van Os, Bram

Working Paper

Information-Theoretic Time-Varying Density Modeling

Tinbergen Institute Discussion Paper, No. TI 2023-037/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: van Os, Bram (2023) : Information-Theoretic Time-Varying Density Modeling, Tinbergen Institute Discussion Paper, No. TI 2023-037/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/273848

TI 2023-037/III
Tinbergen Institute Discussion Paper

Information-Theoretic Time-Varying Density Modeling

Bram van Os[1]

1 Econometric Institute, Erasmus University Rotterdam

Information-Theoretic Time-Varying Density Modeling*

Bram van Os†

Econometric Institute, Erasmus University Rotterdam

April 23, 2023

Abstract

We present a comprehensive framework for constructing dynamic density models by combining optimization with concepts from information theory. Specifically, we propose to recursively update a time-varying conditional density by maximizing the log-likelihood contribution of the latest observation subject to a Kullback-Leibler divergence (KLD) regularization centered at the one-step ahead predicted density. The resulting Relative Entropy Adaptive Density (READY) update has attractive optimality properties, is reparametrization invariant and can be viewed as an intuitive regularized estimator of the pseudo-true density. Popular existing models, such as the ARMA(1,1) and GARCH(1,1), can be retrieved as special cases. Furthermore, we show that standard score-driven models with inverse Fisher scaling can be derived as convenient local approximations of the READY update. Empirical usefulness is illustrated by the modeling of employment growth and asset volatility.

*I gratefully acknowledge helpful comments and suggestions from Dick van Dijk, Maria Grith, Lukas Hoesch, Rutger-Jan Lange, André Lucas, Onno Kleen, Anne Opschoor, Andreas Pick and Ramon de Punder. Any remaining errors are my own.

†vanos@ese.eur.nl

1 Introduction

The distribution of many empirically relevant time series is well known to be time-varying. For example, the distribution of macroeconomic variables is generally characterized by dependence on the business-cycle (Stock and Watson, 1999) and the dispersion of asset returns displays so-called volatility clustering (Engle and Bollerslev, 1986). Many existing models bring about time-variation of the conditional density by innovating the model parameters using an appropriate ex-post estimator. In the context of asset volatility, for example, ARCH-type models are innovated using the squared shock, see e.g. Teräsvirta (2009). However, such approaches are often hard to generalize, and, more importantly, it is often unclear to what extent such a construction yields efficient updates of the conditional density.

We propose an information-theoretically motivated filter framework that allows for a time-varying conditional density by alternating between prediction and update steps. The key contribution is the Relative Entropy Adaptive Density (READY) update, which maximizes the log-likelihood contribution of the most recent observation subject to a Kullback-Leibler divergence (KLD) penalization centered around the one-step ahead predicted density. As a result, we update towards a density that would have been more likely to have generated the latest observation, while simultaneously incorporating persistence by requiring that we do not stray too far from our predicted density. This optimization structure allows the READY update to (i) fully exploit all information in the log likelihood contribution, (ii) automatically produce a joint update of multiple time-varying characteristics and, (iii) be reparameterization invariant. Popular existing time-series models, such as the ARMA, GARCH and absolute-value GARCH, can be retrieved as special cases.

We demonstrate that the READY update can be viewed as a regularized estimator of the pseudo-true density, that is, the density closest to the true density in a KLD sense. The READY update can also be understood as a global information-theoretic version or density-level equivalent of the class of score-driven models, acting in distribution space rather than in Euclidean parameter space. In particular, approximating the KLD penalty of the READY

update using a second-order Taylor approximation around the prediction yields the Proximal Parameter (ProPar) framework of Lange et al. (2022) with penalty proportional to the Fisher information matrix. When additionally linearizing the log-likelihood contribution, we obtain the popular score-driven update used in generalized autoregressive score (GAS; Creal et al., 2013) and dynamic conditional score (DCS; Harvey, 2013) models. This connection provides new insights on the efficiency of score-driven models as well as on the ideal choice of scaling.

The form of the READY update allows for the derivation of powerful theoretical properties. Specifically, for a concave postulated log density controlled by a single time-varying parameter, the READY update in expectation provides a global linear contraction into a noise-dominated region around the pseudo-truth in a KLD sense. Only predictions that already accurate may therefore in expectation not benefit from updating, while very bad predictions are expected to benefit dramatically. This global optimality is on top of favorable local information-theoretic optimality properties, which in turn further strengthen those of the GAS framework (Blasques et al., 2015) by removing the infinitesimally small step-size constraint. Furthermore, we derive filter invertibility (i.e. differences due to initialization disappear exponentially fast) for location-scale distributions with a log concave density and either a time-varying location or scale. Interestingly, these theoretical results impose relatively little demands on the form of the true distribution, such that the READY update remains attractive even when the postulated distribution is substantially misspecified.

We demonstrate the usefulness of the READY framework in two short empirical illustrations. Specifically, we consider filtering the mean of US employment growth and the volatility of Dow Jones returns. For both illustrations, we use a conditional t-distribution such that the READY update automatically downweights the effects of large shocks. For example, this allows our model for US employment to effectively deal with the COVID-19 period. Comparing against the GAS/DCS equivalents, we find the READY update to provide a superior fit for the employment data. For the volatility illustration, we find that the READY model is virtually indistinguishable from the Beta-t-EGARCH model of Harvey and Sucarrat (2014).

Our paper is most closely related to two strands of literature. First, it is associated with the literature on extensions of (stochastic) proximal-point methods (Rockafellar, 1976; Ryu and Boyd, 2016; Asi and Duchi, 2019). In particular, our paper is connected to the Kullback proximal-point (KPP) algorithm, which iteratively optimizes the value function of interest subject to a KLD regularization centered around a previous iterate, see Chrétien and Hero (2000, 2008). The KPP structure has close ties with the natural-gradient method of Amari (1998), which scales the gradient with the inverse Fisher matrix, motivated by arguments from information geometry. Natural gradient methods often present strong efficiency gains (Martens, 2020) and are popular in a wide variety of fields, including reinforcement learning (Kakade, 2001), variational Bayes (Khan et al., 2015), optimizing neural networks (Desjardins et al., 2015) and estimating non-linear state-space models (Courts et al., 2023). The READY update uses the KPP structure to construct a novel filter.

Second, our paper is related to a large literature on score-driven models which innovate time-varying parameters using the log-likelihood score, see Creal et al. (2013) and Harvey (2013). This allows the parameter update mechanism to exploit the distributional information in a straightforward manner. Particularly relevant for economic and financial data this leads to more robust update mechanics for heavy-tailed distributions (e.g. Creal et al., 2011; Lucas and Zhang, 2016; Harvey and Lange, 2017; Opschoor et al., 2018; Gorgi, 2020). Recently, Lange et al. (2022) propose an implicit score-driven method which yields favorable theoretical advantages over the GAS update in terms of stability and optimality. The current paper can be seen as a further extension of the ProPar method of Lange et al. (2022), replacing the weighted ℓ_2 proximal term with a KLD.

This paper is structured as follows. First, Section 2 presents the READY methodology. The link with the class of score-driven models as well as further theoretical properties are discussed in Section 3. Estimation and empirical illustrations are considered in Section 4 and 5, respectively. Section 6 concludes. Finally, proofs are contained in the Online Appendix.

2 Methodology

2.1 Relative entropy adaptive density update

Let y_t for $t = 1, 2, \ldots, T$ denote a possibly vector-valued observation generated by some unknown continuous distribution with time-varying density $p_t^0 := p^0(\cdot|\theta_t^0)$ with real-valued time-varying parameters $\theta_t^0 \in \Theta^0 \subseteq \mathbb{R}^N$. Allowing for dependence on static parameters or available exogenous information is possible, but omitted for clarity of exposition. Our aim is to best track this true density p_t^0, which is of direct interest for short-term forecasting.

We propose a new information-theoretically motivated filtering framework for constructing time-varying density models that alternates between update and prediction steps. Specifically, let $p_{t|t-1} := p(\cdot|\theta_{t|t-1})$ denote our (possibly misspecified) postulated predicted density for time t constructed at time $t-1$. Here $p_{t|t-1}$ is parameterized by some $K \times 1$ time-varying parameter vector $\theta_{t|t-1}$ taking its value in some non-empty convex set $\Theta \subseteq \mathbb{R}^K$. We assume that our postulated density is identified, such that $p(\cdot|\theta_1) = p(\cdot|\theta_2)$ if and only if $\theta_1 = \theta_2 \in \Theta$. The update step at time t uses the observation y_t and the predicted density $p_{t|t-1}$ to construct a nowcast or updated density $p_{t|t} := p(\cdot|\theta_{t|t})$, $\theta_{t|t} \in \Theta$. In turn, the prediction step uses the update $p_{t|t}$ to construct a density forecast for the next period $p_{t+1|t} := p(\cdot|\theta_{t+1|t})$, $\theta_{t+1|t} \in \Theta$.

The main difficulty lies in devising an updating procedure that incorporates efficiently the information found in the most recent observation. Naturally, we want our updated density $p_{t|t}$ to fit our newest observation y_t as best as possible. However, because y_t is inherently noisy, we also want to limit the total update magnitude from our prediction $p_{t|t-1}$ to control the variability of our filter. To tackle this problem head-on, we propose a dual-objective optimization setup that i) measures the observation's fit using the log likelihood, which has attractive information-theoretical advantages (e.g. Akaike, 1973; Roulston and Smith, 2002; Grünwald and Dawid, 2004) and, ii) quantifies the update magnitude using the Kullback-Leibler divergence (KLD), which can be viewed as the appropriate corresponding measure of closeness (e.g. Kullback and Leibler, 1951; Kullback, 1959; Gneiting and Raftery, 2007).

Specifically, we propose the Relative Entropy Adaptive Density (READY) update, given as

$$\begin{aligned}\theta_{t|t} &:= \underset{\theta\in\Theta}{\operatorname{argmax}} \Big\{ \log\, p(y_t|\theta) - \rho\mathcal{D}_{t|t-1}(\theta) \Big\}, \\ &:= \underset{\theta\in\Theta}{\operatorname{argmax}} \Big\{ \log\, p(y_t|\theta) + \rho\mathbb{E}_{\substack{y\\ t|t-1}}[\log\, p(y|\theta)] \Big\},\end{aligned} \tag{1}$$

where $\mathbb{E}_{\substack{y\\ t|t-1}}[\cdot]$ denotes the expectation over y with respect to the predicted density $p_{t|t-1}$, $\mathcal{D}_{t|t-1}(\theta) := \mathcal{D}(p(\cdot|\theta_{t|t-1})\|p(\cdot|\theta)) := \int_{-\infty}^{\infty} \log\left(\frac{p(y|\theta_{t|t-1})}{p(y|\theta)}\right) p(y|\theta_{t|t-1})dy = \mathbb{E}_{\substack{y\\ t|t-1}}[\log\, p(y|\theta_{t|t-1})] - \mathbb{E}_{\substack{y\\ t|t-1}}[\log\, p(y|\theta)]$ denotes the KLD of the prediction $p_{t|t-1}$ from the density $p(\cdot|\theta)$ and $\rho > 0$ is a penalization parameter that controls the persistence. The equivalence between the two forms of (1) follows from the fact that the entropy part of $\mathcal{D}_{t|t-1}(\theta)$ only depends on $\theta_{t|t-1}$, not on the argument θ, and is therefore irrelevant for maximization.

Because $-\rho\mathcal{D}_{t|t-1}(\theta)$ is strictly negative for all $\theta \neq \theta_{t|t-1}$, we have that this term can be interpreted as a penalty for deviating from the prediction $p_{t|t-1}$. The READY update thus fits the latest observation y_t as best as possible, while not straying too far from the prediction $p_{t|t-1}$. Here ρ controls the relative weight of each component, where larger values of ρ lead to larger penalties and therefore smaller-sized updates. From the second form, one may interpret the READY update as maximum likelihood estimation (MLE) with the observation y_t as well as synthetic observations drawn from the predicted density $p_{t|t-1}$ of which the log-likelihood contributions are weighted by the probability of said observation occurring according to the predicted density. This expression demonstrates a clear division of the information available at time t in two parts. Namely, the observation y_t captures the newly arrived information, whereas the prediction $p_{t|t-1}$ encapsulates all information available at time $t-1$ about time t.

The structure of the READY update presents three powerful advantages. First, by formulating the update in terms of log likelihoods the parameter update dynamics fully exploit all available distributional information. In particular, it makes efficient use of the information in the observation y_t via $\log\, p(y_t|\theta)$, thus adhering to the likelihood principle.

For example, this produces robustified update dynamics for heavy-tailed distributions as will be demonstrated in the empirical illustrations in Section 5.

Second, the update of multiple time-varying characteristics is automatically coordinated in such a way to best update the density as a whole. Therefore, we do not need large system matrices (such as e.g. the scaling matrix in the score-driven models of Creal et al., 2013) to approximate these interactions; the READY update only requires one static parameter ρ regardless of the dimension of θ. This keeps the model highly parsimonious.

Third, the READY update is unaffected by the choice of parameterization, similar to MLE. The READY update (1) could therefore also be formulated as a functional optimization problem over a distribution space. We opt for the parameter level notation both for notational simplicity and to facilitate the comparison with existing methods later.

Proposition 1 (Parameter invariance) *Define $\psi := g(\theta)$ for some invertible mapping $g : \Theta \to \Psi$, $\Theta, \Psi \subseteq \mathbb{R}^K$. If $\theta_{t|t} = \underset{\theta\in\Theta}{\operatorname{argmax}} f(\theta|y_t, \theta_{t|t-1}, \rho)$, where $f(\theta|y_t, \theta_{t|t-1}, \rho) := \log p(y_t|\theta) - \rho\mathcal{D}_{t|t-1}(\theta)$, then,*

$$\psi_{t|t} := g(\theta_{t|t}) = \underset{\psi\in\Psi}{\operatorname{argmax}}\, f^*(\psi|y_t, \psi_{t|t-1}, \rho), \tag{2}$$

where $f^*(\psi|y_t, \psi_{t|t-1}, \rho) = f(g^{-1}(\psi)|y_t, g^{-1}(\psi_{t|t-1}), \rho) = f(g^{-1}(\psi)|y_t, \theta_{t|t-1}, \rho)$.

Proposition 1 implies that the READY framework eliminates the arbitrariness of parameterization choice. For example, for a time-varying volatility model it does not matter if we formulate our model in terms of σ, σ^2, $\frac{1}{\sigma}$ or $\frac{1}{\sigma^2}$ and optimize over $(0, \infty)$ or formulate our model in terms of $\log(\sigma)$ and optimize over $\mathbb{R}$; all yield the exact same density update $p_{t|t}$.

The structure of the READY update is similar to that of the Kullback proximal point (KPP) algorithm of Chrétien and Hero (2000) used to accelerate expectation-maximization algorithms. Comparable mathematical structures are also used in Bayesian variational inference, see e.g. Blei et al. (2017). By using the KLD as a regularizer we take into account the geometry of the distributions, which provides a more appropriate notion of distance between distributions than the Euclidean distance at the parameter level. This in turn yields efficient

algorithms that have been found useful in a variety of applications (see e.g. Ravikumar et al., 2010; Khan et al., 2015; Cen et al., 2022). While the mathematical structure of the READY update is similar, we are interesting in filtering rather than estimating a set of static parameters. This leads to important differences in the specification of the penalty. In particular, for estimation one requires a sufficiently fast increase of the penalty parameter ρ in order to reach convergence. In contrast, we leave ρ untouched in order to remain responsive.

We make the following assumption regarding the existence and uniqueness of a solution to the READY problem in (1).

Assumption 1 (Existence and uniqueness) *The solution set of* $\underset{\theta\in\Theta}{\operatorname{argmax}} \left\{ \log\, p(y_t|\theta) - \rho\mathcal{D}_{t|t-1}(\theta) \right\}$ *has exactly one element with probability one.*

Assumption 1 is common in the optimization literature and may be verified in practice by showing that $\theta_{t|t}$ is found in some compact set in which $\log\, p(y_t|\theta) - \rho\mathcal{D}_{t|t-1}(\theta)$ is upper-semi continuous and strictly concave. The former provides existence by Weierstrass' theorem, while the latter implies uniqueness. Existence is therefore generally not an issue when working with continuous distributions. Uniqueness could theoretically be violated. Practically, on the rare occasion that multiple solutions are found one could always consider a simple tie-breaker, such as selecting the smallest-sized update, or consider a mixture solution.

To further understand the theoretical appeal of the READY update, we may write it as an intuitive regularized estimator of the pseudo-true density. That is, a natural objective we ideally would like to solve is given by

$$\begin{aligned} \theta_t^* &:= \underset{\theta\in\Theta}{\operatorname{argmin}}\, \mathcal{D}_t^0(\theta) \\ &:= \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}_{t}^{0}[\log\, p(y|\theta)] \\ &:= \underset{\theta\in\Theta}{\operatorname{argmax}} \int_{-\infty}^{\infty} \log\, p(y|\theta)\, \mathrm{d}F_t^0(y) \end{aligned} \tag{3}$$

where $\mathcal{D}_t^0(\theta) := \mathcal{D}(p^0(\cdot|\theta_t^0)\|p(\cdot|\theta))$ denotes the KLD of the true density p_t^0 from the density $p(\cdot|\theta)$, $\mathbb{E}_{t}^{0}$ is the expectation over y with respect to p_t^0 and $F_t^0(y) := \int_{-\infty}^{y} p^0(q|\theta_t^0)\mathrm{d}q$ the true

cumulative distribution function (CDF) at time t. The equivalence between the first two expressions in (3) follows from the fact that the entropy component of $\mathcal{D}_t^0(\theta)$ does not depend on θ, while the third line expresses the expectation using a Riemann-Stieltjes integral. We refer to $p_t^* := p(\cdot|\theta_t^*)$ as the pseudo-true density with pseudo-true parameter θ_t^*, because it is closest to the true density p_t^0 in a KLD sense, but may be misspecified (i.e. we allow for $p_t^0(\cdot) \neq p(\cdot|\theta), \forall \theta \in \Theta$). In the correctly specified case, the identification implies $\theta_t^* = \theta_t^0$. However, optimization (3) is generally infeasible because it depends on the true density p_t^0. The READY update will provide a feasible alternative with the information that is available.

To draw the connection between (3) and the READY update (1), we write the latter as

$$\begin{aligned}
\theta_{t|t} &:= \operatorname*{argmax}_{\theta\in\Theta} \Big\{ \log p(y_t|\theta) + \rho \mathbb{E}_{y\, t|t-1}[\log p(y|\theta)] \Big\} \\
&:= \operatorname*{argmax}_{\theta\in\Theta} \left\{ \frac{1}{1+\rho} \mathbb{E}_{y\, t}^{e}[\log p(y|\theta)] + \frac{\rho}{1+\rho} \mathbb{E}_{y\, t|t-1}[\log p(y|\theta)] \right\} \\
&:= \operatorname*{argmax}_{\theta\in\Theta} \int_{-\infty}^{\infty} \log p(y|\theta)\ \mathrm{d}\left(\frac{1}{1+\rho} F_t^e + \frac{\rho}{1+\rho} F_{t|t-1} \right)(y),
\end{aligned} \tag{4}$$

where $\mathbb{E}_{y\, t}^{e}[\cdot]$ denotes the expectation over y with respect to the empirical distribution at time t with CDF $F_t^e(y)$, which has all probability mass on $y = y_t$. The second line of (4) follows from the definition of $\mathbb{E}_{y\, t}^{e}[\cdot]$ and multiplication with $\frac{1}{1+\rho} > 0$, which is a positive scalar and therefore does not affect the location of the maximum. The third line of (4) expresses the expectations using Riemann-Stieltjes integrals and combines them in a single expression using the linearity properties of the integrator. The READY update can thus be seen to approximate the true unknown distribution F_t^0 in (3) with the discrete-continuous mixture distribution $\frac{1}{1+\rho} F_t^e + \frac{\rho}{1+\rho} F_{t|t-1}$. The intuition is clear: the empirical CDF provides an unbiased, yet noisy, estimator of the true CDF F_t^0, while the predicted CDF $F_{t|t-1}(y) := \int_{-\infty}^{y} p(q|\theta_{t|t-1}) \mathrm{d}q$ acts as a regularizer to control the variability. These CDFs are then linearly combined with weights $\frac{1}{1+\rho}$ and $\frac{\rho}{1+\rho}$, respectively, such that the result is again a proper CDF. The READY update finds the density within the postulated distribution space closest to this discrete-continuous mixture, producing a simple regularized estimator of the pseudo-truth.

2.2 Prediction step

Because the one-step ahead prediction $p_{t+1|t}$ uses the same information set as the update $p_{t|t}$, we have that the quality of the prediction step depends entirely on its ability to mimic the dynamics of the DGP. For example, if we believe that the true process evolves according to some type of random walk, then we may set $p_{t+1|t} = p_{t|t}$. Empirically, however, we often observe a degree of mean-reversion. We therefore present two different prediction steps that allow for a type of mean-reversion, where each offers distinct advantages.

First, we propose a density-level prediction step as follows

$$\begin{aligned} \theta_{t+1|t} &:= \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \mathbb{E}_{y_{t|t}}[\log\, p(y|\theta)] + \tau \bar{\mathbb{E}}_{y}[\log\, p(y|\theta)] \right\} \\ &:= \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \mathcal{D}_{t|t}(\theta) + \tau \bar{\mathcal{D}}(\theta) \right\}, \end{aligned} \tag{5}$$

where $\mathbb{E}_{y_{t|t}}[\cdot]$ and $\bar{\mathbb{E}}_{y}[\cdot]$ denote the expectation over y with respect to the updated density $p_{t|t}$ and some static density $\bar{p} := \tilde{p}(\cdot|\bar{\theta})$ with $\bar{\theta} \in \tilde{\Theta}$, respectively. Similarly, $\mathcal{D}_{t|t}(\theta) := \mathcal{D}(p(\cdot|\theta_{t|t}) \| p(\cdot|\theta))$ and $\bar{\mathcal{D}}(\theta) := \mathcal{D}(p(\cdot|\bar{\theta}) \| p(\cdot|\theta))$ denote the KLD of the update $p_{t|t}$ and $\bar{p}$ from $p(\cdot|\theta)$. In addition, $\tau > 0$ is a penalization parameter that controls the level of mean-reversion. Here the density $\bar{p}$ may informally be understood as a long-run density in the sense that the filter will be centered around it, while repeated application of the prediction step (5) yields converge towards it. A comparable idea of 'mixing' the updated distribution with a static long-run distribution is used in dynamic kernel density estimation, see Harvey and Oryshchenko (2012). The prediction step in (5) shares two key advantages with the update step in (1) and requires an assumption similar to Assumption 1. Specifically, it is parsimonious, requiring only one parameter τ, and it is reparameterization invariant.

Second, we propose a more standard linear prediction step at the parameter level, i.e.,

$$\theta_{t+1|t} = \omega + \Phi \theta_{t|t}, \tag{6}$$

where ω denotes a $K \times 1$ parameter vector of constants and Φ a $K \times K$ autoregressive matrix. Two advantages of this linear prediction step over the density-level prediction step in (5) are that the former is easier to execute and allows for more flexible mean-reversion dynamics through the specification of Φ. This may be useful if different time-varying characteristics display very distinct levels of mean-reversion. On the other hand, the linear prediction step in (6) may need a link function to remain in the appropriate range, is generally less parsimonious and is dependent on the parameterization chosen.

2.3 Examples of READY models

Applying the READY framework to the location of a normal distribution yields a stationary ARMA(1,1) model, while applying it to the variance gives the stationary GARCH(1,1) model of Bollerslev (1986). For the scale of a Laplace distribution the READY model coincides with the dynamics of the absolute value GARCH(1,1) (AV-GARCH(1,1)) model of Taylor (2008). Interestingly, the density-level and linear prediction steps in (5) and (6) produce the same dynamics for these examples, provided that $\Phi \in (0, 1)$. Appendix A contains additional examples for the Poisson, exponential and binomial distributions.

Example 1 (ARMA) *Consider a normal distribution $p(\cdot|\mu_t, \sigma)$ with dynamic mean μ_t and static variance $\sigma^2 > 0$. Then the READY update (1) with penalty parameter $\rho > 0$ combined with the linear prediction (6) with $\omega \in \mathbb{R}$ and $\Phi \in (-1, 1)$ from $\mu_{t|t-1}$ to $\mu_{t+1|t}$ using the observation y_t yields*

$$y_{t+1} = \omega + \Phi y_t + \psi \varepsilon_t + \varepsilon_{t+1}, \tag{7}$$

where $\varepsilon_t := y_t - \mu_{t|t-1}$ and $\psi := -\Phi \frac{\rho}{1+\rho} \in (-1, 1)$.

Example 2 (GARCH) *Consider a normal distribution $p(\cdot|\mu, \sigma_t)$ with static mean $\mu \in \mathbb{R}$ and dynamic variance σ_t^2. Then the READY update (1) with penalty parameter $\rho > 0$ combined with the linear prediction (6) with $\omega \in \mathbb{R}$ and $\Phi \in [0, 1)$ from $\sigma^2_{t|t-1}$ to $\sigma^2_{t+1|t}$ using*

the observation y_t is given as

$$\sigma^2_{t+1|t} = \omega + \alpha(y_t - \mu)^2 + \beta\sigma^2_{t|t-1}, \tag{8}$$

where $\alpha := \Phi\frac{1}{1+\rho}$ and $\beta := \Phi\frac{\rho}{1+\rho}$ with $\alpha > 0$, $\beta > 0$ and $\alpha + \beta < 1$.

Example 3 (AV-GARCH) *Consider a Laplace distribution $p(\cdot|\mu, \sigma_t)$ with static location $\mu \in \mathbb{R}$ and dynamic scale σ_t. Then the READY update (1) with penalty parameter $\rho > 0$ combined with the linear prediction (6) with $\omega \in \mathbb{R}$ and $\Phi \in [0, 1)$ from $\sigma_{t|t-1}$ to $\sigma_{t+1|t}$ using the observation y_t is given as*

$$\sigma_{t+1|t} = \omega + \alpha|y_t - \mu| + \beta\sigma_{t|t-1}, \tag{9}$$

where $\alpha := \Phi\frac{1}{1+\rho}$ and $\beta := \Phi\frac{\rho}{1+\rho}$ with $\alpha > 0$, $\beta > 0$ and $\alpha + \beta < 1$.

3 Theory

3.1 Connection with score-driven updates

In order to further characterize the READY update in (1) and to facilitate the derivation of theoretical properties, we make assumptions regarding boundary solutions, differentiability and the interchangability of the expectation and the differential operator.

Assumption 2 (Interior solution) $\theta_{t|t-1}, \theta_{t|t} \in \text{Int}(\Theta)$ *with probability one.*

Assumption 3 (Differentiability) $\log p(y|\theta)$ *and* $\mathcal{D}_{t|t-1}(\theta)$ *are at least twice continuously differentiable in* θ, $\forall\theta, \theta_{t|t-1} \in \text{Int}(\Theta)$ *and* $\forall y \in \text{Dom}(y)$.

Assumption 4 (Interchangeability of expectation and derivative) $\forall\theta_{t|t-1}, \theta_{t|t} \in \text{Int}(\Theta)$

$$\frac{\partial}{\partial\theta}\mathbb{E}_{y_{t|t-1}}[\log\, p(y|\theta)]\Big|_{\theta=\theta_{t|t}} = \mathbb{E}_{y_{t|t-1}}\left[\frac{\partial}{\partial\theta}\log\, p(y|\theta)\Big|_{\theta=\theta_{t|t}}\right], \tag{10}$$

$$\frac{\partial^2}{\partial\theta\partial\theta'}\mathbb{E}_{y_{t|t-1}}[\log\, p(y|\theta)]\Big|_{\theta=\theta_{t|t}} = \mathbb{E}_{y_{t|t-1}}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log\, p(y|\theta)\Big|_{\theta=\theta_{t|t}}\right]. \tag{11}$$

Under Assumptions 1-3 the READY update can be found by solving its first-order condition (FOC), which using Assumption 4 may be written as a gradient-type update.

Proposition 2 (READY update as an implicit gradient update) *For a given $t > 0$ let Assumptions 1-4 hold, then with probability one,*

$$\nabla(y_t|\theta_{t|t}) = \rho\mathcal{I}_{t|t-1}(\theta_{t|t})(\theta_{t|t} - \theta_{t|t-1}), \tag{12}$$

where $\nabla(y_t|\theta_{t|t}) := \frac{\partial}{\partial\theta}log\, p(y_t|\theta)\Big|_{\theta=\theta_{t|t}}$ is the log-likelihood score at time t with respect to θ evaluated in the update $\theta_{t|t}$ and $\mathcal{I}_{t|t-1}(\theta_{t|t})$ is the negative expected average Hessian between $\theta_{t|t-1}$ and $\theta_{t|t}$ where the expectation is with respect to the predictive density, that is,

$$\mathcal{I}_{t|t-1}(\theta_{t|t}) := -\mathbb{E}_{y_{t|t-1}}\left[\int_0^1 \frac{\partial^2}{\partial\theta\partial\theta'}\log p(y|\theta)\Big|_{\theta_{t|t-1}+q(\theta_{t|t}-\theta_{t|t-1})}\mathrm{d}q\right]. \tag{13}$$

If $\mathcal{I}_{t|t-1}(\theta_{t|t})$ is invertible, then the READY update takes the form of a second-order implicit stochastic gradient update,

$$\theta_{t|t} = \theta_{t|t-1} + \rho^{-1}\mathcal{I}^{-1}_{t|t-1}(\theta_{t|t})\nabla(y_t|\theta_{t|t}). \tag{14}$$

The form in (14) reveals that the READY update can be written as an implicit stochastic gradient update (see e.g. Bianchi, 2016; Patrascu and Necoara, 2018; Toulis et al., 2021). The update is implicit because $\theta_{t|t}$ is present on both sides of the equation and it is stochastic because it depends on the realization y_t. Furthermore, (14) can be viewed as a second-order gradient update as it uses the information on the expected curvature between the prediction and the update $\mathcal{I}_{t|t-1}(\theta_{t|t})$ to scale the score, see also Toulis et al. (2016). The quantity $\mathcal{I}_{t|t-1}(\theta_{t|t})$ can be understood as a measure of the information content of the prediction $p_{t|t-1}$ about the parameter θ. As a result, 'larger' ('smaller') values of $\mathcal{I}_{t|t-1}(\theta_{t|t})$ will lead to

'smaller' ('larger') updates.

Because the READY update is invariant to the choice of parameterization, one could interpret the READY update as a gradient-type update that occurs in distribution space rather than in Euclidean parameter space. The READY update is therefore closely related to the class of natural gradient methods, which use gradients scaled by a Fisher information matrix, see Amari (1998). In particular, note that $\mathcal{I}_{t|t-1}(\theta_{t|t})$ is similar to the Fisher information matrix $\mathcal{I}_{t|t-1}(\theta_{t|t-1}) := -\mathbb{E}_{y_t|t-1}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log p(y|\theta)\Big|_{\theta_{t|t-1}}\right] = \mathbb{E}_{y_t|t-1}[\nabla(y|\theta_{t|t-1})\nabla(y|\theta_{t|t-1})']$ via the information matrix equality. Because natural gradients are (locally) invariant to the choice of parameterization they enjoy attractive information-theoretical properties and are known to generally present substantial efficiency gains over standard gradient ascent methods (Yang and Amari, 1997; Martens, 2020).

To provide additional intuition for the READY gradient update form in (14), we consider first-order and second-order Taylor approximations of the READY problem in (1). In combination with a linear prediction step as in (6) this yields score-driven models of the type as introduced in Creal et al. (2013), Harvey (2013) and recently in Lange et al. (2022). Specifically, suppose we approximate the KLD penalty $\mathcal{D}_{t|t-1}(\theta)$ in the READY problem in (1) using a second-order Taylor expansion around the prediction $\theta_{t|t-1}$, that is under Assumptions 3 and 4 we have

$$\mathcal{D}_{t|t-1}(\theta) \approx \frac{1}{2}\|\theta - \theta_{t|t-1}\|^2_{\mathcal{I}_{t|t-1}(\theta_{t|t-1})}. \tag{15}$$

Using this approximation, we directly obtain the proximal parameter (ProPar) framework of Lange et al. (2022) with penalty matrix $\rho\mathcal{I}_{t|t-1}(\theta_{t|t-1})$. Note that the constant and first-order terms of the approximation in (15) vanish because $\mathcal{D}_{t|t-1}(\theta_{t|t-1}) = 0$ and $\mathbb{E}_{y_t|t-1}[\nabla(y|\theta_{t|t-1})] = 0$, where the latter follows by differentiability and the fact that $\theta_{t|t-1}$ is the unique minimizer of $\mathcal{D}_{t|t-1}(\theta)$.

The associated FOC of the READY problem using the approximation in (15) reads

$$\theta_{t|t}^{\text{ProPar}} = \theta_{t|t-1} + \rho^{-1}\mathcal{I}_{t|t-1}^{-1}(\theta_{t|t-1})\nabla(y_t|\theta_{t|t}^{\text{ProPar}}), \tag{16}$$

provided $\theta_{t|t}^{\text{ProPar}} \in \Theta$ and where $\mathcal{I}_{t|t-1}(\theta_{t|t-1})$ is invertible by the assumption of identification. The ProPar framework with a penalty matrix proportional to the Fisher matrix therefore yields a similar implicit-gradient update as the READY framework. However, instead of fully accounting for the curvature between the prediction and the update, the ProPar model uses the local approximation $\mathcal{I}_{t|t-1}(\theta_{t|t-1})$. That is, $\mathcal{I}_{t|t-1}(\theta_{t|t-1})$ measures the curvature at the prediction only and, unlike $\mathcal{I}_{t|t-1}(\theta_{t|t})$, does not use the information contained in the new observation y_t. This lightens computational demands, but may also lead to a slight efficiency loss. Additionally, we introduce dependence on the particular parameterization.

Furthermore, if we additionally approximate the log-likelihood contribution $\log p(y_t|\theta)$ in (1) using a first-order Taylor approximation around the prediction $\theta_{t|t-1}$, that is,

$$\log p(y_t|\theta) \approx \log p(y_t|\theta_{t|t-1}) + \langle \nabla(y_t|\theta_{t|t-1}), \theta - \theta_{t|t-1} \rangle, \tag{17}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, then the associated FOC yields the score-driven update used in generalized autoregressive score (GAS; Creal et al., 2013) and dynamic conditional score (DCS; Harvey, 2013) models with inverse Fisher scaling,

$$\theta_{t|t}^{\text{GAS}} = \theta_{t|t-1} + \rho^{-1}\mathcal{I}_{t|t-1}^{-1}(\theta_{t|t-1})\nabla(y_t|\theta_{t|t-1}), \tag{18}$$

provided $\theta_{t|t}^{\text{GAS}} \in \text{Int}(\Theta)$. The approximation of the log-likelihood contribution is thus reflected by the approximation of the implicit score $\nabla(y_t|\theta_{t|t})$ with its explicit counterpart $\nabla(y_t|\theta_{t|t-1})$. Because we now no longer use all the information in the log-likelihood contribution, this may lead to an efficiency loss. In particular, as all new information flows via this term (i.e. y_t is not found in the penalty term), this will turn out to be a more

severe issue for optimality than the second-order approximation of the penalty in (15), see Section 3.2. However, practically, we have that (18) is an explicit relationship that can be executed directly and does not require solving unlike the READY and ProPar updates. The above derivation essentially provides a global information-theoretic intuition for the use of GAS models as well as a direct motivation for using the inverse predictive Fisher matrix $\mathcal{I}_{t|t-1}^{-1}(\theta_{t|t-1})$ for scaling the score, as suggested by Creal et al. (2013). Namely, by using this scaling, we approximately remove the dependence on the chosen parameterization.

In sum, the READY update can be viewed as a global information-theoretic version or density-level equivalent of the score-driven models of Creal et al. (2013), Harvey (2013) and Lange et al. (2022), where the learning rate is proportional to the inverse Fisher information. The accuracy of the approximations in (15) and (17) depend on the update magnitude, such that the differences between the updates of the three methods will be small if $\theta_{t|t}$ and $\theta_{t|t-1}$ are very close, but can grow substantially as the distance between $\theta_{t|t}$ and $\theta_{t|t-1}$ increases. If the signal-to-noise ratio of the data is low, we expect small update steps and therefore close agreement between the three methods. Conversely, if the signal-to-noise ratio is large, we expect substantial benefits from using the READY update over the GAS update, with the ProPar update somewhere in between. In Section 5, we provide two illustrations of the READY model and compare them against standard score-driven equivalents. The first illustration demonstrates a case where the READY update outperforms the GAS update and the second illustration displays a scenario where they produce similar estimates.

3.2 Optimality

Naturally, we are interested in determining whether the READY update yields an updated density that is closer to the (pseudo-)true density compared the predicted density. To this end, we first consider local optimality and afterwards consider the global case. Proposition 3 presents the local information-theoretic properties of the READY update.

Proposition 3 (Local improvement of the READY update) *Let Assumption 1 hold and assume that* $\theta_{t|t} \neq \theta_{t|t-1}$. *Then, with probability one, the update is strictly likelihood concordant, that is,*

$$p(y_t|\theta_{t|t}) > p(y_t|\theta_{t|t-1}). \tag{19}$$

In addition, if the postulated density $p(y|\theta)$ *is continuous in* y, *then, with probability one, the update yields a local KLD improvement. That is,* $\exists \delta > 0$ *such that for* $\mathcal{Y} := \{y \in \text{Dom}(y) | \, \|y - y_t\|^2 \leq \delta \, \}$ *we have that* $\Pr(y \in \mathcal{Y}|\theta_t^0) := \int_{\mathcal{Y}} p^0(y|\theta_t^0)\text{d}y > 0$ *and*

$$\Delta_t(\mathcal{Y}) := \mathbb{E}_{y_t}^0[\log p(y|\theta_{t|t}) - \log p(y|\theta_{t|t-1})|y \in \mathcal{Y}] > 0, \tag{20}$$

where $\Delta_t(\mathcal{Y})$ *denotes the difference in local KLD divergencies from the prediction and the update to the truth over the set* $\mathcal{Y} \subseteq \text{Dom}(y)$.

The first result of Proposition 3 shows that updating always improves the fit of the newly arrived observation y_t. Namely, it makes it so that the information is incorporated in accordance with the likelihood. Conversely, it completely eliminates the possibility of a maladaptation, i.e., when updating using the observation y_t decreases the models capacity to fit y_t. This is a key advantage of the optimization setup.

The second result of Proposition 3 strengthens the first result to a set of positive probability containing y_t, which yields an improvement in terms of the local realized KLD $\Delta_t(\mathcal{Y})$. Proposition 3 is similar to the one of Lange et al. (2022) and can be viewed as a globalized version of the local realized KLD optimality of Blasques et al. (2015) for GAS models. Namely, GAS models generally additionally require infinitesimally small stepsizes to achieve the same result. The local optimality of the GAS update is therefore local in terms of both y and θ, while the READY and ProPar update are only local in y and global in θ. This is a direct result of the fact that the READY and ProPar updates both use the full log-likelihood contribution in the optimization procedure, while the GAS framework uses a first-order approximation, see again (17).

Next, we consider the global optimality properties of the READY framework. Because the information of the observation y_t about the (pseudo-)true density comes with a certain amount of noise, we have that fully globalizing the result of Proposition 3 (i.e. selecting $\mathcal{Y} = \text{Dom}(y)$) is fundamentally impossible. That is, if our prediction is already very close to the (pseudo-)truth, this noise will dominate and a possible deterioration is unpreventable. We refer to this zone as the noise-dominated region (NDR), similar to e.g. Ryu and Boyd (2016). On the other hand, if the prediction is far from the truth then the signal is expected to outweigh the noise, such that an improvement is likely, precisely when we need it most. Our global optimality result formalizes such a contraction of the update towards the NDR.

To simplify matters, we restrict ourselves to the scalar time-varying parameter case ($K = 1$) here and leave the multi-parameter case for future research. To control global behavior, Assumption 5 poses that the log density is concave in its parameter. While this assumption is likely too strong it yields the cleanest expressions and is commonly used to prove the convergence of gradient-based algorithms (e.g. Toulis et al., 2014). Note that this concerns a property of the model and not one of the DGP and can therefore be easily verified in practice. An immediate consequence is that the penalty $\mathcal{D}_{t|t-1}(\theta)$ is convex in θ. Also, due to the parameterization invariance of the READY update (see Proposition 1), it follows that there simply needs to exist a suitable concave parameterization; one does not need to analyze the form of the update in this parameterization in detail. In order to quantify optimality, Assumption 6 poses the existence of a unique pseudo-truth p_t^* (see again (3)) with finite-valued KLD $\mathcal{D}_t^*(\theta) := \mathcal{D}(p(\cdot|\theta_t^*)\|p(\cdot|\theta))$, $\forall \theta \in \Theta$.

Assumption 5 (Concave log likelihood) $\log p(y|\theta)$ *is concave in* θ, $\forall y \in \text{Dom}(y)$.

Assumption 6 (Existence pseudo-truth) *There exists a unique* p_t^* *that solves* (3) *and* $\mathcal{D}_t^*(\theta) < \infty$, $\forall \theta \in \Theta$.

Lemma 1 indicates that the KLD from the update to the pseudo-truth $\mathcal{D}_t^*(\theta_{t|t})$ is at most equal to a convex combination of the KLD from the prediction and the KLD from the

one-period maximum likelihood (ML) density $\hat{p}_t := p(\cdot|\hat{\theta}_t)$ to the pseudo-truth, denoted by $\mathcal{D}_t^*(\theta_{t|t-1})$ and $\mathcal{D}_t^*(\hat{\theta}_t)$, respectively. Therefore, if the density $\hat{p}_t$ is closer to the pseudo-truth p_t^* than the prediction $p_{t|t-1}$, then the update yields an improvement for any choice of penalty parameter $\rho > 0$.

Lemma 1 (READY update as smoothed MLE) *Let $\Theta \subseteq \mathbb{R}$ and let Assumptions 1-6 hold. In addition, assume that the one-period ML estimator $\hat{\theta}_t(y_t)$ exists, is unique and that $\hat{\theta}_t \in \text{Int}(\Theta)$. Then, with probability one,*

$$\mathcal{D}_t^*(\theta_{t|t}) \leq A(y_t|\theta_{t|t}, \theta_{t|t-1})\mathcal{D}_t^*(\theta_{t|t-1}) + [1 - A(y_t|\theta_{t|t}, \theta_{t|t-1})]\mathcal{D}_t^*(\hat{\theta}_t), \tag{21}$$

with smoothing coefficient $A(y_t|\theta_{t|t}, \theta_{t|t-1}) \in (0, 1)$ given as

$$A(y_t|\theta_{t|t}, \theta_{t|t-1}) := [\rho\mathcal{I}_{t|t-1}(\theta_{t|t}) - \mathcal{H}(y_t|\theta_{t|t})]^{-1}\rho\mathcal{I}_{t|t-1}(\theta_{t|t}), \tag{22}$$

where $\mathcal{H}(y_t|\theta_{t|t})$ is the realized average Hessian between $\hat{\theta}_t$ and $\theta_{t|t}$, that is,

$$\mathcal{H}(y_t|\theta_{t|t}) := \int_0^1 \frac{\partial^2}{\partial\theta\partial\theta'} \log p(y_t|\theta)\Big|_{\hat{\theta}_t + q(\theta_{t|t} - \hat{\theta}_t)} \mathrm{d}q. \tag{23}$$

The mixture weight $A(y_t|\theta_{t|t}, \theta_{t|t-1})$ admits an intuitive form by measuring the relative curvature of the penalty between the prediction and update, $\rho\mathcal{I}_{t|t-1}(\theta_{t|t})$, to the realized curvature of the observation between the one-period ML estimator and the update, $\mathcal{H}(y_t|\theta_{t|t})$. These curvatures can informally be understood as measures of the information content of the prediction $p_{t|t-1}$ and the observation y_t about the parameter θ, respectively. 'Larger' ('smaller') $\mathcal{H}(y_t|\theta_{t|t})$ relative to $\rho\mathcal{I}_{t|t-1}(\theta_{t|t})$ will therefore lead to 'larger' ('smaller') update steps.

Theorem 1 formalizes the idea of a contraction when the prediction is far from the pseudo-truth by specifying upper and lower bounds on the curvature of the log-likelihood contribution and the penalty using strong concavity or convexity and Lipschitz continuity of the gradients. Theorem 1 bears some resemblance to the contraction result of Lange et al.

(2022); the difference being that the latter formulates the contraction at the parameter level.

Theorem 1 (Global linear contraction to the NDR) *Let the conditions of Lemma 1 hold. In addition, let* $\log p(y_t|\theta)$ *be* $\alpha_t(y_t)$*-strongly concave and* $\beta_t(y_t)$*-smooth in* θ *and let* $\mathcal{D}_{t|t-1}(\theta)$ *be* $\tilde{\alpha}_t$*-strongly convex and* $\tilde{\beta}_t$*-smooth in* θ*, then,*

$$\mathop{\mathbb{E}}_{y_t}^{0}[\mathcal{D}_t^*(\theta_{t|t})] \leq \eta_t \mathcal{D}_t^*(\theta_{t|t-1}) + \lambda_t \sigma_t, \tag{24}$$

$$\eta_t := \mathop{\mathbb{E}}_{y_t}^{0}[\frac{\rho\tilde{\beta}_t}{\rho\tilde{\beta}_t + \alpha_t}] \in (0,1), \quad \lambda_t := \frac{\mathop{\mathbb{E}}_{y_t}^{0}[\frac{\beta_t}{\rho\tilde{\alpha}_t+\beta_t}\mathcal{D}_t^*(\hat{\theta}_t)]}{\mathop{\mathbb{E}}_{y_t}^{0}[\mathcal{D}_t^*(\hat{\theta}_t)]} \in (0,1), \quad \sigma_t := \mathop{\mathbb{E}}_{y_t}^{0}[\mathcal{D}_t^*(\hat{\theta}_t)] \in (0,\infty). \tag{25}$$

Theorem 1 demonstrates that the READY update with a concave log density yields a contraction to a NDR around the pseudo-true density, in line with the intuition of the READY update as an estimator of the pseudo-truth, see again (4). Specifically, we have that the expected KLD from the update to the pseudo-truth $\mathop{\mathbb{E}}_{y_t}^{0}[\mathcal{D}_t^*(\theta_{t|t})]$ is upper-bounded by a linear function of the KLD from the prediction $\mathcal{D}_t^*(\theta_{t|t-1})$ (to the pseudo-truth). The slope η_t quantifies the rate of contraction, while λ_t measures the exposure to the irreducible noise σ_t. Specifically, σ_t denotes the expected KLD from the one-period ML density $\hat{p}_t$ to the pseudo-truth and reflects the inherent expected error from using y_t to update in a ML setting. If our prediction is already highly accurate, then the additive noise term $\lambda_t\sigma_t$ may dominate and an expected deterioration is possible. In contrast, if our prediction is bad (i.e. $\mathcal{D}_t^*(\theta_{t|t-1}) > \frac{\lambda_t}{1-\eta_t}\sigma_t$), then we expect a linear contraction to the pseudo-truth.

The contraction rate η_t is governed by the strength of concavity of the log-likelihood contribution α_t relative to the smoothness of the gradient of the penalty measured by $\tilde{\beta}_t$. Furthermore, η_t is positively related to the penalty parameter ρ, while λ_t is negatively related. In the limit, we have $\lim_{\rho\to\infty}\eta_t = 1$ and $\lim_{\rho\to\infty}\lambda_t = 0$ and $\lim_{\rho\downarrow 0}\eta_t = 0$ and $\lim_{\rho\downarrow 0}\lambda_t = 1$. Consequently, we have that the choice of ρ presents a trade-off between the rate of contraction towards the NDR (governed by η_t) and the exposure to the irreducible noise σ_t (measured

by λ_t), which determines the size of the NDR. A low value of ρ leads to a model with little persistence and thus larger improvements when the predictions are 'bad', but also to more deteriorations when the prediction are 'good'. The appropriate value of ρ can be determined using the entire estimation sample as will be elaborated upon in Section 4.

3.3 Stability

We now consider the stability aspect of our proposed framework. In particular, we are interested in proving that differences due to initialization disappear sufficiently fast. We proceed with the case of a scalar time-varying parameter and a concave log model density. In addition, we make the following assumption about the form of the KLD penalty.

Assumption 7 (Translationally invariant penalty) *Consider $\Theta \subseteq \mathbb{R}$ and let*

$$\mathcal{D}(p(\cdot|\theta_{t|t-1}+c)\|p(\cdot|\theta+c)) = \mathcal{D}(p(\cdot|\theta_{t|t-1})\|p(\cdot|\theta)) =: \mathcal{D}_{t|t-1}(\theta), \tag{26}$$

$\forall c \in \mathbb{R}$ and $\forall \theta_{t|t-1}, \theta \in \Theta$ such that $\theta_{t|t-1}+c, \theta+c \in \Theta$.

Assumption 7 asserts that the KLD only depends on the Euclidean distance from the prediction, i.e. $\|\theta - \theta_{t|t-1}\|$, not on the location of that distance. While Assumption 7 may appear rather strong, it encompasses the entire family of location-scale distributions with either a time-varying location or a time-varying scale. Namely, for this family of distributions the KLD depends only on the difference in locations or the scale ratio, see Nielsen (2019) for proofs. It immediately follows that the KLD is translationally invariant in the location parameter and the logarithmic scale parameter.

Lemma 2 (Update stability) *Let $\Theta \subseteq \mathbb{R}$ and let Assumptions 1-5 and Assumption 7 hold. Consider two predictions $\theta_{t|t-1}, \tilde{\theta}_{t|t-1} \in \Theta$ that are updated using the READY update in (1) to two updates $\theta_{t|t}, \tilde{\theta}_{t|t} \in \Theta$. Then, with probability one,*

$$\|\theta_{t|t} - \tilde{\theta}_{t|t}\| \leq \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|. \tag{27}$$

If $\log p(y_t|\theta)$ *is* α_t*-strongly concave in* θ *and* $\mathcal{D}_{t|t-1}(\theta)$ *is* $\tilde{\beta}_t$*-smooth in* θ*, then, with probability one,*

$$\|\theta_{t|t} - \tilde{\theta}_{t|t}\| \leq \zeta_t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|, \quad \zeta_t := \frac{\rho\tilde{\beta}_t}{\rho\tilde{\beta}_t + \alpha_t} \in (0,1). \tag{28}$$

Lemma 2 shows that the READY update at time t from $\theta_{t|t-1}$ to $\theta_{t|t}$ is non-expansive. That is, updating cannot increase the Euclidean distance between two different filtered parameter paths. The second result of Lemma 2 shows that this result can be strengthened to a strict contraction with Lipschitz coefficient $\zeta_t \in (0,1)$, comparable to η_t in Theorem 1, in case the log-likelihood contribution is α_t-strongly concave and the penalty is $\tilde{\beta}_t$-smooth. In line with intuition, the dependence of ζ_t on ρ is positive, such that larger values of ρ will yield more persistent dynamics and thus slower contractions.

Theorem 2 demonstrates that composing a non-expansive READY update with a strictly contracting linear prediction map (i.e. $|\phi| < 1$) yields a strictly contracting composed mapping from prediction to prediction. It follows that if this holds for all points in time that the differences due to initialization disappear exponentially fast almost surely. This so-called filter invertibility is a crucial ingredient for the consistency of the ML estimator of the static parameters such as ρ, see Straumann and Mikosch (2006) and Blasques et al. (2022) for detailed discussions on this matter.

Theorem 2 (Invertibility) *Let* $\Theta \subseteq \mathbb{R}$ *and let Assumptions 1-5 and Assumption 7 hold for all* $t > 0$*. Then the filter composed of the READY update (1) and the linear prediction step (6) with* $|\phi| < 1$ *is invertible. That is, for any two initial values* $\theta_{0|0}, \tilde{\theta}_{0|0} \in \Theta$ *producing two sequences* $\{\theta_{t|t-1}\}_{t\geq 1}$ *and* $\{\tilde{\theta}_{t|t-1}\}_{t\geq 1}$*, we have that there exists a constant* $c_{(\cdot)} > 1$ *such that, with probability one,*

$$\lim_{t\to\infty} c_{(\cdot)}^t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|_{(\cdot)}^2 \to 0, \tag{29}$$

for any norm $(\cdot)$*.*

Interestingly, Theorem 2 makes almost no assumptions regarding the DGP. In particular, if Assumptions 1-2 hold for all y, then it is completely independent. This means that the

READY filter is invertible at a fundamental level, which is a direct result of the inherent stability of regularized maximization with a concave target and a convex penalty. For the ℓ_2 penalty this is the well-known non-expansiveness property of the proximal operator (e.g. Parikh and Boyd, 2014), which is why a similar result as Theorem 2 is available for the ProPar framework (see Theorem 1 of Lange et al., 2022). Besides the READY and ProPar frameworks this invertibility result is generally rare. For example, GAS models, like all explicit gradient methods, require careful tuning of the learning rate parameter (ρ^{-1}) to avoid divergence, whereby the maximum stable learning rate is usually linked to the true unknown DGP, see Blasques et al. (2022).

4 Estimation

The optimal values of the penalty parameter ρ and the prediction parameters (ω and Φ for the linear prediction (6) or τ for the density prediction (5)) are generally unknown and require estimation. Because the predicted density for time $t+1$ is available at time t, the proposed framework is observation-driven, see Cox et al. (1981). As a result, the log likelihood of the entire sample can be obtained using the prediction-error decomposition and is closed-form given the sequence $\{\theta_{t|t-1}\}$. We therefore propose to estimate ρ and the prediction parameters jointly with possible static parameters using MLE, similar to the GAS methodology. In the two empirical illustrations of the next section, we combine the READY update with the linear prediction step (6). This simplifies estimation and facilitates the comparison with existing methods, which also use a linear specification.

In terms of the theoretical properties of this ML estimator, we borrow from the theory for GAS models as recently established in Blasques et al. (2022). The bottleneck of these proofs is usually filter invertibility. For the class of log-concave location-scale distributions Theorem 2 provides this required filter invertibility. The empirical volatility illustration falls in this category. Under standard assumptions, such as stationarity and ergodicity of the

DGP in combination with sufficient moments, we therefore obtain the desired consistency and asymptotic normality, see Blasques et al. (2022) for details. The other illustration regarding employment falls within the location-scale setting but does not possess a concave log density. In this case, the asymptotic validity of the ML estimator is a conjecture and may require assumptions about the exclusion of badly behaved DGPs. Note that there generally always exists a value of the autoregressive parameter Φ sufficiently close to 0 for which the filter is invertible. Empirically, we find our method to work well even in this non-concave case, similar to how gradient-based optimization methods also tend to work well in practice outside of stylized settings such as the case of global concavity.

5 Empirical illustrations

5.1 Employment

Employment has a rich history as a key indicator of the economy. Furthermore, it is well established that the conditional mean changes alongside the business cycle characterized by periods of slow growth and shorter periods of more rapid decline, see e.g. Stock and Watson (1999). We therefore propose the following dynamic location model for the monthly log growth rates of employment y_t,

$$y_t = \mu_t + \sigma\varepsilon_t, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} t(\nu), \tag{30}$$

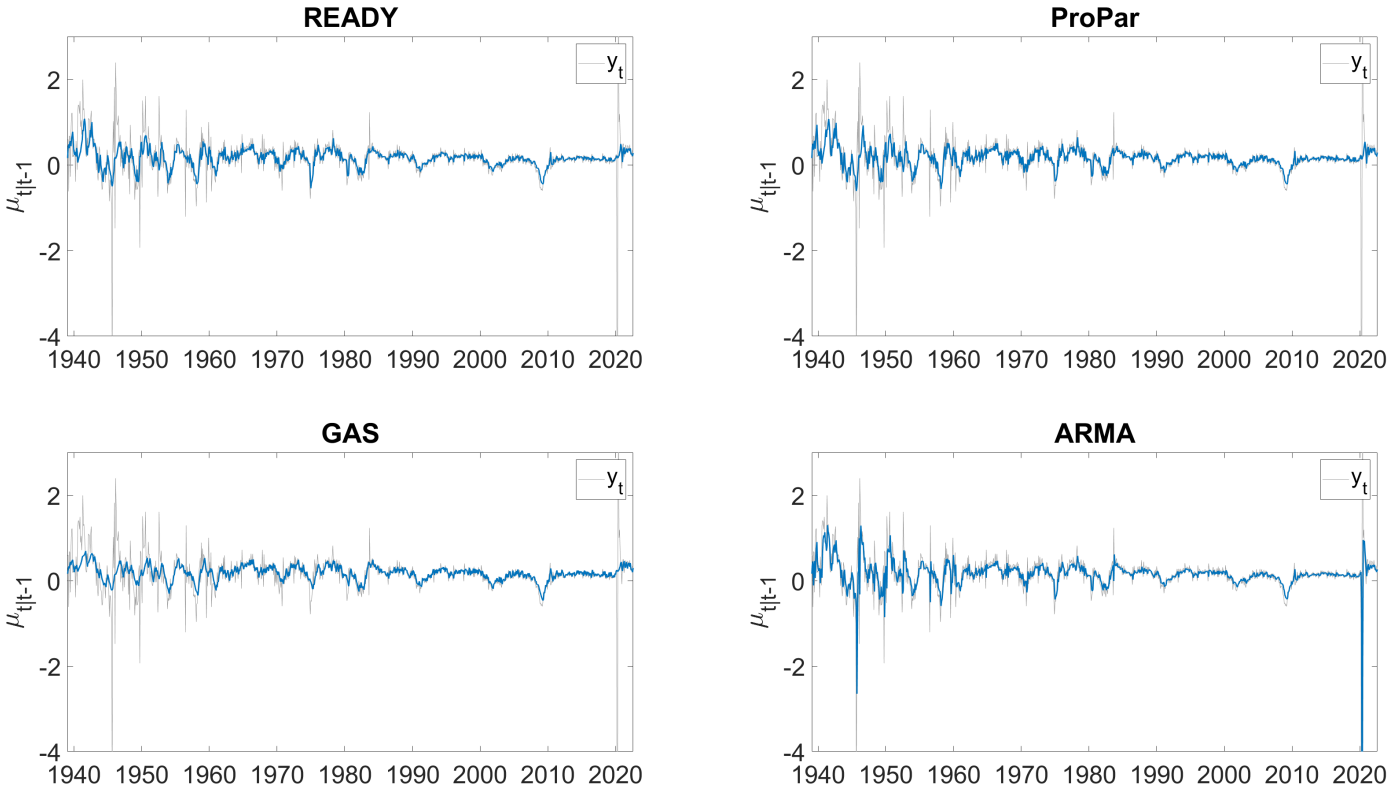
where μ_t denotes the time-varying location, σ a static scale parameter and ε_t an i.i.d. disturbance that follows a Student's t-distribution with degrees of freedom $\nu > 1$. For the prediction step we use a linear specification, that is,

$$\mu_{t+1|t} = \omega + \phi\mu_{t|t}, \tag{31}$$

where $\omega \in \mathbb{R}$ and $\phi \in (-1, 1)$ are parameters.

We consider monthly US employment log growth rates from January 1939 until August 2022 retrieved from the FRED database[1]. Figure 1 displays the evolution of the predictions $\mu_{t|t-1}$ for the READY, ProPar and GAS approaches for the model outlined in (30)-(31). We note that the associated Fisher matrix relevant for the latter two approaches is constant. For comparison purposes, we also include the ARMA(1,1) model with a normal distribution, which is the READY, ProPar and GAS equivalent of (30)-(31) for the normal distribution.

Figure 1: Time-evolution of $\mu_{t|t-1}$ for the READY, ProPar, GAS and ARMA(1,1) models for monthly US employment, January 1939 until August 2022.



Note: the grey line line reflects the evolution of monthly employment itself to compare the models against.

In Figure 1, we observe that all three likelihood-driven models making use of the t-distribution correctly downweight the large shock in April 2020 associated with the onset of the COVID-19 pandemic. The ARMA(1,1) model, however, possesses no such innate

[1] All Employees, Total Nonfarm from https://fred.stlouisfed.org/series/PAYEMS

robustness and makes a large adjustment, which leads to a sizeable forecast error for the next month. This is the strength of linking the evolution of the time-varying parameters to the density as long been recognized by the score-driven literature.

Further comparing the three methods, we find that the READY and ProPar frameworks are highly similar with some more differences relative to the GAS approach. Namely, the READY and ProPar models are more responsive for the earlier decades compared to the GAS model, while all three methods tend to agree on the later dates. This is in line with intuition as the early years are characterized by strong (but sufficiently persistent) time-variation, leading to larger update magnitudes, such that the optimization problems used by the READY, ProPar and GAS methods tend to differ more. Because the fit of the READY and ProPar models is quite close (log likelihoods 59.26 and 58.33, respectively), it appears that a second-order approximation is accurate enough overall, while a first-order approximation by the GAS model leaves some efficiency on the table (log likelihood 16.50). Another possible explanation for the gain of the READY and ProPar models over the GAS method may be due to the robustness of implicit updates against mispecification of the learning rate (e.g. Toulis et al., 2014). As a result, a constant learning rate (ρ^{-1}) for long samples may be less impactful for the READY and ProPar update then for the GAS method.

5.2 Volatility

The workhorse model commonly employed for volatility forecasting is the celebrated GARCH model by Bollerslev (1986). However, it is well known that the conditional distribution of asset returns deviates substantially from the normal, which in turn could make innovating using the squared shock inefficient. The Beta-t-EGARCH model of Harvey and Sucarrat (2014) considers a conditional t-distribution for the error term and updates the logaritmic scale using a score-driven update recursion. As a result, information on the fat-tailed nature of the density is used in the update of the volatility, effectively reducing the impact of very large shocks. We entertain a similar setup but update the volatility using the READY

strategy instead. Specifically, we consider

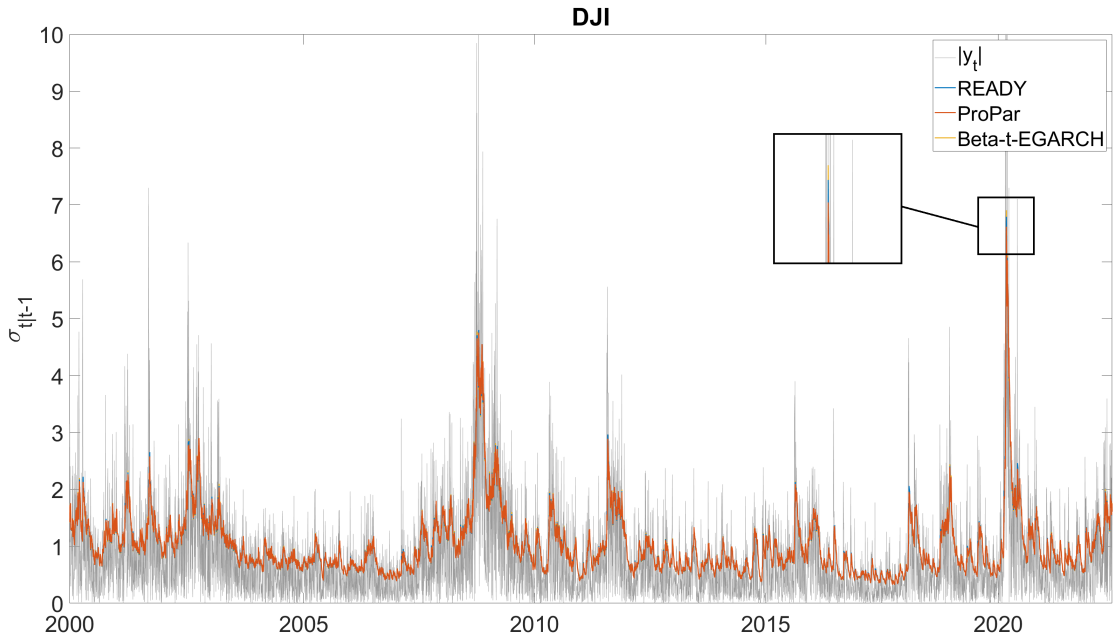
$$y_t = \mu + \exp(\lambda_t)\sqrt{\frac{\nu-2}{\nu}}z_t, \quad z_t \overset{\text{i.i.d.}}{\sim} \mathrm{t}(\nu), \tag{32}$$

$$\lambda_{t+1|t} = \omega + \phi\lambda_{t|t}, \tag{33}$$

where μ denotes a static mean, λ_t the dynamic log volatility, z_t an i.i.d. Student's t-distributed error term with degrees of freedom $\nu > 2$ and $\omega \in \mathbb{R}$ and $\phi \in (-1, 1)$ parameters.

We implement the update portion from $\lambda_{t|t-1}$ to $\lambda_{t|t}$ using the READY, ProPar and Beta-t-EGARCH (i.e. the GAS/DCS equivalent) frameworks, where we note that the associated Fisher information is again constant. Specifically, Figure 2 demonstrates the evolution of the volatility predictions $\sigma_{t|t-1} := \exp(\lambda_{t|t-1})$ of the three models using daily Dow Jones log returns from January 2000 until June 2022, retrieved from the Oxford-Man library[2].

Figure 2: Time-evolution of $\sigma_{t|t-1}$ for the READY, ProPar and Beta-t-EGARCH models for daily Dow Jones index log returns, January 2000 until June 2022.



Note: the grey line represents the absolute returns $|y_t|$, which serve as crude ex-post measures of the daily volatility.

[2]https://realized.oxford-man.ox.ac.uk/

We observe in Figure 2 that the estimates of all three methods are almost identical, with only very minor differences around peaks. This indicates that in this scenario the approximations used by the ProPar and GAS/DCS frameworks are highly accurate, producing updates virtually indistinguishable from the READY update. The differences with the employment illustration may be due to the fact that it is generally harder to estimate a second moment compared to a first moment from a single observation, which is evident from the squared nature of variance proxies. This volatility illustration clearly demonstrates a practical case where the GAS/DCS approach is computationally cheap yet effectively efficient.

6 Conclusion

We propose a new framework for constructing time-varying density models by combining concepts from information theory with optimization techniques. The resulting Relative Entropy Adaptive Density (READY) update maximizes the log-likelihood contribution of the latest observation subject to a Kullback-Leibler divergence (KLD) regularization that penalizes deviations from a one-step ahead predicted density. Because the optimization occurs at the density level, we have that the READY update of multiple time-varying characteristics is automatically joint and unaffected by the choice of parameterization. We demonstrate that the READY update can be viewed as an intuitive regularized estimator of the pseudo-true density. For a single time-varying parameter we derive conditions for a new type of global optimality as well as filter invertibility. Furthermore, we show that the READY framework nests several popular existing time-series models, such as the ARMA and GARCH, and allows for rich connections with the score-driven methods of Creal et al. (2013), Harvey (2013) and Lange et al. (2022). Empirical effectiveness is illustrated by the modeling of employment growth and asset volatility.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. Ed. by B.N. Petrov and F. Caski. Budapest: Akademiai Kiado, 267–81.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Computation* **10**, 251–276.

Asi, H. and J.C. Duchi (2019). Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization* **29**, 2257–2290.

Bianchi, P. (2016). Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization* **26**, 2235–2260.

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

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

Blei, D.M., A. Kucukelbir, and J.D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877.

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

Cen, S., C. Cheng, Y. Chen, Y. Wei, and Y. Chi (2022). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research* **70**, 2563–2578.

Chrétien, S. and A.O. Hero (2000). Kullback proximal algorithms for maximum-likelihood estimation. *IEEE Transactions on Information Theory* **46**, 1800–1810.

— (2008). On EM algorithms and their proximal generalizations. *ESAIM: Probability and Statistics* **12**, 308–326.

Courts, J., A. Wills, T. Schön, and B. Ninness (2023). Variational system identification for nonlinear state-space models. *Automatica* **147**, 110687.

Cox, D.R., G. Gudmundsson, G. Lindgren, L. Bondesson, E. Harsaae, P. Laake, K. Juselius, and S.L. Lauritzen (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics* **8**, 93–115.

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

— (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics* **28**, 777–795.

Desjardins, G., K. Simonyan, R. Pascanu, et al. (2015). Natural neural networks. *Advances in Neural Information Processing Systems* **28**.

Engle, R.F. and T. Bollerslev (1986). Modelling the persistence of conditional variances. *Econometric Reviews* **5**, 1–50.

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

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

Grünwald, P.D. and A.P. Dawid (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics* **32**, 1367–1433.

Harvey, A. (2013). Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series. Vol. **52**. Cambridge University Press.

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

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

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

Kakade, S.M. (2001). A natural policy gradient. *Advances in Neural Information Processing Systems* **14**.

Khan, M.E.E., P. Baqué, F. Fleuret, and P. Fua (2015). Kullback-Leibler proximal variational inference. *Advances in Neural Information Processing Systems* **28**.

Kullback, S. (1959). Information Theory and Statistics. New York: Wiley.

Kullback, S. and R.A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

Lange, R.-J., B. van Os, and D. van Dijk (2022). Robust observation-driven models using proximal-parameter updates. *TI Discussion Papers: Available at Link*.

Lucas, A. and X. Zhang (2016). Score-driven exponentially weighted moving averages and value-at-risk forecasting. *International Journal of Forecasting* **32**, 293–302.

Martens, J. (2020). New insights and perspectives on the natural gradient method. *The Journal of Machine Learning Research* **21**, 5776–5851.

Nielsen, F. (2019). On the Kullback-Leibler divergence between location-scale densities. *ArXiv preprint: Available at Link*.

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

Parikh, N. and S. Boyd (2014). Proximal algorithms. *Foundations and Trends® in Optimization* **1**, 127–239.

Patrascu, A. and I. Necoara (2018). Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *The Journal of Machine Learning Research* **18**, 7204–7245.

Ravikumar, P., A. Agarwal, and M.J. Wainwright (2010). Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *Journal of Machine Learning Research* **11**, 1043–1080.

Rockafellar, R.T. (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* **14**, 877–898.

Roulston, M.S. and L.A. Smith (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review* **130**, 1653–1660.

Ryu, E.K. and S. Boyd (2016). Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent. *Unpublished Manuscript: Available at Link*.

Stock, J.H. and M.W. Watson (1999). Business cycle fluctuations in US macroeconomic time series. *Handbook of Macroeconomics*. Ed. by J. Taylor and M. Woodford. Vol. **1**. Elsevier, 3–64.

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**, 2449–2495.

Taylor, S.J. (2008). Modelling financial time series. New York: Wiley.

Teräsvirta, T. (2009). An introduction to univariate GARCH models. *Handbook of Financial Time Series*. Ed. by T. Mikosch, J. Kreiss, R. Davis, and T. Andersen. Springer, 17–42.

Toulis, P., E.M. Airoldi, and J. Rennie (2014). Statistical analysis of stochastic gradient methods for generalized linear models. *Proceedings of the 31st International Conference on Machine Learning* **32**, 667–675.

Toulis, P., T. Horel, and E.M. Airoldi (2021). The proximal Robbins-Monro method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83**, 188–212.

Toulis, P., D. Tran, and E.M. Airoldi (2016). Towards stability and optimality in stochastic gradient descent. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* **51**, 1290–1298.

Yang, H. and S.-I. Amari (1997). The efficiency and the robustness of natural gradient descent learning rule. *Advances in Neural Information Processing Systems* **10**.

Online Appendix to Information-Theoretic Time-varying Density Modeling

A Examples **1**

A.1 Example 1: ARMA(1,1) 1

A.2 Example 2: GARCH(1,1) 1

A.3 Example 3: AV-GARCH(1,1) 2

A.4 Example 4: Poisson 3

A.5 Example 5: Exponential 3

A.6 Example 6: Binomial 4

B Proofs **5**

B.1 Proposition 1 5

B.2 Proposition 2 5

B.3 Proposition 3 6

B.4 Lemma 1 7

B.5 Theorem 1 8

B.6 Lemma 2 9

B.7 Theorem 2 13

A Examples

A.1 Example 1: ARMA(1,1)

The READY update at time t for a normal distribution with mean μ_t and static variance σ^2 takes the following form

$$\begin{aligned}\mu_{t|t} &= \underset{\mu\in\mathbb{R}}{\text{argmax}} \left\{-\log\left(\sigma\sqrt{2\pi}\right) - \frac{(y_t-\mu)^2}{2\sigma^2} - \rho\left[\log\left(\frac{\sigma}{\sigma}\right) + \frac{\sigma^2+(\mu-\mu_{t|t-1})^2}{2\sigma^2} - \frac{1}{2}\right]\right\} \\ &= \underset{\mu\in\mathbb{R}}{\text{argmax}} \left\{-\frac{(y_t-\mu)^2}{2\sigma^2} - \rho\frac{(\mu-\mu_{t|t-1})^2}{2\sigma^2}\right\},\end{aligned} \tag{A.1}$$

where the second characterization removes all arguments that do not depend on μ. The associated FOC multiplied by σ^2 reads

$$(y_t - \mu_{t|t}) + \rho(\mu_{t|t-1} - \mu_{t|t}) = 0, \tag{A.2}$$

such that rearranging yields

$$\mu_{t|t} = \frac{1}{1+\rho}y_t + \frac{\rho}{1+\rho}\mu_{t|t-1}. \tag{A.3}$$

A.2 Example 2: GARCH(1,1)

The READY update at time t for a normal distribution with mean μ and variance σ_t^2 takes the following form

$$\begin{aligned}\sigma_{t|t}^2 &= \underset{\sigma^2\in\mathbb{R}^+}{\text{argmax}} \left\{-\log\left(\sigma\sqrt{2\pi}\right) - \frac{(y_t-\mu)^2}{2\sigma^2} - \rho\left[\log\left(\frac{\sigma}{\sigma_{t|t-1}}\right) + \frac{\sigma_{t|t-1}^2+(\mu-\mu)^2}{2\sigma^2} - \frac{1}{2}\right]\right\} \\ &= \underset{\sigma^2\in\mathbb{R}^+}{\text{argmax}} \left\{-\frac{1}{2}\log\left(\sigma^2\right) - \frac{(y_t-\mu)^2}{2\sigma^2} - \rho\left[\frac{1}{2}\log\left(\frac{\sigma^2}{\sigma_{t|t-1}^2}\right) + \frac{\sigma_{t|t-1}^2}{2\sigma^2}\right]\right\}, \\ &= \underset{\sigma^2\in\mathbb{R}^+}{\text{argmax}} \left\{-\log\left(\sigma^2\right) - \frac{(y_t-\mu)^2}{\sigma^2} - \rho\left[\log\left(\sigma^2\right) + \frac{\sigma_{t|t-1}^2}{\sigma^2}\right]\right\},\end{aligned} \tag{A.4}$$

where the final line multiplies by a factor two and removes terms that do not depend on σ^2. The associated FOC reads

$$-\frac{1}{\sigma_{t|t}^2}(1+\rho)+\frac{(y_t-\mu)^2}{\sigma_{t|t}^4}+\rho\frac{\sigma_{t|t-1}^2}{\sigma_{t|t}^4}=0, \tag{A.5}$$

such that multiplying by $\sigma_{t|t}^4$ and rearranging yields

$$\sigma_{t|t}^2=\frac{1}{1+\rho}(y_t-\mu)^2+\frac{\rho}{1+\rho}\sigma_{t|t-1}^2. \tag{A.6}$$

A.3 Example 3: AV-GARCH(1,1)

The READY update at time t for a Laplace distribution with static location μ and scale σ_t is given as

$$\begin{aligned}\sigma_{t|t}&=\underset{\sigma\in\mathbb{R}^+}{\text{argmax}}\left\{-\log(2\sigma)-\frac{|y_t-\mu|}{\sigma}-\rho\left[\log\left(\frac{\sigma}{\sigma_{t|t-1}}\right)+\frac{\sigma_{t|t-1}}{\sigma}-1\right]\right\}\\&=\underset{\sigma\in\mathbb{R}^+}{\text{argmax}}\left\{-\log(\sigma)-\frac{|y_t-\mu|}{\sigma}-\rho\left[\log(\sigma)+\frac{\sigma_{t|t-1}}{\sigma}\right]\right\},\end{aligned} \tag{A.7}$$

where the associated FOC is given by

$$-\frac{1}{\sigma_{t|t}}(1+\rho)+\frac{|y_t-\mu|}{\sigma_{t|t}^2}+\rho\frac{\sigma_{t|t-1}}{\sigma_{t|t}^2}=0. \tag{A.8}$$

Multiplying the FOC with $\sigma_{t|t}^2$ and rearranging yields

$$\sigma_{t|t}=\frac{1}{1+\rho}|y_t-\mu|+\frac{\rho}{1+\rho}\sigma_{t|t-1}. \tag{A.9}$$

A.4 Example 4: Poisson

The READY update at time t for a Poisson distribution with parameter λ_t using the observation $y_t \in \mathbb{N}_0$ is given as

$$\begin{aligned}\lambda_{t|t} &= \underset{\lambda\in\mathbb{R}^+}{\operatorname{argmax}} \left\{-\log(y_t!) - \lambda + y_t\log(\lambda) - \rho\left[\lambda_{t|t-1}\log\Big(\frac{\lambda_{t|t-1}}{\lambda}\Big) + \lambda - \lambda_{t|t-1}\right]\right\} \\ &= \underset{\lambda\in\mathbb{R}^+}{\operatorname{argmax}} \left\{-\lambda + y_t\log(\lambda) - \rho\left[\lambda - \lambda_{t|t-1}\log(\lambda)\right]\right\},\end{aligned} \tag{A.10}$$

where the second form removes terms that do not depend on λ. The associated FOC reads

$$-(1+\rho) + \frac{y_t + \rho\lambda_{t|t-1}}{\lambda_{t|t}} = 0. \tag{A.11}$$

Multiplying with $\lambda_{t|t}$ and rearranging yields

$$\lambda_{t|t} = \frac{1}{1+\rho}y_t + \frac{\rho}{1+\rho}\lambda_{t|t-1}. \tag{A.12}$$

A.5 Example 5: Exponential

The READY update at time t for an exponential distribution with scale λ_t using the observation $y_t \geq 0$ is given as

$$\begin{aligned}\theta_{t|t} &= \underset{\theta\in\mathbb{R}^+}{\operatorname{argmax}} \left\{-\log(\theta) - \frac{y_t}{\theta} - \rho\left[\log\Big(\frac{\theta}{\theta_{t|t-1}}\Big) + \frac{\theta_{t|t-1}}{\theta} - 1\right]\right\} \\ &= \underset{\theta\in\mathbb{R}^+}{\operatorname{argmax}} \left\{-\log(\theta) - \frac{y_t}{\theta} - \rho\left[\log(\theta) + \frac{\theta_{t|t-1}}{\theta}\right]\right\},\end{aligned} \tag{A.13}$$

where the second form removes terms that do not depend on θ. The associated FOC reads

$$-\frac{1+\rho}{\theta_{t|t}} + \frac{y_t + \rho\theta_{t|t-1}}{\theta_{t|t}^2} = 0. \tag{A.14}$$

Multiplying with $\theta_{t|t}^2$ and rearranging yields

$$\theta_{t|t} = \frac{1}{1+\rho} y_t + \frac{\rho}{1+\rho} \theta_{t|t-1}. \tag{A.15}$$

A.6 Example 6: Binomial

The READY update at time t for a binomial distribution with fixed number of trails n and probability of success θ_t using the observation $y_t \in \{0, 1, \ldots, n\}$, directly suppressing terms that do not depend on θ for brevity, is given as

$$\theta_{t|t} = \underset{\theta \in \in [0,1]}{\operatorname{argmax}} \left\{ y_t \log(\theta) + (n - y_t) \log(1-\theta) + \rho \left[\log(\theta) n \theta_{t|t-1} + \log(1-\theta) n (1 - \theta_{t|t-1}) \right] \right\}. \tag{A.16}$$

The associated FOC reads

$$\frac{y_t + \rho n \theta_{t|t-1}}{\theta_{t|t}} - \frac{(n - y_t) + \rho n (1 - \theta_{t|t-1})}{1 - \theta_{t|t}} = 0. \tag{A.17}$$

Multiplying with $\theta_{t|t}(1 - \theta_{t|t})$ yields

$$\theta_{t|t} = \frac{y_t + \rho n \theta_{t|t-1}}{y_t + \rho n \theta_{t|t-1} + (n - y_t) + \rho n (1 - \theta_{t|t-1})}, \tag{A.18}$$

which can be written as

$$\theta_{t|t} = \frac{1}{1+\rho} \frac{y_t}{n} + \frac{\rho}{1+\rho} \theta_{t|t-1}. \tag{A.19}$$

B Proofs

B.1 Proposition 1

This result is similar to the reparameterization invariance of the ML estimator and follows directly from the definitions.

B.2 Proposition 2

Using Assumptions 1-3 the READY update solves the first-order condition (FOC) given as

$$\nabla(y_t|\theta_{t|t}) = -\rho\frac{\partial}{\partial\theta}\mathbb{E}_{\substack{y\\ t|t-1}}[\log\, p(y|\theta)]\Big|_{\theta=\theta_{t|t}}. \tag{B.20}$$

Using the interchangability of the partial derivative operator and the expectation (Assumption 4) and the fact that $\mathbb{E}_{\substack{y\\ t|t-1}}[\nabla(y_t|\theta_{t|t-1})] = 0$ (which follows from differentiability and identification), we obtain

$$\nabla(y_t|\theta_{t|t}) = -\rho\mathbb{E}_{\substack{y\\ t|t-1}}[\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_{t|t-1})]. \tag{B.21}$$

Using second-order continuous differentiability (Assumption 3), we may then write the difference in gradients on the right-hand side as an integral involving the Jacobian of the gradient (i.e. the Hessian),

$$\nabla(y_t|\theta_{t|t}) = -\rho\mathbb{E}_{\substack{y\\ t|t-1}}\left[\int_0^1 \frac{\partial^2}{\partial\theta\partial\theta'}\log p(y|\theta)\Big|_{\theta_{t|t-1}+q(\theta_{t|t}-\theta_{t|t-1})}\mathrm{d}q\right](\theta_{t|t}-\theta_{t|t-1}). \tag{B.22}$$

Using the definition of $\mathcal{I}_{t|t-1}(\theta_{t|t})$ this gives the first result. Left-multiplying on both sides by the inverse of $\mathcal{I}_{t|t-1}(\theta_{t|t})$, provided it exists, and rearranging gives the second result.

B.3 Proposition 3

By the assumption of identification, we have that $\mathcal{D}_{t|t-1}(\theta)$ is uniquely minimized at $\theta = \theta_{t|t-1}$. Therefore, if $\theta_{t|t} \neq \theta_{t|t-1}$, we have $\mathcal{D}_{t|t-1}(\theta_{t|t}) > \mathcal{D}_{t|t-1}(\theta_{t|t-1}) = 0$. Because $\theta_{t|t}$ is the unique maximizer of the objective $\log p(y_t|\theta) - \rho\mathcal{D}_{t|t-1}(\theta)$ by Assumption 1, we have

$$\log p(y_t|\theta_{t|t}) - \rho\mathcal{D}_{t|t-1}(\theta_{t|t}) > \log p(y_t|\theta_{t|t-1}) - \rho\mathcal{D}_{t|t-1}(\theta_{t|t-1}) = \log p(y_t|\theta_{t|t-1}), \tag{B.23}$$

from which it directly follows that $\log p(y_t|\theta_{t|t}) > \log p(y_t|\theta_{t|t-1})$. In other words, because $\theta_{t|t}$ is strictly worse than $\theta_{t|t-1}$ in terms of penalty, it must have a strictly higher likelihood fit (otherwise it would not be the unique maximizer of $\log p(y_t|\theta) - \rho\mathcal{D}_{t|t-1}(\theta)$). Monotonicity of the logarithm then provides $p(y_t|\theta_{t|t}) > p(y_t|\theta_{t|t-1})$, which is the first result.

Having obtained that $\log p(y_t|\theta_{t|t}) > \log p(y_t|\theta_{t|t-1})$ if $\theta_{t|t} \neq \theta_{t|t-1}$, we may use the assumed continuity of the density in its first argument to extend this improvement to a neighbourhood of y_t. Specifically, using an ε-δ argument we have that $\exists\delta > 0$ such that for $\mathcal{Y} := \{y \in \text{Dom}(y)|\ \|y - y_t\|^2 \leq \delta\ \}$ we have that $\forall\tilde{y} \in \mathcal{Y}$,

$$\log p(\tilde{y}|\theta_{t|t}) - \log p(\tilde{y}|\theta_{t|t-1}) > \varepsilon > 0. \tag{B.24}$$

In addition, we also have that $\Pr(y \in \mathcal{Y}|\theta_t^0) := \int_{\mathcal{Y}} p^0(y|\theta_t^0)\mathrm{d}y > 0$. This follows directly from the fact that $\mathcal{Y}$ has non-zero Lebesgue measure. In other words, if $\Pr(y \in \mathcal{Y}|\theta_t^0) = 0$, y_t would not be a possible outcome. Combining these two insights we obtain the final result,

$$\Delta_t(\mathcal{Y}) := \mathbb{E}_{y_t}^0[\log p(y|\theta_{t|t}) - \log p(y|\theta_{t|t-1})|y \in \mathcal{Y}] > 0. \tag{B.25}$$

B.4 Lemma 1

From Proposition 2, we have that

$$\nabla(y_t|\theta_{t|t}) = \rho\mathcal{I}_{t|t-1}(\theta_{t|t})(\theta_{t|t} - \theta_{t|t-1}). \tag{B.26}$$

Using the assumption of the existence of the one-period ML estimator $\hat{\theta}_t \in \text{Int}(\Theta)$ and differentiability (Assumption 3), we have that $\nabla(y_t|\hat{\theta}_t) = 0$. Subtracting $\nabla(y_t|\hat{\theta}_t)$ on the left-hand side and 0 on the right-hand side and again writing the difference in gradients as an integral involving the Hessian we obtain,

$$\mathcal{H}(y_t|\theta_{t|t})(\theta_{t|t} - \hat{\theta}_t) = \rho\mathcal{I}_{t|t-1}(\theta_{t|t})(\theta_{t|t} - \theta_{t|t-1}), \tag{B.27}$$

where $\mathcal{H}(y_t|\theta_{t|t})$ is given as

$$\mathcal{H}(y_t|\theta_{t|t}) := \left[\int_0^1 \frac{\partial^2}{\partial\theta\partial\theta'}\log p(y_t|\theta)\Big|_{\hat{\theta}_t + q(\theta_{t|t} - \hat{\theta}_t)} \mathrm{d}q\right]. \tag{B.28}$$

Collecting all terms involving $\theta_{t|t}$ on the left, we obtain

$$[\mathcal{H}(y_t|\theta_{t|t}) - \rho\mathcal{I}_{t|t-1}(\theta_{t|t})]\theta_{t|t} = \hat{\mathcal{H}}_t\hat{\theta}_t - \rho\mathcal{I}_{t|t-1}(\theta_{t|t})\theta_{t|t-1}. \tag{B.29}$$

Finally, multiplying with $[\mathcal{H}(y_t|\theta_{t|t}) - \rho\mathcal{I}_{t|t-1}(\theta_{t|t})]^{-1}$ gives

$$\theta_{t|t} = [I_K - A(y_t|\theta_{t|t}, \theta_{t|t-1})]\hat{\theta}_t + A(y_t|\theta_{t|t}, \theta_{t|t-1})\theta_{t|t-1}, \tag{B.30}$$

where I_K denotes the K-dimensional identity matrix and $A(y_t|\theta_{t|t}, \theta_{t|t-1}) := [\rho\mathcal{I}_{t|t-1}(\theta_{t|t}) - \mathcal{H}(y_t|\theta_{t|t})]^{-1}\rho\mathcal{I}_{t|t-1}(\theta_{t|t})$. Note that invertibility of $\rho\mathcal{I}_{t|t-1}(\theta_{t|t}) - \mathcal{H}(y_t|\theta_{t|t})$ follows from the fact that $-\mathcal{H}(y_t|\theta_{t|t}) > 0$ and $\mathcal{I}_{t|t-1}(\theta_{t|t}) > 0$. This is a result of the concavity of the log

density in θ (Assumption 5) in combination with the following second-order conditions

$$\frac{\partial^2}{\partial\theta\partial\theta'}\log p(y_t|\theta)\Big|_{\hat{\theta}_t} < 0, \quad \frac{\partial^2}{\partial\theta\partial\theta'}\mathbb{E}_{y_t|t-1}\left[\log p(y|\theta)\right]\Big|_{\theta_{t|t-1}} = -\mathcal{I}_{t|t-1}(\theta_{t|t-1}) < 0, \tag{B.31}$$

where the second equation uses the interchangability of derivative and expectation (Assumption 4). These second-order conditions are due to the fact that $\hat{\theta}_t$ and $\theta_{t|t-1}$ uniquely maximize $\log p(y_t|\theta)$ and $\mathbb{E}_{y_t|t-1}[\log p(y|\theta)]$ in combination with second-order differentiability and interior values (Assumption 2, $\hat{\theta}_t \in \text{Int}(\Theta)$ and Assumption 3). By continuity of the Hessian (Assumption 3) it follows that $-\mathcal{H}(y_t|\theta_{t|t}) > 0$ and $\mathcal{I}_{t|t-1}(\theta_{t|t}) > 0$, such that $\rho\mathcal{I}_{t|t-1}(\theta_{t|t}) - \mathcal{H}(y_t|\theta_{t|t}) > 0$, which yields $0 < A(y_t|\theta_{t|t}, \theta_{t|t-1}) < 1$.

Because the log likelihood is concave in θ for all y, it follows that any KLD from $p(\cdot|\theta)$ is convex in θ, such that,

$$\mathcal{D}_t^*(\theta_{t|t}) \leq A(y_t|\theta_{t|t}, \theta_{t|t-1})\mathcal{D}_t^*(\theta_{t|t-1}) + [1 - A(y_t|\theta_{t|t}, \theta_{t|t-1})]\mathcal{D}_t^*(\hat{\theta}_t), \tag{B.32}$$

where the quantities are finite-valued by Assumption 6. This completes the proof.

B.5 Theorem 1

By quantifying the minimum and maximum curvature of the log-likelihood contribution and the penalty, we may obtain bounds for $A(y_t|\theta_{t|t}, \theta_{t|t-1})$ found in Lemma 1. That is, assuming that $\log p(y_t|\theta)$ is $\alpha_t(y_t)$-strongly concave and $\beta_t(y_t)$-smooth in θ and $\mathcal{D}_{t|t-1}(\theta)$ is $\tilde{\alpha}_t$-strongly convex and $\tilde{\beta}_t$-smooth in θ, we obtain,

$$A(y_t|\theta_{t|t}, \theta_{t|t-1}) = \frac{\rho\mathcal{I}_{t|t-1}(\theta_{t|t})}{\rho\mathcal{I}_{t|t-1}(\theta_{t|t}) - \mathcal{H}(y_t|\theta_{t|t})} \leq \frac{\rho\tilde{\beta}_t}{\rho\tilde{\beta}_t + \alpha_t} \in (0, 1), \tag{B.33}$$

and, similarly,

$$A(y_t|\theta_{t|t}, \theta_{t|t-1}) \geq \frac{\beta_t}{\rho\tilde{\alpha}_t + \beta_t} \in (0, 1). \tag{B.34}$$

Filling in these bounds for $A(y_t|\theta_{t|t}, \theta_{t|t-1})$ and taking the expectation over y_t using the true distribution, we obtain

$$\mathbb{E}^0_{y_t}[\mathcal{D}^*_t(\theta_{t|t})] \leq \mathbb{E}^0_{y_t}[\frac{\rho\tilde{\beta}_t}{\rho\tilde{\beta}_t + \alpha_t}]\mathcal{D}^*_t(\theta_{t|t-1}) + \mathbb{E}^0_{y_t}[\frac{\beta_t}{\rho\tilde{\alpha}_t + \beta_t}\mathcal{D}^*_t(\hat{\theta}_t)], \tag{B.35}$$

which in turn provides the final result:

$$\mathbb{E}^0_{y_t}[\mathcal{D}^*_t(\theta_{t|t})] \leq \eta_t\mathcal{D}^*_t(\theta_{t|t-1}) + \lambda_t\sigma_t, \tag{B.36}$$

$$\eta_t := \mathbb{E}^0_{y_t}[\frac{\rho\tilde{\beta}_t}{\rho\tilde{\beta}_t + \alpha_t}] \in (0,1), \quad \lambda_t := \frac{\mathbb{E}^0_{y_t}[\frac{\beta_t}{\rho\tilde{\alpha}_t+\beta_t}\mathcal{D}^*_t(\hat{\theta}_t)]}{\mathbb{E}^0_{y_t}[\mathcal{D}^*_t(\hat{\theta}_t)]} \in (0,1), \quad \sigma_t := \mathbb{E}^0_{y_t}[\mathcal{D}^*_t(\hat{\theta}_t)] \in (0,\infty). \tag{B.37}$$

B.6 Lemma 2

We first prove the non-expansiveness result of Lemma 2 and afterwards use strong concavity and smoothness to quantify the contraction rate. Consider two predictions $\theta_{t|t-1}, \tilde{\theta}_{t|t-1} \in \text{Int}(\Theta) \subseteq \mathbb{R}$. By concavity (Assumption 5) and differentiability (Assumption 3), there is at most a single interval $I = (a, b) \subseteq \Theta$, with $a \leq b$ for which $\nabla(y_t|\theta) = 0$, $\forall\theta \in I$. This is because multiple such intervals would contradict concavity.

Assume that this interval I exists, then we need to consider four subcases. First, if both $\theta_{t|t-1} \in I$ and $\tilde{\theta}_{t|t-1} \in I$, we obtain $\theta_{t|t} = \theta_{t|t-1}$ and $\tilde{\theta}_{t|t} = \tilde{\theta}_{t|t-1}$, such that the first result of Lemma 2 trivially holds. Second, suppose that one of the two predictions is within the interval and one is not. That is, without loss of generality assume that $\theta_{t|t-1} < a$ and $\tilde{\theta}_{t|t-1} \in I$. From the objective function used to obtain $\theta_{t|t}$, we can see that $\forall\theta \in \Theta$, $\theta > \tilde{\theta}_{t|t-1}$ and $\forall\theta \in \Theta$, $\theta < \theta_{t|t-1}$ are strictly dominated by $\tilde{\theta}_{t|t-1}$ and $\theta_{t|t-1}$, respectively. This because these points have both an equal or higher log likelihood as well as a strictly smaller penalty (by concavity and identification a strictly larger update yields a strictly larger penalty). Combined with the fact that $\tilde{\theta}_{t|t} = \tilde{\theta}_{t|t-1}$ it follows that $\theta_{t|t-1} \leq \theta_{t|t} \leq \tilde{\theta}_{t|t} = \tilde{\theta}_{t|t-1}$, such that

$\|\theta_{t|t} - \tilde{\theta}_{t|t}\| \leq \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|$. Third, suppose that neither point is within the interval I and one of the two is on the left of I, while the other is on the right of I. It easy to see that both predictions move towards I, but cannot 'cross', such that again the updated distance is smaller than those of the predictions. Fourth, neither point is within I and both are on the same side. This subcase can be treated the same as if the interval I does not exist.

Assuming that either I does not exist or that neither prediction is contained in I and on the same side, then we have that both predictions have a non-zero gradient of the same sign. Without loss of generality assume that $\tilde{\theta}_{t|t-1} < \theta_{t|t-1}$ and $\nabla(y_t|\tilde{\theta}_{t|t-1}) > 0$, $\nabla(y_t|\theta_{t|t-1}) > 0$. Under Assumptions 2 and 3, it follows that the updates solve the following FOCs:

$$\nabla(y_t|\tilde{\theta}_{t|t}) = -\rho \frac{\partial}{\partial \theta} \tilde{\mathbb{E}}_{y_{t|t-1}}[\log p(y|\theta)]\Big|_{\theta=\tilde{\theta}_{t|t}}, \quad \nabla(y_t|\theta_{t|t}) = -\rho \frac{\partial}{\partial \theta} \mathbb{E}_{y_{t|t-1}}[\log p(y|\theta)]\Big|_{\theta=\theta_{t|t}}, \tag{B.38}$$

where $\tilde{\mathbb{E}}_{y_{t|t-1}}$ denotes the expectation over y using the prediction $p(\cdot|\tilde{\theta}_{t|t-1})$. Using both concavity and the translational invariance of the penalty (Assumption 7), we may deduce that the ordering is maintained, that is, $\tilde{\theta}_{t|t} \leq \theta_{t|t}$. Namely, looking at the FOCs these assumptions imply that the left-hand sides are non-increasing in θ, while the right-hand sides are non-decreasing. Specifically, if $\tilde{\theta}_{t|t} \geq \theta_{t|t}$, then $\nabla(y_t|\tilde{\theta}_{t|t}) \leq \nabla(y_t|\theta_{t|t})$ by concavity and similarly $-\rho \frac{\partial}{\partial \theta} \tilde{\mathbb{E}}_{y_{t|t-1}}[\log\ p(y|\theta)]\Big|_{\theta=\tilde{\theta}_{t|t}} \geq -\rho \frac{\partial}{\partial \theta} \mathbb{E}_{y_{t|t-1}}[\log\ p(y|\theta)]\Big|_{\theta=\theta_{t|t}}$ by concavity and translational invariance. The latter inequality is due the fact that the penalty only depends on the Euclidean distance between update and prediction combined with that if $\tilde{\theta}_{t|t} > \theta_{t|t}$ and $\tilde{\theta}_{t|t-1} < \theta_{t|t-1}$ then $\|\tilde{\theta}_{t|t} - \tilde{\theta}_{t|t-1}\| > \|\theta_{t|t} - \theta_{t|t-1}\|$. By the Assumption of uniqueness of the updates (Assumption 1), we must have that at least one of the two is a strict inequality if $\tilde{\theta}_{t|t} > \theta_{t|t}$. Together with $\nabla(y_t|\theta_{t|t}) = -\rho \frac{\partial}{\partial \theta} \mathbb{E}_{y_{t|t-1}}[\log p(y|\theta)]\Big|_{\theta=\theta_{t|t}}$, it follows that $\forall \theta \in \Theta$, $\theta > \theta_{t|t}$ are not solutions of the FOC of $\tilde{\theta}_{t|t}$, i.e., $\theta_{t|t}$ dominates all points above it. Conversely, if $\tilde{\theta}_{t|t} > \theta_{t|t}$ this would directly invalidate $\theta_{t|t}$ as the unique maximizer of its objective.

It remains to be proven that the movement of $\tilde{\theta}_{t|t-1}$ to $\tilde{\theta}_{t|t}$ upwards is at least as large as the movement of $\theta_{t|t-1}$ to $\theta_{t|t}$. In other words, we require that $\tilde{\theta}_{t|t} - \tilde{\theta}_{t|t-1} \geq \theta_{t|t} - \theta_{t|t-1}$. Consider

the candidate point $\bar{\theta} = \tilde{\theta}_{t|t-1} + \theta_{t|t} - \theta_{t|t-1}$. It follows that if and only if $\tilde{\theta}_{t|t} \geq \bar{\theta}$ that we have the required $\tilde{\theta}_{t|t} - \tilde{\theta}_{t|t-1} \geq \theta_{t|t} - \theta_{t|t-1}$. Because of the translational invariance of the penalty and the fact that $\|\bar{\theta} - \tilde{\theta}_{t|t-1}\| = \|\theta_{t|t} - \theta_{t|t-1}\|$ it follows that $-\rho \frac{\partial}{\partial \theta} \tilde{\mathbb{E}}_{y\,t|t-1}[\log\, p(y|\theta)]\Big|_{\theta=\bar{\theta}} = -\rho \frac{\partial}{\partial \theta} \mathbb{E}_{y\,t|t-1}[\log\, p(y|\theta)]\Big|_{\theta=\theta_{t|t}}$. Combined with $\nabla(y_t|\tilde{\theta}_{t|t}) \geq \nabla(y_t|\theta_{t|t})$ by concavity and the uniqueness of the updates, we conclude that $\forall \theta \in \Theta$, $\theta < \bar{\theta}$ are not solutions of the FOC of $\tilde{\theta}_{t|t}$, i.e., $\bar{\theta}$ strictly dominates all points below it. Conversely, if $\tilde{\theta}_{t|t} < \bar{\theta}$ this would invalidate $\theta_{t|t}$ as the unique maximizer of its objective. Together with the order preserving ($\tilde{\theta}_{t|t} \leq \theta_{t|t}$), this means again that also in this final scenario that $\|\theta_{t|t} - \tilde{\theta}_{t|t}\| \leq \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|$.

Having established the non-expansiveness of the update, we may construct an upper-bound for the contraction rate using α_t-strong concavity of the log-likelihood contribution and $\tilde{\beta}_t$-smoothness of the penalty. First, note that should the interval I now exist, then it consists of only a single point. In particular, this point in I is then the one-period ML estimator $\hat{\theta}_t$. From the proof of Lemma 1, we have that

$$\theta_{t|t} = [I_K - A(y_t|\theta_{t|t}, \theta_{t|t-1})]\hat{\theta}_t + A(y_t|\theta_{t|t}, \theta_{t|t-1})\theta_{t|t-1}, \tag{B.39}$$

where $A(y_t|\theta_{t|t}, \theta_{t|t-1}) = [\rho\mathcal{I}_{t|t-1}(\theta_{t|t}) - \mathcal{H}(y_t|\theta_{t|t})]^{-1}\rho\mathcal{I}_{t|t-1}(\theta_{t|t})$ and the invertibility of $[\rho\mathcal{I}_{t|t-1}(\theta_{t|t}) - \mathcal{H}(y_t|\theta_{t|t})]$ is guaranteed by strong concavity. Similar to the argument used in Theorem 1, we may upperbound $A(y_t|\theta_{t|t}, \theta_{t|t-1})$ as

$$A(y_t|\theta_{t|t}, \theta_{t|t-1}) = \frac{\rho\mathcal{I}_{t|t-1}(\theta_{t|t})}{\rho\mathcal{I}_{t|t-1}(\theta_{t|t}) - \mathcal{H}(y_t|\theta_{t|t})} \leq \frac{\rho\tilde{\beta}_t}{\rho\tilde{\beta}_t + \alpha_t} =: \zeta_t \in (0, 1). \tag{B.40}$$

Using this bound we obtain that

$$\|\theta_{t|t} - \hat{\theta}_t\| = |A(y_t|\theta_{t|t}, \theta_{t|t-1})|\|\theta_{t|t-1} - \hat{\theta}_t\| \leq \zeta_t\|\theta_{t|t-1} - \hat{\theta}_t\|, \tag{B.41}$$

with a similar expression for $\tilde{\theta}_{t|t}$. It is not hard to see that therefore if $\tilde{\theta}_{t|t-1} \leq \hat{\theta}_t \leq \theta_{t|t-1}$ or

$\theta_{t|t-1} \leq \hat{\theta}_t \leq \tilde{\theta}_{t|t-1}$ that

$$\|\theta_{t|t} - \tilde{\theta}_{t|t}\| \leq \zeta_t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|. \tag{B.42}$$

We now only need to consider the case where both predictions have a non-zero gradient of the same sign (i.e. either $\hat{\theta}_t$ does not exist or if it exists then both predictions are either strictly smaller or strictly larger than it). Without loss of generality again assume that $\tilde{\theta}_{t|t-1} < \theta_{t|t-1}$ and $\nabla(y_t|\tilde{\theta}_{t|t-1}) > 0$, $\nabla(y_t|\theta_{t|t-1}) > 0$ and consider the candidate point $\bar{\theta} = \tilde{\theta}_{t|t-1} + \theta_{t|t} - \theta_{t|t-1}$. In the proof of non-expansiveness above we have established that $\bar{\theta} \leq \tilde{\theta}_{t|t} \leq \theta_{t|t}$. Using the FOC, α_t-strong concavity of the log-likelihood contribution $\tilde{\beta}_t$-smoothness of the penalties we obtain,

$$\begin{aligned} \nabla(y_t|\theta_{t|t}) + \alpha_t(\theta_{t|t} - \tilde{\theta}_{t|t}) \leq \nabla(y_t|\tilde{\theta}_{t|t}) &= -\rho \frac{\partial}{\partial \theta} \tilde{\mathbb{E}}_{y\, t|t-1}[\log\, p(y|\theta)]\Big|_{\theta=\tilde{\theta}_{t|t}} \\ &\leq -\rho \frac{\partial}{\partial \theta} \tilde{\mathbb{E}}_{y\, t|t-1}[\log\, p(y|\theta)]\Big|_{\theta=\bar{\theta}} + \rho\tilde{\beta}_t(\tilde{\theta}_{t|t} - \bar{\theta}). \end{aligned} \tag{B.43}$$

Note that $-\rho \frac{\partial}{\partial \theta} \tilde{\mathbb{E}}_{y\, t|t-1}[\log\, p(y|\theta)]\Big|_{\theta=\bar{\theta}} = -\rho \frac{\partial}{\partial \theta} \mathbb{E}_{y\, t|t-1}[\log\, p(y|\theta)]\Big|_{\theta=\theta_{t|t}}$ because of translational invariance of the penalty and the fact that $\|\bar{\theta} - \tilde{\theta}_{t|t-1}\| = \|\theta_{t|t} - \theta_{t|t-1}\|$. Also using the other FOC, we have that $\nabla(y_t|\theta_{t|t}) = -\rho \frac{\partial}{\partial \theta} \mathbb{E}_{y\, t|t-1}[\log\, p(y|\theta)]\Big|_{\theta=\theta_{t|t}}$. Together this means that, $\nabla(y_t|\theta_{t|t}) = -\rho \frac{\partial}{\partial \theta} \tilde{\mathbb{E}}_{y\, t|t-1}[\log\, p(y|\theta)]\Big|_{\theta=\bar{\theta}}$. Combining this with (B.43), we obtain

$$\alpha_t(\theta_{t|t} - \tilde{\theta}_{t|t}) \leq \rho\tilde{\beta}_t(\tilde{\theta}_{t|t} - \bar{\theta}), \tag{B.44}$$

where filling in the definition of $\bar{\theta}$ yields

$$\alpha_t(\theta_{t|t} - \tilde{\theta}_{t|t}) \leq \rho\tilde{\beta}_t(\tilde{\theta}_{t|t} - \tilde{\theta}_{t|t-1} - \theta_{t|t} + \theta_{t|t-1}). \tag{B.45}$$

Further rearranging in turn gives,

$$(\theta_{t|t} - \tilde{\theta}_{t|t}) \leq \frac{\rho\tilde{\beta}_t}{\alpha_t + \rho\tilde{\beta}_t}(\theta_{t|t-1} - \tilde{\theta}_{t|t-1}), \tag{B.46}$$

where both sides are positive because $\tilde{\theta}_{t|t-1} < \theta_{t|t-1}$ and $\tilde{\theta}_{t|t} < \theta_{t|t}$. Taking the absolute value and using the definition of ζ_t gives the final result

$$\|\theta_{t|t} - \tilde{\theta}_{t|t}\| \leq \zeta_t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|. \tag{B.47}$$

B.7 Theorem 2

Using the result of Lemma 2 and the submultiplicativity property of Lipschitz mappings, it follows that the composed mapping from prediction to prediction for each point in time has at most Lipschitz coefficient $|\phi| < 1$. The composition of all these $|\phi|$-Lipschitz contracting mappings trivially produces exponential almost sure convergence of the parameter paths. By norm equivalence it follows that this convergence happens in every norm.