QUASI MAXIMUM LIKELIHOOD ESTIMATION FOR STRONGLY MIXING STATE SPACE MODELS AND MULTIVARIATE LÉVY-DRIVEN CARMA PROCESSES

SCHLEMM, E. AND STELZER, R.

Abstract. We consider quasi maximum likelihood (QML) estimation for general non-Gaussian discrete-time linear state space models and equidistantly observed multivariate Lévy-driven continuous-time autoregressive moving average (MCARMA) processes. In the discrete-time setting, we prove strong consistency and asymptotic normality of the QML estimator under standard moment assumptions and a strong-mixing condition on the output process of the state space model. In the second part of the paper, we investigate probabilistic and analytical properties of equidistantly sampled continuous-time state space models and apply our results from the discrete-time setting to derive the asymptotic properties of the QML estimator of discretely recorded MCARMA processes. Under natural identifiability conditions, the estimators are again consistent and asymptotically normally distributed for any sampling frequency. We also demonstrate the practical applicability of our method through a simulation study and a data example from econometrics.

1. Introduction

Linear state space models have been used in time series analysis and stochastic modelling for many decades because of their wide applicability and analytical tractability (see, e. g., Brockwell and Davis, 1991; Hamilton, 1994, for a detailed account). In discrete time they are defined by the equations

$$\boldsymbol{X}_n = F\boldsymbol{X}_{n-1} + \boldsymbol{Z}_{n-1}, \quad \boldsymbol{Y}_n = H\boldsymbol{X}_n + \boldsymbol{W}_n, \quad n \in \mathbb{Z}, \tag{1.1}$$

where $\boldsymbol{X} = (\boldsymbol{X}_n)_{n\in\mathbb{Z}}$ is a latent state process, F, H are coefficient matrices and, $\boldsymbol{Z} = (\boldsymbol{Z}_n)_{n\in\mathbb{Z}}$, $\boldsymbol{W} = (\boldsymbol{W}_n)_{n\in\mathbb{Z}}$ are sequences of random variables, see Definition 2.1 for a precise formulation of this model. In this paper we investigate the problem of estimating the coefficient matrices F, H as well as the covariances of $\boldsymbol{Z}$ and $\boldsymbol{W}$ from a sample of observed values of the output process $\boldsymbol{Y} = (\boldsymbol{Y}_n)_{n\in\mathbb{Z}}$, using a quasi maximum likelihood (QML) or generalized least squares approach. Given the importance of this problem in practice, it is surprising that a proper mathematical analysis of the quasi maximum likelihood estimation for the model (1.1) has only been performed in cases where the model is in the so-called innovations form

$$\boldsymbol{X}_n = F\boldsymbol{X}_{n-1} + K\boldsymbol{\varepsilon}_{n-1}, \quad \boldsymbol{Y}_n = H\boldsymbol{X}_n + \boldsymbol{\varepsilon}_n, \quad n \in \mathbb{Z}, \tag{1.2}$$

where the innovations $\boldsymbol{\varepsilon}$ form a martingale difference sequence (Hannan and Deistler, 1988, Chapter 4). This includes state space models in which the noise sequences $\boldsymbol{Z}, \boldsymbol{W}$ are Gaussian, because then the innovations, which are uncorrelated by definition, form an i. i. d. sequence. Restriction to these special cases excludes, however, the state space representations of aggregated linear processes, as well as of equidistantly observed continuous-time linear state space models.

In the first part of the present paper we shall prove consistency and asymptotic normality of the quasi maximum likelihood estimator for the general linear state space model (1.1) under the assumptions that the noise sequences $\boldsymbol{Z}, \boldsymbol{W}$ are ergodic, and that the output process $\boldsymbol{Y}$ satisfies a strong-mixing condition in the sense of Rosenblatt (1956). This assumption is not very restrictive, and is, in particular, satisfied if the noise sequence $\boldsymbol{Z}$ is i. i. d. with an absolutely continuous component, and $\boldsymbol{W}$ is strongly mixing. Our results are a multivariate generalization of Francq and Zakoïan (1998), who considered the quasi maximum likelihood estimation for univariate strongly mixing ARMA processes. The very recent paper Boubacar Mainassara and Francq (2011), which deals with the structural estimation of weak vector ARMA processes, instead makes a mixing assumption about the innovations sequence $\boldsymbol{\varepsilon}$ of the process under consideration, which is very difficult to verify for state space models; their results can therefore not be used for the estimation of general discretely-observed linear continuous-time state space models. More importantly, their proof

2010 *Mathematics Subject Classification.* Primary: 62F10, 62F12, 62M09; Secondary: 60G51, 60G10 .

Key words and phrases. asymptotic normality, linear state space model, multivariate CARMA process, quasi maximum likelihood estimation, strong consistency, strong mixing.

appears to be incomplete, because a crucial step in the proof of their Lemma 4 is claimed by the authors to be analogous to the corresponding step in the proof of Francq and Zakoïan (1998, Lemma 3). It is, however, not clear how the argument given there can be modified in order to be compatible with the assumption of strongly mixing innovations, which is weaker than the assumption of a strongly mixing output process as employed in Francq and Zakoïan (1998).

As alluded to above, one advantage of relaxing the assumption of i. i. d. innovations in a discrete-time state space model is the inclusion of sampled continuous-time state space models. These were introduced in the form of continuous-time ARMA (CARMA) models in Doob (1944) as stochastic processes satisfying the formal analogue of the familiar autoregressive moving average equations of discrete-time ARMA processes, namely

$$a(\mathrm{D})Y(t) = b(\mathrm{D})\mathrm{D}W(t), \quad \mathrm{D} = \mathrm{d}/\mathrm{d}t, \tag{1.3}$$

where a and b are suitable polynomials, and W denotes a Brownian motion. In the recent past, a considerable body of research has been devoted to these processes (see, e. g., Brockwell, 2001a, and references therein). One particularly important extension of the model (1.3) was introduced in Brockwell (2001b), where the driving Brownian motion was replaced by a Lévy process with finite logarithmic moments. This allowed for a wide range of possibly heavy-tailed marginal distribution of the process Y as well as the occurrence of jumps in the sample paths, both characteristic features of many observed time series, e. g. in finance (Cont, 2001). Recently, Marquardt and Stelzer (2007) further generalized Eq. (1.3) to the multivariate setting, which gave researchers the possibility to model several dependent time series jointly by one linear continuous-time process. This extension is important, because many time series, exhibit strong dependencies and can therefore not be modelled adequately on an individual basis. In that paper, the multivariate non-Gaussian equivalent of Eq. (1.3), namely $P(\mathrm{D})\boldsymbol{Y}(t) = Q(\mathrm{D})\mathrm{D}\boldsymbol{L}(t)$, for matrix-valued polynomials P and Q and a Lévy process $\boldsymbol{L}$, was interpreted by spectral techniques as a continuous-time state space model of the form

$$\mathrm{d}\boldsymbol{G}(t) = \mathcal{A}\boldsymbol{G}(t)\mathrm{d}t + \mathcal{B}\mathrm{d}\boldsymbol{L}(t), \quad \boldsymbol{Y}(t) = \mathcal{C}\boldsymbol{G}(t); \tag{1.4}$$

see Eq. (3.4) for an expression of the matrices $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. The structural similarity between Eq. (1.1) and Eq. (1.4) is apparent, and it is essential for many of our arguments. Taking a different route, multivariate CARMA processes can be defined as the continuous-time analogue of discrete-time vector ARMA models, described in detail in Hannan and Deistler (1988); Lütkepohl (2005). As continuous-time processes, CARMA processes are suited particularly well to model irregularly spaced and high-frequency data, which makes them a flexible and efficient tool for building stochastic models of time series arising in the natural sciences, engineering and finance (e. g. Benth and Šaltytė Benth, 2009; Fan et al., 1998; Na and Rhee, 2002; Todorov and Tauchen, 2006).

In the univariate Gaussian setting, several different approaches to the estimation problem of CARMA processes have been investigated (see, e. g., Larsson et al., 2006; Nielsen et al., 2000, and references therein). Maximum likelihood estimation based on a continuous record was considered in Brown and Hewitt (1975); Feigin (1976); Pham (1977). Due to the fact that processes are typically not observed continuously and the limitations of digital computer processing, inference based on discrete observations has become more important in recent years; these approaches include variants of the Yule–Walker algorithm for time-continuous autoregressive processes (Hyndman, 1993), maximum likelihood methods (Brockwell et al., 2011; Duncan et al., 1999), and randomized sampling (Rivoira et al., 2002) to overcome the aliasing problem. Alternative methods include discretization of the differential operator (Larsson and Söderström, 2002; Söderström et al., 1997), and spectral estimation (Gillberg and Ljung, 2009; Lahalle et al., 2004; Lii and Masry, 1995; Masry, 1978). For the special case of Ornstein–Uhlenbeck processes, least squares and moment estimators have also been investigated without the assumptions of Gaussianity (Hu and Long, 2009; Spiliopoulos, 2009).

In the second part of this paper we consider the estimation of general multivariate CARMA processes with finite second moments based on equally spaced discrete observations exploiting the results about the quasi maximum likelihood estimation of general linear discrete-time state space models. Under natural identifiability assumptions we obtain strongly consistent and asymptotically normal estimators for the coefficient matrices of a second-order MCARMA process and the covariance matrix of the driving Lévy process, which determine the second-order structure of the process. It is a natural restriction of the quasi maximum likelihood method that distributional properties of the driving Lévy process which are not determined by its covariance matrix cannot be estimated. However, once the autoregressive and moving average coefficients of a CARMA process are (approximately) known, and if high-frequency observations are available, a parametric model for the driving Lévy process can be estimated by the methods described in Brockwell and Schlemm (2011).

Outline of the paper. The organization of the paper is as follows. In Section 2 we develop a quasi maximum likelihood estimation theory for general non-Gaussian discrete-time linear stochastic state space models with finite second moments. In Section 2.1 we precisely define the class of linear stochastic state space models as well as the quasi maximum likelihood estimator. The following two sections 2.3 and 2.4 contain the proofs that, under a set of technical conditions, this estimator is strongly consistent and asymptotically normally distributed as the number of observations tends to infinity, see Theorems 2.5 and 2.6.

In Section 3 we use the results from Section 2 to establish asymptotic properties of a quasi maximum likelihood estimator for multivariate CARMA processes which are observed on a fixed equidistant time grid. As a first step, we review in Section 3.1 their definition as well as their relation to the class of continuous-time state space models. This is followed by an investigation of the probabilistic properties of a sampled MCARMA process in Section 3.2 and an analysis of the important issue of identifiability in Section 3.3. Finally, we are able to state and prove our main result, Theorem 3.17, about the strong consistency and asymptotic normality of the quasi maximum likelihood estimator for equidistantly sampled multivariate CARMA processes in Section 3.4.

In the final Section 4, we present canonical parametrizations, and we demonstrate the applicability of the quasi maximum likelihood estimation for continuous-time state space models with a simulation study and a data example from economics.

Notation. We use the following notation: The space of $m \times n$ matrices with entries in the ring $\mathbb{K}$ is denoted by $M_{m,n}(\mathbb{K})$ or $M_m(\mathbb{K})$ if $m = n$. The set of symmetric matrices is denoted by $\mathbb{S}_m(\mathbb{K})$, and the symbols $\mathbb{S}_m^+(\mathbb{R})$ ($\mathbb{S}_m^{++}(\mathbb{R})$) stand for the subsets of positive semidefinite (positive definite) matrices, respectively. A^T denotes the transpose of the matrix A, $\operatorname{im} A$ its image, $\ker A$ its kernel, $\sigma(A)$ its spectrum, and $\mathbf{1}_m \in M_m(\mathbb{K})$ is the identity matrix. The vector space $\mathbb{R}^m$ is identified with $M_{m,1}(\mathbb{R})$ so that $\boldsymbol{u} = (u^1, \ldots, u^m)^T \in \mathbb{R}^m$ is a column vector. $\|\cdot\|$ represents the Euclidean norm, $\langle\cdot,\cdot\rangle$ the Euclidean inner product, and $\mathbf{0}_m \in \mathbb{R}^m$ the zero vector. $\mathbb{K}[X]$ ($\mathbb{K}\{X\}$) denotes the ring of polynomial (rational) expressions in X over $\mathbb{K}$, $I_B(\cdot)$ the indicator function of the set B, and $\delta_{n,m}$ the Kronecker symbol. The symbols $\mathbb{E}$, $\mathbb{V}$ar, and $\mathbb{C}$ov stand for the expectation, variance and covariance operators, respectively. Finally, we write ∂_m for the partial derivative operator with respect to the mth coordinate and $\nabla = \begin{pmatrix} \partial_1 & \cdots & \partial_r \end{pmatrix}$ for the gradient operator. When there is no ambiguity, we use $\partial_m f(\boldsymbol{\vartheta}_0)$ and $\nabla_{\boldsymbol{\vartheta}} f(\boldsymbol{\vartheta}_0)$ as shorthands for $\partial_m f(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0}$ and $\nabla_{\boldsymbol{\vartheta}} f(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0}$, respectively. A generic constant, the value of which may change from line to line, is denoted by C.

2. Quasi maximum likelihood estimation for discrete-time state space models

In this section we investigate quasi maximum likelihood (QML) estimation for general linear state space models in discrete time, and prove consistency and asymptotic normality. On the one hand, due to the wide applicability of state space systems in stochastic modelling and control, these results are interesting and useful in their own right. In the present paper they will be applied in Section 3 to prove asymptotic properties of the QML estimator for discretely observed multivariate continuous-time ARMA processes.

Our theory extends existing results from the literature, in particular concerning the QML estimation of Gaussian state space models, of state space models whose innovations sequences are martingale differences (Hannan, 1969, 1975; Reinsel, 1997), and of weak univariate ARMA processes which satisfy a strong mixing condition (Francq and Zakoïan, 1998). The techniques used in this section are similar to Boubacar Mainassara and Francq (2011).

2.1. Preliminaries and definition of the QML estimator.

The general linear stochastic state space model is defined as follows.

Definition 2.1 (State space model)**.** An $\mathbb{R}^d$-valued discrete-time linear stochastic state space model $(F, H, \boldsymbol{Z}, \boldsymbol{W})$ of dimension N is characterized by a strictly stationary $\mathbb{R}^{N+d}$-valued sequence $\begin{pmatrix} \boldsymbol{Z}^T & \boldsymbol{W}^T \end{pmatrix}^T$ with mean zero and finite covariance matrix

$$\mathbb{E}\left[\begin{pmatrix} \boldsymbol{Z}_n \\ \boldsymbol{W}_n \end{pmatrix}\begin{pmatrix} \boldsymbol{Z}_m^T & \boldsymbol{W}_m^T \end{pmatrix}\right] = \delta_{m,n}\begin{pmatrix} Q & R \\ R^T & S \end{pmatrix}, \quad n,m \in \mathbb{Z}, \tag{2.1}$$

for some matrices $Q \in \mathbb{S}_N^+(\mathbb{R})$, $S \in \mathbb{S}_d^+(\mathbb{R})$, and $R \in M_{N,d}(\mathbb{R})$; a state transition matrix $F \in M_N(\mathbb{R})$; and an observation matrix $H \in M_{d,N}(\mathbb{R})$. It consists of a state equation

$$\boldsymbol{X}_n = F\boldsymbol{X}_{n-1} + \boldsymbol{Z}_{n-1}, \quad n \in \mathbb{Z}, \tag{2.2a}$$

and an observation equation

$$\boldsymbol{Y}_n = H\boldsymbol{X}_n + \boldsymbol{W}_n, \quad n \in \mathbb{Z}. \tag{2.2b}$$

The $\mathbb{R}^N$-valued autoregressive process $\boldsymbol{X} = (\boldsymbol{X}_n)_{n\in\mathbb{Z}}$ is called the *state vector process*, and $\boldsymbol{Y} = (\boldsymbol{Y}_n)_{n\in\mathbb{Z}}$ is called the *output process*.

The assumption that the processes $\boldsymbol{Z}$ and $\boldsymbol{W}$ are centred is not essential for our results, but simplifies the notation considerably. The following lemma collects important probabilistic properties of the output process $\boldsymbol{Y}$ of such a state space model. Its proof is standard (Brockwell and Davis, 1991, §12.1).

Lemma 2.1. *Assume that $(F, H, \boldsymbol{Z}, \boldsymbol{W})$ is a state space model according to Definition 2.1, and that the eigenvalues of F are less than unity in absolute value.*

i) There exists a unique stationary process $\boldsymbol{Y}$ solving Eqs. (2.2). This process has the moving average representation $\boldsymbol{Y}_n = \boldsymbol{W}_n + H\sum_{\nu=1}^{\infty} F^{\nu-1}\boldsymbol{Z}_{n-\nu}$.

ii) If both $\boldsymbol{Z}$ and $\boldsymbol{W}$ have a finite kth moments for some $k > 0$, then $\boldsymbol{Y}$ has finite kth moments as well.

iii) If the expected value of $\boldsymbol{Y}_n$ is finite, it is given by $\mathbb{E}\boldsymbol{Y}_n = \mathbb{E}\boldsymbol{W}_1 + H\sum_{\nu=1}^{\infty} F^{\nu-1}\mathbb{E}\boldsymbol{Z}_1$. In particular, if both $\boldsymbol{Z}$ and $\boldsymbol{W}$ have mean zero, then $\boldsymbol{Y}$ has mean zero as well.

Before we turn our attention to the estimation problem for this class of state space models, we review the necessary aspects of the theory of Kalman filtering, see Kalman (1960) for the original control-theoretic account and Brockwell and Davis (1991, §12.2) for a treatment in the context of time series analysis. The linear innovations of the output process $\boldsymbol{Y}$ are of particular importance for the quasi maximum likelihood estimation of state space models.

Definition 2.2 (Linear innovations)**.** Let $\boldsymbol{Y} = (\boldsymbol{Y}_n)_{n\in\mathbb{Z}}$ be an $\mathbb{R}^d$-valued stationary stochastic process with finite second moments. The *linear innovations* $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_n)_{n\in\mathbb{Z}}$ of $\boldsymbol{Y}$ are then defined by

$$\boldsymbol{\varepsilon}_n = \boldsymbol{Y}_n - P_{n-1}\boldsymbol{Y}_n, \quad P_n = \text{orthogonal projection onto } \overline{\operatorname{span}}\left\{\boldsymbol{Y}_\nu : -\infty < \nu \leqslant n\right\}, \tag{2.3}$$

where the closure is taken in the Hilbert space of square-integrable random variables with inner product $(X, Y) \mapsto \mathbb{E}\langle X, Y\rangle$.

This definition immediately implies that the innovations $\boldsymbol{\varepsilon}$ of a stationary stochastic process $\boldsymbol{Y}$ are stationary and uncorrelated. The following proposition is a combination of Brockwell and Davis (1991, Proposition 12.2.3) and Hamilton (1994, Proposition 13.2).

Proposition 2.2. *Assume that $\boldsymbol{Y}$ is the output process of the state space model (2.2), that at least one of the matrices Q and S is positive definite, and that the absolute values of the eigenvalues of F are less than unity. Then the following hold.*

i) The discrete-time algebraic Riccati equation

$$\Omega = F\Omega F^T + Q - \left[F\Omega H^T + R\right]\left[H\Omega H^T + S\right]^{-1}\left[F\Omega H^T + R\right]^T \tag{2.4}$$

has a unique positive semidefinite solution $\Omega \in \mathbb{S}_N^+(\mathbb{R})$.

ii) The absolute values of the eigenvalues of the matrix $F - KH \in M_N(\mathbb{R})$ are less than one, where

$$K = \left[F\Omega H^T + R\right]\left[H\Omega H^T + S\right]^{-1} \in M_{N,d}(\mathbb{R}) \tag{2.5}$$

is the steady-state Kalman gain matrix.

iii) The linear innovations $\boldsymbol{\varepsilon}$ *of* $\boldsymbol{Y}$ *are the unique stationary solution to*

$$\hat{\boldsymbol{X}}_n = (F - KH)\,\hat{\boldsymbol{X}}_{n-1} + K\boldsymbol{Y}_{n-1}, \quad \boldsymbol{\varepsilon}_n = \boldsymbol{Y}_n - H\hat{\boldsymbol{X}}_n, \quad n \in \mathbb{Z}. \tag{2.6a}$$

Using the backshift operator B*, which is defined by* $\mathrm{B}\,\boldsymbol{Y}_n = \boldsymbol{Y}_{n-1}$*, this can be written equivalently as*

$$\boldsymbol{\varepsilon}_n = \left\{\mathbf{1}_d - H\left[\mathbf{1}_N - (F-KH)\,\mathrm{B}\right]^{-1} K\,\mathrm{B}\right\}\boldsymbol{Y}_n = \boldsymbol{Y}_n - H\sum_{\nu=1}^{\infty}(F-KH)^{\nu-1}K\boldsymbol{Y}_{n-\nu}. \tag{2.6b}$$

The covariance matrix $V = \mathbb{E}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^T \in \mathbb{S}_d^+(\mathbb{R})$ *of the innovations* $\boldsymbol{\varepsilon}$ *is given by*

$$V = \mathbb{E}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^T = H\Omega H^T + S. \tag{2.7}$$

iv) The process $\boldsymbol{Y}$ *has the innovations representation*

$$\hat{\boldsymbol{X}}_n = F\boldsymbol{X}_{n-1} + K\boldsymbol{\varepsilon}_{n-1}, \quad \boldsymbol{Y}_n = H\boldsymbol{X}_n + \boldsymbol{\varepsilon}_n, \quad n \in \mathbb{Z}, \tag{2.8a}$$

which, similar to Eqs. (2.6), allows for the moving average representation

$$\boldsymbol{Y}_n = \left\{\mathbf{1}_d - H\left[\mathbf{1}_N - F\,\mathrm{B}\right]^{-1} K\,\mathrm{B}\right\}\boldsymbol{Y}_n = \boldsymbol{\varepsilon}_n + H\sum_{\nu=1}^{\infty}F^{\nu-1}K\boldsymbol{\varepsilon}_{n-\nu}, \quad n \in \mathbb{Z}. \tag{2.8b}$$

We now consider parametric families of state space models. For some parameter space $\Theta \subset \mathbb{R}^r$, $r \in \mathbb{N}$, the mappings

$$F_{(\cdot)} : \Theta \to M_N(\mathbb{R}), \qquad H_{(\cdot)} : \Theta \to M_{d,N}, \tag{2.9a}$$

together with a collection of strictly stationary stochastic processes $\boldsymbol{Z}_{\boldsymbol{\vartheta}}$, $\boldsymbol{W}_{\boldsymbol{\vartheta}}$, $\boldsymbol{\vartheta} \in \Theta$, with finite second moments determine a parametric family $(F_{\boldsymbol{\vartheta}}, H_{\boldsymbol{\vartheta}}, \boldsymbol{Z}_{\boldsymbol{\vartheta}}, \boldsymbol{W}_{\boldsymbol{\vartheta}})_{\boldsymbol{\vartheta}\in\Theta}$ of linear state space models according to Definition 2.1. For the variance and covariance matrices of the noise sequences $\boldsymbol{Z}, \boldsymbol{W}$ we use the notation (cf. Eq. (2.1)) $Q_{\boldsymbol{\vartheta}} = \mathbb{E}\boldsymbol{Z}_{\boldsymbol{\vartheta},n}\boldsymbol{Z}_{\boldsymbol{\vartheta},n}^T$, $S_{\boldsymbol{\vartheta}} = \mathbb{E}\boldsymbol{W}_{\boldsymbol{\vartheta},n}\boldsymbol{W}_{\boldsymbol{\vartheta},n}^T$, and $R_{\boldsymbol{\vartheta}} = \mathbb{E}\boldsymbol{Z}_{\boldsymbol{\vartheta},n}\boldsymbol{W}_{\boldsymbol{\vartheta},n}^T$, which defines the functions

$$Q_{(\cdot)} : \Theta \to \mathbb{S}_N^+(\mathbb{R}), \qquad S_{(\cdot)} : \Theta \to \mathbb{S}_d^+, \qquad R_{(\cdot)} : \Theta \to M_{N,d}(\mathbb{R}). \tag{2.9b}$$

It is well known (Brockwell and Davis, 1991, Eq. (11.5.4)) that for this model, minus twice the logarithm of the Gaussian likelihood of $\boldsymbol{\vartheta}$ based on a sample $\boldsymbol{y}^L = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_L)$ of observations can be written as

$$\mathscr{L}(\boldsymbol{\vartheta}, \boldsymbol{y}^L) = \sum_{n=1}^{L} l_{\boldsymbol{\vartheta},n} = \sum_{n=1}^{L}\left[d\log 2\pi + \log\det V_{\boldsymbol{\vartheta}} + \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right], \tag{2.10}$$

where $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}$ and $V_{\boldsymbol{\vartheta}}$ are given by analogues of Eqs. (2.6a) and (2.7), namely

$$\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} = \left\{\mathbf{1}_d - H_{\boldsymbol{\vartheta}}\left[\mathbf{1}_N - (F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}}H_{\boldsymbol{\vartheta}})\,\mathrm{B}\right]^{-1} K_{\boldsymbol{\vartheta}}\,\mathrm{B}\right\}\boldsymbol{Y}_n, \quad n \in \mathbb{Z}, \qquad V_{\boldsymbol{\vartheta}} = H_{\boldsymbol{\vartheta}}\Omega_{\boldsymbol{\vartheta}}H_{\boldsymbol{\vartheta}}^T + S_{\boldsymbol{\vartheta}}, \tag{2.11}$$

and $K_{\boldsymbol{\vartheta}}, \Omega_{\boldsymbol{\vartheta}}$ are defined in the same way as K, Ω in Eqs. (2.4) and (2.5). In the following we always assume that $\boldsymbol{y}^L = (\boldsymbol{Y}_{\boldsymbol{\vartheta}_0,1}, \ldots, \boldsymbol{Y}_{\boldsymbol{\vartheta}_0,L})$ is a sample from the output process of the state space model $(F_{\boldsymbol{\vartheta}_0}, H_{\boldsymbol{\vartheta}_0}, \boldsymbol{Z}_{\boldsymbol{\vartheta}_0}, \boldsymbol{W}_{\boldsymbol{\vartheta}_0})$ corresponding to the parameter value $\boldsymbol{\vartheta}_0$. We therefore call $\boldsymbol{\vartheta}_0$ the *true parameter value*. It is important to note that $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0}$ are the true innovations of $\boldsymbol{Y}_{\boldsymbol{\vartheta}_0}$, and that therefore $\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}^T = V_{\boldsymbol{\vartheta}_0}$, but that this relation fails to hold for other values of $\boldsymbol{\vartheta}$. This is due to the fact that $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ is not the true innovations sequence of the state space model corresponding to the parameter value $\boldsymbol{\vartheta}$. We therefore call the sequence $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ *pseudo-innovations*.

The goal of this section is to investigate how the value $\boldsymbol{\vartheta}_0$ can be estimated from $\boldsymbol{y}^L$ by maximizing Eq. (2.10). The first difficulty one is confronted with is that the pseudo-innovations $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ are defined in terms of the full history of the process $\boldsymbol{Y} = \boldsymbol{Y}_{\boldsymbol{\vartheta}_0}$, which is not observed. It is therefore necessary to use an approximation to these innovations which can be computed from the finite sample $\boldsymbol{y}^L$. One such approximation is obtained if, instead of using the steady-state Kalman filter described in Proposition 2.2, one initializes the filter at $n = 1$ with some prescribed values. More precisely, we define the approximate pseudo-innovations $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta}}$ via the recursion

$$\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},n} = (F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}}H_{\boldsymbol{\vartheta}})\,\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},n-1} + K_{\boldsymbol{\vartheta}}\boldsymbol{Y}_{n-1}, \quad \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n} = \boldsymbol{Y}_n - H_{\boldsymbol{\vartheta}}\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},n}, \quad n \in \mathbb{N}, \tag{2.12}$$

and the prescription $\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},1} = \hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},\mathrm{initial}}$. The initial values $\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},\mathrm{initial}}$ are usually either sampled from the stationary distribution of $\boldsymbol{X}_{\boldsymbol{\vartheta}}$, if that is possible, or set to some deterministic value. Alternatively, one can additionally define a positive semidefinite matrix $\Omega_{\boldsymbol{\vartheta},\mathrm{initial}}$ and compute Kalman gain matrices $K_{\boldsymbol{\vartheta},n}$ recursively via Brockwell and Davis (1991, Eq. (12.2.6)). While this procedure might be advantageous for small sample sizes, the computational burden is significantly smaller when the steady-state Kalman gain is used. The asymptotic properties which we are dealing with in this paper are expected to be the same for both choices because the Kalman gain matrices $K_{\vartheta,n}$ converge to their steady state values as n tends to infinity (Hamilton, 1994, Proposition 13.2).

The quasi maximum likelihood estimator $\hat{\vartheta}^L$ for the parameter ϑ based on the sample y^L is defined as

$$\hat{\vartheta}^L = \operatorname{argmin}_{\vartheta\in\Theta} \widehat{\mathscr{L}}(\vartheta, y^L), \tag{2.13}$$

where $\widehat{\mathscr{L}}(\vartheta, y^L)$ is obtained from $\mathscr{L}(\vartheta, y^L)$ by substituting $\hat{\varepsilon}_{\vartheta,n}$ from Eq. (2.12) for $\varepsilon_{\vartheta,n}$, that is

$$\widehat{\mathscr{L}}(\vartheta, y^L) = \sum_{n=1}^{L} \hat{l}_{\vartheta,n} = \sum_{n=1}^{L} \left[d \log 2\pi + \log\det V_\vartheta + \hat{\varepsilon}_{\vartheta,n}^T V_\vartheta^{-1} \hat{\varepsilon}_{\vartheta,n}\right]. \tag{2.14}$$

2.2. **Technical assumptions and main results.** Our main results about the quasi maximum likelihood estimation for discrete-time state space models are Theorem 2.5, stating that the estimator $\hat{\vartheta}^L$ given by Eq. (2.13) is strongly consistent, which means that $\hat{\vartheta}^L$ converges to ϑ_0 almost surely, and Theorem 2.6, which asserts the asymptotic normality of $\hat{\vartheta}^L$ with the usual $L^{1/2}$ scaling. In order to prove these results, we need to impose the following conditions.

Assumption D1. The parameter space Θ is a compact subset of $\mathbb{R}^r$.

Assumption D2. The mappings $F_{(\cdot)}$, $H_{(\cdot)}$, $Q_{(\cdot)}$, $S_{(\cdot)}$, and $R_{(\cdot)}$ in Eqs. (2.9) are continuous.

The next condition guarantees that the models under consideration describe stationary processes.

Assumption D3. For every $\vartheta \in \Theta$, the following hold:

i) the eigenvalues of F_ϑ have absolute values less than unity,
ii) at least one of the two matrices Q_ϑ and S_ϑ is positive definite,
iii) the matrix V_ϑ is non-singular.

The next lemma shows that the assertions of Assumption D3 hold in fact uniformly in ϑ.

Lemma 2.3. *Suppose that Assumptions D1 to D3 are satisfied. Then the following hold.*

i) There exists a positive number $\rho < 1$ such that, for all $\vartheta \in \Theta$, it holds that

$$\max\{|\lambda| : \lambda \in \sigma(F_\vartheta)\} \leqslant \rho. \tag{2.15a}$$

ii) There exists a positive number $\rho < 1$ such that, for all $\vartheta \in \Theta$, it holds that

$$\max\{|\lambda| : \lambda \in \sigma(F_\vartheta - K_\vartheta H_\vartheta)\} \leqslant \rho, \tag{2.15b}$$

where K_ϑ is defined by Eqs. (2.4) and (2.5).

iii) There exists a positive number C such that $\left\|V_\vartheta^{-1}\right\| \leqslant C$ for all ϑ.

Proof. Assertion i) is a direct consequence of Assumption D3, i), the assumed smoothness of $\vartheta \mapsto F_\vartheta$ (Assumption D2), the compactness of Θ (Assumption D1) and the fact (Bernstein, 2005, Fact 10.11.2) that the eigenvalues of a matrix are continuous functions of its entries. The claim ii) follows with the same argument from Proposition 2.2, ii) and the fact that the solution of a discrete-time algebraic Riccati equation is a continuous function of the coefficient matrices (Lancaster and Rodman, 1995, Chapter 14),(Sun, 1998). Moreover, by Eq. (2.7) and what was already said, the function $\vartheta \mapsto V_\vartheta$ is continuous, which shows that Assumption D3, iii) holds uniformly in ϑ as well, and so iii) is proved. □

For the following assumption about the noise sequences Z and W we use the usual notion of ergodicity (see, e. g., Durrett, 2010, Chapter 6).

Assumption D4. The process $\begin{pmatrix} W_{\vartheta_0}^T & Z_{\vartheta_0}^T \end{pmatrix}^T$ is ergodic.

The assumption that the processes Z_{ϑ_0} and W_{ϑ_0} are ergodic implies via the moving average representation in Lemma 2.1, i) and Krengel (1985, Theorem 4.3) that the output process $Y = Y_{\vartheta_0}$ is ergodic. As a consequence, the pseudo-innovations ε_ϑ defined in Eq. (2.11) are ergodic for every $\vartheta \in \Theta$.

Our first identifiability assumption precludes redundancies in the parametrization of the state space models under consideration and is therefore necessary for the true parameter value ϑ_0 to be estimated consistently. It will be used in Lemma 2.11 to show that the quasi likelihood function given by Eq. (2.14) asymptotically has a unique global minimum at $\boldsymbol{\vartheta}_0$.

Assumption D5. For all $\boldsymbol{\vartheta}_0 \neq \boldsymbol{\vartheta} \in \Theta$, there exists a $z \in \mathbb{C}$ such that

$$H_{\boldsymbol{\vartheta}} \left[\mathbf{1}_N - (F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}) z\right]^{-1} K_{\boldsymbol{\vartheta}} \neq H_{\boldsymbol{\vartheta}_0} \left[\mathbf{1}_N - (F_{\boldsymbol{\vartheta}_0} - K_{\boldsymbol{\vartheta}_0} H_{\boldsymbol{\vartheta}_0}) z\right]^{-1} K_{\boldsymbol{\vartheta}_0}, \tag{2.16a}$$

or

$$V_{\boldsymbol{\vartheta}} \neq V_{\boldsymbol{\vartheta}_0}. \tag{2.16b}$$

Assumption D5 can be rephrased in terms of the spectral densities $f_{\boldsymbol{Y}_{\boldsymbol{\vartheta}}}$ of the output processes $\boldsymbol{Y}_{\boldsymbol{\vartheta}}$ of the state space models $(F_{\boldsymbol{\vartheta}}, H_{\boldsymbol{\vartheta}}, \boldsymbol{Z}_{\boldsymbol{\vartheta}}, \boldsymbol{W}_{\boldsymbol{\vartheta}})$. This characterization will be very useful when we apply the estimation theory developed in this section to state space models that arise from sampling a continuous-time ARMA process.

Lemma 2.4. *If, for all $\boldsymbol{\vartheta}_0 \neq \boldsymbol{\vartheta} \in \Theta$, there exists an $\omega \in [-\pi, \pi]$ such that $f_{\boldsymbol{Y}_{\boldsymbol{\vartheta}}}(\omega) \neq f_{\boldsymbol{Y}_{\boldsymbol{\vartheta}_0}}(\omega)$, then Assumption D5 holds.*

Proof. We recall from Hamilton (1994, Eq. (10.4.43)) that the spectral density $f_{\boldsymbol{Y}_{\boldsymbol{\vartheta}}}$ of the output process $\boldsymbol{Y}_{\boldsymbol{\vartheta}}$ of the state space model $(F_{\boldsymbol{\vartheta}}, H_{\boldsymbol{\vartheta}}, \boldsymbol{Z}_{\boldsymbol{\vartheta}}, \boldsymbol{W}_{\boldsymbol{\vartheta}})$ is given by the expression $f_{\boldsymbol{Y}_{\boldsymbol{\vartheta}}}(\omega) = (2\pi)^{-1} \mathscr{H}_{\boldsymbol{\vartheta}}\left(\mathrm{e}^{\mathrm{i}\omega}\right) V_{\boldsymbol{\vartheta}} \mathscr{H}_{\boldsymbol{\vartheta}}\left(\mathrm{e}^{-\mathrm{i}\omega}\right)^T$, $\omega \in [-\pi, \pi]$, where $\mathscr{H}_{\boldsymbol{\vartheta}}(z) \coloneqq H_{\boldsymbol{\vartheta}} \left[\mathbf{1}_N - (F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}) z\right]^{-1} K_{\boldsymbol{\vartheta}} + z$. If Assumption D5 does not hold, we have that both $\mathscr{H}_{\boldsymbol{\vartheta}}(z) = \mathscr{H}_{\boldsymbol{\vartheta}_0}(z)$ for all $z \in \mathbb{C}$, and $V_{\boldsymbol{\vartheta}} = V_{\boldsymbol{\vartheta}_0}$, and, consequently, that $f_{\boldsymbol{Y}_{\boldsymbol{\vartheta}}}(\omega) = f_{\boldsymbol{Y}_{\boldsymbol{\vartheta}_0}}(\omega)$, for all $\omega \in [-\pi, \pi]$, contradicting the assumption of the lemma. □

Under the assumptions described so far we obtain the following consistency result.

Theorem 2.5 (Consistency of $\hat{\boldsymbol{\vartheta}}^L$)**.** *Assume that $(F_{\boldsymbol{\vartheta}}, H_{\boldsymbol{\vartheta}}, \boldsymbol{Z}_{\boldsymbol{\vartheta}}, \boldsymbol{W}_{\boldsymbol{\vartheta}})_{\boldsymbol{\vartheta} \in \Theta}$ is a parametric family of state space models according to Definition 2.1, and let $\mathbf{y}^L = (\boldsymbol{Y}_{\boldsymbol{\vartheta}_0,1}, \ldots, \boldsymbol{Y}_{\boldsymbol{\vartheta}_0,L})$ be a sample of length L from the output process of the model corresponding to $\boldsymbol{\vartheta}_0$. If Assumptions D1 to D5 hold, then the quasi maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}^L = \operatorname{argmin}_{\boldsymbol{\vartheta} \in \Theta} \widehat{\mathscr{L}}(\boldsymbol{\vartheta}, \mathbf{y}^L)$ is strongly consistent, that is $\hat{\boldsymbol{\vartheta}}^L \to \boldsymbol{\vartheta}_0$ almost surely, as $L \to \infty$.*

We now describe the conditions which we need to impose in addition to Assumptions D1 to D5 for the asymptotic normality of the quasi maximum likelihood estimator to hold. The first one excludes the case that the true parameter value $\boldsymbol{\vartheta}_0$ lies on the boundary of the domain Θ.

Assumption D6. The true parameter value $\boldsymbol{\vartheta}_0$ is an element of the interior of Θ.

Next we need to impose a higher degree of smoothness than stated in Assumption D2 and a stronger moment condition than Assumption D4.

Assumption D7. The mappings $F_{(\cdot)}$, $H_{(\cdot)}$, $Q_{(\cdot)}$, $S_{(\cdot)}$, and $R_{(\cdot)}$ in Eqs. (2.9) are three times continuously differentiable.

By the results of the sensitivity analysis of the discrete-time algebraic Riccati equation in Sun (1998), the same degree of smoothness, namely C^3, also carries over to the mapping $\boldsymbol{\vartheta} \mapsto V_{\boldsymbol{\vartheta}}$.

Assumption D8. The process $\begin{pmatrix} \boldsymbol{W}_{\boldsymbol{\vartheta}_0}^T & \boldsymbol{Z}_{\boldsymbol{\vartheta}_0}^T \end{pmatrix}^T$ has finite $(4+\delta)$th moments for some $\delta > 0$.

By Lemma 2.1, ii), Assumption D8 implies that the process $\boldsymbol{Y}$ has finite $(4+\delta)$th moments. In the definition of the general linear stochastic state space model and in Assumption D4, it was only assumed that the sequences $\boldsymbol{Z}$ and $\boldsymbol{W}$ are stationary and ergodic. This structure alone does not entail a sufficient amount of asymptotic independence for results like Theorem 2.6 to be established We assume that the process $\boldsymbol{Y}$ is strongly mixing in the sense of Rosenblatt (1956), and we impose a summability condition on the strong mixing coefficients, which is known to be sufficient for a Central Limit Theorem for $\boldsymbol{Y}$ to hold (Bradley, 2007; Ibragimov, 1962).

Assumption D9. Denote by $\alpha_{\boldsymbol{Y}}$ the strong mixing coefficients of the process $\boldsymbol{Y} = \boldsymbol{Y}_{\boldsymbol{\vartheta}_0}$. There exists a constant $\delta > 0$ such that $\sum_{m=0}^{\infty} \left[\alpha_{\boldsymbol{Y}}(m)\right]^{\frac{\delta}{2+\delta}} < \infty$.

In the case of exponential strong mixing, Assumption D9 is always satisfied, and it is no restriction to assume that the δ appearing in Assumptions D8 and D9 are the same. It has been shown in Mokkadem (1988); Schlemm and Stelzer (2011) that, because of the autoregressive structure of the state equation (2.2a), exponential strong mixing of the output process $\boldsymbol{Y}_{\boldsymbol{\vartheta}_0}$ can be assured by imposing the condition that the process $\boldsymbol{Z}_{\boldsymbol{\vartheta}_0}$ is an i. i. d. sequence whose marginal distributions possess a non-trivial absolutely continuous component in the sense of Lebesgue's decomposition theorem, see e. g., Halmos (1950, §31, Theorem C) or Lebesgue (1904).

Finally, we require another identifiability assumption, that will be used to ensure that the Fisher information matrix of the quasi maximum likelihood estimator is non-singular. This is necessary because the asymptotic covariance matrix in the asymptotic normality result for $\hat{\boldsymbol{\vartheta}}^L$ is directly related to the inverse of that matrix. Assumption D10 is formulated in terms of the first derivative of the parametrization of the model only, which makes it relatively easy to check in practice; the Fisher information matrix, in contrast, is related to the second derivative of the logarithmic Gaussian likelihood.

For $j \in \mathbb{N}$ and $\boldsymbol{\vartheta} \in \Theta$, the vector $\psi_{\boldsymbol{\vartheta},j} \in \mathbb{R}^{(j+2)d^2}$ is defined as

$$\psi_{\boldsymbol{\vartheta},j} = \begin{pmatrix} \left[\mathbf{1}_{j+1} \otimes K_{\boldsymbol{\vartheta}}^T \otimes H_{\boldsymbol{\vartheta}}\right]\left[\begin{array}{cccc} (\operatorname{vec} \mathbf{1}_N)^T & (\operatorname{vec} F_{\boldsymbol{\vartheta}})^T & \cdots & \left(\operatorname{vec} F_{\boldsymbol{\vartheta}}^j\right)^T \end{array}\right]^T \\ \operatorname{vec} V_{\boldsymbol{\vartheta}} \end{pmatrix}, \tag{2.17}$$

where $\otimes$ denotes the Kronecker product of two matrices, and vec is the linear operator that transforms a matrix into a vector by stacking its columns on top of each other.

Assumption D10. There exists an integer $j_0 \in \mathbb{N}$ such that the $[(j_0 + 2)d^2] \times r$ matrix $\nabla_{\boldsymbol{\vartheta}}\psi_{\boldsymbol{\vartheta}_0,j_0}$ has rank r.

Our main result about the asymptotic distribution of the quasi maximum likelihood estimator for discrete-time state space models is the following theorem. Equation (2.19) shows in particular that this asymptotic distribution is independent of the choice of the initial values $\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},\text{initial}}$.

Theorem 2.6 (Asymptotic normality for $\hat{\boldsymbol{\vartheta}}^L$)**.** *Assume that $(F_{\boldsymbol{\vartheta}}, H_{\boldsymbol{\vartheta}}, \boldsymbol{Z}_{\boldsymbol{\vartheta}}, \boldsymbol{W}_{\boldsymbol{\vartheta}})_{\boldsymbol{\vartheta}\in\Theta}$ is a parametric family of state space models according to Definition 2.1, and let $\boldsymbol{y}^L = (\boldsymbol{Y}_{\boldsymbol{\vartheta}_0,1}, \ldots, \boldsymbol{Y}_{\boldsymbol{\vartheta}_0,L})$ be a sample of length L from the output process of the model corresponding to $\boldsymbol{\vartheta}_0$. If Assumptions D1 to D10 hold, then the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}^L = \operatorname{argmin}_{\boldsymbol{\vartheta}\in\Theta} \widehat{\mathscr{L}}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)$ is asymptotically normally distributed with asymptotic covariance matrix $\Xi = J^{-1}IJ^{-1}$, that is*

$$\sqrt{L}\left(\hat{\boldsymbol{\vartheta}}^L - \boldsymbol{\vartheta}_0\right) \xrightarrow[L\to\infty]{d} \mathscr{N}(\mathbf{0}, \Xi), \tag{2.18}$$

where

$$I = \lim_{L\to\infty} L^{-1}\, \mathbb{V}\mathrm{ar}\left(\nabla_{\boldsymbol{\vartheta}}\mathscr{L}\left(\boldsymbol{\vartheta}_0, \boldsymbol{y}^L\right)\right), \quad J = \lim_{L\to\infty} L^{-1}\nabla_{\boldsymbol{\vartheta}}^2\mathscr{L}\left(\boldsymbol{\vartheta}_0, \boldsymbol{y}^L\right). \tag{2.19}$$

2.3. **Proof of Theorem 2.5 – Strong consistency.** In this section we prove the strong consistency of the quasi maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}^L$. As a first step we show that the stationary pseudo-innovations processes defined by the steady-state Kalman filter are uniformly approximated by their counterparts based on the finite sample $\boldsymbol{y}^L$.

Lemma 2.7. *Under Assumptions D1 to D3, the pseudo-innovations sequences $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ and $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta}}$ defined by the Kalman filter equations (2.6a) and (2.12) have the following properties.*

i) If the initial values $\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},initial}$ are such that $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},initial}\right\|$ is almost surely finite, then, with probability one, there exist a positive number C and a positive number $\rho < 1$, such that $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} - \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}\right\| \leqslant C\rho^n$, $n \in \mathbb{N}$. In particular, $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta}_0,n}$ converges to the true innovations $\boldsymbol{\varepsilon}_n = \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}$ at an exponential rate.

ii) The sequences $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ are linear functions of $\boldsymbol{Y}$, that is there exist matrix sequences $(c_{\boldsymbol{\vartheta},\nu})_{\nu\geqslant 1}$, such that

$$\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} = \boldsymbol{Y}_n + \sum_{\nu=1}^{\infty} c_{\boldsymbol{\vartheta},\nu}\boldsymbol{Y}_{n-\nu}, \quad n \in \mathbb{Z}. \tag{2.20}$$

The matrices $c_{\boldsymbol{\vartheta},\nu}$ are uniformly exponentially bounded, that is there exist a positive constant C and a positive constant $\rho < 1$, such that $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|c_{\boldsymbol{\vartheta},\nu}\right\| \leqslant C\rho^\nu$, $\nu \in \mathbb{N}$.

Proof. We first prove part i) about the uniform exponential approximation of $\boldsymbol{\varepsilon}$ by $\hat{\boldsymbol{\varepsilon}}$. Iterating the Kalman equations (2.6a) and (2.12), we find that, for $n \in \mathbb{N}$,

$$\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} = \boldsymbol{Y}_n - H_{\boldsymbol{\vartheta}} \left(F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}\right)^{n-1} \hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},1} - \sum_{\nu=1}^{n-1} H_{\boldsymbol{\vartheta}} \left(F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}\right)^{\nu-1} K_{\boldsymbol{\vartheta}} \boldsymbol{Y}_{n-\nu},$$

and

$$\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n} = \boldsymbol{Y}_n - H_{\boldsymbol{\vartheta}} \left(F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}\right)^{n-1} \hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},\text{initial}} - \sum_{\nu=1}^{n-1} H_{\boldsymbol{\vartheta}} \left(F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}\right)^{\nu-1} K_{\boldsymbol{\vartheta}} \boldsymbol{Y}_{n-\nu}.$$

Thus, using the fact that, by Lemma 2.3, the spectral radii of $F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}$ are bounded by $\rho < 1$, it follows that

$$\begin{aligned}\sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} - \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}\right\| &= \sup_{\boldsymbol{\vartheta}\in\Theta} \left\|H_{\boldsymbol{\vartheta}} \left(F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}\right)^{n-1} \left(\boldsymbol{X}_{\boldsymbol{\vartheta},0} - \boldsymbol{X}_{\boldsymbol{\vartheta},\text{initial}}\right)\right\| \\ &\leqslant \|H\|_{L^\infty(\Theta)}\, \rho^{n-1} \sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\boldsymbol{X}_{\boldsymbol{\vartheta},0} - \boldsymbol{X}_{\boldsymbol{\vartheta},\text{initial}}\right\|,\end{aligned}$$

where $\|H\|_{L^\infty(\Theta)} := \sup_{\boldsymbol{\vartheta}\in\Theta} \|H_{\boldsymbol{\vartheta}}\|$ denotes the supremum norm of $H_{(\cdot)}$, which is finite by the Extreme Value Theorem. Since the last factor is almost surely finite by assumption, the claim follows. For part ii), we observe that Eq. (2.6a) and Lemma 2.3, ii) imply that $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ has the infinite-order moving average representation $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} = \boldsymbol{Y}_n - H_{\boldsymbol{\vartheta}} \sum_{\nu=1}^{\infty} \left(F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}\right)^{\nu-1} K_{\boldsymbol{\vartheta}} \boldsymbol{Y}_{n-\nu}$, with uniformly exponentially bounded coefficients $c_{\boldsymbol{\vartheta},\nu} := -H_{\boldsymbol{\vartheta}} \left(F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}}\right)^{\nu-1} K_{\boldsymbol{\vartheta}}$. Explicitly, $\|c_{\boldsymbol{\vartheta},\nu}\| \leqslant \|H\|_{L^\infty(\Theta)} \|K\|_{L^\infty(\Theta)}\, \rho^{n-1}$. This completes the proof. □

Lemma 2.8. *Let $\mathscr{L}$ and $\widehat{\mathscr{L}}$ be given by Eqs. (2.10) and (2.14). If Assumptions D1 to D3 are satisfied, then the sequence $L^{-1} \sup_{\boldsymbol{\vartheta}\in\Theta} \left|\widehat{\mathscr{L}}(\boldsymbol{\vartheta}, \boldsymbol{y}^L) - \mathscr{L}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)\right|$ converges to zero almost surely, as $L \to \infty$.*

Proof. We first observe that

$$\begin{aligned}\left|\widehat{\mathscr{L}}(\boldsymbol{\vartheta}, \boldsymbol{y}^L) - \mathscr{L}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)\right| &= \sum_{n=1}^{L} \left[\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1} \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n} - \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1} \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right] \\ &= \sum_{n=1}^{L} \left[\left(\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n} - \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right)^T V_{\boldsymbol{\vartheta}}^{-1} \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n} + \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1} \left(\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n} - \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right)\right].\end{aligned}$$

The fact that, by Lemma 2.3, iii), there exists a constant C such that $\left\|V_{\boldsymbol{\vartheta}}^{-1}\right\| \leqslant C$, for all $\boldsymbol{\vartheta} \in \Theta$, implies that

$$\frac{1}{L} \sup_{\boldsymbol{\vartheta}\in\Theta} \left|\widehat{\mathscr{L}}(\boldsymbol{\vartheta}, \boldsymbol{y}^L) - \mathscr{L}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)\right| \leqslant \frac{C}{L} \sum_{n=1}^{L} \rho^n \left[\sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}\right\| + \sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\|\right]. \tag{2.21}$$

Lemma 2.7, ii) and the assumption that $\boldsymbol{Y}$ has finite second moments imply the finiteness of the expectation $\mathbb{E} \sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\|$. Applying Markov's inequality, one sees that, for every positive ϵ,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\rho^n \sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\| \geqslant \epsilon\right) \leqslant \mathbb{E} \sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}\right\| \sum_{n=1}^{\infty} \frac{\rho^n}{\epsilon} < \infty,$$

because $\rho < 1$. The Borel–Cantelli Lemma shows that $\rho^n \sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\|$ converges to zero almost surely, as $n \to \infty$. In an analogous way one can show that $\rho^n \sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}\right\|$ converges to zero almost surely, and, consequently, so does the Cesàro mean in Eq. (2.21). The claim thus follows. □

Lemma 2.9. *Assume that Assumptions D3 and D4 as well as the first part of Assumption D5, Eq. (2.16a), hold. If $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1} = \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}$ almost surely, then $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$.*

Proof. Assume, for the sake of contradiction, that $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}_0$. By Assumption D5, there exist matrices $C_j \in M_d(\mathbb{R})$, $j \in \mathbb{N}_0$, such that, for $|z| \leqslant 1$,

$$H_{\boldsymbol{\vartheta}} \left[\mathbf{1}_N - (F_{\boldsymbol{\vartheta}} - K_{\boldsymbol{\vartheta}} H_{\boldsymbol{\vartheta}})z\right]^{-1} K_{\boldsymbol{\vartheta}} - H_{\boldsymbol{\vartheta}_0} \left[\mathbf{1}_N - (F_{\boldsymbol{\vartheta}_0} - K_{\boldsymbol{\vartheta}_0} H_{\boldsymbol{\vartheta}_0} z\right]^{-1} K_{\boldsymbol{\vartheta}_0} = \sum_{j=j_0}^{\infty} C_j z^j, \tag{2.22}$$

where $C_{j_0} \neq 0$, for some $j_0 \geqslant 0$. Using Eq. (2.6b) and the assumed equality of $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}$ and $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}$, this equation implies that $\mathbf{0}_d = \sum_{j=j_0}^{\infty} C_j \boldsymbol{Y}_{j_0-j}$ almost surely; in particular, the random variable $C_{j_0}\boldsymbol{Y}_0$ is almost surely equal to a linear combination of the components of $\boldsymbol{Y}_n$, $n<0$. It thus follows from the interpretation of the innovations sequence $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0}$ as linear prediction errors for the process $\boldsymbol{Y}$ that $C_{j_0}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,0}$ is equal to zero, which implies that $\mathbb{E}C_{j_0}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,0}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,0}^T C_{j_0}^T = C_{j_0}V_{\boldsymbol{\vartheta}_0}C_{j_0}^T = 0_d$. Since $V_{\boldsymbol{\vartheta}_0}$ is assumed to be non-singular, this implies that the matrix C_{j_0} is the null matrix, a contradiction to Eq. (2.22). □

Lemma 2.10. *If Assumptions D1 to D4 hold, then, with probability one, the sequence of random functions* $\boldsymbol{\vartheta} \mapsto L^{-1}\widehat{\mathscr{L}}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)$ *converges, as L tends to infinity, uniformly in* $\boldsymbol{\vartheta}$ *to the limiting function* $\mathscr{Q}: \Theta \to \mathbb{R}$ *defined by*

$$\mathscr{Q}(\boldsymbol{\vartheta}) = d\log(2\pi) + \log\det V_{\boldsymbol{\vartheta}} + \mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}. \tag{2.23}$$

Proof. In view of the approximation results in Lemma 2.8, it is enough to show that the sequence of random functions $\boldsymbol{\vartheta} \mapsto L^{-1}\mathscr{L}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)$ converges uniformly to $\mathscr{Q}$. The proof of this assertion is based on the observation following Assumption D4 that for each $\boldsymbol{\vartheta} \in \Theta$ the sequence $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ is ergodic and its consequence that, by Birkhoff's Ergodic Theorem (Durrett, 2010, Theorem 6.2.1), the sequence $L^{-1}\mathscr{L}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)$ converges to $\mathscr{Q}(\boldsymbol{\vartheta})$ point-wise. The stronger statement of uniform convergence follows from Assumption D1 that Θ is compact by an argument that is inspired by the proof of Ferguson (1996, Theorem 16): for $\delta > 0$, we write $B_\delta(\boldsymbol{\vartheta}) = \left\{\boldsymbol{\vartheta}' \in \Theta : \left\|\boldsymbol{\vartheta}' - \boldsymbol{\vartheta}\right\| < \delta\right\}$ for the open ball of radius δ around $\boldsymbol{\vartheta}$. The sequences $\underline{\sigma}_{\boldsymbol{\vartheta}}^\delta = \left(\underline{\sigma}_{\boldsymbol{\vartheta},n}^\delta\right)_{n\in\mathbb{Z}}$ and $\overline{\sigma}_{\boldsymbol{\vartheta}}^\delta = \left(\overline{\sigma}_{\boldsymbol{\vartheta},n}^\delta\right)_{n\in\mathbb{Z}}$, which are defined by

$$\underline{\sigma}_{\boldsymbol{\vartheta},n}^\delta = \inf_{\boldsymbol{\vartheta}' \in B_\delta(\boldsymbol{\vartheta})}\left[\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n} - \mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1}\right] \quad\text{and}\quad \overline{\sigma}_{\boldsymbol{\vartheta},n}^\delta = \sup_{\boldsymbol{\vartheta}' \in B_\delta(\boldsymbol{\vartheta})}\left[\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n} - \mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1}\right],$$

are strictly stationary, ergodic and monotone in δ. By Lemma 2.7, ii) there exists an integrable random variable Z such that $\underline{\sigma}_{\boldsymbol{\vartheta},1}^\delta < Z$ for all δ and all $\boldsymbol{\vartheta} \in \Theta$. Since, moreover, $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}$ is almost surely a continuous function of $\boldsymbol{\vartheta}$, and $\underline{\sigma}_{\boldsymbol{\vartheta},n}^\delta$ thus converges to $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} - \mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}$ almost surely, as $\delta \to 0$, it follows from the Ergodic Theorem and Lebesgue's Dominated Convergence Theorem (Klenke, 2008, Corollary 6.26) that

$$\frac{1}{L}\sum_{n=1}^{L}\underline{\sigma}_{\boldsymbol{\vartheta},n}^\delta \xrightarrow[L\to\infty]{\text{a. s.}} \mathbb{E}\underline{\sigma}_{\boldsymbol{\vartheta},1}^\delta \xrightarrow[\delta\to 0]{} \mathbb{E}\left[\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} - \mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}\right] = 0, \tag{2.24}$$

and similarly for $\overline{\sigma}_{\boldsymbol{\vartheta}}^\delta$. Since, for any $\boldsymbol{\vartheta}' \in B_\delta(\boldsymbol{\vartheta})$, it holds that

$$\frac{1}{L}\sum_{n=1}^{L}\underline{\sigma}_{\boldsymbol{\vartheta},n}^\delta \leqslant \frac{1}{L}\sum_{n=1}^{L}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n} - \mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1} \leqslant \frac{1}{L}\sum_{n=1}^{L}\overline{\sigma}_{\boldsymbol{\vartheta},n}^\delta,$$

it follows that

$$\sup_{\boldsymbol{\vartheta}' \in B_\delta(\boldsymbol{\vartheta})}\left|\frac{1}{L}\sum_{n=1}^{L}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n} - \mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1}\right| \leqslant \left|\frac{1}{L}\sum_{n=1}^{L}\underline{\sigma}_{\boldsymbol{\vartheta},n}^\delta\right| + \left|\frac{1}{L}\sum_{n=1}^{L}\overline{\sigma}_{\boldsymbol{\vartheta},n}^\delta\right|.$$

Letting L tend to infinity on both sides of this inequality, we see that, almost surely,

$$\limsup_{L\to\infty}\sup_{\boldsymbol{\vartheta}' \in B_\delta(\boldsymbol{\vartheta})}\left|\frac{1}{L}\sum_{n=1}^{L}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',n} - \mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}',1}\right| \leqslant \left|\mathbb{E}\underline{\sigma}_{\boldsymbol{\vartheta},1}^\delta\right| + \left|\mathbb{E}\overline{\sigma}_{\boldsymbol{\vartheta},1}^\delta\right|.$$

By Eq. (2.24) one finds, for every $\epsilon > 0$ and every $\boldsymbol{\vartheta} \in \Theta$, a $\delta(\epsilon, \boldsymbol{\vartheta}) > 0$ such that $\left|\mathbb{E}\underline{\sigma}_{\boldsymbol{\vartheta},1}^\delta\right| + \left|\mathbb{E}\overline{\sigma}_{\boldsymbol{\vartheta},1}^\delta\right| < \epsilon$, for all $\delta < \delta(\epsilon, \boldsymbol{\vartheta})$. The collection of balls $\{B_{\delta(\epsilon,\boldsymbol{\vartheta})}(\boldsymbol{\vartheta})\}_{\boldsymbol{\vartheta}\in\Theta}$ covers Θ, and since the domain Θ is assumed to be compact, one can extract a finite subcover $\{B_{\delta(\epsilon,\boldsymbol{\vartheta}_i)}(\boldsymbol{\vartheta}_i)\}_{i=1,\ldots,k}$. Defining $\delta(\epsilon)$ to be the minimum of the radii

$\delta(\epsilon,\boldsymbol{\vartheta}_i)$, $i=1,\ldots,k$, it follows that, with probability one,

$$\begin{aligned}
&\limsup_{L\to\infty}\sup_{\boldsymbol{\vartheta}\in\Theta}\left|\frac{1}{L}\sum_{n=1}^{L}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}-\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}^T V_{\boldsymbol{\vartheta}'}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}\right|\\
=&\limsup_{L\to\infty}\max_{i=1,\ldots,k}\sup_{\boldsymbol{\vartheta}\in B_{\delta(\epsilon,\boldsymbol{\vartheta}_i)}}\left|\frac{1}{L}\sum_{n=1}^{L}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}-\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}\right|\\
\leqslant&\max_{i=1,\ldots,k}\left\{\left|\mathbb{E}\underline{\sigma}_{\boldsymbol{\vartheta}_i,1}^{\delta(\epsilon,\boldsymbol{\vartheta}_i)}\right|+\left|\mathbb{E}\overline{\sigma}_{\boldsymbol{\vartheta}_i,1}^{\delta(\epsilon,\boldsymbol{\vartheta}_i)}\right|\right\}\leqslant\epsilon,
\end{aligned}$$

for all $\delta\leqslant\delta(\epsilon)$. Intersecting over a sequence ϵ_n which converges to zero proves the result. □

Lemma 2.11. *Under Assumptions D1 to D3 and D5, the function $\mathcal{Q}:\Theta\to\mathbb{R}$, as defined in Eq. (2.23), has a unique global minimum at $\boldsymbol{\vartheta}_0$.*

Proof. We first observe that the difference $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}-\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}$ is an element of the Hilbert space spanned by the random variables $\{\boldsymbol{Y}_n, n\leqslant 0\}$, and that $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}$ is, by definition, orthogonal to this space. This implies that the expectation $\mathbb{E}\left(\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}-\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\right)^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}$ is equal to zero and, consequently, that $\mathcal{Q}(\boldsymbol{\vartheta})$ can be written as

$$\mathcal{Q}(\boldsymbol{\vartheta})=d\log(2\pi)+\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}^T V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}+\mathbb{E}\left(\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}-\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\right)^T V_{\boldsymbol{\vartheta}}^{-1}\left(\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}-\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\right)+\log\det V_{\boldsymbol{\vartheta}}.$$

In particular, since $\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}^T V_{\boldsymbol{\vartheta}_0}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}=\operatorname{tr}\left[V_{\boldsymbol{\vartheta}_0}^{-1}\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}^T\right]=d$, it follows that $\mathcal{Q}(\boldsymbol{\vartheta}_0)=\log\det V_{\boldsymbol{\vartheta}_0}+d(1+\log(2\pi))$. The elementary inequality $x-\log x\geqslant 1$, for $x>0$, implies that $\operatorname{tr}M-\log\det M\geqslant d$ for all symmetric positive definite $d\times d$ matrices $M\in\mathbb{S}_d^{++}(\mathbb{R})$ with equality if and only if $M=\mathbf{1}_d$. Using this inequality for $M=V_{\boldsymbol{\vartheta}_0}^{-1}V_{\boldsymbol{\vartheta}}$, we thus obtain that, for all $\boldsymbol{\vartheta}\in\Theta$,

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\vartheta})-\mathcal{Q}(\boldsymbol{\vartheta}_0)=&d+\operatorname{tr}\left[V_{\boldsymbol{\vartheta}}^{-1}\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}^T\right]-\log\det\left(V_{\boldsymbol{\vartheta}_0}^{-1}V_{\boldsymbol{\vartheta}}\right)\\
&+\mathbb{E}\left(\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}-\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\right)^T V_{\boldsymbol{\vartheta}}^{-1}\left(\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}-\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\right)-\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}^T V_{\boldsymbol{\vartheta}_0}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\\
\geqslant&\mathbb{E}\left(\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}-\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\right)^T V_{\boldsymbol{\vartheta}}^{-1}\left(\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}-\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\right)\geqslant 0.
\end{aligned}$$

It remains to argue that this chain of inequalities is in fact a strict inequality if $\boldsymbol{\vartheta}\neq\boldsymbol{\vartheta}_0$. If $V_{\boldsymbol{\vartheta}}\neq V_{\boldsymbol{\vartheta}_0}$, the first inequality is strict, and we are done. If $V_{\boldsymbol{\vartheta}}=V_{\boldsymbol{\vartheta}_0}$, the first part of Assumption D5, Eq. (2.16a), is satisfied. The second inequality is an equality if and only if $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},1}=\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}$ almost surely, which, by Lemma 2.9, implies that $\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0$. Thus, the function $\mathcal{Q}$ has a unique global minimum at $\boldsymbol{\vartheta}_0$. □

of Theorem 2.5. We shall first show that the sequence $L^{-1}\widehat{\mathscr{L}}(\hat{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L)$, $L\in\mathbb{N}$, converges almost surely to $\mathcal{Q}(\boldsymbol{\vartheta}_0)$ as the sample size L tends to infinity. Assume that, for some positive number ϵ, it holds that $\sup_{\boldsymbol{\vartheta}\in\Theta}\left|L^{-1}\widehat{\mathscr{L}}(\boldsymbol{\vartheta},\boldsymbol{y}^L)-\mathcal{Q}(\boldsymbol{\vartheta})\right|\leqslant\epsilon$. It then follows that

$$L^{-1}\widehat{\mathscr{L}}(\hat{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L)\leqslant L^{-1}\widehat{\mathscr{L}}(\boldsymbol{\vartheta}_0,\boldsymbol{y}^L)\leqslant\mathcal{Q}(\boldsymbol{\vartheta}_0)+\epsilon\quad\text{and}\quad L^{-1}\widehat{\mathscr{L}}(\hat{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L)\geqslant\mathcal{Q}(\hat{\boldsymbol{\vartheta}}^L)-\epsilon\geqslant\mathcal{Q}(\boldsymbol{\vartheta}_0)-\epsilon,$$

where it was used that $\hat{\boldsymbol{\vartheta}}^L$ is defined to minimize $\widehat{\mathscr{L}}(\cdot,\boldsymbol{y}^L)$ and that, by Lemma 2.11, $\boldsymbol{\vartheta}_0$ minimizes $\mathcal{Q}(\cdot)$. In particular, it follows that $\left|L^{-1}\widehat{\mathscr{L}}(\hat{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L)-\mathcal{Q}(\boldsymbol{\vartheta}_0)\right|\leqslant\epsilon$. This observation and Lemma 2.10 immediately imply that

$$\mathbb{P}\left(\frac{1}{L}\widehat{\mathscr{L}}(\hat{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L)\xrightarrow[L\to\infty]{}\mathcal{Q}(\boldsymbol{\vartheta}_0)\right)\geqslant\mathbb{P}\left(\sup_{\boldsymbol{\vartheta}\in\Theta}\left|\frac{1}{L}\widehat{\mathscr{L}}(\boldsymbol{\vartheta},\boldsymbol{y}^L)-\mathcal{Q}(\boldsymbol{\vartheta})\right|\xrightarrow[L\to\infty]{}0\right)=1.\tag{2.25}$$

To complete the proof of the theorem, it suffices to show that, for every neighbourhood U of $\boldsymbol{\vartheta}_0$, with probability one, $\hat{\boldsymbol{\vartheta}}^L$ will eventually lie in U. For every such neighbourhood U of $\boldsymbol{\vartheta}_0$, we define the real number $\delta(U):=\inf_{\boldsymbol{\vartheta}\in\Theta\backslash U}\mathcal{Q}(\boldsymbol{\vartheta})-\mathcal{Q}(\boldsymbol{\vartheta}_0)$, which is strictly positive by Lemma 2.11. Then the following sequence of inequalities holds:

$$\begin{aligned}
\mathbb{P}\left(\hat{\boldsymbol{\vartheta}}^L \xrightarrow[L\to\infty]{} \boldsymbol{\vartheta}_0\right) =& \mathbb{P}\left(\forall U\, \exists L_0 : \hat{\boldsymbol{\vartheta}}^L \in U \quad \forall L > L_0\right)\\
\geqslant& \mathbb{P}\left(\forall U\, \exists L_0 : \mathscr{Q}(\hat{\boldsymbol{\vartheta}}^L) - \mathscr{Q}(\boldsymbol{\vartheta}_0) < \delta(U) \quad \forall L > L_0\right)\\
\geqslant& \mathbb{P}\left(\forall U\, \exists L_0 : \left|\frac{1}{L}\widehat{\mathscr{L}}(\hat{\boldsymbol{\vartheta}}^L, \boldsymbol{y}^L) - \mathscr{Q}(\boldsymbol{\vartheta}_0)\right| < \frac{\delta(U)}{2}\right.\\
&\left.\quad \text{and} \quad \left|\frac{1}{L}\widehat{\mathscr{L}}(\hat{\boldsymbol{\vartheta}}^L, \boldsymbol{y}^L) - \mathscr{Q}(\hat{\boldsymbol{\vartheta}}^L)\right| < \frac{\delta(U)}{2} \quad \forall L > L_0\right)\\
\geqslant& \mathbb{P}\left(\forall U\, \exists L_0 : \left|\frac{1}{L}\widehat{\mathscr{L}}(\hat{\boldsymbol{\vartheta}}^L, \boldsymbol{y}^L) - \mathscr{Q}(\boldsymbol{\vartheta}_0)\right| < \frac{\delta(U)}{2} \quad \forall L > L_0\right)\\
&+ \mathbb{P}\left(\forall U\, \exists L_0 : \sup_{\boldsymbol{\vartheta}\in\Theta}\left|\frac{1}{L}\widehat{\mathscr{L}}(\boldsymbol{\vartheta}, \boldsymbol{y}^L) - \mathscr{Q}(\boldsymbol{\vartheta})\right| < \frac{\delta(U)}{2} \quad \forall L > L_0\right) - 1.
\end{aligned}$$

The first probability in the last line is equal to one by Eq. (2.25), the second because, by Lemma 2.10, the random functions $\boldsymbol{\vartheta} \mapsto L^{-1}\widehat{\mathscr{L}}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)$ converge almost surely uniformly to the function $\boldsymbol{\vartheta} \mapsto \mathscr{Q}(\boldsymbol{\vartheta})$. It thus follows that $\mathbb{P}\left(\hat{\boldsymbol{\vartheta}}^L \xrightarrow[L\to\infty]{} \boldsymbol{\vartheta}_0\right) = 1$, which proves the theorem. □

2.4. **Proof of Theorem 2.6 – Asymptotic normality.** In this section we prove the assertion of Theorem 2.6, namely that the distribution of $L^{1/2}\left(\hat{\boldsymbol{\vartheta}}^L - \boldsymbol{\vartheta}_0\right)$ converges to a normal random variable with mean zero and covariance matrix $\Xi = J^{-1}IJ^{-1}$, an expression for which is given in Eq. (2.19). First, we collect basic properties of $\partial_m\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}$ and $\partial_m\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}$, where $\partial_m = \partial/\partial\vartheta^m$ denotes the partial derivative with respect to the mth component of $\boldsymbol{\vartheta}$; the following lemma mirrors Lemma 2.7.

Lemma 2.12. *Assume that Assumptions D1 to D3 and D7 hold. The pseudo-innovations sequences $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ and $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta}}$ defined by the Kalman filter equations (2.6a) and (2.12) have the following properties.*

i) If, for $k \in \{1,\dots,r\}$, the initial values $\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},initial}$ are defined such that both $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},initial}\right\|$ and $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\partial_k\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},initial}\right\|$ are almost surely finite, then, with probability one, there exist a positive number C and a positive number $\rho < 1$, such that $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\partial_k\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} - \partial_k\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}\right\| \leqslant C\rho^n$, $n \in \mathbb{N}$.

ii) For each $k \in \{1,\dots,r\}$, the random sequences $\partial_k\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ are linear functions of $\boldsymbol{Y}$, that is there exist matrix sequences $\left(c^{(k)}_{\boldsymbol{\vartheta},\nu}\right)_{\nu\geqslant 1}$, such that

$$\partial_k\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} = \sum_{\nu=1}^{\infty} c^{(k)}_{\boldsymbol{\vartheta},\nu}\boldsymbol{Y}_{n-\nu}, \quad n \in \mathbb{Z}. \tag{2.26}$$

The matrices $c^{(k)}_{\boldsymbol{\vartheta},\nu}$ are uniformly exponentially bounded, that is there exist a positive constant C and a positive constant $\rho < 1$, such that $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|c^{(k)}_{\boldsymbol{\vartheta},\nu}\right\| \leqslant C\rho^\nu, \nu \in \mathbb{N}$.

iii) If, for $k,l \in \{1,\dots,r\}$, the initial values $\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},initial}$ are defined such that $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},initial}\right\|$, as well as $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\partial_i\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},initial}\right\|$, $i \in \{k,l\}$, and $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\partial^2_{k,l}\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},initial}\right\|$ are almost surely finite, then, with probability one, there exist a positive number C and a positive number $\rho < 1$, such that $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\partial^2_{k,l}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} - \partial^2_{k,l}\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}\right\| \leqslant C\rho^n$, $n \in \mathbb{N}$.

iv) For each $k,l \in \{1,\dots,r\}$, the random sequences $\partial^2_{k,l}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ are linear functions of $\boldsymbol{Y}$, that is there exist matrix sequences $\left(c^{(k,l)}_{\boldsymbol{\vartheta},\nu}\right)_{\nu\geqslant 1}$,such that

$$\partial^2_{k,l}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} = \sum_{\nu=1}^{\infty} c^{(k,l)}_{\boldsymbol{\vartheta},\nu}\boldsymbol{Y}_{n-\nu}, \quad n \in \mathbb{Z}. \tag{2.27}$$

The matrices $c^{(k,l)}_{\boldsymbol{\vartheta},\nu}$ are uniformly exponentially bounded, that is there exist a positive constant C and a positive constant $\rho < 1$, such that $\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|c^{(k,l)}_{\boldsymbol{\vartheta},\nu}\right\| \leqslant C\rho^\nu$, $\nu \in \mathbb{N}$.

Proof. Analogous to the proof of Lemma 2.7, repeatedly interchanging differentiation and summation and using the fact that, as a consequence of Assumptions D1 to D3 and D7, both $\partial_k\left[H_{\boldsymbol{\vartheta}}\left(F_{\boldsymbol{\vartheta}}-K_{\boldsymbol{\vartheta}}H_{\boldsymbol{\vartheta}}\right)^{\nu-1}K_{\boldsymbol{\vartheta}}\right]$ and $\partial^2_{k,l}\left[H_{\boldsymbol{\vartheta}}\left(F_{\boldsymbol{\vartheta}}-K_{\boldsymbol{\vartheta}}H_{\boldsymbol{\vartheta}}\right)^{\nu-1}K_{\boldsymbol{\vartheta}}\right]$ are uniformly exponentially bounded in ν. □

Lemma 2.13. *For each $\boldsymbol{\vartheta}\in\Theta$ and every $m=1,\ldots,r$, the random variable $\partial_m\mathscr{L}(\boldsymbol{\vartheta},\boldsymbol{y}^L)$ has finite variance.*

Proof. By Assumption D8 and the exponential decay of the coefficient matrices $c_{\boldsymbol{\vartheta},\nu}$ and $c^{(m)}_{\boldsymbol{\vartheta},\nu}$ proved in Lemma 2.7, ii) and Lemma 2.12, ii), it follows that

$$\mathbb{E}\left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\|^4\leqslant\left[C\sum_{\nu=0}^{\infty}\rho^\nu\right]^4\mathbb{E}\left\|\boldsymbol{Y}_1\right\|^4<\infty\quad\text{and}\quad\mathbb{E}\left\|\partial_m\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\|^4\leqslant\left[C\sum_{\nu=0}^{\infty}\rho^\nu\right]^4\mathbb{E}\left\|\boldsymbol{Y}_1\right\|^4<\infty.$$

Consequently, the Cauchy–Schwarz inequality implies that, for some constant C,

$$\begin{aligned}\mathbb{E}\left|\partial_m\left(\boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n}V^{-1}_{\boldsymbol{\vartheta}}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right)\right|^2=&\mathbb{E}\left|-\operatorname{tr}\left[V^{-1}_{\boldsymbol{\vartheta}}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n}\left(\partial_mV_{\boldsymbol{\vartheta}}\right)\right]+2\left(\partial_m\boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n}\right)V^{-1}_{\boldsymbol{\vartheta}}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right|^2\\ \leqslant&C\left\{\mathbb{E}\left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\|^4+\left(\mathbb{E}\left\|\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\|^4\mathbb{E}\left\|\partial_m\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\|^4\right)^{1/2}\right\}<\infty,\end{aligned}$$

which proves that $\partial_m\mathscr{L}(\boldsymbol{\vartheta},\boldsymbol{y}^L)$ has finite second moments. □

We need the following multivariate covariance inequality which is a consequence of Davydov's inequality and the multidimensional generalization of an inequality used in the proof of Francq and Zakoïan (1998, Lemma 3). For a positive real number α, we denote by $\lfloor\alpha\rfloor$ the greatest integer smaller than or equal to α.

Lemma 2.14. *Let $\boldsymbol{X}$ be a strictly stationary, strongly mixing d-dimensional stochastic process with finite $(4+\delta)$th moments for some $\delta>0$. Then there exists a constant C, such that for all $d\times d$ matrices A, B, every $n\in\mathbb{Z}$, $\Delta\in\mathbb{N}$, and time indices $\nu,\nu'\in\mathbb{N}_0$, $\mu,\mu'=0,1\ldots,\lfloor\Delta/2\rfloor$, it holds that*

$$\operatorname{\mathbb{C}ov}\left(\boldsymbol{X}^T_{n-\nu}A\boldsymbol{X}_{n-\nu'};\boldsymbol{X}^T_{n+\Delta-\mu}B\boldsymbol{X}_{n+\Delta-\mu'}\right)\leqslant C\left\|A\right\|\left\|B\right\|\left[\alpha_{\boldsymbol{X}}\left(\left\lfloor\frac{\Delta}{2}\right\rfloor\right)\right]^{\delta/(\delta+2)},\tag{2.28}$$

where $\alpha_{\boldsymbol{X}}$ denotes the strong mixing coefficients of the process $\boldsymbol{X}$.

Proof. We first note that the bilinearity of $\operatorname{\mathbb{C}ov}(\cdot;\cdot)$ and the elementary inequality $M_{ij}\leqslant\|M\|$, $M\in M_d(\mathbb{R})$, imply that

$$\begin{aligned}\operatorname{\mathbb{C}ov}\left(\boldsymbol{X}^T_{n-\nu}A\boldsymbol{X}_{n-\nu'};\boldsymbol{X}^T_{n+\Delta-\mu}B\boldsymbol{X}_{n+\Delta-\mu'}\right)=&\sum_{i,j,s,t=1}^{d}A_{ij}B_{st}\operatorname{\mathbb{C}ov}\left(X^i_{n-\nu}X^j_{n-\nu'};X^s_{n+\Delta-\mu}X^t_{n+\Delta-\mu'}\right)\\ \leqslant&d^4\left\|A\right\|\left\|B\right\|\max_{i,j,s,t=1,\ldots,d}\operatorname{\mathbb{C}ov}\left(X^i_{n-\nu}X^j_{n-\nu'};X^s_{n+\Delta-\mu}X^t_{n+\Delta-\mu'}\right).\end{aligned}$$

Since the projection which maps a vector to one of its components is measurable, it follows that the random variable $X^i_{n-\nu}X^j_{n-\nu'}$ is measurable with respect to $\mathscr{F}^{n-\min\{\nu,\nu'\}}_{-\infty}$, the σ-algebra generated by $\{\boldsymbol{X}_k:-\infty<k\leqslant n-\min\{\nu,\nu'\}\}$. Similarly, the random variable $X^s_{n+\Delta-\mu}X^t_{n+\Delta-\mu'}$ is measurable with respect to $\mathscr{F}^{\infty}_{n+\Delta-\max\{\mu,\mu'\}}$. Davydov's inequality (Davydov, 1968, Lemma 2.1) implies that there exists a universal constant K such that

$$\begin{aligned}\operatorname{\mathbb{C}ov}\left(X^i_{n-\nu}X^j_{n-\nu'};X^s_{n+\Delta-\mu}X^t_{n+\Delta-\mu'}\right)\leqslant&K\left(\mathbb{E}\left|X^i_{n-\nu}X^j_{n-\nu'}\right|^{2+\delta}\right)^{1/(2+\delta)}\left(\mathbb{E}\left|X^s_{n+\Delta-\mu}X^t_{n+\Delta-\mu'}\right|^{2+\delta}\right)^{1/(2+\delta)}\\ &\times\left[\alpha_{\boldsymbol{X}}\left(\Delta-\max\{\mu,\mu'\}+\min\{\nu,\nu'\}\right)\right]^{\delta/(2+\delta)}\\ \leqslant&C\left[\alpha_{\boldsymbol{X}}\left(\left\lfloor\frac{\Delta}{2}\right\rfloor\right)\right]^{\delta/(2+\delta)},\end{aligned}$$

where it was used that $\Delta-\max\{\mu,\mu'\}+\min\{\nu,\nu'\}\geqslant\lfloor\Delta/2\rfloor$, and that strong mixing coefficients are non-increasing. By the Cauchy–Schwarz inequality the constant C satisfies

$$C=K\left(\mathbb{E}\left|X^i_{n-\nu}X^j_{n-\nu'}\right|^{2+\delta}\right)^{1/(2+\delta)}\left(\mathbb{E}\left|X^s_{n+\Delta-\mu}X^t_{n+\Delta-\mu'}\right|^{2+\delta}\right)^{1/(2+\delta)}\leqslant K\left(\mathbb{E}\left\|\boldsymbol{X}_1\right\|^{4+2\delta}\right)^{\frac{2}{2+\delta}},$$

and thus does not depend on $n,\nu,\nu',\mu,\mu',\Delta$, nor on i,j,s,t. □

The next lemma is a multivariate generalization of Francq and Zakoïan (1998, Lemma 3). In the proof of Boubacar Mainassara and Francq (2011, Lemma 4) this generalization is used without providing details and, more importantly without imposing Assumption D9 about the strong mixing of $\boldsymbol{Y}$. In view of the derivative terms $\partial_m \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}$ in Eq. (2.30) it is not clear how the result of the lemma can be proved under the mere assumption of strong mixing of the innovations sequence $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0}$. We therefore think that a detailed account, properly generalizing the arguments in the original paper (Francq and Zakoïan, 1998) to the multidimensional setting, is justified.

Lemma 2.15. *Suppose that Assumptions D1 to D3, D8 and D9 hold. Then, for every $\boldsymbol{\vartheta} \in \Theta$, the sequence $L^{-1}\,\mathbb{V}\mathrm{ar}\,\nabla_{\boldsymbol{\vartheta}} \mathscr{L}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)$ of deterministic matrices converges to a limit $I(\boldsymbol{\vartheta})$ as $L \to \infty$.*

Proof. It is enough to show that, for each $\boldsymbol{\vartheta} \in \Theta$, and all $k, l = 1, \ldots, r$, the sequence of real-valued random variables $I^{(k,l)}_{\boldsymbol{\vartheta},L}$, defined by

$$I^{(k,l)}_{\boldsymbol{\vartheta},L} = \frac{1}{L}\sum_{n=1}^{L}\sum_{t=1}^{L} \mathbb{C}\mathrm{ov}\left(\ell^{(k)}_{\boldsymbol{\vartheta},n}, \ell^{(l)}_{\boldsymbol{\vartheta},t}\right), \tag{2.29}$$

converges to a limit as L tends to infinity, where $\ell^{(m)}_{\boldsymbol{\vartheta},n} = \partial_m l_{\boldsymbol{\vartheta},n}$ is the partial derivative of the nth term in expression (2.10) for $\mathscr{L}(\boldsymbol{\vartheta}, \boldsymbol{y}^L)$. It follows from well-known differentiation rules for matrix functions (see, e. g. Horn and Johnson, 1994, Sections 6.5 and 6.6) that

$$\ell^{(m)}_{\boldsymbol{\vartheta},n} = \mathrm{tr}\left[V_{\boldsymbol{\vartheta}}^{-1}\left(\mathbf{1}_d - \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n} V_{\boldsymbol{\vartheta}}^{-1}\right)(\partial_m V_{\boldsymbol{\vartheta}})\right] + 2\left(\partial_m \boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n}\right) V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}. \tag{2.30}$$

By the assumed stationarity of the processes $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$, the covariances in the sum (2.29) depend only on the difference $n-t$. For the proof of the lemma it suffices to show that the sequence $\mathfrak{c}^{(k,l)}_{\boldsymbol{\vartheta},\Delta} = \mathbb{C}\mathrm{ov}\left(\ell^{(k)}_{\boldsymbol{\vartheta},n}, \ell^{(l)}_{n+\Delta,\boldsymbol{\vartheta}}\right)$, $\Delta \in \mathbb{Z}$, is absolutely summable for all $k, l = 1, \ldots, r$, because then the Dominated Convergence Theorem implies that

$$I^{(k,l)}_{\boldsymbol{\vartheta},L} = \frac{1}{L}\sum_{\Delta=-L}^{L}(L - |\Delta|)\,\mathfrak{c}^{(k,l)}_{\boldsymbol{\vartheta},\Delta} \xrightarrow[L\to\infty]{} \sum_{\Delta\in\mathbb{Z}} \mathfrak{c}^{(k,l)}_{\boldsymbol{\vartheta},\Delta} < \infty. \tag{2.31}$$

In view of the of the symmetry $\mathfrak{c}^{(k,l)}_{\boldsymbol{\vartheta},\Delta} = \mathfrak{c}^{(k,l)}_{\boldsymbol{\vartheta},-\Delta}$, it is no restriction to assume that $\Delta \in \mathbb{N}$. In order to show that $\sum_\Delta \left|\mathfrak{c}^{(k,l)}_{\boldsymbol{\vartheta},\Delta}\right|$ is finite, we first use the bilinearity of $\mathbb{C}\mathrm{ov}(\cdot;\cdot)$ to estimate

$$\begin{aligned}
\left|\mathfrak{c}^{(k,l)}_{\boldsymbol{\vartheta},\Delta}\right| \leqslant & 4\left|\mathbb{C}\mathrm{ov}\left(\left(\partial_k \boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n}\right) V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}; \left(\partial_l \boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n+\Delta}\right) V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n+\Delta}\right)\right| \\
& + \left|\mathbb{C}\mathrm{ov}\left(\mathrm{tr}\left[V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n}V_{\boldsymbol{\vartheta}}^{-1}\partial_k V_{\boldsymbol{\vartheta}}\right]; \mathrm{tr}\left[V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n+\Delta}\boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n+\Delta}V_{\boldsymbol{\vartheta}}^{-1}\partial_l V_{\boldsymbol{\vartheta}}\right]\right)\right| + \\
& + 2\left|\mathbb{C}\mathrm{ov}\left(\mathrm{tr}\left[V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n}V_{\boldsymbol{\vartheta}}^{-1}\partial_k V_{\boldsymbol{\vartheta}}\right]; \left(\partial_l \boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n+\Delta}\right) V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n+\Delta}\right)\right| + \\
& + 2\left|\mathbb{C}\mathrm{ov}\left(\left(\partial_k \boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n}\right) V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}; \mathrm{tr}\left[V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n+\Delta}\boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n+\Delta}V_{\boldsymbol{\vartheta}}^{-1}\partial_l V_{\boldsymbol{\vartheta}}\right]\right)\right|.
\end{aligned}$$

Each of these four terms can be analysed separately. We give details only for the first one, the arguments for the other three terms being similar. Using the moving average representations (2.20) and (2.26) for $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$, $\partial_k \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$ and $\partial_l \boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}}$, it follows that

$$\begin{aligned}
&\left|\mathbb{C}\mathrm{ov}\left(\left(\partial_k \boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n}\right) V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}; \left(\partial_l \boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta},n+\Delta}\right) V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n+\Delta}\right)\right| \\
&= \sum_{\nu,\nu',\mu,\mu'=0}^{\infty}\left|\mathbb{C}\mathrm{ov}\left(\boldsymbol{Y}^T_{n-\nu}c^{(k),T}_{\boldsymbol{\vartheta},\nu}V_{\boldsymbol{\vartheta}}^{-1}c_{\boldsymbol{\vartheta},\nu'}\boldsymbol{Y}_{n-\nu'}, \boldsymbol{Y}^T_{n+\Delta-\mu}c^{(l),T}_{\boldsymbol{\vartheta},\mu}V_{\boldsymbol{\vartheta}}^{-1}c_{\boldsymbol{\vartheta},\mu'}\boldsymbol{Y}_{n+\Delta-\mu'}\right)\right|.
\end{aligned}$$

This sum can be split into one part I^+ in which at least one of the summation indices ν, ν', μ and μ' exceeds $\Delta/2$, and one part I^- in which all summation indices are less than or equal to $\Delta/2$. Using the fact that, by the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\left|\mathbb{C}\mathrm{ov}\left(\boldsymbol{Y}^T_{n-\nu}c^{(k),T}_{\boldsymbol{\vartheta},\nu}V_{\boldsymbol{\vartheta}}^{-1}c_{\boldsymbol{\vartheta},\nu'}\boldsymbol{Y}_{n-\nu'}; \boldsymbol{Y}^T_{n+\Delta-\mu}c^{(l),T}_{\boldsymbol{\vartheta},\mu}V_{\boldsymbol{\vartheta}}^{-1}c_{\boldsymbol{\vartheta},\mu'}\boldsymbol{Y}_{n+\Delta-\mu'}\right)\right| \\
&\leqslant \left\|V_{\boldsymbol{\vartheta}}^{-1}\right\|^2\left\|c^{(k)}_{\boldsymbol{\vartheta},\nu}\right\|\left\|c_{\boldsymbol{\vartheta},\nu'}\right\|\left\|c^{(l)}_{\boldsymbol{\vartheta},\mu'}\right\|\left\|c_{\boldsymbol{\vartheta},\mu'}\right\|\mathbb{E}\left\|\boldsymbol{Y}_n\right\|^4,
\end{aligned}$$

it follows from Assumption D8 and the uniform exponential decay of $\left\|c_{\boldsymbol{\vartheta},\nu}\right\|$ and $\left\|c_{\boldsymbol{\vartheta},\nu}^{(m)}\right\|$ proved in Lemma 2.7, ii) and Lemma 2.12, ii) that there exist constants C and $\rho<1$ such that

$$I^{+}=\sum_{\substack{\nu,\nu',\mu,\mu'=0\\ \max\{\nu,\nu',\mu,\mu'\}>\Delta/2}}^{\infty}\left|\mathbb{C}\mathrm{ov}\left(\boldsymbol{Y}_{n-\nu}^{T}c_{\boldsymbol{\vartheta},\nu}^{(k),T}V_{\boldsymbol{\vartheta}}^{-1}c_{\boldsymbol{\vartheta},\nu'}\boldsymbol{Y}_{n-\nu'},\boldsymbol{Y}_{n+\Delta-\mu}^{T}c_{\boldsymbol{\vartheta},\mu}^{(l),T}V_{\boldsymbol{\vartheta}}^{-1}c_{\boldsymbol{\vartheta},\mu'}\boldsymbol{Y}_{n+\Delta-\mu'}\right)\right|\leqslant C\rho^{\Delta/2}. \tag{2.32}$$

For the contribution from all indices smaller than or equal to $\Delta/2$, Lemma 2.14 implies that

$$I^{-}=\sum_{\nu,\nu',\mu,\mu'=0}^{\lfloor\Delta/2\rfloor}\left|\mathbb{C}\mathrm{ov}\left(\boldsymbol{Y}_{n-\nu}^{T}c_{\boldsymbol{\vartheta},\nu}^{(k),T}V_{\boldsymbol{\vartheta}}^{-1}c_{\boldsymbol{\vartheta},\nu'}\boldsymbol{Y}_{n-\nu'},\boldsymbol{Y}_{n+\Delta-\mu}^{T}c_{\boldsymbol{\vartheta},\mu}^{(l),T}V_{\boldsymbol{\vartheta}}^{-1}c_{\boldsymbol{\vartheta},\mu'}\boldsymbol{Y}_{n+\Delta-\mu'}\right)\right|\leqslant C\left[\alpha_{\boldsymbol{Y}}\left(\left\lfloor\frac{\Delta}{2}\right\rfloor\right)\right]^{\delta/(2+\delta)}. \tag{2.33}$$

It thus follows from Assumption D9 that the sequences $\left|c_{\boldsymbol{\vartheta},\Delta}^{(k,l)}\right|$, $\Delta\in\mathbb{N}$, are summable, and Eq. (2.31) completes the proof of the lemma. □

Lemma 2.16. *Let $\mathscr{L}$ and $\widehat{\mathscr{L}}$ be given by Eqs.* (2.10) *and* (2.14)*. Assume that Assumptions D1 to D3 and D7 are satisfied. Then the following hold.*

i) For each $m=1,\ldots,r$, the sequence $L^{-1/2}\sup_{\boldsymbol{\vartheta}\in\Theta}\left|\partial_m\widehat{\mathscr{L}}(\boldsymbol{\vartheta},\boldsymbol{y}^L)-\partial_m\mathscr{L}(\boldsymbol{\vartheta},\boldsymbol{y}^L)\right|$ converges to zero in probability, as $L\to\infty$.

ii) For all $k,l=1,\ldots,r$, the sequence $L^{-1}\sup_{\boldsymbol{\vartheta}\in\Theta}\left|\partial_{k,l}^2\widehat{\mathscr{L}}(\boldsymbol{\vartheta},\boldsymbol{y}^L)-\partial_{k,l}^2\mathscr{L}(\boldsymbol{\vartheta},\boldsymbol{y}^L)\right|$ converges to zero almost surely, as $L\to\infty$.

Proof. Similar to the proof of Lemma 2.8. □

Lemma 2.17. *Under Assumptions D1, D3 and D7 to D9, the random variable $L^{-1/2}\nabla_{\boldsymbol{\vartheta}}\widehat{\mathscr{L}}(\boldsymbol{\vartheta}_0,\boldsymbol{y}^L)$ is asymptotically normally distributed with mean zero and covariance matrix $I(\boldsymbol{\vartheta}_0)$.*

Proof. Because of Lemma 2.16, i) it is enough to show that $L^{-1/2}\nabla_{\boldsymbol{\vartheta}}\mathscr{L}\left(\boldsymbol{\vartheta}_0,\boldsymbol{y}^L\right)$ is asymptotically normally distributed with mean zero and covariance matrix $I(\boldsymbol{\vartheta}_0)$. We begin the proof by recalling the equation

$$\partial_i\mathscr{L}(\boldsymbol{\vartheta},\boldsymbol{y}^L)=\sum_{n=1}^{L}\left\{\mathrm{tr}\left[V_{\boldsymbol{\vartheta}}^{-1}\left(\mathbf{1}_d-\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^{T}V_{\boldsymbol{\vartheta}}^{-1}\right)\partial_iV_{\boldsymbol{\vartheta}}\right]+2\left(\partial_i\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^{T}\right)V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right\}, \tag{2.34}$$

which holds for every component $i=1,\ldots,r$. The facts that $\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}^{T}$ equals $V_{\boldsymbol{\vartheta}_0}$, and that $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}$ is orthogonal to the Hilbert space generated by $\{\boldsymbol{Y}_t,t<n\}$, of which $\partial_i\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^{T}$ is an element, show that $\mathbb{E}\partial_i\mathscr{L}\left(\boldsymbol{\vartheta}_0,\boldsymbol{y}^L\right)=0$. Using Eq. (2.20), expression (2.34) can be rewritten as

$$\partial_i\mathscr{L}\left(\boldsymbol{\vartheta}_0,\boldsymbol{y}^L\right)=\sum_{n=1}^{L}\left[Y_{m,n}^{(i)}-\mathbb{E}Y_{m,n}^{(i)}\right]+\sum_{n=1}^{L}\left[Z_{m,n}^{(i)}-\mathbb{E}Z_{m,n}^{(i)}\right],$$

where, for every $m\in\mathbb{N}$, the processes $Y_m^{(i)}$ and $Z_m^{(i)}$ are defined by

$$Y_{m,n}^{(i)}=\mathrm{tr}\left[V_{\boldsymbol{\vartheta}_0}^{-1}(\partial_iV_{\boldsymbol{\vartheta}_0})\right]+\sum_{\nu,\nu'=0}^{m}\left\{-\mathrm{tr}\left[V_{\boldsymbol{\vartheta}_0}^{-1}c_{\boldsymbol{\vartheta}_0,\nu}\boldsymbol{Y}_{n-\nu}\boldsymbol{Y}_{n-\nu'}^{T}c_{\boldsymbol{\vartheta},\nu'}^{T}V_{\boldsymbol{\vartheta}_0}^{-1}(\partial_iV_{\boldsymbol{\vartheta}_0})\right]+2\boldsymbol{Y}_{n-\nu}^{T}c_{\boldsymbol{\vartheta}_0,\nu}^{(i),T}V_{\boldsymbol{\vartheta}_0}^{-1}c_{\boldsymbol{\vartheta}_0,\nu'}\boldsymbol{Y}_{n-\nu'}\right\}, \tag{2.35a}$$

$$Z_{m,n}^{(i)}=U_{m,n}^{(i)}+V_{m,n}^{(i)}, \tag{2.35b}$$

and

$$U_{m,n}^{(i)}=\sum_{\nu=0}^{\infty}\sum_{\nu'=m+1}^{\infty}\left\{-\mathrm{tr}\left[V_{\boldsymbol{\vartheta}_0}^{-1}c_{\boldsymbol{\vartheta}_0,\nu}\boldsymbol{Y}_{n-\nu}\boldsymbol{Y}_{n-\nu'}^{T}c_{\boldsymbol{\vartheta},\nu'}^{T}V_{\boldsymbol{\vartheta}_0}^{-1}(\partial_iV_{\boldsymbol{\vartheta}_0})\right]+2\boldsymbol{Y}_{n-\nu}^{T}c_{\boldsymbol{\vartheta}_0,\nu}^{(i),T}V_{\boldsymbol{\vartheta}_0}^{-1}c_{\boldsymbol{\vartheta}_0,\nu'}\boldsymbol{Y}_{n-\nu'}\right\},$$

$$V_{m,n}^{(i)}=\sum_{\nu=m+1}^{\infty}\sum_{\nu'=0}^{m}\left\{-\mathrm{tr}\left[V_{\boldsymbol{\vartheta}_0}^{-1}c_{\boldsymbol{\vartheta}_0,\nu}\boldsymbol{Y}_{n-\nu}\boldsymbol{Y}_{n-\nu'}^{T}c_{\boldsymbol{\vartheta},\nu'}^{T}V_{\boldsymbol{\vartheta}_0}^{-1}(\partial_iV_{\boldsymbol{\vartheta}_0})\right]+2\boldsymbol{Y}_{n-\nu}^{T}c_{\boldsymbol{\vartheta}_0,\nu}^{(i),T}V_{\boldsymbol{\vartheta}_0}^{-1}c_{\boldsymbol{\vartheta}_0,\nu'}\boldsymbol{Y}_{n-\nu'}\right\}.$$

It is convenient to also introduce the notations

$$\mathcal{Y}_{m,n}=\begin{pmatrix}Y_{m,n}^{(1)}&\cdots&Y_{m,n}^{(r)}\end{pmatrix}^{T}\quad\text{and}\quad\mathcal{Z}_{m,n}=\begin{pmatrix}Z_{m,n}^{(1)}&\cdots&Z_{m,n}^{(r)}\end{pmatrix}^{T}. \tag{2.36}$$

The rest of the proof proceeds in three steps: first we show that, for each natural number m, the sequence $L^{-1/2}\sum_n[\mathcal{Y}_{m,n}-\mathbb{E}\mathcal{Y}_{m,n}]$ is asymptotically normally distributed with asymptotic covariance matrix I_m, and that I_m converges to $I(\boldsymbol{\vartheta}_0)$ as m tends to infinity. In the second step we prove that $L^{-1/2}\sum_n[\mathcal{Z}_{m,n}-\mathbb{E}\mathcal{Z}_{m,n}]$ goes to zero uniformly in L, as $m\to\infty$, and the last step is devoted to combining the first two steps to prove the asymptotic normality of $L^{-1/2}\nabla_{\boldsymbol{\vartheta}}\mathscr{L}\left(\boldsymbol{\vartheta}_0,\boldsymbol{y}^L\right)$.

Step 1. Since $\boldsymbol{Y}$ is stationary, it is clear that $\mathcal{Y}_m$ is a stationary process. Moreover, the strong mixing coefficients $\alpha_{\mathcal{Y}_m}(k)$ of $\mathcal{Y}_m$ satisfy $\alpha_{\mathcal{Y}_m}(k)\leqslant\alpha_{\boldsymbol{Y}}(\max\{0,k-m\})$ because $\mathcal{Y}_{m,n}$ depends only on the finitely many values $\boldsymbol{Y}_{n-m},\dots,\boldsymbol{Y}_n$ of $\boldsymbol{Y}$ (see Bradley, 2007, Remark 1.8 b)). In particular, by Assumption D9, the strong mixing coefficients of the processes $\mathcal{Y}_m$ satisfy the summability condition $\sum_k[\alpha_{\mathcal{Y}_m}(k)]^{\delta/(2+\delta)}<\infty$. Since, by the Cramér–Wold device, weak convergence of the sequence $L^{-1/2}\sum_{n=1}^L[\mathcal{Y}_{m,n}-\mathbb{E}\mathcal{Y}_{m,n}]$ to a multivariate normal distribution with mean zero and covariance matrix Σ is equivalent to the condition that, for every vector $\boldsymbol{u}\in\mathbb{R}^r$, the sequence $L^{-1/2}\boldsymbol{u}^T\sum_{n=1}^L[\mathcal{Y}_{m,n}-\mathbb{E}\mathcal{Y}_{m,n}]$ converges to a one-dimensional normal distribution with mean zero and variance $\boldsymbol{u}^T\Sigma\boldsymbol{u}$, we can apply the Central Limit Theorem for univariate strongly mixing processes (Herrndorf, 1984),(Ibragimov, 1962, Theorem 1.7) to obtain that

$$\frac{1}{\sqrt{L}}\sum_{n=1}^{L}[\mathcal{Y}_{m,n}-\mathbb{E}\mathcal{Y}_{m,n}]\xrightarrow[L\to\infty]{d}\mathscr{N}(\boldsymbol{0}_r,I_m),\quad\text{where}\quad I_m=\sum_{\Delta\in\mathbb{Z}}\mathbb{C}\mathrm{ov}\left(\mathcal{Y}_{m,n};\mathcal{Y}_{m,n+\Delta}\right).\tag{2.37}$$

The claim that I_m converges to $I(\boldsymbol{\vartheta}_0)$ will follow if we can show that

$$\mathbb{C}\mathrm{ov}\left(Y_{m,n}^{(k)};Y_{m,n+\Delta}^{(l)}\right)\xrightarrow[m\to\infty]{}\mathbb{C}\mathrm{ov}\left(\ell_{\boldsymbol{\vartheta}_0,n}^{(k)};\ell_{\boldsymbol{\vartheta}_0,n+\Delta}^{(l)}\right),\quad\forall\Delta\in\mathbb{Z},\tag{2.38}$$

and that $\left|\mathbb{C}\mathrm{ov}\left(Y_{m,n}^{(k)};Y_{m,n+\Delta}^{(l)}\right)\right|$ is dominated by an absolutely summable sequence. For the first condition, we note that the bilinearity of $\mathbb{C}\mathrm{ov}(\cdot;\cdot)$ implies that

$$\mathbb{C}\mathrm{ov}\left(Y_{m,n}^{(k)};Y_{m,n+\Delta}^{(l)}\right)-\mathbb{C}\mathrm{ov}\left(\ell_{\boldsymbol{\vartheta}_0,n}^{(k)};\ell_{\boldsymbol{\vartheta}_0,n+\Delta}^{(l)}\right)=\mathbb{C}\mathrm{ov}\left(Y_{m,n}^{(k)};Y_{m,n+\Delta}^{(l)}-\ell_{\boldsymbol{\vartheta}_0,n+\Delta}^{(l)}\right)+\mathbb{C}\mathrm{ov}\left(Y_{m,n}^{(k)}-\ell_{\boldsymbol{\vartheta}_0,n}^{(k)};\ell_{\boldsymbol{\vartheta}_0,n+\Delta}^{(l)}\right).$$

These two terms can be treated in a similar manner so we restrict our attention to the second one. The definitions of $Y_{m,n}^{(i)}$ (Eq. (2.35a)) and $\ell_{\boldsymbol{\vartheta},n}^{(i)}$ (Eq. (2.29)) allow us to compute

$$Y_{m,n}^{(k)}-\ell_{\boldsymbol{\vartheta}_0,n}^{(k)}=\sum_{\substack{\nu,\nu'\\\max\{\nu,\nu'\}>m}}\left[\mathrm{tr}\left[V_{\boldsymbol{\vartheta}_0}^{-1}c_{\boldsymbol{\vartheta}_0,\nu}\boldsymbol{Y}_{n-\nu}\boldsymbol{Y}_{n-\nu'}^Tc_{\boldsymbol{\vartheta},\nu'}^TV_{\boldsymbol{\vartheta}_0}^{-1}\partial_iV_{\boldsymbol{\vartheta}_0}\right]-2\boldsymbol{Y}_{n-\nu}^Tc_{\boldsymbol{\vartheta}_0,\nu}^{(i),T}V_{\boldsymbol{\vartheta}_0}^{-1}c_{\boldsymbol{\vartheta}_0,\nu'}\boldsymbol{Y}_{n-\nu'}\right].$$

As a consequence of the Cauchy–Schwarz inequality, Assumption D8 and the exponential bounds in Lemma 2.7, i), we therefore obtain that $\mathbb{V}\mathrm{ar}\left(Y_{m,n}^{(k)}-\ell_{\boldsymbol{\vartheta}_0,n}^{(k)}\right)\leqslant C\rho^m$ independent of n. The L^2-continuity of $\mathbb{C}\mathrm{ov}(\cdot;\cdot)$ thus implies that the sequence $\mathbb{C}\mathrm{ov}\left(Y_{m,n}^{(k)}-\ell_{\boldsymbol{\vartheta}_0,n}^{(k)};\ell_{\boldsymbol{\vartheta}_0,n+\Delta}^{(l)}\right)$ converges to zero as m tends to infinity at an exponential rate uniformly in Δ. The existence of a summable sequence dominating $\left|\mathbb{C}\mathrm{ov}\left(Y_{m,n}^{(k)};Y_{m,n+\Delta}^{(l)}\right)\right|$ is ensured by the arguments given in the proof of Lemma 2.15, reasoning as in the derivation of Eqs. (2.32) and (2.33).

Step 2. In this step we shall show that there exist positive constants C and $\rho<1$, independent of L, such that

$$\mathrm{tr}\,\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{L}}\sum_{n=1}^{L}\mathcal{Z}_{m,n}\right)\leqslant C\rho^m,\quad\mathcal{Z}_{m,n}\text{ given in Eq. (2.36).}\tag{2.39}$$

Since

$$\mathrm{tr}\,\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{L}}\sum_{n=1}^{L}\mathcal{Z}_{m,n}\right)\leqslant2\left[\mathrm{tr}\,\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{L}}\sum_{n=1}^{L}\mathcal{U}_{m,n}\right)+\mathrm{tr}\,\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{L}}\sum_{n=1}^{L}\mathcal{V}_{m,n}\right)\right],\tag{2.40}$$

it suffices to consider the latter two terms. We first observe that

$$\begin{aligned}\mathrm{tr}\,\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{L}}\sum_{n=1}^{L}\mathcal{U}_{m,n}\right)&=\frac{1}{L}\,\mathrm{tr}\sum_{n,n'=1}^{L}\mathbb{C}\mathrm{ov}\left(\mathcal{U}_{m,n};\mathcal{U}_{m,n'}\right)\\&=\frac{1}{L}\sum_{k,l=1}^{r}\sum_{\Delta=-L+1}^{L-1}(L-|\Delta|)\,\mathfrak{u}_{m,\Delta}^{(k,l)}\leqslant\sum_{k,l=1}^{r}\sum_{\Delta\in\mathbb{Z}}\left|\mathfrak{u}_{m,\Delta}^{(k,l)}\right|,\end{aligned}\tag{2.41}$$

where

$$\begin{aligned}\mathfrak{u}_{m,\Delta}^{(k,l)} &= \mathbb{C}\mathrm{ov}\left(U_{m,n}^{(k)}; U_{m,n+\Delta}^{(l)}\right)\\ &= \sum_{\substack{\nu,\mu=0\\ \nu',\mu'=m+1}}^{m} \mathbb{C}\mathrm{ov}\Big(-\operatorname{tr}\left[V_{\vartheta_0}^{-1} c_{\vartheta_0,\nu} \boldsymbol{Y}_{n-\nu} \boldsymbol{Y}_{n-\nu'}^T c_{\vartheta,\nu'}^T V_{\vartheta_0}^{-1} \partial_k V_{\vartheta_0}\right] + \boldsymbol{Y}_{n-\nu}^T c_{\vartheta_0,\nu}^{(k),T} V_{\vartheta_0}^{-1} c_{\vartheta_0,\nu'} \boldsymbol{Y}_{n-\nu'} ;\\ &\qquad - \operatorname{tr}\left[V_{\vartheta_0}^{-1} c_{\vartheta_0,\mu} \boldsymbol{Y}_{n+\Delta-\mu} \boldsymbol{Y}_{n+\Delta-\mu'}^T c_{\vartheta,\mu'}^T V_{\vartheta_0}^{-1} \partial_l V_{\vartheta_0}\right] + \boldsymbol{Y}_{n+\Delta-\mu}^T c_{\vartheta_0,\mu}^{(l),T} V_{\vartheta_0}^{-1} c_{\vartheta_0,\mu'} \boldsymbol{Y}_{n+\Delta-\mu'}\Big).\end{aligned}$$

As before, under Assumption D8, the Cauchy–Schwarz inequality and the exponential bounds for $\left\|c_{\vartheta_0,\nu}\right\|$ and $\left\|c_{\vartheta_0,\nu}^{(k)}\right\|$ imply that $\left|\mathfrak{u}_{m,\Delta}^{(k,l)}\right| < C\rho^m$. By arguments similar to the ones used in the proof of Lemma 2.14 it can be shown that Davydov's inequality implies that for $m < \lfloor\Delta/2\rfloor$ it holds that

$$\begin{aligned}\left|\mathfrak{u}_{m,\Delta}^{(k,l)}\right| &\leqslant C \sum_{\nu=0}^{\infty} \sum_{\nu'=m+1}^{\infty} \sum_{\mu,\mu'=0}^{\lfloor\Delta/2\rfloor} \rho^{\nu+\nu'+\mu+\mu'} \left[\alpha_{\boldsymbol{Y}}\left(\left\lfloor\frac{\Delta}{2}\right\rfloor\right)\right]^{\delta/(2+\delta)} + C \sum_{\nu,\nu'=0}^{\infty} \sum_{\substack{\mu,\mu'\\ \max\{\mu,\mu'\}>\lfloor\Delta/2\rfloor}} \rho^{\nu+\nu'+\mu+\mu'}\\ &\leqslant C\rho^m \left\{\left[\alpha_{\boldsymbol{Y}}\left(\left\lfloor\frac{\Delta}{2}\right\rfloor\right)\right]^{\delta/(2+\delta)} + \rho^{\Delta/2}\right\}.\end{aligned}$$

It thus follows that, independent of the value of k and l,

$$\sum_{\Delta=0}^{\infty} \left|\mathfrak{u}_{m,\Delta}^{(k,l)}\right| = \sum_{\Delta=0}^{2m} \left|\mathfrak{u}_{m,\Delta}^{(k,l)}\right| + \sum_{\Delta=2m+1}^{\infty} \left|\mathfrak{u}_{m,\Delta}^{(k,l)}\right| \leqslant C\rho^m \left\{m + \sum_{\Delta=0}^{\infty} \left[\alpha_{\boldsymbol{Y}}(\Delta)\right]^{\delta/(2+\delta)}\right\},$$

and therefore, by Eq. (2.41), that $\operatorname{tr}\mathbb{V}\mathrm{ar}\left(L^{-1/2}\sum_{n=1}^{L}\mathcal{U}_{m,n}\right) \leqslant C\rho^m$. In an analogous way one can show that $\operatorname{tr}\mathbb{V}\mathrm{ar}\left(L^{-1/2}\sum_{n=1}^{L}\mathcal{V}_{m,n}\right) \leqslant C\rho^m$, and thus the claim (2.39) follows with Eq. (2.40).

Step 3. In step 1 it has been shown that $L^{-1/2}\sum_n\left[\mathcal{Y}_{m,n} - \mathbb{E}\mathcal{Y}_{m,n}\right] \xrightarrow[L\to\infty]{d} \mathcal{N}(\mathbf{0}_r, I_m)$, and that $I_m \xrightarrow[m\to\infty]{} I(\vartheta_0)$. In particular, the limiting normal random variables with covariances I_m converge weakly to a normal random variable with covariance matrix $I(\vartheta_0)$. Step 2 together with the multivariate Chebyshev inequality implies that, for every $\epsilon > 0$,

$$\begin{aligned}&\lim_{m\to\infty}\limsup_{L\to\infty}\mathbb{P}\left(\left\|\frac{1}{\sqrt{L}}\nabla_{\vartheta}\mathscr{L}\left(\vartheta_0, \boldsymbol{y}^L\right) - \frac{1}{\sqrt{L}}\sum_{n=1}^{L}\left[\mathcal{Y}_{m,n} - \mathbb{E}\mathcal{Y}_{m,n}\right]\right\| > \epsilon\right)\\ &\leqslant \lim_{m\to\infty}\limsup_{L\to\infty}\frac{r}{\epsilon^2}\operatorname{tr}\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{L}}\sum_{n=1}^{L}\mathcal{Z}_{m,n}\right) \leqslant \lim_{m\to\infty}\frac{Cr}{\epsilon^2}\rho^m = 0.\end{aligned}$$

Brockwell and Davis (1991, Proposition 6.3.9) thus completes the proof. □

A very important step in the proof of asymptotic normality for quasi maximum likelihood estimators is to establish that the Fisher information matrix J, evaluated at the true parameter value, is non-singular. We shall now show that Assumption D10 is sufficient to ensure that J^{-1} exists for linear state space models. For vector ARMA processes, formulae similar to Eqs. (2.43) below have been derived in the literature (see, e. g., Klein et al., 2008; Klein and Neudecker, 2000), but have not been used to derive criteria for J being non-singular. Our arguments are similar to Boubacar Mainassara and Francq (2011, Lemma 4).

Lemma 2.18. *Assume that Assumptions D1 to D4, D7 and D10 hold. With probability one, the matrix* $J = \lim_{L\to\infty} L^{-1}\nabla_{\vartheta}^2\widehat{\mathscr{L}}(\vartheta_0, \boldsymbol{y}^L)$ *exists and is non-singular.*

Proof. We note that, by Lemma 2.16, ii), it is enough to show that $\lim_{L\to\infty} L^{-1}\nabla_{\vartheta}^2\mathscr{L}(\vartheta_0, \boldsymbol{y}^L)$ exists and is non-singular. As seen earlier, for every $i = 1, \ldots, r$,

$$\partial_i l_{\vartheta,n} = \operatorname{tr}\left[V_{\vartheta}^{-1}\left(\mathbf{1}_d - \boldsymbol{\varepsilon}_{\vartheta,n}\boldsymbol{\varepsilon}_{\vartheta,n}^T V_{\vartheta}^{-1}\right)\partial_i V_{\vartheta}\right] + 2\left(\partial_i\boldsymbol{\varepsilon}_{\vartheta,n}^T\right)V_{\vartheta}^{-1}\boldsymbol{\varepsilon}_{\vartheta,n}. \tag{2.42}$$

Consequently, the second partial derivatives are given by

$$\begin{aligned}\partial_{i,j}^2 l_{\boldsymbol{\vartheta},n} = \operatorname{tr}\Big[&V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_{i,j}^2 V_{\boldsymbol{\vartheta}}\right) - V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_i V_{\boldsymbol{\vartheta}}\right)V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_j V_{\boldsymbol{\vartheta}}\right) - V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_{i,j}^2 V_{\boldsymbol{\vartheta}}\right)\\
&+V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_i V_{\boldsymbol{\vartheta}}\right)V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_j V_{\boldsymbol{\vartheta}}\right) + V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_i V_{\boldsymbol{\vartheta}}\right)V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_j V_{\boldsymbol{\vartheta}}\right)\\
&-V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_i V_{\boldsymbol{\vartheta}}\right)V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_i\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T\right)\Big] + 2\left(\partial_{i,j}^2\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T\right)V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n} + 2\left(\partial_i\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T\right)V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_j\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right)\\
&- 2\operatorname{tr}\left[V_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\left(\partial_i\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T\right)V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_j\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right)\right].\end{aligned}$$

By Lemma 2.1, iii), $\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n} = \mathbf{0}_d$, and by Eq. (2.7), $\mathbb{E}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}^T = V_{\boldsymbol{\vartheta}_0}$. The sequence $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0}$ being the innovations of the process $\boldsymbol{Y}$ implies that $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}$ is orthogonal to the Hilbert space spanned by $\{\boldsymbol{Y}_t, t<n\}$, of which, by Eq. (2.20), both $\partial_i\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}$ and $\partial_{i,j}^2\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}$ are elements. It thus follows that

$$\mathbb{E}\left[\partial_{i,j}^2 l_{\boldsymbol{\vartheta}_0,n}\right] = \operatorname{tr}\left[V_{\boldsymbol{\vartheta}_0}^{-1}\left(\partial_i V_{\boldsymbol{\vartheta}_0}\right)V_{\boldsymbol{\vartheta}_0}^{-1}\left(\partial_j V_{\boldsymbol{\vartheta}_0}\right)\right] + 2\mathbb{E}\left[\left(\partial_i\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}^T\right)V_{\boldsymbol{\vartheta}}^{-1}\left(\partial_j\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta},n}\right)\right].$$

Equations (2.20), (2.26) and (2.27), the ergodicity of $\boldsymbol{Y}$, and Krengel (1985, Theorem 4.3) imply that the sequence $\partial_{i,j}^2 l_{\boldsymbol{\vartheta}_0}$ is ergodic, and Birkhoff's Ergodic Theorem shows that

$$\frac{1}{L}\nabla_{\boldsymbol{\vartheta}}^2\mathscr{L}\left(\boldsymbol{\vartheta}_0,\boldsymbol{y}^L\right) = \frac{1}{L}\sum_{n=1}^L\nabla_{\boldsymbol{\vartheta}}^2 l_{\boldsymbol{\vartheta}_0,n}\xrightarrow[L\to\infty]{\text{a.s}}\mathbb{E}\left[\nabla_{\boldsymbol{\vartheta}}^2 l_{\boldsymbol{\vartheta}_0,n}\right] =: J_1 + J_2,$$

where

$$J_1 = 2\mathbb{E}\left[\left(\nabla_{\boldsymbol{\vartheta}}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\right)^T V_{\boldsymbol{\vartheta}_0}^{-1}\left(\nabla_{\boldsymbol{\vartheta}}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,1}\right)\right]\quad\text{and}\quad J_2 = \left(\operatorname{tr}\left[V_{\boldsymbol{\vartheta}_0}^{-1/2}\left(\partial_i V_{\boldsymbol{\vartheta}_0}\right)V_{\boldsymbol{\vartheta}_0}^{-1}\left(\partial_j V_{\boldsymbol{\vartheta}_0}\right)V_{\boldsymbol{\vartheta}_0}^{-1/2}\right]\right)_{ij}. \tag{2.43}$$

J_2 is positive semidefinite because it can be written as $J_2 = \begin{pmatrix}\boldsymbol{b}_1 & \ldots & \boldsymbol{b}_r\end{pmatrix}^T\begin{pmatrix}\boldsymbol{b}_1 & \ldots & \boldsymbol{b}_r\end{pmatrix}$, where $\boldsymbol{b}_m = \left(V_{\boldsymbol{\vartheta}_0}^{-1/2}\otimes V_{\boldsymbol{\vartheta}_0}^{-1/2}\right)\operatorname{vec}\left(\partial_m V_{\boldsymbol{\vartheta}_0}\right)$. Since J_1 is positive semidefinite as well, proving that J is non-singular is equivalent to proving that for any non-zero vector $\boldsymbol{c}\in\mathbb{R}^r$, the numbers $\boldsymbol{c}^T J_i\boldsymbol{c}$, $i=1,2$, are not both zero. Assume, for the sake of contradiction, that there exists such a vector $\boldsymbol{c} = (c_1,\ldots,c_r)^T$. The condition $\boldsymbol{c}^T J_1\boldsymbol{c}$ implies that, almost surely, $\sum_{k=1}^r c_k\partial_k\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n} = \mathbf{0}_d$ for all $n\in\mathbb{Z}$. It thus follows from the infinite-order moving average representation (2.8b) that

$$\sum_{\nu=1}^{\infty}\sum_{k=1}^{r} c_k\left(\partial_k\mathscr{M}_{\boldsymbol{\vartheta}_0,\nu}\right)\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,-\nu} = \mathbf{0}_d,$$

where the Markov parameters $\mathscr{M}_{\boldsymbol{\vartheta},\nu}$ are given by $\mathscr{M}_{\boldsymbol{\vartheta},\nu} = -H_{\boldsymbol{\vartheta}}F_{\boldsymbol{\vartheta}}^{\nu-1}K_{\boldsymbol{\vartheta}}$, $\nu\geqslant 1$. Since the sequence $\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0}$ is uncorrelated with positive definite covariance matrix, it follows that $\sum_{k=1}^r c_k\left(\partial_k\mathscr{M}_{\boldsymbol{\vartheta}_0,\nu}\right) = \mathbf{0}_d$, for every $\nu\in\mathbb{N}$. Using the relation $\operatorname{vec}(ABC) = \left(C^T\otimes A\right)\operatorname{vec}B$ (Bernstein, 2005, Proposition 7.1.9), we see that the last display is equivalent to $\nabla_{\boldsymbol{\vartheta}}\left(\left[K_{\boldsymbol{\vartheta}_0}^T\otimes H_{\boldsymbol{\vartheta}_0}\right]\operatorname{vec}F_{\boldsymbol{\vartheta}_0}^{\nu-1}\right)\boldsymbol{c} = \mathbf{0}_{d^2}$ for every $\nu\in\mathbb{N}$. The condition $\boldsymbol{c}^T J_2\boldsymbol{c} = 0$ implies that $\left(\nabla_{\boldsymbol{\vartheta}}\operatorname{vec}V_{\boldsymbol{\vartheta}_0}\right)\boldsymbol{c} = \mathbf{0}_{d^2}$. By the definition of $\psi_{\boldsymbol{\vartheta},j}$ in Eq. (2.17) it thus follows that $\nabla_{\boldsymbol{\vartheta}}\psi_{\boldsymbol{\vartheta}_0,j}\boldsymbol{c} = \mathbf{0}_{(j+2)d^2}$, for every $j\in\mathbb{N}$, which, by Assumption D10, is equivalent to the contradiction that $\boldsymbol{c} = \mathbf{0}_r$. □

of Theorem 2.6. Since $\hat{\boldsymbol{\vartheta}}^L$ converges almost surely to $\boldsymbol{\vartheta}_0$ by the consistency result proved in Theorem 2.5, and $\boldsymbol{\vartheta}_0$ is an element of the interior of Θ by Assumption D6, the estimate $\hat{\boldsymbol{\vartheta}}^L$ is an element of the interior of Θ eventually almost surely. The assumed smoothness of the parametrization (Assumption D7) implies that the extremal property of $\hat{\boldsymbol{\vartheta}}^L$ can be expressed as the first order condition $\nabla_{\boldsymbol{\vartheta}}\widehat{\mathscr{L}}(\hat{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L) = \mathbf{0}_r$. A Taylor expansion of $\boldsymbol{\vartheta}\mapsto\nabla_{\boldsymbol{\vartheta}}\widehat{\mathscr{L}}(\boldsymbol{\vartheta},\boldsymbol{y}^L)$ around the point $\boldsymbol{\vartheta}_0$ shows that there exist parameter vectors $\boldsymbol{\vartheta}_i\in\Theta$ of the form $\boldsymbol{\vartheta}_i = \boldsymbol{\vartheta}_0 + c_i(\hat{\boldsymbol{\vartheta}}^L - \boldsymbol{\vartheta}_0)$, $0\leqslant c_i\leqslant 1$, such that

$$\mathbf{0}_r = L^{-1/2}\nabla_{\boldsymbol{\vartheta}}\widehat{\mathscr{L}}(\boldsymbol{\vartheta}_0,\boldsymbol{y}^L) + \frac{1}{L}\nabla_{\boldsymbol{\vartheta}}^2\widehat{\mathscr{L}}(\underline{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L)L^{1/2}\left(\hat{\boldsymbol{\vartheta}}^L - \boldsymbol{\vartheta}_0\right), \tag{2.44}$$

where $\nabla_{\boldsymbol{\vartheta}}^2\widehat{\mathscr{L}}(\underline{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L)$ denotes the matrix whose ith row, $i=1,\ldots,r$, is equal to the ith row of $\nabla_{\boldsymbol{\vartheta}}^2\widehat{\mathscr{L}}(\boldsymbol{\vartheta}_i,\boldsymbol{y}^L)$. By Lemma 2.17 the first term on the right hand side converges weakly to a multivariate normal random variable with mean zero and covariance matrix $I = I(\boldsymbol{\vartheta}_0)$. As in Lemma 2.10 one can show that the sequence $\left(\boldsymbol{\vartheta}\mapsto L^{-1}\nabla_{\boldsymbol{\vartheta}}^3\widehat{\mathscr{L}}(\boldsymbol{\vartheta},\boldsymbol{y}^L)\right)_{L\in\mathbb{N}}$ of random functions converges almost surely uniformly to the continuous function $\boldsymbol{\vartheta}\mapsto\nabla_{\boldsymbol{\vartheta}}^3\mathscr{Q}(\boldsymbol{\vartheta})$ taking values in the space $\mathbb{R}^{r\times r\times r}$. Since on the compact space Θ this function is bounded in the operator norm obtained from identifying $\mathbb{R}^{r\times r\times r}$ with the space of linear functions from $\mathbb{R}^r$ to $M_r(\mathbb{R})$, that sequence is almost surely uniformly bounded, and we obtain that

$$\left\|\frac{1}{L}\nabla^2_{\boldsymbol{\vartheta}}\widehat{\mathscr{L}}(\underline{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L)-\frac{1}{L}\nabla^2_{\boldsymbol{\vartheta}}\widehat{\mathscr{L}}(\boldsymbol{\vartheta}_0,\boldsymbol{y}^L)\right\|\leqslant\sup_{\boldsymbol{\vartheta}\in\Theta}\left\|\frac{1}{L}\nabla^3_{\boldsymbol{\vartheta}}\widehat{\mathscr{L}}(\boldsymbol{\vartheta},\boldsymbol{y}^L)\right\|\left\|\underline{\boldsymbol{\vartheta}}^L-\boldsymbol{\vartheta}_0\right\|\xrightarrow[L\to\infty]{\text{a.s.}}0,$$

because, by Theorem 2.5, the second factor almost surely converges to zero as L tends to infinity. It follows from Lemma 2.18 that $L^{-1}\nabla^2_{\boldsymbol{\vartheta}}\widehat{\mathscr{L}}(\underline{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L)$ converges to the matrix J almost surely, and thus from Eq. (2.44) that $L^{1/2}\left(\hat{\boldsymbol{\vartheta}}^L-\boldsymbol{\vartheta}_0\right)\xrightarrow{d}\mathscr{N}\left(\mathbf{0}_r,J^{-1}IJ^{-1}\right)$, as $L\to\infty$. This shows Eq. (2.18) and completes the proof. □

In practice, one is interested in also estimating the asymptotic covariance matrix Ξ, which is useful in constructing confidence regions for the estimated parameters or in performing statistical tests. This problem has been considered in the framework of estimating weak VARMA processes in Boubacar Mainassara and Francq (2011) where the following procedure has been suggested, which is also applicable in our setup. First, $J(\boldsymbol{\vartheta}_0)$ is estimated consistently by $\hat{J}^L=L^{-1}\nabla^2\widehat{\mathscr{L}_{\boldsymbol{\vartheta}}}\left(\hat{\boldsymbol{\vartheta}}^L,\boldsymbol{y}^L\right)$. For the computation of $\hat{J}^L$ we rely on the fact that the Kalman filter can not only be used to evaluate the Gaussian log-likelihood of a state space model but also its gradient and Hessian. The most straightforward, but computationally burdensome way of achieving this is by direct differentiation of the Kalman filter equations, which results in increasing the number of passes through the filter to $r+1$ and $r(r+3)/2$ for the gradient and the Hessian, respectively. More sophisticated algorithms, including the Kalman smoother and/or the backward filter have been devised and can be found in Kulikova and Semoushin (2006); Segal and Weinstein (1989). The construction of a consistent estimator of $I=I(\boldsymbol{\vartheta}_0)$ is based on the observation that $I=\sum_{\Delta\in\mathbb{Z}}\mathbb{C}\mathrm{ov}(\ell_{\boldsymbol{\vartheta}_0,n},\ell_{\boldsymbol{\vartheta}_0,n+\Delta})$, where $\ell_{\boldsymbol{\vartheta}_0,n}=\nabla_{\boldsymbol{\vartheta}}\left[\log\det V_{\boldsymbol{\vartheta}_0}+\boldsymbol{\varepsilon}^T_{\boldsymbol{\vartheta}_0,n}V^{-1}_{\boldsymbol{\vartheta}_0}\boldsymbol{\varepsilon}_{\boldsymbol{\vartheta}_0,n}\right]$. Assuming that $(\ell_{\boldsymbol{\vartheta}_0,n})_{n\in\mathbb{N}^+}$ admits an infinite-order AR representation $\Phi(\mathrm{B})\ell_{\boldsymbol{\vartheta}_0,n}=\boldsymbol{U}_n$, where $\Phi(z)=\mathbf{1}_r+\sum_{i=1}^\infty\Phi_iz^i$ and $(\boldsymbol{U}_n)_{n\in\mathbb{N}^+}$ is a weak white noise with covariance matrix $\Sigma_{\boldsymbol{U}}$, it follows from the interpretation of $I/(2\pi)$ as the value of the spectral density of $(\ell_{\boldsymbol{\vartheta}_0,n})_{n\in\mathbb{N}^+}$ at frequency zero that I can also be written as $I=\Phi^{-1}(1)\Sigma_{\boldsymbol{U}}\Phi(1)^{-1}$. The idea is to fit a long autoregression to $(\ell_{\hat{\boldsymbol{\vartheta}}^L,n})_{n=1,\ldots L}$, the empirical counterparts of $(\ell_{\boldsymbol{\vartheta}_0,n})_{n\in\mathbb{N}^+}$ which are defined by replacing $\boldsymbol{\vartheta}_0$ with the estimate $\hat{\boldsymbol{\vartheta}}^L$ in the definition of $\ell_{\boldsymbol{\vartheta}_0,n}$. This is done by choosing an integer $s>0$, and performing a least-squares regression of $\ell_{\hat{\boldsymbol{\vartheta}}^L,n}$ on $\ell_{\hat{\boldsymbol{\vartheta}}^L,n-1},\ldots,\ell_{\hat{\boldsymbol{\vartheta}}^L,n-s}$, $s+1\leqslant n\leqslant L$. Denoting by $\hat{\Phi}^L_s(z)=\mathbf{1}_r+\sum_{i=1}^s\hat{\Phi}^L_{i,s}z^i$ the obtained empirical autoregressive polynomial and by $\hat{\Sigma}^L_s$ the empirical covariance matrix of the residuals of the regression, it was claimed in Boubacar Mainassara and Francq (2011, Theorem 4) that under the additional assumption $\mathbb{E}\left[\|\boldsymbol{\varepsilon}_n\|^{8+\delta}\right]<\infty$ the spectral estimator $\hat{I}^L_s=\left(\hat{\Phi}^L_s(1)\right)^{-1}\hat{\Sigma}^L_s\left(\hat{\Phi}^L_s(1)\right)^{T,-1}$ converges to I in probability as $L,s\to\infty$ if $s^3/L\to0$. The covariance matrix of $\hat{\boldsymbol{\vartheta}}^L$ is then estimated consistently as

$$\widehat{\Xi}^L_s=\frac{1}{L}\left(\hat{J}^L\right)^{-1}\hat{I}^L_s\left(\hat{J}^L\right)^{-1}. \tag{2.45}$$

In the simulation study performed in Section 4.2, this estimator for Ξ performs convincingly.

3. Quasi maximum likelihood estimation for Lévy-driven multivariate continuous-time ARMA processes

In this section we pursue the second main topic of the present paper, a detailed investigation of the asymptotic properties of the quasi maximum likelihood estimator of discretely observed multivariate continuous-time autoregressive moving average processes. We will make use of the equivalence between MCARMA and continuous-time linear state space models, as well as of the important observation that the state space structure of a continuous-time process is preserved under equidistant sampling, which allows for the results of the previous section to be applied. The conditions we need to impose on the parametrization of the models under consideration are therefore closely related to the assumptions made in the discrete-time case, except that the mixing and ergodicity assumptions D4 and D9 are automatically satisfied (Marquardt and Stelzer, 2007, Proposition 3.34).

We start the section with a short recapitulation of the definition and basic properties of Lévy-driven continuous-time ARMA processes; this is followed by a discussion of the second-order properties of discretely observed CARMA process, leading to a set of accessible identifiability conditions. Section 3.4 contains our main result about the consistency and asymptotic normality of the quasi maximum likelihood estimator for equidistantly sampled MCARMA processes.

3.1. **Lévy-driven multivariate CARMA processes and continuous-time state space models.** A natural source of randomness in the specification of continuous-time stochastic processes are Lévy processes. For a thorough discussion of these processes we refer the reader to the monographs Applebaum (2004); Bertoin (1996); Sato (1999).

Definition 3.1 (Lévy process)**.** A two-sided $\mathbb{R}^m$-valued *Lévy process* $(\boldsymbol{L}(t))_{t\geqslant 0}$ is a stochastic process, defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, with stationary, independent increments, continuous in probability, and satisfying $\boldsymbol{L}(0) = \boldsymbol{0}_m$ almost surely.

The class of Lévy processes includes many important processes such as Brownian motions, stable processes, and compound Poisson processes as special cases, which makes them very useful in stochastic modelling. Another advantage is that the property of having stationary independent increments implies that Lévy process have a rather particular structure which makes many problems analytically tractable. More precisely, the Lévy–Itô decomposition theorem asserts that every Lévy process can be additively decomposed into a Brownian motion, a compound Poisson process, and a square-integrable pure-jump martingale, where the three terms are independent. This is equivalent to the statement that the characteristic function of a Lévy process $\boldsymbol{L}$ has the special form $\mathbb{E}e^{i\langle \boldsymbol{u}, \boldsymbol{L}(t)\rangle} = \exp\{t\psi^{\boldsymbol{L}}(\boldsymbol{u})\}$, $\boldsymbol{u} \in \mathbb{R}^m$, $t \in \mathbb{R}^+$, where the characteristic exponent $\psi^{\boldsymbol{L}}$ is given by

$$\psi^{\boldsymbol{L}}(\boldsymbol{u}) = i\langle \boldsymbol{\gamma}^{\boldsymbol{L}}, \boldsymbol{u}\rangle - \frac{1}{2}\langle \boldsymbol{u}, \Sigma^{\mathcal{G}} \boldsymbol{u}\rangle + \int_{\mathbb{R}^m} \left[e^{i\langle \boldsymbol{u}, \boldsymbol{x}\rangle} - 1 - i\langle \boldsymbol{u}, \boldsymbol{x}\rangle I_{\{\|x\|\leqslant 1\}}\right] \nu^{\boldsymbol{L}}(d\boldsymbol{x}). \tag{3.1}$$

$\boldsymbol{\gamma}^{\boldsymbol{L}} \in \mathbb{R}^m$ is called the *drift vector*, $\Sigma^{\mathcal{G}}$ is a non-negative definite, symmetric $m \times m$ matrix called the *Gaussian covariance matrix*, and the *Lévy measure* $\nu^{\boldsymbol{L}}$ satisfies the two conditions $\nu^{\boldsymbol{L}}(\{\boldsymbol{0}_m\}) = 0$ and $\int_{\mathbb{R}^m} \min(\|\boldsymbol{x}\|^2, 1)\nu^{\boldsymbol{L}}(d\boldsymbol{x}) < \infty$. For the present purpose it is enough to know that a Lévy process $\boldsymbol{L}$ has finite kth absolute moments, $k > 0$, that is $\mathbb{E}\|\boldsymbol{L}(t)\|^k < \infty$, if and only if $\int_{\|\boldsymbol{x}\|\geqslant 1} \|\boldsymbol{x}\|^k \nu^{\boldsymbol{L}}(d\boldsymbol{x}) < \infty$ (Sato, 1999, Corollary 25.8), and that the covariance matrix $\Sigma^{\boldsymbol{L}}$ of $\boldsymbol{L}(1)$, if it exists, is given by $\Sigma^{\mathcal{G}} + \int_{\|\boldsymbol{x}\|\geqslant 1} \boldsymbol{x}\boldsymbol{x}^T \nu^{\boldsymbol{L}}(d\boldsymbol{x})$ Sato (1999, Example 25.11).

Assumption L1. The Lévy process $\boldsymbol{L}$ has mean zero and finite second moments, which means that $\boldsymbol{\gamma}^{\boldsymbol{L}} + \int_{\|\boldsymbol{x}\|\geqslant 1} \boldsymbol{x}\nu^{\boldsymbol{L}}(d\boldsymbol{x})$ is zero, and that the integral $\int_{\|\boldsymbol{x}\|\geqslant 1} \|\boldsymbol{x}\|^2 \nu^{\boldsymbol{L}}(d\boldsymbol{x})$ is finite.

Just like i. i. d. sequences are used in time series analysis to define ARMA processes, Lévy processes can be used to construct (multivariate) continuous-time autoregressive moving average processes, called (M)CARMA processes. If $\boldsymbol{L}$ is a two-sided Lévy process with values in $\mathbb{R}^m$ and $p > q$ are integers, the d-dimensional $\boldsymbol{L}$-driven MCARMA(p, q) process with autoregressive polynomial

$$z \mapsto P(z) \coloneqq \boldsymbol{1}_d z^p + A_1 z^{p-1} + \ldots + A_p \in M_d(\mathbb{R}[z]) \tag{3.2a}$$

and moving average polynomial

$$z \mapsto Q(z) \coloneqq B_0 z^q + B_1 z^{q-1} + \ldots + B_q \in M_{d,m}(\mathbb{R}[z]) \tag{3.2b}$$

is defined as the solution to the formal differential equation $P(D)\boldsymbol{Y}(t) = Q(D)D\boldsymbol{L}(t)$, $D \equiv (d/dt)$. It is often useful to allow for the dimensions of the driving Lévy process $\boldsymbol{L}$ and the $\boldsymbol{L}$-driven MCARMA process to be different, which is a slight extension of the original definition of Marquardt and Stelzer (2007). The results obtained in that paper remain true if our definition is used. In general, the paths of a Lévy process are not differentiable, so we interpret the defining differential equation as being equivalent to the *state space representation*

$$d\boldsymbol{G}(t) = \mathcal{A}\boldsymbol{G}(t)dt + \mathcal{B}d\boldsymbol{L}(t), \quad \boldsymbol{Y}(t) = \mathcal{C}\boldsymbol{G}(t), \quad t \in \mathbb{R}, \tag{3.3}$$

where $\mathcal{A}$,$\mathcal{B}$, and C are given by

$$\mathcal{A} = \begin{pmatrix} 0 & \mathbf{1}_d & 0 & \dots & 0 \\ 0 & 0 & \mathbf{1}_d & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & \mathbf{1}_d \\ -A_p & -A_{p-1} & \dots & \dots & -A_1 \end{pmatrix} \in M_{pd}(\mathbb{R}), \tag{3.4a}$$

$$\mathcal{B} = \begin{pmatrix} \beta_1^T & \cdots & \beta_p^T \end{pmatrix}^T \in M_{pd,m}(\mathbb{R}), \quad \beta_{p-j} = -I_{\{0,\dots,q\}}(j)\left[\sum_{i=1}^{p-j-1} A_i\beta_{p-j-i} - B_{q-j}\right], \tag{3.4b}$$

$$C = (\mathbf{1}_d, 0, \dots, 0) \in M_{d,pd}(\mathbb{R}). \tag{3.4c}$$

It follows from representation (3.3) that MCARMA processes are special cases of linear multivariate continuous-time state space models, and in fact, the class of linear state space models is equivalent to the class of MCARMA models (Schlemm and Stelzer, 2011, Corollary 3.4). By considering the class of linear state space models, one can define representations of MCARMA processes which are different from Eq. (3.3) and better suited for the purpose of estimation.

Definition 3.2 (State space model)**.** A continuous-time linear state space model $(A, B, C, \boldsymbol{L})$ of dimension N with values in $\mathbb{R}^d$ is characterized by an $\mathbb{R}^m$-valued driving Lévy process $\boldsymbol{L}$, a state transition matrix $A \in M_N(\mathbb{R})$, an input matrix $B \in M_{N,m}(\mathbb{R})$, and an observation matrix $C \in M_{d,N}(\mathbb{R})$. It consists of a state equation of Ornstein–Uhlenbeck type

$$\mathrm{d}\boldsymbol{X}(t) = A\boldsymbol{X}(t)\mathrm{d}t + B\mathrm{d}\boldsymbol{L}(t), \quad t \in \mathbb{R}, \tag{3.5a}$$

and an observation equation

$$\boldsymbol{Y}(t) = C\boldsymbol{X}(t), \quad t \in \mathbb{R}. \tag{3.5b}$$

The $\mathbb{R}^N$-valued process $\boldsymbol{X} = (\boldsymbol{X}(t))_{t\in\mathbb{R}}$ is the *state vector process*, and $\boldsymbol{Y} = (\boldsymbol{Y}(t))_{t\in\mathbb{R}}$ the *output process*.

A solution $\boldsymbol{Y}$ to Eq. (3.5) is called *causal* if, for all t, $\boldsymbol{Y}(t)$ is independent of the σ-algebra generated by $\{\boldsymbol{L}(s) : s > t\}$. Every solution to Eq. (3.5a) satisfies

$$\boldsymbol{X}(t) = \mathrm{e}^{A(t-s)}\boldsymbol{X}(s) + \int_s^t \mathrm{e}^{A(t-u)}B\mathrm{d}\boldsymbol{L}(u), \quad \forall s, t \in \mathbb{R}, \quad s < t, \tag{3.6}$$

where the stochastic integral with respect to $\boldsymbol{L}$ is well-defined by Protter (1990, Theorem 3.9). The independent-increment property of Lévy processes implies that $\boldsymbol{X}$ is a Markov process. The following can be seen as the multivariate extension of Brockwell et al. (2011, Proposition 1) and recalls conditions for the existence of a stationary causal solution of the state equation (3.5a) for easy reference. We always work under the following assumption.

Assumption E. The eigenvalues of the matrix A have strictly negative real parts.

Proposition 3.1 (Sato and Yamazato (1983, Theorem 5.1))**.** *If Assumptions E and L1 hold, then Eq. (3.5a) has a unique strictly stationary, causal solution $\boldsymbol{X}$ given by $\boldsymbol{X}(t) = \int_{-\infty}^t \mathrm{e}^{A(t-u)}B\mathrm{d}\boldsymbol{L}(u)$, which, for fixed $t \in \mathbb{R}$, has the same distribution as $\int_0^\infty \mathrm{e}^{Au}B\mathrm{d}\boldsymbol{L}(u)$. Moreover, $\boldsymbol{X}(t)$ has mean zero and second-order structure*

$$\mathbb{V}\mathrm{ar}(\boldsymbol{X}(t)) =:\Gamma_0 = \int_0^\infty \mathrm{e}^{Au}B\Sigma^{\boldsymbol{L}}B^T\mathrm{e}^{A^Tu}\mathrm{d}u, \tag{3.7a}$$

$$\mathbb{C}\mathrm{ov}\left(\boldsymbol{X}(t+h), \boldsymbol{X}(t)\right) =:\gamma_{\boldsymbol{Y}}(h) = \mathrm{e}^{Ah}\Gamma_0, \quad h \geqslant 0, \tag{3.7b}$$

where the variance Γ_0 satisfies $A\Gamma_0 + \Gamma_0 A^T = -B\Sigma^{\boldsymbol{L}}B^T$.

It is an immediate consequence that the output process $\boldsymbol{Y}$ has mean zero and autocovariance function $\mathbb{R} \ni h \mapsto \gamma_{\boldsymbol{Y}}(h)$ given by $\gamma_{\boldsymbol{Y}}(h) = C\mathrm{e}^{Ah}\Gamma_0 C^T$, $h \geqslant 0$, and that $\boldsymbol{Y}$ itself can be written succinctly as a moving average of the driving Lévy process as

$$\boldsymbol{Y}(t) = \int_{-\infty}^\infty g(t-u)\mathrm{d}\boldsymbol{L}(u), \quad t \in \mathbb{R}; \quad g(t) = C\mathrm{e}^{At}BI_{[0,\infty)}(t). \tag{3.8}$$

As in Marquardt and Stelzer (2007, Proposition 3.30) one shows the following result about the existence of moments.

Proposition 3.2. *Let $\boldsymbol{Y}$ be the output process of the state space model (3.5) driven by the Lévy process $\boldsymbol{L}$. If $\boldsymbol{L}(1)$ is in $L^r(\Omega, \mathbb{P})$ for some $r > 0$, then so are $\boldsymbol{Y}(t)$ and the state vector $\boldsymbol{X}(t)$, $t \in \mathbb{R}$.*

Equation (3.8) which, in conjunction with Eq. (3.4), serves as the definition of a multivariate CARMA process with autoregressive and moving average polynomials given by Eq. (3.2), shows that the behaviour of the process $\boldsymbol{Y}$ depends on the values of the individual matrices A, B, and C only through the products $C\mathrm{e}^{At}B$, $t \in \mathbb{R}$. The following lemma relates this analytical statement to an algebraic one about rational matrices, allowing us to draw a connection to the identifiability theory of discrete-time state space models.

Lemma 3.3. *Two matrix triplets (A, B, C), $(\tilde{A}, \tilde{B}, \tilde{C})$ of appropriate dimensions satisfy $C\mathrm{e}^{At}B = \tilde{C}\mathrm{e}^{\tilde{A}t}\tilde{B}$ for all $t \in \mathbb{R}$ if and only if $C(z\mathbf{1} - A)^{-1}B = \tilde{C}(z\mathbf{1} - \tilde{A})^{-1}\tilde{B}$ for all $z \in \mathbb{C}$.*

Proof. If we start at the first equality and replace the matrix exponentials by their spectral representations (see Lax, 2002, Theorem 17.5), we obtain

$$\int_\gamma \mathrm{e}^{zt}C(z\mathbf{1} - A)^{-1}B\mathrm{d}z = \int_{\tilde{\gamma}} \mathrm{e}^{zt}\tilde{C}(z\mathbf{1} - \tilde{A})^{-1}\tilde{B}\mathrm{d}z, \quad \forall t \in \mathbb{R}, \tag{3.9}$$

where γ is a closed contour in $\mathbb{C}$ winding around each eigenvalue of A exactly once, and likewise for $\tilde{\gamma}$. Since we can always assume that $\gamma = \tilde{\gamma}$ by taking γ to be R times the unit circle, $R > \max\{|\lambda| : \lambda \in \sigma_A \cup \sigma_{\tilde{A}}\}$, we can write Eq. (3.9) as

$$\int_\gamma \mathrm{e}^{zt}\left[C(z\mathbf{1} - A)^{-1}B - \tilde{C}(z\mathbf{1} - \tilde{A})^{-1}\tilde{B}\right]\mathrm{d}z = 0, \quad \forall t \in \mathbb{R}. \tag{3.10}$$

Since the rational matrix function $\Delta(z) = C(z\mathbf{1}-A)^{-1}B-\tilde{C}(z\mathbf{1}-\tilde{A})^{-1}\tilde{B}$ has only poles with modulus less than R, it has an expansion around infinity, $\Delta(z) = \sum_{n=0}^\infty A_n z^{-n}$, $A_n \in M_d(\mathbb{C})$, which converges in a region $\{z \in \mathbb{C} : |z| > r\}$ containing γ. Using the fact that this series converges uniformly on the compact set γ and applying the Residue Theorem from complex analysis (Dieudonné, 1968, 9.16.1), which implies $\int_\gamma \mathrm{e}^{zt}z^{-n}\mathrm{d}z = t^n/n!$, Eq. (3.10) becomes $\sum_{n=0}^\infty \frac{t^n}{n!}A_{n+1} \equiv 0_N$. Consequently, by the Identity Theorem (Dieudonné, 1968, Theorem 9.4.3), A_n is the zero matrix for all $n > 1$, and since $\Delta(z) \to 0$ as $z \to \infty$, it follows that $\Delta(z) \equiv 0_{d,m}$. □

Because of its importance for the following discussion, the rational matrix function $H : z \mapsto C(z\mathbf{1}_N - A)^{-1}B$ is given a special name: it is called the *transfer function* of the state space model (3.5) and is closely related to the spectral density $f_{\boldsymbol{Y}}$ of the output process $\boldsymbol{Y}$, which is defined as $f_{\boldsymbol{Y}}(\omega) = \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}\omega h}\gamma_{\boldsymbol{Y}}(h)\mathrm{d}h$ – the matrix transform of $\gamma_{\boldsymbol{Y}}$. Before we make this relation explicit, we prove the following lemma.

Lemma 3.4. *For any real number v, and matrices $A, B, \Sigma^{\boldsymbol{L}}, \Gamma_0$ as in Eq. (3.7a), it holds that*

$$\int_{-v}^\infty \mathrm{e}^{Au}B\Sigma^{\boldsymbol{L}}B^T\mathrm{e}^{A^Tu}\mathrm{d}u = \mathrm{e}^{-Av}\Gamma_0\mathrm{e}^{-A^Tv}. \tag{3.11}$$

Proof. We define the functions $l, r : \mathbb{R} \to M_N(\mathbb{R})$ by $l(v) = \int_{-v}^\infty \mathrm{e}^{Au}B\Sigma^{\boldsymbol{L}}B^T\mathrm{e}^{A^Tu}\mathrm{d}u$ and $r(v) = \mathrm{e}^{-Av}\Gamma_0\mathrm{e}^{-A^Tv}$. Clearly, both $l : v \mapsto l(v)$ and $r : v \mapsto r(v)$ are differentiable functions of v; taking the derivatives yields

$$\frac{\mathrm{d}}{\mathrm{d}v}l(v) = \mathrm{e}^{-Av}B\Sigma^{\boldsymbol{L}}B^T\mathrm{e}^{-A^Tv} \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}v}r(v) = -A\mathrm{e}^{-Av}\Gamma_0\mathrm{e}^{-A^Tv} - \mathrm{e}^{-Av}\Gamma_0A^T\mathrm{e}^{-A^Tv}.$$

Using Proposition 3.1 one sees immediately that $(\mathrm{d}/\mathrm{d}v)l(v) = (\mathrm{d}/\mathrm{d}v)r(v)$, for all $v \in \mathbb{R}$. Hence, l and r differ only by an additive constant. Since $l(0)$ equals $r(0)$ by the definition of Γ_0, the constant is zero, and $l(v) = r(v)$ for all real numbers v. □

Proposition 3.5. *Let $\boldsymbol{Y}$ be the output process of the state space model (3.5), and denote by $H : z \mapsto C(z\mathbf{1}_N - A)^{-1}B$ its transfer function. Then the relation $f_{\boldsymbol{Y}}(\omega) = (2\pi)^{-1}H(\mathrm{i}\omega)\Sigma^{\boldsymbol{L}}H(-\mathrm{i}\omega)^T$ holds for all real ω; in particular, $\omega \mapsto f_{\boldsymbol{Y}}(\omega)$ is a rational matrix function.*

Proof. First, we recall (Bernstein, 2005, Proposition 11.2.2) that the Laplace transform of any matrix A is given by its resolvent, that is, $(zI-A)^{-1}=\int_0^\infty \mathrm{e}^{-zu}\mathrm{e}^{Au}\mathrm{d}u$, for any complex number z. We are now ready to compute

$$\begin{aligned}\frac{1}{2\pi}H(\mathrm{i}\omega)\Sigma^{\boldsymbol{L}}H(-\mathrm{i}\omega)^T =&\frac{1}{2\pi}C(\mathrm{i}\omega\mathbf{1}_N-A)^{-1}B\Sigma^{\boldsymbol{L}}B^T(-\mathrm{i}\omega\mathbf{1}_N-A^T)^{-1}C^T\\ =&\frac{1}{2\pi}C\left[\int_0^\infty \mathrm{e}^{-\mathrm{i}\omega u}\mathrm{e}^{Au}\mathrm{d}uB\Sigma^{\boldsymbol{L}}B^T\int_0^\infty \mathrm{e}^{\mathrm{i}\omega v}\mathrm{e}^{A^Tv}\mathrm{d}v\right]\mathrm{d}hC^T.\end{aligned}$$

Introducing the new variable $h=u-v$, and using Lemma 3.4, this becomes

$$\begin{aligned}&\frac{1}{2\pi}C\left[\int_0^\infty\int_{-v}^\infty \mathrm{e}^{-\mathrm{i}\omega h}\mathrm{e}^{Ah}\mathrm{e}^{Av}B\Sigma^{\boldsymbol{L}}B^T\mathrm{e}^{A^Tv}\mathrm{d}h\mathrm{d}v\right]C^T\\ =&\frac{1}{2\pi}C\left[\int_0^\infty\int_0^\infty \mathrm{e}^{-\mathrm{i}\omega h}\mathrm{e}^{Ah}\mathrm{e}^{Av}B\Sigma^{\boldsymbol{L}}B^T\mathrm{e}^{A^Tv}\mathrm{d}h\mathrm{d}v+\int_0^\infty\int_{-v}^0 \mathrm{e}^{-\mathrm{i}\omega h}\mathrm{e}^{Ah}\mathrm{e}^{Av}B\Sigma^{\boldsymbol{L}}B^T\mathrm{e}^{A^Tv}\mathrm{d}h\mathrm{d}v\right]C^T\\ =&\frac{1}{2\pi}C\left[\int_0^\infty \mathrm{e}^{-\mathrm{i}\omega h}\mathrm{e}^{Ah}\Gamma_0\mathrm{d}h+\int_{-\infty}^0 \mathrm{e}^{-\mathrm{i}\omega h}\Gamma_0\mathrm{e}^{-A^Th}\mathrm{d}h\right]C^T.\end{aligned}$$

By Eq. (3.7b) and the fact that the spectral density and the autocovariance function of a stochastic process are Fourier duals of each other, the last expression is equal to $(2\pi)^{-1}\int_{-\infty}^\infty \mathrm{e}^{-\mathrm{i}\omega h}\gamma_{\boldsymbol{Y}}(h)\mathrm{d}h=f_{\boldsymbol{Y}}(\omega)$, which completes the proof. □

A converse of Proposition 3.5, which will be useful in our later discussion of identifiability, is the Spectral Factorization Theorem. Its proof can be found in Rozanov (1967, Theorem 1.10.1) and also in Caines (1988, Theorem 4.1.4).

Theorem 3.6. *Every positive definite rational matrix function $f\in\mathbb{S}_d^+(\mathbb{C}\{\omega\})$ of full rank can be factorized as $f(\omega)=(2\pi)^{-1}W(\mathrm{i}\omega)W(-\mathrm{i}\omega)^T$, where the rational matrix function $z\mapsto W(z)\in M_{d,N}(\mathbb{R}\{z\})$, called a spectral factor, has full rank. For fixed N, the spectral factor W is uniquely determined up to an orthogonal transformation $W(z)\mapsto W(z)O$, for some orthogonal $N\times N$ matrix O.*

3.2. **Equidistant observations.** We now turn to properties of the sampled process $\boldsymbol{Y}^{(h)}=(\boldsymbol{Y}_n^{(h)})_{n\in\mathbb{Z}}$ which is defined by $\boldsymbol{Y}_n^{(h)}=\boldsymbol{Y}(nh)$ and represents observations of the process $\boldsymbol{Y}$ at equally spaced points in time. A very fundamental observation is that the linear state space structure of the continuous-time process is preserved under sampling, as detailed in the following proposition. Of particular importance is the explicit formula (3.14) for the spectral density of the sampled process $\boldsymbol{Y}^{(h)}$.

Proposition 3.7. *Assume that $\boldsymbol{Y}$ is the output process of the state space model (3.5). Then the sampled process $\boldsymbol{Y}^{(h)}$ has the state space representation*

$$\boldsymbol{X}_n=\mathrm{e}^{Ah}\boldsymbol{X}_{n-1}+\boldsymbol{N}_n^{(h)},\quad \boldsymbol{N}_n^{(h)}=\int_{(n-1)h}^{nh}\mathrm{e}^{A(nh-u)}B\mathrm{d}\boldsymbol{L}(u),\quad \boldsymbol{Y}_n^{(h)}=C\boldsymbol{X}_n^{(h)}. \tag{3.12}$$

The sequence $\left(\boldsymbol{N}_n^{(h)}\right)_{n\in\mathbb{Z}}$ is i. i. d. with mean zero and covariance matrix

$$\Sigma\!\!\!/^{(h)}=\int_0^h \mathrm{e}^{Au}B\Sigma^{\boldsymbol{L}}B^T\mathrm{e}^{A^Tu}\mathrm{d}u. \tag{3.13}$$

Moreover, the spectral density of $\boldsymbol{Y}^{(h)}$, denoted by $f_{\boldsymbol{Y}}^{(h)}$, is given by

$$f_{\boldsymbol{Y}}^{(h)}(\omega)=C\left(\mathrm{e}^{\mathrm{i}\omega}\mathbf{1}_N-\mathrm{e}^{Ah}\right)^{-1}\Sigma\!\!\!/^{(h)}\left(\mathrm{e}^{-\mathrm{i}\omega}\mathbf{1}_N-\mathrm{e}^{A^Th}\right)^{-1}C^T; \tag{3.14}$$

in particular, $f_{\boldsymbol{Y}}^{(h)}:[-\pi,\pi]\to\mathbb{S}_d^+\left(\mathbb{R}\left\{\mathrm{e}^{\mathrm{i}\omega}\right\}\right)$ is a rational matrix function.

Proof. Eqs. (3.12) follow from setting $t=nh$, $s=(n-1)h$ in Eq. (3.6). That the sequence $(\boldsymbol{Z}_n)_{n\in\mathbb{Z}}$ is i. i. d. as well as expression (3.13) for $\Sigma\!\!\!/^{(h)}$ are immediate consequences of the Lévy process $\boldsymbol{L}$ having independent, homogeneous increments. Expression (3.14) is an application of Hamilton (1994, Eq. (10.4.43)). □

In the following we analyse further the sampled state space model (3.12), in particular we will derive conditions for it to be minimal in the sense that the process $\boldsymbol{Y}^{(h)}$ is not the output process of any state space model of dimension less than N, and for the noise covariance matrix $\Sigma^{(h)}$ given in Eq. (3.13) to be non-singular. We begin by recalling some well-known notions from discrete-time realization and control theory. For a detailed account we refer to Åström (1970); Caines (1988); Sontag (1998), which also explain the origin of the terminology.

Definition 3.3 (Algebraic realization)**.** Let $H \in M_{d,m}(\mathbb{R}\{z\})$ be a rational matrix function. A matrix triple (A, B, C) is called an *algebraic realization of* H *of dimension* N if $H(z) = C(z\mathbf{1}_N - A)^{-1}B$, where $A \in M_N(\mathbb{R})$, $B \in M_{N,m}(\mathbb{R})$, and $C \in M_{d,N}(\mathbb{R})$.

Every rational matrix function has many algebraic realizations of various dimensions. A particularly convenient class are the ones of minimal dimension, which have a number of useful properties.

Definition 3.4 (Minimality)**.** Let $H \in M_{d,m}(\mathbb{R}\{z\})$ be a rational matrix function. A *minimal realization* of H is an algebraic realization of H of dimension smaller than 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.

Two other important properties of algebraic realizations, which are intimately related to the notion of minimality and play a key role in the study of identifiability, are introduced in the following definitions.

Definition 3.5 (Controllability)**.** An algebraic realization (A, B, C) of dimension N is *controllable* if the *controllability* matrix $\mathscr{C} = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \in M_{m,mN}(\mathbb{R})$ has full rank, i. e., if $\operatorname{rank} \mathscr{C} = N$.

Definition 3.6 (Observability)**.** An algebraic realization (A, B, C) of dimension N is *observable* if the *observability* matrix $\mathscr{O} = \begin{bmatrix} C^T & (CA)^T & \cdots & (CA^{n-1})^T \end{bmatrix}^T \in M_{dN,N}(\mathbb{R})$ has full rank, i. e., if $\operatorname{rank} \mathscr{O} = N$.

Remark 3.7. We will often say that a state space system (3.5) is minimal, controllable or observable if the corresponding transfer function has this property.

The next theorem characterizes minimality in a useful way in terms of controllability and observability.

Theorem 3.8 (Hannan and Deistler (1988, Theorem 2.3.3))**.** *A realization* (A, B, C) *is minimal if and only if it is both controllable and observable.*

Lemma 3.9. *For all matrices* $A \in M_N(\mathbb{R})$, $B \in M_{N,m}(\mathbb{R})$, $\Sigma \in \mathbb{S}_m^{++}(\mathbb{R})$, *and every real number* $t > 0$, *the linear subspaces* $\operatorname{im}\left[B, AB, \ldots, A^{N-1}B\right]$ *and* $\operatorname{im} \int_0^t \mathrm{e}^{Au} B\Sigma B^T \mathrm{e}^{A^T u}\mathrm{d}u$ *are equal.*

Proof. The assertion is a straightforward generalization of Bernstein (2005, Lemma 12.6.2). □

Corollary 3.10. *If the triple* (A, B, C) *is minimal of dimension* N, *and* Σ *is positive definite, then the* $N \times N$ *matrix* $\breve{\Sigma} = \int_0^h \mathrm{e}^{Au} B\Sigma B^T \mathrm{e}^{A^T u}\mathrm{d}u$ *has full rank* N.

Proof. By Theorem 3.8, minimality of (A, B, C) implies controllability, i. e. full rank of the controllability matrix $\begin{bmatrix} B & AB & \cdots & A^{N-1}B \end{bmatrix}$. By Lemma 3.9, this is equivalent to $\breve{\Sigma}$ having full rank. □

Proposition 3.11. *Assume that* $\boldsymbol{Y}$ *is the* d*-dimensional output process of the state space model* (3.5) *with* (A, B, C) *being a minimal realization of McMillan degree* N. *Then a sufficient condition for the sampled process* $\boldsymbol{Y}^{(h)}$ *to have the same McMillan degree, is the Kalman–Bertram criterion*

$$\lambda - \lambda' \neq 2h^{-1}\pi i k, \qquad \forall(\lambda, \lambda') \in \sigma(A) \times \sigma(A), \qquad \forall k \in \mathbb{Z}\backslash\{0\}. \tag{3.15}$$

Proof. We will prove the assertion by showing that the N-dimensional state space representation (3.12) is both controllable and observable, and thus, by Theorem 3.8, minimal. Observability has been shown in Sontag (1998, Proposition 5.2.11) using the Hautus criterion (Hautus, 1969). The key ingredient in the proof of controllability is Corollary 3.10, where we showed that the autocovariance matrix $\breve{\Sigma}^{(h)}$ of $\boldsymbol{N}_n^{(h)}$, given by Eq. (3.13), has full rank; this shows that the representation (3.12) is indeed minimal and completes the proof. □

Remark 3.8. Since, by Hannan and Deistler (1988, Theorem 2.3.4), minimal realizations are unique up to a change of basis $(A, B, C) \mapsto (TAT^{-1}, TB, CT^{-1})$, for some non-singular $N \times N$ matrix T, and such a transformation does not change the eigenvalues of A, the criterion (3.15) does not depend on what particular triple (A, B, C) one chooses.

Uniqueness of the principal logarithm (Higham, 2008, Theorem 1.31) implies the following.

Lemma 3.12. *Assume that the matrices $A, B \in M_N(\mathbb{R})$ satisfy $\mathrm{e}^{hA} = \mathrm{e}^{hB}$ for some $h > 0$. If the spectra σ_A, σ_B of A, B satisfy $|\operatorname{Im}\lambda| < \pi/h$ for all $\lambda \in \sigma_A \cup \sigma_B$, then $A = B$.*

Lemma 3.13. *Assume that $A \in M_N(\mathbb{R})$ satisfies Assumption E. For every $h > 0$, the linear map $\mathscr{M}$: $M_N(\mathbb{R}) \to M_N(\mathbb{R})$, $M \mapsto \int_0^h \mathrm{e}^{Au} M \mathrm{e}^{A^T u}\mathrm{d}u$ is injective.*

Proof. If we apply the vectorization operator $\operatorname{vec} : M_N(\mathbb{R}) \to \mathbb{R}^{N^2}$ and use the well-known identity (Bernstein, 2005, Proposition 7.1.9) $\operatorname{vec}(UVW) = (W^T \otimes U)\operatorname{vec}(V)$ for matrices U, V and W of appropriate dimensions, we obtain the induced linear operator

$$\operatorname{vec}\circ\mathscr{M}\circ\operatorname{vec}^{-1} : \mathbb{R}^{N^2} \to \mathbb{R}^{N^2}, \quad \operatorname{vec} M \mapsto \int_0^h \mathrm{e}^{Au} \otimes \mathrm{e}^{Au}\mathrm{d}u \operatorname{vec} M.$$

To prove the claim that the operator $\mathscr{M}$ is injective, it is thus sufficient to show that the matrix $\mathscr{A} := \int_0^h \mathrm{e}^{Au} \otimes \mathrm{e}^{Au}\mathrm{d}u \in M_{N^2}(\mathbb{R})$ is non-singular. We write $A \oplus A := A \otimes \mathbf{1}_N + \mathbf{1}_N \otimes A$. By Bernstein (2005, Fact 11.14.37), $\mathscr{A} = \int_0^h \mathrm{e}^{(A\oplus A)u}\mathrm{d}u$ and since $\sigma(A \oplus A) = \{\lambda + \mu : \lambda, \mu \in \sigma(A)\}$ (Bernstein, 2005, Proposition 7.2.3), Assumption E implies that all eigenvalues of the matrix $A \oplus A$ have strictly negative real parts; in particular, $A \oplus A$ is invertible. Consequently, it follows from Bernstein (2005, Fact 11.13.14) that $\mathscr{A} = (A \oplus A)^{-1}\left[\mathrm{e}^{(A\oplus A)h} - \mathbf{1}_{N^2}\right]$. Since, for any matrix M, it holds that $\sigma(\mathrm{e}^M) = \{\mathrm{e}^\lambda, \lambda \in \sigma(M)\}$ (Bernstein, 2005, Proposition 11.2.3), the spectrum of $\mathrm{e}^{(A\oplus A)h}$ is a subset of the open unit disk, and it follows that $\mathscr{A}$ is invertible. □

3.3. **Overcoming the aliasing effect.** One goal in this paper is the estimation of multivariate CARMA processes or, equivalently, continuous-time state space models, based on discrete observations. In this brief section we concentrate on the issue of identifiability, and we derive sufficient conditions that prevent redundancies from being introduced into an otherwise properly specified model by the process of sampling, an effect known as aliasing (Hansen and Sargent, 1983; McCrorie, 2003).

For ease of notation we choose to parametrize the state matrix, the input matrix, and the observation matrix of the state space model (3.5), as well as the driving Lévy process $\boldsymbol{L}$; from these one can always obtain an autoregressive and a moving average polynomial which describe the same process by applying a left matrix fraction decomposition to the corresponding transfer function, see Patel (1981) and the upcoming Theorems 4.2 and 4.3. We hence assume that there is some compact parameter set $\Theta \subset \mathbb{R}^r$, and that, for each $\boldsymbol{\vartheta} \in \Theta$, one is given matrices $A_{\boldsymbol{\vartheta}}$, $B_{\boldsymbol{\vartheta}}$ and $C_{\boldsymbol{\vartheta}}$ of matching dimensions, as well as a Lévy process $\boldsymbol{L}_{\boldsymbol{\vartheta}}$. A basic assumption is that we always work with second order processes (cf. Assumption L1).

Assumption C1. For each $\boldsymbol{\vartheta} \in \Theta$, it holds that $\mathbb{E}\boldsymbol{L}_{\boldsymbol{\vartheta}} = \mathbf{0}_m$, that $\mathbb{E}\,\|\boldsymbol{L}_{\boldsymbol{\vartheta}}(1)\|^2$ is finite, and that the covariance matrix $\Sigma^{\boldsymbol{L}}_{\boldsymbol{\vartheta}} = \mathbb{E}\boldsymbol{L}_{\boldsymbol{\vartheta}}(1)\boldsymbol{L}_{\boldsymbol{\vartheta}}(1)^T$ is non-singular.

To ensure that the model corresponding to $\boldsymbol{\vartheta}$ describes a stationary output process we impose the analogue of Assumption E.

Assumption C2. For each $\boldsymbol{\vartheta} \in \Theta$, the eigenvalues of $A_{\boldsymbol{\vartheta}}$ have strictly negative real parts.

Next, we restrict the model class so as to only contain minimal algebraic realizations of a fixed McMillan degree.

Assumption C3. For all $\boldsymbol{\vartheta} \in \Theta$, the triple $(A_{\boldsymbol{\vartheta}}, B_{\boldsymbol{\vartheta}}, C_{\boldsymbol{\vartheta}})$ is minimal with McMillan degree N.

Since we shall base the inference on a quasi maximum likelihood approach and thus on second-order properties of the observed process, we require the model class to be identifiable from these available information according to the following definitions.

Definition 3.9 (L^2-equivalence)**.** 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 3.10. A family $(\boldsymbol{Y}_{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta} \in \Theta)$ of continuous-time stochastic processes is *identifiable from the spectral density* if, for every $\boldsymbol{\vartheta}_1 \neq \boldsymbol{\vartheta}_2$, the two processes $\boldsymbol{Y}_{\boldsymbol{\vartheta}_1}$ and $\boldsymbol{Y}_{\boldsymbol{\vartheta}_2}$ are not L^2-observationally equivalent. It is *h-identifiable from the spectral density*, $h > 0$, if, for every $\boldsymbol{\vartheta}_1 \neq \boldsymbol{\vartheta}_2$, the two sampled processes $\boldsymbol{Y}^{(h)}_{\boldsymbol{\vartheta}_1}$ and $\boldsymbol{Y}^{(h)}_{\boldsymbol{\vartheta}_2}$ are not L^2-observationally equivalent.

Assumption C4. The collection of output processes $K(\Theta) := (\boldsymbol{Y}_{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta} \in \Theta)$ corresponding to the state space models $(A_{\boldsymbol{\vartheta}}, B_{\boldsymbol{\vartheta}}, C_{\boldsymbol{\vartheta}}, \boldsymbol{L}_{\boldsymbol{\vartheta}})$ is identifiable from the spectral density.

Since we shall use only observations of $\boldsymbol{Y}$ at discrete points in time separated by a sampling interval h, it would seem more natural to impose the stronger requirement that $K(\Theta)$ be h-identifiable. We will see, however, that this is implied by the previous assumptions if we additionally assume that the following holds.

Assumption C5. For all $\boldsymbol{\vartheta} \in \Theta$, the spectrum of $A_{\boldsymbol{\vartheta}}$ is a subset of $\{z \in \mathbb{C} : -\pi/h < \operatorname{Im} z < \pi/h\}$.

Theorem 3.14 (Identifiability)**.** *Assume that $\Theta \supset \boldsymbol{\vartheta} \mapsto \left(A_{\boldsymbol{\vartheta}}, B_{\boldsymbol{\vartheta}}, C_{\boldsymbol{\vartheta}}, \Sigma^{\boldsymbol{L}}_{\boldsymbol{\vartheta}}\right)$ is a parametrization of continuous-time state space models satisfying Assumptions C1 to C5. Then the corresponding collection of output processes $K(\Theta)$ is h-identifiable from the spectral density.*

Proof. We will show that for every $\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2 \in \Theta$, $\boldsymbol{\vartheta}_1 \neq \boldsymbol{\vartheta}_2$, the sampled output processes $\boldsymbol{Y}^{(h)}_{\boldsymbol{\vartheta}_1}$ and $\boldsymbol{Y}^{(h)}_{\boldsymbol{\vartheta}_2}(h)$ are not L^2-observationally equivalent. Suppose, for the sake of contradiction, that the spectral densities of the sampled output processes were the same. Then the Spectral Factorization Theorem (Theorem 3.6) would imply that there exists an orthogonal $N \times N$ matrix O such that

$$C_{\boldsymbol{\vartheta}_1}(\mathrm{e}^{\mathrm{i}\omega}\mathbf{1}_N - \mathrm{e}^{A_{\boldsymbol{\vartheta}_1}h})\Sigma^{(h),1/2}_{\boldsymbol{\vartheta}_1} O = C_{\boldsymbol{\vartheta}_2}(\mathrm{e}^{\mathrm{i}\omega}\mathbf{1}_N - \mathrm{e}^{A_{\boldsymbol{\vartheta}_2}h})\Sigma^{(h),1/2}_{\boldsymbol{\vartheta}_2}, \quad -\pi \leqslant \omega \leqslant \pi,$$

where $\Sigma^{(h),1/2}_{\boldsymbol{\vartheta}_i}$, $i = 1, 2$, are the unique positive definite matrix square roots of the covariance matrices $\int_0^h \mathrm{e}^{A_{\boldsymbol{\vartheta}_i}u} B_{\boldsymbol{\vartheta}_i} \Sigma^{\boldsymbol{L}}_{\boldsymbol{\vartheta}_i} B^T_{\boldsymbol{\vartheta}_i} \mathrm{e}^{A^T_{\boldsymbol{\vartheta}_i}u}\mathrm{d}u$, defined by spectral calculus. This means that the two triples

$$\left(\mathrm{e}^{A_{\boldsymbol{\vartheta}_1}h}, \Sigma^{(h),1/2}_{\boldsymbol{\vartheta}_1} O, C_{\boldsymbol{\vartheta}_1}\right) \quad \text{and} \quad \left(\mathrm{e}^{A_{\boldsymbol{\vartheta}_2}h}, \Sigma^{(h),1/2}_{\boldsymbol{\vartheta}_2}, C_{\boldsymbol{\vartheta}_2}\right)$$

are algebraic realizations of the same rational matrix function. Since Assumption C5 clearly implies the Kalman–Bertram criterion (3.15), it follows from Proposition 3.11 in conjunction with Assumption C3 that these realizations are minimal, and hence from Hannan and Deistler (1988, Theorem 2.3.4) that there exists an invertible matrix $T \in M_N(\mathbb{R})$ satisfying

$$\mathrm{e}^{A_{\boldsymbol{\vartheta}_1}h} = T^{-1}\mathrm{e}^{A_{\boldsymbol{\vartheta}_2}h}T, \qquad \Sigma^{(h),1/2}_{\boldsymbol{\vartheta}_1} O = T^{-1}\Sigma^{(h),1/2}_{\boldsymbol{\vartheta}_2}, \qquad C_{\boldsymbol{\vartheta}_1} = C_{\boldsymbol{\vartheta}_2}T. \tag{3.16}$$

It follows from the power series representation of the matrix exponential that $T^{-1}\mathrm{e}^{A_{\boldsymbol{\vartheta}_2}h}T$ equals $\mathrm{e}^{T^{-1}A_{\boldsymbol{\vartheta}_2}Th}$. Under Assumption C5, the first equation in conjunction with Lemma 3.12 therefore implies that $A_{\boldsymbol{\vartheta}_1} = T^{-1}A_{\boldsymbol{\vartheta}_2}T$. Using this, the second of the three equations (3.16) gives

$$\Sigma^{(h)}_{\boldsymbol{\vartheta}_1} = \int_0^h \mathrm{e}^{A_{\boldsymbol{\vartheta}_1}u}\left(T^{-1}B_{\boldsymbol{\vartheta}_2}\right)\Sigma^{\boldsymbol{L}}_{\boldsymbol{\vartheta}_2}\left(T^{-1}B_{\boldsymbol{\vartheta}_2}\right)^T \mathrm{e}^{A^T_{\boldsymbol{\vartheta}_1}u}\mathrm{d}u,$$

which, by Lemma 3.13, implies that $(T^{-1}B_{\boldsymbol{\vartheta}_2})\Sigma^{\boldsymbol{L}}_{\boldsymbol{\vartheta}_2}(T^{-1}B_{\boldsymbol{\vartheta}_2})^T = B_{\boldsymbol{\vartheta}_1}\Sigma^{\boldsymbol{L}}_{\boldsymbol{\vartheta}_1}B^T_{\boldsymbol{\vartheta}_1}$. Together with the last of the equations (3.16) and Proposition 3.7 it follows that, for every $\omega \in [-\pi, \pi]$,

$$\begin{aligned} f_{\boldsymbol{\vartheta}_1}(\omega) =& C_{\boldsymbol{\vartheta}_1}(\mathrm{i}\omega\mathbf{1}_N - A_{\boldsymbol{\vartheta}_1})^{-1}B_{\boldsymbol{\vartheta}_1}\Sigma^{\boldsymbol{L}}_{\boldsymbol{\vartheta}_1}B^T_{\boldsymbol{\vartheta}_1}(-\mathrm{i}\omega\mathbf{1}_N - A^T_{\boldsymbol{\vartheta}_1})^{-1}C^T_{\boldsymbol{\vartheta}_1} \\ =& C_{\boldsymbol{\vartheta}_2}(\mathrm{i}\omega\mathbf{1}_N - A_{\boldsymbol{\vartheta}_2})^{-1}B_{\boldsymbol{\vartheta}_2}\Sigma^{\boldsymbol{L}}_{\boldsymbol{\vartheta}_2}B^T_{\boldsymbol{\vartheta}_2}(-\mathrm{i}\omega\mathbf{1}_N - A^T_{\boldsymbol{\vartheta}_2})^{-1}C^T_{\boldsymbol{\vartheta}_2} = f_{\boldsymbol{\vartheta}_2}(\omega); \end{aligned}$$

this contradicts Assumption C4 that $\boldsymbol{Y}_{\boldsymbol{\vartheta}_1}$ and $\boldsymbol{Y}_{\boldsymbol{\vartheta}_2}$ are not L^2-observationally equivalent. □

3.4. **Asymptotic properties of the QML estimator.** In this section we apply the theory that we developed in Section 2 for the quasi maximum likelihood estimation of general discrete-time linear state space models to the estimation of continuous-time linear state space models or, equivalently, multivariate CARMA processes. We have already seen that a discretely observed MCARMA process can be represented by a discrete-time state space model and that, thus, a parametric family of MCARMA processes induces a parametric family of discrete-time state space models. More precisely, Eqs. (3.12) show that the process of sampling with spacing h maps the continuous-time state space models $(A_{\boldsymbol{\vartheta}}, B_{\boldsymbol{\vartheta}}, C_{\boldsymbol{\vartheta}}, \boldsymbol{L}_{\boldsymbol{\vartheta}})_{\boldsymbol{\vartheta}\in\Theta}$ to the discrete-time state space models

$$\left(\mathrm{e}^{A_{\boldsymbol{\vartheta}}h}, C_{\boldsymbol{\vartheta}}, \boldsymbol{N}^{(h)}_{\boldsymbol{\vartheta}}, \boldsymbol{0}\right)_{\boldsymbol{\vartheta}\in\Theta}, \quad \boldsymbol{N}^{(h)}_{\boldsymbol{\vartheta},n} = \int_{(n-1)h}^{nh} \mathrm{e}^{A_{\boldsymbol{\vartheta}}u}B_{\boldsymbol{\vartheta}}\mathrm{d}\boldsymbol{L}_{\boldsymbol{\vartheta}}(u), \tag{3.17}$$

which are not in the innovations form (1.2). The quasi maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}^{L,(h)}$ is defined by Eq. (2.14), applied to the state space model (3.17), that is

$$\hat{\boldsymbol{\vartheta}}^{L,(h)} = \operatorname{argmin}_{\boldsymbol{\vartheta}\in\Theta} \widehat{\mathscr{L}}^{(h)}(\boldsymbol{\vartheta}, \boldsymbol{y}^{L,(h)}), \tag{3.18a}$$

$$\widehat{\mathscr{L}}^{(h)}(\boldsymbol{\vartheta}, \boldsymbol{y}^{L,(h)}) = \sum_{n=1}^{L} \left[d \log 2\pi + \log\det V_{\boldsymbol{\vartheta}}^{(h)} + \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}^{(h),T} V_{\boldsymbol{\vartheta}}^{(h),-1} \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}^{(h)} \right], \tag{3.18b}$$

where $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta}}^{(h)}$ are the pseudo-innovations of the observed process $\boldsymbol{Y}^{(h)} = \boldsymbol{Y}_{\boldsymbol{\vartheta}_0}^{(h)}$, which are computed from the sample $\boldsymbol{y}^{L,(h)} = (\boldsymbol{Y}_1^{(h)}, \ldots, \boldsymbol{Y}_L^{(h)})$ via the recursion

$$\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},n} = \left(\mathrm{e}^{A_{\boldsymbol{\vartheta}} h} - K_{\boldsymbol{\vartheta}}^{(h)} C_{\boldsymbol{\vartheta}}\right) \hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},n-1} + K_{\boldsymbol{\vartheta}}^{(h)} \boldsymbol{Y}_{n-1}^{(h)}, \quad \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{\vartheta},n}^{(h)} = \boldsymbol{Y}_n^{(h)} - C_{\boldsymbol{\vartheta}} \hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},n}, \quad n \in \mathbb{N}.$$

The initial value $\hat{\boldsymbol{X}}_{\boldsymbol{\vartheta},1}$ may be chosen in the same ways as in the discrete-time case. The steady-state Kalman gain matrices $K_{\boldsymbol{\vartheta}}^{(h)}$ and pseudo-covariances $V_{\boldsymbol{\vartheta}}^{(h)}$ are computed as functions of the unique positive definite solution $\Omega_{\boldsymbol{\vartheta}}^{(h)}$ to the discrete-time algebraic Riccati equation

$$\Omega_{\boldsymbol{\vartheta}}^{(h)} = \mathrm{e}^{A_{\boldsymbol{\vartheta}} h} \Omega_{\boldsymbol{\vartheta}}^{(h)} \mathrm{e}^{A_{\boldsymbol{\vartheta}}^T h} + \Sigma_{\boldsymbol{\vartheta}}^{(h)} - \left[\mathrm{e}^{A_{\boldsymbol{\vartheta}} h} \Omega_{\boldsymbol{\vartheta}}^{(h)} C_{\boldsymbol{\vartheta}}^T\right] \left[C_{\boldsymbol{\vartheta}} \Omega_{\boldsymbol{\vartheta}}^{(h)} C_{\boldsymbol{\vartheta}}^T\right]^{-1} \left[\mathrm{e}^{A_{\boldsymbol{\vartheta}} h} \Omega_{\boldsymbol{\vartheta}}^{(h)} C_{\boldsymbol{\vartheta}}^T\right]^T,$$

namely

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

In order to obtain the asymptotic normality of the quasi maximum likelihood estimator for multivariate CARMA processes, it is therefore only necessary to make sure that Assumptions D1 to D10 hold for the model (3.17). The discussion of identifiability in the previous section allows us to specify accessible conditions on the parametrization of the continuous-time model under which the quasi maximum likelihood estimator is strongly consistent. In addition to the identifiability assumptions C3 to C5, we impose the following conditions.

Assumption C6. The parameter space Θ is a compact subset of $\mathbb{R}^r$.

Assumption C7. The functions $\boldsymbol{\vartheta} \mapsto A_{\boldsymbol{\vartheta}}$, $\boldsymbol{\vartheta} \mapsto B_{\boldsymbol{\vartheta}}$, $\boldsymbol{\vartheta} \mapsto C_{\boldsymbol{\vartheta}}$, and $\boldsymbol{\vartheta} \mapsto \Sigma_{\boldsymbol{\vartheta}}^{\boldsymbol{L}}$ are continuous. Moreover, for each $\boldsymbol{\vartheta} \in \Theta$, the matrix $C_{\boldsymbol{\vartheta}}$ has full rank.

Lemma 3.15. *Assumptions C1 to C3, C6 and C7 imply that the family* $\left(\mathrm{e}^{A_{\boldsymbol{\vartheta}} h}, C_{\boldsymbol{\vartheta}}, \boldsymbol{N}_{\boldsymbol{\vartheta}}^{(h)}, \boldsymbol{0}\right)_{\boldsymbol{\vartheta}\in\Theta}$ *of discrete-time state space models satisfies Assumptions D1 to D4.*

Proof. Assumption D1 is clear. Assumption D2 follows from the observation that the functions $A \mapsto \mathrm{e}^A$ and $(A, B, \Sigma) \mapsto \int_0^h \mathrm{e}^{Au} B\Sigma B^T \mathrm{e}^{A^T u} \mathrm{d}u$ are continuous. By Assumptions C2, C6 and C7, and the fact that the eigenvalues of a matrix are continuous functions of its entries, it follows that there exists a positive real number ϵ such that, for each $\boldsymbol{\vartheta} \in \Theta$, the eigenvalues of $A_{\boldsymbol{\vartheta}}$ have real parts less than or equal to $-\epsilon$. The observation that the eigenvalues of e^A are given by the exponentials of the eigenvalues of A thus shows that Assumption D3, i) holds with $\rho := \mathrm{e}^{-\epsilon h} < 1$. Assumption C1 that the matrices $\Sigma_{\boldsymbol{\vartheta}}^{\boldsymbol{L}}$ are non-singular and the minimality assumption C3 imply by Corollary 3.10 that the noise covariance matrices $\Sigma_{\boldsymbol{\vartheta}}^{(h)} = \mathbb{E}\boldsymbol{N}_{\boldsymbol{\vartheta},n}^{(h)} \boldsymbol{N}_{\boldsymbol{\vartheta},n}^{(h),T}$ are non-singular, and thus Assumption D3, ii) holds. Further, by Proposition 2.2, the matrices $\Omega_{\boldsymbol{\vartheta}}$ are non-singular, and so are, because the matrices $C_{\boldsymbol{\vartheta}}$ are assumed to be of full rank, the matrices $V_{\boldsymbol{\vartheta}}$; this means that Assumption D3, iii) is satisfied. Assumption D4 is a consequence of Proposition 3.7, which states that the noise sequences $\boldsymbol{N}_{\boldsymbol{\vartheta}}$ are i. i. d. and in particular ergodic; their second moments are finite because of Assumption C1. □

In order to be able to show that the quasi maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}^{L,(h)}$ is asymptotically normally distributed, we impose the following conditions in addition to the ones described so far.

Assumption C8. The true parameter value $\boldsymbol{\vartheta}_0$ is an element of the interior of Θ.

Assumption C9. The functions $\boldsymbol{\vartheta} \mapsto A_{\boldsymbol{\vartheta}}$, $\boldsymbol{\vartheta} \mapsto B_{\boldsymbol{\vartheta}}$, $\boldsymbol{\vartheta} \mapsto C_{\boldsymbol{\vartheta}}$, and $\boldsymbol{\vartheta} \mapsto \Sigma_{\boldsymbol{\vartheta}}^{\boldsymbol{L}}$ are three times continuously differentiable.

Assumption C10. There exists a positive number δ such that $\mathbb{E}\left\|\boldsymbol{L}_{\boldsymbol{\vartheta}_0}(1)\right\|^{4+\delta} < \infty$.

Lemma 3.16. *Assumptions C8 to C10 imply that Assumptions D6 to D8 hold for the model* (3.17).

Proof. Assumption D6 is clear. Assumption D7 follows from the fact that the functions $A \mapsto \mathrm{e}^A$ and $(A, B, \Sigma) \mapsto \int_0^h \mathrm{e}^{Au} B\Sigma B^T \mathrm{e}^{A^T u}\mathrm{d}u$ are not only continuous, but infinitely often differentiable. For Assumption D8 we need to show that the random variables $\boldsymbol{N} := \boldsymbol{N}_{\boldsymbol{\vartheta}_0,1}$ have bounded $(4+\delta)$th absolute moments. It follows from Rajput and Rosiński (1989, Theorem 2.7) that $\boldsymbol{N}$ is infinitely divisible with characteristic triplet $(\boldsymbol{\gamma}, \Sigma, \nu)$ and that

$$\int_{\|\boldsymbol{x}\|\geqslant 1} \|\boldsymbol{x}\|^{4+\delta}\, \nu(\mathrm{d}\boldsymbol{x}) \leqslant \int_0^1 \left\|\mathrm{e}^{A_{\boldsymbol{\vartheta}_0}(h-s)} B_{\boldsymbol{\vartheta}}\right\|^{4+\delta} \mathrm{d}s \int_{\|\boldsymbol{x}\|\geqslant 1} \|\boldsymbol{x}\|^{4+\delta}\, \nu^{\boldsymbol{L}_{\boldsymbol{\vartheta}_0}}\boldsymbol{\vartheta}(\mathrm{d}\boldsymbol{x}).$$

The first factor on the right side is finite by Assumptions C6 and C9, the second by Assumption C10 and the well known equivalence of finiteness of the αth absolute moment of an infinitely divisible distribution and finiteness of the αth absolute moments of the corresponding Lévy measure restricted to the exterior of the unit ball (Sato, 1999, Corollary 25.8). The same corollary shows that $\mathbb{E}\|\boldsymbol{N}\|^{4+\delta} < \infty$ and thus Assumption D8. □

Our final assumption is the analogue of Assumption D10. It will ensure that the Fisher information matrix of the quasi maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}^{L,(h)}$ is non-singular by imposing a non-degeneracy condition on the parametrization of the model.

Assumption C11. There exists a positive index j_0 such that the $\left[(j_0+2)d^2\right] \times r$ matrix

$$\nabla_{\boldsymbol{\vartheta}}\left(\left[\boldsymbol{1}_{j_0+1} \otimes K_{\boldsymbol{\vartheta}}^{(h),T} \otimes C_{\boldsymbol{\vartheta}}\right]\left[\begin{array}{cccc}\left(\operatorname{vec}\mathrm{e}^{\boldsymbol{1}_N h}\right)^T & \left(\operatorname{vec}\mathrm{e}^{A_{\boldsymbol{\vartheta}}h}\right)^T & \cdots & \left(\operatorname{vec}\mathrm{e}^{A_{\boldsymbol{\vartheta}}^{j_0}h}\right)^T\end{array}\right]^T \atop \operatorname{vec} V_{\boldsymbol{\vartheta}}\right)_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0}$$

has rank r.

Theorem 3.17 (Consistency and asymptotic normality for $\hat{\boldsymbol{\vartheta}}^{L,(h)}$). *Assume that $(A_{\boldsymbol{\vartheta}}, B_{\boldsymbol{\vartheta}}, C_{\boldsymbol{\vartheta}}, \boldsymbol{L}_{\boldsymbol{\vartheta}})_{\boldsymbol{\vartheta}\in\Theta}$ is a parametric family of continuous-time state space models, and denote by $\boldsymbol{y}^{L,(h)} = (\boldsymbol{Y}^{(h)}_{\boldsymbol{\vartheta}_0,1}, \ldots, \boldsymbol{Y}^{(h)}_{\boldsymbol{\vartheta}_0,L})$ a sample of length L from the discretely observed output process corresponding to the parameter value $\boldsymbol{\vartheta}_0 \in \Theta$. Under Assumptions C1 to C7 the quasi maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}^{L,(h)} = \operatorname{argmin}_{\boldsymbol{\vartheta}\in\Theta} \widehat{\mathscr{L}}(\boldsymbol{\vartheta}, \boldsymbol{y}^{L,(h)})$ is strongly consistent, that is*

$$\hat{\boldsymbol{\vartheta}}^{L,(h)} \xrightarrow[L\to\infty]{a.\,s.} \boldsymbol{\vartheta}_0. \tag{3.19}$$

If, moreover, Assumptions C8 to C11 hold, then $\hat{\boldsymbol{\vartheta}}^{L,(h)}$ is asymptotically normally distributed, that is

$$\sqrt{L}\left(\hat{\boldsymbol{\vartheta}}^{L,(h)} - \boldsymbol{\vartheta}_0\right) \xrightarrow[L\to\infty]{d} \mathscr{N}(\boldsymbol{0}, \Xi), \tag{3.20}$$

where the asymptotic covariance matrix $\Xi = J^{-1}IJ^{-1}$ is given by

$$I = \lim_{L\to\infty} L^{-1}\, \mathbb{V}\mathrm{ar}\left(\nabla_{\boldsymbol{\vartheta}}\mathscr{L}\left(\boldsymbol{\vartheta}_0, \boldsymbol{y}^L\right)\right), \quad J = \lim_{L\to\infty} L^{-1}\nabla^2_{\boldsymbol{\vartheta}}\mathscr{L}\left(\boldsymbol{\vartheta}_0, \boldsymbol{y}^L\right). \tag{3.21}$$

Proof. Strong consistency is a consequence of Theorem 2.5 if we can show that the parametric family $\left(\mathrm{e}^{A_{\boldsymbol{\vartheta}}h}, C_{\boldsymbol{\vartheta}}, \boldsymbol{N}_{\boldsymbol{\vartheta}}, \boldsymbol{0}\right)_{\boldsymbol{\vartheta}\in\Theta}$ of discrete-time state space models satisfies Assumptions D1 to D5. The first four of these are shown to hold in Lemma 3.15. For the last one, we observe that, by Lemma 2.4, Assumption D5 is equivalent to the family of state space models (3.17) being identifiable from the spectral density. Under Assumptions C3 to C5 this is guaranteed by Theorem 3.14.

In order to prove Eq. (3.20), we shall apply Theorem 2.6 and therefore need to verify Assumptions D6 to D10 for the state space models $\left(\mathrm{e}^{A_{\boldsymbol{\vartheta}}h}, C_{\boldsymbol{\vartheta}}, \boldsymbol{N}_{\boldsymbol{\vartheta}}, \boldsymbol{0}\right)_{\boldsymbol{\vartheta}\in\Theta}$. The first three hold by Lemma 3.16, the last one as a reformulation of Assumption C11. Assumption D9, that the strong mixing coefficients α of a sampled multivariate CARMA process satisfy $\sum_m [\alpha(m)]^{\delta/(2+\delta)} < \infty$, follows from Assumption C1 and Marquardt and Stelzer (2007, Proposition 3.34), where it was shown that MCARMA processes with a finite logarithmic moment are exponentially strongly mixing. □

4. Practical applicability

In this section we complement the theoretical results from Sections 2 and 3 by commenting on their applicability in practical situations. Canonical parametrizations are a classical subject of research about discrete-time dynamical systems, and most of the results carry over to the continuous-time case; without going into great detail we present the basic notions and results about these parametrizations. The assertions of Theorem 3.17 are confirmed by means of a simulation study for a bivariate non-Gaussian CARMA process. Finally, we estimate the parameters of a CARMA model for a bivariate time series from economics using our quasi maximum likelihood approach.

4.1. **Canonical parametrizations.** We present parametrizations of multivariate CARMA processes that satisfy the identifiability conditions C3 and C4, as well as the smoothness conditions C7 and C9; if, in addition, the parameter space Θ is restricted so that Assumptions C2, C5, C6 and C8 hold, and the driving Lévy process satisfies Assumption C1, the canonically parametrized MCARMA model can be estimated consistently. In order for this estimate to be asymptotically normally distributed, one must additionally impose Assumption C10 on the Lévy process and check that Assumption C11 holds – a condition which we are unable to verify analytically for the general model; for explicit parametrizations, however, it can be checked numerically with moderate computational effort. The parametrizations we are to present are well-known from the discrete-time setting; detailed descriptions with proofs can be found in Deistler (1983); Hannan and Deistler (1988); Lütkepohl and Poskitt (1996); Reinsel (1997) or, from a slightly different perspective, in the control theory literature Gevers (1986); Gevers and Wertz (1984); Guidorzi (1975). We begin with a canonical decomposition for rational matrix functions.

Theorem 4.1 (Bernstein (2005, Theorem 4.7.5))**.** *Let $H \in M_{d,m}(\mathbb{R}\{z\})$ be a rational matrix function of rank r. There exist matrices $S_1 \in M_d(\mathbb{R}[z])$ and $S_2 \in M_m(\mathbb{R}[z])$ with constant determinant, such that $H = S_1 M S_2$, where*

$$M = \begin{bmatrix} \operatorname{diag}\{\epsilon_i/\psi_i\}_{i=1}^r & 0_{r,m-r} \\ 0_{d-r,r} & 0_{d-r,m-r} \end{bmatrix} \in M_{d,m}(\mathbb{R}\{z\}), \tag{4.1}$$

and $\epsilon_1,\ldots\epsilon_r$, $\psi_1,\ldots,\psi_r \in \mathbb{R}[z]$ are monic polynomials uniquely determined by H satisfying the following conditions:

i) for each $i = 1,\ldots,r$, the polynomials ϵ_i and ψ_i have no common roots,
ii) for each $i = 1,\ldots,r-1$, the polynomial ϵ_i divides the polynomial ϵ_{i+1}, and
iii) for each $i = 1,\ldots,r-1$, 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.

The degrees ν_i of the denominator polynomials ψ_i in the Smith–McMillan decomposition of a rational matrix function H are called the Kronecker indices of H, and they define the vector $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_d) \in \mathbb{N}^d$, where we set $\nu_k = 0$ for $k = r+1,\ldots,d$. They satisfy the important relation $\sum_{i=1}^d \nu_i = \delta_M(H)$, where $\delta_M(H)$ denotes the McMillan degree of H, that is the smallest possible dimension of an algebraic realization of H, see Definition 3.4. For $1 \leqslant i,j \leqslant d$, we also define the integers $\nu_{ij} = \min\{\nu_i + I_{\{i>j\}}, \nu_j\}$, and if the Kronecker indices of the transfer function of an MCARMA process $\boldsymbol{Y}$ are $\boldsymbol{\nu}$, we call $\boldsymbol{Y}$ an MCARMA$_{\boldsymbol{\nu}}$ process.

Theorem 4.2 (Echelon state space realization, Guidorzi (1975, Section 3))**.** *For positive integers d and m, let $H \in M_{d,m}(\mathbb{R}\{z\})$ be a rational matrix function with Kronecker indices $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_d)$. Then a unique minimal algebraic realization (A,B,C) of H of dimension $N = \delta_M(H)$ is given by the following structure.*

(i) The matrix $A = (A_{ij})_{i,j=1,\ldots,d} \in M_N(\mathbb{R})$ is a block matrix with blocks $A_{ij} \in M_{\nu_i,\nu_j}(\mathbb{R})$ given by

$$A_{ij} = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \alpha_{ij,1} & \cdots & \alpha_{ij,\nu_{ij}} & 0 & \cdots & 0 \end{pmatrix} + \delta_{i,j} \begin{pmatrix} 0 & & \\ \vdots & \mathbf{1}_{\nu_i-1} & \\ 0 & & \\ 0 & \cdots & 0 \end{pmatrix}, \tag{4.2a}$$

(ii) $B = (b_{ij}) \in M_{N,m}(\mathbb{R})$ unrestricted,

(iii) if $\nu_i > 0$, $i = 1, \ldots, d$, *then*

$$C = \left(\begin{array}{cccc:cccc:c:cccc} 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & & & & & \\ & & & & 1 & 0 & \ldots & 0 & & & 0_{(d-1),\nu_d} & & \\ & 0_{(d-1),\nu_1} & & & & 0_{(d-2),\nu_2} & & & & 1 & 0 & \ldots & 0 \end{array}\right). \tag{4.2b}$$

If $\nu_i = 0$, the elements of the ith row of C are also freely varying, but we concentrate here on the case where all Kronecker indices ν_i are positive. To compute $\boldsymbol{\nu}$ as well as the coefficients $\alpha_{ij,k}$ and b_{ij} for a given rational matrix function H, several numerically stable and efficient algorithms are available in the literature (see, e. g., Rózsa and Sinha, 1975, and the references therein). The orthogonal invariance inherent in spectral factorization (see Theorem 3.6) implies that this parametrization alone does not ensure identifiability. In the case $m = d$, one remedy is to restrict the parametrization to those transfer functions H satisfying $H(0) = H_0$, for a non-singular matrix H_0. To see how one must constrain the parameters $\alpha_{ij,k}, b_{ij}$ in order to ensure this normalization, we work in terms of left matrix fraction descriptions.

Theorem 4.3 (Echelon MCARMA realization, Guidorzi (1975, Section 3))**.** *For positive integers d and m, let $H \in M_{d,m}(\mathbb{R}\{z\})$ be a rational matrix function with Kronecker indices $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)$. Assume that (A, B, C) is a realization of H, parametrized as in Eqs. (4.2). Then a unique left matrix fraction description $P^{-1}Q$ of H is given by*

$$P(z) = \left[p_{ij}(z)\right]_{i,j=1,\ldots,d}, \quad Q(z) = \left[q_{ij}(z)\right]_{\substack{i=1,\ldots,d \\ j=1,\ldots,m}}, \tag{4.3}$$

where

$$p_{ij}(z) = \delta_{i,j} z^{\nu_i} - \sum_{k=1}^{\nu_{ij}} \alpha_{ij,k} z^{k-1}, \quad q_{ij}(z) = \sum_{k=1}^{\nu_i} \kappa_{\nu_1+\ldots+\nu_{i-1}+k,j} z^{k-1}, \tag{4.4}$$

and the coefficient $\kappa_{i,j}$ is the (i, j)th entry of the matrix $K = TB$, where the matrix $T = (T_{ij})_{i,j=1,\ldots,d} \in M_N(\mathbb{R})$ is a block matrix with blocks $T_{ij} \in M_{\nu_i,\nu_j}(\mathbb{R})$ given by

$$T_{ij} = \begin{pmatrix} -\alpha_{ij,2} & \ldots & -\alpha_{ij,\nu_{ij}} & 0 & \ldots & 0 \\ \vdots & ⋰ & & & & \vdots \\ & & & & & \vdots \\ -\alpha_{ij,\nu_{ij}} & & & & & \vdots \\ 0 & & & & & \vdots \\ \vdots & & & & & \vdots \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 \end{pmatrix} + \delta_{i,j} \begin{pmatrix} 0 & 0 & \ldots & \ldots & 0 & 1 \\ 0 & 0 & \ldots & & 1 & 0 \\ \vdots & \vdots & & ⋰ & & \vdots \\ \vdots & & ⋰ & & \vdots & \vdots \\ 0 & 1 & & \ldots & 0 & 0 \\ 1 & 0 & \ldots & \ldots & 0 & 0 \end{pmatrix}. \tag{4.5}$$

The orders p, q of the polynomials P, Q satisfy $p = \max\{\nu_1, \ldots, \nu_d\}$ and $q \leqslant p - 1$. Using this parametrization, there are different ways to impose the normalization $H(0) = H_0 \in M_{d,m}(\mathbb{R})$. One first observes that the special structure of the polynomials P and Q implies that $H(0) = P(0)^{-1}Q(0) = -(\alpha_{ij,1})_{ij}^{-1}(\kappa_{\nu_1+\ldots+\nu_{i-1}+1,j})_{ij}$. The canonical state space parametrization (A, B, C) given by Eqs. (4.2) therefore satisfies $H(0) = -CA^{-1}B = H_0$ if one makes the coefficients $\alpha_{ij,1}$ functionally dependent on the free parameters $\alpha_{ij,m}$, $m = 1, \ldots \nu_{ij}$ and b_{ij} by setting $\alpha_{ij,1} = -[(\kappa_{\nu_1+\ldots+\nu_{k-1}+1,l})_{kl} H_0^{\sim 1}]_{ij}$, where κ_{ij} are the entries of the matrix K appearing in Theorem 4.3 and $H_0^{\sim 1}$ is a right inverse of H_0. Another possibility, which has the advantage of preserving the multi-companion structure of the matrix A, is to keep the $\alpha_{ij,1}$ as free parameters, and to restrict some of the entries of the matrix B instead. Since $|\det K| = 1$ and the matrix T is thus invertible, the coefficients b_{ij} can be written as $B = T^{-1}K$. Replacing the $(\nu_1 + \ldots + \nu_{i-1} + 1, j)$th entry of K by the (i, j)th entry of the matrix $-(\alpha_{kl,1})_{kl} H_0$ makes some of the b_{ij} functionally dependent on the entries of the matrix A, and results in a state space representation with prescribed Kronecker indices and satisfying $H(0) = H_0$. This latter method has also the advantage that it does not require the matrix H_0 to possess a right inverse. In the special case that $d = m$ and $H_0 = -\mathbf{1}_d$, it suffices to set $\kappa_{\nu_1+\ldots+\nu_{i-1}+1,j} = \alpha_{ij,1}$, for $i, j = 1, \ldots, d$. Examples of normalized low-order canonical parametrizations are given in Tables 1 and 2.

$\boldsymbol{\nu}$	$n(\boldsymbol{\nu})$	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	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ \vartheta_1 & \vartheta_2 & \vartheta_3 & \vartheta_4 \\ 0 & 0 & 0 & 1 \\ \vartheta_5 & \vartheta_6 & \vartheta_7 & \vartheta_8 \end{pmatrix}$	$\begin{pmatrix} \vartheta_9 & \vartheta_{10} \\ \vartheta_1 + \vartheta_4\vartheta_{11} + \vartheta_2\vartheta_9 & \vartheta_3 + \vartheta_2\vartheta_{10} + \vartheta_4\vartheta_{12} \\ \vartheta_{11} & \vartheta_{12} \\ \vartheta_5 + \vartheta_8\vartheta_{11} + \vartheta_6\vartheta_9 & \vartheta_7 + \vartheta_6\vartheta_{10} + \vartheta_8\vartheta_{12} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

TABLE 1. Canonical state space realizations (A, B, C) of normalized ($H(0) = -\mathbf{1}_2$) rational transfer functions in $M_2(\mathbb{R}\{z\})$ with different Kronecker indices $\boldsymbol{\nu}$; the number of parameters, $n(\boldsymbol{\nu})$, includes three parameters for a covariance matrix $\Sigma^{\boldsymbol{L}}$.

$\boldsymbol{\nu}$	$n(\boldsymbol{\nu})$	$P(z)$	$Q(z)$	(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	$\begin{pmatrix} z^2-\vartheta_1 z-\vartheta_2 & -\vartheta_3 z-\vartheta_4 \\ -\vartheta_5 z-\vartheta_6 & z^2-\vartheta_7 z-\vartheta_8 \end{pmatrix}$	$\begin{pmatrix} \vartheta_9 z+\vartheta_2 & \vartheta_{10} z+\vartheta_4 \\ \vartheta_{11} z+\vartheta_6 & \vartheta_{12} z+\vartheta_8 \end{pmatrix}$	$(2,1)$

TABLE 2. Canonical MCARMA realizations (P, Q) with order (p, q) of normalized ($H(0) = -\mathbf{1}_2$) rational transfer functions in $M_2(\mathbb{R}\{z\})$ with different Kronecker indices $\boldsymbol{\nu}$; the number of parameters, $n(\boldsymbol{\nu})$, includes three parameters for a covariance matrix $\Sigma^{\boldsymbol{L}}$.

4.2. **A simulation study.** In order to get a better feeling for how the quasi maximum likelihood estimation procedure performs in reality, we present a simulation study for a bivariate CARMA process with Kronecker indices $(1,2)$, i.e. CARMA indices $(p,q) = (2,1)$. As the driving Lévy process we chose a zero-mean normal-inverse Gaussian (NIG) process $(\boldsymbol{L}(t))_{t\in\mathbb{R}}$. Such processes have been found to be useful in the modelling of stock returns and stochastic volatility, as well as turbulence data (see, e.g., Barndorff-Nielsen, 1997, 1998; Barndorff-Nielsen et al., 2004; Rydberg, 1997). The distribution of the increments $\boldsymbol{L}(t) - \boldsymbol{L}(t-1)$ of a bivariate normal-inverse Gaussian Lévy process is characterized by the density

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

where

$$g(\boldsymbol{x}) = \sqrt{\delta^2 + \langle \boldsymbol{x} - \boldsymbol{\mu}, \Delta(\boldsymbol{x} - \boldsymbol{\mu})\rangle}, \quad \kappa^2 = \alpha^2 - \langle \boldsymbol{\beta}, \Delta\boldsymbol{\beta}\rangle > 0,$$

and $\boldsymbol{\mu} \in \mathbb{R}^2$ is a location parameter, $\alpha \geqslant 0$ is a shape parameter, $\boldsymbol{\beta} \in \mathbb{R}^2$ is a symmetry parameter, $\delta \geqslant 0$ is a scale parameter and $\Delta \in M_2^+(\mathbb{R})$, $\det\Delta = 1$, determines the dependence between the two components of $(\boldsymbol{L}(t))_{t\in\mathbb{R}}$. For our simulation study we chose parameters

$$\delta = 1, \quad \alpha = 3, \quad \boldsymbol{\beta} = (1,1)^T, \quad \Delta = \begin{pmatrix} 5/4 & -1/2 \\ -1/2 & 1 \end{pmatrix}, \quad \boldsymbol{\mu} = -\frac{1}{2\sqrt{31}}(3,2)^T, \tag{4.6}$$

resulting in a skewed, semi-heavy-tailed distribution with mean zero and covariance matrix

$$\Sigma^{\boldsymbol{L}} = \frac{1}{31^{3/2}}\begin{pmatrix} 82 & -28 \\ -28 & 64 \end{pmatrix} \approx \begin{pmatrix} 0.4751 & -0.1622 \\ -0.1622 & 0.3708 \end{pmatrix}. \tag{4.7}$$

A sample of 350 independent replicates of the bivariate CARMA$_{1,2}$ process $(\boldsymbol{Y}(t))_{t\in\mathbb{R}}$ driven by a normal-inverse Gaussian Lévy process $(\boldsymbol{L}(t))_{t\in\mathbb{R}}$ with parameters given in Eq. (4.6) were simulated on the equidistant grid $0, 0.01, \ldots, 2000$ by applying an Euler scheme to the stochastic differential equation (3.5) making use of the canonical parametrization given in Table 1. For the simulation, the initial value $\boldsymbol{X}(0) = \boldsymbol{0}_3$ and $\vartheta_{1:7} = (-1, -2, 1, -2, -3, 1, 2)$ was used. Each realization was sampled at integer times ($h = 1$), and quasi maximum likelihood estimates of $\vartheta_1, \ldots, \vartheta_7$ as well as $(\vartheta_8, \vartheta_9, \vartheta_{10}) := \operatorname{vech}\Sigma^{\boldsymbol{L}}$ were computed by numerical maximization of the quasi log-likelihood function using a differential evolution optimization routine (Price et al., 2005) in conjunction with a subspace trust-region method (Branch et al., 1999; Byrd et al., 1988). In Table 3 the sample means and sampled standard deviations of the estimates are reported. Moreover, the standard deviations were estimated using the square roots of the diagonal entries of the asymptotic covariance matrix (2.45) with $s(L) = \lfloor L/\log L\rfloor^{1/3}$, and the estimates are also displayed in Table 3. One

parameter	sample mean	bias	sample std. dev.	mean est. std. dev.
ϑ_1	-1.0001	0.0001	0.0354	0.0381
ϑ_2	-2.0078	0.0078	0.0479	0.0539
ϑ_3	1.0051	-0.0051	0.1276	0.1321
ϑ_4	-2.0068	0.0068	0.1009	0.1202
ϑ_5	-2.9988	-0.0012	0.1587	0.1820
ϑ_6	1.0255	-0.0255	0.1285	0.1382
ϑ_7	2.0023	-0.0023	0.0987	0.1061
ϑ_8	0.4723	-0.0028	0.0457	0.0517
ϑ_9	-0.1654	0.0032	0.0306	0.0346
ϑ_{10}	0.3732	0.0024	0.0286	0.0378

TABLE 3. Quasi maximum likelihood estimates for the parameters of a bivariate NIG-driven CARMA$_{1,2}$ process observed at integer times over the time horizon $[0, 2000]$. The second column reports the empirical mean of the estimators as obtained from 350 independent simulated paths; the third and fourth columns contain the resulting bias and the sample standard deviation of the estimators, respectively, while the last column reports the average of the expected standard deviations of the estimators as obtained from the asymptotic normality result Theorem 3.17.

sees that the bias, the difference between the sample mean and the true parameter value, is very small in accordance with the asymptotic consistency of the estimator. Moreover, the estimated standard deviation is always slightly larger than the sample standard deviation, yet close enough to provide a useful approximation for, e. g., the construction of confidence regions. In order not to underestimate the uncertainty in the estimate, such a conservative approximation to the true standard deviations is desirable in practice. Overall, the estimation procedure performs very well in the simulation study.

4.3. **Application to weekly bond yields.** In this section we provide an illustrative data example and apply the techniques established in the preceding sections to the bivariate weekly series of Moody's seasoned Aaa and Baa corporate bond yields from October 1966 through April 2009; these data are available from the Federal Reserve Bank of St. Louis. We first took the logarithm of the data and the resulting series was seen to have a unit root in each component, so the next step in the data preparation was differencing at lag 1. Using a moving window of length 52 — corresponding to a period of one year — we removed the stochastic volatility effects displayed by the differenced time series to obtain data with no obvious departure from stationarity. Figure 1 shows the weekly bond log-yields after differencing and devolatilization. We have fitted bivariate CARMA processes of McMillan degrees $n = 2, 3, 4$ using the quasi maximum likelihood method described in Section 3.4 and employing the canonical parametrizations of Section 4.1. The numerical values of $\hat{\boldsymbol{\vartheta}}$ as well as their standard errors estimated by the square root of the diagonal entries in the approximate asymptotic covariance matrix $\hat{\Xi}_s^L$, defined in Eq. (2.45), can be found in Table 4. The last row

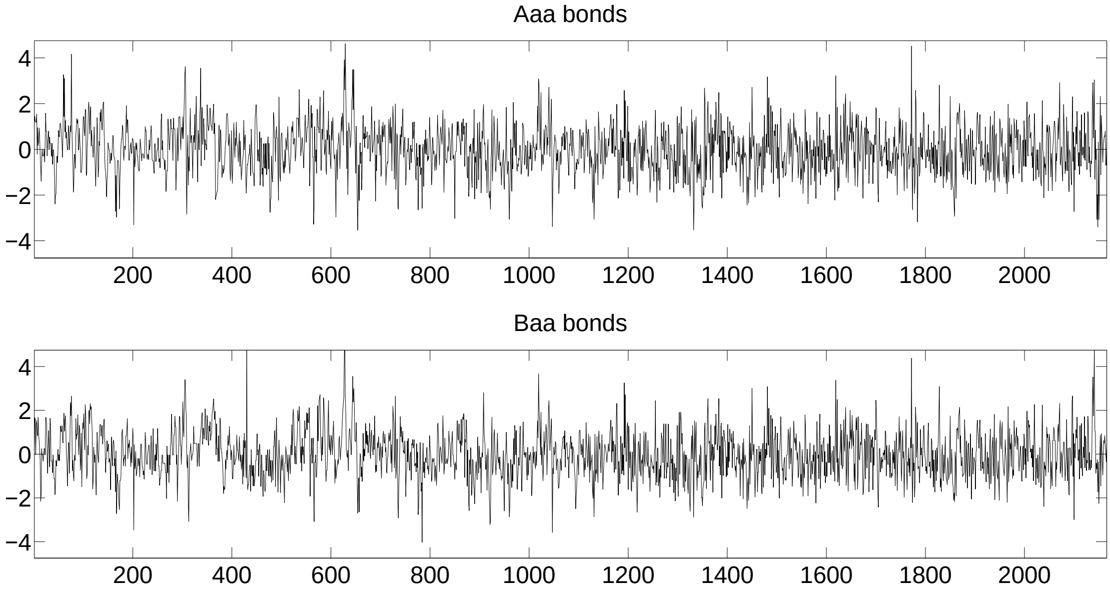

FIGURE 1. Weekly series of Moody's seasoned Aaa and Baa corporate bond yields after differencing and devolatilization

displays the value of twice the negative logarithm of the Gaussian likelihood of the observations under the model corresponding to the estimated parameter value $\hat{\boldsymbol{\vartheta}}$. The quality of the fit can be assessed from Fig. 2

(α,β)	$(1,1)$		$(1,2)$		$(2,1)$		$(2,2)$	
	$\hat{\vartheta}_i$	$\sigma(\vartheta_i)$	$\hat{\vartheta}_i$	$\sigma(\vartheta_i)$	$\hat{\vartheta}_i$	$\sigma(\vartheta_i)$	$\hat{\vartheta}_i$	$\sigma(\vartheta_i)$
$\hat{\vartheta}_1$	-1.1326	0.1349	-1.1538	0.1401	-1.3776	0.0320	**-0.0010**	**0.0336**
$\hat{\vartheta}_2$	**0.2054**	**0.1171**	0.2307	0.1008	-2.4033	0.0197	-1.1601	0.5964
$\hat{\vartheta}_3$	0.3316	0.1206	**-0.2528**	**0.1716**	0.0228	0.0050	**-0.0098**	**0.0268**
$\hat{\vartheta}_4$	-1.0935	0.1065	**-0.0362**	**0.0472**	-4.9948	0.1096	**0.1829**	**0.7429**
$\hat{\vartheta}_5$	2.4105	0.2324	-1.2516	0.1286	-4.6276	0.1538	1.4646	0.3931
$\hat{\vartheta}_6$	2.2483	0.2061	-2.5747	0.4595	**-0.0153**	**0.0108**	1.3662	0.4039
$\hat{\vartheta}_7$	2.7055	0.2116	1.6345	0.2940	-1.2442	0.0391	-0.7438	0.2387
$\hat{\vartheta}_8$			2.8552	0.1966	0.2573	0.0492	-1.7563	0.7209
$\hat{\vartheta}_9$			3.5702	0.2151	2.4302	0.1370	-2.6936	0.6694
$\hat{\vartheta}_{10}$			4.9076	0.3888	2.9784	0.2766	1.7369	0.5381
$\hat{\vartheta}_{11}$					4.1571	0.5043	**-3.6136**	**3.0265**
$\hat{\vartheta}_{12}$							**2.8483**	**2.5122**
$\hat{\vartheta}_{13}$							4.4848	0.3327
$\hat{\vartheta}_{14}$							5.5079	0.1803
$\hat{\vartheta}_{15}$							7.0218	1.4357
$\mathscr{L}(\hat{\boldsymbol{\vartheta}}, \boldsymbol{y}^L)$	**9,893.8**		**9,850.4**		**9,853.0**		**9,840.7**	

TABLE 4. QML estimates of the parameters of an $\mathrm{MCARMA}_{\alpha,\beta}$ model for weekly yields of Moody's seasoned corporate bonds. The marginal standard deviations $\sigma(\vartheta_i)$ are estimated from the asymptotic covariance matrix in Theorem 3.17. The parameters whose confidence region to the level 5% contains zero are marked in bold.

where we compare the autocorrelation functions of the fitted models with the empirical autocorrelation function of the data. One sees how the fit becomes better as one increases the model order in accordance with an increasing value of the Gaussian likelihood; in particular, the autocorrelations of the second component at higher lags are better captured by the higher order models. This phenomenon is well known from the estimation of discrete-time parametric processes where penalty terms in the likelihood together with order selection criteria like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) are used to formalize the trade-off between goodness of fit and model complexity. Understanding their applicability in a continuous-time set-up remains a problem for future research.

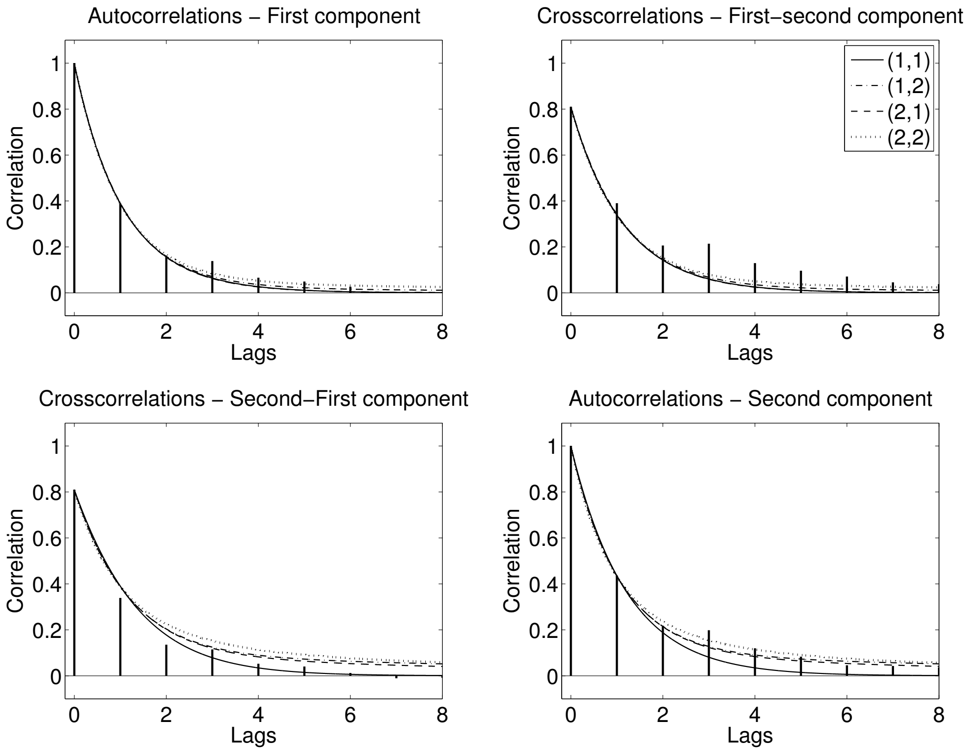

FIGURE 2. Empirical auto- and crosscorrelations of the weekly bond data from Fig. 1 compared to the theoretical auto- and crosscorrelations of estimated MCARMA$_{\alpha,\beta}$ models, for different Kronecker indices (α,β)

ACKNOWLEDGEMENTS

ES acknowledges financial support from the International Graduate School of Science and Engineering of the Technische Universität München and the TUM Institute for Advanced Study, funded by the German Excellence Initiative.

REFERENCES

Applebaum, D. (2004). *Lévy processes and stochastic calculus*, Volume 93 of *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press.

Åström, K. J. (1970). *Introduction to stochastic control theory*, Volume 70 of *Mathematics in Science and Engineering*. New York: Academic Press.

Barndorff-Nielsen, O. E. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scand. J. Stat. 24*(1), 1–13.

Barndorff-Nielsen, O. E. (1998). Processes of normal inverse Gaussian type. *Finance Stoch. 2*(1), 41–68.

Barndorff-Nielsen, O. E., P. Blæsild, and J. Schmiegel (2004). A parsimonious and universal description of turbulent velocity increments. *Eur. Phys. J. B Condens. Matter Phys. 41*(3), 345–363.

Benth, F. E. and J. Šaltytė Benth (2009). Dynamic pricing of wind futures. *Energy Economics 31*(1), 16–24.

Bernstein, D. S. (2005). *Matrix mathematics*. Princeton: Princeton University Press. Theory, facts, and formulas with application to linear systems theory.

Bertoin, J. (1996). *Lévy processes*, Volume 121 of *Cambridge Tracts in Mathematics*. Cambridge: Cambridge University Press.

Boubacar Mainassara, B. and C. Francq (2011). Estimating structural VARMA models with uncorrelated but non-independent error terms. *J. Multivar. Anal. 102*(3), 496–505.

Bradley, R. C. (2007). *Introduction to strong mixing conditions. Vol. 1.* Kendrick Press, Heber City.

Branch, M. A., T. F. Coleman, and Y. Li (1999). A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM J. Sci. Comput. 21*(1), 1–23 (electronic).

Brockwell, P. J. (2001a). Continuous-time ARMA processes. In *Stochastic processes: theory and methods*, Volume 19 of *Handbook of Statistics*, pp. 249–276. Amsterdam: North-Holland.

Brockwell, P. J. (2001b). Lévy-driven CARMA processes. *Ann. Inst. Stat. Math. 53*(1), 113–124.

Brockwell, P. J. and R. A. Davis (1991). *Time series: theory and methods* (Second ed.). Springer Series in Statistics. New York: Springer-Verlag.

Brockwell, P. J., R. A. Davis, and Y. Yang (2011). Estimation for nonnegative Lévy-driven CARMA processes. *J. Bus. Econ. Stat. 29*(2), 250–259.

Brockwell, P. J. and E. Schlemm (2011). Parametric estimation of the driving Lévy process of multivariate CARMA processes from discrete observations. Preprint: Available at `http://www-m4.ma.tum.de`.

Brown, B. M. and J. I. Hewitt (1975). Asymptotic likelihood theory for diffusion processes. *J. Appl. Probab. 12*(2), 228–238.

Byrd, R. H., R. B. Schnabel, and G. A. Shultz (1988). Approximate solution of the trust region problem by minimization over two-dimensional subspaces. *Math. Program. 40*(1), 247–263.

Caines, P. E. (1988). *Linear stochastic systems*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quant. Financ. 1*(2), 223–236.

Davydov, Y. A. (1968). Convergence of distributions generated by stationary stochastic processes. *Theory Probab. Appl. 13*(4), 691–696.

Deistler, M. (1983). The properties of the parameterization of ARMAX systems and their relevance for structural estimation and dynamic specification. *Econometrica 51*(4), 1187–1207.

Dieudonné, J. (1968). *Éléments d'analyse. Tome II: Chapitres XII à XV*. Cahiers Scientifiques, Fasc. XXXI. Éditeur, Paris: Gauthier-Villars.

Doob, J. L. (1944). The elementary Gaussian processes. *Ann. Math. Statistics 15*(3), 229–282.

Duncan, T. E., P. Mandl, and B. Pasik-Duncan (1999). A note on sampling and parameter estimation in linear stochastic systems. *IEEE Trans. Autom. Control 44*(11), 2120–2125.

Durrett, R. (2010). *Probability: theory and examples* (Fourth ed.). Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Fan, H., T. Söderström, M. Mossberg, B. Carlsson, and Y. Zon (1998). Continuous-time AR process parameter estimation from discrete-time data. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, Volume 4, pp. 1232–1244.

Feigin, P. D. (1976). Maximum likelihood estimation for continuous-time stochastic processes. *Adv. Appl. Probab. 8*(4), 712–736.

Ferguson, T. S. (1996). *A course in large sample theory*. Texts in Statistical Science Series. London: Chapman & Hall.

Francq, C. and J.-M. Zakoïan (1998). Estimating linear representations of nonlinear processes. *J. Stat. Plan. Infer. 68*(1), 145–165.

Gevers, M. (1986). ARMA models, their Kronecker indices and their McMillan degree. *Int. J. Control 43*(6), 1745–1761.

Gevers, M. and V. Wertz (1984). Uniquely identifiable state-space and ARMA parametrizations for multivariable linear systems. *Automatica J. IFAC 20*(3), 333–347.

Gillberg, J. and L. Ljung (2009). Frequency-domain identification of continuous-time ARMA models from sampled data. *Automatica 45*(6), 1371–1378.

Guidorzi, R. P. (1975). Canonical structures in the identification of multivariable systems. *Automatica—J. IFAC 11*(4), 361–374.

Halmos, P. R. (1950). *Measure Theory*. New York: D. Van Nostrand Company, Inc.

Hamilton, J. D. (1994). *Time series analysis*. Princeton: Princeton University Press.

Hannan, E. J. (1969). The estimation of mixed moving average autoregressive systems. *Biometrika 56*(3), 579–593.

Hannan, E. J. (1975). The estimation of ARMA models. *Ann. Statist. 3*(4), 975–981.

Hannan, E. J. and M. Deistler (1988). *The statistical theory of linear systems*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.

Hansen, L. P. and T. J. Sargent (1983). The dimensionality of the aliasing problem in models with rational spectral densities. *Econometrica 51*(2), 377–387.

Hautus, M. L. J. (1969). Controllability and observability conditions of linear autonomous systems. *Indag. Math. 31*, 443–448.

Herrndorf, N. (1984). A functional central limit theorem for weakly dependent sequences of random variables. *Ann. Probab. 12*(1), 141–153.

Higham, N. J. (2008). *Functions of matrices*. Philadelphia: Society for Industrial and Applied Mathematics. Theory and computation.

Horn, R. A. and C. R. Johnson (1994). *Topics in matrix analysis*. Cambridge: Cambridge University Press. Corrected reprint of the 1991 original.

Hu, Y. and H. Long (2009). Least squares estimator for Ornstein–Uhlenbeck processes driven by α-stable motions. *Stoch. Process. Their Appl. 119*(8), 2465–2480.

Hyndman, R. J. (1993). Yule–Walker estimates for continuous-time autoregressive models. *J. Time Ser. Anal. 14*(3), 281–296.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory Probab. Appl. 7*, 349–382.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*(1), 35–45.

Klein, A., G. Mélard, and A. Saidi (2008). The asymptotic and exact Fisher information matrices of a vector ARMA process. *Stat. Probab. Lett. 78*(12), 1430–1433.

Klein, A. and H. Neudecker (2000). A direct derivation of the exact Fisher information matrix of Gaussian vector state space models. *Linear Alg. Appl. 321*(1-3), 233–238.

Klenke, A. (2008). *Probability theory*. Universitext. London: Springer-Verlag London Ltd.

Krengel, U. (1985). *Ergodic theorems*, Volume 6 of *de Gruyter Studies in Mathematics*. Berlin: Walter de Gruyter & Co. With a supplement by Antoine Brunel.

Kulikova, M. V. and I. V. Semoushin (2006). Score evaluation within the extended square-root information filter. *Lecture Notes in Computer Science 3991*, 473–481.

Lahalle, E., G. Fleury, and A. Rivoira (2004). Continuous ARMA spectral estimation from irregularly sampled observations. In *Proceedings of the 21st IEEE Instrumentation and Measurement Technology Conference*, Volume 2, pp. 923– 927.

Lancaster, P. and L. Rodman (1995). *Algebraic Riccati equations*. New York – Oxford: The Clarendon Press/Oxford University Press.

Larsson, E. K., M. Mossberg, and T. Söderström (2006). An overview of important practical aspects of continuous-time ARMA system identification. *Circuits Systems Signal Process. 25*(1), 17–46.

Larsson, E. K. and T. Söderström (2002). Identification of continuous-time AR processes from unevenly sampled data. *Automatica J. IFAC 38*(4), 709–718.

Lax, P. D. (2002). *Functional analysis*. Pure and Applied Mathematics. New York: Wiley-Interscience.

Lebesgue, H. (1904). Une propriété caractéristique des fonctions de classe 1. *Bull. Soc. Math. France 32*, 229–242.

Lii, K. S. and E. Masry (1995). On the selection of random sampling schemes for the spectral estimation of continuous time processes. *J. Time Ser. Anal. 16*(3), 291–311.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Berlin: Springer-Verlag.

Lütkepohl, H. and D. S. Poskitt (1996). Specification of echelon-form VARMA models. *J. Bus. Econ. Stat. 14*(1), 69–79.

Marquardt, T. and R. Stelzer (2007). Multivariate CARMA processes. *Stoch. Process. Their Appl. 117*(1), 96–120.

Masry, E. (1978). Poisson sampling and spectral estimation of continuous-time processes. *IEEE Trans. Inf. Theory 24*(2), 173–183.

McCrorie, J. R. (2003). The problem of aliasing in identifying finite parameter continuous time stochastic models. In *Proceedings of the Eighth Vilnius Conference on Probability Theory and Mathematical Statistics, Part II (2002)*, Volume 79, pp. 9–16.

Mokkadem, A. (1988). Mixing properties of ARMA processes. *Stoch. Process. Their Appl. 29*(2), 309–315.

Na, S. S. and H. K. Rhee (2002). An experimental study for property control in a continuous styrene polymerization reactor using a polynomial ARMA model. *Chem. Eng. Sci. 57*(7), 1165–1173.

Nielsen, J. N., H. Madsen, and P. C. Young (2000). Parameter estimation in stochastic differential equations: an overview. *Annu. Rev. Control 24*, 83–94.

Patel, R. V. (1981). Computation of matrix fraction descriptions of linear time-invariant systems. *IEEE Trans. Automat. Control 26*(1), 148–161.

Pham, T. D. (1977). Estimation of parameters of a continuous time Gaussian stationary process with rational spectral density. *Biometrika 64*(2), 385–399.

Price, K. V., R. M. Storn, and J. A. Lampinen (2005). *Differential evolution*. Natural Computing Series. Berlin: Springer-Verlag. A practical approach to global optimization.

Protter, P. (1990). *Stochastic integration and differential equations*, Volume 21 of *Applications of Mathematics*. Berlin: Springer-Verlag. A new approach.

Rajput, B. S. and J. Rosiński (1989). Spectral representations of infinitely divisible processes. *Probab. Theory Relat. Field 82*(3), 451–487.

Reinsel, G. C. (1997). *Elements of multivariate time series analysis* (Second ed.). Springer Series in Statistics. New York: Springer-Verlag.

Rivoira, A., Y. Moudden, and G. Fleury (2002). Real time continuous AR parameter estimation from randomly sampled observations. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 2, pp. 1725–1728.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U. S. A. 42*(1), 43–47.

Rozanov, Y. A. (1967). *Stationary random processes*. San Francisco: Holden-Day Inc. Translated from the Russian by A. Feinstein.

Rózsa, P. and N. K. Sinha (1975). Minimal realization of a transfer function matrix in canonical forms. *Internat. J. Control 21*(2), 273–284.

Rydberg, T. (1997). The normal inverse Gaussian Lévy process: simulation and approximation. *Stoch. Models 13*(4), 887–910.

Sato, K. (1999). *Lévy processes and infinitely divisible distributions*, Volume 68 of *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press.

Sato, K. and M. Yamazato (1983). Stationary processes of Ornstein–Uhlenbeck type. In *Probability theory and mathematical statistics (Tbilisi, 1982)*, Volume 1021 of *Lecture Notes in Math.*, pp. 541–551. Berlin: Springer.

Schlemm, E. and R. Stelzer (2011). Multivariate CARMA processes, continuous-time state space models and complete regularity of the innovations of the sampled processes. *Bernoulli*. To appear.

Segal, M. and E. Weinstein (1989). A new method for evaluating the log-likelihood gradient, the Hessian, and the Fisher information matrix for linear dynamic systems. *IEEE Trans. Inform. Theory 35*(3), 682–687.

Söderström, T., H. Fan, B. Carlsson, and M. Mossberg (1997). Some approaches on how to use the delta operator when identifying continuous-time processes. In *Proceedings of the 36th IEEE Conference on Decision and Control*, Volume 1, pp. 890–895.

Sontag, E. D. (1998). *Mathematical control theory* (Second ed.), Volume 6 of *Texts in Applied Mathematics*. New York: Springer-Verlag. Deterministic finite-dimensional systems.

Spiliopoulos, K. (2009). Method of moments estimation of Ornstein–Uhlenbeck processes driven by general Lévy process. *Ann. I.S.U.P. 53*(2-3), 3–17.

Sun, J. (1998). Sensitivity analysis of the discrete-time algebraic Riccati equation. *Linear Alg. Appl. 275/276*, 595–615.

Todorov, V. and G. Tauchen (2006). Simulation methods for Lévy-driven continuous-time autoregressive moving average (CARMA) stochastic volatility models. *J. Bus. Econ. Stat. 24*(4), 455–469.

TUM Institute for Advanced Study & Fakultät für Mathematik, Technische Universität München, Germany
E-mail address: schlemm@ma.tum.de
URL: http://www-m4.ma.tum.de/pers/schlemm

TUM Institute for Advanced Study & Fakultät für Mathematik, Technische Universität München, Germany
Current address: Institute of Mathematical Finance, Ulm University
E-mail address: robert.stelzer@uni-ulm.de
URL: http://www.uni-ulm.de/mawi/finmath