Bernoulli **23**(4A), 2017, 2860–2886
DOI: 10.3150/16-BEJ830

Information criteria for multivariate CARMA processes

VICKY FASEN and SEBASTIAN KIMMIG*

Institute of Stochastics, Englerstraße 2, D-76131 Karlsruhe, Germany. E-mail: *sebastian.kimmig@kit.edu

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. 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. In this paper, 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.

Keywords: AIC; BIC; CARMA process; consistency; information criteria; law of iterated logarithm; Kronecker index; quasi maximum likelihood estimation

1. Introduction

In this paper, we study necessary and sufficient conditions for weak and strong consistency of information criteria for multivariate continuous-time ARMA(p, q) (MCARMA(p, q)) processes. One-dimensional Gaussian CARMA processes were already investigated by Doob [12] in 1944 and Lévy-driven CARMA processes were propagated at the beginning of this century by Peter Brockwell, see [7] for an overview. 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 [2,4,26], for example. A formal definition of an MCARMA process was recently given in [23]; see Section 2 of this paper. The idea behind it is that for a two-sided $\mathbb{R}^s$-valued Lévy process $L = (L(t))_{t\in\mathbb{R}}$, that is, $L(t) = L(t)\mathbb{1}_{\{t\geq 0\}} - \widetilde{L}(t-)\mathbb{1}_{\{t<0\}}$ 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

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

where D is the differential operator,

$$P(z) := I_{d\times d}z^p + A_1 z^{p-1} + \cdots + A_{p-1}z + A_p \tag{1.2}$$

with $A_1, \dots, A_p \in \mathbb{R}^{d\times d}$ is the autoregressive polynomial and

$$Q(z) := B_0 z^q + B_1 z^{q-1} + \cdots + B_{q-1} z + B_q \tag{1.3}$$

with $B_0, \dots, B_q \in \mathbb{R}^{d\times s}$ is the moving average polynomial. There are a few papers studying the statistical inference of MCARMA processes, for example, [9,13,14,27,28]. In particular, [28] 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 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; in the multivariate setup 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? That is the point where we require model selection criteria or, synonymously, information criteria (cf. [11,22]). The most prominent model selection criteria are the Akaike Information Criterion (AIC) introduced in [1] by Akaike, the Schwarz Information Criterion (SIC), also known as BIC (Bayesian Information Criterion), going back to [29], and the Hannan–Quinn criterion in [19]. The AIC approximates the Kullback–Leibler discrepancy, whereas the BIC approximates the Bayesian a posteriori distribution of the different candidate models. The Hannan–Quinn criterion is based on the AIC of Akaike but with a different penalty term to obtain a strongly consistent information criterion. Information criteria for multivariate ARMAX processes and their statistical inference are well-studied in the monograph [18]; see also [8] for an overview of model selection criteria for ARMA processes. An extension of the AIC to multivariate weak ARMA processes is given in [5]. There exist only a few papers investigating information criteria independent of the underlying model, for example, [31] presents very general likelihood-based information criteria and their properties, and [10] derives 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. They are of the form

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

In our setup $Y^n = (Y(h), \dots, Y(hn))$ is a sample of length n from an MCARMA process, $\widehat{\mathcal{L}}$ 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 $\mathrm{IC}_n(\Theta_1) < \mathrm{IC}_n(\Theta_2)$. 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 more parameters into the model. Without the penalty term, the criterion would always choose the model with more parameters if we compare two parameter spaces both containing a parameter that generates the data. 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. The rest of the paper is structured in the following way. In Section 2 we present basic facts on MCARMA processes and state space models. Since our information criteria are based on quasi maximum likelihood estimation we define first, in Section 3.1, the quasi log-likelihood function for MCARMA processes and in Section 3.2 the model assumptions. Then, in Section 3.3, we derive the asymptotic normality of the QMLE extending the results given in [28] to a misspecified parameter space. For the proof of strong consistency of the information criteria, we require some knowledge about the asymptotic behavior of the quasi log-likelihood function $\widehat{\mathcal{L}}$ as well. For this reason, we prove in Section 3.4 a law of the iterated logarithm for $\widehat{\mathcal{L}}$. Section 4 contains the main results of the paper: necessary and sufficient conditions for strong and weak consistency of information criteria. In particular, we investigate Gaussian MCARMA processes where the results are explicit. Special information criteria are the AIC and the BIC which are the topic of Section 5. Finally, we conclude with a simulation study in Section 6.

Notation

We use the notation $\xrightarrow{\mathcal{D}}$ for weak convergence and $\xrightarrow{\mathbb{P}}$ for convergence in probability. For two random vectors Z_1, Z_2 the notation $Z_1 \stackrel{\mathcal{D}}{=} Z_2$ means equality in distribution. We use as norms the Euclidean norm $\|\cdot\|$ in $\mathbb{R}^d$ and the spectral norm $\|\cdot\|$ for matrices, which is submultiplicative and induced by the Euclidean norm. The matrix $0_{d\times s}$ is the zero matrix in $\mathbb{R}^{d\times s}$ and $I_{d\times d}$ is the identity matrix in $\mathbb{R}^{d\times d}$. For a vector $x \in \mathbb{R}^d$ we write x^T for its transpose. For a matrix $A \in \mathbb{R}^{d\times d}$ we denote by $\mathrm{tr}(A)$ its trace, by $\det(A)$ its determinant and by $\lambda_{\max}(A)$ its largest eigenvalue. If A is symmetric and positive semidefinite we write $A^{\frac{1}{2}}$ for the principal square root, i.e. the symmetric, positive semidefinite matrix satisfying $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$. For a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ we say that X_n is $o_{\text{a.s.}}(a_n)$ if $|X_n/a_n| \to 0$ as $n \to \infty$ $\mathbb{P}$-a.s. and likewise that X_n is $O_{\text{a.s.}}(a_n)$ if $\limsup_{n\to\infty} |X_n/a_n| < \infty$ $\mathbb{P}$-a.s. We write ∂_i for the partial derivative operator with respect to the ith coordinate and $\nabla = (\partial_1, \dots, \partial_r)$ for the gradient operator in $\mathbb{R}^r$. Finally, by $\partial^2_{i,j}$ we denote the second partial derivative with respect to the coordinates i and j, and by $\nabla^2_\vartheta f$ we denote the Hessian matrix of the function f. When there is no ambiguity, we use $\partial_i f(\vartheta_0)$, $\nabla_\vartheta f(\vartheta_0)$ and $\nabla^2_\vartheta f(\vartheta_0)$ as shorthands for $\partial_i f(\vartheta)|_{\vartheta=\vartheta_0}$, $\nabla_\vartheta f(\vartheta)|_{\vartheta=\vartheta_0}$ and $\nabla^2_\vartheta f(\vartheta)|_{\vartheta=\vartheta_0}$, respectively. We interpret $\nabla_\vartheta f(\vartheta)$ as a column vector. In general C denotes a constant which may change from line to line.

2. MCARMA processes and state space processes

We start with a formal definition of an MCARMA process which can be interpreted as solution of (1.1).

Definition 2.1. *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 (1.2) and (1.3) with $p, q \in \mathbb{N}_0$, $q < p$, and $B_0 \neq 0_{d\times s}$.*

Moreover, define

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

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

$$\beta_1 := \cdots := \beta_{p-q-1} := 0_{d\times s} \quad \text{and} \quad \beta_{p-j} := -\sum_{i=1}^{p-j-1} A_i \beta_{p-j-i} + B_{q-j}, \qquad j = 0, \dots, q.$$

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*

$$Y(t) = CX(t) \qquad \text{for } t \in \mathbb{R}, \tag{2.1}$$

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

$$\mathrm{d}X(t) = AX(t)\,\mathrm{d}t + B\,\mathrm{d}L(t). \tag{2.2}$$

In particular, MCARMA$(1,0)$ processes and X in (2.2) are multivariate Ornstein–Uhlenbeck processes. For more details on the well-definedness of MCARMA(p,q) processes see [23]. The class of MCARMA processes is huge. Schlemm and Stelzer [27], Corollary 3.4, showed that the class of continuous-time state space models of the form

$$Y(t) = CX(t) \quad \text{and} \quad \mathrm{d}X(t) = AX(t)\,\mathrm{d}t + B\,\mathrm{d}L(t), \tag{2.3}$$

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 $\mathbb{E}[L(1)] = 0_s$. In general, 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 (2.3) and shortly write $Y = \text{MCARMA}(A, B, C, L)$.

In this paper, we observe the MCARMA process only on a discrete equidistant time-grid with grid distance $h > 0$. It is well known that the Ornstein–Uhlenbeck process $(X(t))_{t\in\mathbb{R}}$ sampled at $h\mathbb{Z}$ is an AR(1)-process with

$$X(kh) = \mathrm{e}^{Ah} X\big((k-1)h\big) + N_{h,k}, \qquad k \in \mathbb{Z},$$

where $N_{h,k} = \int_{(k-1)h}^{kh} \mathrm{e}^{A(kh-t)} B\,\mathrm{d}L(t)$ is a sequence of i.i.d. random vectors. We denote its covariance matrix by $\mathrm{Cov}(N_{h,k}) = \Sigma_h$. Hence, $(Y(kh))_{k\in\mathbb{Z}}$ is the output process of the discrete-time state space model

$$Y(kh) = CX(kh) \qquad \text{where } X(kh) = e^{Ah}X\big((k-1)h\big) + N_{h,k}. \tag{2.4}$$

This discrete-time state space representation is basic for quasi maximum likelihood estimation.

3. Quasi maximum likelihood estimation

3.1. Definition

Since the MCARMA process observed at discrete equidistant time points is a discrete-time state space model as given in (2.4), we use quasi maximum likelihood estimation for discrete-time state space models with respect to identification issues. We now review the most important aspects of estimation as it is done in [28] for MCARMA processes. The estimation is based on the Kalman filter, which calculates the linear innovations of a Gaussian discrete-time state space model; originally introduced in [21] and described in a time series context in [8], Section 12.2. If we observe data, we unfortunately do not know the model parameter behind it and hence, we have to calculate the so-called pseudo-innovations. In the following, we assume that our data set is generated by a continuous-time state space model (A, B, C, L), that is, $Y = \text{MCARMA}(A, B, C, L)$. Moreover, we have a parametric family of MCARMA models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)$ with ϑ in the parameter space $\Theta \subset \mathbb{R}^{N(\Theta)}$, $N(\Theta) \in \mathbb{N}$. The aim is to find $\vartheta_0 \in \Theta$ such that $\text{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$. Therefore, we calculate for every $\vartheta \in \Theta$ the steady-state Kalman gain matrix K_ϑ via the discrete-time Riccati equation

$$\Omega_\vartheta = e^{A_\vartheta h}\Omega_\vartheta e^{A_\vartheta^T h} + \Sigma_{\vartheta,h} - \big(e^{A_\vartheta h}\Omega_\vartheta C_\vartheta^T\big)\big(C_\vartheta \Omega_\vartheta C_\vartheta^T\big)^{-1}\big(e^{A_\vartheta h}\Omega_\vartheta C_\vartheta^T\big)^T,$$

as

$$K_\vartheta = \big(e^{A_\vartheta h}\Omega_\vartheta C_\vartheta^T\big)\big(C_\vartheta \Omega_\vartheta C_\vartheta^T\big)^{-1} \quad \text{and set} \quad V_\vartheta = C_\vartheta \Omega_\vartheta C_\vartheta^T.$$

Based on this, the *pseudo-innovations* are defined as

$$\varepsilon_{\vartheta,k} = Y(kh) - C_\vartheta \widehat{X}_{\vartheta,k} \qquad \text{with} \qquad \widehat{X}_{\vartheta,k} = \big(e^{A_\vartheta h} - K_\vartheta C_\vartheta\big)\widehat{X}_{\vartheta,k-1} + K_\vartheta Y\big((k-1)h\big),$$

where $\widehat{X}_{\vartheta,0} = \sum_{j=1}^\infty (e^{A_\vartheta h} - K_\vartheta C_\vartheta)^{j-1} K_\vartheta Y(-jh)$. For ϑ_0 so that $\text{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ the pseudo-innovations $(\varepsilon_{\vartheta_0,k})_{k\in\mathbb{N}}$ are the innovations, that is, $\varepsilon_{\vartheta_0,k} = Y_k - P_{k-1}Y_k$, where P_k denotes the orthogonal projection onto the space $\overline{\text{span}}\{Y_j : -\infty < j \le k\}$, the closure is taken in L^2 and $V_{\vartheta_0} = \mathbb{E}[\varepsilon_{\vartheta_0,1}\varepsilon_{\vartheta_0,1}^T]$. With this, $-2/n$ times the Gaussian log-likelihood of the model associated to ϑ is

$$\mathcal{L}\big(\vartheta, Y^n\big) = \frac{1}{n}\sum_{k=1}^n \big(d\log(2\pi) + \log\big(\det(V_\vartheta)\big) + \varepsilon_{\vartheta,k}^T V_\vartheta^{-1}\varepsilon_{\vartheta,k}\big) =: \frac{1}{n}\sum_{k=1}^n l_{\vartheta,k}. \tag{3.1}$$

The expectation of this random variable is $\mathcal{Q}(\vartheta) := \mathbb{E}[\mathcal{L}(\vartheta, Y^n)]$. In practical scenarios, it is not possible to calculate the pseudo-innovations, as they are defined in terms of the full history of the process Y but we have only finitely many observations. Suppose now that we have n observations of the output process Y, contained in the sample $Y^n = (Y(h), \dots, Y(nh))$. Therefore, we initialize the filter at $k = 1$ by prescribing $\widehat{X}_{\vartheta,1} = \widehat{X}_{\vartheta,\text{initial}}$ and use the recursion

$$\widehat{\varepsilon}_{\vartheta,k} = Y(kh) - C_\vartheta \widehat{X}_{\vartheta,k} \qquad \text{with} \qquad \widehat{X}_{\vartheta,k} = \left(\mathrm{e}^{A_\vartheta h} - K_\vartheta C_\vartheta\right)\widehat{X}_{\vartheta,k-1} + K_\vartheta Y\left((k-1)h\right).$$

The $\widehat{\varepsilon}_{\vartheta,k}$ are denoted as *approximate pseudo-innovations*. Substituting the approximate pseudo-innovations for their theoretical counterparts in (3.1), we obtain the quasi log-likelihood function

$$\widehat{\mathcal{L}}\left(\vartheta, Y^n\right) := \frac{1}{n}\sum_{k=1}^{n}\left(d\log(2\pi) + \log\left(\det(V_\vartheta)\right) + \widehat{\varepsilon}_{\vartheta,k}^T V_\vartheta^{-1}\widehat{\varepsilon}_{\vartheta,k}\right). \tag{3.2}$$

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

$$\widehat{\vartheta}^n := \underset{\vartheta\in\Theta}{\arg\min}\, \widehat{\mathcal{L}}\left(\vartheta, Y^n\right). \tag{3.3}$$

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

$$\vartheta^* := \underset{\vartheta\in\Theta}{\arg\min}\, \mathcal{Q}(\vartheta). \tag{3.4}$$

The function $\mathcal{Q}$ attains its minimum at ϑ^* in the space Θ. However, if we minimize only over Θ and Θ does not contain a parameter generating Y then it is not clear that the minimum, and hence ϑ^*, is uniquely defined. On the other hand, if there is a $\vartheta_0 \in \Theta$ with $\text{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ then $\vartheta^* = \vartheta_0$. The latter case was investigated in [28].

3.2. Assumptions

In this section, we give the model assumptions which we require for the asymptotic results on the QMLE $\widehat{\vartheta}^n$.

Assumption B.

- B.1 *The parameter space* Θ *is a compact subset of* $\mathbb{R}^{N(\Theta)}$.
- B.2 $\mathbb{E}[L_\vartheta] = 0$, $\mathbb{E}\|L_\vartheta(1)\|^2 < \infty$ *and* $\Sigma_\vartheta^L = \mathbb{E}[L_\vartheta(1)L_\vartheta^T(1)]$ *is non-singular for each* $\vartheta \in \Theta$.
- B.3 *For each* $\vartheta \in \Theta$, *the eigenvalues of* A_ϑ *have strictly negative real parts and are elements of* $\{z \in \mathbb{C} : -\frac{\pi}{h} < \operatorname{Im}(z) < \frac{\pi}{h}\}$.
- B.4 *The pseudo-true parameter* ϑ^* *as defined in* (3.4) *is an element of the interior of* Θ.
- B.5 *For the true Lévy process L there exists a* $\delta > 0$ *such that* $\mathbb{E}\|L(1)\|^{4+\delta} < \infty$.
- B.6 *For every* $\varepsilon > 0$ *there exists a* $\delta(\varepsilon) > 0$ *such that* $\mathcal{Q}(\vartheta^*) \le \min_{\vartheta\in B_\varepsilon(\vartheta^*)^c\cap\Theta} \mathcal{Q}(\vartheta) - \delta(\varepsilon)$, *where* $B_\varepsilon(\vartheta^*)$ *is the open ball with center* ϑ^* *and radius* ε.
- B.7 *The Fisher information matrix of the QMLE is non-singular.*
- B.8 *The maps* $\vartheta \mapsto A_\vartheta$, $\vartheta \mapsto B_\vartheta$, $\vartheta \mapsto C_\vartheta$ *and* $\vartheta \mapsto \Sigma_\vartheta^L$ *are three times continuously differentiable. Moreover, for each* $\vartheta \in \Theta$, *the matrix* C_ϑ *has full rank.*

B.9 *For all* $\vartheta \in \Theta$, *the triple* $(A_\vartheta, B_\vartheta, C_\vartheta)$ *is minimal with McMillan degree* N, *i.e.* $H_\vartheta(z) := C_\vartheta(zI_{N\times N} - A_\vartheta)^{-1}B_\vartheta$ *is minimal in the sense that for any other representation* $H_\vartheta(z) = \widetilde{C}_\vartheta(zI_{\widetilde{N}\times\widetilde{N}} - \widetilde{A}_\vartheta)^{-1}\widetilde{B}_\vartheta$ *the inequality* $\widetilde{N} \geq N$ *holds.*

B.10 *The family* $(\mathrm{MCARMA}(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta))_{\vartheta\in\Theta}$ *is identifiable from the spectral density.*

Remark 3.1.

(a) Two parameter spaces Θ and Θ' both satisfying B.9 with different McMillan degrees generate different processes.

(b) Assumption B.6 is a property called identifiable uniqueness. It makes sure that ϑ^* is the unique minimum of $\mathcal{Q}(\vartheta)$ in Θ (cf. [32], page 28). In the correctly specified case, that is, when the space Θ contains ϑ_0 with $\mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$, the identifiable uniqueness follows from some properties satisfied by the innovations associated to ϑ_0 so that Assumption B.6 can be dropped.

Remark 3.2. An MCARMA process (A, B, C, L) in Echelon form with Kronecker index $m = (m_1, \ldots, m_d)$ has the property that $A = (A_{ij})_{i,j=1,\ldots,d} \in \mathbb{R}^{N\times N}$ is a block matrix with blocks $A_{ij} \in \mathbb{R}^{m_i\times m_j}$ given by

$$A_{ij} = \begin{pmatrix} 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ \vdots & & & & & \vdots \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ \alpha_{ij,1} & \ldots & \alpha_{ij,\min(m_i+\mathbb{1}_{\{i>j\}},m_j)} & 0 & \ldots & 0 \end{pmatrix} + \delta_{i,j}\begin{pmatrix} 0 & & \\ \vdots & I_{(m_i-1)\times(m_i-1)} & \\ 0 & & \\ 0 & \ldots & 0 \end{pmatrix}$$

and

$$C = \begin{pmatrix} 1 & 0 & \ldots & 0 & \vdots & 0 & 0 & \ldots & 0 & \vdots & \vdots & & & \\ & & & & & & & & & & & & 0_{(d-1)\times m_d} & \\ & & & & \vdots & 1 & 0 & \ldots & 0 & \vdots & \vdots & & & \\ & 0_{(d-1)\times m_1} & & & & & & & & & & & & \\ & & & & \vdots & & 0_{(d-2)\times m_2} & & & \vdots & \vdots & 1 & 0 \ \ldots & 0 \end{pmatrix}.$$

The matrix $B = (b_{ij}) \in \mathbb{R}^{N\times s}$ is unrestricted. Moreover, the polynomials $P(z) = [p_{ij}(z)]$ and $Q(z) = [q_{ij}(z)]$ are of the form

$$p_{ij}(z) = \delta_{i,j}z^{m_i} - \sum_{k=1}^{\min(m_i+\mathbb{1}_{\{i>j\}},m_j)} \alpha_{ij,k}z^{k-1} \quad \text{and} \quad q_{ij}(z) = \sum_{k=1}^{m_i} \kappa_{m_1+\cdots+m_{i-1}+k,j}z^{k-1},$$

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

$$T_{ij} = \begin{pmatrix} -\alpha_{ij,2} & \dots & -\alpha_{ij,\min(m_i+\mathbb{1}_{\{i>j\}},m_j)} & 0 & \dots & 0 \\ \vdots & \iddots & & & & \vdots \\ -\alpha_{ij,\min(m_i+\mathbb{1}_{\{i>j\}},m_j)} & & & & & \vdots \\ 0 & & & & & \vdots \\ \vdots & & & & & \vdots \\ 0 & \dots & \dots & & \dots\ \dots & 0 \end{pmatrix}$$

$$+ \delta_{i,j} \begin{pmatrix} 0 & 0 & \dots & \dots & 0 & 1 \\ 0 & 0 & \dots & & 1 & 0 \\ \vdots & \vdots & & \iddots & & \vdots \\ \vdots & & \iddots & & \vdots & \vdots \\ 0 & 1 & & \dots & 0 & 0 \\ 1 & 0 & \dots & \dots & 0 & 0 \end{pmatrix}.$$

This means that the Kronecker index specifies the degrees of the polynomials on the diagonal of the autoregressive polynomial $P(z)$; the polynomials on the secondary line have a degree of at most $\min(m_i + \mathbb{1}_{\{i>j\}}, m_j)$. In particular, we can calculate the degree $p = \max_{i=1,\dots,d} m_i$ of the autoregressive polynomial. Moreover, the polynomials P and Q can be calculated explicitly from A, B and C. Important is that an MCARMA process in Echelon form fulfills the smoothness and identifiability assumptions B.8, B.9 and B.10. A special subclass of MCARMA processes in Echelon form are the one-dimensional CARMA processes, for which the degree p of the autoregressive polynomial is fixed and the zeros of P and Q are distinct. This class corresponds to the class of CARMA processes in Echelon form with Kronecker index p. For more details on MCARMA processes in Echelon form, we refer to [28], Section 4.1.

3.3. Asymptotic normality

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 3.3. *Let Θ with associated family of continuous-time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ be given.*

(a) *Suppose Assumptions B.1 to B.3 as well as B.5 are satisfied. Then, there exists a pseudo-true parameter $\vartheta^* \in \Theta$ as defined in equation (3.4) and for every $n \in \mathbb{N}$, there exists*

$$\vartheta_n^* = \arg\min_{\vartheta\in\Theta} \mathbb{E}\big[\widehat{\mathcal{L}}\big(\vartheta, Y^n\big)\big] \tag{3.5}$$

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) *Suppose Assumptions B.1 to B.9 are satisfied. Then $\widehat{\mathcal{L}}(\vartheta, Y^n) \to \mathcal{Q}(\vartheta)$ $\mathbb{P}$-a.s. holds uniformly in ϑ as $n \to \infty$.*

(c) *Suppose Assumption B is satisfied. Then,*

$$\sqrt{n}\nabla_\vartheta \widehat{\mathcal{L}}\big(\vartheta^*, Y^n\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \mathcal{I}\big(\vartheta^*\big)\big) \quad \text{as} \quad n \to \infty,$$

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

(d) *Suppose Assumptions B.1 to B.9 are satisfied. Then the convergence $\nabla_\vartheta^2 \widehat{\mathcal{L}}(\vartheta, Y^n) \to \mathcal{H}(\vartheta)$ $\mathbb{P}$-a.s. holds uniformly in ϑ as $n \to \infty$, where $\mathcal{H}(\vartheta) := \mathbb{E}[\nabla_\vartheta^2 l_{\vartheta,1}]$ with $l_{\vartheta,1}$ as in (3.1).*

(e) *Suppose Assumption B is satisfied. Then there exist $\varepsilon, \alpha > 0$ such that for almost all ω and for every $n > n_1(\omega)$ and $\vartheta \in B_\varepsilon(\vartheta^*) \cap \Theta$ we have $\det(\nabla_\vartheta^2 \widehat{\mathcal{L}}(\vartheta, Y^n)(\omega)) \geq \alpha$.*

Proof. (a) The existence statements follow directly from [31], Proposition 3.1. As in the proof of [28], Lemma 2.7, we have $\sup_{\vartheta\in\Theta} \mathbb{E}\|\widehat{\varepsilon}_{\vartheta,k}\| < \infty$, $\sup_{\vartheta\in\Theta} \mathbb{E}\|\varepsilon_{\vartheta,k}\| < \infty$ and for some $\rho \in (0, 1)$

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

Hence, the convergence $\vartheta_n^* \to \vartheta^*$ follows.

(b) This is exactly [28], Lemma 2.8, taking [28], Lemma 3.14, into account.

(c) Note that under Assumption B we have

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

Next, we use dominated convergence to interchange expectation and derivative, giving $\mathbb{E}[\nabla_\vartheta \mathcal{L}(\vartheta, Y^n)]|_{\vartheta=\vartheta^*} = 0$. The rest of the proof can be carried out as [28], Lemma 2.16.

(d) The pointwise convergence can be proved as in [28], Lemma 2.17, taking [28], Lemma 3.14, into account, respectively, [6], Lemma 2 and Lemma 3. The stronger statement of uniform convergence can be shown by using the compactness of the parameter space analogous to the proof of [28], Lemma 2.16, respectively [16], Theorem 16.

(e) Assumption B.7 says that the Fisher information matrix $\mathbb{E}[\nabla_\vartheta^2 l_{\vartheta^*,1}]$ is invertible and hence, $\det(\mathbb{E}[\nabla_\vartheta^2 l_{\vartheta^*,1}]) > 0$. Moreover, by Assumption B.8 the map $\vartheta \mapsto \mathbb{E}[\nabla_\vartheta^2 l_{\vartheta,1}]$ is continuous. Thus, there exist $\varepsilon, \alpha > 0$ such that $\inf_{\vartheta\in B_\varepsilon(\vartheta^*)\cap\Theta} \det(\mathbb{E}[\nabla_\vartheta^2 l_{\vartheta,1}]) > \alpha$. Since by (d) as $n \to \infty$,

$$\sup_{\vartheta\in B_\varepsilon(\vartheta^*)\cap\Theta} \big\|\nabla_\vartheta^2 \widehat{\mathcal{L}}\big(\vartheta, Y^n\big) - \mathbb{E}\big[\nabla_\vartheta^2 l_{\vartheta,1}\big]\big\| \to 0 \qquad \mathbb{P}\text{-a.s.},$$

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

We can now state the desired central limit theorem, which basically combines [31], Proposition 4.1, and [28], Theorem 3.16.

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, $\widehat{\vartheta}^n \to \vartheta^*$ $\mathbb{P}$-a.s. as $n\to\infty$, and*

$$\sqrt{n}\big(\widehat{\vartheta}^n - \vartheta^*\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \mathcal{H}^{-1}\big(\vartheta^*\big)\mathcal{I}\big(\vartheta^*\big)\mathcal{H}^{-1}\big(\vartheta^*\big)\big),$$

where

$$\mathcal{I}\big(\vartheta^*\big) = \lim_{n\to\infty} n \operatorname{Var}\big(\nabla_\vartheta \mathcal{L}\big(\vartheta^*, Y^n\big)\big) \quad \textit{and} \quad \mathcal{H}\big(\vartheta^*\big) = \lim_{n\to\infty} \nabla_\vartheta^2 \mathcal{L}\big(\vartheta^*, Y^n\big). \tag{3.6}$$

Proof. The proof can be carried out in the same way as [28], Theorem 3.16, Theorem 2.4 and Theorem 2.5, respectively, replacing ϑ_0 by ϑ^* wherever it appears. Note that we have the additional assumption B.6 concerning identifiable uniqueness, which ensures that the estimator converges to a unique limit, see also [32], Theorem 3.4. □

Remark 3.5.
(a) For the strong consistency part of the theorem, Assumption B.3 can be relaxed requiring only continuity instead of three times differentiability.
(b) In the case that we are in a correctly specified parameter space, this theorem corresponds exactly to [28], Theorem 3.16.
(c) Suppose L is a Brownian motion, then some straightforward but lengthy calculations give $\mathcal{I}(\vartheta^*) = 2\mathcal{H}(\vartheta^*)$; details can be found in [15] as well (cf. [5], Remark 2, for VARMA processes).

3.4. Law of the iterated logarithm

This section is devoted to the development of various forms of the law of the iterated logarithm which we need to study the consistency properties of the information criteria. In the following proposition, we start by establishing a law of the iterated logarithm for linear combinations of partial derivatives of the quasi log-likelihood function.

Proposition 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. Then, for every $x \in \mathbb{R}^{N(\Theta)} \setminus \{0_{N(\Theta)}\}$ it holds $\mathbb{P}$-a.s.*

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

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 B.7. We use the representation $\mathcal{L}(\vartheta, Y^n) = \frac{1}{n}\sum_{k=1}^n l_{\vartheta,k}$ from (3.1). The aim is now to apply the law of the iterated logarithm for dependent random variables as it is given in [24], Theorem 8. Therefore, we need to check the following three conditions:

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

(b) $\mathbb{E}[|x^T \nabla_\vartheta l_{\vartheta^*,k} - \mathbb{E}[x^T \nabla_\vartheta l_{\vartheta^*,k} \mid \sigma(Y((k-m)h), \dots, Y(kh), \dots, Y((k+m)h))]|^2] = O(m^{-2-\delta_2})$ 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 $(Y(kh))_{k\in\mathbb{Z}}$.

These conditions are satisfied by similar arguments as in [28], Lemma 2.16, so that we give only a short sketch: (a) The first statement is already given in the proof of [28], Lemma 2.16. For the second statement, we use the representation given in [28], equation (2.24), that for any $i \in \{1, \dots, N(\Theta)\}$

$$\partial_i l_{\vartheta^*,k} = \operatorname{tr}\big(V_{\vartheta^*}^{-1}\big(I_{d\times d} - \varepsilon_{\vartheta^*,k}\varepsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\big)\partial_i V_{\vartheta^*}\big) + 2\partial_i \varepsilon_{\vartheta^*,k}^T V_{\vartheta^*}^{-1}\varepsilon_{\vartheta^*,k}.$$

Then we obtain with the Cauchy–Schwarz inequality

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

where we have used the compactness of Θ in the last line. From Assumption B.5, 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 $\varepsilon_{\vartheta^*,k}$ and $\partial_i\varepsilon_{\vartheta^*,k}$. With this, we obtain that the right-hand side is finite if $\delta_1 < \frac{\delta}{2}$ and finally, we get $\mathbb{E}|x^T\nabla_\vartheta l_{\vartheta^*,k}|^{2+\delta_1} < \infty$. (b) follows from Step 2 in the proof of [28], Lemma 2.16, and (c) because Y is strongly mixing with geometric rate by [23], Proposition 3.34. Then a consequence of (a)–(c) and [24], Theorem 8, is

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

Since $\log(\log(nx^T\mathcal{I}(\vartheta^*)x)) = O(\log(\log(n)))$ we can therefore deduce the statement for $\mathcal{L}$ by symmetry. Finally, [28], Lemma 2.11 and Lemma 3.14, give $\sqrt{n}\sup_{\vartheta\in\Theta}|\partial_i\widehat{\mathcal{L}}(\vartheta, Y^n) - \partial_i\mathcal{L}(\vartheta, Y^n)| \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ so that we can transfer the result to $\widehat{\mathcal{L}}$ as well. □

The next theorem builds upon this and presents a version of the law of the iterated logarithm for the gradient of the quasi log-likelihood function.

Theorem 3.7. *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 any $\Xi \in \mathbb{R}^{N(\Theta)\times N(\Theta)}$*

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

Proof. An application of Proposition 3.6 gives

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

for every $x \in \mathbb{R}^{N(\Theta)} \setminus \{0_{N(\Theta)}\}$. Just as in the proof of [17], Lemma 2, we can conclude from this that

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

□

Having this theorem allows us to derive a variant of the law of the iterated logarithm for $\widehat{\mathcal{L}}$.

Theorem 3.8. *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*

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

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 $\|\overline{\vartheta}^n - \vartheta^*\| \le \|\widehat{\vartheta}^n - \vartheta^*\|$. Since by Theorem 3.4 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 3.3(e) 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

$$\widehat{\vartheta}^n - \vartheta^* = -(\nabla^2_\vartheta \widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n))^{-1} \nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n) \qquad \mathbb{P}\text{-a.s.} \tag{3.7}$$

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 $\|\check{\vartheta}^n - \widehat{\vartheta}^n\| \le \|\widehat{\vartheta}^n - \vartheta^*\|$, 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 and plugging in (3.7), we arrive at

$$\widehat{\mathcal{L}}(\vartheta^*, Y^n) - \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) = \frac{1}{2}\| \nabla^2_\vartheta \widehat{\mathcal{L}}(\check{\vartheta}^n, Y^n)^{\frac{1}{2}} (\nabla^2_\vartheta \widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n))^{-1} \nabla_\vartheta \widehat{\mathcal{L}}(\vartheta^*, Y^n)\|^2. \tag{3.8}$$

An application of Theorem 3.7 with $\Xi = \mathcal{H}(\vartheta^*)^{-\frac{1}{2}}$ (which is symmetric) yields

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

With $\nabla^2_\vartheta \widehat{\mathcal{L}}(\check{\vartheta}^n, Y^n)^{\frac{1}{2}} \nabla^2_\vartheta \widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n)^{-1} \to \mathcal{H}(\vartheta^*)^{-\frac{1}{2}}$ $\mathbb{P}$-a.s. (cf. Proposition 3.3(d)) and (3.8) we can derive the statement. □

Remark 3.9. This result is an analog to [31], 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 [31] it is only shown that convergence occurs.

4. Likelihood-based information criteria

Throughout the remainder of this paper, we denote by Θ and Θ_0 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 and by $\widehat{\vartheta}^n$ the QMLE based on Y^n in Θ as defined in (3.3). In this main section, we derive properties for likelihood-based information criteria of the following form.

Definition 4.1. *Let $C(n)$ be a positive, nondecreasing function of n with* $\lim_{n\to\infty} C(n)/n = 0$. *Then*

$$\mathrm{IC}_n(\Theta) := \widehat{\mathcal{L}}\big(\widehat{\vartheta}^n, Y^n\big) + N(\Theta)\frac{C(n)}{n}. \tag{4.1}$$

These information criteria have the property that $\mathrm{IC}_n(\Theta) \xrightarrow{\mathbb{P}} \mathcal{Q}(\vartheta^*)$. Since $\mathcal{Q}$ attains its minimum at ϑ_0 for which $\mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ (cf. [28], Lemma 2.10) we choose the parameter space for which the information criterion is minimal. The condition $C(n)/n \to 0$ guarantees that underfitting is not possible, that is, 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, that is, a 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 4.2. *Assume that there is a $\vartheta_0 \in \Theta_0$ with* $\mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$. *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 $F^T F = I_{N(\Theta_0)\times N(\Theta_0)}$ 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 $F^T F = I_{N(\Theta_0)\times N(\Theta_0)}$ guarantees that we have a bijective map from $\Theta_0 \to F\Theta_0 + c \subset \Theta$.

For MCARMA processes parametrized in Echelon form, a parameter space Θ that satisfies Assumption B contains only processes that have the same Kronecker index $m = (m_1, \dots, m_d)$ and hence, fixed degree $p = \max_{i=1,\dots,d} m_i$ of the AR polynomial. However, for the MA polynomial there is only the restriction that the degree is less than or equal to $p - 1$. In this context, 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 for some $q_0 < q \le p_0 - 1$. 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. Let the Kronecker index, 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. Then Θ_0^E is defined as the parameter space generating all MCARMA processes with Kronecker index 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 $\mathrm{MCARMA}(A_{\vartheta_0^E}, B_{\vartheta_0^E}, C_{\vartheta_0^E}, L_{\vartheta_0^E}) = Y$. Note that ϑ_0^E is then the true parameter in Θ_0^E. Next, we define under which circumstances IC_n is consistent; we distinguish two different types of consistency.

Definition 4.3.

(a) *The information criterion* IC_n *is called strongly consistent if for any parameter spaces* Θ_0 *and* Θ *with* $\vartheta_0 \in \Theta_0$ *so that* $\mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$, *and either* $\mathrm{MCARMA}(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta) \neq Y$ *for every* $\vartheta \in \Theta$ *or* Θ_0 *being nested in* Θ *we have*

$$\mathbb{P}\Big(\limsup_{n\to\infty}\big(\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta)\big) < 0\Big) = 1.$$

(b) *The information criterion* IC_n *is called weakly consistent if for any parameter spaces* Θ_0 *and* Θ *with* $\vartheta_0 \in \Theta_0$ *so that* $\mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$, *and either* $\mathrm{MCARMA}(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta) \neq Y$ *for every* $\vartheta \in \Theta$ *or* Θ_0 *being nested in* Θ *we have*

$$\lim_{n\to\infty} \mathbb{P}\big(\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) < 0\big) = 1.$$

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 4.4.

(a) *The criterion* IC_n *is strongly consistent if*

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

The information criterion is not strongly consistent if $\limsup_{n\to\infty} C(n)/\log(\log(n)) = 0$.

(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* $\vartheta_0 \in \Theta_0$ *such that* $\mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ *and* Θ_0 *is nested in* Θ *with map* F. *Moreover, suppose* $\limsup_{n\to\infty} C(n) = C < \infty$. *Define*

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

Then

$$\lim_{n\to\infty} \mathbb{P}\big(\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) > 0\big) = \mathbb{P}\left(\sum_{i=1}^{N(\Theta)-N(\Theta_0)} \lambda_i \chi_i^2 > 2\big[N(\Theta) - N(\Theta_0)\big]C\right) > 0,$$

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

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

Proof. For the whole proof, we denote by ϑ_0 the parameter in Θ_0 with $\mathrm{MCARMA}(A_{\vartheta_0}, B_{\vartheta_0}, C_{\vartheta_0}, L_{\vartheta_0}) = Y$ and by ϑ^* the pseudo-true parameter in Θ. Moreover, denote by $\widehat{\vartheta}_0^n$ the QMLE based on Y^n in Θ_0, by $\widehat{\vartheta}^n$ the QMLE based on Y^n in Θ and by $\widehat{\vartheta}_0^E$ 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: $\mathrm{MCARMA}(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta) \neq Y$ for every $\vartheta \in \Theta$. Then

$$\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta) = \widehat{\mathcal{L}}_0\big(\widehat{\vartheta}_0^n, Y^n\big) - \widehat{\mathcal{L}}\big(\widehat{\vartheta}^n, Y^n\big) + \big[N(\Theta_0) - N(\Theta)\big]\frac{C(n)}{n}. \tag{4.2}$$

On the one hand, by Theorem 3.8 we have that

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

and on the other hand, by Proposition 3.3(b)

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

Finally, in this case $\mathcal{Q}(\vartheta^*) - \mathcal{Q}(\vartheta_0) \geq \delta > 0$ by [28], Lemma 2.10, so that for some $\delta > 0$

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

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 the statement follows.

Case 2: Θ_0 is nested in Θ with map F. Note that Θ_0 and Θ are nested in Θ_0^E as well, implying

$$\widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) = \min_{\vartheta \in \Theta} \widehat{\mathcal{L}}(\vartheta, Y^n) \geq \min_{\vartheta \in \Theta_0^E} \widehat{\mathcal{L}}_E(\vartheta, Y^n) = \widehat{\mathcal{L}}_E(\widehat{\vartheta}_E^n, Y^n). \tag{4.3}$$

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

$$\widehat{\mathcal{L}}_0(\vartheta_0, Y^n) = \widehat{\mathcal{L}}(\vartheta^*, Y^n) = \widehat{\mathcal{L}}_E(\vartheta_0^E, Y^n). \tag{4.4}$$

With this and (4.3), we receive

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

Now, Theorem 3.8 tells us that

$$\begin{aligned}\limsup_{n\to\infty} & \frac{n}{\log(\log(n))}\bigl(\widehat{\mathcal{L}}_E(\vartheta_0^E, Y^n) - \widehat{\mathcal{L}}_E(\widehat{\vartheta}_0^E, Y^n)\bigr) \\ &= \lambda_{\max}\bigl(\mathcal{H}(\vartheta_0^E)^{-\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{H}(\vartheta_0^E)^{-\frac{1}{2}}\bigr) \qquad \mathbb{P}\text{-a.s.}\end{aligned}$$

Turning to the information criterion, this gives

$$\begin{aligned}\limsup_{n\to\infty} & \frac{n}{\log(\log(n))}\bigl(\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta)\bigr) \\ &\leq \limsup_{n\to\infty} \frac{n}{\log(\log(n))}\left(\widehat{\mathcal{L}}_E(\vartheta_0^E, Y^n) - \widehat{\mathcal{L}}_E(\widehat{\vartheta}_0^E, Y^n) + \bigl[N(\Theta_0) - N(\Theta)\bigr]\frac{C(n)}{\log(\log(n))}\right) \\ &\leq \lambda_{\max}\bigl(\mathcal{H}(\vartheta_0^E)^{-\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{H}(\vartheta_0^E)^{-\frac{1}{2}}\bigr) - \limsup_{n\to\infty}\frac{C(n)}{\log(\log(n))} \qquad \mathbb{P}\text{-a.s.,}\end{aligned}$$

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

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

we obtain strong consistency.

Finally, if $\limsup_{n\to\infty} C(n)/\log(\log(n)) = 0$, then from $\widehat{\mathcal{L}}_0(\widehat{\vartheta}_0^n, Y^n) - \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) \geq 0$ it clearly follows that strong consistency cannot hold.

(b) Again we distinguish the two cases from part (a). Case 1 can be 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 $\widehat{\mathcal{L}}(f(\widehat{\vartheta}_0^n), Y^n)$ around $\widehat{\vartheta}^n$ results in

$$\begin{aligned}\widehat{\mathcal{L}}_0(\widehat{\vartheta}_0, Y^n) &= \widehat{\mathcal{L}}(f(\widehat{\vartheta}_0^n), Y^n) \\ &= \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + \frac{1}{2}\bigl(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\bigr)^T \nabla_\vartheta^2 \widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n)\bigl(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\bigr)\end{aligned} \tag{4.5}$$

with $\overline{\vartheta}^n$ such that $\|\overline{\vartheta}^n - \widehat{\vartheta}^n\| \le \|f(\widehat{\vartheta}_0^n) - \widehat{\vartheta}^n\|$. Plugging (4.5) into (4.2) gives

$$\begin{aligned}&\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta)\\ &\quad = \frac{1}{2}\big(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\big)^T \nabla_\vartheta^2 \widehat{\mathcal{L}}\big(\overline{\vartheta}^n, Y^n\big)\big(\widehat{\vartheta}^n - f(\widehat{\vartheta}_0^n)\big) + \big[N(\Theta_0) - N(\Theta)\big]\frac{C(n)}{n}.\end{aligned} \tag{4.6}$$

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 $\widehat{\mathcal{L}}_0(\vartheta, Y^n) = \widehat{\mathcal{L}}(f(\vartheta), Y^n)$ for $\vartheta \in \Theta_0$, so that by the chain rule

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

Moreover, $f(\widehat{\vartheta}_0^n) - \vartheta^* = f(\widehat{\vartheta}_0^n) - f(\vartheta_0) = F(\widehat{\vartheta}_0^n - \vartheta_0)$. As in (3.7), we also have

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

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

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

An application of Proposition 3.3(c) and (d) results in

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

Since by the chain rule $\mathcal{H}(\vartheta_0) = F^T\mathcal{H}(\vartheta^*)F$ 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 (4.6), Proposition 3.3(d) and $C(n) \to \infty$ as $n \to \infty$,

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

Using [20], equation (1.1), gives $\mathbf{N}_F^T\mathcal{H}(\vartheta^*)\mathbf{N}_F \overset{\mathcal{D}}{=} \sum_{i=1}^{N(\Theta)} \lambda_i\chi_i^2$, 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 $\mathcal{H}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_F(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{M}_F(\vartheta^*)\mathcal{H}(\vartheta^*)^{\frac{1}{2}}$. Since $\operatorname{rank}(\mathcal{M}_F(\vartheta^*)) = N(\Theta) - N(\Theta_0)$ and $\mathcal{H}(\vartheta^*)^{\frac{1}{2}}$ and $\mathcal{I}(\vartheta^*)$ have full rank, the number of strictly positive eigenvalues of $\mathcal{H}(\vartheta^*)^{\frac{1}{2}}\mathcal{M}_F(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{M}_F(\vartheta^*)\mathcal{H}(\vartheta^*)^{\frac{1}{2}}$ is $N(\Theta) - N(\Theta_0)$. Hence, the result follows.

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

Remark 4.5.

(a) A conclusion of Theorem 4.4(a) is that strong consistency of the information criterion always holds, independent of the process Y 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 3.8 that

$$\limsup_{n\to\infty} \frac{n}{\log(\log(n))}\big(\mathrm{IC}_n(\Theta_0) - \mathrm{IC}_n(\Theta)\big)$$
$$= \lambda_{\max}\big(\mathcal{M}_F\big(\vartheta^*\big)^{\frac{1}{2}}\mathcal{I}\big(\vartheta^*\big)\mathcal{M}_F\big(\vartheta^*\big)^{\frac{1}{2}}\big) + \limsup_{n\to\infty}\big[N(\Theta_0) - N(\Theta)\big]\frac{C(n)}{\log(\log(n))}.$$

This implies that the information criterion IC_n is strongly consistent iff $\limsup_{n\to\infty} C(n)/\log(\log(n)) > C^*$, where

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

Since the structure of $\mathcal{H}(\vartheta^*)$ 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 4.6).

(c) We would like to note that these results are similar to the statement of [31], 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 3.9 as well).

(d) The proof of Theorem 4.4(a), Case 1, shows that if Θ satisfies MCARMA$(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta) \neq Y$ for every $\vartheta \in \Theta$, then a necessary and sufficient condition for choosing the correct parameter space Θ_0 instead of Θ 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 4.4(c) 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.

Corollary 4.6. *Assume that the Lévy process L which drives the observed process Y is a Brownian motion. Then* IC_n *is strongly consistent iff* $\limsup_{n\to\infty} C(n)/\log(\log(n)) > 2$.

Proof. It is straightforward to construct a space Θ_0 which is nested in Θ_0^E with $N(\Theta_0) = N(\Theta_0^E) - 1$ so that $\lambda_{\max}(\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}\mathcal{I}(\vartheta_0^E)\mathcal{M}_F(\vartheta_0^E)^{\frac{1}{2}}) = 2$; see [15] as well. Additionally, a conclusion of Remark 3.5(c) is that

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

Therefore, the statement follows directly from Theorem 4.4(a) and Remark 4.5(b). □

The results of this section are analogous to the ones obtained for ARMAX processes with i.i.d. noise in [18], Theorem 5.5.1.

5. AIC and BIC

In this section, we transfer the two most well-known information criteria, the AIC and BIC, to the MCARMA framework, highlight the main ideas in their development and apply the results of Section 4 to them. In the following, we assume again that Θ is a parameter space with associated family of continuous-time state space models $(A_\vartheta, B_\vartheta, C_\vartheta, L_\vartheta)_{\vartheta\in\Theta}$ satisfying Assumption B and sequence of QMLE $(\widehat{\vartheta}_n)$ in Θ.

5.1. The Akaike Information Criterion (AIC)

Historically, Akaike's idea was to study the Kullback–Leibler discrepancy of different models and choose the one which minimizes this quantity. In this section, we give arguments why this approach is also reasonable in the case of MCARMA models. As a starting point, let g, f be probability densities on $\mathbb{R}^n$. Then the Kullback–Leibler discrepancy between g and f is

$$K(g \mid f) := \int_{\mathbb{R}^n} f(x) \log\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x = \mathbb{E}_f\bigl[\log(f)\bigr] - \mathbb{E}_f\bigl[\log(g)\bigr] \geq 0.$$

Equality holds only for $g = f$ (cf. [8], page 302). Let now $(f_\vartheta)_{\vartheta\in\Theta}$ be a family of densities on $\mathbb{R}^n$ and fix one "true" density f_{ϑ_0}. With $\mathbb{E}_{\vartheta_0}$ we denote the expectation regarding the distribution with density f_{ϑ_0}. Then, the density that comes closest to f_{ϑ_0} in the Kullback–Leibler sense is given by the one associated to

$$\operatorname*{arg\,min}_{\vartheta\in\Theta} K(f_\vartheta \mid f_{\vartheta_0}) = \operatorname*{arg\,min}_{\vartheta\in\Theta}\bigl\{\mathbb{E}_{\vartheta_0}\bigl[\log(f_{\vartheta_0})\bigr] - \mathbb{E}_{\vartheta_0}\bigl[\log(f_\vartheta)\bigr]\bigr\} = \operatorname*{arg\,min}_{\vartheta\in\Theta}\left\{-\frac{2}{n}\mathbb{E}_{\vartheta_0}\bigl[\log(f_\vartheta)\bigr]\right\}.$$

In our context, f_ϑ denotes the density of the observations Y^n. The problem is that the right-hand side is not directly calculable so that we have to approximate it. To this end, let $\mathcal{Y}^n$ be an independent copy of Y^n and $\widehat{\vartheta}^n(Y^n)$ be the QMLE in Θ based on the observation Y^n. Then we use the approximation

$$\begin{aligned}\min_{\vartheta\in\Theta}\left[-\frac{2}{n}\mathbb{E}_{\vartheta_0}\bigl[\log(f_\vartheta)\bigr]\right] &\approx -\frac{2}{n}\mathbb{E}_{\vartheta_0}\bigl[\log(f_{\widehat{\vartheta}^n(Y^n)}) \mid Y^n\bigr] = -\frac{2}{n}\mathbb{E}\bigl[\log\bigl(f_{\widehat{\vartheta}^n(Y^n)}(\mathcal{Y}^n)\bigr) \mid Y^n\bigr] \\ &\approx \mathbb{E}\bigl[\widehat{\mathcal{L}}\bigl(\widehat{\vartheta}^n(Y^n), \mathcal{Y}^n\bigr) \mid Y^n\bigr].\end{aligned} \tag{5.1}$$

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

Theorem 5.1. *As $n \to \infty$ it holds that*

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

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

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

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

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

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 3.4, as $n \to \infty$,

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

where $\mathcal{N}_1, \mathcal{N}_2$ are independent, $\mathcal{N}(0, \mathcal{H}^{-1}(\vartheta^*)\mathcal{I}(\vartheta^*)\mathcal{H}^{-1}(\vartheta^*))$-distributed random vectors. A conclusion of Proposition 3.3(d) is $\nabla_\vartheta^2 \widehat{\mathcal{L}}(\overline{\vartheta}^n, Y^n) \to \mathcal{H}(\vartheta^*)$ $\mathbb{P}$-a.s. Hence, a continuous mapping theorem gives

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

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

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

The statement follows then since the expectation of the trace is the trace of the expectation. $\square$

As a consequence of (5.1) and Theorem 5.1, we receive the approximation

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

which becomes our information criterion via the following definition:

$$\mathrm{AIC}_n(\Theta) := \widehat{\mathcal{L}}\big(\widehat{\vartheta}^n, Y^n\big) + \frac{\operatorname{tr}(\mathcal{I}(\vartheta^*)\mathcal{H}^{-1}(\vartheta^*))}{n}. \tag{5.2}$$

In general, $\mathcal{I}(\vartheta^*)$ and $\mathcal{H}(\vartheta^*)$ are not known. For practical purposes, they have to be estimated. For both, estimators are known and can be found at the end of [28], Section 2.2, for example.

Remark 5.2. If the Lévy process L which drives the observed process Y is a Brownian motion and $\mathrm{MCARMA}(A_{\vartheta^*}, B_{\vartheta^*}, C_{\vartheta^*}, L_{\vartheta^*}) = Y$, we have $\mathcal{I}(\vartheta^*) = 2\mathcal{H}(\vartheta^*)$ by Remark 3.5(c) and hence, the AIC reduces to $\mathrm{AIC}_n(\Theta) = \widehat{\mathcal{L}}(\widehat{\vartheta}^n, Y^n) + \frac{2N(\Theta)}{n}$; for further details see [15].

The form of the AIC given in this remark coincides with Akaike's original definition (cf. [1]). This suggests to define an alternative version of the AIC, the Classical Akaike Information Criterion (CAIC), as follows:

$$\mathrm{CAIC}_n(\Theta) := \widehat{\mathcal{L}}\big(\widehat{\vartheta}^n, Y^n\big) + \frac{2N(\Theta)}{n}. \tag{5.3}$$

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. For both versions of the AIC, we can immediately make a statement about consistency.

Theorem 5.3. *Both the AIC and the CAIC are neither strongly nor weakly consistent.*

Proof. The CAIC is a special case of IC_n with $C(n) = 2$ such that the assertion follows from Theorem 4.4(b). For the AIC, the proof of Theorem 4.4(b) can directly be adapted. □

5.2. The Bayesian Information Criterion (BIC)

Another information criterion which appears often in the literature is the so-called *Bayesian Information Criterion* (BIC), sometimes also called SIC, an abbreviation for Schwarz Information Criterion, named after the author who originally introduced it in [29]. Another often-cited article in this context is [25], which introduces an equivalent criterion in a slightly different context based on coding theory. As the name Bayesian Information Criterion already suggests, the approach of the definition is based on Bayesian statistics. Our derivation is based on [10], relying on properties of the likelihood function. Suppose that π is a discrete prior probability distribution over the set of candidate spaces Θ and $\pi(\Theta) > 0$ for every parameter space Θ which will be considered. Moreover, suppose that $g(\cdot \mid \Theta)$ is a prior probability distribution over the parameter space Θ. For g, we require the following assumption.

Assumption C. *For every space Θ there exist two constants b and B with $0 < b \le B < \infty$ such that $0 \le g(\vartheta \mid \Theta) \le B$ for all $\vartheta \in \Theta$ and $b \le g(\vartheta \mid \Theta)$ for all ϑ in some neighborhood of the pseudo-true parameter $\vartheta^* \in \Theta$.*

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

$$f\big(\Theta, \vartheta \mid Y^n\big) = \frac{\pi(\Theta) g(\vartheta \mid \Theta) f(Y^n \mid \Theta, \vartheta)}{h(Y^n)}, \tag{5.4}$$

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

$$\mathbb{P}\big(\Theta \mid Y^n\big) = \int_\Theta f\big(\Theta, \vartheta \mid Y^n\big)\,\mathrm{d}\vartheta. \tag{5.5}$$

The idea is to choose the most probable model for the data at hand, that is, the space Θ which maximizes the a posteriori probability. Similar to the derivation of the AIC, the task is now to find a good approximation of (5.5) which is directly calculable from the data. For this note first that maximization of (5.5) is equivalent to minimizing $-2/n$ times the logarithm of $\mathbb{P}(\Theta \mid Y^n)$. Applying this transformation and plugging in (5.4) gives

$$\begin{aligned} &-\frac{2}{n}\log\bigl(\mathbb{P}(\Theta \mid Y^n)\bigr) \\ &\quad = \frac{2}{n}\log\bigl(h\bigl(Y^n\bigr)\bigr) - \frac{2}{n}\log\bigl(\pi(\Theta)\bigr) - \frac{2}{n}\log\biggl(\int_{\Theta} f\bigl(Y^n \mid \Theta, \vartheta\bigr) g(\vartheta \mid \Theta)\,\mathrm{d}\vartheta\biggr). \end{aligned} \tag{5.6}$$

We choose the parameter space Θ with the lowest value of $-\frac{2}{n}\log(\mathbb{P}(\Theta \mid Y^n))$. 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 and use the following theorem.

Theorem 5.4. *Suppose the a priori density g satisfies Assumption C. Then*

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

Proof. By Assumption B, Assumption C, Proposition 3.3 and [31], Proposition 3.1, the regularity assumptions in [10] are satisfied so that the statement follows from there. □

The term $\frac{2}{n}\log(h(Y^n))$ is the same across all parameter spaces and therefore not relevant for model selection. Based on these ideas, we define the BIC:

$$\mathrm{BIC}_n(\Theta) := \widehat{\mathcal{L}}\bigl(\widehat{\vartheta}^n, Y^n\bigr) + N(\Theta)\frac{\log(n)}{n}. \tag{5.7}$$

As with the AIC, we can immediately make a statement about consistency of the BIC by Theorem 4.4(a):

Theorem 5.5. *The BIC is a strongly consistent information criterion.*

6. Simulation study

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

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

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

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

which result in a zero-mean process with covariance matrix

$$\Sigma_{\mathrm{NIG}}^L \approx \begin{pmatrix} 0.4571 & -0.1622 \\ -0.1622 & 0.3708 \end{pmatrix}.$$

In the case of the Brownian motion, the covariance matrix Σ_{BM}^L is equal to the covariance matrix Σ_{NIG}^L 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 (2.2) and then evoking (2.1). 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. After obtaining the discrete samples of the MCARMA process we calculate the AIC, CAIC and BIC as defined in (5.2), (5.3) and (5.7), respectively. In the calculation of the AIC, we estimate the penalty term $\mathrm{tr}(\mathcal{I}(\vartheta^*)\mathcal{H}^{-1}(\vartheta^*))$ by the methods presented in [28], Section 2.2, as well since in general there is no explicit form of $\mathcal{I}(\vartheta^*)$ and $\mathcal{H}(\vartheta^*)$.

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 $\vartheta_0^{(1)} = (-1, -2, 1, -2, -3, 1, 2)$ 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 of 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 4.5(d) is given. Next, we change the true parameter slightly to $\vartheta_0^{(2)} = (-1, -2, 1, -2, -3, 0, 0)$, that is, the data-generating process is now a MCARMA(2, 0) process, while $m_0 = (1, 2)$ remains the same. The results of 50 replications for the true parameter $\vartheta_0^{(2)}$ in space 2 are summarized in Table 1.

As expected because of the strong consistency the BIC performs convincingly and has a high accuracy for both driving Lévy processes. It even achieves a perfect score in the case where

Table 1. Results for the true parameter $\vartheta_0^{(2)}$ in space 2

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	36	42	49	40	46	50
3	$(1,2)$	2	1	10	14	8	1	10	4	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					88%			88%		

the driving noise is a NIG process and makes one wrong decision in the BM scenario. 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 $\operatorname{tr}(\mathcal{I}(\vartheta^*)\mathcal{H}^{-1}(\vartheta^*))$ in the AIC. We realize that in the Gaussian model the estimation error of the penalty term is usually higher for model number 2 than for model 3 (relative to the true values), which results in a higher overfitting rate for the AIC. We also calculate the overfitting probability in the Brownian motion case as given in Theorem 4.4(c). 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 $8/50 = 0.16$ of overfitting in the CAIC is very close.

Finally, 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

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

Here, we consider 7 candidate spaces in total among which there are two parameter spaces in which the true one is nested (spaces 6 and 7); the true parameter space is number 5. We conduct the study with $n = 5000$. The results of 100 repetitions are given in Table 2. The results of this simulation study resemble the ones of the study with $\vartheta_0^{(2)}$ as true parameter – the AIC is the criterion most prone to overfitting, while both the CAIC and BIC perform well. However, the BIC's performance is worse, only slightly outperforming the CAIC in the NIG case and scoring even in the Brownian motion case. The agreement of the AIC and CAIC is now higher in both cases. In light of the CAIC's lesser overfitting rate, it might therefore be reasonable

Table 2. Results for the true parameter $\vartheta_0^{(3)}$ in space 5

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	89	93	93	78	86	87
6	(3, 2)	3	1	17	9	6	6	16	14	13
7	(3, 2)	3	2	19	1	1	1	1	0	0
Agreement						96%			92%	

to prefer the use of the CAIC over the AIC in these scenarios. The approximated overfitting probability of space number 6 for the CAIC is 0.1326, showing that the empirical overfitting rate is very close to the theoretical probability in the NIG case and even a bit lower in the Brownian motion case. The most notable difference to the situation in which $\vartheta_0^{(2)}$ was used is that we now have $n = 5000$ instead of $n = 2000$. This choice was made because we observe that the order selection procedures do not yield sufficiently satisfying results for $\vartheta_0^{(3)}$ when we let $n = 2000$. Upon increasing n to 5000, we observe a decidedly better performance of all the criteria. Because of this, we also expect that the accuracy of the BIC would improve further when more observations are added. These phenomena can be explained by the fact that we now consider larger parameter spaces, containing more parameters. This is a hint that for larger parameter spaces more observations are necessary to obtain the asymptotic results from the theory. Thus, in situations with less observations, alternative information criteria are necessary, for example, based on bootstrap methods (cf. [3,30]), which will be investigated in some future research.

Acknowledgements

Vicky Fasen was supported by the Deutsche Forschungsgemeinschaft through the research grant FA 809/2-2.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) 267–281. Budapest: Akadémiai Kiadó. MR0483125

[2] Applebaum, D. (2009). *Lévy Processes and Stochastic Calculus*, 2nd ed. *Cambridge Studies in Advanced Mathematics* **116**. Cambridge: Cambridge Univ. Press. MR2512800

[3] Bengtsson, T. and Cavanaugh, J.E. (2006). An improved Akaike information criterion for state-space model selection. *Comput. Statist. Data Anal.* **50** 2635–2654. MR2227324

[4] Bertoin, J. (1996). *Lévy Processes*. *Cambridge Tracts in Mathematics* **121**. Cambridge: Cambridge Univ. Press. MR1406564

[5] Boubacar Maïnassara, Y. (2012). Selection of weak VARMA models by modified Akaike's information criteria. *J. Time Series Anal*. **33** 121–130. MR2877612

[6] Boubacar Mainassara, Y. and Francq, C. (2011). Estimating structural VARMA models with uncorrelated but non-independent error terms. *J. Multivariate Anal*. **102** 496–505. MR2755011

[7] Brockwell, P.J. (2014). Recent results in the theory and applications of CARMA processes. *Ann. Inst. Statist. Math*. **66** 647–685. MR3224604

[8] Brockwell, P.J. and Davis, R.A. (1991). *Time Series*: *Theory and Methods*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR1093459

[9] Brockwell, P.J. and Schlemm, E. (2013). Parametric estimation of the driving Lévy process of multivariate CARMA processes from discrete observations. *J. Multivariate Anal*. **115** 217–251. MR3004556

[10] Cavanaugh, J.E. and Neath, A.A. (1999). Generalizing the derivation of the Schwarz information criterion. *Comm. Statist. Theory Methods* **28** 49–66. MR1669504

[11] Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge Univ. Press. MR2431297

[12] Doob, J.L. (1944). The elementary Gaussian processes. *Ann. Math. Stat*. **15** 229–282. MR0010931

[13] Fasen, V. (2014). Limit theory for high frequency sampled MCARMA models. *Adv. in Appl. Probab.* **46** 846–877. MR3254345

[14] Fasen, V. (2016). Dependence estimation for high frequency sampled multivariate CARMA models. *Scand. J. Stat*. **46** 292–320.

[15] Fasen, V. and Kimmig, S. (2015). Information criteria for multivariate CARMA processes. Available at arXiv:1505.00901.

[16] Ferguson, T.S. (1996). *A Course in Large Sample Theory*. *Texts in Statistical Science Series*. London: Chapman & Hall. MR1699953

[17] Finkelstein, H. (1971). The law of the iterated logarithm for empirical distributions. *Ann. Math. Stat*. **42** 607–615. MR0287600

[18] Hannan, E.J. and Deistler, M. (2012). *The Statistical Theory of Linear Systems*. *Classics in Applied Mathematics* **70**. Philadelphia, PA: SIAM. Reprint of the 1988 original [MR0940698]. MR3397291

[19] Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **41** 190–195. MR0547244

[20] Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48** 419–426. MR0137199

[21] Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng*. **82** 35–45.

[22] Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. *Springer Series in Statistics*. New York: Springer. MR2367855

[23] Marquardt, T. and Stelzer, R. (2007). Multivariate CARMA processes. *Stochastic Process. Appl*. **117** 96–120. MR2287105

[24] Oodaira, H. and Yoshihara, K. (1971). The law of the iterated logarithm for stationary processes satisfying mixing conditions. *Kōdai Math. Semin. Rep*. **23** 311–334. MR0307311

[25] Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

[26] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. *Cambridge Studies in Advanced Mathematics* **68**. Cambridge: Cambridge Univ. Press. Translated from the 1990 Japanese original, revised by the author. MR1739520

[27] Schlemm, E. and Stelzer, R. (2012). Multivariate CARMA processes, continuous-time state space models and complete regularity of the innovations of the sampled processes. *Bernoulli* **18** 46–63. MR2888698

[28] Schlemm, E. and Stelzer, R. (2012). Quasi maximum likelihood estimation for strongly mixing state space models and multivariate Lévy-driven CARMA processes. *Electron*. *J*. *Stat*. **6** 2185–2234. MR3020261

[29] Schwarz, G. (1978). Estimating the dimension of a model. *Ann*. *Statist*. **6** 461–464. MR0468014

[30] Shibata, R. (1997). Bootstrap estimate of Kullback–Leibler information for model selection. *Statist*. *Sinica* **7** 375–394. MR1466687

[31] Sin, C.-Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *J*. *Econometrics* **71** 207–225. MR1381082

[32] White, H. (1994). *Estimation*, *Inference and Specification Analysis*. *Econometric Society Monographs* **22**. Cambridge: Cambridge Univ. Press. MR1292251

Received May 2015 and revised January 2016