper bound is a.s. finite from Assumptions 1 and 13. Hence, Assumption 13 implies $t^\delta |\hat{\varepsilon}_t - \varepsilon_t|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ for some $\delta > 2$ following arguments as in the proof of Lemma 8 in Robinson and Zaffaroni (2006). □

Lemma B.2 implies $|\hat{\varepsilon}_t - \varepsilon_t|_{\boldsymbol{\Theta}} \leq C(1+t)^{-\delta}$ for any $t \in \mathbb{N}$ and where C is a positive and a.s. finite random variable. We can now state and prove the desired result.

Proposition B.2. *Under Assumptions 1, 2, 3, 5 (i) and 13:*

(i) If $0 < \beta < 1$, *there exists* $\gamma > 1$ *such that* $t^{\gamma} \left| \hat{h}_t - h_t \right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$, *as* $t \to \infty$ *for* h_t *and* $\hat{h}_t$ *as in* (13) *and* (16), *respectively.*

(ii) If $0 < \beta < 1$ *and* $\mathbb{E}[Y_t^2] < \infty$, *there exists* $\gamma > 1$ *such that* $t^{\gamma} \left| \hat{h}_t - h_t \right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$, *as* $t \to \infty$ *for* h_t *and* $\hat{h}_t$ *as in* (14) *and* (17), *respectively.*

(iii) If $|\beta| < 1$, *there exists* $\gamma > 1$ *such that* $t^{\gamma} \left| \hat{h}_t - h_t \right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$, *as* $t \to \infty$ *for* h_t *and* $\hat{h}_t$ *as in* (15) *and* (18), *respectively.*

Proof. For point (i) we have that

$$\hat{h}_t^2(\boldsymbol{\theta}) = \bar{h}\frac{1-\beta^t}{1-\beta} + \beta^{t-1}\hat{h}_1^2 + \alpha\sum_{s=0}^{t-1}\beta^{t-s}\hat{\varepsilon}_s^2(\boldsymbol{\theta}_\omega),$$

so that:

$$\left|\hat{h}_t^2 - h_t^2\right|_{\boldsymbol{\Theta}} \leq \beta_u^{t-1}\left|\hat{h}_1^2 - h_1^2\right| + \alpha_u\sum_{s=0}^{t-1}\beta_u^{t-s}\left|\hat{\varepsilon}_s^2 - \varepsilon_s^2\right|_{\boldsymbol{\Theta}}$$

for $\alpha_u \coloneqq |\alpha|_{\boldsymbol{\Theta}}$ and $\beta_u \coloneqq |\beta|_{\boldsymbol{\Theta}}$. The first term converges to zero exponentially fast, whence multiplying it for t^{γ} has no influence on its limit behaviour. For the second one, the mean value theorem yields

$$\sum_{s=0}^{t-1}\beta_u^{t-s}\left|\hat{\varepsilon}_s^2 - \varepsilon_s^2\right|_{\boldsymbol{\Theta}} \leq 2\sum_{s=0}^{t-1}\beta_u^{t-s}\left|\breve{\varepsilon}_s\right|_{\boldsymbol{\Theta}}\left|\hat{\varepsilon}_s - \varepsilon_s\right|_{\boldsymbol{\Theta}},$$

where $\breve{\varepsilon}_s(\boldsymbol{\theta}_\omega) = \alpha\varepsilon_s(\boldsymbol{\theta}_\omega) + (1-\alpha)\hat{\varepsilon}_s(\boldsymbol{\theta}_\omega)$ for $\alpha \in (0,1)$. Thus, $|\breve{\varepsilon}_s|_{\boldsymbol{\Theta}} \leq |\varepsilon_s|_{\boldsymbol{\Theta}} + |\hat{\varepsilon}_s - \varepsilon_s|_{\boldsymbol{\Theta}}$

so that

$$t^{\gamma}\sum_{s=0}^{t-1}\beta_u^{t-s}\left|\hat{\varepsilon}_s^2-\varepsilon_s^2\right|_{\boldsymbol{\Theta}}\leq 2t^{\gamma}\sum_{s=0}^{t-1}\beta_u^{t-s}\left|\varepsilon_s\right|_{\boldsymbol{\Theta}}\left|\hat{\varepsilon}_s-\varepsilon_s\right|_{\boldsymbol{\Theta}}+2t^{\gamma}\sum_{s=0}^{t-1}\beta_u^{t-s}\left|\hat{\varepsilon}_s-\varepsilon_s\right|_{\boldsymbol{\Theta}}^2$$
$$\leq 2t^{\gamma}C\sum_{s=0}^{t-1}\beta_u^{t-s}\left(1+s\right)^{-\delta}\left|\varepsilon_s\right|_{\boldsymbol{\Theta}}+2t^{\gamma}\tilde{C}\sum_{s=0}^{t-1}\beta_u^{t-s}\left(1+s\right)^{-2\delta},$$

for some $\delta>2$ from Lemma B.2 (see the remark right after the Lemma). Assumption 5 (i) implies $\mathbb{E}\left[\left|\varepsilon_s\right|_{\boldsymbol{\Theta}}\right]<\infty$, whence Borel Cantelli Lemma entails that the upper bound goes to zero a.s. if $\sum_{t=1}^{\infty}t^{\gamma}\sum_{s=0}^{t-1}\beta^{t-s}\left(1+s\right)^{-\delta}<\infty$. For the generic term of the outer sum we have that $t^{\gamma}\sum_{s=0}^{t-1}\beta^{t-s}\left(1+s\right)^{-\delta}\leq O\left(t^{\gamma}\beta^{t}\right)+O\left(t^{\gamma-\delta}\right)$, so that the sum over t converges as long as $\delta>\gamma+1$. Because $\delta>2$, we can find a $\gamma>1$ such that $\delta>\gamma+1$. Thus we have shown that $t^{\gamma}\left|\hat{h}_{t+1}^2-h_{t+1}^2\right|_{\boldsymbol{\Theta}}\xrightarrow{a.s.}0$ for some $\gamma>1$ as $t\to\infty$. Applying the mean value theorem to $f\left(x\right)=\sqrt{x}$:

$$\left|\sqrt{\hat{h}_{t+1}^2}-\sqrt{h_{t+1}^2}\right|_{\boldsymbol{\Theta}}\leq\left|\frac{1}{2\sqrt{\breve{h}_{t+1}^2}}\right|_{\boldsymbol{\Theta}}\left|\hat{h}_{t+1}^2-h_{t+1}^2\right|_{\boldsymbol{\Theta}}\leq D\left|\hat{h}_{t+1}^2-h_{t+1}^2\right|_{\boldsymbol{\Theta}}$$

for $\breve{h}_t\left(\boldsymbol{\theta}\right)$ a mean value between $h_t\left(\boldsymbol{\theta}\right)$ and $\hat{h}_t\left(\boldsymbol{\theta}\right)$ and where the second inequality is due to Assumption 4. This concludes the proof of point (i).

For point (ii) we have that

$$\left|\hat{h}_t^2-h_t^2\right|_{\boldsymbol{\Theta}}\leq\beta_u^{t-1}\left|\hat{h}_1^2-h_1^2\right|+\alpha_u\sum_{s=0}^{t-1}\beta_u^{t-s}\left|\hat{\varepsilon}_s^2-\varepsilon_s^2\right|_{\boldsymbol{\Theta}}+\gamma_u\sum_{s=0}^{t-1}\beta_u^{t-s}\left|G\left(\hat{\varepsilon}_s\right)\hat{\varepsilon}_s^2-G\left(\varepsilon_s\right)\varepsilon_s^2\right|_{\boldsymbol{\Theta}},$$

where the proof for (i) implies that we only have to show convergence of the last term. Applying the mean value theorem to $f(x)=G(x)x^2$:

$$\left|G\left(\hat{\varepsilon}_s\right)\hat{\varepsilon}_s^2-G\left(\varepsilon_s\left(\boldsymbol{\theta}_{\omega}\right)\right)\varepsilon_s^2\right|_{\boldsymbol{\Theta}}\leq\left|d_xf(x)|_{x=\breve{\varepsilon}_s}\right|_{\boldsymbol{\Theta}}\left|\hat{\varepsilon}_s-\varepsilon_s\right|_{\boldsymbol{\Theta}},$$

where

$$\left|d_x f(x)|_{x=\breve{\varepsilon}_s}\right|_{\boldsymbol{\Theta}} = \left|2G(\breve{\varepsilon}_s)\breve{\varepsilon}_s + G'(\breve{\varepsilon}_s)\breve{\varepsilon}_s^2\right|_{\boldsymbol{\Theta}} < 2\left|\breve{\varepsilon}_s\right|_{\boldsymbol{\Theta}} + \left|\breve{\varepsilon}_s^2\right|_{\boldsymbol{\Theta}}(4c)^{-1},$$

for $\breve{\varepsilon}_s(\boldsymbol{\theta}_\omega)$ a mean value between $\varepsilon_s(\boldsymbol{\theta}_\omega)$ and $\hat{\varepsilon}_s(\boldsymbol{\theta}_\omega)$ and where $G'(x) < (4c)^{-1}$ for any $x \in \mathbb{R}$. Since, $\breve{\varepsilon}_s^2 \leq \varepsilon_s^2 + (\hat{\varepsilon}_s - \varepsilon_s)^2 + 2\left|\varepsilon_s\right|_{\boldsymbol{\Theta}}\left|\hat{\varepsilon}_s - \varepsilon_s\right|_{\boldsymbol{\Theta}}$, we can proceed as in the proof of point (i) but under the more restrictive condition that $\mathbb{E}[Y_t^2] < \infty$, which entails $\mathbb{E}[\varepsilon_t^2(\boldsymbol{\theta}_\omega)] < \infty$.

For point (iii) we start from

$$\begin{aligned}\left|\hat{\bar{h}}_t - \bar{h}_t(\boldsymbol{\theta})\right|_{\boldsymbol{\Theta}} \leq \beta_u^{t-1}\left|\hat{\bar{h}}_1 - \bar{h}_1\right| + (\alpha_u + \gamma_u)\sum_{s=0}^{t-1}\beta_u^{t-s}\left|\hat{u}_s - u_s\right|_{\boldsymbol{\Theta}} + 2\gamma_u\sum_{s=0}^{t-1}\beta_u^{t-s}\left|G(\hat{\varepsilon}_s) - G(\varepsilon_s)\right|_{\boldsymbol{\Theta}} \\ + 2\gamma_u\sum_{s=0}^{t-1}\beta_u^{t-s}\left|G(\hat{\varepsilon}_s)\hat{u}_s - G(\varepsilon_s)u_s\right|_{\boldsymbol{\Theta}},\end{aligned}$$

and apply a mean value argument based on $f(x) = \frac{(\nu+1)x^2}{\nu+x^2} - 1$, $g(x) = G(x)$ and $H(x) = G(x)f(x)$. Because these functions have bounded derivatives, Lemma B.2 implies existence of a $\gamma > 1$ such that $t^\gamma\left|\hat{\bar{h}}_t - \bar{h}_t\right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$. Combining the mean value theorem with the fact that $\bar{h}_t(\boldsymbol{\theta})$ has a bounded support we get the desired convergence for $\hat{h}_t(\boldsymbol{\theta}) = \exp\left(\hat{\bar{h}}_t(\boldsymbol{\theta})\right)$.

□

Remark B.1 (Convergence under exponentially decaying weights)**.** *The convergences in Lemma B.2 and Proposition B.2 all take place exponentially fast under exponentially decaying weights. That is, there exists $\delta > 1$ such that $\delta^t\left|\hat{\varepsilon}_t - \varepsilon_t\right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$, as $t \to \infty$ and similarly for the bandwidth processes. Moreover, $\mathbb{E}[|Y_t|^\rho] < \infty$ for some $\rho > 0$ suffices for these exponentially fast almost sure (e.a.s.) convergences. Indeed, one can show that if $\left|\hat{\varepsilon}_t - \varepsilon_t\right|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0$ and $\mathbb{E}[|Y_t|^\rho] < \infty$, then the following results hold:*

$$\left|\hat{\varepsilon}_t^2 - \varepsilon_t^2\right|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0; \quad \left|G(\hat{\varepsilon}_t)\hat{\varepsilon}_t^2 - G(\varepsilon_t)\varepsilon_t^2\right|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0; \quad \left|\hat{u}_t - u_t\right|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0;$$

$$\left|G(\hat{\varepsilon}_t)\hat{u}_t - G(\varepsilon_t)u_t\right|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0; \quad \left|G(\hat{\varepsilon}_t) - G(\varepsilon_t)\right|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0.$$

B.3 Validity of Assumption 10 (i) and (iv)

Assumption 10 (i) strengthens Assumption 5 (ii) by requiring $\gamma > 2$ instead of $\gamma > 1$ in the convergence condition $t^{\gamma} \left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$. This stricter condition is satisfied by strengthening Assumption 13 as follows:

Assumption 14. *There exists $\pi > 3$ such that $t^{\pi} \sum_{i=t}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} \to 0$ as $t \to \infty$.*

This stronger condition suffices for the next lemma, whose proof is identical to that of Lemma B.2 and is therefore omitted.

Lemma B.3. *Under Assumptions 1, 2, 3 and 14, there exists $\delta > 3$ such that $t^{\delta} |\hat{\varepsilon}_t - \varepsilon_t|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$, as $t \to \infty$.*

Using Lemma B.3 instead of Lemma B.2 in the proof of Proposition B.2 entails the desired quicker convergence, i.e. Assumption 10 (i). Proving the validity of Assumption 14 (iv) requires new arguments as it involves dealing with the gradient of the bandwidth process. Hence, we report its proof in detail.

Proposition B.3. *Under Assumptions 1 to 10 (iii) and 14:*

(i) If $0 < \beta < 1$, there exists $\tilde{\gamma} > 1/2$ such that $t^{\tilde{\gamma}} \left\|\nabla \hat{h}_t - \nabla h_t\right\|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ as $t \to \infty$ for h_t and $\hat{h}_t$ as in (13) and (16), respectively.

(ii) If $0 < \beta < 1$, there exists $\tilde{\gamma} > 1/2$ such that $t^{\tilde{\gamma}} \left\|\nabla \hat{h}_t - \nabla h_t\right\|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ as $t \to \infty$ for h_t and $\hat{h}_t$ as in (14) and (17), respectively.

(iii) If $|\beta| < 1$, there exists $\tilde{\gamma} > 1/2$ such that $t^{\tilde{\gamma}} \left\|\nabla \hat{h}_t - \nabla h_t\right\|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ as $t \to \infty$ for h_t and $\hat{h}_t$ as in (15) and (18), respectively.

Proof. For point (i) we have that:

$$\left\|\nabla\hat{h}_t - \nabla\hat{h}_t\right\|_{\boldsymbol{\Theta}} = \left\|\frac{\nabla\hat{h}_t^2}{\hat{h}_t} - \frac{\nabla h_t^2}{h_t}\right\|_{\boldsymbol{\Theta}} \leq D\left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}}\left\|\nabla h_t^2\right\|_{\boldsymbol{\Theta}} + D\left\|\nabla\hat{h}_t^2 - \nabla h_t^2\right\|_{\boldsymbol{\Theta}}, \tag{19}$$

where the expression for $\nabla h_t^2(\boldsymbol{\theta})$ (see (20) to (22) below) implies that $t^{\tilde{\gamma}}\left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}}\left\|\nabla h_t^2\right\|_{\boldsymbol{\Theta}} = o_{a.s.}(1)$ under Assumptions 8 (i) and 10 (i) for some $\tilde{\gamma} > 1/2$ using Borel Cantelli Lemma. For the second term we have that

$$\left\|\nabla\hat{h}_t^2 - \nabla h_t^2\right\|_{\boldsymbol{\Theta}} \leq \left|\hat{h}_1^2 - h_1^2\right|\left\|\nabla\beta^{t-1}\right\|_{\boldsymbol{\Theta}} + \left\|\nabla\left\{\alpha\sum_{s=0}^{t-1}\beta^{t-s}\left(\hat{\varepsilon}_s^2 - \varepsilon_s^2\right)\right\}\right\|_{\boldsymbol{\Theta}},$$

so that

$$\left|\partial_\alpha\hat{h}_t^2 - \partial_\alpha h_t^2\right|_{\boldsymbol{\Theta}} \leq \sum_{s=0}^{t-1}\beta_u^{t-s}\left|\hat{\varepsilon}_s^2 - \varepsilon_s^2\right|_{\boldsymbol{\Theta}}; \tag{20}$$

$$\left|\partial_\beta\hat{h}_t^2 - \partial_\beta h_t^2\right|_{\boldsymbol{\Theta}} \leq t\left|\hat{h}_1^2 - h_1^2\right|\beta_u^{t-2} + \alpha_u\sum_{s=0}^{t-1}(t-s)\,\beta_u^{t-s-1}\left|\hat{\varepsilon}_s^2 - \varepsilon_s^2\right|_{\boldsymbol{\Theta}}; \tag{21}$$

$$\left\|\nabla_\omega\hat{h}_t^2 - \nabla_\omega h_t^2\right\|_{\boldsymbol{\Theta}} \leq \alpha_u\sum_{s=0}^{t-1}\beta_u^{t-s}\left\|\nabla_\omega\hat{\varepsilon}_s^2 - \nabla_\omega\varepsilon_s^2\right\|_{\boldsymbol{\Theta}}, \tag{22}$$

for α_u and β_u as in the proof of Proposition B.2, and where $\partial_\alpha := \partial/\partial\alpha$ and similarly for the other derivatives. The same steps as in the proof of Proposition B.2 imply that $t^{\tilde{\gamma}}\left|\partial_\alpha\hat{h}_t^2 - \partial_\alpha h_t^2\right|_{\boldsymbol{\Theta}} = o_{a.s.}(1)$ for some $\tilde{\gamma} > 1/2$. Similar passages imply that $\alpha_u t^{\tilde{\gamma}}\sum_{s=0}^{t-1}(t-s)\,\beta_u^{t-s-1}\left|\hat{\varepsilon}_s^2 - \varepsilon_s^2\right|_{\boldsymbol{\Theta}} = o_{a.s.}(1)$ as long as $t^\delta\left|\hat{\varepsilon}_t - \varepsilon_t\right| \xrightarrow{a.s.} 0$ for some $\delta > 3/2$, which is the case from Lemma B.2. Hence, $t^{\tilde{\gamma}}\left|\partial_\beta\hat{h}_t^2 - \partial_\beta h_t^2\right|_{\boldsymbol{\Theta}} = o_{a.s.}(1)$

For (22) we have that

$$\begin{aligned}\left\|\nabla_\omega \hat{\varepsilon}_s^2 - \nabla_\omega \varepsilon_s^2\right\|_{\boldsymbol{\Theta}} &\leq 2\left\{|\varepsilon_s|_{\boldsymbol{\Theta}} \left\|\nabla_\omega \hat{\varepsilon}_s - \nabla_\omega \varepsilon_s\right\|_{\boldsymbol{\Theta}} + \left\|\nabla_\omega \varepsilon_s\right\|_{\boldsymbol{\Theta}} |\hat{\varepsilon}_s - \varepsilon_s|_{\boldsymbol{\Theta}} + |\hat{\varepsilon}_s - \varepsilon_s|_{\boldsymbol{\Theta}} \left\|\nabla_\omega \hat{\varepsilon}_s - \nabla_\omega \varepsilon_s\right\|_{\boldsymbol{\Theta}}\right\} \\ &\leq 2\left\{|\hat{\varepsilon}_s|_{\boldsymbol{\Theta}} \sum_{i=s}^{\infty} \left\|\nabla_\omega \omega_i\right\|_{\boldsymbol{\Theta}} |Y_{s-1-i}| + \left(\sum_{t=0}^{\infty} \left\|\nabla_\omega \omega_i\right\|_{\boldsymbol{\theta}} |Y_{s-1-i}|\right) s^{-\delta} C \right. \\ &\quad \left. + s^{-\delta} C \sum_{i=s}^{\infty} \left\|\nabla_\omega \omega_i\right\|_{\boldsymbol{\Theta}} |Y_{s-1-i}|\right\}\end{aligned}$$

where we have used the fact that $\nabla_\omega \varepsilon_s(\boldsymbol{\theta}) = \sum_{i=0}^{\infty} \nabla_\omega \omega_i(\boldsymbol{\theta}) Y_{s-1-i}$ (and similarly for $\nabla_\omega \hat{\varepsilon}_s(\boldsymbol{\theta})$) and the existence of some $\delta > 3$ and $C \in (0, \infty)$ a.s. such that $|\hat{\varepsilon}_s - \varepsilon_s|_{\boldsymbol{\Theta}} \leq s^{-\delta} C$ for every t following Lemma B.3. Following the same steps as in the proof of Proposition B.2, we have that $t^{\tilde{\gamma}} \left\|\nabla_\omega \hat{h}_t^2 - \nabla_\omega h_t^2\right\|_{\boldsymbol{\Theta}} = o_{a.s.}(1)$ for some $\tilde{\gamma} > 1/2$ as long as there exists some $\tilde{\pi} > 3/2$ such that $s^{\tilde{\pi}} \sum_{i=s}^{\infty} \left\|\nabla_\omega \omega_i\right\|_{\boldsymbol{\Theta}} = o_{a.s.}(1)$. The latter holds under Assumption 14 for all weighting schemes that we consider. Hence, we have shown point (i) of the proposition. The remaining points can be shown using similar arguments and by considering the same points in the proof of Proposition B.2. Hence, we omit their proofs. □

B.4 Validity of Assumption 12

We verify Assumption 12 for the processes in (13) to (15) and for their initialized counterparts

$$\tilde{h}_{t+1}^2(\boldsymbol{\theta}) = \bar{h} + \beta \tilde{h}_t^2(\boldsymbol{\theta}) + \alpha \tilde{\varepsilon}_t^2(\boldsymbol{\theta}_\omega); \tag{23}$$

$$\tilde{h}_{t+1}^2(\boldsymbol{\theta}) = \bar{h} + \beta \tilde{h}_t^2(\boldsymbol{\theta}) + \{\alpha + \gamma G(\tilde{\varepsilon}_t(\boldsymbol{\theta}_\omega))\} \tilde{\varepsilon}_t^2(\boldsymbol{\theta}_\omega); \tag{24}$$

$$\tilde{\tilde{h}}_{t+1}(\boldsymbol{\theta}) = \bar{h} + \beta \tilde{\tilde{h}}_t(\boldsymbol{\theta}) + \alpha \tilde{u}_t(\boldsymbol{\theta}) + \gamma \{2G(\tilde{\varepsilon}_t(\boldsymbol{\theta}_\omega)) - 1\}(\tilde{u}_t(\boldsymbol{\theta}) + 1), \tag{25}$$

for $\tilde{\varepsilon}_t(\boldsymbol{\theta}_\omega)$ as in Section 3.2 and $\tilde{u}_t(\boldsymbol{\theta}) = \frac{(\nu+1)\tilde{\varepsilon}_t^2(\boldsymbol{\theta}_\omega)}{\nu + \tilde{\varepsilon}_t^2(\boldsymbol{\theta}_\omega)} - 1$.

Before proceeding, note that we can write $\omega_i(\boldsymbol{\theta}_\omega) = c(\boldsymbol{\theta}_\omega) a_i(\boldsymbol{\theta}_\omega)$ and $\tilde{\omega}_{i,t}(\boldsymbol{\theta}_\omega) = c_t(\boldsymbol{\theta}_\omega) a_i(\boldsymbol{\theta}_\omega)$, for $c(\boldsymbol{\theta}_\omega) \coloneqq \left[\sum_{i=1}^{\infty} a_i(\boldsymbol{\theta}_\omega)\right]^{-1}$ and $c_t(\boldsymbol{\theta}_\omega) \coloneqq \left[\sum_{i=1}^{t} a_i(\boldsymbol{\theta}_\omega)\right]^{-1}$. Hence, the difference between time-varying and time-invariant weights reads

$$|\tilde{\omega}_{i,t}(\boldsymbol{\theta}_\omega) - \omega_i(\boldsymbol{\theta}_\omega)| = |c_t(\boldsymbol{\theta}_\omega) - c(\boldsymbol{\theta}_\omega)| a_i(\boldsymbol{\theta}_\omega),$$

for $i = 0, \ldots, t$. The next Assumption and Lemma are required to verify Assumption 12.

Assumption 15. *There exists $\tilde{\pi} > 2$ such that $t^{\tilde{\pi}} |c_t - c|_{\boldsymbol{\Theta}} \to 0$ as $t \to \infty$.*

Exponentially decaying weights always satisfy Assumption 15 while hyperbolically decaying ones do it as long as $\theta > 3$.

Lemma B.4. *Under Assumptions 1, 2, 3, 5 (i) and 15, there exists $\tilde{\delta} > 1$ such that $t^{\tilde{\delta}} |\tilde{\varepsilon}_t - \varepsilon_t|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ as $t \to \infty$.*

Proof. Observe that:

$$|\tilde{\varepsilon}_t - \varepsilon_t|_{\boldsymbol{\Theta}} \leq |c_{t-1} - c|_{\boldsymbol{\Theta}} \sum_{i=0}^{t-1} |a_i|_{\boldsymbol{\Theta}} |Y_{t-1-i}| + \sum_{i=t}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} |Y_{t-1-i}|$$

where the second term goes to zero a.s. as in the proof of Lemma B.2. For the first one, Assumption 15 implies

$$|c_{t-1} - c|_{\boldsymbol{\Theta}} \sum_{i=0}^{t-1} |a_i|_{\boldsymbol{\Theta}} |Y_{t-1-i}| \leq t^{-\tilde{\pi}} K \sum_{i=0}^{t-1} |a_i|_{\boldsymbol{\Theta}} |Y_{t-1-i}|,$$

for some $\tilde{\pi} > 2$ and scalar $K > 0$. Hence, the random variable $\sum_{t=1}^{\infty} t^{\tilde{\delta}-\tilde{\pi}} \sum_{i=0}^{t-1} |a_i|_{\boldsymbol{\Theta}} |Y_{t-1-i}|$ is integrable so that $t^{\tilde{\delta}} |c_t - c|_{\boldsymbol{\Theta}} \sum_{i=0}^{t-1} |a_i|_{\boldsymbol{\Theta}} |Y_{t-1-i}| \xrightarrow{a.s.} 0$ as $t \to \infty$ by Borel Cantelli lemma. □

Proposition B.4. *Under Assumptions 1, 2, 3, 5 (i) and 15:*

(i) If $0 < \beta < 1$, *there exists* $\tilde{\gamma} > 0$ *such that* $t^{\tilde{\gamma}} \left| \tilde{h}_t - h_t \right|_{\mathbf{\Theta}} \xrightarrow{a.s.} 0$ *as* $t \to \infty$ *for* h_t *and* $\tilde{h}_t$ *as in* (13) *and* (23), *respectively.*

(ii) If $0 < \beta < 1$ *and* $\mathbb{E}\left[Y_t^2\right] < \infty$, *there exists* $\tilde{\gamma} > 0$ *such that* $t^{\tilde{\gamma}} \left| \tilde{h}_t - h_t \right|_{\mathbf{\Theta}} \xrightarrow{a.s.} 0$ *as* $t \to \infty$ *for* h_t *and* $\tilde{h}_t$ *as in* (14) *and* (24), *respectively.*

(iii) If $|\beta| < 1$, *there exists* $\tilde{\gamma} > 0$ *such that* $t^{\tilde{\gamma}} \left| \tilde{h}_t - h_t \right|_{\mathbf{\Theta}} \xrightarrow{a.s.} 0$ *as* $t \to \infty$ *for* h_t *and* $\tilde{h}_t$ *as in* (15) *and* (25), *respectively.*

Proof. All points can be shown using arguments equivalent to those of Proposition B.2 but replacing Lemma B.2 with Lemma B.4. □

C Derivations

C.1 k-step ahead predictive moments

Proposition C.1 extends Proposition 2.2 in the main body to the general forecast horizon $k > 0$. As for the one-step ahead case, similar expressions could be derived for higher order moments at the cost of a more cumbersome notation and kernel-specific formulas.

Proposition C.1. *If* $\frac{1}{h_{t+1}} \int_{\mathbb{R}} y \mathcal{K}\left(\frac{y - y_{t-i}}{h_{t+1}}\right) dy = y_{t-i}$ *and* $\frac{1}{h_{t+1}} \int_{\mathbb{R}} (y - y_{t-i})^2 \mathcal{K}\left(\frac{y - y_{t-i}}{h_{t+1}}\right) dy = h_{t+1}^2$:

$$\hat{\mu}_{t+k|t} \coloneqq \mathbb{E}_{\widehat{P}}\left[Y_{t+k} \middle| \mathcal{F}_t\right] = \sum_{i=0}^{\infty} g^{(k)}\left(\omega_i\right) y_{t-i}; \tag{26}$$

$$\hat{\sigma}^2_{t+k|t} \coloneqq \mathrm{Var}_{\widehat{P}}\left[Y_{t+k} \middle| \mathcal{F}_t\right] = \sum_{s=0}^{k-1} g^{(s)}\left(\omega_0\right) \hat{h}^2_{t+k-s|t} + \sum_{i=0}^{\infty} g^{(k)}\left(\omega_i\right) y^2_{t-i} - \hat{\mu}^2_{t+k|t}, \tag{27}$$

where $\hat{h}^2_{t+k|t} \coloneqq \mathbb{E}_{\widehat{P}}\left[h^2_{t+k}\middle|\mathcal{F}_t\right]$ *and the functions* $g^{(k)}(\cdot)$ *are given by the difference equation*

$$\begin{aligned} g^{(1)}(\omega_i) &= \omega_i, \\ g^{(k)}(\omega_i) &= g^{(k-1)}(\omega_0)\,\omega_i + g^{(k-1)}(\omega_{i+1}), \quad \text{if } k > 1, \end{aligned} \tag{28}$$

with $g^{(k)}(\omega_i) > 0$ *for any* $i \in \mathbb{N}$ *and* $k \geq 1$*, and* $\sum_{i=0}^{\infty} g^{(k)}(\omega_i) = 1$ *for any* $k \geq 1$*.*

Proof. The proof proceeds by induction with respect to the forecast horizon $k > 0$. Let us start from $\hat{\mu}_{t+k|t} \coloneqq \mathbb{E}_{\widehat{P}}[Y_{t+k}|\mathcal{F}_t] =: \mathbb{E}^{\widehat{P}}_t[Y_{t+k}]$. When $k = 1$, (1) implies that

$$\begin{aligned} \mathbb{E}^{\widehat{P}}_t[Y_{t+1}] &= \int_{\mathbb{R}} y \frac{1}{h_{t+1}} \sum_{i=0}^{\infty} \omega_i \mathcal{K}\left(\frac{y - y_{t-i}}{h_{t+1}}\right) dy \\ &= \sum_{i=0}^{\infty} \omega_i \int_{\mathbb{R}} y \frac{1}{h_{t+1}} \mathcal{K}\left(\frac{y - y_{t-i}}{h_{t+1}}\right) dy \\ &= \sum_{i=0}^{\infty} \omega_i y_{t-i}, \end{aligned}$$

where the last equality is due to the definition of the kernel functions (see the statement of Proposition C.1). Let us assume that $\hat{\mu}_{t+k|t}$ is given by (26) and (28) in the main body; this will be our induction hypothesis. The law of iterated expectation entails

$$\begin{aligned} \mathbb{E}^{\widehat{P}}_t[Y_{t+k+1}] &= \mathbb{E}^{\widehat{P}}_t\left[\mathbb{E}^{\widehat{P}}_{t+1}[Y_{t+k+1}]\right] \\ &= \mathbb{E}^{\widehat{P}}_t\left[\sum_{i=0}^{\infty} g^{(k)}(\omega_i)\, Y_{t+1-i}\right] \\ &= g^{(k)}(\omega_0)\,\mathbb{E}^{\widehat{P}}_t[Y_{t+1}] + \sum_{i=1}^{\infty} g^{(k)}(\omega_i)\, y_{t+1-i} \\ &= g^{(k)}(\omega_0) \sum_{i=0}^{\infty} \omega_i y_{t-i} + \sum_{j=0}^{\infty} g^{(k)}(\omega_{j+1})\, y_{t-j} \\ &= \sum_{i=0}^{\infty} \left\{g^{(k)}(\omega_0)\,\omega_i + g^{(k)}(\omega_{i+1})\right\} y_{t-i} \\ &= \sum_{i=0}^{\infty} g^{(k+1)}(\omega_i)\, y_{t-i}, \end{aligned}$$

where we used the induction hypothesis between the first and the second line, the result for $k = 1$ between the third and the fourth line, and the change of index $j = i - 1$ between the fourth and the fifth line. Since the desired expression holds for $k = 1$ and because the induction hypothesis implies the thesis for any $k > 1$, the proof for $\hat{\mu}_{t+k|t}$ is concluded.

Moving to $\hat{\sigma}^2_{t+k|t} \coloneqq \mathrm{Var}_{\widehat{P}}\left[Y_{t+k} \middle| \mathcal{F}_t\right]$, if $k = 1$ we have that

$$\begin{aligned}
\hat{\sigma}^2_{t+1|t} &= \mathbb{E}^{\widehat{P}}_t\left[Y^2_{t+1}\right] - \hat{\mu}^2_{t+1|t} \\
&= \int_{\mathbb{R}} y^2 \sum_{i=0}^{\infty} \omega_i \frac{1}{h_{t+1}} \mathcal{K}\left(\frac{y - y_{t-i}}{h_{t+1}}\right) dy - \hat{\mu}^2_{t+1|t} \\
&= \hat{h}^2_{t+1|t} + \sum_{i=0}^{\infty} \omega_i y^2_{t-i} - \hat{\mu}^2_{t+1|t},
\end{aligned}$$

where the third line is due to the definition of the the kernel densities. Let us assume that (27) and (28) hold for $\hat{\sigma}^2_{t+k|t}$; again, this will be our induction hypothesis. For $\hat{\sigma}^2_{t+k+1|t}$ we have that

$$\begin{aligned}
\hat{\sigma}^2_{t+k+1|t} &= \mathbb{E}^{\widehat{P}}_t\left[Y^2_{t+k+1}\right] - \hat{\mu}^2_{t+k+1|t} \\
&= \mathbb{E}^{\widehat{P}}_t\left[\mathbb{E}^{\widehat{P}}_{t+1}\left[Y^2_{t+k+1}\right]\right] - \hat{\mu}^2_{t+k+1|t} \\
&= \mathbb{E}^{\widehat{P}}_t\left[\sum_{s=0}^{k-1} g^{(s)}(\omega_0)\, \hat{h}^2_{t+1+k-s|t+1} + \sum_{i=0}^{\infty} g^{(k)}(\omega_i)\, Y^2_{t+1-i}\right] - \hat{\mu}^2_{t+k+1|t} \\
&= \sum_{s=0}^{k-1} g^{(s)}(\omega_0)\, \hat{h}^2_{t+1+k-s|t} + \sum_{i=0}^{k} g^{(k)}(\omega_i)\, \mathbb{E}^{\widehat{P}}_t\left[Y^2_{t+1-i}\right] - \hat{\mu}^2_{t+k+1|t} \\
&= \sum_{s=0}^{k-1} g^{(s)}(\omega_0)\, \hat{h}^2_{t+1+k-s|t} + g^{(k)}(\omega_0)\left\{\hat{h}^2_{t+1|t} + \sum_{i=0}^{\infty} \omega_i y^2_{t-i}\right\} \\
&\quad + \sum_{i=1}^{k} g^{(k)}(\omega_i)\, y^2_{t+1-i} - \hat{\mu}^2_{t+k+1|t} \\
&= \sum_{s=0}^{k} g^{(s)}(\omega_0)\, \hat{h}^2_{t+1+k-s|t} + \sum_{i=0}^{\infty} \left\{g^{(k)}(\omega_0)\, \omega_i + g^{(k)}(\omega_{i+1})\right\} y^2_{t-i} - \hat{\mu}^2_{t+k+1|t}
\end{aligned}$$

$$= \sum_{s=0}^{k} g^{(s)}(\omega_0)\, \hat{h}^2_{t+1+k-s|t} + \sum_{i=0}^{\infty} g^{(k+1)}(\omega_i)\, y^2_{t-i} - \hat{\mu}^2_{t+k+1|t},$$

where we used the law of iterated expectations between the first and the second line, the induction hypothesis between the second and the third line, and the result for $k = 1$ between the fourth and the fifth line. Since the desired expression holds for $k = 1$ and because the induction hypothesis implies the thesis for any $k > 1$, the proof for $\hat{\sigma}^2_{t+k|t}$ is completed.

Finally, $\omega_i > 0$ implies $g^{(k)}(\omega_i) > 0$ for any $i \in \mathbb{N}$ and $k > 0$. Summing both sides of (28) over i and noting that $\sum_{i=0}^{\infty} \omega_i = 1$ gives us $\sum_{i=0}^{\infty} g^{(k)}(\omega_i) = \sum_{i=0}^{\infty} g^{(k-1)}(\omega_i)$ for any $k > 1$. Because $\sum_{i=0}^{\infty} g^{(1)}(\omega_i) = 1$, we have $\sum_{i=0}^{\infty} g^{(k)}(\omega_i) = 1$ for any $k > 0$. □

Remark C.1. *The expression for the k-step ahead variance is particularly relevant when* $\hat{h}^2_{t+k|t}$ *is available in closed form. While this is the case for the GARCH-like approach, the presence of indicators based on asymmetrically distributed innovations (i.e.* ε_t*) makes closed forms unavailable under the GJR and the DCS-EGARCH dynamics. Nevertheless, k-step ahead predictions of these bandwidths can be obtained by numerical integration or simulation.*

C.2 Proof of Theorem 1 (consistency)

The following lemma is instrumental to the consistency proof.

Lemma C.1. *Under Assumptions 1 to 5 (i)-(ii),* $\left|\hat{f}_{t|t-1}(y) - f_{t|t-1}(y)\right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ *for any* $y \in \mathbb{R}$ *as* $t \to \infty$.

Proof. Let us introduce $x_{i,t}(\boldsymbol{\theta}) \coloneqq (y - Y_{t-1-i})/h_t(\boldsymbol{\theta})$, $\hat{x}_{i,t}(\boldsymbol{\theta}) \coloneqq (y - Y_{t-1-i})/\hat{h}_t(\boldsymbol{\theta})$ and

$\Delta_t(y) \coloneqq \left|\hat{f}_{t|t-1}(y) - f_{t|t-1}(y)\right|_{\boldsymbol{\Theta}}$. Then we can write

$$\begin{aligned}\Delta_t(y) &\leq \left|\frac{1}{\hat{h}_t}\sum_{i=0}^{t-1}\omega_i\mathcal{K}(\hat{x}_{i,t}) - \frac{1}{h_t}\sum_{i=0}^{t-1}\omega_i\mathcal{K}(x_{i,t})\right|_{\boldsymbol{\Theta}} + \left|\frac{1}{h_t}\sum_{i=t}^{\infty}\omega_i\mathcal{K}(x_{i,t}(\boldsymbol{\theta}))\right|_{\boldsymbol{\Theta}} \\ &=: A_t(y) + D_{\mathcal{K}}D\sum_{i=t}^{\infty}|\omega_i|_{\boldsymbol{\Theta}},\end{aligned} \tag{29}$$

where $D_{\mathcal{K}} \coloneqq \left|\sup_{x\in\mathbb{R}}\mathcal{K}(x)\right|_{\boldsymbol{\Theta}} < \infty$ thanks to Assumptions 2 - 3 and the fact that $\int_{\mathbb{R}}\mathcal{K}(x;\boldsymbol{\theta}_{\mathcal{K}})\,dx < \infty$ for any $\boldsymbol{\theta}_{\mathcal{K}}$. The same assumptions also imply $D_{\mathcal{K}}D\sum_{i=t}^{\infty}|\omega_i|_{\boldsymbol{\Theta}} \to 0$ as $t \to \infty$. For $A_t(y)$ we apply the mean value theorem to the function $g : \mathcal{H} \to \mathbb{R}^+$

$$g(h) = \frac{1}{h}\sum_{i=0}^{t-1}\omega_i\mathcal{K}\left(\frac{y - Y_{t-1-i}}{h}\right),$$

to get

$$\begin{aligned}A_t(y) &\leq \left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}}\left|\breve{h}_t^{-2}\right|_{\boldsymbol{\Theta}}\sum_{i=0}^{t-1}|\omega_i|_{\boldsymbol{\Theta}}\left|\mathcal{K}(\breve{x}_{i,t}) - \frac{1}{\breve{h}_t}\mathcal{K}'(\breve{x}_{i,t})(y - Y_{t-1-i})\right|_{\boldsymbol{\Theta}} \\ &\leq \left(D^2D_{\mathcal{K}} + D^3M_{\mathcal{K}}\sum_{i=0}^{\infty}|\omega_i|_{\boldsymbol{\Theta}}|y - Y_{t-1-1}|\right)\left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} \\ &=: v_t(y)\left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}}\end{aligned}$$

for $\breve{h}_t(\boldsymbol{\theta})$ a point on the segment joining $h_t(\boldsymbol{\theta})$ and $\hat{h}_t(\boldsymbol{\theta})$, and $\breve{x}_{i,t}(\boldsymbol{\theta}) \coloneqq (y - Y_{t-1-i})/\breve{h}_t(\boldsymbol{\theta})$. Note that we have also used the fact that $\mathcal{K}(x;\cdot) \in \mathcal{C}^2(\mathbb{R})$ and bounded implies existence of $M_{\mathcal{K}} \coloneqq \left|\sup_{x\in\mathbb{R}}\mathcal{K}'(x)\right|_{\boldsymbol{\Theta}}$ under Assumption 2.

Proposition 4.3 in Krengel (1985) based on Assumption 1, and Assumption 5 (i) imply that $\{v_t(y);\, t \in \mathbb{Z}\}$ is SE with finite first moment for any $y \in \mathbb{R}$. At the same time, Assumption 5 (ii) entails that for almost any $\omega \in \Omega$ there exists a time $t_N(\omega)$ such that $t^{\gamma}\left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} < 1$ for any $t \geq t_N(\omega)$. Hence, $A_t(y) \leq t^{-\gamma}v_t(y)$ eventually and a.s., and

where the upper bound converges to zero a.s. using Borel-Cantelli Lemma. Equation (29) implies that we have shown the desired convergence for any $y \in \mathbb{R}$. □

Remark C.2 (Convergence under exponentially decaying weights)**.** *From Remark B.1,* $\gamma^t \left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ *for some* $\gamma > 1$ *under exponentially decaying weights. Hence, we can find* $\breve{\gamma} > 1$ *such that* $\breve{\gamma}^t \left|\hat{f}_{t|t-1}(y) - f_{t|t-1}(y)\right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$*, as* $t \to \infty$*, for any* $y \in \mathbb{R}$ *and under the milder condition* $\mathbb{E}\left[|Y_t|^{\rho}\right] < \infty$*.*

Proof of Theorem 1. We first show that $\left|(T-m)^{-1}\sum_{t=m+1}^{T}\hat{\varphi}_t - \mathbb{E}\left[\varphi_t\right]\right|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ as $T \to \infty$, for $\hat{\varphi}_t(\boldsymbol{\theta})$ and $\varphi_t(\boldsymbol{\theta})$ as defined in (10) and (11), respectively. Then, we verify that $\boldsymbol{\theta}^*$ is the identifiable unique maximizer of $\mathbb{E}\left[\varphi_t(\boldsymbol{\theta})\right]$.

For the first point, let us fix $m = 0$ without loss of generality and write

$$\left|T^{-1}\sum_{t=1}^{T}\hat{\varphi}_t - \mathbb{E}\left[\varphi_t\right]\right|_{\boldsymbol{\Theta}} \leq \frac{1}{T}\sum_{t=1}^{T}\left|\hat{\varphi}_t - \varphi_t\right|_{\boldsymbol{\Theta}} + \left|\frac{1}{T}\sum_{t=1}^{T}\varphi_t - \mathbb{E}\left[\varphi_t\right]\right|_{\boldsymbol{\Theta}}. \tag{30}$$

For the first term, the mean value theorem yields

$$\begin{aligned}
\left|\hat{\varphi}_t - \varphi_t\right|_{\boldsymbol{\Theta}} &\leq \left|\frac{1}{\breve{f}_{t|t-1}(Y_t)}\right|_{\boldsymbol{\Theta}} \left|\hat{f}_{t|t-1}(Y_t) - f_{t|t-1}(Y_t)\right|_{\boldsymbol{\Theta}} \\
&< \frac{1}{\alpha}\left|\frac{1}{f_{t|t-1}(Y_t)}\right|_{\boldsymbol{\Theta}} \left|\hat{f}_{t|t-1}(Y_t) - f_{t|t-1}(Y_t)\right|_{\boldsymbol{\Theta}}, \\
&\leq \frac{1}{\alpha}\left|\frac{h_t}{\omega_0 \mathcal{K}\left(\frac{Y_t - Y_{t-1}}{h_t}\right)}\right|_{\boldsymbol{\Theta}} \left|\hat{f}_{t|t-1}(Y_t) - f_{t|t-1}(Y_t)\right|_{\boldsymbol{\Theta}}
\end{aligned}$$

where $\breve{f}_{t|t-1}(\boldsymbol{\theta}; Y_t) = \alpha f_{t|t-1}(\boldsymbol{\theta}; Y_t) + (1-\alpha)\hat{f}_{t|t-1}(\boldsymbol{\theta}; Y_t)$ for $\alpha \in (0,1)$ so that

$\breve{f}_{t|t-1}(\boldsymbol{\theta}; Y_t) > \alpha f_{t|t-1}(\boldsymbol{\theta}; Y_t)$. Looking at the proof of Lemma C.1, we see that

$$
\begin{aligned}
|\hat{\varphi}_t - \varphi_t|_{\boldsymbol{\Theta}} \leq \frac{1}{\alpha} \left| \frac{h_t}{\omega_0 \mathcal{K}\left(\frac{Y_t - Y_{t-1}}{h_t}\right)} \right|_{\boldsymbol{\Theta}} \left| \hat{h}_t - h_t \right|_{\boldsymbol{\Theta}} \frac{1}{\breve{h}_t} \left(D\, D_{\mathcal{K}} + D^2 M_{\mathcal{K}} \sum_{i=0}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} |y - Y_{t-1-i}| \right) \\
+ \frac{1}{\alpha} \left| \frac{D_{\mathcal{K}}}{\omega_0 \mathcal{K}\left(\frac{Y_t - Y_{t-1}}{h_t}\right)} \right|_{\boldsymbol{\Theta}} \sum_{i=t}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} \\
\leq \frac{\left| \hat{h}_t - h_t \right|_{\boldsymbol{\Theta}}}{\left| \mathcal{K}\left(\frac{Y_t - Y_{t-1}}{h_t}\right) \right|_{\boldsymbol{\Theta}}} \left(c_1 + c_2 \sum_{i=0}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} |y - Y_{t-1-i}| \right) + c_3 \left| \frac{1}{\mathcal{K}\left(\frac{Y_t - Y_{t-1}}{h_t}\right)} \right|_{\boldsymbol{\Theta}} \sum_{i=t}^{\infty} |\omega_i|_{\boldsymbol{\Theta}}
\end{aligned}
$$

where we used the fact that $\breve{h}_t \geq \beta h_t$ for some $\beta \in (0,1)$. From the proof of Lemma C.1, $\left| \hat{h}_t - h_t \right|_{\boldsymbol{\Theta}} (c_1 + c_2 \sum_{i=0}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} |y - Y_{t-1-i}|) = o_{a.s.}(1)$, while $\sum_{i=t}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} = o(1)$ by construction. Moreover, $\left| K\left(\frac{Y_t - Y_{t-1}}{h_t}\right)^{-1} \right|_{\boldsymbol{\Theta}}$ is an SE sequence with finite first moment from Assumption 5 (iii). Hence, Toeplitz Lemma (see Theorem 1.1 in Linero and Rosalsky, 2013) implies that $T^{-1} \sum_{t=1}^{T} |\hat{\varphi}_t - \varphi_t|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0$ as $T \to \infty$.

For the second term in (30), note that $\varphi_t(\boldsymbol{\theta})$ forms an SE sequence of continuous functions on the compact set $\boldsymbol{\Theta}$. Hence, for any $\theta \in \boldsymbol{\Theta}$, $|\varphi_t(\boldsymbol{\theta})| \leq \left| \log\left(f_{t|t-1}(Y_t)\right) \right|_{\boldsymbol{\Theta}} < \infty$ a.s. where the upper bound is integrable from Assumption 5 (iii).[14] Thus, this term converges to zero almost surely from the uniform law of large numbers of Rao (1962).

Point (ii) is a consequence of Assumption 6 after noting that $\mathbb{E}[\varphi_t(\boldsymbol{\theta})]$ is continuous over the compact space $\boldsymbol{\Theta}$. □

Remark C.3 (Consistency under exponentially decaying weights)**.** *Under exponentially*

[14] In particular,

$$\log\left(f_{t|t-1}(Y_t;\boldsymbol{\theta})\right) = \log^+\left(f_{t|t-1}(Y_t;\boldsymbol{\theta})\right) - \log^-\left(f_{t|t-1}(Y_t;\boldsymbol{\theta})\right) = \log^+\left(f_{t|t-1}(Y_t;\boldsymbol{\theta})\right) + \log^+\left(\frac{1}{f_{t|t-1}(Y_t;\boldsymbol{\theta})}\right),$$

so that $\left|\log\left(f_{t|t-1}(Y_t;\boldsymbol{\theta})\right)\right| \leq \log^+\left(f_{t|t-1}(Y_t;\boldsymbol{\theta})\right) + \log^+\left(\frac{1}{f_{t|t-1}(Y_t;\boldsymbol{\theta})}\right)$ where $x^+ = \max(x, 0)$ and $x^- = -x^+$. Assumptions 3 and 4 imply boundedness of the first quantity while the second one has finite expectation under Assumption 5 (ii).

decaying weights, $\mathbb{E}\left[\log^{+}\left(\mathcal{K}\left(\frac{Y_t - Y_{t-1}}{h_{t+1}^2(\boldsymbol{\theta})}; \boldsymbol{\theta}_{\mathcal{K}}\right)^{-1}\right)\right] < \infty$ *suffices for* $\widehat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}^*$. *Indeed,* $\left|f_{t|t-1}^{-1}(y)\right|_{\boldsymbol{\Theta}} \left|\hat{f}_{t|t-1}(y) - f_{t|t-1}(y)\right|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0$ *as* $t \to \infty$ *under this milder condition thanks to Lemma 2.1 in Straumann and Mikosch (2006). This e.a.s. convergence is enough for* $\frac{1}{T-m}\sum_{t=m+1}^{T} |\hat{\varphi}_t - \varphi_t|_{\boldsymbol{\Theta}} \xrightarrow{a.s.} 0, \qquad as\ T \to \infty.$

C.3 Proof of Theorem 2

The proof of of Theorem 2 requires three intermediate results.

Lemma C.2. *Under Assumptions 1, 3, 4 and 8,* $\mathbb{E}\left[\left\|\nabla\varphi_t(\boldsymbol{\theta}^*)\right\|^{2+\delta}\right] < \infty$.

Proof. We prove the lemma for each of the three sub-vectors of $\nabla\varphi_t(\boldsymbol{\theta}^*)$. To improve readability, we suppress the dependence of $f_{t|t-1}(\boldsymbol{\theta}; Y_t)$ on Y_t. From the expression in Appendix F we have that

$$\begin{aligned}\left\|\nabla_{\mathcal{K}}\varphi_t(\boldsymbol{\theta})\right\| &\leq \frac{1}{\sum_{i=0}^{\infty}\omega_i(\boldsymbol{\theta}_\omega)\mathcal{K}_{i,t}(\boldsymbol{\theta})}\sum_{i=0}^{\infty}\omega_i(\boldsymbol{\theta}_\omega)\left\|\nabla_{\mathcal{K}}\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\| \\ &= \frac{1}{\sum_{i=0}^{\infty}\omega_i(\boldsymbol{\theta}_\omega)\mathcal{K}_{i,t}(\boldsymbol{\theta})}\sum_{i=0}^{\infty}\omega_i(\boldsymbol{\theta}_\omega)\mathcal{K}_{i,t}(\boldsymbol{\theta})\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\right\|,\end{aligned}$$

where $\mathcal{K}_{i,t}(\boldsymbol{\theta}) := \mathcal{K}\left(\frac{Y_t - Y_{t-1-i}}{h_t(\boldsymbol{\theta})}; \boldsymbol{\theta}_{\mathcal{K}}\right)$ and we used the fact that $\nabla_{\mathcal{K}}\mathcal{K}_{i,t}(\boldsymbol{\theta}) = \mathcal{K}_{i,t}(\boldsymbol{\theta})\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}$. Let us define the random variable

$$\xi_{i,t}(\boldsymbol{\theta}) := \frac{\omega_i(\boldsymbol{\theta}_\omega)\mathcal{K}_{i,t}(\boldsymbol{\theta})}{\sum_{i=0}^{\infty}\omega_i(\boldsymbol{\theta}_\omega)\mathcal{K}_{i,t}(\boldsymbol{\theta})},$$

which is such that $\xi_{i,t}(\boldsymbol{\theta}) \in (0,1)$ a.s. and $\sum_{i=0}^{\infty}\xi_{i,t}(\boldsymbol{\theta}) = 1$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $t \in \mathbb{Z}$. Thus,

$$\mathbb{E}\left[\left\|\nabla_{\mathcal{K}}\varphi_t(\boldsymbol{\theta})\right\|^{2+\delta}\right] = \mathbb{E}\left[\left(\sum_{i=0}^{\infty}\xi_{i,t}(\boldsymbol{\theta})\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\right\|\right)^{2+\delta}\right],$$

where the right hand side is finite as long as $\mathbb{E}\left[\sum_{i=0}^{\infty}\xi_{i,t}^{2+\delta}(\boldsymbol{\theta})\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\right\|^{2+\delta}\right]<\infty$. Since $\xi_{i,t}(\boldsymbol{\theta})<\xi_{i,t}^{2+\delta}(\boldsymbol{\theta})$ by definition, we have that

$$\begin{aligned}\mathbb{E}\left[\sum_{i=0}^{\infty}\xi_{i,t}(\boldsymbol{\theta})^{2+\delta}\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\right\|^{2+\delta}\right] &< \mathbb{E}\left[\sum_{i=0}^{\infty}\xi_{i,t}(\boldsymbol{\theta})\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\right\|^{2+\delta}\right]\\ &< \sum_{i=0}^{\infty}\mathbb{E}\left[\xi_{i,t}(\boldsymbol{\theta})\right]\left(\mathbb{E}\left[\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\right\|^{(2+\delta)\left(1+\tilde{\delta}\right)}\right]\right)^{1/(1+\tilde{\delta})},\end{aligned}$$

where we applied Hölder's inequality with Hölder's conjugates $\left(1+\tilde{\delta}\right)$ and $\frac{1+\tilde{\delta}}{\tilde{\delta}}$ for $\tilde{\delta}>0$. Hence, Assumption 8 (i) and $\sum_{i=0}^{\infty}\xi_{i,t}(\boldsymbol{\theta})\equiv 1$ imply $\mathbb{E}\left[\left\|\nabla_{\mathcal{K}}\varphi_t(\boldsymbol{\theta})\right\|^{2+\delta}\right]<\infty$.

Assumptions 3, 4 and 8 (ii), and the expression for $\nabla_{\omega}\varphi_t(\boldsymbol{\theta})$ imply

$$\begin{aligned}\left\|\nabla_{\omega}\varphi_t(\boldsymbol{\theta})\right\| \leq D\left\|\nabla_{\omega}h_t(\boldsymbol{\theta})\right\| + D^2\left\|\nabla_{\omega}h_t(\boldsymbol{\theta})\right\|\sum_{i=0}^{\infty}\xi_{i,t}(\boldsymbol{\theta})\left|Y_t-Y_{t-1-i}\right|\left|d_x\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\right|\\ + \frac{D_{\mathcal{K}}S}{\omega_0\mathcal{K}_{0,t}(\boldsymbol{\theta})},\end{aligned}$$

for $\xi_{i,t}(\boldsymbol{\theta})$ as before and where we used the fact that $d_x\mathcal{K}_{i,t}(\boldsymbol{\theta})=\mathcal{K}_{i,t}(\boldsymbol{\theta})\,d_x\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}$ for $d_x\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\coloneqq\frac{d}{dx}\log\left\{\mathcal{K}(x;\boldsymbol{\theta}_{\mathcal{K}})\right\}$ when evaluated at $x=\frac{Y_t-Y_{t-1-i}}{h_t(\boldsymbol{\theta})}$.

Taking powers and applying a Hölder's argument we get

$$\begin{aligned}\left\|\nabla_{\omega}\varphi_t(\boldsymbol{\theta})\right\|^{2+\delta} \leq 3^{1+\delta}\Bigg\{D^{4+\delta}\left\|\nabla_{\omega}h_t(\boldsymbol{\theta})\right\|^{2+\delta}\left(\sum_{i=0}^{\infty}\xi_{i,t}(\boldsymbol{\theta})\left|Y_t-Y_{t-1-i}\right|\left|d_x\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\right|\right)^{2+\delta}\\ + \left\|\nabla_{\omega}h_t(\boldsymbol{\theta})\right\|^{2+\delta} + \frac{D_{\mathcal{K}}^{2+\delta}S^{2+\delta}}{\omega_0^{2+\delta}\mathcal{K}_{0,t}^{2+\delta}(\boldsymbol{\theta})}\Bigg\}.\end{aligned}$$

Using the same steps as in the proof for $\nabla_{\mathcal{K}}\varphi_t(\boldsymbol{\theta})$, Assumption 8 (i) implies that the upper bound has finite expectation when $\boldsymbol{\theta}=\boldsymbol{\theta}^*$. An identical approach shows that $\mathbb{E}\left[\left\|\nabla_h\varphi_t(\boldsymbol{\theta}^*)\right\|^{2+\delta}\right]<\infty$. Thus we have proven the lemma. □

Lemma C.3. *Under Assumptions 1, 2, 3, 4, 8 and 9 (i)-(iv),* $\mathbb{E}\left[\left\|\nabla^2\varphi_t\right\|_{\boldsymbol{\Theta}}\right]<\infty$.

Proof. The proof proceeds by considering each of the blocks defining the Hessian matrix. For the sake of readability, we omit Y_t from the arguments of $f_{t|t-1}(Y_t, \boldsymbol{\theta})$. Before starting, it is convenient to recall that $\|\mathbf{x}\mathbf{y}'\| \leq \|\mathbf{x}\| \|\mathbf{y}\| n$ for any $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$.

Let us consider $\nabla^2_{\mathcal{K}\mathcal{K}}\varphi_t(\boldsymbol{\theta})$. From the expressions in Appendix F we get:

$$\left\|\nabla^2_{\mathcal{K}\mathcal{K}}\varphi_t(\boldsymbol{\theta})\right\| \leq \frac{1}{\left|\omega_0(\boldsymbol{\theta})\mathcal{K}_{0,t}(\boldsymbol{\theta})\right|_{\boldsymbol{\Theta}}} \sum_{i=0}^{\infty} \omega_i(\boldsymbol{\theta}_\omega) \left\|\nabla^2_{\mathcal{K}\mathcal{K}}\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\| + \left\|\nabla_{\mathcal{K}}\varphi_t(\boldsymbol{\theta})\right\|^2 d_{\mathcal{K}},$$

where both terms are integrable under Assumption 9 (i) – (ii) and following the proof of Lemma C.2.

For $\nabla^2_{\mathcal{K}\omega}\varphi_t(\boldsymbol{\theta})$ we have:

$$\begin{aligned}
\left\|\nabla^2_{\mathcal{K}\omega}\varphi_t(\boldsymbol{\theta})\right\| \leq\ & d_\omega \left\|\nabla_{\mathcal{K}}\varphi_t(\boldsymbol{\theta})\right\| \left\|\nabla_\omega\varphi_t(\boldsymbol{\theta})\right\| + \frac{d_\omega}{f_{t|t-1}(\boldsymbol{\theta})} \frac{1}{h_t(\boldsymbol{\theta})} \sum_{i=0}^{\infty} \left\|\nabla_{\mathcal{K}}\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\| \left\|\nabla_\omega\omega_i(\boldsymbol{\theta})\right\| \\
& + \frac{d_\omega}{f_{t|t-1}(\boldsymbol{\theta})} \frac{\left\|\nabla_\omega h_t(\boldsymbol{\theta})\right\|}{h_t^3(\boldsymbol{\theta})} \sum_{i=0}^{\infty} \omega_i(\boldsymbol{\theta}_\omega) \left\|d_x\nabla_{\mathcal{K}}\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\| \left|Y_t - Y_{t-1-i}\right| \\
& + \frac{d_\omega}{f_{t|t-1}(\boldsymbol{\theta})} \frac{\left\|\nabla_\omega h_t(\boldsymbol{\theta})\right\|}{h_t^2(\boldsymbol{\theta})} \sum_{i=0}^{\infty} \omega_i(\boldsymbol{\theta}_\omega) \left\|\nabla_{\mathcal{K}}\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\| \\
=:\ & A_t(\boldsymbol{\theta}) + B_t(\boldsymbol{\theta}) + C_t(\boldsymbol{\theta}) + D_t(\boldsymbol{\theta}).
\end{aligned}$$

Proceeding as in the proof of Lemma C.2, Assumption 9 (ii) implies $\mathbb{E}[A_t(\boldsymbol{\theta})] < \infty$. For $B_t(\boldsymbol{\theta})$ we note that

$$B_t(\boldsymbol{\theta}) \leq \frac{c_0}{\mathcal{K}_{0,t}(\boldsymbol{\theta})} \sum_{i=0}^{\infty} \left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}(\boldsymbol{\theta})\right\}\right\| \left\|\nabla_\omega\omega_i(\boldsymbol{\theta}_\omega)\right\|,$$

where the upper bound is integrable from Assumptions 8 (ii) and 9 (ii). For $C_t(\boldsymbol{\theta})$, first

note that

$$\begin{aligned}\left\|d_x\nabla_{\mathcal{K}}\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\| &= \left\|\mathcal{K}'_{i,t}\left(\boldsymbol{\theta}\right)\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}+\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)d_x\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\| \\ &\leq \left|\mathcal{K}'_{i,t}\left(\boldsymbol{\theta}\right)\right|\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\|+\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\left\|d_x\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\| \\ &= \mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\left\{\left|d_x\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right|\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\|+\left\|d_x\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\|\right\}\end{aligned}$$

so that

$$\begin{aligned}C_t\left(\boldsymbol{\theta}\right) &\leq c_0\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|\sum_{i=0}^{\infty}\xi_{i,t}\left(\boldsymbol{\theta}\right)\left|Y_t-Y_{t-1-i}\right|\left|d_x\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right|\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\| \\ &\quad + c_0\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|\sum_{i=0}^{\infty}\xi_{i,t}\left(\boldsymbol{\theta}\right)\left|Y_t-Y_{t-1-i}\right|\left\|d_x\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\| \\ &=: C_{1,t}\left(\boldsymbol{\theta}\right)+C_{2,t}\left(\boldsymbol{\theta}\right).\end{aligned}$$

For $C_{1,t}\left(\boldsymbol{\theta}\right)$, repeated use of Hölder's inequality yields (omitting powers outside of the expectations, for the sake of readability)

$$\begin{aligned}\mathbb{E}\left[C_{1,t}\left(\boldsymbol{\theta}\right)\right] < \mathbb{E}\left[\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|^4\right]\sum_{i=0}^{\infty}\Big\{&\xi_{i,t}\left(\boldsymbol{\theta}\right)\mathbb{E}\left[\left|Y_t-Y_{t-1-i}\right|^{8+\delta}\right]\mathbb{E}\left[\left|d_x\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right|^8\right] \\ &\mathbb{E}\left[\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\|^{2,}\right]\Big\}\end{aligned}$$

where the right hand side is finite thanks to Assumptions 8 (i) and 9 (ii). A similar argument entails $\mathbb{E}\left[C_{2,t}\left(\boldsymbol{\theta}\right)\right] < \infty$. Moving on to $D_t\left(\boldsymbol{\theta}\right)$ we get

$$\begin{aligned}D_t\left(\boldsymbol{\theta}\right) &\leq c_0\frac{\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|}{\sum_{i=0}^{\infty}\omega_i\left(\boldsymbol{\theta}_\omega\right)\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)}\sum_{i=0}^{\infty}\omega_i\left(\boldsymbol{\theta}\omega\right)\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\| \\ &= c_0\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|\sum_{i=0}^{\infty}\xi_{i,t}\left(\boldsymbol{\theta}\right)\left\|\nabla_{\mathcal{K}}\log\left\{\mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right)\right\}\right\|\end{aligned}$$

so that $\mathbb{E}\left[D_t\left(\boldsymbol{\theta}\right)\right] < \infty$ from Assumption 9 (ii). Hence, $\mathbb{E}\left[\left\|\nabla^2_{\mathcal{K}\omega}\varphi_t\left(\boldsymbol{\theta}\right)\right\|\right] < \infty$. The proof for $\nabla^2_{\mathcal{K}h}\varphi_t\left(\boldsymbol{\theta}\right)$ is identical and we omit it. For $\nabla^2_{\omega\omega}\varphi_t\left(\boldsymbol{\theta}\right)$ we get

$$
\begin{aligned}
\left\|\nabla^2_{\omega\omega}\varphi_t\left(\boldsymbol{\theta}\right)\right\| \leq\ & d_\omega \left\|\nabla_\omega \varphi_t\left(\boldsymbol{\theta}\right)\right\|^2 + \frac{1}{f_{t|t-1}\left(\boldsymbol{\theta}\right)} \frac{\left\|\nabla^2_{\omega\omega} h_t\left(\boldsymbol{\theta}\right)\right\|}{h_t^2\left(\boldsymbol{\theta}\right)} \sum_{i=0}^{\infty} \omega_i\left(\boldsymbol{\theta}_\omega\right) \left|\mathcal{K}'_{i,t}\left(\boldsymbol{\theta}\right)\right| \left|Y_t - Y_{t-1-i}\right| \\
& + \frac{2d_\omega}{f_{t|t-1}\left(\boldsymbol{\theta}\right)} \frac{\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|}{h_t^2\left(\boldsymbol{\theta}\right)} \sum_{i=0}^{\infty} \left\|\nabla_\omega \omega_i\left(\boldsymbol{\theta}_\omega\right)\right\| \mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right) \\
& + \frac{1}{f_{t|t-1}\left(\boldsymbol{\theta}\right)} \frac{\left\|\nabla^2_{\omega\omega} h_t\left(\boldsymbol{\theta}\right)\right\|}{h_t^2\left(\boldsymbol{\theta}\right)} \sum_{i=0}^{\infty} \omega_i\left(\boldsymbol{\theta}_\omega\right) \mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right) \\
& + \frac{1}{f_{t|t-1}\left(\boldsymbol{\theta}\right)} \sum_{i=0}^{\infty} \frac{\left\|\nabla^2_{\omega\omega} \omega_i\left(\boldsymbol{\theta}_\omega\right)\right\|}{h_t\left(\boldsymbol{\theta}\right)} \mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right) \\
& + \frac{2d_\omega}{f_{t|t-1}\left(\boldsymbol{\theta}\right)} \frac{\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|}{h_t^3\left(\boldsymbol{\theta}\right)} \sum_{i=0}^{\infty} \left\|\nabla_\omega \omega_i\left(\boldsymbol{\theta}_\omega\right)\right\| \left|\mathcal{K}'_{i,t}\left(\boldsymbol{\theta}\right)\right| \left|Y_t - Y_{t-1-i}\right| \\
& + \frac{4d_\omega}{f_{t|t-1}\left(\boldsymbol{\theta}\right)} \frac{\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|^2}{h_t^4\left(\boldsymbol{\theta}\right)} \sum_{i=0}^{\infty} \omega_i\left(\boldsymbol{\theta}_\omega\right) \left|\mathcal{K}'_{i,t}\left(\boldsymbol{\theta}\right)\right| \left|Y_t - Y_{t-1-i}\right| \\
& + \frac{d_\omega}{f_{t|t-1}\left(\boldsymbol{\theta}\right)} \frac{\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|^2}{h_t^5\left(\boldsymbol{\theta}\right)} \sum_{i=0}^{\infty} \omega_i\left(\boldsymbol{\theta}_\omega\right) \left|\mathcal{K}''_{i,t}\left(\boldsymbol{\theta}\right)\right| \left|Y_t - Y_{t-1-i}\right|^2 \\
& + \frac{2d_\omega}{f_{t|t-1}\left(\boldsymbol{\theta}\right)} \frac{\left\|\nabla_\omega h_t\left(\boldsymbol{\theta}\right)\right\|^2}{h_t^3\left(\boldsymbol{\theta}\right)} \sum_{i=0}^{\infty} \omega_i\left(\boldsymbol{\theta}_\omega\right) \mathcal{K}_{i,t}\left(\boldsymbol{\theta}\right),
\end{aligned}
$$

for $\mathcal{K}''_{i,t}\left(\boldsymbol{\theta}\right) \coloneqq \frac{d^2}{dx^2}\mathcal{K}\left(x;\boldsymbol{\theta}_\mathcal{K}\right)$ evaluated at $x = \frac{Y_t - Y_{t-1-i}}{h_t(\boldsymbol{\theta})}$. All these terms can be shown to have finite expectation under Assumptions 8 and 9 (i)-(iv). The proofs always rely on arguments already seen and therefore are omitted. Showing $\mathbb{E}\left[\left\|\nabla^2_{\omega h}\varphi_t\left(\boldsymbol{\theta}\right)\right\|\right] < \infty$ and $\mathbb{E}\left[\left\|\nabla^2_{hh}\varphi_t\left(\boldsymbol{\theta}\right)\right\|\right] < \infty$ implies writing an upper bound similar to that for $\left\|\nabla^2_{\omega\omega}\varphi_t\left(\boldsymbol{\theta}\right)\right\|$. Hence, both properties can be shown using similar arguments as before.

All the previous results hold for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Moreover, $\varphi_t\left(\boldsymbol{\theta}\right) \in \mathcal{C}^2\left(\boldsymbol{\Theta}\right)$ for $\boldsymbol{\Theta}$ compact. Hence, $\left\|\nabla^2\varphi_t\right\|_{\boldsymbol{\Theta}}$ exists finite and has finite first moment. This concludes the proof. □

Lemma C.4. *Under Assumptions 1 to 10,* $\frac{1}{\sqrt{T-m}} \left\|\sum_{t=m+1}^{T}\left(\nabla\varphi_t - \nabla\widehat{\varphi}_t\right)\right\|_{\boldsymbol{\Theta}} \xrightarrow{p} 0.$

Proof. Without loss of generality, assume that $m = 0$ and write

$$\begin{aligned}\frac{1}{\sqrt{T}}\left\|\sum_{t=1}^{T}\nabla\varphi_t - \sum_{t=1}^{T}\nabla\widehat{\varphi}_t\right\|_{\boldsymbol{\Theta}} &\leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\|\nabla\varphi_t - \nabla\widehat{\varphi}_t\right\|_{\boldsymbol{\Theta}} \\ &\leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(H_{1,t} + H_{2,t} + H_{3,t}\right),\end{aligned}$$

for

$$\begin{aligned}H_{1,t} &:= \frac{\left\|\nabla f_{t|t-1}\right\|_{\boldsymbol{\Theta}}}{\tilde{\alpha}^2\left|f_{t|t-1}^2\right|_{\boldsymbol{\Theta}}}\left|f_{t|t-1} - \hat{f}_{t|t-1}\right|_{\boldsymbol{\Theta}} \\ H_{2,t} &:= \left|\frac{1}{f_{t|t-1}}\right|_{\boldsymbol{\Theta}}\left\|\nabla f_{t|t-1} - \nabla\hat{f}_{t|t-1}\right\|_{\boldsymbol{\Theta}} \\ H_{3,t} &:= \left|\frac{1}{\tilde{\alpha}^2 f_{t|t-1}^2}\right|_{\boldsymbol{\Theta}}\left|f_{t|t-1} - \hat{f}_{t|t-1}\right|_{\boldsymbol{\Theta}}\left\|\nabla f_{t|t-1} - \nabla\hat{f}_{t|t-1}\right\|_{\boldsymbol{\Theta}},\end{aligned}$$

where the term $\left|\tilde{\alpha}^2 f_{t|t-1}^2\right|_{\boldsymbol{\Theta}}$ is due to a mean value argument as in the proof of Theorem 1. From the proof of Lemma C.1, Assumption 10 (i) implies existence of some $\alpha > 1$ such that $t^{\alpha}\left|f_{t|t-1} - \hat{f}_{t|t-1}\right|_{\boldsymbol{\Theta}} = o_{a.s.}(1)$. Because the random variable $\left|f_{t|t-1} - \hat{f}_{t|t-1}\right|_{\boldsymbol{\Theta}}$ has bounded support, there exists a positive scalar c_0 such that $\left|f_{t|t-1} - \hat{f}_{t|t-1}\right|_{\boldsymbol{\Theta}} \leq t^{-\alpha}c_0$ for any t, so that

$$H_{1,t} \leq c_0 t^{-\alpha}\frac{\left\|\nabla\varphi_{t|t-1}\right\|_{\boldsymbol{\Theta}}}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}},$$

whence

$$
\begin{aligned}
P\left\{T^{-1/2}\sum_{t=1}^{T}H_{1,t}>\varepsilon\right\} &\leq P\left\{T^{-1/3}\sum_{t=1}^{T}H_{1,t}^{2/3}>\varepsilon^{2/3}\right\} \\
&\leq \frac{c_1}{T^{1/3}}\sum_{t=1}^{T}t^{-2\alpha/3}\mathbb{E}\left[\frac{\left\|\nabla\varphi_{t|t-1}\right\|_{\boldsymbol{\Theta}}^{2/3}}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}^{2/3}}\right] \\
&\leq \frac{c_1}{T^{1/3}}\sum_{t=1}^{T}t^{-2\alpha/3}\left(\mathbb{E}\left[\left\|\nabla\varphi_{t|t-1}\right\|_{\boldsymbol{\Theta}}^{2}\right]\right)^{1/3}\left(\mathbb{E}\left[\frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}\right]\right)^{2/3} \\
&\leq \frac{c_2}{T^{1/3}}\int_{1}^{T}t^{-2\alpha/3}dt \\
&\leq \frac{c_3}{T^{1/3}}T^{1-2\alpha/3},
\end{aligned}
$$

where we used Hölder's inequality between the second and the third line, and the fact that $\sum_{t=1}^{T}t^{-2\alpha/3}\leq\int_{t}^{T}t^{-2\alpha/3}dt=c_4\left(T^{1-2\alpha/3}-1\right)$ along with Assumptions 9 (ii) and 10 (ii) between the third and the fourth one. Note that the upper bound converges to zero since $\alpha>1$, whence $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}H_{1,t}=o_p(1)$. We study the limit behaviour of $P\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}H_{2,t}>\varepsilon\right\}$ separately for each entry of $\nabla f_{t|t-1}\left(\boldsymbol{\theta}\right)$. For $\nabla_{\mathcal{K}}f_{t|t-1}\left(\boldsymbol{\theta}\right)$ we have that

$$
\begin{aligned}
\left\|\nabla_{\mathcal{K}}f_{t|t-1}-\nabla_{\mathcal{K}}\hat{f}_{t|t-1}\right\|_{\boldsymbol{\Theta}} &\leq \left\|\frac{1}{h_t}\sum_{i=t}^{\infty}\omega_i\nabla_{\mathcal{K}}\mathcal{K}\left(x_{i,t}\right)\right\|_{\boldsymbol{\Theta}}+\left\|\frac{1}{h_t}\sum_{i=0}^{t-1}\omega_i\nabla_{\mathcal{K}}\mathcal{K}\left(x_{i,t}\right)-\frac{1}{\hat{h}_t}\sum_{i=0}^{t-1}\omega_i\nabla_{\mathcal{K}}\mathcal{K}\left(\hat{x}_{i,t}\right)\right\|_{\boldsymbol{\Theta}} \\
&\leq U_{\mathcal{K}}D\sum_{i=t}^{\infty}\left|\omega_i\right|_{\boldsymbol{\Theta}}+\left(D^2U_{\mathcal{K}}+D^3\breve{U}_{\mathcal{K}}\sum_{i=0}^{\infty}\left|\omega_i\right|_{\boldsymbol{\Theta}}\left|Y_t-Y_{t-1-i}\right|\right)\left|\hat{h}_t-h_t\right|_{\boldsymbol{\Theta}} \\
&= c_0\sum_{i=t}^{\infty}\left|\omega_i\right|_{\boldsymbol{\Theta}}+v_t\left|\hat{h}_t-h_t\right|_{\boldsymbol{\Theta}},
\end{aligned}
$$

for $x_{i,t}\left(\boldsymbol{\theta}\right)$ and $\hat{x}_{i,t}\left(\boldsymbol{\theta}\right)$ as in the proof of Lemma C.1, $v_t\coloneqq(c_1+c_2\sum_{i=0}^{\infty}\left|\omega_i\right|_{\boldsymbol{\Theta}}\left|Y_t-Y_{t-1-i}\right|)$, and we used a mean value expansion of $g(h)=\frac{1}{h}\sum_{i=0}^{t-1}\omega_i\mathcal{K}\left(\frac{Y_t-Y_{t-1-i}}{h}\right)$. Thus, we get that

$$
\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}\left\|\nabla_{\mathcal{K}}f_{t|t-1}-\nabla_{\mathcal{K}}\hat{f}_{t|t-1}\right\|_{\boldsymbol{\Theta}}\leq\frac{c_1}{\sqrt{T}}\sum_{t=1}^{T}\frac{t^{-\pi}}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}+\frac{C}{\sqrt{T}}\sum_{t=1}^{T}t^{-\gamma}\frac{v_t}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}},\quad(31)
$$

where Assumptions 10 (i) – (iii) imply $\left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} \leq t^{-\gamma} C$ for some $\gamma > 2$ and an a.s. positive and finite random variable C, and $\sum_{i=t}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} \leq t^{-\pi} c_0$ for some $\pi > 1/2$. For the first term, Markov inequality entails

$$P\left\{\frac{c_1}{\sqrt{T}}\sum_{t=1}^{T}\frac{t^{-\pi}}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} > \varepsilon\right\} < \frac{c_2}{\sqrt{T}}\sum_{t=1}^{T} t^{-\pi}\mathbb{E}\left[\frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}\right] \leq \frac{c_3}{\sqrt{T}}T^{1-\pi},$$

where we used Assumptions 9 (ii) and 10 (ii), and the relation $\sum_{t=1}^{T} t^{-\pi} \leq \int_1^T t^{-\pi} dt$ to get an upper bound that converges to zero when $\pi > 1/2$. For the second term in (31), we see that

$$\begin{aligned} P\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} t^{-\gamma}\frac{v_t}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} > \varepsilon\right\} &\leq P\left\{\frac{1}{T^{1/4}}\sum_{t=1}^{T} t^{-\gamma/2}\frac{v_t^{1/2}}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}^{1/2}} > \varepsilon^{1/2}\right\} \\ &< \frac{c_0}{T^{1/4}}\sum_{t=1}^{T} t^{-\gamma/2}\mathbb{E}\left[\frac{v_t^{1/2}}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}^{1/2}}\right] \\ &\leq \frac{c_1}{T^{1/4}}T^{1-\gamma/2}, \end{aligned}$$

where we used Cauchy-Schwarz inequality, Assumptions 8 (i) and 9 (ii), and the previous relation between sums and integrals to derive the upper bound. The latter converges to zero under Assumption 10 (i), so that $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}\left\|\nabla_{\mathcal{K}} f_{t|t-1} - \nabla_{\mathcal{K}}\hat{f}_{t|t-1}\right\|_{\boldsymbol{\Theta}} = o_p(1)$.

For $\nabla_\omega f_{t|t-1}(\boldsymbol{\theta})$ we can write

$$\begin{aligned} \left\|\nabla_\omega f_{t|t-1} - \nabla_\omega \hat{f}_{t|t-1}\right\|_{\boldsymbol{\Theta}} \leq & \left\|\frac{1}{h_t}\sum_{i=t}^{\infty}\mathcal{K}(x_{i,t})\nabla_\omega\omega_i\right\|_{\boldsymbol{\Theta}} + \left\|\frac{\nabla_\omega h_t}{h_t^2}\sum_{i=t}^{\infty}\omega_i\mathcal{K}(x_{i,t})\right\|_{\boldsymbol{\Theta}} \\ & + \left\|\frac{\nabla_\omega h_t}{h_t^3}\sum_{i=t}^{\infty}\omega_i\mathcal{K}'(x_{i,t})(Y_t - Y_{t-1-i})\right\|_{\boldsymbol{\Theta}} \\ & + \left\|\frac{1}{\hat{h}_t}\sum_{i=0}^{t-1}\mathcal{K}(\hat{x}_{i,t})\nabla_\omega\omega_i - \frac{1}{h_t}\sum_{i=0}^{t-1}\mathcal{K}(x_{i,})\nabla_\omega\omega_i\right\|_{\boldsymbol{\Theta}} \\ & + \left\|\frac{\nabla_\omega\hat{h}_t}{\hat{h}_t^2}\sum_{i=0}^{t-1}\omega_i\mathcal{K}(\hat{x}_{i,t}) - \frac{\nabla_\omega h_t}{h_t^2}\sum_{i=0}^{t-1}\omega_i\mathcal{K}(x_{i,t})\right\|_{\boldsymbol{\Theta}} \\ & + \left\|\sum_{i=0}^{t-1}\omega_i(Y_t - Y_{t-1-i})\left\{\frac{\nabla_\omega\hat{h}_t}{\hat{h}_t^3}\mathcal{K}'(\hat{x}_{i,t}) - \frac{\nabla_\omega h_t}{h_t^3}\mathcal{K}'(x_{i,t})\right\}\right\|_{\boldsymbol{\Theta}} \end{aligned}$$

$$= A_t + B_t + C_t + D_t + E_t + F_t.$$

We examine one term at a time starting from A_t:

$$\begin{aligned} P\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}A_t > \varepsilon\right\} &\leq P\left\{\frac{c_0}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\left|\omega_0\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}}\right\} \\ &\leq \frac{c_1}{\sqrt{T}}\sum_{t=1}^{T} t^{-\pi}\mathbb{E}\left[\frac{1}{\left|\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}}\right] \\ &\leq \frac{c_2}{\sqrt{T}}T^{1-\pi}, \end{aligned}$$

thanks to Markov inequality, and Assumptions 8 (ii) and 9 (ii). The upper bound converges to zero under Assumption 10 (iii). A similar proof based on the same assumptions shows that $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}B_t = o_p(1)$.

For C_t, the same arguments as in the proof of Lemma B.2 imply

$$\begin{aligned} \frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}\left\|\frac{\nabla_\omega h_t}{h_t^3}\sum_{i=t}^{\infty}\omega_i\mathcal{K}'\left(x_{i,t}\right)\left(Y_t - Y_{t-1-i}\right)\right\|_{\boldsymbol{\Theta}} &\leq c_0\frac{\left\|\nabla_\omega h_t\right\|_{\boldsymbol{\Theta}}}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}\sum_{i=t}^{\infty}\left|\omega_i\right|_{\boldsymbol{\Theta}}\left|Y_t - Y_{t-1-i}\right| \\ &\leq c_0\frac{\left\|\nabla_\omega h_t\right\|_{\boldsymbol{\Theta}}}{\left|\omega_0\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}}t^{-\pi}, \end{aligned}$$

so that

$$\begin{aligned} P\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}}C_t > \varepsilon\right\} &\leq P\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}c_0\frac{\left\|\nabla_\omega h_t\right\|_{\boldsymbol{\Theta}}}{\left|\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}}t^{-\pi} > \varepsilon\right\} \\ &\leq \frac{c_1}{\sqrt{T}}\sum_{t=1}^{T} t^{-\pi}\mathbb{E}\left[\frac{\left\|\nabla_\omega h_t\right\|_{\boldsymbol{\Theta}}}{\left|\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}}\right] \\ &\leq \frac{c_2}{\sqrt{T}}T^{1-\pi}, \end{aligned}$$

where we used Cauchy-Schwarz inequality along with Assumption 9 (ii). For D_t we have

that:

$$\begin{aligned}
\frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} D_t &\leq \frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\theta}}} \left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} \sum_{i=0}^{t-1} \left\|\nabla \omega_i\right\|_{\boldsymbol{\Theta}} \left| \frac{\mathcal{K}\left(\breve{x}_{i,t}\right)}{\breve{h}_t^2} + \frac{\mathcal{K}'\left(\breve{x}_{i,t}\right)}{\breve{h}_t^3} \left(Y_t - Y_{t-1-i}\right) \right|_{\boldsymbol{\Theta}} \\
&\leq \frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} \frac{\left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}}}{\left|\breve{h}_t\right|_{\boldsymbol{\Theta}}} \sum_{i=0}^{t-1} \left\|\nabla \omega_i\right\|_{\boldsymbol{\Theta}} \left(c_0 + c_1 \left|Y_t - Y_{t-1-i}\right|\right) \\
&\leq \frac{c_0}{\left|\omega_0 \mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}} \left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} \sum_{i=0}^{t-1} \left\|\nabla \omega_i\right\|_{\boldsymbol{\Theta}} \left(c_1 + c_2 \left|Y_t - Y_{t-1-i}\right|\right), \\
&\leq \frac{c_3 C t^{-\gamma}}{\left|\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}} \sum_{i=0}^{\infty} \left\|\nabla \omega_i\right\|_{\boldsymbol{\Theta}} \left(c_1 + c_2 \left|Y_t - Y_{t-1-i}\right|\right)
\end{aligned}$$

where the first inequality is due to a mean value expansion around $\breve{h}_t\left(\boldsymbol{\theta}\right) = \alpha h_t\left(\boldsymbol{\theta}\right) + (1 - \alpha)\hat{h}_t\left(\boldsymbol{\theta}\right)$ for some $\alpha \in (0,1)$, the fact that $\breve{h}_t\left(\boldsymbol{\theta}\right) \geq \alpha h_t\left(\boldsymbol{\theta}\right)$ to move from the second to the third line, and Assumption 10 (i) to obtain the last upper bound for $C \in (0, \infty)$ a.s. and some $\gamma > 2$. Let $v_t \coloneqq \sum_{i=0}^{\infty} \left\|\nabla \omega_i\right\|_{\boldsymbol{\Theta}} \left(c_1 + c_2 \left|Y_t - Y_{t-1-i}\right|\right)$, so that

$$P\left\{ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{t^{-\gamma}}{\left|\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}} v_t > \varepsilon \right\} \leq \frac{c_0}{\sqrt{T}} \sum_{t=1}^{T} t^{-\gamma} \mathbb{E}\left[\frac{v_t}{\left|\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}}\right] \leq \frac{c_1}{\sqrt{T}} T^{1-\gamma},$$

where we used Markov and Caquchy-Schwarz inequality along with Assumptions 8 and 9 (ii). Hence we get $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} D_t = o_p(1)$. Moving to E_t we see that

$$\begin{aligned}
E_t &\leq \left\|\nabla_\omega h_t\right\|_{\boldsymbol{\Theta}} \left| \frac{1}{\hat{h}_t^2} \sum_{i=0}^{t-1} \omega_i \mathcal{K}\left(\hat{x}_{i,t}\right) - \frac{1}{h_t^2} \sum_{i=0}^{t-1} \omega_i \mathcal{K}\left(x_{i,t}\right) \right|_{\boldsymbol{\Theta}} + \left\|\nabla_\omega \hat{h}_t - \nabla_\omega h_t\right\|_{\boldsymbol{\Theta}} \left| \frac{1}{\hat{h}_t^2} \sum_{i=0}^{t-1} \omega_i \mathcal{K}\left(\hat{x}_{i,t}\right) \right|_{\boldsymbol{\Theta}} \\
&\leq c_0 \frac{\left\|\nabla_\omega h_t\right\|_{\boldsymbol{\Theta}}}{\left|h_t\right|_{\boldsymbol{\Theta}}} \left|\hat{h}_t - h_t\right|_{\boldsymbol{\Theta}} \sum_{i=0}^{\infty} \left|\omega_i\right|_{\boldsymbol{\Theta}} \left(c_1 + c_2 \left|Y_t - Y_{t-1-i}\right|\right) + c_3 \frac{\left\|\nabla_\omega \hat{h}_t - \nabla_\omega h_t\right\|_{\boldsymbol{\Theta}}}{\left|\hat{h}_t\right|_{\boldsymbol{\Theta}}} \\
&\leq c_0 C \frac{\left\|\nabla_\omega h_t\right\|_{\boldsymbol{\Theta}}}{\left|h_t\right|_{\boldsymbol{\Theta}}} t^{-\gamma} v_t + c_3 \tilde{C} t^{-\tilde{\gamma}} \left|\frac{h_t}{\hat{h}_t}\right|_{\boldsymbol{\Theta}} \frac{1}{\left|h_t\right|_{\boldsymbol{\Theta}}}
\end{aligned}$$

thanks to a mean value argument similar to that for D_t, and where the last upper bound is

due to Assumption 10 (i) – (iv) for C and $\tilde{C}$ two a.s. positive and finite random variables. The above upper bound implies that:

$$\begin{aligned}\frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} E_t &\leq \frac{c_0 C}{\left|\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}} \left\|\nabla_\omega h_t\right\|_{\boldsymbol{\Theta}} t^{-\gamma} v_t + \frac{c_3 \tilde{C}}{\left|\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}} t^{-\tilde{\gamma}} \left(C t^{-\gamma} + 1\right) \\ &= c_0 C E_{1,t} + c_3 \tilde{C} C E_{2,t} + c_3 \tilde{C} C E_{3,t},\end{aligned}$$

where we used the fact that Assumption 10 (i) implies $\left|\frac{h_t}{\hat{h}_t}\right|_{\boldsymbol{\Theta}} \leq (Ct^{-\gamma} + 1)$ for C and γ as before. For $E_{1,t}$ we get that

$$\begin{aligned}P\left\{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} E_{1,t} > \varepsilon\right\} &\leq P\left\{\frac{1}{T^{1/4}} \sum_{t=1}^{T} E_{1,t}^{1/2} > \varepsilon^{1/2}\right\} \\ &\leq \frac{c_0}{T^{1/4}} \sum_{t=1}^{T} t^{-\gamma/2} \mathbb{E}\left[\frac{\left\|\nabla_\omega h_t\right\|^{1/2} v_t^1/2}{\left|\mathcal{K}_{0,t}\right|_{\boldsymbol{\Theta}}^{1/2}}\right] \\ &\leq \frac{c_1}{T^{1/4}} T^{1-\gamma/2},\end{aligned}$$

where we applied Markov and Cauchy-Schwarz inequalities (the latter repeatedly) along with Assumptions 8 (i) and 9 (ii). This upper bound entails $\frac{c_0 C}{\sqrt{T}} \sum_{t=1}^{T} E_{1,t} = o_p(1)$. Similar passages imply that $\frac{c_3 C}{\sqrt{T}} \sum_{t=1}^{T} E_{2,t} = o_p(1)$ and $\frac{c_3 C \tilde{C}}{\sqrt{T}} \sum_{t=1}^{T} E_{3,t} = o_p(1)$, so that $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} E_t = o_p(1)$. Similar steps also imply that $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} F_t = o_p(1)$ under Assumptions 8 (i), 9 (ii) and 10. Hence, we showed that $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} \left\|\nabla_\omega f_{t|t-1} - \nabla_\omega \hat{f}_{t|t-1}\right\|_{\boldsymbol{\Theta}} = o_p(1)$. The proof for $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{1}{\left|f_{t|t-1}\right|_{\boldsymbol{\Theta}}} \left\|\nabla_h f_{t|t-1} - \nabla_h \hat{f}_{t|t-1}\right\|_{\boldsymbol{\Theta}}$ is identical and we omit it. Thus, we get that $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} H_{2,t} = o_p(1)$.

For $H_{3,t} = \left|\frac{1}{\alpha^2 f_{t|t-1}^2}\right|_{\boldsymbol{\Theta}} \left|f_{t|t-1} - \hat{f}_{t|t-1}\right|_{\boldsymbol{\Theta}} \left\|\nabla f_{t|t-1} - \nabla \hat{f}_{t|t-1}\right\|_{\boldsymbol{\Theta}}$ we note that $\left|\frac{1}{\alpha^2 f_{t|t-1}}\right|_{\boldsymbol{\Theta}} \left|f_{t|t-1} - \hat{f}_{t|t-1}\right|_{\boldsymbol{\Theta}} = o_{a.s.}(1)$ using Borel Cantelli lemma. Hence,there exists

a random variable $C \in (0, \infty)$ a.s. such that:

$$H_{3,t} \leq \left| \frac{C}{f_{t|t-1}} \right|_{\boldsymbol{\Theta}} \left\| \nabla f_{t|t-1} - \nabla \hat{f}_{t|t-1} \right\|_{\boldsymbol{\Theta}} = C H_{2,t}.$$

Thus, $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} H_{3,t} = o_p(1)$ and we have shown the lemma. □

Remark C.4 (Convergence of the score under exponentially decaying weights)**.** *Remark C.3 and Assumption 5 (ii) imply* $\left| \frac{1}{\hat{f}_{t|t-1}} - \frac{1}{f_{t|t-1}} \right|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0$ *as* $t \to \infty$ *under exponentially decaying weights. Similarly,* $\left\| \nabla \hat{f}_{t|t-1} - \nabla f_{t|t-1} \right\|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0$ *as* $t \to \infty$ *under this setting. Since this property is implied by Remark B.1, Assumption 10 (i, iii, iv) is redundant in this case. Lemma TA.14 in Blasques et al. (2022) then implies that* $\|\nabla \varphi_t - \nabla \hat{\varphi}_t\|_{\boldsymbol{\Theta}} \xrightarrow{e.a.s.} 0$ *so that Lemma C.4 holds almost surely rather than in probability.*

Proof of Theorem 2. We first derive the asymptotic distribution of $\boldsymbol{\theta}_T$ and then show that it coincides with that of $\widehat{\boldsymbol{\theta}}_T$. Let us define the function

$$\varphi_T(\boldsymbol{\theta}) \coloneqq \frac{1}{T-m} \sum_{t=m+1}^{T} \varphi_t(\boldsymbol{\theta})$$

so that $\boldsymbol{\theta}_T \coloneqq \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\, \varphi_T(\boldsymbol{\theta})$. As before, assume $m = 0$ without loss of generality. A mean value expansion of $\nabla \varphi_T(\boldsymbol{\theta})$ yields

$$\nabla \varphi_T(\boldsymbol{\theta}_T) - \nabla \varphi_T(\boldsymbol{\theta}^*) = \nabla^2 \varphi_T\left(\widetilde{\boldsymbol{\theta}}\right)(\boldsymbol{\theta}_T - \boldsymbol{\theta}^*),$$

for $\widetilde{\boldsymbol{\theta}}$ a mean value between $\boldsymbol{\theta}_T$ and $\boldsymbol{\theta}^*$. Since $\nabla \varphi_T(\boldsymbol{\theta}_T) = \mathbf{0}$ by definition, we have that

$$\sqrt{T}(\boldsymbol{\theta}_T - \boldsymbol{\theta}^*) = -\left(\nabla^2 \varphi_T\left(\widetilde{\boldsymbol{\theta}}\right)\right)^{-1} \sqrt{T} \nabla \varphi_T(\boldsymbol{\theta}^*).$$

Lemma C.2 and Assumption 11 allow us to apply the central limit theorem for near epoch

dependent from Pötscher and Prucha (1997), so that $\sqrt{T}\nabla\varphi_T(\boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, V)$ as $T \to \infty$ and for V as in the main body. Lemma C.3 and the fact that $\boldsymbol{\theta}_T \xrightarrow{a.s.} \boldsymbol{\theta}^*$ imply: $\nabla^2\varphi_T\left(\widetilde{\boldsymbol{\theta}}\right) \xrightarrow{a.s.} \mathbb{E}\left[\nabla^2\varphi_t(\boldsymbol{\theta}^*)\right]$, where the limit is negative definite from Assumption 9 (v). Combining these results we get $\sqrt{T}(\boldsymbol{\theta}_T - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, for $\boldsymbol{\Sigma}$ as in the main body.

Let us now consider the function

$$\widehat{\varphi}_T(\boldsymbol{\theta}) \coloneqq \frac{1}{T-m}\sum_{t=m+1}^{T}\widehat{\varphi}_t(\boldsymbol{\theta})$$

so that $\widehat{\boldsymbol{\theta}}_T \coloneqq \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, \widehat{\varphi}_T(\boldsymbol{\theta})$. Again, let $m = 0$. A mean value expansion of $\nabla\varphi_T(\boldsymbol{\theta})$ yields:

$$\nabla\varphi_T(\boldsymbol{\theta}_T) - \nabla\varphi_T\left(\widehat{\boldsymbol{\theta}}_T\right) = \nabla^2\varphi_T\left(\overline{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}_T - \widehat{\boldsymbol{\theta}}_T\right),$$

for $\overline{\boldsymbol{\theta}}$ a mean value between $\boldsymbol{\theta}_T$ and $\widehat{\boldsymbol{\theta}}_T$, and where $\nabla\varphi_T(\boldsymbol{\theta}_T) = \nabla\widehat{\varphi}_T\left(\widehat{\boldsymbol{\theta}}_T\right) = \mathbf{0}$. Thus,

$$\sqrt{T}\left(\nabla\widehat{\varphi}_T\left(\widehat{\boldsymbol{\theta}}_T\right) - \nabla\varphi_T\left(\widehat{\boldsymbol{\theta}}_T\right)\right) = \nabla^2\varphi_T\left(\overline{\boldsymbol{\theta}}\right)\sqrt{T}\left(\boldsymbol{\theta}_T - \widehat{\boldsymbol{\theta}}_T\right),$$

where $\nabla^2\varphi_T\left(\overline{\boldsymbol{\theta}}\right) \xrightarrow{a.s.} \mathbb{E}\left[\nabla^2\varphi_T(\boldsymbol{\theta}^*)\right]$ from Lemma C.3 and Theorem 1. Lemma C.4 implies

$$\sqrt{T}\left\|\nabla\widehat{\varphi}_T\left(\widehat{\boldsymbol{\theta}}_T\right) - \nabla\varphi_T\left(\widehat{\boldsymbol{\theta}}_T\right)\right\| \xrightarrow{p} 0,$$

so that $\sqrt{T}\left\|\boldsymbol{\theta}_T - \widehat{\boldsymbol{\theta}}_T\right\| \xrightarrow{p} 0$ and the two estimators share the same asymptotic distribution.

□

C.4 Proof of Proposition 3.1

Lemma C.5 shows that the difference between $f_{t+1|t}(y;\boldsymbol{\theta})$ and $\tilde{f}_{t+1|t}(y;\boldsymbol{\theta})$ converges to zero in probability, uniformly over $\boldsymbol{\Theta}$, and for any $y \in \mathbb{R}$ as $t \to \infty$. This result is instrumental

for Proposition 3.1 and its proof is similar to that of C.1.

Lemma C.5. *Under Assumptions 1, 3, 4, 5 (i) and 12,* $\left|\tilde{f}_{t+1|t}(y) - f_{t+1|t}(y)\right|_{\boldsymbol{\Theta}} \xrightarrow{p} 0$ *as* $t \to \infty$ *and for any* $y \in \mathbb{R}$.

Proof. Let us define the quantities $x_{i,t}(\boldsymbol{\theta}) := (y - Y_{t-1-i})/h_t(\boldsymbol{\theta})$, $\tilde{x}_{i,t}(\boldsymbol{\theta}) := (y - Y_{t-1-i})/\tilde{h}_t(\boldsymbol{\theta})$ and $\Delta_t(y) := \left|\tilde{f}_{t|t-1}(y) - f_{t|t-1}(y)\right|_{\boldsymbol{\Theta}}$. Then we have

$$
\begin{aligned}
\Delta_t(y) &\leq \sum_{i=0}^{t-1} \left| \frac{\omega_i}{h_t} \mathcal{K}(x_{i,t}) - \frac{\tilde{\omega}_{i,t-1}}{\tilde{h}_t} \mathcal{K}(\tilde{x}_{i,t}) \right|_{\boldsymbol{\Theta}} + \sum_{i=t}^{\infty} \left| \frac{\omega_i}{h_t} \mathcal{K}(x_{i,t}) \right|_{\boldsymbol{\Theta}} \\
&\leq \sum_{i=0}^{t-1} \left| \frac{\omega_i}{h_t} \mathcal{K}(x_{i,t}) - \frac{\omega_i}{\tilde{h}_t} \mathcal{K}(\tilde{x}_{i,t}) \right|_{\boldsymbol{\Theta}} + D_{\mathcal{K}} D' \sum_{i=t}^{\infty} |\omega_i|_{\boldsymbol{\Theta}} + \sum_{i=0}^{t-1} \left| \frac{\omega_i}{\tilde{h}_t} \mathcal{K}(\tilde{x}_{i,t}) - \frac{\tilde{\omega}_{i,t-1}}{\tilde{h}_t} \mathcal{K}(\tilde{x}_{i,t}) \right|_{\boldsymbol{\Theta}} \\
&=: A_t(y) + B_t + C_t(y),
\end{aligned}
$$

where $A_t \xrightarrow{p} 0$ for any y and $B_t \to 0$ as $t \to \infty$ following similar steps as in the proof of Lemma C.1. For C_t we have that

$$
C_t(y) \leq D_{\mathcal{K}} D \left| c_{t-1} - c \right|_{\boldsymbol{\Theta}} \sum_{i=0}^{t-1} |a_i|_{\boldsymbol{\Theta}},
$$

for $c_{t-1}(\boldsymbol{\theta}_\omega)$, $c(\boldsymbol{\theta}_\omega)$ and $a_i(\boldsymbol{\theta}_\omega)$ as in Section B.4, and where the upper bound converges to zero in probability under Assumption 15. □

Proposition 3.1 readily follows by combining Lemma C.5 with a continuity argument as in the proof of Proposition 3.2 in Blasques et al. (2018). Hence, we omit its proof.

D Monte Carlo analysis

In this section, we carry out a Monte Carlo analysis to study how effectively Dynamic Kernel models can predict one-step ahead distributions under different data generating

processes (DGP). We simulate data from the model:

$$\begin{aligned} y_t &= \rho y_{t-1} + \sigma_t z_t, \qquad z_t \overset{i.i.d.}{\sim} D, \\ \sigma_t^2 &= \omega + \alpha \left(y_{t-1} - \mu_{t-1}\right)^2 + \beta \sigma_{t-1}^2, \end{aligned} \tag{32}$$

where $(\rho, \omega, \alpha, \beta) = (0.8, 0.05, 0.2, 0.7)$ and D is a distribution with zero mean and unit variance. In particular, we consider the cases where D is: Gaussian, Student's t with $\nu = 4$ degrees of freedom, and Skew-t as in Fernández and Steel (1998) with skewness parameter $\xi = 0.9$ and $\nu = 4$ degrees of freedom (these values imply that innovations are left skewed and fat-tailed). We consider $M = 1000$ Monte Carlo samples of length $T = 2000$ and split each sample into two sub-samples of equal length. Models are estimated on the first $T_1 = 1000$ observations and density forecasts are made for the remaining points in time.

For each DGP, we estimate two Dynamic Kernel models based on EWMA weights, a GARCH-like bandwidth and either a Gaussian or a Student's t kernel. Experiments with other weighting schemes and bandwidth dynamics returned very similar results.[15] As in Engle (2002) and Koopman et al. (2016) (among others), we first study how well we can predict a time-varying parameter of interest. Because the interest is in predictive distributions, we consider the mean absolute error (MAE) for a set of one-step ahead quantiles. That is, we look at

$$\text{MAE}_j = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \left| \hat{q}_{j,t|t-1} - q^0_{j,t|t-1} \right|,$$

where $\hat{q}_{j,t|t-1}$ $\left(q^0_{j,t|t-1}\right)$ is the estimated (true) one-step ahead predictive quantile for probability level $\tau_j \in (0,1)$. We consider probability levels $\tau_j = 1\%, 10\%, 20\%, \ldots, 80\%, 90\%, 99\%$

[15] For hyperbolically decaying weights, estimates θ were always above five: a clear hint that exponential weights sufficed for this DGP. Similarly, leverage parameters in the GJR and in the DCS case were always negligible. Hence, we opted for the more parsimonious GARCH-like dynamics.

Table 6: Monte Carlo results, Mean Absolute Errors.

Gaussian data generating process											
Kernel/τ	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Gaussian	0.621	0.756	0.858	0.935	0.982	0.998	0.981	0.933	0.857	0.755	0.619
Student's t	0.645	0.751	0.854	0.932	0.982	0.999	0.981	0.931	0.852	0.749	0.644
Student's t data generating process											
Kernel/τ	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Gaussian	0.633	0.715	0.801	0.877	0.947	0.979	0.946	0.876	0.801	0.715	0.630
Student's t	0.546	0.567	0.662	0.796	0.927	0.987	0.926	0.794	0.660	0.567	0.546
Skew-t data generating process											
Kernel/τ	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Gaussian	0.837	0.737	0.842	0.915	0.968	0.981	0.928	0.847	0.775	0.713	0.698
Student's t	0.521	0.601	0.693	0.814	0.931	0.987	0.930	0.789	0.642	0.552	0.517

Note: Average MAEs across $M = 1000$ Monte Carlo samples. The DGP is that in (32) with Gaussian, Student's t and Skew-t innovations. Results are reported as a fraction of the average MAE for the approach of Harvey and Oryshchenko (2012). Models are estimated over $T_1 = 1000$ observations and quantiles are predicted for the subsequent $T_2 = 1000$ points in time.

and present results in Table 6. In all cases, we show the average MAE across Monte Carlo samples. Numbers are reported with respect to the approach of Harvey and Oryshchenko (2012) based on a Gaussian kernel function. For the Gaussian DGP (upper panel), there is no benefit in using a Student's t kernel instead of a Gaussian one. Introducing a dynamic bandwidth improves upon the approach of Harvey and Oryshchenko (2012) for all probability levels of interest. The improvement is more sizeable when we consider probability levels below (above) 20% (80%), i.e. in the tails of the distribution. The central and the lower panels suggest that using a Student's t kernel is useful when the DGP is fat-tailed, i.e. in the Student's t and Skew-t cases, especially for the tails of the predictive sdistribution. Adopting a dynamic bandwidth is also relevant for these fat-tailed DGPs.

To understand the goodness-of-fit of our models, we consider unconditional and condi-

Table 7: Monte Carlo results, empirical rejection frequencies of coverage tests.

		Gaussian data generating process										
Coverage	Kernel/τ	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Unconditional	Gaussian	0.085	0.061	0.026	0.006	0.003	0.003	0.002	0.009	0.028	0.049	0.067
Unconditional	Student's t	0.087	0.061	0.021	0.007	0.002	0.002	0.001	0.008	0.031	0.043	0.080
Conditional	Gaussian	0.046	0.073	0.064	0.047	0.048	0.048	0.050	0.058	0.059	0.064	0.033
Conditional	Student's	0.039	0.074	0.068	0.048	0.052	0.050	0.056	0.056	0.058	0.066	0.040
		Student's t data generating process										
Coverage	Kernel/τ	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Unconditional	Gaussian	0.424	0.620	0.876	0.786	0.280	0.007	0.292	0.787	0.878	0.654	0.422
Unconditional	Student's t	0.098	0.053	0.037	0.015	0.006	0.004	0.006	0.015	0.036	0.061	0.096
Conditional	Gaussian	0.287	0.569	0.821	0.725	0.305	0.113	0.296	0.705	0.815	0.591	0.307
Conditional	Student's t	0.053	0.072	0.079	0.072	0.067	0.060	0.052	0.076	0.069	0.080	0.065
		Skew-t data generating process										
Coverage	Kernel/τ	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Unconditional	Gaussian	0.701	0.406	0.842	0.894	0.693	0.073	0.058	0.630	0.892	0.835	0.153
Unconditional	Student's t	0.212	0.065	0.033	0.042	0.049	0.082	0.064	0.046	0.045	0.140	0.150
Conditional	Gaussian	0.563	0.373	0.792	0.823	0.601	0.138	0.140	0.580	0.847	0.770	0.092
Conditional	Student's t	0.121	0.090	0.099	0.094	0.103	0.124	0.130	0.087	0.080	0.140	0.083

Note: The nominal size is 5%, and the DGP is that in (32) with Gaussian, Student's t and Skew-t innovations. The number of Monte Carlo samples is $M = 1000$. Results are presented for models based on a Gaussian or on a Student's t kernel. Models are estimated over $T_1 = 1000$ observations and quantiles are predicted for the subsequent $T_2 = 1000$ points in time.

tional coverage tests for the previous eleven quantiles (see Section 4.4 in the main body for a discussion of these tests). Tests are performed at the 5% level of significance. Table 7 reports empirical rejection frequencies for the null hypotheses of correct unconditional or conditional coverage. In particular, each entry is obtained as $\frac{1}{M}\sum_{m=1}^{M}\mathbb{1}\left(p_{j,m} < 0.05\right)$ for $p_{j,m}$ the p-value of the test for the j-th quantile over the m-th sample. Results based on the method of Harvey and Oryshchenko (2012) were always worse than those based on our models and we do not report them.

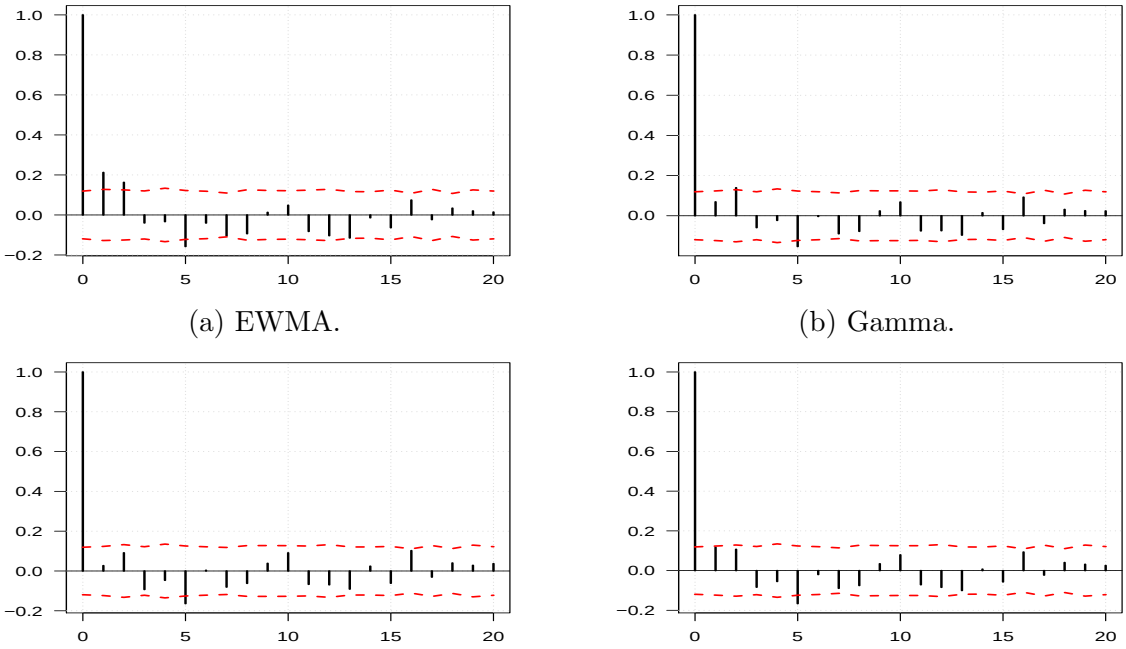
As in Table 6, the Gaussian and the Student's t kernel perform similarly under the

Gaussian GDP. The conditional coverage seems to be particularly good, with empirical rejection frequencies being very close to the nominal size. The Gaussian kernel provides poor coverage when the DGP is fat-tailed, as we can see from the central panel of Table 7. Using a Student's t kernel drastically improves the goodness-of-fit of the predicted distributions, as we implied by empirical rejection frequencies being much closer to 5%. Adding skewness further hampers results based on Gaussian kernels. Models based on Student's t density still provide a good unconditional coverage for probability levels between ten and eighty percent. The conditional coverage is satisfying for all quantiles of interest, as we can see from empirical rejection frequencies always being lower than 14%.

E Further empirical results

E.1 Further results on model diagnostics

In this section, we carry out model diagnostics in the form of residual analysis, thus complementing the results of Section 4.4 in the main body. To do it, we study the sample auto-correlation function (acf) of standardized residuals $\hat{\varepsilon}_t := \left(Y_t - \hat{\mu}_{t|t-1}\right)/\hat{\sigma}_{t|t-1}$ and of their squares. Both processes are uncorrelated at any lead or lag when $\hat{\mu}_{t|t-1}$ and $\hat{\sigma}_{t|t-1}$ are correctly specified. Figure 10 shows the sample acf for $\hat{\varepsilon}_t$ implied by EWMA (upper-left panel), Gamma (upper-right panel), hyperbolic (lower-left panel) and flexible hyperbolic (lower-right panel) weights. Results are based on a GJR bandwidth and are robust with respect to this modelling choice. Red dashed lines are (non-parametric) 95% confidence bands for non-linear processes as detailed by Francq and Zakoïan (2009). While residuals based on EWMA weights exhibit significantly positive first and second order auto-correlations, no such result is observable for hyperbolically decaying weights. Gamma weights improve upon EWMA ones but still imply some statistically significant auto-correlation at lag two.

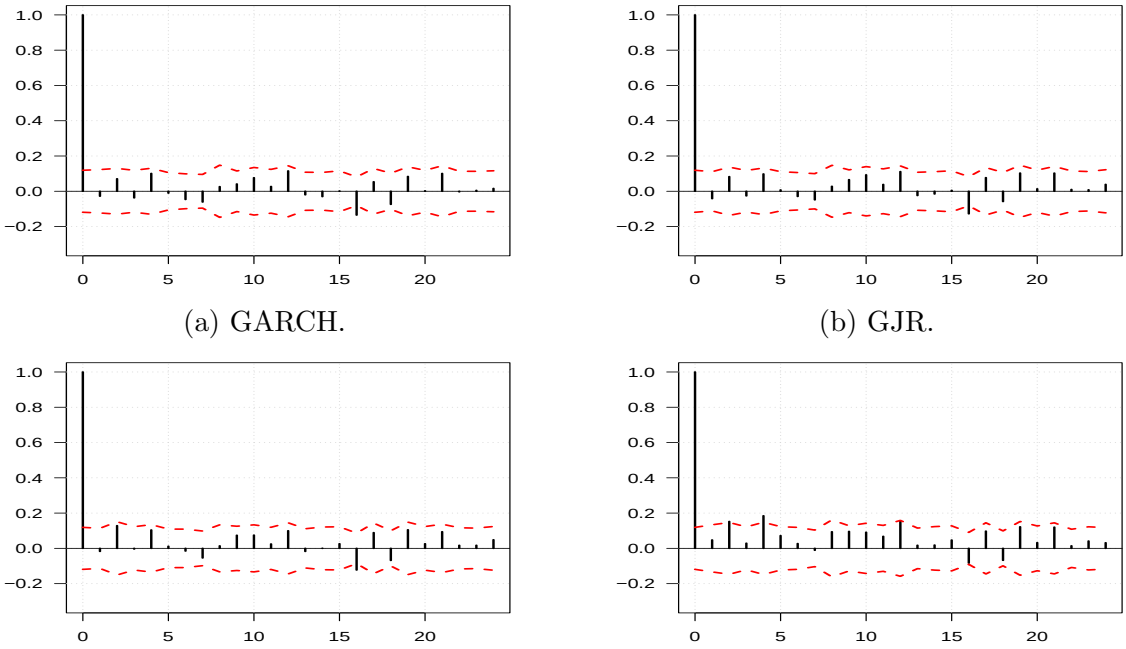


(a) EWMA. (b) Gamma.

(c) Hyperbolic. (d) Flexible hyperbolic.

Figure 10: In clockwise order: sample auto-correlation functions of standardized residuals implied by EWMA, Gamma, flexible hyperbolic and hyperbolic weights with GJR-like bandwith process. Red dashed lines are 95% confidence bands for non-linear processes as in Francq and Zakoïan (2009).

Sample acf of $\hat{\varepsilon}_t^2$ are in Figure 11. They are based on four models with hyperbolically decaying weights and different bandwidth processes. The heteroskedasticity in the data is not properly captured with a fixed bandwidth. Indeed, sample auto-correlations in Panel (d) are larger than those in the other panels. The null hypothesis of no auto-correlation is rejected for some lags. Conversely, a dynamic bandwidth returns serially uncorrelated squared residuals for any bandwidth process. Identical conclusions hold for the other weighting schemes.



(a) GARCH. (b) GJR.

(c) DCS-EGARCH. (d) Fixed.

Figure 11: In clockwise order: sample auto-correlation functions of squared standardized residuals implied by hyperbolic weights with GARCH, GJR, fixed and DCS-EGARCH bandwidths. Red dashed lines are 95% confidence bands for non-linear processes as in Francq and Zakoïan (2009).

F Score and Hessian of Dynamic Kernel models

Consider the SE sequence $\{\varphi_t(\boldsymbol{\theta}); t \in \mathbb{Z}\}$, with generic element

$$\varphi_t(\boldsymbol{\theta}) := \log\left\{\frac{1}{h_t(\boldsymbol{\theta})}\sum_{i=0}^{\infty}\omega_i(\boldsymbol{\theta}_\omega)\,\mathcal{K}(x_{i,t}(\boldsymbol{\theta});\boldsymbol{\theta}_\mathcal{K})\right\},$$

for $x_{i,t}(\boldsymbol{\theta}) := (Y_t - Y_{t-1-i})/h_t(\boldsymbol{\theta})$. Let us partition the score as

$$\nabla\varphi_t(\boldsymbol{\theta}) = (\nabla'_\mathcal{K}\varphi_t(\boldsymbol{\theta}), \nabla'_\omega\varphi_t(\boldsymbol{\theta}), \nabla'_h\varphi_t(\boldsymbol{\theta}))',$$

where $\nabla_{\mathcal{K}}\varphi_t(\boldsymbol{\theta})$ denotes the gradient with respect to $\boldsymbol{\theta}_{\mathcal{K}}$ and similarly for the remaining sub-vectors. We then have that

$$\nabla_i \varphi_t(\boldsymbol{\theta}) = \frac{1}{f_{t|t-1}(\boldsymbol{\theta}; Y_t)} \nabla_i f_{t|t-1}(\boldsymbol{\theta}; Y_t),$$

for $i \in \{\mathcal{K}, \omega, h\}$ and

$$\begin{aligned}
\nabla_{\mathcal{K}} f_{t|t-1}(\boldsymbol{\theta}) =& \frac{1}{h_t(\boldsymbol{\theta})} \sum_{i=0}^{\infty} \omega_i(\boldsymbol{\theta}_\omega) \nabla_{\mathcal{K}} \mathcal{K}(x_{i,t}(\boldsymbol{\theta}); \boldsymbol{\theta}_{\mathcal{K}}); \\
\nabla_\omega f_{t|t-1}(\boldsymbol{\theta}) =& -\frac{\nabla_\omega h_t(\boldsymbol{\theta})}{h_t^2(\boldsymbol{\theta})} \left\{ \sum_{i=0}^{\infty} \omega_i(\boldsymbol{\theta}_\omega) \left[\mathcal{K}(x_{i,t}(\boldsymbol{\theta}); \boldsymbol{\theta}_{\mathcal{K}}) + \mathcal{K}'(x_{i,t}(\boldsymbol{\theta}); \boldsymbol{\theta}_{\mathcal{K}}) \frac{(Y_t - Y_{t-1-i})}{h_t(\boldsymbol{\theta})} \right] \right\} \\
&+ \frac{1}{h_t(\boldsymbol{\theta})} \sum_{i=0}^{\infty} \mathcal{K}(x_{i,t}(\boldsymbol{\theta}); \boldsymbol{\theta}_{\mathcal{K}}) \nabla_\omega \omega_i(\boldsymbol{\theta}_\omega); \\
\nabla_h f_{t|t-1}(\boldsymbol{\theta}) =& -\frac{\nabla_h h_t(\boldsymbol{\theta})}{h_t^2(\boldsymbol{\theta})} \left\{ \sum_{i=0}^{\infty} \omega_i(\boldsymbol{\theta}_\omega) \left[\mathcal{K}(x_{i,t}(\boldsymbol{\theta}); \boldsymbol{\theta}_{\mathcal{K}}) + \mathcal{K}'(x_{i,t}(\boldsymbol{\theta}); \boldsymbol{\theta}_{\mathcal{K}}) \frac{(Y_t - Y_{t-1-i})}{h_t(\boldsymbol{\theta})} \right] \right\},
\end{aligned}$$

where we omitted the dependence on Y_t. Doing similarly for the Hessian matrix we get

$$\nabla^2_{i,j} \varphi_t(\boldsymbol{\theta}) = \frac{1}{f_{t|t-1}(\boldsymbol{\theta})} \nabla^2_{ij} f_{t|t-1}(\boldsymbol{\theta}) - \frac{1}{f^2_{t|t-1}(\boldsymbol{\theta})} \nabla_i f_{t|t-1}(\boldsymbol{\theta}) \nabla'_j f_{t|t-1}(\boldsymbol{\theta}),$$

for $i \in \{\mathcal{K}, \omega, h\}$, similarly for j and

$$\begin{aligned}
\nabla^2_{\mathcal{K}\mathcal{K}} f_{t|t-1} &= \frac{1}{h_t} \sum_{i=0}^{\infty} \omega_i \nabla^2_{\mathcal{K}\mathcal{K}} \mathcal{K}(x_{i,t}); \\
\nabla^2_{\mathcal{K}\omega} f_{t|t-1} &= \sum_{i=0}^{\infty} \left\{ \nabla_{\mathcal{K}} \mathcal{K}(x_{i,t}) \frac{\nabla'_\omega \omega_i}{h_t} - \omega_i \left[\nabla_{\mathcal{K}} \mathcal{K}(x_{i,t}) + (\nabla_{\mathcal{K}} \mathcal{K}(x_{i,t}))' \frac{Y_t - Y_{t-1-i}}{h_t} \right] \frac{\nabla'_\omega h_t}{h_t^2} \right\} \\
\nabla^2_{\mathcal{K}h} f_{t|t-1} &= -\left\{ \sum_{i=0}^{\infty} \omega_i \left[\nabla_{\mathcal{K}} \mathcal{K}(x_{i,t}) + (\nabla_{\mathcal{K}} \mathcal{K}(x_{i,t}))' \frac{(Y_t - Y_{t-1-i})}{h_t} \right] \right\} \frac{\nabla'_h h_t}{h_t^2}; \\
\nabla^2_{\omega\omega} f_{t|t-1} &= \frac{\nabla_\omega h_t \nabla'_\omega h_t}{h_t^3} \left\{ \sum_{i=0}^{\infty} \omega_i \left[2\mathcal{K}(x_{i,t}) + 4\mathcal{K}'(x_{i,t}) \frac{(Y_t - Y_{t-1-i})}{h_t} + \mathcal{K}''(x_{i,t}) \frac{(Y_t - Y_{t-1-i})^2}{h_t^2} \right] \right\}
\end{aligned}$$

$$
\begin{aligned}
&- \frac{\nabla^2_{\omega\omega} h_t}{h_t^2} \left\{ \sum_{i=0}^{\infty} \omega_i \left[\mathcal{K}\left(x_{i,t}\right) + \mathcal{K}'\left(x_{i,t}\right) \frac{\left(Y_t - Y_{t-1-i}\right)}{h_t} \right] \right\} + \frac{1}{h_t} \sum_{i=0}^{\infty} \mathcal{K}\left(x_{i,t}\right) \nabla^2_{\omega\omega} \omega_i \\
&- 2 \left\{ \sum_{i=0}^{\infty} \nabla'_{\omega} \omega_i \left[\mathcal{K}\left(x_{i,t}\right) + \mathcal{K}'\left(x_{i,t}\right) \frac{\left(Y_t - Y_{t-1-i}\right)}{h_t} \right] \right\} \\
\nabla^2_{\omega h} f_{t|t-1} = &\frac{\nabla_{\omega} h_t \nabla'_h h_t}{h_t^3} \left\{ \sum_{i=0}^{\infty} \omega_i \left[2\mathcal{K}\left(x_{i,t}\right) + 4\mathcal{K}'\left(x_{i,t}\right) \frac{\left(Y_t - Y_{t-1-i}\right)}{h_t} + \mathcal{K}''\left(x_{i,t}\right) \frac{\left(Y_t - Y_{t-1-i}\right)^2}{h_t^2} \right] \right\} \\
&- \left\{ \sum_{i=0}^{\infty} \nabla_{\omega} \omega_i \left[\mathcal{K}\left(x_{i,t}\right) + \mathcal{K}'\left(x_{i,t}\right) \frac{\left(Y_t - Y_{t-1-i}\right)}{h_t} \right] \right\} \frac{\nabla'_h h_t}{h_t^2} \\
&- \frac{\nabla^2_{\omega h} h_t}{h_t^2} \left\{ \sum_{i=0}^{\infty} \omega_i \left[\mathcal{K}\left(x_{i,t}\right) + \mathcal{K}'\left(x_{i,t}\right) \frac{\left(Y_t - Y_{t-1-i}\right)}{h_t} \right] \right\} \\
\nabla^2_{hh} f_{t|t-1} = &\frac{\nabla_h h_t \nabla'_h h_t}{h_t^3} \left\{ \sum_{i=0}^{\infty} \omega_i \left[2\mathcal{K}\left(x_{i,t}\right) + \mathcal{K}'\left(x_{i,t}\right) \frac{\left(Y_t - Y_{t-i-1}\right)}{h_t} + .\mathcal{K}''\left(x_{i,t}\right) \frac{\left(Y_t - Y_{t-1-i}\right)^2}{h_t^2} \right] \right\} \\
&- \frac{\nabla^2_{hh} h_t}{h_t^2} \left\{ \sum_{i=0}^{\infty} \omega_i \left[\mathcal{K}\left(x_{i,t}\right) + \mathcal{K}'\left(x_{i,t}\right) \frac{\left(Y_t - Y_{t-1-i}\right)}{h_t} \right] \right\},
\end{aligned}
$$

where we have omitted the dependences on $\boldsymbol{\theta}$ to improve readability and $\left(\nabla_{\mathcal{K}} \mathcal{K}\left(x_{i,t}\right)\right)'$ is shorthand notation for $\frac{d}{d\,x} \nabla_{\mathcal{K}} \mathcal{K}\left(x\right)\big|_{x=x_{i,t}}$, with the derivative being taken w.r.t. to x for each entry of $\nabla_{\mathcal{K}} \mathcal{K}$.