UNIVERSITÀ DEGLI STUDI DI PADOVA

VRIJE UNIVERSITEIT AMSTERDAM

Sede Amministrativa: Università degli Studi di Padova
Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIX

Vrije Universiteit Amsterdam
Faculty of Economics and Business Administration
Department of Econometrics and Operations Research

On observation-driven time series modeling

Direttore della Scuola: Prof.ssa Monica Chiogna

Supervisori: Prof.ssa Luisa Bisaglia and Prof. Siem Jan Koopman

Cosupervisore: Prof. Francisco Blasques

Dottorando: Paolo Gorgi

20 Gennaio 2017

Acknowledgements

First of all, I would like to thank my supervisors Luisa Bisaglia, Francisco Blasques and Siem Jan Koopman. In particular, Luisa for supporting me throughout these three years and also for pushing me to start the PhD after my master graduation. Without her, I would not have undertaken this PhD. Francisco for his support and guidance during these years. His help and thoughtful advices guided me through all the difficult moments I encountered during the completion of this thesis. Siem Jan for his support in these years and the fruitful meetings we had. I am also very grateful to him for giving me the opportunity to spend two years of my PhD at the VU University in Amsterdam.

I also wish to thank the director of the PhD school in Padua, Monica Chiogna, that, together with Siem Jan, made possible the double degree program between Padua and Amsterdam.

I thank my colleagues and the administrative staff at the department of Statistical Sciences in Padua. In particular, my colleagues of the XXIX cycle Andrea, Claudia, Davide, Elisa, Lucia, Khanh and Mirko for the time we spent together during the first year of our PhD.

I thank my colleagues and the people I met at the VU University in Amsterdam. In particular, my PhD colleagues Agnieszka, Artem, Marc, Leopoldo and Lorenzo for the time we spent together and the discussions about research we had.

A special thank to my family and my extended family. In particular, my sisters Elena and Mary, my brothers Andrea, Giovanni and Marco, my father Adriano and my uncle Francesco for their support.

Finally, I wish to thank all my friends from outside academia for the happy moments we had together in these years.

Amsterdam, August 19, 2016. *Paolo Gorgi*

Summary

This thesis addresses different aspects of observation-driven time series modeling. The main contributions concern the reliability of likelihood-based inference and the specification of dynamic models to capture complex behaviors observed in time series data.

As concerns inference, the main focus of the thesis is on invertibility conditions for observation-driven time series models. Invertibility plays a key role in ensuring the consistency of likelihood-based estimators. However, the invertibilty conditions typically employed in the literature are often unfeasible to be checked. Therefore, the reliability of inference fails to be guaranteed in practice. This thesis contributes to the literature by deriving feasible conditions that ensure the consistency of the maximum likelihood estimator for a wide class of models. One of the most appealing features of our consistency results is that they hold for both correctly specified and misspecified models. Several empirical examples covering different observation-driven models are presented. These examples highlight the practical relevance of the theoretical results.

As concerns model specification, we cover two lines of research. The first is related to integer-valued time series data. We propose an extension to the class of Integer-valued Autoregressive models that allows the survival probability to vary over time. We show how our model can be easily estimated by maximum likelihood and we prove the consistency of the estimator. The flexibility of the proposed approach is shown through a simulation experiment and an application to a real time series of crime reports. Finally, the second line of research on model specification is an extension of the Generalized Autoregressive Score framework. We propose a class of models that updates time-varying parameters at different speeds in different time periods. The new updating equation can be employed to describe time series where the amount of information contained in the data is changing over time. This peculiarity is highlighted through a simulation study and we provide theoretical foundations for the proposed approach. Furthermore, two empirical applications to S&P 500 stock returns and US inflation illustrate how our method can be useful in practice.

Summary in Italian

Questa tesi tratta diversi aspetti della modellazione di serie storiche attraverso modelli *observation-driven*. I principali contributi della tesi riguardano l'inferenza basata sulla verosimiglianza e la specificazione di modelli per serie storiche con comportamenti dinamici complessi.

Per quanto riguarda l'inferenza, la tesi si focalizza su condizioni di invertibilità per modelli *observation-driven*. Assicurare invertibilità è importante per poter assicurare la consistenza degli stimatori di massima verosimiglianza. Le condizioni di invertibilità tipicamente considerate in letteratura non sono testabili in situazioni pratiche. Il nostro contributo consiste nella derivazione di condizioni di invertibilità testabili che garantiscono la consistenza dello stimatore per un'ampia classe di modelli. Una delle principali caratteristiche dei nostri risultati è che sono applicabili sia a modelli correttamente specificati che a modelli non correttamente specificati. Diversi esempi empirici sono presentati che illustrano la rilevanza pratica dei nostri risultati teorici.

Per quanto riguarda la specificazione di modelli, due linee di ricerca sono state considerate. La prima riguarda serie storiche a valori interi. Proponiamo un estensione dei modelli *Integer-valued Autoregressive* che consente alla probabilità di sopravvivenza di variare nel tempo. Mostriamo come questi modelli siano facilmente stimabili attraverso lo stimatore di massima verosimiglianza per il quale viene anche dimostrata la consistenza. La flessibilità dell'approccio considerato è mostrata attraveso uno studio di simulazione e un applicazione a una serie storica reale sul crimine. Infine, il secondo ramo di ricerca sulla specificazione è un estensione dei modelli *Generalized Autoregressive Score*. La specificazione che proponiamo consente la variazione della velocità di aggiornamento del parametro dinamico in diversi istanti temporali. Questo nuovo sistema di aggiornamento è in grado di descrivere situazioni dove l'informazione contenuta nei dati cambia nel tempo. Questa peculiarità è illustrata attraverso uno studio di simulazione e il sistema di aggiornameto proposto è giustificato da alcune proprietà di ottimalità. Inoltre, due applicazioni empiriche sui rendimenti azionari dell'indice S&P 500 e l'inflazione degli Stati Uniti illustrano come l'approccio presentato possa essere utile nella pratica.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Time series data are encountered in most fields of empirical science as phenomena are typically observed sequentially over time. Examples range from the number of sunspots, or the water flow of a river in natural sciences to the number of inhabitants of a city, or the returns of a financial index in social sciences. The main assumption behind time series analysis is that past observations contain information about future observations. The idea is therefore to exploit this information and obtain more accurate predictions of future outcomes. Statistical modeling plays a key role in time series analysis as it summarizes the relevant information in the data and provides a probabilistic representation of the phenomenon of interest.

Statistical modeling of time series data has a long history. The first applications of autoregressive models go back to Yule (1927). Box and Jenkins (1970) provided a unified approach to specification, estimation, diagnostic checking and forecasting of Integrated Autoregressive Moving Average (ARIMA) models. ARIMA models represent a milestone for time series modeling and their main justification rests on the Wold decomposition theorem (Wold, 1938). Several extensions of the ARIMA framework have been proposed over the years. Examples include the vector autoregressive model, Sims (1980), and the cointegration analysis of Engle and Granger (1987). A limitation of the ARIMA approach and its extensions is that they describe the linear dependence in the data but they do not explicitly take into account possible nonlinearities. This may be too restrictive in some situations of practical interest. For this reason, in the late 70s researchers started focusing on nonlinear time series models. One of the firsts to consider a nonlinear model was Tong (1978), introducing the class of Threshold Autoregressive (TAR) models. TAR

models allow the conditional mean of the process to depend on past observations in a nonlinear fashion. Nonlinear models come in different forms and shapes as nonlinearities can be introduced in different ways. A typical approach that produces nonlinear specifications is to allow time variation in some features of the probability distribution of interest, i.e. some parameters. A well know example is to have time dependence in the variance of the observations. Popular models with dynamic variance are the Generalized Autoregressive Conditional Heteroscedastic (GARCH) model of Engle (1982) and Bollerslev (1986) and the Stochastic Volatility (SV) model of Taylor (1986). These models have been successfully employed in Econometrics and Finance to describe the well known volatility clustering often observed in financial asset returns.

Most time-varying parameter models can be classified in two categories: observation-driven and parameter-driven models (Cox, 1981). In observation-driven models, the parameter of interest is made time-varying considering a stochastic processes where the source of randomness comes from past observations. Whereas, in parameter-driven models, the time-varying parameter is specified as a stochastic process with its own source of error. In the context of volatility models, the GARCH model is an example of observation-driven model as the source of randomness is provided by past squared observations. On the other hand, the SV model is an example of parameter-driven model as the dynamic variance is driven by a latent autoregressive process. In most situations, as also in the case of the GARCH and the SV model, these two classes of models play equivalent roles. Their goal is to enable some features of the distribution of the variable of interest to change over time and, in this way, capture some form of dependence in the data. However, their statistical properties are quite different. Observation-driven models have the great advantage that they can be easily estimated since the likelihood function is available in closed form through a prediction error decomposition. Therefore, only standard optimization methods are needed to perform likelihood-based inference. Instead, in parameter-driven models, the likelihood function is usually not in closed form as it contains integrals with no analytical solutions. Therefore, estimation is much more challenging from a computational point of view and time-consuming simulation-based methods are usually required. Some rare exceptions with close form solutions exist, see for example the Markov Switching models where the Hamilton filter can be employed (Hamilton, 1989).

In parameter-driven models, the time-varying parameter is typically specified as an autoregressive process where the innovation is an independently and identically distributed (i.i.d.) sequence of Gaussian random variables. On the other hand, the specification of observation-driven models is often based on intuition. For instance, to make the variance time-varying it makes sense to consider a linear combination of squared past observations;

this leads to the well known GARCH model. However, sometimes it is not clear which function of the past observations to use and an intuitive choice may not always be the best option. Creal et al. (2013) and Harvey (2013) proposed an updating equation where the innovation is given by the score of the conditional distribution of the observations. This approach provides a general framework to specify the time-varying parameter in an observation-driven setting. The resulting class of models is known as Generalized Autoregressive Score (GAS) models. Besides being intuitive, the use of the score as driving mechanism to update time-varying parameters is also justified by an optimality reasoning (Blasques et al., 2015). Since its introduction, the GAS framework has been successfully employed to develop dynamic models in econometrics and time series analysis, see for instance Salvatierra and Patton (2015), Harvey and Luati (2014) and Creal et al. (2011). It also turns out that many existing observation-driven models are in fact GAS models. Examples include the GARCH model and, in the context of integer-valued time series, the Poisson autoregressive model of Davis et al. (2003). For a more detailed discussion see Creal et al. (2013).

In this thesis, we address different aspects of observation-driven time series modeling including model specification and statistical inference. These two aspects are particularly relevant from a practical perspective as an appropriate specification of the model and a reliable inferential procedure are two of the main ingredients to obtain an accurate probabilistic representation of the time series of interest. The focus of the thesis is mostly on score-driven models though general results for observation-driven models are considered in Chapter 2. In particular, the second chapter of the thesis is concerned with model estimation of observation-driven models, whereas the third and fourth chapters are concerned with model specification in the setting of score-driven models. The 3 main chapters of the thesis are self contained and they can be read separately. In the following, we provide a brief outline for each chapter of the thesis. More detailed outlines can be found at the beginning of each chapter.

The first line of research, Chapter 2, concerns the consistency of likelihood-based inference for observation-driven models. One of the key steps to ensure the reliability of the Maximum Likelihood (ML) estimator is the study of the asymptotic behavior of the filtered time-varying parameter, i.e. the time-varying parameter recovered using the observed data. In the context of Quasi Maximum Likelihood (QML) estimation of GARCH-type models, Straumann and Mikosch (2006) proposed to rely on Theorem 3.1 of Bougerol (1993) to ensure the asymptotic stability of the filtered parameter, which is known as invertibility. Compared to previous research, their approach allows us to handle nonlinearities in the recursion of the filtered parameter. However, the required invert-

ibility conditions often impose restrictions on the parameter space that are unfeasible to be checked in practice. This occurs because these invertibility conditions depend on the properties of the Data Generating Process (DGP) that are unknown. Wintenberger (2013) noted this problem for the EGARCH model of Nelson (1991) and proposed to replace the unfeasible conditions with a feasible empirical invertibility condition. This method delivers a consistent QML estimator for the EGARCH model. We note that this problem is not a peculiarity of the EGARCH model but a general problem for observation-driven models with nonlinearities in the filtered parameter recursion. Therefore, often, the asymptotic theory can be ensured only for either degenerate or very small parameter regions that are unrealistic in empirical applications. As examples, we consider the Beta-t-GARCH model of Harvey (2013) and Creal et al. (2013), the location model of Harvey and Luati (2014) and the autoregressive model of Blasques et al. (2014b) and Delle Monache and Petrella (2016). We build on the work of Wintenberger (2013) and deliver a general theory for observation-driven models that ensures the consistency of the ML estimator under feasible invertibility conditions. The resulting theory is shown to cover applications of practical interest such as modeling of financial stock returns and macroeconomic variables. An appealing feature of our theoretical results is that they hold also in the case of model misspecification. In this situation, the consistency is proved with respect to a pseudo-true parameter.

The second line of research, Chapter 3, concerns integer-valued time series modeling. Over the last few years, there has been an increasing interest in modeling time series with non-continuous response variables. This due to the fact that many observed variables take values in a discrete support and models for continuous variables are not suited in these situations. One of the most popular class of models for count time series data is the class of Integer-valued Autoregressive (INAR) models introduced by Al-Osh and Alzaid (1987) and McKenzie (1988). INAR models can be seen as a discrete version of the continuous response AR models as they share several common properties. An appealing feature of INAR models is their interpretation as birth-death processes: at each time period the count is given by the sum between the number of new born elements and the number of elements surviving from the previous period. Assuming a constant survival probability can be too restrictive in many situations as real time series often exhibit changes in their behavior over time. Therefore, allowing different persistence levels in different time periods can be useful to better describe the observed variable and enhance the forecasting performance of the model. We propose a novel dynamic specification for the surviving probability. The peculiarity of our approach is to consider an observation-driven dynamic for the surviving probability based on the GAS framework. The resulting class of mod-

els is appealing from several prospectives. First, the proposed dynamic coefficient is very effective in capturing smooth changes in the survival probability. We illustrate this through a simulation study designed in a misspecified setting where the survival probability follows different deterministic paths. Second, the estimation of the model can be easily performed by maximum likelihood using standard optimization algorithms as for the classic INAR model. Finally, the proposed class of models allows us to consider general distributions for the new born process without additional difficulties in the derivation of the model specification and estimation. One of the main contributions of this chapter is the study of some statistical properties of the proposed model. In particular, we show the consistency of ML estimation for the static parameters and for the predictive probability mass function. Furthermore, we also provide an empirical application to a crime time series to illustrate how our class of models can be useful in practice.

The third and last line of research, Chapter 4, concerns model specification in the framework of GAS models. As mentioned before, in the GAS framework, the time-varying parameter is specified as an autoregressive process where the innovation is given by the score of the predictive likelihood. As discussed in Blasques et al. (2015), the GAS updating mechanism can be seen as a sort of Newton Raphson algorithm where the score provides the direction of the updating step. We propose to allow the magnitude of the updating step to be time-varying. The idea behind having time variation in the size of the step is related to the amount of local information in the data. In some time periods, the most recent observations can be very informative to predict future observations, whereas, in other periods, this may not be the case. Therefore, in such situations, we would like the time-varying parameter to be updated quickly when the data is informative and slowly when the data is not informative. The specification we introduce to capture time variation in the magnitude of the GAS updating step is given by a weighted autocorrelation of past GAS innovations. This has an intuitive interpretation: the amount of local information in the data is determined by the dependence of past score innovations. We perform a simulation study as an illustrative example of this idea and show the benefits that our approach can provide. Furthermore, in the spirit of Blasques et al. (2015), we derive an optimality justification for the proposed method in terms of Kullback-Leibler (KL) divergence reduction between the true and unknown conditional distribution and the postulated statistical model. Finally, some empirical examples considering volatility and location models are presented. In particular, in the context of volatility models, we derive an extension of the GARCH model and perform an empirical study using the stock returns of the S&P 500 financial index. Whereas, in the context of location models, we specify a fat tailed model and illustrate an empirical application to the US consumer price inflation series. Overall,

the empirical results show promising results both in-sample and out-of-sample.

1.2 Main contributions of the thesis

In the following, we summarize the main original contributions of the thesis.

Chapter 2.

1. Feasible consistency conditions for ML estimation of a wide class of observation-driven time series models are derived.

2. The consistency of the ML estimator for the Beta-t-GARCH model is proved under a testable invertibility condition.

3. The theory developed in the chapter is shown to be useful also outside the framework of GARCH-type models. This is done by means of two examples in the context of location models.

Chapter 3.

1. A new class of observation-driven INAR models with dynamic survival probability is introduced. Estimation and forecasting procedures of the proposed class of models is presented.

2. The consistency of ML estimation of the static parameter vector and the conditional probability mass function is proved.

3. The flexibility of the proposed class of models is illustrated through a simulation study. Furthermore, an empirical application to a crime time series is provided.

Chapter 4.

1. An extension of the GAS framework is proposed. This extension introduces time variation in the updating equation of score-driven models.

2. An optimality argument that justifies the proposed specification of the time-varying parameter update is derived.

3. Empirical illustrations in economics and finance to show how the proposed approach can be useful in practice are presented. More specifically, applications to financial stock returns and the US inflation series are considered.

Chapter 2

Feasible Invertibility Conditions for Maximum Likelihood Estimation of Observation-Driven Models

2.1 Introduction

Observation-driven models are widely employed in time series analysis and econometrics. These models feature time-varying parameters that are specified through a Stochastic Recurrence Equation (SRE) driven by past observed elements of the time series. A well know example of observation-driven models is the class of GARCH-type models. Observation-driven models are widely used also outside the context of volatility models; see for instance the Dynamic Conditional Correlation (DCC) model of Engle (2002), the time-varying quantile model of Engle and Manganelli (2004), the dynamic copula models of Patton (2006), the score models of Creal et al. (2013) and the time-varying location model of Harvey and Luati (2014).

The asymptotic theory of the QML estimator for GARCH-type models has attracted much attention. Lumsdaine (1996) and Lee and Hansen (1994) obtained the consistency and asymptotic normality of the QML estimator for the GARCH(1,1). Berkes et al. (2003) generalized their results to the GARCH(p,q). Among others, Francq and Zakoian (2004) and Robinson and Zaffaroni (2006) weakened the conditions for consistency and asymptotic normality and extended the results to a larger class of models. Straumann and Mikosch (2006) provided a very general approach to handle nonlinearities in the variance recursion. Their theory relies on the work of Bougerol (1993) to ensure the invertibility of the filtered time-varying variance and delivers asymptotic results that are subject to

some restrictions on the parameter region where the QML estimator is defined. The severity of these restrictions typically depends on the degree of nonlinearity in the recurrence equation.

We note that, in practical applications, the invertibility conditions of Straumann and Mikosch (2006) often fail to be guaranteed. We will illustrate this issue through some empirical examples featuring the Beta-t-GARCH model of Harvey (2013) and Creal et al. (2013), the autoregressive model with dynamic coefficient of Blasques et al. (2014b) and Delle Monache and Petrella (2016) and the fat-tailed location model of Harvey and Luati (2014). The main problem lies on the fact that these conditions are empirically unfeasible as they depend on the unknown DGP. This leads researchers to rely on feasible conditions that are typically only satisfied in either degenerate or very small parameter regions that are too restrictive for practical situations. To handle this issue and ensure the asymptotic theory of the QML estimator of the EGARCH(1,1) model of Nelson (1991), Wintenberger (2013) proposed to stabilize the inferential procedure by restricting the optimization of the quasi-likelihood function to a parameter region that satisfies an empirical version of the required invertibility conditions considered in Straumann and Mikosch (2006). This method provides a consistent QML estimator for the EGARCH(1,1) model.

In the literature, there are also consistency proofs for observation-driven models with nonlinear filters that do not rely on the invertibility concept of Straumann and Mikosch (2006), see for instance Harvey (2013), Harvey and Luati (2014) and Ito (2016). However, these results appeal to Lemma 2.1 of Jensen and Rahbek (2004) and rely on the very restrictive and non-standard assumption that the true value of the unobserved time-varying parameter is known at time $t = 0$. Unlike Jensen and Rahbek (2004), who carefully show that they do not need to impose this assumption in their non-stationary GARCH paper, this crucial issue is typically not addressed. As discussed in Wintenberger (2013) and Sorokin (2011), invertibility is not just a technical assumption as the lack of knowledge of the true initial value of the time-varying parameter at $t = 0$ can lead to the impossibility of recovering asymptotically the true time-varying parameter even knowing the true vector of static parameters. Furthermore, besides the invertibility issue, the results based on Lemma 2.1 of Jensen and Rahbek (2004) are only valid under the correct specification of the model and assuming that the likelihood function is maximized on an arbitrary small neighborhood around the true parameter value.

In this chapter, we extend the stabilization method of Wintenberger (2013) to a large class of observation-driven models and prove the consistency of the resulting ML estimator. The resulting theory provides feasible invertibility conditions that allow us to drop the unrealistic assumption that the time-varying parameter is known at $t = 0$. Our consistency

results hold for both correctly specified and misspecified models. In the latter case consistency is considered with respect to a pseudo-true parameter that has the interpretation of minimizing a marginal KL divergence between the true unknown conditional distribution and the conditional distribution of the postulated model. Additionally, we derive a test and confidence bounds for the "true" unfeasible parameter region. Our results cover a very wide class of models including ML estimation of GARCH-type models. In financial applications, maximum likelihood estimation for GARCH-type models is often preferred to QML estimation as the time series exhibit fat-tails and asymmetry. In this context, we provide an example of how our results can be useful in practice. In particular, we prove the consistency of the ML estimator for the Beta-t-GARCH model of Harvey (2013) and Creal et al. (2013). The usefulness of our theoretical results is further illustrated considering two example in the context of dynamic location model. In particular, we discuss the implications of our results considering the dynamic autoregressive model of Blasques et al. (2014b) and Delle Monache and Petrella (2016) and the fat-tailed location model of Harvey and Luati (2014).

The chapter is structured as follows. Section 2.2 motivates the theory presented in the chapter with an empirical application for which the invertibility conditions used in Straumann and Mikosch (2006) are too restrictive. Section 2.3 introduces the notion of invertibility of the filter and analyzes it in the context of the class of observation-driven models studied in this chapter. Section 2.4 presents the asymptotic results. Section 2.5 derives an invertibility test for the filter and obtains confidence bounds for the parameter space of interest. Section 2.6 shows the practical importance of the asymptotic results through some empirical illustrations. Section 2.7 concludes.

2.2 Motivation

Consider the Beta-t-GARCH model introduced by Harvey (2013) and Creal et al. (2013) for a sequence of financial returns $\{y_t\}_{t\in\mathbb{Z}}$ with time-varying conditional volatility and leverage effects,

$$y_t = \sqrt{f_t}\varepsilon_t \quad \text{and} \quad f_{t+1} = \omega + \beta f_t + (\alpha + \gamma d_t)\frac{(v+1)y_t^2}{(v-2) + y_t^2/f_t}, \tag{2.1}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of standard Student-t random variables with $v > 2$ degrees of freedom and d_t is a dummy variable that takes value $d_t = 1$ if $y_t \leq 0$ and $d_t = 0$ otherwise. In order to perform ML estimation of the model, the observed data

$\{y_t\}_{t=1}^T$ are used to obtain the filtered time-varying parameter $\hat{f}_t(\theta)$ as

$$\hat{f}_{t+1}(\theta) = \omega + \beta \hat{f}_t(\theta) + (\alpha + \gamma d_t) \frac{(v+1)y_t^2}{(v-2) + y_t^2/\hat{f}_t(\theta)}, \quad t \in \mathbb{N},$$

where the recursion is initialized at $\hat{f}_0(\theta) \in [0, +\infty)$. The invertibility concept of Straumann and Mikosch (2006) is concerned with the stability of $\hat{f}_t(\theta)$. In particular, it ensures that asymptotically the filtered parameter $\hat{f}_t(\theta)$ does not depend on the initialization $\hat{f}_0(\theta)$. Figure 2.2.1 illustrates the importance of the invertibility of the filter. The plots show differences between filtered volatility paths obtained from the S&P 500 returns for different initializations $\hat{f}_0(\theta)$. The left panel shows a situation where the filter is invertible and hence the effect of the initialization $\hat{f}_0(\theta)$ on $\hat{f}_t(\theta)$ vanishes as t increases. The right panel shows that the effect of the initialization does not vanish when the filter is not invertible.

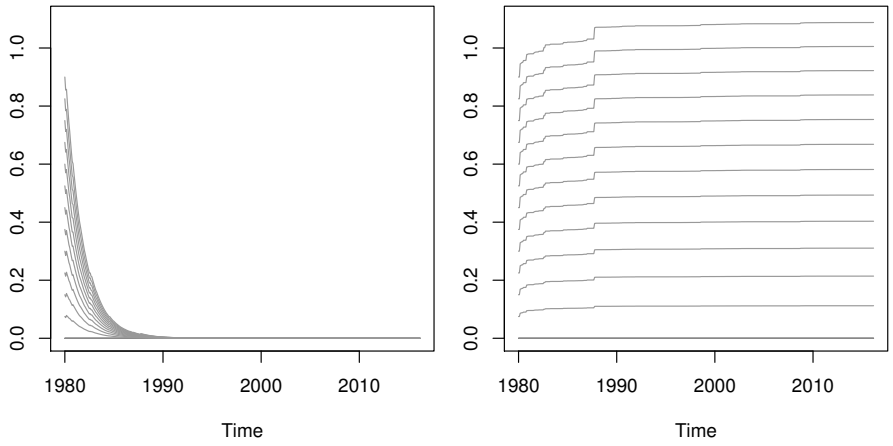

Figure 2.2.1: *Filtered variance paths for different initializations and using the S&P 500 time series. Differences are with respect to the filter initialized at* $\hat{f}_0(\theta) = 0.1$. *In the first plot, the vector of static parameters is selected to satisfy the invertibility conditions. In the second plot, a vector of static parameters that does not satisfy the invertibility conditions is considered.*

From a ML estimation perspective, the lack of invertibility of the filter also poses fundamental problems. Without invertibility, even asymptotically, the likelihood function depends on the initialization and hence this may lead the ML estimator to converge to different points when different initializations are considered. Furthermore, we may also be in a situation where we have a consistent estimator for the static parameter vector θ but we may not be able to consistently estimate the time-varying parameter. This consideration comes naturally from the fact that lack of invertibility can lead to the impossibility

of recovering the true path of the time-varying parameter even when the true vector of static parameters θ_0 is known, see Wintenberger (2013) and Sorokin (2011) for a more detailed discussion. As we shall see, the following condition on the parameter region Θ is sufficient for invertibility, and hence ensures the reliability of the ML estimator,

$$E \log \left| \beta + (\alpha + \gamma d_t) \frac{(v+1) y_t^4}{\left((v-2)\bar{\omega} + y_t^2\right)^2} \right| < 0, \ \forall \ \theta \in \Theta, \tag{2.2}$$

where $\bar{\omega} = \omega/(1-\beta)$. However, in practice, it is not possible to evaluate the expectation in (2.2) as it depends on the unknown DGP. Note that this is true even when the model is correctly specified as the true parameter vector θ_0 is unknown. Therefore, the derivation of the region Θ has to rely on feasible sufficient conditions to ensure (2.2). As we shall see in Section 2.6, assuming either correct specification or that y_t has a symmetric probability distribution around zero[1], we can obtain the following sufficient invertibility condition that does not depend on y_t

$$\frac{1}{2} \log |\beta + (\alpha + \gamma)(v+1)| + \frac{1}{2} \log |\beta + \alpha(v+1)| < 0.$$

Unfortunately, Figure 2.2.2 suggests that the set Θ obtained from such a sufficient condition is too small for empirical applications. In particular, Figure 2.2.2 highlights that a typical ML point estimate lies far outside Θ. This specific point estimate is obtained from monthly log-differences of the S&P 500 financial index from January 1980 to April 2016. Figure 2.2.2 might indicate that the filter is not stable or invertible. However, as we shall see in Section 2.6, this seems not to be the case. This point estimate lies well inside the estimated region for an invertible filter. The tests and confidence bounds developed in Section 2.5 further confirm this claim.

As we will discuss in Section 2.6, the problem illustrated in Figure 2.2.2 is not specific to this sample of data. Different samples of financial returns produce similar point estimates that lie also outside Θ. This problem is also not specific for the class of conditional heteroscedastic models. We illustrate this point considering the autoregressive model of Blasques et al. (2014b) and Delle Monache and Petrella (2016) and the location model of Harvey and Luati (2014). We find that, in general, the typical invertibility conditions needed to ensure the consistency of the ML estimator, which are considered for instance in Straumann and Mikosch (2006), Straumann (2005) and Blasques et al. (2014a), lead often to a parameter region that is too small for practical purposes. In contrary, the estimation method of Wintenberger (2013), proposed for the QML estimator of the EGARCH(1,1)

[1]Note that without this assumption the feasible invertibility condition would be even more restrictive.

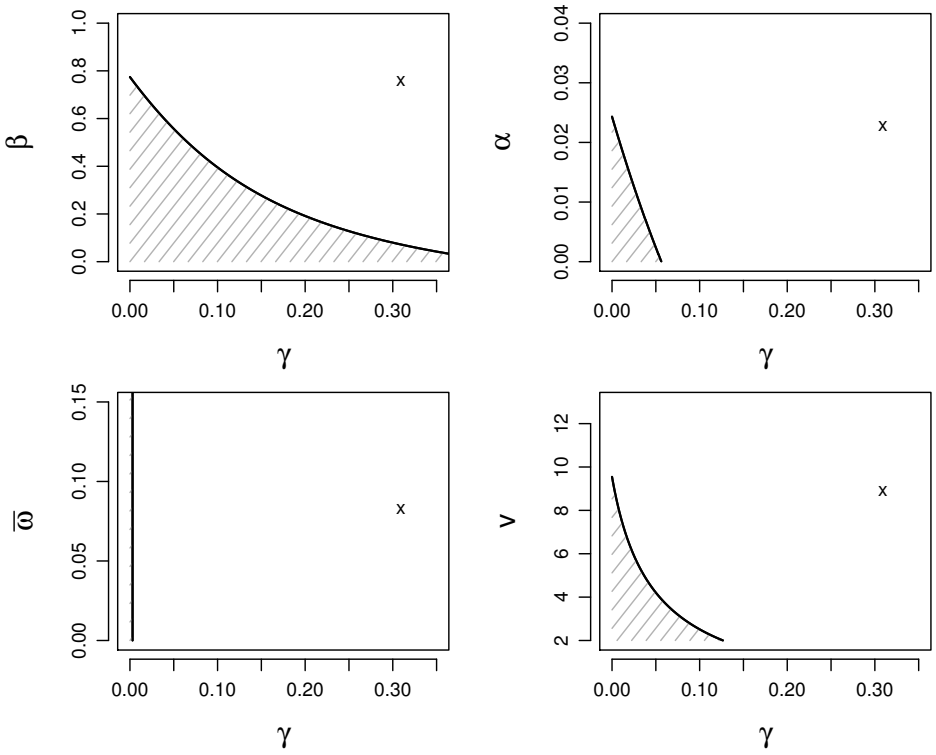

Figure 2.2.2: *The shaded area identifies the parameter region Θ that satisfies sufficient conditions for invertibility. Crosses locate the point estimate of the parameters of the Beta-t-GARCH model.*

model, can provide a parameter region large enough for practical applications. In Section 2.3 and Section 2.4, we will generalize the method of Wintenberger (2013) to ML estimation of a wide class of observation-driven models.

2.3 Invertibility of observation-driven filters

Let the observed sample of data $\{y_t\}_{t=1}^T$ be a subset of the realized path of a random sequence $\{y_t\}_{t\in\mathbb{Z}}$ with elements taking values in $\mathcal{Y} \subseteq \mathbb{R}$ and having and unknown conditional density $p^o(y_t|y^{t-1})$, where y^{t-1} denotes the entire past of the process $y^{t-1} = \{y_{t-1}, y_{t-2}, ...\}$. Consider now the following parametric observation-driven time-varying parameter model postulated by the researcher

$$y_t|f_t \sim p(y_t|f_t, \theta), \tag{2.3}$$

$$f_{t+1} = \phi(f_t, Y_t^k, \theta), \quad t \in \mathbb{Z}, \tag{2.4}$$

where $\theta \in \Theta \subseteq \mathbb{R}^p$ is a vector of static parameters, f_t is a time-varying parameter that takes values in $\mathcal{F}_\theta \subseteq \mathbb{R}$, ϕ is a continuous function from $\mathcal{F}_\theta \times \mathcal{Y}^k \times \Theta$ into $\mathcal{F}_\theta$, Y_t^k is a vector containing k lags of the observed time series $Y_t^k = (y_t, y_{t-1}, ..., y_{t-k})^T$, and $p(\cdot|f_t, \theta)$ is a conditional density function such that $(y, f, \theta) \mapsto p(y|f, \theta)$ is continuous on $\mathcal{Y} \times \mathcal{F}_\theta \times \Theta$.

As mentioned before, we also address the possibility of having a misspecified model. More specifically, we allow the parametric model in (2.3) and (2.4) to be fully misspecified. This means that both the dynamic specification of f_t and the conditional density $p(\cdot|f_t, \theta)$ can be misspecified. Note that a true time-varying parameter f_t may not even exist as we only assume that a true conditional density $p^o(\cdot|y^{t-1})$ exists. When we assume correct specification, the DGP $\{y_t\}_{t\in\mathbb{Z}}$ satisfies the model's equations (2.3) and (2.4) for $\theta = \theta_0$ and we denote with f_t^o the true time-varying parameter. In this situation, we have that $p^o(\cdot|y^{t-1}) = p(\cdot|f_t^o, \theta_0)$. Despite the possibility of model misspecification, it is worth noting that the model in (2.3) and (2.4) is very general and it covers a wide range of observation-driven models. Besides many GARCH-type models, this class of models includes location models as in Harvey and Luati (2014), Multiplicative Error Memory (MEM) models as in Engle (2002), Autoregressive Conditional Duration models as in Engle and Russell (1998), Autoregressive Conditional Intensity models as in Russell (2001) and Poisson autoregressive models as in Davis et al. (2003).

An important advantage of observation-driven models is that the likelihood function is analytically tractable and can be written in closed form as the product of conditional density functions. We consider the convention that the observations are available from time $t = 1 - k$. Using the observed data, the filtered parameter $\hat{f}_t(\theta)$ that enters in the likelihood function is obtained through the following SRE

$$\hat{f}_{t+1}(\theta) = \phi(\hat{f}_t(\theta), Y_t^k, \theta), \quad t \in \mathbb{N}, \tag{2.5}$$

where the recursion is initialized at $t = 0$ with $\hat{f}_0(\theta) \in \mathcal{F}_\theta$. Note that the set $\mathcal{F}_\theta$, where the time-varying parameter takes values, is indexed by $\theta \in \Theta$. As we will see for the Beta-t-GARCH model, this can be relevant in practice when dealing with specific models to weaken invertibility conditions. The ML estimator is formally defined as

$$\hat{\theta}_T = \arg\max_{\theta\in\Theta} \hat{L}_T(\theta), \tag{2.6}$$

where $\hat{L}_T(\theta)$ denotes the log-likelihood function evaluated at $\theta \in \Theta$,

$$\hat{L}_T(\theta) \;=\; T^{-1} \sum_{t=1}^{T} \hat{l}_t(\theta), \tag{2.7}$$

and $\hat{l}_t(\theta) = \log p(y_t|\hat{f}_t(\theta), \theta)$.

One of the difficulties in ensuring the consistency of the ML estimator is related to the recursive nature of the time-varying parameter and the consequent need of initializing the recursion in (2.5). In particular, it is important to note that the sequence $\{\hat{f}_t(\theta)\}_{t\in\mathbb{N}}$ as well as the sequence $\{\hat{l}_t(\theta)\}_{t\in\mathbb{N}}$ are both non-stationary. Therefore, the study of the limit behavior of $\{\hat{f}_t(\theta)\}_{t\in\mathbb{N}}$ is a natural requirement to ensure an appropriate form of convergence of the log-likelihood function $\hat{L}_T(\theta)$. The required stability of $\{\hat{f}_t(\theta)\}_{t\in\mathbb{N}}$ is known as invertibilty.

Bougerol (1993) provides well known conditions for the filtered sequence $\{\hat{f}_t(\theta)\}_{t\in\mathbb{N}}$ initialized at time $t = 0$ to converge exponentially fast almost surely (e.a.s.)[2] to a unique stationary and ergodic sequence $\{\tilde{f}_t(\theta)\}_{t\in\mathbb{Z}}$ as $t \to \infty$. In essence, this means that the effect of the initialization vanishes asymptotically at an exponential rate.[3] More formally, for any given $\theta \in \Theta$ and under appropriate conditions, Theorem 3.1 in Bougerol (1993) shows that

$$|\hat{f}_t(\theta) - \tilde{f}_t(\theta)| \xrightarrow{e.a.s.} 0, \;\; t \to \infty,$$

for any initialization $\hat{f}_0(\theta) \in \mathcal{F}_\theta$. Straumann and Mikosch (2006) make use of Bougerol's theorem and note that the e.a.s. convergence stated above is sufficient for the invertibility of the filter[4]. Their definition of invertibility is closely related to the definition of invertibility in Granger and Andersen (1978) as it implies that f_t^o is y^{t-1} measurable.

We mention that the stationary and ergodic limit sequence is denoted by $\tilde{f}_t(\theta)$ and not $f_t(\theta)$ to stress the fact that the stochastic properties of $\tilde{f}_t(\theta)$ are different from the stochastic properties of the sequence $f_t(\theta)$ that follows the model's equations (2.3) and (2.4). This because $\tilde{f}_t(\theta)$ is driven by past random variables of the DGP, which does not follow the model's equations. Under correct specification, we have that $\tilde{f}_t(\theta)$ has the same stochastic properties of $f_t(\theta)$ only when $\theta = \theta_0$ as the DGP follows the model equations only at θ_0. For more details see Straumann and Mikosch (2006) and Wintenberger (2013).

[2] A sequence of non-negative random variables $\{x_t\}_{t\in\mathbb{N}}$ is said to converge e.a.s. to zero if there exists a constant $\gamma > 1$ such that $\gamma^t x_t \xrightarrow{a.s.} 0$ as t diverges.

[3] In the context of correctly specified models this implies that the true path $\{f_t^o\}_{t\in\mathbb{Z}}$ can be asymptotically recovered as $\hat{f}_t(\theta_0)$ converges to $\tilde{f}_t(\theta_0) = f_t^o$ a.s. as t diverges.

[4] Straumann and Mikosch (2006) say that the model is invertible if $\hat{f}_t(\theta_0)$ converges in probability to $\tilde{f}_t^o$ and use Theorem 3.1 of Bougerol (1993) precisely to obtain the desired convergence.

It is also worth stressing the fact that, even if the model is assumed to be well specified, different conditions are required to establish invertibility and stationarity. As shown by Sorokin (2011) for some GARCH-type models, we can have that for a given θ_0, the model in (2.4) admits a stationary solution but lacks invertibility. In these situations, the true sequence $\{\hat{f}_t(\theta_0)\}_{t\in\mathbb{N}}$ can exhibit chaotic behaviors and the true path of f_t^o cannot be recovered asymptotically even when the true vector of static parameters θ_0 is known. See also the discussion in Wintenberger (2013). For this reason, ensuring the invertibility of the filtered parameter is not merely a technical requirement but an important ingredient to ensure the reliability of the inferential procedure.

The invertibility of the sequence $\{\hat{f}_t(\theta)\}_{t\in\mathbb{N}}$ evaluated at a single parameter value $\theta \in \Theta$ is not enough to ensure an appropriate convergence of the log-likelihood function over Θ. This happens naturally because the likelihood function depends on the functional sequence $\{\hat{f}_t\}_{t\in\mathbb{N}}$. In this regard, Wintenberger (2013) introduced the notion of continuous invertibility for GARCH-type models to ensure the uniform convergence of the filtered volatility. In our case, accounting for the continuity of the function ϕ, the elements of the sequence $\{\hat{f}_t\}_{t\in\mathbb{N}}$ can be considered as random elements in the space of continuous functions $\mathbb{C}(\Theta, \mathcal{F}_\Theta)$ that map from Θ into $\mathcal{F}_\Theta$, $\mathcal{F}_\Theta := \bigcup_{\theta\in\Theta} \mathcal{F}_\theta$, equipped with the uniform norm $\|\cdot\|_\Theta$, where $\|f\|_\Theta = \sup_{\theta\in\Theta} |f(\theta)|$ for any $f \in \mathbb{C}(\Theta, \mathcal{F}_\Theta)$. We say that the filter $\{\hat{f}_t\}_{t\in\mathbb{N}}$ is invertible if for any initialization $\hat{f}_0 \in \mathbb{C}(\Theta, \mathcal{F}_\Theta)$

$$\|\hat{f}_t - \tilde{f}_t\|_\Theta \xrightarrow{e.a.s.} 0, \quad t \to \infty,$$

where $\{\tilde{f}_t\}_{t\in\mathbb{Z}}$ is a stationary and ergodic sequence of random functions. Also in this case, note the relation with the invertibility concept in Granger and Andersen (1978) as the invertibility implies that the stochastic function $\tilde{f}_t$ is y^{t-1} measurable.

Proposition 2.3.1 presents sufficient conditions for the invertibility of $\{\hat{f}_t\}_{t\in\mathbb{N}}$. As in Straumann (2005), Straumann and Mikosch (2006) and Wintenberger (2013), the conditions we consider are based on Theorem 3.1 of Bougerol (1993). First, we define the stochastic Lipschitz coefficient $\Lambda_t(\theta)$ as

$$\Lambda_t(\theta) := \sup_{f\in\mathcal{F}_\theta} \left|\dot{\phi}(f, Y_t^k, \theta)\right|,$$

where $\dot{\phi}(f, Y_t^k, \theta) = \partial\phi(f, Y_t^k, \theta)/\partial f$.

Proposition 2.3.1. *Assume $\{y_t\}_{t\in\mathbb{Z}}$ is a stationary and ergodic sequence of random variables. Moreover, let the following conditions hold*

(i) There exists $\bar{f} \in \mathcal{F}_\Theta$ such that $E\log^+ \|\phi(\bar{f}, Y_t^k, \cdot)\|_\Theta < \infty$.

(ii) $E \sup_{\theta\in\Theta} \sup_{f\in\mathcal{F}_\Theta} \log^+ \left|\dot{\phi}(f, Y_t^k, \theta)\right| < \infty$.

(iii) $\log \Lambda_0(\theta)$ *is a.s. continuous on* Θ *and* $E \log \Lambda_0(\theta) < 0$ *for any* $\theta \in \Theta$.

Then, the functional sequence $\{\hat{f}_t\}_{t\in\mathbb{N}}$ *defined in (2.5) converges exponentially almost surely and uniformly to a unique stationary and ergodic sequence* $\{\tilde{f}_t\}_{t\in\mathbb{Z}}$*, i.e.*

$$\|\hat{f}_t - \tilde{f}_t\|_\Theta \xrightarrow{e.a.s.} 0 \ \ as \ \ t \to \infty,$$

for any initialization $\hat{f}_0 \in \mathbb{C}(\Theta, \mathcal{F}_\Theta)$.

Proposition 2.3.1 not only ensures the convergence of $\{\hat{f}_t\}_{t\in\mathbb{N}}$ to a stationary and ergodic sequence $\{\tilde{f}_t\}_{t\in\mathbb{Z}}$ but also that this sequence is unique and therefore the initialization $\hat{f}_0$ is irrelevant asymptotically. Note also that Proposition 2.3.1 holds irrespective of the correct specification of the model as it only requires that the data are generated by a stationary and ergodic process. Often, in practical situations, the so-called 'contraction condition' stated in (iii) is the most restrictive condition and it also imposes the most severe constraints on the parameter space Θ.

Remark 2.3.1. *When the model is correctly specified and conditions (i)-(iii) of Proposition 2.3.1 hold, then the filter evaluated at* $\theta_0 \in \Theta$ *converges to the true unobserved time-varying parameter* $\{f_t^o\}_{t\in\mathbb{Z}}$*, i.e.*

$$|\hat{f}_t(\theta_0) - f_t^o| \xrightarrow{e.a.s.} 0 \ \ as \ \ t \to \infty,$$

for any initialization $\hat{f}_0(\theta_0) \in \mathcal{F}_{\theta_0}$.

Remark 2.3.1 highlights an important implication of Proposition 2.3.1 under correct specification. We obtain that, knowing the vector of static parameters θ_0, the true path of f_t^o can be recovered asymptotically.

2.4 Maximum likelihood estimation

The invertibility of the filter obtained from Proposition 2.3.1 can be used to establish the consistency of the ML estimator defined in (2.6) over the parameter space Θ. We also discuss how the invertibility allows us to ensure the consistency of the plug-in estimators $\hat{f}_t(\hat{\theta}_T)$ and $p(y|\hat{f}_t(\hat{\theta}_T), \hat{\theta}_T)$, $y \in \mathcal{Y}$, for the time-varying parameter and the conditional density function. After the derivation of these results, we obtain the consistency of the ML estimator replacing the unfeasible parameter region Θ with an estimated set $\hat{\Theta}_T$ that

ensures an empirical version of the contraction condition $E \log \Lambda_0(\theta) < 0$. Finally, we study the case of model misspecification for the ML estimator based on the feasible parameter region $\hat{\Theta}_T$.

The subsequent results are subject to the stationarity and ergodicity of the data generating process. In the case of correct specification, stationarity and ergodicity can be checked studying the properties of the DGP, see Blasques et al. (2014c) for sufficient conditions for a wide class of observation-driven processes. In the case of misspecification, instead of imposing that the data are generated by a specific stationary and ergodic process, we allow the data generating process to be any stationary ad ergodic process.

2.4.1 Consistency of ML estimation

The first consistency result we obtain is under the assumption of correct specification. We denote the log-likelihood function evaluated at the stationary limit of the filtered parameter $\tilde{f}_t$ as $L_T(\theta) = T^{-1} \sum_{t=1}^{T} l_t(\theta)$, where $l_t(\theta) = \log p(y_t | \tilde{f}_t(\theta), \theta)$, and we denote by L the function $L(\theta) = E l_0(\theta)$. The following conditions are considered.

C1: The data generating process, which satisfies the equations (2.3) and (2.4) with $\theta = \theta_0 \in \Theta$, admits a stationary and ergodic solution and $E|l_0(\theta_0)| < \infty$.

C2: For any $\theta \in \Theta$, $l_0(\theta_0) = l_0(\theta)$ a.s. if and only if $\theta = \theta_0$.

C3: Conditions (i)-(iii) of Proposition 2.3.1 are satisfied for the compact set $\Theta \subset \mathbb{R}^p$.

C4: There exists a stationary sequence of random variables $\{\eta_t\}_{t \in \mathbb{Z}}$ with $E \log^+ |\eta_0| < \infty$ such that almost surely $\|\hat{l}_t - l_t\|_\Theta \leq \eta_t \|\hat{f}_t - \tilde{f}_t\|_\Theta$ for any $t \geq N$, $N \in \mathbb{N}$.

C5: $E\|l_0 \vee 0\|_\Theta < \infty$.

Condition **C1** ensures that the data are generated by a stationary and ergodic process and imposes an integrability condition on predictive log-likelihood, which is needed to apply an ergodic theorem. Condition **C2** is a standard identifiability condition. Conditions **C3** and **C4** ensure the a.s. uniform convergence of $\hat{L}_T$ to L_T. Finally, Condition **C5** ensures that L_n converges to an upper semicontinuous function L. As also considered in Straumann and Mikosch (2006), this final argument replaces the well known uniform convergence argument, namely, the uniform convergence of L_T to L. Note that Condition **C5** is weaker than the conditions typically needed for uniform convergence and in many cases it holds automatically as $l_0(\theta)$ is bounded from above with probability 1. Theorem 2.4.1 guarantees the strong consistency of the ML estimator.

Theorem 2.4.1. *Let the conditions **C1**-**C5** hold, then the ML estimator defined in (2.6) is strongly consistent, i.e.*

$$\hat{\theta}_T \xrightarrow{a.s.} \theta_0, \quad T \to \infty$$

for any initialization $\hat{f}_0 \in \mathbb{C}(\Theta, \mathcal{F}_\Theta)$.

The proof is in the appendix. In Section 2.6, the strong consistency of the Beta-t-GARCH model is proved by checking these conditions.

Often, the main objective of time series modeling is to describe the dynamic behaviour of the observed data and predict future observations. For this reason, it is interesting to study the consistency of the estimation of the time-varying parameter f_t^o and the conditional density function $p(y|f_t^o, \theta_0)$, $y \in \mathcal{Y}$. This further highlights the importance of the invertibility of the filter as without invertibility it may be possible to estimate consistently the static parameters, as shown by Jensen and Rahbek (2004) for the non-stationary GARCH(1,1), but it may not be possible to estimate consistently the time-varying parameter and the conditional density function. We consider plug-in estimates for the time-varying parameter, given by $\hat{f}_t(\hat{\theta}_T)$, and for the conditional density function, given by $p(y|\hat{f}_t(\hat{\theta}_T), \hat{\theta}_T)$, $y \in \mathcal{Y}$. The next result shows the consistency of these plug-in estimators. The consistency is obtained when both t and T go to infinity. This is needed because as T grows we obtain the consistency of the static parameter estimator and as t grows, thanks to the invertibility of the filter, we obtain that the effect of the initialization $\hat{f}_0$ becomes negligible. To obtain the desired result, besides the consistency conditions employed in Theorem 2.4.1, we additionally impose some Lipschitz conditions.

L1: There is a stationary sequence of random variables $\{v_t\}_{t\in\mathbb{Z}}$ such that almost surely

$$|\tilde{f}_t(\theta_1) - \tilde{f}_t(\theta_2)| \leq v_t \|\theta_1 - \theta_2\|, \quad \forall \ \theta_1, \theta_2 \in \Theta, t \in \mathbb{Z}.$$

L2: For any $y \in \mathcal{Y}$ there is a constant $c_y > 0$ such that

$$c_y \big|p(y|f_1, \theta_1) - p(y|f_2, \theta_2)\big| \leq \|\theta_1 - \theta_2\| + |f_1 - f_2|, \quad \forall \ \theta_1, \theta_2 \in \Theta \text{ and } f_1, f_2 \in \mathcal{F}_\Theta.$$

The vector norm $\|\cdot\|$ can be any vector norm. Corollary 2.4.1 below follows immediately from the Lipschitz condition on the filter **L1** and the Lipschitz condition on the conditional density function **L2**.

Corollary 2.4.1. *Let the conditions **C1**-**C5** and **L1** hold, then the plug-in estimator* $\hat{f}_t(\hat{\theta}_T)$ *is consistent, i.e.*

$$|\hat{f}_t(\hat{\theta}_T) - f_t^o| \xrightarrow{pr} 0, \quad T \to \infty, t \to \infty.$$

*Assume furthermore that also **L2** holds, then the plug-in estimator $p(y|\hat{f}_t(\hat{\theta}_T), \hat{\theta}_T)$ is consistent, i.e.*

$$\left|p(y|\hat{f}_t(\hat{\theta}_T), \hat{\theta}_T) - p(y|f_t^o, \theta_0)\right| \xrightarrow{pr} 0, \qquad T \to \infty, t \to \infty,$$

for any $y \in \mathcal{Y}$ and any initialization $\hat{f}_0 \in \mathbb{C}(\Theta, \mathcal{F}_\Theta)$.

Corollary 2.4.1 shows that the time-varying parameter f_t^o and the conditional density function $p(y|f_t^o, \theta_0)$, $y \in \mathcal{Y}$, can be consistently estimated.

2.4.2 ML on an estimated parameter region

As discussed before, the Lyapunov condition $E \log \Lambda_0(\theta) < 0$ imposes some restrictions on the parameter region Θ. Furthermore, in situations where $\Lambda_0(\theta)$ depends on Y_0^k, these restrictions cannot be checked as the expectation depends on the unknown DGP. Note that this is true even in the case of correct specification as the true parameter θ_0 is unknown. A possible solution is to obtain testable sufficient conditions such that $E \log \Lambda_0(\theta) < 0$ and define the set Θ accordingly. However, as discussed before, this often leads to very severe restrictions, reducing the set Θ to a small region that is usually too small for practical applications. Therefore, a better alternative consists in checking the condition $E \log \Lambda_0(\theta) < 0$ empirically and define the ML estimator as the maximizer of the log-likelihood on an estimated parameter region. In the context of QML estimation, this approach have been proposed by Wintenberger (2013) to stabilize the QML estimator of the EGARCH(1,1) model of Nelson (1991). In this section we formally define this ML estimator and we prove its consistency for the general class of observation-driven models defined in (2.3). In Section 2.6, we show how these results can be relevant in practical applications.

We define a compact set $\hat{\Theta}_T$ that satisfies an empirical version of the Lyapunov condition $E \log \Lambda_0(\theta) < 0$ as

$$\hat{\Theta}_T = \left\{ \theta \in \bar{\Theta} : \frac{1}{T} \sum_{t=1}^{T} \log \Lambda_t(\theta) \leq -\delta \right\}, \tag{2.8}$$

where $\bar{\Theta} \subset \mathbb{R}^p$ is a compact set and $\delta > 0$ is an arbitrary small constant. We assume that the compact set $\bar{\Theta}$ is chosen in such a way that $(f, y, \theta) \mapsto \phi(f, y, \theta)$ is a continuous on $\mathcal{F}_{\bar{\Theta}} \times \mathcal{Y}^k \times \bar{\Theta}$ and $(y, f, \theta) \mapsto p(y|f, \theta)$ is continuous on $\mathcal{Y} \times \mathcal{F}_{\bar{\Theta}} \times \bar{\Theta}$. For notational convenience, we also define the set $\Theta_c = \{\theta \in \bar{\Theta} : E \log \Lambda_0(\theta) < -c\}$, $c \in \mathbb{R}$. The ML

estimator on this empirical region $\hat{\Theta}_T$ is formally defined as

$$\hat{\hat{\theta}}_T = \underset{\theta \in \hat{\Theta}_T}{\arg\max}\, \hat{L}_T(\theta). \tag{2.9}$$

To ensure the consistency of this ML estimator in the case of correct specification the following conditions are considered.

A1: The DGP, which is given by the model in (2.3) and (2.4) with $\theta_0 \in \Theta_\delta$, admits a stationary and ergodic solution and $E|l_0(\theta_0)| < \infty$.

A2: Condition *(i)* and *(ii)* of Proposition 2.3.1 are satisfied for any compact subset $\Theta \subseteq \Theta_0$. Moreover, the map $\theta \mapsto \log \Lambda_0(\theta)$ is almost surely continuous on $\bar{\Theta}$ and $E\| \log \Lambda_0\|_{\bar{\Theta}} < \infty$.

A3: Conditions **C2**, **C4** and **C5** are satisfied for any compact subset $\Theta \subseteq \Theta_0$.

Note that **A1** ensures stationarity, ergodicity and invertibility of the data generating process. This condition can be seen as the equivalent of the condition **C1** in Theorem 2.4.1. The condition **A2** imposes some assumptions on $\log \Lambda_0(\theta)$. These assumptions are needed to guarantee a certain form of convergence for the set $\hat{\Theta}_T$ and consequently ensure the continuous invertibility $\|\hat{f}_t - \tilde{f}_t\|_{\hat{\Theta}_T} \xrightarrow{\text{e.a.s.}} 0$ as $t \to 0$ for large enough T. Therefore, **A2** can be seen as the equivalent of **C3** in Theorem 4.1. Finally, **A3**, together with **A2**, is sufficient to ensure that asymptotically the identifiability condition **C2**, the regularity condition **C4** and the integrability condition **C5** holds. The next theorem states the strong consistency of the ML estimator in (2.9) under correct specification.

Theorem 2.4.2. *Let conditions **A1**-**A3** hold, then the ML estimator defined in (2.9) is strongly consistent, i.e.*

$$\hat{\hat{\theta}}_T \xrightarrow{a.s.} \theta_0, \quad T \to \infty$$

for any initialization $\hat{f}_0 \in \mathbb{C}(\bar{\Theta}, \mathcal{F}_{\bar{\Theta}})$.

Theorem 2.4.2 generalizes Theorem 5 of Wintenberger (2013), which is specific to QML estimation of the EGARCH(1,1) model, to ML estimation of the wide class of observation-driven models specified in (2.3) and (2.4). The conditions required to ensure the strong consistency in Theorem 2.4.2 are feasible to be checked. This differs from other results in the literature such as Straumann and Mikosch (2006), Harvey (2013), Harvey and Luati (2014) and Ito (2016).

We now switch our focus to the possibility of having a misspecified model. This case is probably the most interesting from a practical point of view as the assumption that the

observed data are actually generated by the postulated model may be unreasonable. In the following, we show that, under misspecification, the ML estimator in (2.9) converges to a pseudo-true parameter θ^* that minimizes an average Kullback-Leibler (KL) divergence between the true conditional density $p^o(y_t|y^{t-1})$ and the postulated conditional density $p(y_t|\tilde{f}_t(\theta), \theta)$. Studies on consistency results with respect to pseudo true parameter for misspecified models go back to White (1982). We define the conditional KL divergence $KL_t(\theta)$ as

$$KL_t(\theta) = \int_{\mathcal{Y}} \log \frac{p^o(x|y^{t-1})}{p(x|\tilde{f}_t(\theta), \theta)} p^o(x|y^{t-1}) dx \tag{2.10}$$

and the average (marginal) KL divergence $KL(\theta)$ as $KL(\theta) = EKL_t(\theta)$. The pseudo true parameter θ^* is defined as the minimizer of $KL(\theta)$. The consistency result in this misspecified framework follows in a similar way as in the case of correct specification. This because Proposition 2.3.1 ensures the uniform convergence of $\hat{f}_t$ with no regards of the correct specification. The differences concern the stationarity and ergodicity of the DGP and the identifiability of the model. The following conditions are considered.

M1: The observed data are generated by a stationary and ergodic process $\{y_t\}_{t\in\mathbb{Z}}$ with conditional density function $p^o(y_t|y^{t-1})$ and the condition $E|\log p^o(y_0|y^{-1})| < \infty$ is satisfied.

M2: There is a parameter vector $\theta^* \in \Theta_\delta$ that is the unique maximizer of L, i.e. $L(\theta^*) > L(\theta)$ for any $\theta \in \Theta_0$, $\theta \neq \theta^*$.

M3: Condition **A2** is satisfied and **C4** and **C5** are satisfied for any compact set $\Theta \subseteq \Theta_0$.

Condition **M1** imposes the stationarity and ergodicity of the generating process and some moment conditions. Condition **M2** ensures identifiability in this misspecified setting. The continuous invertibility is ensured by **M3** as it imposes that **A2** holds and the results of Proposition 2.3.1 are irrespective of the correct specification of the model. Finally, in the same way as in **A3**, **M3** ensures that the conditions **C4** and **C5** hold for large enough T.

Theorem 2.4.3. *Let the conditions **M1**-**M3** hold, then the average KL divergence $KL(\theta)$ is well defined and the pseudo true parameter θ^* is its unique minimizer. Furthermore, the ML estimator defined in (2.9) is strongly consistent, i.e.*

$$\hat{\hat{\theta}}_T \xrightarrow{a.s.} \theta^*, \quad T \to \infty$$

for any initialization $\hat{f}_0 \in \mathbb{C}(\bar{\Theta}, \mathcal{F}_{\bar{\Theta}})$.

This result further highlights the relevance of ensuring invertibility. In this case, it is not possible to assume correct initialization of the filtered parameter as in Harvey (2013), Harvey and Luati (2014) and Ito (2016) since the true time-varying parameter does not even exists. The requirement that the filtered parameter asymptotically does not have to depend on the arbitrary chosen initialization is very intuitive as otherwise different initialization could provide different results.

We also note that situations of correctly-specified non-invertible models can be thought as a particular case of misspecification. This because, under non-invertibility, the true parameter value θ_0 is such that $E \log \Lambda_0(\theta_0) \geq 0$ and therefore asymptotically outside the parameter region $\hat{\Theta}_T$ with probability 1. In such situations, indeed, the ML estimator constrained on the empirical region $\hat{\Theta}_T$ is inconsistent with respect to θ_0 but we can ensure that asymptotically the initialization is not affecting the parameter estimate.

2.5 Confidence bounds for the parameter region

For a given sample $\{y_t\}_{t=1}^T$, some of the elements of the empirical region $\hat{\Theta}_T$ may not satisfy the required contraction condition $E \log \Lambda_0(\theta) < 0$. Therefore, for a given point $\theta \in \bar{\Theta}$, it may be of interest to test whether the condition is satisfied. Proposition 2.5.1 establishes the asymptotic normality of test statistic T_T defined below under the null hypothesis that $H_0 : E \log \Lambda_0(\theta) = 0$. Furthermore, we note that the statistic diverges under the alternative $H_1 : E \log \Lambda_0(\theta) \neq 0$. This result can naturally be used to produce interesting confidence bounds. Below we let σ_T^2 denote the variance of $T^{-\frac{1}{2}} \sum_{t=1}^T \log \Lambda_t(\theta)$.

Proposition 2.5.1. *Let* $\{y_t\}_{t\in\mathbb{Z}}$ *be stationary, ergodic and* α*-mixing of size* $-2r/(r-2)$*,* $r > 2$*, with* $E|\log \Lambda_0(\theta)|^r < \infty$ *for any* $\theta \in \bar{\Theta}$*. Then, under the null hypothesis* $H_0 : E \log \Lambda_0(\theta) = 0$ *we have*

$$T_T := \frac{T^{-\frac{1}{2}} \sum_{t=1}^T \log \Lambda_t(\theta)}{\hat{\sigma}_T} \xrightarrow{d} N(0,1) \quad \text{as } T \to \infty,$$

where $\hat{\sigma}_T^2$ *is a consistent estimator of* σ_T^2*. Furthermore,* $T_T \to -\infty$ *as* $T \to \infty$ *when* $E \log \Lambda_0(\theta) < 0$*, and* $T_T \to \infty$ *as* $T \to \infty$ *when* $E \log \Lambda_0(\theta) > 0$*.*

The variance σ_T^2 can be consistently estimated using the Newey-West estimator; see Newey and West (1987). Proposition 2.5.1 shows that, for any given θ and at any given confidence level α, we ascertain asymptotically if θ is a boundary point satisfying $E \log \Lambda_0(\theta) =$

0. If the null hypothesis is rejected with negative values of T_T, then the evidence suggests that the contraction condition is satisfied for that θ, i.e. that $E \log \Lambda_0(\theta) < 0$. If the null hypothesis is rejected with positive values of T_T, then the evidence suggests that $E \log \Lambda_0(\theta) > 0$. On the basis of the asymptotic result in Proposition 2.5.1, we can also obtain level α confidence sets for $\Theta_0 = \left\{\theta \in \bar{\Theta} : E \log \Lambda_0(\theta) < 0\right\}$. More specifically, we consider the set $\hat{\Theta}_\alpha^{up} = \left\{\theta \in \bar{\Theta} : T_T < z_{1-\alpha}\right\}$ such that for any $\theta \in \Theta_0$ we have

$$\lim_{n\to\infty} P\{\theta \in \hat{\Theta}_\alpha^{up}\} \geq 1 - \alpha.$$

This means that any element in the set Θ_0 has an asymptotic probability of at least $1 - \alpha$ of being contained in the set $\hat{\Theta}_\alpha^{up}$. Similarly, we also consider the set $\hat{\Theta}_\alpha^{lo} = \left\{\theta \in \bar{\Theta} : T_T < z_\alpha\right\}$ and for this set for that any $\theta \in \Theta_0^c$, where $\Theta_0^c = \left\{\theta \in \bar{\Theta} : E \log \Lambda_0(\theta) \geq 0\right\}$, we have that

$$\lim_{n\to\infty} P\{\theta \in \hat{\Theta}_\alpha^{lo}\} \leq \alpha.$$

The set $\hat{\Theta}_\alpha^{lo}$ can be seen as a lower bound confidence set of level α for Θ_0. This because, $\hat{\Theta}_\alpha^{lo}$ is a conservative set in the sense that we fix the maximum asymptotic probability α such that a θ not contained in Θ_0 can be in $\hat{\Theta}_\alpha^{lo}$. In an equivalent way, the set $\hat{\Theta}_\alpha^{up}$ can be seen as an upper bound confidence set for Θ_0. In this case, the maximum asymptotic probability of having an element $\theta \in \Theta_0$ not in $\hat{\Theta}_\alpha^{up}$ is fixed at a level α.

2.6 Some practical examples

2.6.1 The Beta-t-GARCH model

Consider first the properties of the Beta-t-GARCH model as a DGP. The process equation in (2.1) with $\theta = \theta_0$ can be expressed as

$$\begin{aligned} f_{t+1}^o &= \omega_0 + f_t^o c_t, \\ c_t &= \beta_0 + (\alpha_0 + \gamma_0 d_t)(v_0 + 1) b_t, \end{aligned}$$

where $b_t = \varepsilon_t^2/(v_0 - 2 + \varepsilon_t^2)$ has a beta distribution with parameters $1/2$ and $v_0/2$, see Chapter 3 of Harvey (2013). In order to ensure that f_t^o is positive with probability 1 and that f_t^o is the conditional variance of y_t given y^{t-1}, the parameter vector $\theta_0 = (\omega_0, \beta_0, \alpha_0, \gamma_0, v_0)^T$ has to satisfy the following conditions $\omega_0 > 0$, $\beta_0 \geq 0$, $\alpha_0 > 0$, $\gamma_0 \geq -\alpha_0$ and $v_0 > 2$. Letting $v_0 \to \infty$, the Student-t distribution approaches the Gaus-

sian distribution and the recursion of f_t^o in (2.1) becomes

$$f_{t+1}^o = \omega_0 + \beta_0 f_t^o + (\alpha_0 + \gamma_0 d_t) y_t^2.$$

Therefore, in the limit case $v_0 \to \infty$, this model is equivalent to the GJR-GARCH model of Glosten et al. (1993), and to the GARCH(1,1) model when $\gamma_0 = 0$.

Theorem 2.6.1. *The model in (2.1) admits a unique stationary and ergodic solution* $\{f_t^o\}_{t\in\mathbb{Z}}$ *if and only if* $E \log c_t < 0$.

Theorem 2.6.1 above derives a necessary and sufficient moment condition for the Beta-t-GARCH model to generate stationary ergodic paths. A simpler restriction on the parameters of the model that is sufficient for obtaining stationary and ergodic paths is

$$\beta_0 + \alpha_0 + \gamma_0/2 < 1.$$

Theorem 2.6.2 complements Theorem 2.6.1 by providing additional restrictions which ensure that the variance of the Beta-t-GARCH process is not only strictly stationary and ergodic but also has some bounded moments.

Theorem 2.6.2. *Let* $Ec_t^z < 1$*, where* $z \in \mathbb{R}^+$*, then (2.1) admits a unique stationary and ergodic solution* $\{f_t^o\}_{t\in\mathbb{Z}}$ *that satisfies* $E|f_t^o|^z < \infty$.

Having analyzed some properties of the Beta-t-GARCH as a DGP, we now turn to the properties of the model as a filter that is fitted to the data.

Invertibility of the filter

Let us analyze invertibility of the functional filtered parameter $\hat{f}_t$. The filter equation of the Beta-t-GARCH is given by

$$\hat{f}_{t+1}(\theta) = \omega + \beta \hat{f}_t(\theta) + (\alpha + \gamma d_t) \frac{(v+1) y_t^2}{(v-2) + y_t^2 / \hat{f}_t(\theta)}, \quad t \in \mathbb{N}, \tag{2.11}$$

where the recursion is initialized at a point $\hat{f}_0(\theta) \in \mathcal{F}_\theta = [\bar{\omega}, \infty), \bar{\omega} = \omega/(1-\beta)$. The observations $\{y_t\}_{t=1}^T$ are considered to be a realization from a random process. If we assume correct specification, then the generating process is given by (2.1) and there exists some true unknown parameter θ_0 that defines the properties of the data. It is straightforward to see that the set $\mathcal{F}_\theta$ where the SRE in (2.11) lies is given by $[\bar{\omega}, \infty)$. This is true irrespective

of the correct specification of the model as the last summand on the right hand side of the equation in (2.11) is positive with probability 1.

Corollary 2.6.1 follows immediately from Proposition 2.3.1 and provides sufficient conditions for the desired invertibility result.

Corollary 2.6.1. *Let $\{y_t\}_{t\in\mathbb{N}}$ be a stationary and ergodic sequence of random variables, and let Θ be a compact set such that*

$$E\log\left|\beta+(\alpha+\gamma d_0)\frac{(v+1)y_0^4}{\left((v-2)\bar{\omega}+y_0^2\right)^2}\right|<0,\ \forall\,\theta\in\Theta.$$

Then, the sequence $\{\hat{f}_t\}_{t\in\mathbb{N}}$ defined in (2.11) converges exponentially almost surely and uniformly to a unique stationary and ergodic sequence $\{\tilde{f}_t\}_{t\in\mathbb{Z}}$, i.e.

$$\|\hat{f}_t-\tilde{f}_t\|_\Theta \xrightarrow{e.a.s.} 0 \ \ as\ \ t\to\infty,$$

for any initialization $\hat{f}_0\in\mathbb{C}(\Theta,\mathcal{F}_\Theta)$.

As we can see from Corollary 2.6.1, the Lipschitz coefficient $\Lambda_0(\theta)$ depends on the DGP through y_0. Therefore, in practice, the parameter region Θ cannot be explicitly obtained from the condition $E\log\Lambda_0(\theta)<0$. As mentioned in Section 2.2, assuming either correct specification or that y_0 has a symmetric distribution around zero, the unfeasible contraction condition $E\log\Lambda_0(\theta)<0$ is ensured by the following feasible sufficient condition

$$\frac{1}{2}\log|\beta+\alpha(v+1)|+\frac{1}{2}\log|\beta+(\alpha+\gamma)(v+1)|<0. \tag{2.12}$$

This is obtained from the fact that, taking the supremum over y_0, it results that with probability 1

$$E\log\left|\beta+(\alpha+\gamma d_0)\frac{(v+1)y_0^4}{\left((v-2)\bar{\omega}+y_0^2\right)^2}\right|\leq E\log\left|\beta+(\alpha+\gamma d_0)(v+1)\right|.$$

Thus, assuming that the median of y_0 is equal to zero, the feasible condition in (2.12) follows immediately. Now, building on the theory developed in Sections 2.3 and 2.4, we are ready to consider as an alternative to (2.12) the estimated region $\hat{\Theta}_T$ that satisfies an empirical version of $E\log\Lambda_0(\theta)<0$, namely

$$T^{-1}\sum_{t=1}^{T}\log\left|\beta+(\alpha+\gamma d_t)\frac{(v+1)y_t^4}{\left((v-2)\bar{\omega}+y_t^2\right)^2}\right|<0. \tag{2.13}$$

Clearly, this empirical condition imposes weaker restrictions on the parameter region. In the following, we discuss how the difference between the condition (2.12) and (2.13) can be relevant in practice. Figure 2.6.1 complements Figure 2.2.2 by showing that our empirical region is significantly larger than the region obtained from (2.12). Most importantly, Figure 2.6.1 reveals that the ML point estimates obtained from the S&P 500 index lie well inside the empirical region. .

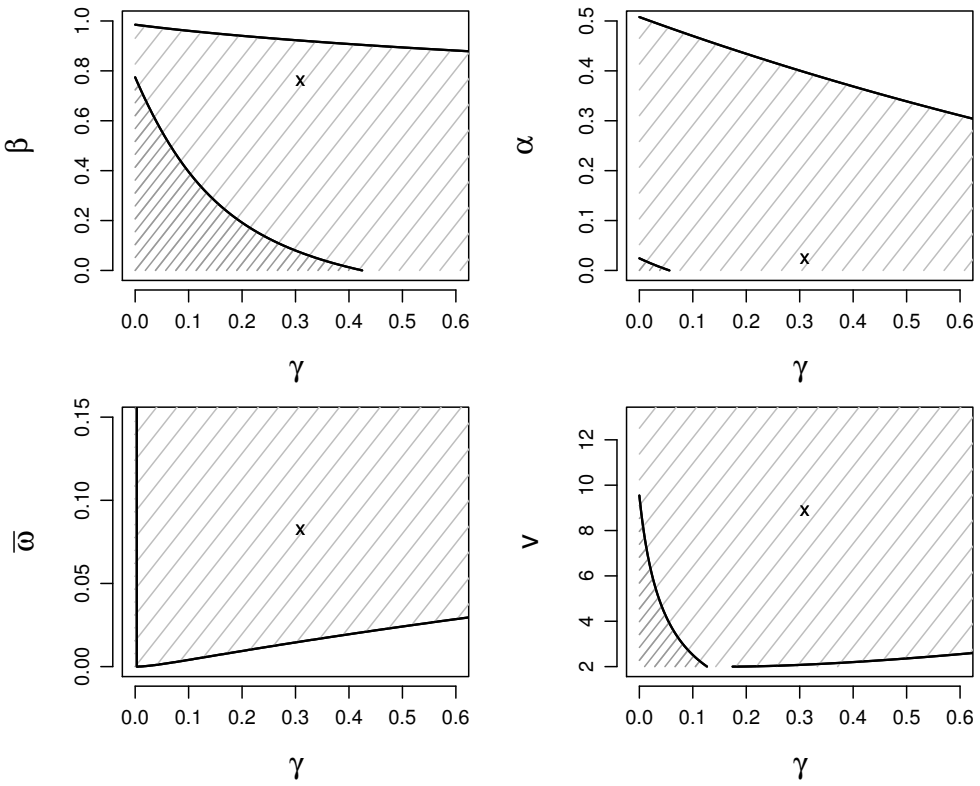

Figure 2.6.1: *The light gray area represents the parameter region obtained from (2.13) for the log-returns of the S&P 500. In the 2-dimensional plots the other parameters are fixed at their estimated value. The dark gray area is the region obtained from (2.12). The crosses denote the estimated value of the parameter.*

From the theory developed in Section 2.5, we can also obtain confidence bounds for the unfeasible parameter region. Note also that the conditions needed to apply Proposition 2.5.1 and thus obtain the confidence bounds are easily met in this case. In particular, the condition $E|\log \Lambda_0(\theta)|^r < \infty$ is satisfied for any $r > 0$ as long as $\beta > 0$. Whereas, from the results in Francq and Zakoïan (2006), it follows that the strong mixing assumption is always satisfied when the model is correctly specified. Figure 2.6.2 provides an high degree of confidence that the Beta-t-GARCH filter is in fact invertible. In particular, Figure 2.6.2 plots 95% confidence bounds for the invertibility region. We highlight that the point estimate lies well inside the 95% lower bound.

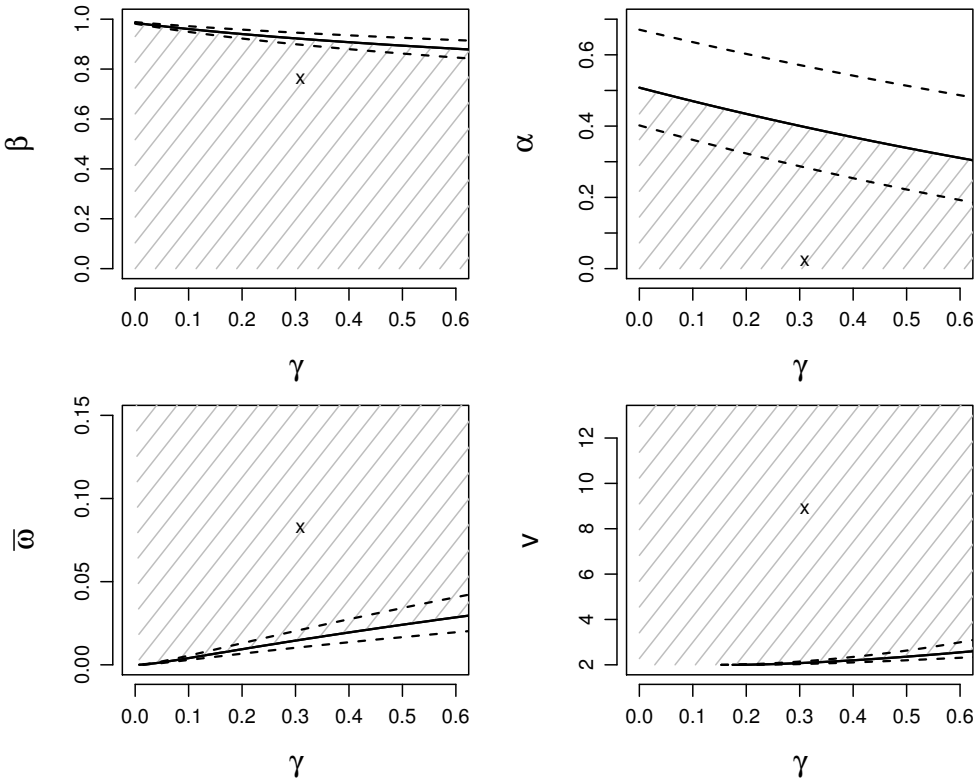

Figure 2.6.2: *Confidence bounds of 95% level for the invertibility region are represented by the dashed lines. The light gray areas represent the parameter region obtained from (2.13) for the log-returns of the S&P 500. Crosses denote the estimated value of the parameter.*

Table 2.6.1 reveals that the importance of our empirical invertibility condition is not specific to the S&P 500 index. In particular, for each time series in Table 2.6.1, we obtain the unrestricted maximizer of the likelihood function $\hat{\theta}$ and we show that inequality (2.12) evaluated at $\theta = \hat{\theta}$ fails whereas inequality (2.13) holds. This suggests that condition (2.12) is too restrictive in practice and that condition (2.13) can be used to define a reasonably large region of the parameter space on which we can maximize the log-likelihood function. The last column of Table 2.6.1 also shows that the null hypothesis that the point estimate is a boundary point of the invertibility region is strongly rejected.

Having discussed the invertibility of the Beta-t-GARCH filter, we are now ready to derive some consistency results for the ML estimator.

	ω	β	α	γ	v	(2.12)	(2.13)	p-value
DJIA	0.058	0.554	0.000	0.371	7.417	0.357	-0.507	0.000
	(0.019)	(0.160)	(0.047)	(0.116)	(2.339)			
S&P 500	0.020	0.759	0.023	0.309	8.893	0.691	-0.181	0.000
	(0.013)	(0.114)	(0.046)	(0.111)	(2.640)			
NASDAQ	0.026	0.754	0.106	0.198	9.865	1.022	-0.109	0.000
	(0.010)	(0.077)	(0.033)	(0.071)	(3.396)			
NI 225	0.088	0.637	0.000	0.230	26.552	0.746	-0.416	0.000
	(0.010)	(0.000)	(0.010)	(0.037)	(1.083)			
FTSE 100	0.042	0.595	0.059	0.332	7.621	0.737	-0.378	0.000
	(0.012)	(0.134)	(0.049)	(0.107)	(2.255)			
DAX	0.046	0.731	0.050	0.212	7.932	0.642	-0.218	0.000
	(0.013)	(0.088)	(0.046)	(0.073)	(2.905)			

Table 2.6.1: *Estimate of the model specified in (2.1) for the log-returns of some of the most popular stock indexes. Monthly time series from January 1980 to April 2016 are considered. The columns labeled (2.12) and (2.13) contain the values of respectively condition (2.12) and (2.13) evaluated at the estimated parameter value. The last column contains the p-value of the test to see whether the point estimate is in a boundary point of the "true" invertibility region.*

Consistency of the ML estimator

The log-likelihood function $\hat{L}_T$ is defined as in (2.7) with $\hat{l}_t(\theta)$ given by

$$\hat{l}_t(\theta) = \log\left(\frac{\Gamma\left(2^{-1}(v+1)\right)}{\sqrt{(v-2)\pi}\Gamma\left(2^{-1}v\right)}\right) - \frac{1}{2}\log\hat{f}_t(\theta) - \frac{v+1}{2}\log\left(1+\frac{y_t^2}{(v-2)\hat{f}_t(\theta)}\right),$$

where Γ denotes the gamma function.

Here, we obtain the consistency results for the Beta-t-GARCH model. The first result follows by an application of Theorem 2.4.1.

Theorem 2.6.3. *Let the observed data be generated by a stochastic process $\{y_t\}_{t\in\mathbb{Z}}$ that satisfies the model equations in (2.1) at $\theta = \theta_0 \in \Theta$ and such that $E\log c_t < 0$. Furthermore, let Θ be a compact set that satisfies the condition in (2.2) and such that $\omega > 0$, $\beta \geq 0$, $\alpha \geq 0$, $\gamma \geq -\alpha$ and $v > 2$ for any $\theta \in \Theta$. Then the ML estimator $\hat{\theta}_T$ defined in (2.6) is strongly consistent.*

Theorem 2.6.3, besides considering a more general model, extends the asymptotic results of Ito (2016) in several directions. In particular, Theorem 2.6.3 does not impose the assumption that the time-varying parameter f_t^o is observed at $t = 0$ and furthermore it does not consider that the likelihood function is maximized on an arbitrarily small

neighborhood around the true parameter θ_0. The next result shows the consistency of the ML estimator in (2.9) for the Beta-t-GARCH model.

Theorem 2.6.4. *Let the observed data be generated by a stochastic process $\{y_t\}_{t\in\mathbb{Z}}$ that satisfies the model equations in (2.1) at $\theta_0 \in \Theta_\delta$ and such that $E\log c_t < 0$. Furthermore, let $\bar{\Theta}$ be a compact set such that $\omega > 0$, $\beta > 0$, $\alpha \geq 0$, $\gamma \geq -\alpha$ and $v > 2$ for any $\theta \in \bar{\Theta}$. Then the ML estimator $\hat{\hat{\theta}}_T$ defined in (2.9) is strongly consistent.*

Unlike Theorem 2.6.3, Theorem 2.6.4 does not require the unfeasible invertibility condition in (2.2) to be satisfied as the optimization of the likelihood is on a region that satisfies an empirical version of (2.2).

2.6.2 Autoregressive model with dynamic coefficient

The practical relevance of the empirical invertibility conditions discussed in this chapter is not restricted to volatility models. On the contrary, it applies to the general class of observation-driven models. Consider the first-order autoregressive model with dynamic coefficient and fat tails of Blasques et al. (2014b) and Delle Monache and Petrella (2016). This model is specified through the following equations

$$\begin{aligned} y_t &= f_t y_{t-1} + \sigma\varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} t_v, \\ f_{t+1} &= \omega + \beta f_t + \alpha \frac{(y_t - f_t y_{t-1})y_{t-1}}{1 + v^{-1}\sigma^{-2}(y_t - f_t y_{t-1})^2}, \end{aligned}$$

where σ, ω, β, α and v are static parameters to be estimated and t_v denotes a Student-t distribution. This model is not exactly of the form in (2.3) and (2.4) as the conditional density of y_t given f_t depends also on the lagged value y_{t-1}. However, the extension needed to include this situation and possibly exogenous variables in the conditional density in (2.3) is trivial.

This autoregressive model allows for time-varying autocorrelation. In particular, it can describe time series that exhibit periods of strong temporal persistence, or near-unit-root dynamics, and periods of low dependence, or strong mean reverting behavior. There is evidence that many time series in economics feature such complex nonlinear dynamics; see Bec et al. (2008) for an example in real exchange rates. Following the results of Proposition 2.3.1 and taking into account that

$$\dot{\phi}(f, Y_t^k, \theta) = \beta + \alpha \frac{(y_t - f y_{t-1})^2 - v\sigma^2}{\left((y_t - f y_{t-1})^2 + v\sigma^2\right)^2} v\sigma^2 y_{t-1}^2,$$

we obtain that the stochastic coefficient $\Lambda_t(\theta)$ is given by

$$\Lambda_t(\theta) = \max\left\{|\beta - \alpha y_{t-1}^2|, |\beta + \frac{1}{8}\alpha y_{t-1}^2|\right\}.$$

In this case there is not a clear way to derive sufficient conditions to ensure that $E\log\Lambda_t(\theta) < 0$. A trivial solution is to impose that $\alpha = 0$ and $|\beta| < 1$. However, in this way, we get a degenerate parameter region and f_t becomes a static parameter. This situation is not of practical interest and the only possibility is to rely on the results of Section 2.4 and estimate the parameter region $\hat{\Theta}_T$.

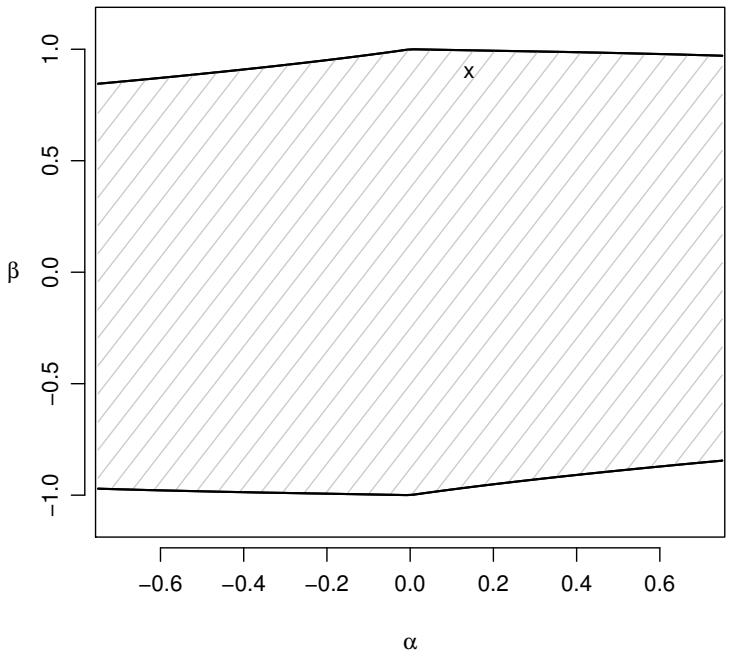

Figure 2.6.3: *Estimated parameter region and ML estimate obtained using the US unemployment claims time series.*

To show how the results of the previous sections can be useful in this situation, we derive the estimated region considering the weekly time series of log-differences of US unemployment claims. Note that this series is the the same considered in the empirical application of Blasques et al. (2014b). As we can see from Figure 2.6.3, the maximizer of the likelihood function is contained in the estimated region. This suggests that the empirical invertibility condition is not too restrictive. Therefore, we can conclude that the results obtained in the previous sections may be useful in this case.

2.6.3 Fat-tailed location model

As a final example, we consider the Student-t location model of Harvey and Luati (2014). This model is specified through the following equations

$$y_t = f_t + \sigma\varepsilon_t, \;\; \varepsilon_t \overset{\text{iid}}{\sim} t_v,$$
$$f_{t+1} = \omega + \beta f_t + \alpha \frac{y_t - f_t}{1 + v^{-1}\sigma^{-2}(y_t - f_t)^2},$$

where σ, ω, β, α and v are static parameters to be estimated and t_v denotes a Student-t distribution.

In an application to rail travel data of the United Kingdom, Harvey and Luati (2014) show that their Student-t location model is capable of extracting a smooth and robust trend from the rail travel series. Harvey and Luati (2014) also provide an asymptotic theory for the ML estimator of the static parameters of the model. Unfortunately, by relying on Lemma 1 of Jensen and Rahbek (2004), the ML estimator properties are obtained under the restrictive and non-standard assumption that the true time-varying mean f_t^o at time $t = 0$ is known. In addition, the asymptotic results derived in Harvey and Luati (2014) are only valid under correct model specification and assuming that the likelihood is maximized on an arbitrarily small parameter space containing θ_0. Therefore, also in this case, the results derived in this chapter can be useful to obtain the consistency of the ML estimator under weaker conditions. In the following, we only discuss invertibility conditions and provide an empirical example where our theory can be useful in practice.

First note that, as long as $|\beta| < 1$, the sequence $\{\hat{f}_t(\theta)\}_{t\in\mathbb{N}}$ takes values in $[\bar{\omega}_l, \bar{\omega}_u]$, where $\bar{\omega}_l = (\omega - c)/(1-\beta)$ and $\bar{\omega}_u = (\omega + c)/(1-\beta)$, with $c = |\alpha|\sqrt{3v\sigma^2}/4$. Defining the function $s_\theta(x) := v\sigma^2(x^2 - v\sigma^2)/(x^2 + v\sigma^2)^2$, we obtain that the stochastic coefficient $\Lambda_t(\theta)$ is

$$\Lambda_t(\theta) = \max\left\{|z_{1t}|, |z_{2t}|\right\},$$

where z_{1t} and z_{2t} are respectively given by

$$z_{1t} = \begin{cases} \beta - \alpha & \text{if} \quad y_t \in [\bar{\omega}_l, \bar{\omega}_u], \\ \beta + \alpha \min\left(s_\theta(y_t - \bar{\omega}_u), s_\theta(y_t - \bar{\omega}_l)\right) & \text{otherwise,} \end{cases}$$

and

$$z_{2t} = \begin{cases} \beta + \alpha/8 & \text{if} \quad y_t \pm \sqrt{3v\sigma^2} \in [\bar{\omega}_l, \bar{\omega}_u], \\ \beta + \alpha \max\left(s_\theta(y_t - \bar{\omega}_u), s_\theta(y_t - \bar{\omega}_l)\right) & \text{otherwise.} \end{cases}$$

We note that it is possible to obtain an upper bound for $\Lambda_t(\theta)$ independent of the observations. This is given by

$$\Lambda_t(\theta) \leq \max(|\beta - \alpha|, |\beta + \alpha/8|). \tag{2.14}$$

Unfortunately, this condition can be too restrictive. Figure 2.6.4 shows an example where this more restrictive condition fails to hold while, on the other hand, the empirical condition is satisfied. We consider the US monthly series of changes in the consumer price inflation index from January 1947 to February 2016. As we can see in Figure 2.6.4, the estimated parameter region $\hat{\Theta}_T$ is larger than the region obtained from the upper bound in (2.14). Furthermore, the empirical region is also large enough to contain the parameter estimate.

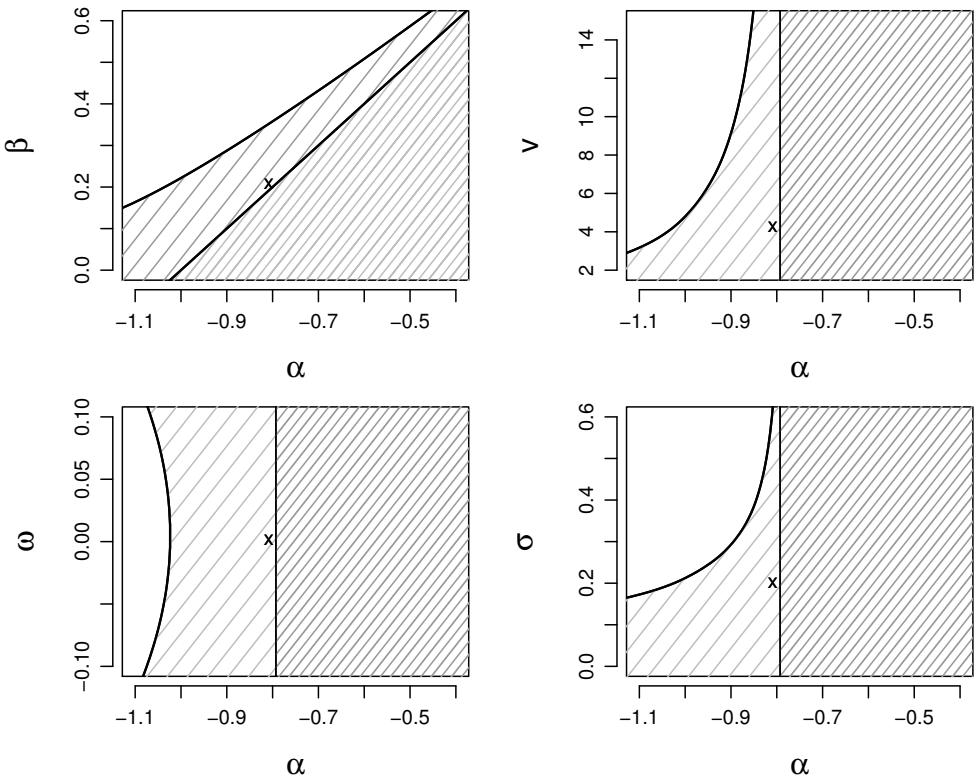

Figure 2.6.4: *The light gray area denotes the estimated parameter region and the dark gray are denotes the region obtained from (2.14). The crosses denote the ML parameter estimate. The plots are obtained using the US consumer price index time series from January 1947 to February 2016.*

2.7 Conclusion

In this chapter, we have proposed considerably weaker conditions that can be used in practice to ensure the consistency of the ML estimator. These results are applicable to a wide class of observation-driven models. Furthermore, we have shown that our consistency results hold for both correctly specified and misspecified models. Additionally, we have also derived an asymptotic test and confidence bounds for the unfeasible "true" invertibility region of the parameter space. The practical usefulness of the theory developed in the chapter has been highlighted by analyzing a number of popular observation-driven models with real datasets.

Appendix

2.A Proofs

Proof of Proposition 2.3.1. To prove this proposition, we first rely on the results of Proposition 3.12 of Straumann and Mikosch (2006) and we then employ the same argument as in the proof of Theorem 2 of Wintenberger (2013) to relax the uniform contraction condition. This proposition is closely related to Theorem 2 of Wintenberger (2013), the main difference is that we explicitly allow the set $\mathcal{F}_\theta$ to depend on θ.

Consider the functional SRE

$$\hat{f}_{t+1} = \Phi_t(\hat{f}_t),\ t \in \mathbb{N},$$

where the random map Φ_t is such that $\Phi_t(f) = \phi(f(\cdot), Y_t^k, \cdot)$ for any $f \in \mathbb{C}(C, \mathcal{F}_C)$, where C denotes a compact set. This SRE lies in the separable Banach space $\mathbb{C}(C, \mathcal{F}_C)$ equipped with the uniform norm $\|\cdot\|_C$. Therefore, taking into account that by the mean value theorem

$$\sup_{f_1, f_2 \in \mathcal{F}_C, f_1 \neq f_2} \frac{|\phi(f_1, Y_t^k, \theta) - \phi(f_2, Y_t^k, \theta)|}{|f_1 - f_2|} \leq \sup_{f \in \mathcal{F}_C} |\dot{\phi}(f, Y_t^k, \theta)|,$$

from Proposition 3.12 of Straumann and Mikosch (2006), it results that the conditions

(a) $E \log^+ \|\phi(\bar{f}, Y_t^k, \cdot)\|_C < \infty$ for some $\bar{f} \in \mathcal{F}_C$.

(b) $E \sup_{\theta \in C} \sup_{f \in \mathcal{F}_C} \log^+ |\dot{\phi}(f, Y_t^k, \theta)| < \infty$.

(c) $E \sup_{\theta \in C} \sup_{f \in \mathcal{F}_C} \log |\dot{\phi}(f, Y_t^k, \theta)| < 0$.

are sufficient to apply Theorem 3.1 of Bougerol (1993) and obtain the convergence result $\|\hat{f}_t - \tilde{f}_t\|_C \xrightarrow{e.a.s.} 0$. Note that this is true for any given compact set C that satisfies (a)-(c).

Now, we define the following stochastic function

$$\Lambda_t^*(\theta_1, \theta_2) := \sup_{f \in \mathcal{F}_{\theta_1}} |\dot{\phi}(f, Y_t^k, \theta_2)|,$$

and, we define a compact neighborhood of $\theta \in \Theta$ with radius $\epsilon > 0$ as $B_\epsilon(\theta) = \{\tilde{\theta} \in \Theta : \|\theta - \tilde{\theta}\| \leq \epsilon\}$. Then, for any non-increasing sequence of constants $\{\epsilon_i\}_{i \in \mathbb{N}}$ such that $\lim_{i \to \infty} \epsilon_i = 0$, the sequence $\left\{\sup_{(\theta_1,\theta_2) \in B_{\epsilon_i}(\theta) \times B_{\epsilon_i}(\theta)} \log \Lambda_0^*(\theta_1, \theta_2)\right\}_{i \in \mathbb{N}}$ is a non-increasing sequence of random variables and by continuity, which is ensured by (iii), we have that

$$\lim_{i \to \infty} \sup_{(\theta_1,\theta_2) \in B_{\epsilon_i}(\theta) \times B_{\epsilon_i}(\theta)} \log \Lambda_0^*(\theta_1, \theta_2) = \log \Lambda_0(\theta).$$

Condition (ii) implies that $E \sup_{(\theta_1,\theta_2) \in \Theta \times \Theta} \log \Lambda_0^*(\theta_1, \theta_2) \in \mathbb{R} \cup \{-\infty\}$. As a result, we can apply the monotone convergence theorem and obtain

$$E \lim_{i \to \infty} \sup_{(\theta_1,\theta_2) \in B_{\epsilon_i}(\theta) \times B_{\epsilon_i}(\theta)} \log \Lambda_0^*(\theta_1, \theta_2) = E \log \Lambda_0(\theta).$$

Therefore, for any $\theta \in \Theta$ such that $E \log \Lambda_0(\theta) < 0$ there exists an $\epsilon_\theta > 0$ such that

$$E \sup_{(\theta_1,\theta_2) \in B_{\epsilon_\theta}(\theta) \times B_{\epsilon_\theta}(\theta)} \log \Lambda_0^*(\theta_1, \theta_2) < 0.$$

From this and noting that

$$\sup_{\theta \in B_{\epsilon_\theta}(\theta)} \sup_{f \in \mathcal{F}_{B_{\epsilon_\theta}(\theta)}} \log |\dot{\phi}(f, Y_t^k, \theta)| = \sup_{(\theta_1,\theta_2) \in B_{\epsilon_\theta}(\theta) \times B_{\epsilon_\theta}(\theta)} \log \Lambda_0^*(\theta_1, \theta_2),$$

we obtain that the conditions (a)-(c) are satisfied for the compact set $B_{\epsilon_\theta}(\theta)$ as (i) implies (a), (ii) implies (b) and (iii) implies (c). Therefore, we conclude that

$$\|\hat{f}_t - \tilde{f}_t\|_{B_{\epsilon_\theta(\theta)}} \xrightarrow{e.a.s.} 0.$$

The desired result follows as Θ is compact and $\Theta = \bigcup_{\theta \in \Theta} B_{\epsilon_\theta}(\theta)$. Therefore, there exists a finite set of points $\{\theta_1, \ldots, \theta_K\}$ such that $\Theta = \bigcup_{k=1}^K B_{\epsilon_k}(\theta_k)$ and it follows that

$$\|\hat{f}_t - \tilde{f}_t\|_\Theta = \bigvee_{k=1}^K \|\hat{f}_t - \tilde{f}_t\|_{B_{\epsilon_k(\theta_k)}} \xrightarrow{e.a.s.} 0.$$

□

Proof of Theorem 2.4.1. We prove the theorem from the following intermediate steps:

(S1) The model is identifiable, i.e. $L(\theta_0) > L(\theta)$ for any $\theta \in \Theta$, $\theta \neq \theta_0$.

(S2) The function $\hat{L}_T$ converges a.s. uniformly to L_T as $T \to \infty$, i.e. $\|\hat{L}_T - L_T\|_\Theta \xrightarrow{\text{a.s.}} 0$ as $T \to \infty$.

(S3) For any $\epsilon > 0$, the following inequality holds with probability 1

$$\limsup_{T \to \infty} \sup_{\theta \in B^c(\theta_0, \epsilon)} \hat{L}_T(\theta) < L(\theta_0), \tag{2.15}$$

where $B^c(\theta_0, \epsilon) = \Theta \setminus B(\theta_0, \epsilon)$ with $B(\theta_0, \epsilon) = \{\theta \in \Theta : \|\theta_0 - \theta\| < \epsilon\}$;

(S4) The result in (S3) implies strong consistency.

(S1) First note that, by **C1**, $L(\theta_0)$ exists and is finite and, by **C5**, $L(\theta)$ exists for any $\theta \in \Theta$ with either $L(\theta) = -\infty$ or $L(\theta) \in \mathbb{R}$. For the values $\theta \in \Theta$ such that $L(\theta) = -\infty$, the result $L(\theta_0) > L(\theta)$ follows immediately as $L(\theta_0)$ is finite. Hence, from now on, we consider only the values $\theta \in \Theta$ such that $L(\theta)$ is finite. It is well known that $\log(x) \leq x - 1$ for any $x \in \mathbb{R}^+$ with the equality only in the case $x = 1$. This implies that almost surely

$$l_0(\theta) - l_0(\theta_0) \leq \frac{p(y_0|\tilde{f}_0(\theta), \theta)}{p(y_0|f_0^o, \theta_0)} - 1. \tag{2.16}$$

Moreover, we have that the inequality in (2.16) holds as a strict inequality with positive probability as the possibility that $p(y_0|\tilde{f}_0(\theta), \theta) = p(y_0|f_0^o, \theta_0)$ a.s. is ruled out by **C2** for any $\theta \neq \theta_0$. As a result

$$E\left[E\left[l_0(\theta) - l_0(\theta_0)|y^{-1}\right]\right] < E\left[E\left[\frac{p(y_0|\tilde{f}_0(\theta), \theta)}{p(y_0|f_0^o, \theta_0)}\Big|y^{-1}\right]\right] - 1 = 0, \quad \forall\, \theta \neq \theta_0$$

where the right hand side of the inequality is equal to zero as $p(y_0|f_0^o, \theta_0)$ is the true conditional density function. The desired result $L(\theta_0) > L(\theta)$ follows as $l_0(\theta) - l_0(\theta_0)$ is integrable and therefore by the law of total expectation

$$L(\theta) - L(\theta_0) = E[E[l_0(\theta) - l_0(\theta_0)|y^{-1}]] < 0 \quad \forall\, \theta \neq \theta_0.$$

This concludes the proof of step (S1).

(S2) First, note that $\|\hat{f}_t - \tilde{f}_t\|_\Theta \xrightarrow{\text{e.a.s.}} 0$ as $t \to \infty$ by an application of Proposition 2.3.1 as conditions *(i)-(iii)* hold by **C3** and $\{y_t\}_{t \in \mathbb{Z}}$ is stationary and ergodic by **C1**. Second, by Lemma 2.1 of Straumann and Mikosch (2006) the series $\sum_{t=N}^{\infty} \eta_t \|\hat{f}_t - \tilde{f}_t\|_\Theta$

converges a.s. and therefore the inequality in **C4** implies $\sum_{t=N}^{\infty} \|\hat{l}_t - l_t\|_\Theta < \infty$ a.s.. As a result $T^{-1}\sum_{t=1}^{T} \|\hat{l}_t - l_t\|_\Theta \xrightarrow{\text{a.s.}} 0$ and $\|\hat{L}_T - L_T\|_\Theta \xrightarrow{\text{a.s.}} 0$ follows as $\|\hat{L}_T - L_T\|_\Theta \leq T^{-1}\sum_{t=1}^{T} \|\hat{l}_t - l_t\|_\Theta$ for any $T \in \mathbb{N}$. This concludes the proof of (S2).

(S3) First, note that in virtue of (S2) $\hat{L}_T$ is asymptotically equivalent to L_T and therefore we just need to prove that (S3) holds for L_T. To show this, a similar argument as in the proof of Lemma 3.11 of Pfanzagl (1969) is employed. Consider any decreasing sequence of real numbers $\{\epsilon_i\}_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} \epsilon_i = 0$, then $\{\sup_{\theta^*\in B(\theta,\epsilon_i)} l_0(\theta^*)\}_{i\in\mathbb{N}}$ defines a non-increasing sequence of random variables and, by continuity, we have that $\lim_{i\to\infty} \sup_{\theta^*\in B(\theta,\epsilon_i)} l_0(\theta^*) = l_0(\theta)$. As **C5** implies $E\sup_{\theta\in\Theta} l_0(\theta) < \infty$ we can apply the monotone convergence theorem and we get

$$\lim_{i\to\infty} E \sup_{\theta^*\in B(\theta,\epsilon_i)} l_0(\theta^*) = L(\theta).$$

Recalling that $L(\theta_0) > L(\theta)$ by (S1), we have that for any $\theta \neq \theta_0$ there exists an $\epsilon_\theta > 0$ such that

$$\limsup_{T\to\infty} \sup_{\theta^*\in B(\theta,\epsilon_\theta)} L_T(\theta^*) \leq E \sup_{\theta^*\in B(\theta,\epsilon_\theta)} l_0(\theta^*) < L(\theta_0).$$

Finally, by compactness of $B^c(\theta_0,\epsilon)$ and by $B^c(\theta_0,\epsilon) \subseteq \bigcup_{\theta\in B^c(\theta_0,\epsilon)} B(\theta,\epsilon_\theta)$, there is a finite set of points $\{\theta_1,\ldots,\theta_K\}$ such that $B^c(\theta_0,\epsilon) \subseteq \bigcup_{k=1}^{K} B(\theta_k,\epsilon_k)$. Therefore, for any $T \in \mathbb{N}$ we have

$$\sup_{\theta\in B^c(\theta_0,\epsilon)} L_T(\theta) \leq \bigvee_{k=1}^{K} T^{-1}\sum_{t=1}^{T} \sup_{\theta\in B(\theta_k,\epsilon_k)} l_t(\theta),$$

and taking the limit in both sides of the inequality it results

$$\limsup_{T\to\infty} \sup_{\theta\in B^c(\theta_0,\epsilon)} L_T(\theta) \leq \bigvee_{k=1}^{K} E \sup_{\theta\in B(\theta_k,\epsilon_k)} l_0(\theta) < L(\theta_0).$$

This concludes the proof of (S3).

(S4) This last step follows from standard arguments due to Wald (1949). From the definition of the ML estimator, we have $\hat{L}_T(\hat{\theta}_T) \geq \hat{L}_T(\theta_0)$ for any $T \in \mathbb{N}$. Therefore, given the result in (S3), we have that

$$\liminf_{T\to\infty} \hat{L}_T(\hat{\theta}_T) \geq L(\theta_0). \tag{2.17}$$

Now, if we assume that there exists an $\epsilon > 0$ such that $\limsup_{T\to\infty} \|\hat{\theta}_T - \theta_0\| \geq \epsilon$, then

in virtue of (2.17) it must hold that

$$\limsup_{T\to\infty} \sup_{\theta\in B^c(\theta_0,\epsilon)} \hat{L}_T(\theta) \geq L(\theta_0),$$

but because of (2.15) this event has probability zero. As a result, $\limsup_{T\to\infty} \|\hat{\theta}_T - \theta_0\| < \epsilon$ with probability 1 for any $\epsilon > 0$. This concludes the proof of the theorem. □

Proof of Corollary 2.4.1. First, we consider the consistency of the time-varying parameter estimator $\hat{f}_t(\hat{\theta}_T)$. From the Lipschitz condition **L1**, it follows that with probability 1

$$|\hat{f}_t(\hat{\theta}_T) - f_t^o| \leq \|\hat{f}_t - \tilde{f}_t\|_\Theta + v_t\|\hat{\theta}_T - \theta_0\|.$$

Therefore, the consistency result is obtained as both terms on the right hand side of the inequality go to zero in probability when both t and T go to infinity. In particular, the first term goes to zero a.s. from the invertibility of the filter and the second term goes to zero as $\{v_t\}_{t\in\mathbb{Z}}$ is stationary, thus $O_p(1)$, and $\|\hat{\theta}_T - \theta_0\|$ is $o_p(1)$ as ensured by Theorem 2.4.1. Finally, the consistency of the plug-in density function estimator follows immediately from the additional Lipschitz condition **L2** as it implies that with probability 1

$$\left|p(y|\hat{f}_t(\hat{\theta}_T), \hat{\theta}_T) - p(y|f_t^o, \theta_0)\right| \leq c_y^{-1} \left(\|\hat{\theta}_T - \theta_0\| + |\hat{f}_t(\hat{\theta}_T) - f_t^o|\right),$$

and the right hand side of the inequality goes to zero in probability from the consistency of $\hat{\theta}_T$ and $\hat{f}_t(\hat{\theta}_T)$ as T and t go to infinity. □

Proof of Theorem 2.4.2. To prove this theorem we show that the steps (S1)-(S4) in the proof of Theorem 2.4.1 hold replacing the set Θ with the set $\hat{\Theta}_T$.

First we show that the following results hold true

(a) Almost surely, for large enough T, the true parameter vector θ_0 is contained in the set $\hat{\Theta}_T$.

(b) Almost surely, for large enough T, the set $\hat{\Theta}_T$ is contained in the compact set $\bar{\Theta}_{\delta/2}$ defined as $\bar{\Theta}_{\delta/2} := \{\theta \in \bar{\Theta} : E \log \Lambda_0(\theta) \leq -\delta/2\}$.

By the a.s. continuity of $\log \Lambda_t(\theta)$ in $\bar{\Theta}$ ensured by **A2**, the sequence $\{\log \Lambda_t\}_{t\in\mathbb{N}}$ is a stationary and ergodic sequence of elements in the separable Banach space $\mathbb{C}(\bar{\Theta}, \mathbb{R})$ equipped with the uniform norm $\|\cdot\|_{\bar{\Theta}}$. The uniform integrability condition $E\|\log \Lambda_0\|_{\bar{\Theta}} <$

∞ in **A2** enables us to apply the ergodic theorem of Rao (1962) and it follows that

$$\left\| T^{-1} \sum_{t=1}^{T} \log \Lambda_t - E \log \Lambda_0 \right\|_{\bar{\Theta}} \xrightarrow{\text{a.s.}} 0, \; T \to \infty. \tag{2.18}$$

This implies that for a large enough T all the points $\theta \in \bar{\Theta}$ such that $E \log \Lambda_0(\theta) < -\delta$ are contained in $\hat{\Theta}_T$. Therefore, the result (a) holds as condition **A1** ensures that $E \log \Lambda_0(\theta_0) < -\delta$. As concerns the result (b), the application of the uniform ergodic theorem implies that the map $\theta \mapsto E \log \Lambda_0(\theta)$ is continuous in $\bar{\Theta}$. This yields that the set $\bar{\Theta}_{\delta/2}$ is compact. Finally, $\hat{\Theta}_T \subset \bar{\Theta}_{\delta/2}$ almost surely for large enough T follows immediately from definition of $\hat{\Theta}_T$ and $\bar{\Theta}_{\delta/2}$ and the uniform convergence in (2.18).

Indeed, $\bar{\Theta}_{\delta/2}$ is a compact set contained in $\bar{\Theta}$ and such that $E \log \Lambda_0(\theta) < 0$ for any $\theta \in \bar{\Theta}_{\delta/2}$. Therefore, from the result (b) together with **A1**-**A3**, it is easy to see that (S1) is a.s. satisfied for large enough T as it holds for the set $\bar{\Theta}_{\delta/2}$. We also have that (S2) and (S3) are satisfied for the set $\hat{\Theta}_T$ as they hold for the set $\bar{\Theta}_{\delta/2}$. Finally, the step (S4) follows in the same way as in the proof of Theorem 2.4.1 by noting that (a) implies that

$$\hat{L}_T(\hat{\hat{\theta}}_T) \geq \hat{L}_T(\theta_0)$$

almost surely for large enough T.

□

Proof of Theorem 2.4.3. The expectation $E \log p^o(y_0|y^{-1})$ exists and is finite by **M1** and moreover $E \log p(y_0|\tilde{f}_0(\theta), \theta)$ exists for any $\theta \in \Theta_0$ by **M3**. This implies that the marginal KL divergence $KL(\theta)$ is well defined for any $\theta \in \Theta_0$. The condition **M2** guarantees that $L(\theta)$ has a unique maximizer in Θ_0, which is denoted by θ^*. This implies that θ^* is the unique minimizer of the average KL divergence $KL(\theta)$. As concerns the consistency result, replacing θ_0 with θ^*, the proof is equivalent to the the proof of Theorem 2.4.2. This can be easily seen as the step (S1) holds by assumption replacing θ_0 with θ^*. Then, the steps (S2)-(S4) do not rely on the correct specification of the model and the consistency is obtained with respect to maximizer of the limit function L, which in this case is given by θ^*.

□

Proof of Proposition 2.5.1. For any $\theta \in \Theta$, the random coefficient $\Lambda_t(\theta)$ is a measurable function of Y_t^k for any given $k \in \mathbb{N}$. Therefore, as $\{y_t\}_{t\in\mathbb{Z}}$ is α-mixing of size $-2r/(r-2)$, it results that $\{\log \Lambda_t(\theta)\}_{t\in\mathbb{Z}}$ is α-mixing of size $-2r/(r-2)$ as well, see for instance

Theorem 14.1 in Davidson (1994). Given the convergence in probability of $\hat{\sigma}_T^2$ to

$$\lim_{T\to\infty} \text{Var}\left(T^{-1/2}\sum_{i=1}^{T}\log\Lambda_t(\theta)\right)$$

and accounting that $E|\log\Lambda_t(\theta)|^r < \infty$, the asymptotic normality result then follows immediately by an application of a central limit theorem for strong mixing processes (see for instance Theorem 7.8 of Durrett (2004)) together with an application of Slutsky's theorem. □

Proof of Theorem 2.6.1. First note that the model equation $f_{t+1}^o = \omega_0 + f_t^o c_t$ is a SRE of the form $f_{t+1}^o = \psi_t(f_t^o)$, where $\psi_t(x) := \omega_0 + xc_t$ for any $x \in [0,\infty)$. Therefore, $\{\psi_t\}_{t\in\mathbb{Z}}$ is a stochastic sequence of maps from $[0,\infty)$ into $[0,\infty)$. The proof of the if part of the theorem follows noting that the condition $E\log c_t < 0$ is sufficient to satisfy the assumptions of Theorem 3.1 in Bougerol (1993). In particular, the first assumption is satisfied as $E|\omega_0 + xc_t| < \infty$ for any $x \in [0,\infty)$ whereas the second assumption immediately holds by $E\log c_t < 0$.

As concerns the only if part, we consider a similar argument as in Bougerol and Picard (1992). In particular, we show that if $\{f_t^o\}_{t\in\mathbb{Z}}$ is a stationary and ergodic solution of (2.1), then $E\log c_t$ has to be strictly negative. From the recursion

$$f_t^o = \omega_0\left(1 + \sum_{k=1}^{n-1}\prod_{i=1}^{k} c_{t-i}\right) + \prod_{i=1}^{n} c_{t-i} f_{t-n}^o,$$

it follows that almost surely the following inequality holds

$$\sum_{k=1}^{n-1}\prod_{i=1}^{k} c_{t-i} \leq f_t^o, \quad \forall\, n \in \mathbb{N}.$$

This means that $\lim_{n\to\infty}\sum_{k=1}^{n-1}\prod_{i=1}^{k} c_{t-i}$ has to be finite almost surely and therefore $\prod_{i=1}^{k} c_{t-i}$ has to converge almost surely to zero as $k \to \infty$. As $\{c_t\}_{t\in\mathbb{Z}}$ is an i.i.d sequence of random variables, the almost sure convergence to zero of $\prod_{i=1}^{k} c_{t-i}$ implies that $E\log c_t$ is strictly negative by lemma 2.1 of Bougerol and Picard (1992). This concludes the proof of the theorem. □

Proof of Theorem 2.6.2. When the process admits a stationary solution, the following representation holds true

$$f_t^o = \omega_0\left(1 + \sum_{k=1}^{\infty}\prod_{i=1}^{k} c_{t-i}\right).$$

In the case $z \in [1, \infty)$, by the Minkowski inequality and considering that $\{c_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of positive random variables, we have that

$$(E(f_t^o)^z)^{1/z} \leq \omega_0 \left(1 + \sum_{k=1}^{\infty} (Ec_{t-i}^z)^{k/z}\right).$$

Therefore, when $Ec_{t-i}^z < 1$, the result $E(f_t^o)^z < \infty$ follows from the convergence of the series $\sum_{k=1}^{n}(Ec_{t-i}^z)^{k/z}$. As concerns the case $z \in [0, 1)$, by sub-additivity we have that

$$E(f_t^o)^z \leq \omega_0^z \left(1 + \sum_{k=1}^{\infty} (Ec_{t-i}^z)^{k}\right).$$

Then, as before, the desired result follows from the convergence of the series $\sum_{k=1}^{n}(Ec_{t-i}^z)^k$. □

Proof of Theorem 2.6.3. First note that the expression of the probability density function of a student-t random variable with v degrees of freedom is

$$k_v(x) = s(v)(1 + v^{-1}x^2)^{-(v+1)/2},$$

where

$$s(v) = \frac{\Gamma\left(2^{-1}(v+1)\right)}{\sqrt{v\pi}\Gamma\left(2^{-1}v\right)},$$

and where Γ denotes the gamma function.

In the following we check that the conditions **C1-C5** are satisfied, then the proof follows by an application of Theorem 2.4.1.

(C1) The stationarity and ergodicity of the sequence $\{y_t\}_{t\in\mathbb{Z}}$ is a direct consequence of Theorem 2.6.1. In the following, we prove that the integrability condition $E|l_0(\theta_0)| \leq \infty$ is satisfied. First, note that $l_0(\theta_0)$ is given by

$$l_0(\theta_0) = \log s(v_0) - \frac{1}{2}\log f_0^o - \frac{v_0+1}{2}\log\left(1 + v_0^{-1}\varepsilon_0^2\right),$$

therefore we just need to show that $E|\log f_0^o| < \infty$ holds. Consider a decreasing sequence of numbers $\{\epsilon_i\}_{i\in\mathbb{N}}$, $\epsilon_i > 0$, such that $\lim_{i\to\infty}\epsilon_i = 0$, then $\{(c_t^{\epsilon_i} - 1)/\epsilon_i\}_{i\in\mathbb{N}}$ is a decreasing sequence of random variables such that $\lim_{i\to\infty}(c_t^{\epsilon_i} - 1)/\epsilon_i = \log c_t$. An application of the monotone convergence theorem leads to

$$\lim_{i\to\infty} E\left(\frac{c_t^{\epsilon_i} - 1}{\epsilon_i}\right) = E\log c_t.$$

Therefore if $E\log c_t < 0$, then there exists an $\bar{\epsilon} > 0$ such that $E(c_t^{\bar{\epsilon}} - 1)/\bar{\epsilon} < 0$ and thus $Ec_t^{\bar{\epsilon}} < 1$. In virtue of Theorem 2.6.2, $E(f_t^o)^{\bar{\epsilon}} < \infty$ and thus we have that $E\log^+ f_t^o < \infty$. The desired result follows as $f_t^o \geq \omega_0/(1-\beta_0) > 0$ a.s. and therefore $E\log^+ f_t^o < \infty$ implies $E|\log f_t^o| < \infty$.

(C2) Note that $a_1 k_{v_1}(a_1 x) = a_2 k_{v_2}(a_2 x)$ for any $x \in \mathbb{R}$ if and only if $(v_1, a_1) = (v_2, a_2)$. Therefore, if $\varepsilon_0 \sim t_v$ then $a_1 k_{v_1}(a_1\varepsilon_0) = a_2 k_{v_2}(a_2\varepsilon_0)$ a.s. if and only if $(v_1, a_1) = (v_2, a_2)$ as ε_0 is an absolutely continuous random variable with a positive density function on $\mathbb{R}$. As a result, considering that $l_0(\theta_0) = l_0(\theta)$ a.s. if and only if

$$k_{v_0}(\varepsilon_0) = \sqrt{\frac{f_0^o}{\tilde{f}_0(\theta)}} k_v\left(\sqrt{\frac{f_0^o}{\tilde{f}_0(\theta)}}\varepsilon_0\right) \text{ a.s.},$$

we have that $l_0(\theta_0) = l_0(\theta)$ a.s. if and only if $v = v_0$ and $f_0^o = \tilde{f}_0(\theta_0)$ a.s.. This means that the non-trivial implication $l_0(\theta_0) = l_0(\theta)$ a.s. only if $\theta = \theta_0$ is satisfied if we can show that, given $v = v_0$, $f_0^o = \tilde{f}_0(\theta)$ a.s. only if $\theta = \theta_0$. Considering that the sequence $\{\tilde{f}_t\}_{t\in\mathbb{Z}}$ is stationary, we have that $f_0^o = \tilde{f}_0(\theta)$ a.s. is the same as $f_t^o = \tilde{f}_t(\theta)$ a.s. for any $t \in \mathbb{Z}$. Assuming $f_t^o = \tilde{f}_t(\theta)$ a.s., the difference $f_{t+1}^o - \tilde{f}_{t+1}(\theta)$ satisfies

$$f_{t+1}^o - \tilde{f}_{t+1}(\theta) = \omega_0 - \omega + f_t^o z_t,$$

$$z_t = \beta_0 - \beta + \Big(\alpha_0 - \alpha + (\gamma_0 - \gamma)d_t\Big)(v_0 + 1)b_t,$$

where $b_t = \varepsilon_t^2/(v_0 - 2 + \varepsilon_t^2)$. Now, the first step is to show that if $f_{t+1}^o - \tilde{f}_{t+1}(\theta) = 0$ a.s., then $\omega_0 = \omega$, the proof is by contradiction. Assume that $\omega_0 \neq \omega$ and $f_{t+1}^o - \tilde{f}_{t+1}(\theta) = 0$ a.s., then it must be that $f_t^o z_t = \omega - \omega_0 \neq 0$ a.s.. Noting that f_t^o is independent of z_t, the only way this is possible is if both f_t^o and z_t are constants different from zero. However, the possibility that f_t^o has a degenerate distribution is ruled out by $\alpha_0 > 0$, therefore $\omega = \omega_0$. As $\omega = \omega_0$ and f_{t+1}^o is non-zero with probability 1, the only way to have $f_{t+1}^o - \tilde{f}_{t+1}(\theta)$ a.s. is if $z_t = 0$ a.s.. The second step is to show that we need also $\beta = \beta_0$. Using the same argument as before, to have $\beta \neq \beta_0$ and $z_t = 0$ a.s. the random variable b_t has to be constant as b_t is independent of d_t. However, b_t is non-constant for any $v_0 \in (2, +\infty)$. Therefore, we have that $\beta = \beta_0$. Finally, having $\beta = \beta_0$, to have $z_t = 0$ a.s. it must be that $\big(\alpha_0 - \alpha + (\gamma_0 - \gamma)d_t\big) = 0$ a.s.. Indeed, as d_t is non-constant, this is possible only if $\alpha = \alpha_0$ and $\gamma = \gamma_0$. This concludes the proof.

(C3) This condition is immediately satisfied by Corollary 2.6.1.

(C4) From the expression of $l_t(\theta)$ and by an application of the mean value theorem, it

results that

$$|\hat{l}_t(\theta) - l_t(\theta)| \leq |r_t(\theta)||\hat{f}_t(\theta) - \tilde{f}_t(\theta)|,$$

for any $\theta \in \Theta$ and any $t \in \mathbb{N}$. The stochastic coefficient $r_t(\theta)$ has the following expression

$$r_t(\theta) = 2^{-1} f_t^*(\theta)^{-1} \left(\frac{(v+1)v^{-1} f_t^*(\theta) y_t^2}{1 + v^{-1} f_t^*(\theta) y_t^2} - 1 \right),$$

where $f_t^*(\theta)$ a point between $\tilde{f}_t(\theta)$ and $\hat{f}_t(\theta)$. Considering that $\tilde{f}_t(\theta)$ and $\hat{f}_t(\theta)$ lie in the set $[c, +\infty)$, $c = \inf_{\theta \in \Theta} \omega/(1-\beta) > 0$, it results that

$$\|\hat{l}_t - l_t\|_\Theta \leq \|r_t\|_\Theta \|\hat{f}_t - \tilde{f}_t\|_\Theta \leq \bar{r} \|\hat{f}_t - \tilde{f}_t\|_\Theta,$$

where

$$\bar{r} = 2^{-1} c^{-1} \left(1 + c^{-1} \left(\max_{\theta \in \Theta} v + 1 \right) \right).$$

This shows that C4 is satisfied setting $\eta_t = \bar{r}$ for any $t \in \mathbb{N}$.

(C5) In view of $\tilde{f}_0(\theta) \geq \inf_{\theta \in \Theta} \omega/(1-\beta) > 0$ a.s. for any $\theta \in \Theta$, it results that

$$\sup_{\theta \in \Theta} l_0(\theta) \leq \sup_{\theta \in \Theta} s(v) - \frac{1}{2} \log \left(\inf_{\theta \in \Theta} \omega/(1-\beta) \right) < \infty,$$

with probability 1. This proves the desired result $E\|l_0 \vee 0\|_\Theta < \infty$. □

Proof of Theorem 2.6.4. The proof follows by showing that conditions **A1**-**A3** hold true. Condition **A1** is satisfied as the stationarity and ergodicity of the sequence $\{y_t\}_{t \in \mathbb{Z}}$ is ensured by Theorem 2.6.1 and the integrability condition $E|l_0(\theta_0)| < \infty$ can be shown in the same way as in the proof of Theorem 2.6.3, see the step C1. Condition **A2** is satisfied as the conditions (i) and (ii) hold by Corollary 2.6.1 and the continuity assumption follows immediately from the functional form of the Lipschitz coefficient in Corollary 2.6.1 and the constraints imposed on the compact set $\bar{\Theta}$. Finally, Condition **A3** is satisfied as the steps **C2**, **C4** and **C5** in the proof of Theorem 2.6.3 hold for any compact set satisfying the contraction condition in Corollary 2.6.1. □

Chapter 3

INAR models with Dynamic Survival Probability driven by a Stochastic Recurrence Equation

3.1 Introduction

Over the last few years, there has been an increasing interest in modeling and forecasting integer-valued time series. The reason being that many observed time series are not continuous and the use of specific models to take this into account allows us to better describe the time series behavior. One of the most popular models for time series of counts is the INAR model of Al-Osh and Alzaid (1987) and McKenzie (1988). Its specification is based on the thinning operator '$\circ$' of Steutel and Van Harn (1979). For a given $N \in \mathbb{N}$ and $\alpha \in (0,1)$, the thinning operator is defined to satisfy $\alpha \circ N = \sum_{i=1}^{N} x_i$, where $\{x_i\}_{i=1}^{N}$ is a sequence of independent Bernoulli random variables with success probability α. The thinning operator enables the specification of integer-valued time series models in an autoregressive fashion. In fact, INAR models can be seen as a discrete response version of the well known linear autoregressive model. The first order INAR model is described by the following equation

$$y_t = \alpha \circ y_{t-1} + \varepsilon_t, \ t \in \mathbb{Z}, \tag{3.1}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of integer-valued random variables. As in the original formulation of Al-Osh and Alzaid (1987) and McKenzie (1988), the error term ε_t is typically assumed to be Poisson distributed. Other distributions have also been considered in

the literature as the Poisson imposes equidispersion and this is can be restrictive in practice, see Al-Osh and Aly (1992) and Jazi et al. (2012). Besides the distribution of the error term, the INAR specification in (3.1) has been generalized in several directions. Among others, Alzaid and Al-Osh (1990) and Jin-Guan and Yuan (1991) extended the first order INAR model to a general order p, Alzaid and Al-Osh (1990) considered a generalized thinning operator and Pedeli and Karlis (2011) introduced a bivariate INAR model.

Real time series data often exhibit changing dynamic behaviors. As a result, employing more flexible specifications for the dynamic component of the model can provide a better description the underlying behavior of the time series and produce better forecasts. The contribution of this chapter is in this direction: we introduce a new class of INAR models with time-varying coefficient. The peculiarity of our approach is that the dynamics of the INAR coefficient is specified through a SRE driven by the score of the predictive likelihood. The use of the score to update time-varying parameters has been recently proposed by Creal et al. (2013) and Harvey (2013). Since then, the score framework has been successfully employed to develop dynamic models in econometrics and time series analysis.

In the literature, time variation of the INAR survival probability has also been considered by Zheng et al. (2007) and Zheng and Basawa (2008). In Zheng et al. (2007) the random coefficient is specified as a sequence of i.i.d. random variables. This approach provides a more flexible class of conditional distributions but, because of the i.i.d. assumption, it does not lead to a dynamic specification of the INAR coefficient. Zheng and Basawa (2008) allowed the INAR coefficient to depend on past observations. Their method introduces a dynamic structure but, as we will see discuss later in this chapter, it is not able to properly capture smooth changes of the INAR coefficient.

The class of models we introduce in this chapter should not be interpreted as a DGP but as filter to approximate a more complex and unknown DGP (Blasques et al., 2015). In this direction, we illustrate the flexibility of the proposed dynamic specification for the INAR coefficient by means of a simulation study in a misspecified framework. The results show how the model is able to capture complex dynamic behaviors and well approximate the true distribution of different DGPs. Furthermore, we derive some statistical properties of the ML estimator: we prove the consistency of the ML estimator in a misspecified framework and show that also the conditional predictive pmf can be consistently estimated through a plug-in estimator. In particular, the plug-in pmf estimator is shown to converge to a pseudo-true conditional pmf that has the interpretation of minimizing on average the KL divergence with the true pmf of the DGP. These results are useful to ensure the reliability of inference and forecasting. Finally, the practical usefulness of the proposed

model is shown thorough an application to a real time series dataset of crime reports. The results are promising and show how the dynamic survival probability can enhance both in sample and out-of-sample performances of INAR models.

The chapter is structured as follows. Section 3.2 introduces the class of models. Section 3.3 discusses the consistency of ML estimation. Section 3.4 presents the Monte Carlo simulation experiments. Section 3.5 illustrates the empirical application. Section 3.6 concludes.

3.2 INAR models with autoregressive coefficient

3.2.1 The class of models

In this section, we extend the class of INAR models in (3.1) by allowing the survival probability α to change over time. The dynamics of the time-varying coefficient α_t is specified on the basis of the score framework of Creal et al. (2013) and Harvey (2013). The tv-INAR model is described by the following equations

$$y_t = \alpha_t \circ y_{t-1} + \varepsilon_t, \tag{3.2}$$

$$\operatorname{logit} \alpha_{t+1} = \omega + \beta \operatorname{logit} \alpha_t + \tau s_t, \tag{3.3}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of random variables with probability mass function (pmf) $p_e(x,\xi)$, $\xi \in \Xi \subseteq \mathbb{R}^k$, for $x \in \mathbb{N}$, the vector $\theta = (\omega, \beta, \tau, \xi)^T$ is a $k+3$ dimensional parameter vector to be estimated and $s_t = s_t(\alpha_t, \xi)$ denotes the score of the predictive log-likelihood $\partial \log p(y_t|\alpha_t, y_{t-1}, \xi)/\partial \operatorname{logit} \alpha_t$. The functional form of the predictive likelihood $p(y_t|\alpha_t, y_{t-1}, \xi)$ can be obtained by the convolution between the conditional pmf of $\alpha_t \circ y_{t-1}$ and the pmf of the error term ε_t, i.e.

$$p(y_t|\alpha_t, y_{t-1}, \xi) = \sum_{k=0}^{\max\{y_t, y_{t-1}\}} p_b(k, y_{t-1}, \alpha_t) p_e(y_t - k, \xi),$$

where $p_b(x, y_{t-1}, \alpha_t)$ for $x \in \{0, \dots, y_{t-1}\}$ is the pmf of a Binomial random variable with size y_{t-1} and success probability α_t. An analytical expression of the score innovation s_t is given in the Appendix. The logit link function in equation (3.3) is considered to ensure that the dynamic coefficient α_t is between zero and one.

The dynamic tv-INAR model in (3.2) and (3.3) retains the well known interpretation of INAR models as death-birth processes. In particular, the observed number of elements

y_t alive at time t is given by the sum between the number of surviving elements from time $t-1$ and the new birth elements ε_t. In our dynamic specification, each of the elements alive at time $t-1$ has a probability α_t of surviving at time t. We also note that the proposed model is observation-driven as the dynamic probability α_t is driven solely by past observations. The score s_t can be seen as the innovation of the dynamic system in (3.3) as it provides the new information that becomes available at time t observing y_t. The interpretation of s_t as an innovation is further justified by the fact that its conditional expectation $E(s_t|y_{t-1}, \alpha_t)$ is equal to zero.

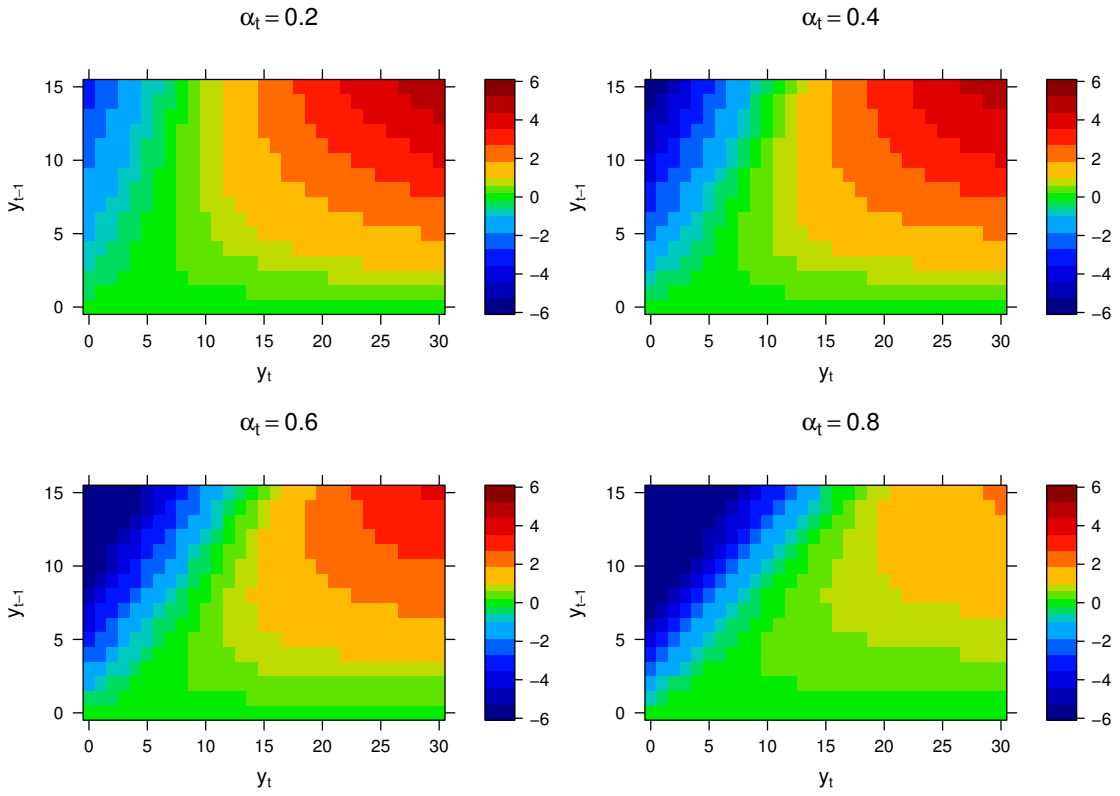

Figure 3.2.1: *The plots represent the impact of y_t and y_{t-1} on the score innovation s_t for different values of the survival probability α_t. A Poisson distribution with mean equal to five is considered as distribution of the error term ε_t.*

It is interesting to see how the information obtained observing y_t is processed through the score to update the dynamic coefficient α_t. Figure 3.2.1 describes the impact of y_t on s_t for different values of y_{t-1} and α_t. As we can see from the plots, the survival probability α_t gets a negative update when y_t is small and y_{t-1} is large. This has an intuitive explanation: the information about α_t we get observing a small y_t after a large y_{t-1} is that the survival probability should be small as otherwise with a large α_t we wold expect many elements from time $t-1$ to survive and thus a large y_t as well. As a result,

α_t should get a negative update to discount this information. Similarly, observing a large y_t following a large y_{t-1} suggests an high survival probability. Thus, the probability α_t should be updated accordingly and get a positive innovation s_t. Finally, an innovation s_t close to zero may be an indication of either a lack of information or that the observed value of y_t is compatible with the value y_{t-1} and the current state of the survival probability α_t. The former case reflects situations when y_{t-1} is zero (or close to zero). This because observing y_t provides no information on the survival probability of the elements y_{t-1} as there are no elements alive at $t-1$. On the other hand, the latter case of observing a value y_t compatible with y_{t-1} and α_t can be interpreted as the green area that separates the red and the blue areas in Figure 3.2.1.

This line of reasoning concerning the direction of the update s_t is subject to the current value of α_t. For instance, in a situation where α_t is close to zero perhaps observing a small y_t after a large y_{t-1} is exactly what we would expect. Thus the score update s_t may be close to zero in this case. This dependence of the score update s_t on the current survival probability α_t can be noted across the different plots in Figure 3.2.1.

It is also worth mentioning that the functional form of the score innovation s_t depends on the specification of the pmf of the error term ε_t as the predictive likelihood depends on it. In practice, the pmf $p_e(x, \xi)$ can be chosen in such a way to take into account the main features observed in the data. For instance, as we will consider in the application in Section 3.5, a Negative Binomial distribution may be considered instead of a Poisson when the data suggest overdispersion. Alternatively, a zero inflated Poisson or Negative Binomial distributions may be employed when dealing with time series with a large number of zeros.

3.2.2 Parameter estimation

The static parameter vector θ of the tv-INAR model can be estimated by ML. The log likelihood function is available in closed form through a prediction error decomposition, namely

$$\hat{L}_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t | \hat{\alpha}_t(\theta), y_{t-1}, \xi),$$

where $\hat{\alpha}_t(\theta)$ is obtained recursively using the observed data $\{y_t\}_{t=0}^{T}$ as

$$\operatorname{logit} \hat{\alpha}_{t+1}(\theta) = \omega + \beta \operatorname{logit} \hat{\alpha}_t(\theta) + \tau s_t(\hat{\alpha}_t(\theta), \xi), \tag{3.4}$$

where the recursion is initialized at a fixed point $\operatorname{logit}\hat{\alpha}_0(\theta) \in \mathbb{R}$. A reasonable choice for the initialization is $\operatorname{logit}\hat{\alpha}_0(\theta) = \omega/(1-\beta)$. That is the unconditional mean $E\operatorname{logit}\alpha_t$ implied by the tv-INAR model under the parametric assumption θ. This follows as the expected value of the score is equal to zero. The ML estimator is finally given by

$$\hat{\theta}_T = \arg\sup_{\theta\in\Theta} \hat{L}_T(\theta), \tag{3.5}$$

where Θ is a compact parameter set contained in $\mathbb{R} \times (-1,1) \times \mathbb{R} \times \Xi$.

The asymptotic stability of the filtered parameter $\operatorname{logit}\hat{\alpha}_t(\theta)$ and the consistency of the ML estimator in as well as the predictive distribution are studied in Section 3.3. Furthermore, in Section 3.4, a simulation experiment is performed to study the finite sample behavior of the ML estimator and to further confirm its reliability.

3.2.3 Forecasting

One of the advantages of properly modeling count time series taking into account the discreteness of the data is that it is possible to obtain coherent forecasts of the entire probability mass function. As shown in Freeland and McCabe (2004), forecasts h steps ahead are typically available in closed form for INAR models as (3.1). The conditional pmf h steps ahead can be obtained by repeated applications of the convolution formula. Similarly, for point forecasts, a closed form expression is available as the conditional expectation h steps ahead is $E(y_{T+h}|y_T) = \alpha^h y_T + \mu$, with $\mu = E(\varepsilon_t)$.

In the following, we illustrate a possible way to obtain h steps ahead forecasts from the tv-INAR model. A closed form expression for the conditional pmf h steps ahead $p_{T+h}(x)$ is only available for $h = 1$. In particular, it is given by

$$p_{T+1}(x) = \sum_{k=0}^{\min\{x,y_T\}} p_b(k, y_T, \alpha_T) p_e(x-k).$$

Numerical methods are required to obtain $p_{T+h}(x)$ for $h \geq 2$. A possibility is to approximate $p_{T+h}(x)$ considering the following simulation scheme. First, simulate B realization for y_{T+h}, $y_{T+h}^{(i)}$, $i = 1, \ldots, B$. Then, obtain an approximation of $p_{T+h}(x)$ as $\hat{p}_{T+h}(x) = n_x^h/B$, where n_x^h denotes the number of draws $y_{T+h}^{(i)}$, $i = 1, \ldots, B$, equal to x. The simulations of $y_{T+h}^{(i)}$, $i = 1, \ldots, B$ can be performed considering the following procedure. For $k = 1, \ldots, h$

1. Simulate $\varepsilon_k^{(i)}$ from the distribution $p_e(x, \xi)$ and $\alpha_{T+k}^{(i)} \circ y_{T+k-1}^{(i)}$ from a Binomial

distribution with size $y_{T+k-1}^{(i)}$ and success probability $\alpha_{T+k}^{(i)}$.

2. Compute $y_{T+k}^{(i)} = \alpha_{T+k}^{(i)} \circ y_{T+k-1}^{(i)} + \varepsilon_k^{(i)}$ and update $\alpha_{T+k}^{(i)}$ to $\alpha_{T+k+1}^{(i)}$ using the equation $\operatorname{logit} \alpha_{T+k+1}^{(i)} = \omega + \beta \operatorname{logit} \alpha_{T+k}^{(i)} + \tau s_{t+k}^{(i)}$.

Similarly, point forecasts h steps ahead can be obtained approximating the conditional expectation $E(y_{T+h}|y_T, \alpha_t)$ with the sample average $B^{-1} \sum_{i=1}^{B} y_{T+h}^{(i)}$. Alternatively, the sample median of $y_{T+h}^{(i)}$, $i = 1, \ldots, B$ can be considered to obtain integer forecasts that are coherent with the discreteness of the data, see Freeland and McCabe (2004).

3.3 Some statistical properties

3.3.1 Stability of the filter

In this section, we discuss the reliability of the ML estimator defined in (3.5). In particular, we show that the static parameter vector as well as the conditional pmf can be consistently estimated. We focus our asymptotic results on the case of model misspecification. Consistency is therefore obtained with respect to a pseudo-true parameter that has the interpretation of minimizing an average KL divergence between the postulated INAR model and a true unknown DGP. Score-driven models are typically not interpreted as DGP but as filters to approximate a more complex and unknown true DGP. In this regard, Blasques et al. (2015) provided a theoretical justification to score-driven models by showing their optimality in a misspecified setting. In the following, we only assume that the observed data are generated by a stationary and ergodic count process. Therefore, we do not impose a specific DGP for the observed data.

A key ingredient to ensure the reliability of the ML estimator for observation-driven models is the stability of the filtered time-varying parameter. This stability is typically referred in the literature as the invertibility of the model, see Wintenberger (2013) and Straumann and Mikosch (2006). In the following, we derive conditions to ensure that the filtered parameter in (3.4) converges to a unique stationary sequence irrespective of the initialization $\hat{\alpha}_0(\theta)$. This result is particularly important as it implies that the initialization is irrelevant asymptotically and provides the basis to ensure the consistency of the ML estimator.

First, we impose some regularity conditions on the pmf of the error term $p_e(x, \xi)$.

Assumption 3.3.1. *The function $\xi \mapsto p_e(x, \xi)$ is continuous in Ξ for any $x \in \mathbb{N}$ and $p_e(x, \xi) > 0$ for any $(x, \xi) \in \mathbb{N} \times \Xi$.*

Assumption 3.3.1 requires the pmf $p_e(x, \xi)$ to have full support in $\mathbb{N}$ and to be continuous with respect to ξ. These conditions are satisfied for most parametric pmf such as the Poisson, the zero inflated Poisson and the Negative Binomial. However, it is worth mentioning that distributions with limited support such as the Binomial are ruled out by this assumption.

The next result ensures the stability of the filtered parameter $\{\hat{\alpha}_t(\theta)\}_{t\in\mathbb{N}}$ specified in (3.4). In particular, it shows the almost sure uniform convergence of the functional sequence $\{\hat{\alpha}_t\}_{t\in\mathbb{N}}$ to a unique stationary and ergodic functional sequence $\{\tilde{\alpha}_t\}_{t\in\mathbb{Z}}$. The convergence is considered with respect to the uniform norm $\|\cdot\|_\Theta$, where $\|f\|_\Theta = \sup_{\theta\in\Theta}|f(\theta)|$ for any function f that maps from Θ into $\mathbb{R}$.

Proposition 3.3.1. *Assume that $\{y_t\}_{t\in\mathbb{Z}}$ is a stationary and ergodic sequence of random variables that take values in $\mathbb{N}$ and such that $Ey_t^2 < \infty$. Moreover, let Assumption 3.3.1 be satisfied and the following condition hold*

$$E \log \sup_{\alpha\in(0,1)} |\beta + \tau \dot{s}_t(\alpha, \xi)| < 0, \quad \forall\, \theta \in \Theta, \tag{3.6}$$

where $\dot{s}_t(\alpha, \xi) = \partial s_t(\alpha, \xi)/\partial \operatorname{logit}\alpha$. Then, the filtered parameter $\{\hat{\alpha}_t(\theta)\}_{t\in\mathbb{N}}$ defined in (3.4) converges exponentially almost surely and uniformly in Θ to a unique stationary and ergodic sequence $\{\tilde{\alpha}_t(\theta)\}_{t\in\mathbb{Z}}$, i.e.

$$\|\operatorname{logit}\hat{\alpha}_t - \operatorname{logit}\tilde{\alpha}_t\|_\Theta \xrightarrow{e.a.s.} 0 \quad as \quad t \to \infty.$$

The proof is given in the appendix. Proposition 3.3.1 does not require correct specification of the model. The observed data can be generated by any stationary and ergodic count process.

The contraction condition in (3.6) can be checked empirically using the observed data. It is not possible to obtain a closed form expression for (3.6) as it depends on the DGP and on the specification of $p_e(x, \xi)$. However, with the next proposition, we show that the parameter region Θ that satisfies (3.6) is not degenerate.

Proposition 3.3.2. *The contraction condition* (3.6) *of Proposition 3.3.1 is implied by the following sufficient condition*

$$E \log \max(|\beta - \tau y_{t-1}/4|, |\beta + \tau m_t^2|) < 0, \quad \forall\, \theta \in \Theta,$$

where $m_t = \min\{y_{t-1}, y_t\}$.

Proposition 3.3.2 guarantees that the parameter region Θ is not degenerate as for small enough $|\beta|$ and $|\tau|$ the inequality is always satisfied.

3.3.2 Consistency of ML estimation

We assume the observed data to be a realized path from an unknown DGP $\{y_t\}_{t\in\mathbb{Z}}$. Furthermore, we denote with $p^o(x|y^{t-1})$, $x \in \mathbb{N}$, the true pmf of y_t conditionally on the past observations $y^{t-1} = \{y_{t-1}, y_{t-2}, \dots\}$. The KL divergence between the true conditional pmf $p^o(x|y^{t-1})$ and the postulated one $p(x|\tilde{\alpha}_t(\theta), y_{t-1}, \xi)$ is given by

$$KL_t(\theta) = \sum_{x=0}^{\infty} \log\left(\frac{p^o(x|y^{t-1})}{p(x|\tilde{\alpha}_t(\theta), y_{t-1}, \xi)}\right) p^o(x|y^{t-1}).$$

We define the pseudo-true parameter $\theta^* \in \Theta$ as the minimizer of the average KL divergence $KL(\theta) = EKL_t(\theta)$ in the parameter set Θ. We also denote with $\alpha_t^* = \tilde{\alpha}_t(\theta^*)$ the pseudo-true time-varying coefficient and with $p_t^*(x) = p(x|\alpha_t^*, y_{t-1}, \xi^*)$, $x \in \mathbb{N}$, the pseudo-true conditional pmf. In the following, we treat also the consistency of the plug-in estimators $\hat{\alpha}_t(\hat{\theta}_T)$ and $\hat{p}_t(x, \hat{\theta}_T) = p(x|y_{t-1}, \hat{\alpha}_t(\hat{\theta}_T), \hat{\xi}_T)$ for α_t^* and $p_t^*(x)$ respectively. This is of particular interest in practice as the main objective of INAR models is not the interpretation of the static parameter estimates but approximating the true pmf for forecasting purposes.

We start considering the following assumption, which imposes some moment conditions and the contraction condition of Proposition 3.3.1.

Assumption 3.3.2. *The moment conditions $Ey_t^2 < \infty$ and $E\sup_{\theta\in\Theta} |\log p_e(y_t, \xi)| < \infty$ hold true. Furthermore, the contraction condition in (3.6) is satisfied.*

Assumption 3.3.2 enables us to ensure the uniform a.s. convergence of the likelihood function $\hat{L}_T(\theta)$ to a well defined deterministic function $L(\theta) = El_0(\theta)$, where $l_t(\theta) = \log p(y_t|\tilde{\alpha}_t(\theta), y_{t-1}, \xi)$ denotes the t-th contribution to the likelihood function when the limit filter $\tilde{\alpha}_t(\theta)$ is considered. Note that the uniform moment condition $E\sup_{\theta\in\Theta} |\log p_e(y_t, \xi)|$ is needed only because we are considering a general class of pmf for the error term. For most pmf this condition is always satisfied. For instance, it holds true immediately as long as $Ey_t^2 < \infty$ if $p_e(x, \xi)$ is a Poisson or a Negative Binomial pmf.

Finally, we impose the following identifiability condition.

Assumption 3.3.3. *The function $L(\theta) = El_0(\theta)$ has a unique maximizer in the set Θ.*

Assumption 3.3.3 is needed to ensure the uniqueness of the pseudo-true parameter θ^*. In general, if this assumption is not satisfied, we obtain that the limit points of the ML estimator belong to the set of points that minimize the average KL divergence $KL(\theta)$.

We are now ready to deliver the strong consistency of the ML estimator with respect to the pseudo-true parameter θ^*.

Theorem 3.3.1. *Let the observed data $\{y_t\}_{t=1}^T$ be generated by a stationary and ergodic count process $\{y_t\}_{t\in\mathbb{Z}}$ and let Assumption 3.3.1-3.3.3 be satisfied. Then the ML estimator defined in (3.5) is strongly consistent with respect to the pseudo-true parameter θ^*, i.e.*

$$\hat{\theta}_T \xrightarrow{a.s.} \theta^*, \qquad T \to \infty.$$

As special case of Theorem 3.3.1, we could also obtain the strong consistency of the ML estimator when the model is correctly specified.

Remark 3.3.1. *If we assume that the observed data $\{y_t\}_{t=1}^T$ are generated by a stationary and ergodic process $\{y_t\}_{t\in\mathbb{Z}}$ that satisfies the model's equations (3.2) and (3.3) for $\theta = \theta_0$, $\theta_0 \in \Theta$. It can be easily shown that under Assumptions 3.3.1-3.3.3 the ML estimator is strongly consistent, i.e.*

$$\hat{\theta}_T \xrightarrow{a.s.} \theta_0, \qquad T \to \infty.$$

In the next section, the finite sample properties of the ML estimator under correct specification are investigated through a simulation study.

We now turn our attention to the study of the consistency of the plug-in estimators $\hat{\alpha}_t(\hat{\theta}_T)$ and $\hat{p}_t(x, \hat{\theta}_T)$. Note that the consistency of these estimators do not follow trivially from the consistency of $\hat{\theta}_T$. This because these estimators are random functions of $\hat{\theta}_T$ that change at different times t without converging. Therefore, it is not possible to apply a continuous mapping theorem and immediately obtain the desired consistency. The results we obtain require that both t and the sample size T go to infinity. This because $T \to \infty$ is needed for the consistency of the ML estimator and $t \to \infty$ is needed to make the effect of the initialization of the filter to vanish.

The next results show that the plug-in estimator $\hat{\alpha}_t(\hat{\theta}_T)$ is strongly consistent with respect to the pseudo-true time-varying coefficient.

Lemma 3.3.1. *Let the conditions of Theorem 3.3.1 hold. Then, the plug-in estimator logit $\hat{\alpha}_t(\hat{\theta}_T)$ is strongly consistent, i.e.*

$$\left| \textit{logit}\, \hat{\alpha}_t(\hat{\theta}_T) - \textit{logit}\, \alpha_t^* \right| \xrightarrow{a.s.} 0, \qquad t \to \infty,\ T \to \infty.$$

In order to ensure the consistency of the plug-in estimator $\hat{p}_t(x, \hat{\theta}_T)$, we need the following additional regularity condition on the pmf of the error term.

Assumption 3.3.4. *The function $\xi \mapsto p_e(x, \xi)$ is continuously differentiable in Ξ for any $x \in \mathbb{N}$.*

Assumption 3.3.4 is a standard regularity condition that is satisfied for most popular pmf such as the Poisson and the Negative Binomial. The next result delivers the consistency of the conditional pmf estimator. In this case we are only able to ensure consistency and not strong consistency.

Theorem 3.3.2. *Let the observed data $\{y_t\}_{t=1}^T$ be generated by a stationary and ergodic count process $\{y_t\}_{t\in\mathbb{Z}}$ and let Assumption 3.3.1-3.3.4 be satisfied. Then the conditional pmf plug-in estimator $\hat{p}_t(x, \hat{\theta}_T)$ is consistent, i.e.*

$$|\hat{p}_t(x, \hat{\theta}_T) - p_t^*(x)| \xrightarrow{pr.} 0, \qquad t \to \infty,\ T \to \infty,$$

for any $x \in \mathbb{N}$.

3.4 Monte Carlo experiment

3.4.1 Finite sample behavior of the ML estimator

We perform a Monte Carlo simulation experiment to test the reliability of the ML estimator in finite samples. We consider the dynamic INAR model specified in (3.2) and (3.3) with a Poisson error distribution having mean μ. The experiment consists on generating 1000 time series of size T from the tv-INAR model and estimating the parameter vector $\theta = (\omega, \beta, \tau, \mu)^T$ by ML. Different parameter values θ and different sample sizes T are considered. The simulation results are collected in Table 3.4.1. In particular, Table 3.4.1 reports the mean, the bias, the standard deviation (SD) and the square root of the mean square error (MSE) of the ML estimator obtained from the 1000 Monte Carlo replications.

The simulation results in Table 3.4.1 further confirm that the parameter vector θ can be consistently estimated by maximum likelihood. This can be elicited from the fact that the MSE of the estimator is decreasing as the sample size T increases. We also note that the estimator of the parameter β tends to be negatively biased in finite samples. In all the cases considered, the parameter β is underestimated on average. The magnitude of the bias seems also to be relevant as, especially for $T = 250$, the square root of the MSE is considerably larger then the SD. Therefore, this indicates that the bias contribution to the

		ω	β	τ	μ	ω	β	τ	μ
True Value		**-0.50**	**0.90**	**0.15**	**6.00**	**-0.50**	**0.95**	**0.15**	**6.00**
$T = 250$	Mean	-0.505	0.825	0.161	5.985	-0.496	0.896	0.159	5.996
	Bias	-0.005	-0.075	0.011	-0.015	0.004	-0.054	0.009	-0.004
	SD	0.326	0.175	0.100	0.588	0.411	0.117	0.097	0.570
	$\sqrt{\text{MSE}}$	0.326	0.190	0.101	0.588	0.411	0.129	0.097	0.570
$T = 500$	Mean	-0.496	0.868	0.153	5.986	-0.503	0.927	0.154	5.997
	Bias	0.004	-0.032	0.003	-0.014	-0.003	-0.023	0.004	-0.003
	SD	0.213	0.093	0.062	0.407	0.246	0.053	0.053	0.393
	$\sqrt{\text{MSE}}$	0.213	0.098	0.062	0.407	0.246	0.058	0.053	0.392
$T = 1000$	Mean	-0.494	0.885	0.151	5.987	-0.499	0.939	0.150	5.992
	Bias	-0.006	-0.015	0.001	-0.013	-0.001	-0.011	0.000	-0.008
	SD	0.152	0.050	0.042	0.295	0.171	0.034	0.035	0.279
	$\sqrt{\text{MSE}}$	0.152	0.052	0.042	0.295	0.171	0.036	0.035	0.279
True Value		**-0.50**	**0.90**	**0.30**	**6.00**	**-0.50**	**0.95**	**0.30**	**6.00**
$T = 250$	Mean	-0.481	0.862	0.304	5.943	-0.502	0.916	0.302	5.945
	Bias	0.019	-0.038	0.004	-0.057	-0.002	-0.034	0.002	-0.055
	SD	0.361	0.095	0.101	0.512	0.501	0.066	0.097	0.473
	$\sqrt{\text{MSE}}$	0.361	0.103	0.101	0.514	0.500	0.075	0.097	0.476
$T = 500$	Mean	-0.495	0.883	0.297	5.971	-0.492	0.935	0.298	5.971
	Bias	0.005	-0.017	-0.003	-0.029	0.008	-0.015	-0.002	-0.055
	SD	0.221	0.044	0.057	0.338	0.361	0.030	0.052	0.310
	$\sqrt{\text{MSE}}$	0.221	0.048	0.057	0.339	0.361	0.033	0.052	0.311
$T = 1000$	Mean	-0.490	0.891	0.299	5.978	-0.502	0.943	0.298	5.981
	Bias	0.010	-0.019	-0.001	-0.022	-0.002	-0.007	0.002	-0.019
	SD	0.156	0.029	0.040	0.242	0.233	0.019	0.035	0.219
	$\sqrt{\text{MSE}}$	0.156	0.031	0.040	0.243	0.233	0.020	0.035	0.220

Table 3.4.1: *Summary statistics of the sample ML estimator distribution for different parameter values θ and different sample sizes T. The statistics in the table are obtained from 1000 Monte Carlo replications.*

MSE is not negligible compared to the variance contribution. The negative bias for β is not surprising as the values of β considered in the simulations are close to 1 and similar results on the bias are well known for ML estimation of linear autoregressive models. As concerns the other parameters, the results suggest that the bias can be considered negligible as the SD is almost equal to the square root of the MSE in all the scenario considered.

3.4.2 Filtering under misspecification

Score-driven updates for time-varying parameters have been shown to be optimal in a misspecified framework where the aim is to reduce the Kullback Leibler divergence between the postulated model and the true unknown DGP, see Blasques et al. (2015). This section illustrates the flexibility of the proposed specification through a simulation study. In this experiment, we consider different DGPs of the form

$$y_t = \alpha_t^o \circ y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{P}(5),$$

where $\mathcal{P}(5)$ denotes a Poisson distribution with mean $\mu = 5$. The DGPs differ on the basis of the specification of the sequence $\{\alpha_t^o\}_{t\in\mathbb{Z}}$. The following four dynamics are considered.

1. Fast sine: $\alpha_t^o = 0.5 + 0.25\sin(\pi t/100)$.
2. Slow sine: $\alpha_t^o = 0.5 + 0.25\sin(\pi t/250)$.
3. Fast steps: $\alpha_t^o = 0.25 I_{[-1,0]}\left(\sin(\pi t/100)\right) + 0.75 I_{(0,1]}\left(\sin(\pi t/100)\right)$.
4. Slow steps: $\alpha_t^o = 0.25 I_{[-1,0]}\left(\sin(\pi t/250)\right) + 0.75 I_{(0,1]}\left(\sin(\pi t/250)\right)$.

where $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ otherwise. The DGPs are thus Poisson INAR models where the coefficient α_t^o is allowed to change in different ways. The red lines in Figure 3.4.1 show the path of α_t^o, $t = 1, \ldots, 500$, for the four different DGPs. As we can see, the fast sine and the slow sine specifications allow the coefficient to change smoothly over time, whereas, the fast step and slow step specifications exhibit abrupt changes over time.

The simulation experiment consists on generating 1000 Monte Carlo time series draws of size $T = 500$ from the different DGPs. For each draw, the following models are estimated: a Poisson INAR model with static coefficient, the tv-INAR model with Poisson innovation and a Poisson INAR model with dynamic coefficient as considered in Zheng and Basawa (2008). For the latter model, the dynamic coefficient is given by $\operatorname{logit}\alpha_t = \omega + \tau y_{t-1}$, where ω and τ are parameters to be estimated. The model of Zheng and Basawa (2008) is denoted as rc-INAR. The performances of the models is measured in terms of approximation of the true condition pmf and the true survival probability α_t^o. As concerns pmf approximation, we compute the KL divergence between the true pmf and the estimated one. Whereas, as concerns α_t^o, we consider the mean square error (MSE) between α_t^o and the estimated survival probability. Table 3.4.2 reports the results of the simulation experiment. As we can see, the tv-INAR model has the pest performance for

	Square root MSE			
	Fast sine	Slow sine	Fast steps	Slow steps
INAR	0.242	0.257	0.322	0.356
rc-INAR	0.112	0.111	0.145	0.132
tv-INAR	**0.077**	**0.060**	**0.101**	**0.072**

	KL divergence			
	Fast sine	Slow sine	Fast steps	Slow steps
INAR	0.238	0.253	0.412	0.442
rc-INAR	0.117	0.114	0.212	0.185
tv-INAR	**0.053**	**0.029**	**0.128**	**0.057**

Table 3.4.2: *Average MSE and KL divergence between the true DGP and the different models.*

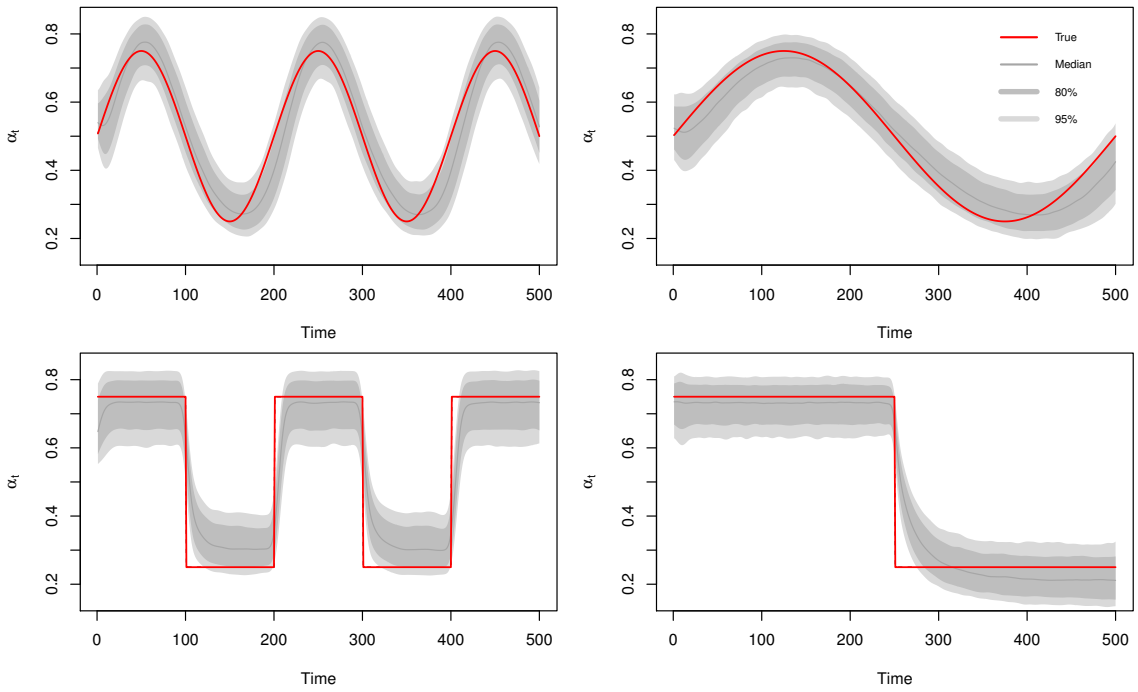

Figure 3.4.1: *The red line denotes the true path α_t^o. The gray area represents confidence bounds of the filtered path of α_t for the tv-INAR model obtained from the* 1000 *Monte Carlo replications. The first plot is for the fast sine configuration, the second is for the slow sine, the third is for the fast steps and the last is for the slow steps specification.*

both KL divergence and MSE. This is true for all the DGPs considered. We also note that the performance difference is relevant in relative terms. The KL divergence and MSE from the tv-INAR model are about half of those from the rc-INAR model and about one

third of those from the INAR model. These results show the flexibility of the tv-INAR model and its ability to approximate complex DGPs.

Figure 3.4.1 further illustrates the ability of the tv-INAR specification to capture the dynamic behavior of the true α_t^o in the different settings considered. The gray areas in the plots represents 95% confidence bounds for the filtered path of α_t and the red lines denotes the true paths α_t^o. As we can see, in the fast sine and slow sine configurations, the true path α_t^o is always inside the 95% confidence bounds. This shows the ability of the tv-INAR model capture smooth changes in α_t^o. On the other hand, in the fast steps and slow steps configurations, the true α_t^o is not inside the confidence bounds right after the sudden changes in the level of α_t^o. This is natural as the filtered path requires some time periods before adapting to break in the level of α_t^o. However, also in this situation, we can see how the filtered path from the tv-INAR model is able to approximate reasonably well the true α_t^o.

3.5 Application to crime data

3.5.1 In-sample results

We present an empirical illustration of the proposed methodology to the monthly number of offensive conduct reports in the city of Blacktown, Australia, from January 1995 to December 2014. The time series is from the New South Wales dataset of police reports and is available at `http://data.gov.au/`. Figure 3.5.1 shows the plot of the series. As we can see, there are two time periods with a particular high level of criminal activities. The first is around 2002 and the second is around 2010. During these periods we expect the estimated survival probability α_t to be higher as they can be seen as periods of high persistence. As discussed in Jin-Guan and Yuan (1991), INAR(p) models have the same autocorrelation structure of continuous-valued AR(p) models. The sample autocorrelation functions in Figure 3.5.1 suggest that a first-order INAR model should be appropriate for this dataset. We consider several model specifications: the INAR and the tv-INAR model with Poisson and Negative Binomial error distribution. The sample mean of the data is 9.3 and the sample variance is 24.3. This is an indication that there is overdispersion in the data and thus a Negative Binomial distribution for the error term may be more suited. The different specifications employed are summarized in Table 3.5.1.

The ML estimation results are collected in Table 3.5.2. We consider the likelihood ratio test to check the significance of the dynamic coefficient α_t. Given its meaningful interpretation in a misspecified framework, we also report the Akaike information criterion

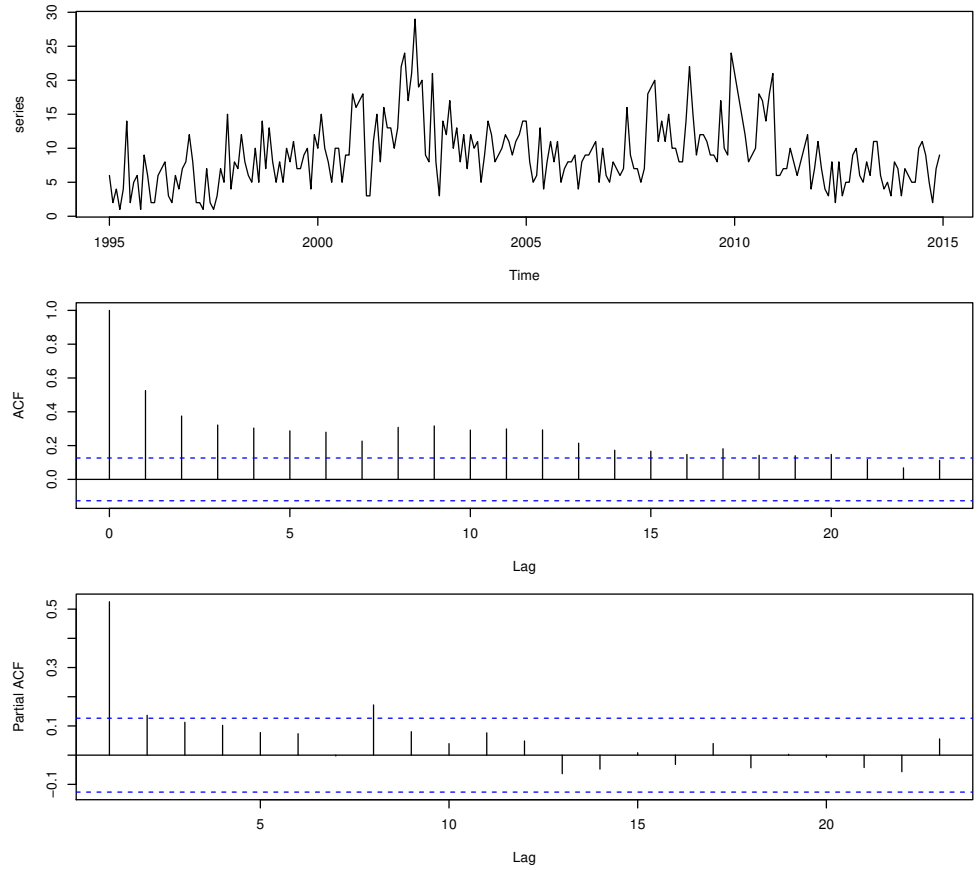

Figure 3.5.1: *The first plot shows the monthly number of offensive conduct reports in Blacktown from January 1995 to December 2014. The second and third plots represent the sample autocorrelation functions of the series.*

	Model description
tv-NBINAR	Model in (3.2) and (3.3) with Negative Binomial error of mean μ and variance σ^2.
NBINAR	Model in (3.1) with Negative Binomial error of mean μ and variance σ^2.
tv-PoINAR	Model in (3.2) and (3.3) with Poisson error of mean μ.
PoINAR	Model in (3.1) with Poisson error of mean μ.

Table 3.5.1: *The table describes the specification of each model.*

(AIC) as a means of comparison between non-nested models. The results suggest that the inclusion of the dynamic specification for α_t plays a relevant role as confirmed by the likelihood ratio test and the AIC. The likelihood ratio test shows that the dynamic coefficient is highly significant for both the Poisson and the Negative Binomial specification. Overall the model with the smallest AIC is tv-INAR model. Furthermore, for both the Negative Binomial models, the estimated variance of the error term is more than double the estimated mean. We can thus conclude that the Negative Binomial distribution provides a

	ω	β	τ	μ	σ^2	log-lik	pvalue	AIC
tv-NBINAR	-0.907 (0.338)	0.965 (0.027)	0.135 (0.055)	6.083 (0.481)	14.155 (1.853)	-662.91	0.002	1335.82
NBINAR	-0.401 (0.176)	-	-	5.586 (0.456)	15.265 (2.125)	-669.03	-	1344.07
tv-PoINAR	-1.258 (0.294)	0.967 (0.019)	0.141 (0.033)	6.539 (0.313)	-	-695.04	0.000	1398.24
PoINAR	-0.613 (0.140)	-	-	6.046 (0.323)	-	-714.58	-	1433.21

Table 3.5.2: *ML estimate of the models in Table 3.5.1. The last three columns contain respectively the log-likelihood, the pvalue of the likelihood ratio test between the tv-INAR models and their static INAR counterparts and the AIC.*

better fitting than the Poisson. This is also coherent with the overdispersion observed in the data. We can conclude that the results indicate a better fitting for the tv-INAR model. This shows that the tv-INAR model can be useful in practical applications.

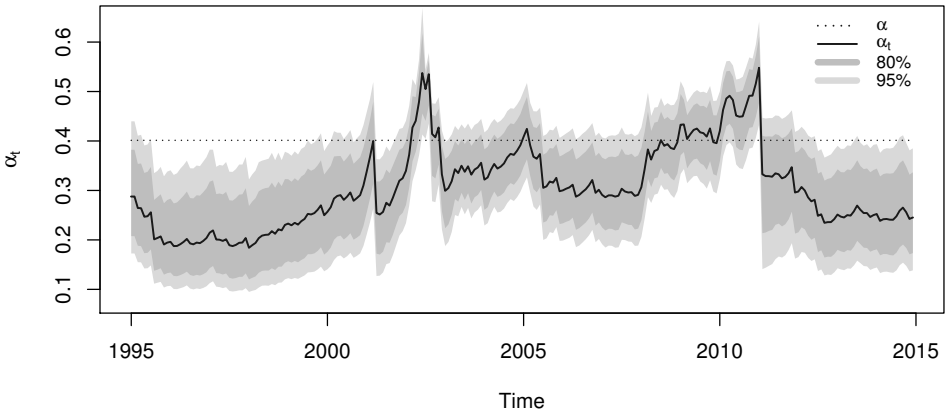

Figure 3.5.2: *Filtered time-varying coefficient α_t with 80% and 95% confidence bounds obtained from the tv-NBINAR model. The dashed line represents the static coefficient α estimated from the NBINAR model.*

From Table 3.5.2, we also note that the time-varying parameter α_t is highly persistent as the estimated β is close to 1. The filtered path of α_t together with 80% and 95% confidence bounds[1] is plotted in Figure 3.5.2. As expected, the survival probability is particularly high around 2002 and around 2010. This reflects the high level of criminal activities that can be interpreted as an higher survival probability of past elements. The plot in Figure 3.5.2 also highlights that there is a relevant difference in considering a static α instead of a dynamic α_t. This can be noted from the dashed line, which represents the

[1]The confidence bounds are obtained simulating from the asymptotic distribution of the ML estimator as proposed by Blasques et al. (2016).

static α estimate, that lies outside the 95% confidence bounds for α_t in some time periods.

3.5.2 Forecasting results

We perform a pseudo out-of-sample experiment to compare the forecasting performances of the models. The full sample size of the series is 240 observations. We split it into two subsamples: the first 140 observations are considered as a training sample and the last 100 observations as a forecasting evaluation sample. The training sample is then expanded recursively. We evaluate the forecast performance of the models in terms of both point forecast and pmf forecast. The point forecast accuracy is evaluated by the forecast MSE $100^{-1}\sum_{i=1}^{100}(\hat{y}_i - y_i)^2$. Whereas, the pmf forecast accuracy is evaluated by the log score criterion, i.e. $100^{-1}\sum_{i=1}^{100}\log\hat{p}_i(y_i)$. The log score criterion provides a means of comparison based on the KL divergence between the true DGP and the estimated models.

	Mean squared error					
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
tv-NBINAR	**15.77**	**20.15**	**20.56**	**21.51**	**21.36**	**21.23**
NBINAR	16.51	21.47	22.61	23.70	23.85	23.72
tv-PoINAR	16.33	20.66	21.18	21.98	21.82	21.52
PoINAR	17.00	21.82	22.86	23.79	23.91	23.78
	Log score criterion					
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
tv-NBINAR	**-2.73**	**-2.82**	**-2.83**	**-2.85**	**-2.85**	**-2.85**
NBINAR	-2.75	-2.85	-2.88	-2.91	-2.91	-2.91
tv-PoINAR	-2.83	-2.96	-2.98	-3.00	-3.00	-2.98
PoINAR	-2.88	-3.08	-3.12	-3.18	-3.19	-3.18

Table 3.5.3: *Forecast MSE and log score criterion computed using the last 100 observations for different forecast horizons.*

The results are collected in Table 3.5.3. As we can see the inclusion of the dynamic coefficient α_t provides better forecast performances in the subsample considered. In particular, the tv-NBINAR model outperforms the NBINAR model in terms of both point forecasts and pmf forecasts. The same happens for the tv-PoINAR compared to the PoINAR. This holds true for all forecast horizons considered. Furthermore, the use of the Negative Binomial distribution is particularly relevant to improve the pmf forecasts.

In particular, the Negative Binomial models dominate the Poisson models in terms of log-score criterion. This result is quite natural as the the Negative Binomial models take into account the overdispersion in the data. On the other hand, as concerns the point forecasts, the dynamic parameter α_t seems to play a major role in improving the forecast performances. This can be noted as the models with dynamic α_t dominate the models with static α in terms on MSE. The best performing model is the tv-NBINAR for both criteria and all forecast horizons. This suggests that the flexibility introduced by α_t as well as the choice of an appropriate distribution for the error term can be important to better predict future observations. Overall, these out-of-sample results together with the in sample results show that the tv-INAR models can be useful in practical application.

3.6 Conclusion

We have proposed a flexible INAR model with dynamic coefficient. This model may be interpreted as a filter to approximate more complex DGPs. The empirical results are promising for both simulated and real data. Future research may include the extension of the first-order dynamic INAR model to a general order p. Other work to be done concerns the asymptotic theory of the ML estimator. At the moment, we have only proved the consistency of the estimator. The asymptotic normality requires the study of the first two derivatives of the log likelihood. In this regard, we encountered some difficulties concerning the existence of some moments for the derivative processes.

Appendix

3.A Derivatives of the predictive log-likelihood

Defining $s_t(\bar{\alpha}, \xi) := \partial \log p(y_t|\bar{\alpha}, y_{t-1}, \xi)/\partial \operatorname{logit} \bar{\alpha}$ and $\dot{s}_t(\bar{\alpha}, \xi) := \partial s_t(\bar{\alpha}, \xi)/\partial \operatorname{logit} \bar{\alpha}$, by elementary calculus we obtain that

$$s_t(\bar{\alpha}, \xi) = \left(\sum_{k=0}^{m_t} p_{kt}(\bar{\alpha}, \xi) \right)^{-1} \left(\sum_{k=0}^{m_t} p_{kt}(\bar{\alpha}, \xi)(k - y_{t-1}\bar{\alpha}) \right), \tag{3.7}$$

and

$$\begin{aligned} \dot{s}_t(\bar{\alpha}, \xi) = &\left(\sum_{j=0}^{m_t} \sum_{k=0}^{m_t} p_{kt}(\bar{\alpha}, \xi) p_{jt}(\bar{\alpha}, \xi) \right)^{-1} \times \\ &\left(\sum_{j=0}^{m_t} \sum_{k=0}^{m_t} p_{kt}(\bar{\alpha}, \xi) p_{jt}(\bar{\alpha}, \xi) \left(k(k-j) - \bar{\alpha}(1-\bar{\alpha}) y_{t-1} \right) \right), \end{aligned} \tag{3.8}$$

where $m_t = \min(y_t, y_{t-1})$ and

$$p_{kt}(\bar{\alpha}, \xi) = \binom{y_{t-1}}{k} \bar{\alpha}^k (1-\bar{\alpha})^{y_{t-1}-k} p_e(y_t - k, \xi).$$

3.B Proofs

Proof of Proposition 3.3.1. The stability conditions we consider to obtain the convergence result are based on Theorem 3.1 of Bougerol (1993). Straumann and Mikosch (2006) applied Bougerol's theorem in the space of continuous functions $\mathbb{C}(\Theta, \mathbb{R})$ equipped with the uniform norm $\|\cdot\|_\Theta$. In particular, they provide stability conditions for functional SRE of the form

$$x_{t+1}(\theta) = \phi_t(x_t(\theta), \theta), \ t \in \mathbb{N}, \tag{3.9}$$

where $x_0(\theta) \in \mathbb{R}$, the map $(x, \theta) \mapsto \phi_t(x, \theta)$ from $\mathbb{R} \times \Theta$ into $\mathbb{R}$ is almost surely continuous and the sequence $\{\phi_t(x, \theta)\}_{t\in\mathbb{Z}}$ is stationary and ergodic for any $(x, \theta) \in \mathbb{R} \times \Theta$. Wintenberger (2013) weakened Straumann and Mikosch (2006) conditions replacing a uniform contraction condition with a pointwise condition. The uniform exponential almost sure convergence of a filter satisfying the SRE in (3.9) can be obtained on the basis of Theorem 2 of Wintenberger (2013) from the following conditions:

(a) There exists an $x \in \mathbb{R}$ such that $E \log^+ (\sup_{\theta\in\Theta} |\phi_0(x, \theta)|) < \infty$,

(b) $E \log^+ (\sup_{\theta\in\Theta} \Lambda_0(\theta)) < \infty$,

(c) $E \log (\Lambda_0(\theta)) < 0$ for any $\theta \in \Theta$,

where the random coefficient $\Lambda_t(\theta)$ is defined as

$$\Lambda_t(\theta) = \sup_{(x_1,x_2)\in\mathbb{R}^2, x_1\neq x_2} \frac{|\phi_0(x_1, \theta) - \phi_0(x_2, \theta)|}{|x_1 - x_2|}.$$

In our case, the random function ϕ_t that defines the SRE in (3.9) has the following form

$$\phi_t(x, \theta) = \omega + \beta x + \tau s_t \left(\text{logit}^{-1}(x), \xi\right).$$

First we note that our SRE satisfies the stationarity and continuity requirements to apply Wintenberger's results. In particular, we obtain that the a.s. continuity of $\phi_t(x, \theta)$ follows immediately from the a.s. continuity of $(x, \theta) \mapsto s_t \left(\text{logit}^{-1}(x), \xi\right)$, which is implied by Assumption 3.3.1, and the continuity of the Binomial likelihood (see the functional form of s_t in (3.7)). Furthermore, the stationarity and ergodicity of $\{\phi_t\}_{t\in\mathbb{Z}}$ follows from the stationarity and ergodicity of $\{y_t\}_{t\in\mathbb{Z}}$ together with an application of Proposition 4.3 of Krengel (1985) as $s_t \left(\text{logit}^{-1}(x), \xi\right)$ is a measurable function of y_t and y_{t-1}. In the following, we will prove the proposition by showing that conditions (a)-(c) are satisfied.

As concerns (a), setting $x = 0$ and accounting that $Ey_0^2 < 0$, by an application of Lemma 3.C.1, we obtain that

$$\begin{aligned} E \log^+ \left(\sup_{\theta\in\Theta} |\phi_0(x, \theta)|\right) &\leq \sup_{\theta\in\Theta} |\omega| + \sup_{\theta\in\Theta} |\tau| E \sup_{\theta\in\Theta} |s_t (0.5, \xi)| \\ &\leq \sup_{\theta\in\Theta} |\omega| + \sup_{\theta\in\Theta} |\tau| E|y_{t-1}| < \infty. \end{aligned}$$

Thus (a) is proved.

As concerns (b), by an application of Lemma 3.C.1, we have that

$$E\log^+\left(\sup_{\theta\in\Theta}\Lambda_0(\theta)\right) \leq E\sup_{\theta\in\Theta}\sup_{x\in\mathbb{R}}|\partial\phi_0(x,\theta)/\partial x| \leq \sup_{\theta\in\Theta}|\beta| + \sup_{\theta\in\Theta}|\tau|E\sup_{\theta\in\Theta}\sup_{\bar{\alpha}\in(0,1)}|\dot{s}(\bar{\alpha},\xi)|$$
$$\leq \sup_{\theta\in\Theta}|\beta| + \sup_{\theta\in\Theta}|\tau|E|y_{t-1}^2| < \infty,$$

as $Ey_0^2 < \infty$. This shows that (b) holds true.

Finally, as concerns (c), by condition (3.6) we obtain for any $\theta \in \Theta$

$$E\log\left(\Lambda_0(\theta)\right) \leq E\sup_{x\in\mathbb{R}}|\partial\phi_0(x,\theta)/\partial x| \leq E\sup_{\bar{\alpha}\in(0,1)}|\beta + \tau\dot{s}(\bar{\alpha},\xi)| < 0.$$

This proves (c) and concludes the proof of the proposition.

□

Proof of Proposition 3.3.2. The result follows immediately by an application of Lemma 3.C.1, which provides an upper bound for the derivative of the score.

□

Proof of Theorem 3.3.1. Assumption 3.3.3 ensures that $L(\theta) = El_t(\theta)$ has a unique maximizer in the compact set Θ, which indeed corresponds to the pseudo-true parameter θ^* that minimizes the marginal KL divergence. In the following, we show that the log likelihood function $L_T(\theta)$ converges almost surely uniformly in Θ to $L(\theta)$, namely

$$\|\hat{L}_T - L\|_\Theta \xrightarrow{a.s.} 0, \quad T \to \infty. \tag{3.10}$$

Then, given the compactness of Θ and the identifiability of θ^*, the almost sure convergence $\hat{\theta}_T \xrightarrow{a.s.} \theta^*$ follows by well known standard arguments due to Wald (1949).

Defining $L_T(\theta) = T^{-1}\sum_{t=1}^T l_t(\theta)$, with $l_t(\theta) = \log p(y_t|\tilde{\alpha}_t(\theta), y_{t-1}, \xi)$, an application of the triangle inequality yields

$$\|\hat{L}_T - L\|_\Theta \leq \|\hat{L}_T - L_T\|_\Theta + \|L_T - L\|_\Theta. \tag{3.11}$$

Therefore, the uniform convergence in (3.10) follows if both terms on the right hand side of the inequality (3.11) converge almost surely to zero.

First we show that $\|\hat{L}_T - L_T\|_\Theta \xrightarrow{a.s.} 0$. An application of the mean value theorem

together with Lemma 3.C.1 yields

$$\begin{aligned}|\hat{l}_t(\theta) - l_t(\theta)| &\leq \sup_{\bar{\alpha}\in(0,1)} |s_t(\bar{\alpha}, \xi)||\operatorname{logit}\hat{\alpha}_t(\theta) - \operatorname{logit}\tilde{\alpha}_t(\theta)| \\ &\leq y_{t-1}|\operatorname{logit}\hat{\alpha}_t(\theta) - \operatorname{logit}\tilde{\alpha}_t(\theta)|\end{aligned}$$

for any $\theta \in \Theta$ and $t \in \mathbb{N}$. Furthermore, taking into account that $\|\operatorname{logit}\hat{\alpha}_t - \operatorname{logit}\tilde{\alpha}_t\|_\Theta \xrightarrow{e.a.s.} 0$ by Proposition 3.3.1 and that $E|y_{t-1}| < \infty$ holds true by assumption, an application of Lemma 2.1 of Straumann and Mikosch (2006) yields

$$\sum_{t=1}^{\infty} y_{t-1}\|\operatorname{logit}\hat{\alpha}_t - \operatorname{logit}\tilde{\alpha}_t\|_\Theta < \infty$$

almost surely. As a result, we have that $T^{-1}\sum_{t=1}^{T}\|\hat{l}_t - l_t\|_\Theta \xrightarrow{a.s.} 0$ and therefore we conclude that the desired result $\|\hat{L}_T - L_T\|_\Theta \xrightarrow{a.s.} 0$ is proved as

$$\|\hat{L}_T - L_T\|_\Theta \leq T^{-1}\sum_{t=1}^{T}\|\hat{l}_t - l_t\|_\Theta.$$

We are now left with showing that $\|L_T - L\|_\Theta \xrightarrow{a.s.} 0$. Note that $\{l_t\}_{t\in\mathbb{N}}$ is a stationary and ergodic sequence of random elements that takes values in the space continuous functions $\mathbb{C}(\Theta, \mathbb{R})$ equipped with the uniform norm $\|\cdot\|_\Theta$. Therefore, the desired convergence result follows by an application of the ergodic theorem of Rao (1962) provided that the uniform integrability condition $E\|l_t\|_\Theta < \infty$ is satisfied. In the following, we show that this condition holds true. First, note that $l_t(\theta) \leq 0$ with probability 1 for any $\theta \in \Theta$ as $p(y_1|\bar{\alpha}, y_2, \xi) \leq 1$ for any $(y_1, y_2, \xi, \bar{\alpha}) \in \mathbb{N}^2 \times \Xi \times (0, 1)$. Thus, accounting that $\log(1 + \exp(x)) \leq 1 + |x|$ for any $x \in \mathbb{R}$, we obtain

$$\begin{aligned}|l_t(\theta)| = -l_t(\theta) &= -\log\sum_{k=0}^{m_t} p_{kt}(\tilde{\alpha}_t(\theta), \xi) \leq -\log p_{0t}(\tilde{\alpha}_t(\theta), \xi) \\ &\leq -y_{t-1}\log(1 - \tilde{\alpha}_t(\theta)) - \log p_e(y_{t-1}, \xi) \\ &\leq y_{t-1}\log(1 + \exp(\operatorname{logit}\tilde{\alpha}_t(\theta))) - \log p_e(y_{t-1}, \xi) \\ &\leq y_{t-1}(1 + |\operatorname{logit}\tilde{\alpha}_t(\theta)|) - \log p_e(y_{t-1}, \xi)\end{aligned}$$

almost surely for any $\theta \in \Theta$. Finally, an application of the Cauchy-Schwarz inequality yields

$$\|l_t\| \leq Ey_t + Ey_t^2 + \|\operatorname{logit}\tilde{\alpha}_t\|_\Theta^2 + E\sup_{\theta\in\Theta}|\log p_e(y_{t-1}, \xi)| < \infty,$$

where $Ey_t^2 < \infty$ and $E\sup_{\theta\in\Theta} |\log p_e(y_{t-1},\xi)| < \infty$ are satisfied by assumption and $\|\operatorname{logit}\tilde{\alpha}_t\|_\Theta^2 < \infty$ follows by an application of Lemma 3.C.2. □

Proof of Lemma 3.3.1. The proof of this result is an immediate consequence of Theorem 3 of Wintenberger (2013). We simply sketch the main steps only to illustrate that all conditions needed are satisfied. The same notation and definitions as in the proof of Proposition 3.3.1 are considered. First note that it is sufficient to show that $|\operatorname{logit}\tilde{\alpha}_t(\hat{\theta}_T) - \operatorname{logit}\tilde{\alpha}_t^*| \xrightarrow{a.s.} 0$ as $T\to\infty$. This because we have

$$|\operatorname{logit}\hat{\alpha}_t(\hat{\theta}_T) - \operatorname{logit}\tilde{\alpha}_t^*| \le |\operatorname{logit}\tilde{\alpha}_t(\hat{\theta}_T) - \operatorname{logit}\tilde{\alpha}_t^*| + \|\operatorname{logit}\hat{\alpha}_t - \operatorname{logit}\tilde{\alpha}_t\|_\Theta,$$

and $\|\operatorname{logit}\hat{\alpha}_t - \operatorname{logit}\tilde{\alpha}_t\|_\Theta \xrightarrow{a.s.} 0$ from Proposition 3.3.1. From the results in Theorem 2 of Wintenberger (2013) and the assumptions considered in Proposition 3.3.1, we have that for any $\theta\in\Theta$ there exists a compact neighborhood $B(\theta)$ of θ such that the contraction condition holds uniformly, namely $E\log(\|\Lambda_t\|_{B(\theta)}) < 0$. Therefore, this is true also for the pseudo-true parameter $\theta^*\in\Theta$. As in the proof of Theorem 3 of Wintenberger (2013), repeated applications of the mean value theorem yield

$$\|\operatorname{logit}\tilde{\alpha}_t(\cdot) - \operatorname{logit}\tilde{\alpha}_t^*\|_{B(\theta^*)} \le \sum_{k=1}^{\infty}\prod_{i=1}^{k}\|\Lambda_{t-i}\|_{B(\theta^*)}\|\phi_{t-k}(\operatorname{logit}\tilde{\alpha}_{t-k}^*,\cdot) - \operatorname{logit}\tilde{\alpha}_{t-k+1}^*\|_{B(\theta^*)}$$

for any $\theta\in B(\theta^*)$ with probability 1. The existence of the limit on the right hand side is obtained from Lemma 2.1 of Straumann and Mikosch (2006) together with the integrability condition $E\log^+\|\operatorname{logit}\tilde{\alpha}_t\|_{B(\theta^*)}$ implied by Lemma 3.C.2 and $\prod_{i=1}^k\|\Lambda_{t-i}\|_{B(\theta^*)} \xrightarrow{e.a.s.} 0$ as $k\to\infty$ implied by the uniform contraction condition. Finally, the desired result $|\operatorname{logit}\tilde{\alpha}_t(\hat{\theta}_T) - \operatorname{logit}\tilde{\alpha}_t^*| \xrightarrow{a.s.} 0$ follows as in Theorem 3 of Wintenberger (2013) taking into account that the ML estimator $\hat{\theta}_T$ is strongly consistent by Theorem 3.3.1. □

Proof of Theorem 3.3.2. An application of the mean value theorem together with Lemma 3.C.3 yields that for any $x\in\mathbb{N}$ there is a $C_x > 0$ and a stationary sequence of random variables $\{\eta_t\}_{t\in\mathbb{N}}$ such that the following inequalities hold true with probability 1

$$\begin{aligned}|\hat{p}_t(x,\hat{\theta}_T) - p_t^*(x)| \le & \sup_{(\alpha,\theta)\in(0,1)\times\Theta}\left|\frac{\partial p(x|y_{t-1},\alpha,\xi)}{\partial\operatorname{logit}\alpha}\right|\left|\operatorname{logit}\hat{\alpha}_t(\hat{\theta}_T) - \operatorname{logit}\alpha_t^*\right| + \\ & + \sup_{(\alpha,\theta)\in(0,1)\times\Theta}\left\|\frac{\partial p(x|y_{t-1},\alpha,\xi)}{\partial\xi}\right\|_1\|\hat{\xi}_T - \xi^*\|_1 \\ \le & \eta_t|\operatorname{logit}\hat{\alpha}_t(\hat{\theta}_T) - \operatorname{logit}\alpha_t^*| + C_x\|\hat{\xi}_T - \xi^*\|_1.\end{aligned}$$

The desired convergence to zero in probability of $|\hat{p}_t(x, \hat{\theta}_T) - p_t^*(x)|$ then follows immediately as $\|\hat{\xi}_T - \xi^*\|_1$ is $o_p(1)$ by Theorem 3.3.1 and $|\operatorname{logit} \hat{\alpha}_t(\hat{\theta}_T) - \operatorname{logit} \alpha_t^*|$ is $o_p(1)$ by Lemma 3.3.1. □

3.C Technical lemmas

Lemma 3.C.1. *Let Assumption 3.3.1 hold, then the following inequalities are satisfied with probability 1 for any $\bar{\alpha} \in (0, 1)$ and $\xi \in \Xi$*

(i) $|s_t(\bar{\alpha}, \xi)| \leq 2y_{t-1}$.

(ii) $-y_{t-1}/4 \leq \dot{s}_t(\bar{\alpha}, \xi) \leq m_t^2$.

Proof. Assumption 3.3.1 implies that $p_{kt}(\bar{\alpha}, \xi) > 0$ with probability 1 for any $\bar{\alpha} \in (0, 1)$ and $\xi \in \Xi$. This ensures that $s_t(\bar{\alpha}, \xi)$ and $\dot{s}_t(\bar{\alpha}, \xi)$ are well defined as their denominator, see expressions (3.7) and (3.8), is almost surely larger then zero for any $\bar{\alpha} \in (0, 1)$ and $\xi \in \Xi$.

To show that (i) is satisfied, we note that

$$|s_t(\bar{\alpha}, \xi)| \leq \left(\sum_{k=0}^{m_t} p_{kt}(\bar{\alpha}, \xi) \right)^{-1} \left(\sum_{k=0}^{m_t} p_{kt}(\bar{\alpha}, \xi)(k + y_{t-1}\bar{\alpha}) \right) \leq (1 + \bar{\alpha}) y_{t-1},$$

therefore (i) immediately holds true as $\bar{\alpha} \in (0, 1)$.

As concerns (ii), taking into account that $y_t \geq 0$ almost surely, we obtain that the numerator of expression (3.8) has the following upper bound

$$\left(\sum_{j=0}^{m_t} \sum_{k=0}^{m_t} p_{kt}(\bar{\alpha}, \xi) p_{jt}(\bar{\alpha}, \xi) k(k - j) \right) \leq \left(\sum_{j=0}^{m_t} \sum_{k=0}^{m_t} p_{kt}(\bar{\alpha}, \xi) p_{jt}(\bar{\alpha}, \xi) \right) m_t^2,$$

therefore it follows immediately that $\dot{s}_t(\bar{\alpha}, \xi) \leq m_t^2$. Similarly, we obtain that the numerator of (3.8) is larger or equal than

$$\left(\sum_{j=0}^{m_t} \sum_{k=0}^{m_t} p_{kt}(\bar{\alpha}, \xi) p_{jt}(\bar{\alpha}, \xi) \right) (-\bar{\alpha}(1 - \bar{\alpha}) y_{t-1}),$$

therefore $\dot{s}_t(\bar{\alpha}, \xi) \geq -y_{t-1}/4$ as $\bar{\alpha} \in (0, 1)$ and, as a result, it follows that (ii) is satisfied. □

Lemma 3.C.2. *Let the conditions of Proposition 3.3.1 hold, then $E \sup_{\theta \in \Theta} |\operatorname{logit} \tilde{\alpha}_t(\theta)|^2 < \infty$.*

Proof. The lemma is proved by showing that there exists a stationary and ergodic sequence $\{\tilde{\nu}_t\}_{t\in\mathbb{Z}}$ such that $E\tilde{\nu}_t^2 < \infty$ and that $\|\operatorname{logit}\tilde{\alpha}_t\|_\Theta < (\tilde{\nu}_t + 1)$ with probability 1. Then, it is immediate to conclude that $E\|\operatorname{logit}\tilde{\alpha}_t\|_\Theta^2 < \infty$.

First, we define the sequence $\{\hat{v}_t\}_{t\in\mathbb{N}}$ through the following stochastic recurrence equation

$$\hat{v}_{t+1} = \omega_u + \beta_u\hat{v}_t + 2\tau_u y_t,\ t \in \mathbb{N},$$

which is initialized at $\hat{v}_0 = \omega_u/(1-\beta_u)$ and where $\omega_u = \sup_{\theta\in\Theta}|\omega|$, $\beta_u = \sup_{\theta\in\Theta}|\beta|$ and $\tau_u = \sup_{\theta\in\Theta}|\tau|$. Considering that $\beta_u < 1$ from the specification of Θ and that $\{y_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic, an application of Theorem 3.1 of Bougerol (1993) yields that $|\hat{v}_t - \tilde{v}_t| \xrightarrow{a.s.} 0$ as t goes to infinity, where $\{\tilde{v}_t\}_{t\in\mathbb{N}}$ is a stationary and ergodic sequence that admits the following representation

$$\tilde{v}_t = \omega_u/(1-\beta_u) + 2\tau_u\sum_{k=1}^{\infty}\beta_u^k y_{t-k}.$$

From this expression, it is straightforward to obtain that $Ey_t^2 < \infty$, together with $\beta_u < 1$, entails $E\tilde{v}_t^2 < \infty$.

In the following, we show that $\|\operatorname{logit}\tilde{\alpha}_t\|_\Theta < (\tilde{\nu}_t + 1)$ with probability 1. Taking into account the definition of the sequence $\{\operatorname{logit}\hat{\alpha}(\theta)\}_{t\in\mathbb{N}}$ in (3.4) and the fact that $\sup_{\theta\in\Theta}|s_t(\bar{\alpha},\xi)| < 2y_{t-1}$ almost surely for any $\bar{\alpha} \in (0,1)$ by Lemma 3.C.1, it follows immediately that $\|\operatorname{logit}\hat{\alpha}_t\|_\Theta \leq \hat{v}_t$ with probability 1 for any $t \in \mathbb{N}$. Therefore, we have that for a large enough $t \in \mathbb{N}$ with probability 1

$$\|\operatorname{logit}\tilde{\alpha}_t\|_\Theta - \tilde{v}_t - 1 \leq \|\operatorname{logit}\hat{\alpha}_t\|_\Theta - \hat{v}_t - 1 + \|\operatorname{logit}\tilde{\alpha}_t - \operatorname{logit}\hat{\alpha}_t\|_\Theta + |\tilde{v}_t - \hat{v}_t| < 0,$$

as $\|\operatorname{logit}\tilde{\alpha}_t - \operatorname{logit}\hat{\alpha}_t\|_\Theta$ and $|\tilde{v}_t - \hat{v}_t|$ go to zero almost surely. As a result, given the stationarity of $\{\|\operatorname{logit}\tilde{\alpha}_t\|_\Theta - \tilde{v}_t\}$ we infer that $\|\operatorname{logit}\tilde{\alpha}_t\|_\Theta < (\tilde{v}_t + 1)$ with probability 1 for any $t \in \mathbb{Z}$. This concludes the proof. □

Lemma 3.C.3. *Let the conditions of Theorem 3.3.2 hold. Then, for any $x \in \mathbb{N}$ there exists a stationary sequence of random variables $\{\eta_t\}_{t\in\mathbb{N}}$ and a constant $C_x > 0$ such that almost surely*

(i) $\sup_{(\alpha,\theta)\in(0,1)\times\Theta}\left|\frac{\partial p(x|y_{t-1},\alpha,\xi)}{\partial\, logit\,\alpha}\right| \leq \eta_t.$

(ii) $\sup_{(\alpha,\theta)\in(0,1)\times\Theta}\left\|\frac{\partial p(x|y_{t-1},\alpha,\xi)}{\partial\xi}\right\|_1 \leq C_x.$

Proof. First we show that (i) holds true. From standard calculus, we obtain that

$$\frac{\partial p(x|y_{t-1},\alpha,\xi)}{\partial \operatorname{logit}\alpha} = \sum_{k=0}^{m_{xt}} p_{kt}(x,\alpha,\xi)(k-\alpha y_{t-1}),$$

where $m_{xt} = \min(x, y_{t-1})$ and

$$p_{kt}(x,\alpha,\xi) = \binom{y_{t-1}}{k}\alpha^k(1-\alpha)^{y_{t-1}-k}p_e(x-k,\xi).$$

As a result, taking into account that $0 \leq p_{kt}(x,\alpha,\xi) \leq 1$ with probability 1 for any $(x,\alpha,\xi) \in \mathbb{N}\times(0,1)\times\Xi$, it follows that

$$\left|\frac{\partial p(x|y_{t-1},\alpha,\xi)}{\partial \operatorname{logit}\alpha}\right| \leq \sum_{k=0}^{m_{xt}} p_{kt}(x,\alpha,\xi)(k+y_{t-1}) \leq \sum_{k=0}^{y_{t-1}}(k+y_{t-1}) \leq 2(1+y_{t-1})y_{t-1}.$$

Therefore, the result (i) is proved setting $\eta_t = 2(1+y_{t-1})y_{t-1}$ and recalling that $\{y_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic and thus $\{\eta_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic as well.

As concerns (ii), we have that

$$\frac{\partial p(x|y_{t-1},\alpha,\xi)}{\partial \xi} = \sum_{k=0}^{m_{xt}} \binom{y_{t-1}}{k}\alpha^k(1-\alpha)^{y_{t-1}-k}\frac{\partial p_e(x-k,\xi)}{\partial \xi}.$$

As a result, we obtain that the following inequalities are satisfied almost surely

$$\left\|\frac{\partial p(x|y_{t-1},\alpha,\xi)}{\partial \operatorname{logit}\alpha}\right\|_1 \leq \sum_{k=0}^{m_{xt}} \binom{y_{t-1}}{k}\alpha^k(1-\alpha)^{y_{t-1}-k}\left\|\frac{\partial p_e(x-k,\xi)}{\partial \xi}\right\|_1 \leq \sum_{k=0}^{x}\left\|\frac{\partial p_e(x-k,\xi)}{\partial \xi}\right\|_1.$$

Therefore, from the continuity of the derivative provided by Assumption 3.3.4 and the compactness of Θ, we obtain that for any given $x-k \in \mathbb{N}$ there is a constant $C_{kx} > 0$ such that

$$\sup_{\theta\in\Theta}\left\|\frac{\partial p_e(x-k,\xi)}{\partial \xi}\right\|_1 \leq C_{kx}.$$

This shows that the result in (ii) holds as $C_x = \sum_{k=0}^{x} C_{kx} < \infty$. □

Chapter 4

Accelerating GARCH and Score-Driven Models: Optimality, Estimation and Forecasting

4.1 Introduction

In time series analysis, a widely adopted approach to model the temporal dependence in the data is to consider a parametric distribution for the observations and allow some of the parameters to vary over time. The specification of the time-varying parameters plays a central role in determining the dynamic properties of the model. Depending on the specification of the time-varying parameters, most of the models in this setting can be classified in two categories: observation-driven and parameter-driven models, see Cox (1981). The main advantage of observation-driven models is that the likelihood function is available in closed form. This allows us to avoid time-consuming simulation-based methods and facilitates likelihood-based inference. The GAS updating mechanism provides a general framework to specify time-varying parameters in an observation-driven setting. The use of the score as driving mechanism to update time-varying parameters is also justified by an optimality reasoning, see Blasques et al. (2015). GAS models have been widely used in statistics and econometrics. They have a comparable predictive ability to parameter-driven models but with the additional advantage of being easy to estimate, see Koopman et al. (2016).

Time series data often exhibit complex dynamic behaviors. A possible situation is to have that the amount of information contained in past observations is changing over time, i.e. large in some time periods and small in others. In such a situation, we would like

to have a dynamic specification that updates the time-varying parameter quickly when the data is informative and slowly when the data is not informative. Within the GAS framework, this can be achieved introducing a dynamic parameter that determines the magnitude of the score innovation at each time period. On the basis of this idea, we propose a generalization of the class of GAS models: the accelerating GAS (aGAS) models. A special case of our approach is the accelerating GARCH (aGARCH) model, which is an extension of the GARCH model. We illustrate the intuition behind this specification and provide an empirical study to the S&P 500 stock returns. The results show how the proposed accelerating volatility framework can be useful to enhance in-sample and out-of-sample performances of GARCH models. Besides the volatility case, we also discuss the general aGAS case and provide a theoretical line of reasoning in the spirit of Blasques et al. (2015) to justify the proposed method. Furthermore, we present a simulation example to show the role that this approach can play and how it can produce flexible models to better approximate an unknown DGP. Finally, in the context of location models, we consider an empirical application to the quarterly US consumer price inflation series by specifying a fat-tailed model with dynamic conditional mean and volatility. The accelerating updating equation renders our aGAS model capable of describing not only the the fast changes in the inflation level during the 1970's and 1980's but also the smooth and flat dynamics of the conditional mean during the great moderation of the two decades that followed.

The chapter is structured as follows. Section 4.2 introduces the aGARCH model. Section 4.3 presents the general aGAS framework. Section 4.4 derives the theoretical justification for the accelerating models. Section 4.5 illustrates the simulation experiment. Section 4.6 presents the application to the S&P 500 stock returns. Section 4.7 presents the application to the US inflation series. Section 4.8 concludes.

4.2 Accelerating GARCH model

The GARCH(1,1) model of Engle (1982) and Bollerslev (1986) is given by

$$y_t = \sigma_t \varepsilon_t, \quad \sigma_{t+1}^2 = \omega + \alpha y_t^2 + \beta \sigma_t^2,$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of random variables with zero mean and unit variance. We propose an extension of the GARCH model: the aGARCH model. The aGARCH

model is described by the following equations

$$y_t = \sigma_t \varepsilon_t, \quad \sigma^2_{t+1} = \omega + \beta \sigma^2_t + \alpha_t (y^2_t - \sigma^2_t), \tag{4.1}$$

$$\alpha_t = \beta \, \text{logit}^{-1}(f_{t+1}), \quad f_{t+1} = \omega_f + \beta_f f_t + \alpha_f (\varepsilon^2_t - 1)(\varepsilon^2_{t-1} - 1). \tag{4.2}$$

The logit link function rescaled by β is needed to ensure the positivity of the variance. In this way, the dynamic parameter α_t is constrained to take values between zero and β. The aGARCH variance equation can be written as a GARCH model with time-varying parameters, namely

$$\sigma^2_{t+1} = \omega + \alpha_t y^2_t + \beta_t \sigma^2_t,$$

where $\beta_t = \beta - \alpha_t$. This formulation further highlights why α_t needs to be between zero and β.

The term $y^2_t - \sigma^2_t$ in equation (4.1) is the innovation of the variance recursion. The parameter α_t is particularly important as it determines the amount of information about σ^2_{t+1} contained in the last observation y_t. The idea of having a time-varying α_t is that in some time periods the data may be more informative than in others. For instance, this could be due to a break in the level of the variance. Before the break, the variance may be changing slowly and therefore the magnitude of the innovations should be small. Whereas, right after the break, the new observations are very informative about the new variance level and thus the parameter α_t should increase to update quickly σ^2_t. In the following, we provide an illustration of this idea and show the role that a dynamic α_t can play. Assume that we are interested in approximating a true variance path that is observed with an error disturbance. The true variance is represented by the red line in Figure 4.2.1. The observed variance is filtered considering the GARCH and the aGARCH recursions. Figure 4.2.1 illustrates the filtered variance paths using a small α, a large α and a dynamic α_t. As we can see, having a fixed α leads to a trade-off between being exposed to the disturbance component and updating quickly the variance after the break. This can be noted observing the gray line and the green line in Figure 4.2.1. On the other hand, the advantage of a dynamic α_t is shown by the black line. We can update quickly the variance after the break and, at the same time, we can be robust against the disturbance component in periods when the true variance is constant. Figure 4.2.2 shows how the dynamic α_t is evolving over time. This plot further illustrates that the filtered variance is updated quickly only after the brake when the aGARCH recursion is employed.

We also note that as α_t approaches β the aGARCH model becomes a first order ARCH model. This means that the variance depends only on the most recent observations when

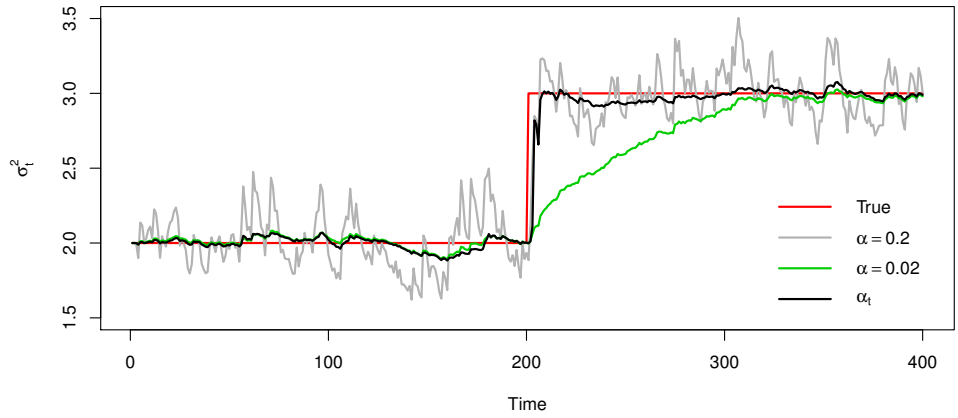

Figure 4.2.1: *Filtered variance for different parameters* α*. The green line is obtained setting* $\alpha = 0.02$*, the gray line setting* $\alpha = 0.20$ *and the black line considering a time-varying* α_t*.*

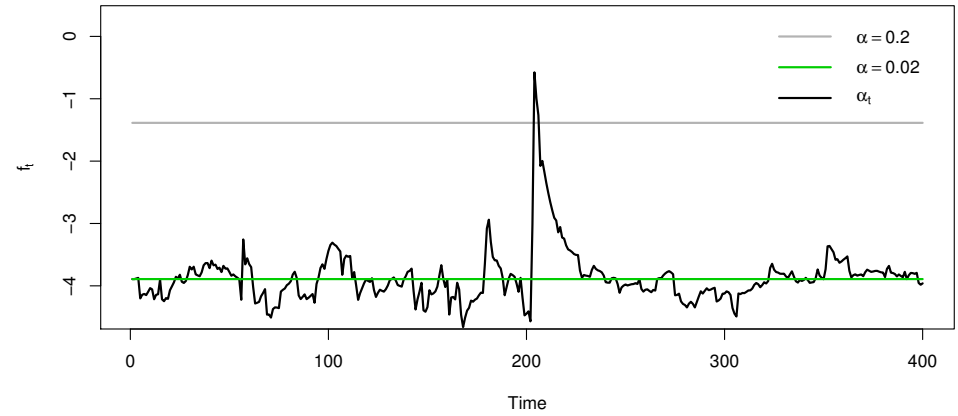

Figure 4.2.2: *The black line denotes the filtered* α_t *from the aGARCH model. The green and gray lines represent the constant* α *of the GARCH model.*

α_t is large. Whereas, when α_t is close to zero, the impact of the last observation is lower as its effect is averaged with that of the other observations. As a result, a large α_t after the break leads to a shorter memory of the filtered variance. This is quite natural as the new observations are very informative about the new level, whereas, the past level of σ_t^2, obtained filtering observations before the break, is not very informative.

The updating mechanism of the dynamic α_t described in equation (4.2) has also an intuitive interpretation. In particular, α_t is driven by products of standardized past innovations. Therefore, it increases when past innovations are positively correlated, decreases when the correlation is negative and remains constant when the correlation is zero. A positive correlation indicates that for repeated observations the innovation tends to be either above or below its expectation. This is indeed an indication that the variance should be updated more quickly. In the same way, a negative correlation indicates that consecutive

innovations tend to have opposite sign. Clearly, this can indicate that the variance is being updated too quickly as the disturbance component affect too much the path of the variance and thus innovations are more likely to have opposite sign. Finally, a correlation equal to zero may suggest a situation of equilibrium where the variance is being updated in the right way. In Section 4.4, we will show that the updating mechanism considered for α_t is justified by an optimality reasoning.

Time variation in the parameters of the GARCH(1,1) model has also been considered by Engle and Lee (1999). Their model presents time variation in ω. The dynamic ω is interpreted as a long run variance component. The Engle and Lee GARCH model can be written as a GARCH(2,2) model. In our case, the aGARCH model does not have an higher order GARCH representation. This is due to the fact that the variance recursion becomes a nonlinear function of past y_t^2 when α_t is time varying.

In the next section, we will see that the aGARCH specification is a special case of the more general aGAS framework for time-varying parameter models.

4.3 Accelerating Score-Driven models

The GAS framework of Creal et al. (2013) and Harvey (2013) provides a general approach to specify time-varying parameter models. GAS models have been successfully applied to a large number of problems in time series analysis. Examples include the location and scale fat-tailed models of Harvey and Luati (2014); Andres (2014), the dynamic factor models in Creal et al. (2014), and the time-varying copula models of Oh and Patton (2016), Creal et al. (2011) and Salvatierra and Patton (2015). We propose a class of models that extends the GAS framework by introducing time variation in the GAS updating equation. The idea is the same as illustrated for the GARCH model. In particular, the aGARCH model presented in the previous section is a special case of aGAS model with time-varying variance and Gaussian conditional distribution.

The aGAS model is described by the following equations

$$y_t \sim p(y_t|\lambda_t;\theta), \quad \lambda_{t+1} = \omega_\lambda + \beta_\lambda \lambda_t + \alpha_t s_{\lambda,t}, \tag{4.3}$$

$$\alpha_t = h(f_{t+1}), \quad f_{t+1} = \omega_f + \beta_f f_t + \alpha_f s_{f,t}, \tag{4.4}$$

where $p(\cdot|\lambda_t,\theta)$ is a parametric conditional density, h is an increasing link function, ω_λ, ω_f, β_λ, β_f and α_f are unknown parameters to be estimated and $\theta \in \Theta$ is a vector containing all the static parameters of the model. The innovation terms $s_{\lambda,t}$ and $s_{f,t}$ are specified

on the basis of the score of the predictive log-likelihood

$$s_{\lambda,t} = S_{\lambda,t} u_{\lambda,t}, \quad u_{\lambda,t} = \partial \log p(y_t|\lambda_t;\theta)/\partial \lambda_t,$$
$$s_{f,t} = S_{f,t} u_{f,t}, \quad u_{f,t} = \partial \log p(y_t|\lambda_t;\theta)/\partial f_t,$$

where $S_{\lambda,t}$ and $S_{f,t}$ are positive scaling factors. Note that the time index t of the time-varying parameters λ_t and f_t denotes that they are functions of past observation up to time $t-1$, namely functions of $\{y_{t-1}, y_{t-2}, \dots\}$. It is also easy to see that the aGAS specification in (4.3) and (4.4) is a generalization of the GAS framework. In particular, the GAS model is given by the equations in (4.3) and setting α_t equal to a static parameter α_λ to be estimated.

By straightforward calculations, we obtain that the innovation $s_{f,t}$ in (4.4) has the following expression

$$s_{f,t} = C_{f,t} u_{\lambda,t} u_{\lambda,t-1}, \tag{4.5}$$

where $C_{f,t}$ is a positive scaling factor. The formula in (4.5) provides a more explicit form for $s_{f,t}$, which can be directly derived from $u_{\lambda,t}$ without the need of calculating any other derivative. The innovation $s_{f,t}$ of the dynamic α_t is therefore given by rescaled products of past score innovations. From this expression it is also straightforward to see that the aGARCH model is a special case of aGAS model for time-varying variance. Furthermore, the same intuitive interpretation as for the aGARCH case applies here for the innovation $s_{f,t}$ of the dynamic α_t. In particular, α_t tends to increase when there is positive autocorrelation in past score innovations and decreases when there is negative correlation.

We also mention that the use of scaling factors for score innovations is very popular in GAS modeling and the choice of which scaling to use may depend on the model at hand. Creal et al. (2013) proposed the use of the Fisher information I_t to account for the curvature of the score. Typical choices for the scaling factor are the inverse of the Fisher Information, the square root of the Fisher Information inverse and the identity matrix. Note that considering $I_t^{-1/2}$ as a scaling factor leads the conditional variance of the score innovations to be equal to 1. Therefore, the variability of the innovation of the autoregressive process in (4.3) is determined solely by α_t.

4.4 Optimality properties

In this section, we provide a theoretical justification for the aGAS specification in (4.3) and (4.4). Blasques et al. (2015) developed a line of reasoning to show some optimality features of the GAS updating mechanism. We build on Blasques et al. (2015) and show that the use of the score-based innovation in (4.5) for α_t has an optimality justification. Furthermore, we also show how, under certain conditions, the updating mechanism of the aGAS model outperforms the classic GAS update in terms of local Kullback Leibler divergence reduction. The results are based on a misspecified model setting where the objective is to consider the dynamic specification that minimizes the KL divergence between a postulated conditional distribution and the unknown true distribution of the DGP.

4.4.1 A general updating mechanism

Assume that the sequence of observed data $\{y_t\}_{t=1}^T$ with values in $\mathcal{Y} \subseteq \mathbb{R}$ is generated by an unknown stochastic process that satisfies

$$y_t \sim p_t^o(y_t), \quad t \in \mathbb{N},$$

where p_t^o is the true unknown conditional density. We consider a conditional density for the observations as in (4.3), $y_t \sim p(y_t|\lambda_t;\theta)$, where $\theta \in \Theta$ is a static parameter and λ_t a time-varying parameter that takes values in $\Lambda \subseteq \mathbb{R}$. Note that also the model density $p(\cdot|\lambda_t;\theta)$ is allowed to be misspecified and a true λ_t^o and θ_0 such that $p_t^o = p(\cdot|\lambda_t^o;\theta_0)$ may not even exist.

The objective is to specify the dynamics of the time-varying parameter λ_t in such a way that the conditional density $p(\cdot|\lambda_t;\theta)$ implied by the model is as close as possible to the true conditional density p_t^o. To evaluate the distance between these two conditional densities, a classical approach is to consider the Kullback-Leibler (KL) divergence introduced in Kullback and Leibler (1951) as a measure of divergence, or distance, between probability distributions. The KL divergence plays an important role in information theoretic settings (Jaynes, 1957, 2003) as well as in the world of statistics (Kullback, 1959; Akaike, 1973). The importance of the KL divergence in econometric applications is reviewed in Maasoumi (1986) and Ullah (1996, 2002).

The ideal specification of λ_t minimizes the KL divergence between the true conditional density p_t^o and the model-implied conditional density $p(\cdot|\lambda_t;\theta)$. In other words, a sequence $\{\lambda_t\}_{t\in\mathbb{N}}$ is optimal if for each $t \in \mathbb{N}$, the value of λ_t minimizes the following

KL divergence

$$\mathrm{KL}_Y\left(p_t^o, p(\cdot|\lambda_t;\theta)\right) = \int_Y p_t^o(y) \log \frac{p_t^o(y)}{p(y|\lambda_t;\theta)} dy,$$

where Y denotes the set over which the local KL divergence is evaluated; see Hjort and Jones (1996), Ullah (2002) and Blasques et al. (2015) for applications of the local KL divergence. Assuming that $\{\lambda_t^*\}_{t\in\mathbb{N}}$ is an optimal sequence that minimizes the KL divergence for any $t \in \mathbb{N}$, we would like our model to deliver a filtered time-varying parameter $\{\lambda_t\}_{t\in\mathbb{N}}$ that approximates arbitrarily well the trajectory of $\{\lambda_t^*\}_{t\in\mathbb{N}}$.

Of course, from the outset, there is no reason to suppose that the classic GAS recursion

$$\lambda_t = \omega_\lambda + \beta_\lambda \lambda_{t-1} + \alpha_\lambda s_{\lambda,t-1}$$

would ever deliver such a result. Lemma 4.4.1 reminds us that a time-varying update of the type

$$\lambda_t(f_t) = \omega_\lambda + \beta_\lambda \lambda_{t-1} + h(f_t) s_{\lambda,t-1},$$

could deliver a better approximation to $\{\lambda_t^*\}_{t\in\mathbb{N}}$.

Lemma 4.4.1. *If an optimal sequence $\{\lambda_t^*\}_{t\in\mathbb{N}}$ exists, then for any given initialization $\lambda_0 \in \Lambda$ there exists a sequence $\{f_t\}_{t\in\mathbb{N}}$ of points such that $\lambda_t(f_t) = \lambda_t^*\ \forall t \in \mathbb{N}$. Moreover, f_t is almost surely constant if and only if there is some $c \in \mathbb{R}$ such that $s_{\lambda,t-1} = (\lambda_t^* - \omega_\lambda - \beta_\lambda \lambda_{t-1})/h(c)$ almost surely for every $t \in \mathbb{N}$.*

In practice, however, the problem is how to specify the dynamics of f_t. Below, we will address the issue by providing a theoretical justification for the score-based update of f_t.

We also note that in this section for notational convenience we write λ_t as a function of f_t and discuss the update of f_t and not $\alpha_{t-1} = h(f_t)$. However, this change of notation does not lead to any practical difference as h is defined to be a monotone increasing link function.

4.4.2 Optimality of score innovations

We build on the work of Blasques et al. (2015) that provides optimality arguments for a score-based updating equation. Specifically, Blasques et al. (2015) show that considering an updating scheme of the form

$$\lambda_{t+1} = \lambda_t + \alpha_\lambda s_{\lambda,t}$$

reduces locally the KL divergence between the model density and the true probability density. In particular, they show that the variation in the KL divergence obtained by updating the time-varying parameter from λ_t to λ_{t+1} satisfies

$$\mathrm{KL}_Y\left(p_t^o, p(\cdot|\lambda_{t+1};\theta)\right) - \mathrm{KL}_Y\left(p_t^o, p(\cdot|\lambda_t;\theta)\right) < 0,$$

when the update is local $\lambda_t \approx \lambda_{t+1}$ and the set Y is a neighborhood of y_t. This result is subject to the fact that the parameter α_λ has to be positive because otherwise the information provided by the score is distorted. Clearly, as this optimality concept regards only the direction of the update, we can conclude that the optimality holds also when α_λ is time varying as long as it is positive. This justifies the use of a positive link function h in (4.3) that ensures the positivity of $h(f_t)$.

It is also worth mentioning that the optimality concept in Blasques et al. (2015) is shown to hold for $(\omega_\lambda, \beta_\lambda) \approx (0, 1)$. This because the reduction of local KL divergence from the update is considered with respect to p_t^o. In practice, what we really want is to reduce the KL divergence with respect to p_{t+1}^o as the updated time-varying parameter λ_{t+1} is used to specify the conditional probability measure of y_{t+1}. The problem is that λ_t is updated using information from p_t^o and therefore, without imposing any restriction on the true sequence of conditional densities, it is impossible to say whether the updating scheme makes any sense with respect to p_{t+1}^o. Blasques et al. (2015) show that having $(\omega_\lambda, \beta_\lambda) \approx (0, 1)$ is optimal also with respect to the density p_{t+1}^o only if the true conditional density varies sufficiently smoothly over time. This justifies the possibility that in practice it may be reasonable to consider also $(\omega_\lambda, \beta_\lambda) \neq (0, 1)$.

We now add to the results of Blasques et al. (2015) by considering the updating scheme in (4.4) for the time-varying parameter f_t and showing that it has a similar optimality justification. More specifically, we provide an optimality reasoning for the updating scheme in (4.4) setting $(\omega_f, \beta_f) \approx (0, 1)$,

$$f_{t+1} = f_t + \alpha_f s_{f,t}. \tag{4.6}$$

At time $t-1$, a given parameter value $f_t \in \mathcal{F} \subseteq \mathbb{R}$ is used to update a given $\lambda_{t-1} \in \Lambda$ by the recursion in (4.3), namely

$$\lambda_t(f_t) = \omega_\lambda + \beta_\lambda \lambda_{t-1} + h(f_t) s_{\lambda,t-1}.$$

Then, at time t we observe y_t and the parameter f_t is updated to f_{t+1}. We consider optimal an updating mechanism that processes properly the information provided by y_t. The idea

is that f_t has to be updated in such a way that the model density with the updated f_{t+1} is closer to the true density p_t^o than the model density $p(\cdot|\lambda_t(f_t);\theta)$. We consider the following definition.

Definition 4.4.1. *The realized KL variation for the parameter update from f_t to f_{t+1} is*

$$\Delta_{f,t}^{t+1} = KL_Y\big(p_t^o, p(\cdot|\lambda_t(f_{t+1});\theta)\big) - KL_Y\big(p_t^o, p(\cdot|\lambda_t(f_t);\theta)\big).$$

A parameter update for f_t is said to be optimal in local realized KL divergence if and only if $\Delta_{f,t}^{t+1} < 0$ almost surely for any $(f_t, \theta) \in \mathcal{F} \times \Theta$.

The results we present are local in the sense that we will show that at each step the score update gives the right direction to reduce a local realized KL divergence. As in Blasques et al. (2015), we focus on sets of the form

$$Y = B(y_t, \epsilon_y) = \{y \in \mathcal{Y} : |y_t - y| < \epsilon_y\},$$
$$F = B(f_t, \epsilon_f) = \{f_{t+1} \in \mathcal{F} : |f_t - f_{t+1}| < \epsilon_f\}.$$

First, we impose some regularity assumptions on the score $s_{\lambda,t}$. In particular, we impose that the score has some differentiability properties and also that it is nonzero with probability 1 to ensure that the parameter f_t is always updated.

Assumption 4.4.1. *The score $u_{\lambda,t} = u_\lambda(y_t, \lambda_t, \theta)$ is continuously differentiable in y_t and λ_t, and almost surely $u_\lambda(y_t, \lambda_t, \theta) \neq 0$ for any $(\lambda_t, \theta) \in \Lambda \times \Theta$ and $t \in \mathbb{N}$.*

The next proposition states that the score update for f_t is optimal in the sense of Definition 4.4.1.

Proposition 4.4.1. *Let Assumption 4.4.1 hold, then the update from f_t to f_{t+1} in (4.6) is optimal in terms of local realized KL divergence as long as α_f is positive.*

The next proposition stresses the fact that only the score $s_{f,t}$ provides the right direction to update f_t.

Proposition 4.4.2. *Let Assumption 4.4.1 hold, then any parameter update from f_t to f_{t+1} is optimal in local realized KL divergence if and only if sign$(f_{t+1} - f_t)$ = sign$(s_{f,t})$ almost surely for any $f_t \in \mathcal{F}$.*

4.4.3 Relative optimality

The optimality concept developed in the previous section is only related to the update of f_t, but, in practice, the update of f_t is only a tool to improve the update of $\lambda_t(f_t)$. The idea is to compare the score update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ with the score update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$. Indeed, the former corresponds to an aGAS update and the latter corresponds to a GAS update as f_t is maintained constant. As before, the quality of the updates is measured in terms of KL reduction. We are thus interested in comparing the variation in KL divergence obtained by updating the parameter from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$,

$$\Delta^{t+1}_{\lambda,t+1} = \mathrm{KL}_Y\big(p^o_t, p(\cdot|\lambda_{t+1}(f_{t+1});\theta)\big) - \mathrm{KL}_Y\big(p^o_t, p(\cdot|\lambda_t(f_t);\theta)\big),$$

against the variation in KL divergence obtained under the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$

$$\Delta^{t}_{\lambda,t+1} = \mathrm{KL}_Y\big(p^o_t, p(\cdot|\lambda_{t+1}(f_t);\theta)\big) - \mathrm{KL}_Y\big(p^o_t, p(\cdot|\lambda_t(f_t);\theta)\big).$$

Clearly, the first type of update is better if it can ensure a greater reduction in KL divergence.

Definition 4.4.2. *The parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ is said to dominate the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ in local realized KL divergence, if and only if*

$$\Delta^{t+1}_{\lambda,t+1} - \Delta^{t}_{\lambda,t+1} < 0.$$

The notion of dominance in local realized KL divergence in Definition 4.4.2 provides a line of comparison for the parameter updates. We can say that the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ outperforms the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ if $\Delta^{t+1}_{\lambda,t+1} < \Delta^{t}_{\lambda,t+1}$. The results we obtain are local in the sense that the KL divergence is evaluated locally and the innovations $s_{\lambda,t-1}$ and $s_{\lambda,t}$ are in a neighborhood of zero. Moreover, we also impose that the observation y_t lies in a neighborhood of y_{t-1}. More formally, the realized KL divergence in Definition 4.4.1 is evaluated is a sets of the form

$$Y = B(y_t, \epsilon_y) = \{y \in \mathcal{Y} : |y_t - y| < \epsilon_y\},$$

with $y_t \in B(y_{t-1}, \epsilon_y)$ and $s_{\lambda,t-1}, s_{\lambda,t} \in B(0, \epsilon_\lambda)$. The result is stated in the following proposition.

Proposition 4.4.3. *Let Assumption 4.4.1 hold. Then, the parameter update from $\lambda_t(f_t)$*

to $\lambda_{t+1}(f_{t+1})$ generated by (4.6) dominates almost surely the the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ in local realized KL reduction for every $\lambda_{t-1} \in \Lambda$ and $f_t \in \mathcal{F}$.

The result in Proposition 4.4.3 is related to the fact that when the updating steps are small enough and the information provided by the data changes smoothly, y_{t-1} is close to y_t, then the the update from λ_{t-1} to $\lambda_t(f_t)$ and the update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ are in the same direction. In this situation, the score update for f_t leads to $f_{t+1} > f_t$ and therefore an update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ in the same direction as the update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ but larger in absolute value. This means that for some small enough $s_{\lambda,t-1}$ and $s_{\lambda,t}$ the update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ reduces the local KL divergence more than the update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$.

4.5 Monte Carlo experiment

In this section, we present a simulation exercise as an intuitive example of the role that the time-varying parameter α_t can play. The simulation study consists on generating time series from a stochastic process and comparing the predictive ability of GAS and aGAS models. The time series are generated by the following DGP

$$y_t = \mu_t^o + \eta_t, \quad t \in \mathbb{Z}, \tag{4.7}$$

where μ_t^o is a deterministic mean and $\{\eta_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Gaussian random variables with zero mean and unit variance. The deterministic mean μ_t^o takes values in $\{0, \delta\}$, $\delta > 0$, and is defined to switch every $\gamma \times 10^2$ time periods from 0 to δ and vice versa. More formally, μ_t^o is specified as

$$\mu_t^o = \begin{cases} 0 & \text{if} \quad \sin\left(\gamma^{-1}10^{-2}(\pi t - 1)\right) \geq 0 \\ \delta & \text{if} \quad \sin\left(\gamma^{-1}10^{-2}(\pi t - 1)\right) < 0. \end{cases}$$

Figure 4.5.1 shows a realization from the DGP with $\delta = 3$ and $\gamma = 2$. We consider this particular DGP to provide an intuition of why the time-varying α_t of the aGAS model can be relevant. The idea is that, in time periods where the true μ_t^o is constant, we would like the noise component η_t not to affect too much the filtered path of the mean. This reflects a situation with a small α_t. On the other hand, we would like the filtered mean to react when the breaks in the level occur to attain quickly the new level of μ_t^o. This reflects a situation with a large α_t.

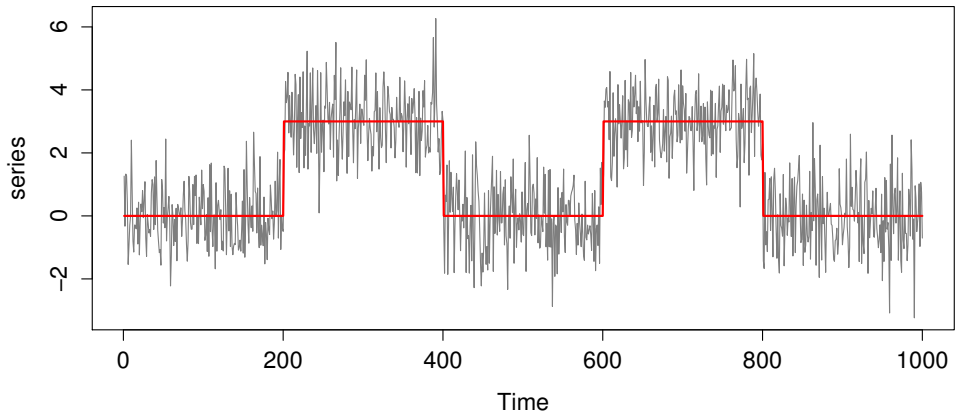

Figure 4.5.1: *Realization of size* $T = 1000$ *from the DGP with* $\delta = 3$ *and* $\gamma = 2$*. The red line represent the true* μ_t^o.

To filter the simulated series, the following aGAS model is considered

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{4.8}$$

The time-varying mean μ_t is specified as

$$\mu_{t+1} = \mu_t + \alpha_t s_{\mu,t},$$
$$\alpha_t = \exp(f_{t+1}/2), \; f_{t+1} = \omega_f + \beta_f f_t + \alpha_f s_{f,t},$$

where $s_{\mu,t} = y_t - \mu_t$ and $s_{f,t} = s_{\mu,t} s_{\mu,t-1}$. The expressions for the innovations of μ_t and α_t are obtained from the score of the predictive likelihood as in (4.3) and (4.4). Note that, in this case, the Fisher information is constant and therefore the scaling of the score is irrelevant as it only leads to a reparametrization of the model. We consider also the GAS model obtained treating α_t as a static parameter to be estimated, i.e. $\alpha_t = \alpha_\mu$. This GAS model is equivalent to an ARIMA(0,1,1) model. In particular, taking first differences we obtain an MA(1) model $y_t - y_{t-1} = (1 - \alpha_\mu)\varepsilon_{t-1} + \varepsilon_t$.

We generate 1000 Monte Carlo replications of sample size $T = 1000$ from the process in (4.7) for different values of δ and γ. For each of the 1000 replications, we estimate by ML the aGAS model in (4.8) and its standard GAS counterpart. In order to evaluate the performance of the models, the filtered mean μ_t from these two models is compared with the true mean μ_t^o. We compute the square root of the mean square error (MSE) between the filtered μ_t and true mean μ_t^o. The results of the experiment are collected in Table 4.5.1. The results show that the aGAS model outperforms the GAS model. In particular, the MSE of the aGAS model is smaller for all DGPs except for the DGP with $\delta = 0$. This indicates that the aGAS filter is able to better approximate the true μ_t^o in terms of

	$\gamma = 1.0$		$\gamma = 1.5$		$\gamma = 2.0$		$\gamma = 2.5$	
	GAS	aGAS	GAS	aGAS	GAS	aGAS	GAS	aGAS
$\delta = 0.0$	**3.86**	3.99	**3.86**	3.99	**3.86**	3.99	**3.86**	3.99
$\delta = 0.5$	22.34	**22.33**	20.19	**20.19**	18.17	**18.13**	17.05	**16.94**
$\delta = 1.0$	31.69	**31.40**	28.57	**28.07**	25.70	**24.91**	23.99	**22.89**
$\delta = 1.5$	39.21	**38.13**	35.25	**33.56**	31.66	**29.14**	29.48	**26.31**
$\delta = 2.0$	45.78	**43.50**	41.05	**37.62**	36.81	**31.97**	34.21	**28.47**
$\delta = 2.5$	51.78	**48.29**	46.30	**41.26**	41.45	**34.64**	38.47	**30.60**
$\delta = 3.0$	57.38	**53.09**	51.18	**45.02**	45.75	**37.58**	42.40	**32.83**
$\delta = 3.5$	62.71	**58.05**	55.80	**48.98**	49.79	**40.91**	46.08	**35.54**

Table 4.5.1: *Square root of the MSE between the true μ_t^o and the filtered parameter μ_t for different values of δ and γ.*

quadratic error. The fact that the GAS performs better than the aGAS for $\delta = 0$ is quite natural as $\delta = 0$ means that the true mean μ_t^o is constant in all time periods. Therefore, there are no benefits from using a dynamic α_t but only the drawback of having a more parametrized model that leads to an higher parameter estimation uncertainty. Similarly, from Table 4.5.1, we can also note that the improvement due to the dynamic parameter α_t tends to increase as the size of the jumps increases.

To better understand the effect of the dynamic parameter α_t, Figure 4.5.2 reports the simulation results for the DGP with $\delta = 3$ and $\gamma = 2$. As we can see from the first plot, the 90% variability bounds for the aGAS are narrower than those of the GAS in time periods when μ_t^o is constant. This shows that the true mean is predicted with greater accuracy and the filter is less exposed to the noise component. The opposite situation can be noted after the breaks: the variability bounds of the aGAS are larger for a few time periods. This is consistent with the fact that after the brakes the aGAS filter is reacting faster to handle the changes in the level and thus it is also more exposed to the disturbance component. From the second plot in Figure 4.5.2, we can note how, in different time periods, the squared errors tends to be larger for the GAS model. Furthermore, the 90% level confidence bounds show that aGAS model seems to outperform the GAS not only on average but for almost all Monte Carlo random draws. Finally, the third plot in Figure 4.5.2 illustrates the behavior of the time-varying α_t. In particular, the dashed line represents the average filtered α_t from the aGAS model and the continuous line the average estimate of the static α_μ from the GAS model. As expected, the dynamic α_t is close to zero when μ_t^o is constant and it increases after the breaks. This allows the filtered mean to be updated at different speeds in different time periods and leads to the advantages illustrated in the first two

plots of Figure 4.5.2.

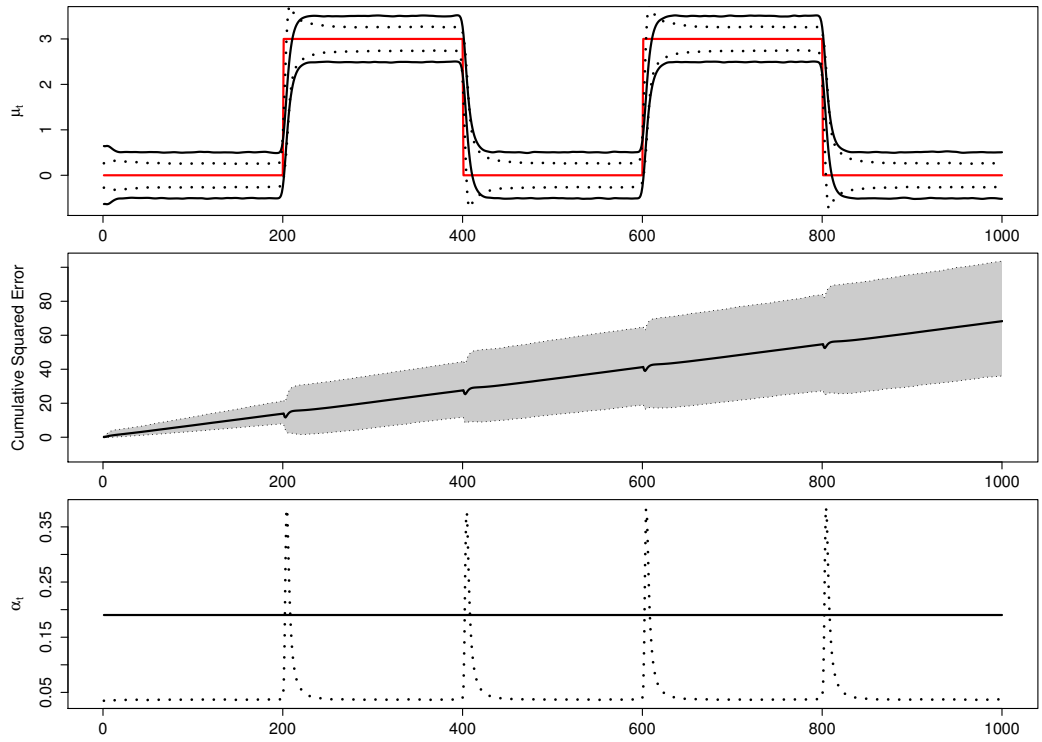

Figure 4.5.2: *First plot: the red line represents μ_t^o, the continuous lines represent 90% variability bounds for the GAS μ_t, and the dashed lines represent 90% variability bounds for the aGAS μ_t. Second plot: cumulative squared error difference between the aGAS and the GAS. The shadowed area denotes a 90% confidence region. Third plot: the continuous line is the average estimate of α for the GAS, and the dashed line is the average estimate of α_t for the aGAS.*

4.6 Empirical application to US stock returns

In this section, we evaluate the performance of the aGARCH model through a comparison using the stocks that are currently in the S&P 500 index. Daily stock returns from 2008 to 2015 are considered. The series of the S&P 500 that are not available since 2008 are excluded from the study. The resulting number of time series is 460. The performances of the models are evaluated both in-sample and out-of-sample. The in-sample evaluation is based on the AIC. This choice is due to the fact that GAS models can be seen as filters in a misspecified framework and the the AIC provides a meaningful interpretation in this case. The out-of-sample evaluation is based on the log-score criterion: $n^{-1}\sum_{t=1}^{n}\log p_{T+i}(y_{T+i})$, where $p_t(\cdot)$ denotes the conditional density of y_t given the past

observations up to $t-1$. This criterion is widely known and used in the literature for evaluating density forecasts. The out-of-sample period consists on the daily observations in 2015. The training sample is from 2008 to 2014. The static parameters are estimated only once, i.e. no expanding or rolling windows are used.

	Full dataset				Top 10% Kurtosis			
	In-sample		Out-of-sample		In-sample		Out-of-sample	
	No.	Pct.	No.	Pct.	No.	Pct.	No.	Pct.
GARCH	72	15.6%	89	19.3%	0	0.0%	13	28.3%
ELGARCH	264	57.4%	268	58.3%	11	23.9%	15	32.6%
aGARCH	124	27.0%	103	22.4%	35	76.1%	18	39.1%
Total	460	100.0%	460	100.0%	46	100.0%	46	100.0%

Table 4.6.1: *Number and the percentage of series in the S&P 500 index where each Gaussian model outperforms the others.*

We first perform the comparison considering a Gaussian error distribution. The models we consider are the GARCH, ELGARCH and aGARCH, where ELGARCH indicates the GARCH model of Engle and Lee (1999). Table 4.6.1 reports the number of series in the S&P 500 index where each model outperforms the others. The table also contains the results considering the 10% of the S&P 500 series with the highest Kurtosis. The aGARCH has the smallest AIC for 27.0% of the series, whereas the ELGARCH has the smallest AIC in 57.4% of the cases. We note that the aGARCH model seems to perform particularly well with series that present heavy tails. In fact, considering only the 10% of the series with highest kurtosis, the aGARCH is the model that performs best for the majority of the series. This peculiarity is further highlighted by Figures 4.6.1 and 4.6.2. Figure 4.6.1 shows that the aGARCH model performs better than the other models more often when fat-tailed series are considered. In particular, we see that the performance increases as we condition the comparison on series with fatter tails. The Boxplots in Figure 4.6.2 shows the distribution of the Kurtosis for the S&P 500 series grouped according to the best performing model. This plot indicates that the series where the aGARCH performs best tend to present fat tails, whereas, the series where the GARCH performs best tend to have a low Kurtosis. We thus conclude that, in general, the more complex models, i.e. the ELGARCH and the aGARCH, seem to outperform the standard GARCH model when fat tails are present.

The aGARCH model in (4.1) and (4.2) is obtained from the aGAS framework with a Gaussian distribution for the error term. Other distributions can be employed. Considering a Student-t distribution, we can extend the Beta-t-GARCH model of Creal et al.

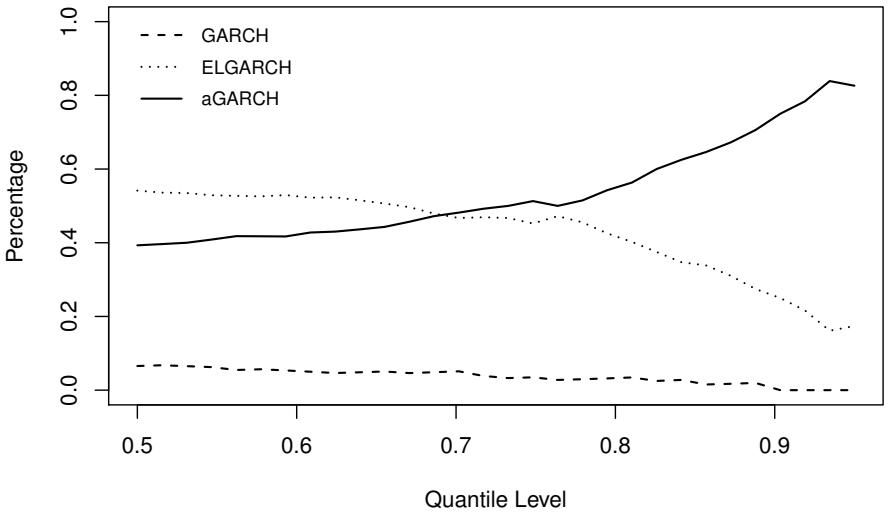

Figure 4.6.1: *Percentage of series where each model outperforms the others in terms of AIC. The percentage is computed only for the series with skewness above a certain quantile. The quantile levels are indicated on the horizontal axis.*

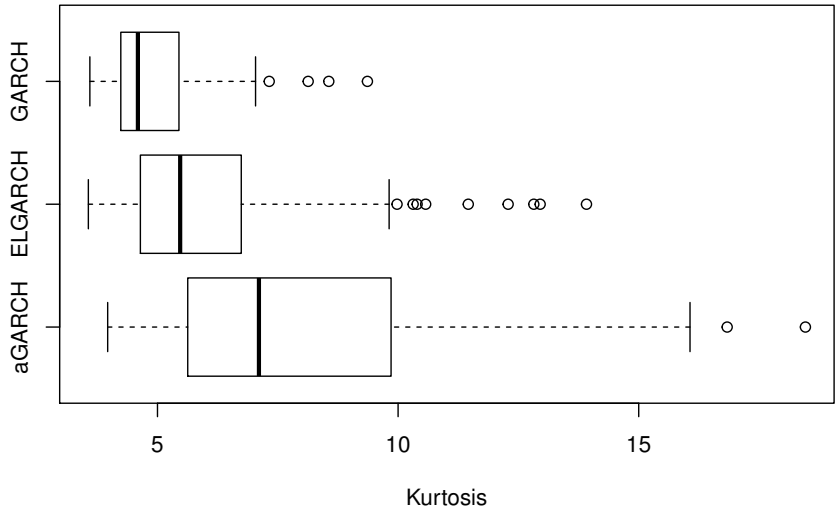

Figure 4.6.2: *Boxplots of the Kurtosis distribution of the series grouped by model. The series in each model are those where that model has the best in-sample performance.*

(2013) and Harvey (2013). The accelerating Beta-t-GARCH (aBeta-t-GARCH) model is described by the following equations

$$y_t = \sigma_t \varepsilon_t, \quad \sigma_{t+1}^2 = \omega + \beta \sigma_t^2 + \alpha_t \sigma_t^2 s_{\sigma,t},$$
$$\alpha_t = \beta \, \text{logit}^{-1}(f_{t+1}), \quad f_{t+1} = \omega_f + \beta_f f_t + \alpha_f s_{\sigma,t} s_{\sigma,t-1},$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Student-t distributed random variables with zero mean unit variance and ν degrees of freedom. As in Creal et al. (2013), the score innovation $s_{\sigma,t}$ has the following expression

$$s_{\sigma,t} = \frac{(\nu+1)\varepsilon_t^2}{(\nu-2)+\varepsilon_t^2} - 1.$$

It is easy to see that the limit case $\nu \to \infty$ of the aBeta-t-GARCH model coincides with the aGARCH model. Furthermore, setting $\alpha_t = \alpha$ to be a static parameter leads to the Beta-t-GARCH model of Creal et al. (2013) and Harvey (2013).

In the following, we perform a second empirical study where we compare models with a Student-t error distribution. The models considered are the Beta-t-GARCH and aBeta-t-GARCH as well as the GARCH, ELGARCH and aGARCH with Student-t error distribution, which we denote as tGARCH, ELtGARCH and atGARCH respectively. We note that the Beta-t-GARCH specification takes into account the fat tails not only in the error distribution but also in the updating mechanism of the variance σ_t^2. Namely, the impact of extreme observations on σ_t^2 is attenuated. As discussed in Creal et al. (2013) and Harvey (2013) this can provide benefits when dealing with fat tailed time series. Similarly as before, Table 4.6.2 reports the number of series in the S&P 500 index where each model is outperforming the others. The in-sample results shows that the Beta-t-GARCH is the best model for 67.6% of the series. However, this result seems not to be very consistent with the out-of-sample results where the Beta-t-GARCH is the best in only the 22.4% of cases. The atGARCH model and the aBeta-t-GARCH are the best models for a significant proportion of the series. As before, we can look at the results for the 10% of the series with highest Kurtosis. The Beta-t-GARCH and the aBeta-t-GARCH are the best in-sample specifications for all series. The out-of-sample results are also rather coherent with this finding. Overall the aGARCH and aBeta-t-GARCH models are the best models for a large proportion of the series.

We can conclude that, for a relevant number of the S&P 500 series, the inclusion of the dynamic α_t can enhance the in-sample and the out-of-sample performances of GARCH-type models. This is true for the Normal experiment as well as the Student-t experiment. Moreover, in both cases, the effect of the dynamic α_t seems particularly relevant for fat tailed time series. These results suggest that different specifications can be useful to better approximate the dynamics of different series. The accelerating volatility framework thus provides a flexible class of models that can be useful in practical applications.

	Full dataset				Top 10% Kurtosis			
	In-sample		Out-of-sample		In-sample		Out-of-sample	
	No.	Pct.	No.	Pct.	No.	Pct.	No.	Pct.
tGARCH	29	6.3%	14	3.0%	0	0.0%	0	0.0%
ELtGARCH	49	10.7%	186	40.4%	0	0.0%	4	8.7%
atGARCH	21	4.6%	66	14.3%	0	0.0%	6	13.0%
Beta-t-GARCH	311	67.6%	103	22.4%	39	84.8%	19	41.3%
aBeta-t-GARCH	50	10.9%	91	19.8%	7	15.2%	17	37.0%
Total	460	100.0%	460	100.0%	46	100.0%	46	100.0%

Table 4.6.2: *Number and the percentage of series in the S&P 500 index where each Student-t model outperforms the others.*

4.7 Application to US inflation

4.7.1 A fat tailed aGAS location model

Relying on the aGAS framework, we propose a fat-tailed model where the parameter that determines the magnitude of the update of the mean process is allowed to vary over time. More specifically, we consider a Student-t conditional distribution for y_t where both the mean and the variance are time varying. As we will see, the resulting model has some similarities with the stochastic volatility model of Stock and Watson (2007). The Student-t distribution in a GAS framework allows us to handle outliers by attenuating their impact on the filtered parameters. Applications in the literature of Student-t GAS models for location and scale parameters can be found in Creal et al. (2013), Harvey (2013) and Harvey and Luati (2014). In particular, Harvey (2013) considered a Student-t model with both a time-varying mean and the variance. The novelty of the model we propose in the following is the inclusion of the time-varying parameter α_t to enable the time-varying mean to capture more complex dynamics.

We consider the following aGAS model with time-varying conditional mean and volatility

$$y_t = \mu_t + \sigma_t \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} t_v, \quad t \in \mathbb{Z}, \tag{4.9}$$

The time-varying parameters are described by the following equations

$$\begin{aligned} \mu_{t+1} &= \mu_t + \alpha_t s_{\mu,t}, \\ \alpha_t = \exp(f_{t+1}/2), \; f_{t+1} &= \omega_f + \beta_f f_t + \alpha_f s_{f,t}, \\ \log \sigma^2_{t+1} &= \omega_\sigma + \beta_\sigma \log \sigma^2_t + \alpha_\sigma s_{\sigma,t}, \end{aligned}$$

where ω_f, β_f, α_f, ω_σ, β_σ and α_σ are static parameters to be estimated and $s_{\mu,t}$, $s_{f,t}$ and $s_{\sigma,t}$ are the score-based innovations of the processes. In the following, the functional form of these innovations is illustrated. A graphical representation is presented in Figure 4.7.1. The innovation $s_{\mu,t}$ of the mean process μ_t is obtained setting the scaling factor $S_{\mu,t}$ equal to the square root of the inverse Fisher information, $s_{\mu,t}$ takes the form

$$s_{\mu,t} = \frac{(v+1)(y_t - \mu_t)\sigma_t^{-1}}{(v-2) + (y_t - \mu_t)^2\sigma_t^{-2}}.$$

The first plot in Figure 4.7.1 shows the effect of a standardized observation $\varepsilon_t = (y_t - \mu_t)/\sigma_t$ on $s_{\mu,t}$. As we can notice the relationship between ε_t and $s_{\mu,t}$ is nonlinear and the impact of extreme values of ε_t on $s_{\mu,t}$ is attenuated. The degree of attenuation depends on the parameter v: the smaller the parameter v, the lower the sensitivity of $s_{\mu,t}$ to outliers; see Harvey and Luati (2014) for more details. The innovation $s_{f,t}$ is derived from expression (4.5) setting $C_{f,t} = S_{\mu,t}S_{\mu,t-1}$

$$s_{f,t} = s_{\mu,t}s_{\mu,t-1}.$$

The second plot in Figure 4.7.1 shows the effect of ε_t and ε_{t-1} on $s_{f,t}$. As we can see $s_{f,t}$ is positive when ε_t and ε_{t-1} have the same sign and negative when ε_t and ε_{t-1} have opposite sign. Also in this case extreme values of ε_t and ε_{t-1} are detected as outliers and their impact on $s_{f,t}$ is attenuated. Finally, the innovation of the process $\log\sigma_t^2$ takes the form

$$s_{\sigma,t} = \frac{(v+1)(y_t - \mu_t)^2\sigma_t^{-2}}{(v-2) + (y_t - \mu_t)^2\sigma_t^{-2}} - 1.$$

Note that in this case the Fisher information is constant and so it does not affect the functional form of $s_{\sigma,t}$. The impact of ε_t on $s_{\sigma,t}$ is shown in the third plot of Figure 4.7.1. This update $s_{\sigma,t}$ is the same as for the Beta-t-EGARCH model of Harvey (2013).

As the degrees of freedom of the Student-t distribution goes to infinity, the Student-t distribution approaches the standard Gaussian distribution. In this limit case, the model in (4.9) becomes a Gaussian score-driven model where the innovation for μ_t is given by $s_{\mu,t} = (y_t - \mu_t)\sigma_t^{-1}$ and the innovation for σ_t^2 is given by $s_{\sigma,t} = (y_t - \mu_t)^2\sigma_t^{-2} - 1$. The impact function of the standardized observation $(y_t - \mu_t)\sigma_t^{-1}$ on $s_{\mu,t}$ and $s_{\sigma,t}$ can be seen in Figure 4.7.1.

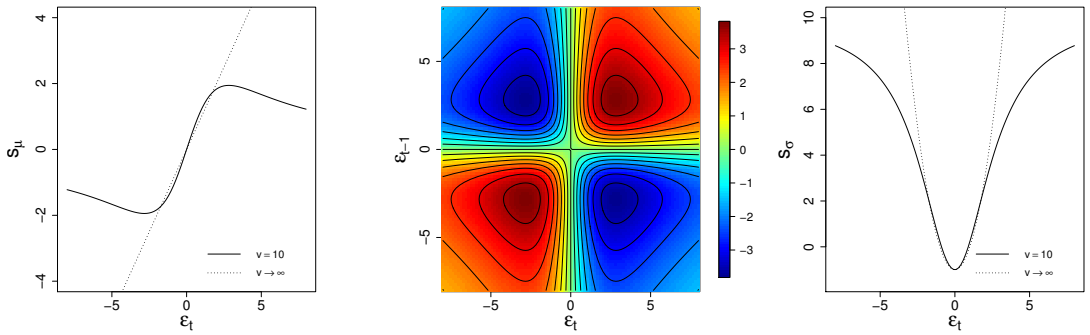

Figure 4.7.1: *In the first image the values taken by $s_{\mu,t}$ as a function of ε_t are plotted. In the second image there is a contour plot that shows the values taken by $s_{f,t}$ as a function of ε_t and ε_{t-1}. In the third image the values taken by $s_{\sigma,t}$ as a function of ε_t are plotted. In all plots the degrees of freedom of the Student-t is set equal to 10.*

4.7.2 Empirical application

In our empirical analysis, we consider the US quarterly consumer price index, which is obtained from the FRED dataset. As standard procedure adopted in the literature, the inflation time series y_t is computed as the annualized log-difference of the price index series p_t, namely, the transformation $y_t = 400 \log(p_t/p_{t-1})$ is considered. The resulting inflation series is from the first quarter of 1952 to the first quarter of 2015. The series is plotted in Figure 4.7.2. We consider several specifications of the aGAS model. These specifications are listed in Table 4.7.1.

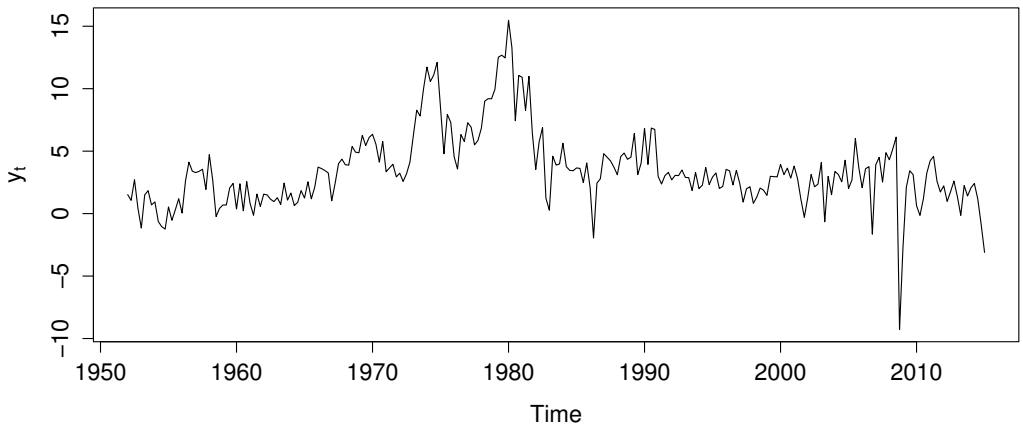

Figure 4.7.2: *Quarterly consumer price US inflation series.*

	Description	Reference
Model t.1	The full model in (4.9)	
Model t.2	$\beta_\sigma = 0$ and $\alpha_\sigma = 0$	
Model t.3	$\beta_f = 0$ and $\alpha_f = 0$	Harvey (2013)
Model t.4	$\beta_\sigma = 0$, $\alpha_\sigma = 0$, $\beta_f = 0$ and $\alpha_f = 0$	Harvey and Luati (2014)
Model n.1	Limit case of Model t.1 with $v \to \infty$	
Model n.2	Limit case of Model t.2 with $v \to \infty$	
Model n.3	Limit case of Model t.3 with $v \to \infty$	
Model n.4	Limit case of Model t.4 with $v \to \infty$	

Table 4.7.1: *The second column describes the specification of the model. The third column provides some references for the specific models obtained constraining the parameters of the full model in (4.9).*

	δ_f	β_f	α_f	δ_σ	β_σ	α_σ	v	loglik	LRT	AIC
Model t.1	-1.518 (0.799)	0.967 (0.027)	0.258 (0.113)	1.055 (0.236)	0.861 (0.092)	0.215 (0.089)	5.571 (1.572)	-475.3	-	**964.6**
Model t.2	-1.493 (0.402)	0.914 (0.028)	0.294 (0.071)	1.182 (0.178)	-	-	3.826 (0.553)	-482.7	0.001	975.4
Model t.3	-0.468 (0.280)	-	-	1.080 (0.207)	0.869 (0.126)	0.163 (0.099)	7.583 (2.399)	-481.8	0.002	973.6
Model t.4	-0.305 (0.213)	-	-	1.111 (0.134)	-	-	5.639 (1.431)	-488.8	0.000	983.6
Model n.1	-1.366 (0.618)	0.969 (0.022)	0.182 (0.072)	1.169 (0.203)	0.937 (0.030)	0.088 (0.033)	-	-504.2	-	1020.4
Model n.2	-0.304 (0.416)	0.971 (0.028)	0.060 (0.036)	1.251 (0.089)	-	-	-	-515.3	0.000	1038.6
Model n.3	-0.231 (0.314)	-	-	1.213 (0.161)	0.939 (0.026)	0.054 (0.021)	-	-510.2	0.002	1028.4
Model n.4	-0.080 (0.266)	-	-	1.264 (0.089)	-	-	-	-516.8	0.000	1037.6

Table 4.7.2: *Estimate of the models in Table 4.7.1. Standard errors are in brackets. The last three columns contain respectively the log-likelihood, the pvalue of the likelihood ratio test with respect to the full models and the AIC. The parameters δ_f and δ_σ are given by $\delta_f = \omega_f/(1-\beta_f)$ and $\delta_\sigma = \omega_\sigma/(1-\beta_\sigma)$.*

The estimation results of Model t.1-t.4 and Model n.1-n.4 are collected in Table 4.7.2. The table reports the pvalue of the likelihood ratio test between each model and the corresponding full model. The results show that the inclusion of the dynamic variance σ_t as well as α_t are highly significant. In particular, we obtain that the null hypothesis of the of the likelihood ratio test is rejected at a 1% level in all cases. Furthermore, we report that the model with the lowest AIC is Model t.1. The AIC also indicates that the Student-t specifications, Model t.1-t.4, have a better fitting than the Normal ones, Model n.1-n.4. This is also confirmed by the fact that the estimated degrees of freedom v are small for all

four Student-t models.

Figure 4.7.3 contains the plots of the filtered parameters μ_t, σ_t and α_t for Model t.1. From the plot of μ_t, we can see how the model is effectively able to handle outliers. This is particularly clear in the fourth quarter of 2008 where the extreme peak of inflation does not dramatically affect μ_t. From the plot of α_t, we can note that during the period of exceptional high inflation, approximately between 1972 and 1983, also the filtered α_t takes high values. This is consistent with the fact that during periods of persistent and quick changes in the level of y_t the parameter μ_t has to be updated quickly to capture these changes and α_t plays a key role in this. Finally, as we can see in the third plot of Figure 4.7.3, the variability σ_t seems to increase in periods of economic recession. See the NBER recession index in the first plot.

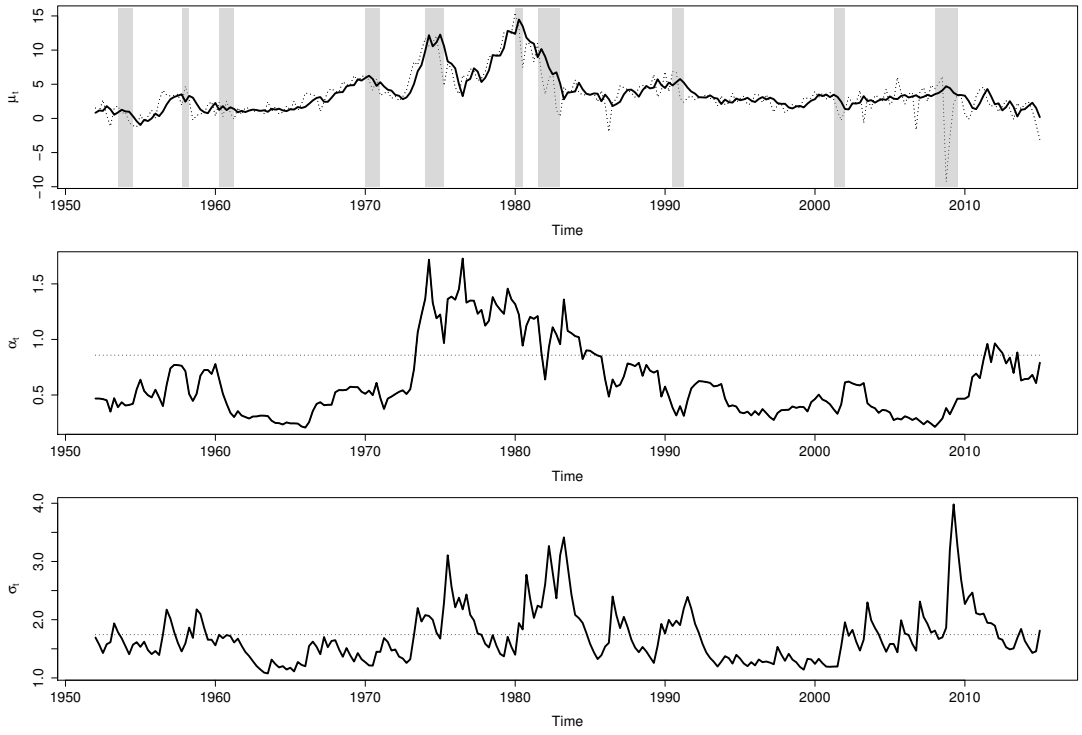

Figure 4.7.3: *Estimated time-varying parameters from Model t.1. First plot:* μ_t*. Second plot:* α_t*. Third plot:* σ_t*.*

In order to better appreciate the effect of the inclusion of the time-varying parameter α_t on the filtered μ_t, we compare the filtered μ_t obtained from Model t.1 and Model t.3. Note that both Model t.1 and t.3 include a time-varying variability σ_t, the only difference between the two models is that in Model t.3 α_t is not time varying. We consider two periods where the inflation series exhibits different behaviors: the period from 1973 to 1982, first plot in Figure 4.7.4, and the the period from 1999 to 2008, second plot in

Figure 4.7.4. In the period between 1973 and 1982, we note that the time series seems to change level quickly, denoting an high persistence in the inflation process. In this period the time-varying α_t takes large values, see the second plot in Figure 4.7.3. This allows the μ_t of Model t.1 to react more promptly to the changes in the level of the series. This fact can be easily noted from the plot as the μ_t of Model t.1 goes above the μ_t of Model t.3 when the inflation level is increasing and vice versa when the inflation level is decreasing. As concerns the period between 1999 and 2008, the second plot in Figure 4.7.4 shows that the inflation series seems to change level slowly, a slight increasing trend with a lot of noise around it. In this situation, the small values of the time-varying α_t, see the second plot in Figure 4.7.3, allow the μ_t of Model t.1 to change slowly, capturing the increasing trend but not being too much affected by the noise. The benefit of the time-varying α_t can be noted from the plot as the filtered μ_t of Model t.3 is more noisy than the filtered μ_t of Model t.1. These two plots in Figure 4.7.4 show how the inclusion of the time-varying α_t allows the model to be more flexible and better adapt to changing behaviors of the series. The improvement in terms of in-sample fitting is also confirmed by the likelihood ratio test and the AIC.

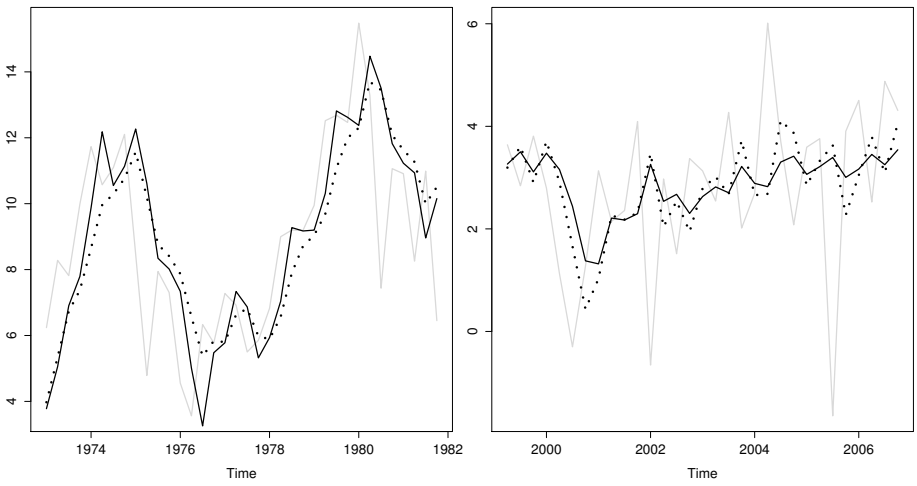

Figure 4.7.4: *Filtered μ_t from Model t.1 and Model t.3 for different time periods. The gray line is the inflation series, the dashed line is the filtered μ_t from Model t.3 and the continuous line is the filtered μ_t from Model t.1.*

4.7.3 Pseudo out-of-sample forecasts

Finally, we perform a pseudo out-of-sample study to compare the forecasting performance of the models in Table 4.7.1. In this study we include also other models: a local level model, an ARIMA(4,1,0) and an ARIMA(1,1,1). The forecast mean square error (FMSE)

and the forecast mean absolute error (FMAE) are computed using the last 100 observations and the estimation of the models is performed using a fixed rolling window. We consider forecasts from 1 to 4 steps ahead. Differences in forecast accuracy are tested by Diebold and Mariano (DM) test, Diebold and Mariano (1995). The DM test is used to test the null hypothesis that Model t.1 has the same FMSE as the other models against the alternative of different FMSE. Note that the DM test is performed for both nested and non-nested models; the asymptotic normal distribution of the DM test statistic for nested models is ensured by the fixed rolling window, see Giacomini and White (2006).

As we can see from the results collected in Table 4.7.3, Model n.1 has the smallest FMSE and FMAE and Model t.1 has the best forecasting performance among the fat-tailed models. This suggests that the inclusion of the time-varying α_t tends to enhance the forecasting performance of the GAS models. For forecasting horizon of 1 year ($h = 4$), we obtain that Model t.1 significantly outperforms most of the models at a 5% or 10% significance level. As concerns the other forecasting horizons, we conclude that we cannot reject the hypothesis that the differences in terms of forecast accuracy observed in the subsample are just due by chance.

	FMSE				FMAE			
	$h=1$	$h=2$	$h=3$	$h=4$	$h=1$	$h=2$	$h=3$	$h=4$
Model t.1	1.00	1.00	1.00	**1.00**	**1.00**	1.00	1.00	**1.00**
Model t.2	1.02	1.02	1.02	1.05	1.01	1.02	1.01	1.03
Model t.3	1.11	1.12	1.09	1.14**	1.04	1.04	1.03	1.07**
Model t.4	1.13*	1.14*	1.09	1.16**	1.05	1.06*	1.02	1.08**
Model n.1	**0.96**	**0.99**	**0.98**	**1.00**	**1.00**	**0.99**	**0.99**	**1.00**
Model n.2	1.02	1.20	1.18	1.15*	1.02	1.09	1.07	1.05
Model n.3	1.03	1.09	1.07	1.10	1.04	1.03	1.03	1.04
Model n.4	1.02	1.20	1.18	1.15*	1.02	1.09	1.07	1.05*
Local level model	1.02	1.20	1.19	1.16*	1.02	1.09	1.07	1.06*
ARIMA(4,1,0)	1.06	1.25	1.33	1.25**	1.02	1.07	1.10	1.10**
ARIMA(1,1,1)	0.98	1.16	1.14	1.12	**1.00**	1.06	1.04	1.03

Table 4.7.3: *FMSE and FMAE ratio from the last 100 observations of the quarterly US consumer price inflation series. The benchmark is Model t.1. The FMSE and FMAE of Model t.1 is at the denominator of the ratio.*

4.8 Conclusion

This chapter has introduced a novel class of models for time-varying parameters capable of describing complex dynamics. We have provided both theoretical and simulation-based evidence that the aGAS formulation can outperform GAS models with a time-invariant structure for the updating equation. The real data applications to the S&P 500 and the US inflation series have illustrated that the proposed accelerating approach is capable of improving in-sample and out-of-sample performances of GAS models.

Appendix

4.A Proofs

Proof of Lemma 4.4.1. The first statement follows by noting that $\lambda_t(f_t) = \lambda_t^*$ if $\{f_t\}_{t\in\mathbb{N}}$ is a random sequence such that $f_t = h^{-1}\big((\lambda_t^* - \omega_\lambda - \beta_\lambda\lambda_{t-1})/s_{\lambda,t-1}\big)$ for any $t \in \mathbb{N}$. As concerns the second statement, the if part is immediately proved by noting that $s_{\lambda,t-1} = (\lambda_t^* - \omega_\lambda - \beta_\lambda\lambda_{t-1})/h(c)$ implies $f_t = c$. Finally, to prove the only if part of the statement, suppose that, for some $t \in \mathbb{N}$, there exists no $c \in \mathbb{R}$ such that $s_{\lambda,t-1} = (\lambda_t^* - \omega_\lambda - \beta_\lambda\lambda_{t-1})/h(c)$. Then, setting $f_t = c\ \forall\ t \in \mathbb{N}$ implies that $\lambda_t(f_t) \neq \lambda_t^*$ for some $t \in \mathbb{N}$ and any possible $c \in \mathbb{R}$. □

Proof of Proposition 4.4.1. The proof follows the same argument as in Blasques et al. (2015). By an application of the mean value theorem, the local realized KL divergence can be expressed as

$$
\begin{aligned}
\Delta_{f,t}^{t+1} &= \int_{B(y_t,\epsilon_y)} p_t^o(y) \log \frac{p(y|\lambda_t(f_t))}{p(y|\lambda_t(f_{t+1}))} dy = \\
&= \int_{B(y_t,\epsilon_y)} p_t^o(y) \frac{\partial \log p(y|\lambda_t(\dot{f}_t))}{\partial \dot{f}_t} (f_t - f_{t+1}) dy = \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \alpha_f C_{f,t} S_{\lambda,t-1} u_\lambda\big(y_{t-1},\lambda_{t-1}\big)^2 u_\lambda\big(y,\lambda_t(\dot{f}_t)\big) u_\lambda\big(y_t,\lambda_t(f_t)\big) dy = \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y,\lambda_t(\dot{f}_t)\big) u_\lambda\big(y_t,\lambda_t(f_t)\big) dy,
\end{aligned}
$$

where $\tilde{C}_t = \alpha_f C_{f,t} S_{\lambda,t-1} u_\lambda(y_{t-1},\lambda_{t-1})^2$ and $\dot{f}_t$ is a point between f_t and f_{t+1}. Applying

again the mean value theorem it results

$$\Delta_{f,t}^{t+1} = -\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y,\lambda_t(\dot{f}_t)\big)u_\lambda\big(y_t,\lambda_t(f_t)\big)dy =$$

$$= -\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y_t,\lambda_t(f_t)\big)^2 dy + \tag{4.10}$$

$$-\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y_t,\lambda_t(f_t)\big)\frac{\partial u_\lambda\big(\dot{y}_t,\lambda_t(\ddot{f}_t)\big)}{\partial \dot{y}_t}(y-y_t)dy + \tag{4.11}$$

$$-\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y_t,\lambda_t(f_t)\big)\frac{\partial u_\lambda\big(\dot{y}_t,\lambda_t(\ddot{f}_t)\big)}{\partial \ddot{f}_t}(\dot{f}_t-f_t)dy, \tag{4.12}$$

where $\ddot{f}_t$ is a point between $\dot{f}_t$ and f_t, and $\dot{y}_t$ is a point between y and y_t. The desired result follows from the fact that the term (4.10) is a.s. negative and the terms (4.11) and (4.12) can be made arbitrary small in absolute value compared to the first term by selecting the ball radius ϵ_y and ϵ_f small enough. □

Proof of Proposition 4.4.2. The if part of the proposition follows immediately from a similar argument as in the proof of Proposition 4.4.1. As concerns the only if part, if $sign(f_{t+1}-f_t) = sign(s_{f,t})$ does not hold with probability 1 for any $f_t \in \mathcal{F}$, it means that there exists an $f_t \in \mathcal{F}$ such that $sign(f_{t+1}-f_t) \neq sign(s_{f,t})$ holds with positive probability. Following a similar argument as in the proof of Proposition 4.4.1, this implies that there is a positive probability to have an y_t such that

$$\Delta_{f,t}^{t+1} = -\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y_t,\lambda_t(\dot{f}_t)\big)(f_{t+1}-f_t)dy > 0,$$

for small enough $\epsilon_y > 0$ and $\epsilon_f > 0$. This concludes the proof. □

Proof of Proposition 4.4.3. The line of argument is similar as in the proof of Proposition 4.4.2, the result follows by repeated applications of the mean value theorem. The difference in local KL variation can be expressed as

$$\Delta_{\lambda,t+1}^{t+1} - \Delta_{\lambda,t+1}^{t} = \int_{B(y_t,\epsilon_y)} p_t^o(y)\log\frac{p(y|\lambda_{t+1}(f_t))}{p(y|\lambda_{t+1}(f_{t+1}))}dy =$$

$$= \int_{B(y_t,\epsilon_y)} p_t^o(y)\frac{\partial \log p(y|\lambda_{t+1}(\dot{f}_t))}{\partial \dot{f}_t}(f_t-f_{t+1})dy =$$

$$= -\int_{B(y_t,\epsilon_y)} p_t^o(y)\alpha_f C_{f,t} S_{\lambda,t-1} u_\lambda\big(y_t,\lambda_t(f_t)\big)^2 u_\lambda\big(y_{t-1},\lambda_{t-1}\big)u_\lambda\big(y,\lambda_{t+1}(\dot{f}_t)\big)dy =$$

$$= -\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y_{t-1},\lambda_{t-1}\big)u_\lambda\big(y,\lambda_t(\dot{f}_t)\big)dy,$$

where $\tilde{C}_t = \alpha_f C_{f,t} S_{\lambda,t-1} u_\lambda\big(y_t, \lambda_t(f_t)\big)^2$ and $\dot{f}_t$ is a point between f_t and f_{t+1}. Applying again the mean value theorem it results

$$\begin{aligned}\Delta^{t+1}_{\lambda,t+1} - \Delta^{t}_{\lambda,t+1} &= -\int_{B(y_t,\epsilon_y)} p^o_t(y)\tilde{C}_t u_\lambda\big(y, \lambda_t(\dot{f}_t)\big) u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) dy = \\ &= -\int_{B(y_t,\epsilon_y)} p^o_t(y)\tilde{C}_t U_{1,t} U_{2,t} dy,\end{aligned}$$

where $U_{1,t}$ and $U_{2,t}$ are respectively given by

$$U_{1,t} = u_\lambda\big(y_t, \lambda_t(f_t)\big) + \frac{\partial u_\lambda(\dot{y}_t, \dot{\lambda}_t)}{\partial \dot{\lambda}_t}(\lambda_{t+1}(\dot{f}_t) - \lambda_t(f_t)) + \frac{\partial u_\lambda(\dot{y}_t, \dot{\lambda}_t)}{\partial \dot{y}_t}(y - y_t)$$

and

$$U_{2,t} = u_\lambda\big(y_t, \lambda_t(f_t)\big) + \frac{\partial u_\lambda(\ddot{y}_t, \ddot{\lambda}_t)}{\partial \ddot{\lambda}_t}(\lambda_{t-1} - \lambda_t(f_t)) + \frac{\partial u_\lambda(\ddot{y}_t, \ddot{\lambda}_t)}{\partial \ddot{y}_t}(y_{t-1} - y_t),$$

with $\dot{y}_t$ a point between y_t and y, $\dot{\lambda}_t$ a point between $\lambda_t(f_t)$ and $\lambda_{t+1}(\dot{f}_t)$, $\ddot{y}_t$ a point between y_{t-1} and y_t and $\ddot{\lambda}_t$ a point between λ_{t-1} and $\lambda_t(f_t)$. Taking into account that by Assumption 4.4.1 the score $u_\lambda\big(y_t, \lambda_t(f_t)\big)$ is nonzero with probability 1, we have that the second and the third term in the expressions of $U_{1,t}$ and $U_{2,t}$ can be made arbitrary small in absolute value with respect to the first term by selecting the ball radius ϵ_y and ϵ_λ small enough. As a result, the product $U_{1,t}U_{2,t}$ can be made positive for any $\dot{y}_t, y \in B(y_t, \epsilon_y)$. This, together with the positivity of $p^o_t(y)$ and $\tilde{C}_t$, implies that $\Delta^{t+1}_{\lambda,t+1} - \Delta^{t}_{\lambda,t+1}$ is negative. □

Chapter 5

Conclusion

In this thesis several aspects of observation-driven time series modeling have been discussed. In Chapter 2, the theoretical results we obtained are useful in practical situations as the invertibility conditions can be checked empirically. This approach allows us to cover both correctly specified and misspecified models as the conditions depend only on the DGP, which is partially observable through the data. The only assumption needed is the stationarity and ergodicity of the DGP. This assumption may be restrictive in some situations. However, departures from this assumption are difficult to tackle in a general framework and model-specific studies are usually required. Furthermore, we also note that there are few results in the literature that handle non-stationarity for observation-driven models and they also usually rest on very restrictive assumptions. A possible future line of research may be the derivation of the asymptotic normality of the ML estimator. The main challenge here is to handle the general case without imposing either very high level assumptions or too restrictive conditions that are unreasonable in practical situations. The main difficulty we encountered in the derivation of asymptotic normality for this general case is related to moment conditions on the derivatives of the likelihood function. In Chapter 3, we developed a flexible class of models for count time series data. The Monte Carlo experiment and the empirical application to the crime data show that the model can outperform existing models in predicting future outcomes. The model we proposed should be interpreted as a filter and not a DGP. In this direction, we derived the consistency of the ML estimator under a general distribution of the error term. A future line of research may be the derivation of the asymptotic normality. As for the general case discussed in Chapter 2, the difficulties lie on obtaining moment conditions on the derivative processes of the likelihood function. Another possible future extension is to consider a general order p for the INAR models with dynamic coefficients. Finally,

in Chapter 4, we introduced a novel class of models that are an extension of the GAS framework. The proposed models have an intuitive interpretation as illustrated in the simulation study. They can describe changes in the amount of local information contained in the data. The optimality reasoning we presented justifies the approach in a misspecified setting. The empirical examples confirm that these models can be useful in some practical situations. Overall, we can conclude that the thesis provides several advances in observation-driven modeling that may be considered relevant from both a theoretical and an empirical perspective.

Bibliography

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In Petrov, B. and Caski, F., editors, *Proceedings of the Second International Symposium on Information Theory, Armenian SSR*, pages 267–281. Akademiai Kiado, Budapest.

Al-Osh, M. A. and Aly, E.-E. A. (1992). First order autoregressive time series with negative binomial and geometric marginals. *Communications in Statistics-Theory and Methods*, 21(9):2483–2492.

Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8(3):261–275.

Alzaid, A. and Al-Osh, M. (1990). An integer-valued pth-order autoregressive structure (INAR (p)) process. *Journal of Applied Probability*, 27(2):314–324.

Andres, P. (2014). Maximum likelihood estimates for positive valued dynamic score models; The DySco package. *Computational Statistics and Data Analysis*, 76(2):34–42.

Bec, F., Rahbek, A., and Shephard, N. (2008). The ACR Model: A Multivariate Dynamic Mixture Autoregression. *Oxford Bulletin of Economics and Statistics*, 70(5):583–618.

Berkes, I., Horváth, L., and Kokoszka, P. (2003). GARCH processes: structure and estimation. *Bernoulli*, 9(2):201–227.

Blasques, F., Koopman, S. J., Lasak, K., and Lucas, A. (2016). In-sample confidence bands and out-of-sample forecast bands for time-varying parameters in observation-driven models. *International Journal of Forecasting*, 32(3):875–887.

Blasques, F., Koopman, S. J., and Lucas, A. (2014a). Maximum Likelihood Estimation for Generalized Autoregressive Score Models. *Tinbergen Institute Discussion Paper 14-029/III*.

Blasques, F., Koopman, S. J., and Lucas, A. (2014b). Optimal Formulations for Nonlinear Autoregressive Processes. *Tinbergen Institute Discussion Paper 14-103/III*.

Blasques, F., Koopman, S. J., and Lucas, A. (2014c). Stationarity and ergodicity of univariate generalized autoregressive score processes. *Electronic Journal of Statistics*, 8(1):1088–1112.

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

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

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

Bougerol, P. and Picard, N. (1992). Strict Stationarity of Generalized Autoregressive Processes. *The Annals of Probability*, 20(4):1714–1730.

Box, G. E. P. and Jenkins, G. (1970). *Time Series Analysis, Forecasting, and Control*. Holden-Day, San Francisco.

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

Creal, D., Koopman, S. J., and Lucas, A. (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics*, 29(4):552–563.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized Autoregressive Score Models with Applications. *Journal of Applied Econometrics*, 28(5):777–795.

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

Davidson, J. (1994). *Stochastic Limit Theory*. Advanced Texts in Econometrics, Oxford University Press.

Davis, R. A., Dunsmuir, W. T. M., and Streett, S. B. (2003). Observation-driven models for Poisson counts. *Biometrika*, 90(4):777–790.

Delle Monache, D. and Petrella, I. (2016). Adaptive Models and Heavy Tails. *Bank of England Working Paper No. 577*.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–265.

Durrett, R. (2004). *Probability: theory and examples*. Duxbury Press.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007.

Engle, R. F. (2002). Dynamic Conditional Correlation. *Journal of Business & Economic Statistics*, 20(3):339–350.

Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, 55(2):251–276.

Engle, R. F. and Lee, G. G. J. (1999). A long-run and short-run component model of stock return volatility. In *Cointegration, causality and forecasting: A festschrift in honor of Clive W. J. Granger*. New York: Oxford University Press.

Engle, R. F. and Manganelli, S. (2004). Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381.

Engle, R. F. and Russell, J. R. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica*, 66(5):1127–1162.

Francq, C. and Zakoian, J. M. (2004). Maximum Likelihood Estimation of Pure GARCH and ARMA-GARCH Processes. *Bernoulli*, 10(4):605–637.

Francq, C. and Zakoïan, J.-M. (2006). Mixing properties of a general class of GARCH(1,1) models without moment assumptions on the observed process. *Econometric Theory*, 22(5):815–834.

Freeland, R. and McCabe, B. (2004). Forecasting discrete valued low count time series. *International Journal of Forecasting*, 20(3):427–434.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance*, 48(5):1779–1801.

Granger, C. and Andersen, A. (1978). On the invertibility of time series models. *Stochastic Processes and their Applications*, 8(1):87–92.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

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

Harvey, A. and Luati, A. (2014). Filtering With Heavy Tails. *Journal of the American Statistical Association*, 109(507):1112–1122.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *Annals of Statistics*, 24(4):1433–1854.

Ito, R. (2016). Asymptotic Theory for Beta-t-GARCH. *Cambridge Working Papers in Economics CWPE1607*.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Reviews*, 106(4):620–630.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.

Jazi, M. A., Jones, G., and Lai, C.-D. (2012). First-order integer valued AR processes with zero inflated Poisson innovations. *Journal of Time Series Analysis*, 33(6):954–963.

Jensen, S. T. and Rahbek, A. (2004). Asymptotic Inference for Nonstationary GARCH. *Econometric Theory*, 20(6):1203–1226.

Jin-Guan, D. and Yuan, L. (1991). The Integer-valued Autoregressive (INAR (p)) Model. *Journal of Time Series Analysis*, 12(2):129–142.

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

Krengel, U. (1985). *Ergodic theorems*. de Gruyter, Berlin.

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

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lee, S. and Hansen, B. (1994). Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory*, 10(1):29–52.

Lumsdaine, R. L. (1996). Consistency and Asymptotic Normality of the Quasi-Maximum Likelihood Estimator in IGARCH(1,1) and Covariance Stationary GARCH(1,1) Models. *Econometrica*, 64(3):575–596.

Maasoumi, E. (1986). The Measurement and Decomposition of Multidimensional Inequality. *Econometrica*, 54(4):991–997.

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability*, 20(4):822–835.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns : A New Approach. *Econometrica*, 59(2):347–370.

Newey, W. and West, K. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–08.

Oh, D. H. and Patton, A. J. (2016). Time-Varying Systemic Risk: Evidence from a Dynamic Copula Model of CDS Spreads. *Journal of Business & Economic Statistics*, Forthcoming.

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

Pedeli, X. and Karlis, D. (2011). A bivariate INAR (1) process with application. *Statistical modelling*, 11(4):325–349.

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika*, 14(1):249–272.

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

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

Russell, J. R. (2001). Econometric modeling of multivariate irregularly-spaced high-frequency data. *Graduate School of Business, University of Chicago*.

Salvatierra, I. D. L. and Patton, A. J. (2015). Dynamic copula models and high frequency data. *Journal of Empirical Finance*, 30:120–135.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Sorokin, A. (2011). Non-invertibility in some heteroscedastic models. *Arvix preprint 1104.3318*.

Steutel, F. and Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 7(5):893–899.

Stock and Watson (2007). Why Has U.S. Inflation Become Harder to Forecast? *Journal of Money, Credit and Banking*, 39(1):3–33.

Straumann, D. (2005). Estimation in Conditionally Heteroschedastic Time Series Models. *Springer, New York*, 181.

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

Taylor, S. J. (1986). *Modelling Financial Time Series*. Wiley, New York.

Tong, H. (1978). *On a threshold model*. Sijthoff & Noordhoff.

Ullah, A. (1996). Entropy, Divergence and Distance Measures with Econometric Applications. *Journal of Statistical Planning and Inference*, 49(1):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.

Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.

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

Wintenberger, O. (2013). Continuous Invertibility and Stable QML Estimation of the EGARCH(1,1) Model. *Scandinavian Journal of Statistics*, 40(4):846–867.

Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist and Wiksel.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226:267–298.

Zheng, H. and Basawa, I. V. (2008). First-order observation-driven integer-valued autoregressive processes. *Statistics & Probability Letters*, 78(1):1–9.

Zheng, H., Basawa, I. V., and Datta, S. (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference*, 137(1):212–229.

Paolo Gorgi

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: gorgi@stat.unipd.it

Current Position

Since January 2014; (expected completion: February 2017)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: On observation-driven time series modeling
Supervisor: Prof. Luisa Bisaglia and Prof. Siem Jan Koopman
Co-supervisors: Prof. Francisco Blasques.

Research interests

- Statistical inference for dynamic models
- Stochastic processes
- Score-driven models
- Time series analysis

Education

September 2011 – September 2013
Master degree (*laurea specialistica/magistrale***) in Statistical Sciences** .
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "Estimation and forecasting INAR(1) models with Binomial and Negative Binomial error distributions"
Supervisor: Prof. Luisa Bisaglia
Final mark: 110 *cum laude*

September 2008 – July 2011
Bachelor degree (*laurea triennale***) in Statistics Economics and Finance**.
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "From CAPM to Conditional CAPM: an empirical study on Italian equities"
Supervisor: Prof. Massimiliano Caporin
Final mark: 110 *cum laude*

Visiting periods

February 2015 – November 2016
VU University,
Amsterdam, The Netherlands.
Supervisor: Prof. Siem Jan Koopman

Computer skills

- R (advanced)
- Stata (Intermediate)

Language skills

Italian: native; English: fluent

Publications

Working papers
Blasques, F., Gorgi, P., Koopman, S. J. and Wintenberger, O. (2016). Feasible Invertibility Conditions and Maximum Likelihood Estimation for Observation-Driven Models. *Tinbergen Institute Discussion Paper.* TI 2016-082/III.

Gorgi, P. (2016). Integer-valued autoregressive models with survival probability driven by a stochastic recurrence equation. *Arvix preprint.* arXiv:1609.01910.

Blasques, F., Gorgi, P., Koopman, S. J. and Wintenberger, O. (2015). A Note on Continuous Invertibility and Stable QML Estimation of the EGARCH(1,1) Model. *Tinbergen Institute Discussion Paper.* TI 2015-131/III.

Conference presentations

Gorgi, P. (2016). Integer-valued autoregressive models with survival probability driven by a stochastic recurrence equation (oral presentation), *COMPSTAT 2016*, Oviedo, Spain, 23-26 August 2016.

Blasques, F., Gorgi, P. and Koopman, S. J. (2016). Accelerating Score-Driven Models: Optimality, Estimation and Forecasting (poster presentation), *NESG 2016*, Leuven, Belgium, 17-18 June 2016.

Teaching experience

September 2016 – November 2016
International Business Mathematics
Bachelor Degree in International Business Administration
VU University, Amsterdam
Instructor: Prof. Reinout Heijungs

References

Prof. Siem Jan Koopman
VU University, Amsterdam, The Netherlands
Phone: +31 205986019
e-mail: s.j.koopman@vu.nl

Prof. Francisco Blasques
VU University, Amsterdam, The Netherlands
Phone: +31 205985621
e-mail: f.blasques@vu.nl

Prof. Luisa Bisaglia
University of Padova, Italy
Phone: +39 0498274180
e-mail: bisaglia@stat.unipd.it