TI 2014-103/III
Tinbergen Institute Discussion Paper

Optimal Formulations for Nonlinear Autoregressive Processes

Francisco Blasques
Siem Jan Koopman
André Lucas

Faculty of Economics and Business Administration, VU University Amsterdam, and Tinbergen Institute.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Optimal Formulations for Nonlinear Autoregressive Processes

Francisco Blasques[a,b,c], Siem Jan Koopman[b,c,d], and André Lucas[a,c]

(a) *Department of Finance, VU University Amsterdam*

(b) *Department of Econometrics, VU University Amsterdam*

(c) *Tinbergen Institute*

(d) *CREATES, Aarhus University*

August 7, 2014

Abstract

We develop optimal formulations for nonlinear autoregressive models by representing them as linear autoregressive models with time-varying temporal dependence coefficients. We propose a parameter updating scheme based on the score of the predictive likelihood function at each time point. The resulting time-varying autoregressive model is formulated as a nonlinear autoregressive model and is compared with threshold and smooth-transition autoregressive models. We establish the information theoretic optimality of the score driven nonlinear autoregressive process and the asymptotic theory for maximum likelihood parameter estimation. The performance of our model in extracting the time-varying or the nonlinear dependence for finite samples is studied in a Monte Carlo exercise. In our empirical study we present the in-sample and out-of-sample performances of our model for a weekly time series of unemployment insurance claims.

1 Introduction

Forecasting and policy analysis in economics and finance are often successfully based on linear dynamic regression models with autoregressive structures. The models are based on parsimonious formulations of temporal dependence and are effective in describing the dynamic salient features of the time series. In many forecasting studies the linear Gaussian autoregressive (AR) model is taken as

the benchmark model and it is found regularly that more elaborate models cannot improve the benchmark forecasting accuracy for different forecast horizons. A favorable aspect of the AR model is that its autoregressive coefficients can be estimated by standard regression methods. The statistical properties of the maximum likelihood estimator are well documented.

Many dynamic features of empirical relevance in fields such as biology, medicine, economics, finance and engineering can however not be appropriately addressed by the linear Gaussian AR model, see the discussions in Teräsvirta, Tjostheim, and Granger (2010). For example, in physics, laws of motions and gravitations are typically nonlinear while in economics and finance economic agents typical interact in a nonlinear way, partly implied by restrictions such as capacity utilization and non-zero unemployment. For this purpose, various nonlinear dynamic models have been proposed in the literature including the threshold AR model of Tong (1983) and the smooth transition AR model of Chan and Tong (1986) and Teräsvirta (1994). A general representation of a nonlinear AR model with additive innovations takes the form

$$y_t = \varphi(y^{t-1}; \boldsymbol{\theta}) + u_t \ , \qquad u_t \sim p_u(\boldsymbol{\theta}), \tag{1}$$

for an observed time series process $\{y_t\}$, with some function φ of the infinite past $y^{t-1} := (y_{t-1}, y_{t-2}, ...)$, parameter vector $\boldsymbol{\theta}$ and a sequence of zero-mean independent innovations $\{u_t\}$ with density $p_u(\boldsymbol{\theta})$. The choice for a function φ is somewhat arbitrary and is often based on convenience and feasibility.

It is easy to show that any process $\{y_t\}$ generated by the nonlinear AR model in (1) also admits the representation,

$$y_t = b_t y_{t-1} + u_t \ , \qquad u_t \sim p_u(\boldsymbol{\theta}) \ , \qquad b_t = \xi(y^{t-1}) \tag{2}$$

where b_t is a random temporal dependence parameter that can be written as a measurable function ξ of the infinite past y^{t-1}. The representations of the data generating process for $\{y_t\}$ in (1) and (2) can be used interchangeably: the time-varying dependence parameter b_t in (2) implies the autoregressive function φ and vice-versa.

More immediate motivations to adopt the latter representation of a linear autoregressive model with a time-varying temporal dependence coefficient can also be provided. For example, the financial crisis that started in 2007 has clearly led to fundamental changes in economic and financial interrelationships. The dynamic structures in time series of economic growth, inflation and interest rates have been affected as a result. To accommodate possible changes in the dynamic properties of economic time series, we can impose time-varying functions for the autoregressive coefficients in the autoregressive model. It implies that dynamic properties of the time series change over time. Earlier contributions in the econometrics literature have considered time-varying parameters in an autoregressive model, most often in the context of vector autoregressive models. Doan, Litterman, and Sims (1984) have been the first to explore the estimation of time-varying coefficients in AR models via the representation of the model in state space form and the application of the Kalman filter. More elaborate Markov chain Monte Carlo methods have been explored by Kadiyala and Karlsson (1993) and Clark and McCracken (2010).

In our study we adopt the AR model with time-varying temporal dependence as a representation of the nonlinear AR model. It enable us to develop *optimal* formulations of nonlinear AR processes. We then compare our resulting nonlinear AR model with the existing threshold and smooth transition AR models in detail. We further establish the information theoretic optimality of the AR parameter updating scheme. In particular, we show that for each parameter update we

reduce the Kullback-Leibler divergence between the true unknown conditional density of the data and the model implied conditional density. We show that only our model with parameter updating based on the score function can have the optimality property. The asymptotic properties of the maximum likelihood estimates are also explored. General conditions for consistency and asymptotic normality of the estimates are documented. We further analyze the finite-sample performance for estimating the time-varying AR coefficient in a Monte Carlo study. Finally, we illustrate the empirical relevance of the model for a time series of growth rates in US unemployment insurance claims. We show that our model outperforms its most direct competitors, in both minimizing in-sample fit and out-of-sample forecasting errors.

The time-varying parameter process is specified as an observation driven model. In particular, we adopt the approach of Creal, Koopman, and Lucas (2013) and Harvey (2013) in which the parameter update is determined by the score function of the predictive loglikelihood function for observation y_t. Blasques, Koopman, and Lucas (2014a) have shown that updating time-varying parameters based on the predictive score function is optimal in an information theoretic sense. In a similar and independent development, but for a different purpose, Delle Monache and Petrella (2014) also propose to use the score function for updating parameters in a linear AR model.

The remainder of the paper is organized as follows. Section 2 introduces the model and represents it in state space form. In Section 3 we formulate the model in reduced form and compare its main features with those of other reduced form models. Section 4 shows that the model is information theoretic optimal, regardless whether the model is correctly or incorrectly specified. In Section 5 we review the stochastic properties of the filter for the time-varying parameter. Section 6 establishes the consistency and asymptotic normality of the maximum

likelihood estimator. Section 7 presents the results of a Monte Carlo exercise that analyses the time-varying parameter estimation in a finite sample setting. In Section 8 we illustrate the model by presenting our empirical analysis. Section 9 concludes.

2 Autoregressive Model with Score Driven Coefficient

Consider the autoregressive process of order one, the AR(1) model, with a time-varying temporal dependence coefficient as given by

$$y_t = h(f_t; \boldsymbol{\theta}) y_{t-1} + u_t, \qquad u_t \sim p_u(u_t; \boldsymbol{\theta}), \tag{3}$$

where $\{y_t\}$ is a time series, $\{f_t\}$ is the time-varying parameter that determines the temporal dependence or the mean-reverting behavior of $\{y_t\}$ via the link function $h(f_t; \boldsymbol{\theta})$, $\{u_t\}$ is the disturbance that is identically and independently distributed with density $p_u(u_t; \boldsymbol{\theta})$, and $\boldsymbol{\theta}$ is a vector of fixed, unknown parameters. We assume that realizations are available for y_t for $t = 1, \ldots, T$. The time-varying parameter f_t will be specified as an observation-driven process which is formally defined by Cox (1981). In effect, the time-varying parameter represents a function of past observations, that is $f_t = f_t(y^{t-1}, f_1; \boldsymbol{\theta})$. The transformation function $h()$ can be used to rule out negative temporal dependence (if $h(f; \cdot) \geq 0 \ \forall \ f$) or explosive behavior (if $-1 \leq h(f; \cdot) \leq 1 \ \forall \ f$) or even unit-root behavior (if $-1 < h(f; \cdot) < 1 \ \forall \ f$) for all t. In any case, this model allows also for the temporal dependence of $\{y_t\}$ and its mean-reverting behavior to change over time. In effect, by letting $h(f_t; \cdot) = 1$, or even $h(f_t; \cdot) > 1$ in some occasions, we can confer $\{y_t\}$ with a 'transient' unit-root or even explosive behavior during specific time periods. In general, these specifications do not rule out the possibility that $\{y_t\}$ is strictly stationary and ergodic (SE). Indeed, following Bougerol (1993),

under appropriate regularity conditions, $\{y_t\}$ is SE as long as $\mathbb{E}|h(f_t;\cdot)| < 1$.

REMARK 1. *All results derived in this paper extend trivially to the autoregressive model with intercept $a \in \mathbb{R}$ as given by $y_t = a + h(f_t;\boldsymbol{\theta})y_{t-1} + u_t$. For simplicity, we set $a = 0$ and treat the case of the de-meaned sequence $\{y_t\}$.*

The AR(1) model with the time-varying temporal dependence coefficient is defined by equation (3). The specification of the AR(1) model relies on the two functions $h()$ and $p_u()$. The motivation for the link function $h()$ is briefly discussed above. Typical examples are the unity function $h(f;\cdot) = f$ and the logistic function $h(f;\cdot) = [1+\exp(-f)]^{-1}$ which will be considered below. Other appropriate link functions $h()$ can be adopted as well. The choice of the density function $p_u()$ is implied by the distribution assumption for the observations. Here the typical examples are the normal and the Student's t densities. A convenient feature of our modelling framework is that no further econometric complexities arise when one departs from the normal density function $p_u()$.

Observation driven models are essentially 'filters' for $\{f_t\}$; they update the parameter f_t using the information provided by the most recent observations of the process $\{y_t\}$. In general, they take the form of

$$f_t = \phi(f_{t-1}, y_{t-1};\boldsymbol{\theta}), \tag{4}$$

for a given initial value f_1. It implicitly follows that $f_t = f_t(y^{t-1}, f_1;\boldsymbol{\theta})$ for the infinite past $y^{t-1} := (y_{t-1}, y_{t-2}, ...)$. Any function $\phi()$ can be considered for the updating of f_t but it is not immediately obvious what the optimal choice is for $\phi()$. The challenge is to find the best way to combine past information in y_{t-1} and f^{t-1} for producing the parameter update f_t. For the same but more general purpose, Creal, Koopman, and Lucas (2013) introduce a class of observation driven models based on a particular choice of $\phi()$ which partly depends on the score function

of the predictive loglikelihood function for observation y_t. Blasques, Koopman, and Lucas (2014a) show that this choice of $\phi()$ is optimal from an information theoretical perspective.

The predictive loglikehood function for y_t in the AR(1) model of equation (3) conditional on y_{t-1} and f_t is given by

$$\ell(y_t|y_{t-1}, f_t; \boldsymbol{\theta}) = \log p_u(\tilde{u}_t; \boldsymbol{\theta}), \tag{5}$$

where $\tilde{u}_t$ is the *prediction error* or *residual*

$$\tilde{u}_t := \tilde{u}_t(y_t, y_{t-1}, f_t; \boldsymbol{\theta}) = y_t - \mathbb{E}_{t-1} y_t = y_t - h(f_t; \boldsymbol{\theta}) y_{t-1}, \tag{6}$$

written as a function of y_t, y_{t-1} and f_t, where $\mathbb{E}_{t-1}$ is the expectation conditional on y^{t-1}. The observation driven update for f_t as proposed by Creal, Koopman, and Lucas (2013) is given by

$$f_t = \omega + \alpha s_{t-1} + \beta f_{t-1}, \tag{7}$$

where ω, α and β are unknown coefficients and

$$s_t = s_t(y_t, y_{t-1}, f_t; \boldsymbol{\theta}) := \nabla_t(y_t, y_{t-1}, f_t; \boldsymbol{\theta}) \,/\, S(f_t), \tag{8}$$

is the scaled score with

$$\nabla_t = \nabla_t(y_t, y_{t-1}, f_t; \boldsymbol{\theta}) := \frac{\partial \log \ell(y_t|y_{t-1}, f_t; \boldsymbol{\theta})}{\partial f_t}, \tag{9}$$

and $S(f_t)$ is some scaling function. For this specification of the function $\phi()$, we have taken $p = q = 1$ in (4) and the parameter vector $\boldsymbol{\theta}$ in (5) includes the coefficients ω, α and β from (7). The update equation (7) formulates a highly

nonlinear function for f_t in terms of the past observations y^{t-1}. The functional form is partly determined by the observation density (3) while the impact of past observations on f_t is mostly determined by the coefficients ω, α and β. Given the updating equation (7) and the role of the score function, this update is designed to take a step in the direction that maximizes the likelihood contribution at time t, given the past information y^{t-1}. It is argued that different scaling functions $S(\cdot)$ can be considered. Here we focus on a purely 'score driven' update for convenience and we set $S(f) = 1 \; \forall \; f$.

2.1 Model I : Affine Gaussian updating

Consider the case where $h(f;\cdot) = f \; \forall \; f$ and where u_t is normally independently distributed with mean zero, that is $p_u(u_t;\boldsymbol{\theta}) = N(0,\sigma^2)$ with $\boldsymbol{\theta} = \sigma^2$. For this case, we have

$$y_t = f_t y_{t-1} + u_t, \qquad u_t \sim N(0,\sigma^2),$$

with the predictive loglikelihood function at time t given by

$$\ell(y_t|y_{t-1},f_t;\boldsymbol{\theta}) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\tilde{u}_t^2,$$

where prediction error $\tilde{u}_t$, as defined in (6), reduces to $\tilde{u}_t = y_t - f_t y_{t-1}$. The score function is given by

$$\nabla_t = \frac{\partial \ell(y_t|y_{t-1},f_t;\boldsymbol{\theta})}{\partial f_t} = \tilde{u}_t \frac{y_{t-1}}{\sigma^2} \tag{10}$$

and hence the update for the dependence parameter is given by

$$f_t = \omega + \alpha \tilde{u}_{t-1}\frac{y_{t-2}}{\sigma^2} + \beta f_{t-1}. \tag{11}$$

The update of f_t responds to the prediction error $\tilde{u}_{t-1}$ which is multiplied by the step size y_{t-2}/σ^2. It follows that the score update brings parameter f_t closer to

zero since $|f_t| < |f_{t-1}|$, when the change in y_{t-1} is reverting to its mean, that is when $\tilde{u}_t < 0$. On the other hand, it moves f_t away from its mean, $|f_t| > |f_{t-1}|$, when the change in y_{t-1} is not reverting to the mean, that is when $\tilde{u}_t > 0$.

We will argue below that the multiplication by y_{t-2} in (11) plays two crucial roles in Model I. First, it signals if the process is below or above its mean. Second, it allows the update to distinguish between those changes in observed data that are driven by the innovations u_t, from those that are driven by the time-varying parameter f_t.

It follows from (11) that the strength of the score in the update is determined both by α and the term y_{t-2}/σ^2. If $\beta = 1$, the score is the only determinant of the parameter update. However, when $0 < \beta < 1$, the parameter will move away from its mean only when we have sufficiently strong evidence that y_{t-1} is not 'mean-reverting'. It implies that, for a given α, a β smaller than unity will make f_t revert to its mean more often, when compared to the case that β equals unity. Also, if σ^2 is large, one expects large variations in y_t only because of the large variability in u_t; it is not because the conditional mean has changed.

2.2 Model II : Logistic updating

The function $h()$ can be used to ensure that the dependence in $\{y_t\}$ satisfies certain properties. In Model II we adopt a logistic function $h()$ that allows for transient unit-root dynamics but rules out negative dependence and explosive behavior as $0 \leq h(f;\cdot) \leq 1$.

Let $h(f;\cdot) = 1/(1+\exp(-f)) \ \forall \ f$ and $p_u(u_t;\boldsymbol{\theta}) = N(0,\sigma^2)$ with $\boldsymbol{\theta} = \sigma^2$. In this case the score is given by

$$\nabla_t = \frac{\partial \ell(y_t|y_{t-1},f_t;\boldsymbol{\theta})}{\partial f_t} = \tilde{u}_t h'(f_t;\boldsymbol{\theta})\, y_{t-1} \,/\, \sigma^2, \tag{12}$$

where

$$h'(f;\boldsymbol{\theta}) := \frac{\partial h(f;\boldsymbol{\theta})}{\partial f},$$

which in this case reduces to $h'(f;\cdot) = h(f;\cdot)^2 \exp(-f)$. Hence the parameter update becomes

$$\begin{aligned} f_t &= \omega + \alpha \tilde{u}_{t-1} h(f_{t-1};\cdot)^2 \exp(-f_{t-1}) y_{t-2} \,/\, \sigma^2 + \beta f_{t-1} \\ &= \omega + \alpha\Big(y_{t-1} - \frac{1}{1+\exp(-f_{t-1})} y_{t-2}\Big) \frac{\exp(-f_{t-1})}{(1+\exp(-f_{t-1}))^2} (y_{t-2} \,/\, \sigma^2) + \beta f_{t-1}. \end{aligned} \tag{13}$$

The resulting updating function for f_t restricts the autoregressive coefficient to satisfy $0 \leq h(f_t;\cdot) \leq 1$ which sometimes can be motivated by economic theory. In other cases, the econometrician simply wants to bound the possible influence of outlying observations on the actual value of f_t and its future counterparts.

2.3 Model III : Robust updating

Robustness to outliers can be obtained via the specification of $h()$ as above or by explicitly recognizing that the density function of u_t requires fat tails. The dependence of the update equation for f_t on both $h(f_t;\boldsymbol{\theta})$ and $p_u(u_t;\boldsymbol{\theta})$ is a convenient feature of our score driven approach. Let $h(f;\cdot) = f \;\forall\; f$ and $p_u(u_t;\boldsymbol{\theta}) = \tau(0,1,\lambda)$ which is the density function of the standardized Students' t distribution with zero mean, unity variance and degrees of freedom λ. We have $\boldsymbol{\theta} = \lambda$. The predictive loglikelihood function at time t is given by

$$\ell(y_t|y_{t-1}, f_t;\boldsymbol{\theta}) = c - \frac{\lambda+1}{2} \log\left(1 + \frac{\tilde{u}_t^2}{\lambda}\right),$$

Three Nonlinear Dynamic Model Specifications

Model		Equations	$h(f_t)$	s_t	$p_u(u_t;\boldsymbol{\theta})$
I	Affine Gaussian	(3) + (11)	f_t	(10)	$N(0,\sigma^2)$
II	Logistic Gaussian	(3) + (13)	$1\,/\,(1+\exp(-f_t))$	(12)	$N(0,\sigma^2)$
III	Affine Student's t	(3) + (15)	f_t	(14)	Student's $t(\lambda)$

Table 1: Model is $y_t = h(f_t)y_{t-1}+u_t$ and $f_t = \omega+\alpha s_{t-1}+\beta f_{t-1}$ with $u_t \sim p_u(u_t;\boldsymbol{\theta})$.

where c is a constant that does not depend on y_t nor on f_t. In this case, the score function becomes

$$\nabla_t = \frac{\partial \ell(y_t|y_{t-1},f_t)}{\partial f_t} = (\lambda+1)\tilde{u}_t \frac{y_{t-1}}{\lambda+\tilde{u}_t^2}. \tag{14}$$

The updating function is then given by

$$f_t = \omega + \alpha(\lambda+1)\frac{(y_{t-1}-f_t y_{t-2})y_{t-2}}{\lambda+(y_{t-1}-f_t y_{t-2})^2} + \beta f_{t-1}. \tag{15}$$

The updating function for f_t reveals that it is now less sensitive to large variations in observed data compared to the updating for the affine Gaussian case. In particular, our robust updating is a bounded function of $\tilde{u}_{t-1}$. The intuition follows straightforwardly. When the innovations u_t are coming from a fat tailed density, large variations in observed data are relatively more likely to correspond to draws from the tail of $p_u(u_t;\boldsymbol{\theta})$, rather than from changes in the conditional expectation. On the contrary, when the innovations in f_t are bounded by a small variance, large variations in y_t are likely to correspond to changes in the conditional expectation.

2.4 Assessment of Three Model Specifications

Table 1 reviews our main three nonlinear dynamic model specifications. Figure 1 compares the different updating functions for f_t in their responses to variations in the observation y_t when the innovations are generated by Gaussian and Student's t distributions. Figure 1 reveals several interesting features of the updating functions. First, it shows that the update tends to decrease f_t when y_{t-1} shows mean-reverting behavior and increase f_t otherwise. Second, it shows that parameter updates with $\beta = 0.5$ tend to bring f_t faster to its unconditional mean, here it is zero as $\omega = 0$ when compared to the updates with $\beta = 1$. In the former case, the score information dominates the update. Third, it shows how the specified distribution for u_t is incorporated in the parameter updating step. In particular, it shows that for a fat tailed distribution, large variations in the observed data do not necessarily reflect a strong change in the conditional expectation. As a result, f_t is bounded in y_{t-1}. This is in sharp contrast with the linear response shown for the affine Gaussian case.

Figure 1 further reveals how the updating function uses the value y_{t-2} as a crucial guidance mechanism to distinguish between changes in observed data that provide information about the conditional expectation and those that do not. For example, consider the case where observed data is very close to its mean (left graph). Then, there is no reason to strongly update the conditional expectation, regardless of y_{t-1} being large or small; the observation y_{t-1} does not contain much information about the dependence of the process $\{y_t\}$. Indeed, in the case where $y_{t-2} = 0$, then y_{t-1} contains no information about the dependence of the process because the mean-reverting force is simply inactive. Any value of y_{t-1} would have to reflect only the innovation draw. On the contrary, in the case where y_{t-2} is large (in absolute value), the observed y_{t-1} will carry more information about f_t. In particular, consider the case where $y_{t-2} = 4$ (right graph). Then, if y_{t-1} is

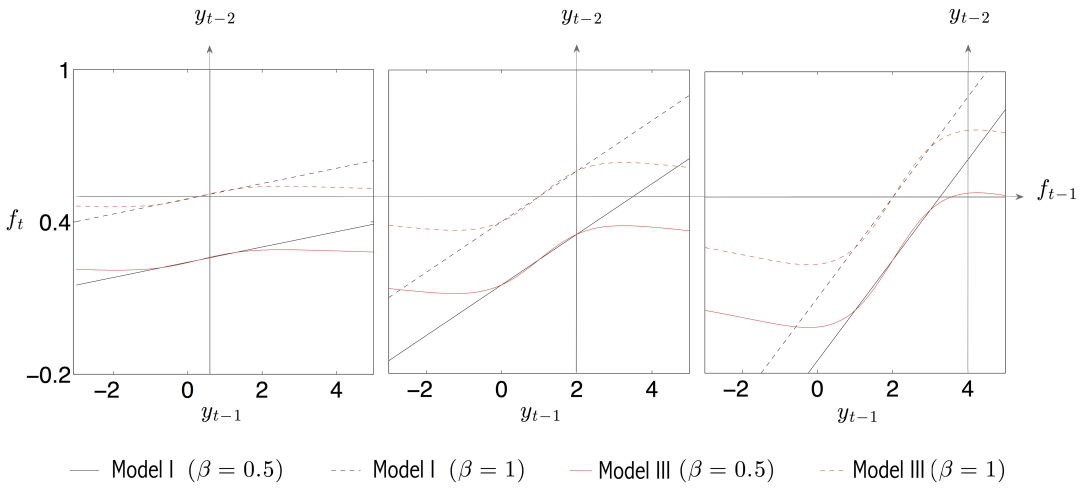

Figure 1: Shape of Normal (black) and Student's t (red) updating functions. The updated parameter $f_t = \phi(f_{t-1}, y_{t-1}, y_{t-2}; \boldsymbol{\theta})$ is plotted as a function of y_{t-1} for given $f_{t-1} = 0.5$ and given low initial state $y_{t-2} = 0.5$ (left) high initial state $y_{t-2} = 2$ (middle) and very high initial state $y_{t-2} = 4$ (right). All plots obtained with $\omega = 0$ and $\alpha = 0.1$. Solid lines have $\beta = 0.5$ and dashed lines have $\beta = 1$.

also large, these observations provide strong evidence that the process has strong dependence and hence that f_t is close to one. Only the unlikely event of two consecutive draws from the right tail of u_t could potentially indicate a low value of f_{t-1}.

2.5 Maximum likelihood estimation and forecasting

Maximum likelihood (ML) estimation of parameters in the AR(1) model with a time-varying autoregressive coefficient is similar as for autoregressive moving average models. The conditional likelihood function can be evaluated in closed-form given that we have both an updating equation for f_t and an expression for the score function s_t explicitly available. The conditional loglikelihood function is then simply obtained via the prediction error decomposition and is given by

$$\ell_T(\boldsymbol{\theta}, f_1) = \sum_{t=2}^{T} \ell(y_t | y_{t-1}, f_t; \boldsymbol{\theta}).$$

The maximization of this loglikelihood function with respect to $\boldsymbol{\theta}$ is typically carried out using a quasi-Newton optimization method. The prediction errors $\tilde{u}_t$ evaluated at the maximum likelihood estimate of $\boldsymbol{\theta}$ can be used for diagnostic checking procedures.

The forecasting of future values of y_{T+j} and f_{T+j} for $j = 1, 2, \ldots$, can be obtained as follows. The forecast for y_{T+1} is based on (3) with f_{T+1} computed by (7) given a value for y_T. The other forecasts for $j = 2, 3, \ldots$ are obtained in the same way where future y_t's are replaced by their forecasts. It follows that all future score values, that is $s_{T+1}, s_{T+2}, \ldots$ in (8), can be set to zero.

3 Nonlinear Autoregressive Representations

The nonlinear autoregressive (AR) model can be generally expressed as in (1), that is $y_t = \varphi(y^{t-1}; \boldsymbol{\theta}) + u_t$. The relation between a nonlinear AR model and a linear AR(1) model with a time-varying autoregressive coefficient is established straightforwardly. The nonlinear AR model (1) can be expressed as $y_t = h(f_t; \boldsymbol{\theta}) y_{t-1} + u_t$, given the equivalence $h(f_t; \boldsymbol{\theta}) \equiv \varphi(y^{t-1}; \boldsymbol{\theta}) \,/\, y_{t-1}$ since $f_t = f_t(y^{t-1}; \boldsymbol{\theta})$. This specification is well defined a.s. because y_{t-1} is present in both the nominator and denominator. The converse is also true: for all distributions considered for u_t, the dynamic model (3)-(4) can be expressed as a nonlinear autoregressive moving average (ARMA) model. We discuss this in more detail and provide illustrations below.

The AR(1) model (3) implies that

$$h(f_t; \boldsymbol{\theta}) = \frac{y_t - u_t}{y_{t-1}}.$$

In the case of an unity function for $h()$, we have $f_t = (y_t - u_t) y_{t-1}^{-1}$ and the

score-driven updating function becomes

$$f_t = \omega + \alpha s_{t-1}(y_{t-1}, y_{t-2}, \frac{y_{t-1} - u_{t-1}}{y_{t-2}}; \boldsymbol{\theta}) + \beta \frac{y_{t-1} - u_{t-1}}{y_{t-2}}.$$

By substituting this expression into (3), that is $y_t = f_t\, y_{t-1} + u_t$, we obtain

$$y_t = \omega y_{t-1} + \alpha s_{t-1}\Big(y_{t-1}, y_{t-2}, \frac{y_{t-1} - u_{t-1}}{y_{t-2}}; \boldsymbol{\theta}\Big) y_{t-1} + \beta \frac{y_{t-1} - u_{t-1}}{y_{t-2}} y_{t-1} + u_t,$$

which we can recognize as an ARMA model with two lags for the dependent variable and one lag for the innovation u_t, that is a nonlinear ARMA$(2, 1)$ model.

3.1 Illustration: Models I and III

In the case of the affine Gaussian AR(1) model in Section 2.1, we have $s_t = u_t\, y_{t-1} \,/\, \sigma^2$ and the nonlinear ARMA$(2, 1)$ model becomes

$$y_t = \omega y_{t-1} + \alpha \frac{y_{t-1} y_{t-2} u_{t-1}}{\sigma^2} + \beta \frac{y_{t-1} - u_{t-1}}{y_{t-2}} y_{t-1} + u_t.$$

In the case of the Student's t AR(1) model with λ as the degrees of freedom in Section 2.3, we have $s_t = (\lambda + 1)\, u_t\, y_{t-1} \,/\, (\lambda + u_t^2)$ so that the nonlinear ARMA representation for y_t is given by

$$y_t = \omega y_{t-1} + \alpha(\lambda + 1) \frac{y_{t-1} y_{t-2} u_{t-1}}{\lambda + u_{t-1}^2} + \beta \frac{y_{t-1} - u_{t-1}}{y_{t-2}} y_{t-1} + u_t.$$

It is interesting that these extensive nonlinear ARMA model representations originate from a basic linear AR(1) model with a time-varying autoregressive coefficient based on the observation-driven process $f_t = \omega + \alpha s_{t-1} + \beta f_{t-1}$. While the original model is relatively simple, it implies an extensive but parsimonious nonlinear ARMA model. We emphasize that we do not base our analysis on

the nonlinear ARMA framework: we analyze the time series by means of the time-varying parameter AR(1) model representation as discussed in Section 2.5.

3.2 Illustration: Model II

The illustrations above are based on the unity function for $h(f) = f$. The Gaussian AR(1) model with a logistic function for the time-varying autoregression coefficient can also be represented as a nonlinear ARMA$(2, 1)$ model. In case of Model II in Section 2.2, we have $y_t = h(f_t)y_{t-1}+u_t$ where $h(f_t) = [1+\exp(-f_t)]^{-1}$ and $p_u(u_t)$ is a Gaussian density. The corresponding score function is given by

$$s_t = h'(f_t)\,\frac{y_{t-1}\,u_t}{\sigma^2} \;=\; h(f_t)[1-h(f_t)]\,\frac{y_{t-1}\,u_t}{\sigma^2} \;=\; \frac{(y_t-u_t)\,(u_t-\Delta y_t)\,u_t}{\sigma^2\,y_{t-1}},$$

with $\Delta y_t = y_t - y_{t-1}$, since $h(f_t) = (y_t - u_t)\,/\,y_{t-1}$ and $h'(f_t) = h(f_t)^2\exp(-f_t)$. The updating equation is

$$f_t = \omega + \alpha s_{t-1} + \beta \log\Big(\frac{y_{t-1}-u_{t-1}}{u_{t-1}-\Delta y_{t-1}}\Big), \tag{16}$$

since $f_t = -\log[h(f_t)^{-1}-1]$ and $h(f_t) = (y_t-u_t)\,/\,y_{t-1}$. After some minor algebra we obtain the nonlinear ARMA model representation as

$$\begin{aligned} y_t &= h(f_t)y_{t-1} + u_t \\ &= \left[1+\exp(-\omega-\alpha s_{t-1})\Big(\frac{u_{t-1}-\Delta y_{t-1}}{y_{t-1}-u_{t-1}}\Big)^{\beta}\right]^{-1} y_{t-1} + u_t. \end{aligned}$$

The intuition behind these nonlinear ARMA representations are not so clear but we have shown that our original modeling framework leads effectively to a nonlinear ARMA model. We therefore compare our model with some other well-known nonlinear dynamic models next.

3.3 Comparison with other nonlinear models

Two well-known nonlinear AR models are the threshold AR (TAR) model of Tong (1983) and the smooth transition AR (STAR) model of Chan and Tong (1986) and Teräsvirta (1994). We relate our nonlinear dynamic models with the basic versions of these two nonlinear AR models. In the next section we show that our modeling framework has favorable optimality properties. Such proofs of optimality are not available for other models such as the TAR and STAR.

When written as a nonlinear autoregressive model, the TAR model takes the form

$$y_t = \gamma_1 y_{t-1} + \gamma_2 \, \mathrm{I}(y_{t-2} < \gamma_3) y_{t-1} + u_t,$$

where I() is an indicator function that returns unity if the condition in the argument holds, and otherwise zero. The TAR model can be expressed and generalized in various ways. The STAR model can be specified as

$$y_t = \frac{\gamma_4 y_{t-1}}{1 + \exp(-\gamma_6 \, y_{t-2})} - \frac{\exp(-\gamma_6 \, y_{t-2}) \gamma_5 y_{t-1}}{1 + \exp(-\gamma_6 \, y_{t-2})} + u_t.$$

Both TAR and STAR models are effectively nonlinear ARMA$(2, 0)$ models with four parameters when we assume that $u_t \sim N(0, \sigma^2)$. These models have the same number of parameters as our Models I and II of Sections 2.1 and 2.2, respectively.

Naturally, the TAR and STAR can also be represented as a simple linear AR(1) model with a time-varying parameter. The TAR model represents a class of models that let the coefficient change when a certain value crosses a certain benchmark.

$$y_t = \rho_t \, y_{t-1} + u_t \ , \qquad \rho_t = \gamma_1 + \gamma_2 \, \mathrm{I}(y_{t-2} < \gamma_3).$$

The STAR model admits the time-varying parameter representation

$$y_t = \rho_t \, y_{t-1} + u_t \ , \qquad \rho_t = \gamma_4 x_{t-2} + \gamma_5 (1 - x_{t-2}), \qquad x_t := \frac{1}{1 + \exp(-\gamma_6 \, y_t)}.$$

In Figure 2 we present the response to y_t for different values of y_{t-1} and y_{t-2}, for different parameter settings in case of the TAR and STAR models, and for different ranges of values for u_{t-1} and for $\beta = 0.5, 1$ in case of Model II. Figure 2 compares the shapes of the nonlinear ARMA representation of the Model II of Section 2 for different parameter settings with those of the TAR and STAR models. Although the latter two models are nonlinear AR(2) models and Model II is a nonlinear ARMA(2,1) model, the nonlinear functions are quite similar in many respects and over a range of values for u_{t-1}.

In all cases, there are two regimes that are clearly identified. One regime with large slope over the y_{t-1} axis, which occurs for positive values of y_{t-2}, and another regime with small slope over the y_{t-1} axis that occurs for negative values of y_{t-2}. In both the TAR and STAR models these regimes are linear in y_{t-1}, and hence, in each regime, the slope is constant over y_{t-1}. The cross-section over the y_{t-2} axis shows however the difference between the TAR and STAR models. Namely, in the TAR case, the transition from one regime to the other is discontinuous, whereas in the STAR case it is smooth. The response behaviour of Model II is similar to the TAR given that the transition is discontinuous. It is similar to the STAR given that the responses to the range of y_{t-2} values is nonlinear in each regime. The responses of Model II are nonetheless unique given that the two regimes are nonlinear in y_{t-1}. In particular, the low regimes shows an increasing slope in y_{t-1}, while the high regime shows a decreasing slope in y_{t-1}. Most importantly, and in contrast to the TAR and STAR models, Model II is the result of a score driven time-varying parameter model which possesses a number of optimality properties described in the next section.

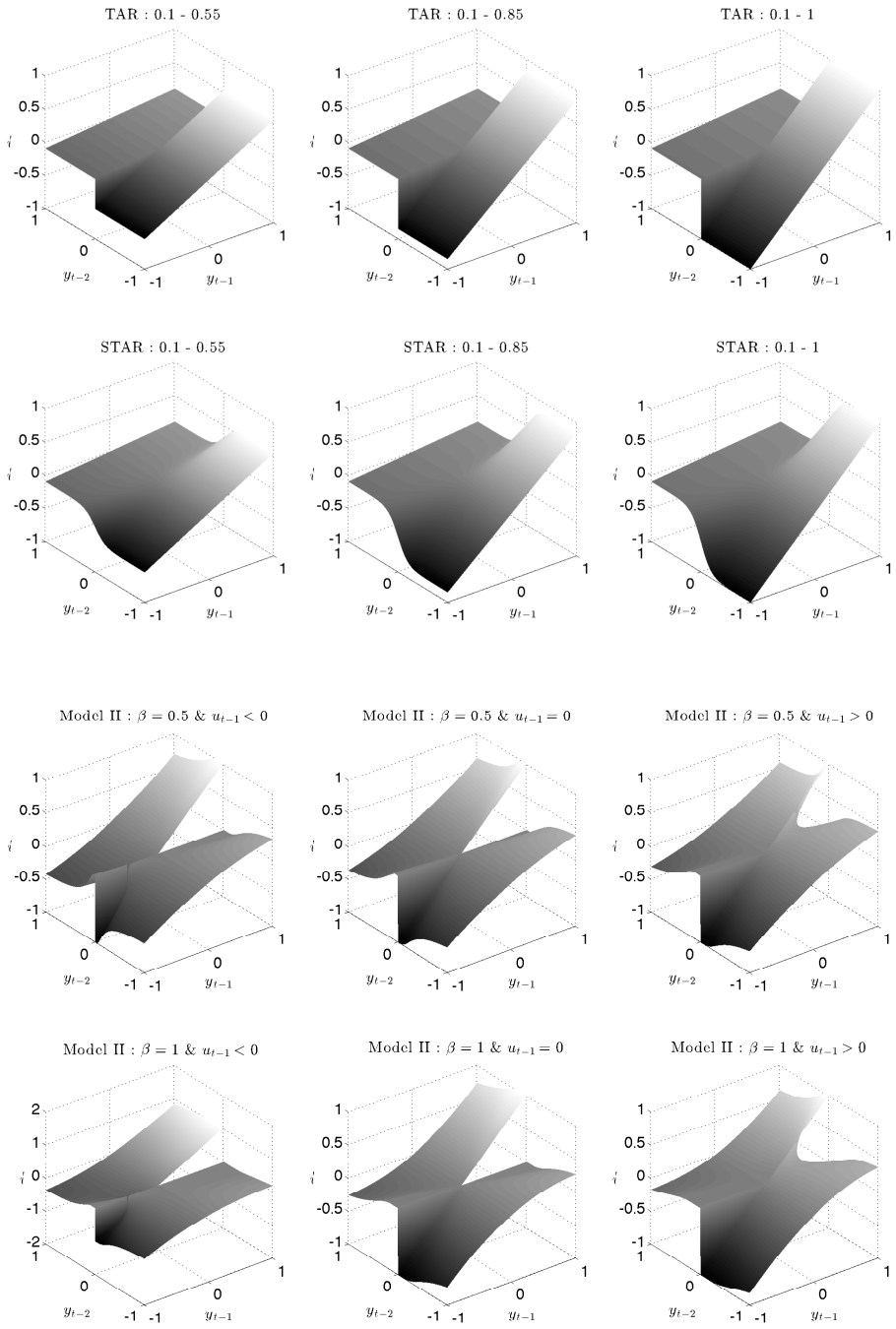

Figure 2: Response functions for TAR and STAR models (top 2 rows) are presented for different slopes in each regime. The response functions for Model II (bottom 2 rows) are presented for different values of β and innovations u_{t-1}.

4 Optimality of Autoregressive Parameter Update

The score driven update for the time-varying parameter f_t is not only intuitively appealing, but it can also be shown to be optimal. This section provides a simple extension of the results in Blasques, Koopman, and Lucas (2014a) to the context of Markov autoregressive models with time-varying dependence parameters. In particular, it shows that that the update for f_t is the optimal observation driven parameter update, in an information theoretic sense, as it reduces locally the Kullback-Leibler (KL) divergence between the true conditional density and the conditional density implied by the model. Just as in Blasques, Koopman, and Lucas (2014a), we also show that only the score update can possess these properties. These results can be obtained whether the model is correctly or incorrectly specified.

The KL divergence is an important measure of distance in various fields from information theory to statistics and econometrics; see, for example, Ullah (1996, 2002) for several applications of the KL divergence in econometrics. In the developments below, we have $p_t := p(\cdot|f_t, y_{t-1})$ as the true conditional density of y_t indexed by the true time-varying parameter f_t and the lagged observation y_{t-1}, and we define $\tilde{p}_t := \tilde{p}(\cdot|\tilde{f}_t, y_{t-1})$ as the postulated density indexed by the filtered parameter $\tilde{f}_t$ and y_{t-1}.

Definition 1. (RKL Optimality) *The realized KL (RKL) variation of a parameter update from $\tilde{f}_t \in \tilde{\mathcal{F}}$ to $\tilde{f}_{t+1} \in \tilde{\mathcal{F}}$ is given by*

$$\begin{aligned}\Delta_{t|t} &= \mathcal{D}_{\mathrm{KL}}\Big(p_t, \tilde{p}_{t+1}\Big) - \mathcal{D}_{\mathrm{KL}}\Big(p_t, \tilde{p}_t\Big) \\ &= \int_Y p(y|f_t, y_{t-1}) \ \Big(\ln \tilde{p}(y|\tilde{f}_t, y_{t-1}; \boldsymbol{\theta}) - \ln \tilde{p}(y|\tilde{f}_{t+1}, y_{t-1})\Big) \ \mathrm{d}y.\end{aligned}$$

where $\mathcal{D}_{\mathrm{KL}}\Big(p_t, \tilde{p}_t\Big)$ denotes the KL divergence between p_t and $\tilde{p}_t$. For a given p_t, a parameter update is RKL optimal given y_{t-1} if and only if $\Delta_{t|t} < 0$.

Definition 2. (CKL Optimality) *Define the conditional expected KL (CKL) variation of a parameter update from* $\tilde{f}_t \in \tilde{\mathcal{F}}$ *to a random* $\tilde{f}_{t+1}$ *taking values in* $\tilde{\mathcal{F}}$ *as*

$$\Delta_{t|t-1} = \int_F q(\tilde{f}_{t+1}|\tilde{f}_t, f_t, y_{t-1}; \boldsymbol{\theta}) \left[\int_Y p(y|f_t, y_{t-1}) \ln \frac{\tilde{p}(y|\tilde{f}_t, y_{t-1}; \boldsymbol{\theta})}{\tilde{p}(y|\tilde{f}_{t+1}, y_{t-1}; \boldsymbol{\theta})} \mathrm{d}y \right] d\tilde{f}_{t+1},$$

with $q(\tilde{f}_{t+1}|\tilde{f}_t, f_t; \boldsymbol{\theta})$ *denoting the density of* $\tilde{f}_{t+1}$ *conditional on both* $\tilde{f}_t$ *and* f_t*. For a given* p_t*, an update is CKL optimal given* y_{t-1} *if and only if* $\Delta_{t|t-1} < 0$*.*

The RKL optimality measures the change in the KL divergence between the true conditional density $p(\cdot|f_t, y_{t-1})$ and the model's conditional densities $\tilde{p}(\cdot|\tilde{f}_t, y_{t-1})$ and $\tilde{p}(\cdot|\tilde{f}_{t+1}, y_{t-1})$ for given points $\tilde{f}_t \in \tilde{\mathcal{F}}$ and $\tilde{f}_{t+1} \in \tilde{\mathcal{F}}$ since it is conditional on information until time t. On the contrary, $\tilde{f}_{t+1}$ is random in the CKL optimality because this measure is conditional on the information until time $t-1$. Hence, CKL optimality measures the *expected* change in the KL divergence between the true conditional density $p(\cdot|f_t, y_{t-1})$ and the model's conditional density $\tilde{p}(\cdot|\tilde{f}_t, y_{t-1})$ and the random density $\tilde{p}(\cdot|\tilde{f}_{t+1}, y_{t-1})$.

Definition 3. *A nonlinear autoregressive model as in (1) is said to be RKL (CKL) optimal if it admits a time-varying parameter representation (2) with RKL (CKL) optimal parameter update.*

Next we show that the score update for f_t in (7) is locally optimal for any p_t. Local results focus on the 'direction' of the updating step. Intuitively, an update is locally KLV optimal if the updating step is in the correct direction, that is, if the update is in a direction that reduces the KL divergence. By 'local' we mean that the results hold for every $\tilde{f}_{t+1}$ in a neighborhood of $\tilde{f}_t \in \tilde{\mathcal{F}}$ and every y in a neighborhood of $y_t \in \mathcal{Y}$. In other words we show that $\exists \delta_f > 0 \wedge \delta_y > 0$ such that

$\sup_{p_t} \Delta_{t-1|t-1} < 0$ and $\sup_{p_t} \Delta_{t|t-1} < 0$ hold on the sets

$$F = F_{\delta_f}(\tilde{f}_t) := \{\tilde{f} \in \tilde{\mathcal{F}} : |\tilde{f} - \tilde{f}_t| < \delta_f\}, \quad Y = Y_{\delta_y}(y_t) := \{y \in \mathcal{Y} : |y - y_t| < \delta_y\}.$$

ASSUMPTION 1. *$p(y|f, y') > 0 \ \forall \ (y, y', f) \in \mathbb{R} \times \mathbb{R} \times \mathcal{F}$ and $\tilde{\nabla}(\tilde{f}, y, y', ; \boldsymbol{\theta}) \neq 0$ for every $(\tilde{f}, \boldsymbol{\theta}) \in \tilde{\mathcal{F}} \times \Theta$ and almost every $(y, y') \in \mathbb{R} \times \mathbb{R}$.*

ASSUMPTION 2. *$\alpha > 0$ and $S(\tilde{f}, y; \boldsymbol{\theta}) > 0 \ \forall \ (\tilde{f}, y, \boldsymbol{\theta}) \in \tilde{\mathcal{F}} \times \mathbb{R} \times \Theta$.*

The proofs of Lemmas 1 and 2 below can be easily obtained by extending the proofs of Propositions 1-5 in Blasques, Koopman, and Lucas (2014a) so as to allow the lagged y_{t-1} to enter the conditioning set in both p and $\tilde{p}$. For this reason, the proofs of the two Lemmas below are deferred to the main appendix. Lemma 1 below shows that the score update for f_t is locally optimal.

LEMMA 1. *Let Assumptions 1 and 2 hold and let $(\omega, \beta) = (0, 1)$. Then, the score update for f_t is locally RKL optimal and CKL optimal given y_{t-1} for any p_t.*

Lemma 2 shows that only a 'score-equivalent' update can have this optimality property. An update is said to be 'score-equivalent' if it essentially 'mimics' the score update for f_t in (7) locally.

DEFINITION 4. (Score-Equivalent Update) *An observation driven parameter update $\tilde{f}_{t+1} = \phi(\tilde{f}_t, y_t, y_{t-1})$ is said to be 'score-equivalent' if and only if the condition $\text{sign}(\Delta\phi(f, y, y'; \boldsymbol{\theta})) = \text{sign}(\tilde{\nabla}(f, y, y'; \boldsymbol{\theta}))$ holds for almost every $(y, y', f) \in \mathcal{Y} \times \mathcal{Y} \times \mathcal{F}$ and every $\boldsymbol{\theta}$.*

LEMMA 2. *Let Assumptions 1 and 2 hold. For any given p_t, a parameter update is locally RKL optimal and CKL optimal given given y_{t-1} if and only if the parameter update is score-equivalent.*

These properties of the score update hold also when we allow for $(\omega, \beta) \neq (0, 1)$ as long as the "forces away" from the optimal direction at $\tilde{f}_t$, which are determined by the autoregressive component $\omega + (\beta - 1)\tilde{f}_t$, are weaker than the "forces towards" the optimal direction, which are determined by the score component $\alpha s(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta})$; see the main appendix for such extensions.

5 Stochastic Properties of the Filter

Next we obtain the strict stationarity and ergodicity (SE) of the filter for the time-varying parameter. The SE properties are required for deriving the asymptotic properties of the ML estimator in Section 6. The proofs of the Theorems are presented in the main appendix.

For notational simplicity, we define the score update as

$$f_{t+1} := \phi(f_t, y_t, y_{t-1}; \boldsymbol{\theta}) := \omega + \alpha s(f_t, y_t, y_{t-1}; \boldsymbol{\theta}) + \beta f_t$$

and we define the random supremum

$$\bar{\phi}_{t,k}(\boldsymbol{\theta}) := \sup_{f \in \mathcal{F}} \left| \alpha \frac{\partial s(f, y_t, y_{t-1}; \boldsymbol{\theta})}{\partial f} + \beta \right|^k.$$

When convenient, we shall also state explicitly the dependence of the filtered parameter f_{t+1} on the initialization $f_1 \in \mathcal{F}$, the data $y^{1:t} = \{y_s\}_{s=1}^t$ and the parameter vector $\boldsymbol{\theta} \in \Theta$. The score update equation is then expressed as

$$f_{t+1}(y^{1:t}, \boldsymbol{\theta}, f_1) = \omega + \alpha s\Big(f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1), y_t, y_{t-1}; \boldsymbol{\theta}\Big) + \beta f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1),$$

$\forall\ t \in \mathbb{N}$. Theorem 1 states sufficient conditions for the stochastic sequence $\{f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1)\}_{t \in \mathbb{N}}$ initialized at $f_1 \in \mathbb{R}$ to converge almost surely, uniformly on

$\mathcal{F}$, and exponentially fast to stationary and ergodic (SE) sequence $\{f_t(y^{t-1}, \boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ that has n_f bounded moments, where $y^{t-1} := \{y_s\}_{s=-\infty}^{s=t-1}$. It establishes the convergence to an SE limit of the sequence $\{f_t(y^{1:t-1}, \cdot, f_1)\}_{t\in\mathbb{N}}$ with random elements taking values in the Banach space $(\mathbb{C}(\Theta, \mathcal{F}), \|\cdot\|^{\Theta})$ for every $t \in \mathbb{N}$, where $\|\cdot\|^{\Theta}$ denotes the supremum norm on Θ.

THEOREM 1. *Let $\mathcal{F}$ be convex, Θ be compact, $\{y_t\}_{t\in\mathbb{Z}}$ be SE, $s \in \mathbb{C}(\mathcal{F} \times \mathcal{Y}^2 \times \Theta)$ and assume there exists a non-random $f_1 \in \mathcal{F}$ such that*

(i) $\mathbb{E} \ln^+ \sup_{\boldsymbol{\theta}\in\Theta} |s(f_1, y_t, y_{t-1}; \boldsymbol{\theta})| < \infty$*; and*

(ii) $\mathbb{E} \ln \sup_{\boldsymbol{\theta}\in\Theta} \bar{\phi}'_{1,1}(\boldsymbol{\theta}) < 0$.

Then $\{f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1)\}_{t\in\mathbb{N}}$ converges exponentially almost surely to the limit SE process $\{f_t(y^{t-1}, \boldsymbol{\theta})\}_{t\in\mathbb{Z}}$; i.e. we have $\sup_{\boldsymbol{\theta}\in\Theta} |f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1) - f_t(y^{t-1}, \boldsymbol{\theta})| \overset{e.a.s.}{\to} 0$ as $t \to \infty$. If furthermore $\exists\ n_f \geq 1$ such that

(iii) $\mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} |s(f_1, y_t, y_{t-1}; \boldsymbol{\theta})|^{n_f} < \infty$*; and either*

(iv) $\sup_{\boldsymbol{\theta}\in\Theta} |s(f, \mathbf{y}, f; \boldsymbol{\theta}) - s(f', \mathbf{y}, f; \boldsymbol{\theta})| < |f - f'| \ \forall\ (f, f', \mathbf{y}) \in \mathcal{F} \times \mathcal{F} \times \mathcal{Y}^2$*; or*

(v) $\mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} \bar{\phi}'_{1,n_f}(\boldsymbol{\theta}) < 1$ *and* $f_t(y^{t-1}, \boldsymbol{\theta}, f_1) \perp \bar{\phi}'_{t+1,n_f}(\boldsymbol{\theta})\ \forall\ (t, f_1) \in \mathbb{N} \times \mathcal{F}$.

Then both $\{f_t(y^{t-1}, \boldsymbol{\theta}, f_1)\}_{t\in\mathbb{N}}$ and the limit SE process $\{f_t(y^{t-1}, \boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ have n_f bounded moments; i.e. $\sup_t \mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} |f_t(y^{t-1}, \boldsymbol{\theta}, f_1)|^{n_f} < \infty$, and furthermore, $\mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} |f_t(y^{t-1}, \boldsymbol{\theta})|^{n_f} < \infty$.

Verification of these conditions is typically straightforward as the following illustration for Model II of Section 2.2 shows. The updating equation for f_t is given by (13). If $\{y_t\}_{t\in\mathbb{Z}}$ is SE and satisfies $\mathbb{E}|y|^{n_y} < \infty$, then the SE condition

reduces to

$$\mathbb{E}\ln\Big|\alpha\frac{y_{t-1}y_{t-2}}{\sigma^2}\frac{\exp(2f_{t-1})-\exp(f_{t-1})}{(1+\exp(f_t))^3}-\alpha\frac{y_{t-2}^2}{\sigma^2}\frac{\exp(2f_{t-1})(\exp(f_{t-1})-2)}{(\exp(f_{t-1})+1)^4}+\beta\Big|<0. \tag{17}$$

It follows that $\{f_t\}$ is SE with $\mathbb{E}|f_t|^2<\infty$ for every triplet (α,β,σ^2) satisfying (17). A subset of this region that is analytically tractable is given by the diamond-shaped set,

$$\frac{c}{\sigma^2}\mathbb{E}|y_{t-2}^2|<\frac{1-|\beta|}{|\alpha|}\quad\text{ where }\quad c=\frac{39+55\sqrt{33}}{4608}\approx 0.077. \tag{18}$$

6 Asymptotic Properties of Maximum Likelihood

The observation driven structure of our model allows for a simple implementation of a maximum likelihood estimation procedure. The ML estimator of the updating parameters is defined as an element of the arg min set of the sample loglikelihood function $\ell_T(\boldsymbol{\theta},f_1)$,

$$\hat{\boldsymbol{\theta}}_T\in\arg\min_{\boldsymbol{\theta}\in\Theta}\ell_T(\boldsymbol{\theta},f_1)$$

where

$$\ell_T(\boldsymbol{\theta},f_1)=\frac{1}{T}\sum_{t=2}^{T}\ell_t(\boldsymbol{\theta},f_1)=\frac{1}{T}\sum_{t=2}^{T}\log p_u\Big(y_t-h\Big(f_t(y^{t-1};\boldsymbol{\theta})\Big)y_{t-1};\boldsymbol{\theta}\Big).$$

The results of Section 5 can now be used to establish the existence, consistency and asymptotic normality of the ML estimators of the updating parameters. In what follows $(\Omega,\mathcal{F},\mathbb{P})$ denotes the underlying complete probability space. Observed data $\{y_t\}_{t=1}^T$ is thus a subset of the realized path of a real-valued stochastic process $\mathbf{y}:\Omega\to\mathbb{R}^\infty$ where $\mathbb{R}^\infty:=\times_{t\in\mathbb{Z}}\mathbb{R}$ denotes the infinite Cartesian product of copies of $\mathbb{R}$.

ASSUMPTION 3. $(\Theta, \mathfrak{B}(\Theta))$ *is a measurable space and* Θ *is a compact set. Furthermore both* $h : \mathbb{R} \to \mathbb{R}$ *and* $p_u : \mathbb{R} \times \Theta \to \mathbb{R}$ *are continuously differentiable in their arguments.*

Theorem 1 below establishes the existence and measurability of the ML estimator. Essentially, it ensures that there exists a random variable that lies in the $\arg\max$ set of $L_T(\cdot, f_1)$ for every f_1.

THEOREM 2. (Existence) *Let Assumption 3 hold. Then there exists a.s. an* $\mathbb{R}/\mathfrak{B}(\Theta)$*-measurable map* $\hat{\boldsymbol{\theta}}_T : \Omega \times \mathbb{R} \to \Theta$ *satisfying* $\hat{\boldsymbol{\theta}}_T(f_1) \in \arg\max_{\boldsymbol{\theta} \in \Theta} \ell_T(\boldsymbol{\theta}, f_1)$ *for all* $T \in \mathbb{N}$ *and every initialization* $f_1 \in \mathbb{R}$.

In order to establish consistency, we now impose enough conditions to ensure that the likelihood function satisfies a uniform law of large numbers for SE processes. Assumption 4 establishes conditions that make the arguments of the likelihood function SE.

ASSUMPTION 4. $\exists\ (n_f, f) \in [1, \infty) \times \mathbb{R}$ *such that*

(i) $\sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}|s(f, y_t, y_{t-1}; \boldsymbol{\theta})|^{n_f} < \infty$ *and either*

(ii) $\sup_{(f^*, y, y', \boldsymbol{\theta}) \in \mathbb{R} \times \mathcal{Y} \times \mathcal{Y} \times \Theta} |\beta + \alpha\, \partial s(f^*, y, y'; \boldsymbol{\theta})/\partial f| < 1$ *or*

(iii) $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \bar{\phi}'_{1,n_f}(\boldsymbol{\theta}) = \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\beta + \alpha\, \partial s(f^*, y_t, y_{t-1}; \boldsymbol{\theta})/\partial f| < 1$ *and* $f_t(y^{t-1}, \boldsymbol{\theta}, f_1) \perp \bar{\phi}'_{t+1,n_f}(\boldsymbol{\theta})\ \forall\ (t, f_1) \in \mathbb{N} \times \mathcal{F}$.

The moment conditions stated in Assumption 5 below require the definition of 'moment preserving map'. This allows us essentially to obtain moment conditions for the likelihood function from primitive conditions. In particular, we derive the number of bounded moments of the likelihood from the number of bounded moments of both the data and the filtered process.

DEFINITION 5. (Moment Preserving Maps) *A function* $H : \mathbb{R} \times \Theta \to \mathbb{R}$ *is said to be* n/m*-moment preserving, denoted as* $H \in \mathbb{M}_\Theta(n, m)$*, if and only if* $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |x_t(\boldsymbol{\theta})|^n < \infty$ *implies* $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |H(x_t(\boldsymbol{\theta}); \boldsymbol{\theta})|^m < \infty$.

ASSUMPTION 5. $h \in \mathbb{M}_\Theta(n_f, n_h)$ *and* $\log p_u \in \mathbb{M}_\Theta\left(n, n_{\log p_u}\right)$ *with* $n_{\log p_u} \geq 1$ *for* $n = \min\{n_y, n_y n_h/(n_y + n_h)\}$.

Theorem 3 obtains the consistency and asymptotic normality of the ML estimators of the parameters for SE data.

THEOREM 3. (Consistency) *Let* $\{y_t\}_{t\in\mathbb{Z}}$ *be an SE sequence satisfying* $\mathbb{E}|y_t|^{n_y} < \infty$ *for some* $n_y > 0$ *and assume that Assumptions 3, 4 and 5 hold. Furthermore, let* $\boldsymbol{\theta}_0 \in \Theta$ *be the unique maximizer of* $\ell_\infty(\boldsymbol{\theta})$ *on the parameter space* Θ*. Then the MLE satisfies* $\hat{\boldsymbol{\theta}}_T(f_1) \xrightarrow{a.s.} \boldsymbol{\theta}_0$ *as* $T \to \infty$ *for every* $f_1 \in \mathbb{R}$.

Theorem 4 obtains the asymptotic normality of the ML estimator. In this theorem we impose moment bounds directly on the derivatives of the likelihood function. As for Theorem 3, these moment bounds can be derived from primitive conditions concerning the moment preserving properties of h and p_u; see the technical appendix and Blasques, Koopman, and Lucas (2014b) for further details. However, for simplicity, Theorem 4 adopts the more usual approach of imposing moment conditions directly on the derivatives of the likelihood; see Straumann and Mikosch (2006).

Below we let $\mathcal{I}(\boldsymbol{\theta}_0) := \mathbb{E}\ell''_T(\boldsymbol{\theta}_0)$ be the Fisher information matrix evaluated at $\boldsymbol{\theta}_0 \in \Theta$, and $\mathcal{J}(\boldsymbol{\theta}_0) := \mathbb{E}\ell'_T(\boldsymbol{\theta}_0)\ell'_T(\boldsymbol{\theta}_0)^\top$ denote the expected outer product of gradients also evaluated at the point $\boldsymbol{\theta}_0 \in \Theta$.

THEOREM 4. (Asymptotic Normality) *Let* $\{y_t\}_{t\in\mathbb{Z}}$ *be an SE sequence satisfying* $\mathbb{E}|y_t|^{n_y} < \infty$ *for some* $n_y > 0$ *and let Assumptions 3, 4 and 5 hold. Furthermore, let* $\mathbb{E}|\ell'_T(\boldsymbol{\theta}_0)|^2 < \infty$, $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |\ell''_T(\boldsymbol{\theta})| < \infty$ *and* $\boldsymbol{\theta}_0 \in \text{int}(\Theta)$ *be the unique*

maximizer of ℓ_∞ *on* Θ*. Then the ML estimator* $\hat{\boldsymbol{\theta}}_T(f_1)$ *satisfies*

$$\sqrt{T}\Big(\hat{\boldsymbol{\theta}}_T(f_1) - \boldsymbol{\theta}_0\Big) \xrightarrow{d} \mathcal{N}\Big(0, \mathcal{I}^{-1}(\boldsymbol{\theta}_0)\mathcal{J}(\boldsymbol{\theta}_0)\mathcal{I}^{-1}(\boldsymbol{\theta}_0)\Big) \quad as \quad T \to \infty.$$

7 Small Sample Properties of the Score Filter

We analyze the filtering properties of our nonlinear ARMA modeling framework in a simulation setting. In particular, our Monte Carlo illustration focuses on the finite-sample performance of the filter when the data generating process for $\{y_t\}$ is given by

$$y_t = f_t y_{t-1} + u_t \quad , \quad u_t \sim N(0, \lambda)$$

with $\{f_t\}$ following a sigmoid path in the interval $[0, 1]$, that is

$$f_t = 0.5 + 0.5\sin(t/150).$$

Figure 3 is based on 1000 Monte Carlo simulated paths for a sample size of $T = 500$ and $T = 1500$. The Figure displays the true path of $\{f_t\}$ as well as the paths filtered by the Models I and II. The cloud of dots are filtered points $\{\tilde{f}_t\}$. The dashed lines are bounds containing 95% of the mass of the filter over the 1000 Monte Carlo repetitions. The presented results in Figure 3 illustrate how different specifications of the autoregressive model for $\{y_t\}$ can lead to score filters for $\{f_t\}$ with different properties. While Figure 3 shows that both models perform well, it also reveals that Model II (with the logistic link function) outperforms Model I (with the unit link function). In particular, Model I tends to do worse during periods of low dependence.

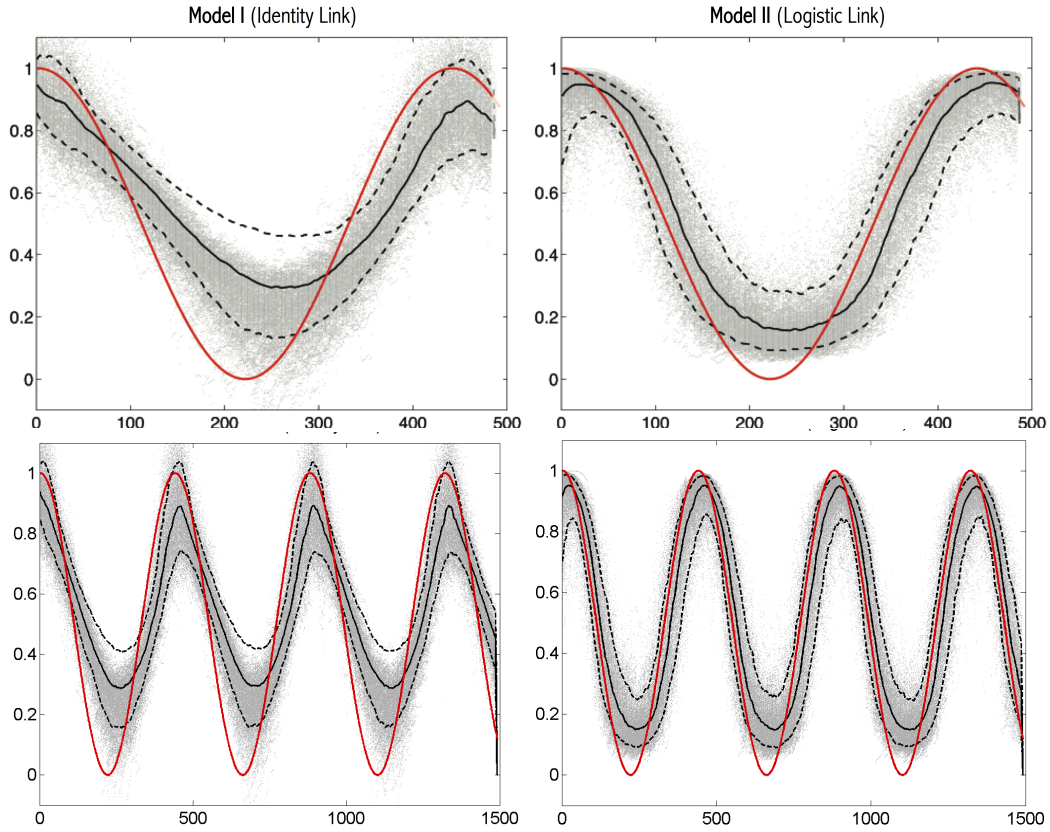

Figure 3: We present 1000 draws in a Monte Carlo performance comparison for Model I (left) and Model II (right) for sample size $T = 500$ (top) and $T = 1500$ (bottom).

8 Application to Unemployment Insurance Claims

We illustrate the empirical relevance of our new nonlinear dynamic models by analyzing the time-varying behaviour of weekly unemployment insurance claims (UIC) in the U.S. In particular, we consider our Affine Gaussian Model I of Section 2.1 with its nonlinear ARMA representation presented in Section 3.1. Another illustration for U.S. industrial production is provided in the technical appendix.

The development of economic models for unemployment insurance claims (UIC) and related macroeconomic variables, together with the empiricial econo-

metric analysis of UIC time series, have received considerable attention in the literature; see, amongst others, McMurrer and Chasanov (1995), Meyer (1995), Anderson and Meyer (1997, 2000),, Hopenhayn and Nicolini (1997) and Ashenfelter, Ashmore, and Deschenes (2005). Furthermore, the importance of forecasting weekly UIC time series data has been highlighted by Gavin and Kliesen (2002) where they show that UIC is highly effective as a leading indicator for labor market conditions and hence for the forecasting of GDP growth rates.

We analyse the weekly time series of growth rates for US seasonally adjusted UIC from 1960 towards 2013 by means of Model I. Figure 4 presents our observed UIC time series together with the filtered estimates of the time-varying autoregressive parameter from Model I. The autoregressive parameter estimates fluctuate considerably over time: ranging from a minimum of roughly 0.2 in the late 1960s where UIC data is shown to have low temporal dependence, to a maximum of over 0.6 in the 1980s where the process deviates from its overall mean persistently over an extended number of weeks. Furthermore, we observe that the temporal dependence in weekly UIC started to increase prior to the financial crisis unleashed by the fall of Lehman Brothers in 2008 reaching a peak of almost 0.6 in that year, followed by a steady decline until late 2010.

Table 2 compares the performance of Model I against that of the TAR, STAR and linear AR models. In the latter case, the order of the AR model is established by the general-to-specific methodology that selects the lag length based on the minimum corrected Akaike's information criterion (AICc) of Hurvich and Tsai (1991).

As in many economic time series, linear AR models tend to outperform nonlinear ones in modeling the UIC; see Clark and McCracken (2010). The reported log likelihood confirms that the autoregressive model of order 5 outperforms all models in maximizing the likelihood and minimizing the AICc. Table 2 reveals

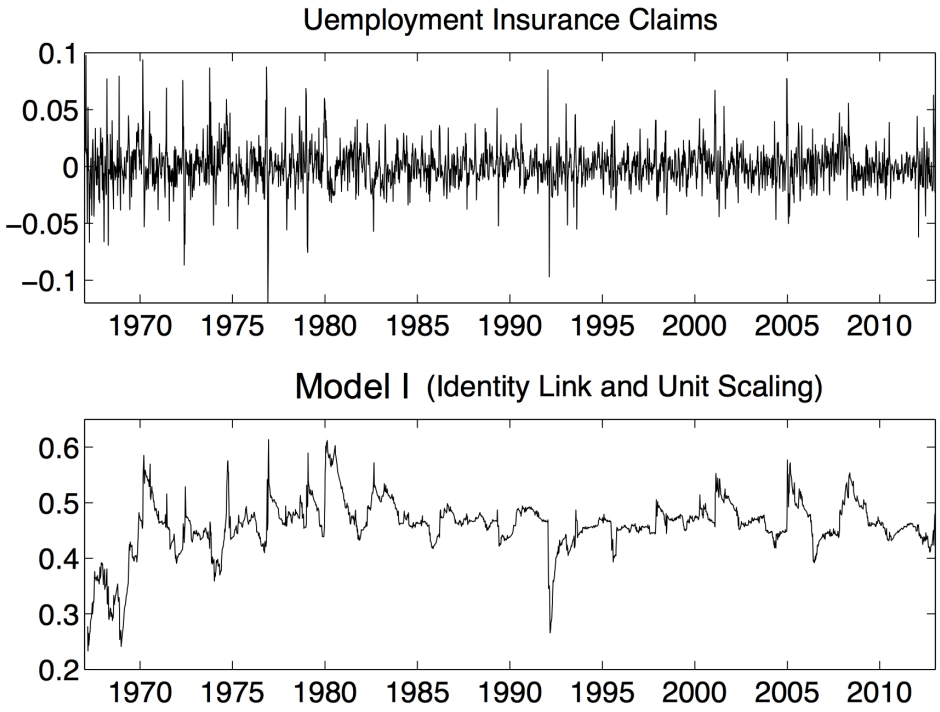

Figure 4: Growth rate of US seasonally adjusted weekly unemployment insurance claims reported by the Federal Reserve Bank of St. Louis.

however that the lower AICc of the AR(5) is not translated into a much better forecasting performance. On the contrary, among the possible linear AR(p) models, the AR(2) achieves the lowest RMSE of one-step ahead in-sample predictions. Furthermore, our preferred Model I outperforms the TAR and STAR models. Model I does not only outperform its nonlinear TAR and STAR rivals in terms of it maximized log likelihood value and its corresponding AICc value, it also outperforms the linear AR models in terms of forecasting accuracy.

Unemployment Insurance Claims: Model Comparison

	Model I	TAR	STAR	AR(2)	AR(5)
Log Lik	6743.96	6736.22	6736.86	6438.89	6967.71
AICc	-13477.901	-13462.41	-13463.70	-12869.76	-13921.39
F-RMSE	0.7502	0.7522	0.7521	0.8484	1.2081

Table 2: The values for log likelihood (Log Lik), Akaike's information criterion with finite sample correction (AICc) and root mean squared errors for 1, 2 and 3 step-ahead forecasts of the growth rate of US seasonally adjusted weekly unemployment insurance claims reported by the Federal Reserve Bank of St. Louis.

9 Final Remarks

We have introduced a new nonlinear dynamic model specification together with the corresponding linear autoregressive model that has time-varying temporal dependence parameters. Our nonlinear dynamic model class is based on recently developed observation driven score models for time-varying parameters. We have shown that the nonlinear dynamic model is optimal in an information theoretic updating sense and performs well in finite sample Monte Carlo exercises. We also have developed the asymptotic theory for the maximum likelihood estimator of the static parameter vector. In an empirical illustration for a macroeconomic time series, our most basic nonlinear dynamic model outperforms the well-known threshold and smooth-transition autoregressive models, in both in-sample fit and out-of-sample forecasting.

A Proofs of Theorems and Propositions

A.1 Proof of Theorem 1

Proof. Let $(\mathbb{C}(\Theta, \mathcal{F}), \|\cdot\|_\Theta)$ be a separable Banach space under compact Θ by application of the Arzelá–Ascoli theorem to obtain completeness and the Stone-Weierstrass theorem for separability. We obtain the uniform e.a.s. convergence of the filter

$$\sup_{\boldsymbol{\theta}\in\Theta} |f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1) - f_t(y^{t-1}, \boldsymbol{\theta})| \overset{e.a.s.}{\to} 0$$

from Bougerol (1993, Theorem 3.1) which applies to the sequence $\{f_t(y^{1:t-1}, \cdot, f_1^\Theta)\}_{t\in\mathbb{N}}$ with elements $f_t(y^{1:t-1}, \cdot, f_1^\Theta)$ taking values in the separable Banach space $\mathcal{F}_\Theta \subseteq (\mathbb{C}(\Theta, \mathcal{F}), \|\cdot\|_\Theta)$ with initialization f_1^Θ in $\mathbb{C}(\Theta, \mathcal{F})$ at $t = 1$, $f_1^\Theta(\boldsymbol{\theta}) = f_1 \ \forall\ \boldsymbol{\theta} \in \Theta$, and generated according to

$$f_t(y^{1:t}, \cdot, f_1^\Theta) = \phi_t\big(f_t(y^{1:t-1}, \cdot, f_1^\Theta)\big) \ \forall\ t \in \mathbb{N},$$

where $\{\phi_t\}_{t\in\mathbb{Z}}$ is here a sequence of stochastic recurrence equations $\phi_t : \Xi \times \mathbb{C}(\Theta, \mathcal{F}) \to \mathbb{C}(\Theta, \mathcal{F}) \ \forall\ t$ as in Straumann and Mikosch (2006, Proposition 3.12). The SE nature of $\{y_t\}_{t\in\mathbb{Z}}$ and the continuity of ϕ on $\mathcal{F} \times \mathcal{Y}^2 \times \Theta$, implied by $s \in \mathbb{C}(\mathcal{F} \times \mathcal{Y}^2 \times \Theta)$, ensures that $\{\phi_t\}_{t\in\mathbb{Z}}$ is SE by Krengel (1985, Proposition 4.3).

Condition C1 in Bougerol (1993, Theorem 3.1) follows immediately from the moment bound $\mathbb{E}\ln^+ \sup_{\boldsymbol{\theta}\in\Theta} |s(f_1, y_t, y_{t-1}, \boldsymbol{\theta})| < \infty$ by a simple norm sub-additivity. Similarly, the log moment bound $\mathbb{E}\ln^+ \sup_{\boldsymbol{\theta}\in\Theta} |s(f_1, y_t, y_{t-1}, \boldsymbol{\theta})| < \infty$ implies $\mathbb{E}\log^+ \|\phi_0(f^\Theta) - f^\Theta\|_\Theta^{n_f} < \infty$.

For any pair $(f^\Theta, f'^\Theta) \in \mathbb{C}(\Theta) \times \mathbb{C}(\Theta)$, define

$$\rho_t = \rho(\phi_t) = \sup_{(f^\Theta, f'^\Theta)\in\mathcal{F}_\Theta\times\mathcal{F}_\Theta} \frac{\|\phi_t(f^\Theta) - \phi_t(f'^\Theta)\|_\Theta}{\|f^\Theta - f'^\Theta\|_\Theta}.$$

Condition C2 in Bougerol (1993, Theorem 3.1) holds if $\mathbb{E}\ln\rho_t < 0$ and this is ensured by

$\mathbb{E}\ln\sup_{\boldsymbol{\theta}\in\Theta}\bar{\phi}'_{t,1}(\boldsymbol{\theta}) = \mathbb{E}\ln\sup_{\boldsymbol{\theta}\in\Theta}\bar{\phi}'_{1,1}(\boldsymbol{\theta}) < 0$ because

$$\begin{aligned}
\mathbb{E}\ln\rho(\phi_t) &:= \mathbb{E}\ln \sup_{(f^\Theta, f'^\Theta)\in\mathcal{F}_\Theta\times\mathcal{F}_\Theta\subseteq\mathbb{C}(\Theta,\mathcal{F})\times\mathbb{C}(\Theta,\mathcal{F}):\|f^\Theta - f'^\Theta\|>0} \frac{\|\phi_t(f^\Theta) - \phi_t(f'^\Theta)\|_\Theta}{\|f^\Theta - f'^\Theta\|_\Theta} \\
&= \mathbb{E}\ln \sup_{(f^\Theta, f'^\Theta)\in\mathcal{F}_\Theta\times\mathcal{F}_\Theta:\|f^\Theta - f'^\Theta)\|>0} \frac{\sup_{\boldsymbol{\theta}\in\Theta}|\phi(f(\boldsymbol{\theta},f_1),y_t,y_{t-1},\boldsymbol{\theta}) - \phi(f'(\boldsymbol{\theta},f_1),y_t,y_{t-1},\boldsymbol{\theta})|}{\sup_{\boldsymbol{\theta}\in\Theta}|f(\boldsymbol{\theta},f_1) - f'(\boldsymbol{\theta},f_1)|} \\
&\leq \mathbb{E}\ln \sup_{(f,f')\in\mathcal{F}_\Theta\times\mathcal{F}_\Theta:\|f^\Theta - f'^\Theta)\|>0} \frac{\sup_{\boldsymbol{\theta}\in\Theta}\bar{\phi}'_{t,1}(\boldsymbol{\theta})\sup_{\boldsymbol{\theta}\in\Theta}|f(\boldsymbol{\theta},f_1) - f'(\boldsymbol{\theta},f_1)|}{\sup_{\boldsymbol{\theta}\in\Theta}|f(\boldsymbol{\theta},f_1) - f'(\boldsymbol{\theta},f_1)|} \\
&= \mathbb{E}\ln\sup_{\boldsymbol{\theta}\in\Theta}\bar{\phi}'_{t,1}(\boldsymbol{\theta}) < 0.
\end{aligned}$$

and since $\mathbb{E}\ln\rho_1 \leq \mathbb{E}\ln\sup_{\boldsymbol{\theta}\in\Theta}\bar{\phi}'_{1,1}(\boldsymbol{\theta}) < 0 < \infty$ and $\mathbb{E}\ln\rho(\phi_J\circ\ldots\circ\phi_1) \leq \mathbb{E}\ln\prod_{j=1}^J\rho_j \leq \mathbb{E}\sum_{j=1}^J\ln\bar{\phi}'_{1,1}(\boldsymbol{\theta}) < 0$. As a result, we conclude that

$$\sup_{\boldsymbol{\theta}\in\Theta}|f_t(y^{1:t-1},\boldsymbol{\theta},f_1) - f_t(y^{1:t-1},\boldsymbol{\theta})| \overset{e.a.s.}{\to} 0;$$

i.e. $\{f_t(\cdot,f_1)\}_{t\in\mathbb{N}}$ converges e.a.s. to an SE solution $\{f_t(\cdot)\}_{t\in\mathbb{Z}}$ in $\|\cdot\|_\Theta$-norm. Uniqueness and e.a.s. convergence is obtained in Straumann and Mikosch (2006, Theorem 2.8).

Finally, the moment bounds

$$\sup_t\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|f_t(y^{1:t-1},\boldsymbol{\theta},f_1)|^{n_f} < \infty \quad\text{and}\quad \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|f_t(y^{1:t-1},\boldsymbol{\theta})|^{n_f} < \infty$$

are obtained by noting that

$$\sup_t\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|f_t(y^{1:t-1},\boldsymbol{\theta},f_1)|^{n_f} < \infty$$

if and only if

$$\sup_t(\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|f_t(y^{1:t-1},\boldsymbol{\theta},f_1)|^{n_f})^{1/n_f} = \sup_t\|f_t(\cdot,f_1)\|^\Theta_{n_f} < \infty,$$

and that, for any $f^\Theta\in\mathbb{C}(\Theta,\mathcal{F})$, having $\|f_t(\cdot,f_1) - f^\Theta\|^\Theta_{n_f} < \infty$ implies $\|f_t(\cdot,f_1^\Theta)\|^\Theta_{n_f} < \infty$ since continuity on the compact Θ implies $\sup_{\boldsymbol{\theta}\in\Theta}|f(\boldsymbol{\theta})| < \infty$. The moment bound can then be obtained by a simple adaptation of the proof of Propositions SA.1 and SA.2 in Blasques at al. (2014b). In particular, obtain the moment bound by noting that there exists $\bar{\bar{\phi}} < \infty$ and $\bar{f} < \infty$ such that

$$\sup_t\|f_t(\cdot,f_1^\Theta) - f^\Theta\|^\Theta_{n_f} \leq \bar{c}\sup_t\|f_t(\cdot,f_1^\Theta) - f^\Theta\|^\Theta_{n_f} + A$$

with $\bar{c} = \sup_{\boldsymbol{\theta}\in\Theta} \bar{\phi}'(\boldsymbol{\theta})$, $A = \bar{c}\sup_{\boldsymbol{\theta}\in\Theta} |f_t - f^{\Theta}(\boldsymbol{\theta})| + \bar{\bar{\phi}} + \bar{f} = (\bar{c}+1)\bar{f} + \bar{\bar{\phi}}$, and hence,

$$\begin{aligned}\sup_t \|f_t(\cdot, f_1^{\Theta}) - f^{\Theta}\|_{n_f}^{\Theta} &\leq \sum_{j=0}^{t} (\bar{c})^j ((\bar{c}+1)\bar{f} + \bar{\bar{\phi}}) + \bar{c}^{t+1} \sup_t \|f_1^{\Theta} - f^{\Theta}\|_{n_f}^{\Theta} \\ &\leq \frac{(\bar{c}+1)\bar{f} + \bar{\bar{\phi}}}{1-\bar{c}} + \|f_1^{\Theta} - f^{\Theta}\|_{n_f}^{\Theta} < \infty.\end{aligned}$$

For the convenience of the reader, a detailed proof is made available in the technical appendix. □

A.2 Proof of Theorem 2

Proof. Assumption 3 implies that the $\ell_T(\boldsymbol{\theta}, f_1) = (1/T)\sum_{t=2}^{T} \ell_t(\boldsymbol{\theta}, f_1)$ is a.s. continuous in $\boldsymbol{\theta} \in \Theta$ through continuity of each

$$\ell_t(\boldsymbol{\theta}, f_1) = \ell(y_t, f_t(y^{t-1}, f_1, \boldsymbol{\theta}), \boldsymbol{\theta}) = p_u(y_t - h(f_t(y^{t-1}, f_1, \boldsymbol{\theta}))y_{t-1}, \boldsymbol{\theta})$$

ensured in turn by the differentiability of p_u and h and the implied a.s.c. of

$$\nabla(y_t, f_t(y^{t-1}, f_1, \boldsymbol{\theta}); \boldsymbol{\theta}) = \partial \log p_u(y_t - h(f_t)y_{t-1})/\partial f$$

in $(f_t(y^{t-1}, f_1, \boldsymbol{\theta}); \boldsymbol{\theta})$ and the resulting c. of $f_t(y^{t-1}, f_1, \boldsymbol{\theta})$ in $\boldsymbol{\theta}$ as a composition of t continuous maps. Together with the compactness of Θ this implies by Weierstrass' theorem that the arg max set is non-empty a.s. and hence that $\hat{\boldsymbol{\theta}}_T$ exists a.s. $\forall\, T \in \mathbb{N}$. Assumption 3 implies also by a similar argument that $\ell_T(\boldsymbol{\theta}, f_1) = \ell\big(y^t, f^T(y^{t-1}, f_1, \boldsymbol{\theta}), \boldsymbol{\theta}\big)$ is continuous in y^t $\forall\, \boldsymbol{\theta} \in \Theta$ and hence measurable w.r.t. a Borel σ-algebra. The measurability of $\hat{\boldsymbol{\theta}}_T$ follows from White (1994, Theorem 2.11) or Gallant and White (1988, Lemma 2.1, Theorem 2.2). □

A.3 Proof of Theorem 3

Proof. We obtain $\hat{\boldsymbol{\theta}}_T(f_1) \overset{a.s.}{\to} \boldsymbol{\theta}_0$ from the uniform convergence of the log likelihood function

$$\sup_{\boldsymbol{\theta}\in\Theta} |\ell_T(\boldsymbol{\theta}, f_1) - \ell_\infty(\boldsymbol{\theta})| \overset{a.s.}{\to} 0 \;\forall\; f_1 \in \mathcal{F} \;\text{ as }\; T \to \infty \tag{19}$$

and the identifiable uniqueness of the maximizer $\boldsymbol{\theta}_0 \in \Theta$ introduced in White (1994),

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\epsilon} \ell_\infty(\boldsymbol{\theta}) < \ell_\infty(\boldsymbol{\theta}_0) \ \forall \ \epsilon > 0. \tag{20}$$

See White (1994, Theorem 3.4) or Theorem 3.3 in Gallant and White (1988) for a proof of consistency based on (19) and (20).

We obtain(19) by noting that

$$\sup_{\boldsymbol{\theta}\in\Theta} |\ell_T(\boldsymbol{\theta}, f_1) - \ell_\infty(\boldsymbol{\theta})| \ \leq \ \sup_{\boldsymbol{\theta}\in\Theta} |\ell_T(\boldsymbol{\theta}, f_1) - \ell_T(\boldsymbol{\theta})| + \sup_{\boldsymbol{\theta}\in\Theta} |\ell_T(\boldsymbol{\theta}) - \ell_\infty(\boldsymbol{\theta})| \tag{21}$$

and then showing that both terms on the right hand side of the inequality vanish.

By continuity, the first term in (21) vanishes almost surely if

$$\sup_{\boldsymbol{\theta}\in\Theta} |\ell_t(\boldsymbol{\theta}, f_1) - \ell_t(\boldsymbol{\theta})| \overset{a.s.}{\to} 0 \quad \text{as} \quad t \to \infty.$$

The continuity of p_u ensures that

$$\ell_t(\cdot, f_1) = \ell(f_t(y^t, \cdot, f_1), y_t, y_{t-1}, \cdot)$$

has ℓ continuous in $(f_t(y^t, \cdot, f_1), y_t, y_{t-1})$. Since all the assumptions of Theorem 1 are satisfied we know that there exists a unique SE sequence $\{f_t(y^t, \cdot))\}_{t\in\mathbb{Z}}$ with elements taking values in $\mathbb{C}(\Theta, \mathcal{F})$ such that

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|(f_t(y^{t-1}, f_1, \boldsymbol{\theta}), y_t, y_{t-1}) - (f_t(y^t, \boldsymbol{\theta}), y_t, y_{t-1})\right| \overset{a.s.}{\to} 0$$

and $\sup_t \mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} |f_t(y^{t-1}, f_1, \boldsymbol{\theta})|^{n_f} < \infty$ and $\mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} |f_t(y^t, \boldsymbol{\theta})|^{n_f} < \infty$ with $n_f \geq 1$. Hence, the application of a continuous mapping theorem for $\ell : \mathbb{C}(\Theta, \mathcal{F}) \to \mathbb{C}(\Theta, \mathcal{F})$ yields the desired result.

The second term in (21) vanishes by applying the ergodic theorem for separable Banach spaces in Rao (1962) to the likelihood sequence $\{\ell_T(\cdot)\}$ with elements taking values in $\mathbb{C}(\Theta, \mathbb{R})$. Subsequently, the SE nature of $\{\ell_T\}_{t\in\mathbb{Z}}$ is implied by the continuity of ℓ on the SE sequence $\{(y_t, y_{t-1}, f_t(y^{t-1}, \cdot))\}_{t\in\mathbb{Z}}$ and by the Proposition 4.3 in Krengel (1985) The moment bound $\mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} |\ell_t(\boldsymbol{\theta})| < \infty$ is implied by $\mathbb{E}|y_t|^{n_y} < \infty$, $\sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}|f_t(y^{t-1}, \boldsymbol{\theta})|^{n_f} < \infty \ \forall \ \boldsymbol{\theta} \in \Theta$ and the

moment preserving conditions of Assumption 5 which ensures

$$\sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}|h(f_t(y^{t-1},\boldsymbol{\theta}))|^{n_h} < \infty \;\forall\; \boldsymbol{\theta}\in\Theta$$

and hence

$$\sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}|y_t - h(f_t(y^{t-1},\boldsymbol{\theta}))y_{t-1}|^{n} < \infty \;\forall\; \boldsymbol{\theta}\in\Theta$$

with $n = \min\{n_y, n_y n_h/(n_y+n_h)\}$, by a generalized Holder inequality. As a result, by Assumption 5,

$$\sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}|\log p_u(y_t - h(f_t(y^{t-1},\boldsymbol{\theta}))y_{t-1})| < \infty \;\forall\; \boldsymbol{\theta}\in\Theta$$

and the ULLN $\sup_{\boldsymbol{\theta}\in\Theta}|\ell_T(\boldsymbol{\theta}) - \mathbb{E}\ell_t(\boldsymbol{\theta})| \overset{a.s.}{\to} 0$ as $T\to\infty$ follows.

Finally, the identifiable uniqueness (see e.g. White (1994)) of $\boldsymbol{\theta}_0\in\Theta$ in (20) is ensured by the uniqueness of $\boldsymbol{\theta}_0$ as the maximizer of the likelihood, the compactness of Θ, and the continuity of the limit likelihood function $\mathbb{E}\ell_t(\boldsymbol{\theta})$ in $\boldsymbol{\theta}\in\Theta$ which is obtained from the continuity of ℓ_T in $\boldsymbol{\theta}\in\Theta\;\forall\; T\in\mathbb{N}$ and the uniform convergence in (19). □

A.4 Proof of Theorem 4

Proof. Asymptotic normality is obtain from (see e.g. White (1994, Theorem 6.2) for a proof): (i) the consistency of $\hat{\boldsymbol{\theta}}_T \overset{a.s.}{\to} \boldsymbol{\theta}_0 \in \text{int}(\Theta)$; (ii) the a.s. twice continuous differentiability of $\ell_T(\boldsymbol{\theta}, f_1)$ in $\boldsymbol{\theta}\in\Theta$; (iii) the asymptotic normality of the score

$$\sqrt{T}\ell'_T\big(\boldsymbol{\theta}_0, \boldsymbol{f}_1^{(0:1)}\big) \overset{d}{\to} N\big(0, J(\boldsymbol{\theta}_0)\big) \;\text{ as }\; T\to\infty \;\text{ where }\; J(\boldsymbol{\theta}_0) = \mathbb{E}\big(\ell'_t\big(\boldsymbol{\theta}_0\big)\big)^2; \tag{22}$$

(iv) the uniform convergence of the likelihood's second derivative,

$$\sup_{\boldsymbol{\theta}\in\Theta} \big\|\ell''_T(\boldsymbol{\theta}, \boldsymbol{f}_1^{(0:2)}) - \ell''_\infty(\boldsymbol{\theta})\big\|_{\mathbb{R}^{16}} \overset{a.s.}{\to} 0 \;\text{ as }\; T\to\infty\;; \tag{23}$$

and finally, (v) the non-singularity of the limit $\ell''_\infty(\boldsymbol{\theta}) = \mathbb{E}\ell''_t(\boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\theta})$.

(i) follows by Theorem 3 and the additional assumption that $\boldsymbol{\theta}_0\in\text{int}(\Theta)$.

(ii) follows from Assumption 3.

(iii) follows by Theorem 18.10[iv] in van der Vaart (2000) by showing that,

$$\|\ell'_T\big(\boldsymbol{\theta}_0, \boldsymbol{f}_1^{(0:1)}\big) - \ell'_T\big(\boldsymbol{\theta}_0\big)\|_{\mathbb{R}^4} \overset{e.a.s.}{\to} 0 \;\text{ as }\; T\to\infty \tag{24}$$

to conclude that

$$\|\sqrt{T}\ell'_T(\boldsymbol{\theta}_0, \boldsymbol{f}_1^{(0:1)}) - \sqrt{T}\ell'_T(\boldsymbol{\theta}_0)\|_{\mathbb{R}^4} = \sqrt{T}\|\ell'_T(\boldsymbol{\theta}_0, \boldsymbol{f}_1^{(0:1)}) - \ell'_T(\boldsymbol{\theta}_0)\|_{\mathbb{R}^4} \overset{a.s.}{\to} 0 \quad \text{as } T \to \infty$$

and applying the CLT for SE martingales in Billingsley (1961) to obtain

$$\sqrt{T}\ell'_T(\boldsymbol{\theta}_0) \overset{d}{\to} N(0, \mathcal{I}(\boldsymbol{\theta}_0)) \text{ as } T \to \infty \text{ where } \mathcal{I}(\boldsymbol{\theta}_0) = \mathbb{E}\big(\ell'_t(\boldsymbol{\theta}_0, f_1)\big)^2. \tag{25}$$

The e.a.s. convergence in (24) follows from

$$|f_t(y^{t-1}, \boldsymbol{\theta}_0, f_1) - f_t(y^{t-1}, \boldsymbol{\theta}_0)| \overset{e.a.s.}{\to} 0$$

and

$$\|\boldsymbol{f}_t^{(1)}(y^{t-1}, \boldsymbol{\theta}_0, \boldsymbol{f}_1^{(0:1)}) - \boldsymbol{f}_t^{(1)}(y^{t-1}, \boldsymbol{\theta}_0)\|_{\mathbb{R}^4} \overset{e.a.s.}{\to} 0$$

which are both implied by the conditions of Theorem 1; see the technical appendix and Blasques et al. (2014b) for further details on the e.a.s. convergence of the derivative process. The convexity of $\mathcal{F}$ and the differentiability of the likelihood function

$$\ell'_t(\boldsymbol{\theta}, \boldsymbol{f}_1^{(0:1)}) = \ell'(y^t, \boldsymbol{f}_t^{(0:1)}(y^{t-1}, \boldsymbol{\theta}, \boldsymbol{f}_1^{(0:1)}))$$

allows us to employ the mean-value theorem

$$\begin{aligned}\|\ell'_T(\boldsymbol{\theta}_0, \boldsymbol{f}_1^{(0:1)}) - \ell'_T(\boldsymbol{\theta}_0)\|_{\mathbb{R}^4} &\leq \sum_{i=1}^{5} \Big|\frac{\partial \ell'(y^t, \hat{\boldsymbol{f}}_t^{(0:1)})}{\partial f}\Big| \big|\boldsymbol{f}_{j,t}^{(0:1)}(y^{t-1}, \boldsymbol{\theta}_0, \boldsymbol{f}_1^{(0:1)}) - \boldsymbol{f}_{j,t}^{(0:1)}(y^{t-1}, \boldsymbol{\theta}_0)\big| \\ &= \sum_{i=1}^{5} O_p(1) o_{e.a.s}(1) = o_{e.a.s.}(1)\end{aligned}$$

where $\boldsymbol{f}_{j,t}^{(0:1)}$ denotes the j-th element of $\boldsymbol{f}_t^{(0:1)}$ and hence vanishes e.a.s.. The probability bound

$$\big|\partial \ell'(y^t, \hat{\boldsymbol{f}}_t^{(0:1)})/\partial f\big| = O_p(1)$$

is implied by the moment bound $\mathbb{E}\big|\partial \ell'(y^t, \hat{\boldsymbol{f}}_t^{(0:1)})/\partial f\big| < \infty$. The CLT in (25) finally follows from Billingsley (1961) since $\{\ell'_T(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is an SE martingale sequence with $\mathbb{E}\big(\ell'_T(\boldsymbol{\theta}_0)\big)^2 < \infty$.

(iv) is obtained by noting that

$$\sup_{\boldsymbol{\theta}\in\Theta}\|\ell_T''(\boldsymbol{\theta},f_1)-\ell_\infty''(\boldsymbol{\theta})\|_{\mathbb{R}^{16}}\leq\sup_{\boldsymbol{\theta}\in\Theta}\|\ell_T''(\boldsymbol{\theta},f_1)-\ell_T''(\boldsymbol{\theta})\|_{\mathbb{R}^{16}}+\sup_{\boldsymbol{\theta}\in\Theta}\|\ell_T''(\boldsymbol{\theta})-\ell_\infty''(\boldsymbol{\theta})\|_{\mathbb{R}^{16}}$$

then showing that both terms on the right hand side vanish a.s. and applying the ergodic theorem for separable Banach spaces in Rao (1962) (see also Straumann and Mikosch (2006, Theorem 2.7)) to the sequence $\{\ell_T''(\cdot)\}$ with elements taking values in $\mathbb{C}(\Theta,\mathbb{R}^{16})$ so that $\sup_{\boldsymbol{\theta}\in\Theta}\|\ell_T''(\boldsymbol{\theta})-\ell_\infty''(\boldsymbol{\theta})\|_{\mathbb{R}^{16}}\overset{a.s.}{\to}0$ where $\ell_\infty''(\boldsymbol{\theta})=\mathbb{E}\ell_t''(\boldsymbol{\theta})\ \forall\ \boldsymbol{\theta}\in\Theta$.

The first term satisfies $\sup_{\boldsymbol{\theta}\in\Theta}\|\ell_t''(\boldsymbol{\theta},f_1)-\ell_t''(\boldsymbol{\theta})\|_{\mathbb{R}^{16}}\overset{a.s.}{\to}0$ as $t\to\infty$. The smoothness conditions in Assumption 4 ensure that $\ell_t''(\cdot,f_1)=\ell''(y_t,\boldsymbol{f}_t^{(0:2)}(y^{t-1},\cdot,\boldsymbol{f}_{0:2}),\cdot)$ with ℓ'' continuous in $(y_t,\boldsymbol{f}_t^{(0:2)}(y^{t-1},\cdot,\boldsymbol{f}_{0:2}))$. Under the conditions of Theorem 1, we know that there exists a unique SE sequence $\{\boldsymbol{f}_t^{(0:2)}(y^{t-1},\cdot)\}_{t\in\mathbb{Z}}$ with elements taking values in $\mathbb{C}(\Theta,\mathcal{F}^{(0:i)})$ such that $\sup_{\boldsymbol{\theta}\in\Theta}\left\|(y_t,\boldsymbol{f}_t^{(0:2)}(y^{t-1},\boldsymbol{\theta},\boldsymbol{f}_{0:2}))-(y_t,\boldsymbol{f}_t^{(0:2)}(y^{t-1},\boldsymbol{\theta})\right\|\overset{a.s.}{\to}0$; see the technical appendix and Blasques et al. (2014b) for further details. Hence, the desired result is obtained by application of a continuous mapping theorem for $\ell'':\mathbb{C}(\Theta,\mathcal{F}^{(0:2)})\to\mathbb{C}(\Theta,\mathcal{F}^{(0:2)})$.

The ULLN $\sup_{\boldsymbol{\theta}\in\Theta}\|\ell_T''(\boldsymbol{\theta})-\mathbb{E}\ell_t''(\boldsymbol{\theta})\|_{\mathbb{R}^{16}}\overset{a.s.}{\to}0$ as $T\to\infty$ follows, under the moment bound $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\ell_t''(\boldsymbol{\theta})\|_{\mathbb{R}^{16}}<\infty$, by the SE nature of $\{\ell_T''\}_{t\in\mathbb{Z}}$ which is implied by continuity of ℓ'' on the SE sequence $\{(y_t,f_t(y^{t-1},\cdot))\}_{t\in\mathbb{Z}}$ and Proposition 4.3 in Krengel (1985).

(v) is implied by the uniqueness of $\boldsymbol{\theta}_0$ as a maximum of the limit likelihood.

□

B Extensions to Non-Local Optimality

Lemma 3. *Let Assumptions 1 and 2 hold. Then, the score update is locally RKL optimal given y_{t-1} for every p_t if*

$$\alpha>\frac{|\omega+(\beta-1)\tilde{f}_t|}{S(\tilde{f}_t,y_{t-1};\boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t,y_t,y_{t-1};\boldsymbol{\theta})|}. \tag{26}$$

Furthermore, the score update is locally CKL optimal given y_{t-1} for every p_t if

$$\alpha>\frac{\mathbb{E}_{Y_f}^{t-1}|\tilde{\nabla}(\tilde{f}_t,y_t,y_{t-1};\boldsymbol{\theta})|}{S(\tilde{f}_t,y_{t-1};\boldsymbol{\theta})\mathbb{E}_{Y_f}^{t-1}|\tilde{\nabla}(\tilde{f}_t,y_t,y_{t-1};\boldsymbol{\theta})|^2}|\omega+(\beta-1)\tilde{f}_t|, \tag{27}$$

$$\textit{with}\quad \mathbb{E}_{Y_f}^{t-1}|\tilde{\nabla}(\tilde{f}_t,y_t,y_{t-1};\boldsymbol{\theta})|=\int_{Y_f}p(y_t|f_t)\,|\tilde{\nabla}(\tilde{f}_t,y,y_{t-1};\boldsymbol{\theta})|\mathrm{d}y$$

$$\textit{and}\quad Y_f:=\{y\in\mathbb{R}:|\phi(\tilde{f}_t,y,y_{t-1};\boldsymbol{\theta})-\tilde{f}_t|<\delta_f\}.$$

For any given value of α, the larger the absolute value of the scaled score $S(\tilde{f}_t;\boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta})|$, the more likely it is that the realized step is locally optimal. Similarly, the closer ω is to 0 or β is to one (which corresponds to the Newton-score update), the easier it is to obtain local optimality.

Example 1. For Model I the local RKL optimality condition in (26) reduces to

$$\alpha > \sigma^2 \frac{|\omega + (\beta - 1)\tilde{f}_t|}{|(y_{t-1} - \tilde{f}_{t-1}y_{t-2})y_{t-2}|}, \tag{28}$$

and the local CKL optimality condition in (27) is given by

$$\alpha > \frac{\mathbb{E}_{Y_f}^{t-1}|(y_{t-1} - \tilde{f}_{t-1}y_{t-2})y_{t-2}|}{\mathbb{E}_{Y_f}^{t-1}|(y_{t-1} - \tilde{f}_{t-1}y_{t-2})y_{t-2}|^2}|\omega + (\beta - 1)\tilde{f}_t|. \tag{29}$$

C Proofs of Optimality Lemmas

Proof of Proposition 1. This proof follows closely that in Blasques et al. (2014), but extends it by allowing the scaling, the score and all conditional densities to depend on y_{t-1}. In this sense, the derived result is also different as it obtains optimality conditional on y_{t-1}. Note also that this proof could be made considerably shorter by adopting a different method of proof. However, the present method of proof allows us to reduce the length of all subsequent proofs since we can borrow substantially from the expressions derived here.

By a repeated application of the mean value theorem to $\tilde{p}(y|\tilde{f}_{t+1}, y_{t-1};\boldsymbol{\theta})$ and $\tilde{\nabla}_t(\tilde{f}^*_{t+1}, y_t, y_{t-1};\boldsymbol{\theta})$, and using the form of the Newton-GAS update $\tilde{f}_{t+1} - \tilde{f}_t = \alpha S(\tilde{f}_t, y_{t-1};\boldsymbol{\theta})\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta})$, we

obtain CKL optimality if

$$
\begin{aligned}
&\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \ln \frac{\tilde{p}(y|\tilde{f}_t;\boldsymbol{\theta})}{\tilde{p}(y|\tilde{f}_{t+1}, y_{t-1};\boldsymbol{\theta})} \mathrm{d}y \\
&\quad = \quad -\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \frac{\partial \ln \tilde{p}(y|\tilde{f}^*_{t+1};\boldsymbol{\theta})}{\partial f} (\tilde{f}_{t+1} - \tilde{f}_t) \mathrm{d}y \\
&\quad = \quad -\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \tilde{\nabla}(\tilde{f}^*_{t+1}, y, y_{t-1};\boldsymbol{\theta})\ \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta}) \mathrm{d}y \\
&\quad = \quad -\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \big(\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta})\big)^2 \mathrm{d}y \qquad (30) \\
&\qquad\quad -\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta}) \frac{\partial \tilde{\nabla}(\tilde{f}^{**}_{t+1}, y^{**}_t;\boldsymbol{\theta})}{\partial y} (y_t - y) \mathrm{d}y \\
&\qquad\quad -\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta}) \frac{\partial \tilde{\nabla}(\tilde{f}^{**}_{t+1}, y^{**}_t;\boldsymbol{\theta})}{\partial f} (\tilde{f}^*_{t+1} - \tilde{f}_t) \mathrm{d}y < 0, \\
&\quad =: \quad -\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \big(\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta})\big)^2 \mathrm{d}y + A(\delta_f, \delta_y) + B(\delta_f, \delta_y), \qquad (31)
\end{aligned}
$$

where $\tilde{f}^*_{t+1}$ is a point between $\tilde{f}_{t+1}$ and $\tilde{f}_t$, $\tilde{f}^{**}_{t+1}$ is a point between $\tilde{f}^*_{t+1}$ and $\tilde{f}_t$, y^{**}_t is a point between y_t and y, and $A(\delta_f, \delta_y)$ and $B(\delta_f, \delta_y)$ in (31) are equal to the second and third term of (30), respectively. From Assumptions 1 and 2 we obtain $\alpha S(\tilde{f}_t;\boldsymbol{\theta})\big(\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta})\big)^2 > 0$ almost surely, such that for every $\tilde{f}_t$ and p_t, $\exists \gamma < 0$ such that

$$
-\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \alpha S(\tilde{f}_t;\boldsymbol{\theta}) \big(\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta})\big)^2 \mathrm{d}y \leq \gamma < 0.
$$

The desired result now follows by noting that second and third term in (30) can be made arbitrarily small compared to the first term due to the differentiability of the score and the compactness of $Y_{\delta_y}(y_t)$; see the working paper version for more details.

The proof for local CKL-optimality follows immediately by a similar argument using the assumption that $\tilde{f}_{t+1}$ is a continuous random variable with a density. More details can be found in the working paper version of this paper. □

Proof of Proposition 2. Let $\tilde{f}_{t+1} - \tilde{f}_t = \phi(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta}) - \phi(\tilde{f}_{t-1}, y_{t-1};\boldsymbol{\theta}) = \Delta\phi(\tilde{f}_t, y_t, y_{t-1};\boldsymbol{\theta})$. We follow the same line of proof as for Proposition 1. To prove the 'if' part, we write the local

RKL variation is for any given p_t as

$$-\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \frac{\partial \ln \tilde{p}(y|\tilde{f}^*_{t+1}; \boldsymbol{\theta})}{\partial f} (\tilde{f}_{t+1} - \tilde{f}_t) \mathrm{d}y$$
$$= -\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \tilde{\nabla}(\tilde{f}^*_{t+1}, y, y_{t-1}; \boldsymbol{\theta}) \Delta\phi(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta}) \mathrm{d}y.$$

Using the definition of a score-equivalent update and by the same argument as in the proof of Proposition 1, we have that for sufficiently small δ_y

$$\mathrm{sign}(\tilde{\nabla}(\tilde{f}^*_{t+1}, y, y_{t-1}; \boldsymbol{\theta})) = \mathrm{sign}(\Delta\phi(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta})) \quad \forall\ (f, y) \in F_{\delta_f}(\tilde{f}_t) \times Y_{\delta_y}(y_t),$$

and hence

$$\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \tilde{\nabla}(\tilde{f}^*_{t+1}, y, y_{t-1}; \boldsymbol{\theta}) \Delta\phi(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta}) \mathrm{d}y > 0. \tag{32}$$

This implies that (C) is strictly negative under the regularity conditions of Assumptions 1 and 2. A similar argument holds for CKL variation by taking a subsequent expectation over $\tilde{f}_{t+1}$ given $\tilde{f}_t$ and f_t.

To prove the 'only if' part, suppose that the update $\tilde{f}_{t+1} = \phi(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta})$ is not score-equivalent. Then, by Assumption 1, there must exists an open set $FY \subseteq \mathcal{F} \times \mathbb{R}$ such that $\mathrm{sign}(\tilde{\nabla}(f, y, y_{t-1}; \boldsymbol{\theta})) \neq \mathrm{sign}(\Delta\phi(f, y, y_{t-1}; \boldsymbol{\theta}))$ for all $(f, y) \in FY$. This implies in turn that for sufficiently small δ_y we get $\tilde{\nabla}(\tilde{f}^*_{t+1}, y, y_{t-1}; \boldsymbol{\theta}) \Delta\phi(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta}) < 0$ for all $(f, y) \in FY$. Hence, by Assumption 1, $\exists \delta_y > 0$ such that

$$\int_{Y_{\delta_y}(y_t)} p(y|f_t, y_{t-1}) \tilde{\nabla}(\tilde{f}^*_{t+1}, y, y_{t-1}; \boldsymbol{\theta}) \Delta\phi(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta}) \mathrm{d}y < 0.$$

We thus conclude that an update that is not score equivalent is not RKL optimal regardless of p_t. By Assumption 1, the result extends immediately to CKL optimality. □

Proof of Proposition 3. As in the proof of Proposition 1, we require

$$\begin{aligned}
\Delta_{t|t} &= -\int_Y p(y|f_t, y_{t-1})\tilde{\nabla}(\tilde{f}_t^*, y, y_{t-1}; \boldsymbol{\theta})(\tilde{f}_{t+1} - \tilde{f}_t)\mathrm{d}y \\
&= -\int_Y p(y|f_t, y_{t-1})\tilde{\nabla}(y|\tilde{f}_t^*, y_{t-1})(\omega + \alpha S(\tilde{f}_t; \boldsymbol{\theta})\tilde{\nabla}(f_t, y, y_{t-1}; \boldsymbol{\theta}) + (\beta - 1)\tilde{f}_t)\mathrm{d}y \\
&= -\int_Y p(y|f_t, y_{t-1})\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta})(\omega + (\beta - 1)\tilde{f}_t)\mathrm{d}y \qquad (33) \\
&\quad -\int_Y p(y|f_t, y_{t-1})\alpha S(\tilde{f}_t; \boldsymbol{\theta})\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta})^2\mathrm{d}y + A(\delta_y, \delta_f) \ < \ 0
\end{aligned}$$

where $A(\delta_y, \delta_f)$ is an appropriate remainder term as in the proof of Proposition 1, which can be made arbitrarily small by the selecting small enough values for (δ_y, δ_f). The second term in (33) is surely strictly negative, whereas the first term may not be. As a result, for small enough (δ_y, δ_f) we obtain the desired inequality if

$$\alpha S(\tilde{f}_t; \boldsymbol{\theta})\tilde{\nabla}(y_t|\tilde{f}_t))^2 > |\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta})|\,|\omega + (\beta - 1)\tilde{f}_t| \quad \Leftrightarrow \quad \alpha > \frac{|\omega + (\beta - 1)\tilde{f}_t|}{S(\tilde{f}_t; \boldsymbol{\theta})|\tilde{\nabla}(\tilde{f}_t, y_t, y_{t-1}; \boldsymbol{\theta})|}.$$

The proof for local CKL optimality follows by a similar same argument and taking additional local expectations with respect to $\tilde{f}_{t+1}$ given $\tilde{f}_t$ and f_t. □

D Additional Results on the Stochastic Properties of the Filtered Sequence

Below we provide some details on the derivation of the moment bounds that have been omitted in the proof of Theorem 1.

Proof. The moment bounds $\sup_t \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1)|^{n_f} < \infty$ and $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|f_t(y^{1:t-1}, \boldsymbol{\theta})|^{n_f} < \infty$ are obtained by noting that

$$\sup_t \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1)|^{n_f} < \infty$$

if and only if

$$\sup_t(\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1)|^{n_f})^{1/n_f} = \sup_t \|f_t(\cdot, f_1)\|_{n_f}^{\Theta} < \infty,$$

and that, for any $f^{\Theta} \in \mathbb{C}(\Theta, \mathcal{F})$, having $\|f_t(\cdot, f_1) - f^{\Theta}\|_{n_f}^{\Theta} < \infty$ implies $\|f_t(\cdot, f_1^{\Theta})\|_{n_f}^{\Theta} < \infty$ since continuity on the compact Θ implies $\sup_{\boldsymbol{\theta}\in\Theta}|f(\boldsymbol{\theta})| < \infty$. Now, since we have shown

above that $\exists\ f^{\Theta} = \phi(y, f, f_*^{\Theta}, \cdot) \in \mathbb{C}(\Theta, \mathcal{F})$ satisfying $\|\phi(y_t, y_{t-1}, f^{\Theta}, \cdot)\|_{n_f}^{\Theta} \leq \bar{\bar{\phi}} < \infty$ and $\|f_1^{\Theta} - f^{\Theta}\|_{n_f}^{\Theta} = \|f_1^{\Theta} - \phi(f_*^{\Theta}, y, f, \cdot)\|_{n_f}^{\Theta} < \infty$, then

$$
\begin{aligned}
&\sup_t \|f_t(\cdot, f_1^{\Theta}) - f^{\Theta}\|_{n_f}^{\Theta} = \\
&\quad \sup_t \|\phi(y_t, y_{t-1}, f_t(\cdot, f_1^{\Theta}), \cdot) - \phi(y, f, f_*^{\Theta}, \cdot)\|_{n_f}^{\Theta} \\
&\quad \leq \sup_t \|\phi(y_t, y_{t-1}, f_t(\cdot, f_1^{\Theta}), \cdot) - \phi(y_t, y_{t-1}, f_*^{\Theta}, \cdot)\|_{n_f}^{\Theta} \\
&\qquad + \sup_t \|\phi(y_t, y_{t-1}, f_*^{\Theta}, \cdot)\|_{n_f}^{\Theta} + \sup_{\boldsymbol{\theta} \in \Theta} |\phi(y, f, f_*^{\Theta}, \cdot)| \\
&\quad \leq \sup_t \Big(\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1) - f^{\Theta}(\boldsymbol{\theta})|^{n_f} \\
&\qquad \times \sup_{\boldsymbol{\theta} \in \Theta} \frac{|\phi(y_t, y_{t-1}, f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1^{\Theta}), \boldsymbol{\theta}) - \phi(y_t, y_{t-1}, f_*^{\Theta}(\boldsymbol{\theta}), \boldsymbol{\theta})|^{n_f}}{|f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1^{\Theta}) - f^{\Theta}(\boldsymbol{\theta})|^{n_f}} \Big)^{1/n_f} \\
&\qquad + \sup_t \|\phi(y_t, y_{t-1}, f_*^{\Theta}, \cdot)\|_{n_f}^{\Theta} + \sup_{\boldsymbol{\theta} \in \Theta} |f^{\Theta}(\boldsymbol{\theta})| \\
&\quad \leq \sup_t \Big(\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1) - f^{\Theta}(\boldsymbol{\theta})|^{n_f} \\
&\qquad \times \sup_{\boldsymbol{\theta} \in \Theta} \sup_{(f^{\Theta}, f'^{\Theta}) \in \mathbb{C}(\Theta, \mathcal{F}) \times \mathbb{C}(\Theta, \mathcal{F}) : \|f^{\Theta} - f'^{\Theta}\|_{\Theta} > 0} \frac{|\phi(y_t, y_{t-1}, f^{\Theta}(\boldsymbol{\theta}), \boldsymbol{\theta}) - \phi(y_t, y_{t-1}, f'^{\Theta}(\boldsymbol{\theta}), \boldsymbol{\theta})|^{n_f}}{|f^{\Theta}(\boldsymbol{\theta}) - f'^{\Theta}(\boldsymbol{\theta})|^{n_f}} \Big)^{1/n_f} \\
&\qquad + \sup_t \|\phi(y_t, y_{t-1}, f_*^{\Theta}, \cdot)\|_{n_f}^{\Theta} + \sup_{\boldsymbol{\theta} \in \Theta} |f^{\Theta}(\boldsymbol{\theta})| \\
&\quad \leq \sup_t \Big(\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1) - f^{\Theta}(\boldsymbol{\theta})|^{n_f} \\
&\qquad \times \sup_{\boldsymbol{\theta} \in \Theta} \sup_{(f^{\Theta}, f'^{\Theta}) \in \mathbb{C}(\Theta, \mathcal{F}) \times \mathbb{C}(\Theta, \mathcal{F}) : \|f^{\Theta} - f'^{\Theta}\|_{\Theta} > 0} \frac{\bar{\phi}'_{t+1, n_f}(\boldsymbol{\theta}) |f^{\Theta}(\boldsymbol{\theta}) - f'^{\Theta}(\boldsymbol{\theta})|^{n_f}}{|f^{\Theta}(\boldsymbol{\theta}) - f'^{\Theta}(\boldsymbol{\theta})|^{n_f}} \Big)^{1/n_f} \\
&\qquad + \sup_t \|\phi(y_t, y_{t-1}, f_*^{\Theta}, \cdot)\|_{n_f}^{\Theta} + \sup_{\boldsymbol{\theta} \in \Theta} |f^{\Theta}(\boldsymbol{\theta})| \\
&\quad \leq \sup_t \Big(\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |f_t(y^{1:t-1}, \boldsymbol{\theta}, f_1) - f^{\Theta}(\boldsymbol{\theta})|^{n_f} \Big)^{1/n_f} \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \bar{\phi}'_{t+1, n_f}(\boldsymbol{\theta}) \\
&\qquad + \sup_t \|\phi(y_t, y_{t-1}, f_*^{\Theta}, \cdot)\|_{n_f}^{\Theta} + \sup_{\boldsymbol{\theta} \in \Theta} |f^{\Theta}(\boldsymbol{\theta})| \\
&\quad \leq \Big(\sup_t \|f_t(\cdot, x_1^{\Theta}) - f^{\Theta}\|_{n_f}^{\Theta} + \|f^{\Theta} - f^{\Theta}(\boldsymbol{\theta})(\cdot)\|_{n_f}^{\Theta} \Big) \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \bar{\phi}'(\boldsymbol{\theta}) + \bar{\bar{\phi}} + \bar{f},
\end{aligned}
$$

with $\bar{\bar{\phi}} < \infty$ and $\bar{f} < \infty$ and the thus yielding the recursion $\sup_t \|f_t(\cdot, f_1^{\Theta}) - f^{\Theta}\|_{n_f}^{\Theta} \leq \bar{c} \sup_t \|f_t(\cdot, f_1^{\Theta}) - f^{\Theta}\|_{n_f}^{\Theta} + A$, with $\bar{c} = \sup_{\boldsymbol{\theta} \in \Theta} \bar{\phi}'(\boldsymbol{\theta})$, $A = \bar{c} \sup_{\boldsymbol{\theta} \in \Theta} |f_t - f^{\Theta}(\boldsymbol{\theta})| + \bar{\bar{\phi}} + \bar{f} =$

$(\bar{c}+1)\bar{f}+\bar{\bar{\phi}}$, and hence,

$$\sup_t \|f_t(\cdot, f_1^\Theta) - f^\Theta\|_{n_f}^\Theta \leq \sum_{j=0}^{t} (\bar{c})^j ((\bar{c}+1)\bar{f}+\bar{\bar{\phi}}) + \bar{c}^{t+1} \sup_t \|f_1^\Theta - f^\Theta\|_{n_f}^\Theta$$
$$\leq \frac{(\bar{c}+1)\bar{f}+\bar{\bar{\phi}}}{1-\bar{c}} + \|f_1^\Theta - f^\Theta\|_{n_f}^\Theta < \infty.$$

□

The following proposition derives a subset of Θ over which the contractions on $\bar{\phi}_{t,k}(\boldsymbol{\theta})$ hold true. These contractions are used in the paper to establish the exponential fast convergence of the filter to an asymptotic SE sequence.

PROPOSITION 1. *Let $\bar{s}'_{t,k}(\lambda)$ denote the expected score supremum*

$$\bar{s}'_{t,k}(\lambda) := \mathbb{E} \sup_{f\in\mathcal{F}} \Big| \frac{\partial s(f, y_t, y_{t-1}, \boldsymbol{\theta})}{\partial f} \Big|.$$

If $\bar{s}'_{t,k}(\lambda) < \infty$ holds for every $\lambda \in \Lambda \subseteq \mathbb{R}$, then the contraction conditions

$$\mathbb{E} \sup_{(f,f')\in\mathcal{F}\times\mathcal{F}: f\neq f'} \frac{|\phi_t(f) - \phi_t(f')|}{|f-f'|} < 1$$

$$\textit{and} \quad \mathbb{E}\bar{\phi}_{t,k}(\boldsymbol{\theta}) < 1 = \mathbb{E} \sup_{f\in\mathcal{F}} \Big| \alpha \frac{\partial s(f, y_t, y_{t-1}, \lambda)}{\partial f} + \beta \Big|^k < 1$$

holds, for $k=1$, on the subset Θ^* of Θ given by,

$$\left\{ (\omega, \alpha, \beta, \lambda) \in \mathbb{R}\times\mathbb{R}\times(-1,1)\times\mathbb{R} \ : \ |\alpha| < \frac{1-|\beta|}{\bar{s}'_{t,1}(\lambda)} \right\}.$$

and for $k = n \geq 1$ on the subset Θ^{**} of Θ given by,

$$\left\{ (\omega, \alpha, \beta, \lambda) \in \mathbb{R}\times\mathbb{R}\times(-1,1)\times\mathbb{R} \ : \ \sum_{k=0}^{n} \binom{n}{k} |\alpha|^k \bar{s}'_{u,k}(\lambda) |\beta|^{n-k} < 1 \right\}.$$

Proof. Since $s \in \mathbb{C}^{(0,1,0)}(\mathcal{U}\times\mathcal{F}\times\Lambda)$ and $\mathcal{F}$ is convex, by the mean value theorem, $\forall\ (f,f') \in \mathcal{F}\times\mathcal{F}\ \exists\ f^* \in \mathcal{F}$ such that

$$\frac{|\phi(y_t, f, \boldsymbol{\theta}) - \phi(y_t, y_{t-1}, f', \boldsymbol{\theta})|}{|f-f'|} = \Big| \frac{\partial \phi(y_t, y_{t-1}, f^*, \boldsymbol{\theta})}{\partial f} \Big|$$

and hence for $\forall\ (k, f, f')\ \exists\ f^*$ such that

$$\frac{|\phi(y_t, f, \boldsymbol{\theta}) - \phi(y_t, f', \boldsymbol{\theta})|^k}{|f - f'|^k} = \Big|\frac{\partial \phi(y_t, f^*, \boldsymbol{\theta})}{\partial f}\Big|^k$$

and

$$\mathbb{E} \sup_{(f,f') \in \mathcal{F} \times \mathcal{F}} \frac{|\phi(y_t, f, \boldsymbol{\theta}) - \phi(y_t, y_{t-1}, f', \boldsymbol{\theta})|^k}{|f - f'|^k} \leq \mathbb{E} \sup_{f^* \in \mathcal{F}} \Big|\frac{\partial \phi(y_t, y_{t-1}, f^*, \boldsymbol{\theta})}{\partial f}\Big|^k.$$

As a result, condition (ii) in Proposition 1 stating that

$$\mathbb{E} \sup_{(f,f') \in \mathcal{F} \times \mathcal{F}} \frac{|\phi(y_t, y_{t-1}, f, \boldsymbol{\theta}) - \phi(y_t, y_{t-1}, f', \boldsymbol{\theta})|^k}{|f - f'|^k} \leq \bar{\phi}'_k(\boldsymbol{\theta}) < 1\ \forall\ \boldsymbol{\theta} \in \Theta \tag{34}$$

is trivially implied by having,

$$\mathbb{E} \sup_{f^* \in \mathcal{F}} \Big|\frac{\partial \phi(y_t, y_{t-1}, f^*, \boldsymbol{\theta})}{\partial f}\Big|^k = \mathbb{E} \sup_{f^* \in \mathcal{F}} \Big|\alpha \frac{\partial s(y_t, y_{t-1}, f^*; \lambda)}{\partial f} + \beta\Big|^k \leq \bar{s}_{t,k}(\lambda) < 1 \tag{35}$$

$\forall\ \boldsymbol{\theta} \in \Theta$, and which, for $k = 1$, is surely satisfied by every $\boldsymbol{\theta}$ in the set

$$\left\{(\omega, \alpha, \beta, \lambda) \in \mathbb{R} \times \mathbb{R} \times (-1, 1) \times \mathbb{R}\ :\ |\alpha| < \frac{1 - |\beta|}{\bar{s}''_{t,1}(\lambda)}\right\}.$$

Now for $k = n \geq 1$ we have by the Binomial theorem that,

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \Big|\alpha \frac{\partial s(y_t, y_{t-1}, f; \lambda)}{\partial f} + \beta\Big|^n &\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left(|\alpha| \Big|\frac{\partial s(y_t, y_{t-1}, f; \lambda)}{\partial f}\Big| + |\beta|\right)^n \\
&\leq \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k=0}^{n} \binom{n}{k} |\alpha|^k \Big|\frac{\partial s(y_t, y_{t-1}, f; \lambda)}{\partial f}\Big|^k |\beta|^{n-k} \\
&\leq \sum_{k=0}^{n} \binom{n}{k} |\alpha|^k \Big|\mathbb{E} \sup_{f \in \mathcal{F}} \frac{\partial s(y_t, y_{t-1}, f; \lambda)}{\partial f}\Big|^k |\beta|^{n-k} \\
&\leq \sum_{k=0}^{n} \binom{n}{k} |\alpha|^k \bar{s}'_{t,k}(\lambda) |\beta|^{n-k}
\end{aligned}$$

Hence, the inequality in (35) holds for every $\boldsymbol{\theta}$ such that

$$\sum_{k=0}^{n} \binom{n}{k} |\alpha|^k \bar{s}'_{t,k}(\lambda) |\beta|^{n-k} < 1.$$

□

E Technical Appendix

E.1 Supporting Results for Asymptotic Normality

For the asymptotic normality of the ML estimator we will use uniform laws of large numbers for the first two derivative of the likelihood, and a central limit theorem for the score. The first and second derivatives of the likelihood are given by,

$$\ell'_T(\boldsymbol{\theta}, \mathbf{f}_1^{0:1}) = -\frac{1}{T}\sum_{t=1}^{T}\Big(y_{t-1}h'_t\boldsymbol{f}_t^{(1)}p_{u,t}^{(1,0)} + p_{u,t}^{(0,1)}\Big)(p_{u,t})^{-1}$$

and

$$\begin{aligned}
\ell''_T(\boldsymbol{\theta}, \mathbf{f}_1^{0:2}) = -\frac{1}{T}\sum_{t=1}^{T}\Big(&y_{t-1}h''_t\boldsymbol{f}_t^{(1)}\boldsymbol{f}_t^{(1)}p_{u,t}^{(1,0)} + y_{t-1}h'_t\boldsymbol{f}_t^{(2)}p_{u,t}^{(1,0)} - \big(y_{t-1}h'_t\boldsymbol{f}_t^{(1)}\big)^2p_{u,t}^{(2,0)} \\
&+ y_{t-1}h'_t\boldsymbol{f}_t^{(1)}p_{u,t}^{(1,1)} - y_{t-1}h'_t\boldsymbol{f}_t^{(1)}p_{u,t}^{(1,1)} + p_{u,t}^{(0,2)}\Big)(p_{u,t})^{-1} \\
&- \Big(y_{t-1}h'_t\boldsymbol{f}_t^{(1)}p_{u,t}^{(1,0)} + p_{u,t}^{(0,1)}\Big)(p_{u,t})^{-2}\Big(y_{t-1}h'_t\boldsymbol{f}_t^{(1)}p_{u,t}^{(1,0)} + p_{u,t}^{(0,1)}\Big) \\
= -\frac{1}{T}\sum_{t=1}^{T}\Bigg[&\Big(y_{t-1}h''_t\boldsymbol{f}_t^{(1)}\boldsymbol{f}_t^{(1)}p_{u,t}^{(1,0)} + y_{t-1}h'_t\boldsymbol{f}_t^{(2)}p_{u,t}^{(1,0)} - \big(y_{t-1}h'_t\boldsymbol{f}_t^{(1)}\big)^2p_{u,t}^{(2,0)} + p_{u,t}^{(0,2)}\Big)(p_{u,t})^{-1} \\
&-\Big(y_{t-1}h'_t\boldsymbol{f}_t^{(1)}p_{u,t}^{(1,0)} + p_{u,t}^{(0,1)}\Big)(p_{u,t})^{-2}\Big(y_{t-1}h'_t\boldsymbol{f}_t^{(1)}p_{u,t}^{(1,0)} + p_{u,t}^{(0,1)}\Big)\Bigg]
\end{aligned}$$

where $h'_t = h(f_t)$, $p_{u,t} = p_u(u_t;\boldsymbol{\theta}) = p_u(y_t - h(f_t)y_{t-1};\boldsymbol{\theta})$, $p_{u,t}^{(1,0)} = \partial p_u(u_t;\boldsymbol{\theta})/\partial u_t$, $p_{u,t}^{(0,1)} = \partial p_u(u_t;\boldsymbol{\theta})/\partial\lambda$, $p_{u,t}^{(2,0)} = \partial^2 p_u(u_t;\boldsymbol{\theta})/\partial u_t^2$, $p_{u,t}^{(0,2)} = \partial^2 p_u(u_t;\boldsymbol{\theta})/\partial u_t^2$, $p_{u,t}^{(1,1)} = \partial^2 p_u(u_t;\boldsymbol{\theta})/\partial u_t\partial\lambda$, and where $\boldsymbol{f}_t^{(0:2)} = (f_t, \boldsymbol{f}_t^{(1)}, \boldsymbol{f}_t^{(2)})$ and $\boldsymbol{f}_t^{(0:1)} = (f_t, \boldsymbol{f}_t^{(1)})$ are collections of partial derivatives $\boldsymbol{f}_t^{(1)} = \partial f_t(y^{t-1}, \boldsymbol{\theta}, f_1)/\partial\boldsymbol{\theta}$ and $\boldsymbol{f}_t^{(2)} = \partial^2 f_t(y^{t-1}, \boldsymbol{\theta}, f_1)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$.

In the Appendix, we show that the ith-derivative process $\{\boldsymbol{f}_t^{(i)}(y^{t-1}, \boldsymbol{\theta}, \boldsymbol{f}_1^{(0:i)})\}$ of $f_t(y^{t-1}, \boldsymbol{\theta}, f_1)$ with respect to $\boldsymbol{\theta}$ satisfies the stochastic recurrence equation

$$\boldsymbol{f}_t^{(i)}(y^{t-1}, \boldsymbol{\theta}, \boldsymbol{f}_1^{(0:i)}) = A_{i,t}(\boldsymbol{\theta}) + B_{i,t}(\boldsymbol{\theta})\boldsymbol{f}_t^{(i-1)}(y^{t-1}, \boldsymbol{\theta}, \boldsymbol{f}_1^{(0:i)}) \quad \forall\ (i, \boldsymbol{\theta}, T), \tag{36}$$

where $A_{i,t}(\boldsymbol{\theta})$ and $B_{i,t}(\boldsymbol{\theta})$ are functions of $\boldsymbol{f}t^{(0:i-1)}$, h and p_u. Equation (36) allows us to establish the SE properties of $\{f_t^{(i)}(y^{t-1}, \boldsymbol{\theta}, \boldsymbol{f}_1^{(0:i)})\}$ by studying the properties of $\{A_{i,t}(\boldsymbol{\theta})\}$ and $\{B_{i,t}(\boldsymbol{\theta})\}$.

For notational simplicity, we define the random derivative $s^{(1,0)}(f^*;\lambda) = \partial s(f^*, y_t;\lambda)/\partial f$,

and its kth power supremum as

$$r_t^k(\boldsymbol{\theta}) = \sup_{f^* \in f^*} |\beta + \alpha\, s^{(1,0)}(f^*; \lambda)|^k,$$

where $\mathcal{F} \subseteq f^* \subset \mathbb{R}$. Theorem 2 also uses the definition

$$s^{(\mathbf{k})}(f, y_t; \lambda) = \partial^{k_1+k_2+k_3} s(f, y_t; \lambda)/(\partial f^{k_1} \partial y_t^{k_2} \partial \lambda^{k_3}),$$

with $\mathbf{k} = (k_1, k_2, k_3)$. We also adopt the convention that $n_s^f := n_s^{(1,0,0)}$, $n_s^{ff} := n_s^{(2,0,0)}$, $n_s^\lambda := n_s^{(0,0,1)}$, $n_s^{\lambda\lambda} := n_s^{(0,0,2)}$ and $n_s^{f\lambda} := n_s^{(1,0,1)}$, and define $n_f^{(1)} = \min\left\{n_f, n_s, n_s^\lambda\right\}$ and

$$n_f^{(2)} = \min\left\{n_f^{(1)}\ ,\ n_s^\lambda\ ,\ n_s^{\lambda\lambda}\ ,\ \frac{n_s^f n_f^{(1)}}{n_s^f + n_f^{(1)}}\ ,\ \frac{n_s^{ff} n_f^{(1)}}{2n_s^{ff} + n_f^{(1)}}\ ,\ \frac{n_s^{f\lambda} n_f^{(1)}}{n_s^{f\lambda} + n_f^{(1)}}\right\}.$$

Using this notation, we have the following proposition.

PROPOSITION 2. *Let $\Theta^* \subset \mathbb{R}^{3+n_\lambda}$ be compact and $\{y_t\}_{T\in\mathbb{Z}}$ be an SE sequence satisfying $\mathbb{E}|y_t|^{n_y} < \infty$ for some $n_y > 0$. Suppose that $s \in \mathbb{C}^{(2,0,2)}(\mathcal{F}\times\mathcal{Y}\times\Lambda^*)$ and $s^{(\mathbf{k})} \in \mathbb{M}_{\Theta^*,\Theta^*}(\mathbf{n}, n_s^{(\mathbf{k})})$, with $\Lambda^* = \Theta^* \cap \Lambda$, and $\mathbf{n} := (n_f, n_y)$. Assume $n_f^{(1)} > 0$, $n_f^{(2)} > 0$, and $\exists\ f \in \mathcal{F}$ such that*

(i) $\mathbb{E}\ln^+ \sup_{\lambda\in\Lambda^*} |s(f, y_t; \lambda)| < \infty$;

(ii) $\mathbb{E}\ln \sup_{\boldsymbol{\theta}\in\Theta^*} r_1^1(\boldsymbol{\theta}) < 0$.

Then, for $i = 0, 1, 2$, there exists a unique SE sequence $\{f_t^{(i)}(y^{t-1}, \boldsymbol{\theta})\}_{T\in\mathbb{Z}}$, such that

$$\sup_{\boldsymbol{\theta}\in\Theta^*} \|f_t^{(i)}(y^{t-1}, \boldsymbol{\theta}, \boldsymbol{f}_1^{(0:i)}) - f_t^{(i)}(y^{t-1}, \boldsymbol{\theta})\|^r \overset{e.a.s.}{\to} 0 \quad \text{as } T \to \infty.$$

If furthermore $\exists\ n_f \geq 1$ such that $n_f d \geq 1$, $n_{ff} \geq 1$ and

(iii) $\|s(f, y_t; \lambda)\|_{\Lambda^*}^{n_f} < \infty$;

(iv) $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta^*} r n_f(\boldsymbol{\theta}) < 1$;

(v) $f_t(y^{t-1}, \boldsymbol{\theta}, f_1) \perp r^{n_f}(\boldsymbol{\theta})\ \forall\ (T, \boldsymbol{\theta}, f_1)$.

Then $\sup_T \|f_t(y^{t-1}, \boldsymbol{\theta}, f_1)\|_{Theta^*}^{n_f} < \infty$, $\sup_T \|f_t(y^{t-1}, f_1)\|_{Theta^*}^{n_f} < \infty$,
$\sup_T \|f_t^{(i)}(y^{t-1}, \boldsymbol{\theta}, \boldsymbol{f}_1^{(0:i)})\|_{\Theta^*}^{n_f^{(i)}} < \infty$, *and* $\|f_t^{(i)}(y^{t-1}, \boldsymbol{\theta})\|_{\Theta^*}^{n_f^{(i)}} < \infty$ *for* $i = 1, 2$.

We have suppressed the dependence of s on λ in the moment preserving properties by defining $\mathbf{n} := (n_f, n_y)$. We can do so without loss of generality, as we have assumed λ to be non-stochastic such that all moments of λ exist.

We can simplify the moment results in Proposition 2 substantially by writing the moments $n_f d$ and n_{ff} for the first and second derivative processes in terms of a common minimum moment bound that holds for all derivatives of s. We state this as a separate remark.

REMARK 2. Let the assumptions of Proposition 2 hold with $m = \min\{n_s^{(i,0,j)} : (i,j) \in \mathbb{N}^2,\ i + j \leq 2\}$, then the moment bounds on the derivative processes hold with $n_f d = m$ and $n_{ff} = m/3$.

The bound in expectation in *(ii)* is sometimes difficult to handle. Remark 3 states that we can ensure condition *(ii)* if we assume a uniform bound on the derivative of the score process.

REMARK 3. If $\sup_{(f^*, y_t; \lambda) \in f^* \times \mathcal{Y} \times \Theta^*} |\beta + \alpha\, \partial s(f^*, y_t; \lambda)/\partial f| < 1$, we can drop conditions *(iv)* and *(v)* in Proposition 2.

We also find that conditions *(iii)* and *(iv)* imply *(i)* and *(ii)*, respectively.

E.2 Application to Industrial Production

Nonlinear models are known to perform better than linear models in explaining industrial production data; see e.g. Teräsvirta, Tjøstheim, and Granger (2010). The relative better fit of nonlinear models can be explained by a number of economic factors and empirical regularities that have been the focus of study of economists since very early on. Indeed, as pointed out by Granger and Terasvirta (1993, Chapters 8 and 9), already in the 30s Keynes (1936, p.314) argued that economic contractions were shorter and more violent than economic expansions. Burns and Mitchell (1946) took this as an empirical fact that linear models cannot reproduce. Nonlinear autoregressive models such as TAR and STAR can however reproduce these empirical regularities.

We adopt Model II to analyze the growth rate of US seasonally adjusted industrial production index (2007=100) spanning from 1919 to 2013. Table **??** shows that Model II outperforms both its linear and nonlinear competitors in log likelihood and AICc fit as we all the one-step-ahead forecast RMSE.

Industrial Production: Model Comparison

	Model II	TAR	STAR	AR(3)
Log Lik	3025.94	3020.07	3020.50	3020.84
AICc	-6041.83	-6030.09	-6030.95	-6031.62
F-RMSE	0.560	0.564	0.563	0.880

Table 3: Log-likelihood, Akaike's information criterion with finite sample correction and root mean squared errors for 1, 2 and 3 step-ahead forecasts of the growth rate of US monthly seasonally adjusted industrial production index (2007=100) reported by the Federal Reserve Bank of St. Louis.

References

Anderson, P. M. and B. D. Meyer (1997). Unemployment insurance takeup rates and the after-tax value of benefits. *The Quarterly Journal of Economics 112*(3), 913–37.

Anderson, P. M. and B. D. Meyer (2000). The effects of the unemployment insurance payroll tax on wages, employment, claims and denials. *Journal of Public Economics 78*(1-2), 81–106.

Ashenfelter, O., D. Ashmore, and O. Deschenes (2005). Do unemployment insurance recipients actively seek work? evidence from randomized trials in four us states. *Journal of Econometrics 125*, 53–75.

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

Blasques, F., S. J. Koopman, and A. Lucas (2014a). Information theoretic optimality of observation driven time series models. *Discussion Paper, Tinbergen Institute* (14-046/III).

Blasques, F., S. J. Koopman, and A. Lucas (2014b). Maximum likelihood estimation for generalized autoregressive score models. *Discussion Paper, Tinbergen Institute* (14-029/III).

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

Burns, A. F. and W. C. Mitchell (1946). *Measuring Business Cycles*. Number burn46-1 in NBER Books. National Bureau of Economic Research, Inc.

Chan, K. S. and H. Tong (1986). On Estimating Thresholds in Autoregressive Models. *Journal of Time Series Analysis 7*(3), 179–190.

Clark, T. E. and M. W. McCracken (2010). Averaging forecasts from VARs with uncertain instabilities. *Journal of Applied Econometrics 25*, 5–29.

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

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

Delle Monache, D. and I. Petrella (2014). Adaptive models and heavy tails. Technical report, Working Paper, Birkbeck University.

Doan, T., R. B. Litterman, and C. A. Sims (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews 3*, 1–144.

Gallant, R. and H. White (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Cambridge University Press.

Gavin, W. T. and K. L. Kliesen (2002). Unemployment insurance claims and economic activity. *Review* (May), 15–28.

Granger, C. W. J. and T. Terasvirta (1993, Decembrie). *Modelling Non-Linear Economic Relationships*. Number 9780198773207 in OUP Catalogue. Oxford University Press.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails*. Cambridge University Press.

Hopenhayn, H. A. and J. P. Nicolini (1997). Optimal unemployment insurance. *Journal of Political Economy 105*(2), 412–38.

Hurvich, C. M. and C. Tsai (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika 78*(3), 499–509.

Kadiyala, K. R. and S. Karlsson (1993). Forecasting with generalized Bayesian vector autoregressions. *Journal of Forecasting 12*, 365–378.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. London: Macmillan.

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

McMurrer, D. and A. Chasanov (1995). Trends in unemployment insurance benefits. *Monthly Labor Review 118*(9), 30–39.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics 13*(2), 151–61.

Rao, R. R. (1962). Relations between Weak and Uniform Convergence of Measures with Applications. *The Annals of Mathematical Statistics 33*(2), 659–680.

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

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association 89*, 208–218.

Teräsvirta, T., D. Tjostheim, and C. W. J. Granger (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press.

Teräsvirta, T., D. Tjøstheim, and C. W. J. Granger (2010). *Modelling Nonlinear Economic Time Series*. Oxford: Oxford University Press.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*. New York: Springer-Verlag.

Ullah, A. (1996). Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference 69*, 137–162.

Ullah, A. (2002). Uses of entropy and divergence measures for evaluating econometric approximations and inference. *Journal of Econometrics 107*(1-2), 313–326.

van der Vaart, A. W. (2000). *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge Books. Cambridge University Press.