Statistical Inference for MCARMA Processes

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Dipl.-Math. Sebastian Florian Werner Kimmig
aus Oberkirch

Tag der mündlichen Prüfung: 22. Juni 2016

Referent: Prof. Dr. Vicky Fasen–Hartmann
Korreferent: Prof. Dr. Robert Stelzer

Abstract

Multivariate continuous–time ARMA(p,q) (MCARMA(p,q)) processes are the continuous–time analog of the well-known vector ARMA(p,q) processes. They have attracted interest over the last years. This thesis contributes to the field of statistical inference of MCARMA processes in two ways.

In the first part, we study information criteria, which provide a method to select a suitably MCARMA process as a model for given data. Their background is that methods to estimate the parameters of an MCARMA process require an identifiable parametrization, such as the Echelon form with a fixed Kronecker index, which is in the one–dimensional case the degree p of the autoregressive polynomial. Thus, the Kronecker index has to be known in advance before parameter estimation can be done. When this is not the case information criteria can be used to estimate the Kronecker index and the degrees (p,q), respectively. We investigate information criteria for MCARMA processes based on quasi maximum likelihood estimation. Therefore, we first derive the asymptotic properties of quasi maximum likelihood estimators for MCARMA processes in a misspecified parameter space. Then, we present necessary and sufficient conditions for information criteria to be strongly and weakly consistent, respectively. In particular, we study the well-known Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as special cases.

The second part of the thesis is concerned with robust estimation of the parameters of MCARMA processes. The estimators in the present literature do not work when the data contains outliers. Therefore, robust estimators of the CARMA parameters, which are able to deal with different types of outliers in the data, are necessary. We first extend the class of M–estimators to the MCARMA case. Although these estimators provide a class of strongly consistent and asymptotically normally distributed parameter estimators, they, too, are not robust in the MCARMA setup. We then restrain ourselves to the special case of univariate CARMA processes and use an indirect estimation procedure similar to the one proposed by de Luna and Genton [2001] for ARMA processes. This is motivated by the fact that generalized M-estimators are robust estimators for AR processes, but in general not for ARMA processes.

For the indirect estimator we first approximate the discretely observed CARMA(p,q) process by an auxiliary AR(r) representation, $r \geq 2p-1$. The parameters of this AR(r) process are estimated by a generalized M–estimator. Due to identifiability, there exists a unique, injective map linking the parameters of the AR(r) process to the parameters of the underlying CARMA(p,q) process. Unfortunately, an analytic representation of this map does not exist. To overcome this we have to estimate this map by an additional simulation study. Coupling both estimators gives a robust estimator for the CARMA process. We present the asymptotic behavior as well as the robustness properties of this estimator and develop model selection criteria based on it, too. In both parts, the results are illustrated by a simulation study.

Acknowledgments

First and foremost, I wholeheartedly thank my PhD advisor Prof. Dr. Vicky Fasen–Hartmann for providing me with the opportunity to carry out my PhD studies over the past four years and her guidance during this time. Without the countless discussions we had, the helpful remarks, constructive criticism, reliance, encouragement and patience provided by her, completing this work would not have been remotely possible. I also thank her for the chance of working as a scientific employee, enabling me to learn a lot by and about teaching mathematics to students. I thank Prof. Dr. Robert Stelzer for agreeing to act as referee on this thesis.

Thanks are also due to all my current and former colleagues at the Institute of Stochastics at the Faculty of Mathematics at KIT. I always felt that the working atmosphere at "our" institute was truly pleasant. Specifically, I would like to thank my former office mates Dr. Franziska Häfner, Anton Popp, Dr. Viola Riess, Markus Scholz and Jan Weis for all the laughs and anecdotes we shared over our time in office 5A-10.

I thank all the people I have the joy of calling my friends. Namely, I want to mention Daniel Beese, Andreas Harter, Dominik Kohler, Patrick Müller, Manuel Roth, Till Schulte–Rebbelmund and Stefan Urschel. Not only have I known most of these guys for more than 10 years, but also life certainly would have been less fun without all the Tuesday evenings (sometimes also occurring on Monday, Wednesday or Thursday) we shared over the last 5 years (and counting).

I thank my family, in particular my sisters Sandra Maier and Valeska Kalpers and especially my parents Ursula and Werner Kimmig, for always being there and their loving support, not only during my PhD studies but during my entire life.

Last, but certainly not least, I thank my girlfriend Anne Kathrin Siebert for her unconditional love through all the highs and lows of life and simply everything else.

Contents

CHAPTER 1

Introduction

Historical overview and motivation

Ever since the 1971 first edition of the book by Box et al. [2015], autoregressive moving average (ARMA) processes and their vector–valued (VARMA) counterparts have been among the most popular time series models in both applications and theory (see e.g. Brockwell and Davis [1991] and Hannan and Deistler [2012]). However, they are inherently a class of discrete–time models and nowadays there is a growing interest in stochastic models with a continuous time parameter. Two reasons for this interest are, on the one hand, that many phenomena, which one typically models by time series, evolve continuously in time. Possible examples come from physics, for example the temperature over time at a certain location, but also from economics, where the price of a stock can be seen as a continuous–time process since nowadays price data is available at very high frequency. On the other hand, continuous–time models often lead to a rich and satisfying mathematical theory, for example again in the field of finance, where the modern theory is very much based on continuous–time models, the most famous example probably being the option pricing model by Black and Scholes.

If one wishes to study time series for which the time parameter is continuous, using a continuous–time analog of the discrete–time VARMA models might be attractive. This analog are the multivariate continuous–time autoregressive moving average (MCARMA) processes, which are the core object of attention in this thesis. One-dimensional Gaussian CARMA processes were already investigated by Doob [1944]

and Lévy-driven CARMA processes were propagated at the beginning of this century by Peter Brockwell, see Brockwell [2014] for an overview. This generalization allows for a multitude of different marginal distributions of the CARMA process by varying the driving Lévy process. The formal extension to the multivariate setup was recently given in Marquardt and Stelzer [2007]. This generalization is important because it allows to consider the simultaneous behavior of multiple time series and their interdependence using just one collective model. In the past few years, this class of models has found application in many fields, including but not limited to, finance (Barndorff-Nielsen and Shephard [2001], Todorov [2009], Andresen et al. [2014], Benth et al. [2014]), econometrics (Bergstrom [1990]) or signal processing (Larsson and Söderström [2002], Larsson et al. [2006], Garnier and Wang [2008]). (M)CARMA processes are able to capture phenomena like jumps in the sample paths and heavy tails (see e. g. Todorov and Tauchen [2006] and Benth and Šaltytė-Benth [2009]) and are also suitable for the modeling of unequally spaced (Jones [1981], Jones [1985]) and high–frequency data, which is a very ongoing topic. They are also important as building block of more complicated models as introduced for example in Brockwell and Marquardt [2005], Haug and Stelzer [2011] and Barndorff-Nielsen and Stelzer [2011], where again applications in finance play a central role.

From the mathematical perspective, an $\mathbb{R}^s$-valued Lévy process $(L(t))_{t\geq 0}$ is a stochastic process in $\mathbb{R}^s$ with independent and stationary increments, $L(0) = 0_s$ $\mathbb{P}$-a.s. and càdlàg (continue à droite, limite à gauche) sample paths. Special cases of Lévy processes are Brownian motions and (compound) Poisson processes. Further information on Lévy processes can be found in Applebaum [2009], Bertoin [1998], and Sato [1999], for example. The fundamental idea is now that that for a two-sided $\mathbb{R}^s$-valued Lévy process $L = (L(t))_{t\in\mathbb{R}}$, i.e. $L(t) = L(t)\mathbb{1}_{\{t\geq 0\}} - \widetilde{L}(t-)\mathbb{1}_{\{t<0\}}$ where $(\widetilde{L}(t))_{t\geq 0}$ is an independent copy of the Lévy process $(L(t))_{t\geq 0}$, and positive integers $p > q$, a d-dimensional MCARMA(p, q) process is the solution to the stochastic differential equation

$$P(D)Y(t) = Q(D)DL(t) \quad \text{for } t \in \mathbb{R}, \tag{1.1}$$

where $D = \frac{\partial}{\partial t}$ is the differential operator,

$$P(z) := I_{d\times d}z^p + A_1 z^{p-1} + \ldots + A_{p-1}z + A_p,\ z \in \mathbb{C},$$

with $A_1, \ldots, A_p \in \mathbb{R}^{d\times d}$ being the autoregressive (AR) polynomial and

$$Q(z) := B_0 z^q + B_1 z^{q-1} + \ldots + B_{q-1} z + B_q,\ z \in \mathbb{C},$$

with $B_0, \ldots, B_q \in \mathbb{R}^{d\times s}$ being the moving average (MA) polynomial. Of course, Lévy processes are in general not differentiable, so that this is a purely formal definition. To make sense of it, one uses an equivalent definition in terms of a continuous–time state space model, which is of the form

$$\mathrm{d}X(t) = AX(t)\mathrm{d}t + B\mathrm{d}L(t) \ \text{ and } \ Y(t) = CX(t) \quad \text{for } t \in \mathbb{R},$$

where $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times s}$ and $C \in \mathbb{R}^{d\times N}$ are suitably defined matrices, since Schlemm and Stelzer [2011] have shown that the class of MCARMA processes and the class of continuous–time state space models are equivalent

For practical purposes, e.g. when one is interested in fitting a model to data, statistical inference of MCARMA processes is a topic of great importance. There are a few papers concerned with this, e.g. Brockwell and Schlemm [2013], Fasen [2014], Fasen [2016], Schlemm and Stelzer [2011] and Schlemm and Stelzer [2012]. In particular, Schlemm and Stelzer [2012] derive the asymptotic behavior of the quasi maximum likelihood estimator (QMLE) under the assumption that the underlying parameter space Θ with $N(\Theta)$ parameters contains the true data–generating parameter and satisfies some identifiability assumptions. These are typical assumptions for estimation procedures. For a one–dimensional CARMA process we only obtain identifiability when the degree p of the autoregressive polynomial is fixed in the parameter space (cf. Brockwell et al. [2011]); in the multivariate setup an identifiable parametrization is, for example, provided by the Echelon form. When this form is used, the Kronecker index, which specifies the order of the coefficients of the multivariate autoregressive polynomial, has to be fixed. If we know the Kronecker index, we know the degree p of the autoregressive polynomial as well. But if we observe data, how do we know what is the true Kronecker index of the data, so that we do the parameter estimation in a suitable parameter space Θ? Put differently, how can one arrive at a suitable choice of the AR degree p and the MA degree q for the MCARMA process, since statistical procedures typically assume these parameters as known? This problem is known as the problem of model selection or identification and has been studied extensively in the literature, including, but not limited to, the framework of discrete–time time series and especially VARMA processes.

One popular approach to solving the problem of model selection is by using information or, synonymously, model selection criteria (cf. Burnham and Anderson

[2002], Claeskens and Hjort [2008], Konishi and Kitagawa [2008]). The most prominent model selection criteria are the Akaike Information Criterion (AIC) introduced in Akaike [1973] by Akaike, the Schwarz Information Criterion (SIC), also known as BIC (Bayesian Information Criterion), going back to Schwarz [1978], and the Hannan–Quinn criterion in Hannan and Quinn [1979]. The AIC approximates the Kullback–Leibler discrepancy of a candidate model and the true model, which is a measure for the information that is lost when the candidate model is used instead of the true model. The deciding idea is then to minimize the approximation of the discrepancy to find the most well–fitting model. The BIC approximates the Bayesian a posteriori distribution of different candidate models and aims to maximize this probability to find the best model. The Hannan–Quinn criterion is based on the AIC of Akaike but with a different penalty term to obtain a strongly consistent information criterion, contrary to the AIC which is in general not consistent. Information criteria for multivariate ARMAX processes, which constitute an extension of ARMA processes that includes exogenous variables, and their statistical inference are well–studied in the monograph Hannan and Deistler [2012]; see also Brockwell and Davis [1991] for an overview of model selection criteria for ARMA processes. An extension of the AIC to multivariate weak ARMA processes is given in Boubacar Maïnassara [2012]. There exist only a few papers investigating information criteria independent of the underlying model, e.g. Sin and White [1996] present very general likelihood–based information criteria and their properties, and Cavanaugh and Neath [1999] derive the BIC. All of these information criteria have in common that they are likelihood–based and choose as candidate model the model for which the information criterion attains the lowest value.

So far, to the best of our knowledge, no attempts to study information criteria in the framework of MCARMA processes have been made. This is the motivation for the first part of the thesis, in which this problem is approached. Extending the existing results in the literature to the MCARMA framework, we define a general class of information criteria, which are based on the pseudo–Gaussian log–likelihood function and of the form

$$\mathrm{IC}_n(\Theta) := \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + N(\Theta)\frac{C(n)}{n}.$$

In our setup $Y^n = (Y(h), \ldots, Y(hn))$ is a sample of length n from an MCARMA process observed at discrete, equidistant time points, $\widehat{\mathcal{L}}$ is the properly normalized quasi log-likelihood function, $\widehat{\vartheta}^n$ is the QMLE and $C(n)$ is a penalty term. We choose the parameter space as the most suitable for which the information criterion is lowest. This means that for two parameter spaces Θ_1, Θ_2 we say that Θ_1 fits

better than Θ_2 to the data if we have $IC_n(\Theta_1) < IC_n(\Theta_2)$. A strongly consistent information criterion chooses the correct space asymptotically with probability 1, and for a weakly consistent information criterion the convergence to the true space holds in probability. The sequence $C(n)$ can be interpreted as a penalty term for the inclusion of additional parameters into the model. Without the penalty term, the criterion would always choose the model with more parameters if we compare two nested parameter spaces. However, this is not feasible, since the inclusion of too many parameters ultimately leads to an interpolation of the data, such that the model would not provide information about the process generating the data anymore. The employment of an information criterion can therefore be seen as seeking a trade–off between accuracy and complexity. It can also be interpreted as implementation of the principle of parsimony in model building, cf. [Box et al. 2015, Subsection 1.3.1]. We will study the consistency properties of this family of information criteria and show how the derivation of the well–known AIC and BIC naturally leads to special members of this family. The theoretical results are then complemented by a simulation study, which illustrates the theoretical results in practice.

The second topic treated in this thesis is robust estimation of MCARMA processes. The concept of robustness has been advocated in statistics for over 50 years, starting with the works of Tukey [1960], Huber [1964] and Hampel [1971]. It is motivated by the fact that mathematical results on the performance of statistical procedures, e.g. estimation of the parameters in a parametric model, often strongly depend on the fact that the underlying model assumptions are exactly fulfilled. However, in data, one often is confronted with the phenomenon that a majority of the data does satisfy suitable assumptions while there are also some data points, so–called outliers, which are, in some sense, "very different" from the bulk of the data and violate the assumptions. In the context of parameter estimation, it is known for very wide classes of models that only a few of these outliers suffice to impact the performance of classical estimators, such as the QMLE or the least squares estimator, greatly. For example, when estimating the parameter of a stationary AR(1) process, the size of a single observation going to infinity can cause the estimate to either converge to 0, to 1 or to -1 (cf. [Maronna et al. 2006, Subsection 8.1.3]). This is problematic because if the parameter is equal to 1 or -1, the AR(1) process is no longer stationary and if the parameter is 0, the process is indistinguishable from its driving noise. For this reason, robust estimators have been developed, which are able to deal with the presence of outliers, e.g. abnormally large observations, in the data and avoid these problems. For an overview that treats mainly the cases of i.i.d. data

and linear regression models, see the monographs Huber and Ronchetti [2009] and Hampel et al. [2005].

Generalizing the results on i.i.d. data, the topic of robust estimation in the framework of time series has been a very active one in research since the end of the 20th century. The treatment of outliers in time series comes with the additional difficulty that the temporal structure of the outliers has to be taken into account, which is not relevant in the i.i.d. case. For this reason, estimators that are robust for independent data may fail to be so in the time series context, which is for example the case for the M–estimators introduced by Huber [1964]. Moreover, the definition of an outlier and the way of modeling the presence of outliers used in the i.i.d. case typically do not make much sense in the time series setup. Therefore, different models and methods are needed. In the case of pure autoregressive processes, the class of generalized M–estimators (GM estimators) (first appearing in Hampel [1975], Mallows [1975], Kleiner and Martin [1979], Krasker and Welsch [1982]) was used successfully, since for this class of processes these estimators combine two desirable properties: they allow for the development of an asymptotic theory, as shown by Bustos [1982], and are robust (cf. Künsch [1984]). For the more general ARMA processes, the GM estimators again fail to be robust due to the structure of the innovations sequence of these processes. While there are approaches that allow for robust estimation of ARMA processes, for the longest time most of them suffered from the problem that an asymptotic theory was not readily available. For an overview, see Martin and Yohai [1985] or [Maronna et al. 2006, Chapter 8]. Recently, this problem has been overcome, e.g. by the methods of Muler et al. [2009] and de Luna and Genton [2001]. The idea in the latter paper is to make use of a simulation–based indirect estimation as advocated by Smith [1993], Gouriéroux et al. [1993] and Gouriéroux and Monfort [1997]. The core idea is to not estimate the parameter of interest directly, but to take a detour and use a, in some sense, "more readily handled model", which is called the auxiliary model. In conjunction with simulated data, one can then construct an estimator for the parameter of interest in terms of estimators of the auxiliary parameters. By choosing the auxiliary model as a pure AR process and estimating its parameters with a GM estimator, de Luna and Genton [2001] were able to construct a robust estimator for the parameters of an ARMA process. For one–dimensional CARMA processes, we will proceed analogously and eventually obtain a robust estimator with a tractable asymptotic distribution.

Outline of the thesis

The thesis is structured as follows. In Chapter 2, the basic framework for the rest of the thesis is established. We start by recalling the formal definitions of Lévy processes and multivariate CARMA processes in Section 2.1. In Section 2.2, QML estimation of the parameters of an MCARMA process from discrete–time observations in the spirit of Schlemm and Stelzer [2012] is reviewed in detail, since it is fundamental for much of what follows. In Subsection 2.2.3, we especially extend the results on the QMLE obtained by Schlemm and Stelzer [2012] to the case of misspecified parameter spaces. This is necessary, because in the subsequent study of information criteria, it will be required to know the behavior of the QMLE in this kind of parameter space.

Consistency of information criteria is then the topic of Chapter 3. We first describe the structure of the parametrizations we consider with respect to model selection in Section 3.1. Afterwards, we prove a law of the iterated logarithm for the pseudo–Gaussian log likelihood function in Section 3.2. This will be the most important tool for eventually showing strong consistency of information criteria. The main results of this chapter are contained in Section 3.3, in which we first define our family of information criteria as well as the notion of consistency and then characterize the consistency of the criteria in terms of the penalty function $C(n)$.

Section 3.4 considers the derivation of the AIC in the framework of MCARMA processes. By making use of Akaike's original idea of approximating the Kullback–Leibler discrepancy of a given parametric model and the true model, we obtain particular members of our general family of information criteria as special cases in Subsection 3.4.1. In Subsection 3.4.2 we study the consistency properties of these special criteria. In Subsection 3.4.3 we follow the ideas of Boubacar Maïnassara [2012] and approximate the Kullback–Leibler discrepancy in a different way, leading to a different form of the AIC which is similar to the corrected AIC of Hurvich and Tsai [1993]. Subsection 3.4.4 serves to study the consistency properties of this variant of the AIC. In Subsection 3.4.5, we take yet another route and introduce a version of the AIC which is based on a bootstrap procedure for discrete–time state space models, analogous to Cavanaugh and Shumway [1997]. In Section 3.5, we treat the BIC and elaborate on the approximation of the Bayesian a posteriori probability of parameter spaces. This approach again leads us naturally to a special case of the general information criteria considered in Section 3.3. Section 3.6 closes the chapter with an extensive simulation study in which we apply the various criteria for bivariate MCARMA processes. We consider different true Kronecker indices, different driving Lévy processes, different candidate spaces and different sample sizes, examining in detail the performance of the different criteria in each setup and comparing the

results to the theory.

Chapter 4 constitutes the second part of the thesis in which robust estimation of MCARMA processes is treated. As a first step, in Section 4.1, we introduce the general replacement model, which serves to model the occurrence of outliers in discrete–time observations of an MCARMA process. Afterwards, in Section 4.2, we treat the class of M–estimators, which generalize the QMLE by replacing the pseudo log–likelihood function by a different, general loss function. We will then show that this class provides us with strongly consistent and asymptotically normally distributed estimators when there are no outliers in the data. However, just as for ARMA processes, this class of estimators fails to be robust when outliers are present. We therefore move on and, restricting us to one–dimensional CARMA processes, study in Section 4.3 the indirect estimator, which does achieve the desired robustness.

In its treatment, we first start by introducing the auxiliary AR(r) representation with correlated noise of a discretely sampled CARMA process by means of a set of Yule–Walker equations in Subsection 4.3.1. Although this representation does not enable us to carry out inference for the parameters of the data–generating CARMA process directly, it is a fundamental building block of the indirect estimator. In Subsection 4.3.2 we define the indirect estimator and derive its asymptotic behavior in the absence of outliers. We make use of the fact that there is an injective correspondence between the parameters of a CARMA(p,q) process and the parameters of the auxiliary AR(r) process defined in Subsection 4.3.1 if $r \geq 2p-1$. Then, we calculate two estimates of the parameters of the auxiliary AR(r) process: one is calculated from the actual, observed data and one is calculated from simulated data. By minimizing a suitable distance between these two estimates, we then obtain an estimator for the parameters of the data–generating CARMA process. We show that this estimator is strongly consistent and asymptotically normally distributed in the absence of outliers if both of the estimators that are applied to the AR(r) representation also possess these properties. Since the auxiliary AR(r) process is driven by correlated noise, it is not immediately clear how to obtain such estimators. In Subsection 4.3.3, we first introduce the class of GM estimators and study their asymptotic behavior. We will obtain that they are strongly consistent and asymptotically normal under suitable assumptions. Moreover, in this section we also treat the well known least squares and pseudo–Gaussian maximum likelihood estimator and show that also for an AR(r) process with correlated noise, they are strongly consistent and asymptotically normally distributed. We can therefore use all three of these estimators to construct the indirect estimator.

In Subsection 4.3.4, we study the behavior of the indirect estimator when outliers are present in the data. To this end, we select a GM estimator as the estimator that is applied to the contaminated data, because it is well known that those estimators are robust when applied to pure autoregressive processes. For the estimator in the simulation part the use of the LS or QMLE estimator is most convenient. In particular, we study three measures of robustness: qualitative robustness and resistance in Subsubsection 4.3.4.1, the breakdown point in Subsubsection 4.3.4.2 and the influence functional in Subsubsection 4.3.4.3. The quintessence is that the robustness properties of GM estimators for autoregressive processes are preserved by the indirect estimator and thus also hold for the estimation of the CARMA parameters.

In Section 4.4, we come back to the topic of model selection, albeit this time in the context of outlier–afflicted data. We define information criteria which are the analog of those from Section 3.3, but based on the indirect estimator and not the QMLE. Similar consistency assertions as before can then be derived. In Section 4.5, we conduct various simulations to assess the practical performance of the indirect estimator. For CARMA processes of different orders p and q and different driving Lévy processes, we employ a variety of contamination configurations for the parameter estimation. A short study of the information criteria based on the indirect estimator is also included.

Chapter 5 concludes the thesis, mentioning some open problems and directions for future research. The appendix contains some technical results and proofs, which were excluded from the main text.

CHAPTER 2

FUNDAMENTALS

In this chapter, the groundwork for the rest of the thesis will be laid. We will introduce terms and concepts used extensively throughout the entirety of the work. In particular, we will first define multivariate CARMA processes and an equivalent model class, the continuous–time state space models (CSSMs), and then treat quasi maximum likelihood estimation in this context.

2.1. MULTIVARIATE CARMA PROCESSES AND CONTINUOUS–TIME STATE SPACE MODELS

We start by defining multivariate CARMA processes. To this end, we first introduce the class of Lévy processes, which will be the source of randomness in the process. They can be seen as the analogue to i.i.d. white noise typically used in discrete–time time series models. We will keep the introduction brief and not discuss these processes in great detail here, further information can be found in Applebaum [2009], Sato [1999] or Bertoin [1998], for example.

Definition 2.1. *An $\mathbb{R}^s$-valued **Lévy process** $(L(t))_{t\geq 0}$ is a stochastic process, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with stationary, independent increments, continuous in probability and satisfying $L(0) = 0$ almost surely. We assume without loss of generality that the paths of the Lévy process are right–continuous and have left limits (càdlàg). A two–sided Lévy process $(L(t))_{t\in\mathbb{R}}$ is then defined as $L(t) = L(t)\mathbb{1}_{\{t\geq 0\}} - \widetilde{L}(t-)\mathbb{1}_{\{t<0\}}$, where $(\widetilde{L}(t))_{t\geq 0}$ is an independent copy of the Lévy*

process $(L(t))_{t\geq 0}$.

Special cases of Lévy processes include Brownian Motion and (compound) Poisson processes amongst others. One of the most important results about Lévy processes is the Lévy–Itô–decomposition, which states that every Lévy process is the sum of a Brownian Motion, a compound Poisson process and a square-integrable pure-jump martingale, where the three processes are independent. Equivalently, this can also be stated in terms of the characteristic function of the Lévy process, which satisfies $\mathbb{E}\left[\exp(i\langle u, L(t)\rangle)\right] = \exp(t\psi^L(u))$ for $u \in \mathbb{R}^s$ and $t \in \mathbb{R}_+$, where $\langle\cdot,\cdot\rangle$ is the standard Euclidean scalar product and the so-called characteristic exponent ψ^L is given by

$$\psi^L(u) = i\langle\gamma^L, u\rangle - \frac{1}{2}\langle u, \Sigma^G u\rangle + \int_{\mathbb{R}^s} \left[\exp(i\langle u, x\rangle) - 1 - i\langle u, x\rangle \mathbb{1}_{\{\|x\|\leq 1\}}\right] \nu^L(\mathrm{d}x).$$

This is known as the Lévy–Khintchine formula. It is composed of a drift vector $\gamma^L \in \mathbb{R}^s$, a non-negative definite, symmetric matrix $\Sigma^G \in \mathbb{R}^{s\times s}$ called the Gaussian covariance matrix, and the Lévy measure ν^L, which satisfies the conditions

$$\nu^L(\{0\}) = 0 \quad \text{and} \quad \int_{\mathbb{R}^s} \min(\|x\|^2, 1)\nu^L(\mathrm{d}x) < \infty.$$

Regarding the absolute moments of the Lévy process, it holds for every $k > 0$ that

$$\mathbb{E}[\|L(t)\|^k] < \infty \iff \int_{\|x\|\geq 1} \|x\|^k \nu^L(\mathrm{d}x) < \infty$$

by [Sato 1999, Corollary 25.8]. Furthermore, for the covariance matrix it holds that

$$\Sigma^L = \mathbb{E}[L(1)L(1)^T] = \Sigma^G + \int_{\|x\|\geq 1} xx^T \nu^L(\mathrm{d}x)$$

if it exists ([Sato 1999, Example 25.11]). We will always operate with Lévy processes that have finite second moments, which we formalize in the following assumption:

Assumption L. The Lévy process L has mean zero and finite second moments, i. e. $\gamma^L + \int_{\|x\|\geq 1} x\nu^L(\mathrm{d}x)$ is zero, and the integral $\int_{\|x\|\geq 1} \|x\|^2\nu^L(\mathrm{d}x)$ is finite.

With this in mind, we are now able to define multivariate continuous-time autoregressive moving average (MCARMA) processes. In the univariate case, these have been introduced in Doob [1944] with a Brownian motion as source of randomness. In Brockwell [2001] this was generalized to the case of a driving Lévy process and in Marquardt and Stelzer [2007] the one–dimensional processes were generalized to

multivariate ones. To this end, we first define the autoregressive polynomial (AR) P and the moving average (MA) polynomial Q as follows:

Definition 2.2. *Let $p > q$ be integers. Furthermore, let $A_1, \ldots, A_p \in \mathbb{R}^{d\times d}$, $B_0, \ldots, B_q \in \mathbb{R}^{d\times s}$ and define the matrix polynomial $P(z)$ by*

$$z \mapsto P(z) := I_{d\times d} z^p + A_1 z^{p-1} + \ldots + A_p \tag{2.1}$$

and the matrix polynomial $Q(z)$ by

$$z \mapsto Q(z) := B_0 z^q + B_1 z^{q-1} + \ldots + B_q, \tag{2.2}$$

where $I_{d\times d}$ is the d-dimensional identity matrix.

For a two-sided Lévy process with values in $\mathbb{R}^s$ satisfying $\mathbb{E}\|L(1)\|^2 < \infty$ we would like to interpret a d–dimensional, L-driven **MCARMA(p,q) process** $(Y(t))_{t\in\mathbb{R}}$ with AR polynomial P and MA polynomial Q as the solution to the differential equation

$$P(D)Y(t) = Q(D)DL(t) \text{ where } D := \frac{\partial}{\partial t}, \tag{2.3}$$

analogous to the difference equation satisfied by ARMA processes in discrete time (cf. [Brockwell and Davis 1991, Equation (3.1.5)]). However, defining MCARMA processes via the differential equation (2.3) bears one problem: the paths of a Lévy process are not differentiable in general, which implies that the differential equation cannot be solved. It therefore only acts as a formal motivation. The way out is to interpret the MCARMA process as a special multivariate continuous-time linear state space model:

Definition 2.3. *Let $(L(t))_{t\in\mathbb{R}}$ be an $\mathbb{R}^s$-valued Lévy process with $\mathbb{E}\|L(1)\|^2 < \infty$ and let the polynomials $P(z), Q(z)$ be defined as in (2.1) and (2.2) with $p, q \in \mathbb{N}_0$, $q < p$. Moreover, define*

$$A = \begin{pmatrix} 0_{d\times d} & I_{d\times d} & 0_{d\times d} & \cdots & 0_{d\times d} \\ 0_{d\times d} & 0_{d\times d} & I_{d\times d} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0_{d\times d} \\ 0_{d\times d} & \cdots & \cdots & 0_{d\times d} & I_{d\times d} \\ -A_p & -A_{p-1} & \cdots & \cdots & -A_1 \end{pmatrix} \in \mathbb{R}^{pd\times pd},$$

$C = (I_{d\times d}, 0_{d\times d}, \ldots, 0_{d\times d}) \in \mathbb{R}^{d\times pd}$ *and* $B = (\beta_1^T \cdots \beta_p^T)^T \in \mathbb{R}^{pd\times s}$ *with*

$$\beta_{p-j} := -\mathbb{1}_{\{0,\ldots,q\}}(j) \left(\sum_{i=1}^{p-j-1} A_i \beta_{p-j-i} + B_{q-j} \right), \quad j = 0, \ldots, p-1.$$

Assume that the eigenvalues of A have strictly negative real parts. Then the $\mathbb{R}^d$-valued causal MCARMA(p, q) process $Y = (Y(t))_{t\in\mathbb{R}}$ *is defined by the state space equation*

$$Y(t) = CX(t) \quad \text{for } t \in \mathbb{R}, \tag{2.4}$$

where X is the stationary unique solution to the pd-dimensional stochastic differential equation

$$\mathrm{d}X(t) = AX(t)\mathrm{d}t + B\mathrm{d}L(t). \tag{2.5}$$

This definition slightly extends the one given in Marquardt and Stelzer [2007] because it is allowed for the dimensions of the driving Lévy process and the MCARMA process to be different, however, this changes nothing about the validity of the results in that paper. In particular, MCARMA$(1, 0)$ processes and X in (2.5) are multivariate Ornstein-Uhlenbeck processes. The assumptions about the moments of the Lévy process and the eigenvalues of A are necessary for the solution X of (2.5) to be unique, stationary and causal ([Sato and Yamazato [1983], Theorem 5.1]). Schlemm and Stelzer [2011, Corollary 3.4] shows that the class of continuous-time state space models of the form

$$Y(t) = CX(t) \quad \text{and} \quad \mathrm{d}X(t) = AX(t)\mathrm{d}t + B\mathrm{d}L(t), \tag{2.6}$$

where $A \in \mathbb{R}^{N\times N}$ has only eigenvalues with strictly negative real parts, $B \in \mathbb{R}^{N\times s}$ and $C \in \mathbb{R}^{d\times N}$, and the class of causal MCARMA processes are equivalent if $\mathbb{E}\|L(1)\|^2 < \infty$ and $\mathbb{E}[L(1)] = 0_s$. Note that in contrast to Definition 2.3, it is not required that A, B or C have any kind of special structure. For the purposes of statistical inference, it turned out that working with general state state models is advantageous. Therefore, when we talk about an MCARMA process or a state space model Y, respectively, corresponding to (A, B, C, L), we mean that the MCARMA process Y is defined as in (2.6) and shortly write $Y = \text{MCARMA}(A, B, C, L)$. The covariance matrix of L is always denoted by Σ^L. To finish up this subsection, we introduce the the so–called transfer function of a continuous–time state space model. It will play a key role later when we turn to a MCARMA process observed on a discrete time grid.

Definition 2.4. *Let A, B and C be matrices of appropriate dimensions that define a continuous-time state space model as in Definition* 2.3 *for a Lévy process L. Then the function $H : \mathbb{R} \to \mathbb{R}^{d\times s}(\{z\})$ with $H(z) = C(zI_N - A)^{-1}B$ is called the transfer function of the continuous-time state space model defined by (A, B, C, L). Here $\mathbb{R}^{d\times s}(\{z\})$ denotes the space of $d \times s$ matrices with rational functions as entries.*

2.2. Estimating the parameters of MCARMA processes

In this section, we investigate how the parameters of a MCARMA process can be estimated from equidistant, discrete–time observations. Furthermore, we then investigate the properties and asymptotic behavior of the estimators. To this end, one would typically first introduce a parametric family of MCARMA processes. However, as mentioned at the end of the last section, MCARMA processes are equivalent to continuous–time state space models. Especially in the multivariate case, working with state space models is more convenient. We therefore consider a parameter space Θ and assume that for each $\vartheta \in \Theta$ we are given matrices A_ϑ, B_ϑ, C_ϑ of matching dimensions as well as a Lévy process L_ϑ as in Definition 2.3.
In Schlemm and Stelzer [2012], parameter estimation is done via quasi–maximum likelihood estimation (QMLE) in the case that there exists a true parameter in Θ in the sense that for an observed process Y and some $\vartheta_0 \in \Theta$ it holds that $Y = \text{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$. The authors were then able to show that the QMLE approach leads to a strongly consistent and asymptotically normally distributed estimator. We want to use this QMLE approach to ultimately construct likelihood–based information criteria for the orders of the MCARMA process Y. For this, we will also have to deal with the case in which the above–mentioned assumption fails when we are in a misspecified parameter space. Another argument for inspecting the misspecified case is that if we want to fit MCARMA processes to empirical data, it is questionable if something as a true parameter exists at all, since we only develop a mathematical model that approximates reality. With this understanding, all possible models are then misspecified. The theory of maximum likelihood estimation in possibly misspecified spaces has received a considerable amount of attention in the past, see e. g. White [1982], White [1996] and Sin and White [1996]. This general theory will be very useful for us and allow us to extend the theory of Schlemm and Stelzer [2012] to the more general, possibly misspecified, case with very similar results under very similar assumptions.

2.2.1. Observation and identification

In the rest of the thesis, when we do statistical inference about an MCARMA process, we observe it only on a discrete equidistant time-grid with grid distance $h > 0$. In this section, we first investigate the probabilistic structure of an MCARMA process observed in such a way. Furthermore, we discuss how it can be ensured that an observed process corresponds to exactly one continuous–time state space model when we are given a parametric family of state space models to choose from, reviewing the

most important results from Schlemm and Stelzer [2012].
The first fundamental observation is that the discrete–time stochastic process which results from observing corresponds to a discrete–time state space model as shown in the following proposition:

Proposition 2.5 (Schlemm and Stelzer [2012], Proposition 3.1 and Proposition 3.3)**.** *Assume that $Y = MCARMA(A, B, C, L)$. Then the sampled process $(Y(kh))_{k\in\mathbb{Z}}$ has a discrete–time linear state space representation, i. e. it holds that*

$$Y(kh) = CX(kh) \quad \text{where} \quad X(kh) = \mathrm{e}^{Ah} X((k-1)h) + N_{h,k},\ k \in \mathbb{Z}, \tag{2.7}$$

and $N_{h,k} = \int_{(k-1)h}^{kh} \mathrm{e}^{A(kh-t)} B\mathrm{d}L(t)$ is a sequence of i.i.d. random vectors with covariance matrix

$$\Sigma_h = \int_0^h \exp(Au)B\Sigma^L B^T \exp(A^T u)\mathrm{d}u. \tag{2.8}$$

The spectral density of $(Y(kh))_{k\in\mathbb{Z}}$, denoted by f_Y^h, is defined as the Fourier transform of the autocovariance function, i.e.

$$f_Y^h(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-i\omega h}\,\gamma_Y(h)\mathrm{d}h,\ \omega \in [-\pi, \pi],$$

where $\gamma_Y(h) := \mathrm{Cov}(Y(t+h), Y(t)) = C\,\mathrm{e}^{Ah}\,\Gamma_0 C^T$, $h \geq 0$, with

$$\Gamma_0 := \mathrm{Var}(X(t)) = \int_0^\infty \mathrm{e}^{Au}\,B\Sigma^L B^T\,\mathrm{e}^{A^T u}\,\mathrm{d}u.$$

f_Y^h is given by

$$f_Y^h(\omega) = C\left(\exp(i\omega)I_N - \exp(Ah)\right)^{-1}\Sigma_h\left(\exp(i\omega)I_N - \exp(A^T h)\right)^{-1}C^T,$$

and the resulting $d \times d$ matrix is positive semidefinite.

Proof. See cited references. □

In principle, it could be possible for the process $(Y(kh))_{k\in\mathbb{Z}}$ to be the output process of a discrete–time linear state space model with higher dimension as the one appearing in (2.7) and also for the matrix Σ_h to be singular. In order for parameter estimation to be possible, these effects must be excluded, which is why we next explore conditions that ensure this. For this, we will need the following definitions:

Definition 2.6. *Let H be a $d \times s$ rational matrix function, i. e. a $d \times s$ matrix whose entries are rational functions of the variable $z \in \mathbb{R}$. A matrix triple (A, B, C),*

where $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times s}$ and $C \in \mathbb{R}^{d\times N}$, is called an ***algebraic realization*** *of H of dimension N if $H(z) = C(zI_N - A)^{-1}B$ for every $z \in \mathbb{R}$.*

Algebraic realizations are generally not unique, not even the dimension needs to be unique. This is the reason why we introduce the concept of a minimal realization and the so–called McMillan degree:

Definition 2.7. *Let H be a $d \times s$ rational matrix function. A* ***minimal realization*** *of H is an algebraic realization of H of dimension smaller or equal to the dimension of every other algebraic realization of H. The dimension of a minimal realization of H is the* ***McMillan degree*** *of H.*

This definition alone is not very useful, because it does not state an accessible way of checking if a given algebraic realization is minimal. To state a theorem that provides this, we need the following two definitions:

Definition 2.8. *An algebraic realization (A, B, C) of dimension N is* ***controllable*** *if the controllability matrix $\mathcal{C} = \begin{pmatrix} B & AB & \dots & A^{n-1}B \end{pmatrix} \in \mathbb{R}^{s\times sN}$ has full rank.*

Definition 2.9. *An algebraic realization (A, B, C) of dimension N is* ***observable*** *if the observability matrix $\mathcal{O} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} \in \mathbb{R}^{dN\times N}$ has full rank.*

Theorem 2.10 (Hannan and Deistler [2012], Theorem 2.3.3)**.** *An algebraic realization is minimal if and only if it is both controllable and observable.*

Proof. See cited reference. □

Remark 2.11. *We call a continuous–time state space model controllable, observable or minimal if the corresponding transfer function as defined in Definition 2.4 has these properties.*

We now have all the ingredients necessary to give criteria for the covariance matrix of our sampled process to be regular and for the sampled process to have the same McMillan degree as our MCARMA process:

Proposition 2.12 (Schlemm and Stelzer [2012], Corollary 3.1)**.** *If the triple (A, B, C) is a minimal realization of the transfer function of dimension N of a continuous–time state space model (A, B, C, L), and Σ^L is positive definite, then the $N \times N$ matrix*

$$\mathcal{Z} = \int_0^h \exp(Au)B\Sigma^L B^T \exp(A^T u)\mathrm{d}u$$

has full rank N.

Proof. See cited reference. □

Proposition 2.13 (Schlemm and Stelzer [2012], Proposition 3.4)**.** *Assume that $Y = MCARMA(A, B, C, L)$ with (A, B, C) being a minimal realization of the associated transfer function of McMillan degree N. Then a sufficient condition for the sampled process $(Y(kh))_{k\in\mathbb{Z}}$ to have the same McMillan degree is the Kalman–Bertram criterion*

$$\lambda - \lambda' \neq \frac{2\pi i k}{h} \quad \forall(\lambda, \lambda') \in \sigma(A) \times \sigma(A),\ \forall k \in \mathbb{Z} \setminus \{0\}, \tag{2.9}$$

where $\sigma(A)$ denotes the spectrum of A.

Proof. See cited reference. □

In the next two definitions we explain what exactly we understand under the equality of two MCARMA processes:

Definition 2.14. *Two stochastic processes, irrespective of whether their index sets are continuous or discrete, are L^2-**observationally equivalent** if their spectral densities are the same.*

Definition 2.15. *A family $(Y_\vartheta)_{\vartheta\in\Theta}$ of continuous–time stochastic processes is **identifiable from the spectral density** if, for every $\vartheta_1 \neq \vartheta_2$, the two processes $(Y_{\vartheta_1}(t))_{t\in\mathbb{R}}$ and $(Y_{\vartheta_2}(t))_{t\in\mathbb{R}}$ are not L^2-observationally equivalent. It is h–**identifiable from the spectral density**, $h > 0$, if, for every $\vartheta_1 \neq \vartheta_2$, the two sampled processes $(Y_{\vartheta_1}(kh))_{k\in\mathbb{Z}}$ and $(Y_{\vartheta_2}(kh))_{k\in\mathbb{Z}}$ are not L^2-observationally equivalent.*

This completes the tools we need to have at hand to describe under which conditions it can be ensured that an observed process belongs to exactly one member of the parametric family $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$. The obvious idea would be to make the assumption that the family $(Y_\vartheta)_{\vartheta\in\Theta}$ of output processes is h–identifiable from the spectral density for the observation distance h we chose. However, we shall see in the next section that this can be derived from identifiability assumptions on the family of continuous–time processes and some conditions that exclude so–called aliasing effects.

2.2.2. Canonical parametrizations

Up until now, we have not specified exactly how the matrices A_ϑ, B_ϑ and C_ϑ depend on the parameter $\vartheta \in \Theta$ when we talk about a family of continuous–time state space models. It seems logical to demand that a parametrization should at the very

least provide output processes that are identifiable from the spectral density. A parametrization that achieves this is the so–called Echelon MCARMA parametrization, which shall be presented here and used throughout the rest of the thesis. This section is essentially [Schlemm and Stelzer 2012, Section 4.1], but since the results are fundamental we explicitly repeat them here. The principal idea is to concentrate on the transfer function H of a CSSM as given in Definition 2.4, find a unique parametrization for it and then establish a connection between H, P and Q. We start with a canonical decomposition for rational matrix functions:

Theorem 2.16 (Bernstein [2009], Theorem 4.7.5)**.** *Let H be a $d \times s$ rational matrix function of rank r. There exist matrices $S_1 \in \mathbb{R}^{d\times d}[z]$ and $S_2 \in \mathbb{R}^{s\times s}[z]$ with constant determinant, such that $H = S_1 M S_2$, where*

$$M = \begin{pmatrix} diag(\frac{\epsilon_1}{\psi_1}, \dots, \frac{\epsilon_r}{\psi_r}) & 0_{r,s-r} \\ 0_{d-r,r} & 0_{d-r,s-r} \end{pmatrix} \in \mathbb{R}^{d\times s}(\{z\}).$$

Here $\mathbb{R}^{d\times d}[z]$ is the space of $d\times d$ matrices with polynomials in the variable z as entries. Moreover, $\epsilon_1, \dots, \epsilon_r, \psi_1, \dots, \psi_r \in \mathbb{R}[z]$ are monic polynomials uniquely determined by H satisfying the following conditions:
For each $i = 1, \dots, r$ the polynomials ϵ_i and ψ_i have no common roots and for each $i = 1, \dots, r-1$ the polynomial ϵ_i divides the polynomial ϵ_{i+1} while the polynomial ψ_{i+1} divides the polynomial ψ_i. The triple (S_1, M, S_2) is called the Smith-McMillan decomposition of H.

An important role is played by the degrees of the denominator polynomials ψ_i in the Smith-McMillan decomposition of a rational matrix function H. We call these degrees Kronecker indices and denote them by m_i for $i = 1, \dots, r$ and define the vector $m = (m_1, \dots, m_d) \in \mathbb{N}^d$. Here $m_k = 0$ for $k = r+1, \dots, d$, however in the following we only concentrate on the case where all Kronecker indices are strictly positive.
The Kronecker indices have the important property that $\sum_{i=1}^d m_i = N$, where N is the McMillan degree of H, i. e. the smallest possible dimension of an algebraic realization of H, see Definition 2.7. For $1 \le i, j \le d$ we furthermore define $m_{ij} = \min\{m_i + \mathbb{1}_{\{i>j\}}, m_j\}$. With these terms we can now give a unique, minimal algebraic realization of a transfer function:

Theorem 2.17 (Echelon state space realization, Guidorzi [1975], Section 3)**.** *Let H be a $d \times s$ rational matrix function with Kronecker indices $m = (m_1, \dots, m_d)$ and assume that $m_i > 0$ for $i = 1, \dots, d$. Then a unique, minimal algebraic realization (A, B, C) of H of dimension $N = \sum_{i=1}^d m_i$ is given by the following structure:*

i) The matrix $A = (A_{ij})_{i,j=1,\dots,d} \in \mathbb{R}^{N\times N}$ *is a block matrix with blocks* $A_{ij} \in \mathbb{R}^{m_i\times m_j}$ *given by*

$$A_{ij} = \begin{pmatrix} 0 & \dots & \dots & \dots & \dots & 0 \\ \vdots & & & & & \vdots \\ 0 & \dots & \dots & \dots & \dots & 0 \\ \alpha_{ij,1} & \dots & \alpha_{ij,m_{ij}} & 0 & \dots & 0 \end{pmatrix} + \delta_{i,j} \begin{pmatrix} 0 & & \\ \vdots & I_{m_i-1} & \\ 0 & & \\ 0 & \dots & 0 \end{pmatrix}.$$

ii) $B = (b_{ij}) \in \mathbb{R}^{N\times s}$ *unrestricted.*

iii)

$$C = \begin{pmatrix} 1 & 0 & \dots & 0 & \vdots & 0 & 0 & \dots & 0 & \vdots & \vdots & & & \\ & & & & & & & & & & & & & 0_{d-1,m_d} \\ & & & & \vdots & 1 & 0 & \dots & 0 & \vdots & \vdots & & & \\ & 0_{d-1,m_1} & & & & & & & & & & & & \\ & & & & \vdots & & 0_{d-2,m_2} & & & \vdots & \vdots & 1 & 0 & \dots & 0 \end{pmatrix}$$

However, this form alone does not ensure identifiability, since there is an orthogonal invariance in the spectral factorization, see [Schlemm and Stelzer 2012, Theorem 3.5]). One way to get rid of this problem is to require $H(0) = H_0$ for a non-singular matrix H_0, often the negative of the identity matrix. To see the consequences this has on the parameters $\alpha_{ij,k}$ and b_{ij} and also show how one can obtain the MCARMA polynomials P and Q from A, B and C, we work in terms of left matrix fraction descriptions:

Theorem 2.18 (Echelon MCARMA realization, Guidorzi [1975], Section 3)**.** *Let* H *be a* $d\times s$ *rational matrix function with Kronecker indices* $m = (m_1, \dots, m_d)$*. Assume that* (A, B, C) *is a realization of* H*, parametrized as in Theorem 2.17. Then a unique left matrix fraction description* $P^{-1}Q$ *of* H *is given by* $P = [p_{ij}(z)]$*,* $Q = [q_{ij}(z)]$*, where*

$$p_{ij}(z) = \delta_{i,j} z^{m_i} - \sum_{k=1}^{m_{ij}} \alpha_{ij,k} z^{k-1}, \quad q_{ij}(z) = \sum_{k=1}^{m_i} \kappa_{m_1+\dots+m_{i-1}+k,j} z^{k-1}, \tag{2.10}$$

where $\kappa_{i,j}$ *is the* (i,j)*th entry of the matrix* $K = TB$*, where* $T = (T_{ij})_{i,j=1,\dots,d} \in \mathbb{R}^{N\times N}$

is a block matrix with blocks $T_{ij} \in \mathbb{R}^{m_i \times m_j}$ *given by*

$$T_{ij} = \begin{pmatrix} -\alpha_{ij,2} & \dots & -\alpha_{ij,m_{ij}} & 0 & \dots & 0 \\ \vdots & \iddots & & & & \vdots \\ -\alpha_{ij,m_{ij}} & & & & & \vdots \\ 0 & & & & & \vdots \\ \vdots & & & & & \vdots \\ 0 & \dots & \dots & \dots & \dots & 0 \end{pmatrix} + \delta_{i,j} \begin{pmatrix} 0 & 0 & \dots & \dots & 0 & 1 \\ 0 & 0 & \dots & & 1 & 0 \\ \vdots & \vdots & & \iddots & & \vdots \\ \vdots & & \iddots & & \vdots & \vdots \\ 0 & 1 & & \dots & 0 & 0 \\ 1 & 0 & \dots & \dots & 0 & 0 \end{pmatrix}.$$

Using this form, there are several ways to enforce the normalization $H(0) = H_0$. Since $H(0) = P(0)^{-1}Q(0) = -(\alpha_{ij,1})_{ij}^{-1}(\kappa_{m_1+\ldots+m_{i-1}+1,j})_{ij}$ we can restrict some of the entries of the matrix B to achieve our goal. Note that $|\det K| = 1$ and hence T is invertible, which is why B can be written as $B = T^{-1}K$. If we now replace the $(m_1+\ldots+m_{i-1}+1, j)$th entry of K by the (i,j)th entry of the matrix $-(\alpha_{kl,1})_{kl}H_0$, we have made some of the b_{ij} dependent on the entries of the matrix A and achieved the normalization. If $d = s$ and $H_0 = -I_{d\times d}$, then it suffices to set $\kappa_{m_1+\ldots+m_{i-1}+1,j} = \alpha_{ij,1}$. In the rest of the thesis, this normalization was always used in practical situations, i.e. when simulations were carried out.

As in [Schlemm and Stelzer 2012, Tables 1 and 2], we give examples for the case of $d = s = 2$ and $H(0) = -I_{2\times 2}$. The number of free parameters, $N(\Theta)$, includes three parameters for the covariance matrix Σ^L of the Lévy process.

Table 2.1.: Canonical state space realizations (A, B, C) of rational transfer functions with different Kronecker indices m

m	$N(\Theta)$	A	B	C
$(1,1)$	7	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_3 & \vartheta_4 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_3 & \vartheta_4 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$(1,2)$	10	$\begin{pmatrix} \vartheta_1 & \vartheta_2 & 0 \\ 0 & 0 & 1 \\ \vartheta_3 & \vartheta_4 & \vartheta_5 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_6 & \vartheta_7 \\ \vartheta_3+\vartheta_5\vartheta_6 & \vartheta_4+\vartheta_5\vartheta_7 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
$(2,1)$	11	$\begin{pmatrix} 0 & 1 & 0 \\ \vartheta_1 & \vartheta_2 & \vartheta_3 \\ \vartheta_4 & \vartheta_5 & \vartheta_6 \end{pmatrix}$	$\begin{pmatrix} \vartheta_7 & \vartheta_8 \\ \vartheta_1+\vartheta_2\vartheta_7 & \vartheta_3+\vartheta_2\vartheta_8 \\ \vartheta_4+\vartheta_5\vartheta_7 & \vartheta_6+\vartheta_5\vartheta_8 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
$(2,2)$	15	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ \vartheta_1 & \vartheta_2 & \vartheta_3 & \vartheta_4 \\ 0 & 0 & 0 & 1 \\ \vartheta_5 & \vartheta_6 & \vartheta_7 & \vartheta_8 \end{pmatrix}$	$\begin{pmatrix} \vartheta_9 & \vartheta_{10} \\ \vartheta_1+\vartheta_4\vartheta_{11}+\vartheta_2\vartheta_9 & \vartheta_3+\vartheta_2\vartheta_{10}+\vartheta_4\vartheta_{12} \\ \vartheta_{11} & \vartheta_{12} \\ \vartheta_5+\vartheta_8\vartheta_{11}+\vartheta_6\vartheta_9 & \vartheta_7+\vartheta_6\vartheta_{10}+\vartheta_8\vartheta_{12} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

If we consider a parameter space Θ that is parametrized as just described, then the Kronecker indices m of Θ are closely connected to the AR order of the MCARMA

Table 2.2.: Canonical MCARMA realizations (P, Q) with order (p, q) of rational transfer functions with different Kronecker indices m

m	$N(\Theta)$	$P(z)$	$Q(z)$	(p, q)
$(1,1)$	7	$\begin{pmatrix} z-\vartheta_1 & -\vartheta_2 \\ -\vartheta_3 & z-\vartheta_4 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_3 & \vartheta_4 \end{pmatrix}$	$(1,0)$
$(1,2)$	10	$\begin{pmatrix} z-\vartheta_1 & -\vartheta_2 \\ -\vartheta_3 & z^2-\vartheta_4 z-\vartheta_5 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_6 z+\vartheta_3 & \vartheta_7 z+\vartheta_5 \end{pmatrix}$	$(2,1)$
$(2,1)$	11	$\begin{pmatrix} z^2-\vartheta_1 z-\vartheta_2 & -\vartheta_3 \\ -\vartheta_4 z-\vartheta_5 & z-\vartheta_6 \end{pmatrix}$	$\begin{pmatrix} \vartheta_7 z+\vartheta_2 & \vartheta_8 z+\vartheta_3 \\ \vartheta_5 & \vartheta_6 \end{pmatrix}$	$(2,1)$
$(2,2)$	15	$\begin{pmatrix} z^2-\vartheta_1 z-\vartheta_2 & -\vartheta_3 z-\vartheta_4 \\ -\vartheta_5 z-\vartheta_6 & z^2-\vartheta_7 z-\vartheta_8 \end{pmatrix}$	$\begin{pmatrix} \vartheta_9 z+\vartheta_2 & \vartheta_{10} z+\vartheta_4 \\ \vartheta_{11} z+\vartheta_6 & \vartheta_{12} z+\vartheta_8 \end{pmatrix}$	$(2,1)$

processes with $Y = \text{MCARMA}(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)$ for $\vartheta \in \Theta$, because it holds that $p = \max_{i=1,\dots,d} m_i$ for every such process.

However, for the MA order q things look a bit different. In fact, for the models in Θ it holds that $0 \leq q \leq p-1$, i. e. by fixing Θ and thus m the MA order of the contained models is not uniquely defined.

Later on, the Kronecker indices will be what our order selection criteria estimate. As pointed out, this would not provide information about the MA degree, which is why we will need to refine the parametrizations. This will be explained in detail at the start of the chapter dealing with order selection.

2.2.3. Quasi–maximum likelihood estimation for MCARMA processes

With the knowledge from the previous sections, we can now turn to the QML estimation for MCARMA processes. In the following we assume that our data set is generated by a continuous–time state space model (A, B, C, L), i.e. for $(Y(t))_{t\in\mathbb{R}}$ with $Y = \text{MCARMA}(A, B, C, L)$ the discretely sampled process $(Y(kh))_{k\in\mathbb{Z}}$ is the data–generating process. Moreover, we have a parametric family of MCARMA models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)$ with ϑ in a parameter space $\Theta \subset \mathbb{R}^{N(\Theta)}$, $N(\Theta) \in \mathbb{N}$. The aim is to find $\vartheta_0 \in \Theta$ such that $\text{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$, if such a thing exists. As we have seen in Proposition 2.5, observing a parametric family of MCARMA processes at equidistant points in time induces a parametric family of discrete–time state space models. Therefore, QML estimation of a MCARMA process observed in this way is equivalent to QML estimation of a discrete–time state space model. We now review the most important aspects of the estimation of such a model, heavily relying on Schlemm and Stelzer [2012]. A key role in the QML estimation of this kind of processes is played by the so–called linear innovations. These are defined in

the following way:

Definition 2.19. *The **linear innovations** $\epsilon = (\epsilon_k)_{k\in\mathbb{Z}}$ of the process $(Y(nh))_{n\in\mathbb{Z}}$ are defined by $\epsilon_k = Y(kh) - P_{k-1}Y(kh)$, where P_k denotes the orthogonal projection onto the space $\overline{span}\{Y(mh) : -\infty < m \leq k\}$ and the closure is taken in L^2.*

Note that this definition ensures that the innovations are stationary, uncorrelated and have mean 0.
What is of particular importance is the fact that these innovations can be calculated through the so-called Kalman filter, originally introduced in Kalman [1960] and described in a time series context in [Brockwell and Davis 1991, §12.2]. This is summarized in [Schlemm and Stelzer 2012, Proposition 2.1], who in turn combined [Brockwell and Davis 1991, Proposition 12.2.3] and [Hamilton 1994, Proposition 13.2] to obtain it. We employ the fundamental ideas of the Kalman filter in the following to calculate the so–called pseudo–innovations. To this end, we proceed as follows:
For every $\vartheta \in \Theta$, the steady-state Kalman gain matrices K_ϑ and covariances V_ϑ are computed as functions of the unique positive definite solution Ω_ϑ to the discrete–time Riccati equation

$$\Omega_\vartheta = \mathrm{e}^{A_\vartheta h}\,\Omega_\vartheta\,\mathrm{e}^{A_\vartheta^T h} + \Sigma_{\vartheta,h} - \left(\mathrm{e}^{A_\vartheta h}\,\Omega_\vartheta C_\vartheta^T\right)\left(C_\vartheta\Omega_\vartheta C_\vartheta^T\right)^{-1}\left(\exp^{A_\vartheta h}\,\Omega_\vartheta C_\vartheta^T\right)^T, \tag{2.11}$$

via

$$K_\vartheta = \left(\mathrm{e}^{A_\vartheta h}\,\Omega_\vartheta C_\vartheta^T\right)\left(C_\vartheta\Omega_\vartheta C_\vartheta^T\right)^{-1},\ V_\vartheta = C_\vartheta\Omega_\vartheta C_\vartheta^T. \tag{2.12}$$

Based on this, the **pseudo–innovations** $(\epsilon_{\vartheta,k})_{k\in\mathbb{Z}}$ are defined by

$$\begin{aligned}\widehat{X}_{\vartheta,k} &= (\mathrm{e}^{A_\vartheta h} - K_\vartheta C_\vartheta)\widehat{X}_{\vartheta,k-1} + K_\vartheta Y((k-1)h),\\ \epsilon_{\vartheta,k} &= Y(kh) - C_\vartheta\widehat{X}_{\vartheta,k}\\ &= \left[I_{d\times d} - C_\vartheta\left(I_N - (\mathrm{e}^{A_\vartheta h} - K_\vartheta C_\vartheta)B\right)^{-1}K_\vartheta B\right]Y(kh),\ k\in\mathbb{Z}.\end{aligned} \tag{2.13}$$

where B denotes the backshift operator, i.e. $BY_k = Y_{k-1}$. Note that we call them pseudo–innovations because, in general, they will not coincide with the innovations of the process $(Y_\vartheta(nh))_{n\in\mathbb{Z}}$. If, however, $Y = \mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$, then $\epsilon_{\vartheta_0,k} = \epsilon_k$ and it holds that $\mathbb{E}[\epsilon_{\vartheta_0,k}\epsilon_{\vartheta_0,k}^T] = V_{\vartheta_0}$, which is in general not true for $\vartheta \neq \vartheta_0$.
Suppose now that we have n observations of the process Y at discrete, equidistant times, contained in the sample $Y^n = (Y(h), \ldots, Y(nh))$. In this situation, [Brockwell and Davis 1991, Eq. (11.5.4)] tells us that $\frac{-2}{n}$ times the logarithm of the Gaussian

likelihood of ϑ can be written as

$$\mathcal{L}(\vartheta, Y^n) = \frac{1}{n}\sum_{k=1}^{n}(d\log(2\pi) + \log(\det(V_\vartheta)) + \epsilon_{\vartheta,k}^T V_\vartheta^{-1}\epsilon_{\vartheta,k}) \tag{2.14}$$

$$=: \frac{1}{n}\sum_{k=1}^{n} l_{\vartheta,k} \tag{2.15}$$

The expectation of this random variable is

$$\mathscr{Q}(\vartheta) := \mathbb{E}\left[\mathcal{L}(\vartheta, Y^n)\right]. \tag{2.16}$$

Gaussian likelihood in this situation means that this is the exact likelihood function if the innovations are normally distributed. If this assumption fails to hold $\mathcal{L}$ is not the true likelihood function of the innovations, hence the name quasi maximum likelihood estimation.

Moreover, in practical scenarios it will not even be possible to calculate the pseudo-innovations, as they are defined in terms of the full history of the process $(Y(nh))_{n\in\mathbb{Z}}$ and we only have a finite amount of observations at our disposal. Therefore we need a method to approximate them based on this finite sample. For example, one could initialize the Kalman filter at $k = 1$ by prescribing $\widehat{X}_{\vartheta,1} = \widehat{X}_{\vartheta,\text{initial}}$ and then use the recursion

$$\begin{aligned} \widehat{X}_{\vartheta,k} &= (\mathrm{e}^{A_\vartheta h} - K_\vartheta C_\vartheta)\widehat{X}_{\vartheta,k-1} + K_\vartheta Y((k-1)h),\ k \geq 2, \\ \widehat{\epsilon}_{\vartheta,k} &= Y(kh) - C_\vartheta \widehat{X}_{\vartheta,k},\ k \in \mathbb{N}. \end{aligned} \tag{2.17}$$

$\widehat{X}_{\vartheta,\text{initial}}$ can be sampled from the stationary distribution of X_ϑ if possible or simply set to some deterministic value. We call the $\widehat{\epsilon}_{\vartheta,k}$ obtained in this way the *approximate pseudo–innovations*.

Substituting the approximate pseudo-innovations for their theoretical counterparts in (2.14), we obtain the quantity that will be minimized to obtain the Gaussian QMLE, namely

$$\widehat{\mathcal{L}}(\vartheta, Y^n) = \frac{1}{n}\sum_{k=1}^{n}(d\log(2\pi) + \log(\det(V_\vartheta)) + \widehat{\epsilon}_{\vartheta,k}^T V_\vartheta^{-1}\widehat{\epsilon}_{\vartheta,k}) \tag{2.18}$$

The QMLE based on the sample Y^n is then given by

$$\widehat{\vartheta}^n := \operatorname*{arg\,min}_{\vartheta\in\Theta} \widehat{\mathcal{L}}(\vartheta, Y^n). \tag{2.19}$$

The idea is that $\widehat{\vartheta}^n$ is an estimator for the *pseudo–true parameter*

$$\vartheta^* := \arg\min_{\vartheta\in\Theta} \mathscr{Q}(\vartheta). \tag{2.20}$$

We call this parameter pseudo–true parameter since we do not necessarily assume that Θ contains some ϑ_0 with $Y = \mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$. Also, strictly speaking, argmin is a slight abuse of notation here, as we cannot know if the pseudo-true parameter is unique, and in fact it will turn out this is something that we will have to assume. Since we will always make this assumption later, we have introduced this notation here already. If Θ does contain such a ϑ_0 as just described, we always have $\vartheta^* = \vartheta_0$, i. e. this definition is compatible with that special case.
We now need to address several points: First off, we must ensure that the collection of output processes $(Y_\vartheta)_{\vartheta\in\Theta}$ is h-identifiable from the spectral density in order to prevent the introduction of aliasing effects into our model by the discrete observations.
Secondly, we should consider the implications of using the approximate pseudo–innovations $(\widehat{\epsilon}_{\vartheta,k})_{k\in\mathbb{N}}$ instead of their theoretical counterparts, which also affects the log-likelihood function.
And lastly, we need to clarify some points which arise when we are in a misspecified parameter space: under which conditions does the QMLE converge to the pseudo–true parameter ϑ^*, if at all? Are those conditions any different in the correctly and incorrectly specified case?
Before we answer these questions, we will list all the assumptions we use to develop the asymptotic theory of the QMLE in one place for easy reference:

Assumption B.

- B.1 The parameter space Θ is a compact subset of $\mathbb{R}^{N(\Theta)}$.
- B.2 For each $\vartheta \in \Theta$, it holds that $\mathbb{E}[L_\vartheta] = 0$, $\mathbb{E}[\|L_\vartheta(1)\|^2] < \infty$ and the covariance matrix $\Sigma^L_\vartheta = \mathbb{E}\left[L_\vartheta(1)L^T_\vartheta(1)\right]$ is non-singular.
- B.3 For each $\vartheta \in \Theta$, the eigenvalues of A_ϑ have strictly negative real parts.
- B.4 The functions $\vartheta \mapsto A_\vartheta$, $\vartheta \mapsto B_\vartheta$, $\vartheta \mapsto C_\vartheta$ and $\vartheta \mapsto \Sigma^L_\vartheta$ are three times continuously differentiable. Moreover, for each $\vartheta \in \Theta$, the matrix C_ϑ has full rank.
- B.5 For all $\vartheta \in \Theta$, the triple $(A_\vartheta, B_\vartheta, C_\vartheta)$ is minimal with McMillan degree N.
- B.6 The family $(\mathrm{MCARMA}(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta))_{\vartheta\in\Theta}$ is identifiable from the spectral density.

B.7 For all $\vartheta \in \Theta$, the spectrum of A_ϑ is a subset of $\{z \in \mathbb{C} : -\frac{\pi}{h} < \mathrm{Im} z < \frac{\pi}{h}\}$.

B.8 The pseudo–true parameter ϑ^* is an element of the interior of Θ.

B.9 For the true Lévy process L there exists a $\delta > 0$ such that $\mathbb{E}[\|L(1)\|^{4+\delta}] < \infty$.

B.10 For every $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that

$$\mathscr{Q}(\vartheta^*) \leq \min_{\vartheta \in B_\epsilon(\vartheta^*)^c \cap \Theta} \mathscr{Q}(\vartheta) - \delta(\epsilon),$$

where $B_\epsilon(\vartheta^*)$ is the open ball with center ϑ^* and radius ϵ.

B.11 The Fisher information matrix of the quasi maximum likelihood estimator is non-singular.

Remark 2.20.

a) At the beginning of [Schlemm and Stelzer 2012, Chapter 4.1] it is explained that the Echelon MCARMA realization from Subsection 2.2.2 fulfills the smoothness and identifiability assumptions desired in Assumption B automatically, namely B.4, B.5 and B.6. The rest of the assumptions can be easily imposed by restricting Θ and the driving Lévy process in a suitable way. For this reason we always use the Echelon form as parametrization.

b) Note that imposing these assumptions on a parametric family of continuous–time state space models achieves the following:
Two processes in the family $(MCARMA(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta))_{\vartheta \in \Theta}$ can not possess the same spectral density, since by B.6 this is excluded. Moreover, by B.5 it is also ensured that no process in this family can be described equivalently by a process from another family with different McMillan degree that fulfills the assumptions as well (again in the sense that they possess the same spectral density). Hence, for two parameter spaces Θ and Θ' both satisfying Assumption B with different McMillan degrees they will always contain different processes.
Moreover, Assumption B.2 is the equivalent of Assumption L, Assumption B.3 guarantees the existence and uniqueness of a stationary solution, and Assumption B.7 implies the Kalman–Bertram criterion (Eq. (2.9)).

c) Assumption B.10 is a property called identifiable uniqueness: it makes sure that ϑ^* is the unique minimum of $\mathscr{Q}(\vartheta)$ in Θ (White [1996, p. 28]). In the situation we investigated, i.e. for the Echelon form, we were not able to find a way to guarantee the uniqueness to hold besides simply assuming it. In light of

(2.20) *this does not seem to be very restrictive, however. In conjunction with what was explained in the previous paragraph the assumption just means that in every parametric family we can find one continuous–time state space model (or MCARMA process) which is the best approximation to the true one, i. e. the notation in* (2.20) *is justified.*

d) In the correctly specified case, i. e. when the space Θ *contains* ϑ_0 *with* $MCARMA(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$*, the identifiable uniqueness follows from some properties satisfied by the innovations associated to the true parameter* ϑ_0*, see [Schlemm and Stelzer 2012, Lemma 2.9 and 2.10], i. e. Assumption* $B.10$ *can then be dropped without any replacement.*

e) In case of a correctly specified parameter space, we can replace Assumption $B.11$ *by the assumption that there exists a positive index* i_0 *such that the* $[(i_0+2)d^2] \times r$ *matrix*

$$\nabla_\vartheta \begin{pmatrix} \left[I_{i_0+1\times i_0+1} \otimes K_\vartheta^T \otimes C_\vartheta\right] \begin{pmatrix} \operatorname{vec}\exp(I_{N\times N}h) \\ \operatorname{vec}\exp(A_\vartheta h) \\ \vdots \\ \operatorname{vec}\exp(A_\vartheta^{i_0} h) \end{pmatrix} \\ \operatorname{vec} V_\vartheta \end{pmatrix}_{\vartheta=\vartheta_0}$$

has rank $N(\Theta)$*. This condition is used in Schlemm and Stelzer [2012] as Assumption C11 and guarantees the desired non-singularity.*

We will now put some of these assumptions to use in order to show that we do not need to assume h-identifiability, but can deduce it from the given assumptions:

Theorem 2.21 (Schlemm and Stelzer [2012], Theorem 3.13)**.** *Assume that the space* Θ *with associated family of continuous–time state space models* $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ *satisfies Assumptions* $B.2$*,* $B.3$*,* $B.5$*,* $B.6$ *and* $B.7$*. Then the corresponding collection of output processes* $(MCARMA(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta))_\vartheta, \vartheta \in \Theta)$ *is h-identifiable from the spectral density.*

Proof. See cited reference. □

As a next step, we will answer the questions raised by the use of the approximate pseudo–innovations. The answers are quite pleasant, because under mild conditions it does not matter if we consider the empirical approximate pseudo-innovations or their theoretical counterparts, the pseudo-innovations:

Lemma 2.22. *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumptions B.1 to B.5.*

a) There exists a matrix sequence $(c_{\vartheta,k})_{k\in\mathbb{N}}$ such that

$$\epsilon_{\vartheta,k} = Y(kh) + \sum_{\nu=1}^{\infty} c_{\vartheta,\nu} Y((k-\nu)h), \quad k \in \mathbb{Z}.$$

Furthermore, there exists a positive constant C and a constant $\rho \in (0,1)$ such that

$$\sup_{\vartheta\in\Theta} \|c_{\vartheta,k}\| \leq C\rho^k, \quad k \in \mathbb{N}.$$

If the initial values $\widehat{X}_{\vartheta,initial}$ are such that $\sup_{\vartheta\in\Theta} \|\widehat{X}_{\vartheta,initial}\|$ is almost surely finite, then there exist $C > 0$ and $\rho \in (0,1)$ such that

$$\sup_{\vartheta\in\Theta} \|\epsilon_{\vartheta,k} - \widehat{\epsilon}_{\vartheta,k}\| \leq C\rho^k \quad \mathbb{P}\text{-}a.s.$$

b) For each $i \in \{1,\ldots,N(\Theta)\}$, there exists a matrix sequence $(c^{(i)}_{\vartheta,k})_{k\in\mathbb{N}}$ such that

$$\partial_i \epsilon_{\vartheta,k} = \sum_{\nu=1}^{\infty} c^{(i)}_{\vartheta,\nu} Y((k-\nu)h), \quad k \in \mathbb{Z}.$$

Furthermore, there exists a positive constant C and a constant $\rho \in (0,1)$ such that

$$\sup_{\vartheta\in\Theta} \|c^{(i)}_{\vartheta,k}\| \leq C\rho^k, \quad k \in \mathbb{N}.$$

If for $i \in \{1,\ldots,N(\Theta)\}$ the initial values $\widehat{X}_{\vartheta,initial}$ are such that $\sup_{\vartheta\in\Theta} \|\widehat{X}_{\vartheta,initial}\|$ and $\sup_{\vartheta\in\Theta} \|\partial_i \widehat{X}_{\vartheta,initial}\|$ are almost surely finite, then there exist $C > 0$ and $\rho \in (0,1)$ such that

$$\sup_{\vartheta\in\Theta} \|\partial_i \epsilon_{\vartheta,k} - \partial_i \widehat{\epsilon}_{\vartheta,k}\| \leq C\rho^k \quad \mathbb{P}\text{-}a.s.$$

c) For each $i,j \in \{1,\ldots,N(\Theta)\}$, there exists a matrix sequence $(c^{(i,j)}_{\vartheta,k})_{k\in\mathbb{N}}$ such that

$$\partial^2_{i,j} \epsilon_{\vartheta,k} = \sum_{\nu=1}^{\infty} c^{(i,j)}_{\vartheta,\nu} Y((k-\nu)h), \quad k \in \mathbb{Z}.$$

Furthermore, there exists a positive constant C and a constant $\rho \in (0,1)$ such

that

$$\sup_{\vartheta\in\Theta} \|c_{\vartheta,k}^{(i,j)}\| \leq C\rho^k, \quad k \in \mathbb{N}.$$

If for some $i, j \in \{1, \ldots, N(\Theta)\}$ *the initial values* $\widehat{X}_{\vartheta,initial}$ *are such that* $\sup_{\vartheta\in\Theta} \|\widehat{X}_{\vartheta,initial}\|$, $\sup_{\vartheta\in\Theta} \|\partial_l \widehat{X}_{\vartheta,initial}\|$, $l \in \{i, j\}$, *and* $\sup_{\vartheta\in\Theta} \|\partial_{i,j}^2 \widehat{X}_{\vartheta,initial}\|$ *are almost surely finite, then there exist* $C > 0$ *and* $\rho \in (0, 1)$ *such that*

$$\sup_{\vartheta\in\Theta} \|\partial_{i,j}\epsilon_{\vartheta,k} - \partial_{i,j}\widehat{\epsilon}_{\vartheta,k}\| \leq C\rho^k \quad \mathbb{P}\text{-}a.s.$$

Proof. Part a) is Schlemm and Stelzer [2012, Lemma 2.6], part b) is Schlemm and Stelzer [2012, Lemma 2.11i) and ii)] and part c) is Schlemm and Stelzer [2012, Lemma 2.11iii) and iv)] where we additionally use Schlemm and Stelzer [2012, Lemma 3.14]. □

A consequence of this lemma is especially that the approximate pseudo–innovations converge to the pseudo–innovations at an exponential rate if the assumptions on the initial values are satisfied. In the rest of the thesis, we always assume this to be the case (it suffices to set the initial values to 0 for example). Furthermore, the convergence of the approximate pseudo–innovations also carries over to the likelihood function, the approximate likelihood function and their respective derivatives:

Lemma 2.23. *Assume that the space* Θ *with associated family of continuous–time state space models* $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ *satisfies Assumptions B.1 to B.5. If for* $i, j \in \{1, \ldots, N(\Theta)\}$ *the initial values* $\widehat{X}_{\vartheta,initial}$ *are such that* $\sup_{\vartheta\in\Theta} \|\widehat{X}_{\vartheta,1}\|$, $\sup_{\vartheta\in\Theta} \|\partial_i \widehat{X}_{\vartheta,1}\|$ *and* $\sup_{\vartheta\in\Theta} \|\partial_{i,j}^2 \widehat{X}_{\vartheta,1}\|$ *are almost surely finite, then it holds:*

a) $\sup_{\vartheta\in\Theta} \left|\widehat{\mathcal{L}}(\vartheta, Y^n) - \mathcal{L}(\vartheta, Y^n)\right| \to 0$ *as* $n \to \infty$ $\mathbb{P}$-*a.s.*

b) $\sqrt{n}\sup_{\vartheta\in\Theta} \left|\partial_i\widehat{\mathcal{L}}(\vartheta, Y^n) - \partial_i\mathcal{L}(\vartheta, Y^n)\right| \xrightarrow{\mathbb{P}} 0$ *as* $n \to \infty$.

c) $\sup_{\vartheta\in\Theta} \left|\partial_{i,j}^2\widehat{\mathcal{L}}(\vartheta, Y^n) - \partial_{i,j}^2\mathcal{L}(\vartheta, Y^n)\right| \to 0$ *as* $n \to \infty$ $\mathbb{P}$-*a.s.*

d) $\sup_{\vartheta\in\Theta} \mathbb{E}\left[\left|\widehat{\mathcal{L}}(\vartheta, Y^n) - \mathcal{L}(\vartheta, Y^n)\right|\right] \to 0$ *as* $n \to \infty$.

Proof. a) This is Schlemm and Stelzer [2012, Lemma 2.7] taking Schlemm and Stelzer [2012, Lemma 3.14] into account.

b) & c) The assertions can be shown as in a), using Schlemm and Stelzer [2012, Lemma 2.11].

d) We have $\sup_{\vartheta\in\Theta}\mathbb{E}\,\|\widehat{\epsilon}_{\vartheta,k}\|<\infty$, $\sup_{\vartheta\in\Theta}\mathbb{E}\,\|\epsilon_{\vartheta,k}\|<\infty$ as in the proof of Schlemm and Stelzer [2012, Lemma 2.7], and for some $\rho\in(0,1)$ the behavior

$$\sup_{\vartheta\in\Theta}\mathbb{E}\left[\left|\widehat{\mathcal{L}}(\vartheta,Y^n)-\mathcal{L}(\vartheta,Y^n)\right|\right]\leq\frac{C}{n}\sum_{k=1}^{n}\rho^k\sup_{\vartheta\in\Theta}(\mathbb{E}\|\widehat{\epsilon}_{\vartheta,k}\|+\mathbb{E}\|\epsilon_{\vartheta,k}\|)\overset{n\to\infty}{\to}0.$$

□

The rest of this sections is devoted to proving a central limit theorem for the quasi maximum likelihood estimator under Assumption B. To this end, first, we show that the output processes are exponentially strong mixing and discuss the implications of this:

Proposition 2.24. *Suppose that the parametric family $(A_\vartheta,B_\vartheta,C_\vartheta,L_\vartheta)_{\vartheta\in\Theta}$ of continuous–time state space models satisfies Assumptions B.1 and B.2. Then each output processes Y_ϑ is exponentially strongly mixing.*

Proof. Assumption $B.1$ guarantees that the Lévy processes L_ϑ all have finite second moments and Assumption $B.2$ guarantees that the solutions of the state space models all are causal. Therefore, the requirements of [Marquardt and Stelzer 2007, Proposition 3.34] are satisfied and the assertion follows. □

The next proposition collects auxiliary results which are used in the proof of the asymptotic normality of the QMLE. They are highlighted here separately for easier reference, because they will appear again later in a different context.

Proposition 2.25.

a) Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta,B_\vartheta,C_\vartheta,L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumptions B.1 to B.3 as well as B.9. Then, there exists a pseudo–true parameter $\vartheta^*\in\Theta$ as defined in Equation (2.20) and for every $n\in\mathbb{N}$, there exists

$$\vartheta_n^*=\operatorname*{arg\,min}_{\vartheta\in\Theta}\mathbb{E}\left[\widehat{\mathcal{L}}(\vartheta,Y^n)\right]\tag{2.21}$$

as well. If Θ also satisfies the other parts of Assumption B, then $\vartheta_n^*\to\vartheta^*$ as $n\to\infty$. In particular, for n sufficiently large ϑ_n^* is in the interior of Θ as well.

b) Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta,B_\vartheta,C_\vartheta,L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumptions B.1 to B.5. Then the strong law of large numbers

$$\widehat{\mathcal{L}}(\vartheta,Y^n)\to\mathscr{Q}(\vartheta)\quad\mathbb{P}\text{-a.s.}$$

holds uniformly in ϑ as $n \to \infty$.

c) *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B. Then, as $n \to \infty$,*

$$\sqrt{n}\nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{I}(\vartheta^*)),$$

where $\mathcal{I}(\vartheta^*) = \lim_{n\to\infty} n \operatorname{Var}(\nabla_\vartheta \mathcal{L}(\vartheta^*, Y^n))$.

d) *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumptions B.1 to B.5. Then the convergence*

$$\nabla^2_\vartheta \widehat{\mathcal{L}}(\vartheta, Y^n) \to \mathcal{J}(\vartheta) \quad \mathbb{P}\text{-}a.s.$$

holds uniformly in ϑ as $n \to \infty$, where $\mathcal{J}(\vartheta) := \mathbb{E}\left[\nabla^2_\vartheta l_{\vartheta,1}\right]$.

e) *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B. Then there exist $\epsilon, \alpha > 0$ such that for almost all ω and for every $n > n_1(\omega)$ and $\vartheta \in B_\epsilon(\vartheta^*) \cap \Theta$ we have*

$$\det\left(\nabla^2_\vartheta \widehat{\mathcal{L}}(\vartheta, Y^n)(\omega)\right) \geq \alpha.$$

Proof. a) The existence statements follow directly from Sin and White [1996, Proposition 3.1]. The convergence $\vartheta^*_n \to \vartheta^*$ follows from Lemma 2.23d).

b) This is exactly Schlemm and Stelzer [2012, Lemma 2.8] taking Schlemm and Stelzer [2012, Lemma 3.14] into account.

c) Note that under Assumption B we have

$$\nabla_\vartheta \mathbb{E}\left[\mathcal{L}(\vartheta, Y^n)\right]\Big|_{\vartheta=\vartheta^*} = 0.$$

Next, we use dominated convergence to interchange the expectation and derivation, giving

$$\mathbb{E}\left[\nabla_\vartheta \mathcal{L}(\vartheta, Y^n)\right]\Big|_{\vartheta=\vartheta^*} = 0. \tag{2.22}$$

The rest of the proof can now be carried out as that of [Schlemm and Stelzer 2012, Lemma 2.16]. That proof makes use of the fact that the above expectation is zero, but obtains this result in a different way, which is why we pointed it out separately here.

d) The sequence $\left(\partial^2_{i,j} l_{\vartheta,n}\right)_{i,j}$ is ergodic for every $i,j \in \{1,\dots,r\}$. This follows from the representation in (2.15), the moving average representations of the innovations and their derivatives as given in (2.22), (2.22), (2.22), the strong mixing of the output process Y (Proposition 2.24) and lastly [Krengel 1985, Theorem 4.3].
Hence, by Birkhoff's Ergodic Theorem, it follows that the matrix of second derivatives satisfies a strong law of large numbers, i. e. it holds pointwise

$$\nabla^2_\vartheta \mathcal{L}(\vartheta, Y^n) = \sum_{k=1}^{n} \nabla^2_\vartheta l_{\vartheta,k} \xrightarrow{\mathbb{P}\text{-a.s.}} \mathbb{E}\left[\nabla^2_\vartheta l_{\vartheta,1}\right] = \mathcal{J}(\vartheta),\ n \to \infty.$$

The stronger notion of uniform convergence can be shown by using the compactness of the parameter space and applying [Ferguson 1996, Theorem 16a)].

e) Assumption $B.11$ says that the Fisher information matrix $\mathbb{E}[\nabla^2_\vartheta l_{\vartheta^*,1}]$ is invertible and hence, $\det(\mathbb{E}[\nabla^2_\vartheta l_{\vartheta^*,1}]) > 0$. Moreover, by Assumption B.4 the map $\vartheta \mapsto \mathbb{E}[\nabla^2_\vartheta l_{\vartheta,1}]$ is continuous. Thus, there exist $\epsilon, \alpha > 0$ such that $\inf_{\vartheta \in B_\epsilon(\vartheta^*)\cap\Theta} \det(\mathbb{E}[\nabla^2_\vartheta l_{\vartheta,1}]) > \alpha$. Since by d) as $n \to \infty$,

$$\sup_{\vartheta \in B_\epsilon(\vartheta^*)\cap\Theta} \|\nabla^2_\vartheta \widehat{\mathcal{L}}(\vartheta, Y^n) - \mathbb{E}\left[\nabla^2_\vartheta l_{\vartheta,1}\right]\| \to 0 \quad \mathbb{P}\text{-a.s.},$$

we finally get $\lim_{n\to\infty} \inf_{\vartheta \in B_\epsilon(\vartheta^*)\cap\Theta} \det(\nabla^2_\vartheta \widehat{\mathcal{L}}(\vartheta, Y^n)) > \alpha$ $\mathbb{P}$-a.s..

□

The following lemma gives the result that the function $\mathscr{Q}$ attains its minimum at parameter values ϑ_0 with $Y = \text{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$, a statement which we will use later in the proof of consistency of information criteria again (see also [Schlemm and Stelzer 2012, Lemma 2.10] and [Boubacar Maïnassara 2012, Lemma 1], although the former only treats the case of a space Θ with $\vartheta_0 \in \Theta$):

Lemma 2.26. *Let Θ satisfy Assumption B in its entirety and ϑ_0 be such that $Y = MCARMA(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$. Then it holds for every $\vartheta \in \Theta$:*

$$\mathscr{Q}(\vartheta) \geq \mathscr{Q}(\vartheta_0).$$

Furthermore, for every $\vartheta \in \Theta$ with $Y_\vartheta \neq MCARMA(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$ there exists a $\delta > 0$ such that

$$\mathscr{Q}(\vartheta) - \mathscr{Q}(\vartheta_0) > \delta. \tag{2.23}$$

Proof. By definition we have

$$\begin{aligned}\mathscr{Q}(\vartheta) &= \log(2\pi) + \log(\det(V_\vartheta)) + \mathbb{E}\left[\epsilon_{\vartheta,1}^T V_\vartheta^{-1}\epsilon_{\vartheta,1}\right] \\ &= \log(2\pi) + \log(\det(V_\vartheta)) + \operatorname{tr}\left(V_\vartheta^{-1}\mathbb{E}\left[\epsilon_{\vartheta,1}\epsilon_{\vartheta,1}^T\right]\right) \\ &= \log(2\pi) + \log(\det(V_\vartheta)) + \operatorname{tr}\left(V_\vartheta^{-1}\left(\mathbb{E}\left[\epsilon_{\vartheta_0,1}\epsilon_{\vartheta_0,1}^T\right] + 2\mathbb{E}\left[\epsilon_{\vartheta_0,1}\left(\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}\right)^T\right]\right.\right. \\ &\quad \left.\left. + \mathbb{E}\left[\left(\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}\right)\left(\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}\right)^T\right]\right)\right) \qquad (2.24)\end{aligned}$$

Note that $\epsilon_{\vartheta,1}$ and $\epsilon_{\vartheta_0,1}$ are defined in terms of the same random variables Y, hence the difference $\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}$ is an element of the linear past of $\epsilon_{\vartheta_0,1}$, to which the latter is orthogonal by definition. This means the second expectation in (2.24) is zero (since $\mathbb{E}[\epsilon_{\vartheta_0,1}] = 0$), which we can use together with the fact that $\mathbb{E}\left[\epsilon_{\vartheta_0,1}\epsilon_{\vartheta_0,1}^T\right] = V_{\vartheta_0}$ to obtain

$$(2.24) = \log(2\pi) + \log(\det(V_\vartheta)) + \operatorname{tr}\left(V_\vartheta^{-1}\left(V_{\vartheta_0} + \mathbb{E}\left[\left(\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}\right)\left(\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}\right)^T\right]\right)\right).$$

For $\vartheta = \vartheta_0$ things simplify even more and we have

$$\begin{aligned}\mathscr{Q}(\vartheta_0) &= \log(2\pi) + \log(\det(V_{\vartheta_0})) + \operatorname{tr}\left(V_{\vartheta_0}^{-1}V_{\vartheta_0}\right) \\ &= \log(2\pi) + \log(\det(V_{\vartheta_0})) + d\end{aligned}$$

If we then regard the difference we find

$$\begin{aligned}\mathscr{Q}(\vartheta) - \mathscr{Q}(\vartheta_0) &= \log(\det(V_\vartheta)) + \operatorname{tr}\left(V_\vartheta^{-1}V_{\vartheta_0}\right) - \log(\det(V_{\vartheta_0})) \\ &\quad - d + \operatorname{tr}\left(V_\vartheta^{-1}\mathbb{E}\left[\left(\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}\right)\left(\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}\right)^T\right]\right) \\ &= -\log\left(\det(V_\vartheta^{-1}V_{\vartheta_0})\right) + \operatorname{tr}\left(V_\vartheta^{-1}V_{\vartheta_0}\right) - d + \delta, \qquad (2.25)\end{aligned}$$

where

$$\delta := \operatorname{tr}\left(V_\vartheta^{-1}\mathbb{E}\left[\left(\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}\right)\left(\epsilon_{\vartheta,1}-\epsilon_{\vartheta_0,1}\right)^T\right]\right).$$

Note that this definition implies that $\delta = 0$ if and only if $\epsilon_{\vartheta,1} = \epsilon_{\vartheta_0,1}$ almost surely. Otherwise, $\delta > 0$.

Since the matrices V_ϑ and V_{ϑ_0} are symmetric and positive definite, the same carries over to the product $V_\vartheta^{-1}V_{\vartheta_0}$. Denote by $\lambda_{\vartheta,i}$, $i = 1,\ldots,d$ the d strictly positive eigenvalues of this product. For $x > 0$ we have the elementary inequality $x - \log(x) \geq 1$, which we can use to establish

$$\sum_{i=1}^{d}\left(\lambda_{\vartheta,i} - \log(\lambda_{\vartheta,i})\right) \geq d$$

$$\begin{aligned}
&\iff \sum_{i=1}^{d} \lambda_{\vartheta,i} - \log\left(\prod_{i=1}^{d} \lambda_{\vartheta,i}\right) \geq d \\
&\iff \operatorname{tr}\left(V_\vartheta^{-1} V_{\vartheta_0}\right) - \log\left(\det\left(V_\vartheta^{-1} V_{\vartheta_0}\right)\right) \geq d. \qquad (2.26)
\end{aligned}$$

Applying (2.26) to (2.25), we see that

$$\mathscr{Q}(\vartheta) - \mathscr{Q}(\vartheta_0) \geq d - d + \delta = \delta \geq 0, \qquad (2.27)$$

which is exactly what we wanted to prove.
For the strictly positive lower bound in the case of $Y_\vartheta \neq \text{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$, we noted above that $\epsilon_{\vartheta,1} \neq \epsilon_{\vartheta_0,1}$ if and only if $\delta > 0$. Therefore, if $\delta = 0$, then $\epsilon_{\vartheta,1} = \epsilon_{\vartheta_0,1}$ holds almost surely. Since Assumptions B.1 - B.3 and B.6 hold, [Schlemm and Stelzer 2012, Lemma 2.9] is applicable and tells us that $\epsilon_{\vartheta,1} = \epsilon_{\vartheta_0,1}$ almost surely implies $\vartheta = \vartheta_0$. Thus (2.23) follows directly from (2.27). □

Remark 2.27. *In Lemma 2.26, the strict inequality with a strictly positive δ will always hold if we consider a parameter space whose vector of Kronecker indices m is not equal to the Kronecker indices of the output process Y, since then the data–generating process cannot be generated by a parameter in Θ by the assumptions on the parametrization (see also Remark 2.20b)). However, for a space Θ that contains a ϑ^* with $Y = MCARMA(A_{\vartheta^*}, B_{\vartheta^*}, C_{\vartheta^*}, L_{\vartheta^*})$, equality in the statement of the lemma will be attained for ϑ^*, which is then also the pseudo–true parameter in Θ as explained at the start of this section.*

We can now state the desired central limit theorem, which basically combines [Sin and White 1996, Proposition 4.1] and [Schlemm and Stelzer 2012, Theorem 3.4]:

Theorem 2.28. *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B. Then, as $n \to \infty$,*

$$\widehat{\vartheta}^n \to \vartheta^* \quad \mathbb{P}\text{-}a.s.,$$

and

$$\sqrt{n}\left(\widehat{\vartheta}^n - \vartheta^*\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Xi(\vartheta^*)),$$

where

$$\Xi(\vartheta^*) = \mathcal{J}^{-1}(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{J}^{-1}(\vartheta^*)$$

with

$$\mathcal{I}(\vartheta^*) = \lim_{n\to\infty} n \operatorname{Var}(\nabla_\vartheta \mathcal{L}(\vartheta^*, Y^n)) \quad \text{and} \quad \mathcal{J}(\vartheta^*) = \lim_{n\to\infty} \nabla^2_\vartheta \mathcal{L}(\vartheta^*, Y^n). \qquad (2.28)$$

Proof. The proof can be carried out in the same way as Schlemm and Stelzer [2012, Theorem 3.16, Theorem 2.4 and Theorem 2.5], respectively, replacing ϑ_0 by ϑ^* wherever it appears. For the strong consistency, the key idea is to make use of the convergence result given in Proposition 2.25b) and the fact that ϑ^* is the unique minimizer of $\mathscr{Q}$ in Θ, which follows from $B.10$ and ensures that the estimator converges to a unique limit, see also White [1996, Theorem 3.4]. For the asymptotic normality, one uses a Taylor expansion of $\nabla_\vartheta \widehat{\mathcal{L}}$ around the value ϑ^* and then uses the results from Proposition 2.25c)-e) to obtain the statement. Detailed steps are omitted here, since the arguments from the proofs of Schlemm and Stelzer [2012, Theorem 3.16, Theorem 2.4 and Theorem 2.5] apply completely analogously. The only difference is that instead of the true parameter we have the pseudo–true parameter here, but due to the additional assumption $B.10$ the same techniques still can be used. □

Remark 2.29. *a) For the strong consistency part of the theorem, Assumption $B.3$ can be relaxed to only require continuity instead of three times differentiability.*

b) In the case that we are in a correctly specified parameter space, this theorem corresponds exactly to [Schlemm and Stelzer 2012, Theorem 3.16].

CHAPTER 3

Consistency of information criteria for MCARMA processes

This chapter is devoted to the study of a class of information criteria for MCARMA processes and their asymptotic properties. The motivation behind information criteria is the following: Assume that we are given discrete–time, equidistant observations $Y^n = (Y(h), \ldots, Y(nh))$ of a d–dimensional MCARMA process Y. Let the Kronecker indices of the Echelon form, the degree of the AR polynomial and the degree of the MA polynomial, respectively, belonging to Y be denoted by m_0, p_0 and q_0, respectively. As we have seen in Section 2.2, the QMLE is consistent and asymptotically normal, even when we operate in a parameter space Θ which is misspecified as long as Assumption B is satisfied. However, in order to truly estimate the parameters of the data–generating process, it is necessary to operate in a space which contains ϑ_0 with $Y = \text{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$, because otherwise the QMLE will estimate a pseudo–true parameter instead. This especially means that we need to know the true Kronecker indices m_0. This is exactly the point where we require model selection, or, synonymously, information criteria. These criteria are characterized by the fact that they allow to compare different parameter spaces and enable us to say which of the spaces fits best to given data. Moreover, one would ideally hope to be able to identify the true parameter space eventually if such a thing exists. We will study a family of information criteria, which all have in common that they build upon the pseudo–Gaussian likelihood function and the QMLE. The main result of this chapter will be a theorem that gives conditions for the criteria to be asymptotically

weakly or strongly consistent, respectively. Another important fact is that these criteria generalize well–known information criteria for other model classes, e.g. the AIC and BIC. Since the structure of these criteria laid the groundwork for the more general definition we work with, we will study them and their derivation separately, highlighting how the idea of approximating the Kullback–Leibler discrepancy (for the AIC) or the Bayesian a posteriori probability (for the BIC) of different parameter spaces naturally leads to particular members of our more general family of criteria.

3.1. SETUP OF THE PARAMETER SPACES FOR ORDER SELECTION

As first step towards our information criteria, we will have a closer look at the relevant parameter spaces. Assume that different parameter spaces, each containing continuous–time state space models in Echelon form and differing by their Kronecker indices, are given. As explained at the end of Subsection 2.2.2, if we now estimate the true Kronecker indices m_0 by, say, an information criterion, we would (indirectly) obtain an estimate for p_0, but not for q_0. In order to alleviate this, we will "decompose" a space Θ, with fixed Kronecker indices m, and thus fixed AR degree p, into several smaller, not necessarily disjoint (in terms of their output processes) spaces and obtain so–called nested spaces.
To motivate this further and explain it with an illustrative example, let us consider the one–dimensional case first.

Example 3.1. *The main reason to use the Echelon form is that it provides an identifiable parametrization of multivariate CARMA models. In one dimension, an identifiable parametrization can be obtained by much simpler means, namely by using the coefficients of the AR and MA polynomial as defined in Definition 2.2 as parameters and demanding that those polynomials have no common zeros. In particular, this means that we can choose a parameter space as*

$$\Theta_0 = \{\vartheta = (a_1, \dots, a_p, b_0, \dots, b_{p-1}) : P \text{ and } Q \text{ have no common zeros}\} \subseteq \mathbb{R}^{p+p-1}$$

and impose the conditions from Assumption B in order to obtain a space suitable for quasi maximum likelihood estimation. We have $q = p - 1$ with no further restrictions, except those that are necessary from a technical perspective, i.e. mandated by Assumption B. In the space Θ_0, there therefore are $p + (p - 1) = 2p - 1$ free parameters. However, Assumption B does not imply that $b_0 \neq 0$ for every $\vartheta \in \Theta_0$. In this sense, the MA degree of the processes parametrized by Θ_0 is not uniquely defined. For example, if $p = 2$ and $b_0 = 0$, $b_1 \neq 0$ then the corresponding output process is a

CARMA(2,1) process where one of the MA coefficients is 0*. Alternatively, we can also interpret it as a CARMA(2,0) process and omit the superfluous coefficient that is equal to zero. Thus, the MA degree of processes parametrized by elements in* Θ_0 *as defined above can vary between* 0 *and* $p-1$*. This is exactly the effect that also occurs with the Echelon form, since fixing the Kronecker indices does not uniquely define the MA degree.*

We now "partition" the space Θ_0 *into a sequence of nested models, each of which is also suitable for maximum lilkelihood estimation. This comes at the price that these spaces will then not necessarily be disjoint anymore (in the sense that they do not all parametrize truly different output processes). In order to not violate Assumption* $B.5$*, it is necessary that we keep* p *fixed and only vary* q*. For* $q < p-1$*, we can then define*

$$\Theta = \{(a_1, \dots, a_p, b_0, \dots, b_q) : P \text{ and } Q \text{ have no common zeros}\} \subseteq \mathbb{R}^{p+q+1}. \quad (3.1)$$

Note that the MA order of the contained processes is now less or equal to q*. When doing order selection, instead of considering only the space* Θ_0*, we can then introduce* $p-1$ *additional spaces for each* p*, namely those with* $0 \leq q \leq p-2$*. The advantage is that we obtain more information that way: in this case we have* $p+q$ *free parameters in* Θ*. If we ultimately minimize an information criterion over all the spaces constructed in this way, we also obtain information about the order of the MA polynomial, not only about that of the AR polynomial.*
We can observe another phenomenon in this scenario: If $p = p_0$ *and* $q_0 \neq p_0 - 1$*, then for every* $q > q_0$ *and* Θ *as in* (3.1)*, we can find an element which generates the same output process as* ϑ_0*, namely*

$$\vartheta^* = (a_1^*, \dots, a_{p_0}^*, \underbrace{0, \dots, 0}_{q-q_0 \text{ times}}, b_0^*, \dots, b_{q_0}^*).$$

The notation ϑ^* *is not a coincidence. Remember that we denoted by* ϑ^* *the pseudo–true parameter in a misspecified space in Theorem 2.28, defined by* (2.20)*. Since* ϑ^* *differs from* ϑ_0 *only by some added zeros, which do not influence the innovation sequence* $(\epsilon_{\vartheta^*,k})_{k\in\mathbb{Z}}$*, we have* $\epsilon_{\vartheta^*,k} = \epsilon_{\vartheta_0,k}$ *for every* $k \in \mathbb{Z}$*. This in turn implies that the expected log-likelihood function* $\mathcal{Q}$*, defined in* (2.16)*, attains the same value at* ϑ^* *as at* ϑ_0*, while Lemma 2.26 tells us that* $\mathcal{Q}(\vartheta_0)$ *is the global minimum of* $\mathcal{Q}$*, i. e.* ϑ^* *therefore is a global minimum of* $\mathcal{Q}$ *in* Θ *and hence the pseudo–true parameter in* Θ*. This is an important observation and will play a crucial role in some of the results to come.*

For the multivariate case, the idea remains in principle exactly the same. However,

due to the complexity of the Echelon form, we cannot explicitly write down the structure of the "smaller spaces". Instead, we can only give an abstract definition by saying that a nested Θ contains parameter vectors that parametrize a continuous–time state space model in Echelon form of which some parameters are a priori restricted to prescribed values. The reason for this is that there is no general formula for the number of free parameters associated to given Kronecker indices. Moreover, unlike in the one-dimensional case, even if we chose to restrict parameters to 0, this does not automatically lead to a decrease in the degree of the MA polynomial of the corresponding MCARMA processes. However, it is possible (and reasonable in practical scenarios) to choose the spaces in such a way that each Θ contains MCARMA processes with Kronecker indices m and MA degree q or less for some $q \in \{0, \dots, p-1\}$ in order to obtain additional information when doing order selection. On the other hand, requiring this special "decomposition" is not necessary, which is why we proceed in a more general setting.

Also noteworthy is the fact that, in contrast to the one-dimensional case, we cannot simply remove parameters from a space to obtain the nested spaces. Instead we have to set certain parameters to a fixed value in the Echelon form. However, if a space Θ is nested in Θ', we still think of Θ as a subset of a vector space with dimension strictly less than $N(\Theta')$, since the parameter vectors should only contain the free parameters on which we have not imposed any a priori restrictions.

Let us complement these explanations by an example. We again look at the canonical parametrizations for $d = 2$ as presented in Table 2.1 and Table 2.2 for $d = s = 2$ and $H(0) = H_0 = -I_{2\times 2}$. We now want to decompose the spaces given in those tables into smaller spaces in such a way that the processes in two different spaces not only differ by their Kronecker indices, but also by the degree of their respective MA polynomial. Of course this is easy because we have already written down the parametrizations explicitly and therefore know exactly on which parameters we must impose which restrictions. The results are given in the following tables, where we repeat the entries of Table 2.1 and Table 2.2 for the sake of easier comparison:

Table 3.1.: Canonical state space realizations (A, B, C) of rational transfer functions with different Kronecker indices m and additional decomposition of the parameter space

m	(p,q)	$N(\Theta)$	A	B	C
$(1,1)$	$(1,0)$	7	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_3 & \vartheta_4 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_3 & \vartheta_4 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$(1,2)$	$(2,0)$	8	$\begin{pmatrix} \vartheta_1 & \vartheta_2 & 0 \\ 0 & 0 & 1 \\ \vartheta_3 & \vartheta_4 & \vartheta_5 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ 0 & 0 \\ \vartheta_3 & \vartheta_4 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
$(1,2)$	$(2,1)$	10	$\begin{pmatrix} \vartheta_1 & \vartheta_2 & 0 \\ 0 & 0 & 1 \\ \vartheta_3 & \vartheta_4 & \vartheta_5 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_6 & \vartheta_7 \\ \vartheta_3+\vartheta_5\vartheta_6 & \vartheta_4+\vartheta_5\vartheta_7 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
$(2,1)$	$(2,0)$	9	$\begin{pmatrix} 0 & 1 & 0 \\ \vartheta_1 & \vartheta_2 & \vartheta_3 \\ \vartheta_4 & \vartheta_5 & \vartheta_6 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ \vartheta_1 & \vartheta_3 \\ \vartheta_4 & \vartheta_6 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
$(2,1)$	$(2,1)$	11	$\begin{pmatrix} 0 & 1 & 0 \\ \vartheta_1 & \vartheta_2 & \vartheta_3 \\ \vartheta_4 & \vartheta_5 & \vartheta_6 \end{pmatrix}$	$\begin{pmatrix} \vartheta_7 & \vartheta_8 \\ \vartheta_1+\vartheta_2\vartheta_7 & \vartheta_3+\vartheta_2\vartheta_8 \\ \vartheta_4+\vartheta_5\vartheta_7 & \vartheta_6+\vartheta_5\vartheta_8 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
$(2,2)$	$(2,0)$	11	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ \vartheta_1 & \vartheta_2 & \vartheta_3 & \vartheta_4 \\ 0 & 0 & 0 & 1 \\ \vartheta_5 & \vartheta_6 & \vartheta_7 & \vartheta_8 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ \vartheta_1 & \vartheta_3 \\ 0 & 0 \\ \vartheta_5 & \vartheta_7 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
$(2,2)$	$(2,1)$	15	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ \vartheta_1 & \vartheta_2 & \vartheta_3 & \vartheta_4 \\ 0 & 0 & 0 & 1 \\ \vartheta_5 & \vartheta_6 & \vartheta_7 & \vartheta_8 \end{pmatrix}$	$\begin{pmatrix} \vartheta_9 & \vartheta_{10} \\ \vartheta_1+\vartheta_4\vartheta_{11}+\vartheta_2\vartheta_9 & \vartheta_3+\vartheta_2\vartheta_{10}+\vartheta_4\vartheta_{12} \\ \vartheta_{11} & \vartheta_{12} \\ \vartheta_5+\vartheta_8\vartheta_{11}+\vartheta_6\vartheta_9 & \vartheta_7+\vartheta_6\vartheta_{10}+\vartheta_8\vartheta_{12} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

Table 3.2.: Canonical MCARMA realizations (P, Q) with order (p, q) of rational transfer functions with different Kronecker indices m and additional decomposition of the parameter space

m	(p,q)	$N(\Theta)$	$P(z)$	$Q(z)$
$(1,1)$	$(1,0)$	7	$\begin{pmatrix} z-\vartheta_1 & -\vartheta_2 \\ -\vartheta_3 & z-\vartheta_4 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_3\vartheta_4 & \end{pmatrix}$
$(1,2)$	$(2,0)$	8	$\begin{pmatrix} z-\vartheta_1 & -\vartheta_2 \\ -\vartheta_3 & z^2-\vartheta_4 z-\vartheta_5 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_3 & \vartheta_5 \end{pmatrix}$
$(1,2)$	$(2,1)$	10	$\begin{pmatrix} z-\vartheta_1 & -\vartheta_2 \\ -\vartheta_3 & z^2-\vartheta_4 z-\vartheta_5 \end{pmatrix}$	$\begin{pmatrix} \vartheta_1 & \vartheta_2 \\ \vartheta_6 z+\vartheta_3 & \vartheta_7 z+\vartheta_5 \end{pmatrix}$
$(2,1)$	$(2,0)$	9	$\begin{pmatrix} z^2-\vartheta_1 z-\vartheta_2 & -\vartheta_3 \\ -\vartheta_4 z-\vartheta_5 & z-\vartheta_6 \end{pmatrix}$	$\begin{pmatrix} \vartheta_2 & \vartheta_3 \\ \vartheta_5 & \vartheta_6 \end{pmatrix}$
$(2,1)$	$(2,1)$	11	$\begin{pmatrix} z^2-\vartheta_1 z-\vartheta_2 & -\vartheta_3 \\ -\vartheta_4 z-\vartheta_5 & z-\vartheta_6 \end{pmatrix}$	$\begin{pmatrix} \vartheta_7 z+\vartheta_2 & \vartheta_8 z+\vartheta_3 \\ \vartheta_5 & \vartheta_6 \end{pmatrix}$
$(2,2)$	$(2,0)$	11	$\begin{pmatrix} z^2-\vartheta_1 z-\vartheta_2 & -\vartheta_3 z-\vartheta_4 \\ -\vartheta_5 z-\vartheta_6 & z^2-\vartheta_7 z-\vartheta_8 \end{pmatrix}$	$\begin{pmatrix} \vartheta_2 & \vartheta_4 \\ \vartheta_6 & \vartheta_8 \end{pmatrix}$
$(2,2)$	$(2,1)$	15	$\begin{pmatrix} z^2-\vartheta_1 z-\vartheta_2 & -\vartheta_3 z-\vartheta_4 \\ -\vartheta_5 z-\vartheta_6 & z^2-\vartheta_7 z-\vartheta_8 \end{pmatrix}$	$\begin{pmatrix} \vartheta_9 z+\vartheta_2 & \vartheta_{10} z+\vartheta_4 \\ \vartheta_{11} z+\vartheta_6 & \vartheta_{12} z+\vartheta_8 \end{pmatrix}$

Returning to the general procedure, we can then consider all of these spaces, which differ by the Kronecker indices and the number of free parameters, instead of the spaces which only differ by the Kronecker indices. Of course, we will still require the nested spaces to fulfill Assumption B in order to be able to do QMLE in them, but this is possible. After having explored the structure of the parameter spaces under consideration in detail, we progress further towards the information criteria. As already mentioned, one goal later on is to study their asymptotic behavior. To this end, the results and tools from Chapter 2 will of course be necessary. However, we will need some additional tools. In particular, the most important result in the study of consistency will be the law of the iterated logarithm for the function $\widehat{\mathcal{L}}$ as defined in (2.18). For this reason, we first derive this result.

3.2. The law of the iterated logarithm

The goal of this section is to establish a law of the iterated logarithm for $\widehat{\mathcal{L}}$. This is done because in order to deduce strong consistency of information criteria later, we will need to study the limit superior of $\widehat{\mathcal{L}}$. The law of the iterated logarithm will provide us with a suitable scaling sequence that, when multiplied with $\widehat{\mathcal{L}}$, will lead to a non–zero limit superior, which is exactly what we will need.
To make the derivation more manageable, it is broken down into three separate steps, each building upon the former and culminating in the desired result. In the

following proposition we begin by establishing a law of the iterated logarithm for linear combinations of partial derivatives of the quasi log–likelihood function.

Proposition 3.2. *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B. Then, for every $x \in \mathbb{R}^{N(\Theta)} \setminus \{0_{N(\Theta)}\}$ it holds $\mathbb{P}$-a.s.*

$$\begin{aligned}\limsup_{n\to\infty} \frac{-\sqrt{n}}{\sqrt{\log(\log(n))}} x^T \nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n) &= \limsup_{n\to\infty} \frac{\sqrt{n}}{\sqrt{\log(\log(n))}} x^T \nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n) \\ &= \sqrt{2 \cdot x^T \mathcal{I}(\vartheta^*) x}.\end{aligned}$$

Proof. Let $x \in \mathbb{R}^{N(\Theta)} \setminus \{0_{N(\Theta)}\}$. First, it can be deduced that $x^T\mathcal{I}(\vartheta^*)x$ is finite and positive from Schlemm and Stelzer [2012, Lemma 2.16]. Moreover, by Schlemm and Stelzer [2012, Eq. (2.24)] the representation

$$\partial_i l_{\vartheta^*,k} = \operatorname{tr}\left(V_{\vartheta^*}^{-1}\left(I_{d\times d} - \epsilon_{\vartheta^*,k}\epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\right)\partial_i V_{\vartheta^*}\right) + 2\partial_i \epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k} \tag{3.2}$$

holds. By Lemma 2.22 we know that both the pseudo-innovations and their partial derivatives can be expressed as moving averages of the true output process via

$$\epsilon_{\vartheta^*,k} = \sum_{\nu=0}^{\infty} c_{\vartheta^*,\nu} Y((k-\nu)h), \qquad \partial_i \epsilon_{\vartheta^*,k} = \sum_{\nu=0}^{\infty} c_{\vartheta^*,\nu}^{(i)} Y((k-\nu)h) \tag{3.3}$$

and the inequalities $\sup_{\vartheta\in\Theta}\|c_{\vartheta,\nu}\| \le C\rho^\nu$ and $\sup_{\vartheta\in\Theta}\|c_{\vartheta,\nu}^{(i)}\| \le C\rho^\nu$ are satisfied for some $C > 0$ and $\rho \in (0,1)$ for $i \in \{1,\dots,N(\Theta)\}$. Thus, $x^T\nabla_\vartheta l_{\vartheta^*,k} = \sum_{i=1}^{N(\Theta)} x_i \partial_i l_{\vartheta^*,k}$ can be written as $f(Y(kh), Y((k-1)h), \dots)$ for a suitable function f.
The aim is now to apply the law of the iterated logarithm for dependent random variables as it's given in Oodaira and Yoshihara [1971, Theorem 8], for which we need to check the following three conditions:

a) $\mathbb{E}\left[x^T\nabla_\vartheta l_{\vartheta^*,k}\right] = 0$ and $\mathbb{E}\left|x^T\nabla_\vartheta l_{\vartheta^*,k}\right|^{2+\delta_1} < \infty$ for some $\delta_1 > 0$.

b) $\mathbb{E}\left[\left|x^T\nabla_\vartheta l_{\vartheta^*,k} - \mathbb{E}\left[x^T\nabla_\vartheta l_{\vartheta^*,k} \mid \sigma\left(Y((k-m)h), \dots, Y(kh), \dots, Y((k+m)h)\right)\right]\right|^2\right]$ $= O(m^{-2-\delta_2})$ for some $\delta_2 > 0$ and $m \in \mathbb{N}$.

c) $\sum_{k=1}^{\infty} \alpha_{Y^{(h)}}(k)^{\frac{\delta_3}{2+\delta_3}} < \infty$ for some $0 < \delta_3 < \delta_1$, where $(\alpha_{Y^{(h)}}(k))_{k\in\mathbb{Z}}$ denotes the strong mixing coefficients of the process $(Y(kh))_{k\in\mathbb{Z}}$.

a) We start with the first condition. For the first part it follows as in (2.22) that $\mathbb{E}\left[\partial_i l_{\vartheta^*,k}\right] = 0$ for every $i \in \{1,\dots,N(\Theta)\}$, hence $\mathbb{E}\left[x^T\nabla_\vartheta l_{\vartheta^*,k}\right] = 0$. For the second part, for any $i \in \{1,\dots,N(\Theta)\}$ we employ (3.2) and the Cauchy-Schwarz inequality

to obtain

$$\begin{aligned}\mathbb{E}\left|\partial_i l_{\vartheta^*,k}\right|^{2+\delta_1} &\leq C\mathbb{E}\left|\operatorname{tr}\left(V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*}\right)\right|^{2+\delta_1} + C\mathbb{E}\left|\partial_i\epsilon_{\vartheta^*,k}^T V_{\vartheta}^{-1}\epsilon_{\vartheta^*,k}\right|^{2+\delta_1} \\ &\leq C\left(\mathbb{E}\|\epsilon_{\vartheta^*,k}\|^{4+2\delta_1} + \left(\mathbb{E}\|\epsilon_{\vartheta^*,k}\|^{4+2\delta_1}\mathbb{E}\|\partial_i\epsilon_{\vartheta^*,k}\|^{4+2\delta_1}\right)^{\frac{1}{2}}\right),\end{aligned}$$

where we have used the compactness of Θ in the last line. From Assumption $B.9$ we know that the driving Lévy process L of Y has finite $(4+\delta)$th moment for some $\delta > 0$, which carries over to the $(4+\delta)$th moment of $Y(kh)$, $k \in \mathbb{Z}$, and hence to $\epsilon_{\vartheta^*,k}$ and $\partial_i\epsilon_{\vartheta^*,k}$. With this, we obtain that the right-hand side is finite if $\delta_1 < \frac{\delta}{2}$. Since $i \in \{1,\dots,N(\Theta)\}$ is arbitrary and $x^T\nabla_\vartheta l_{\vartheta^*,k}$ is a linear combination of those components, we get $\mathbb{E}\left|x^T\nabla_\vartheta l_{\vartheta^*,k}\right|^{2+\delta_1} < \infty$.

b) For the second condition, we begin by decomposing the partial derivative as in the proof of Schlemm and Stelzer [2012, Lemma 2.16]. For $m \in \mathbb{N}$ we write

$$\partial_i l_{\vartheta^*,k} = Y_{m,k}^{(i)} - \mathbb{E}\left[Y_{m,k}^{(i)}\right] + Z_{m,k}^{(i)} - \mathbb{E}\left[Z_{m,k}^{(i)}\right],$$

where

$$\begin{aligned}Y_{m,k}^{(i)} =& \operatorname{tr}\left(V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*}\right) + \sum_{\nu,\nu'=0}^{m}\Big(-\operatorname{tr}\left(V_{\vartheta^*}^{-1}c_{\vartheta^*,\nu}Y((k-\nu)h)Y_{\vartheta_0}^T((k-\nu')h)c_{\vartheta^*,\nu'}^T V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*}\right) \\ &+ 2Y_{\vartheta_0}^T((k-\nu)h)c_{\vartheta^*,\nu}^{(i),T}V_{\vartheta^*}^{-1}c_{\vartheta^*,\nu'}Y((k-\nu')h)\Big),\end{aligned}$$

$$Z_{m,k}^{(i)} = \partial_i l_{\vartheta^*,k} - Y_{m,k}^{(i)}.$$

Hence, we obtain

$$\begin{aligned}&\mathbb{E}\left[\left|x^T\nabla_\vartheta l_{\vartheta^*,k} - \mathbb{E}\left[x^T\nabla_\vartheta l_{\vartheta^*,k}\right|\sigma\left(Y((k-m)h),\dots,Y(kh),\dots,Y((k+m)h)\right)\right]\right|^2\right] \\ &\leq \mathbb{E}\left[\left|\sum_{i=1}^{N(\Theta)} x_i Z_{m,k}^{(i)} - \mathbb{E}\left[\sum_{i=1}^{N(\Theta)} x_i Z_{m,k}^{(i)}\right]\right|^2\right] \\ &= \sum_{i=1}^{N(\Theta)} x_i^2\operatorname{Var}(Z_{m,k}^{(i)}) + 2\sum_{\substack{i,j=1\\ i\neq j}}^{N(\Theta)} x_i x_j\operatorname{Cov}(Z_{m,k}^{(i)}, Z_{m,k}^{(j)}).\end{aligned}$$

From step 2 of the proof of Schlemm and Stelzer [2012, Lemma 2.16] we know that $\operatorname{Cov}(Z_{m,k}^{(i)}, Z_{m,k}^{(j)}) \leq C\rho^m$ for a positive constant C and $\rho \in (0,1)$, and every $i,j \in \{1,\dots,N(\Theta)\}$. Thus, the second condition is satisfied as well.

c) Lastly, we turn to the third condition. By Proposition 2.24 the strong mixing coefficients $\alpha_Y(t)$ of $(Y(t))_{t\in\mathbb{R}}$ are $O(\mathrm{e}^{-at})$ for some $a > 0$, which carries over to those

of the sampled process $(Y(kh))_{k\in\mathbb{Z}}$. Thus, we can choose $\delta_3 < \delta_1 < \frac{\delta}{2}$ to obtain $\sum_{k=1}^{\infty} \alpha_{Y^{(h)}}(k)^{\frac{\delta_3}{2+\delta_3}} < \infty$ as desired.

Then a consequence of a)-c) and Oodaira and Yoshihara [1971, Theorem 8] is the law of the iterated logarithm

$$\limsup_{n\to\infty} \frac{\left|\sum_{k=1}^{n}(\sum_{i=1}^{N(\Theta)} x_i \partial_i l_{\vartheta^*,k})\right|}{\sqrt{2nx^T\mathcal{I}(\vartheta^*)x\log(\log(nx^T\mathcal{I}(\vartheta^*)x))}} = 1 \quad \mathbb{P}\text{-a.s.}$$

Since $\log(\log(nx^T\mathcal{I}(\vartheta^*)x)) = O(\log(\log(n)))$ we can therefore deduce the statement by symmetry (the driving Lévy process has expectation 0_s) for $\mathcal{L}$. Finally, by Lemma 2.23b) we can transfer the result to $\widehat{\mathcal{L}}$ as well. □

In the next step, we use this result to establish a law of the iterated logarithm for the norm of the gradient of $\widehat{\mathcal{L}}$, multiplied by an arbitrary matrix:

Theorem 3.3. *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B. Moreover, let $\Xi \in \mathbb{R}^{N(\Theta)\times N(\Theta)}$ be an arbitrary matrix. Then it holds that*

$$\limsup_{n\to\infty} \frac{\sqrt{n}}{\sqrt{\log(\log(n))}} \|\Xi\nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n)\| = \sqrt{2\cdot\lambda_{max}(\Xi\mathcal{I}(\vartheta^*)\Xi^T)} \quad \mathbb{P}\text{-a.s.}$$

Proof. An application of Proposition 3.2 gives

$$\limsup_{n\to\infty} \frac{\sqrt{n}}{\sqrt{\log(\log(n))}} x^T\Xi\nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n) = \sqrt{2\cdot x^T\Xi\mathcal{I}(\vartheta^*)\Xi^T x} \quad \mathbb{P}\text{-a.s.}$$

for every $x \in \mathbb{R}^{N(\Theta)} \setminus \{0_{N(\Theta)}\}$. Using the fact that $\mathbb{R}^{N(\Theta)}$ is its own dual space and viewing the mapping $x \mapsto x^T\Xi\nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n_{\vartheta_0})$ as the application of the linear functional x to $\Xi\nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n_{\vartheta_0})$, this means that a law of the iterated logarithm holds for every univariate process of the form $x^T\Xi\nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n_{\vartheta_0})$. Just as in the proof of Finkelstein [1971, Lemma 2], we can conclude from this that $\mathbb{P}$-a.s.

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{\sqrt{n}}{\sqrt{\log(\log(n))}} \|\Xi\nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n)\| \\
&= \limsup_{n\to\infty} \frac{\sqrt{n}}{\sqrt{\log(\log(n))}} \sup_{\|x\|=1} \left|x^T\Xi\nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n)\right| \\
&= \sup_{\|x\|=1} \sqrt{2\cdot x^T\Xi\mathcal{I}(\vartheta^*)\Xi^T x} \\
&= \sqrt{2\cdot\lambda_{\max}(\Xi\mathcal{I}(\vartheta^*)\Xi^T)}
\end{aligned}$$

where we additionally use Zhulenev [1991, Eq. (22)] since the covariance matrix of $\Xi\nabla_\vartheta\mathcal{L}(\vartheta^*, Y^n_{\vartheta_0})$ is not necessarily the identity (since we are in a finite-dimensional Hilbert space, see also Ledoux and Talagrand [1991, pp. 222 and 232] for the identification of the limit). □

Finally, having this theorem allows us to derive the law of the iterated logarithm for $\widehat{\mathcal{L}}$:

Theorem 3.4. *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B. Then*

$$\limsup_{n\to\infty} \frac{n}{\log(\log(n))} \left(\widehat{\mathcal{L}}(\vartheta^*, Y^n) - \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n)\right) = \lambda_{max}(\mathcal{J}(\vartheta^*)^{-\frac{1}{2}}\mathcal{I}(\vartheta^*)\mathcal{J}(\vartheta^*)^{-\frac{1}{2}})\ \mathbb{P}\text{-}a.s.$$

Proof. A first-order Taylor expansion of $\nabla_\vartheta\widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n)$ around ϑ^* gives

$$0 = \nabla_\vartheta\widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) = \nabla_\vartheta\widehat{\mathcal{L}}(\vartheta^*, Y^n) + \nabla^2_\vartheta\widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n)(\widehat{\vartheta}^n - \vartheta^*),$$

for some $\overline{\vartheta}^n$ with $\|\overline{\vartheta}^n - \vartheta^*\| \leq \|\widehat{\vartheta}^n - \vartheta^*\|$. Since by Theorem 2.28 we know that $\widehat{\vartheta}^n \to \vartheta^*$ $\mathbb{P}$-a.s., $\overline{\vartheta}^n \to \vartheta^*$ $\mathbb{P}$-a.s. as well. A conclusion of Proposition 2.25e) is that $\lim_{n\to\infty}\det(\nabla^2_\vartheta\widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n)) > 0$ $\mathbb{P}$-a.s., so that

$$\widehat{\vartheta}^n - \vartheta^* = -\left(\nabla^2_\vartheta\widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n)\right)^{-1}\nabla_\vartheta\widehat{\mathcal{L}}(\vartheta^*, Y^n) \quad \mathbb{P}\text{-a.s.} \tag{3.4}$$

is well-defined. Now we employ a Taylor expansion again, albeit this time we expand $\widehat{\mathcal{L}}(\vartheta^*, Y^n)$ around $\widehat{\vartheta}^n$ and use a second-order expansion. This gives us

$$\widehat{\mathcal{L}}(\vartheta^*, Y^n) = \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + \frac{1}{2}(\widehat{\vartheta}^n - \vartheta^*)^T\nabla^2_\vartheta\widehat{\mathcal{L}}(\check{\vartheta}^n, Y^n)(\widehat{\vartheta}^n - \vartheta^*),$$

for some $\check{\vartheta}^n$ with $\|\check{\vartheta}^n - \widehat{\vartheta}^n\| \leq \|\widehat{\vartheta}^n - \vartheta^*\|$, where we have used $\nabla_\vartheta\widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) = 0$. As above we have $\check{\vartheta}^n \to \vartheta^*$ $\mathbb{P}$-a.s. Rearranging the terms, we arrive at

$$\begin{aligned}\widehat{\mathcal{L}}(\vartheta^*, Y^n) - \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) &= \frac{1}{2}\|\nabla^2_\vartheta\widehat{\mathcal{L}}(\check{\vartheta}^n, Y^n)^{\frac{1}{2}}(\widehat{\vartheta}^n - \vartheta^*)\|^2 \\ &= \frac{1}{2}\|\nabla^2_\vartheta\widehat{\mathcal{L}}(\check{\vartheta}^n, Y^n)^{\frac{1}{2}}(\nabla^2_\vartheta\widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n))^{-1}\nabla_\vartheta\widehat{\mathcal{L}}(\vartheta^*, Y^n)\|^2. \end{aligned} \tag{3.5}$$

An application of Theorem 3.3 with $\Xi = \mathcal{J}(\vartheta^*)^{-\frac{1}{2}}$ (which is symmetric) yields

$$\begin{aligned}&\limsup_{n\to\infty} \frac{\sqrt{n}}{\sqrt{\log(\log(n))}}\|\mathcal{J}(\vartheta^*)^{-\frac{1}{2}}\nabla_\vartheta\widehat{\mathcal{L}}(\vartheta^*, Y^n)\| \\ =&\sqrt{2\cdot\lambda_{\max}(\mathcal{J}(\vartheta^*)^{-\frac{1}{2}}\mathcal{I}(\vartheta^*)\mathcal{J}(\vartheta^*)^{-\frac{1}{2}})} \quad \mathbb{P}\text{-a.s.}\end{aligned}$$

With $\nabla^2_\vartheta \widehat{\mathcal{L}}(\check{\vartheta}^n, Y^n)^{\frac{1}{2}} \nabla^2_\vartheta \widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n)^{-1} \to \mathcal{J}(\vartheta^*)^{-\frac{1}{2}}$ $\mathbb{P}$-a.s. (cf. Proposition 2.25d)) and (3.5) we can derive the statement. □

Remark 3.5. *This result is an analog to Sin and White [1996, Proposition 5.1] which investigates consistency of information criteria under some different model assumptions. However, it is stronger than the one in the cited article, since we are able to specify the limit superior exactly while in Sin and White [1996] it is only shown that convergence occurs.*

We are now at the point where we have studied all the auxiliary tools necessary for the treatment of order selection criteria, so that we can now introduce and analyze them.

3.3. Likelihood-based information criteria

This main section of the chapter will contain our main results on consistency of information criteria. First, we give their definition:

Definition 3.6. *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B. Furthermore, let $\widehat{\vartheta}^n$ be the QMLE based on Y^n in Θ as defined in* (2.19) *and let $C(n)$ be a positive, nondecreasing function of n with*

$$\lim_{n\to\infty} \frac{C(n)}{n} = 0.$$

Then a likelihood-based information criterion has the form

$$IC_n(\Theta) := \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + N(\Theta)\frac{C(n)}{n}. \tag{3.6}$$

These information criteria have the property that $\mathrm{IC}_n(\Theta) \xrightarrow{\mathbb{P}} \mathscr{Q}(\vartheta^*)$. Since $\mathscr{Q}$ attains its minimum at ϑ_0 for which $\mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ (cf. Lemma 2.26), when comparing different parameter spaces we choose that one for which the information criterion is minimal as the most suitable.

Remark 3.7. *$C(n)$ can be interpreted as penalty term for the inclusion of more parameters into the model. This penalty is needed to obtain meaningful information, since the inclusion of more parameters always leads to an improved fit of the model to the data, i. e. without the penalty term the criterion would always choose the model with the most parameters. However, this is not feasible, since the inclusion of too many parameters ultimately leads to an interpolation of the data, such that*

the model would not provide information about the phenomenon generating the data anymore. The employment of an information criterion can therefore be seen as seeking a trade-off between accuracy and complexity.

The condition $C(n)/n \to 0$ guarantees that underfitting is not possible, i. e. asymptotically there is no positive probability of choosing a parameter space which cannot generate the process underlying the data. However, $C(n)/n \to 0$ is not sufficient to exclude overfitting, i.e. an asymptotically positive probability to choose a space with more parameters than necessary. In the following we will give necessary and sufficient conditions to exclude this case. To this end we need some notation.

Definition 3.8. *Let Θ and Θ_0 be parameter spaces with associated families of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta_0}$ and $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$, respectively, satisfying Assumption B. Assume that there is a $\vartheta_0 \in \Theta_0$ with $MCARMA(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$. We say that Θ_0 is nested in Θ if $N(\Theta_0) < N(\Theta)$ and there exist a matrix $F \in \mathbb{R}^{N(\Theta)\times N(\Theta_0)}$ with $F^TF = I_{N(\Theta_0)\times N(\Theta_0)}$ as well as a $c \in \mathbb{R}^{N(\Theta)}$ such that*

$$(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta_0} = (A_{F\vartheta+c}, B_{F\vartheta+c}, C_{F\vartheta+c}, L_{F\vartheta+c})_{\vartheta\in\Theta_0}.$$

The interpretation of nested is that all processes generated by a parameter in Θ_0 can also be generated by a parameter in Θ. However, there are also processes which can be generated by a parameter in Θ, but not by a parameter in Θ_0. In this sense Θ_0 is contained in Θ. The condition $F^TF = I_{N(\Theta_0)\times N(\Theta_0)}$ guarantees that we have a bijective map from $\Theta_0 \to F\Theta_0 + c \subset \Theta$. For MCARMA processes parametrized in Echelon form, as explained in Section 3.1 a parameter space Θ that satisfies Assumption B contains only processes that have the same Kronecker indices $m = (m_1, \ldots, m_d)$ and hence, fixed degree $p = \max_{i=1,\ldots,d} m_i$ of the AR polynomial. For the MA polynomial we only know that the degree is less than or equal to $p-1$. However, in Section 3.1 we explained how one can further partition such a parameter space. In the context of Definition 3.8, Θ_0 could be a parameter space generating processes with Kronecker index m_0 and MA degree not exceeding q_0, where Θ generates processes with Kronecker index m_0 and MA degree not exceeding q, $q_0 < q \leq p_0 - 1$. Then Θ_0 is nested in Θ. In this way our information criteria can be used to estimate the Kronecker index, the degree of the AR polynomial and the degree of the MA polynomial.

In the following we investigate only parameter spaces with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)$ in Echelon form. Remember that we denoted the Kronecker indices, the degree of the AR polynomial and the degree of

the MA polynomial, respectively, belonging to Y by m_0, p_0 and q_0, respectively. Then Θ_0^E denotes the parameter space generating all MCARMA processes with Kronecker indices m_0. The degree of the AR polynomial of those processes is then p_0, the degree of the MA polynomial is between 0 and $p_0 - 1$. The space Θ_0^E is the biggest parameter space generating MCARMA processes in Echelon form, satisfying Assumption B and containing a parameter ϑ_0^E with $\text{MCARMA}(A_{\vartheta_0^E}, B_{\vartheta_0^E}, C_{\vartheta_0^E}, L_{\vartheta_0^E}) = Y$. Note that ϑ_0^E is then the pseudo–true parameter in Θ_0^E.

Next, we define under which circumstances IC_n is consistent; we distinguish two different types of consistency.

Definition 3.9.

a) The information criterion IC_n is called strongly consistent if for any parameter spaces Θ_0 and Θ with associated families of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta_0}$ and $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$, respectively, satisfying Assumption B and with a $\vartheta_0 \in \Theta_0$ such that $MCARMA(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$, and either $MCARMA(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta) \neq Y$ for every $\vartheta \in \Theta$ or Θ_0 being nested in Θ we have

$$\mathbb{P}\left(\limsup_{n\to\infty} \left(IC_n(\Theta_0) - IC_n(\Theta)\right) < 0\right) = 1.$$

b) The information criterion IC_n is called weakly consistent if for any parameter spaces Θ_0 and Θ with associated families of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta_0}$ and $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$, respectively, satisfying Assumption B and with a $\vartheta_0 \in \Theta_0$ such that $MCARMA(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$, and either $MCARMA(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta) \neq Y$ for every $\vartheta \in \Theta$ or Θ_0 being nested in Θ we have

$$\lim_{n\to\infty} \mathbb{P}\left(IC_n(\Theta_0) - IC_n(\Theta) < 0\right) = 1.$$

If the information criterion is strongly consistent, then the chosen parameter space converges almost surely to the true parameter space. For a weakly consistent information criterion we only have convergence in probability. Moreover, if we compare two parameter spaces both containing a parameter that generates the true output process, then we choose the parameter space with less parameters asymptotically almost surely in the strongly consistent case, whereas in the weakly consistent case we have convergence in probability. This especially means overfitting is asymptotically excluded. With these notions we characterize consistency of IC_n for MCARMA processes in terms of the penalty term $C(n)$.

Theorem 3.10.

a) The criterion IC_n is strongly consistent if

$$\limsup_{n\to\infty} \frac{C(n)}{\log(\log(n))} > \lambda_{max}(\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}}).$$

If $\limsup_{n\to\infty} C(n)/\log(\log(n)) = 0$, then the information criterion is not strongly consistent.

b) The criterion IC_n is weakly consistent if $\limsup_{n\to\infty} C(n) = \infty$.
If $\limsup_{n\to\infty} C(n) < \infty$ then IC_n is neither weakly nor strongly consistent.

c) Let Θ and Θ_0 be parameter spaces with associated families of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta_0}$ and $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$, respectively, satisfying Assumption B. Assume that there is a $\vartheta_0 \in \Theta_0$ with MCARMA$(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ and Θ_0 is nested in Θ with map F and ϑ^* is the pseudo–true parameter in Θ. Moreover, suppose $\limsup_{n\to\infty} C(n) = C < \infty$. Define

$$\mathcal{M}_F(\vartheta^*) := -\mathcal{J}^{-1}(\vartheta^*) + F(F^T\mathcal{J}(\vartheta^*)F)^{-1}F^T.$$

Then

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}(IC_n(\Theta_0) - IC_n(\Theta) > 0)\\ &= \mathbb{P}\left(\sum_{i=1}^{N(\Theta)-N(\Theta_0)} \lambda_i\chi_i^2 > 2[N(\Theta) - N(\Theta_0)]C\right) > 0,\end{aligned}$$

where (χ_i^2) is a sequence of independent χ^2 random variables with one degree of freedom and the λ_i are the $N(\Theta) - N(\Theta_0)$ strictly positive eigenvalues of

$$\mathcal{J}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_F(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{M}_F(\vartheta^*)\mathcal{J}(\vartheta^*)^{\frac{1}{2}}.$$

Proof. For the whole proof, we denote by ϑ_0 the parameter in Θ_0 with MCARMA$(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ and by ϑ^* the pseudo–true parameter in Θ. Moreover, let $\widehat{\vartheta}_0^n$ denote the QMLE based on Y^n in Θ_0, $\widehat{\vartheta}^n$ the QMLE based on Y^n in Θ and $\widehat{\vartheta}_0^E$ the QMLE based on Y^n in Θ_0^E. The corresponding quasi log–likelihood functions are denoted by $\widehat{\mathcal{L}}_0$, $\widehat{\mathcal{L}}$ and $\widehat{\mathcal{L}}_E$, respectively.

a) We distinguish two different cases.

Case 1: MCARMA$(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta) \neq Y$ for every $\vartheta \in \Theta$. Then

$$\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) = \widehat{\mathcal{L}}_0\left(\widehat{\vartheta}_0^n, Y^n\right) - \widehat{\mathcal{L}}\left(\widehat{\vartheta}^n, Y^n\right) + [N(\Theta_0) - N(\Theta)]\frac{C(n)}{n}. \quad (3.7)$$

On the one hand, by Theorem 3.4 we have that

$$\begin{aligned}\widehat{\mathcal{L}}\left(\widehat{\vartheta}^n, Y^n\right) &= \widehat{\mathcal{L}}\left(\vartheta^*, Y^n\right) + O_{\text{a.s.}}\left(\frac{\log(\log(n))}{n}\right),\\ \widehat{\mathcal{L}}_0\left(\widehat{\vartheta}_0^n, Y^n\right) &= \widehat{\mathcal{L}}_0\left(\vartheta_0, Y^n\right) + O_{\text{a.s.}}\left(\frac{\log(\log(n))}{n}\right),\end{aligned}$$

and on the other hand, by Proposition 2.25b)

$$\widehat{\mathcal{L}}\left(\vartheta^*, Y^n\right) = \mathscr{Q}(\vartheta^*) + o_{\text{a.s.}}(1) \quad \text{and} \quad \widehat{\mathcal{L}}_0\left(\vartheta_0, Y^n\right) = \mathscr{Q}(\vartheta_0) + o_{\text{a.s.}}(1).$$

Finally, in this case the inequality from eq. (2.23) is strict, so that for some $\delta > 0$

$$\begin{aligned}\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) &= \mathscr{Q}(\vartheta_0) - \mathscr{Q}(\vartheta^*) + \widehat{r}(n) + [N(\Theta_0) - N(\Theta)]\frac{C(n)}{n}\\ &< -\delta + \widehat{r}(n) + [N(\Theta_0) - N(\Theta)]\frac{C(n)}{n},\end{aligned}$$

where $\widehat{r}(n)$ is $o_{\text{a.s.}}(1)$. By assumption it holds that $C(n)/n \to 0$ as $n \to \infty$, so that we get

$$\mathbb{P}\left(\limsup_{n\to\infty}\left(\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta)\right) < -\delta\right) = 1.$$

Case 2: Θ_0 is nested in Θ with map F. Note that Θ_0 is also nested in Θ_0^E by definition, which then in turn means that Θ is nested in Θ_0^E, implying

$$\widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) = \min_{\vartheta\in\Theta}\widehat{\mathcal{L}}(\vartheta, Y^n) \geq \min_{\vartheta\in\Theta_0^E}\widehat{\mathcal{L}}_E(\vartheta, Y^n) = \widehat{\mathcal{L}}_E(\widehat{\vartheta}_E^n, Y^n). \quad (3.8)$$

Moreover, $\widehat{\epsilon}_{\vartheta_0,k} = \widehat{\epsilon}_{\vartheta^*,k} = \widehat{\epsilon}_{\vartheta_0^E,k}$ and hence,

$$\widehat{\mathcal{L}}_0\left(\vartheta_0, Y^n\right) = \widehat{\mathcal{L}}\left(\vartheta^*, Y^n\right) = \widehat{\mathcal{L}}_E\left(\vartheta_0^E, Y^n\right). \quad (3.9)$$

With this and (3.8) we receive

$$\widehat{\mathcal{L}}_0(\widehat{\vartheta}_0^n, Y^n) - \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) \leq \widehat{\mathcal{L}}_E\left(\vartheta_0^E, Y^n\right) - \widehat{\mathcal{L}}_E(\widehat{\vartheta}_E^n, Y^n).$$

Now, Theorem 3.4 tells us that

$$\begin{aligned}&\limsup_{n\to\infty}\frac{n}{\log(\log(n))}\left(\widehat{\mathcal{L}}_E\left(\vartheta_0^E,Y^n\right)-\widehat{\mathcal{L}}_E(\widehat{\vartheta}_E^n,Y^n)\right)\\&=\lambda_{\max}(\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}})\quad\mathbb{P}\text{-a.s.}\end{aligned}$$

Turning to the information criterion, this gives

$$\begin{aligned}&\limsup_{n\to\infty}\frac{n}{\log(\log(n))}\left(\mathrm{IC}_n(\Theta_0)-\mathrm{IC}_n(\Theta)\right)\\\leq&\limsup_{n\to\infty}\frac{n}{\log(\log(n))}\Big(\widehat{\mathcal{L}}_E\left(\vartheta_0^E,Y^n\right)-\widehat{\mathcal{L}}_E(\widehat{\vartheta}_E^n,Y^n)\\&\qquad+[N(\Theta_0)-N(\Theta)]\frac{C(n)}{\log(\log(n))}\Big)\\\leq&\lambda_{\max}(\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}})-\limsup_{n\to\infty}\frac{C(n)}{\log(\log(n))}\quad\mathbb{P}\text{-a.s.},\end{aligned}$$

since $N(\Theta_0)-N(\Theta)\leq -1$. Hence, if

$$\limsup_{n\to\infty}\frac{C(n)}{\log(\log(n))}>\lambda_{\max}(\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}}),$$

we obtain

$$\mathbb{P}\left(\limsup_{n\to\infty}\frac{n}{\log(\log(n))}\left(\mathrm{IC}_n(\Theta_0)-\mathrm{IC}_n(\Theta)\right)<0\right)=1.$$

Finally, if $\limsup_{n\to\infty}C(n)/\log(\log(n))=0$, then from

$$\widehat{\mathcal{L}}_0\left(\widehat{\vartheta}_0^n,Y^n\right)-\widehat{\mathcal{L}}\left(\widehat{\vartheta}^n,Y^n\right)\geq 0$$

it clearly follows that

$$\mathbb{P}\left(\limsup_{n\to\infty}\frac{n}{\log(\log(n))}\left(\mathrm{IC}_n(\Theta_0)-\mathrm{IC}_n(\Theta)\right)>0\right)=1,$$

so that strong consistency cannot hold.

b) Again we distinguish the two cases from part a). Case 1 is dealt with analogously as in a), so that we only need to give detailed arguments for case 2. Suppose therefore that Θ_0 is nested in Θ. Define the map $f:\Theta_0\to\Theta$ by $f(\vartheta)=F\vartheta+c$, where F and c are as in the definition of nested spaces. Then, a Taylor expansion

of $\widehat{\mathcal{L}}\left(f(\widehat{\vartheta}_0^n), Y^n\right)$ around $\widehat{\vartheta}^n$ results in

$$\begin{aligned}\widehat{\mathcal{L}}_0\left(\widehat{\vartheta}_0^n, Y^n\right) &= \widehat{\mathcal{L}}\left(f(\widehat{\vartheta}_0^n), Y^n\right) = \widehat{\mathcal{L}}\left(\widehat{\vartheta}^n, Y^n\right) \\ &\quad + \frac{1}{2}\left(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\right)^T \nabla_\vartheta^2 \widehat{\mathcal{L}}\left(\overline{\vartheta}^n, Y^n\right)\left(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\right) \qquad (3.10)\end{aligned}$$

with $\overline{\vartheta}^n$ such that $\|\overline{\vartheta}^n - \widehat{\vartheta}^n\| \leq \|f(\widehat{\vartheta}_0^n) - \widehat{\vartheta}^n\|$. Plugging (3.10) into (3.7) gives

$$\begin{aligned}\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) &= \frac{1}{2}\left(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\right)^T \nabla_\vartheta^2 \widehat{\mathcal{L}}\left(\overline{\vartheta}^n, Y^n\right)\left(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\right) \\ &\quad + [N(\Theta_0) - N(\Theta)]\frac{C(n)}{n}. \qquad (3.11)\end{aligned}$$

In order to be able to show weak consistency, we will study the behavior of the random variable $\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)$. Note that $\widehat{\mathcal{L}}_0\left(\vartheta, Y^n\right) = \widehat{\mathcal{L}}\left(f(\vartheta), Y^n\right)$ for $\vartheta \in \Theta_0$, so that by the chain rule

$$\nabla_\vartheta \widehat{\mathcal{L}}_0(\vartheta_0, Y^n) = F^T \nabla_\vartheta \widehat{\mathcal{L}}(f(\vartheta_0), Y^n) = F^T \nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n).$$

Moreover,

$$f(\widehat{\vartheta}_0^n) - \vartheta^* = f(\widehat{\vartheta}_0^n) - f(\vartheta_0) = F(\widehat{\vartheta}_0^n - \vartheta_0).$$

As in (3.4), we also have

$$\begin{aligned}\widehat{\vartheta}^n - \vartheta^* &= -\left(\nabla_\vartheta^2 \widehat{\mathcal{L}}(\check{\vartheta}^n, Y^n)\right)^{-1} \nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n), \\ \widehat{\vartheta}_0^n - \vartheta_0 &= -\left(\nabla_\vartheta^2 \widehat{\mathcal{L}}_0(\tilde{\vartheta}^n, Y^n)\right)^{-1} \nabla_\vartheta \widehat{\mathcal{L}}_0(\vartheta_0, Y^n),\end{aligned}$$

where $\check{\vartheta}^n$ is such that $\|\check{\vartheta}^n - \vartheta^*\| \leq \|\widehat{\vartheta}^n - \vartheta^*\|$ and $\tilde{\vartheta}^n$ is such that $\|\tilde{\vartheta}^n - \vartheta_0\| \leq \|\widehat{\vartheta}_0^n - \vartheta_0\|$. In particular, $\check{\vartheta}^n \to \vartheta^*$ and $\tilde{\vartheta}^n \to \vartheta_0$ $\mathbb{P}$-a.s. as $n \to \infty$. To summarize,

$$\begin{aligned}\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n) &= \widehat{\vartheta}^n - \vartheta^* - F(\widehat{\vartheta}_0^n - \vartheta_0) \\ &= \left[-\left(\nabla_\vartheta^2 \widehat{\mathcal{L}}(\check{\vartheta}^n, Y^n)\right)^{-1} + F\left(\nabla_\vartheta^2 \widehat{\mathcal{L}}_0(\tilde{\vartheta}^n, Y^n)\right)^{-1} F^T\right] \nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n).\end{aligned}$$

An application of Proposition 2.25c) and d) results in

$$\sqrt{n}(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)) \xrightarrow{\mathcal{D}} \left[-\mathcal{J}(\vartheta^*)^{-1} + F\mathcal{J}(\vartheta_0)^{-1}F^T\right] \mathcal{N}(0_{N(\Theta)}, \mathcal{I}(\vartheta^*)) =: \mathbf{N}_F. \qquad (3.12)$$

Since by the chain rule $\mathcal{J}(\vartheta_0) = F^T \mathcal{J}(\vartheta^*) F$ the random vector $\mathbf{N}_F$ is distributed as $\mathcal{N}(0_{N(\Theta)}, \mathcal{M}_F(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{M}_F(\vartheta^*))$ (note that $\mathcal{M}_F(\vartheta^*)$ is symmetric). Finally, by (3.11), Proposition 2.25d) and $C(n) \to \infty$ as $n \to \infty$,

$$\begin{aligned}
&\mathbb{P}(\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) < 0) \\
&= \mathbb{P}\left(\frac{1}{2}\sqrt{n}\left(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\right)^T \nabla_\vartheta^2 \widehat{\mathcal{L}}\left(\overline{\vartheta}^n, Y^n\right)\sqrt{n}\left(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\right)\right. \\
&\qquad \left. < -[N(\Theta_0) - N(\Theta)]C(n)\right) \overset{n\to\infty}{\longrightarrow} \mathbb{P}\left(\mathbf{N}_F^T \mathcal{J}(\vartheta^*)\mathbf{N}_F < \infty\right). \qquad (3.13)
\end{aligned}$$

Using Imhof [1961, Eq. (1.1)] gives

$$\mathbf{N}_F^T \mathcal{J}(\vartheta^*)\mathbf{N}_F \overset{\mathcal{D}}{=} \sum_{i=1}^{N(\Theta)} \lambda_i \chi_i^2, \qquad (3.14)$$

where (χ_i^2) is a sequence of independent χ^2 random variables with one degree of freedom and the λ_i are the eigenvalues of $\mathcal{J}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_F(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{M}_F(\vartheta^*)\mathcal{J}(\vartheta^*)^{\frac{1}{2}}$. Since $\mathrm{rank}(\mathcal{M}_F(\vartheta^*)) = N(\Theta) - N(\Theta_0)$ and $\mathcal{J}(\vartheta^*)^{\frac{1}{2}}$ and $\mathcal{I}(\vartheta^*)$ have full rank, the number of strictly positive eigenvalues of $\mathcal{J}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_F(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{M}_F(\vartheta^*)\mathcal{J}(\vartheta^*)^{\frac{1}{2}}$ is $N(\Theta) - N(\Theta_0)$. Hence, the result follows.

c) With the arguments in b) we obtain the statement.

□

Remark 3.11.

a) A conclusion of Theorem 3.10a) is that strong consistency of the information criterion always holds, independent of the process Y generating the observed data and hence ϑ_0^E, if $\limsup_{n\to\infty} C(n)/\log(\log(n)) = \infty$.

b) Let Θ_0 be nested in Θ with map F. Then it can be shown as in the proof of Theorem 3.4 that

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{n}{\log(\log(n))}\left(IC_n(\Theta_0) - IC_n(\Theta)\right) \\
&\quad = \lambda_{max}(\mathcal{M}_F(\vartheta^*)^{\frac{1}{2}}\mathcal{I}(\vartheta^*)\mathcal{M}_F(\vartheta^*)^{\frac{1}{2}}) + \limsup_{n\to\infty}[N(\Theta_0) - N(\Theta)]\frac{C(n)}{\log(\log(n))}.
\end{aligned}$$

This implies that the information criterion IC_n is not strongly consistent iff

$\limsup_{n\to\infty} C(n)/\log(\log(n)) < C^*$, *where*

$$C^* := \max_F \frac{\lambda_{max}(\mathcal{M}_F(\vartheta^*)^{\frac{1}{2}}\mathcal{I}(\vartheta^*)\mathcal{M}_F(\vartheta^*)^{\frac{1}{2}})}{N(\Theta) - N(\Theta_0)} \leq \lambda_{max}(\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{J}(\vartheta_0^E)^{-\frac{1}{2}}).$$

Since the structure of $\mathcal{J}(\vartheta^*)$ and $\mathcal{I}(\vartheta^*)$ is in general not known, it is difficult to calculate C^* explicitly. However, in the Gaussian case we will derive that $C^* = 2$ (cf. Corollary 3.12).

c) *We would like to note that these results are similar to the statement of Sin and White [1996, Corollary 5.3] under different model assumptions. However, the authors present only sufficient conditions for strong consistency, where we also have a necessary condition (see Remark 3.5 as well).*

d) *As the proof of Theorem 3.10a), Case 1, shows, for spaces Θ with MCARMA$(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta) \neq Y$ for every $\vartheta \in \Theta$ a necessary and sufficient condition for choosing the correct parameter space asymptotically with probability 1 is $\lim_{n\to\infty} C(n)/n = 0$. Only if we allow nested models as well the additional condition $\limsup_{n\to\infty} C(n)/\log(\log(n)) > C^*$ becomes necessary. The probability in Theorem 3.10c) is the overfitting probability.*

To wrap up this section, we want to study the special case where the observed MCARMA process is driven by a Brownian motion. Some of the technical auxiliary results for the proof are given in the appendix.

Corollary 3.12. *Assume that the Lévy process L which drives the observed process Y is a Brownian motion. Then:*

a) *IC_n is strongly consistent iff $\limsup_{n\to\infty} C(n)/\log(\log(n)) > 2$.*

b) *If $\limsup C(n) = C < \infty$, then the overfitting probability of IC_n for a space Θ in which Θ_0 is nested is*

$$\mathbb{P}(\chi^2_{N(\Theta)-N(\Theta_0)} > [N(\Theta) - N(\Theta_0)]C),$$

where $\chi^2_{N(\Theta)-N(\Theta_0)}$ denotes a χ^2-distributed random variable with $N(\Theta) - N(\Theta_0)$ degrees of freedom.

Proof. a) From Lemma A.2b) we know that there exists a space Θ_0 such that there is a $\vartheta_0 \in \Theta_0$ with MCARMA$(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ and Θ_0 is nested in Θ_0^E with map F. Moreover, $N(\Theta_0) = N(\Theta_0^E) - 1$ and

$$\lambda_{\max}(\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}) = 2.$$

Additionally, a conclusion of Lemma A.2a) is that

$$\lambda_{\max}(\mathcal{H}(\vartheta_0^E)^{-\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{H}(\vartheta_0^E)^{-\frac{1}{2}}) = 2\lambda_{\max}(I_{N(\Theta_0^E)\times N(\Theta_0^E)}) = 2.$$

Therefore the statement follows directly from Theorem 3.10a) and Remark 3.11b).

b) Lemma A.2c) tells us that those eigenvalues of $\mathcal{H}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_F(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{M}_F\mathcal{H}(\vartheta^*)^{\frac{1}{2}}$ can only be 0 or 2 where 2 appears exactly $N(\Theta) - N(\Theta_0)$ times. Hence,

$$\sum_{i=1}^{N(\Theta)} \lambda_i\chi_i^2 = 2 \sum_{i=1}^{N(\Theta)-N(\Theta_0)} \chi_i^2 \overset{\mathcal{D}}{=} 2\chi^2_{N(\Theta)-N(\Theta_0)}.$$

The statement now follows from the definition of the overfitting probability in Theorem 3.10c).

□

We have now completed the investigation of consistency for our information criteria. Note that the results of this section are analogous to the ones obtained for ARMAX processes with i.i.d. noise in Hannan and Deistler [2012, Theorem 5.5.1]. As our next topics, we will treat some particular information criteria in more detail, namely the AIC and BIC. Our study will follow the ideas that historically led to their definition. At the end, however, it will turn out that we precisely arrive at special cases of IC_n as defined in Definition 3.6, enabling us to use the results of this section to immediately draw conclusions about the consistency of these two prominent criteria.

3.4. AIC FOR MULTIVARIATE CARMA PROCESSES

3.4.1. DERIVATION OF THE AIC

The underlying idea of the Akaike Information Criterion (AIC) was first introduced in Akaike [1973] (see also de Leeuw [1992]) and has since been derived and motivated in a lot of ways, see e. g. Shibata [1976], Bozdogan [1987] or Burnham and Anderson [2002]. Moreover, it has been used for a wide range of models, including one-dimensional and multivariate or vector ARMA (VARMA) processes (see [Brockwell and Davis 1991, §9.3]) and Boubacar Maïnassara [2012], respectively). The extension of the guiding ideas to the context of MCARMA processes is the purpose of this subsection.

The fundamental idea behind the AIC is to estimate the so–called Kullback–Leibler information or Kullback–Leibler discrepancy, which was originally introduced in Kullback and Leibler [1951] and can be seen as a measure for the difference between two probability distributions. In order to write it down, suppose that we are given a random vector $\mathcal{X}$, a set Θ and parametric family of possible probability distributions of $\mathcal{X}$, represented by the corresponding densities $(f_\vartheta)_{\vartheta\in\Theta}$. The Kullback–Leibler discrepancy between the distributions corresponding to the parameters ϑ and ϑ_0 is then given by

$$K(f_\vartheta \mid f_{\vartheta_0}) = \int f_{\vartheta_0}(x) \log\left(\frac{f_{\vartheta_0}(x)}{f_\vartheta(x)}\right) \mathrm{d}x \tag{3.15}$$

An alternative interpretation of $K(f_\vartheta \mid f_{\vartheta_0})$ is that it denotes the amount of information lost when f_ϑ is used to approximate f_{ϑ_0}.

Note that in the literature this is sometimes also called the Kullback–Leibler distance, which is technically not correct since $K(f_\vartheta \mid f_{\vartheta_0}) \neq K(f_{\vartheta_0} \mid f_\vartheta)$ in general. Also noteworthy is the fact that this is just the negative of Boltzmann's entropy (Boltzmann [1877]), such that minimizing the Kullback–Leibler discrepancy (which will be the guiding principle later) is essentially just maximizing entropy, i. e. applying the second law of thermodynamics.

For our purposes it will be very helpful to consider a slight variation of this object, which arises from multiplying this quantity by two, expressing the integral as an expectation and rearranging some terms with help of the laws of the logarithm. Doing that we arrive at

$$d(f_\vartheta \mid f_{\vartheta_0}) := 2K(f_\vartheta \mid f_{\vartheta_0}) = \Delta(f_\vartheta \mid f_{\vartheta_0}) - \Delta(f_{\vartheta_0} \mid f_{\vartheta_0}) \tag{3.16}$$

where

$$\Delta(f_\vartheta \mid f_{\vartheta_0}) = \mathbb{E}_{\vartheta_0}\left[-2\log(f_\vartheta)\right]. \tag{3.17}$$

Note that the original definition of the Kullback–Leibler discrepancy does not assume that there is some sort of "true" model or parameter, it is merely a method to compare two distributions. However, if we suppose there indeed **is** a true, unknown model corresponding to the parameter ϑ_0, the motivation for the use of the Kullback–Leibler discrepancy as a tool in the context of model selection becomes clear: The value of $K(f_\vartheta \mid f_{\vartheta_0})$ (or, equivalently, $d(f_\vartheta \mid f_{\vartheta_0})$) decreases the better the true distribution is described by the one associated to the parameter ϑ. Hence, the density that comes closest to f_{ϑ_0} in the Kullback–Leibler sense is given by the one associated to

$$\begin{aligned}\arg\min_{\vartheta\in\Theta} 2K(f_\vartheta|f_{\vartheta_0}) &= \arg\min_{\vartheta\in\Theta}\{\Delta(f_\vartheta \mid f_{\vartheta_0}) - \Delta(f_{\vartheta_0} \mid f_{\vartheta_0})\} \\ &= \arg\min_{\vartheta\in\Theta}\left\{-2\mathbb{E}_{\vartheta_0}[\log(f_\vartheta)]\right\},\end{aligned}$$

where we have used that $\Delta(f_{\vartheta_0} \mid f_{\vartheta_0})$ in (3.16) cannot be influenced in any way, which is why minimizing $\Delta(f_\vartheta \mid f_{\vartheta_0})$ is the relevant notion. In our context $\mathcal{X} = Y^n = (Y(h), \ldots, Y(nh))$ is the sample of length n, containing equidistant observations of realizations of the MCARMA process of which we want to estimate m_0, p_0 and q_0. f_ϑ denotes the density of the observations Y^n for $\vartheta \in \Theta$, a parameter space which may or may not contain a parameter ϑ_0 with $Y = \mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$, i. e. it is explicitly allowed for Θ to be misspecified. Moreover, we assume that Θ is set up as explained in Section 3.1, i. e. the Echelon form is used for processes in Θ. On top of that, to make use of the results we obtained in previous chapters we assume that Θ satisfies Assumption B, which especially implies that Theorem 2.28 is applicable. As it is not possible to calculate $d(f_\vartheta \mid f_{\vartheta_0})$ for every ϑ directly if we only have some observations of the process and not full information about the parameters, we will need to approximate it.

The approximation is done by first replacing the unknown parameter ϑ in $\Delta(f_\vartheta \mid f_{\vartheta_0})$ by its QMLE, arriving at $\mathbb{E}_{\vartheta_0}\left[-2\log(f_{\widehat{\vartheta}^n(Y^n)}) \mid Y^n\right]$ as the object of interest. This is motivated by the consistency of the maximum likelihood estimator, which ideally converges to ϑ_0 if we are in a correctly specified parameter space and the fact that we regard the sample Y^n as given and fixed here, which is why we consider the conditional expectation.

In the second step of the approximation, we replace the unknown density f_ϑ by the Gaussian quasi–likelihood. This explains why we switched from $K(\vartheta \mid \vartheta_0)$ to

$d(\vartheta \mid \vartheta_0)$: $-2/n$ times the logarithm of the Gaussian quasi-likelihood is per definition equal to $\mathcal{L}$, such that we from now on use its sample–based version $\widehat{\mathcal{L}}$.
As a side remark, consider the following: If $-2/n \log(f_\vartheta) = \mathcal{L}(\vartheta, Y^n)$, then the pseudo–true parameter ϑ^* in a misspecified parameter space Θ as defined in (2.20) is then also a minimizer of (3.17) and (3.16), respectively. This means that in the case of a misspecified model the maximum likelihood estimator converges to the parameter which induces the process that is closest to the true one among all processes in Θ in the information–theoretic sense provided by the Kullback–Leibler discrepancy.
Returning to the approximation procedure, in a third step we now suppose that $\mathcal{Y}^n$ is a second sample of n observations with same spacing as Y^n and satisfying the same MCARMA (or state space) equations, but independent of Y^n. Then, we can also use this sample to calculate the likelihood function. Using that $\mathcal{Y}^n$ and Y^n are independent, we summarize our approximation steps as follows

$$\begin{aligned}\min_{\vartheta\in\Theta} \frac{\Delta(f_\vartheta \mid f_{\vartheta_0})}{n} &= \min_{\vartheta\in\Theta} \mathbb{E}_{\vartheta_0}\left[-\frac{2}{n}\log(f_\vartheta)\right] \approx \mathbb{E}_{\vartheta_0}\left[-\frac{2}{n}\log(f_{\widehat{\vartheta}^n(Y^n)}) \mid Y^n\right] \\ &\approx \mathbb{E}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), Y^n) \mid Y^n\right] \approx \mathbb{E}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), \mathcal{Y}^n) \mid Y^n\right] \qquad (3.18)\end{aligned}$$

The right-hand side can again be approximated by the following theorem:

Theorem 3.13. *As $n \to \infty$ it holds that*

$$n\left(\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), \mathcal{Y}^n) - \left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(\mathcal{Y}^n), \mathcal{Y}^n) - \frac{\operatorname{tr}\left(\mathcal{I}(\vartheta^*)\mathcal{J}^{-1}(\vartheta^*)\right)}{n}\right]\right) \xrightarrow{\mathcal{D}} Z_{\vartheta^*},$$

where Z_{ϑ^*} is a random variable with expectation $\mathbb{E}[Z_{\vartheta^*}] = 0$.

Proof. A second-order Taylor expansion of $\widehat{\mathcal{L}}(\widehat{\vartheta}^n(\mathcal{Y}^n), Y^n)$ around $\widehat{\vartheta}^n(Y^n)$ gives

$$\begin{aligned}\widehat{\mathcal{L}}(\widehat{\vartheta}^n(\mathcal{Y}^n), Y^n) &= \widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), Y^n) \\ &\quad + \frac{1}{2}\left(\widehat{\vartheta}^n(\mathcal{Y}^n) - \widehat{\vartheta}^n(Y^n)\right)^T \nabla^2_\vartheta \widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n)\left(\widehat{\vartheta}^n(\mathcal{Y}^n) - \widehat{\vartheta}^n(Y^n)\right),\end{aligned}$$

where $\|\overline{\vartheta}^n - \widehat{\vartheta}^n(Y^n)\| \leq \|\widehat{\vartheta}^n(\mathcal{Y}^n) - \widehat{\vartheta}^n(Y^n)\|$. Hence,

$$\begin{aligned}&\widehat{\mathcal{L}}(\widehat{\vartheta}^n(\mathcal{Y}^n), Y^n) - \widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), Y^n) \\ &\quad = \frac{1}{2}\operatorname{tr}\left(\nabla^2_\vartheta \widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n)\left(\widehat{\vartheta}^n(\mathcal{Y}^n) - \widehat{\vartheta}^n(Y^n)\right)\left(\widehat{\vartheta}^n(\mathcal{Y}^n) - \widehat{\vartheta}^n(Y^n)\right)^T\right).\end{aligned}$$

On the one hand, since both $\widehat{\vartheta}^n(Y^n)$ and $\widehat{\vartheta}^n(\mathcal{Y}^n)$ converge $\mathbb{P}$-a.s. to ϑ^*, the vector $\overline{\vartheta}^n \to \vartheta^*$ $\mathbb{P}$-a.s. as well. On the other hand, by the independence of Y^n and $\mathcal{Y}^n$, the random vectors $\widehat{\vartheta}^n(\mathcal{Y}^n)$ and $\widehat{\vartheta}^n(Y^n)$ are independent as well. By Theorem 2.28, as

$n \to \infty$,

$$\sqrt{n}\left(\widehat{\vartheta}^n(Y^n) - \vartheta^*, \widehat{\vartheta}^n(\mathcal{Y}^n) - \vartheta^*\right) \xrightarrow{\mathcal{D}} (\mathcal{N}_1, \mathcal{N}_2),$$

where $\mathcal{N}_1, \mathcal{N}_2$ are independent, $\mathcal{N}(0, \mathcal{J}^{-1}(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{J}^{-1}(\vartheta^*))$-distributed random vectors. A conclusion of Proposition 2.25d) is $\nabla^2_\vartheta \widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n) \to \mathcal{J}(\vartheta^*)$ $\mathbb{P}$-a.s. Hence, a continuous mapping theorem gives

$$n\left(\widehat{\mathcal{L}}(\widehat{\vartheta}^n(\mathcal{Y}^n), Y^n) - \widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), Y^n)\right) \xrightarrow{\mathcal{D}} \frac{1}{2}\operatorname{tr}\left(\mathcal{J}(\vartheta^*)(\mathcal{N}_1 + \mathcal{N}_2)(\mathcal{N}_1 + \mathcal{N}_2)^T\right),$$

and by the independence of $\mathcal{N}_1$ and $\mathcal{N}_2$ we have

$$\mathbb{E}\left[\mathcal{J}(\vartheta^*)(\mathcal{N}_1 + \mathcal{N}_2)(\mathcal{N}_1 + \mathcal{N}_2)^T\right] = 2\mathcal{J}(\vartheta^*)\mathbb{E}\left[\mathcal{N}_1\mathcal{N}_1^T\right] = 2\mathcal{I}(\vartheta^*)\mathcal{J}^{-1}(\vartheta^*).$$

The statement follows then since the expectation of the trace is the trace of the expectation. □

As a consequence of (3.18) and Theorem 3.13 we receive the approximation

$$\min_{\vartheta\in\Theta}\left[-\frac{2}{n}\mathbb{E}_{\vartheta_0}[\log(f_\vartheta)]\right] \approx \widehat{\mathcal{L}}(\widehat{\vartheta}^n(\mathcal{Y}^n), \mathcal{Y}^n) + \frac{\operatorname{tr}\left(\mathcal{I}(\vartheta^*)\mathcal{H}^{-1}(\vartheta^*)\right)}{n},$$

which becomes our information criterion via the following definition:

$$\mathrm{AIC}_n(\Theta) = \mathcal{L}(\widehat{\vartheta}^n, Y^n) + \frac{\operatorname{tr}\left(\mathcal{I}(\vartheta^*)\mathcal{J}^{-1}(\vartheta^*)\right)}{n} \tag{3.19}$$

Remark 3.14. *If the Lévy process L which drives the observed process Y is a Brownian motion and MCARMA$(A_{\vartheta^*}, B_{\vartheta^*}, C_{\vartheta^*}, L_{\vartheta^*}) = Y$, we have $\mathcal{I}(\vartheta^*) = 2\mathcal{J}(\vartheta^*)$ by Lemma A.2a) and hence, the AIC reduces to $AIC_n(\Theta) = \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + \frac{2N(\Theta)}{n}$.*

The form of the AIC given in this remark coincides with Akaike's original definition (cf. Akaike [1973]). This suggests to define an alternative version of the AIC, the Classical Akaike Information Criterion (CAIC), as follows:

$$\mathrm{CAIC}_n(\Theta) := \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + \frac{2N(\Theta)}{n}. \tag{3.20}$$

This criterion avoids the additional work of estimating the matrices $\mathcal{I}(\vartheta^*)$ and $\mathcal{H}^{-1}(\vartheta^*)$ appearing in the AIC, which comes at the cost of not being exact when the driving Lévy process is not a Brownian motion. This is appealing because it can be quite difficult to evaluate the trace term in practice, depending on the structure of the models one investigates. Fortunately, for MCARMA processes it turns out that there

are quite accessible methods of estimating the matrices $\mathcal{I}(\vartheta^*)$ and $\mathcal{J}(\vartheta^*)$ from data, which are discussed in Schlemm and Stelzer [2012], but shall be repeated here briefly: The matrix $\mathcal{J}(\vartheta^*)$ can be estimated consistently by $\widehat{\mathcal{J}}^n = \nabla^2_\vartheta \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n)$, which in turn can again be calculated by employing the Kalman filter, which is also able to evaluate the Hessian matrix of the Gaussian log-likelihood ([Schlemm and Stelzer 2012, p. 2197]).
For $\mathcal{I}(\vartheta^*)$ the idea, which originally goes back to Boubacar Maïnassara and Francq [2011], is to use the representation $\mathcal{I}(\vartheta^*) = \sum_{\Delta\in\mathbb{Z}} \operatorname{Cov}(\gamma_{\vartheta^*,0}, \gamma_{\vartheta^*,\Delta})$ where

$$\gamma_{\vartheta^*,m} = \nabla_\vartheta \left[\log(\det(V_{\vartheta^*})) + \epsilon^T_{\vartheta^*,m} V^{-1}_{\vartheta^*} \epsilon_{\vartheta^*,m}\right].$$

One now assumes that $(\gamma_{\vartheta^*,m})_{m\in\mathbb{N}}$ admits an infinite-order AR-representation of the form $\Phi(B)\gamma_{\vartheta^*,m} = U_m$ where $\Phi(z) = I_r + \sum_{i=1}^\infty \Phi_i z^i$ and $(U_m)_{m\in\mathbb{N}}$ is a weak white noise (i. e. uncorrelated, but not necessarily independent) with covariance matrix Σ_U. If this is true, one can interpret $\frac{\mathcal{I}(\vartheta^*)}{2\pi}$ as the value of the spectral density of $(\gamma_{\vartheta^*,m})_{m\in\mathbb{N}}$ at frequency 0 and obtain with the help of [Brockwell and Davis 1991, p. 459] that $\mathcal{I}$ can be written as $\mathcal{I}(\vartheta^*) = \Phi^{-1}(1)\Sigma_U\Phi^{-1}(1)$.
One now replaces the unknown pseudo–true parameter ϑ^* in $\gamma_{\vartheta_0,m}$ by the QML estimate $\widehat{\vartheta}^n$ and then fits a long autoregression to the resulting empirical objects, i. e. one chooses an integer $s > 0$ and performs a least-squares regression of $\gamma_{\widehat{\vartheta}^n,m}$ on $\gamma_{\widehat{\vartheta}^n,m-1}, \ldots, \gamma_{\widehat{\vartheta}^n,m-s}$ where $s+1 \le m \le n$. Denoting the corresponding empirical AR polynomial by $\widehat{\Phi}^n_s(z) = I_r + \sum_{i=1}^s \widehat{\Phi}^L_{i,s} z^i$ and the empirical covariance matrix of the residuals by $\widehat{U}^n_s$, it is claimed in [Boubacar Maïnassara and Francq 2011, Theorem 3] that

$$\widehat{I}^n_s = \left(\widehat{\Phi}^n_s(1)\right)^{-1} \widehat{U}^n_s \left(\left(\widehat{\Phi}^n_s(1)\right)^T\right)^{-1}$$

converges to $\mathcal{I}(\vartheta^*)$ in probability as $n, s \to \infty$ if the conditions $\mathbb{E}\left[\|\epsilon_{\vartheta^*,1}\|^{8+\delta}\right] < \infty$ for some $\delta > 0$ and $\frac{s^3}{n} \to 0$ are satisfied.
An alternative method of estimating $\mathcal{I}(\vartheta^*)\mathcal{J}(\vartheta^*)^{-1}$ for simulated data is the following: first, an estimate $\widehat{\Xi}^n$ of $\Xi(\vartheta^*)$ is calculated as n times the empirical covariance matrix of a suitably large number of independent realizations of $\widehat{\vartheta}^n$. Similarly, again simulating independently a suitably large number of times, one can obtain an estimate $\widehat{\mathcal{J}}^n$ of $\mathcal{J}(\vartheta^*)$ as the arithmetic mean of the realizations of $\nabla^2_\vartheta \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n)$. Because $\Xi(\vartheta^*) = \mathcal{J}(\vartheta^*)^{-1}\mathcal{I}(\vartheta^*)\mathcal{J}(\vartheta^*)^{-1}$, an estimate of $\mathcal{I}(\vartheta^*)\mathcal{J}(\vartheta^*)^{-1}$ is then $\widehat{\mathcal{J}}^n\widehat{\Xi}^n$. This estimator avoids the inversion of matrices at the cost of relying on simulations. This means that if one has only given a data set, it might not be feasible, unless one carries out the simulations using the maximum likelihood estimate obtained from

the data as the true parameter.

3.4.2. PROPERTIES OF THE AIC

In this short section, we want to investigate both versions of the AIC defined at the end of the last section in terms of the results of Chapter 3. In other words, we want to study the question whether those criteria exhibit strong or weak consistency. The answer is given in the following proposition:

Theorem 3.15. *Both the AIC and the CAIC are neither strongly nor weakly consistent.*

Proof. The CAIC is a special case of IC_n with $C(n) \equiv 2$. The assertion immediately follows from that theorem. The penalty term of the AIC does not exactly fit the scheme of IC_n, but the proof of Theorem 3.10c) can directly be adapted to obtain the same result. □

Remark 3.16. *a) By the above Theorem and the explanations in Remark 3.11 we see that the overfitting probability for both versions of the AIC is non-zero asymptotically. As pointed out in Remark 3.11d), this problem is induced by taking spaces in which the true space is nested into consideration. If we stick to strictly disjoint spaces and do not partition them further, then both versions of the AIC will be strongly consistent, again by Remark 3.11d) (for $CAIC_n$ this is obvious, as $\frac{C(n)}{n} \to 0$ for $n \to \infty$ obviously holds, for the other version the proof is the same as for Case 1 of Theorem 3.10a)). This comes at the cost of obtaining fewer information about the number of free parameters.*
Remember that this is directly related to information about the MA degree q_0 of the data-generating MCARMA process if the partitions are chosen suitably. As a consequence, if we only want to estimate the AR degree p_0 (which is uniquely defined by the Kronecker indices) then both AIC versions will be strongly consistent. This could be advantageous if we only want to consider CAR(p) processes of various orders as possible models, for example.
Another immediate consequence of this is that the underfitting probability of the AIC goes to 0 as $n \to \infty$, since the true parameter space can never be nested in a space with too few parameters.

b) The inconsistency of the AIC is widely known for other model classes, amongst others also for ARMA processes, see e. g. Shibata [1976] or [Hannan and Deistler 2012, Theorem 5.6.1], where the latter also shows that the underfitting probability goes to 0 asymptotically for ARMA processes. The results

of Theorem 3.15 therefore match these and show that these properties are in general also found for the class of MCARMA processes.

3.4.3. An alternative approach to the AIC

In this section we will take an alternative route to estimating the Kullback–Leibler discrepancy defined in Subsection 3.4.1. This will result in a different criterion for order selection, which is inspired by Boubacar Maïnassara [2012], who adapted the idea of Hurvich and Tsai [1993] to multivariate ARMA processes driven by weak white noise.
Fundamentally, we still work within the same framework and want to approximate and minimize the quantity $\Delta(f_\vartheta \mid f_{\vartheta_0})$ from (3.17). As before, Y^n denotes a sample of n equidistant observations of the underlying MCARMA(p_0, q_0) process. Also as before, the parameter space Θ in which we operate has been fixed for the moment, however it is still of the form of Section 3.1 and assumed to fulfill Assumption B. In a first step we now argue as before to approximate

$$\begin{aligned}\min_{\vartheta\in\Theta} \frac{\Delta(f_\vartheta \mid f_{\vartheta_0})}{n} &= \min_{\vartheta\in\Theta} \mathbb{E}_{\vartheta_0}\left[-\frac{2}{n}\log(f_\vartheta)\right] \approx \mathbb{E}_{\vartheta_0}\left[-\frac{2}{n}\log(f_{\widehat{\vartheta}^n(Y^n)}) \mid Y^n\right] \\ &\approx \mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), Y^n) \mid Y^n\right]\end{aligned}$$

For our new approach, we now write down the expectation on the right–hand side explicitly. For the sake of readability, we hereby write $\widehat{\vartheta}^n$ for $\widehat{\vartheta}^n(Y^n)$:

$$\begin{aligned}\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), Y^n)\right] &= d\log(2\pi) + \mathbb{E}_{\vartheta_0}\left[\log(\det(V_{\widehat{\vartheta}^n}))\right] + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{\vartheta_0}\left[\epsilon_{\widehat{\vartheta}^n,k}^T V_\vartheta^{-1}\epsilon_{\widehat{\vartheta}^n,k}\right] \\ &= nd\log(2\pi) + n\mathbb{E}_{\vartheta_0}\left[\log(\det(V_{\widehat{\vartheta}^n}))\right] + \operatorname{tr}\left(V_{\widehat{\vartheta}^n}^{-1}\mathbb{E}_{\vartheta_0}\left[\epsilon_{\widehat{\vartheta}^n,1}\epsilon_{\widehat{\vartheta}^n,1}^T\right]\right)\end{aligned} \tag{3.21}$$

We now approximate the right–hand side. To this end, we first drop the constant $nd\log(2\pi)$ from Eq. (3.21), as it is the same across all models and therefore negligible. In the wake of this we arrive at the expression

$$\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), Y^n)\right] \approx \mathbb{E}_{\vartheta_0}\left[\log(\det(V_{\widehat{\vartheta}^n}))\right] + \operatorname{tr}\left(V_{\widehat{\vartheta}^n}^{-1}\mathbb{E}_{\vartheta_0}\left[\epsilon_{\widehat{\vartheta}^n,1}\epsilon_{\widehat{\vartheta}^n,1}^T\right]\right) \tag{3.22}$$

In a first step of further approximation, we will now do a Taylor expansion of the pseudo–innovations $\epsilon_{\vartheta,n}$ around the pseudo–true parameter ϑ^*. From this we obtain

$$\epsilon_{\vartheta,n} = \epsilon_{\vartheta^*,n} + \nabla_\vartheta \epsilon_{\vartheta^*,n}(\vartheta - \vartheta^*) + R_n(\overline{\vartheta}) \tag{3.23}$$

where $\overline{\vartheta}$ is between ϑ^* and ϑ (in the sense of the Euclidean norm on the parameter space). The last summand in the expansion is the rest term of second order, which

is $O_p(\|\vartheta - \vartheta^*\|^2)$. With this expansion we deduce:

$$\begin{aligned}\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta,1}\epsilon_{\vartheta,1}^T\right] = \mathbb{E}_{\vartheta_0}\big[&\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T + \epsilon_{\vartheta^*,1}(\vartheta-\vartheta^*)^T\nabla_\vartheta\epsilon_{\vartheta^*,1}^T + \epsilon_{\vartheta^*,1}R_1^T \\ &+\nabla_\vartheta\epsilon_{\vartheta^*,1}(\vartheta-\vartheta^*)\epsilon_{\vartheta^*,1}^T + \nabla_\vartheta\epsilon_{\vartheta^*,1}(\vartheta-\vartheta^*)(\vartheta-\vartheta^*)^T\nabla_\vartheta\epsilon_{\vartheta^*,1}^T \\ &+\nabla_\vartheta\epsilon_{\vartheta^*,1}(\vartheta-\vartheta^*)R_1^T + R_1\epsilon_{\vartheta^*,1}^T + R_1(\vartheta-\vartheta^*)^T\nabla_\vartheta\epsilon_{\vartheta^*,1}^T + R_1R_1^T\big].\end{aligned} \tag{3.24}$$

Now we observe that the sequence $(\epsilon_{\vartheta^*,n})_{n\in\mathbb{N}}$ is orthogonal by construction, which implies $\epsilon_{\vartheta^0,1}$ is independent of any linear combination of its past values $\epsilon_{\vartheta^0,m}$, $m<1$. By Lemma 2.22 we know that both the innovations and their partial derivatives can be expressed as moving averages of the observations and can conclude that $\epsilon_{\vartheta^*,1}$ is also independent of its derivatives, i. e. every summand in (3.24) that contains two out of three of the objects $\epsilon_{\vartheta^*,1}$, $\nabla_\vartheta\epsilon_{\vartheta^*,1}$ and R_1 disappears because the expectation of the innovations is 0. Furthermore, we employ that R_1 is $O_p(\|\vartheta-\vartheta^*\|^2)$ to simplify (3.24), giving

$$\begin{aligned}\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta,1}\epsilon_{\vartheta,1}^T\right] = \mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T\right] + \underbrace{\mathbb{E}_{\vartheta_0}\left[\nabla_\vartheta\epsilon_{\vartheta^*,1}(\vartheta-\vartheta^*)(\vartheta-\vartheta^*)^T\nabla_\vartheta\epsilon_{\vartheta^*,1}^T\right]}_{=:D(\vartheta)} \\ + O_p(\|\vartheta-\vartheta^*\|^4)\end{aligned}$$

We plug this into (3.22) and obtain

$$\begin{aligned}\mathbb{E}_{\vartheta_0}\left[\mathcal{L}(\widehat{\vartheta}^n(Y^n),Y^n)\right] &\approx \mathbb{E}_{\vartheta_0}\left[\log\left(\det\left(V_{\widehat{\vartheta}^n}\right)\right)\right] + \mathbb{E}_{\vartheta_0}\left[\operatorname{tr}\left(V_{\widehat{\vartheta}^n}^{-1}\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T\right]\right.\right. \\ &\quad \left.\left.+V_{\widehat{\vartheta}^n}^{-1}D(\widehat{\vartheta}^n) + V_{\widehat{\vartheta}^n}^{-1}O_p(\|\widehat{\vartheta}^n-\vartheta^*\|^4)\right)\right] \\ &\approx \mathbb{E}_{\vartheta_0}\left[\log\left(\det\left(V_{\widehat{\vartheta}^n}\right)\right)\right] + \mathbb{E}_{\vartheta_0}\left[\operatorname{tr}\left(V_{\widehat{\vartheta}^n}^{-1}\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T\right]\right)\right] \\ &\quad + \mathbb{E}_{\vartheta_0}\left[\operatorname{tr}\left(V_{\widehat{\vartheta}^n}^{-1}D(\widehat{\vartheta}^n)\right)\right]\end{aligned} \tag{3.25}$$

by using the linearity of the expectation and the trace. Note that the O_p term is negligible for our purposes and thus has been dropped.
Next, we turn our attention to $V_{\widehat{\vartheta}^n}^{-1}$ which appears in both the second and the third term. From (3.23) we obtain the following equation when disregarding the rest term, which is asymptotically negligible:

$$\epsilon_{\vartheta,n} - \nabla_\vartheta\epsilon_{\vartheta^*,n}\vartheta = \epsilon_{\vartheta^*,n} - \nabla_\vartheta\epsilon_{\vartheta^*,n}\vartheta^*$$

Defining $M := -\nabla_\vartheta \epsilon_{\vartheta^*,n}$ and $\mathcal{R}_{\vartheta,n} := \epsilon_{\vartheta,n} + M\vartheta$, we can write this equivalently as

$$\mathcal{R}_{\vartheta,n} = M\vartheta^* + \epsilon_{\vartheta^*,n}.$$

This is a classical, multivariate linear regression model. If we now assume that the innovations are normally distributed with mean 0 and covariance matrix $\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T\right]$, we can use [Anderson 2003, Theorem 8.2.2] to obtain that n times the maximum likelihood estimator of that covariance matrix, which is given by $V_{\widehat{\vartheta}^n} = \frac{1}{n}\sum_{k=1}^n \epsilon_{\widehat{\vartheta}^n,k}\epsilon_{\widehat{\vartheta}^n,k}^T$, is distributed according to a $\mathcal{W}\left(\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T\right], n - N(\Theta)\right)$ distribution, where $\mathcal{W}(A,c)$ denotes the (d-dimensional) Wishart distribution with mean A and c degrees of freedom. A similar result is also found in Hurvich and Tsai [1993] in a purely autoregressive context. Note also that even if the normality of the innovations may not be true, we still assume it here in order to make progress with our derivation, acknowledging that this may entail an approximation error, which we find acceptable since the AIC is an approximative result by itself.
Having this it then holds by the general theory on the Wishart distribution (see e. g. [Muirhead 1982, p. 97]) that $\frac{1}{n}V_{\widehat{\vartheta}^n}^{-1}$ has a $\mathcal{W}^{-1}\left(\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T\right]^{-1}, n - N(\Theta)\right)$ distribution, where $\mathcal{W}^{-1}$ signifies the inverse Wishart distribution. This is useful because we can then deduce that

$$\begin{aligned}\frac{1}{n}\mathbb{E}_{\vartheta_0}\left[V_{\widehat{\vartheta}^n}^{-1}\right] &= \frac{\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T\right]^{-1}}{n - N(\Theta) - d - 1} \\ \Rightarrow \mathbb{E}_{\vartheta_0}\left[V_{\widehat{\vartheta}^n}^{-1}\right] &= \frac{n}{n - N(\Theta) - d - 1}\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T\right]^{-1}. \end{aligned} \tag{3.26}$$

This will be employed multiple times in the following. For now, we turn our attention to the third term in (3.25). We have

$$\begin{aligned}\operatorname{tr}(V_\vartheta^{-1}D(\vartheta)) &= \operatorname{tr}\left(V_\vartheta^{-1}\mathbb{E}_{\vartheta_0}\left[\nabla_\vartheta\epsilon_{\vartheta^*,1}(\vartheta-\vartheta^*)(\vartheta-\vartheta^*)^T\nabla_\vartheta\epsilon_{\vartheta^*,1}^T\right]\right) \\ &= \mathbb{E}_{\vartheta_0}\left[\operatorname{tr}\left(\nabla_\vartheta\epsilon_{\vartheta^*,1}^T V_\vartheta^{-1}\nabla_\vartheta\epsilon_{\vartheta^*,1}(\vartheta-\vartheta^*)(\vartheta-\vartheta^*)^T\right)\right] \\ &= \frac{1}{n}\operatorname{tr}\left(\mathbb{E}_{\vartheta_0}\left[\nabla_\vartheta\epsilon_{\vartheta^*,1}^T V_\vartheta^{-1}\nabla_\vartheta\epsilon_{\vartheta^*,1}\right] n(\vartheta-\vartheta^*)(\vartheta-\vartheta^*)^T\right). \end{aligned} \tag{3.27}$$

Note that there are three separate steps involved in getting from the first to the second line: First we employed the fact that V_ϑ^{-1} is not random to put it inside the expectation, then we interchanged the trace and the expectation and finally we applied the relation $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ with $B = \nabla_\vartheta\epsilon_{\vartheta^*,1}^T$ and $A = V_\vartheta^{-1}\nabla_\vartheta\epsilon_{\vartheta^*,1}(\vartheta-\vartheta^*)(\vartheta-\vartheta^*)^T$. Likewise, we interchanged expectation and trace again from the second to third line and pulled the non-random object $L(\vartheta-\vartheta^*)(\vartheta-\vartheta^*)^T$ out of the expectation.
Having this, we can now plug in the quasi maximum likelihood estimator of ϑ and

the covariance matrix of the innovations and take the expectation again. This leads us to the thing we are interested in, the third term in (3.25):

$$\begin{aligned}\mathbb{E}_{\vartheta_0}\left[\operatorname{tr}\left(V_{\widehat{\vartheta}^n}^{-1}D(\widehat{\vartheta}^n)\right)\right] &= \frac{1}{n}\mathbb{E}_{\vartheta_0}\left[\operatorname{tr}\left(\mathbb{E}_{\vartheta_0}\left[\nabla_\vartheta \epsilon_{\vartheta^*,1}^T V_{\widehat{\vartheta}^n}^{-1}\nabla_\vartheta \epsilon_{\vartheta^*,1}\right] n(\widehat{\vartheta}^n-\vartheta^*)(\widehat{\vartheta}^n-\vartheta^*)^T\right)\right] \\ &= \frac{1}{n}\operatorname{tr}\left(\mathbb{E}_{\vartheta_0}\left[\nabla_\vartheta \epsilon_{\vartheta^*,1}^T V_{\widehat{\vartheta}^n}^{-1}\nabla_\vartheta \epsilon_{\vartheta^*,1}\right]\mathbb{E}_{\vartheta_0}\left[n(\widehat{\vartheta}^n-\vartheta^*)(\widehat{\vartheta}^n-\vartheta^*)^T\right]\right).\end{aligned}$$

Theorem 2.28 gives that $\widehat{\vartheta}^n$ is asymptotically normally distributed. Thanks to this, the second factor in the above expression converges to the asymptotic covariance matrix $\Xi(\vartheta^*)$. Therefore we are able to do the following approximation, also taking into account (3.26):

$$\mathbb{E}_{\vartheta_0}\left[\operatorname{tr}\left(V_{\widehat{\vartheta}^n}^{-1}D(\widehat{\vartheta}^n)\right)\right] \approx \frac{1}{n}\operatorname{tr}\left(\mathbb{E}_{\vartheta_0}\left[\nabla_\vartheta \epsilon_{\vartheta^*,1}^T \frac{n}{n-N(\Theta)-d-1} V_{\vartheta^*}^{-1}\nabla_\vartheta \epsilon_{\vartheta^*,1}\right]\Xi(\vartheta^*)\right).$$

Moreover, the proof of [Schlemm and Stelzer 2012, Lemma 2.17] shows that

$$\nabla_\vartheta^2 \mathcal{L}(\vartheta^*, Y^n) \overset{n\to\infty}{\to} J_1(\vartheta^*) + J_2(\vartheta^*) \ \mathbb{P}\text{-a.s.},$$

where

$$J_1(\vartheta^*) = 2\mathbb{E}_{\vartheta_0}\left[\nabla_\vartheta \epsilon_{\vartheta^*,1}^T V_{\vartheta^*}^{-1}\nabla_\vartheta \epsilon_{\vartheta^*,1}\right],\ J_2 = \left(\operatorname{tr}\left[V_{\vartheta^*}^{-\frac{1}{2}}(\partial_i V_{\vartheta^*})V_{\vartheta^*}^{-1}(\partial_j V_{\vartheta^*})V_{\vartheta^*}^{-\frac{1}{2}}\right]\right)_{ij}.$$

Consequently, we can write

$$\mathbb{E}_{\vartheta_0}\left[\operatorname{tr}\left(V_{\widehat{\vartheta}^n}^{-1}D(\widehat{\vartheta}^n)\right)\right] \approx \frac{1}{2(n-N(\Theta)-d-1)}\operatorname{tr}\left(J_1(\vartheta^*)\Xi(\vartheta^*)\right) \tag{3.28}$$

Now we can combine all these considerations and plug them into (3.25), starting with (3.28):

$$\begin{aligned}\mathbb{E}_{\vartheta_0}\left[\mathcal{L}(\widehat{\vartheta}^n(Y^n), Y^n)\right] &\approx \mathbb{E}_{\vartheta_0}\left[\log(\det(V_{\widehat{\vartheta}^n}))\right] + \mathbb{E}\operatorname{tr}(V_{\widehat{\vartheta}^n}^{-1}\mathbb{E}_{\vartheta_0}\left[\epsilon_{\vartheta^*,1}\epsilon_{\vartheta^*,1}^T\right]) \\ &\quad + \frac{1}{2(n-N(\Theta)-d-1)}\operatorname{tr}\left(J_1(\vartheta^0)\Xi(\vartheta^*)\right) \\ &\overset{(3.26)}{\approx} \mathbb{E}_{\vartheta_0}\left[\log(\det(V_{\widehat{\vartheta}^n}))\right] + \frac{n}{n-N(\Theta)-d-1}\operatorname{tr}(I_d) \\ &\quad + \frac{1}{2(n-N(\Theta)-d-1)}\operatorname{tr}\left(J_1(\vartheta^*)\Xi(\vartheta^*)\right) \\ &= \mathbb{E}_{\vartheta_0}\left[\log(\det(V_{\widehat{\vartheta}^n}))\right] + \frac{dn}{n-N(\Theta)-d-1} \\ &\quad + \frac{1}{2(n-N(\Theta)-d-1)}\operatorname{tr}\left(J_1(\vartheta^*)\Xi(\vartheta^*)\right)\end{aligned}$$

The first term can be approximated by dropping the expectation, so that we arrive at the following, modified (hence the M) version of the AIC:

$$\begin{aligned}\mathrm{AICM}_n(\Theta) &= \log(\det(V_{\widehat{\vartheta}^n})) \\ &\quad + \frac{n}{n - N(\Theta) - d - 1}\left(d + \frac{1}{2}\operatorname{tr}\left(J_1(\vartheta^*)\Xi(\vartheta^*)\right)\right). \end{aligned} \tag{3.29}$$

Looking at this result it remains to answer the question how one can obtain estimators for $V_{\widehat{\vartheta}^n}$, $J_1(\vartheta^*)$ and $\Xi(\vartheta^*)$. Since $\Xi(\vartheta^*) = \mathcal{J}^{-1}(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{J}^{-1}(\vartheta^*)$, seeking an estimator for it really comes down to estimating $\mathcal{I}(\vartheta^*)$ and $\mathcal{J}(\vartheta^*)$. This can be done by the methods explained at the end of Subsection 3.4.1. $V_{\widehat{\vartheta}^n}$ can be estimated by

$$\mathbb{E}\left[\epsilon_{\widehat{\vartheta}^n}\epsilon_{\widehat{\vartheta}^n}^T\right] \approx \frac{1}{n}\sum_{k=1}^{n}\epsilon_{\widehat{\vartheta}^n,k}\epsilon_{\widehat{\vartheta}^n,k}^T,$$

from which we can deduce that an estimate of $\log(\det(V_{\widehat{\vartheta}^n}))$ is given by

$$\log\left(\det\left(\frac{1}{n}\sum_{k=1}^{n}\epsilon_{\widehat{\vartheta}^n,k}\epsilon_{\widehat{\vartheta}^n,k}^T\right)\right).$$

For $J_1(\vartheta^*)$, recall its definition

$$J_1(\vartheta^*) = 2\mathbb{E}_{\vartheta_0}\left[\nabla_\vartheta \epsilon_{\vartheta^*,1}^T V_{\vartheta^*}^{-1}\nabla_\vartheta \epsilon_{\vartheta^*,1}\right]$$

Replacing the expectation by the arithmetic mean, the pseudo-true parameter by the maximum likelihood estimator and the pseudo-innovations by their empirical counterpart, we have an estimator

$$\widehat{J}_1^n = 2\frac{1}{n}\sum_{k=1}^{n}\left(\nabla_\vartheta \widehat{\epsilon}_{\widehat{\vartheta}^n,k}^T V_{\widehat{\vartheta}^n}^{-1}\nabla_\vartheta \widehat{\epsilon}_{\widehat{\vartheta}^n,k}\right).$$

For the calculation of the gradient of the empirical pseudo-innovations we can rely on the fact that Kalman filter can not only be used to evaluate the innovations themselves, but also their derivatives.

3.4.4. PROPERTIES OF THE MODIFIED AIC

As we have seen in Theorem 3.15, both the CAIC and AIC derived in Subsection 3.4.1 suffer from the problem of overfitting if our candidate spaces include spaces in which the true one is nested: Even if the amount of observations becomes very large (i. e.

for $n \to \infty$), the probability of choosing a model with too many parameters is strictly positive. In this section, we want to investigate whether this effect is alleviated when using the criterion AICM or if it is still present. The answer is given in the next proposition:

Proposition 3.17. *Let Θ and Θ_0 be parameter spaces with associated families of continuous-time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta_0}$ and $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$, respectively, satisfying Assumption B. Assume that there is a $\vartheta_0 \in \Theta_0$ with $MCARMA(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ and Θ_0 is nested in Θ with map F and ϑ^* is the pseudo–true parameter in Θ.*

Assume that the driving Lévy process of Y is a Brownian motion. Then it holds:

$$\lim_{n\to\infty} \mathbb{P}\left(AICM_n(\Theta) < AICM_n(\Theta_0)\right) = \mathbb{P}\left(\chi^2_{d(N(\Theta)-N(\Theta_0))} > d(N(\Theta) - N(\Theta_0)\right),$$

where $\chi^2_{d(N(\Theta)-N(\Theta_0))}$ denotes a chi-squared random variable with $d(N(\Theta) - N(\Theta_0))$ degrees of freedom.
This especially implies that $AICM_n$ is neither strongly nor weakly consistent.

Proof. Denote the maximum likelihood estimator in Θ by $\widehat{\vartheta}^n$ and the maximum likelihood estimator in Θ_0 by $\widehat{\vartheta}^n_0$, respectively. The probability we are interested in can then be rewritten as

$$\mathbb{P}\left(\log\left(\frac{nV_{\widehat{\vartheta}^n}}{nV_{\widehat{\vartheta}^n_0}}\right) < \frac{dn}{n - N(\Theta_0)} - \frac{dn}{n - (N(\Theta_0) + N(\Theta) - N(\Theta_0))} + \frac{d_1}{2\left(n - N(\Theta_0)\right)} - \frac{d_2}{2\left(n - (N(\Theta_0) + N(\Theta) - N(\Theta_0))\right)}\right). \quad (3.30)$$

Here we have defined implicitly the constants $d_1 = \operatorname{tr}\left(J_1(\vartheta^*)\Xi(\vartheta^*)\right)$ and $d_2 = \operatorname{tr}\left(J_1(\vartheta_0)\Xi(\vartheta_0)\right)$. To make notation a bit more convenient in the future, we introduce the abbreviations $c_1 := N(\Theta)$ and $c_2 := N(\Theta_0) + N(\Theta) - N(\Theta_0)$.
By the assumption on the driving Lévy process, it holds that $\epsilon_{\vartheta,n} \sim \mathcal{N}(0, V_\vartheta)$ for every $n \in \mathbb{Z}$, i. e. the innovations are now normally distributed. Hence, as in the derivation of the AIC, we notice that $nV_{\widehat{\vartheta}^n_0}$ has a $\mathcal{W}(V_{\vartheta_0}, n - N(\Theta_0))$ distribution. Likewise, $nV_{\widehat{\vartheta}^n}$ is $\mathcal{W}(V_{\vartheta_0}, n - N(\Theta))$-distributed. Note that the two mean matrices are the same, since the innovations of the process associated to ϑ^* and ϑ_0, which are the parameters the maximum likelihood estimators converge to, are the same, hence their covariance matrix (which is just the mean matrix) is the same as well. By this

we obtain

$$\frac{\det\left(nV_{\widehat{\vartheta}^n}\right)}{\det\left(nV_{\widehat{\vartheta}_0^n}\right)} \stackrel{\mathcal{D}}{=} \frac{\det(\mathcal{W}(V_{\vartheta_0}, n-c_2))}{\det(\mathcal{W}(V_{\vartheta_0}, n-c_2)+\mathcal{W}(V_{\vartheta_0}, N(\Theta)-N(\Theta_0)))}, \tag{3.31}$$

where the Wishart variables on the right-hand side are independent.
By [Anderson 2003, Theorem 8.4.1] it then holds

$$\begin{aligned}
&\frac{\det(\mathcal{W}(V_{\vartheta_0}, n-c_2))}{\det(\mathcal{W}(V_{\vartheta_0}, n-c_2)+\mathcal{W}(V_{\vartheta_0}, N(\Theta)-N(\Theta_0)))} \\
&\stackrel{\mathcal{D}}{=} \prod_{i=1}^{d} \beta\left(\frac{n-c_2-i+1}{2}, \frac{N(\Theta)-N(\Theta_0)}{2}\right) \\
&\stackrel{\mathcal{D}}{=} \prod_{i=1}^{d} \frac{\chi^2_{n-c_2-i+1}}{\chi^2_{n-c_2-i+1}+\chi^2_{N(\Theta)-N(\Theta_0)}},
\end{aligned}$$

where $\beta(a,b)$ is the beta distribution with parameters a and b and the β variables are independent. The second equality holds because for independent random variables $X \sim \chi^2_{\theta_1}$ and $Y \sim \chi^2_{\theta_2}$ the ratio $\frac{X}{X+Y}$ has a beta distribution with parameters $\frac{\theta_1}{2}$ and $\frac{\theta_2}{2}$, i. e. the χ^2 variables can also be taken as independent.
As a consequence of this, it holds for the inverse quotient that

$$\frac{LV_{\widehat{\vartheta}_0^n}}{LV_{\widehat{\vartheta}^n}} \stackrel{\mathcal{D}}{=} \prod_{i=1}^{d}\left(1+\frac{\chi^2_{\overline{q}}}{\chi^2_{n-c_2-i+1}}\right). \tag{3.32}$$

Because $\log\left(\frac{nV_{\widehat{\vartheta}^n}}{nV_{\widehat{\vartheta}_0^n}}\right) = -\log\left(\frac{nV_{\widehat{\vartheta}_0^n}}{nV_{\widehat{\vartheta}^n}}\right)$ we can write (3.30) as follows:

$$\begin{aligned}
(3.30) &= \mathbb{P}\left(-\log\left(\frac{nV_{\widehat{\vartheta}_0^n}}{nV_{\widehat{\vartheta}^n}}\right) < \frac{-dn(N(\Theta)-N(\Theta_0))}{(n-c_1)(n-c_2)} - \frac{d_1(N(\Theta)-N(\Theta_0))}{2(n-c_1)(n-c_2)}\right. \\
&\quad \left. + \frac{d_2(N(\Theta)-N(\Theta_0))}{2(n-c_1)(n-c_2)}\right) \\
&\stackrel{(3.32)}{=} \mathbb{P}\left(-\log\left(\prod_{i=1}^{d}\left(1+\frac{\chi^2_{N(\Theta)-N(\Theta_0)}}{\chi^2_{n-c_2-i+1}}\right)\right) < \frac{-dn(N(\Theta)-N(\Theta_0))}{(n-c_1)(n-c_2)}\right. \\
&\quad \left. - \frac{d_1 n(N(\Theta)-N(\Theta_0))}{2(n-c_1)(n-c_2)} + \frac{d_2(N(\Theta)-N(\Theta_0))}{2(n-c_1)(n-c_2)}\right) \\
&= \mathbb{P}\left(-\sum_{i=1}^{d}\log\left(1+\frac{\chi^2_{N(\Theta)-N(\Theta_0)}}{\chi^2_{n-c_2-i+1}}\right) < \frac{-dn(N(\Theta)-N(\Theta_0))}{(n-c_1)(n-c_2)}\right. \\
&\quad \left. - \frac{d_1(N(\Theta)-N(\Theta_0))}{2(n-c_1)(n-c_2)} + \frac{d_2(N(\Theta)-N(\Theta_0))}{2(n-c_1)(n-c_2)}\right)
\end{aligned} \tag{3.33}$$

Before we let $n \to \infty$ we need three more facts: First off, we do a Taylor expansion of $\log(1+x)$ around 0, obtaining

$$\log(1+x) = \log(1) + \frac{1}{1+x}\bigg|_{x=0}(x-0) + o(|x|) = x + o(|x|).$$

Secondly, for every $i \in \{1, \dots, d\}$ the quotient $\frac{\chi^2_{n-c_2-i+1}}{L}$ converges to 1 almost surely for $n \to \infty$. This follows from the fact that a $\chi^2_{n-c_2-i+1}$-distributed random variable can be written as $\sum_{j=1}^{n-c_2-i+1} Z_j^2$, where the Z_j are i.i.d. $\mathcal{N}(0,1)$-distributed. The quotient $\frac{\chi^2_{n-c_2-i+1}}{n}$ then obeys the law of large numbers and converges to $\mathbb{E}Z_j^2 = 1$. Hence

$$\begin{aligned} -n\sum_{i=1}^{d} \log\left(1 + \frac{\chi^2_{N(\Theta)-N(\Theta_0)}}{\chi^2_{n-c_2-i+1}}\right) &= -n\sum_{i=1}^{d}\left(\frac{\chi^2_j}{\chi^2_{L-c_2-i+1}}\right) + o_p(1) \\ &\overset{n\to\infty}{\to} -\sum_{i=1}^{d} \chi^2_{N(\Theta)-N(\Theta_0)} \overset{\mathcal{D}}{=} \chi^2_{d\cdot(N(\Theta)-N(\Theta_0))}, \end{aligned}$$

since the d χ^2-variables in the sum are independent.
Lastly note that after multiplying both sides of the inequality by n, the first summand on the right-hand side of the inequality in (3.33) converges to $-d(N(\Theta) - N(\Theta_0))$ as $n \to \infty$ while the second and third summand tend to 0, which can be seen by regarding the numerator and denominator of these two fractions as polynomials in the variable n.
Eventually we are now able to combine these considerations and obtain

$$\begin{aligned} (3.33) &\overset{n\to\infty}{\to} \mathbb{P}\left(-\chi^2_{N(\Theta)-N(\Theta_0)} < -d(N(\Theta) - N(\Theta_0))\right) \\ &= \mathbb{P}\left(\chi^2_{d(N(\Theta)-N(\Theta_0))} > d(N(\Theta) - N(\Theta_0))\right) \end{aligned}$$

and the proof is complete.
The fact that AICM_n is not consistent immediately follows from the fact that the overfitting probability asymptotically is non–zero, which we have just shown (cp. Remark 3.11d)). □

3.4.5. A BOOTSTRAP VARIANT OF AIC

In this section, we will explore another method of approximating the Kullback–Leibler discrepancy based on bootstrap methods. Remember that throughout Subsection 3.4.1, the motivation was to estimate $\min_{\vartheta\in\Theta} K(f_\vartheta \mid f_{\vartheta_0})$ as defined in (3.15) or, equivalently, $\min_{\vartheta\in\Theta} \Delta(f_\vartheta \mid f_{\vartheta_0})$ as defined in (3.17). We then approximated $1/n$ times this quantity by $\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), Y^n) \mid Y^n\right]$ in (3.18). However, it then turned out that $\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n), Y^n)$ was a biased estimator of this expectation in Theorem 3.13, since we had to subtract a penalty term to obtain an asymptotically unbiased estimator. In defining the CAIC, we then replaced this penalty term by $\frac{2N(\Theta)}{n}$, motivated by the fact that the penalty term is equal to this in the case that the driving Lévy process is a Brownian motion.

The deciding observation is that we obtained an asymptotic result, i.e. if n is large in comparison to $N(\Theta)$, the approach of Subsection 3.4.1 is justified. However, in small samples, this is not necessarily the case as observed by Hurvich and Tsai [1989]. The authors point out that the CAIC is substantially negatively biased in those cases, where small samples are characterized by $N(\Theta) \approx 2n$ for the largest space under consideration. To remedy this, Cavanaugh and Shumway [1997] propose a bootstrap–based variant of CAIC. The authors then show that their criterion, which they call AICb, is asymptotically equivalent to the CAIC and illustrate via simulation studies the better performance in small samples. We follow the ideas of Cavanaugh and Shumway [1997], taking advantage of the fact that they develop their theory in the framework of discrete–time state space models with not necessarily Gaussian noise and use the Gaussian QMLE. This will imply that their theoretical results essentially immediately carry over to our framework of MCARMA processes as we will see in the following.

We operate in the usual framework of this chapter, i.e. $Y^n = (Y(h), \ldots, Y(nh))$ is a sample of length n, containing equidistant observations of realizations of the data–generating MCARMA process $(Y(t))_{t\in\mathbb{R}}$. We also consider a parameter space Θ containing models in Echelon form that fulfills Assumption B and may or may not contain a ϑ_0 with $Y = \text{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$. To define AICb, we now need bootstrap replicates of $\widehat{\vartheta}^n$, the QMLE based on Y^n. To achieve this, just as Cavanaugh and Shumway [1997], we use the bootstrap algorithm proposed by Stoffer and Wall [1991], which is specifically tailored to the discrete–time state space framework. It proceeds in the following steps ([Stoffer and Wall 1991, p. 1025]):

Step 1: Using Y^n, calculate the Gaussian QMLE as in (2.19) and the set of standardized approximate pseudo–innovations $\{V_{\widehat{\vartheta}^n}^{-\frac{1}{2}}\widehat{\epsilon}_{\widehat{\vartheta}^n,1}, \ldots, V_{\widehat{\vartheta}^n}^{-\frac{1}{2}}\widehat{\epsilon}_{\widehat{\vartheta}^n,n}\}$, using the Kalman

filter from Subsection 2.2.3.

Step 2: Draw a sample of size n with replacement from the set of standardized innovations calculated in step 1 to obtain a set of resampled innovations $\{\widehat{\epsilon}^*_{\widehat{\vartheta}^n,1}, \ldots, \widehat{\epsilon}^*_{\widehat{\vartheta}^n,n}\}$.

Step 3: Using (2.11), (2.12) and (2.17) with $\vartheta = \widehat{\vartheta}^n$ and $V^{\frac{1}{2}}_{\widehat{\vartheta}^n}\widehat{\epsilon}^*_{\widehat{\vartheta}^n,k}$ in place of $\widehat{\epsilon}_{\vartheta,k}$, calculate resampled observations $Y^n_* = (Y_*(h), \ldots, Y_*(nh))$. The initial value used in the initialization of the Kalman filter stays fixed.

Step 4: Using the observations Y^n_*, calculate the Gaussian QMLE $\widehat{\vartheta}^n_*$.

Step 5: For a number $b \in \mathbb{N}$ of bootstrap replications, repeat steps 2 to 4 to obtain a set of bootstrap replicates $\{\widehat{\vartheta}^n_*(i), 1 \leq i \leq b\}$.

In the appendix of Stoffer and Wall [1991], it is shown that the relative frequency distribution of the $\widehat{\vartheta}^n_*(i)$ behaves for $n, b \to \infty$ the same as the distribution of $\widehat{\vartheta}^n$ for $n \to \infty$, i.e. the two estimators are asymptotically equivalent. To show this, results of Lennart and Caines [1980] are used, which hold in our context as well as a consequence of Assumption B (cf. Proposition 2.25). We explicitly remark that neither Gaussianity nor independence of the true innovations are required. An immediate consequence is that the results on the bootstrap estimator also hold for our situation.
We can now define the criterion AICb analogous to Cavanaugh and Shumway [1997] by

$$\begin{aligned} \mathrm{AICb}_n(\Theta) &= \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + \frac{2\left(\frac{1}{b}\sum_{i=1}^b \widehat{\mathcal{L}}(\widehat{\vartheta}^n_*(i), Y^n) - \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n)\right)}{n} \\ &= -\widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + \frac{2}{nb}\sum_{i=1}^b \widehat{\mathcal{L}}(\widehat{\vartheta}^n_*(i), Y^n). \end{aligned} \tag{3.34}$$

From the first line, we can see that AICb is built in the same way as other information criteria we studied before: the pseudo Gaussian log–likelihood function, evaluated at the QMLE, is penalized by an additional term. Note that, in contrast to Cavanaugh and Shumway [1997], we divide the term $\frac{2}{b}\sum_{i=1}^b \widehat{\mathcal{L}}(\widehat{\vartheta}^n_*(i), Y^n)$ by the sample size n. The reason is that our definition of $\widehat{\mathcal{L}}$ already incorporates a factor $\frac{1}{n}$, which is not the case in Cavanaugh and Shumway [1997]. Model selection is then again done by comparing the values of AICb_n for different spaces Θ and choosing the one which attains the lowest value. With the terminology of Subsection 3.4.1, specifically (3.18), we have the following property of the AICb:

Proposition 3.18. *For a parameter space Θ that satisfies Assumption B, $AICb_n(\Theta)$ is an asymptotically almost surely unbiased estimator of $\mathbb{E}_{\vartheta_0}\left[\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\mid Y^n\right]\right]$, i.e. as first $b\to\infty$ and then $n\to\infty$ it holds that*

$$|AICb_n(\Theta)-\mathbb{E}_{\vartheta_0}\left[\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\mid Y^n\right]\right]|\longrightarrow 0.$$

Proof. The results of [Cavanaugh and Shumway 1997, Section 2], using [Cavanaugh and Shumway 1997, Lemma 1-3], are directly applicable (keeping the difference in normalization of $\widehat{\mathcal{L}}$ in mind). To prove their lemmas, they need regularity assumptions on the parametrization ([Cavanaugh and Shumway 1997, p. 489]). These assumptions are satisfied in our case, since we assume that the parameter space Θ satisfies Assumption B. Moreover, Cavanaugh and Shumway [1997] use the asymptotic results of Lennart and Caines [1980]. Namely, results on strong consistency and asymptotic normality of the QMLE in discrete–time state space models are needed. These are true in our context as well (cf. Proposition 2.25 and Theorem 2.28). These results are also the results that are used by Stoffer and Wall [1991] in their asymptotic justification of the bootstrap procedure (cf. [Stoffer and Wall 1991, Appendix]), upon which the AICb is fundamentally built.
From the results on [Cavanaugh and Shumway 1997, pp. 475-476] we receive that

$$\begin{aligned}&\left|2\left(\frac{1}{bn}\sum_{i=1}^{b}\widehat{\mathcal{L}}(\widehat{\vartheta}^n_*(i),Y^n)-\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),Y^n)\right)\right.\\&\left.-\mathbb{E}_{\vartheta_0}\left[\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\mid Y^n\right]\right]+\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\right]\right|\to 0\quad\mathbb{P}\text{-a.s.}\end{aligned}$$

as $b\to\infty$ and then $n\to\infty$. We can also write

$$\begin{aligned}&|\mathrm{AICb}_n(\Theta)-\mathbb{E}_{\vartheta_0}\left[\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\mid Y^n\right]\right]|\\&=\left|\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),Y^n)+2\left(\frac{1}{bn}\sum_{i=1}^{b}\widehat{\mathcal{L}}(\widehat{\vartheta}^n_*(i),Y^n)-\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),Y^n)\right)\right.\\&\quad\left.-\mathbb{E}_{\vartheta_0}\left[\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\mid Y^n\right]\right]\right|\\&\le\left|2\left(\frac{1}{bn}\sum_{i=1}^{b}\widehat{\mathcal{L}}(\widehat{\vartheta}^n_*(i),Y^n)-\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),Y^n)\right)\right.\\&\quad\left.-\mathbb{E}_{\vartheta_0}\left[\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\mid Y^n\right]\right]+\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\right]\right|\\&\quad+\left|\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),Y^n)-\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\right]\right|\end{aligned}$$

Moreover, as $n\to\infty$, we have that $\widehat{\vartheta}^n(Y^n)\to\vartheta^*$ $\mathbb{P}$-a.s. and also $\mathcal{L}(\widehat{\vartheta}^n(Y^n),Y^n)\to$

$\mathscr{Q}(\vartheta^*)$ $\mathbb{P}$-a.s. (cp. [Schlemm and Stelzer 2012, p. 2201]). Using this as well as Lemma 2.23a) and the continuity of the function $\mathscr{Q}$, we obtain

$$\begin{aligned}
&\lim_{n\to\infty}\left|\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),Y^n)-\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\right]\right|\\
&\leq \lim_{n\to\infty}\left|\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),Y^n)-\mathscr{Q}(\vartheta^*)\right|+\lim_{n\to\infty}\left|\mathscr{Q}(\vartheta^*)-\mathbb{E}_{\vartheta_0}\left[\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),\mathcal{Y}^n)\right]\right|\\
&= \lim_{n\to\infty}\left|\widehat{\mathcal{L}}(\widehat{\vartheta}^n(Y^n),Y^n)-\mathscr{Q}(\vartheta^*)\right|+\lim_{n\to\infty}\left|\mathscr{Q}(\vartheta^*)-\mathscr{Q}(\widehat{\vartheta}^n(Y^n))\right|=0\ \mathbb{P}\text{-a.s.}
\end{aligned}$$

Therefore, the assertion of the proposition follows. □

A consequence of this, as pointed out on [Cavanaugh and Shumway 1997, p. 477], is that asymptotically, AICb_n performs the same as AIC_n and CAIC_n, i.e. for very large n it does not matter which of these criteria is used, because the results will then be the same. It should be noted that AICb_n is computationally much more expensive than the other two criteria, because for each bootstrap replication a nonlinear optimization problem has to be solved numerically to obtain $\widehat{\vartheta}^n_*$. Therefore, if n is large, usage of AICb_n is probably not recommendable. However, it is a criterion specifically designed for small–sample situations. Cavanaugh and Shumway [1997] present convincing results of simulation studies which confirm that AICb_n is indeed superior to the other AIC–type criteria in small samples. We will do the same in Section 3.6 and perform a simulation study which shows that this also holds true in our context, which justifies AICb_n also from a practical point of view.
Another bootstrap–based information criterion called WIC is introduced in the unpublished research memorandum Ishiguro and Sakamoto [1991] and applied in Ishiguro et al. [1991] to an aperture imaging synthesis problem. It is defined similarly to AICb_n via

$$\mathrm{WIC}_n(\Theta)=\widehat{\mathcal{L}}(\widehat{\vartheta}^n,Y^n)+\frac{2\left(\frac{1}{b}\sum_{i=1}^b\widehat{\mathcal{L}}(\widehat{\vartheta}^n_*(i),Y^n)-\widehat{\mathcal{L}}(\widehat{\vartheta}^n_*(i),Y^n_*)\right)}{n}$$

However, the authors give no explanation why they define their criterion in this particular way. This is in contrast to the AICb, for which it is shown in Cavanaugh and Shumway [1997] that its penalty term serves as an estimator of a bias correction term and the underlying idea is to approximate the Kullback–Leibler discrepancy. Moreover, in our simulation experiment with small sample size we observed no satisfying performance of WIC_n, i.e. the overfitting rate was very high compared to the other criteria and especially AICb_n. Hence, since it is both better founded theoretically and performed better in the simulation study, we recommend the use of AICb_n and only mention WIC_n here briefly for sake of completeness.

3.5. BIC FOR MULTIVARIATE CARMA PROCESSES

3.5.1. DERIVATION OF THE BIC

Besides the AIC, the Bayesian Information Criterion (BIC) is probably the second most well–known and applied "information criterion" in the literature. Note that we have put the term "information criterion" in quotation marks, because, strictly speaking, it is not based on an information theoretic approach and therefore not truly an information criterion as the AIC is. Instead, as the name already suggests, it is based on an approach via Bayesian statistics. It is sometimes also called SIC, an abbreviation for Schwarz Information Criterion, named after the author who originally introduced it in Schwarz [1978]. Another often-cited article in this context is Rissanen [1978], which introduces an equivalent criterion in a slightly different context based on coding theory.
In this section, we shall largely take the same approach as in Cavanaugh and Neath [1999] to apply the ideas from the context of Bayesian statistics to develop the BIC. Ultimately, this will then again lead to a criterion that fits into the framework of our likelihood–based information criteria. As before, we assume that we have n equidistant, discrete–time observations of an MCARMA process Y, contained in the sample Y^n. We operate with parameter spaces Θ which contain continuous-time state space model in the Echelon form and satisfy Assumption B. Each parameter space is again characterized by its unique vector of Kronecker indices and its numbers of free parameters as illustrated in Section 3.1. We allow the spaces to be nested in the sense of Definition 3.8, but do not necessarily require this.
The core idea behind the BIC is that we want to choose the parameter space that is the most probable one given the data at hand. To make this explicit, suppose that π is a discrete prior probability distribution over the set of candidate spaces. The only assumption about π is that it assigns a positive probability to each of the spaces, which is standard in Bayesian theory. Moreover, suppose that $g(\cdot \mid \Theta)$ is a prior probability distribution over the parameter space Θ. For g we assume that it is bounded and bounded away from zero in a neighborhood of the pseudo–true parameter:

Assumption C. For every space Θ there exist two constants b and B with $0 < b \leq B < \infty$ such that $0 \leq g(\vartheta \mid \Theta) \leq B$ for all $\vartheta \in \Theta$ and $b \leq g(\vartheta \mid \Theta)$ for all $\vartheta \in N_0(\vartheta^*)$, where $N_0(\vartheta^*)$ denotes a neighborhood of the pseudo–true parameter ϑ^* contained in Θ.

Now we can apply Bayes' theorem to obtain the joint posterior probability distri-

bution f of Θ and ϑ:

$$f(\Theta, \vartheta \mid Y^n) = \frac{\pi(\Theta) g(\vartheta \mid \Theta) f(Y^n \mid \Theta, \vartheta)}{h(Y^n)}, \tag{3.35}$$

where $h(\cdot)$ denotes the (unknown) marginal density of Y^n. With this, we can calculate the a posteriori probability of the space Θ as

$$\mathbb{P}(\Theta \mid Y^n) = \int_{\Theta} f(\Theta, \vartheta \mid Y^n) \mathrm{d}\vartheta. \tag{3.36}$$

We can now make the notion of "choosing the most probable model for the data at hand" precise: Choose the space Θ which maximizes (3.36). Similar to the derivation of the AIC, the task is now to find a good approximation of (3.36) which is directly calculable from the data.
For this note first that maximization of (3.36) is equivalent to minimizing $-2/n$ times the logarithm of $\mathbb{P}(\Theta \mid Y^n)$. Applying this transformation and plugging in (3.35) gives

$$\begin{aligned} -\frac{2}{n} \log\left(\mathbb{P}(\Theta \mid Y^n)\right) = \frac{2}{n} \log(h(Y^n)) - \frac{2}{n} \log(\pi(\Theta)) \\ - \frac{2}{n} \log\left(\int_{\Theta} f(Y^n \mid \Theta, \vartheta) g(\vartheta \mid \Theta) \mathrm{d}\vartheta \right). \end{aligned} \tag{3.37}$$

We choose the parameter space Θ with the lowest value of $-\frac{2}{n} \log\left(\mathbb{P}(\Theta \mid Y^n)\right)$. Hence, we have to approximate this expression. For this, we approximate the unknown density $f(Y^n \mid \Theta, \vartheta)$ by the pseudo-Gaussian likelihood function, which we denote by $\not{L}$, and establish an upper and a lower bound for the logarithm of the integral on the right-hand side of (3.37). It will turn out that we have, under the condition that Y^n is known,

$$\begin{aligned} \mathcal{L}(\widehat{\vartheta}^n, Y^n) + N(\Theta) \frac{\log(n)}{n} + \frac{R_1(\Theta)}{n} \leq -\frac{2}{n} \log\left(\int \not{L}(\vartheta \mid Y^n) g(\vartheta \mid \Theta) \mathrm{d}\vartheta \right) \\ \leq \mathcal{L}(\widehat{\vartheta}^n, Y^n) + N(\Theta) \frac{\log(n)}{n} + \frac{R_2(\Theta)}{n} \end{aligned} \tag{3.38}$$

where $N(\Theta)$ is, as before, the number of free parameters in Θ and $R_1(\Theta)$ and $R_2(\Theta)$ are rest terms which do not depend on n. Since $2\log(\pi(\Theta))$ does not depend on n as well, we can write

$$-\frac{2}{n} \log\left(\mathbb{P}(\Theta \mid Y^n)\right) = \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + N(\Theta) \frac{\log(n)}{n} + \left[\frac{2}{n} \log(h(Y^n)) + O\left(\frac{\log(n)}{n} \right) \right].$$

Moreover, the term $2\log(h(Y^n))$ is the same across all models and therefore not relevant for model selection. Hence, we drop it from the approximation and choose the model that minimizes the information criterion defined as

$$\mathrm{BIC}_n(\Theta) := \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + N(\Theta)\frac{\log(n)}{n}. \tag{3.39}$$

In order to establish the central inequalities from (3.38), we use two lemmas which we will state in the following. The first one is the analog of [Cavanaugh and Neath 1999, Lemma 1]: Take in mind that the idea is that Y^n is known and hence $\mathcal{L}(\widehat{\vartheta}^n, Y^n)$ is deterministic.

Lemma 3.19. *For any Y^n there exist positive constants λ_1 and λ_2 and an integer n_1 such that the following holds for every ϑ in a neighborhood $N(\vartheta^*)$ of the pseudo–true parameter ϑ^* and for every $n > n_1$:*

$$\begin{aligned}
&\frac{1}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) + \frac{\lambda_2}{2}(\vartheta - \widehat{\vartheta}^n)^T(\vartheta - \widehat{\vartheta}^n) \\
\leq &\frac{1}{2}\mathcal{L}(\vartheta, Y^n) \\
\leq &\frac{1}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) + \frac{\lambda_1}{2}(\vartheta - \widehat{\vartheta}^n)^T(\vartheta - \widehat{\vartheta}^n).
\end{aligned}$$

Proof. Let Y^n be known and fixed. We start by taking a second-order Taylor expansion of $\frac{1}{2}\mathcal{L}$ around $\widehat{\vartheta}^n$, giving

$$\begin{aligned}
\frac{1}{2}\mathcal{L}(\vartheta, Y^n) &= \frac{1}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) + (\vartheta - \widehat{\vartheta}^n)^T\frac{1}{2}\nabla_\vartheta\mathcal{L}(\widehat{\vartheta}^n, Y^n) \\
&\quad + \frac{1}{2}(\vartheta - \widehat{\vartheta}^n)^T\frac{1}{2}\nabla^2_\vartheta\mathcal{L}(\overline{\vartheta}^n, Y^n)(\vartheta - \widehat{\vartheta}^n) \\
&= \frac{1}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) + \frac{1}{2}(\vartheta - \widehat{\vartheta}^n)^T\frac{1}{2}\nabla^2_\vartheta\mathcal{L}(\overline{\vartheta}^n, Y^n)(\vartheta - \widehat{\vartheta}^n),
\end{aligned} \tag{3.40}$$

where $\overline{\vartheta}^n$ is between ϑ and $\widehat{\vartheta}^n$ in the sense of the Euclidean norm on the parameter space Θ (i. e. it can be written as $\overline{\vartheta}^n = t\vartheta + (1-t)\widehat{\vartheta}^n$ for some $t \in [0,1]$) and we have used that $\widehat{\vartheta}^n$ is the point where $\mathcal{L}$ attains its minimum, implying $\nabla_\vartheta\mathcal{L}(\widehat{\vartheta}^n, Y^n) = 0$. Next, denote by $\lambda_n^{\max}(\vartheta)$ and $\lambda_n^{\min}(\vartheta)$ the largest and smallest eigenvalue of the matrix $\frac{1}{2}\nabla^2_\vartheta\mathcal{L}(\vartheta, Y^n)$, respectively. By the general theory on quadratic forms we have the following chain of inequalities:

$$\begin{aligned}
(\vartheta - \widehat{\vartheta}^n)^T(\vartheta - \widehat{\vartheta}^n)\lambda_n^{\min}(\overline{\vartheta}^n) &\leq (\vartheta - \widehat{\vartheta}^n)^T\frac{1}{2}\nabla^2_\vartheta\mathcal{L}(\overline{\vartheta}^n, Y^n)(\vartheta - \widehat{\vartheta}^n) \\
&\leq (\vartheta - \widehat{\vartheta}^n)^T(\vartheta - \overline{\vartheta}^n)\lambda_n^{\max}(\overline{\vartheta}^n).
\end{aligned} \tag{3.41}$$

Because of the almost sure convergence of Hessian matrix of $\mathcal{L}$, as pointed out in Proposition 2.25d), we know that for every $\vartheta \in \Theta$ $\lambda_n^{\max}(\vartheta)$ and $\lambda_n^{\min}(\vartheta)$ converge $\mathbb{P}$-almost surely to the largest and smallest eigenvalue of $\frac{1}{2}\mathcal{J}(\vartheta)$, respectively, and the convergence is uniform in ϑ. Because of Assumption B.11 we know that $J(\vartheta^*)$ is non–singular and therefore, by continuity of the map $\vartheta \mapsto \mathcal{J}(\vartheta)$, the eigenvalues of $\mathcal{J}(\vartheta)$ lie between two positive constants if ϑ is in a neighborhood $N(\vartheta^*)$ of ϑ^*. This allows us to find a $n_2 \in \mathbb{N}$, a neighborhood $N(\vartheta^*)$ of ϑ^* and two constants λ_1, λ_2 with $0 < \lambda_2 < \lambda_1 < \infty$ such that

$$\lambda_2 \leq \inf_{n>n_2} \left\{ \inf_{\vartheta \in N(\vartheta^*)} \lambda_n^{\min}(\vartheta) \right\}$$

and

$$\lambda_1 \geq \sup_{n>n_2} \left\{ \sup_{\vartheta \in N(\vartheta^*)} \lambda_n^{\max}(\vartheta) \right\}.$$

If $\overline{\vartheta}^n$ is in $N(\vartheta^*)$, we can apply these bounds to (3.41), obtaining

$$\begin{aligned} 0 \leq (\vartheta - \widehat{\vartheta}^n)^T(\vartheta - \widehat{\vartheta}^n)\lambda_2 &\leq (\vartheta - \widehat{\vartheta}^n)^T \frac{1}{2} \nabla_\vartheta^2 \mathcal{L}(\overline{\vartheta}^n, Y^n)(\vartheta - \widehat{\vartheta}^n) \\ &\leq (\vartheta - \widehat{\vartheta}^n)^T(\vartheta - \widehat{\vartheta}^n)\lambda_1. \end{aligned} \tag{3.42}$$

It remains to argue that $\overline{\vartheta}^n$ indeed is contained in this neighborhood if L is large enough. To see this, remember that $\widehat{\vartheta}^n$ (regarded as a random variable) converges to ϑ^* $\mathbb{P}$-almost surely by Theorem 2.28. This means that $\widehat{\vartheta}^n \in N(\vartheta^*)$ for every n larger than some constant n_3. Since $\overline{\vartheta}^n$ is between ϑ and $\widehat{\vartheta}^n$, it is an element of the neighborhood $N(\vartheta^*)$ if the latter two are elements of it. If we now choose $n_1 = \max\{n_2, n_3\}$ the statement of the Lemma follows from these considerations and applying the inequalities from (3.42) to equation (3.40). □

The second lemma is a general one, it does not rely on the structure of a MCARMA process or special properties of the maximum likelihood method. Hence it is almost exactly the same as Lemma 2 in Cavanaugh and Neath [1999], only slightly altered to fit our notation better:

Lemma 3.20. *Let $(T_n)_{n\in\mathbb{N}}$ and $(U_n)_{n\in\mathbb{N}}$ be two sequences of positive random variables. Moreover, let $(\rho_n)_{n\in\mathbb{N}}$ be a positive, convergent sequence defined in such a way that $T_n \geq U_n$ holds whenever $U_n > \rho_n$. Finally, suppose there exist two positive constants γ and ϵ such that*

$$\mathbb{P}((T_n - \rho_n) \geq \gamma) \geq \epsilon \; \forall n \in \mathbb{N}.$$

Then there exists a positive integer n_4 such that for every $\delta > 0$ and $n > n_4$ it holds:

$$\log(\mathbb{E}(T_n^n)) - \log(\mathbb{E}(U_n^n)) > -\delta.$$

Proof. See cited reference. □

Now we have all that is required to establish (3.38). For this, we define

$$\begin{aligned}
Z_n(\vartheta) &:= \exp\left(-\frac{1}{2}\mathcal{L}(\vartheta, Y^n)\right), \\
X_{1,n}(\vartheta) &:= \exp\left(-\frac{1}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) - \frac{\lambda_1}{2}(\vartheta - \widehat{\vartheta}^n)^T(\vartheta - \widehat{\vartheta}^n)\right), \\
X_{2,n}(\vartheta) &:= \exp\left(-\frac{1}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) - \frac{\lambda_2}{2}(\vartheta - \widehat{\vartheta}^n)^T(\vartheta - \widehat{\vartheta}^n)\right), \\
X_1(\vartheta) &:= \exp\left(-\frac{1}{2}\mathscr{Q}(\vartheta^*) - \frac{\lambda_1}{2}(\vartheta - \vartheta^*)^T(\vartheta - \vartheta^*)\right), \\
Z(\vartheta) &:= \exp\left(-\frac{1}{2}\mathscr{Q}(\vartheta)\right),
\end{aligned}$$

with $\mathscr{Q}$ as defined in (2.16).
An application of Lemma 3.19 allows us to find a natural number n_1 and a neighborhood $N(\vartheta^*)$ of ϑ^* such that we have

$$X_{1,n}(\vartheta) \leq Z_n(\vartheta) \leq X_{2,n}(\vartheta) \tag{3.43}$$

for every $n > n_1$ and every $\vartheta \in N(\vartheta^*)$. We can also note that both X_1 and Z attain a global maximum at ϑ^* and the value of the maxima is the same. The convergence results from Proposition 2.25b) allow us to conclude

$$\begin{aligned}
X_{1,n}(\vartheta) \to X_1(\vartheta)\ \mathbb{P}\text{-almost surely, uniformly in } \vartheta \text{ and} \\
Z_n(\vartheta) \to Z(\vartheta)\ \mathbb{P}\text{-almost surely, uniformly in } \vartheta.
\end{aligned}$$

The plan is now to apply Lemma 3.20 to the random variables $U_n = X_{1,n}(\vartheta)$ and $T_n = Z_n(\vartheta)$. In order to be allowed to do that, we need to specify the sequence $(\rho_n)_{n\in\mathbb{N}}$ and check the existence of the constants γ and ϵ. Note that the relevant probability measure in this context is $g(\vartheta \mid \Theta)$: the observations Y^n are assumed to be fixed and known, the source of randomness is the parameter ϑ. We shall denote the corresponding probabilities by $\mathbb{P}^{\vartheta|\Theta}$ and expectations by $\mathbb{E}_{\vartheta|\Theta}$ in the following. We will now give an explicit construction of $(\rho_n)_{n\in\mathbb{N}}$ and γ:

First off, choose

$$\epsilon_n^{(1)} := \sup_{\vartheta\in\Theta} |X_{1,n}(\vartheta) - X_1(\vartheta)|,$$
$$\epsilon_n^{(2)} := \sup_{\vartheta\in\Theta} |Z_n(\vartheta) - Z(\vartheta)|,$$

and

$$\epsilon_n := \max\{\epsilon_n^{(1)}, \epsilon_n^{(2)}\}.$$

Note that these quantities do depend on Y^n, but are not random anymore since the observations are fixed.
Now choose a compact neighborhood $M(\vartheta^*)$ of ϑ^* such that $\emptyset \neq M(\vartheta^*) \subset N(\vartheta^*)$ and

$$\sup_{\vartheta\in N(\vartheta_0)^{\mathrm{C}}} X_1(\vartheta) < \sup_{\vartheta\in M(\vartheta_0)^{\mathrm{C}}} X_1(\vartheta) < \sup_{\vartheta\in M(\vartheta_0)^{\mathrm{C}}} Z(\vartheta) < Z(\vartheta^*) = X_1(\vartheta^*).$$

Note that this is possible because we have $X_1(\vartheta) < Z(\vartheta)$ for every $\vartheta \in \Theta \setminus \{\vartheta^*\}$, since $\mathscr{Q}$ attains a global minimum at ϑ^* by definition in (2.20).
Now let

$$2\gamma := \sup_{\vartheta\in M(\vartheta^*)^{\mathrm{C}}} X_1(\vartheta) - \sup_{\vartheta\in N(\vartheta_0)^{\mathrm{C}}} X_1(\vartheta) > 0,$$
$$\rho^* := \sup_{\vartheta\in N(\vartheta^*)^{\mathrm{C}}} X_1(\vartheta) < X_1(\vartheta^*)$$

and

$$\rho_n := \rho^* + \epsilon_n.$$

Note that ρ_n is positive for every $n \in \mathbb{N}$ and the sequence $(\rho_n)_{n\in\mathbb{N}}$ converges to ρ^*.
Suppose now that $X_{1,n}(\vartheta) > \rho_n$ for every n larger than some $n_1 \in \mathbb{N}$ and a fixed ϑ. Then:

$$X_1(\vartheta) = X_1(\vartheta) - X_{1,n}(\vartheta) + X_{1,n}(\vartheta) \geq -\epsilon_n + X_{1,n}(\vartheta) > -\epsilon_n + \rho_n = \rho^*.$$

Hence, $\vartheta \notin N(\vartheta^*)^{\mathrm{C}}$, and

$$\{\vartheta \in \Theta \mid X_{1,n}(\vartheta) > \rho_n \text{ for every } n \text{ larger than } n_1\} \subseteq N(\vartheta^*),$$

which in turn implies $X_{1,n}(\vartheta) \leq Z_n(\vartheta)$ via Equation (3.43). Hence the sequence

$(\rho_n)_{n\in\mathbb{N}}$ fulfills the first part of the conditions in Lemma 3.20. For the second part, consider $\vartheta \in M(\vartheta^*)$. Then, $Z(\vartheta) > \rho^* + 2\gamma = \sup_{\vartheta\in M(\vartheta^*)^C} X_1(\vartheta)$ and consequently

$$\begin{aligned} Z_n(\vartheta) &= Z_n(\vartheta) - Z(\vartheta) + Z(\vartheta) \geq -\epsilon_n + Z(\vartheta) \\ &> -\epsilon_n + \rho^* + 2\gamma = \rho_n + \gamma + \gamma - 2\epsilon_n \geq \rho_n + \gamma \end{aligned}$$

because we will eventually have $\epsilon_n < \frac{\gamma}{2}$ for every n larger than some constant $n_2 \in \mathbb{N}$ ($(\epsilon_n)_{n\in\mathbb{N}}$ converges to 0). This delivers

$$M(\vartheta^*) \subseteq \{\vartheta \in \Theta : Z_n(\vartheta) - \rho_n \geq \gamma\}.$$

Since $M(\vartheta^*)$ was chosen to be non-empty and compact, it has a strictly positive probability, meaning $\mathbb{P}^{\vartheta|\Theta}(Z_n(\vartheta) - \rho_n \geq \gamma) \geq \tilde{\epsilon}$ for some $\tilde{\epsilon} > 0$ and for every $n > n_2$. If n is now larger than $n_3 = \max\{n_1, n_2\}$, both requirements of Lemma 3.20 are fulfilled and for any $\delta_* > 0$ it holds

$$\log(\mathbb{E}_{\vartheta|\Theta}[(Z_n(\vartheta))^n]) - \log(\mathbb{E}_{\vartheta|\Theta}[(X_{1,n}(\vartheta))^n]) > -\frac{\delta_*}{2} \quad \forall n > n_3.$$

In a similar way (reversing the roles of $Z_n(\vartheta)$ and $X_{2,n}(\vartheta)$) we obtain the inequality

$$\log(\mathbb{E}_{\vartheta|\Theta}[(X_{2,n}(\vartheta))^n]) - \log(\mathbb{E}_{\vartheta|\Theta}[(Z_n(\vartheta))^n]) > -\frac{\delta_*}{2} \quad \forall n > n_4$$

by another application of Lemma 3.20. If n is now larger than $n_* = \max\{n_3, n_4\}$ we can combine these results to obtain the following chain of inequalities:

$$\begin{aligned} -2\log(\mathbb{E}_{\vartheta|\Theta}[(X_{2,n}(\vartheta))^L]) - \delta_* &< -2\log(\mathbb{E}_{\vartheta|\Theta}[(Z_n(\vartheta))^L]) \\ &< -2\log(\mathbb{E}_{\vartheta|\Theta}[(X_{1,n}(\vartheta))^L]) + \delta_*. \end{aligned} \tag{3.44}$$

Let us consider the middle term for a moment:

$$\begin{aligned} -2\log(\mathbb{E}_{\vartheta|\Theta}[(Z_n(\vartheta))^n]) &= -2\log\left(\int_\Theta \left(\exp\left(-\frac{1}{2}\mathcal{L}(\vartheta, Y^n)\right)\right)^n g(\vartheta \mid \Theta)\mathrm{d}\vartheta\right) \\ &= -2\log\left(\int \exp\left(-\frac{n}{2}\mathcal{L}(\vartheta, Y^n)\right) g(\vartheta \mid \Theta)\mathrm{d}\vartheta\right) \\ &= -2\log\left(\int \exp\left(-\frac{n}{2}(-\frac{2}{n}\log(\not{L}(\vartheta \mid Y^n)))\right) g(\vartheta \mid \Theta)\mathrm{d}\vartheta\right) \\ &= -2\log\left(\int \not{L}(\vartheta \mid Y^n) g(\vartheta \mid \Theta)\mathrm{d}\vartheta\right). \end{aligned}$$

This shows that the middle term, divided by n, in the above inequality (3.44) is the

same as the middle term in inequality (3.38), meaning we can establish the desired upper and lower bound for this term.
In the following, it is assumed that n_* is large enough to guarantee that $\widehat{\vartheta}^n \in N_0(\vartheta^*)$ for all $n > n_*$, where $N_0(\vartheta^*)$ denotes the neighborhood of ϑ^* from Assumption C, in which $g(\vartheta \mid \Theta)$ is bounded from below (which is not a restriction, if the condition is not met by the initial n_* we can just enlarge it further). We start with the term on the far left of the inequalities in (3.44):

$$\begin{aligned}
&-2\log(\mathbb{E}_{\vartheta|\Theta}[(X_{2,n}(\vartheta))^n]) \\
&= -2\log\left(\int_\Theta \left(\exp\left(-\frac{1}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) - \frac{\lambda_2}{2}\left(\vartheta - \widehat{\vartheta}^n\right)^T\left(\vartheta - \widehat{\vartheta}^n\right)\right)\right)^n \right. \\
&\left. \cdot\, g(\vartheta \mid \Theta)\mathrm{d}\vartheta\right) \\
&\geq -2\log\left(\int_{\mathbb{R}^{N(\Theta)}} B\exp\left(-\frac{n}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) - \frac{n\lambda_2}{2}\left(\vartheta - \widehat{\vartheta}^n\right)^T\left(\vartheta - \widehat{\vartheta}^n\right)\right)\mathrm{d}\vartheta\right) \\
&= n\mathcal{L}(\widehat{\vartheta}^n, Y^n) - 2\log(B) - 2\log\left(\left(\frac{2\pi}{n\lambda_2}\right)^{\frac{N(\Theta)}{2}}\int_{\mathbb{R}^{N(\Theta)}}\left(\frac{n\lambda_2}{2\pi}\right)^{\frac{N(\Theta)}{2}}\right. \\
&\left. \cdot\, \exp\left(-\frac{1}{2}\frac{\left(\vartheta - \widehat{\vartheta}^n\right)^T\left(\vartheta - \widehat{\vartheta}^n\right)}{\frac{1}{n\lambda_2}}\right)\mathrm{d}\vartheta\right) \qquad (3.45)
\end{aligned}$$

Note that the term remaining inside the integral is the density of a $N(\Theta)$-dimensional normal distribution with expectation $\widehat{\vartheta}^n$ and covariance matrix $\frac{1}{n\lambda_2}I_{N(\Theta)\times N(\Theta)}$, i. e. the whole integral is simply equal to 1. Hence we can write

$$\begin{aligned}
(3.45) &= n\mathcal{L}(\widehat{\vartheta}^n, Y^n) - 2\log(B) - 2\log\left(\left(\frac{n\lambda_2}{2\pi}\right)^{\frac{N(\Theta)}{2}}\right) \\
&= n\mathcal{L}(\widehat{\vartheta}^n, Y^n) - 2\log(B) + N(\Theta)\log(L) + N(\Theta)\log(\lambda_2) - N(\Theta)\log(2\pi),
\end{aligned}$$

which is exactly the form claimed in (3.38) if we denote

$$R_1(\Theta) := N(\Theta)\log(\lambda_2) - N(\Theta)\log(2\pi) - 2\log(B).$$

Note that all terms contained in this rest do not depend on the sample size n.
For the right-hand side of the inequalities in (3.44) we proceed in a similar way:

$$
\begin{aligned}
&- 2\log(\mathbb{E}_{\vartheta|\Theta}[(X_{1,n}(\vartheta))^n]) \\
&= -2\log\left(\int_\Theta \left(\exp\left(-\frac{1}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) - \frac{\lambda_1}{2}\left(\vartheta - \widehat{\vartheta}^n\right)^T\left(\vartheta - \widehat{\vartheta}^n\right)\right)\right)^n \right. \\
&\left. \cdot\, g(\vartheta \mid \Theta)\mathrm{d}\vartheta\right) \\
&\leq -2\log\left(\int_{N_0(\vartheta^*)} b\exp\left(-\frac{n}{2}\mathcal{L}(\widehat{\vartheta}^n, Y^n) - \frac{n\lambda_1}{2}\left(\vartheta - \widehat{\vartheta}^n\right)^T\left(\vartheta - \widehat{\vartheta}^n\right)\right)\mathrm{d}\vartheta\right) \\
&= n\mathcal{L}(\widehat{\vartheta}^n, Y^n) - 2\log(b) - 2\log\left(\left(\frac{2\pi}{n\lambda_1}\right)^{\frac{N(\Theta)}{2}}\int_{N_0(\vartheta^*)}\left(\frac{n\lambda_1}{2\pi}\right)^{\frac{N(\Theta)}{2}}\right. \\
&\left. \cdot\, \exp\left(-\frac{1}{2}\frac{\left(\vartheta - \widehat{\vartheta}^n\right)^T\left(\vartheta - \widehat{\vartheta}^n\right)}{\frac{1}{n\lambda_1}}\right)\mathrm{d}\vartheta\right) \qquad (3.46)
\end{aligned}
$$

Now the assumption about n_* comes into play, since it allows us to conclude that we have $\widehat{\vartheta}^n \in N_0(\vartheta^*)$ for $n > n_*$, which means that the integral appearing here is bounded below by some constant c (depending on $N_0(\vartheta^*)$ and λ_1, but not on n). This allows us to estimate

$$
\begin{aligned}
(3.46) &\leq n\mathcal{L}(\widehat{\vartheta}^n, Y^n) - 2\log(b) - 2\log\left(\left(\frac{2\pi}{n\lambda_1}\right)^{\frac{N(\Theta)}{2}}\right) - 2\log(c) \\
&= n\mathcal{L}(\widehat{\vartheta}^n, Y^n) + N(\Theta)\log(n) + R_2(\Theta),
\end{aligned}
$$

where $R_2(\Theta) := -2\log(b) - 2\log(c) + N(\Theta)\log(\lambda_1) - N(\Theta)\log(2\pi)$ is again a rest term independent of n. This completes the proof of the inequalities stated in Equation (3.38) and hence the derivation of the BIC as shown after said equation.

3.5.2. Consistency of the BIC

As is well–known from the literature, the BIC is a strongly consistent criterion in the case of ARMA processes as originally shown in Hannan [1980]. From our general results on IC_n, it follows easily that this is true in our context as well:

Theorem 3.21. *The BIC is a strongly consistent information criterion.*

Proof. Take $C(n) = \log(n)$ in (3.6). The assertion then follows from Theorem 3.10 and Remark 3.11a) since $\lim_{n\to\infty} \frac{\log(n)}{\log(\log(n))} = \infty$. □

3.6. Simulation study of order selection criteria

The results on information criteria obtained in the previous sections will now be illustrated by a simulation study. In this context we would like to thank Eckhard Schlemm and Robert Stelzer, who kindly provided the MATLAB code for the simulation and parameter estimation of the MCARMA process. As before, we use the Echelon MCARMA parametrization in the simulations. Throughout our simulations, we always consider two-dimensional MCARMA processes. As driving Lévy process, we use, on the one hand, a two-dimensional, correlated Brownian motion and, on the other hand, a two-dimensional, normal-inverse Gaussian (NIG) process. For the NIG process the increments $L(t) - L(t-1)$ have the density

$$f_{NIG}(x;\mu,\alpha,\beta,\delta,\Delta) = \frac{\delta\, \mathrm{e}^{\delta\kappa}}{2\pi} \frac{\mathrm{e}^{\langle \beta x\rangle}}{\mathrm{e}^{\alpha g(x)}} \frac{1+\alpha g(x)}{g(x)^3}, \quad x \in \mathbb{R}^2,$$

where $g(x) = \sqrt{\delta^2 + \langle x-\mu, \Delta(x-\mu)\rangle}$, $\kappa^2 = \alpha^2 - \langle \beta, \Delta\beta\rangle$. The parameter $\mu \in \mathbb{R}^2$ is a location parameter, $\alpha \geq 0$ is a shape parameter, $\beta \in \mathbb{R}^2$ is a symmetry parameter, $\delta \geq 0$ is a scale parameter and $\Delta \in \mathbb{R}^{2\times 2}$ is a positive semidefinite matrix with $\det(\Delta) = 1$ that determines the dependence between the components of the Lévy process. In the simulations we use the values

$$\delta = 1, \quad \alpha = 3, \quad \beta = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \Delta = \begin{pmatrix} \frac{5}{4} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}, \quad \mu = -\frac{1}{2\sqrt{31}} \begin{pmatrix} 3 \\ 2 \end{pmatrix},$$

which result in a zero-mean process with covariance matrix

$$\Sigma^L_{NIG} \approx \begin{pmatrix} 0.4571 & -0.1622 \\ -0.1622 & 0.3708 \end{pmatrix}.$$

In the case of the Brownian motion the covariance matrix Σ^L_{BM} is equal to the covariance matrix Σ^L_{NIG} in the NIG case. In the estimation the number of free parameters includes three parameters for the covariance matrix of the driving Lévy process.

The simulation of the continuous–time MCARMA process is done with the initial value $X(0) = 0$, applying the Euler-Maruyama method to the stochastic differential equation (2.5) and then evoking (2.4). For the Euler-Maruyama scheme we operate on the interval $[0, n]$, where n is the number of observations and the step size is 0.01. Afterwards, the simulated process is sampled at discrete points in time with sampling distance $h = 1$, resulting in n observations of the discretely sampled MCARMA

process. Figure 3.1 shows a typical sample path of a bivariate CARMA(2,1) process simulated in this way and Figure 3.2 shows a sample path for the same parameter configuration when the NIG process is used as driving Lévy process. After obtaining the discrete samples of the MCARMA process we calculate the AIC, CAIC and BIC as defined in (3.19), (3.20) and (3.39), respectively. In the calculation of the AIC we estimate the penalty term $\operatorname{tr}\left(\mathcal{I}(\vartheta^*)\mathcal{H}^{-1}(\vartheta^*)\right)$ by the second method explained at the end of Subsection 3.4.1 since in general there is no explicit form of $\mathcal{I}(\vartheta^*)$ and $\mathcal{H}(\vartheta^*)$. This means that we use $\operatorname{tr}(\widehat{\mathcal{J}}^n\widehat{\Xi}^n)$ as estimate, where $\widehat{\Xi}^n$ is the empirical covariance matrix of independent realizations of $\widehat{\vartheta}^n$ and $\widehat{\mathcal{J}}^n$ is the arithmetic mean of independent realizations of $\nabla^2_\vartheta\widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n)$.

For the first part of the study, we simulate a two-dimensional MCARMA process with Kronecker index $m_0 = (1,2)$, $p = 2$ and $q = 1$ with parameter $\vartheta_0^{(1)} = \Big(-1,-2,1,-2,-3,1,2\Big)$ and $n = 2000$. We consider eight different parameter spaces in total with $m_0 \in \{1,2\}^2$, $p \in \{1,2\}$ and $q \in \{0,1\}$. We observe that every information criterion makes the right choice of the parameter space in all 50 replications, independent of the driving Lévy process. There are no effects of overfitting, which is not surprising considering the fact that the true parameter is chosen in such a way that it is only contained in one space, so that the scenario from Remark 3.11d) is given. Next, we change the true parameter slightly to $\vartheta_0^{(2)} = \Big(-1,-2,1,-2,-3,0,0\Big)$, i.e. the data–generating process is now a MCARMA(2,0) process, while $m_0 = (1,2)$ remains the same. The results of 100 replications for the true parameter $\vartheta_0^{(2)}$ in space 2 are summarized in Table 3.3.

Space	Model				BM			NIG		
	m	p	q	$N(\Theta)$	AIC	CAIC	BIC	AIC	CAIC	BIC
1	$(1,1)$	1	0	7	0	0	0	0	0	0
2	$(1,2)$	2	0	8	92	85	100	89	84	100
3	$(1,2)$	2	1	10	8	15	0	11	16	0
4	$(2,1)$	2	0	9	0	0	0	0	0	0
5	$(2,1)$	2	1	11	0	0	0	0	0	0
6	$(2,2)$	2	0	11	0	0	0	0	0	0
7	$(2,2)$	2	1	15	0	0	0	0	0	0
Agreement					93%			95%		

Table 3.3.: Results for the true parameter $\vartheta_0^{(2)}$ in space 2.

As expected because of the strong consistency the BIC performs convincingly and achieves a perfect score for both driving Lévy processes. Furthermore, both versions of the AIC exhibit overfitting. The line "agreement" records the percentage of repetitions in which the CAIC and AIC lead to the same choice, revealing that there is an undeniable difference between the CAIC and the AIC in both cases. From the theory, we know that this should not happen when the driving Lévy process is a Brownian motion since the criteria are then the same. This difference comes from the estimation error by estimating the penalty term $\operatorname{tr}\left(\mathcal{I}(\vartheta^*)\mathcal{J}^{-1}(\vartheta^*)\right)$ in the AIC. We realize that in the Gaussian model the value of the penalty term in the AIC is higher than the value of the penalty term in the CAIC for both space number 2 and space number 3. However, the error is smaller in space number 3 than in space number 2, which results in a higher overfitting rate for the CAIC. This is because $\widehat{\mathcal{L}}(\widehat{\vartheta^n}, Y^n)$ is smaller in model 3 than in model 2 – this difference is compensated for more often in the AIC than in the CAIC because of the larger penalty terms, which leads to a lesser overfitting rate. We also calculate the overfitting probability in the Brownian motion case as given in Theorem 3.10c). For this, note that there is only one parameter space in which the true one is nested (space number 3) and for that space we have $C = 2$ and $N(\Theta) - N(\Theta_0) = 2$. The strictly positive eigenvalues of $\mathcal{H}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_F(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{M}_F(\vartheta^*)\mathcal{H}(\vartheta^*)^{\frac{1}{2}}$ are calculated with the help of MATLAB and turn out to be both equal to 2, so that the overfitting probability simplifies to $\mathbb{P}(\chi_1^2 > 2) \approx 0.1573$. The empirical probability 0.15 of overfitting in the CAIC is very close.

Next, we consider another situation in which the data–generating process is a MCARMA(3,0) process with Kronecker index $m_0 = (3, 2)$ and the true parameter is

$$\vartheta_0^{(3)} = \Big(-3, -6, -5, 2, -3, -0.2, -4, -2.5, -7, -9, 0, 0, 0, 0, 0, 0\Big).$$

Here, we consider 7 candidate spaces in total. Among them are two parameter spaces in which the true space is nested (spaces 6 and 7); the true parameter space is number 5. We conduct the study again with $n = 2000$. The results of 100 repetitions are given in Table 3.4. The results of this simulation study resemble the ones of the study with $\vartheta_0^{(2)}$ as true parameter – the CAIC is the criterion most prone to overfitting, while the AIC fares slightly better and the BIC still performs perfectly. The agreement of the AIC and CAIC is now lower in both cases. We also note that the AIC overfits in favor of model 6 while the CAIC selects model 7 if it makes a wrong choice. The explanation is similar as in the study before: the penalty terms in the AIC all deviate from the values of the penalty terms in the CAIC. The larger the space, however, the less the deviation. This, in combination with the fact that

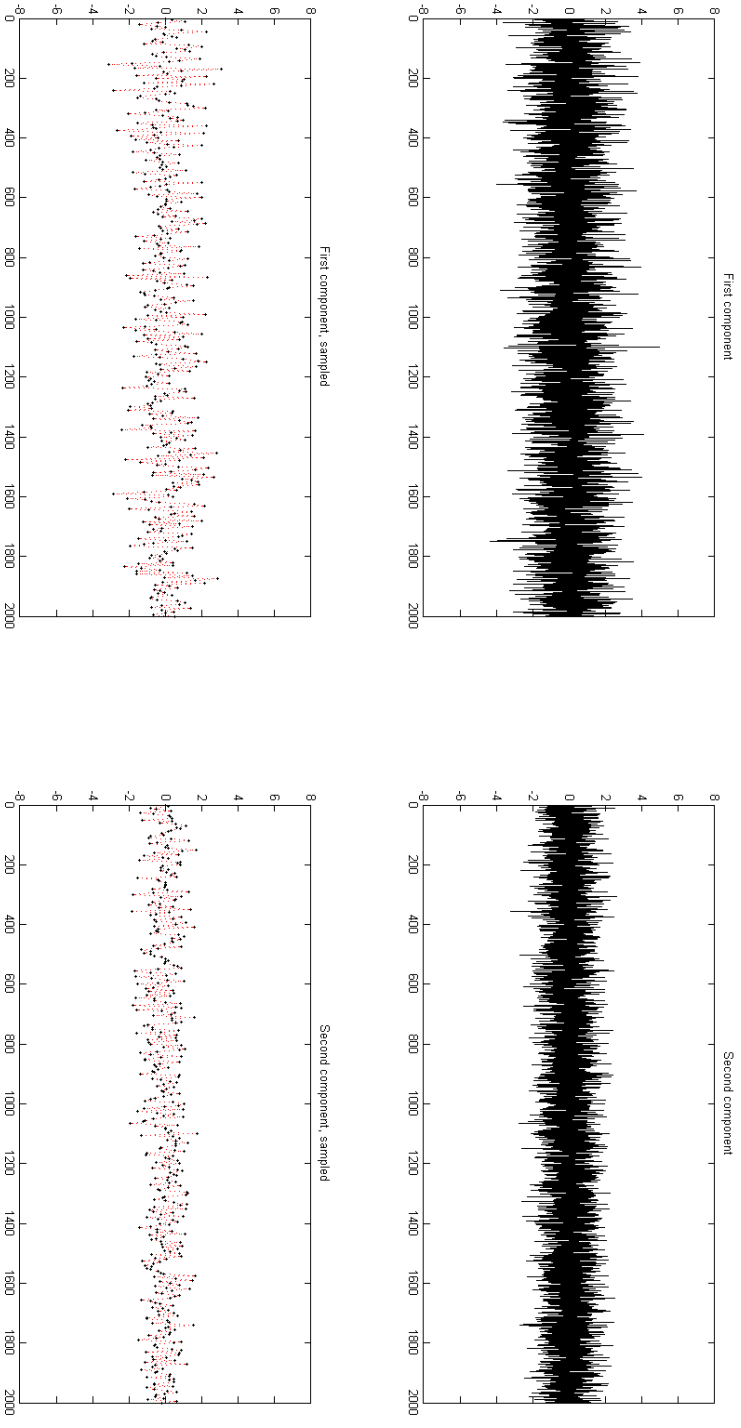

Figure 3.1.: Sample path of a CARMA(2,1) process driven by a Brownian motion and its discretely sampled version.

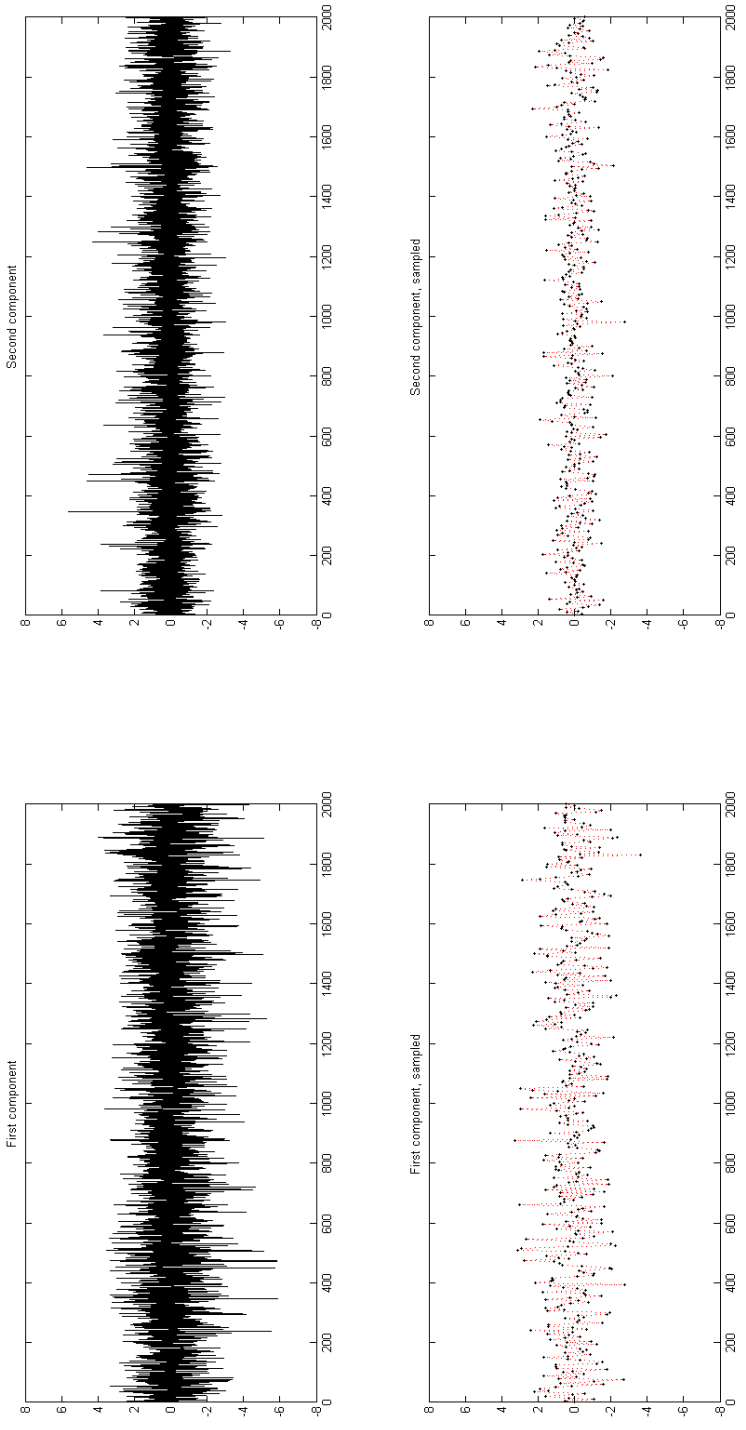

Figure 3.2.: Sample path of a CARMA(2,1) process driven by a NIG process and its discretely sampled version.

Space	Model				BM			NIG		
	m	p	q	$N(\Theta)$	AIC	CAIC	BIC	AIC	CAIC	BIC
1	$(1,1)$	1	0	7	1	0	0	5	0	0
2	$(1,2)$	2	1	10	0	0	0	0	0	0
3	$(2,1)$	2	1	11	0	0	0	0	0	0
4	$(2,2)$	2	1	15	0	0	0	0	0	0
5	$(3,2)$	3	0	13	91	89	100	87	88	100
6	$(3,2)$	3	1	17	9	0	0	13	0	0
7	$(3,2)$	3	2	19	0	11	0	0	12	0
Agreement					84%			79%		

Table 3.4.: Results for the true parameter $\vartheta_0^{(3)}$ in space 5.

the value of $\widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n)$ decreases as the number of parameters increases, leads to the selection of different overfitted spaces. The actual overfitting rate, however, is comparable to the simulation study with $m_0 = (1,2)$, i.e. the actual number of wrong choices has not increased systematically. This is also backed up by the observation that the approximated overfitting probability of space number 7 for the CAIC is 0.1019, showing that the empirical overfitting rate is very close to the theoretical probability in the Brownian motion case and slightly higher in the NIG case.
Lastly, we perform a small–sample simulation study to illustrate the advantages of AICb_n as defined in Subsection 3.4.5 in this situation. For the study, we chose again $m_0 = (3,2)$ and the true parameter $\vartheta_0^{(3)}$. We let $h = 1$ and $n = 15$, the driving Lévy process is a Brownian motion with the same covariance matrix as before. To ensure that our sample, despite its small size, is generated by a stationary process, we simulate the data–generating process on the interval $[0, 750]$ and use $Y_{\vartheta_0}(735), \ldots, Y_{\vartheta_0}(750)$ as the actual observations. For the number of bootstrap replications, we choose $b = 150$, following the recommendations of [Cavanaugh and Shumway 1997, p. 487f], who found that this value yields satisfying results. We considered only 3 candidate spaces here, the true parameter space and the two spaces in which it is nested (spaces 5, 6 and 7 in Table 3.4). The information criteria employed are AIC_n, CAIC_n, AICb_n, WIC_n and BIC_n. The reason for only considering three spaces is that due to the high number of bootstrap repetitions, adding further spaces dramatically increases the computation time. For one iteration of the order selection procedure (i.e. calculating each of the 4 criteria once for each parameter space), we needed about three days of computation time, such that a satisfyingly large number of iterations could only obtained by means of parallel computing. The bootstrap replicates were calculated by the algorithm of Stoffer and Wall [1991] as described in Subsection 3.4.5. We report on the results of 50 iterations here. The results are given in Table 3.5. To be consistent with Table 3.4, we have

kept the numbering of the spaces, although we do not consider 7 spaces in total anymore.

Space	Model				BM				
	m	p	q	$N(\Theta)$	AIC	CAIC	BIC	AICb	WIC
5	$(3,2)$	3	0	13	23	31	43	45	27
6	$(3,2)$	3	1	17	18	15	7	2	3
7	$(3,2)$	3	2	19	9	4	0	3	20

Table 3.5.: Results for the true parameter $\vartheta_0^{(3)}$ in space 5 with $n = 15$.

As we can see, the performance of AICb_n is vastly better than of the other two AIC–type criteria and also better than that of WIC_n. Its performance also better than that of the BIC, but only slightly. This coincides with the results of [Cavanaugh and Shumway 1997, Table 5], in which the authors report that the BIC (which they call SIC, their BIC is defined differently) becomes more and more competitive with the AICb as the number of parameters of the largest model, relative to the amount of observations, decreases. Since we have a maximal number of parameters equal to 19 and 15 observations, we are not quite in the situation with $2n = N(\Theta)$ in which the AICb performs best according to Cavanaugh and Shumway [1997]. Due to the high computation times necessary, we could not evaluate other scenarios and parameter setups for the AICb, but nevertheless these results and the theoretical framework affirm that it is a criterion worth considering in the small–sample setup.
Note that the performance of WIC_n, the second bootstrap–based criterion we briefly mentioned at the end of Subsection 3.4.5, is not satisfying. In fact, the number of correct selections is only slightly higher than 50 %. Due to this and also the fact that no satisfying theoretical foundation is available for WIC_n, we recommend the use of AICb_n and illustrate the results on WIC_n here mainly for sake of completeness.

Remark 3.22. *Except for Subsection 3.4.3, Subsection 3.4.4 and Subsection 3.4.5, the contents of Chapters 2 and 3 have been accepted for publication in Fasen and Kimmig [2016+]. The simulation study of Section 3.6 has also been altered in comparison to the published version, although the general framework remains the same. Most importantly, an error in the code of the simulation has been fixed and the studies have been repeated with the error–free code, which is why the results differ from those in the published article.*

CHAPTER 4

Robust estimation of MCARMA processes

The second part of this thesis is concerned with robust estimation of MCARMA processes. The central aspect of robust statistics is to investigate the behavior of statistical procedures when some of the underlying model assumptions are not satisfied. Historically, among the first scientists studying this problem were Tukey [1960], Huber [1964] and Hampel [1971] in the case of i.i.d. observations. In those articles, deviations from the model assumptions were understood as some observations coming from a different distribution than the one that was assumed for the i.i.d. data. Tukey [1960] then noticed that classical estimators, such as the empirical standard deviation, are highly sensible to this phenomenon and therefore, alternative procedures should be considered. Huber [1964] introduced the class of M–estimators for the location parameter of i.i.d. observations and showed that these estimators fare much better than, for example, the traditional sample mean. In Hampel [1971], the intuitive notion of a procedure being robust was first formalized and the terms qualitative robustness and breakdown point, which we will also study later on, were introduced (still in the case of i.i.d. observations). The M–estimators of Huber were later generalized to the context of linear regression, see e.g. Huber [1973], Yohai and Maronna [1979], Maronna and Yohai [1981] and [Maronna et al. 2006, Chapters 4 and 5]. Subsequently, further generalizations to the context of dependent observations, especially time series, were made, see e.g. Martin and Jong [1977], Denby and Martin [1979], Martin [1980], Bustos [1982] for the treatment of autoregressive processes

and Muler et al. [2009] for the treatment of ARMA processes. In the course of these generalizations, the need for a different class of estimators was recognized, as classical M–estimators fail to be robust in the case of dependent observations. Amongst other procedures (for an overview see [Maronna et al. 2006, Chapter 8]), the class of so–called generalized M–estimators (GM estimators) was considered and quite successfully applied to autoregressive processes. This class will appear also later in this chapter as a building block of a robust estimator for CARMA processes.

The first step in considering robustness consists of clarifying what is meant by the notion of "deviations from the nominal (or true) model". Contrary to the i.i.d. case, assuming that some observations arise from a different distribution than the others does not make much sense when considering time series. The reason is that in this context, the model assumption usually is that the observed data is generated by one particular, "true" time series, say Y, in a parametric family. Therefore, a natural way to model a deviation from the model assumptions would be to assume that the data is not a realization of the true time series Y, but only in some sense "close" to it. Up until now in this thesis, we always assumed that observations were generated by some discretely sampled MCARMA process $(Y(nh))_{n\in\mathbb{Z}}$ which we could observe. In a first step, in Section 4.1 we will define in what way we deviate from this assumption and how to model this appropiately. After establishing that, we will then move to the topic of robust parameter estimation.

As explained in [Huber and Ronchetti 2009, p. 5], a robust statistical procedure can be characterized by three properties. First, it should be reasonably efficient at the optimal model. Secondly, it should produce only slightly different results if there are slight deviations from the underlying model assumptions. And thirdly, larger deviations from the assumptions should, at least up to a certain point, not have a catastrophic effect. In the context of MCARMA processes, the statistical procedure which we will consider is the estimation of the parameters of the data–generating process. In Section 4.2, we will define and study M–estimators for MCARMA processes, which achieve the first property (they perform well at the nominal model), but unfortunately are not robust. Moving to the special case of one–dimensional CARMA processes, in Section 4.3 we will construct a so–called indirect estimator. Studying this estimator in detail will then allow us to show that it achieves all three of the properties outlined above. Subsection 4.3.1 is concerned with the behavior under the nominal model. In Subsection 4.3.4 the performance under deviations from the nominal model will be studied in terms of three robustness measures: qualitative robustness, the breakdown point and the influence functional. In Section 4.4, we come back to a topic already treated earlier, namely model selection in this new

framework. Lastly, we conduct a simulation study in Section 4.5 to illustrate the various theoretical results in practice.

We now start by formally defining what we classify as observations that are only close to, but not directly from, a data–generating process.

4.1. Discretely observed CARMA processes and outliers

The deviations from the true model we consider are characterized by the fact that we do not observe the data–generating process perfectly. Instead, we will only observe a disturbed process that is built from observations of the true time series subject to contamination by so–called outliers. These outliers can be thought of as atypical observations that do not arise because of the model structure, but due to some external influence, e.g. measurement errors. Therefore, a whole sample of observations which contains outliers does not come from the true model anymore, but is still close to it as long as the total number of outliers is not overwhelmingly large. There are several ways how one can formalize the presence of outliers in the data, with three main characterizations that are typically used in the literature, the so–called innovation outliers (IO), additive outliers (AO) and replacement outliers (RO). The latter two can be seen as special cases of the so–called general replacement model. We define the different types of outliers for a discretely sampled MCARMA process as follows:

Definition 4.1. *Let $(Y(nh))_{n\in\mathbb{Z}}$ be a d–dimensional discretely sampled MCARMA process as in* (2.5) *for some fixed $h > 0$.*

a) We say that $(Y(nh))_{n\in\mathbb{Z}}$ is afflicted by innovation outliers (IOs) if the innovations of this process have a heavy–tailed distribution (i.e. a distribution with infinite variance).

b) Let $g : [0,1] \to [0,1]$ be a function that satisfies $g(\gamma) - \gamma = o(\gamma)$ for $\gamma \to 0$. Let $(V_n)_{n\in\mathbb{Z}}$ be a stochastic process taking only the values 0 and 1 with

$$\mathbb{P}(V_n = 1) = g(\gamma) \tag{4.1}$$

and let $(Z_n)_{n\in\mathbb{Z}}$ be an arbitrary, d–dimensional stochastic process. We say that outliers are modeled by the general replacement model if we observe the disturbed process $(\widetilde{Y}_n)_{n\in\mathbb{Z}}$ instead of $(Y(nh))_{n\in\mathbb{Z}}$, where

$$\widetilde{Y}_n = (1 - V_n)Y(nh) + V_n Z_n. \tag{4.2}$$

Remark 4.2. *a) The general replacement model is, for example, also used in Martin and Yohai [1986]. The interpretation is that at each point $n \in \mathbb{Z}$, an outlier is observed with probability $g(\gamma)$, while the true value $Y(nh)$ is observed with probability $1 - g(\gamma)$. It has the advantage that one can obtain both additive and replacement outliers by choosing the processes $(Z_n)_{n\in\mathbb{Z}}$ and $(V_n)_{n\in\mathbb{Z}}$ adequately. Specifically, to model replacement outliers, one assumes that $(Z_n)_{n\in\mathbb{Z}}$, $(V_n)_{n\in\mathbb{Z}}$ and $(Y(nh))_{n\in\mathbb{Z}}$ are jointly independent. Then, if the realization of V_n at time n is equal to 1, the value $Y(nh)$ will be replaced by the realization of Z_n, justifying the use of the name replacement outliers.*
On the other hand, modeling additive outliers can be achieved by taking $Z_n = Y(nh) + W_n$ for some process $(W_n)_{n\in\mathbb{Z}}$ and assuming that $(Y(nh))_{n\in\mathbb{Z}}$ is independent from $(V_n)_{n\in\mathbb{Z}}$. Then we have

$$\widetilde{Y}_n = Y(nh) + V_n W_n,$$

such that the realization of W_n is added to the realization of $Y(nh)$ if V_n realizes as 1, modeling exactly the behavior one wishes to have.

b) Another advantage of the general replacement model is that one can easily model the temporal structure of outliers. On the one hand, for example, if $(V_n)_{n\in\mathbb{Z}}$ is chosen as an i.i.d. process with $\mathbb{P}(V_n = 1) = \gamma$, then outliers typically appear isolated, i.e. between two outliers there is usually a period of time where no outliers are present. On the other hand, one can also model patchy outliers by letting $(B_n)_{n\in\mathbb{Z}}$ be an i.i.d. process of Bernoulli variables with success probability ϵ and setting $V_n = \max(B_{n-k}, \ldots, B_n)$ for a $k \in \mathbb{N}$. Then

$$\mathbb{P}(V_n = 1) = 1 - (1 - \epsilon)^k = k\epsilon + o(\epsilon)$$

for $\epsilon \to 0$, i.e. (4.1) holds with $\gamma := k\epsilon$. For ϵ sufficiently small, outliers then typically appear in a patch or block of size k.

c) The fundamental difference between IOs and the other two types of outliers is that the observations in the case of IOs still come from a discretely sampled MCARMA process, albeit with infinite–variance noise. For the other two types of outliers, the observations do not follow the same structure as in the uncontaminated case, which makes estimation more difficult. Also, in the case of IOs, an outlier influences the future values of the stochastic process $(Y(nh))_{n\in\mathbb{Z}}$, while this is not the case for outliers generated by the general replacement model.

d) In the case of IOs, one needs to estimate a process with infinite second moments. For this reason, the results are typically quite different from those in the case of the other outlier types. For example, in Andrews et al. [2009] the maximum likelihood estimator is used successfully, while it is well–known that MLEs are typically nonrobust in the case of additive or replacement outliers. In Davis et al. [1992], M–estimation is studied in this context. We will not study IOs in detail here, leaving their treatment in the context of MCARMA processes open for future research and focusing on the general replacement model for the rest of the thesis.

4.2. M-estimators for MCARMA processes

In this chapter, the aim is to introduce the notion of M-estimation for the class of MCARMA processes. Since its introduction in Huber [1964] for the estimation of the location parameter of i.i.d. data, this class of estimators has been applied in many problems and fields, including but not limited to parameter estimation in linear regression (Huber [1973], Yohai and Maronna [1979]) and of ARMA processes with finite ([Maronna et al. 2006, Section 8.4]) and infinite (Davis et al. [1992]) variance. The principal idea of M–estimators is to generalize the maximum likelihood procedure by replacing the likelihood function by a more general one (hence the name, M–estimator is a shorthand for "Maximum likelihood type estimator"). Huber [1964] recognized that the likelihood function is generally unbounded and therefore proposed to use a bounded function to achieve robustness. Typically, it is then also possible to show consistency and asymptotic normality for this class of estimators at the nominal model, i.e. in the absence of outliers. We will also take this approach. Since the method is a generalization of maximum likelihood estimation, if we choose suitable functions to replace the likelihood function, the proofs will be very similar to the ones in the study of QMLE for MCARMA processes.

Just as with maximum likelihood estimation, we assume that n, equidistant, discrete–time observations with sampling distance $h > 0$ of a MCARMA processes $(Y(t))_{t\in\mathbb{R}}$ are given in the form of $Y^n = (Y(h), \dots, Y(nh))$. To do parameter estimation, we again consider a parameter space Θ and a parametric family of CSSMs $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ that satisfy Assumptions B.1 to B.9. Moreover, for simplicity's sake, we assume that Θ contains an element ϑ_0 that generates the true output process. Of course in the context of M-estimation we can forgo the assumptions B.10 and B.11, as those were explicitly connected to the maximum likelihood method, while the rest of Assumption B is more generally related to the parametrization under consideration.

Then, similar to (2.19), we define our estimator by

$$\widehat{\vartheta}^n_{\mathrm{M}} = \arg\min_{\vartheta\in\Theta} \frac{1}{n}\widehat{\mathcal{L}}_M(\vartheta, Y^n) := \arg\min_{\vartheta\in\Theta} \frac{1}{n}\sum_{k=1}^{n} \rho(\widehat{\epsilon}_{\vartheta,k}), \tag{4.3}$$

where $\rho : \mathbb{R}^d \to \mathbb{R}$ is a suitably chosen loss function and $(\widehat{\epsilon}_{\vartheta,k})_{k\in\mathbb{N}}$ are the approximate pseudo-innovations as calculated by the Kalman filter. For the loss function, we make the following assumption:

Assumption F.

F.1 The function $\rho : \mathbb{R}^d \to \mathbb{R}$ is three times continuously differentiable.

F.2 The expectation $\mathbb{E}[\rho(\epsilon_{\vartheta,1})]$ exists for every $\vartheta \in \Theta$.

F.3 It holds that

$$|(\partial_i\rho)(x) - (\partial_i\rho)(y)| \le c\|x-y\| \text{ and } |(\partial^2_{i,j}\rho)(x) - (\partial^2_{i,j}\rho)(y)| \le c\|x-y\|$$

for all $i, j \in \{1, \dots, d\}$ and $x, y \in \mathbb{R}^d$ and some constant $c \ge 0$.

We will now proceed similarly as in [Schlemm and Stelzer 2012, Section 2] to show consistency and asymptotic normality of $\widehat{\vartheta}^n_{\mathrm{M}}$. Just as in the maximum likelihood case, it is usually more convenient for the proofs to consider the function $\mathcal{L}_M(\vartheta, Y^n) := \frac{1}{n}\sum_{k=1}^{n} \rho(\epsilon_{\vartheta,k})$ and its derivatives, respectively, where we have replaced the approximate pseudo-innovations by their theoretical counterparts, instead of $\widehat{\mathcal{L}}_M$. Analogous to Lemma 2.23, we therefore show that the approximate quantities converge to their theoretical counterparts:

Lemma 4.3. *Assume that for $i, j \in \{1, \dots, N(\Theta)\}$ the initial values $\widehat{X}_{\vartheta,initial}$ are such that $\sup_{\vartheta\in\Theta} \|\widehat{X}_{\vartheta,1}\|$, $\sup_{\vartheta\in\Theta} \|\partial_i\widehat{X}_{\vartheta,1}\|$ and $\sup_{\vartheta\in\Theta} \|\partial^2_{i,j}\widehat{X}_{\vartheta,1}\|$ are almost surely finite. Then it holds:*

a) $\sup_{\vartheta\in\Theta} \left|\widehat{\mathcal{L}}_M(\vartheta, Y^n) - \mathcal{L}_M(\vartheta, Y^n)\right| \to 0$ *as* $n \to \infty$ $\mathbb{P}$*-a.s.*

b) $\sqrt{n}\sup_{\vartheta\in\Theta} \left|\partial_i\widehat{\mathcal{L}}_M(\vartheta, Y^n) - \partial_i\mathcal{L}_M(\vartheta, Y^n)\right| \xrightarrow{\mathbb{P}} 0$ *as* $n \to \infty$.

c) $\sup_{\vartheta\in\Theta} \left|\partial^2_{i,j}\widehat{\mathcal{L}}_M(\vartheta, Y^n) - \partial^2_{i,j}\mathcal{L}_M(\vartheta, Y^n)\right| \to 0$ *as* $n \to \infty$ $\mathbb{P}$*-a.s.*

Proof. a) First off, note that by a Taylor expansion we have that

$$\sup_{\vartheta\in\Theta} |\widehat{\mathcal{L}}_M(\vartheta, Y^n) - \mathcal{L}_M(\vartheta, Y^n)| \le \sup_{\vartheta\in\Theta} \frac{1}{n}\sum_{k=1}^{n} |\rho(\widehat{\epsilon}_{\vartheta,k}) - \rho(\epsilon_{\vartheta,k})|$$

$$\leq \sup_{\vartheta\in\Theta} \frac{1}{n} \sum_{k=1}^{n} \|\nabla_\vartheta \rho(\overline{\epsilon}_{\vartheta,k})\| \|\widehat{\epsilon}_{\vartheta,k} - \epsilon_{\vartheta,k}\|,$$

where $\|\overline{\epsilon}_{\vartheta,k} - \epsilon_{\vartheta,k}\| \leq \|\widehat{\epsilon}_{\vartheta,k} - \epsilon_{\vartheta,k}\|$,

$$\begin{aligned}
&\overset{F.3}{\leq} \frac{C}{n} \sum_{k=1}^{n} \left((\sup_{\vartheta\in\Theta}(\|\overline{\epsilon}_{\vartheta,k}\| + \|\nabla_\vartheta \rho(0)\|) \sup_{\vartheta\in\Theta} \|\widehat{\epsilon}_{\vartheta,k} - \epsilon_{\vartheta,k}\| \right) \\
&\leq \frac{C}{n} \sum_{k=1}^{n} \left(\sup_{\vartheta\in\Theta}(\|\overline{\epsilon}_{\vartheta,k} - \epsilon_{\vartheta,k} + \epsilon_{\vartheta,k}\| + \|\nabla_\vartheta \rho(0)\|)\rho^k \right) \\
&\leq \frac{C}{n} \sum_{k=1}^{n} \left(\sup_{\vartheta\in\Theta} \|\widehat{\epsilon}_{\vartheta,k} - \epsilon_{\vartheta,k}\|\rho^k + \sup_{\vartheta\in\Theta} \|\epsilon_{\vartheta,k}\|\rho^k + \|\nabla_\vartheta \rho(0)\|\rho^k \right) \\
&\leq \frac{C}{n} \sum_{k=1}^{n} \left(\widetilde{\rho}^k + \rho^k \sup_{\vartheta\in\Theta} \|\epsilon_{\vartheta,k}\| + \|\nabla_\vartheta \rho(0)\|\rho^k \right)
\end{aligned} \tag{4.4}$$

where we have used that $\sup_{\vartheta\in\Theta} \|\widehat{\epsilon}_{\vartheta,k} - \epsilon_{\vartheta,k}\| \leq C\rho^k$ for $C > 0$ and $\rho \in (0,1)$ by Lemma 2.22a) and defined $\widetilde{\rho} = \rho^2$. As in the proof of [Schlemm and Stelzer 2012, Lemma 2.7], we can show that $\rho^k \sup_{\vartheta\in\Theta} \|\epsilon_{\vartheta,k}\|$ converges to 0 almost surely as $k \to \infty$ by using the Markov inequality and the Borel–Cantelli lemma. Since the sequences $(\widetilde{\rho}^k)_{k\in\mathbb{N}}$ and $(\|\nabla_\vartheta \rho(0)\|\rho^k)_{k\in\mathbb{N}}$ converge to zero as well, we obtain that the sum in (4.4) converges to 0 almost surely. The proofs of b) and c) are similar. □

The next lemma deals with the asymptotic behavior of the function $\mathcal{L}_M$:

Lemma 4.4. *For $n \to \infty$, the sequence of random functions $\vartheta \mapsto \mathcal{L}_M(\vartheta, Y^n)$ converges uniformly in ϑ almost surely to the limiting function*

$$\mathscr{Q}_M(\vartheta) := \mathbb{E}\left[\rho(\epsilon_{\vartheta,1})\right].$$

Proof. Since the sampled output process $(Y(nh))_{n\in\mathbb{Z}}$ is ergodic, the same is true for the pseudo-innovation sequence $(\epsilon_{\vartheta,n})_{n\in\mathbb{Z}}$ for every $\vartheta \in \Theta$ as apparent from (2.13). Since ρ is measurable, the sequence $(\rho(\epsilon_{\vartheta,n}))_{n\in\mathbb{Z}}$ is ergodic again by [Durrett 2010, Theorem 7.1.3]. Hence, by Birkhoff's ergodic theorem ([Durrett 2010, Theorem 7.2.1]) it follows that the sequence $\mathcal{L}_M(\vartheta, Y^n)$ converges almost surely pointwise to $\mathscr{Q}_M$. Uniform convergence can be shown by means of the compactness of the parameter space and an application of [Ferguson 1996, Theorem 16a)]. □

This lemma suffices for us to derive the consistency of the M–estimator:

Theorem 4.5. *If the function $\vartheta \mapsto \mathscr{Q}_M(\vartheta)$ has a unique global minimum at ϑ_0, then the estimator $\widehat{\vartheta}_M^n$ is strongly consistent, i. e. it holds that*

$$\widehat{\vartheta}_M^n \overset{n\to\infty}{\longrightarrow} \vartheta_0$$

almost surely.

Proof. The proof of consistency is the same one as that of [Schlemm and Stelzer 2012, Theorem 2.4], making use of Lemma 4.4 with the function $\widehat{\mathcal{L}}_M$ replacing $\widehat{\mathcal{L}}$ and $\mathscr{Q}_M$ replacing $\mathscr{Q}$. □

To show asymptotic normality of the M-estimator $\widehat{\vartheta}_{\mathrm{M}}^n$, we will prove a series of lemmas and then put them together eventually. This way of proceeding is in principle the same as the one in [Schlemm and Stelzer 2012, Section 2.4], with the difference that we have a general loss function ρ whereas Schlemm and Stelzer have the pseudo–Gaussian likelihood function. Therefore, they can use the explicit representation of this likelihood and its properties, whereas we work with the properties that can be deduced from Assumption F. Some ideas of the proofs remain valid under the change of the objective function. However, a key difference to the approach of Schlemm and Stelzer [2012] is how a central limit theorem for the suitably scaled gradient of the objective function is derived, the analogue of [Schlemm and Stelzer 2012, Lemma 2.16], which is a crucial step in the proof of the central limit theorem for $\widehat{\vartheta}_{\mathrm{M}}^n$. We will obtain this central limit theorem by making use of the concept of near–epoch dependent stochastic processes. They are defined as follows:

Definition 4.6. *Let $(R_n)_{n\in\mathbb{Z}}$ be a (vector–valued) stochastic process and denote $\mathcal{F}_{n-m}^{n+m} = \sigma(R_{n-m}, \ldots, R_{n+m})$ for $m \in \mathbb{N}$, $n \in \mathbb{Z}$. Let $(S_n)_{n\in\mathbb{Z}}$ be a one–dimensional stochastic process with $\mathbb{E}[|S_n|] < \infty$ for every $n \in \mathbb{Z}$. $(S_n)_{n\in\mathbb{Z}}$ is said to be near epoch dependent in L^p–norm (L^p–NED) on $(R_n)_{n\in\mathbb{Z}}$ if it holds that*

$$\| S_n - \mathbb{E}\left[S_n \mid \mathcal{F}_{n-m}^{n+m}\right] \|_{L^p} \leq v_m d_n,$$

where $(d_n)_{n\in\mathbb{Z}}$ is a sequence of positive constants and $v_m \to 0$ as $m \to \infty$ holds. $(S_n)_{n\in\mathbb{Z}}$ is said to be of size φ_0 if $(v_m)_{m\in\mathbb{N}}$ is of order $O(m^{-\varphi})$ for every $\varphi > \varphi_0$.

Remark 4.7. *a) Near epoch dependence is a property of the map from $(R_n)_{n\in\mathbb{Z}}$ to $(S_n)_{n\in\mathbb{Z}}$ in the sense that the latter may depend on the full history (and even future) of the former, but the dependence disappears fast enough as the distance between time points increases. This concept is particularly useful if $(R_n)_{n\in\mathbb{Z}}$ is a mixing process, since functions of a mixing process which depend*

on infinitely many of its values are not necessarily mixing anymore. However, under some mild additional conditions, a process that is near epoch dependent on a strongly mixing process inherits enough of its properties to obtain a law of large numbers and a central limit theorem. This is because for each $n \in \mathbb{Z}$ the random variable S_n can be "approximated well" by $\mathbb{E}\left[S_n \mid \mathcal{F}_{n-m}^{n+m}\right]$, which is then by construction a finite–lag function of a mixing process, i.e. itself mixing again. Formally, one can show that $(S_n)_{n\in\mathbb{Z}}$ is a so–called mixingale in this case. For further details on near epoch dependence and mixingales we refer to [Davidson 1994, Chapters 16 and 17] and White [1996].

b) If a process $(S_n)_{n\in\mathbb{Z}}$ is of size φ_0, it is of course also of size φ' for every $\varphi' > \varphi_0$ by definition. Similarly, if $(S_n)_{n\in\mathbb{Z}}$ is L^p–NED on $(R_n)_{n\in\mathbb{Z}}$ for some $p > 0$, it is also L^q–NED on the same process for every $1 \leq q \leq p$ ([Davidson 1994, p. 268]).

The most fundamental observation now is that for both the innovations and their partial derivatives, every component is L^2–NED on the data–generating CARMA process $(Y(nh))_{n\in\mathbb{Z}}$:

Lemma 4.8. *For every $\vartheta \in \Theta$, every $j \in \{1, \ldots, d\}$ and every $i \in \{1, \ldots, N(\Theta)\}$, the processes $((\epsilon_{\vartheta,n})_j)_{n\in\mathbb{Z}}$ and $((\partial_i \epsilon_{\vartheta,n})_j)_{n\in\mathbb{Z}}$ are L^2–NED of size $-\infty$ on $(Y(nh))_{n\in\mathbb{Z}}$.*

Proof. The proof is analogous to [Davidson 1994, Example 17.3], but we give it here in detail for sake of completeness. By Lemma 2.22a) it holds that

$$\epsilon_{\vartheta,k} = Y(kh) + \sum_{\nu=1}^{\infty} c_{\vartheta,\nu} Y((k-\nu)h)$$

for $k \in \mathbb{Z}$. Therefore, for the j–th component, denoting by $(c_{\vartheta,\nu})_j$ the j–th row of the matrix $c_{\vartheta,\nu}$, we have that

$$(\epsilon_{\vartheta,k})_j = \sum_{\nu=0}^{\infty} (c_{\vartheta,\nu})_j^T Y((k-\nu)h).$$

Denoting

$$\mathcal{F}_{k-m}^{k+m} := \sigma(Y((k-m)h), \ldots, Y((k+m)h),\ k \in \mathbb{Z},\ m \in \mathbb{N},$$

we have that

$$
\begin{aligned}
&\| (\epsilon_{\vartheta,k})_j - \mathbb{E}\left[(\epsilon_{\vartheta,k})_j \mid \mathcal{F}_{k-m}^{k+m}\right] \|_{L^2} \\
&= \left\| \sum_{\nu=m+1}^{\infty} (c_{\vartheta,\nu})_j^T \left(Y((k-\nu)h) - \mathbb{E}\left[Y((k-\nu)h) \mid \sigma(Y((k-m)h), \ldots, Y((k+m)h))\right]\right) \right\|_{L^2} \\
&\leq \sum_{\nu=m+1}^{\infty} \left\| (c_{\vartheta,\nu})_j^T (Y((k-\nu)h) - \mathbb{E}\left[Y((k-\nu)h) \mid \sigma(Y((k-m)h), \ldots, Y((k+m)h))\right]) \right\|_{L^2}
\end{aligned}
$$

where we used the Minkowski inequality. By definition of the L^2–norm and the Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
&\left\| (c_{\vartheta,\nu})_j^T (Y((k-\nu)h) - \mathbb{E}\left[Y((k-\nu)h) \mid \sigma(Y((k-m)h), \ldots, Y((k+m)h))\right]) \right\|_{L^2} \\
&= \left(\mathbb{E}\left[\left| (c_{\vartheta,\nu})_j^T (Y((k-\nu)h) - \mathbb{E}\left[Y((k-\nu)h) \mid \sigma(Y((k-m)h), \ldots, Y((k+m)h))\right]) \right|^2 \right] \right)^{\frac{1}{2}} \\
&\leq \| (c_{\vartheta,\nu})_j^T \| \left(\mathbb{E}\left[\| Y((k-\nu)h) - \mathbb{E}\left[Y((k-\nu)h) \mid \sigma(Y((k-m)h), \ldots, Y((k+m)h))\right] \|^2 \right] \right)^{\frac{1}{2}} \\
&= \| (c_{\vartheta,\nu})_j^T \| \| Y((k-\nu)h) - \mathbb{E}\left[Y((k-\nu)h) \mid \sigma(Y((k-m)h), \ldots, Y((k+m)h))\right] \|_{L^2} \\
&\leq \| (c_{\vartheta,\nu})_j^T \| \| Y((k-\nu)h) \|_{L^2} \\
&= \| (c_{\vartheta,\nu})_j^T \| \left(\mathbb{E}\left[\| Y((k-\nu)h) \|^2 \right] \right)^{\frac{1}{2}} \\
&\leq \| c_{\vartheta,\nu} \| \left(\mathbb{E}\left[\| Y(h) \|^2 \right] \right)^{\frac{1}{2}},
\end{aligned}
$$

where we also used that the process $(Y(nh))_{n\in\mathbb{Z}}$ has mean zero and is stationary. Defining $v_m := \sum_{\nu=m+1}^{\infty} \| c_{\vartheta,\nu} \|$, we see that this sequence is of size $-\infty$ since $\| c_{\vartheta,\nu} \| \leq C\rho^{\nu}$ for $\rho \in (0,1)$ by Lemma 2.22a). Letting $d_t = d_1 := (\mathbb{E}\left[\| Y(h) \|^2 \right])^{\frac{1}{2}}$ we obtain the assertion. The proof for $(\partial_i \epsilon_{\vartheta,n})_{n\in\mathbb{Z}}$ is analogous, using Lemma 2.22b). □

A useful property of near epoch dependence is that one can state readily verified conditions under which this property is preserved by transformations. This is the subject of the following lemma:

Lemma 4.9. *Assume that $(S_n)_{n\in\mathbb{Z}}$ and $(T_n)_{n\in\mathbb{Z}}$ are one–dimensional stochastic processes, which are L^2–NED on a process $(R_n)_{n\in\mathbb{Z}}$ of size $-\varphi_S$ and $-\varphi_T$, respectively.*

a) The process $(S_n + T_n)_{n\in\mathbb{Z}}$ is L^2–NED on $(R_n)_{n\in\mathbb{Z}}$ of size $-\min\{\varphi_S, \varphi_T\}$.

b) Assume, in addition, that $\varphi_S = \varphi_T$ and that $\mathbb{E}\left[|S_n|^{2r}\right] < \infty$ and $\mathbb{E}\left[|T_n|^{2r}\right] < \infty$ for every $n \in \mathbb{Z}$ and some $r > 2$. Then, $(S_n T_n)_{n\in\mathbb{Z}}$ is L^2–NED on $(R_n)_{n\in\mathbb{Z}}$ of size $-\frac{\varphi_S(r-2)}{2(r-1)}$.

c) Assume, in addition, that the function $\tau : \mathbb{R}^d \to \mathbb{R}$ satisfies $|\tau(x) - \tau(y)| \leq c\|x - y\|$ for every $x, y \in \mathbb{R}^d$ and some $c > 0$. Moreover, assume that for

$i = 1, \dots, d$, the processes $(S_n^{(i)})_{n\in\mathbb{Z}}$ are all L^2–NED on a process $(R_n)_{n\in\mathbb{Z}}$ of a common size $-\varphi_S$. Then the process $(\tau(S_n^{(1)}, \dots, S_n^{(d)}))_{n\in\mathbb{Z}}$ is L^2–NED on $(R_n)_{n\in\mathbb{Z}}$ of size $-\varphi_S$.

Proof. Part a) is [Davidson 1994, Theorem 17.8]. Part b) is [Davidson 1994, Example 17.17]. Part c) is a special case of [Davidson 1994, Theorem 17.12]. □

We need these three specific results to obtain the following lemma:

Lemma 4.10. *For every $c \in \mathbb{R}^{N(\Theta)}$ and every $\vartheta \in \Theta$, the one–dimensional process $(c^T \nabla_\vartheta(\rho(\epsilon_{\vartheta,n})))_{n\in\mathbb{Z}}$ is L^2–NED of size $-\infty$ on $(Y(nh))_{n\in\mathbb{Z}}$.*

Proof. First, observe that it holds

$$c^T \nabla_\vartheta(\rho(\epsilon_{\vartheta,n})) = \sum_{l=1}^{N(\Theta)} c_l \left(\sum_{j=1}^{d} (\partial_j \rho)(\epsilon_{\vartheta,n})(\partial_l \epsilon_{\vartheta,n})_j \right).$$

By Assumption F.3, Lemma 4.8 and Lemma 4.9c), the process $((\partial_j\rho)(\epsilon_{\vartheta,n}))_{n\in\mathbb{Z}}$ is L^2–NED of size $-\infty$ on $(Y(nh))_{n\in\mathbb{Z}}$ for every $j \in \{1, \dots, d\}$. Again by Lemma 4.8, $((\partial_l \epsilon_{\vartheta,n})_j)_{n\in\mathbb{Z}}$ is also L^2–NED of size $-\infty$ on $(Y(nh))_{n\in\mathbb{Z}}$ for every $j \in \{1, \dots, d\}$. By the moving average representations of Lemma 2.22a) and b) and Assumption B.9 and F.3, there exists a $\delta > 0$ such that $\mathbb{E}\left[|(\partial_j\rho)(\epsilon_{\vartheta,n})|^{4+\delta}\right] < \infty$ and $\mathbb{E}\left[|(\partial_l \epsilon_{\vartheta,n})_j|^{4+\delta}\right] < \infty$. Applying Lemma 4.9b) with $r = 2 + \frac{\delta}{2}$, it follows that the process $((\partial_j\rho)(\epsilon_{\vartheta,n})(\partial_l\epsilon_{\vartheta,n})_j)_{n\in Z}$ is L^2–NED of size $-\infty$ on $(Y(nh))_{n\in\mathbb{Z}}$. Repeatedly applying Lemma 4.9a) allows to deal with the sums and arrive at the assertion of the lemma. □

Before we can obtain the desired central limit theorem, we need to consider the behavior of the variance of the gradient of the objective function. This is the topic of the following two lemmas, in which we first establish that this variance exists for every $n \in \mathbb{N}$ and then consider the asymptotic behavior under suitable scaling. The first lemma is analogous to [Schlemm and Stelzer 2012, Lemma 2.12].

Lemma 4.11. *For each $\vartheta \in \Theta$ and every $i \in \{1, \dots, r\}$ the variance of $\partial_i \mathcal{L}_M(\vartheta, Y^n)$ is finite.*

Proof. For fixed $i \in \{1, \dots, r\}$ it holds that

$$\partial_i \mathcal{L}_M(\vartheta, Y^n) = \frac{1}{n} \sum_{k=1}^{n} \sum_{j=1}^{d} (\partial_j \rho)(\epsilon_{\vartheta,k})(\partial_i \epsilon_{\vartheta,k})_j =: \frac{1}{n} \sum_{k=1}^{n} \partial_i g_{\vartheta,k}.$$

Two applications of the Cauchy–Schwarz inequality and Assumption F.3 give the following chain of estimates:

$$\begin{aligned}
\operatorname{Var}(\partial_i \mathcal{L}_M(\vartheta, Y^n)) &= \frac{1}{n^2} \sum_{k,l=1}^{n} \sum_{j,j'=1}^{d} \operatorname{Cov}\left((\partial_j \rho)(\epsilon_{\vartheta,k})(\partial_i \epsilon_{\vartheta,k})_j; (\partial_{j'} \rho)(\epsilon_{\vartheta,l})(\partial_i \epsilon_{\vartheta,l})_{j'}\right) \\
&\leq \frac{1}{n^2} \sum_{k,l=1}^{n} \sum_{j,j'=1}^{d} \sqrt{\operatorname{Var}\left((\partial_j \rho)(\epsilon_{\vartheta,k})(\partial_i \epsilon_{\vartheta,k})_j\right)} \sqrt{\operatorname{Var}\left((\partial_{j'} \rho)(\epsilon_{\vartheta,l})(\partial_i \epsilon_{\vartheta,l})_{j'}\right)} \\
&\leq \frac{1}{n^2} \sum_{k,l=1}^{n} \sum_{j,j'=1}^{d} \sqrt{\mathbb{E}\left[((\partial_j \rho)(\epsilon_{\vartheta,k})(\partial_i \epsilon_{\vartheta,k})_j)^2\right]} \sqrt{\mathbb{E}\left[((\partial_{j'} \rho)(\epsilon_{\vartheta,l})(\partial_i \epsilon_{\vartheta,l})_{j'})^2\right]} \\
&\leq \frac{1}{n^2} \sum_{k,l=1}^{n} \sum_{j,j'=1}^{d} \Big(\mathbb{E}\left[((\partial_j \rho)(\epsilon_{\vartheta,k}))^4\right]^{\frac{1}{4}} \mathbb{E}\left[((\partial_i \epsilon_{\vartheta,k})_j)^4\right]^{\frac{1}{4}} \cdot \\
&\qquad \cdot \mathbb{E}\left[((\partial_{j'} \rho)(\epsilon_{\vartheta,l}))^4\right]^{\frac{1}{4}} \mathbb{E}\left[((\partial_i \epsilon_{\vartheta,l})_{j'})^4\right]^{\frac{1}{4}} \Big) \\
&\leq C \mathbb{E}\left[\|\epsilon_{\vartheta,1}\|^4\right]^{\frac{1}{2}} \sum_{j,j'=1}^{d} \mathbb{E}\left[((\partial_i \epsilon_{\vartheta,1})_j)^4\right]^{\frac{1}{4}} \mathbb{E}\left[((\partial_i \epsilon_{\vartheta,1})_{j'})^4\right]^{\frac{1}{4}} < \infty
\end{aligned}$$

since the fourth moments appearing here are finite by Assumption B.9 and the moving average representations given in Lemma 2.22. □

As before, the correct scaling of the variance of the gradient to obtain convergence is by a factor n, as the following lemma shows:

Lemma 4.12. *For every $\vartheta \in \Theta$, there exists a deterministic matrix $\mathcal{I}_M(\vartheta)$ such that*

$$n \operatorname{Var}\left(\nabla_\vartheta \mathcal{L}_M(\vartheta, Y^n)\right) \overset{n\to\infty}{\longrightarrow} \mathcal{I}_M(\vartheta).$$

Proof. Since the element in row m and column l of the matrix $n \operatorname{Var}\left(\nabla_\vartheta \mathcal{L}_M(\vartheta, Y^n)\right)$ can be written as

$$I_{\vartheta,n}^{(m,l)} := n \operatorname{Cov}(\partial_m \mathcal{L}_M(\vartheta, Y^n), \partial_l \mathcal{L}_M(\vartheta, Y^n)) = \frac{1}{n} \sum_{k=1}^{n} \sum_{t=1}^{n} \operatorname{Cov}(\partial_m \rho(\epsilon_{\vartheta,k}), \partial_l \rho(\epsilon_{\vartheta,t}))$$

it follows from [Davidson 1994, p. 266] that a sufficient condition for the statement to hold is the summability of the sequence $(\operatorname{Cov}(\partial_m \rho(\epsilon_{\vartheta,k}), \partial_l \rho(\epsilon_{\vartheta,k+\Delta})))_{\Delta \in \mathbb{N}}$ for every $k \in \mathbb{Z}$ and $m, l \in \{1, \ldots, N(\Theta)\}$. Since by Lemma 4.10, both $(\partial_m \rho(\epsilon_{\vartheta,k}))_{k\in\mathbb{Z}}$ and $(\partial_l \rho(\epsilon_{\vartheta,k}))_{k\in\mathbb{Z}}$ are L^2–NED of size $-\infty$ on $(Y(kh))_{k\in\mathbb{Z}}$ and $(Y(kh))_{k\in\mathbb{Z}}$ is an exponentially strongly mixing process by Proposition 2.24 (i.e. its mixing coefficients are of size $-\infty$), the assertion now follows from [Davidson 1994, Theorem 17.7] and the comment right after that theorem. □

We can now put these results together for the central limit theorem:

Lemma 4.13. *It holds that*

$$\sqrt{n}\nabla_\vartheta \widehat{\mathcal{L}}_M(\vartheta_0, Y^n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{I}_M(\vartheta_0)),\ n \to \infty.$$

Proof. By Lemma 4.3b), it suffices to show that the random variable $\sqrt{n}\nabla_\vartheta \mathcal{L}_M(\vartheta_0, Y^n)$ has the limiting normal distribution given in the statement of the lemma. We make use of the Cramér–Wold device and show that

$$\sqrt{n}c^T\nabla_\vartheta \mathcal{L}_M(\vartheta_0, Y^n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, c^T\mathcal{I}_\mathrm{M}(\vartheta_0)c),\ n \to \infty,$$

for every $c \in \mathbb{R}^{N(\Theta)}$. To this end, note that, as already used in the proof of Lemma 4.10, there exists a $\delta > 0$ such that

$$\mathbb{E}\left[|c^T\nabla_\vartheta \mathcal{L}_M(\vartheta_0, Y^n)|^{4+\delta}\right] < \infty$$

holds. Moreover, by Lemma 4.10, $(c^T\nabla_\vartheta(\rho(\epsilon_{\vartheta,n})))_{n\in\mathbb{Z}}$ is L^2–NED of size $-\infty$ on $(Y(nh))_{n\in\mathbb{Z}}$, which by Proposition 2.24 is an exponentially strongly mixing process, i.e. an α–mixing process of size $-\infty$. Lastly, it holds that

$$\mathrm{Var}(c^T\nabla_\vartheta \mathcal{L}_M(\vartheta_0, Y^n)) = c^T \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{k=1}^n \nabla_\vartheta g_{\vartheta_0,k}\right) c$$

is $O_\mathbb{P}(n^{-1})$ by Lemma 4.12, i.e. $\mathrm{Var}(\sum_{k=1}^n \nabla_\vartheta g_{\vartheta_0,k})$ is $O_\mathbb{P}(n)$. Therefore, the conditions of [White 1996, Theorem A.3.7] are satisfied and we obtain that

$$\frac{c^T\nabla_\vartheta \mathcal{L}_M(\vartheta_0, Y^n)}{\sqrt{\mathrm{Var}(c^T\nabla_\vartheta \mathcal{L}_M(\vartheta_0, Y^n))}} = \frac{c^T\sum_{k=1}^n \nabla_\vartheta g_{\vartheta_0,k}}{\sqrt{\mathrm{Var}(c^T\sum_{k=1}^n \nabla_\vartheta g_{\vartheta_0,k})}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),\ n \to \infty,$$

from which we deduce the assertion by Lemma 4.12 and Slutzky's lemma. □

We now need one further lemma which deals with the limiting behavior of the second derivative of the objective function, which will appear again in the asymptotic covariance matrix of our estimator. The statement corresponds to the one of [Schlemm and Stelzer 2012, Lemma 2.17].

Lemma 4.14. *The matrix* $\mathcal{J}_M(\vartheta_0) = \lim_{n\to\infty} \nabla^2_\vartheta \widehat{\mathcal{L}}_M(\vartheta_0, Y^n)$ *exists, for* $i, k = 1, \ldots, r$. *Its components are given by*

$$(\mathcal{J}_M(\vartheta_0))_{i,k} = \mathbb{E}\left[\left(\sum_{j=1}^d \left(\sum_{l=1}^d (\partial^2_{j,l}\rho)(\epsilon_{\vartheta_0,1})(\partial_k\epsilon_{\vartheta_0,1})_k\right)(\partial_i\epsilon_{\vartheta_0,1})_j + (\partial_j\rho)(\epsilon_{\vartheta_0,1})(\partial^2_{i,k}\epsilon_{\vartheta_0,1})_j\right)\right],$$

where $(\partial_k \epsilon_{\vartheta_0,1})_k$ *denotes the* k*–th component of the vector* $\partial_k \epsilon_{\vartheta_0,1}$ *(and likewise for the other vectors involved).*

Proof. By Lemma 4.3c), the limit of $\nabla^2_\vartheta \widehat{\mathcal{L}}_M(\vartheta_0, Y^n)$ is the same as the one of $\nabla^2_\vartheta \mathcal{L}_M(\vartheta_0, Y^n)$. Observe that

$$(\nabla^2_\vartheta \mathcal{L}_M(\vartheta_0, Y^n))_{i,k} = \frac{1}{n}\sum_{k'=1}^{n}\sum_{j=1}^{d}\left(\left(\sum_{l=1}^{d}(\partial^2_{j,l}\rho)(\epsilon_{\vartheta_0,k'})(\partial_k \epsilon_{\vartheta_0,k'})_k\right)(\partial_i \epsilon_{\vartheta_0,k'})_j + (\partial_j \rho)(\epsilon_{\vartheta_0,k'})(\partial^2_{i,k}\epsilon_{\vartheta_0,k'})_j\right),$$

By Lemma 2.22b), each partial derivative $\partial_k \epsilon_{\vartheta_0,n}$ $(n \in \mathbb{Z})$ is a measurable function of the ergodic process $(Y(nh))_{n\in\mathbb{Z}}$. Hence, the process $(\partial_k \epsilon_{\vartheta_0,n})_{n\in\mathbb{Z}}$ is ergodic again. For the same reason, the process $(\partial^2_{i,k}\epsilon_{\vartheta_0,n})_{n\in\mathbb{Z}}$ is also ergodic by the moving average representation of Lemma 2.22c). By two applications of the Cauchy–Schwarz inequality, the moving average representations in Lemma 2.22b) and Lemma 2.22c) and the fact that the data–generating process $Y((nh))_{n\in\mathbb{Z}}$ has finite $(4+\delta)$–th moments we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\partial^2_{j,l}\rho)(\epsilon_{\vartheta_0,k'})(\partial_k\epsilon_{\vartheta_0,k'})_k(\partial_i\epsilon_{\vartheta_0,k'})_j + (\partial_j\rho)(\epsilon_{\vartheta_0,k'})(\partial^2_{i,k}\epsilon_{\vartheta_0,k'})_j\right\|\right] \\
&\overset{F.3}{\leq} C\mathbb{E}\left[(\|\epsilon_{\vartheta_0,k'}\| + \|(\partial^2_{j,l}\rho)(0)\|)\|\partial_k\epsilon_{\vartheta_0,k'}\|\|\partial_i\epsilon_{\vartheta_0,k'}\|\right] + C'\mathbb{E}\left[(\|\epsilon_{\vartheta_0,k'}\| + \|(\partial_j\rho)(0)\|)\|\partial^2_{i,k}\epsilon_{\vartheta_0,k'}\|\right] \\
&\leq C\mathbb{E}\left[(\|\epsilon_{\vartheta_0,k'}\| + \|(\partial^2_{j,l}\rho)(0)\|)^2\right]^{\frac{1}{2}}\mathbb{E}\left[\|\partial_k\epsilon_{\vartheta_0,k'}\|^2\|\partial_i\epsilon_{\vartheta_0,k'}\|^2\right]^{\frac{1}{2}} \\
&+ C'\mathbb{E}\left[(\|\epsilon_{\vartheta_0,k'}\| + \|(\partial_j\rho)(0)\|)^2\right]^{\frac{1}{2}}\mathbb{E}\left[\|\partial^2_{i,k}\epsilon_{\vartheta_0,k'}\|^2\right]^{\frac{1}{2}} \\
&\leq C\mathbb{E}\left[(\|\epsilon_{\vartheta_0,k'}\| + \|(\partial^2_{j,l}\rho)(0)\|)^2\right]^{\frac{1}{2}}\mathbb{E}\left[\|\partial_k\epsilon_{\vartheta_0,k'}\|^4\right]^{\frac{1}{4}}\mathbb{E}\left[\|\partial_i\epsilon_{\vartheta_0,k'}\|^4\right]^{\frac{1}{4}} \\
&+ C'\mathbb{E}\left[(\|\epsilon_{\vartheta_0,k'}\| + \|(\partial_j\rho)(0)\|)^2\right]^{\frac{1}{2}}\mathbb{E}\left[\|\partial^2_{i,k}\epsilon_{\vartheta_0,k'}\|^2\right]^{\frac{1}{2}} < \infty.
\end{aligned}$$

Hence, $(\mathcal{J}_{\mathrm{M}}(\vartheta_0))_{i,k}$ is finite for $i,k \in \{1,\ldots,r\}$. The assertion now follows from Birkhoff's Ergodic Theorem. □

We are now ready to state the theorem about the asymptotic distribution of the M-estimator:

Theorem 4.15. *Assume that the matrix* $\mathcal{J}_M(\vartheta_0)$ *from Lemma 4.14 is invertible. Then the M-estimator* $\widehat{\vartheta}^n_M$ *defined in* (4.3) *is asymptotically normally distributed, i. e.*

$$\sqrt{n}\left(\widehat{\vartheta}^n_M - \vartheta_0\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \Xi_M(\vartheta_0)),$$

where the asymptotic covariance matrix $\Xi_M(\vartheta_0) = \mathcal{J}_M(\vartheta_0)^{-1}\mathcal{I}_M(\vartheta_0)\mathcal{J}_M(\vartheta_0)^{-1}$ *is given*

by

$$\mathcal{I}_M(\vartheta_0) = \lim_{n\to\infty} n \operatorname{Var}(\nabla_\vartheta \mathcal{L}_M(\vartheta_0, Y^n)) \text{ and } \mathcal{J}_M(\vartheta_0) = \lim_{n\to\infty} \nabla^2_\vartheta \mathcal{L}_M(\vartheta_0, Y^n).$$

Proof. From Theorem 4.5 we know that $\widehat{\vartheta}^n_{\mathrm{M}}$ converges almost surely to ϑ_0. Moreover, the true parameter is an element of the interior of the parameter space by B.8 and hence the same is true for $\widehat{\vartheta}^n_{\mathrm{M}}$ for n large enough. Alas, the property that $\widehat{\vartheta}^n_{\mathrm{M}}$ minimizes $\widehat{\mathcal{L}}_M(\vartheta, Y^n)$ can be written as $\nabla_\vartheta \widehat{\mathcal{L}}_M(\widehat{\vartheta}^n_{\mathrm{M}}, Y^n) = 0$. Now, we do a Taylor expansion of the function $\nabla_\vartheta \widehat{\mathcal{L}}_M(\vartheta, Y^n)$ around the true parameter ϑ_0. This gives us the equation

$$0 = \sqrt{n}\nabla_\vartheta \widehat{\mathcal{L}}_M(\vartheta_0, Y^n) + \nabla^2_\vartheta \widehat{\mathcal{L}}_M(\overline{\vartheta}^n, Y^n)\sqrt{n}(\widehat{\vartheta}^n_{\mathrm{M}} - \vartheta_0),$$

where $\overline{\vartheta}^n$ is between ϑ_0 and $\widehat{\vartheta}^n_{\mathrm{M}}$ in the sense of the Euclidean norm. Moreover,

$$\|\nabla^2_\vartheta \widehat{\mathcal{L}}_M(\overline{\vartheta}^n, Y^n) - \nabla^2_\vartheta \mathcal{L}_M(\vartheta_0, Y^n)\| \leq \sup_{\vartheta\in\Theta} \|\nabla^3_\vartheta \mathcal{L}_M(\vartheta, Y^n)\| \|\overline{\vartheta}^n - \vartheta_0\|.$$

As in the proof of [Schlemm and Stelzer 2012, Theorem 2.5], it can be deduced from the compactness of Θ that the right–hand side converges almost surely to 0 as $n \to \infty$. Together with Lemma 4.13 and Lemma 4.14 and the assumed invertability of $\mathcal{J}_{\mathrm{M}}(\vartheta_0)$ this delivers the assertion of the theorem. □

An important special case of an M–estimator is the least squares estimator. Parameter estimation of CARMA processes via the least squares estimator has been studied in Brockwell et al. [2011], but only in the one–dimensional case and for subordinators as driving Lévy processes. We generalize these results to the multivariate framework and to more general Lévy processes. The least squares estimator fits the framework of this section by choosing $\rho(x) = \|x\|_2^2$ in (4.3). Then, it is clear that Assumption F is fulfilled, as

$$(\partial_i \rho)(x) = 2x_i \quad \text{and} \quad (\partial^2_{i,j}\rho)(x) = 2\delta_{i,j},\ \forall i,j \in \{1,\dots,d\},\ \forall x \in \mathbb{R}^d.$$

Next, we check that $\mathcal{Q}_{\mathrm{M}}(\vartheta)$ has a unique minimum at $\vartheta = \vartheta_0$:

Lemma 4.16. *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B and contains ϑ_0 with $Y_{\vartheta_0} = MCARMA(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$. Then, for $\rho(x) = \|x\|_2^2$, the function $\mathcal{Q}_M$ has a unique minimum at ϑ_0.*

Proof. It holds

$$\begin{aligned}
\mathscr{Q}_M(\vartheta) &= \mathbb{E}\left[\|\epsilon_{\vartheta,1} - \epsilon_{\vartheta_0,1} + \epsilon_{\vartheta_0,1}\|_2^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[((\epsilon_{\vartheta,1})_i - (\epsilon_{\vartheta_0,1})_i + (\epsilon_{\vartheta_0,1})_i)^2\right] \\
&= \sum_{i=1}^{d} \left(\mathbb{E}\left[((\epsilon_{\vartheta,1})_i - (\epsilon_{\vartheta_0,1})_i)^2\right] + \mathbb{E}\left[(\epsilon_{\vartheta_0,1})_i^2\right] + \mathrm{Cov}((\epsilon_{\vartheta,1})_i - (\epsilon_{\vartheta_0,1})_i; (\epsilon_{\vartheta_0,1})_i)\right) \\
&= \sum_{i=1}^{d} \left(\mathbb{E}\left[((\epsilon_{\vartheta,1})_i - (\epsilon_{\vartheta_0,1})_i)^2\right] + (V_{\vartheta_0})_{ii}\right) \\
&\geq \mathrm{tr}(V_{\vartheta_0}),
\end{aligned}$$

with equality if and only if $\vartheta = \vartheta_0$ due to B.6. Note that we used that the difference $\epsilon_{\vartheta,1} - \epsilon_{\vartheta_0,1}$ is an element of the Hilbert space spanned by $\{Y(kh), k \leq 0\}$ and that $\epsilon_{\vartheta_0,1}$ is orthogonal to this space by construction to drop out the covariance. This completes the proof. □

We can now easily obtain the following result on the asymptotic behavior of the least squares estimator:

Theorem 4.17. *Assume that the space Θ with associated family of continuous–time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B and contains ϑ_0 with $Y_{\vartheta_0} = MCARMA(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$.*
Denote

$$\widehat{\vartheta}^n_{LS} = \arg\min_{\vartheta\in\Theta} \frac{1}{n}\widehat{\mathcal{L}}_{LS}(\vartheta, Y^n) := \arg\min_{\vartheta\in\Theta} \frac{1}{n}\sum_{k=1}^{n} \|\widehat{\epsilon}_{\vartheta,k}\|^2.$$

Then, as $n \to \infty$,

$$\widehat{\vartheta}^n_{LS} \to \vartheta_0 \quad \mathbb{P}\text{-a.s.},$$

and

$$\sqrt{n}\left(\widehat{\vartheta}^n_{LS} - \vartheta_0\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Xi_{LS}(\vartheta_0)),$$

where

$$\Xi_{LS}(\vartheta_0) = \mathcal{J}^{-1}_{LS}(\vartheta_0)\mathcal{I}_{LS}(\vartheta_0)\mathcal{J}^{-1}_{LS}(\vartheta_0)$$

with

$$\mathcal{I}_{LS}(\vartheta_0) = \lim_{n\to\infty} n\,\mathrm{Var}(\nabla_\vartheta \mathcal{L}_{LS}(\vartheta, Y^n)) \quad \text{and} \quad \mathcal{J}_{LS}(\vartheta_0) = \lim_{n\to\infty} \nabla^2_\vartheta \mathcal{L}_{LS}(\vartheta, Y^n)).$$

Proof. The theorem is an immediate consequence of Theorem 4.5 and Theorem 4.15. □

Remark 4.18. *It is also possible to use the explicit representation of the partial derivatives of the function $\mathcal{L}_{LS}(\vartheta, Y^n)$ to obtain the same result on the asymptotic behavior of $\widehat{\vartheta}^n_{LS}$ by using the same techniques as in [Schlemm and Stelzer 2012]. Namely, Lemma 4.13 can be proven analogous to [Schlemm and Stelzer 2012, Lemma 2.14 and Lemma 2.16]. The principle idea is the same as when using the general theory on near epoch dependence: the process of interest can be approximated by a process that has the desired asymptotic behavior (because it depends on finitely many values of a strongly mixing process) and the approximation error is asymptotically negligible. In this sense, using the theory on NED processes naturally generalizes the ideas of Schlemm and Stelzer [2012] to more general objective functions ρ.*

Another example for an M–estimator is the so–called Tukey bisquare estimator. For a constant $k > 0$ it is defined by choosing the loss function ρ as

$$\rho(x) = \frac{k^2}{6}\left(1 - \left(1 - \frac{\|x\|^2}{k^2}\right)^3\right)\mathbb{1}_{\{\|x\|\leq k\}}.$$

Here, it holds that

$$\partial_i\rho(x) = \|x\|\left(1 - \frac{\|x\|^2}{k^2}\right)^2 x_i\mathbb{1}_{\{\|x\|\leq k\}}.$$

Since all partial derivatives are non–zero only on a compact set, it is obvious that each partial derivative satisfies the Lipschitz condition of Assumption F.3 by the multivariate mean value theorem. The same argument can be made for the second–order partial derivatives. In contrast to the least squares estimator, the loss function of the Tukey bisquare estimator is bounded. It will reappear later when we consider generalized M–estimators in the context of indirect estimation for CARMA processes (cf. Example 4.25).

In summary, we have seen that M–estimators can be defined for MCARMA processes and that the asymptotics can be studied with similar tools as in the maximum likelihood case. While these are pleasant and interesting results, they do not achieve what we truly want: an estimator that is robust towards outliers. This is true because it is well–known from the literature that this class of estimators is not robust towards additive or replacement outliers for ARMA(p,q) processes with $q > 0$. This is due to the fact that an outlier at time t has an effect on all innovations

for every $t' \geq t$ (Muler et al. [2009], [Maronna et al. 2006, Chapter 8]), which, in turn, is caused by the linear innovations of an ARMA(p,q) process being an infinite moving average sequence of all past observations. For discretely sampled MCARMA processes, one can on the one hand see them as weak VARMA processes ([Schlemm and Stelzer 2011, Theorem 4.2]) and on the other hand, their linear innovations are also infinite moving average sequences of the past (cp. Lemma 2.22a)). For these reasons, we immediately realize that the problems from the ARMA scenario carry over. We therefore need yet another ansatz if we want to obtain a robust estimator, which will be the topic of the following section.

4.3. Indirect estimation for CARMA processes

In the rest of this chapter, we restrict ourselves to CARMA processes, i.e. to the case $d = 1$. As mentioned at the end of the last section, constructing a robust estimator directly by using the discrete–time state space model or weak ARMA representation of a sampled CARMA process is difficult because the innovations are a moving average process of infinite order. One way out would be to calculate a different kind of "innovations" by using the robust filter as proposed by Masreliez [1975] and base the estimator on these robust innovations ([Maronna et al. 2006, Section 8.8]). However, this approach has the downside that it yields biased estimators and no asymptotic theory is available (Muler et al. [2009]). For this reason, we take yet another road and will make use of the so–called method of indirect inference, originally proposed by Smith [1993] and extended by Gouriéroux et al. [1993] (see also the overview in Gouriéroux and Monfort [1997]).

The core idea of the method is avoid estimating the parameter of interest directly. Originally, the authors named as reasons for this ansatz, for example, that a suitable likelihood function is not available or too complex to handle. The authors then propose to estimate an auxiliary, different parameter instead and use the resulting estimate in conjunction with simulated data to construct an estimator for the original parameter of interest. This method has been successfully used in different contexts, see e.g. [Gouriéroux and Monfort 1997, Chapter 4], Jiang and Turnbull [2004], de Luna and Genton [2001] and de Luna and Genton [2000]. The latter two papers recognized that it is possible to construct robust estimators via this approach, even for model classes where direct robust estimation is difficult (e.g. ARMA processes). The reason is that the auxiliary parameter, which is estimated from the actual, outlier–contaminated data, can be chosen as a parameter of a simpler model than the original one that admits robust estimation. In de Luna and Genton [2001] and also in our case, the auxiliary model is an autoregressive process, since robust estimation of this class of processes via so–called generalized M–estimators (GM estimators) is well–understood from a theoretical point of view. We therefore start by studying how we can obtain an auxiliary AR representation of a discretely sampled CARMA process.

4.3.1. The AR(r) representation of a CARMA process

In the following, we always assume that we operate in a parameter space Θ which fulfills Assumptions B.1 - B.9 and that there exists a parameter $\vartheta_0 \in \Theta$ with $Y = \text{CARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$. Since we are in one dimension, we do not need to

assume the Echelon form for the models in this parameter space. Instead, in the following we always work with the simpler identifiable parametrization introduced in Example 3.1, i.e. we use the coefficients $a_1, \dots, a_p$ and $b_0, \dots, b_q$ of P and Q as defined in Definition 2.2 as parameters and assume that for $\vartheta \in \Theta$ the polynomials P_ϑ and Q_ϑ have no common zeros. The corresponding matrices A_ϑ, B_ϑ and C_ϑ (which, in truth, is independent of ϑ) are then given as in Definition 2.3. For any $\vartheta \in \Theta$, the auxiliary AR(r) representation of the sampled processes $(Y_\vartheta(nh))_{n\in\mathbb{Z}}$ is defined as follows:

Definition and Proposition 4.19. *For $r \geq 2p-1$ we call*

$$\pi_\vartheta := (\pi_{\vartheta,1}, \dots, \pi_{\vartheta,r}, \sigma_\vartheta)$$

the auxiliary parameter of the AR(r) representation of $(Y_\vartheta(nh))_{n\in\mathbb{Z}}$ if the stationary process $(U_{\vartheta,n})_{n\in Z}$ defined by

$$U_{\vartheta,n} := Y_\vartheta(nh) - \sum_{i=1}^{r} \pi_{\vartheta,r+i-1} Y_\vartheta(h(n-i)). \tag{4.5}$$

with $\mathbb{E}[U_{\vartheta,1}] = 0$ and $\mathrm{Var}(U_{\vartheta,1}) = \sigma_\vartheta^2$ satisfies

$$\mathbb{E}\left[U_{\vartheta,r+1} Y_\vartheta((r+1-j)h)\right] = 0 \quad \forall j = 1, \dots, r. \tag{4.6}$$

For every $\vartheta \in \Theta$ and every $r \geq 2p-1$, π_ϑ exists and is unique.

Proof. First, we need to show that for any $r \in \mathbb{N}$, the covariance matrix of $(Y_\vartheta(h), \dots, Y_\vartheta((r+1)h))$ is non–singular. To see this, note that the autocovariance function of $(Y_\vartheta(nh))_{n\in\mathbb{Z}}$, which is a stationary process, is

$$\gamma_{Y_\vartheta}(z) = C_\vartheta \, \mathrm{e}^{A_\vartheta h z} \, \Gamma_0 C^T, \; z \in \mathbb{Z},$$

by [Schlemm and Stelzer 2012, Proposition 3.1] with $\Gamma_0 = \int_0^\infty \mathrm{e}^{A_\vartheta u} \, B_\vartheta \Sigma_\vartheta^L B_\vartheta^T \, \mathrm{e}^{A_\vartheta u} \, \mathrm{d}u$. Since the covariance matrix Σ_ϑ^L is non–singular by Assumption B.2 and C_ϑ has full rank by Assumption B.4 we have that $\gamma_{Y_\vartheta}(0) > 0$. Moreover, the eigenvalues of A_ϑ have strictly negative real part by B.3 and therefore $\gamma_{Y_\vartheta}(z) \to 0$ as $z \to \infty$ holds. By [Brockwell and Davis 1991, Proposition 5.1.1], it therefore follows that the covariance matrix of $(Y_\vartheta(h), \dots, Y_\vartheta((r+1)h))$ is non–singular for every $r \in \mathbb{N}$.
Having this, it follows from [Brockwell and Davis 1991, §8.1] that there exist unique numbers $\pi_{\vartheta,1}, \dots, \pi_{\vartheta,r}, \sigma_{\vartheta_0}^2$, which solve the set of $r+1$ Yule–Walker equations. These

equations state that $\mathrm{Var}(U_{\vartheta,1}) = \sigma_\vartheta^2$ and

$$0 = \mathbb{E}\left[(Y_\vartheta((r+1)h) - \sum_{i=1}^{r} \pi_{\vartheta,r+i-1} Y_\vartheta(h(r+1-i)) \begin{pmatrix} Y_\vartheta(rh) \\ \vdots \\ Y_\vartheta(h) \end{pmatrix} \right], \tag{4.7}$$

which shows the assertion because of the stationarity of the process $(Y_\vartheta(nh))_{n\in\mathbb{Z}}$. □

Remark 4.20. *Since*

$$Y_\vartheta(nh) = \sum_{i=1}^{r} \pi_{\vartheta,r+i-1} Y_\vartheta(h(n-i)) + U_{\vartheta,n},$$

$(U_{\vartheta,n})_{n\in Z}$ *can be interpreted as the noise sequence driving the auxiliary AR(r) representation of* $(Y_\vartheta(nh))_{n\in\mathbb{Z}}$. *Per construction, however, the sequence* $(U_{\vartheta,n})_{n\in Z}$ *is not an uncorrelated sequence – only uncorrelated with* $Y_\vartheta(h(n-1)), \ldots, Y_\vartheta(h(n-r))$. *From this property, it follows that* $U_{\vartheta,n}$ *can be interpreted as the error of the best linear predictor of* $Y_\vartheta(nh)$ *in terms of* $Y_\vartheta(h(n-1)), \ldots, Y_\vartheta(h(n-r))$.

For the asymptotic results, we will later need to study the map that links the original parameter space to the auxiliary one. The next definition introduces this map and states under which conditions the properties that we require later hold:

Definition and Proposition 4.21. *Define the parameter space* $\Pi \subseteq \mathbb{R}^{r+1}$ *as the set containing all possible parameter vectors of stationary, invertible AR(r) processes. The map* π *from* Θ *to* Π *with* $\vartheta \mapsto \pi_\vartheta$ *and* π_ϑ *as given in Definition 4.19 is called the link function or binding function.* π *is injective and continuously differentiable if* $r \geq 2p-1$.

Proof. We make use of the fact that a discretely observed CARMA(p,q) process $(Y_\vartheta(nh))_{n\in\mathbb{Z}}$ admits a representation as a stationary, invertible ARMA$(p,p-1)$ process with weak white noise of the form

$$P_{\mathrm{ARMA}}(B) Y_\vartheta(nh) = Q_{\mathrm{ARMA}}(B) \epsilon_{\vartheta,n}, \tag{4.8}$$

where $(\epsilon_{\vartheta,k})_{k\in\mathbb{Z}}$ is the innovations sequence of $(Y_\vartheta(kh))_{k\in\mathbb{Z}}$ (cf. Definition 2.19), $P_{\mathrm{ARMA}}(z) = \prod_{i=1}^{p}(1 - \mathrm{e}^{h\lambda_i} z)$, the λ_i being the eigenvalues of the matrix A_ϑ in (2.5) and Q_{ARMA} is a monic, Schur–stable polynomial (cp. [Schlemm and Stelzer 2011, Theorem 4.2] for the case that the eigenvalues of A are all distinct and [Brockwell and Lindner 2009, Lemma 2.1] for the general case). The coefficients of Q_{ARMA} can be calculated by identifying them with the coefficients of the invertible moving average

process whose autocorrelations at lags $1, \dots, p-1$ match those of $P_{\text{ARMA}}(B)Y_\vartheta(nh)$, see [Brockwell et al. 2011, Section 4]. We can now decompose the map $\pi : \Theta \to \Pi$ into three separate maps, for which we define the following spaces:

$$\begin{aligned}\mathcal{M} := \{&(a_1, \dots, a_p, b_1, \dots, b_{p-1}, \sigma) \in \mathbb{R}^{2p} : \text{ The coefficients define a weak} \\ &\text{ARMA}(p, p-1) \text{ model as in (4.8) for which } P_{\text{ARMA}} \text{ and } Q_{\text{ARMA}} \\ &\text{have no common zeros.}\} \subseteq \mathbb{R}^{2p} \\ \mathcal{G} := \{&\gamma = (\gamma_0, \dots, \gamma_r) \in \mathbb{R}^{r+1} : \text{ The coefficients define the autocovariances} \\ &\text{up to order } r \text{ of a stationary stochastic process}\} \subseteq \mathbb{R}^{r+1}\end{aligned}$$

Denote by $\pi_1 : \Theta \to \mathcal{M}$ the map which maps the parameters of a CARMA process to the coefficients of the weak ARMA$(p, p-1)$ representation of its sampled version. Denote by $\pi_2 : \mathcal{M} \to \mathcal{G}$ the map which maps the parameters of a weak ARMA$(p, p-1)$ process to its autocovariances of lags $0, \dots, r$. Lastly, denote by $\pi_3 : \mathcal{G} \to \Pi$ the map which maps a vector of autocovariances γ to the parameters of an AR(r) process of the form in (4.5). Then we have that $\pi = \pi_3 \circ \pi_2 \circ \pi_1$.
Because we assumed that the beginning of this section that Θ satisfies Assumption B, we obtain in particular that the family of processes $((Y_\vartheta(nh))_{n\in\mathbb{Z}}, \vartheta \in \Theta)$ is identifiable (cf. Theorem 2.21). Since each of the sampled processes admits exactly one weak ARMA$(p, p-1)$ representation, the map π_1 is injective.
Next, the map π_2 is injective if $r \geq p + p - 1 = 2p - 1$. The reason is that by the method of [Brockwell and Davis 1991, p. 93], the autocovariances $\gamma(k)$, $k \in \mathbb{Z}$, of an ARMA$(p, p-1)$ process are completely determined as the solution of a difference equation with p boundary conditions, which depend on the coefficient vector $(a_1, \dots, a_p, b_1, \dots, b_{p-1}, \sigma)$. Hence, viewing $(\gamma_0, \dots, \gamma_r)$ as solution of those recursive equations, injectivity of π_2 always holds if $r \geq 2p-1$, since then the number of equations is greater then or equal to the number of variables (see also [de Luna and Genton 2001, Section 4.1]). Finally, the map π_3 is even bijective, because it is defined by the Yule–Walker–equations for an AR(r) process. Hence, π is injective as a composition of three injective maps if $r \geq 2p - 1$. The differentiability of the map π follows from the fact that each of the maps π_1, π_2 and π_3 is differentiable. For π_1, this follows from Assumption B.4 and the fact that the coefficients of the weak ARMA representation are determined by the CARMA parameters. Notably, the AR parameters are determined by the eigenvalues of the matrix A_ϑ and the MA parameters can be obtained from the CARMA parameters as in the proof of [Brockwell and Davis 1991, Proposition 3.2.1], see also [Brockwell et al. 2011, Proposition 3]. From the algorithmic description of π_2 given by [Brockwell and

Davis 1991, p. 93], one sees immediately that this map is differentiable, too. The differentiability of π_3 follows from the fact that it is defined by the Yule–Walker equations, which are differentiable with respect to γ by construction. □

In the following we always assume that $r \geq 2p - 1$.

4.3.2. Definition and asymptotics of the indirect estimator

We will now introduce the main idea of this section. Remember that our aim is to obtain an estimator for the parameters of a CARMA process that is well–performing in the absence of outliers in the data as well as in the presence of outliers, i. e. robust. To achieve this goal, as explained at the start of the section, we will now construct an indirect estimator. Moreover, we assume in this subsection that there are no outliers in the data, i.e. we have observations $Y^n = (Y(h), \ldots, Y(nh))$ from the data–generating process $(Y(kh))_{k\in\mathbb{Z}}$. We do this in order to first study the asymptotic behavior of the indirect estimator in the absence of outliers.
For fixed r, denote by $\widehat{\pi}^n$ an estimator of π_{ϑ_0} that is calculated from the observations Y^n. Possible choices for this estimator will be discussed later in Subsection 4.3.3. With regard to robustness, we will later choose a robust estimator in this step. If we could analytically invert the link function π and calculate $\pi^{-1}(\widehat{\pi}^n)$, then we would be able to obtain an estimator of ϑ_0 since $\widehat{\pi}^n$ estimates π_{ϑ_0}. However, this is not possible in general since no analytic representation of π exists. To overcome this problem, we now perform a second estimation, which is based on simulations and constitutes the other building block of indirect estimation.
We fix a number $s \in \mathbb{N}$ and simulate a sample path of length sn of a Lévy process $(L_S(t))_{t\in\mathbb{R}}$, say. We assume that this Lévy process satisfies $\mathbb{E}[|L_S(1)|^{4+\delta}] < \infty$ for a $\delta > 0$. Then, for a fixed parameter $\vartheta \in \Theta$ we then generate a sample path of the associated CARMA process, using the simulated path of L_S. This gives us a vector of "pseudo–observations" $(Y_\vartheta(h), \ldots, Y_\vartheta(snh))$ of length sn.
Note that, strictly speaking, Y_ϑ is an abuse of notation here, since the parameter $\vartheta \in \Theta$ also contains elements that parametrize the Lévy process L_ϑ which drives Y_ϑ. Of course, if we use the fixed Lévy process L_S, those parameters are not used. However, we chose not to introduce another new notation here and instead keep ϑ as parameter, remembering that we only really parametrize the corresponding matrices A_ϑ, B_ϑ and C_ϑ with it from here on. This of course has the consequence that the indirect estimator will, eventually, only give us an estimator for the entries of ϑ which parametrize A_ϑ, B_ϑ and C_ϑ. If one is also interested in the driving Lévy process, one

would then have to resort to another estimation later on (cf. Brockwell and Schlemm [2013]).
Returning to the pseudo–observations, we can calculate an estimate of π_ϑ by applying an estimator to the pseudo–observations $(Y_\vartheta(h), \dots, Y_\vartheta(snh))$. We do not need to use a robust estimator in this step, because we have simulated the observations and therefore can be sure that no outliers are present in data. We denote this estimator by $\widehat{\pi}^n_{\mathrm{S}}(\vartheta)$. Obviously, varying ϑ, while keeping n and s fixed, will result in different values of $\widehat{\pi}^n_{\mathrm{S}}(\vartheta)$. This is a deciding observation, because the idea is now to choose that value of ϑ as estimate for ϑ_0 which minimizes a suitable distance between $\widehat{\pi}^n$ and $\widehat{\pi}^n_{\mathrm{S}}(\vartheta)$. The formal definition is as follows:

Definition 4.22. *Let $\widehat{\pi}^n$ be an estimator of π_{ϑ_0} calculated from the data Y^n, let $\widehat{\pi}^n_S(\vartheta)$ be an estimator of π_ϑ calculated from the "pseudo–observations" $(Y_\vartheta(h), \dots, Y_\vartheta(snh))$ and let Ω be a symmetric, positive definite weighting matrix. Define the function $\mathcal{L}_{Ind} : \Theta \to [0, \infty)$ by*

$$\mathcal{L}_{Ind}(\vartheta, Y^n) := \operatorname*{arg\,min}_{\vartheta \in \Theta} (\widehat{\pi}^n - \widehat{\pi}^n_S(\vartheta))^T \Omega (\widehat{\pi}^n - \widehat{\pi}^n_S(\vartheta)). \tag{4.9}$$

Then, the indirect estimator of ϑ_0 is defined as

$$\widehat{\vartheta}^n_{Ind} = \operatorname*{arg\,min}_{\vartheta \in \Theta} \mathcal{L}_{Ind}(\vartheta, Y^n). \tag{4.10}$$

Note that the function $\mathcal{L}_{\mathrm{Ind}}$, which we minimize to obtain the indirect estimator, is not a likelihood function as the notation might suggest, but we nevertheless chose this symbol to keep the notation as consistent as possible with the concepts used earlier in the thesis. Figure 4.1 gives an overview of the indirect estimation procedure (inspired by [de Luna and Genton 2001, Figure 1]).

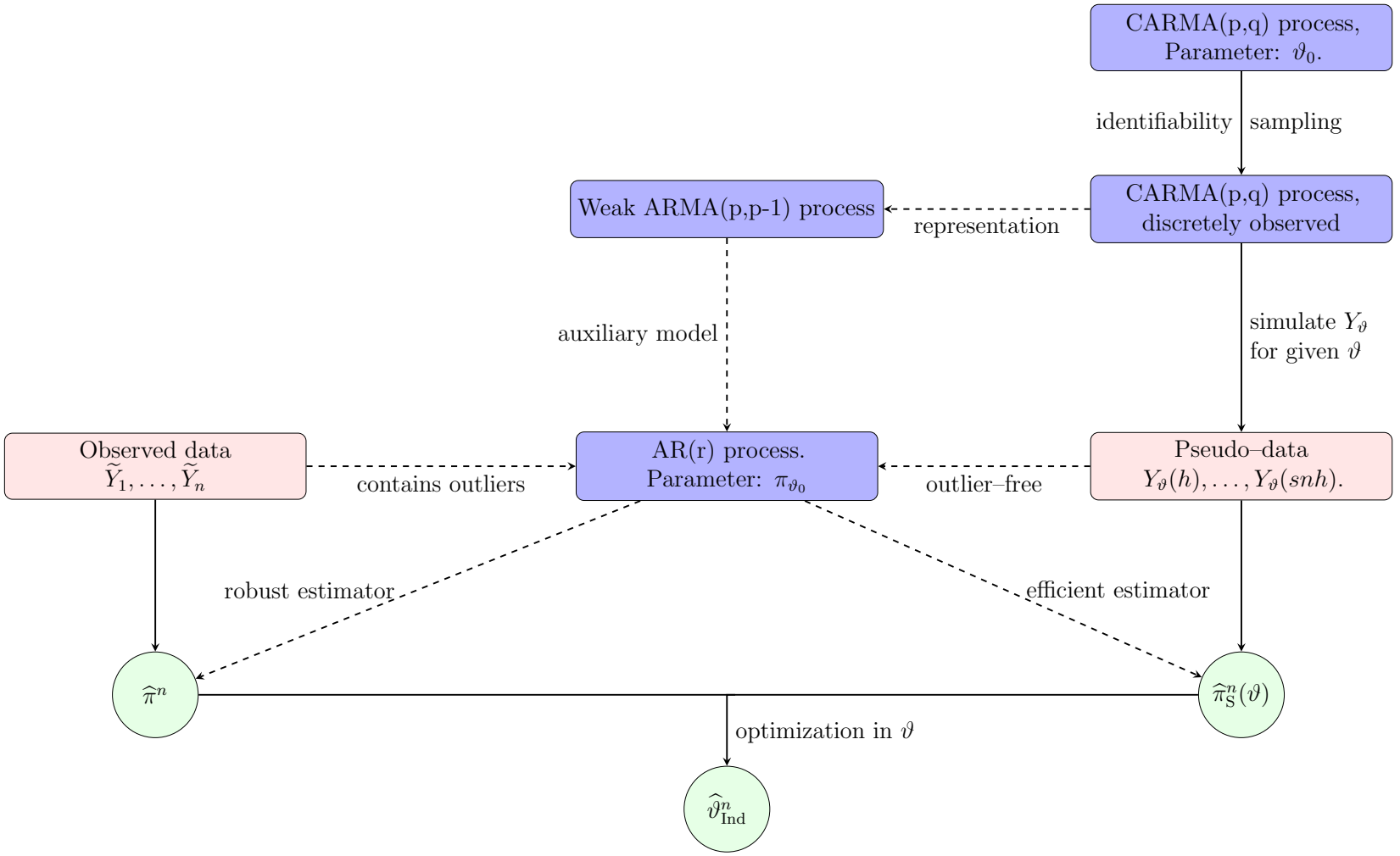

Figure 4.1.: Flowchart of the indirect estimation procedure

With respect to the asymptotic behavior of the indirect estimator, the basic conclusion is now that if the two estimators used in the construction are strongly consistent and asymptotically normally distributed, then the indirect estimator will be strongly consistent for ϑ_0 and also asymptotically normally distributed. More precisely, we have the following theorem:

Theorem 4.23. *a) Assume that $r \geq 2p-1$ and that*

$$\widehat{\pi}^n \overset{n\to\infty}{\longrightarrow} \pi_{\vartheta_0} \ \mathbb{P}\text{-}a.s. \tag{4.11}$$

Analogously, assume that

$$\sup_{\vartheta\in\Theta} \|\widehat{\pi}_S^n(\vartheta) - \pi_\vartheta\| \overset{n\to\infty}{\longrightarrow} 0, \ \mathbb{P}\text{-}a.s. \tag{4.12}$$

Then it holds that

$$\widehat{\vartheta}_{Ind}^n \longrightarrow \vartheta_0 \quad \mathbb{P}\text{-}a.s., \ n\to\infty. \tag{4.13}$$

b) In addition to the assumptions of part a), assume that for each $n \in \mathbb{N}$ the map $\vartheta \mapsto \widehat{\pi}_S^n(\vartheta)$ is continuously differentiable. Moreover, assume for every $\vartheta \in \Theta$ that

$$\sqrt{ns}(\widehat{\pi}_S^n(\vartheta) - \pi_\vartheta) \overset{\mathcal{D}}{\longrightarrow} \mathcal{N}(0, \Xi_S(\pi_\vartheta)), \ n\to\infty, \tag{4.14}$$

that

$$\sqrt{n}(\widehat{\pi}^n - \pi_{\vartheta_0}) \overset{\mathcal{D}}{\longrightarrow} \mathcal{N}(0, \Xi_D(\pi_{\vartheta_0})), \ n\to\infty, \tag{4.15}$$

and that for any sequence $(\overline{\vartheta}^n)_{n\in\mathbb{N}}$ with $\overline{\vartheta}^n \longrightarrow \vartheta_0$ $\mathbb{P}$-a.s. it also holds

$$\nabla_\vartheta \widehat{\pi}_S^n(\overline{\vartheta}^n) \longrightarrow \nabla_\vartheta \pi_{\vartheta_0}, \ \mathbb{P}\text{-}a.s., n\to\infty. \tag{4.16}$$

Then it holds that

$$\sqrt{n}(\widehat{\vartheta}_{Ind}^n - \vartheta_0) \overset{\mathcal{D}}{\longrightarrow} \mathcal{N}(0, \Xi_{Ind}(\vartheta_0)), \ n\to\infty, \tag{4.17}$$

where

$$\Xi_{Ind}(\vartheta_0) = (\mathcal{J}_{Ind}(\vartheta_0))^{-1}\mathcal{I}_{Ind}(\vartheta_0)(\mathcal{J}_{Ind}(\vartheta_0))^{-1}$$

for

$$\mathcal{J}_{Ind}(\vartheta_0) = \nabla_\vartheta \pi_{\vartheta_0}^T \Omega \nabla_\vartheta \pi_{\vartheta_0}$$

and

$$\mathcal{I}_{Ind}(\vartheta_0) = (\nabla_\vartheta \pi_{\vartheta_0})^T \Omega \left(\Xi_D(\pi_{\vartheta_0}) + \frac{1}{s}\Xi_S(\pi_{\vartheta_0}) \right) \Omega \nabla_\vartheta \pi_{\vartheta_0}.$$

Proof. a) We first start by proving the consistency. To this end, we define the function

$$\begin{aligned} \mathscr{Q}_{\mathrm{Ind}} : \Theta &\to [0,\infty) \\ \vartheta &\mapsto (\pi_{\vartheta_0} - \pi_\vartheta)^T \Omega (\pi_{\vartheta_0} - \pi_\vartheta). \end{aligned} \tag{4.18}$$

With this, we than have that

$$\begin{aligned} &\sup_{\vartheta\in\Theta} |\mathcal{L}_{\mathrm{Ind}}(\vartheta, Y^n) - \mathscr{Q}_{\mathrm{Ind}}(\vartheta)| \\ &= \sup_{\vartheta\in\Theta} |(\widehat{\pi}^n - \widehat{\pi}^n_{\mathrm{S}}(\vartheta))^T \Omega (\widehat{\pi}^n - \widehat{\pi}^n_{\mathrm{S}}(\vartheta)) - (\pi_{\vartheta_0} - \pi_\vartheta)^T \Omega (\pi_{\vartheta_0} - \pi_\vartheta)| \\ &\leq |(\widehat{\pi}^n)^T \Omega \widehat{\pi}^n - \pi_{\vartheta_0}^T \Omega \pi_{\vartheta_0}| + \sup_{\vartheta\in\Theta} |(\widehat{\pi}^n_{\mathrm{S}}(\vartheta))^T \Omega \widehat{\pi}^n - \pi_\vartheta^T \Omega \pi_{\vartheta_0}| \\ &+ \sup_{\vartheta\in\Theta} |(\widehat{\pi}^n)^T \Omega \widehat{\pi}^n_{\mathrm{S}}(\vartheta) - \pi_{\vartheta_0}^T \Omega \pi_\vartheta| + \sup_{\vartheta\in\Theta} |(\widehat{\pi}^n_{\mathrm{S}}(\vartheta))^T \Omega \widehat{\pi}^n_{\mathrm{S}}(\vartheta) - \pi_\vartheta^T \Omega \pi_\vartheta| \end{aligned}$$

The four summands on the right–hand side each converge $\mathbb{P}$-a.s. to zero as $n \to \infty$. For the first one, this is a consequence of the assumed strong consistency of $\widehat{\pi}^n$. For the remaining three, the arguments are similar, so that we treat only the second one exemplary. In the following $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product. We have

$$\begin{aligned} &\sup_{\vartheta\in\Theta} |(\widehat{\pi}^n_{\mathrm{S}}(\vartheta))^T \Omega \widehat{\pi}^n - \pi_\vartheta^T \Omega \pi_{\vartheta_0}| \\ &= \sup_{\vartheta\in\Theta} |(\widehat{\pi}^n_{\mathrm{S}}(\vartheta))^T \Omega \widehat{\pi}^n - \pi_\vartheta^T \Omega \widehat{\pi}^n + \pi_\vartheta^T \Omega \widehat{\pi}^n - \pi_\vartheta^T \Omega \pi_{\vartheta_0}| \\ &\leq \sup_{\vartheta\in\Theta} |(\widehat{\pi}^n_{\mathrm{S}}(\vartheta))^T \Omega \widehat{\pi}^n - \pi_\vartheta^T \Omega \widehat{\pi}^n| + \sup_{\vartheta\in\Theta} |\pi_\vartheta^T \Omega \widehat{\pi}^n - \pi_\vartheta^T \Omega \pi_{\vartheta_0}| \\ &= \sup_{\vartheta\in\Theta} |\langle \Omega^{\frac{1}{2}} (\widehat{\pi}^n_{\mathrm{S}}(\vartheta) - \pi_\vartheta), \Omega^{\frac{1}{2}} \widehat{\pi}^n \rangle| + \sup_{\vartheta\in\Theta} |\langle \Omega^{\frac{1}{2}} \pi_\vartheta, \Omega^{\frac{1}{2}} (\widehat{\pi}^n - \pi_{\vartheta_0}) \rangle| \\ &\leq \sup_{\vartheta\in\Theta} \|\Omega^{\frac{1}{2}} (\widehat{\pi}^n_{\mathrm{S}}(\vartheta) - \pi_\vartheta)\|^2 \|\Omega^{\frac{1}{2}} \widehat{\pi}^n\|^2 + \sup_{\vartheta\in\Theta} \|\Omega^{\frac{1}{2}} \pi_\vartheta\|^2 \|\Omega^{\frac{1}{2}} (\widehat{\pi}^n - \pi_{\vartheta_0})\|^2 \\ &\to 0, \ \mathbb{P}\text{-a.s.}, \ n \to \infty. \end{aligned}$$

Here, we used the Cauchy–Schwarz inequality, the fact that $\sup_{\vartheta\in\Theta} \|\Omega^{\frac{1}{2}} \pi_\vartheta\|$ is finite by the continuity of the map π and the compactness of Θ as well as both (4.11) and (4.12).

Therefore, the function $\mathcal{L}_{\mathrm{Ind}}(\vartheta, Y^n)$ converges uniformly in ϑ almost surely to the limiting function $\mathscr{Q}_{\mathrm{Ind}}(\vartheta)$. Per construction, $\widehat{\vartheta}^n_{\mathrm{Ind}}$ minimizes $\mathcal{L}_{\mathrm{Ind}}(\vartheta, Y^n)$ and $\mathscr{Q}_{\mathrm{Ind}}$ has a unique minimum at $\vartheta = \vartheta_0$ (since Ω is positive definite and the map π is injective). Therefore, strong consistency of $\widehat{\vartheta}^n_{\mathrm{Ind}}$ follows by arguing once more as in the proof of [Schlemm and Stelzer 2012, Theorem 2.4].

b) For the asymptotic normality, note that

$$\sqrt{n}(\widehat{\pi}^n - \widehat{\pi}^n_{\mathrm{S}}(\vartheta_0)) = \sqrt{n}(\widehat{\pi}^n - \pi_{\vartheta_0}) + \sqrt{n}(\pi_{\vartheta_0} - \widehat{\pi}^n_{\mathrm{S}}(\vartheta_0)).$$

From (4.15), we know that

$$\sqrt{n}(\widehat{\pi}^n - \pi_{\vartheta_0}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Xi_{\mathrm{D}}(\pi_{\vartheta_0}))$$

and from (4.14) we know that

$$\sqrt{n}(\widehat{\pi}^n_{\mathrm{S}}(\vartheta_0) - \pi_{\vartheta_0})) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{s}\Xi_S(\pi_{\vartheta_0})\right),\ n \to \infty.$$

Since both estimators are independent from each other, it follows that

$$\sqrt{n}(\widehat{\pi}^n - \widehat{\pi}^n_{\mathrm{S}}(\vartheta_0)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Xi_{\mathrm{D}}(\pi_{\vartheta_0}) + \frac{1}{s}\Xi_S(\pi_{\vartheta_0})\right). \tag{4.19}$$

The defining equation (4.10) can also be expressed as

$$0 = \nabla_\vartheta \mathcal{L}_{\mathrm{Ind}}(\vartheta, Y^n)\big|_{\vartheta=\widehat{\vartheta}^n_{\mathrm{Ind}}} = (\nabla_\vartheta \widehat{\pi}^n_{\mathrm{S}}(\widehat{\vartheta}^n_{\mathrm{Ind}}))^T \Omega(\widehat{\pi}^n - \widehat{\pi}^n_{\mathrm{S}}(\widehat{\vartheta}^n_{\mathrm{Ind}})).$$

We now use a Taylor expansion of order 1 around the true value ϑ_0 to obtain:

$$\begin{aligned}
0 &= \sqrt{n}\nabla_\vartheta \mathcal{L}_{\mathrm{Ind}}(\widehat{\vartheta}^n_{\mathrm{Ind}}, Y^n) \\
&= \sqrt{n}\nabla_\vartheta \mathcal{L}_{\mathrm{Ind}}(\vartheta_0, Y^n) + \sqrt{n}\nabla^2_\vartheta \mathcal{L}_{\mathrm{Ind}}(\overline{\vartheta}^n, Y^n)(\widehat{\vartheta}^n_{\mathrm{Ind}} - \vartheta_0) \\
&= (\nabla_\vartheta \widehat{\pi}^n_{\mathrm{S}}(\vartheta_0))^T \Omega \sqrt{n}(\widehat{\pi}^n - \widehat{\pi}^n_{\mathrm{S}}(\vartheta_0)) - (\nabla_\vartheta \widehat{\pi}^n_{\mathrm{S}}(\overline{\vartheta}^n))^T \Omega (\nabla_\vartheta \widehat{\pi}^n_{\mathrm{S}}(\overline{\vartheta}^n)) \sqrt{n}(\widehat{\vartheta}^n_{\mathrm{Ind}} - \vartheta_0).
\end{aligned}$$

Here, $\overline{\vartheta}^n$ is such that $\|\overline{\vartheta}^n - \vartheta_0\| \leq \|\widehat{\vartheta}^n_{\mathrm{Ind}} - \vartheta_0\|$. By the strong consistency of $\widehat{\vartheta}^n_{\mathrm{Ind}}$ we have $\overline{\vartheta}^n \to \vartheta_0$ $\mathbb{P}$-a.s. for $n \to \infty$. We rewrite this equation to

$$\sqrt{n}(\widehat{\vartheta}^n_{\mathrm{Ind}} - \vartheta_0) = ((\nabla_\vartheta \widehat{\pi}^n_{\mathrm{S}}(\overline{\vartheta}^n))^T \Omega (\nabla_\vartheta \widehat{\pi}^n_{\mathrm{S}}(\overline{\vartheta}^n)))^{-1} (\nabla_\vartheta \widehat{\pi}^n_{\mathrm{S}}(\vartheta_0))^T \Omega \sqrt{n}(\widehat{\pi}^n - \widehat{\pi}^n_{\mathrm{S}}(\vartheta_0))$$

By (4.19), the fact that $\widehat{\pi}^n_{\mathrm{S}}(\cdot)$ converges almost surely uniformly to $\pi_{(\cdot)}$ ((4.12)), $\overline{\vartheta}^n \to \vartheta_0$ $\mathbb{P}$-a.s. and (4.16), we obtain as $n \to \infty$

$$\sqrt{n}(\widehat{\vartheta}^n_{\mathrm{Ind}} - \vartheta_0) \xrightarrow{\mathcal{D}} (\nabla_\vartheta \pi^T_{\vartheta_0} \Omega \nabla_\vartheta \pi_{\vartheta_0})^{-1} (\nabla_\vartheta \pi_{\vartheta_0})^T \Omega \cdot \mathcal{N}\left(0, \Xi_{\mathrm{D}}(\pi_{\vartheta_0}) + \frac{1}{s}\Xi_S(\pi_{\vartheta_0})\right). \tag{4.20}$$

This completes the proof.

□

Remark 4.24.

a) The asymptotic covariance matrix can also be written as

$$\Xi_{Ind}(\vartheta_0) = \mathcal{H}(\vartheta_0)\left(\Xi_D(\pi_{\vartheta_0}) + \frac{1}{s}\Xi_S(\pi_{\vartheta_0})\right)\mathcal{H}(\vartheta_0)^T,$$

where

$$\mathcal{H}(\vartheta_0) = (\nabla_\vartheta \pi_{\vartheta_0}^T \Omega \nabla_\vartheta \pi_{\vartheta_0})^{-1} \nabla_\vartheta \pi_{\vartheta_0}^T \Omega. \tag{4.21}$$

This is the analog of the form given in [de Luna and Genton 2001, Eq. (4.4)].

b) Note that the asymptotic results hold for every $r \geq 2p-1$, but increasing the auxiliary AR order does not necessarily yield better results. The results also hold for all $s \in \mathbb{N}$, and the asymptotic covariance matrix of $\widehat{\vartheta}_{Ind}^n$ does explicitly depend on s. A consequence is that for $s \to \infty$, we have that $\Xi_{Ind}(\vartheta_0) \to \mathcal{H}(\vartheta_0)\Xi_D(\pi_{\vartheta_0})\mathcal{H}(\vartheta_0)^T$. From the defining formula, we then see that there is an optimal choice for Ω, namely $\Omega = (\Xi_D(\pi_{\vartheta_0}))^{-1}$, in which case we obtain that

$$\Xi_{Ind}(\vartheta_0) \to (\nabla_\vartheta \pi_{\vartheta_0}^T (\Xi_D(\pi_{\vartheta_0}))^{-1} \nabla_\vartheta \pi_{\vartheta_0})^{-1},\ s \to \infty.$$

An estimator of $\Xi_{Ind}(\vartheta_0)$ can then be obtained by plugging in a consistent estimator of $\Xi_D(\pi_{\vartheta_0})$ and approximating $\nabla_\vartheta \pi_{\vartheta_0}$ numerically (de Luna and Genton [2001], Remark 3-5).

4.3.3. ESTIMATING THE AR(R) REPRESENTATION OF A CARMA PROCESS

As evident from the previous subsection, in order to apply the indirect estimator for a CARMA process, we need strongly consistent and asymptotically normally distributed estimators for its AR(r) representation. In this subsection, we will study three such estimators: the class of generalized M–estimators (GM), the least squares (LS) estimator and the quasi maximum likelihood estimator (QMLE). Ultimately, we have in mind to use an indirect estimator for which a GM estimator is applied to the outlier–affected data and the LS or QMLE estimator is used for the simulated data, because this will give us a robust estimator for the parameters of a CARMA process. Before we can do this, though, we study the asymptotic theory of the three aforementioned estimators and start with the treatment of GM estimators.

4.3.3.1. GENERALIZED M–ESTIMATORS

In this subsection, we largely follow the approach of Bustos [1982], who develops the theory of GM estimators for autoregressive processes. In doing so, we will, however, make some slightly different assumptions as in that paper, since some of them would be too restrictive in our case. Since, ultimately, the GM estimator will be applied to a CARMA process afflicted by outliers, we study the theory of the GM estimators not only for a perfectly observed CARMA process. Instead, we work with the general replacement model as defined in (4.2). We need to make some assumptions:

Assumption G.

G.1 The processes $(V_n)_{n\in\mathbb{Z}}$ and $(Z_n)_{n\in\mathbb{Z}}$ are strictly stationary with $\mathbb{E}[|V_1|^{2+\delta}] < \infty$ and $\mathbb{E}[|Z_1|^{2+\delta}] < \infty$ for some $\delta > 0$.

G.2 $(V_n)_{n\in\mathbb{Z}}$ and $(Z_n)_{n\in\mathbb{Z}}$ are exponentially strong mixing, i. e. $\alpha_V(m) \leq C\rho^m$ and $\alpha_Z(m) \leq C\rho^m$ for some $C > 0$, $\rho \in (0,1)$ and every $m \in \mathbb{N}$.

G.3 Either the processes $(Y(nh))_{n\in\mathbb{Z}}$, $(V_n)_{n\in\mathbb{Z}}$ and $(Z_n)_{n\in\mathbb{Z}}$ are jointly independent or we have that $Z_n = Y(nh) + W_n$ for a process $(W_n)_{n\in\mathbb{Z}}$ such that $(Y(nh))_{n\in\mathbb{Z}}$, $(V_n)_{n\in\mathbb{Z}}$ and $(W_n)_{n\in\mathbb{Z}}$ are jointly independent.

G.4 It holds that

$$\mathbb{P}(a\widetilde{Y}_{r+1} + \pi_r\widetilde{Y}_r + \ldots + \pi_1\widetilde{Y}_1 = 0) = 0$$

for all $a \in \mathbb{R}, \pi \in \mathbb{R}^r$ with $|a| + \|\pi\| > 0$.

These assumptions are the analog to [Bustos 1982, Assumption M1)-M5)]. The biggest difference is that the role of the process $(Y_n)_{n\in\mathbb{Z}}$ used by Bustos [1982] is taken by the sampled CARMA process $(Y(nh))_{n\in\mathbb{Z}}$ in our case. Bustos [1982] requires the process $(Y_n)_{n\in\mathbb{Z}}$ to be an infinite–order moving average of a Φ–mixing noise sequence, which is generally not fulfilled by a sampled CARMA process. However, we already know from Proposition 2.24 that sampled CARMA processes are exponentially strong mixing (Φ–mixing is the stronger notion, i.e. it implies strong mixing, but the converse is not true). Therefore, the exponential strong mixing assumption on the processes $(V_n)_{n\in\mathbb{Z}}$ and $(Z_n)_{n\in\mathbb{Z}}$ is more natural in our scenario. The mixing assumption made by Bustos is used at only one point to obtain a central limit theorem. We will see later that our assumptions give us the same kind of limit theorem and are therefore suitable (see Lemma 4.30). Additionally, we require higher order moments and independence of $(Y(nh))_{n\in\mathbb{Z}}$, $(V_n)_{n\in\mathbb{Z}}$ and $(Z_n)_{n\in\mathbb{Z}}$ and not only independence of the latter two from the noise driving $(Y(nh))_{n\in\mathbb{Z}}$, which are somewhat stronger assumptions than in Bustos [1982].

To define the GM estimators, let now two functions $\phi : \mathbb{R}^r \times \mathbb{R} \to \mathbb{R}$ and $\chi : \mathbb{R} \to \mathbb{R}$ be given. Conditions on these two functions will be imposed later. Moreover, assume that we have observations $\widetilde{Y}^n = (\widetilde{Y}_1, \widetilde{Y}_2, \ldots, \widetilde{Y}_n)$ from the process defined in (4.2). We define the parameters to be estimated as the solutions of the system

$$\mathbb{E}\left[\phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1} - \pi_{0,r}\widetilde{Y}_r - \ldots - \pi_{0,1}\widetilde{Y}_1}{\sigma_0}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}\right] = 0, \tag{4.22}$$

$$\mathbb{E}\left[\chi\left(\left(\frac{\widetilde{Y}_{r+1} - \pi_{0,r}\widetilde{Y}_r - \ldots - \pi_{0,1}\widetilde{Y}_1}{\sigma_0}\right)^2\right)\right] = 0. \tag{4.23}$$

and denote them by

$$\pi_0 = (\pi_{0,1}, \ldots, \pi_{0,r}, \sigma_0).$$

The interpretation is that the $\pi_{0,i}$ are the coefficients of the AR polynomial $1 + \pi_{0,r}z + \ldots + \pi_{0,1}z^r$ and σ_0 is a scale parameter of the innovation sequence of the fitted AR(r) process. Note that π_0 and σ_0 in general depend on the processes $(V_n)_{n\in\mathbb{Z}}$ and $(Z_n)_{n\in\mathbb{Z}}$. We choose not to indicate this in the notation to make the exposition more readable, and will instead highlight this fact explicitly in the text when it is necessary. Now,

the GM estimate $\widehat{\pi}^n$ based on ϕ and χ is defined as the solution of the equations

$$\frac{1}{n-r}\sum_{t=1}^{n-r}\phi\left(\begin{pmatrix}\widetilde{Y}_t\\ \vdots \\ \widetilde{Y}_{t+r-1}\end{pmatrix}, \frac{\widetilde{Y}_{t+r}-\widehat{\pi}_r^n\widetilde{Y}_{t+r-1}-\ldots-\widehat{\pi}_1^n\widetilde{Y}_t}{\widehat{\sigma}^n}\right)\begin{pmatrix}\widetilde{Y}_t\\ \vdots \\ \widetilde{Y}_{t+r-1}\end{pmatrix}=0 \quad (4.24)$$

$$\frac{1}{n-r}\sum_{t=1}^{n-r}\chi\left(\left(\frac{\widetilde{Y}_{t+r}-\widehat{\pi}_r^n\widetilde{Y}_{t+r-1}-\ldots-\widehat{\pi}_1^n\widetilde{Y}_t}{\widehat{\sigma}^n}\right)^2\right)=0 \quad (4.25)$$

Here, $\widehat{\sigma}^n$ is an estimate of the scale of the innovations of the AR(r) process fitted to $\widetilde{Y}^n$ while the $\widehat{\pi}_i^n$ estimate the parameters of the AR polynomial. Next, we give some examples for GM estimators:

Example 4.25.

a) There are two main classes of GM estimators, the so–called Mallows estimators and the Hampel–Krasker–Welsch estimators. More information on them can be found in Bustos [1982], Denby and Martin [1979], Martin [1980] and Martin and Yohai [1986]. In the literature, this kind of estimators sometimes appear under the name BIF (for bounded influence) estimators. The class of Mallows estimators was originally proposed in Mallows [1975] for the regression setup for non–dependent data and later generalized to the time series setting. They are defined by choosing

$$\phi(y,u)=w(y)\psi(u),$$

where w is a strictly positive weight function and ψ is a suitably chosen robustifying function. The other class consists of the Hampel–Krasker–Welsch estimators for which one chooses

$$\phi(y,u)=\frac{\psi(w(y)u)}{w(y)},$$

where w is again a weight function and ψ again is a suitably chosen bounded function. In the original study of this estimator, Hampel [1978] and Krasker and Welsch [1982] used $\psi=\psi_k$ (as defined in part b) below) and $w(y)=\frac{1}{\|y\|}$ as a special case.

b) Typical choices for ψ are the so–called Huber ψ_k–functions, see e.g. [Maronna et al. 2006, Eq. (2.28)]. Those functions are defined by

$$\psi_k(u)=sign(u)\min\{|u|,k\}$$

for a constant $k > 0$. A possibility for w is e.g. to use $w(y) = \frac{\psi_k(|y|)}{|y|}$ for a Huber function ψ_k. Another choice for ψ, which is not of the Huber type, is the so–called Tukey bisquare (or biweight) function, introduced in Beaton and Tukey [1974], which is given by

$$\psi(u) = u\left(1 - \frac{u^2}{k^2}\right)^2 \mathbb{1}_{\{|u| \leq k\}},$$

where k is again a tuning constant.

c) *For the function χ, a typical choice is*

$$\chi(x^2) = \psi^2(x) - \mathbb{E}_Z[\psi^2(Z)]$$

with the same ψ function as in the definition of ϕ. Here, the expectation is taken with respect to a suitably distributed random variable Z. Suitably distributed in this context means that Z is chosen in such a way that the scale parameter σ_0 coincides with the standard deviation of the innovations if the process $(\widetilde{Y}_n)_{n\in\mathbb{Z}}$ actually is a stationary AR(r) process. This can be achieved by choosing Z as having the same distribution as the innovations divided by their standard deviation. For example, if the innovations have a normal distribution and $Z \sim \mathcal{N}(0,1)$, this will be the case. This choice of χ corresponds to Huber's "proposal 2" ([Huber 1964, p. 97]).

In order to develop an asymptotic theory and to obtain a robust estimator, it is necessary to impose assumptions on ϕ and χ, which we will do next analogous to [Bustos 1982, E1) - E6)]:

Assumption H. Assume that $\phi : \mathbb{R}^r \times \mathbb{R} \to \mathbb{R}$ and $\chi : \mathbb{R} \to \mathbb{R}$ are such that

H.1 For each $y \in \mathbb{R}^r$, the map $u \mapsto \phi(y,u)$ is odd, uniformly continuous and it holds that $\phi(y,u) \geq 0$ for $u \geq 0$.

H.2 $(y,u) \mapsto \phi(y,u)y$ is bounded and there exists $c > 0$ such that

$$|\phi(y,u)y - \phi(z,u)z| \leq c\|y - z\|$$

for all $u \in \mathbb{R}$.

H.3 For each $y \in \mathbb{R}^r$, the map $u \mapsto \frac{\phi(y,u)}{u}$ is non–increasing and there exists $u_0 \in \mathbb{R}$ such that $\frac{\phi(y,u_0)}{u_0} > 0$.

H.4 ϕ is differentiable with respect to u and the map $u \mapsto \frac{\partial \phi(y,u)}{\partial u}$ is continuous, while $u \mapsto \frac{\partial \phi(y,u)}{\partial u} y$ is bounded.

H.5 It holds

$$\mathbb{E}\left[\sup_{u\in\mathbb{R}}\left\{u\left(\frac{\partial}{\partial u}\phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots \\ \widetilde{Y}_r\end{pmatrix}, u\right)\right)\left\|\begin{pmatrix}\widetilde{Y}_1\\ \vdots \\ \widetilde{Y}_r\end{pmatrix}\right\|\right\}\right] < \infty.$$

H.6 χ is bounded and increasing on $\{x : -a \leq \chi(x) < b\}$, where $b = \sup_{x\in\mathbb{R}} \chi(x)$ and $a = \chi(0)$. Furthermore, χ is differentiable and the map $x \mapsto x\chi'(x^2)$ is continuous and bounded. Lastly, $\chi(u_0^2) > 0$.

In the following, when talking about GM estimators, we always assume that Assumption G and Assumption H are satisfied.

Remark 4.26. *As pointed out on [Bustos 1982, p. 497], one can deduce under these assumptions from [Maronna and Yohai 1981, Theorem 2.1] that there exists* $\pi_0 \in \mathbb{R}^r \times (0,\infty)$ *such that* (4.22) *and* (4.23) *are fulfilled. More precisely, there exists a compact set* $K \subset \mathbb{R}^r \times (0,\infty)$ *such that* $\pi_0 \in K$ *and for any* $\pi \in K^c$*, equations* (4.22), (4.23), (4.24) *and* (4.25) *do not hold ([Bustos 1982, p. 500]). This means that the relevant parameter space for GM estimation is the compact set* K *and restricting attention to this set does not entail loss of generality.*

The parameter π_0 can be seen as the pseudo–true parameter to which the GM-estimator converges in this scenario. We are dealing with a pseudo–true parameter as we do not explicitly demand that $(\widetilde{Y}_n)_{n\in\mathbb{Z}}$ actually is an AR(r) process, and most of the times that will not be the case when outliers are present. Of course, it is not clear a priori if π_0 is unique. However, if it is, then we will have almost sure convergence of the GM estimator:

Theorem 4.27. *Assume that the solutions of* (4.22) *and* (4.23) *are unique. Then we have that* $\widehat{\pi}^n \overset{n\to\infty}{\longrightarrow} \pi_0$ $\mathbb{P}$*-a.s.*

Proof. The proof of [Bustos 1982, Theorem 2.1] can directly be used, as all of the assumptions required in it are fulfilled in our scenario. □

The assumed uniqueness of the limiting parameters is, in general, not easy to verify. Additionally, one would like to have that $\pi_0 = \pi_\vartheta$ for the auxiliary parameter defined in Definition 4.19 in the case that the GM estimator is applied to realizations of an uncontaminated, sampled CARMA process $(Y(nh))_{n\in\mathbb{Z}}$ with $Y = \text{CARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$. The following proposition gives a sufficient condition for this to hold:

Proposition 4.28. *Assume that* $(\widetilde{Y}_n)_{n\in\mathbb{Z}} = (Y(nh))_{n\in\mathbb{Z}}$*, i.e. there are no outliers present in the data. Moreover, assume that for* $U_{\vartheta_0,r+1}$ *as defined in equation* (4.5) *it holds that*

$$(U_{\vartheta_0,r+1}, Y_{\vartheta_0}(rh), \ldots, Y_{\vartheta_0}(h)) \stackrel{\mathcal{D}}{=} (-U_{\vartheta_0,r+1}, Y_{\vartheta_0}(rh), \ldots, Y_{\vartheta_0}(h)).$$

Assume that the function $u \mapsto \phi(y,u)$ *is nondecreasing and strictly increasing for* $|u| \leq u_0$*, where* u_0 *satisfies Assumptions H.3 and H.6. Moreover, assume that the function* χ *is chosen in such a way that*

$$\mathbb{E}\left[\chi\left(\left(\frac{U_{\vartheta_0,1}}{\sigma_{\vartheta_0}}\right)^2\right)\right] = 0$$

holds. Then the auxiliary parameter π_{ϑ_0} *as defined in Definition 4.19 is the unique solution to* (4.22) *and* (4.23)*.*

Proof. By the analog of [Maronna and Yohai 1981, Lemma 2.1] in the autoregression case, we have that for each fixed $(\pi_1, \ldots, \pi_r) \in \mathbb{R}^r$ there exists a unique solution σ of the equation

$$\mathbb{E}\left[\chi\left(\left(\frac{Y((r+1)h) - \pi_r Y(hr) - \ldots - \pi_1 Y(h)}{\sigma}\right)^2\right)\right] = 0.$$

By assumption, the function χ is chosen in such a way that for $(\pi_{\vartheta_0,1}, \ldots, \pi_{\vartheta_0,r})$ this unique solution is σ_{ϑ_0}. Therefore, we have that π_{ϑ_0} is a solution of (4.23). Next, we show that the auxiliary parameter is a solution of (4.22), too. Since the function $\phi(y,u)$ is odd in u by Assumption H.1, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\phi\left(\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}, \frac{Y_{\vartheta_0}((r+1)h) - \pi_{\vartheta_0,r} Y(rh) - \ldots - \pi_{\vartheta_0,1} Y(h)}{\sigma_{\vartheta_0}}\right)\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}\right] \\
=&\mathbb{E}\left[\phi\left(\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}, \frac{U_{\vartheta_0,r+1}}{\sigma_{\vartheta_0}}\right)\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}\right] \\
=&\mathbb{E}\left[-\phi\left(\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}, -\frac{U_{\vartheta_0,r+1}}{\sigma_{\vartheta_0}}\right)\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}\right]
\end{aligned}$$

$$= - \mathbb{E}\left[\phi\left(\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}, \frac{U_{\vartheta_0,r+1}}{\sigma_{\vartheta_0}}\right)\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}\right], \tag{4.26}$$

where the last equality follows from the assumption on the distribution of $(-U_{\vartheta_0,r+1}, Y_{\vartheta_0}(rh), \ldots, Y_{\vartheta_0}(h))$. From these equations, we see that

$$\mathbb{E}\left[\phi\left(\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}, \frac{U_{\vartheta_0,r+1}}{\sigma_{\vartheta_0}}\right)\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}\right] = 0$$

holds and π_{ϑ_0} therefore is a solution of equation (4.22). We now show, similar to [Maronna and Yohai 1981, Theorem 2.2a)], that π_{ϑ_0} also is the unique solution of (4.22) and (4.23). Assume that another solution $(\pi_1', \ldots, \pi_r', \sigma')$ exists. Note that the arguments in the derivation of (4.26) still hold if we replace σ_{ϑ_0} in the denominator of the second argument of ϕ by σ'. Thus, we obtain that it also holds that

$$\mathbb{E}\left[\phi\left(\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}, \frac{U_{\vartheta_0,r+1}}{\sigma'}\right)\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}\right] = 0$$

and therefore

$$\begin{aligned}\mathbb{E}\Bigg[\Bigg(&\phi\left(\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}, \frac{Y_{\vartheta_0}((r+1)h) - \pi_r' Y_{\vartheta_0}(rh) - \ldots - \pi_1' Y_{\vartheta_0}(h)}{\sigma'}\right) \\ &-\phi\left(\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}, \frac{U_{\vartheta_0,r+1}}{\sigma'}\right)\Bigg)\begin{pmatrix} Y_{\vartheta_0}(h) \\ \vdots \\ Y_{\vartheta_0}(rh) \end{pmatrix}\Bigg] = 0. \end{aligned} \tag{4.27}$$

Since $\mathbb{P}((Y_{\vartheta_0}(h), \ldots, Y_{\vartheta_0}(rh)) = (0, \ldots, 0)) = 0$ and $\phi(y, u)$ is strictly increasing on the interval $(-u_0, u_0)$ for every $y \in \mathbb{R}^r$, we must have that

$$\begin{aligned} 1 =& \mathbb{P}\Bigg(\Bigg(\left\{\frac{Y_{\vartheta_0}((r+1)h) - \pi_r' Y_{\vartheta_0}(rh) - \ldots - \pi_1' Y_{\vartheta_0}(h)}{\sigma'} \geq u_0\right\} \cap \left\{\frac{U_{\vartheta_0,r+1}}{\sigma'} \geq u_0\right\}\Bigg) \\ &\dot\cup \Bigg(\left\{\frac{Y_{\vartheta_0}((r+1)h) - \pi_r' Y_{\vartheta_0}(rh) - \ldots - \pi_1' Y_{\vartheta_0}(h)}{\sigma'} \leq -u_0\right\} \cap \left\{\frac{U_{\vartheta_0,r+1}}{\sigma'} \leq -u_0\right\}\Bigg)\Bigg) \\ \leq& \mathbb{P}\left(\frac{Y_{\vartheta_0}((r+1)h) - \pi_r' Y_{\vartheta_0}(rh) - \ldots - \pi_1' Y_{\vartheta_0}(h)}{\sigma'} \geq u_0\right) \end{aligned}$$

$$+\mathbb{P}\left(\frac{Y_{\vartheta_0}((r+1)h) - \pi_r' Y_{\vartheta_0}(rh) - \ldots - \pi_1' Y_{\vartheta_0}(h)}{\sigma'} \leq -u_0\right)$$
$$=\mathbb{P}\left(\frac{|Y_{\vartheta_0}((r+1)h) - \pi_r' Y_{\vartheta_0}(rh) - \ldots - \pi_1' Y_{\vartheta_0}(h)|}{\sigma'} \geq u_0\right),$$

because otherwise (4.27) cannot hold. Therefore,

$$\left(\frac{Y_{\vartheta_0}((r+1)h) - \pi_r' Y_{\vartheta_0}(rh) - \ldots - \pi_1' Y_{\vartheta_0}(h)}{\sigma'}\right)^2 \geq u_0^2, \ \mathbb{P}\text{-a.s.} \tag{4.28}$$

Now, π' is by assumption also a solution of (4.23) and hence we have that

$$0 = \mathbb{E}\left[\chi\left(\left(\frac{Y_{\vartheta_0}((r+1)h) - \pi_r' Y_{\vartheta_0}(rh) - \ldots - \pi_1' Y_{\vartheta_0}(h)}{\sigma'}\right)^2\right)\right] \overset{(4.28)}{\geq} \chi(u_0^2) \overset{H.6}{>} 0,$$

a contradiction. □

Remark 4.29.

a) The assumption $(U_{\vartheta_0,r+1}, Y_{\vartheta_0}(rh), \ldots, Y_{\vartheta_0}(h)) \overset{\mathcal{D}}{=} (-U_{\vartheta_0,r+1}, Y_{\vartheta_0}(rh), \ldots, Y_{\vartheta_0}(h))$ *is fulfilled, for example, if the distribution of* $U_{\vartheta_0,r+1}$ *is symmetric and* $U_{\vartheta_0,r+1}$ *is independent of* $(Y_{\vartheta_0}(rh), \ldots, Y_{\vartheta_0}(h))$*. This again is fulfilled if the Lévy process driving* $(Y(t))_{t\in\mathbb{R}}$ *is a Brownian motion, because then, by the equation*

$$Y(t) = \int_{-\infty}^{t} C_{\vartheta_0} \, \mathrm{e}^{A_{\vartheta_0}(t-u)} \, B_{\vartheta_0} \mathrm{d}L_{\vartheta_0}(u), \ t \in \mathbb{R}$$

([Schlemm and Stelzer 2012, Proposition 3.1]) the process $(Y_{\vartheta_0}(t))_{t\in\mathbb{R}}$ *is a Gaussian process (see also [Brockwell 2001, p. 155]). Hence, every finite–dimensional marginal of* $(Y_{\vartheta_0}(t))_{t\in\mathbb{R}}$ *has a multivariate normal distribution. Because* $(Y_{\vartheta_0}(h), \ldots, Y_{\vartheta_0}((r+1)h))$ *is exactly the marginal distribution of the CARMA process* $(Y_{\vartheta_0}(t))_{t\in\mathbb{R}}$ *at times* $h, \ldots, (r+1)h$*, this random vector especially has multivariate normal distribution. For normally distributed random variables it is well know that they are independent if and only if they are uncorrelated, hence the property that* $U_{\vartheta_0,r+1}$ *is uncorrelated with* $Y_{\vartheta_0}(h), \ldots, Y_{\vartheta_0}((r+1)h)$ *by construction implies the independence in this case. The symmetry of a normal distribution is obvious.*

b) The monotonicity assumption on ϕ *is fulfilled, for example, for both the Mallows and Hampel–Krasker–Welsch estimators of Example 4.25a) when the function* ψ *is chosen as a Huber* ψ_k*–function as in Example 4.25b) with* $u_0 = k$*.*

c) The assumption about χ *is fulfilled, for example, if* χ *is chosen as in Example 4.25c)*

with $Z \stackrel{\mathcal{D}}{=} \frac{U_{\vartheta_0,1}}{\sqrt{\mathrm{Var}(U_{\vartheta_0,1})}}$. *In the case that the driving Lévy process is a Brownian motion as in part a), this entails that the assumption is fulfilled if* $Z \sim \mathcal{N}(0,1)$.

Next, we would like to deduce a central limit theorem for our GM estimator. Since we have altered the assumptions on the involved stochastic processes, we cannot use all of the results of Bustos [1982] that lead to the CLT directly. In particular, we need the following lemma, which is the analog of [Bustos 1982, Lemma 3.1] under our assumptions:

Lemma 4.30. *For fixed* $y \in \mathbb{R}^{r+1}$, *define the map*

$$\begin{aligned} \Psi : \mathbb{R}^r \times (0,\infty) &\to \mathbb{R}^{r+1} \\ \pi = (\pi_1, \ldots, \pi_r, \sigma) &\mapsto \Psi(y, \pi) \end{aligned}$$

by

$$\Psi(y,\pi) = \begin{pmatrix} \phi\left(\begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix}, \frac{y_{r+1} - \pi_r y_r - \ldots - \pi_1 y_1}{\sigma}\right) \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} \\ \chi\left(\left(\frac{y_{r+1} - \pi_r y_r - \ldots - \pi_1 y_1}{\sigma}\right)^2\right) \end{pmatrix}$$

Furthermore, define the stochastic process $(\Psi(t))_{t\in\mathbb{N}}$ *by* $\Psi(t) = \Psi((\widetilde{Y}_t, \ldots, \widetilde{Y}_{t+r-1}), \pi_0)$. *Then it holds that*

$$\frac{1}{\sqrt{n-r}} \sum_{t=1}^{n-r} \Psi(t) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, \mathcal{I}_{GM}(\pi_0)),$$

where the matrix $\mathcal{I}_{GM}(\pi_0)$ *is defined by*

$$(\mathcal{I}_{GM}(\pi_0))_{ij} = \mathbb{E}\left[\Psi_i(1)\Psi_j(1)\right] + 2\sum_{t=1}^{\infty} \mathbb{E}\left[\Psi_i(1)\Psi_j(1+t)\right] \tag{4.29}$$

and $\Psi_i(t)$ *denotes the* i*–th component of* $\Psi(t)$, $i = 1, \ldots, r+1$. *Especially, each* $(\mathcal{I}_{GM}(\pi_0))_{ij}$ *is finite for* $i, j \in \{1, \ldots, r+1\}$.

Proof. By the Cramer–Wold device, the statement of the Lemma is equivalent to the assertion that $\frac{1}{\sqrt{n-r}} x^T \sum_{t=1}^{n-r} \Psi(t)$ converges to a univariate normal distribution with mean 0 and variance $x^T \mathcal{I}_{\mathrm{GM}}(\pi_0) x$ for every $x \in \mathbb{R}^{r+1}$. According to [Ibragimov 1962, Theorem 1.7], this holds if we can show that $\mathbb{E}|x^T\Psi(t)|^{2+\delta} < \infty$ and that $(x^T\Psi(t))_{t\in\mathbb{N}}$ is strongly mixing with $\sum_{k=1}^{\infty} \alpha_{x^T\Psi}(k)^{\frac{\delta}{2+\delta}} < \infty$ for some $\delta > 0$. The same theorem then also states that $x^T \mathcal{I}_{\mathrm{GM}}(\pi_0) x < \infty$, from which we then can deduce that for $i, j \in \{1, \ldots, r+1\}$ the entry $(\mathcal{I}_{\mathrm{GM}}(\pi_0))_{ij}$ is finite and therefore $\mathcal{I}_{\mathrm{GM}}(\pi_0)$ is

well–defined.
For the existence of the $(2+\delta)$–th moment of $x^T\Psi(t)$, note that

$$\mathbb{E}\left[|x^T\Psi(t)|^{2+\delta}\right] \leq C\|x\|_\infty^{2+\delta} \sum_{i=1}^{r+1} \mathbb{E}\left[\|\Psi_i(t)\|^{2+\delta}\right] < \infty, \tag{4.30}$$

where the last inequality holds since every $\Psi_i(t)$ is bounded by H.2 and H.6. Moreover, the process $(\widetilde{Y}_n)_{n\in\mathbb{Z}}$ is strongly mixing for the following reason: For every $n \in \mathbb{N}$, it holds that $\widetilde{Y}_n = g(V_n, Z_n, Y(nh))$ for some measurable function g if the first part of Assumption G.3 is fulfilled or $\widetilde{Y}_n = g(V_n, W_n, Y(nh))$ if the second part is fulfilled. In either case, the three processes to which g is applied are independent by assumption. Hence, by [Bradley 2007, Theorem 6.6(II)] it holds that

$$\alpha_{\widetilde{Y}}(m) \leq \alpha_V(m) + \alpha_Z(m) + \alpha_{Y^{(h)}}(m) \leq C\rho^m$$

for some $C > 0$ and $\rho \in (0,1)$ by Assumption G.2 and Proposition 2.24. Furthermore, $\Psi(t)$ depends on the finitely many values $\widetilde{Y}_t, \ldots, \widetilde{Y}_{t+r}$ and by [Bradley 2007, Remark 1.8b)] this ensures that $\alpha_\Psi(m) \leq \alpha_{\widetilde{Y}}(m+r) \leq C\rho^m$. Hence, the strong mixing coefficients of $x^T\Psi$ satisfy the summability condition and the lemma is proven. □

The rest of the lemmas that are used in the proof of the central limit theorem are the same as in Bustos [1982]. For sake of completeness, we state them in the appendix (Section A.2). Keeping this in mind, we can now state and prove a central limit theorem for the GM estimator $\widehat{\pi}^n$. Remember that the relevant parameter space is the compact set K from Remark 4.26.

Theorem 4.31. *For $\pi = (\pi_1, \ldots, \pi_r, \sigma) \in K$, define*

$$\mathscr{Q}_{GM}(\pi) = \begin{pmatrix} \mathbb{E}\left[\phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots \\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1} - \pi_r\widetilde{Y}_r - \ldots - \pi_1\widetilde{Y}_1}{\sigma}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots \\ \widetilde{Y}_r\end{pmatrix}\right] \\ \mathbb{E}\left[\chi\left(\left(\frac{\widetilde{Y}_{r+1} - \pi_r\widetilde{Y}_r - \ldots - \pi_1\widetilde{Y}_1}{\sigma}\right)^2\right)\right] \end{pmatrix}.$$

Assume that $\mathcal{J}_{GM}(\pi_0) := \nabla_\pi \mathscr{Q}_{GM}(\pi_0)$, the Jacobian of $\mathscr{Q}_{GM}$ evaluated at the pseudo-true parameter, is non-singular and that $\widehat{\pi}^n \xrightarrow{\mathbb{P}} \pi_0$. Then it holds that

$$\sqrt{n-r}(\widehat{\pi}^n - \pi_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Xi_{GM}(\pi_0)),\ n \to \infty,$$

where the matrix $\Xi_{GM}(\pi_0)$ *is defined by*

$$\Xi_{GM}(\pi_0) := (\mathcal{J}_{GM}(\pi_0))^{-1}\mathcal{I}_{GM}(\pi_0)(\mathcal{J}_{GM}(\pi_0))^{-1}. \tag{4.31}$$

Proof. Note first that for $i, j = 1, \ldots, r$ it holds that

$$\begin{aligned}
&\sup_{\pi\in K}\left|\frac{\partial}{\partial\pi_i}\phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\widetilde{Y}_j\right| \\
&= \sup_{\pi\in K}\left|\left(\frac{\partial}{\partial u}\phi\right)\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\widetilde{Y}_j\frac{\widetilde{Y}_i}{\sigma}\right| \\
&\leq \sup_{\pi\in K} C\left|\left(\frac{\partial}{\partial u}\phi\right)\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\right|\left\|\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}\right\|^2 \\
&\leq \sup_{u\in\mathbb{R}} C\left\|\left(\frac{\partial}{\partial u}\phi\right)\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, u\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}\right\|\left\|\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}\right\| \\
&\leq C\left\|\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}\right\|.
\end{aligned}$$

by Assumption H.4. By Assumption G.1, the expectation of the right–hand side is finite. Similarly,

$$\begin{aligned}
&\sup_{\pi\in K}\left\|\frac{\partial}{\partial\sigma}\phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}\right\| \\
&= \sup_{\pi\in K}\left(\frac{1}{\sigma}\left\|\frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\left(\frac{\partial}{\partial u}\phi\right)\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}\right\|\right) \\
&\leq C\sup_{u\in\mathbb{R}}\left\|u\left(\frac{\partial}{\partial u}\phi\right)\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, u\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}\right\|.
\end{aligned}$$

The expectation of the right–hand side is finite because of Assumption H.5. A similar argument, using Assumption H.6, also shows that $\left|\frac{\partial}{\partial\pi_i}\chi\left(\left(\frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)^2\right)\right|$ for $i=1,\ldots,r$ and $\left|\frac{\partial}{\partial\sigma}\chi\left(\left(\frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)^2\right)\right|$ are uniformly dominated by integrable random variables. Therefore, the existence of $\mathcal{J}_{\mathrm{GM}}(\pi_0)$ follows by an application of the dominated convergence theorem ([Rosenthal 2006, Proposition 9.2.1]). Remember that $\mathscr{Q}_{\mathrm{GM}}(\pi_0)=0$ by (4.22) and (4.23). We do a first-order Taylor expansion of this expression around $\widehat{\pi}^n$ to obtain

$$0=\sqrt{n-r}\mathscr{Q}_{\mathrm{GM}}(\pi_0)=\sqrt{n-r}\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)+\sqrt{n-r}\nabla_\pi\mathscr{Q}_{\mathrm{GM}}(\overline{\pi}^n)(\pi_0-\widehat{\pi}^n), \quad (4.32)$$

where $\|\pi_0-\overline{\pi}^n\|\leq\|\pi_0-\widehat{\pi}^n\|$. Note now that

$$\sqrt{n-p}\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)=-\frac{1}{\sqrt{n-r}}\sum_{t=1}^{n-r}\Psi(t)+\frac{1}{\sqrt{n-r}}\sum_{t=1}^{n-r}(\Psi(t)+\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)). \quad (4.33)$$

Applying Lemma 4.30 to the first summand and Lemma A.5 to the second one shows that

$$\sqrt{n-r}\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\xrightarrow{\mathcal{D}}\mathcal{N}(0,\mathcal{I}_{\mathrm{GM}}(\pi_0)),\ n\to\infty.$$

Moreover, by Assumption H, the consistency of $\widehat{\pi}^n$ and the dominated convergence theorem, we can conclude that $\nabla_\pi\mathscr{Q}_{\mathrm{GM}}(\overline{\pi}^n)$ converges to $\mathcal{J}_{\mathrm{GM}}(\pi_0)$ as $n\to\infty$. Since $\mathcal{J}_{\mathrm{GM}}(\pi_0)$ is non–singular, the Theorem now follows from (4.32). □

Remark 4.32. *The statement of this theorem coincides with the one of [Bustos 1982, Theorem 2.2]. The difference lies in the proof, namely* (4.33)*: for the first summand, we obtain a central limit theorem by means of Lemma 4.30, which has a different proof than the corresponding statement ([Bustos 1982, Lemma 3.1]), since our assumptions are different.*

4.3.3.2. The least squares estimator

In this subsection, we estimate the parameters of the auxiliary AR(r) representation by the least squares estimator. In the treatment, we retain the assumption that we have a parameter space Θ that satisfies Assumptions B.1 to B.9 and that we have observations $Y^n=(Y(h),\ldots,Y(nh))$ of a process $(Y(nh))_{n\in\mathbb{Z}}$ with $Y=\mathrm{CARMA}(A_{\vartheta_0},B_{\vartheta_0},C_{\vartheta_0},L_{\vartheta_0})$ for some $\vartheta_0\in\Theta$. We are interested in estimating the parameter π_{ϑ_0} of the AR(r) representation of $(Y(nh))_{n\in\mathbb{Z}}$ defined in (4.5). To do parameter estimation in a suitable space, we assume that $r\geq 2p-1$ and thus the binding function π as defined in Definition 4.21 is injective and continuously

differentiable. The space $\Pi' = \pi(\Theta) \subseteq \Pi$ is then compact and the relevant space for parameter estimation. We define the estimator as follows:

Definition 4.33. *Based on the sample $Y^n = (Y(h), \ldots, Y(nh))$ of the process $(Y(nh))_{n\in\mathbb{Z}}$, the least squares estimator of π_{ϑ_0} is*

$$\widehat{\pi}^n_{LS} = (\widehat{\pi}^n_{LS,1}, \ldots, \widehat{\pi}^n_{LS,r}, \widehat{\sigma}_{LS,n})$$

where $\widehat{\pi}^n_{LS,1}, \ldots, \widehat{\pi}^n_{LS,r}$ minimize

$$\begin{aligned} S(\pi) &:= \frac{1}{n-r}\sum_{t=1}^{n-r} \left(Y((t+r)h) - \pi_r Y((t+r-1)h) - \ldots - \pi_1 Y(th)\right)^2 \\ &=: \frac{1}{n-r}\sum_{t=1}^{n-r} U_t(\pi)^2 \end{aligned} \tag{4.34}$$

in Π' and $\widehat{\sigma}_{LS,n}$ is obtained from $\widehat{\pi}^n_{LS,1}, \ldots, \widehat{\pi}^n_{LS,r}$ via the equation

$$\widehat{\sigma}^2_{LS,n} = \frac{1}{n-r}\sum_{t=1}^{n-r} \left(Y((t+r)h) - \widehat{\pi}^n_{LS,r} Y((t+r-1)h) - \ldots - \widehat{\pi}^n_{LS,1} Y(th)\right)^2.$$

Note that the right–hand side of (4.34) is differentiable with respect to π. Hence, the derivative evaluated at the minimizer must be equal to zero. Since for $i = 1, \ldots, r$ it holds that

$$\frac{\partial}{\partial \pi_i} U_t(\pi)^2 = -2\left(Y((t+r)h) - \pi_r Y((t+r-1)h) - \ldots - \pi_1 Y(th)\right) Y((t+i-1)h)$$

we obtain that $\widehat{\pi}^n_{\mathrm{LS}}$ is a solution of (4.24) and (4.25) for the special choices $\phi(y,u) = u$ and $\chi(x) = x - 1$, i.e. the least squares estimator can be seen as a special case of a GM estimator. We have the following result on the asymptotic behavior of this estimator:

Theorem 4.34. *The estimator $\widehat{\pi}^n_{LS}$ based on the sample $Y^n = (Y(h), \ldots, Y(nh))$ is strongly consistent for π_{ϑ_0}, i. e. we have*

$$\widehat{\pi}^n_{LS} \xrightarrow{n\to\infty} \pi_{\vartheta_0} \ \mathbb{P}\text{-}a.s.$$

Furthermore, it holds that

$$\sqrt{n}(\widehat{\pi}^n_{LS} - \pi_{\vartheta_0}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Xi_{LS}(\pi_{\vartheta_0})\right), \ n \to \infty,$$

where the matrix $\Xi_{LS}(\pi_{\vartheta_0})$ *is defined via* (4.29) *and* (4.31)*, respectively, for* $\phi(y,u) = u$ *and* $\chi(x) = x - 1$*.*

Proof. For the particular choice $\phi(y,u) = u$ and $\chi(x) = x - 1$, the unique solution to (4.22) and (4.23) is π_{ϑ_0} as shown in Definition 4.19. Per construction, $\widehat{\pi}^n_{\mathrm{LS}}$ minimizes the function $S(\pi)$. Since the sampled CARMA process $(Y(nh))_{n\in\mathbb{Z}}$ is ergodic, the same is then true for the process $(U_n(\pi)^2)_{n\in\mathbb{Z}}$ since each $U_n(\pi)^2$ results from applying a measurable function to finitely many $Y(nh)$. Moreover,

$$\mathbb{E}[U_1(\pi)^2] < \infty$$

since the process $(Y(nh))_{n\in\mathbb{Z}}$ is stationary and has finite second moments. Therefore, it holds by Birkhoff's ergodic theorem that

$$S(\pi) \longrightarrow \mathbb{E}[U_1(\pi)^2] \quad \mathbb{P}\text{-a.s., } n \to \infty.$$

We have just noticed that for fixed ϑ the derivative of the limiting function has a unique zero at $\pi = \pi_{\vartheta_0}$, i.e. the limiting function itself has a unique minimum. Moreover, the space $\Pi' = \pi(\Theta)$ is compact. Therefore, by [Ferguson 1996, Theorem 16a)] we obtain on the one hand that

$$\sup_{\pi\in\Pi'} |S(\pi) - \mathbb{E}[U_1(\pi)^2]| \longrightarrow 0 \quad \mathbb{P}\text{-a.s., } n \to \infty \tag{4.35}$$

and from this by analogous arguments as in [Francq and Zakoïan 2005, Proof of Theorem 12.5] that

$$\widehat{\pi}^n_{\mathrm{LS}} \overset{n\to\infty}{\longrightarrow} \pi_{\vartheta_0} \ \mathbb{P}\text{-a.s.}$$

Note that the authors assume in their Theorem 12.5 that the innovations of the data–generating process are uncorrelated, which is not satisfied in our situation. However, in the proof of strong consistency of the least squares estimator, they use this fact only to establish that the limiting function has a unique minimum. We do not need this argument here, since we can derive the uniqueness as shown above.
The asymptotic normality of $\widehat{\pi}^n_{\mathrm{LS}}$ follows in principle from Theorem 4.31. Here we again make use of the fact that we can interpret the least squares estimator as a particular GM estimator with $\phi(y,u) = u$ and $\chi(x) = x - 1$. Still, we need to be careful because these two functions do not satisfy Assumptions H.2, H.4 and H.6 with respect to boundedness. However, a close inspection of the proof of Theorem 4.31 reveals that the boundedness is only used at two points. The first is at (4.30), where we deduce the finiteness of the expectation by boundedness. However, this is not a

necessary condition. In the situation of this theorem, equation (4.30) reads

$$\Psi_i(t) = \left(Y((t+r)h) - \pi_{\vartheta_0,r}Y((t+r-1)h) - \ldots - \pi_{\vartheta_0,1}Y(th)\right)Y((t+i-1)h)$$

for $i = 1, \ldots, r$ and

$$\Psi_{r+1}(t) = \left(\frac{Y((t+r)h) - \pi_{\vartheta_0,r}Y((t+r-1)h) - \ldots - \pi_{\vartheta_0,1}Y(th)}{\sigma_{\vartheta_0}}\right)^2 - 1.$$

Therefore, inequality (4.30) follows since the Lévy process driving $(Y(t))_{t\in\mathbb{R}}$ has finite $(4+\delta)$–th moment by B.9, which then transfers to $(Y(t))_{t\in\mathbb{R}}$ and subsequently to the finitely many elements of which each $\Psi_i(t)$ is a linear combination (note that this moment condition is also sufficient to obtain the exponential strong mixing of $(Y(t))_{t\in\mathbb{R}}$ used in the proof of Lemma 4.30).

The second point where the boundedness assumptions are used is right at the beginning of the proof of Theorem 4.31 to deduce the existence of the matrix $\mathcal{J}_{\mathrm{GM}^*}(\pi_{\vartheta_0})$. Here we have used the notation GM^* in the index to indicate that we refer to the GM estimator for the particular choices $\phi(y,u) = u$ and $\chi(x) = x - 1$ and not to a general one. However, in our case the expectations defining $\mathcal{Q}_{\mathrm{GM}^*}(\pi)$ are obviously differentiable in π because of the linearity of the expectation.

An assumption of Theorem 4.31 was also that the Jacobian $\nabla_\pi \mathcal{Q}_{\mathrm{GM}^*}(\pi)$ evaluated at the true parameter, i.e. $\mathcal{J}_{\mathrm{GM}^*}(\pi_\vartheta)$, is non–singular. This we can verify by direct calculation. Plugging in the functions ϕ and χ we work with here, one sees immediately that interchanging derivative and expectation is allowed. Denoting by $\mathcal{Q}_{\mathrm{GM}^*}(\pi)$ the function defined in Theorem 4.31 for the special choice $\phi(y,u) = u$ and $\chi(x) = x - 1$ and by $(\mathcal{Q}_{\mathrm{GM}^*}(\pi))_j$ the j–th component, we obtain that

$$\begin{aligned}
\frac{\partial}{\partial \pi_i}(\mathcal{Q}_{\mathrm{GM}^*}(\pi))_j &= -\frac{\mathbb{E}\left[Y(ih)Y(jh)\right]}{\sigma} = -\frac{\mathrm{Cov}(Y(ih), Y(jh))}{\sigma},\ i,j = 1,\ldots,r,\\
\frac{\partial}{\partial \sigma}(\mathcal{Q}_{\mathrm{GM}^*}(\pi))_j &= -\frac{\mathbb{E}\left[Y((r+1)h) - \pi_r Y(rh) - \ldots - \pi_1 Y(h)\right]}{\sigma^2},\ j = 1,\ldots,r,\\
\frac{\partial}{\partial \pi_i}(\mathcal{Q}_{\mathrm{GM}^*}(\pi))_{r+1} &= -\frac{\mathbb{E}\left[2(Y((r+1)h) - \pi_r Y(rh) - \ldots - \pi_1 Y(h))Y(ih)\right]}{\sigma^2},\ i = 1,\ldots,r,\\
\frac{\partial}{\partial \sigma}(\mathcal{Q}_{\mathrm{GM}^*}(\pi))_{r+1} &= -2\frac{\mathbb{E}\left[(Y((r+1)h) - \pi_r Y(rh) - \ldots - \pi_1 Y(h))^2\right]}{\sigma^3}.
\end{aligned}$$

Plugging in $\pi = \pi_{\vartheta_0}$ and using the defining equation (4.7), we can simplify in this

special case to

$$\mathcal{J}_{\mathrm{GM}^*}(\vartheta_0) = -\frac{1}{\sigma_{\vartheta_0}} \begin{pmatrix} \mathrm{Var}(Y(1)) & \mathrm{Cov}(Y(h), Y(2h)) & \dots & \mathrm{Cov}(Y(h), Y(rh)) & 0 \\ \mathrm{Cov}(Y(h), Y(2h)) & \mathrm{Var}(Y(2h)) & \dots & \mathrm{Cov}(Y(2h), Y(rh)) & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{Cov}(Y(h), Y(rh)) & \dots & \dots & \mathrm{Var}(Y(rh)) & 0 \\ 0 & 0 & \dots & 0 & 2 \end{pmatrix}$$

Hence, $\mathcal{J}_{\mathrm{GM}^*}(\pi_\vartheta)$ is non–singular if and only if the upper left $r \times r$ block is. However, the upper left block is, up to a positive factor, the covariance matrix of the random vector $(Y(h), \dots, Y(rh))$, which is of course non–singular because it is positive definite. In particular, none of the proofs of Lemma A.3 - Lemma A.5 do need the boundedness, such that the result follows. □

4.3.3.3. The quasi maximum likelihood estimator

The third estimator for the parameters of the auxiliary AR(r) representation of a CARMA process is the quasi maximum likelihood estimator. We stay in the framework of Subsubsection 4.3.3.2, i.e. we have a parameter space Θ that satisfies Assumptions B.1 to B.9 and observations $Y^n = (Y(h), \dots, Y(nh))$ of a process $(Y(nh))_{n\in\mathbb{Z}}$ with $Y = \mathrm{CARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$ for some $\vartheta_0 \in \Theta$. The link function $\pi : \Theta \to \Pi$ is assumed to be injective (i.e. $r \geq 2p-1$ holds) and continuously differentiable. For observations $(Y(h), \dots, Y(nh))$ and a parameter $\pi \in \Pi'$, the pseudo–innovation in the AR(r) model at time $t+r$ is exactly $U_t(\pi)$ as defined in (4.34), i.e.

$$U_t(\pi) = Y((t+r)h) - \pi_r Y((t+r-1)h) - \dots - \pi_1 Y(th). \tag{4.36}$$

For the definition of the quasi maximum likelihood estimator, we again use the Gaussian likelihood function, which is the exact likelihood if the $U_t(\pi)$ are Gaussian. Taking logarithms and multiplying by $-2/(n-r)$, we define:

Definition 4.35. *Based on the sample $Y^n = (Y(h), \dots, Y(nh))$ of the process $(Y(nh))_{n\in\mathbb{Z}}$, the quasi Gaussian likelihood function for the AR(r) process is defined as*

$$\begin{aligned} \mathcal{L}_{AR}(\pi, Y^n) &= \frac{1}{n-r} \sum_{t=1}^{n-r} l_{AR,t}(\pi) \\ &:= \frac{1}{n-r} \sum_{t=1}^{n-r} \left(\log(\sigma^2) + \frac{U_t(\pi)^2}{\sigma^2} \right) \end{aligned}$$

$$= \frac{1}{n-r}\sum_{t=1}^{n-r}\left(\log(\sigma^2) + \frac{(Y((t+r)h) - \pi_r Y((t+r-1)h) - \ldots - \pi_1 Y(th))^2}{\sigma^2}\right)$$

Based on this, the quasi maximum likelihood estimator $\widehat{\pi}^n_{MLE}(\vartheta)$ *is defined by*

$$\widehat{\pi}^n_{MLE} = \arg\min_{\pi\in\Pi'} \mathcal{L}_{AR}(\pi, Y^n).$$

Remark 4.36. *The quasi maximum likelihood estimator and the least squares estimator are identical in this situation. Remember that by Definition 4.33* $\widehat{\pi}^n_{LS,1}, \ldots, \widehat{\pi}^n_{LS,r}$ *minimize*

$$S(\pi) := \frac{1}{n-r}\sum_{t=1}^{n-r}\left(Y((t+r)h) - \pi_r Y((t+r-1)h) - \ldots - \pi_1 Y(th)\right)^2.$$

Since this function is differentiable with respect to π, $(\widehat{\pi}^n_{LS,1}, \ldots, \widehat{\pi}^n_{LS,r})$ *is the unique zero of the gradient. Differentiating partially gives for* $i \in \{1, \ldots, r\}$*:*

$$\frac{\partial}{\partial \pi_i} S(\pi) = -\frac{2}{n-r}\sum_{t=1}^{n-r}(Y((t+r)h) - \pi_r Y((t+r-1)h) - \ldots - \pi_1 Y(th))Y((t+r-i)h).$$

Additionally, $\widehat{\sigma}^2_{LS,n}$ *is defined as*

$$\widehat{\sigma}^2_{LS,n} = \frac{1}{n-r}\sum_{t=1}^{n-r}\left(Y((t+r)h) - \widehat{\pi}^n_{LS,r} Y((t+r-1)h) - \ldots - \widehat{\pi}^n_{LS,1} Y(th)\right)^2$$

The function $\mathcal{L}_{AR}$ *is differentiable with respect to* π*, too, i.e. we can differentiate* $\mathcal{L}_{AR}$ *partially with respect to each variable and look for a zero of the gradient to obtain* $\widehat{\pi}^n_{MLE}$*. Doing so, we obtain for for* $i \in \{1, \ldots, r\}$ *that*

$$\frac{\partial}{\partial \pi_i}\mathcal{L}_{AR}(\pi, Y^n) = -\frac{2}{(n-r)\sigma^2}\sum_{t=1}^{n-r}(Y((t+r)h) - \pi_r Y((t+r-1)h) - \ldots - \pi_1 Y(th))Y((t+r-i)h) \tag{4.37}$$

and

$$\frac{\partial}{\partial \sigma}\mathcal{L}_{AR}(\pi, Y^n) = -\frac{2}{n-r}\sum_{t=1}^{n-r}\frac{(Y((t+r)h) - \pi_r Y((t+r-1)h) - \ldots - \pi_1 Y(th))^2}{\sigma^3} - \frac{2}{\sigma^2}. \tag{4.38}$$

Since $\frac{\partial}{\partial \pi_i} S(\pi)$ *is a multiple of* $\frac{\partial}{\partial \pi_i}\mathcal{L}_{AR}(\pi, Y^n)$ *for* $i \in \{1, \ldots, r\}$*, the unique zeros of these partial derivatives are the same, i.e.*

$$(\widehat{\pi}^n_{MLE,1}, \ldots, \widehat{\pi}^n_{MLE,r}) = (\widehat{\pi}^n_{LS,1}, \ldots, \widehat{\pi}^n_{LS,r}).$$

With this, it is obvious from (4.38) *that* $\widehat{\sigma}^2_{MLE,n} = \widehat{\sigma}^2_{LS,n}$ *holds, too, and the estimators are indeed the same.*
In the theory of parameter estimation of ARMA processes, it is well–known that least squares and maximum likelihood estimation are asymptotically equivalent, see e.g. [Brockwell and Davis 1991, §8.7]. However here we have the stronger notion that the estimators are identical for every finite sample size.

By this remark, we immediately obtain the following theorem:

Theorem 4.37. *Let* $\widehat{\pi}^n_{MLE}$ *be defined as in Definition 4.35. Then* $\widehat{\pi}^n_{MLE}$ *is strongly consistent for* π_{ϑ_0}*, i. e. we have*

$$\widehat{\pi}^n_{MLE} \xrightarrow{n\to\infty} \pi_{\vartheta_0} \ \mathbb{P}\text{-a.s.}$$

Furthermore, it holds that

$$\sqrt{n}(\widehat{\pi}^n_{MLE} - \pi_{\vartheta_0}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Xi_{MLE}(\pi_{\vartheta_0})\right), \ n \to \infty,$$

where the matrix $\Xi_{MLE}(\pi_{\vartheta_0})$ *is defined via*

$$\Xi_{MLE}(\pi_{\vartheta_0}) = (\mathcal{J}_{MLE}(\pi_{\vartheta_0}))^{-1}\mathcal{I}_{MLE}(\pi_{\vartheta_0})(\mathcal{J}_{MLE}(\pi_{\vartheta_0}))^{-1}$$

for

$$\mathcal{I}_{MLE}(\pi_{\vartheta_0}) = \lim_{n\to\infty} n \operatorname{Var}(\nabla_\pi \mathcal{L}_{AR}(\pi_{\vartheta_0}, Y^n)) \quad \text{and} \quad \mathcal{J}_{MLE}(\pi_{\vartheta_0}) = \lim_{n\to\infty} \nabla^2_\pi \mathcal{L}_{AR}(\pi_{\vartheta_0}, Y^n).$$

Proof. Since the least squares estimator and the QMLE coincide, the theorem immediately follows from Theorem 4.34. □

The asymptotic results derived up until now are not yet sufficient to guarantee that Theorem 4.23 holds when the QMLE or the least squares estimator is used as the estimator in the simulation part. We also need to make sure that (4.16) is satisfied, which is the topic of the next theorem.

Theorem 4.38. *Assume that* Θ *satisfies Assumption B. Let* $(L_S(t))_{t\in\mathbb{R}}$ *be a Lévy process that satisfies Assumptions B.2 and B.9. Assume that for every* $\vartheta \in \Theta$ *the eigenvalues of* A_ϑ *are all distinct. Moreover, assume that* $\mathbb{P}$*-a.s. it holds for every* $t \in \mathbb{R}$ *and every* $j \in \{1, \dots, N(\Theta)\}$ *that*

$$\frac{\partial}{\partial\vartheta_j}\int_{-\infty}^t C_\vartheta \, \mathrm{e}^{A_\vartheta(t-u)} \, B_\vartheta \mathrm{d}L_S(u) = \int_{-\infty}^t \frac{\partial}{\partial\vartheta_j}(C_\vartheta \, \mathrm{e}^{A_\vartheta(t-u)} \, B_\vartheta)\mathrm{d}L_S(u).$$

Denote by $\widehat{\pi}^n_{MLE}(\vartheta)$ the QMLE of the parameter π_ϑ of the AR(r) representation of the CARMA process $(Y_\vartheta(nh))_{n\in\mathbb{Z}}$ driven by L_S based on a sample $Y^n_\vartheta = (Y_\vartheta(h), \ldots, Y_\vartheta(nh))$. Furthermore, assume that the the sequence of random functions $\vartheta \mapsto \nabla^2_\vartheta \widehat{\pi}^n_{MLE}(\vartheta)$ is almost surely uniformly bounded. Then, it holds that

$$\nabla_\vartheta \widehat{\pi}^n_{MLE}(\overline{\vartheta}^n) \longrightarrow \nabla_\vartheta \pi_{\vartheta_0},\ \mathbb{P}\text{-}a.s., n \to \infty$$

for every sequence $(\overline{\vartheta}^n)_{n\in\mathbb{N}}$ with $\overline{\vartheta}^n \to \vartheta_0$ $\mathbb{P}$-a.s. as $n \to \infty$.

Proof. We begin by calculating the derivative of the function $\vartheta \mapsto \widehat{\pi}^n_{\mathrm{MLE}}(\vartheta)$. This function can be characterized as the solution of

$$\nabla_\pi \mathcal{L}_{\mathrm{AR}}(\widehat{\pi}^n_{\mathrm{MLE}}(\vartheta), Y^n_\vartheta) = 0\ \forall \vartheta \in \Theta.$$

By the implicit function theorem, the derivative $\frac{\partial}{\partial \vartheta_j}\widehat{\pi}^n_{\mathrm{MLE}}(\vartheta)$ for $j \in \{1, \ldots, N(\Theta)\}$ is then given by

$$\begin{aligned}\frac{\partial}{\partial \vartheta_j}\widehat{\pi}^n_{\mathrm{MLE}}(\vartheta) &= -\left(\nabla_\pi(\nabla_\pi \mathcal{L}_{\mathrm{AR}}(\widehat{\pi}^n_{\mathrm{MLE}}(\vartheta), Y^n_\vartheta))\right)^{-1} \frac{\partial}{\partial \vartheta_j}\nabla_\pi \mathcal{L}_{\mathrm{AR}}(\widehat{\pi}^n_{\mathrm{MLE}}(\vartheta), Y^n_\vartheta) \\ &= -\left(\nabla^2_\pi \mathcal{L}_{\mathrm{AR}}(\widehat{\pi}^n_{\mathrm{MLE}}(\vartheta), Y^n_\vartheta)\right)^{-1} \frac{\partial}{\partial \vartheta_j}\nabla_\pi \mathcal{L}_{\mathrm{AR}}(\widehat{\pi}^n_{\mathrm{MLE}}(\vartheta), Y^n_\vartheta). \qquad (4.39)\end{aligned}$$

We consider the asymptotic behavior of the two terms on the right–hand side separately. Since $\widehat{\pi}^n_{\mathrm{MLE}}(\vartheta) \to \pi_\vartheta$ holds $\mathbb{P}$-a.s. by Theorem 4.37 (note that ϑ is the "true" parameter in this scenario and the driving Lévy process satisfies Assumptions B.2 and B.9), we obtain that

$$-\left(\nabla^2_\pi \mathcal{L}_{\mathrm{AR}}(\widehat{\pi}^n_{\mathrm{MLE}}(\vartheta), Y^n_\vartheta)\right)^{-1} \longrightarrow -\mathbb{E}\left[\nabla^2_\pi l_{\mathrm{AR},1}(\pi_\vartheta, \vartheta)\right]^{-1},\ n \to \infty,\ \mathbb{P}\text{-a.s.}$$

For the second factor, observe that by (4.37) and (4.38) we have for $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, N(\Theta)\}$

$$\begin{aligned}&\frac{\partial}{\partial \vartheta_j}\frac{\partial}{\partial \pi_i}\mathcal{L}_{\mathrm{AR}}(\pi, Y^n_\vartheta) \\ &= -\frac{2}{(n-r)\sigma^2}\cdot \\ &\sum_{t=1}^{n-r}\Bigg(\left(\frac{\partial}{\partial \vartheta_j}Y_\vartheta((t+r)h) - \pi_r\frac{\partial}{\partial \vartheta_j}Y_\vartheta((t+r-1)h) - \ldots - \pi_1\frac{\partial}{\partial \vartheta_j}Y_\vartheta(th)\right)Y_\vartheta((t+r-i)h) \\ &+(Y_\vartheta((t+r)h) - \pi_r Y_\vartheta((t+r-1)h) - \ldots - \pi_1 Y_\vartheta(th))\frac{\partial}{\partial \vartheta_j}Y_\vartheta((t+r-i)h)\Bigg) \qquad (4.40)\end{aligned}$$

and similarly

$$\begin{aligned}
&\frac{\partial}{\partial\vartheta_j}\frac{\partial}{\partial\sigma}\mathcal{L}_{\mathrm{AR}}(\pi, Y_\vartheta^n)\\
&= -\frac{4}{(n-r)\sigma^3}\sum_{t=1}^{n-r}\Bigg(\left(Y_\vartheta((t+r)h) - \pi_r Y_\vartheta((t+r-1)h) - \ldots - \pi_1 Y_\vartheta(th)\right)\\
&\left(\frac{\partial}{\partial\vartheta_j}Y_\vartheta((t+r)h) - \pi_r\frac{\partial}{\partial\vartheta_j}Y_\vartheta((t+r-1)h) - \ldots - \pi_1\frac{\partial}{\partial\vartheta_j}Y_\vartheta(th)\right)\Bigg). \qquad (4.41)
\end{aligned}$$

For an arbitrary $t \in \mathbb{R}$, we have by assumption that

$$\frac{\partial}{\partial\vartheta_j}Y_\vartheta(t) = \frac{\partial}{\partial\vartheta_j}\int_{-\infty}^{t} C_\vartheta\,\mathrm{e}^{A_\vartheta(t-u)}\,B_\vartheta \mathrm{d}L_S(u) = \int_{-\infty}^{t}\frac{\partial}{\partial\vartheta_j}(C_\vartheta\,\mathrm{e}^{A_\vartheta(t-u)}\,B_\vartheta)\mathrm{d}L_S(u).$$

Since we are in the one–dimensional case and use the parametrization from Example 3.1 (as described at the beginning of Subsection 4.3.1), the vector C_ϑ is independent of ϑ and always equal to the first unit vector in $\mathbb{R}^p$. Therefore, by [Schlemm 2011, Proposition 3.14] it holds that

$$\begin{aligned}
\frac{\partial}{\partial\vartheta_j}(C_\vartheta\,\mathrm{e}^{A_\vartheta(t-u)}\,B_\vartheta) &= C\frac{\partial}{\partial\vartheta_j}\left(\mathrm{e}^{A_\vartheta(t-u)}\,B_\vartheta\right)\\
&= C\left(\frac{\partial}{\partial\vartheta_j}\,\mathrm{e}^{A_\vartheta(t-u)}\right)B_\vartheta + C\,\mathrm{e}^{A_\vartheta(t-u)}\left(\frac{\partial}{\partial\vartheta_j}B_\vartheta\right)
\end{aligned}$$

Note that the derivative $\frac{\partial}{\partial\vartheta_j}\,\mathrm{e}^{A_\vartheta(t-u)}$ of the matrix–valued function $\vartheta \mapsto \mathrm{e}^{A_\vartheta(t-u)}$ exists for each $j \in \{1,\ldots,N(\Theta)\}$ and is a continuous function by Assumption B.4. It can also be interpreted as the partial derivative with respect to ϑ_j of the real–valued function $\vartheta \mapsto Ce^{A_\vartheta(t-u)}B$ evaluated for $B = B_\vartheta$ (an explicit formula can be obtained from [Tsai and Chan 2003, Theorem 4] if the eigenvalues of A_ϑ are all distinct). We now define

$$g_{\vartheta,j}(t) := \left(C\left(\frac{\partial}{\partial\vartheta_j}\,\mathrm{e}^{A_\vartheta t}\right)B_\vartheta + C\,\mathrm{e}^{A_\vartheta t}\left(\frac{\partial}{\partial\vartheta_j}B_\vartheta\right)\right)\mathbb{1}_{[0,\infty)}(t),$$

for $j \in \{1,\ldots,N(\Theta)\}$ and obtain that

$$\frac{\partial}{\partial\vartheta_j}Y_\vartheta(t) = \int_{-\infty}^{\infty} g_{\vartheta,j}(t-u)\mathrm{d}L_S(u).$$

Setting

$$G_\vartheta(t) := \begin{pmatrix} g_\vartheta(t) \\ g_{\vartheta,1}(t) \\ \vdots \\ g_{\vartheta,N(\Theta)}(t) \end{pmatrix},$$

where $g_\vartheta(t) = C_\vartheta \, \mathrm{e}^{A_\vartheta t} \, B_\vartheta \mathbb{1}_{[0,\infty)}(t)$ is the kernel function of the moving average representation of the process $(Y_\vartheta(t))_{t\in\mathbb{R}}$. With this, we can write

$$\begin{pmatrix} Y_\vartheta(t) \\ \nabla_\vartheta Y_\vartheta(t) \end{pmatrix} = \int_{-\infty}^{\infty} G_\vartheta(t-u) \mathrm{d}L_S(u).$$

Therefore, the process $\begin{pmatrix} Y_\vartheta(t) \\ \nabla_\vartheta Y_\vartheta(t) \end{pmatrix}_{t\in\mathbb{R}}$ is a multivariate continuous–time moving average process with kernel function G_ϑ. By [Fuchs and Stelzer 2013, Theorem 3.5], we obtain that this process is mixing and therefore in particular ergodic. From (4.40) and (4.41), we see that for each $j \in \{1, \ldots, N(\Theta)\}$ it holds that

$$\begin{aligned} &\frac{\partial}{\partial \vartheta_j} \nabla_\pi \mathcal{L}_{\mathrm{AR}}(\pi, Y_\vartheta^n) \\ =&\frac{C_j}{n-r} \sum_{t=1}^{n-r} h_j \left(\pi, Y_\vartheta((t+r)h), \ldots, Y_\vartheta(th), \nabla_\vartheta Y_\vartheta((t+r)h), \ldots, \nabla_\vartheta Y_\vartheta(th)\right) \end{aligned}$$

for constants $C_j \in \mathbb{R}$ and measurable functions h_j. By the ergodicity of $\begin{pmatrix} Y_\vartheta(t) \\ \nabla_\vartheta Y_\vartheta(t) \end{pmatrix}_{t\in\mathbb{R}}$ and [Bradley 2007, Proposition 2.10(II)], each of the processes

$$\left(h_j \left(\pi, Y_\vartheta((t+r)h), \ldots, Y_\vartheta(th), \nabla_\vartheta Y_\vartheta((t+r)h), \ldots, \nabla_\vartheta Y_\vartheta(th)\right)\right)_{t\in\mathbb{Z}}, \; j = 1, \ldots, N(\Theta),$$

is ergodic again, so that we obtain by Birkhoff's ergodic theorem that

$$\nabla_\vartheta \nabla_\pi \mathcal{L}_{\mathrm{AR}}(\pi, Y_\vartheta^n) \longrightarrow \mathbb{E}\left[\nabla_\vartheta \nabla_\pi l_{\mathrm{AR},1}(\pi, \vartheta)\right], \; \mathbb{P}\text{-a.s.}, \; n \to \infty. \tag{4.42}$$

From (4.40) and (4.41), it is obvious that $\nabla_\vartheta \nabla_\pi \mathcal{L}_{\mathrm{AR}}(\pi, Y_\vartheta^n)$ is differentiable once more with respect to π. By the same arguments that led to (4.42) and an application of [Ferguson 1996, Theorem 16a)], using the compactness of Π', we can then obtain that

$$\sup_{\pi\in\Pi'} \| \nabla_\pi \nabla_\vartheta \nabla_\pi \mathcal{L}_{\mathrm{AR}}(\pi, Y_\vartheta^n) - \mathbb{E}\left[\nabla_\pi \nabla_\vartheta \nabla_\pi l_{\mathrm{AR},1}(\pi, \vartheta)\right] \| \to 0, \; \mathbb{P}\text{-a.s.}, \; n \to \infty.$$

By arguing as in the proof of [Schlemm and Stelzer 2012, Theorem 2.5], we can infer from this that the sequence of random functions $\pi \mapsto \nabla_\pi \nabla_\vartheta \nabla_\pi \mathcal{L}_{\mathrm{AR}}(\pi, Y_\vartheta^n)$ is almost surely uniformly bounded on the compact set Π'. Therefore,

$$\begin{aligned}&\|\nabla_\vartheta \nabla_\pi \mathcal{L}_{\mathrm{AR}}(\widehat{\pi}_{\mathrm{MLE}}^n(\vartheta), Y_\vartheta^n) - \nabla_\vartheta \nabla_\pi \mathcal{L}_{\mathrm{AR}}(\pi_\vartheta, Y_\vartheta^n)\| \\ &\leq \sup_{\pi \in \Pi'} \|\nabla_\pi \nabla_\vartheta \nabla_\pi \mathcal{L}_{\mathrm{AR}}(\pi, Y_\vartheta^n)\| \|\widehat{\pi}_{\mathrm{MLE}}^n(\vartheta) - \pi_\vartheta\| \to 0 \ \mathbb{P}\text{-a.s.}, \ n \to \infty,\end{aligned}$$

since $\widehat{\pi}_{\mathrm{MLE}}^n(\vartheta) \to \pi_\vartheta$ holds $\mathbb{P}$-a.s. by Theorem 4.37. By (4.42), we can deduce from this that

$$\nabla_\vartheta \nabla_\pi \mathcal{L}_{\mathrm{AR}}(\widehat{\pi}_{\mathrm{MLE}}^n(\vartheta), Y_\vartheta^n) \longrightarrow \mathbb{E}\left[\nabla_\vartheta \nabla_\pi l_{\mathrm{AR},1}(\pi_\vartheta, \vartheta)\right], \ \mathbb{P}\text{-a.s.}, \ n \to \infty.$$

Plugging this into (4.39), we finally obtain that

$$\nabla_\vartheta \widehat{\pi}_{\mathrm{MLE}}^n(\vartheta) \to -\mathbb{E}\left[\nabla_\pi^2 l_{\mathrm{AR},1}(\pi_\vartheta, \vartheta)\right]^{-1} \mathbb{E}\left[\nabla_\vartheta \nabla_\pi l_{\mathrm{AR},1}(\pi_\vartheta, \vartheta)\right] \ \mathbb{P}\text{-a.s.}, \ n \to \infty.$$

Exchanging derivative and expectation by means of the dominated convergence theorem (using that the set $\Theta \times \Pi'$ is compact and the continuity of the respective derivatives), we can also express this as

$$\nabla_\vartheta \widehat{\pi}_{\mathrm{MLE}}^n(\vartheta) \to -\nabla_\pi^2 \mathbb{E}\left[l_{\mathrm{AR},1}(\pi_\vartheta, \vartheta)\right]^{-1} \nabla_\vartheta \nabla_\pi \mathbb{E}\left[l_{\mathrm{AR},1}(\pi_\vartheta, \vartheta)\right] \ \mathbb{P}\text{-a.s.}, \ n \to \infty.$$

Since the function $\vartheta \mapsto \pi_\vartheta$ is characterized implicitly by the equation

$$\nabla_\pi \mathbb{E}\left[l_{\mathrm{AR},1}(\pi_\vartheta, \vartheta)\right] = 0 \ \forall \vartheta \in \Theta,$$

we can deduce via the implicit function theorem that

$$\nabla_\vartheta \pi_\vartheta = -\nabla_\pi^2 \mathbb{E}\left[l_{\mathrm{AR},1}(\pi_\vartheta, \vartheta)\right]^{-1} \nabla_\vartheta \nabla_\pi \mathbb{E}\left[l_{\mathrm{AR},1}(\pi_\vartheta, \vartheta)\right],$$

from which we obtain that

$$\nabla_\vartheta \widehat{\pi}_{\mathrm{MLE}}^n(\vartheta) \to \nabla_\vartheta \pi_\vartheta \ \mathbb{P}\text{-a.s.}, \ n \to \infty \tag{4.43}$$

pointwise for every $\vartheta \in \Theta$. Eventually, for a sequence $(\overline{\vartheta}^n)_{n \in \mathbb{N}}$ with $\overline{\vartheta}^n \to \vartheta_0$ $\mathbb{P}$-a.s. as $n \to \infty$ we have:

$$\begin{aligned}\|\nabla_\vartheta \widehat{\pi}_{\mathrm{MLE}}^n(\overline{\vartheta}^n) - \nabla_\vartheta \pi_{\vartheta_0}\| &\leq \|\nabla_\vartheta \widehat{\pi}_{\mathrm{MLE}}^n(\overline{\vartheta}^n) - \nabla_\vartheta \widehat{\pi}_{\mathrm{MLE}}^n(\vartheta_0)\| + \|\nabla_\vartheta \widehat{\pi}_{\mathrm{MLE}}^n(\vartheta_0) - \nabla_\vartheta \pi_{\vartheta_0}\| \\ &\leq \sup_{\vartheta \in \Theta} \|\nabla_\vartheta^2 \widehat{\pi}_{\mathrm{MLE}}^n(\vartheta)\| \|\overline{\vartheta}^n - \vartheta_0\| + \|\nabla_\vartheta \widehat{\pi}_{\mathrm{MLE}}^n(\vartheta_0) - \nabla_\vartheta \pi_{\vartheta_0}\|.\end{aligned}$$

By assumption, $\sup_{\vartheta\in\Theta} \|\nabla^2_\vartheta \widehat{\pi}^n_{\mathrm{MLE}}(\vartheta)\| \leq C$ holds $\mathbb{P}$-a.s. for a constant $C > 0$ independent of n. By the fact that $\overline{\vartheta}^n \to \vartheta_0$ $\mathbb{P}$-a.s. as $n \to \infty$ and (4.43), we can deduce that the right–hand side converges to 0 $\mathbb{P}$-a.s. as $n \to \infty$ and the proof is complete. □

Remark 4.39. *a) Since the QMLE and the LS estimator coincide (cf. Remark 4.36) in this context, under analogous assumptions as in Theorem 4.38 the results of that theorem also hold for $\widehat{\pi}^n_{LS}$.*

b) Since we assumed that for every $t \in \mathbb{R}$ and every $j \in \{1, \ldots, N(\Theta)\}$ that

$$\frac{\partial}{\partial\vartheta_j} Y_\vartheta(t) = \frac{\partial}{\partial\vartheta_j} \int_{-\infty}^t C_\vartheta \, \mathrm{e}^{A_\vartheta(t-u)} \, B_\vartheta \mathrm{d}L_S(u) = \int_{-\infty}^t \frac{\partial}{\partial\vartheta_j} (C_\vartheta \, \mathrm{e}^{A_\vartheta(t-u)} \, B_\vartheta) \mathrm{d}L_S(u)$$

holds $\mathbb{P}$-a.s., we can interpret the derivative $\frac{\partial}{\partial\vartheta_j} Y_\vartheta(t)$ pathwise as the sensitivity of $Y_\vartheta(t)$ with respect to changes in the kernel function g_ϑ. It is important to notice that the driving Lévy process does not depend on ϑ in this context, i.e. the "randomness" in the paths of $(Y_\vartheta(t))_{t\in\mathbb{R}}$ does not depend on ϑ and is not affected by the derivative, which is why the pathwise interpretation of the derivative makes sense here and the sensitivity of the kernel function with respect to the parameters is the deciding factor in the derivative.

4.3.4. Robustness properties of the indirect estimator

In Subsection 4.3.2, we were able to show that the indirect estimator is strongly consistent and asymptotically normally distributed in the case that there are no outliers, i.e. when we perfectly observe the data–generating CARMA process. Complementing this, as mentioned at the beginning of this chapter, a robust estimator should also perform well in the presence of outliers. The study of the indirect estimator with respect to this is the topic of this subsection.

We work under the following assumptions: We operate in a parameter space Θ satisfying assumptions B.1 - B.9. Θ contains a true parameter ϑ_0 such that $Y = \mathrm{CARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0})$ holds for the data–generating CARMA(p,q) process $(Y(t))_{t\in\mathbb{R}}$. We do not observe the sampled process $(Y(nh))_{n\in\mathbb{Z}}$ directly, but instead the contaminated process $(\widetilde{Y}_n)_{n\in\mathbb{Z}}$ as defined in Definition 4.1 (specifically (4.2)). We assume that Assumption G is satisfied. A sample of length n of this process is denoted by $\widetilde{Y}^n$. For the auxiliary AR(r) representation of Definition 4.19, we assume that $r \geq 2p-1$, so that the binding function π is injective. For the indirect estimator as defined in Definition 4.22, we take $\widehat{\pi}^n$ as GM estimator as defined in Subsubsection 4.3.3.1 that satisfies Assumption H. For the simulation–based estimator $\widehat{\pi}^n_{\mathrm{S}}$, we can use any of the three estimators introduced in Subsection 4.3.3, but it is most convenient to think of either the least squares or the QML estimator, because these are easier to handle (remember that $\widehat{\pi}^n_{\mathrm{S}}$ is applied to outlier–free, simulated data, therefore there it is not needed to use a robust estimator here).

We will consider three different measures of robustness: qualitative resistance, the breakdown point and the influence functional, which all have different interpretations and represent different aspects of robustness. The starting point will be the notion of qualitative resistance.

4.3.4.1. Resistance and qualitative robustness

As outlined in the introduction to this chapter, the most fundamental questions when considering robustness of an estimator is how the estimator behaves when the data does not satisfy the model assumptions. One could intuitively call an estimator robust, when small deviations from the nominal model do not have much effect on the estimator. In the context of a CARMA process and the model we introduced in Section 4.1, this then translates to demanding that a small number of outliers in a data sample should not exert too much influence on the estimator. This property is known as qualitative robustness or resistance of the estimator and was originally introduced in Hampel [1971] for i.i.d. observations, who measured deviations from

the nominal model in terms of the Prokhorov distance on the set of probability distributions. The same article also gives a slight extension to the case of data that is generated by permutation–invariant distributions, introducing the term π–robustness ([Hampel 1971, p.1893]). Of course, time series do not satisfy the assumption of permutation invariance in general. Therefore, there have been various attempts to generalize the concept of qualitative robustness to the time series setting. However, there is no unique, intuitively correct way of doing so. Diverse approaches can be found in Papantoni-Kazakos and Gray [1979], Cox [1981] or Boente et al. [1987], among others.

In Boente et al. [1987, Remark 3.1] it is explained that the concepts of Cox [1981] and Papantoni-Kazakos and Gray [1979] do not seem adequate for the time series context. The reason is that, when using these concepts, estimators which depend only on a fixed set of coordinates can be robust. This is intuitively contradictory, since a small percentage of the observations then completely controls the behavior of the estimator. Moreover, Boente et al. [1987] argue that their concept of π_{d_n}–robustness best generalizes Hampel's original idea. [Boente et al. 1987, Theorem 3.1] proves that π_{d_n}–robustness is equivalent to Hampel's π–robustness for i.i.d. processes and therefore extends it to the setting of time series. They then go ahead and define the term of resistance. Boente et al. [1987, Theorem 4.2] shows that resistance in their sense again implies π_{d_n} robustness under mild conditions. The concept of resistance also has the intuitive appeal of making a statement about changes in the values of the estimator when comparing two deterministic samples, while π_{d_n}–robustness is only a statement concerning the distribution of the estimator, which is in general not easily tractable. For these reasons, we apply the definitions in the sense of Boente et al. [1987] here and explore the respective properties for our indirect estimator.

To this end, let y be a (infinite-length) realization of the discretely sampled, data–generating CARMA process $(Y(nh))_{n\in\mathbb{Z}}$. Formally, we can write that $y \in \mathbb{R}^\infty$, where $\mathbb{R}^\infty$ denotes the infinite cartesian product of $\mathbb{R}$ with itself. On this space, equipped with the Borel σ-field $\mathcal{B}^\infty$ we denote the set of all probability measures by $\mathcal{P}(\mathbb{R}^\infty)$. In the following, we denote for $y \in \mathbb{R}^\infty$ as above by y^n the vector of the first n coordinates, i. e. $y^n = (y(h), y(2h), \ldots, y(nh))$. We can now define resistance:

Definition 4.40. *Let $y \in \mathbb{R}^\infty$ and let $(\widehat{\vartheta}^n)_{n\in\mathbb{N}}$ be a sequence of estimators. Denote by $\widehat{\vartheta}^n(z^n)$ the value of $\widehat{\vartheta}^n$ when it is calculated using the deterministic realization $z^n \in \mathbb{R}^n$.*

a) $(\widehat{\vartheta}^n)_{n\in\mathbb{N}}$ is called resistant at y if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$\sup\left\{\|\widehat{\vartheta}^n(z^n) - \widehat{\vartheta}^n(w^n)\| : z^n, w^n \in B_\delta(y^n)\right\} \leq \epsilon \quad \forall n \in \mathbb{N}, \tag{4.44}$$

where $B_\delta(x)$ denotes the open ball with center x and radius δ with respect to the metric

$$d_n(z^n, w^n) = \inf\left\{\epsilon : \frac{\#\{i \in \{1,\dots,n\} : |z_i^n - w_i^n| \geq \epsilon\}}{n} \leq \epsilon\right\}. \tag{4.45}$$

We say that $(\widehat{\vartheta}^n)_{n\in\mathbb{N}}$ is asymptotically resistant at y if for every $\epsilon > 0$ there exists $\delta > 0$ and $N_0(\epsilon, y) \in \mathbb{N}$ such that (4.44) holds for $n \geq N_0(\epsilon, y)$.

b) For $\mathbb{Q} \in \mathcal{P}(\mathbb{R}^\infty)$ we say that $(\widehat{\vartheta}^n)_{n\in\mathbb{N}}$ is strongly resistant at $\mathbb{Q}$ if

$$\mathbb{Q}\left(\left\{y \in \mathbb{R}^\infty : (\widehat{\vartheta}^n)_{n\in\mathbb{N}} \text{ is resistant at } y\right\}\right) = 1.$$

We say that $(\widehat{\vartheta}^n)_{n\in\mathbb{N}}$ is asymptotically strongly resistant at $\mathbb{Q}$ if

$$\mathbb{Q}\left(\left\{y \in \mathbb{R}^\infty : (\widehat{\vartheta}^n)_{n\in\mathbb{N}} \text{ is asymptotically resistant at } y\right\}\right) = 1.$$

With this definition at hand, we want to study the question whether our indirect estimator for CARMA processes is resistant. We will make use of the fact that it is built out of two blocks, the GM estimator of the auxiliary AR representation, which deals with possible outliers in the observations, and the outlier–free estimator of the AR representation based on simulated data. As it turns out, under our assumptions from Assumption H, GM estimators applied to a certain class of stationary, ergodic processes are already asymptotically strongly resistant; the discretely sampled CARMA process is a special case.

Theorem 4.41. *Let $\widehat{\pi}^n$ be a GM estimator as defined in (4.24) and (4.25), where ϕ and χ fulfill Assumption H and assume that the solutions of (4.22) and (4.23) are unique. Then $(\widehat{\pi}^n)_{n\in\mathbb{N}}$ is asymptotically strongly resistant at the measure $\mathbb{P}_{Y(h)}$, which is the probability measure associated to the distribution of the data–generating CARMA process $(Y(nh))_{n\in\mathbb{Z}}$.*

Proof. The statement follows from [Boente et al. 1987, Theorem 5.1]. The theorem requires that ϕ and χ fulfill Assumption H, that the limiting equation has a unique solution, which we assumed, and that $(Y(nh))_{n\in\mathbb{Z}}$ is ergodic and fulfills G.4, which is automatically given for every sampled stationary CARMA process, meaning it is especially given for the data–generating process. □

The next step now consists of establishing that the asymptotic strong resistance of the GM estimators transfers to the indirect estimator, which is not very hard since

we can again make use of the fact that no outliers are present in the simulation–based estimate used to construct the indirect estimate.

Theorem 4.42. *Let $(Y(nh))_{n\in\mathbb{Z}}$ be the data–generating CARMA process sampled at discrete points in time, let $\widehat{\vartheta}^n_{Ind}$ be defined as in (4.10) and let the assumptions of Theorem 4.41 be satisfied. Then, $(\widehat{\vartheta}^n_{Ind})_{n\in\mathbb{N}}$ is asymptotically strongly resistant at the measure $\mathbb{P}_{Y^{(h)}}$, which is the probability measure associated to the distribution of the data–generating CARMA process $(Y(nh))_{n\in Z}$.*

Proof. From the proof of Theorem 4.23a) we know that there exists a set A_0 with $\mathbb{P}_{Y^{(h)}}(A_0) = 1$ such that $\mathcal{L}_{\text{Ind}}(\vartheta, y^n)$ converges uniformly in ϑ to $\mathscr{Q}_{\text{Ind}}(\vartheta)$ as $n \to \infty$ for every $y \in A_0$. Since the function $\mathscr{Q}_{\text{Ind}}$ has a unique minimum at ϑ_0 and for this minimum it holds that $\mathscr{Q}_{\text{Ind}}(\vartheta_0) = 0$, we have that for every $\epsilon > 0$ there exist an $\eta > 0$ and a $N_1(y) \in \mathbb{N}$ such that for $y \in A_0$ and every $n \geq N_1(y)$ it holds that

$$\inf_{|\vartheta-\vartheta_0|\geq\frac{\epsilon}{2}} |\mathcal{L}_{\text{Ind}}(\vartheta, y^n)| > \eta \tag{4.46}$$

and

$$|\mathcal{L}_{\text{Ind}}(\vartheta_0, y^n)| \leq \frac{\eta}{4}. \tag{4.47}$$

By Theorem 4.23a) we know that $\widehat{\vartheta}^n_{\text{Ind}}(y^n)$ converges to ϑ_0 as $n \to \infty$ for every $y \in A_0$. Therefore, there exists a $N_2(y) \in \mathbb{N}$ such that for every $n \geq N_2(y)$ we have that

$$\|\widehat{\vartheta}^n_{\text{Ind}}(y^n) - \vartheta_0\| < \frac{\epsilon}{2}. \tag{4.48}$$

Under the assumptions of Theorem 4.41, the GM estimator $\widehat{\pi}^n$ is asymptotically strongly resistant at $\mathbb{P}_{Y^{(h)}}$. Therefore, there exists a set $A \subseteq \mathbb{R}^\infty$ with $\mathbb{P}_{Y^{(h)}}(A) = 1$ such that for every $y \in A$ and every $\epsilon' > 0$ there exists a $\delta' > 0$ and a natural number $N_2(y) \in \mathbb{N}$ with

$$\|\widehat{\pi}^n(z^n) - \widehat{\pi}^n(y^n)\| \leq \epsilon' \quad \forall n \geq N_1(y)$$

for every $z^n \in B_{\delta'}(y^n)$. By the definition of $\mathcal{L}_{\text{Ind}}$ in (4.9) and the asymptotic strong resistance of $\widehat{\pi}^n$ at $\mathbb{P}_{Y^{(h)}}$, we have that for η as in (4.46) and (4.47), there exists a $\delta > 0$ and $N_3(y) \in \mathbb{N}$ such that for every $n \geq N_3(y)$ and for every $z^n \in B_\delta(y^n)$ it holds that

$$\begin{aligned}
&\sup_{\vartheta\in\Theta} |\mathcal{L}_{\text{Ind}}(\vartheta, y^n) - \mathcal{L}_{\text{Ind}}(\vartheta, z^n)| \\
&= \sup_{\vartheta\in\Theta} | - (\widehat{\pi}^n(y^n) - \widehat{\pi}^n(z^n))^T \Omega \widehat{\pi}^n_S(\vartheta) - \widehat{\pi}^n_S(\vartheta)^T \Omega (\widehat{\pi}^n(y^n) - \widehat{\pi}^n(z^n)) \\
&\qquad + \widehat{\pi}^n(y^n)^T \Omega \widehat{\pi}^n(y^n) - \widehat{\pi}^n(z^n)^T \Omega \widehat{\pi}^n(z^n)|
\end{aligned}$$

$$\begin{aligned}&\leq \sup_{\vartheta\in\Theta} 2\|\widehat{\pi}_S^n(\vartheta)\|\|\Omega\|\|\widehat{\pi}^n(y^n)-\widehat{\pi}^n(z^n)\| + |\widehat{\pi}^n(y^n)^T\Omega\widehat{\pi}^n(y^n)-\widehat{\pi}^n(z^n)^T\Omega\widehat{\pi}^n(z^n)|\\ &\leq \sup_{\pi\in\Pi'} \|\pi\|\|\Omega\|\|\widehat{\pi}^n(y^n)-\widehat{\pi}^n(z^n)\| + |\widehat{\pi}^n(y^n)^T\Omega\widehat{\pi}^n(y^n)-\widehat{\pi}^n(z^n)^T\Omega\widehat{\pi}^n(z^n)| \leq \frac{\eta}{4}\end{aligned} \tag{4.49}$$

This holds since $\|\widehat{\pi}^n(y^n)-\widehat{\pi}^n(z^n)\|$ can be made arbitrarily small by choosing δ and $N_3(y)$ suitably, because the space $\Pi'=\pi(\Theta)$ is compact, i.e. $\sup_{\pi\in\Pi'}\|\pi\|\leq C$ for some $C>0$, and because the quadratic form $x\mapsto x^T\Omega x$ is continuous.
For $y\in A\cap A_0$, $n\geq\max\{N_1(y),N_2(y)N_3(y)\}$ and for every $z^n\in B_\delta(y^n)$ it then holds by (4.46) and (4.47) that

$$|\mathcal{L}_{\mathrm{Ind}}(\vartheta_0,z^n)| = |\mathcal{L}_{\mathrm{Ind}}(\vartheta_0,z^n)-\mathcal{L}_{\mathrm{Ind}}(\vartheta_0,y^n)+\mathcal{L}_{\mathrm{Ind}}(\vartheta_0,y^n)| \leq \frac{\eta}{4}+\frac{\eta}{4}=\frac{\eta}{2}.$$

Likewise, (4.46) and (4.49) give us that

$$\begin{aligned}\inf_{|\vartheta-\vartheta_0|\geq\frac{\epsilon}{2}} |\mathcal{L}_{\mathrm{Ind}}(\vartheta,z^n)| &\geq \inf_{|\vartheta-\vartheta_0|\geq\frac{\epsilon}{2}} (|\mathcal{L}_{\mathrm{Ind}}(\vartheta,y^n)| - |\mathcal{L}_{\mathrm{Ind}}(\vartheta,z^n)-\mathcal{L}_{\mathrm{Ind}}(\vartheta,y^n)|)\\ &\geq \inf_{|\vartheta-\vartheta_0|\geq\frac{\epsilon}{2}} |\mathcal{L}_{\mathrm{Ind}}(\vartheta,y^n)| - \sup_{|\vartheta-\vartheta_0|\geq\frac{\epsilon}{2}} |\mathcal{L}_{\mathrm{Ind}}(\vartheta,z^n)-\mathcal{L}_{\mathrm{Ind}}(\vartheta,y^n)|\\ &\geq \eta-\frac{\eta}{4}=\frac{3\eta}{4}.\end{aligned}$$

Since $\widehat{\vartheta}^n_{\mathrm{Ind}}(z^n)$ minimizes $\mathcal{L}_{\mathrm{Ind}}(\vartheta,z^n)$ per definition, it must therefore hold that

$$\|\widehat{\vartheta}^n_{\mathrm{Ind}}(z^n)-\vartheta_0\|<\frac{\epsilon}{2}$$

for every $n\geq\max\{N_1(y),N_2(y),N_3(y)\}$. For every $y\in A\cap A_0$ and every $z^n\in B_\delta(y^n)$ we obtain from this and (4.48) that for every $n\geq\max\{N_1(y),N_2(y),N_3(y)\}$ it holds that

$$\|\widehat{\vartheta}^n_{\mathrm{Ind}}(z^n)-\widehat{\vartheta}^n_{\mathrm{Ind}}(y^n)\| \leq \|\widehat{\vartheta}^n_{\mathrm{Ind}}(z^n)-\vartheta_0\| + \|\widehat{\vartheta}^n_{\mathrm{Ind}}(y^n)-\vartheta_0\| < \epsilon.$$

Since $\mathbb{P}_{Y^{(h)}}(A\cap A_0)=1$ and it of course holds for z^n and $w^n\in B_\delta(y^n)$ that

$$\|\widehat{\vartheta}^n_{\mathrm{Ind}}(z^n)-\widehat{\vartheta}^n_{\mathrm{Ind}}(w^n)\| \leq \|\widehat{\vartheta}^n_{\mathrm{Ind}}(z^n)-\widehat{\vartheta}^n_{\mathrm{Ind}}(y^n)\| + \|\widehat{\vartheta}^n_{\mathrm{Ind}}(y^n)-\widehat{\vartheta}^n_{\mathrm{Ind}}(w^n)\|.$$

this proves that $(\widehat{\vartheta}^n_{\mathrm{Ind}})_{n\in\mathbb{N}}$ is asymptotically strongly resistant at $\mathbb{P}_{Y^{(h)}}$.

□

As mentioned in the introduction of this subsection, one could also define qualitative robustness of a sequence of estimators by demanding that the distribution of the estimator does not change too much when the data is changed slightly, i.e. afflicted

by outliers. To make this notion explicit, we first need to define a pseudometric for measures on a metric space, the Prokhorov distance:

Definition 4.43. *For a metric space (M, d) with Borel sets $\mathcal{B}(M)$, the Prokhorov distance π_d between two measures μ, ν on $\mathcal{B}(M)$ with respect to d is defined as*

$$\pi_d(\mu, \nu) := \inf\{\epsilon > 0 : \mu(A) \leq \nu(\{x \in M : d(x, A) < \epsilon\}) + \epsilon \; \forall A \in \mathcal{B}(M)\}.$$

This pseudometric is a key component of the definition of qualitative robustness, which is as follows in our scenario:

Definition 4.44. *Let $\mathbb{P} \in \mathcal{P}(\mathbb{R}^\infty)$, let d_Θ be a metric on Θ, let ρ_n be a pseudometric on $\mathcal{P}(\mathbb{R}^n)$ for all $n \in \mathbb{N}$ and denote by $\mathbb{P}_n$ the n–th order marginal of $\mathbb{P}$. Finally, denote by $\mathbb{P}_{\widehat{\vartheta}_n} \in \mathcal{P}(\Theta)$ the distribution of the estimator $\widehat{\vartheta}_n$ under $\mathbb{P}_n$.*
Then, the sequence of estimators $(\widehat{\vartheta}_n)_{n\in\mathbb{N}}$ is said to be ρ_n–robust at $\mathbb{P}$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for every $\mathbb{Q}_n \in \mathcal{P}(\mathbb{R}^n)$ with $\rho_n(\mathbb{P}_n, \mathbb{Q}_n) < \delta$ it holds that

$$\pi_{d_\Theta}(\mathbb{P}_{\widehat{\vartheta}_n}, \mathbb{Q}_{\widehat{\vartheta}_n}) \leq \epsilon.$$

As shown in [Boente et al. 1987, Theorem 3.1], this is a direct generalization of the definition of π–robustness given by Hampel [1971] for i.i.d. processes. Moreover, for a special choice of ρ_n, this kind of robustness is implied by asymptotic strong resistance, which enables us to obtain:

Theorem 4.45. *Assume that the assumptions of Theorem 4.42 are fulfilled. Choose ρ_n as the Prokhorov distance on $\mathcal{B}(\mathbb{R}^n)$ with respect to d_n as defined in* (4.45)*, i.e.*

$$\begin{aligned}\rho_n(\mathbb{P}_n, \mathbb{Q}_n) &= \pi_{d_n}(\mathbb{P}_n, \mathbb{Q}_n) \\ &= \inf\{\epsilon > 0 : \mathbb{P}_n(A) \leq \mathbb{Q}_n(\{x^n \in \mathbb{R}^n : d_n(x^n, A) < \epsilon\}) + \epsilon \; \forall A \in \mathcal{B}(\mathbb{R}^n)\}.\end{aligned}$$

Then, the sequence of estimators $(\widehat{\vartheta}^n_{Ind})_{n\in\mathbb{N}}$ is π_{d_n}–robust at $\mathbb{P}_{Y^{(h)}}$, which is the probability measure associated to the distribution of the data–generating CARMA process $(Y(nh))_{n\in Z}$.

Proof. By [Cox 1981, Lemma 5], the GM estimator $\widehat{\pi}^n$ is a continuous function of y^n for every $n \in \mathbb{N}$. By definition, $\widehat{\vartheta}^n_{\text{Ind}}$ depends on y^n only through a continuous function applied to $\widehat{\pi}^n(y^n)$ and therefore $\widehat{\vartheta}^n_{\text{Ind}}$ is a continuous function of y^n for every $n \in \mathbb{N}$, too. From Theorem 4.42, it follows that $(\widehat{\vartheta}^n_{\text{Ind}})_{n\in\mathbb{N}}$ is asymptotically strongly resistant at $\mathbb{P}_{Y^{(h)}}$. By [Boente et al. 1987, Proposition 4.2], these two properties imply that $(\widehat{\vartheta}^n_{\text{Ind}})_{n\in\mathbb{N}}$ is strongly resistant at $\mathbb{P}_{Y^{(h)}}$. [Boente et al. 1987, Theorem 4.2a)] then gives the π_{d_n}–robustness of $(\widehat{\vartheta}^n_{\text{Ind}})_{n\in\mathbb{N}}$, since it is implied by strong resistance. □

In conclusion, we can say that our indirect estimator is asymptotically strongly resistant as well as π_{d_n}–robust under the same assumptions we used to derive Theorem 4.23, i.e. no additional assumptions were necessary. This is in contrast to, for example, M–estimators, which are not qualitatively robust even in the case of linear regression ([Maronna and Yohai 1981, p.8]).

4.3.4.2. The breakdown point

Intuitively speaking, the breakdown point is (for a sample of data with fixed length n) the maximum percentage of outliers which can be contained in the data without "ruining" the estimator. In this sense, it measures by how much the observed data can deviate from the nomial model before catastrophic effects in the estimation procedure happen. How this idea should adequately be formalized, however, depends strongly on the model under consideration. The term was coined originally in [Hampel 1971, Section 6] for i.i.d. observations in the asymptotic framework. It was later extended, on the one hand to the case of a finite number of observations in Donoho and Huber [1983] and on the other hand to different models, such as regression models, e.g. in Maronna et al. [1979] or Maronna and Yohai [1991] among many others (these two references deal explicitly with GM estimators) and the time series context, e.g. in Martin and Yohai [1985] and Martin [1980]. Unfortunately, in the literature there is no generally accepted way of defining the breakdown point for time series. The definition for i.i.d. observations is not directly transferable because in time series the configuration of outliers (i.e. at which times and whether they appear in patches or isolated) has a decisive effect on the performance of estimators, which is of course absent in the case of i.i.d. observations, where estimators are typically invariant to permutation of the data. Furthermore, for time series, determining under which conditions an estimator has broken down may depend on the model – for a stationary AR(1) process, for example, one could argue on the one hand that a breakdown has occurred if the parameter estimate is equal to 0, since then the time series is indistinguishable from the driving noise. On the other hand, it would also seem reasonable to say that the estimator has broken if the estimate is equal to 1 or -1, since these values of the parameter do no longer provide a stationary AR(1) model. Defining the breakdown point in this way, specifically tailored to the model under consideration, is somewhat unsatisfactory, especially considering the fact that the edge cases are not always as easily identifiable as in the AR(1) example.
For a very general class of models and estimators, the breakdown point as defined in Genton and Lucas [2003] takes these problems into account. Heuristically speaking, the fundamental idea of that definition is that the breakdown point is the smallest

amount of outlier contamination with the property that the performance of the estimator does not get worse anymore if the contamination is increased further. For a formal definition, see [Genton and Lucas 2003, Definition 1 and Definition 2]. The downside of this definition is that, except for special cases, the defining expression is not analytically tractable and it is therefore very difficult to explicitly calculate the breakdown point for a given estimator and a given contamination model.
For our indirect estimator $\widehat{\vartheta}^n_{\text{Ind}}$ as defined in (4.10), studying the breakdown point means studying the breakdown point of the GM estimator $\widehat{\pi}^n$ which is obtained by applying the method of Subsubsection 4.3.3.1 to the observations $\widetilde{Y}_1, \ldots, \widetilde{Y}_n$. The reason is that the other building block of $\widehat{\vartheta}^n_{\text{Ind}}$, $\widehat{\pi}^n_{\text{S}}$, is obtained by applying an estimator to a simulated, outlier–free sample. Hence, if the rate of contamination is high enough to "ruin the estimator", it must ruin the GM estimator. In the literature, one often finds the fact that the breakdown point of a GM estimator, applied to an AR(r) process is positive, but bounded from above by $\frac{1}{r+1}$. The earliest mention of this in a published article seems to be [Martin 1980, p. 239]. However, no calculation is presented to support this. Later literature often cites this result, too, e.g. [de Luna and Genton 2001, p. 377] and [Genton and Lucas 2003, p. 89]. In Genton and Lucas [2003], the authors mention that this result is consistent with their definition of breakdown, unfortunately without providing a proof as well. The only reference containing explicit calculations to our knowledge seems to be Martin and Jong [1977], which is only availabe in form of an unpublished technical memorandum. For this reason, its content could not be accessed in order to check the calculations and, unfortunately, carrying out the calculations to verify the alluded upper bound of $\frac{1}{r+1}$ proved a task which was not succesful. We can only give an intuitive justification in a special case as follows:
Suppose that our sample of length n is of the form

$$\widetilde{Y}^n = (Y(h), Y(2h), \ldots, Y((r+1)h)+\xi, Y((r+2)h), \ldots, Y(2(r+1)h)+\xi, \ldots, Y(nh)), \tag{4.50}$$

where ξ is a fixed real number. This means that in our realization we have equally spaced additive outliers of size ξ, where there are r outlier–free observations in between. In total, the fraction of contamination in this setup is then $\frac{\lfloor \frac{n}{r+1} \rfloor}{n}$. If we compute $\widehat{\pi}^n$ with this sample, we need to evaluate (4.24) and (4.25). We then can write

$$(4.24) = \frac{1}{n-r} \left(\phi \left(\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}, \frac{Y((r+1)h) + \xi - \widehat{\pi}^n_r Y(rh) - \ldots - \widehat{\pi}^n_1 Y(h)}{\widehat{\sigma}^n} \right) \begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix} \right)$$

$$
\begin{aligned}
&+\phi\left(\begin{pmatrix} Y(2h) \\ \vdots \\ Y((r+1)h)+\xi \end{pmatrix}, \frac{Y((r+2)h) - \widehat{\pi}_r^n(Y((r+1)h)+\xi) - \ldots - \widehat{\pi}_1^n Y(h)}{\widehat{\sigma}^n}\right) \cdot \\
&\begin{pmatrix} Y(2h) \\ \vdots \\ Y((r+1)h)+\xi \end{pmatrix} + \ldots + \\
&\phi\left(\begin{pmatrix} Y((r+1)h)+\xi \\ \vdots \\ Y(2rh) \end{pmatrix}, \frac{Y((2r+1)h) - \widehat{\pi}_r^n Y(2rh) - \ldots - \widehat{\pi}_1^n(Y((r+1)h)+\xi)}{\widehat{\sigma}^n}\right) \cdot \\
&\begin{pmatrix} Y((r+1)h)+\xi \\ \vdots \\ Y(2rh) \end{pmatrix} + \sum_{t=r+2}^{n} \phi\left(\begin{pmatrix} \widetilde{Y}_t \\ \vdots \\ \widetilde{Y}_{t+r-1} \end{pmatrix}, \frac{\widetilde{Y}_{t+r} - \widehat{\pi}_r^n \widetilde{Y}_{t+r-1} - \ldots - \widehat{\pi}_1^n \widetilde{Y}_t}{\widehat{\sigma}^n}\right)\begin{pmatrix} \widetilde{Y}_t \\ \vdots \\ \widetilde{Y}_{t+r-1} \end{pmatrix}\Bigg) \\
&\stackrel{!}{=} 0 \qquad (4.51)
\end{aligned}
$$

and likewise for (4.25). The deciding observation is that the outlier at time $r+1$ influences all summands of the equation from time $t = 1$ until $t = r+1$. Additionally, the remaining summands from time $t = r+2$ to $t = n$ are affected in the same way since a new outlier appears exactly after $r+1$ points in time.
We now consider the special case of the Mallows estimator of Example 4.25a). Hence, $\phi(y,u) = w(y)\psi(u)$ for every $(y,u) \in \mathbb{R}^{r+1}$. Additionally, we assume that the function ψ is non–decreasing, as it is for example satisfied by the Huber ψ_k–functions of Example 4.25b). By Assumption H.2, there exists a constant $C > 0$ such that

$$\|\phi(y,u)y\| = w(y)\|y\||\psi(u)| \leq C \; \forall (y,u) \in \mathbb{R}^{r+1}.$$

By the assumed monotonicity of ψ, it must therefore hold that

$$\lim_{|u|\to\infty} \psi(u) = C'$$

for some constant $C' > 0$. Likewise, the behavior

$$\lim_{\|y\|\to\infty} w(y)\|y\| = \widetilde{C}$$

for some $\widetilde{C} \geq 0$ needs to hold.
If we now let $\xi \to \infty$ in (4.51), then the second argument of ϕ in the summands for $t = 1, \ldots, r+1$ tends to either $-\infty$ or ∞ while the norm of the first argument tends

to ∞ in the summands for $t = 2, \ldots, r+1$. More precisely, this means that for the summand corresponding to $t = 1$ it holds that

$$\begin{aligned}
&\lim_{\xi\to\infty} \phi\left(\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}, \frac{Y((r+1)h) + \xi - \widehat{\pi}_r^n Y(rh) - \ldots - \widehat{\pi}_1^n Y(h)}{\widehat{\sigma}^n}\right)\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix} \\
&= w\left(\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}\right)\left\|\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}\right\| \lim_{\xi\to\infty} \psi\left(\frac{Y((r+1)h) + \xi - \widehat{\pi}_r^n Y(rh) - \ldots - \widehat{\pi}_1^n Y(h)}{\widehat{\sigma}^n}\right) \\
&= C' w\left(\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}\right)\left\|\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}\right\| > 0,
\end{aligned}$$

which is a constant independent of π. Similarly, for the summand with $t = 2$ we have that

$$\begin{aligned}
&\lim_{\xi\to\infty} \phi\left(\begin{pmatrix} Y(2h) \\ \vdots \\ Y((r+1)h) + \xi \end{pmatrix}, \frac{Y((r+2)h) - \widehat{\pi}_r^n(Y((r+1)h) + \xi) - \ldots - \widehat{\pi}_1^n Y(h)}{\widehat{\sigma}^n}\right) \\
&\quad = \lim_{\xi\to\infty} w\left(\begin{pmatrix} Y(2h) \\ \vdots \\ Y((r+1)h) + \xi \end{pmatrix}\right)\left\|\begin{pmatrix} Y(2h) \\ \vdots \\ Y((r+1)h) + \xi \end{pmatrix}\right\| \\
&\quad \cdot \lim_{\xi\to\infty} \psi\left(\frac{Y((r+2)h) - \widehat{\pi}_r^n(Y((r+1)h) + \xi) - \ldots - \widehat{\pi}_1^n Y(h)}{\widehat{\sigma}^n}\right) \\
&\quad = \widetilde{C}C' \geq 0.
\end{aligned}$$

The same behavior holds for the summands with $t = 3, \ldots, r+1$. Summarizing and using that w is a strictly positive function, we have

$$\begin{aligned}
&\lim_{\xi\to\infty} \sum_{t=1}^{r+1} \phi\left(\begin{pmatrix} \widetilde{Y}_t \\ \vdots \\ \widetilde{Y}_{t+r-1} \end{pmatrix}, \frac{\widetilde{Y}_{t+r} - \widehat{\pi}_r^n \widetilde{Y}_{t+r-1} - \ldots - \widehat{\pi}_1^n \widetilde{Y}_t}{\widehat{\sigma}^n}\right)\begin{pmatrix} \widetilde{Y}_t \\ \vdots \\ \widetilde{Y}_{t+r-1} \end{pmatrix} \\
&= C' w\left(\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}\right)\left\|\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}\right\| + (r-1)\widetilde{C}C' > 0.
\end{aligned}$$

We can argue just in the same way for the sums with indices starting at $t = r + 2$. Hence, for $\xi \to \infty$, every summand of the left–hand side of (4.51) converges to a strictly positive constant, independent of π, and therefore (4.24) is not solvable. For (4.25), we can argue just in the same way and obtain that $\widehat{\pi}^n$ does not exist. Note that adding further outliers to components of (4.50) does not have an effect anymore, since all summands of the estimating equation are already afflicted, i.e. we have considered a worst case scenario. Since the fraction of contamination was $\frac{\lfloor \frac{n}{r+1} \rfloor}{n}$, the breakdown point must be less than this ratio. Letting $n \to \infty$ delivers a breakdown point bounded from above by $\frac{1}{r+1}$, just as claimed.

On the other hand, the perhaps more important fact is that the breakdown point of a GM estimator is positive. This is an important distinction from, e.g. the QMLE and least squares estimator, because it is well known that the latter two have a breakdown point of zero, i.e. one single outlier can cause the estimators to break down (see e.g. [Genton and Lucas 2003, Section 5.1]). To see that the breakdown point is positive, we can argue similarly as above: Assume that the contaminated, finite sample of fixed size $n \in \mathbb{N}$ has one single outlier at time $k \in \{1, \ldots, n\}$. This outlier can, at most, affect r summands of the defining equations (4.24) and (4.25) as we have seen above. Since the summands are all bounded by a constant C, therefore, at worst, this outlier causes r summands to be equal to a non–zero constant independent of π. If $n > r$ (which is reasonable to suppose, since the AR order r is typically rather small in comparison to the number of observations), then there are summands remaining which are not afflicted by outliers. At worst, the outlier therefore shifts the left hand side of (4.24) and (4.25) by $\pm\frac{rC}{n-1}$ and a sample of size $n - r$ remains, so that solutions continue to exist if they existed before.

For large r, the upper bound of $\frac{1}{r+1}$ is rather unsatisfactory, because the breakdown point will then be very low. For the indirect estimator, however, it may be necessary to choose r rather large, depending on the order of the CARMA(p, q) process. Therefore, if one suspects that the data contains a percentage of outliers high enough to surpass the breakdown point for an appropriate r, it might be advisable to not use a GM estimator at all.

Luckily, the construction of the indirect estimator allows for modifications which can deal with this problem. Remember that we use the auxiliary AR representation only because it is convenient to apply GM estimators to it and exploit their robustness properties. However, one could also use any other auxiliary model for which a robust estimator is available. One possibility would be to use the weak ARMA$(p, p-1)$ representation of the sampled CARMA process as described in (4.8) as the auxiliary model and use the bounded MM (BMM) estimators for ARMA processes

as introduced in Muler et al. [2009]. In that paper, the authors argue that their estimators form a class of estimators for ARMA models which are again consistent and asymptotically normal when no outliers are present and have good robustness properties. Specifically, in Muler et al. [2009, Section 6] the authors point out that the BMM estimators do not break down in the sense of Genton and Lucas [2003], i.e. they achieve the largest possible breakdown point of $\frac{1}{2}$. The downside of this approach is that the theory in Muler et al. [2009] is only developed for ARMA models driven by a strong white noise sequence, i.e. a white noise that consists of independent and identically distributed random variables. Since the noise in the ARMA$(p, p-1)$ representation coincides with the innovations calculated by the Kalman filter for the state–space representation of the sampled CARMA process (cf. (4.8)), we know that we are only dealing with an uncorrelated, but not independent noise sequence. Therefore, the results of Muler et al. [2009] are not directly applicable. However, if one had the analogs of [Muler et al. 2009, Theorem 4 and Theorem 6], which state that the BMM estimator is strongly consistent and asymptotically normal under suitable conditions for an ARMA process with a noise sequence that is i.i.d., one could replace the auxiliary AR(r) representation by the ARMA$(p, p-1)$ representation, $\widehat{\pi}^n$ by the BMM estimator and choose $\widehat{\pi}^n_{\mathrm{S}}$ as a strongly consistent, asymptotically normal estimator for the ARMA$(p, p-1)$ representation. In this setup the assumptions of Theorem 4.23 would be fulfilled again, so the results of Subsection 4.3.2 would still hold. It is worth mentioning that if the Lévy process driving the observed process $(Y(t))_{t\in\mathbb{R}}$ is a Brownian motion, then the sampled ARMA$(p, p-1)$ process will be driven by an i.i.d. noise, because the innovations are then normally distributed and thus independent. In this special case, the results on the BMM estimator hold and can be applied. Since the construction of the BMM estimator is very involved and lengthy, we do not go into details here and instead refer to Muler et al. [2009] for further information.

4.3.4.3. The influence functional

We continue our investigation of robustness of $\widehat{\vartheta}^n_{\mathrm{Ind}}$ with the study of the influence functional of the indirect estimator. This measure of robustness was originally introduced as the influence curve by Hampel [1974] for i.i.d. processes. Intuitively speaking, it measures the change in the asymptotic bias of an estimator caused by an infinitesimal amount of contamination in the data. It was later generalized to the time series context by Künsch [1984], who explicitly studies the estimation of autoregressive processes. However, in the paper of Künsch, only estimators which depend on a finite–dimensional marginal distribution of the data–generating process

and a very specific form of contamination are considered. To remedy this, a further generalization was then made by Martin and Yohai [1986], who consider the influence functional and explicitly allow for the estimators to depend on the measure for the process. Martin and Yohai also argue that their idea is a generalization of the definition of Künsch which fits better to the time series context, since their definition of contamination makes more sense in the time series setup ([Martin and Yohai 1986, Section 4]. We work with their definition in the following and will point out in what sense it differs from that of Künsch.
Consider now the model as in (4.2). The distribution of $(\widetilde{Y}_n)_{n\in\mathbb{Z}}$ is characterized by the joint distribution of $(Y(nh))_{n\in\mathbb{Z}}$, $(V_n)_{n\in\mathbb{Z}}$ and $(Z_n)_{n\in\mathbb{Z}}$. We denote the probability measure associated to the distribution $(Z_n)_{n\in\mathbb{Z}}$ on $\mathbb{R}^\infty$ by $\mathbb{P}_Z$ and the probability measure associated to the distribution of $(\widetilde{Y}_n)_{n\in\mathbb{Z}}$ by $\mathbb{P}_Y^\gamma$ for $0 \leq \gamma \leq 1$. Note that $\gamma = 0$ corresponds to the case where there are no outliers, i. e. we can observe the nominal process without error and then write $\mathbb{P}_Y^0 = \mathbb{P}_{Y^{(h)}}$, which is the probability measure associated to the distribution of the data–generating CARMA process $(Y(nh))_{n\in Z}$. We denote

$$\{\mathbb{P}_Y^\gamma\} := \{\mathbb{P}_Y^\gamma, 0 \leq \gamma \leq 1\} \subseteq \mathcal{P}(\mathbb{R}^\infty).$$

We assume that the assumptions of Theorem 4.27 and Theorem 4.31 are satisfied for every choice of γ. Remember that the pseudo–true parameter in this case depends on the processes V and Z, i.e. it is different for different values of γ and we denote it by π_0^γ. We assume that $\pi_0^0 = \pi_{\vartheta_0}$, i.e. in the case of no outliers the auxiliary parameter of the data–generating CARMA process is estimated. Sufficient for this are for example the conditions of Proposition 4.28. We introduce the statistical functional T_{GM} by defining

$$\begin{aligned} T_{\mathrm{GM}} : \{\mathbb{P}_Y^\gamma\} &\to \Pi \\ \mathbb{P}_Y^\gamma &\mapsto \pi_0^\gamma. \end{aligned}$$

Then, the definition of the influence functional for the GM estimators as considered in Subsubsection 4.3.3.1 is as follows:

$$IF_{\mathrm{GM}}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}) := \lim_{\gamma\to 0} \frac{T_{\mathrm{GM}}(\mathbb{P}_Y^\gamma) - T_{\mathrm{GM}}(\mathbb{P}_{Y^{(h)}})}{\gamma} = \lim_{\gamma\to 0} \frac{\pi_0^\gamma - \pi_0}{\gamma} \tag{4.52}$$

whenever this limit is well–defined. Note that the influence functional depends on the whole "arc" of contaminated measures $\{\mathbb{P}_Y^\gamma, 0 \leq \gamma \leq 1\}$. This is the most important difference to the definition used by Künsch [1984], because in that paper

the approximation $\mathbb{P}_Y^\gamma = (1-\gamma)\mathbb{P}_{Y^{(h)}} + \gamma\nu$ for some fixed $\nu \in \mathcal{P}(\mathbb{R}^\infty)$ is used ([Künsch 1984, Eq. (1.11)]). The influence functional measures the effect of an infinitesimal contamination of the true process by the process Z on the asymptotic estimate defined via the functional T_{GM}.
In a similar vein, we can define the influence functional for the estimation of the parameter ϑ_0 of our CARMA process, which is what we are truly interested in. Analogous to T_{GM}, we first define a suitable statistical functional T_{Ind} via

$$\begin{aligned} T_{\mathrm{Ind}} : \{\mathbb{P}_Y^\gamma\} &\to \Theta \\ \mathbb{P}_Y^\gamma &\mapsto \vartheta_0^\gamma := \arg\min_{\vartheta\in\Theta} (\pi_0^\gamma - \pi_\vartheta)^T \Omega (\pi_0^\gamma - \pi_\vartheta). \end{aligned} \tag{4.53}$$

This definition is motivated by the fact that $\widehat{\pi}^n$ converges almost surely to π_0^γ in this scenario by Theorem 4.27. With the same arguments as in the uncontaminated case (cf. the proof of Theorem 4.23), we then obtain that $\mathcal{L}_{\mathrm{Ind}}(\vartheta, \widetilde{Y}^n)$ converges uniformly in ϑ almost surely to $(\pi_0^\gamma - \pi_\vartheta)^T\Omega(\pi_0^\gamma - \pi_\vartheta)$, the analog of (4.18). Continuing to argue as in the proof of Theorem 4.23, $\widehat{\vartheta}_{\mathrm{Ind}}^n$ then converges uniformly almost surely to ϑ_0^γ in this case and the definition of T_{Ind} is justified. With this the definition of the influence functional of the indirect estimator is

$$IF_{\mathrm{Ind}}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}) = \lim_{\gamma\to 0} \frac{T_{\mathrm{Ind}}(\mathbb{P}_Y^\gamma) - T_{\mathrm{Ind}}(\mathbb{P}_{Y^{(h)}})}{\gamma} = \lim_{\gamma\to 0} \frac{\vartheta_0^\gamma - \vartheta_0}{\gamma}$$

Of course we are interested in properties of this functional, mainly if and under which conditions it remains bounded. Boundedness of the influence functional implies that the estimate arising from the contaminated process cannot move too far away from the one in the uncontaminated case if the rate of contamination is infinitesimal. This property is well–known for the influence functional for GM estimators of AR processes. Since these estimators are an integral building block of the indirect estimator, one can hope that it carries over to this scenario and indeed it does, since the two functionals are proportional. This follows from de Luna and Genton [2000, Theorem 1] as special case, but we cite the theorem in full and also give its proof in detail here for the sake of completeness:

Theorem 4.46. *Under the assumptions made at the beginning of Subsection 4.3.4, IF_{Ind} exists whenever IF_{GM} exists and*

$$IF_{Ind}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}) = \mathcal{H}(T_{Ind}(\mathbb{P}_{Y^{(h)}}))IF_{GM}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}),$$

where

$$\mathcal{H}(T_{Ind}(\mathbb{P}_{Y^{(h)}})) = \mathcal{H}(\vartheta_0) = (\nabla_\vartheta \pi_{\vartheta_0}^T \Omega \nabla_\vartheta \pi_{\vartheta_0})^{-1} \nabla_\vartheta \pi_{\vartheta_0}^T \Omega$$

as in (4.21).

Proof. By (4.53), it holds that

$$\begin{aligned} &\nabla_\vartheta \pi_{\vartheta_0^\gamma} \Omega(\pi_0^\gamma - \pi_{\vartheta_0^\gamma}) = 0 \\ \Longrightarrow &\nabla_\vartheta \pi_{\vartheta_0^\gamma} \Omega(\pi_0^\gamma - \pi_0 + \pi_0 - \pi_{\vartheta_0^\gamma}) = 0 \\ \Longrightarrow &\nabla_\vartheta \pi_{\vartheta_0^\gamma} \Omega \left(\frac{\pi_0^\gamma - \pi_0}{\gamma} \right) = \nabla_\vartheta \pi_{\vartheta_0^\gamma} \Omega \left(\frac{\pi_{\vartheta_0^\gamma} - \pi_0}{\gamma} \right). \end{aligned} \tag{4.54}$$

By Definition 4.21, the map $\vartheta \mapsto \nabla_\vartheta \pi_\vartheta$ is continuous. Moreover, $\vartheta_0^\gamma \to \vartheta_0$ as $\gamma \to 0$ holds. Therefore, by continuity, $\nabla_\vartheta \pi_{\vartheta_0^\gamma} \to \nabla_\vartheta \pi_{\vartheta_0}$ as $\gamma \to 0$ holds. Furthermore,

$$\lim_{\gamma \to 0} \frac{\pi_{\vartheta_0^\gamma} - \pi_0}{\gamma} = \nabla_\vartheta \pi_{\vartheta_0} \lim_{\gamma \to 0} \frac{\vartheta_0^\gamma - \vartheta_0}{\gamma} = \nabla_\vartheta \pi_{\vartheta_0} IF_{\text{Ind}}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}).$$

Taking the limit $\gamma \to 0$ on both sides of (4.54) thus implies

$$\begin{aligned} &\nabla_\vartheta \pi_{\vartheta_0} \Omega IF_{\text{GM}}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}) = \nabla_\vartheta \pi_{\vartheta_0} \Omega \nabla_\vartheta \pi_{\vartheta_0} IF_{\text{Ind}}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}) \\ \Longrightarrow &IF_{\text{Ind}}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}) = (\nabla_\vartheta \pi_{\vartheta_0} \Omega \nabla_\vartheta \pi_{\vartheta_0})^{-1} \nabla_\vartheta \pi_{\vartheta_0} \Omega IF_{\text{GM}}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}) \\ &= \mathcal{H}(T_{\text{Ind}}(\mathbb{P}_{Y^{(h)}})) IF_{\text{GM}}(\mathbb{P}_Z, \{\mathbb{P}_Y^\gamma\}) \end{aligned}$$

as claimed. □

From this theorem we see that the question of boundedness of the influence functional for the indirect estimator of a CARMA process reduces to the question of boundedness of the influence functional for the GM estimation of the auxiliary AR process. Conditions under which the influence functional is bounded are studied in Martin and Yohai [1986], where it is shown that GM estimators yield a bounded influence functional for AR(1) processes. Those results generalize easily to the setting of GM estimators for AR(r) processes with $r \geq 2$ applied to sampled CARMA processes as we considered them in Subsubsection 4.3.3.1 as we shall see in the following.

For our investigation of boundedness of the influence functional, assume that our contamination model (4.2) is such that we have additive outliers, i.e. we are in the case of Remark 4.2b) and it holds that $\widetilde{Y}_n = Y(nh) + V_n W_n$, where the processes are chosen in such a way that Assumption G is satisfied. Under these assumptions, we can now state the result on the influence functional defined in (4.52):

Theorem 4.47. *Let the additive outlier model hold. Assume furthermore that the matrix $\mathcal{J}_{GM}(\pi_{\vartheta_0})$ as defined in Theorem 4.31 is non–singular. Then there exists a constant $K > 0$ such that it holds:*

$$|IF_{GM}(\mathbb{P}_W, \{\mathbb{P}_Y^\gamma\})| \leq 2(r+1)C \left\|(\mathcal{J}_{GM}(\pi_{\vartheta_0}))^{-1}\right\|.$$

Proof. The plan is to apply [Martin and Yohai 1986, Theorem 4.3]. To this end, note that $T_{\mathrm{GM}}(\mathbb{P}_Y^\gamma)$ (i. e. the parameter π_0^γ defined by (4.22) and (4.23)) only depends on $r+1$ values of the process $(\widetilde{Y}_n)_{n\in Z}$. Moreover, we have that each of those summands is bounded by (say) a constant C by Assumptions H.2 and H.6 and that

$$\mathbb{E}\left[\phi\left(\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}, \frac{Y((r+1)h) - \pi_{\vartheta_0,r}Y(rh) - \ldots - \pi_{\vartheta_0,1}Y(h)}{\sigma_{\vartheta_0}}\right)\begin{pmatrix} Y(h) \\ \vdots \\ Y(rh) \end{pmatrix}\right] = 0,$$

$$\mathbb{E}\left[\chi\left(\left(\frac{Y((r+1)h) - \pi_{\vartheta_0,r}Y(rh) - \ldots - \pi_{\vartheta_0,1}Y(h)}{\sigma_{\vartheta_0}}\right)^2\right)\right] = 0.$$

by (4.22) and (4.23). Note that we have written Y instead of $\widetilde{Y}$ here because $\gamma = 0$ corresponds to the uncontaminated case where there are no outliers present, which implies that $\widetilde{Y}$ then coincides with Y and $\pi_0^\gamma = \pi_0^0 = \pi_{\vartheta_0}$. Since we also assumed that $\mathcal{J}_{\mathrm{GM}}(\pi_{\vartheta_0})$, the derivative of $\mathcal{Q}_{\mathrm{GM}}(\pi)$ at π_{ϑ_0}, exists and is non–singular, Assumptions (a), (b) and (c) of [Martin and Yohai 1986, Theorem 4.3] are satisfied (see also [Martin and Yohai 1986, Comment 4.3]).

However, this is not sufficient yet, as to apply [Martin 1980, Theorem 4.3] it must also be checked that T_{GM} satisfies [Martin and Yohai 1986, Eq. (4.6)]. Sufficient conditions for this equation to hold are given in [Martin and Yohai 1986, Theorem 4.2], which we will now verify. Remember that by Assumption G.3 the process $(V_n)_{n\in\mathbb{Z}}$ is independent from the processes $(W_n)_{n\in\mathbb{Z}}$ and $(Y(nh))_{n\in\mathbb{Z}}$, such that the assumption on the distribution of $(\widetilde{Y}_n)_{n\in\mathbb{Z}}$ of [Martin and Yohai 1986, Theorem 4.2] is satisfied.

As in the proof of [Martin and Yohai 1986, Theorem 5.2], we obtain that Assumption (a) of [Martin and Yohai 1986, Theorem 4.2] is satisfied. Assumption (b) of [Martin and Yohai 1986, Theorem 4.2] is an assumption of Theorem 4.47 as well. For assumption (c), the object $H_m(\pi)$ that must be studied is given in our context by

$$H_m(\pi) = \sup_{a_i\in\{0,1\},\, b_i\in\{0,1\},\, i=1,\dots,r+1-m}$$
$$\left|\mathbb{E}\left[\left(\phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_m\\ Y_{m+1}+a_{r-m+1}Z_{m+1}\\ \vdots\\ Y_r+a_2Z_r\end{pmatrix}, \frac{Y_{r+1}+a_1Z_{r+1}-\pi_r(Y_r+a_2Z_r)-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_m\\ Y_{m+1}+a_{r-m+1}Z_{m+1}\\ \vdots\\ Y_r+a_2Z_r\end{pmatrix}\right)\chi\left(\left(\frac{Y_{r+1}+a_1Z_{r+1}-\pi_rY_r+a_2Z_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)^2\right)\right.$$
$$\left.-\left(\phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_m\\ Y_{m+1}+b_{r-m+1}Z_{m+1}\\ \vdots\\ Y_r+b_2Z_r\end{pmatrix}, \frac{Y_{r+1}+b_1Z_{r+1}-\pi_r(Y_r+b_2Z_r)-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_m\\ Y_{m+1}+b_{m+1}Z_{m+1}\\ \vdots\\ Y_r+b_rZ_r\end{pmatrix}\right)\chi\left(\left(\frac{Y_{r+1}+b_1Z_{r+1}-\pi_r(Y_r+b_2Z_r)-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)^2\right)\right]\right|$$

for $0 \le m \le r+1$ and by $H_m(\pi) = 0$ for $m > r+1$, where we used the notation $Y_t = Y(th)$ for $t = 1, \dots, r+1$ for the sake of readability. Since every component of $H_m(\pi)$ is bounded for every $1 \le m \le r+1$ and every π, it follows that

$$\sum_{m=1}^{\infty} \sup_{\pi\in\Pi'} H_m(\pi) = \sum_{m=1}^{r+1} \sup_{\pi\in\Pi'} H_m(\pi) \le (r+1)C < \infty,$$

and therefore Assumption (c) of [Martin and Yohai 1986, Theorem 4.2] holds. For the same reason it follows from the boundedness assumptions that

$$\sup_{\pi\in\Pi'} \sup_{(\widetilde{Y}_1,\dots,\widetilde{Y}_{r+1})^T} \mathbb{E}\left[\left(\phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}\right)\chi\left(\left(\frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)^2\right)\right] < \infty,$$

which is Assumption (d). Last but not least, it holds that

$$\lim_{\pi\to\pi_{\vartheta_0}} \begin{pmatrix} \phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix} \\ \chi\left(\left(\frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)^2\right)\end{pmatrix}$$
$$= \begin{pmatrix} \phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1}-\pi_{\vartheta_0,r}\widetilde{Y}_r-\ldots-\pi_{\vartheta_0,1}\widetilde{Y}_1}{\sigma_{\vartheta_0}}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix} \\ \chi\left(\left(\frac{\widetilde{Y}_{r+1}-\pi_{\vartheta_0,r}\widetilde{Y}_r-\ldots-\pi_{\vartheta_0,1}\widetilde{Y}_1}{\sigma_{\vartheta_0}}\right)^2\right)\end{pmatrix} \quad \mathbb{P}\text{-a.s.}$$

by the continuity of the functions ϕ and χ. And, again by the boundedness we assumed in H.2 and H.6, we obtain that

$$\mathbb{E}\left[\sup_{\pi\in\Pi'}\left\|\begin{pmatrix} \phi\left(\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix}, \frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)\begin{pmatrix}\widetilde{Y}_1\\ \vdots\\ \widetilde{Y}_r\end{pmatrix} \\ \chi\left(\left(\frac{\widetilde{Y}_{r+1}-\pi_r\widetilde{Y}_r-\ldots-\pi_1\widetilde{Y}_1}{\sigma}\right)^2\right)\end{pmatrix}\right\|\right] < \infty,$$

which is Assumption (e).
This completes the list of required assumptions for [Martin and Yohai 1986, Eq. (4.6)] to hold which is the last ingredient in the proof of the boundedness of the influence functional in this case. □

Remark 4.48. *Note that [Martin and Yohai 1986, Theorem 5.2] proceeds in the same way to show that the influence functional of an AR(1) process with additive outliers is bounded. In contrast to our results however, they do not consider the case of higher order AR processes and also assume that the true, outlier–free process possesses Gaussian innovations, which we do not require.*

4.4. Model selection using the indirect estimator

An important question that we have not addressed so far is the issue of model selection in the robustness framework. A naive idea would be to apply the criteria developed in Chapter 3 to determine the orders of the CARMA process before using a robust estimator for parameter estimation. However, since those information criteria are based on the QMLE and we know that this estimator is not robust towards outliers, it is reasonable to expect that the outliers also greatly affect the information criteria. Instead, it would better to use some kind of robust model selection procedure which takes the outliers into account. In the literature, there exist a few approaches towards this problem. For example, [Martin 1980, Section 6] proposes a modification to the AIC in the framework of determining the order r of an autoregressive process based on GM estimators. [Ronchetti 1997, Section 3.2] also treats this topic and introduces a robust version of the AIC, referencing the results of Behrens [1991]. Similarly, Machado [1993] treats the BIC and proposes a robustified version of it. The problem with these cited references is, however, that they all assume that the parameter of interest is estimated by an M or GM estimator and modify the criteria accordingly. For our problem, this means that we could apply these robust order selection procedures to estimate the order of the auxiliary AR process introduced in Subsection 4.3.1, since the parameters of this process are estimated by a GM estimator as part of the indirect approach. However, this is of limited use, since the auxiliary AR representation is just an appliance in constructing the indirect estimator. In Subsection 4.3.2 we only required that $r \geq 2p-1$ and did not observe in any kind that the indirect estimator works better or worse if r is chosen in a particular way above this threshold. Hence, choosing r optimally does not directly aid us in finding a suitable parameter space for ϑ_0. $\widehat{\vartheta}^n_{\mathrm{Ind}}$, on the other hand, is not an M or GM estimator, i.e. those criteria cannot be applied and we have to look for an alternative method. However, the guiding idea used by the cited references is still useful: they replace the likelihood function by the robust function that is optimized to obtain the respective estimator. We will proceed in the same way.
Similar to Subsection 4.3.2, we consider the case when there are no outliers in the data, i.e. when $\widetilde{Y}^n = Y^n$. Remembering how our information criteria were defined, it suggests itself to define a criterion based on the indirect estimator as

$$\mathrm{IC}^{\mathrm{Ind}}_n(\Theta) := \mathcal{L}_{\mathrm{Ind}}(\widehat{\vartheta}^n_{\mathrm{Ind}}, Y^n) + N(\Theta)\frac{C(n)}{n},$$

where

$$\mathcal{L}_{\mathrm{Ind}}(\vartheta, Y^n) = (\widehat{\pi}^n - \widehat{\pi}^n_{\mathrm{S}}(\vartheta))^T \Omega (\widehat{\pi}^n - \widehat{\pi}^n_{\mathrm{S}}(\vartheta))$$

as in (4.10) can be interpreted as "pseudo–likelihood" function and $C(n)$ is again a penalty function as in Definition 3.6. Given a family of parameter spaces parametrizing CARMA(p,q) processes of different orders, we will again choose the parameter space Θ (or the model represented by this parameter space) as the most suitable for the data for which $\mathrm{IC}_n^{\mathrm{Ind}}(\Theta)$ is minimal. Since we are in the one–dimensional case, we know from Example 3.1 that a parameter space can be uniquely characterized by the order p of the AR polynomials of the CARMA processes it contains, which needs to be fixed by B.6, and the maximal order q of the MA polynomials of the processes. If the data–generating CARMA process $(Y(t))_{t\in\mathbb{R}}$ is a CARMA(p_0,q_0) process, the true parameter space Θ_0 is then the one for which $p=p_0$ and $q=q_0$. Ultimately, just as in Chapter 3, we would like to draw conclusions about consistency of these information criteria in the sense of Definition 3.9. To this end, we have to study the behavior of $\widehat{\vartheta}^n_{\mathrm{Ind}}$ if we have a parameter space Θ such that $Y \neq \mathrm{CARMA}(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)$ for every $\vartheta\in\Theta$, i.e. if we are in a misspecified parameter space. In the one–dimensional case, a space is misspecified precisely if and only if either $p\neq p_0$ or $p=p_0$ and $q<q_0$. Similarly, the true parameter space Θ_0 is nested in Θ if and only if $p=p_0$ and $q\geq q_0$. Obviously, if $\vartheta_0\in\Theta_0$, then for any Θ in which Θ_0 is nested the results on the indirect estimator then also hold in Θ, because there exists $\vartheta^*\in\Theta$ with $\mathrm{CARMA}(A_{\vartheta^*},B_{\vartheta^*},C_{\vartheta^*},L_{\vartheta^*}) = \mathrm{CARMA}(A_{\vartheta_0},B_{\vartheta_0},C_{\vartheta_0},L_{\vartheta_0})$. For non–nested spaces, summarizing Subsection 2.2.3, all the results from the correctly specified case carry over in the case of the QMLE as long as there is a unique pseudo–true parameter $\vartheta^*\in\Theta$, which minimizes the almost sure limit of $\widehat{\mathcal{L}}(\vartheta, Y^n)$, i.e. $\mathscr{Q}$ in the QMLE case. We will see that this is also the case for the indirect estimator. We always assume in the following that the auxiliary parameter space Π and the weighting matrix Ω are the same for every candidate space Θ. This is not a restrictive assumption, it just implies that $r\geq 2p-1$ needs to hold for the maximum p under consideration, i.e. for the largest parameter space.

Proposition 4.49. *Assume that the parameter space Θ is such that $Y\neq CARMA(A_\vartheta,B_\vartheta,C_\vartheta,L_\vartheta)$ for every $\vartheta\in\Theta$ and let the assumptions of Section 4.3 be satisfied. Then, there exists a unique $\vartheta^*\in\Theta$ such that*

$$\vartheta^* = \operatorname*{arg\,min}_{\vartheta\in\Theta} \mathscr{Q}_{Ind}(\vartheta) = \operatorname*{arg\,min}_{\vartheta\in\Theta} (\pi_{\vartheta_0}-\pi_\vartheta)^T\Omega(\pi_{\vartheta_0}-\pi_\vartheta).$$

Moreover, it holds that

$$\mathscr{Q}_{Ind}(\vartheta^*) = (\pi_{\vartheta_0}-\pi_{\vartheta^*})^T\Omega(\pi_{\vartheta_0}-\pi_{\vartheta^*}) > 0$$

and Theorem 4.23 holds with ϑ^* replacing ϑ_0.

Proof. The function $\vartheta \mapsto \mathscr{Q}_{\mathrm{Ind}}(\vartheta)$ is continuous by construction of π. Since Θ is compact, it therefore attains its minimum at some $\vartheta^* \in \Theta$. Moreover, π is injective by assumption, which implies that ϑ^* is unique. Since the parameter space is misspecified by construction, this minimum also has to be strictly positive because $\pi_{\vartheta^*} \neq \pi_{\vartheta_0}$ holds. Theorem 4.23 can then be proven just in the same way as before, replacing ϑ_0 by ϑ^* wherever it appears and noting that the only property of ϑ_0 we used was that it is the unique minimizer of $\mathscr{Q}_{\mathrm{Ind}}(\vartheta)$. □

in the following, we always assume that the pseudo–true parameter ϑ^* is an element of the interior of the corresponding space Θ, analogous to Assumption B.8. Using the previous proposition and this additional assumption, we can draw a conclusion about the consistency of our new information criterion $\mathrm{IC}_n^{\mathrm{Ind}}$:

Theorem 4.50.

a) Assume that the penalty term $C(n)$ fulfills $C(n) \to \infty$ as $n \to \infty$. Then, IC_n^{Ind} is a weakly consistent information criterion in the sense of Definition 3.9b).

b) If $\limsup_{n\to\infty} C(n) < \infty$, then IC_n^{Ind} is neither weakly nor strongly consistent in the sense of Definition 3.9.

c) Consider a parameter space Θ with $p = p_0$ but $q > q_0$, i.e. Θ_0 is nested in Θ with map F in the sense of Definition 3.8. For the pseudo–true parameter $\vartheta^* \in \Theta$, define

$$\mathcal{M}_{F,Ind}(\vartheta^*) := -(\mathcal{J}_{Ind}(\vartheta^*))^{-1} + F(\mathcal{J}_{Ind}(\vartheta_0))^{-1}F^T.$$

If $\limsup_{n\to\infty} C(n) = C < \infty$, we have

$$\lim_{n\to\infty} \mathbb{P}(IC_n^{Ind}(\Theta_0) - IC_n^{Ind}(\Theta) > 0) = \mathbb{P}\left(\sum_{i=1}^{q-q_0} \lambda_i \chi_i^2 > 2[N(\Theta) - N(\Theta_0)]C\right) > 0,$$

where (χ_i^2) is a sequence of independent χ^2 random variables with one degree of freedom and the λ_i are the $q - q_0$ strictly positive eigenvalues of

$$\mathcal{J}_{Ind}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_{F,Ind}(\vartheta^*)\mathcal{I}_{Ind}(\vartheta^*)\mathcal{M}_{F,Ind}(\vartheta^*)\mathcal{J}_{Ind}(\vartheta^*)^{\frac{1}{2}}.$$

Proof. The proof is analogous to the one of Theorem 3.10b), the only difference being that we need to use different asymptotic results. Throughout the proof, we consider

the true parameter space Θ_0 and an alternative space Θ fulfilling Assumption B. We denote the "likelihood" associated to Θ_0 by $\mathcal{L}_{\text{Ind},0}$, the indirect estimator in Θ_0 by $\widehat{\vartheta}^n_{\text{Ind},0}$ and the limiting function of $\mathcal{L}_{\text{Ind},0}$ by $\mathscr{Q}_{\text{Ind},0}$. For the space Θ we use the usual notation without any additional subscripts.

a) We have to distinguish two cases.
Case 1: Consider a parameter space Θ such that Θ_0 is not nested in Θ. Then:

$$\text{IC}^{\text{Ind}}_n(\Theta_0) - \text{IC}^{\text{Ind}}_n(\Theta) = \mathcal{L}_{\text{Ind},0}(\widehat{\vartheta}^n_{\text{Ind},0}, Y^n) - \mathcal{L}_{\text{Ind}}(\widehat{\vartheta}^n_{\text{Ind}}, Y^n) + (N(\Theta_0) - N(\Theta))\frac{C(n)}{n}.$$

We now show that $\mathcal{L}_{\text{Ind},0}(\widehat{\vartheta}^n_{\text{Ind},0}, Y^n) - \mathscr{Q}_{\text{Ind},0}(\vartheta_0) \to 0$ almost surely for $n \to \infty$ (by the same argument as [Schlemm and Stelzer 2012, p. 2201]). For $\epsilon > 0$ and $\omega \in \Omega \setminus N$, where N is a set with $\mathbb{P}(N) = 0$, suppose that $n \in \mathbb{N}$ is such that

$$\sup_{\vartheta \in \Theta_0} |\mathcal{L}_{\text{Ind},0}(\vartheta, Y^n) - \mathscr{Q}_{\text{Ind},0}(\vartheta)| < \epsilon.$$

Then

$$\mathcal{L}_{\text{Ind},0}(\widehat{\vartheta}^n_{\text{Ind},0}, Y^n) \leq \mathcal{L}_{\text{Ind},0}(\vartheta_0, Y^n) \leq \mathscr{Q}_{\text{Ind},0}(\vartheta_0) + \epsilon$$

and

$$\mathcal{L}_{\text{Ind},0}(\widehat{\vartheta}^n_{\text{Ind},0}, Y^n) \geq \mathscr{Q}_{\text{Ind},0}(\widehat{\vartheta}^n_{\text{Ind},0}) - \epsilon \geq \mathscr{Q}_{\text{Ind},0}(\vartheta_0) - \epsilon,$$

which implies $|\mathcal{L}_{\text{Ind},0}(\widehat{\vartheta}^n_{\text{Ind},0}, Y^n) - \mathscr{Q}_{\text{Ind},0}(\vartheta_0)| < \epsilon$. Therefore,

$$\mathbb{P}(\mathcal{L}_{\text{Ind},0}(\widehat{\vartheta}^n_{\text{Ind},0}, Y^n) \overset{n\to\infty}{\longrightarrow} \mathscr{Q}_{\text{Ind},0}(\vartheta_0)) \geq \mathbb{P}\left(\sup_{\vartheta \in \Theta} |\mathcal{L}_{\text{Ind},0}(\vartheta, Y^n) - \mathscr{Q}_{\text{Ind},0}(\vartheta)| \overset{n\to\infty}{\longrightarrow} 0\right) = 1,$$

where the last equality holds by the uniform almost sure convergence of $\mathcal{L}_{\text{Ind},0}$ to $\mathscr{Q}_{\text{Ind},0}$, which we obtained in the proof of Theorem 4.23. By the same argument we also obtain that $\mathcal{L}_{\text{Ind}}(\widehat{\vartheta}^n_{\text{Ind}}, Y^n) - \mathscr{Q}_{\text{Ind}}(\vartheta^*) \to 0$. Therefore,

$$\mathcal{L}_{\text{Ind},0}(\widehat{\vartheta}^n_{\text{Ind},0}, Y^n) - \mathcal{L}_{\text{Ind}}(\widehat{\vartheta}^n_{\text{Ind}}, Y^n) \to \mathscr{Q}_{\text{Ind},0}(\vartheta_0) - \mathscr{Q}_{\text{Ind}}(\vartheta^*) > 0 \quad \mathbb{P}\text{-a.s.}, n \to \infty,$$

since Θ is misspecified. By assumption, we have that $C(n)/n \to 0$ as $n \to \infty$, and hence

$$\mathbb{P}(\text{IC}^{\text{Ind}}_n(\Theta_0) - \text{IC}^{\text{Ind}}_n(\Theta) \geq 0) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(\mathscr{Q}_{\text{Ind},0}(\vartheta_0) - \mathscr{Q}_{\text{Ind}}(\vartheta^*) \geq 0) = 0,$$

showing that the probability of selecting Θ instead of Θ_0 goes to zero.

Case 2: Consider a parameter space Θ in which Θ_0 is nested, i.e. we have $p = p_0$ but $q > q_0$ for Θ. In this case we argue as in the proof of Theorem 3.10b), defining $f : \Theta_0 \to \Theta$ by $f(\vartheta) = F\vartheta + c$, where F and c are as in the definition of nested spaces in Definition 3.8 and writing

$$\begin{aligned}\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) = \frac{1}{2}\left(\widehat{\vartheta}^n_{\mathrm{Ind}} - f(\widehat{\vartheta}^n_{\mathrm{Ind},0})\right)^T \nabla^2_\vartheta \mathcal{L}_{\mathrm{Ind}}\left(\overline{\vartheta}^n, Y^n\right)\left(\widehat{\vartheta}^n_{\mathrm{Ind}} - f(\widehat{\vartheta}^n_{\mathrm{Ind},0})\right) \\ + [N(\Theta_0) - N(\Theta)]\frac{C(n)}{n}\end{aligned}$$

with $\overline{\vartheta}^n$ such that $\|\overline{\vartheta}^n - \widehat{\vartheta}^n_{\mathrm{Ind}}\| \leq \|f(\widehat{\vartheta}^n_{\mathrm{Ind},0}) - \widehat{\vartheta}^n_{\mathrm{Ind}}\|$. The rest of the proof now uses the analogous arguments as the one of Theorem 3.10b) and the asymptotic results derived in the proof of Theorem 4.23. In this way, we first arrive at the analog of (3.12):

$$\begin{aligned}\widehat{\vartheta}^n_{\mathrm{Ind}} - f(\widehat{\vartheta}^n_{\mathrm{Ind},0}) \xrightarrow{\mathcal{D}} \left[-((\nabla_\vartheta \pi_{\vartheta^*})^T \Omega \nabla_\vartheta \pi_{\vartheta^*})^{-1} + F((\nabla_\vartheta \pi_{\vartheta_0})^T \Omega \nabla_\vartheta \pi_{\vartheta_0})^{-1} F^T\right] \\ \cdot \mathcal{N}\left(0, (\nabla_\vartheta \pi_{\vartheta^*})^T \Omega \left(\Xi_{\mathrm{D}}(\vartheta^*) + \frac{1}{s}\Xi_{\mathrm{S}}(\vartheta^*)\right)\Omega \nabla_\vartheta \pi_{\vartheta^*}\right) \\ =: \mathbf{N}_{F,\mathrm{Ind}}\end{aligned}$$

Using this and $C(n) \to \infty$ for $n \to \infty$, we then obtain the analog of (3.13):

$$\begin{aligned}&\mathbb{P}(\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) < 0) \\ &= \mathbb{P}\left(\frac{1}{2}\sqrt{n}\left(\widehat{\vartheta}^n_{\mathrm{Ind}} - f(\widehat{\vartheta}^n_{\mathrm{Ind},0})\right)^T \nabla^2_\vartheta \mathcal{L}_{\mathrm{Ind}}\left(\overline{\vartheta}^n, Y^n\right)\sqrt{n}\left(\widehat{\vartheta}^n_{\mathrm{Ind}} - f(\widehat{\vartheta}^n_{\mathrm{Ind},0})\right)\right. \\ &\quad \left. < -[N(\Theta_0) - N(\Theta)]C(n)\right) \xrightarrow{n\to\infty} \mathbb{P}\left(\mathbf{N}^T_{F,\mathrm{Ind}} \mathcal{J}_{\mathrm{Ind}}(\vartheta^*)\mathbf{N}_{F,\mathrm{Ind}} < \infty\right). \quad (4.55)\end{aligned}$$

As in the proof of Theorem 3.10b), Imhof [1961, Eq. (1.1)] gives

$$\mathbf{N}^T_{F,\mathrm{Ind}} \mathcal{J}_{\mathrm{Ind}}(\vartheta^*)\mathbf{N}_{F,\mathrm{Ind}} \overset{\mathcal{D}}{=} \sum_{i=1}^{N(\Theta)} \lambda_i \chi^2_i,$$

where the λ_i are the eigenvalues of $\mathcal{J}_{\mathrm{Ind}}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_{F,\mathrm{Ind}}(\vartheta^*)\mathcal{I}_{\mathrm{Ind}}(\vartheta^*)\mathcal{M}_{F,\mathrm{Ind}}(\vartheta^*)\mathcal{J}_{\mathrm{Ind}}(\vartheta^*)^{\frac{1}{2}}$ and (χ^2_i) is a sequence of independent χ^2 random variables with one degree of freedom. Since $\mathrm{rank}(\mathcal{M}_{F,\mathrm{Ind}}(\vartheta^*)) = N(\Theta) - N(\Theta_0) = q - q_0$ and $\mathcal{J}_{\mathrm{Ind}}(\vartheta^*)^{\frac{1}{2}}$ and $\mathcal{I}_{\mathrm{Ind}}(\vartheta^*)$ have full rank, the number of strictly positive eigenvalues of

$$\mathcal{J}_{\mathrm{Ind}}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_{F,\mathrm{Ind}}(\vartheta^*)\mathcal{I}_{\mathrm{Ind}}(\vartheta^*)\mathcal{M}_{F,\mathrm{Ind}}(\vartheta^*)\mathcal{J}_{\mathrm{Ind}}(\vartheta^*)^{\frac{1}{2}}$$

is $q - q_0$. Hence, the result follows.

b) & c) The results follow from the arguments given in a), in particular from (4.55), which in this case has the form

$$\mathbb{P}(\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) < 0) \overset{n\to\infty}{\longrightarrow} \mathbb{P}\left(\mathbf{N}_{F,\mathrm{Ind}}^T \mathcal{J}_{\mathrm{Ind}}(\vartheta^*)\mathbf{N}_{F,\mathrm{Ind}} < (q - q_0)C\right).$$

□

By varying the penalty term $C(n)$ one can, as in the QMLE–based case, define a multitude of information criteria. Two particular examples are the analogs of the BIC and the CAIC (cp. (3.39) and (3.20)), which we define as

$$\mathrm{BIC}_n^{\mathrm{Ind}}(\Theta) := \mathcal{L}_{\mathrm{Ind}}(\widehat{\vartheta}_{\mathrm{Ind}}^n, Y^n) + N(\Theta)\frac{\log(n)}{n} \tag{4.56}$$

and

$$\mathrm{CAIC}_n^{\mathrm{Ind}}(\Theta) := \mathcal{L}_{\mathrm{Ind}}(\widehat{\vartheta}_{\mathrm{Ind}}^n, Y^n) + \frac{2N(\Theta)}{n}. \tag{4.57}$$

Note that we gave these criteria the same names as in the QMLE–based case. However, the interpretations derived in Subsection 3.4.1 and Subsection 3.5.1 as approximations of the Kullback–Leibler discrepancy and the Bayesian a posteriori probability, respectively, are no longer valid. These interpretations naturally led to the introduction of the QMLE, which we do not have here. Nevertheless, Theorem 4.50 enables us to easily derive their consistency properties:

Corollary 4.51. *BIC_n^{Ind} is a weakly consistent information criterion, $CAIC_n^{Ind}$ is neither weakly nor strongly consistent.*

Proof. The properties are obvious from Theorem 4.50. □

Remark 4.52. *a) If one wanted to show strong consistency for this class of information criteria, it would be necessary to have a law of the iterated logarithm for $\mathcal{L}_{Ind}(\vartheta^*, Y^n) - \mathcal{L}_{Ind}(\widehat{\vartheta}_{Ind}^n, Y^n)$, analogous to Theorem 3.4. The proof would then be very similar to that of Theorem 3.10a). However, due to the definition of $\mathcal{L}_{Ind}(\vartheta, Y^n)$, such a law is not easily obtained. We conjecture that the analogous result to Theorem 3.10a) would then hold, i.e. strong consistency would especially hold if $\lim_{n\to\infty} C(n) = \infty$.*

b) The new information criteria of this section are robust in the following, heuristic sense: If the observations Y^n are replaced by outlier–afflicted observations $\widetilde{Y}^n$

as defined in (4.2)*, by Theorem 4.42 we know that the value of* $\widehat{\vartheta}^n_{Ind}$ *remains close to its value in the scenario without outliers. Hence,* $\mathcal{L}_{Ind}(\widehat{\vartheta}^n_{Ind}, \widetilde{Y}^n)$ *will then also be close to* $\mathcal{L}_{Ind}(\widehat{\vartheta}^n_{Ind}, Y^n)$*. The information criterion therefore arrives at the same selected parameter space in most situations even though outliers are present. On the contrary, for the QMLE based information criteria, this does bot hold since the QML estimate* $\widehat{\vartheta}^n$ *is not resistant and therefore can be arbitrarily bad in the presence of outliers, leading to frequent wrong model choices by the criteria. Our simulation study in Subsection 4.5.2 confirms this behavior. A more formal treatment is found in Machado [1993], however there a different structure of the estimator upon which the criteria are built is assumed, which is why the results do not apply here.*

4.5. SIMULATION STUDY OF INDIRECT ESTIMATION

In this section, we illustrate the theoretical results from Section 4.3 and Section 4.4 by means of simulation. We consider several situations to assess the performance of our indirect estimator. The simulation methods for the CARMA processes used throughout the study are the same as in Section 3.6. We simulate the CARMA process on the interval $[0, 1000]$ and choose a sampling distance of $h = 1$, resulting in $n = 1000$ observations of the discrete–time process. The simulated processes are driven either by a standard Brownian motion or by a univariate NIG process. For the NIG process we use the parameters $\alpha = 3$, $\beta = 1$, $\Delta = 1$, $\delta = 2.5145$ and $\mu = -0.8890$, where the interpretation of the parameters is analogous to those of the two–dimensional NIG process in Section 3.6. These parameters result in a zero–mean Lévy process with variance approximately 1, which allows for comparison of the results to the Brownian motion case.

For the indirect estimator in Definition 4.22, we take $\widehat{\pi}^n$ as GM estimator as in Subsubsection 4.3.3.1. For calculating the GM estimator, we use the S–Plus software. This is done because it provides a pre–built function `ar.gm` for applying GM estimators to AR processes. This function uses a Mallows estimator as in Example 4.25a). The weight function $w(y)$ is the Tukey bisquare function from Example 4.25b) applied to $\|y\|$, for the function $\psi(u)$ the user can choose between the two classes explained in Example 4.25b), namely the Huber ψ_k–functions and the bisquare function. The function is implemented as an iterative least squares procedure as described by [Martin 1980, p. 231ff.] and therefore also allows to use first some iteration steps with the Huber ψ_k–function and then some steps with the bisquare function. As advocated by Martin [1980], we experience that doing 6 iterations using the Huber function and then 2 steps with the bisquare function works well. In our experiments we choose $k = 4$ for the tuning constant of the ψ_k–function. The order of the auxiliary AR representation is chosen conveniently for the different situations and will be mentioned explicitly in each example. In general, we set $s = 75$ to obtain the simulation–based observations $(Y_\vartheta(h), \ldots, Y_\vartheta(snh))$ in the simulation part of the indirect procedure. The Lévy process used for the simulation is the same as the one driving the observed CARMA process. For the estimator $\widehat{\pi}^n_{\mathrm{S}}$, we choose the least squares estimator of Definition 4.33. For the weighting matrix Ω we choose the identity matrix for convenience reasons. Some (unreported) experiments in which we first estimated the asymptotic covariance matrix of the GM estimator by the empirical covariance matrix of a suitable number of independent realizations of $\widehat{\pi}^n$ and set Ω to be the inverse of that estimate did not significantly affect the procedure positively or negatively, so that the use of the convenient identity matrix seems

justified. For the outlier model as defined in Definition 4.1b) via (4.2), we choose the process $(V_n)_{n\in\mathbb{Z}}$ as i.i.d. Bernoulli variables, where $\mathbb{P}(V_1 = 1) = \gamma$ varies and will be given in detail for each experiment. In all but one of the studies, the process $(Z_n)_{n\in\mathbb{Z}}$ is chosen to be deterministic, i.e. $Z_n = \xi$ for $n \in \mathbb{Z}$. We use varying values of ξ. In the experiment where $(Z_n)_{n\in\mathbb{Z}}$ has a different structure, this will be mentioned explicitly.

4.5.1. Parameter estimation

Case 1: Processes driven by Brownian motion

In each experiment, we calculate the indirect estimator and, for comparison purposes, the QMLE as defined in Subsection 2.2.3 in 50 independent replications and report on the average estimated value, the bias relative to the true parameters and the empirical variance of the parameter estimates.
In a first experiment, we use as true process a CARMA(1,0) process with parameter $\vartheta_0^{(1)} = -2$. This process is of particular interest, because its discretely sampled version admits a weak ARMA(1,0), i.e. an AR(1), representation. For this reason, one would expect the procedure to work very well here as the auxiliary representation is actually exact. Naturally, we choose $r = 1$ in this case. We consider three different scenarios of outlier contamination. In the first case, we set $\xi = 5$ and $\gamma = 0.1$. In the second, we let $\xi = 10$ and $\gamma = 0.1$, while for the last one we choose $\xi = 5$ and $\gamma = 0.15$. Note that already $\xi = 5$ represents quite large outliers, since for a sample path in this situation we typically observe that the values of the discretely sampled process lie between -3.5 and 3.5. Figure 4.2 shows a typical sample path of the CARMA(1,0) process with parameter $\vartheta_0^{(1)}$, the discretely sampled process and the outlier–afflicted process in the situation with $\gamma = 0.1$, $\xi = 5$. The results of the simulation studies are given in Table 4.1, Table 4.2 and Table 4.3, respectively.

	$\xi = 5$, $\gamma = 0.1$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-2.4497	-0.4497	0.0559	-2.0768	-0.0768	0.0513

Table 4.1.: Results for $\vartheta_0^{(1)}$, $\xi = 5$, $\gamma = 0.1$

As we can see, already in the case of $\xi = 5$ and $\gamma = 0.1$, the indirect estimator performs vastly better than the QMLE, giving a much less biased estimate at a similar variance. In the situation of Table 4.2, i.e. when ξ is increased to 10, the QMLE has lost all its information about the true parameter and provides no useful

	$\xi = 10, \gamma = 0.1$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-5.9097	-3.9097	0.2360	-2.0245	-0.0245	0.0663

Table 4.2.: Results for $\vartheta_0^{(1)}$, $\xi = 10$, $\gamma = 0.1$

	$\xi = 5, \gamma = 0.15$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-2.7873	-0.7873	0.3860	-2.1703	-0.1703	0.0973

Table 4.3.: Results for $\vartheta_0^{(1)}$, $\xi = 5$, $\gamma = 0.15$

estimate anymore. On the other hand, the indirect estimator stays close to the true value (we explain the even smaller bias in comparison to the situation with $\xi = 5$ as caused by the relatively small number of 50 iterations. Of course one should not expect the bias to systematically decrease when ξ increases), while the variance has increased only slightly. Increasing γ to 0.15 but keeping $\xi = 5$ shows that both estimators perform worse than in the situation with $\gamma = 0.1$, which is to be expected. But once again, the indirect estimator deals better with higher outlier percentage than the QMLE. Comparing these studies, we see that for the indirect estimator the percentage of outliers has a bigger effect on the estimates than the actual size of the outliers. For the QMLE however, the situation is reversed: its performance, relative to Table 4.1, is worse when ξ is increased to 10 compared to the situation where $\xi = 5$ is kept and γ increases to 0.15.

For our next simulation experiment, we move away from the CARMA(1,0) process to a CARMA(3,1) process. This especially means that the sampled process is not a weak AR process anymore, meaning that we truly make use of all the components of the indirect estimation procedure now. The true parameter is

$$\vartheta_0^{(2)} = \begin{pmatrix} -1 & -2 & -2 & 0 & 1 \end{pmatrix}.$$

For this process, we choose $r = 5$, which is also the minimum order of the auxiliary AR representation for which the assumptions on the indirect procedure are satisfied. We do four experiments in this setup. We estimate $\vartheta_0^{(2)}$ for each of the following contamination configurations: $\xi = 5$ and $\gamma = 0.1$, $\xi = 10$ and $\gamma = 0.1$, $\xi = 5$ and $\gamma = \frac{1}{6}$ and $\xi = 5$ and $\gamma = 0.25$. Remember that in this situation, the breakdown point as defined in Subsubsection 4.3.4.2 has an upper bound of $\frac{1}{6}$ since we have $r = 5$. Hence, $\gamma = 0.25$ lies above the breakdown point and we would expect to encounter problems in the estimation procedure, while for $\gamma \leq \frac{1}{6}$ these problems should not

occur. We will see that this is indeed the case. Figure 4.3 shows a typical sample path of the CARMA(3,1) process with parameter $\vartheta_0^{(2)}$, the discretely sampled process and the outlier–afflicted process in the situation with $\gamma = 0.1$, $\xi = 5$. The results of the four experiments are given in Table 4.4, Table 4.5, Table 4.6 and Table 4.7, respectively.

	$\xi = 5, \gamma = 0.1$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-0.6876	0.3124	0.0294	-1.0174	-0.0174	0.0114
ϑ_2	-2.6307	-0.6307	0.2550	-1.9930	0.0070	0.0068
ϑ_3	-3.3831	-1.3831	0.0573	-1.9954	0.0046	0.0194
ϑ_4	2.5467	2.5467	0.0190	0.0048	0.0048	0.0040
ϑ_5	0.5621	-0.4379	0.0498	1.0007	0.0007	0.0064

Table 4.4.: Results for $\vartheta_0^{(2)}$, $\xi = 5$, $\gamma = 0.1$

	$\xi = 10, \gamma = 0.1$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-0.4614	0.5386	0.0478	-1.0202	-0.0202	0.0110
ϑ_2	-1.4955	0.5045	0.1616	-2.0063	-0.0063	0.0107
ϑ_3	-1.8424	0.1576	0.0953	-1.9802	0.0198	0.0259
ϑ_4	3.3178	3.3178	0.1280	0.0047	0.0047	0.0066
ϑ_5	2.6317	1.6317	0.0477	0.9987	-0.0013	0.0074

Table 4.5.: Results for $\vartheta_0^{(2)}$, $\xi = 10$, $\gamma = 0.1$

	$\xi = 5, \gamma = \frac{1}{6}$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-0.4445	0.5555	0.0029	-1.0052	-0.0052	0.0071
ϑ_2	-2.3450	-0.3450	0.1135	-1.9815	0.0185	0.0136
ϑ_3	-3.5119	-1.5119	0.1823	-2.0210	-0.0210	0.0276
ϑ_4	3.1446	3.1446	0.0376	0.0106	0.0106	0.0057
ϑ_5	0.7903	-0.2097	0.1317	1.0043	0.0043	0.0054

Table 4.6.: Results for $\vartheta_0^{(2)}$, $\xi = 5$, $\gamma = \frac{1}{6}$

	$\xi = 5, \gamma = 0.25$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-0.1022	0.8978	0.0029	-0.0638	0.9362	0.0024
ϑ_2	-1.5663	0.4337	0.1135	-2.9026	-0.9026	0.0008
ϑ_3	-4.2751	-2.2751	0.1823	-2.1834	-0.1834	0.0094
ϑ_4	4.2080	4.2080	0.0376	1.1326	1.1326	0.0018
ϑ_5	0.8238	-0.1762	0.1317	1.8766	0.8766	0.0021

Table 4.7.: Results for $\vartheta_0^{(2)}$, $\xi = 5$, $\gamma = 0.25$

For the first two experiments, i.e. those in which $\gamma = 0.1$, we immediately recognize the maximum likelihood estimate as basically worthless in this setup, being severely biased and very far from the true parameter value. Especially the inclusion of a zero component in the true parameter seems to pose a major problem, since this component is affected by the most bias. On the other hand, the indirect estimator is still very close to the true parameter value in all of the components, including the zero component. This reaffirms that the indirect estimation procedure works in practical scenarios and the results in the former experiments were not due to the use of the CARMA(1,0) process. Increasing ξ to 10 while keeping $\gamma = 0.1$ results in a slightly worse performance of the indirect estimator. However, the increase in the bias is not very substantial and the estimates are still reasonably close to the true values. This of course cannot be said about the maximum likelihood estimator, which again delivers severely biased estimates.

Comparing Table 4.4 to Table 4.6, we see that the increase of γ from 0.1 to $\frac{1}{6}$ also affects the performance of the indirect estimator. For all components of $\vartheta_0^{(2)}$ but the first, the bias of the indirect estimator increases (even though it decreases for the first component, we still have an overall increase). A similar effect occurs for the variances. However, the loss in quality of the indirect estimator is manageable and the calculated estimates still closely resemble the true parameter. This means that even at the breakdown point, the performance of the indirect estimator is satisfying, although of course not as good as for lower contamination probabilities.

The situation is vastly different in the experiment with $\gamma = 0.25$. Here, we see that the indirect estimator, too, gives estimates which are severely biased and quite far away from the true parameters. Although the performance of the QMLE is still rather bad in the estimation of the zero component, its bias is less than that of the indirect estimator in three of the five components, but still substantial. The estimates delivered by the indirect estimator, however, are not satisfying either in this case. This of course is explained by the outlier probability of $\gamma = 0.25$, which causes the

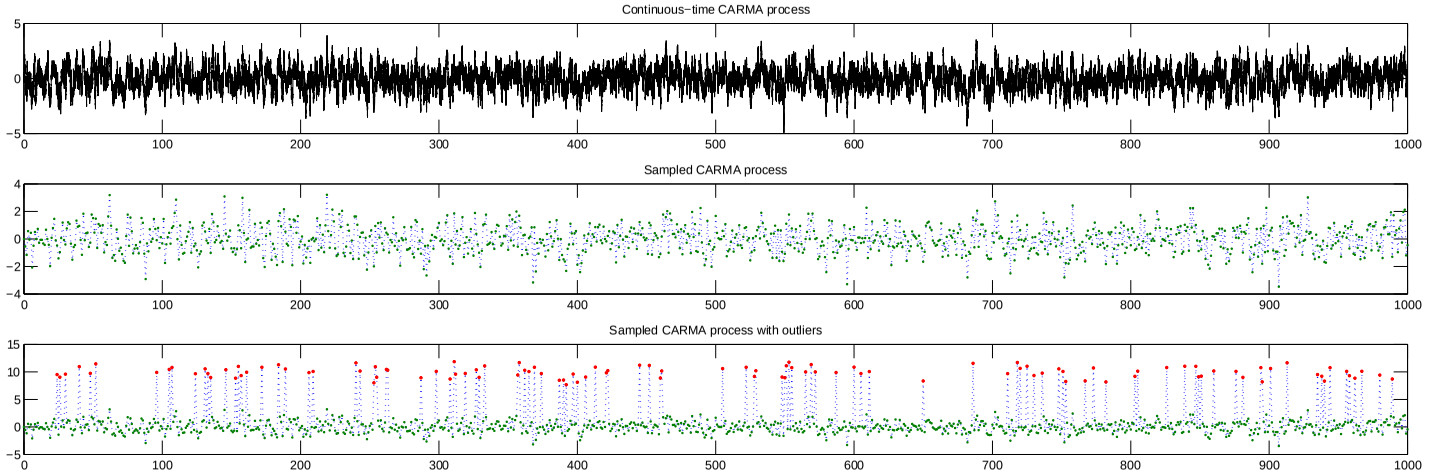

Figure 4.2.: Sample path of a CARMA(1,0) process and its discretely sampled version without and with outliers.

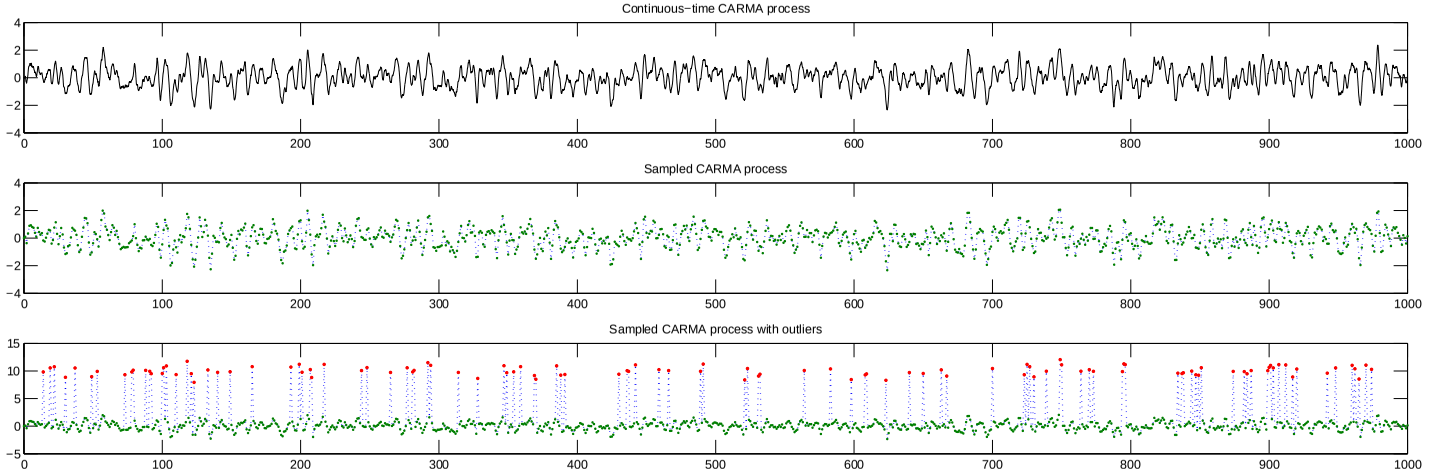

Figure 4.3.: Sample path of a CARMA(3,1) process and its discretely sampled version without and with outliers.

bad performance. We also observe that the numerical procedure used to obtain the parameter estimates in this scenario quite often fails to deliver a result at all because the algorithm terminates with an error. The error occurs when the estimated value of ϑ_0 is not an element of Θ anymore (the algorithm first calculates an estimate and then checks if it is an element of Θ, i.e. if Assumption B is satisfied). The results in the table are averaged over experiments in which the algorithm did deliver a result, the failed attempts were discarded. The ratio of successful to unsuccessful experiments was roughly equal to 1:2, i.e. the algorithm failed about twice as often as it succeeded. In this sense, we can say that the estimator has broken down: for a given outlier–contaminated sample, it either does not return an admissible estimate at all, or, if it does, the estimate is far away from the true parameter. The latter statement is also evident from the fact that the variance of the indirect estimates is far smaller in this case than in the other experiments, i.e. if the algorithm is able to calculate a result, there is very little variance in it, which, intuitively, means that the algorithm typically returns very similar bad estimates if it returns a result at all.

Lastly, we switch to another class of CARMA processes, this time using a CARMA(2,1) process with true parameter

$$\vartheta_0^{(3)} = \begin{pmatrix} -0.5 & -1 & -2 \end{pmatrix}.$$

We report on two simulation studies, which are a bit different than the ones conducted before. In the first of these two studies, we let $\gamma = 0$ to compare the performance of the indirect estimator to that of the QMLE in the uncontaminated case. For the other CARMA processes of this section we also studied this situation, but only choose to report on it once here since the results were very similar and the same conclusions could be drawn. For the second study, we did not use the simple additive outlier model but instead one that is a little more involved. Namely, we chose the process $(Z_n)_{n\in\mathbb{Z}}$ to be a sequence of i.i.d. random variables with $\mathbb{P}(Z_1 = \xi) = \mathbb{P}(Z_1 = -\xi) = \frac{1}{2}$, i.e. every time an outlier appears the sign is chosen randomly with equal probability. This change in model serves to study the performance of the procedure in circumstances that are a bit more complicated than the simple additive outlier model with fixed size of ξ. We choose $\gamma = 0.1$ and $\xi = 10$ in this study. In both studies, $r = 3$. The results are given in Table 4.8 and Table 4.9.

In the situation without outliers, both estimators are very close to the true parameter values. The differences in the bias and variance, which result in the indirect estimator performing even slightly better than the QMLE here, we explain as being not systematic, but due to the approximations in the numerical procedure and the number of only 50 iterations. We see that both estimators can be used

	$\gamma = 0$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-0.5112	-0.0112	0.0071	-0.5080	-0.0080	0.0086
ϑ_2	-0.9969	0.0031	0.0056	-1.0019	-0.0019	0.0050
ϑ_3	-2.0427	-0.0427	0.0246	-2.0156	-0.0156	0.0112

Table 4.8.: Results for $\vartheta_0^{(3)}$, $\gamma = 0$ (no outliers).

	$\mathbb{P}(Z_1 = 10) = \mathbb{P}(Z_1 = -10) = 0.5$, $\gamma = 0.1$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-0.3796	0.1204	0.0585	-0.4908	0.0092	0.0134
ϑ_2	-2.0699	-1.0699	0.2484	-1.0479	-0.0479	0.0132
ϑ_3	-4.0399	-2.0399	0.1408	-2.0015	-0.0015	0.0091

Table 4.9.: Results for $\vartheta_0^{(3)}$, $\mathbb{P}(Z_1 = 10) = \mathbb{P}(Z_1 = -10) = \frac{1}{2}$, $\gamma = 0.1$

to achieve satisfying results, which confirms the theoretical results on the indirect estimator in the uncontaminated case. For the other true parameters, we conducted the same study (i.e. in the outlier–free situation) and observed basically the same results, i.e. both estimators performed well and there was no notable, systematic better performance of the one or the other.
Using the indirect estimator also yields satisfying results under the more complicated outlier model. To a degree, this is not surprising since the GM estimator, which controls the robustness properties of $\widehat{\vartheta}^n_{\mathrm{Ind}}$, is per construction not sensitive to the sign of the outlier, but only to the absolute size. Since the outlier–generating process is symmetric and the total percentage of outliers is still at 10%, it seems intuitive that the performance does not differ much from before. Similarly, it is to be expected that the QMLE does not improve in comparison to the other outlier scenarios, which is confirmed by our study here.

Case 2: NIG process

In the following experiments, we replace the driving Brownian motion by the NIG process and repeat the experiments of Table 4.1, Table 4.4 and Table 4.5, using the same outlier configurations and auxiliary AR orders as in those experiments. The results are given in Table 4.10, Table 4.11 and Table 4.12.

We see that the results do not substantially differ from those in the Brownian motion case. In all three experiments with the NIG process, the QMLE, just like in the Brownian motion case, ceases to be a meaningful estimator in the presence of

	$\xi = 5, \gamma = 0.1$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-2.3851	-0.3851	0.0650	-2.0536	-0.0536	0.0860

Table 4.10.: Results for $\vartheta_0^{(1)}$, $\xi = 5$, $\gamma = 0.1$, driving NIG noise

	$\xi = 5, \gamma = 0.1$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-0.6669	0.3331	0.0311	-1.0113	-0.0113	0.0056
ϑ_2	-2.5668	-0.5668	0.2676	-1.9936	0.0064	0.0042
ϑ_3	-3.3917	-1.3917	0.0626	-1.9967	0.0033	0.0096
ϑ_4	0.5077	-0.4923	0.0535	1.0027	0.0027	0.0043
ϑ_5	2.5503	2.5503	0.0236	-0.0070	-0.0070	0.0035

Table 4.11.: Results for $\vartheta_0^{(2)}$, $\xi = 5$, $\gamma = 0.1$, driving NIG noise

outliers in the data. The indirect estimator on the other hand continues to provide good estimates. Of course, there are some differences in the bias and variance of the experiments with the NIG process in comparison to the experiments with the Brownian motion. Specifically, the bias and variance of the indirect estimator are a bit higher overall (as compared to Table 4.1) in the situation of Table 4.10, while in the situation of Table 4.11 and Table 4.12 we observe that the bias also is, on average, greater than in the corresponding experiments with Brownian motion while the variance is slightly lower. When we consider the bias, this matches the results of the simulation study in Section 3.6, where the performance typically also was slightly worse in the NIG case but otherwise very similar. We therefore conclude that indirect estimator is not limited to Gaussian CARMA processes, but also can be applied successfully for more general Lévy processes.

4.5.2. Model selection

The simulation experiment in this subsection serves to test the performance of our information criterion defined in Section 4.4 in the presence of outliers. We use the CARMA(3,1) process with true parameter $\vartheta_0^{(2)}$ as before as data–generating Lévy process. We compare the true parameter space Θ_0, which is the space of all CARMA(3,1) processes, to the space of all CARMA(3,2) processes and the space of all CARMA(3,0) processes. The true parameter space is nested in the former space, while the latter is itself nested in Θ_0. We consider two different information criteria, namely $\mathrm{BIC}_n^{\mathrm{Ind}}$ and $\mathrm{CAIC}_n^{\mathrm{Ind}}$ as defined in (4.56) and (4.57), respectively. We compare these two criteria to the performance of the QMLE–based BIC_n, of which

	$\xi = 10, \gamma = 0.1$					
	MLE			Indirect		
	Mean	Bias	Variance	Mean	Bias	Variance
ϑ_1	-0.4691	0.5309	0.0689	-1.0030	-0.0030	0.0045
ϑ_2	-1.5286	0.4714	0.1200	-1.9712	0.0288	0.0112
ϑ_3	-1.8365	0.1635	0.0668	-2.0321	-0.0321	0.0178
ϑ_4	2.6651	1.6651	0.0316	1.0056	0.0056	0.0038
ϑ_5	3.3668	3.3668	0.0981	0.0127	0.0127	0.0056

Table 4.12.: Results for $\vartheta_0^{(2)}$, $\xi = 10$, $\gamma = 0.1$, driving NIG noise

we know that it is strongly consistent in the outlier–free case. For this experiment, we chose again $\xi = 5$ and $\gamma = 0.1$. Both the Brownian motion and the NIG process were used as driving Lévy process. For both scenarios, we conducted 20 replications in total, the number being somewhat low because of the very high computation time required (one replication typically took between 60 and 90 minutes to finish). The results are given in Table 4.13.

Space	Model			BM			NIG		
	p	q	$N(\Theta)$	BIC_n	$\mathrm{BIC}_n^{\mathrm{Ind}}$	$\mathrm{CAIC}_n^{\mathrm{Ind}}$	BIC_n	$\mathrm{BIC}_n^{\mathrm{Ind}}$	$\mathrm{CAIC}_n^{\mathrm{Ind}}$
1	3	0	3	0	0	0	0	0	0
2	3	1	4	1	18	16	2	20	19
3	3	2	5	19	2	4	18	0	1

Table 4.13.: Results for the true parameter $\vartheta_0^{(2)}$ in space 2.

Despite the relatively low number of repetitions, the results are insightful. We see that the BIC_n is drastically affected by the outliers and now has an error rate of 95% and 90%, respectively, having lost the strong consistency. On the other hand, $\mathrm{BIC}_n^{\mathrm{Ind}}$, which we showed to be only weakly consistent for the outlier–free case, seems to retain this property when there are outliers in the data. In 90% of the replications in the Brownian motion case, it comes to the correct decision while it overfits in the other 10%. Of course we know that for a weakly consistent criterion the overfitting rate should eventually go to zero as n increases. However, an error rate of 10% seems acceptable, given the fact that the outliers have to be taken into account. For the NIG case, it even achieves a perfect score. We conjecture, however, that this perfect score is due to the low number of total experiments and we would observe overfitting in the NIG case, too, when carrying out more experiments. In comparison, $\mathrm{CAIC}_n^{\mathrm{Ind}}$ exhibits an overfitting rate of 20% and 5%, i.e. in both cases it overfits more often. This is an expected effect, since we know that $\mathrm{CAIC}_n^{\mathrm{Ind}}$ is not consistent due to the deterministic penalty term and thus exhibits a non–zero

asymptotic overfitting probability even in the outlier–free case. This experiment clearly shows that it is advisable to use criteria of the form given by $\mathrm{IC}_n^{\mathrm{Ind}}$ instead of the "classical" information criteria when one suspects the presence of outliers in the data, since they perform much better, even compared to information criteria which are strongly consistent in the absence of outliers.

Summarizing the studies, we can say that our indirect estimator performs convincingly for various orders p and q of the data–generating CARMA process, for different driving Lévy processes and for a variety of outlier configurations. Of course, it is clear that this method, too, has its bounds. We especially saw that both an increase of γ, the proportion of outliers, and of the size of the outliers affect the performance. Increasing γ too far eventually causes the estimator to break down. This is natural, since with an increasing number of outliers we have less "good" observations which provide full information about the data–generating process and it therefore becomes harder to detect the true structure of the process. Nevertheless, as soon as outliers are present, the use of the indirect estimator instead of the QMLE always seems advisable, since there was no situation in which the performance of the QMLE came close to being satisfying. On the other hand, when no outliers are present, one does not lose much, if at all, by using the indirect estimator. However, it should be mentioned that the computation is more involved and slower in comparison to the QMLE method, such that the latter might be preferred due to this when one is certain that no outliers are present. If one wishes to be on the safe side, however, usage of the indirect estimator is probably not a mistake.

CHAPTER 5

Conclusion and outlook

This thesis has contributed to the field of statistical inference for MCARMA processes in two ways. On the one hand, model selection procedures for stationary MCARMA processes based on observations at an equidistant time grid have been studied. Using the Echelon form as parametrization, the approach via likelihood–based information criteria under natural identifiability and regularity conditions provided us with a method of estimating the Kronecker indices and the orders p and q of the continuous–time, data–generating process from discrete–time observations. Moreover, in Theorem 3.10, we were able to obtain results that allowed to explicitly derive the consistency properties of the criteria from the penalty term $C(n)$. They enable us to answer the question whether the true model will be selected asymptotically. We showed in Subsection 3.4.1 and Subsection 3.5.1 that the well–known AIC and BIC fit naturally in this framework and the underlying ideas of approximating the Kullback–Leibler discrepancy and the Bayesian a posteriori probability, respectively, remain valid.

On the other hand, the thesis also contributed to the theory of parameter estimation for MCARMA processes with a special emphasis on robustness in the one–dimensional case. We were first able to introduce a quite general class of strongly consistent, asymptotically normally distributed estimators for the parameters of MCARMA processes, the M–estimators of Section 4.2. In the special case of one–dimensional processes, we also studied the indirect estimator in detail, which is another strongly consistent, asymptotically normally distributed estimator when applied to outlier–free data as evident from Theorem 4.23. Moreover, due to its special structure, we were

able to show in Subsection 4.3.4 that if a GM estimator is applied to estimate the parameters of the auxiliary AR representation of the sampled CARMA process, its robustness properties are inherited by the indirect estimator. Thus, we can provide a robust estimator for the parameters of the CARMA process. Using this estimator as foundation, we were also able to extend the model selection criteria from the first part of the thesis in Section 4.4, establishing a bridge between the two fields.
For both the field of model selection and robust estimation simulation studies showed that, at least for simulated data, the theoretical results can also be observed and verified in practical situations.

There are several directions for further research. For example, there are generalizations of (M)CARMA processes, e.g. the continuous–time threshold ARMA (CTARMA) processes of Stramer et al. [1996] or the non–stationary, cointegrated MCARMA processes treated in Scholz [2016]. These models come with additional parameters that define their structure, in the latter case for example the cointegration rank. At the moment, information criteria that are able to estimate these additional structural parameters, too, have not been studied yet and therefore of course the question of their properties, for example consistency, is open. Another possible way of extending the results on information criteria could be to relax the assumption that there is a "true" model Θ_0 among the candidate models. We always worked under this assumption in Chapter 3, but of course in applications this is an idealization, since real–world data will never be a perfect realization of a theoretical mathematical model. Especially the proofs with regard to consistency relied strongly on this assumption, so that it would be interesting to explore what happens when it is dropped and all spaces under consideration are then misspecified. A starting point in this direction could be the work by Hurvich and Tsai [1991], who examined this question in the context of linear regression and discrete–time autoregressive processes. Additional remarks can also be found in [Burnham and Anderson 2002, Subsection 6.3.4].

In the area of robust estimation, the potential extensions are plentiful, too. First of all, we only studied one estimator of the CARMA parameters that achieves the desired robustness. There is a multitude of other approaches in the literature which one could investigate and transfer to the context of (M)CARMA processes. In particular, one could attempt to take a similar route as with the QMLE of Subsection 2.2.3, but instead of the Kalman filter use a robust filter, as proposed by Masreliez [1975], to calculate robust innovations and base the parameter estimation on these. For AR and ARMA processes, an overview of these methods is given in [Maronna et al.

2006, Section 8.6 and Section 8.8], where additional references can also be found. Another class of estimators for ARMA models, which could possibly be extended to the framework of CARMA processes, are the RA estimators, which are based on the residual autocovariances, as proposed by Bustos and Yohai [1986].
The question how robust estimators (in the sense of estimators that truly are insensitive to outliers) for multivariate CARMA processes can be obtained is also left open at the moment. In particular, a generalization of the indirect estimator as studied in Section 4.3 is not straightforwardly possible, since a theory of GM estimation for multivariate AR processes does not exist to our knowledge. There are several approaches towards robust parameter estimation of vector autoregressive processes, e.g. in Garcia Ben et al. [1999], where the RA estimators of Bustos and Yohai [1986] are generalized, in Croux and Joossens [2008] where a least trimmed squares procedure is used or in Muler and Yohai [2013], where the BMM estimators of Muler et al. [2009] are generalized to the multivariate setting. All these approaches, however, only give robust estimators for VAR and not for VARMA processes. Therefore, for a robust estimator of MCARMA processes, combining the indirect estimation method with one of these procedures could be a promising avenue in future research.
Another possible extension would be to consider a different way of modeling the outliers. In this thesis, the estimators were founded on discrete–time observations of MCARMA processes and we modeled the outliers via the general replacement model for discrete–time stochastic processes. A different approach would be to choose a model in which already the continuous–time process is afflicted by outliers. To our knowledge, no systematic approaches towards modeling outliers in continuous–time processes exist yet, such that a first step could be to define a suitable analog in continuous time of the general replacement model of Section 4.1 and then continue from there, e.g. by again sampling on a discrete time–grid and building estimators based on these observations.

In all of the thesis, inference was based on discrete–time, equidistant observations with sampling distance $h > 0$ fixed. It is also possible to move away from this assumption in order to better treat irregularly spaced or high–frequency data, for example by observing the data–generating MCARMA process at times h_n, $2h_n$, ..., nh_n for a sequence $(h_n)_{n\in\mathbb{N}}$ with $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. In Fasen and Fuchs [2013a] and Fasen and Fuchs [2013b] this observation scheme is used to first study the behavior of the periodogram of a sampled CARMA(p,q) process and subsequently construct an estimator for the parameters of the CARMA(p,q) process based on the periodogram. It is assumed that p and q are known, such that the

development of model selection criteria in this framework is a topic that needs to be considered. It is also worth mentioning that in these two papers it is assumed that the driving Lévy process is symmetric and α–stable for some $\alpha \in (0, 2]$. If $\alpha < 2$, this means that the process has infinite variance. In the thesis at hand, we always assumed that the driving Lévy process has (at least) finite second moment, such that the use of α–stable Lévy processes with $\alpha < 2$ was not allowed. A big question for further research is therefore if similar results can be obtained if such processes are used.

Concluding, it is therefore evident that the field of statistical inference for MCARMA processes is far from completely explored and that there remain a lot of interesting questions to be investigated and eventually answered. Given the intuitive appeal as continuous–time analog of VARMA processes and the current popularity of MCARMA processes amongst scientists, we expect this to be a lively area of research in the years to come.

Appendices

APPENDIX A

Technical appendices

A.1. Auxiliary results for Section 3.3

In this appendix, we give the calculations for the Brownian motion case in Section 3.3.

Lemma A.1. *Let $A, B \in \mathbb{R}^{d\times d}$ be matrices, where B is symmetric. Then*

$$\operatorname{tr}\left((\operatorname{vec}(I_{d\times d}) \otimes \operatorname{vec}(I_{d\times d})^T)(A \otimes B)\right) = \operatorname{tr}(AB).$$

Proof. The structure of the matrix $\operatorname{vec}(I_{d\times d}) \otimes \operatorname{vec}(I_{d\times d})^T$ is as follows:

$$\operatorname{vec}(I_{d\times d}) \otimes \operatorname{vec}(I_{d\times d})^T = \begin{pmatrix} 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 1 \end{pmatrix},$$

where there are exactly d zero rows between the non-zero rows and also exactly d zero columns between the non-zero columns. By the definition of the Kronecker product,

we then have

$$
\begin{aligned}
&(\mathrm{vec}(I_{d\times d}) \otimes \mathrm{vec}(I_{d\times d})^T)(A \otimes B) \\
&= \begin{pmatrix}
a_{11}b_{11} + \ldots + a_{d1}b_{d1} & * & * & * & * & * & * & \ldots & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\
* & * & * & * & a_{12}b_{12} + \ldots + a_{d2}b_{d2} & * & * & \ldots & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\
* & * & * & * & * & * & * & \ldots & a_{1d}b_{1d} + \ldots + a_{dd}b_{dd}
\end{pmatrix}
\end{aligned}
$$

Again, there are exactly d zero rows between the non-zero rows. The asterisks mark entries of the matrix which are (possibly) non-zero, but not relevant in our case, since we are only interested in the trace. Alas, with the assumption that B is symmetric we have

$$
\begin{aligned}
&\mathrm{tr}\left((\mathrm{vec}(I_{d\times d}) \otimes \mathrm{vec}(I_{d\times d})^T)(A \otimes B)\right) \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}b_{ij} = \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}b_{ji} = \sum_{i=1}^{d}(AB)_{ii} = \mathrm{tr}(AB)
\end{aligned}
$$

and the assertion follows. □

Lemma A.2. *Assume that the Lévy process L which drives the observed process Y is a Brownian motion.*

a) Assume that the space Θ with associated family of continuous-time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfies Assumption B and that MCARMA$(A_{\vartheta^*}, B_{\vartheta^*}, C_{\vartheta^*}, L_{\vartheta^*}) = Y$ for the pseudo-true parameter ϑ^*. Then

$$\mathcal{I}(\vartheta^*) = 2\mathcal{J}(\vartheta^*).$$

b) There exists a space Θ_0 with associated family of continuous-time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta_0}$ satisfying Assumption B such that MCARMA$(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ for some $\vartheta_0 \in \Theta_0$. Moreover, Θ_0 is nested in Θ_0^E with map F, $N(\Theta_0) = N(\Theta_0^E) - 1$ and

$$\lambda_{max}(\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}) = 2.$$

c) The matrix $\mathcal{J}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_F\mathcal{I}(\vartheta^*)\mathcal{M}_F\mathcal{J}(\vartheta^*)^{\frac{1}{2}}$ *has only the eigenvalues* 2 *and* 0*, where* 0 *has multiplicity* $N(\Theta_0)$ *and* 2 *has multiplicity* $N(\Theta) - N(\Theta_0)$*.*

Proof. a) An analogous statement for VARMA processes is given in Boubacar Maïnassara and Francq [2011, Remark 2]. However, they state it without a proof. Since the proof is not so obvious we will give it in detail here for MCARMA processes. First, note that since the driving Lévy process is a Brownian motion, it holds per construction that the linear innovations $(\epsilon_k)_{k\in\mathbb{Z}}$ of the process $(Y(kh))_{k\in\mathbb{Z}}$ are i.i.d. $\mathcal{N}(0,V)$–distributed (cf. Definition 2.19). Moreover, per assumption it also holds that $\epsilon_{\vartheta^*,k} = \epsilon_k$ for every $k \in \mathbb{Z}$, hence we also have that $\epsilon_{\vartheta^*,k} \sim \mathcal{N}(0, V_{\vartheta^*})$ and $V_{\vartheta^*} = V$. By definition

$$\mathcal{I}(\vartheta^*) = \lim_{n\to\infty} n \operatorname{Var}\left(\nabla_\vartheta \mathcal{L}(\vartheta^*, Y^n)\right),$$

which means that for $i, j \in \{1, \dots, N(\Theta)\}$ we have to study terms of the form

$$\begin{aligned}
&\operatorname{Var}\left(n\nabla_\vartheta \mathcal{L}(\vartheta^*, Y^n)\right)_{ij} \\
&\overset{(3.2)}{=} \sum_{k=1}^{n} \mathbb{E}\left[\left(\operatorname{tr}\left(V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*}\right) - \operatorname{tr}\left(V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*}\right) + 2\partial_i \epsilon_{\vartheta_*,k}^T V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\right)\right. \\
&\left.\cdot\left(\operatorname{tr}\left(V_{\vartheta^*}^{-1}\partial_j V_{\vartheta^*}\right) - \operatorname{tr}\left(V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\partial_j V_{\vartheta^*}\right) + 2\partial_j \epsilon_{\vartheta_*,k}^T V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\right)\right] \\
&+ \sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n} \mathbb{E}\left[\left(\operatorname{tr}\left(V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*}\right) - \operatorname{tr}\left(V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*}\right) + 2\partial_i \epsilon_{\vartheta_*,k}^T V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\right)\right. \\
&\left.\cdot\left(\operatorname{tr}\left(V_{\vartheta^*}^{-1}\partial_j V_{\vartheta^*}\right) - \operatorname{tr}\left(V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,l}\epsilon_{\vartheta^*,l}^T V_{\vartheta^*}^{-1}\partial_j V_{\vartheta^*}\right) + 2\partial_j \epsilon_{\vartheta_*,l}^T V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,l}\right)\right] \\
&=: \sum_{k=1}^{n} a_k + \sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n} b_{k,l}. \qquad \text{(A.1)}
\end{aligned}$$

We start to investigate a_k. By definition, every innovation $\epsilon_{\vartheta^*,k}$ is orthogonal to $\overline{\operatorname{span}}\{Y(jh) : -\infty < j < k\}$ and by Lemma 2.22b) both $\partial_i\epsilon_{\vartheta^*,k}$ and $\partial_j\epsilon_{\vartheta^*,k}$ are elements of $\overline{\operatorname{span}}\{Y(jh) : -\infty < j < k\}$. Hence, $\epsilon_{\vartheta^*,k}$ is independent of $\partial_i\epsilon_{\vartheta^*,k}$ and $\partial_j\epsilon_{\vartheta^*,k}$. This, together with the independence of the innovation sequence $(\epsilon_{\vartheta^*,k})_{k\in\mathbb{N}}$, the fact that $\mathbb{E}[\partial_i\epsilon_{\vartheta^*,k}] = 0$, $\mathbb{E}[\epsilon_{\vartheta^*,k}\epsilon_{\vartheta^*,k}^T] = V_{\vartheta^*}$ and the interchangeability of trace and expectation, allows us to simplify

$$\begin{aligned}
a_k &= -\operatorname{tr}\left(V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*}\right)\operatorname{tr}\left(V_{\vartheta^*}^{-1}\partial_j V_{\vartheta^*}\right) \\
&\quad + \mathbb{E}\left[\operatorname{tr}\left(V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*}\right)\operatorname{tr}\left(V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\partial_j V_{\vartheta^*}\right)\right] \\
&\quad + 4\mathbb{E}\left[\partial_i \epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\partial_j \epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\epsilon_{\vartheta^*,k}\right] \\
&=: a_k^{(1)} + a_k^{(2)} + a_k^{(3)}. \qquad \text{(A.2)}
\end{aligned}$$

For the second term, we define $\tilde{\epsilon}_{\vartheta^*,k} = V_{\vartheta^*}^{-\frac{1}{2}} \epsilon_{\vartheta^*,k} \sim \mathcal{N}(0, I_{d\times d})$ and have by standard calculation rules for Kronecker products (Bernstein [2009, Proposition 7.1.6 and Proposition 7.1.12]):

$$\begin{aligned} a_k^{(2)} &= \mathbb{E}\left[\operatorname{tr}\left(\left(V_{\vartheta^*}^{-\frac{1}{2}} \tilde{\epsilon}_{\vartheta^*,k} \tilde{\epsilon}_{\vartheta^*,k}^T V_{\vartheta^*}^{-\frac{1}{2}} \partial_i V_{\vartheta^*}\right) \otimes \left(V_{\vartheta^*}^{-\frac{1}{2}} \tilde{\epsilon}_{\vartheta^*,k} \tilde{\epsilon}_{\vartheta^*,k}^T V_{\vartheta^*}^{-\frac{1}{2}} \partial_j V_{\vartheta^*}\right)\right)\right] \\ &= \operatorname{tr}\left(\left(V_{\vartheta^*}^{-\frac{1}{2}} \otimes V_{\vartheta^*}^{-\frac{1}{2}}\right) \cdot \mathbb{E}\left[\tilde{\epsilon}_{\vartheta^*,k} \tilde{\epsilon}_{\vartheta^*,k}^T \otimes \tilde{\epsilon}_{\vartheta^*,k} \tilde{\epsilon}_{\vartheta^*,k}^T\right] \cdot \left(V_{\vartheta^*}^{-\frac{1}{2}} \partial_i V_{\vartheta^*} \otimes V_{\vartheta^*}^{-\frac{1}{2}} \partial_j V_{\vartheta^*}\right)\right). \end{aligned}$$

Since $\tilde{\epsilon}_{\vartheta^*,k} \sim \mathcal{N}(0, I_{d\times d})$, by means of Balestra and Holly [1990, Theorem 1] the expectation appearing in the last line is

$$\mathbb{E}\left[\tilde{\epsilon}_{\vartheta^*,k} \tilde{\epsilon}_{\vartheta^*,k}^T \otimes \tilde{\epsilon}_{\vartheta^*,k} \tilde{\epsilon}_{\vartheta^*,k}^T\right] = K_{d,d} + I_{d^2\times d^2} + \operatorname{vec}(I_{d\times d}) \otimes \operatorname{vec}(I_{d\times d})^T,$$

where $K_{d,d}$ is the $d^2 \times d^2$ Kronecker permutation matrix (Bernstein [2009, Eq. (7.1.20)]). Together with the linearity and the cyclic permutation property of the trace, we use this to obtain

$$\begin{aligned} a_k^{(2)} &= \operatorname{tr}\left(K_{d,d}\left(V_{\vartheta^*}^{-\frac{1}{2}} \otimes V_{\vartheta^*}^{-\frac{1}{2}}\right)\left(V_{\vartheta^*}^{-\frac{1}{2}} \partial_i V_{\vartheta^*} \otimes V_{\vartheta^*}^{-\frac{1}{2}} \partial_j V_{\vartheta^*}\right)\right) \\ &\quad + \operatorname{tr}\left(\left(V_{\vartheta^*}^{-\frac{1}{2}} \otimes V_{\vartheta^*}^{-\frac{1}{2}}\right)\left(V_{\vartheta^*}^{-\frac{1}{2}} \partial_i V_{\vartheta^*} \otimes V_{\vartheta^*}^{-\frac{1}{2}} \partial_j V_{\vartheta^*}\right)\right) \\ &\quad + \operatorname{tr}\left((\operatorname{vec}(I_{d\times d}) \otimes \operatorname{vec}(I_{d\times d})^T)\left(V_{\vartheta^*}^{-\frac{1}{2}} \otimes V_{\vartheta^*}^{-\frac{1}{2}}\right)\left(V_{\vartheta^*}^{-\frac{1}{2}} \partial_i V_{\vartheta^*} \otimes V_{\vartheta^*}^{-\frac{1}{2}} \partial_j V_{\vartheta^*}\right)\right) \\ &= \operatorname{tr}\left(K_{d,d}\left(V_{\vartheta^*}^{-1} \partial_i V_{\vartheta^*} \otimes V_{\vartheta^*}^{-1} \partial_j V_{\vartheta^*}\right)\right) + \operatorname{tr}\left(V_{\vartheta^*}^{-1} \partial_i V_{\vartheta^*} \otimes V_{\vartheta^*}^{-1} \partial_j V_{\vartheta^*}\right) \\ &\quad + \operatorname{tr}\left((\operatorname{vec}(I_{d\times d}) \otimes \operatorname{vec}(I_{d\times d})^T)(V_{\vartheta^*}^{-1} \partial_i V_{\vartheta^*} \otimes V_{\vartheta^*}^{-1} \partial_j V_{\vartheta^*})\right). \end{aligned}$$

We now apply Lemma A.1 as well as Bernstein [2009, Fact 7.4.30 xviii) and Proposition 7.1.12] to get

$$a_k^{(2)} = 2 \operatorname{tr}\left(V_{\vartheta^*}^{-1} \partial_i V_{\vartheta^*} V_{\vartheta^*}^{-1} \partial_j V_{\vartheta^*}\right) + \operatorname{tr}\left(V_{\vartheta^*}^{-1} \partial_i V_{\vartheta^*}\right) \operatorname{tr}\left(V_{\vartheta^*}^{-1} \partial_j V_{\vartheta^*}\right).$$

It remains to consider $a_k^{(3)}$ in (A.2). The independence of $\partial_j \epsilon_{\vartheta^*,k} \partial_i \epsilon_{\vartheta^*,k}^T$ and $\epsilon_{\vartheta^*,k}$, the cyclic permutation property of the trace and the interchangeability of expectation and trace leads to

$$\begin{aligned} a_k^{(3)} &= \mathbb{E}\left[\operatorname{tr}\left(V_{\vartheta^*}^{-1} \epsilon_{\vartheta^*,k} \epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1} \partial_j \epsilon_{\vartheta^*,k} \partial_i \epsilon_{\vartheta^*,k}^T\right)\right] \\ &= \operatorname{tr}\left(\mathbb{E}\left[V_{\vartheta^*}^{-1} \epsilon_{\vartheta^*,k} \epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\right] \mathbb{E}\left[\partial_j \epsilon_{\vartheta^*,k} \partial_i \epsilon_{\vartheta^*,k}^T\right]\right) \\ &= \operatorname{tr}\left(V_{\vartheta^*}^{-1} \mathbb{E}\left[\partial_j \epsilon_{\vartheta^*,k} \partial_i \epsilon_{\vartheta^*,k}^T\right]\right) \\ &= \mathbb{E}\left[\partial_i \epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1} \partial_j \epsilon_{\vartheta^*,k}\right]. \end{aligned}$$

Combining those calculations finally results in

$$a_k = a_k^{(1)} + a_k^{(2)} + a_k^{(3)} = 2\,\mathrm{tr}\left(V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*} V_{\vartheta^*}^{-1}\partial_j V_{\vartheta^*}\right) + 4\mathbb{E}\left[\partial_i \epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\partial_j \epsilon_{\vartheta^*,k}\right].$$

By similar calculations, we can verify that $b_{k,l} = 0$ for $k \neq l$.
Finally, this implies

$$(\mathcal{I}(\vartheta^*))_{ij} = a_k = 2\,\mathrm{tr}\left(V_{\vartheta^*}^{-1}\partial_i V_{\vartheta^*} V_{\vartheta^*}^{-1}\partial_j V_{\vartheta^*}\right) + 4\mathbb{E}\left[\partial_i \epsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\partial_j \epsilon_{\vartheta^*,k}\right].$$

By Schlemm and Stelzer [2012, (2.33a) and (2.33b)], this term is equal to $(2\mathcal{J}(\vartheta^*))_{ij}$ as proclaimed.

b) & c) Denote by $v_1, \ldots, v_{N(\Theta_0^E)}$ the eigenvectors of $\mathcal{H}(\vartheta_0^E)$ which are an orthonormal basis of $\mathbb{R}^{N(\Theta_0^E)}$. Define $F = (v_1, \ldots, v_{N(\Theta_0^E)-1}) \in \mathbb{R}^{N(\Theta_0^E)\times(N(\Theta_0^E)-1)}$ and let $\Theta_0 \subseteq F^T\Theta_0^E$ be compact such that $F\Theta_0 + (\vartheta_0^E - FF^T\vartheta_0^E) \subseteq \Theta_0^E$ and $F^T\vartheta_0^E \in \Theta_0$. Define

$$\begin{aligned}&(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta_0}\\ :=&(A_{F\vartheta+(\vartheta_0^E-FF^T\vartheta_0^E)}, B_{F\vartheta+(\vartheta_0^E-FF^T\vartheta_0^E)}, C_{F\vartheta+(\vartheta_0^E-FF^T\vartheta_0^E)}, L_{F\vartheta+(\vartheta_0^E-FF^T\vartheta_0^E)})_{\vartheta\in\Theta_0}.\end{aligned}$$

Then $\vartheta_0 = F^T\vartheta_0^E$, Θ_0 is nested in Θ_0^E with map F and satisfies Assumption B, and $N(\Theta_0) = N(\Theta_0^E) - 1$. Moreover, the eigenvectors $v_1, \ldots, v_{N(\Theta_0^E)-1}$ are basis vectors of the image of F and $v_{N(\Theta_0^E)}$ is a basis of the kernel of F^T. Then $v_{N(\Theta_0^E)}$ is an eigenvector of $\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}$ for the eigenvalue 2 and $v_1, \ldots, v_{N(\Theta_0^E)-1}$ are eigenvectors of $\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}$ for the eigenvalue 0 as well.

□

A.2. Auxiliary results for Subsection 4.3.3

This appendix contains [Bustos 1982, Lemma 3.3 – Lemma 3.5], which are needed for the study of the asymptotic properties of GM estimators. The proofs do not differ from those in the original article (although we use slightly different assumptions). We list them in this appendix for easier reference, using the notation of Subsection 4.3.3.

Lemma A.3 (Bustos [1982], Lemma 3.3)**.** *Let $b_0 > 0$ be such that $\|\pi - \pi_0\| \leq b_0$ implies that $\|\mathcal{Q}_{GM}(\pi)\| \geq a\|\pi - \pi_0\|$ for some $a > 0$ and define*

$$C = \{\pi \in \mathbb{R}^r \times (0,\infty) : \|\pi - \pi_0\| \leq b_0\}. \tag{A.3}$$

Then it holds that

$$\sup_{\pi\in C}\frac{\left\|\sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi)-\mathscr{Q}_{GM}(\pi)-\Psi((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0))\right\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{GM}(\pi)\|}\overset{n\to\infty}{\to}0\ \mathbb{P}\text{-}a.s.$$

Proof. The existence of a and b_0 as in the assumption follows from the assumed non–singularity of $\mathcal{J}_{\mathrm{GM}}(\pi_0)=\nabla_\pi\mathscr{Q}_{\mathrm{GM}}(\pi_0)$. Define now

$$g_n((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi):=\sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi)-\mathscr{Q}_{\mathrm{GM}}(\pi)).$$

By (4.22) and (4.23) we have that $\mathscr{Q}_{\mathrm{GM}}(\pi_0)=0$. Therefore it holds

$$\begin{aligned}&\frac{\left\|\sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi)-\mathscr{Q}_{\mathrm{GM}}(\pi)-\Psi(\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0))\right\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\pi)\|}\\&=\frac{\left\|g_n((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi)-g((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0)\right\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\pi)\|}\end{aligned}$$

Applying the multivariate mean value theorem to the function $g(\pi)$, we find that there exists a constant $K>0$ such that

$$\begin{aligned}&\sup_{\pi\in C}\left\|g((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi)-g((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0)\right\|\\&\leq K\sup_{\pi\in C}\|\pi-\pi_0\|\\&\cdot\max_{i,j=1,\ldots,r+1}\left\|\sum_{t=1}^{n-r}\left(\partial_i\Psi_j((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi)-\mathbb{E}\left[\partial_i\Psi_j((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi)\right]\right)\right\|\end{aligned}$$

Rearranging some terms, we can conclude from this that

$$\begin{aligned}&\sup_{\pi\in C}\frac{\left\|g_n((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi)-g_n((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0)\right\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\pi)\|}\\&\leq K\sup_{\pi\in C}\frac{\|\pi-\pi_0\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\pi)\|}\\&\cdot\max_{i,j=1,\ldots,r+1}\left(\left\|\sum_{t=1}^{n-r}(\partial_i\Psi_j((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi)-\partial_i\Psi_j((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0))\right.\right.\\&\left.-(n-r)\mathbb{E}\left[\partial_i\Psi_j((\widetilde{Y}_1,\ldots,\widetilde{Y}_{1+r}),\pi)-\partial_i\Psi_j((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0)\right]\right\|\\&\left.+\left\|\sum_{t=1}^{n-r}\partial_i\Psi_j((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0)-(n-r)\mathbb{E}\left[\partial_i\Psi_j((\widetilde{Y}_1,\ldots,\widetilde{Y}_{1+r}),\pi_0)\right]\right\|\right)\end{aligned}\tag{A.4}$$

By using the inequalities defining the set C and introducing the factor $\frac{1}{n-r}$ in front of both sums, we obtain the next inequality:

$$\begin{aligned}\frac{(\text{A.4})}{n-r} \leq K \sup_{\pi\in C} \frac{1}{\frac{1}{\sqrt{n-r}b_0}+a} \\ \cdot \max_{i,j=1,\dots,r+1} \Bigg(\Bigg\| \frac{1}{n-r}\sum_{t=1}^{n-r} (\partial_i\Psi_j((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\pi) - \partial_i\Psi_j((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\pi_0)) \\ -\mathbb{E}\left[\partial_i\Psi_j((\widetilde{Y}_1,\dots,\widetilde{Y}_{1+r}),\pi) - \partial_i\Psi_j((\widetilde{Y}_1,\dots,\widetilde{Y}_{1+r}),\pi_0)\right]\Bigg\| \\ + \left\| \frac{1}{n-r}\sum_{t=1}^{n-r} \partial_i\Psi_j((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\pi_0) - \mathbb{E}\left[\partial_i\Psi_j((\widetilde{Y}_1,\dots,\widetilde{Y}_{1+r}),\pi_0)\right]\right\| \Bigg)\end{aligned}$$

Both summands converge to zero almost surely as $n\to\infty$, the first one by [Bustos 1982, Lemma 3.2], the second one by ergodicity. This completes the proof. □

Lemma A.4. *It holds that*

$$\frac{\left\|\sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\pi_0)+\mathscr{Q}_{GM}(\widehat{\pi}^n))\right\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{GM}(\widehat{\pi}^n)\|} \xrightarrow{\mathbb{P}} 0,\ n\to\infty.$$

Proof. By (4.24) and (4.25), we have that $\sum_{t=1}^{n-r}\Psi((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\widehat{\pi}^n)=0$. Hence, we have

$$\begin{aligned}&\frac{\left\|\sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\pi_0)+\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n))\right\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|} \\ &= \frac{\left\|\sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\pi_0)+\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)-\Psi((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\widehat{\pi}^n))\right\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|}.\end{aligned}$$

By assumption, we have that $\widehat{\pi}^n \xrightarrow{\mathbb{P}} \pi_0$, i.e. it holds that $\mathbb{P}(\widehat{\pi}^n\in C)\to 1$ for $n\to\infty$, where C is the set defined in (A.3). If $\widehat{\pi}^n\in C$ we have that

$$\begin{aligned}&\frac{\left\|\sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\widehat{\pi}^n)+\Psi((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\pi_0)+\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n))\right\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|} \\ &\leq \sup_{\pi\in C}\frac{\left\|\sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\pi)+\Psi((\widetilde{Y}_t,\dots,\widetilde{Y}_{t+r}),\pi_0)+\mathscr{Q}_{\mathrm{GM}}(\pi))\right\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\pi)\|}\end{aligned}$$

and this completes the proof since we now can apply the previous lemma. □

Lemma A.5. *It holds that*

$$\frac{1}{\sqrt{n-r}}\sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0)+\mathscr{Q}_{GM}(\widehat{\pi}^n)) \xrightarrow{\mathbb{P}} 0,\ n\to\infty.$$

Proof. Denote $G_n := \sum_{t=1}^{n-r}(\Psi((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0)+\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n))$. By Lemma 4.30, for $\epsilon>0$ there exists a $M<\infty$ such that $\mathbb{P}(A_n)<\epsilon$ where the set A_n is defined as

$$A_n := \left\{\frac{1}{\sqrt{n-r}}\left\|\sum_{t=1}^{n-r}\Psi((\widetilde{Y}_t,\ldots,\widetilde{Y}_{t+r}),\pi_0)\right\|>M\right\}.$$

Similarly, define the set

$$B_n := \left\{\frac{\|G_n\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|}>\frac{1}{2}\right\}.$$

Now, on $A_n^c\cap B_n^c$ it holds that

$$\begin{aligned}\sqrt{n-r}\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|-M &\le \sqrt{n-r}\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|-\frac{1}{\sqrt{n-r}}\|G_n-(n-r)\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|\\ &\le \frac{\|G_n\|}{\sqrt{n-r}}\le\frac{1+\sqrt{n-r}\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|}{2},\end{aligned}$$

from which we conclude that

$$\sqrt{n-r}\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|\le 2M+1.$$

This inequality in turn implies that

$$\frac{2(1+M)}{1+\sqrt{n-r}\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|}\ge 1.$$

Therefore, we can deduce that

$$\frac{\|G_n\|}{\sqrt{n-r}}\le 2(1+M)\frac{\|G_n\|}{\sqrt{n-r}+(n-r)\|\mathscr{Q}_{\mathrm{GM}}(\widehat{\pi}^n)\|}$$

since we have multiplied by a factor greater than or equal to 1. By the previous lemma, the right–hand side converges to zero for every $\omega\in A_n^c\cap B_n^c$. Hence, by the definition of those sets, $\mathbb{P}\left(\limsup_{n\to\infty}\frac{\|G_n\|}{\sqrt{n-r}}>c\right)\le\epsilon$ for every $c>0$ and the assertion follows. □

Bibliography

[1] Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle". In: *2nd International Symposium on Information Theory*. Ed. by B. N. Petrov and F. Caski. Budapest: Akademiai Kiado, pp. 267–281.

[2] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. 3rd ed. New Jersey: Wiley.

[3] Andresen, A., Benth, F. E., Koekebakker, S. and Zakamulin, V. (2014). "The CARMA interest rate model". In: *Int. J. Theor. Appl. Finance* **17**(2), pp. 1–27.

[4] Andrews, B., Calder, M. and Davis, R. A. (2009). "Maximum likelihood estimation for α-stable autoregressive processes". In: *Ann. Statist.* **37**(4), pp. 1946–1982.

[5] Applebaum, D. (2009). *Lévy Processes and Stochastic Calculus*. 2nd ed. Cambridge Univ. Press.

[6] Balestra, P. and Holly, A. (1990). "A general Kronecker formula for the moments of the multivariate normal distribution". In: *DEEP Cahier No. 9002*.

[7] Barndorff-Nielsen, O. E. and Shephard, N. (2001). "Non–Gaussian Ornstein–Uhlenbeck–based models and some of their uses in financial economics". In: *J. Roy. Statist. Soc. Ser. B* **63**(2), pp. 167–241.

[8] Barndorff-Nielsen, O. E. and Stelzer, R. (2011). "Multivariate supOU processes". In: *Ann. Appl. Probab.* **21**(1), pp. 140–182.

[9] BEATON, A. E. AND TUKEY, J. W. (1974). "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data". In: *Technometrics* **16**(2), pp. 147–185.

[10] BEHRENS, J. (1991). "Robuste Ordnungswahl für autoregressive Prozesse". PhD thesis. Kaiserslautern: Universität Kaiserslautern.

[11] BENTH, F. E. AND ŠALTYTĖ-BENTH, J. (2009). "Dynamic pricing of wind futures". In: *Energ. Econ.* **31**(1), pp. 16–24.

[12] BENTH, F. E., KLÜPPELBERG, C., MÜLLER, G. AND VOS, L. (2014). "Futures pricing in electricity markets based on stable CARMA spot models". In: *Energ. Econ.* **44**, pp. 392–406.

[13] BERGSTROM, A. R. (1990). *Continuous time econometric modelling*. Oxford University Press.

[14] BERNSTEIN, D. S. (2009). *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton Univ. Press.

[15] BERTOIN, J. (1998). *Lévy Processes*. Cambridge Univ. Press.

[16] BOENTE, G., FRAIMAN, R. AND YOHAI, V. J. (1987). "Qualitative robustness for stochastic processes". In: *Ann. Statist.* **15**(3), pp. 1293–1312.

[17] BOLTZMANN, L. (1877). "Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respective den Sätzen über das Wärmegleichgewicht". In: *Wiener Berichte*, pp. 373–435.

[18] BOUBACAR MAÏNASSARA, Y. (2012). "Selection of weak VARMA models by modified Akaike's information criteria". In: *J. Time Ser. Anal.* **33**(1), pp. 121–130.

[19] BOUBACAR MAÏNASSARA, Y. AND FRANCQ, C. (2011). "Estimating structural VARMA models with uncorrelated but non-independent error terms". In: *J. Multivariate Anal.* **102**(3), pp. 496–505.

[20] BOX, G. E., JENKINS, G. M., REINSEL, G. C. AND LJUNG, G. M. (2015). *Time series analysis: forecasting and control*. 5th ed. Wiley.

[21] Bozdogan, H. (1987). "Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions". In: *Psychometrika* **52**(3), pp. 345–370.

[22] Bradley, R. (2007). *Introduction to Strong Mixing Conditions. Volume 1.* Heber City: Kendrick Press.

[23] Brockwell, P. J. (2001). "Lévy-driven CARMA processes". In: *Ann. Inst. Statist. Math.* **53**(1), pp. 113–124.

[24] Brockwell, P. J. (2014). "Recent results in the theory and applications of CARMA processes". In: *Ann. Inst. Statist. Math.* **66**(4), pp. 647–685.

[25] Brockwell, P. J. and Davis, R. (1991). *Time Series: Theory and Methods.* 2nd ed. New York: Springer.

[26] Brockwell, P. J. and Lindner, A. (2009). "Existence and uniqueness of stationary Lévy-driven CARMA processes". In: *Stochastic Process. Appl.* **119**(8), pp. 2660–2681.

[27] Brockwell, P. J. and Marquardt, T. (2005). "Lévy-driven and fractionally integrated ARMA processes with continuous time parameter". In: *Statist. Sinica* **15**(2), pp. 477–494.

[28] Brockwell, P. J. and Schlemm, E. (2013). "Parametric estimation of the driving Lévy process of multivariate CARMA processes from discrete observations". In: *J. Multivariate Anal.* **115**, pp. 217–251.

[29] Brockwell, P. J., Davis, R. A. and Yang, Y. (2011). "Estimation for non-negative Lévy-driven CARMA processes". In: *J. Bus. Econom. Statist.* **29**(2), pp. 250–259.

[30] Burnham, K. and Anderson, D. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach.* 2nd ed. New York: Springer.

[31] Bustos, O. H. (1982). "General M-estimates for contaminated pth-order autoregressive processes: consistency and asymptotic normality". In: *Z. Wahrscheinlichkeit* **59**(4), pp. 491–504.

[32] Bustos, O. H. and Yohai, V. J. (1986). "Robust estimates for ARMA models". In: *J. Amer. Statist. Assoc.* **81**(393), pp. 155–168.

[33] CAVANAUGH, J. E. AND NEATH, A. A. (1999). "Generalizing the derivation of the Schwarz information criterion". In: *Comm. Statist. Theory Methods* **28**(1), pp. 49–66.

[34] CAVANAUGH, J. E. AND SHUMWAY, R. H. (1997). "A bootstrap variant of AIC for state-space model selection". In: *Statist. Sinica* **7**(2), pp. 473–496.

[35] CLAESKENS, G. AND HJORT, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

[36] COX, D. (1981). "Metrics on stochastic processes and qualitative robustness". In: *University of Washington, Dept. of Statistics, Techn. Rep* **3**.

[37] CROUX, C. AND JOOSSENS, K. (2008). "Robust estimation of the vector autoregressive model by a least trimmed squares procedure". In: *COMPSTAT 2008*. Springer, pp. 489–501.

[38] DAVIDSON, J. (1994). *Stochastic Limit Theory*. Advanced Texts in Econometrics. Oxford Univ. Press.

[39] DAVIS, R. A., KNIGHT, K. AND LIU, J. (1992). "M-estimation for autoregressions with infinite variance". In: *Stochastic Process. Appl.* **40**(1), pp. 145–180.

[40] DE LEEUW, J. (1992). "Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle". In: *Breakthroughs in Statistics*. Ed. by S. Kotz and N. L. Johnson. 1. London: Springer, pp. 599–609.

[41] DE LUNA, X. AND GENTON, M. G. (2000). "Robust simulation-based estimation". In: *Statist. Probab. Lett.* **48**(3), pp. 253–259.

[42] DE LUNA, X. AND GENTON, M. G. (2001). "Robust simulation-based estimation of ARMA models". In: *J. Comput. Graph. Stat.* **10**(2), pp. 370–387.

[43] DENBY, L. AND MARTIN, R. D. (1979). "Robust estimation of the first-order autoregressive parameter". In: *J. Amer. Statis. Assoc.* **74**(365), pp. 140–146.

[44] DONOHO, D. L. AND HUBER, P. J. (1983). "The notion of breakdown point". In: *A festschrift for Erich L. Lehmann*. Ed. by P. Bickel, K. Doksum and J. Hodges Jr. Wadsworth, Belmont, CA, pp. 157–184.

[45] DOOB, J. L. (1944). "The Elementary Gaussian Processes". In: *Ann. Math. Stat.* **15**(3), pp. 229–282.

[46] DURRETT, R. (2010). *Probability: Theory and Examples*. 4th ed. Cambridge Univ. Press.

[47] FASEN, V. (2014). "Limit theory for high frequency sampled MCARMA models". In: *Adv. Appl. Probab.* **46**(3), pp. 846–877.

[48] FASEN, V. (2016). "Dependence Estimation for High Frequency Sampled Multivariate CARMA Models". In: *Scand. J. Statist.* **43**(1), pp. 292–320.

[49] FASEN, V. AND FUCHS, F. (2013a). "On the Limit Behavior of the Periodogram of High-Frequency Sampled Stable CARMA Processes". In: *Stochastic Process. Appl.* **123**(1), pp. 229–273.

[50] FASEN, V. AND FUCHS, F. (2013b). "Spectral Estimates for High-Frequency Sampled CARMA Processes". In: *J. Time Series Anal.* **34**(5), pp. 532–551.

[51] FASEN, V. AND KIMMIG, S. (2016+). "Information criteria for multivariate CARMA processes". In: *Accepted for publication in Bernoulli*.

[52] FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. Taylor & Francis.

[53] FINKELSTEIN, H. (1971). "The Law of the Iterated Logarithm for Empirical Distributions". In: *Ann. Math. Statist.* **42**(2), pp. 607–615.

[54] FRANCQ, C. AND ZAKOÏAN, J.-M. (2005). "Recent results for linear time series models with non independent innovations". In: *Statistical modeling and analysis for complex data problems*. Springer, pp. 241–265.

[55] FUCHS, F. AND STELZER, R. (2013). "Mixing conditions for multivariate infinitely divisible processes with an application to mixed moving averages and the supOU stochastic volatility model". In: *ESAIM: Probability and Statistics* **17**, pp. 455–471.

[56] GARCIA BEN, M., MARTINEZ, E. J. AND YOHAI, V. J. (1999). "Robust Estimation in Vector Autoregressive Moving-Average Models". In: *J. Time Ser. Anal.* **20**(4), pp. 381–399.

[57] GARNIER, H. AND WANG, L. (2008). *Identification of continuous-time models from sampled data*. Springer.

[58] GENTON, M. G. AND LUCAS, A. (2003). "Comprehensive definitions of breakdown points for independent and dependent observations". In: *J.R. Stat. Soc. Ser. B Stat. Methodol.* **65**(1), pp. 81–94.

[59] GOURIÉROUX, C. AND MONFORT, A. (1997). *Simulation-based Econometric Methods*. Core Lectures. Oxford University Press, Oxford.

[60] GOURIÉROUX, C., MONFORT, A. AND RENAULT, E. (1993). "Indirect inference". In: *J. Appl. Econometr.* **8**(S1), S85–S118.

[61] GUIDORZI, R. (1975). "Canonical structures in the identification of multivariable systems". In: *Automatica* **11**(4), pp. 361–374.

[62] HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton: Princeton Univ. Press.

[63] HAMPEL, F. R. (1971). "A general qualitative definition of robustness". In: *Ann. Math. Statist.* Pp. 1887–1896.

[64] HAMPEL, F. R. (1974). "The influence curve and its role in robust estimation". In: *J. Amer. Statist. Assoc.* **69**(346), pp. 383–393.

[65] HAMPEL, F. R. (1975). "Beyond location parameters: Robust concepts and methods". In: *Proc. 40th Session I.S.I., Bull. Int. Statist. Inst.* **46**(1), pp. 375–382.

[66] HAMPEL, F. R. (1978). "Modern trends in the theory of robustness". In: *Statistics* **9**(3), pp. 425–442.

[67] HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. AND STAHEL, W. A. (2005). *Robust statistics: the approach based on influence functions*. Wiley.

[68] HANNAN, E. J. AND QUINN, B. G. (1979). "The determination of the order of an autoregression". In: *J. Roy. Statist. Soc. Ser. B* **41**(2), pp. 190–195.

[69] HANNAN, E. (1980). "The estimation of the order of an ARMA process". In: *Ann. Statist.* **8**(5), pp. 1071–1081.

[70] HANNAN, E. AND DEISTLER, M. (2012). *The Statistical Theory of Linear Systems*. Society for Industrial and Applied Mathematics.

[71] Haug, S. and Stelzer, R. (2011). "Multivariate ECOGARCH processes". In: *Econometric Theory* **27**(2), pp. 344–371.

[72] Huber, P. J. (1964). "Robust estimation of a location parameter". In: *Ann. Math. Stat.* **35**(1), pp. 73–101.

[73] Huber, P. J. (1973). "Robust regression: asymptotics, conjectures and Monte Carlo". In: *Ann. Statist.* **1**(5), pp. 799–821.

[74] Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics*. 2nd ed. Wiley.

[75] Hurvich, C. M. and Tsai, C.-L. (1989). "Regression and time series model selection in small samples". In: *Biometrika* **76**(2), pp. 297–307.

[76] Hurvich, C. M. and Tsai, C.-L. (1991). "Bias of the corrected AIC criterion for underfitted regression and time series models". In: *Biometrika* **78**(3), pp. 499–509.

[77] Hurvich, C. M. and Tsai, C.-L. (1993). "A corrected Akaike information criterion for vector autoregressive model selection". In: *J. Time Ser. Anal.* **14**(3), pp. 271–279.

[78] Ibragimov, I. (1962). "Some Limit Theorems for Stationary Processes". In: *Theory Probab. Appl.* **7**(4), pp. 349–382.

[79] Imhof, J. P. (1961). "Computing the Distribution of Quadratic Forms in Normal Variables". In: *Biometrika* **48**(3), pp. 419–426.

[80] Ishiguro, M. and Sakamoto, Y. (1991). "WIC: An estimation-free information criterion". In: *Research Memorandum, Institute of Statistical Mathematics, Tokyo*.

[81] Ishiguro, M., Morita, K.-I. and Ishiguro, M. (1991). "Application of an estimator-free information criterion (WIC) to aperture synthesis imaging". In: *IAU Colloq. 131: Radio Interferometry. Theory, Techniques, and Applications*. Ed. by C. J. Cornwell and R. A. Perley. Vol. 19, pp. 243–247.

[82] Jiang, W. and Turnbull, B. (2004). "The indirect method: inference based on intermediate statistics – a synthesis and examples". In: *Statist. Sci.* **19**(2), pp. 239–263.

[83] JONES, R. H. (1981). "Fitting a continuous time autoregression to discrete data". In: *Applied time series analysis II*. Ed. by D. F. Findley, pp. 651–682.

[84] JONES, R. H. (1985). "Time series analysis with unequally spaced data". In: *Handbook of statistics*. Ed. by E. Hannan, P. Krishnaiah and M. Rao. Vol. 5. Elsevier, pp. 157–178.

[85] KALMAN, R. E. (1960). "A New Approach to Linear Filtering and Prediction Problems". In: *J. Basic Eng.* **82**(1), pp. 35–45.

[86] KLEINER, B. AND MARTIN, R. D. (1979). "Robust estimation of power spectra". In: *J. Roy. Statist. Soc. Ser. B* **41**(3), pp. 313–351.

[87] KONISHI, S. AND KITAGAWA, G. (2008). *Information Criteria and Statistical Modeling*. Springer, New York.

[88] KRASKER, W. S. AND WELSCH, R. E. (1982). "Efficient bounded-influence regression estimation". In: *J. Amer. Statist. Assoc.* **77**(379), pp. 595–604.

[89] KRENGEL, U. (1985). *Ergodic Theorems*. Vol. 6. de Gruyter Stud. Math. Walter de Gruyter & Co., Berlin.

[90] KULLBACK, S. AND LEIBLER, R. A. (1951). "On Information and Sufficiency". In: *Ann. Math. Stat.* **22**(1), pp. 79–86.

[91] KÜNSCH, H (1984). "Infinitesimal robustness for autoregressive processes". In: *Ann. Statist.* **12**(3), pp. 843–863.

[92] LARSSON, E. K. AND SÖDERSTRÖM, T. (2002). "Identification of continuous-time AR processes from unevenly sampled data". In: *Automatica* **38**(4), pp. 709–718.

[93] LARSSON, E. K., MOSSBERG, M. AND SÖDERSTRÖM, T. (2006). "An overview of important practical aspects of continuous-time ARMA system identification". In: *CSSP* **25**(1), pp. 17–46.

[94] LEDOUX, M. AND TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Vol. 23. Springer Science & Business Media.

[95] LENNART, L. AND CAINES, P. E. (1980). "Asymptotic normality of prediction error estimators for approximate system models". In: *Stochastics* **3**(1–4), pp. 29–46.

[96] MACHADO, J. A. (1993). "Robust model selection and M–estimation". In: *Econometric Theory* **9**(3), pp. 478–493.

[97] MALLOWS, C. L. (1975). "On some topics in robustness". In: *Unpublished Bell Labs Technical Memo, Murray Hill, NJ.*

[98] MARONNA, R. A. AND YOHAI, V. J. (1981). "Asymptotic behavior of general M-estimates for regression and scale with random carriers". In: *Z. Wahrscheinlichkeit* **58**(1), pp. 7–20.

[99] MARONNA, R. A. AND YOHAI, V. J. (1991). "The Breakdown Point of Simultaneous General M Estimates of Regression and Scale". In: *J. Amer. Statist. Assoc.* **86**(415), pp. 699–703.

[100] MARONNA, R. A., BUSTOS, O. H. AND YOHAI, V. J. (1979). "Bias-and efficiency-robustness of general M-estimators for regression with random carriers". In: *Smoothing techniques for curve estimation*. Springer, pp. 91–116.

[101] MARONNA, R. A., MARTIN, D. R. AND YOHAI, V. J. (2006). *Robust Statistics*. Wiley.

[102] MARQUARDT, T. AND STELZER, R. (2007). "Multivariate CARMA processes". In: *Stochastic Process. Appl.* **117**(1), pp. 96–120.

[103] MARTIN, D. R. (1980). "Robust estimation of autoregressive models". In: *Directions in time series*. Ed. by D. Brillinger and G. Tiao. Vol. 1. Institute of Mathematical Statistics, Hayward, CA, pp. 228–262.

[104] MARTIN, D. R. AND JONG, J (1977). "Asymptotic properties of robust generalized M-estimates for the first order autoregressive parameter". In: *Unpublished Bell Labs Technical Memo, Murray Hill, NJ.*

[105] MARTIN, D. R. AND YOHAI, V. J. (1985). "Robustness in time series and estimating ARMA models". In: *Handbook of statistics*. Ed. by E. Hannan, P. Krishnaiah and M. Rao. Vol. 5. Elsevier, pp. 119–155.

[106] MARTIN, D. R. AND YOHAI, V. J. (1986). "Influence functionals for time series". In: *Ann. Statist.* **14**(3), pp. 781–818.

[107] MASRELIEZ, C. J. (1975). "Approximate non-Gaussian filtering with linear state and observation relations". In: *IEEE Trans. Automat. Control* **20**(1), pp. 107–110.

[108] MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley Ser. Probab. Stat. New York: Wiley.

[109] MULER, N. AND YOHAI, V. J. (2013). "Robust estimation for vector autoregressive models". In: *Comput. Stat. Data Anal.* **65**, pp. 68–79.

[110] MULER, N., PEÑA, D. AND YOHAI, V. J. (2009). "Robust estimation for ARMA models". In: *Ann. Statist.* **37**(2), pp. 816–840.

[111] OODAIRA, H. AND YOSHIHARA, K.-I. (1971). "The law of the iterated logarithm for stationary processes satisfying mixing conditions". In: *Kodai Math. J.* **23**(3), pp. 311–334.

[112] PAPANTONI-KAZAKOS, P AND GRAY, R. M. (1979). "Robustness of estimators on stationary observations". In: *Ann. Probab.* **7**(6), pp. 989–1002.

[113] RISSANEN, J. (1978). "Modeling by shortest data description". In: *Automatica* **14**(5), pp. 465–471.

[114] RONCHETTI, E. (1997). "Robustness aspects of model choice". In: *Statist. Sinica* **7**(2), pp. 327–338.

[115] ROSENTHAL, J. (2006). *A First Look at Rigorous Probability Theory*. World Scientific.

[116] SATO, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge Univ. Press.

[117] SATO, K.-I. AND YAMAZATO, M. (1983). "Stationary processes of Ornstein-Uhlenbeck type". In: *Teor. Veroyatnost. Mat. Statist.* Pp. 541–551.

[118] SCHLEMM, E. AND STELZER, R. (2012). "Quasi maximum likelihood estimation for strongly mixing state space models and multivariate Lévy-driven CARMA processes". In: *Electron. J. Stat.* **6**, pp. 2185–2234.

[119] SCHLEMM, E. (2011). "Estimation of Continuous-Time ARMA Models and Random Matrices with Dependent Entries". PhD thesis. München: Technische Universität München.

[120] SCHLEMM, E. AND STELZER, R. (2011). "Multivariate CARMA processes, continuous-time state space models and complete regularity of the innovations of the sampled processes". In: *Bernoulli* **18**(1), pp. 46–63.

[121] Scholz, M. (2016). "Estimation of Cointegrated Multivariate Continuous-Time Autoregressive Moving Average Processes". PhD thesis. Karlsruhe: Karlsruhe Institute of Technology.

[122] Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *Ann. Stat.* **6**(2), pp. 461–464.

[123] Shibata, R. (1976). "Selection of the order of an autoregressive model by Akaike's information criterion". In: *Biometrika* **63**(1), pp. 117–126.

[124] Sin, C.-Y. and White, H. (1996). "Information criteria for selecting possibly misspecified parametric models". In: *J. Econometrics* **71**(1), pp. 207–225.

[125] Smith, A. A. (1993). "Estimating nonlinear time-series models using simulated vector autoregressions". In: *J. Appl. Econometr.* **8**(S1), S63–S84.

[126] Stoffer, D. S. and Wall, K. D. (1991). "Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter". In: *J. Amer. Statist. Assoc.* **86**(416), pp. 1024–1033.

[127] Stramer, O, Tweedie, R. L. and Brockwell, P. J. (1996). "Existence and stability of continuous–time threshold ARMA processes". In: *Statist. Sinica* **6**(3), pp. 715–732.

[128] Todorov, V. (2009). "Estimation of continuous–time stochastic volatility models with jumps using high-frequency data". In: *J. Econometrics* **148**(2), pp. 131–148.

[129] Todorov, V. and Tauchen, G. (2006). "Simulation methods for Lévy-driven continuous-time autoregressive moving average (CARMA) stochastic volatility models". In: *Journal of Business & Economic Statistics* **24**(4), pp. 455–469.

[130] Tsai, H. and Chan, K. S. (2003). "A note on parameter differentiation of matrix exponentials, with applications to continuous-time modelling". In: *Bernoulli* **9**(5), pp. 895–919.

[131] Tukey, J. W. (1960). "A survey of sampling from contaminated distributions". In: *Contributions to Probability and Statistics. Essays in honor of Harold Hotelling*. Ed. by I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann, pp. 448–485.

[132] WHITE, H. (1982). "Maximum Likelihood Estimation of Misspecified Models". In: *Econometrica* **50**(1), pp. 1–25.

[133] WHITE, H. (1996). *Estimation, Inference and Specification Analysis*. Cambridge Univ. Press.

[134] YOHAI, V. J. AND MARONNA, R. A. (1979). "Asymptotic behavior of M–estimators for the linear model". In: *Ann. Statist.* **7**(2), pp. 258–268.

[135] ZHULENEV, S. V. (1991). "On the law of the iterated logarithm in the finite-dimensional case". In: *J. Math. Sci. (New York)* **57**(4), pp. 3210–3216.