Closed-form results for vector moving average models with a univariate estimation approach

Federico Poloni[1] & Giacomo Sbrana[2 3]

Abstract

The estimation of a vector moving average (VMA) process represents a challenging task since the likelihood estimator is extremely slow to converge, even for small-dimensional systems. An alternative estimation method is provided, based on computing several aggregations of the variables of the system and applying likelihood estimators to the resulting univariate processes; the VMA parameters are then recovered using linear algebra tools. This avoids the complexity of maximizing the multivariate likelihood directly. Closed-form results are presented and used to compute the parameters of the process as a function of its autocovariances, using linear algebra tools. Then, an autocovariance estimation method based on the estimation of univariate models only is introduced. It is proved that the resulting estimator is consistent and asymptotically normal. A Monte Carlo simulation shows the good performance of this estimator in small samples.

Keywords: VARMA estimation; maximum likelihood; canonical factorization.

1. Introduction

Let $\mathbf{y}_t$ be a vector moving average process of order $q > 0$, (VMA(q))

$$\mathbf{y}_t = \mathbf{v}_t + \sum_{i=1}^{q} \Theta_i \mathbf{v}_{t-i}, \tag{1}$$

where $\mathbf{y}_j \in \mathbb{R}^d$, $\Theta_j \in \mathbb{R}^{d\times d}$, $\mathbf{v}_j \in \mathbb{R}^d$ for each j. Here $\mathbf{v}_t$ is a process of independent and identically distributed (i.i.d.) noises; in particular, $\mathbb{E}[\mathbf{v}_t] = 0$, $\mathbb{E}\left[\mathbf{v}_t\mathbf{v}_t^T\right] = \Sigma_{\mathbf{v}} > 0$, and $\mathbb{E}\left[\mathbf{v}_t\mathbf{v}_s^T\right] = 0$ for any $t \neq s$.

The VMA process represents a relevant framework, widely discussed and employed by the time series literature in the last century (see Reinsel (2003), Lütkepohl

[1]Dipartimento di Informatica, Università di Pisa, Largo Pontecorvo, 3, 56127 Pisa, Italy. E-mail: federico.poloni@unipi.it

[2]NEOMA Business School, 59 Rue Pierre Taittinger, 51100 Reims, France. E-mail: giacomo.sbrana@neoma-bs.fr

[3]Corresponding author. Tel: +33 0326774759.

(2005) and Brockwell and Davis (2009)). Indeed, it represents a potential benchmark in forecasting time series (see Lütkepohl (2012)) and it can be used for impulse response analysis (see for example Plagborg-Møller et al. (2015)). In addition, it appears as the reduced form of DSGE models (see Ravenna (2007)) as well as structural time series models (see Harvey (1990), Durbin and Koopman (2012) and Hyndman et al. (2008)).

The maximum likelihood (ML) estimation of the parameters of a VMA(q) is a challenging task; the maximization of the likelihood function can be computationally intractable even for small-dimensional systems. As a consequence, many statistical and econometric packages estimate vector autoregressive processes (VAR) but not VMA processes or VARMA processes.

To avoid the complications associated to ML estimation, the literature provides several alternative estimation methods based on various approaches (see Dufour and Pelletier (2005); Durbin (1960); Galbraith et al. (2002); Hannan and Kavalieris (1984); Hannan and Rissanen (1982); Kapetanios (2003); Koreisha and Pukkila (1990); Wilson (1973)).

In the first part of our paper, we present several results and algorithms from the numerical linear algebra literature that can be used to compute the parameters of a VMA process as a closed-form function of its autocovariances $\Gamma_k := \mathbb{E}\left[\mathbf{y}_t \mathbf{y}_{t-k}^T\right]$ (spectral factorization). To our knowledge, the only closed-form results for VARMA models appearing in the statistical literature are those given by the Yule-Walker equations (see Lütkepohl (2005)), which deal with the autoregressive part only. All existing methods to deal with the MA part (such as the innovations algorithm (Brockwell and Davis, 2009, Proposition 11.4.2)) require an iterative procedure or are formulated as a minimization problem with no explicit solution. In contrast, the only iterative part of our proposed algorithm lies in the computation of the eigenvalues and eigenvectors of a matrix, which is a well studied problem in numerical linear algebra and is so fast and reliable that it is typically comparable with other non-iterative operations such as matrix multiplication. Hence the first nontrivial contribution of the present paper is introducing this method to the econometrics community.

We then suggest a novel estimation procedure to obtain these autocovariances Γ_k, which works as follows.

1. We choose a vector of weights $\mathbf{w}_0 \in \mathbb{R}^{1\times d}$ (or two vectors $\mathbf{w}_0, \mathbf{w}_1 \in \mathbb{R}^{1\times d}$), and compute the scalar aggregate process $x_t = \mathbf{w}_0\mathbf{y}_t$ (or $x_t = \mathbf{w}_0\mathbf{y}_t + \mathbf{w}_1\mathbf{y}_{t-1}$).
2. We estimate the parameters of the univariate MA process followed by x_t by maximum likelihood. We make use of these parameters to compute estimates for the autocovariances of x_t.
3. We compute, separately, analytic expressions for the autocovariances of the aggregated process in terms of the unknown values of the entries of Γ_k, $k = 0, 1, \dots, q$. Equating these expressions with the values computed in Step 2, we obtain several linear equations in the entries of the matrices Γ_k.
4. We repeat Steps 1–3 for several choices of the aggregation vectors, until we have enough equations to determine the matrices Γ_k completely.

5. We solve these equations using a weighted least-squares procedure, to determine estimates of the Γ_k.
6. We use the spectral factorization technique to recover the parameters Θ_k, $k = 1, 2, \dots, q$, and $\Sigma_{\mathbf{v}}$ from the Γ_k.

Our method generalizes the so-called *META (Moment Estimation Through Aggregation) estimator*, first described for simpler models in Poloni and Sbrana (2015b) and Poloni and Sbrana (2015a). (See Sbrana et al. (2015) for an empirical application of this estimator.) A similar idea of sampling a large-dimensional model several times to simplify it appears in the indirect inference method of Gourieroux et al. (1993) and in the indirect continuous GMM estimator of Kotchoni (2014).

Contrary to most of the alternative estimators mentioned earlier, our method still uses the Gaussian likelihood: however, we replace the multivariate maximum likelihood estimation problem with several univariate ones, with computational advantage.

We provide asymptotic theory for our estimator, proving consistency and normality under the assumption that the noise is i.i.d..

Finally, we present a Monte Carlo simulation to provide evidence of the good performance of the closed-form estimator.

The remainder of the paper is structured as follows. Section 2 describes a closed-form spectral factorization method based on linear algebra computations. Section 3 describes the estimation procedure and its possible variants. A Monte Carlo simulation comparing the small-sample performance of the META approach with those of standard estimation methods is in Section 4. In Section 5 we provide the asymptotic properties of the META estimator, while Section 6 concludes. The proofs of all theorems are relegated to the Appendix.

2. Closed form results

In this section, we describe a method to derive the parameters of a VMA(q) as an analytic function of its autocovariances.

2.1. Autocovariance generating function, transfer function and canonical factorization

In this paper, we rely heavily on the formalism of transfer functions and lag operators (Box and Jenkins (1976); Harvey (1990)), which is a powerful method to derive the properties of linear stochastic processes reducing them to polynomial and rational function manipulation. Given a bi-infinite sequence $\mathbf{v} = (\mathbf{v}_t)_{t=\dots,-1,0,1,\dots}$, we denote by L (*lag operator*) the map defined by $(L\mathbf{v})_t = \mathbf{v}_{t-1}$. Moreover, for any rational function $F \in \mathbb{C}(L)^{m\times n}$ in the parameter L, we set $F(L)^\star := F(L^{-1})^T$. Here and in the rest of the paper, we use the notation A^T to denote the transpose of a matrix A, and use bold letters for vectors and uppercase letters for matrices.

A large family of time series, called *stationary linear models*, can be written as $\mathbf{y} = G(L)\mathbf{v}$, where $\mathbf{v} = (\mathbf{v}_t)_{t=\dots,-1,0,1,\dots}$, (with $\mathbf{v}_t \in \mathbb{R}^m$ for each t), is a family of i.i.d. random variables with $\mathbb{E}[\mathbf{v}_t] = 0$ and $\mathbb{E}\left[\mathbf{v}_t\mathbf{v}_t^T\right] = \Sigma_{\mathbf{v}} > 0$, and $G(L) \in \mathbb{R}(L)^{d\times m}$ is a rational function in L, called *transfer function*. Among them is the VMA(q) as in (1), for which $m = d$ and $G(L) = \Theta(L) := I + \sum_{i=1}^q \Theta_i L^i$. Given a linear model $(\mathbf{y}_t)$, with $\mathbf{y}_t \in \mathbb{R}^d$ for each t, we define its *autocovariances* as $\Gamma_k := \mathbb{E}[\mathbf{y}_t \mathbf{y}_{t-k}] \in \mathbb{R}^{d\times d}$, for $k \in \mathbb{Z}$, and its *autocovariance generating function* as $\Gamma(L) := \sum_{i=-\infty}^{\infty} \Gamma_i L^i \in \mathbb{R}(L)^{d\times d}$. It is immediate to prove that $\Gamma(L) = \Gamma(L)^\star$, i.e., that $\Gamma_{-i} = \Gamma_i^T$ for each $i \in \mathbb{Z}$. Matrix rational functions satisfying this properties are known as *palindromic* (see Chu et al. (2010)). Moreover, the following result holds (Harvey, 1990, Equation 8.1.25).

Lemma 1. *Let* $\mathbf{y} = G(L)\mathbf{v} \in \mathbb{R}^{d\times m}$ *be a stationary linear model; then,*

$$\Gamma(L) = G(L)\Sigma_{\mathbf{v}} G(L)^\star. \tag{2}$$

From this lemma, it is immediate to see that $\Gamma(\mathrm{e}^{i\lambda}) = \Gamma(\mathrm{e}^{-i\lambda})^T \geq 0$ for each $\lambda \in [0, 2\pi]$, where by the notation $\Gamma(\mathrm{e}^{i\lambda})$ we mean replacing the variable L with the complex number $\mathrm{e}^{i\lambda}$, which lies on the unit circle. The function $\Gamma(\mathrm{e}^{i\lambda})$ is also known as the (asymptotic) *spectral density matrix* of the process $\mathbf{y}_t$.

We wish to derive a method to solve the inverse problem, that is, given an autocovariance generating function $\Gamma(L)$, finding a suitable stationary linear model $G(L)$ satisfying (2). We start from a $\Gamma(L)$ satisfying the following assumptions.

Assumption P $\Gamma(\mathrm{e}^{i\lambda})$ is positive definite for each $\lambda \in [0, 2\pi]$.

Assumption Q $\Gamma(L) \in \mathbb{R}(L)^{d\times d}$ is a palindromic Laurent polynomial of degree q, i.e., $\Gamma_k = 0$ whenever $|k| > q$.

If the two assumptions hold, there exists a unique factorization of the form (2) with $m = d$, i.e.,

$$\Gamma(L) = \Theta(L)\Sigma_{\hat{\mathbf{v}}}\Theta(L)^\star, \quad \Sigma_{\hat{\mathbf{v}}} \in \mathbb{R}^{d\times d}, \quad \Theta(L) = I + \sum_{i=1}^{q} \Theta_i L^i \in \mathbb{R}[L]^{d\times d}, \tag{3}$$

where the VMA(q) process $\Theta(L)$ is *invertible*, that is, it holds that $\det\Theta(z) \neq 0$ whenever $|z| \leq 1$.

This means that we can reparametrize any model whose ACGF satisfies Assumptions P and Q as an invertible VMA(q) with uncorrelated noise $\hat{\mathbf{v}}_t$. The factorization (3) has been widely studied, not only for polynomials functions but also for more general forms of $\Gamma(L)$, in several fields, such as operator theory (canonical factorization, Bart et al. (2010)), control theory (J-spectral factorization, Hunt (1993)), and time series (spectral density, Rozanov (1967) and Hamilton (1994), Wold decomposition, Fuller (1996)). For an elementary proof of this result (with minor differences), see for instance Ephremidze (2010).

2.2. Linearization of a palindromic matrix polynomial

A square rational matrix function $F(L) \in \mathbb{R}(L)^{m\times m}$ is called *regular* if $\det F(L)$ is not identically zero. In this case, its *eigenvalues* are defined as the values $\lambda \in \mathbb{C}$ such that $\det F(\lambda) = 0$. The *eigenvectors* associated to λ are defined as vectors $\mathbf{x}$ in $\mathbb{C}^d$ such that $F(\lambda)\mathbf{x} = 0$.

The following result is classical in linear algebra (see Gohberg et al. (1982)), although with several variants on how the blocks of the matrix C are laid out.

Lemma 2 (companion form). *Let $P(L) = \sum_{i=0}^{k} P_i L^i \in \mathbb{C}[L]^{d\times d}$ be a matrix polynomial, with the leading coefficient P_k nonsingular, and let*

$$C = \begin{bmatrix} 0 & 0 & \dots & 0 & -P_k^{-1}P_0 \\ I & 0 & \dots & 0 & -P_k^{-1}P_1 \\ 0 & I & \ddots & 0 & -P_k^{-1}P_2 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & I & -P_k^{-1}P_{k-1} \end{bmatrix}, \tag{4}$$

where each block has size $d \times d$. Let $(\lambda_i, \mathbf{u}_i)$, $i = 1, 2, \dots dk$, be the eigenvalues and associated eigenvectors of C. Then, the eigenvalues of $P(L)$ and their associate eigenvectors are

$$\lambda_i,\ \mathbf{x}_i := \begin{bmatrix} 0 & 0 & \dots & 0 & I \end{bmatrix} \mathbf{u}_i.$$

2.3. From moments to parameters: a simpler version

While spectral factorizations are widely studied from the theoretical point of view, in most references (especially for the multivariate case) the solution is given only as an integral representation or in an abstract form that does not allow an easy computation. Here, we present a practical algorithm to compute them in the case in which $\Gamma(L)$ is a matrix Laurent polynomial.

We start by presenting in this subsection a simpler version of the algorithm. This first version is not fully general as it requires a few nonsingularity and nondegenerateness assumptions, and may suffer from some numerical instability, but still it is (1) simpler to explain for a first approach (2) suitable for implementation in most software packages, as it only requires a function to compute eigenvalues and eigenvectors of a matrix. A more general and rigorous version of this approach, requiring fewer assumptions but more sophisticated linear algebra tools, is in Section 2.4.

Let us consider an invertible VMA(q) process with transfer function

$$\Theta(L) = I + \sum_{i=1}^{q} \Theta_i L^i.$$

Since the factorization (3) is unique, the problem of determining $\Theta(L)$ and $\Sigma_{\mathbf{v}}$ given $\Gamma(L) = \Theta(L)\Sigma_{\mathbf{v}}\Theta(L)^\star$ is equivalent to spectral factorization. The determinant $\det\Gamma(z)$ vanishes if and only if z is an eigenvalue of $\Theta(L)$ or of $\Theta(L)^\star$. Let us suppose that $\Theta(L)$ has qd distinct eigenvalues. Thanks to the invertibility assumption, each of them has modulus greater than 1; hence, we shall denote them as $\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_{qd}^{-1}$, with $|\lambda_i| < 1$ for each $i = 1, 2, \dots, qd$. The eigenvalues of $\Theta(L)^\star$ are then given by λ_i, for $i = 1, 2, \dots, qd$.

We can find numerically the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_{qd}, \lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_{qd}^{-1}$ and eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{2qd}$ of $\Gamma(L)$ by constructing the matrix C in (4) for the matrix polynomial $P(L) = L^q\Gamma(L)$ and applying Lemma 2.

Let now $H(L) := L^q\Theta(L)^\star$, which is a polynomial in L. It follows from the invertibility assumption that, for each $i = 1, 2, \dots, qd$, in the product $\lambda_i^q\Gamma(\lambda_i) = \Theta(\lambda_i)\Sigma_{\hat{\mathbf{v}}}H(\lambda_i)$ the first factor $\Theta(\lambda_i)$ is nonsingular, as well as $\Sigma_{\mathbf{v}}$, and hence $H(\lambda_i)\mathbf{x}_i =$

0. Therefore, $(\lambda_i, \mathbf{x}_i)$ for $i = 1, 2, \dots, qd$ are the eigenvalues and eigenvectors of the matrix polynomial $H(L)$.

Now all we are missing is a method to reconstruct a matrix polynomial given its eigenpairs. Such a method is described in the book Gohberg et al. (1982):

Theorem 3 (Gohberg et al. (1982))**.** *Let $\lambda_1, \lambda_2, \dots, \lambda_{qd}$ and $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{qd}$ be the eigenvalues and eigenvectors of a degree-q matrix polynomial $H(L) \in \mathbb{C}^{d\times d}[L]$, and let*

$$X_1 = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_{qd} \end{bmatrix} \in \mathbb{C}^{d\times qd}, \quad D = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_{qd}) \in \mathbb{C}^{qd\times qd}. \tag{5}$$

Then, the matrix

$$Y = \begin{bmatrix} X_1 \\ X_1 D \\ X_1 D^2 \\ \vdots \\ X_1 D^{q-1} \end{bmatrix} \in \mathbb{C}^{qd\times qd} \tag{6}$$

is nonsingular. Partition Y^{-1} into $qd \times d$ blocks $\begin{bmatrix} V_0 & V_1 & \dots V_{q-1} \end{bmatrix}$; then,

$$H(L) = IL^q - X_1 D^q V_{q-1} L^{q-1} - X_1 D^q V_{q-2} L^{q-2} - \dots - X_1 D^q V_1 L - X_1 D^q V_0. \tag{7}$$

Hence the coefficients Θ_i of the VMA(q) (1) are given by

$$\Theta_i = -(X_1 D^q V_{q-i})^T, \tag{8}$$

with X_1, D as in (5).

Once we have determined the coefficients Θ_i, the value of $\Sigma_{\hat{\mathbf{v}}}$ can be obtained for instance by evaluating (3) in $L = 1$, i.e.,

$$\Sigma_{\hat{\mathbf{v}}} = (I + \Theta_1 + \Theta_2 + \dots + \Theta_q)^{-1} \Gamma(1) \left((I + \Theta_1 + \Theta_2 + \dots + \Theta_q)^{-1} \right)^*. \tag{9}$$

Putting everything together, we obtain Algorithm 1.

2.4. From moments to parameters: a more rigorous version

It is important to identify the steps which lack rigor in the previous discussion.

- To apply Lemma 2, we are assuming that the leading coefficient is nonsingular, i.e., $\det \Gamma_q \neq 0$. This need not be the case.
- We are assuming that the matrix rational function $\Gamma(L)$ has $2qd$ distinct eigenvectors; this need not be the case: it could have multiple eigenvalues and Jordan chains.

The aim of this section is giving a more thorough treatment of this material, including an algorithm that works in a numerically robust way even in case of repeated or clustered eigenvalues. Most of the material is taken from Gohberg et al. (1988) and Gohberg et al. (1982).

Algorithm 1: Spectral factorization of a polynomial function (simpler version).

Input: degree q and coefficients $\Gamma_0 = \Gamma_0^T, \Gamma_1, \dots, \Gamma_q$ of an ACGF satisfying Assumptions P and Q.

Output: coefficients $\Sigma_{\hat{\mathbf{v}}}$ and Θ_i, $i = 1, 2, \dots, q$ of its factorization (3).

Construct the companion matrix C (as in (4)) for the matrix polynomial $P(L) = L^q \Gamma(L)$;

compute its eigenvalues and eigenvectors, which must come in pairs $(\lambda, 1/\lambda)$;

label by $\lambda_1, \lambda_2, \dots, \lambda_{qd}$ and $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{qd}$ the eigenpairs for which $|\lambda_i| < 1$;

form the matrices X, D, Y as in (5) and (6);

invert Y and denote by V_i its blocks, as in Theorem 3;

compute Θ_i, for $i = 1, 2, \dots, q$, using (8), and $\Sigma_{\hat{\mathbf{v}}}$ using (9).

First of all, we wish to get rid of the matrix inverse in (4). Following Gohberg et al. (1988), we define

$$\tilde{C}(L) = \begin{bmatrix} I & 0 & 0 & \dots & 0 & 0 \\ 0 & I & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & I & 0 \\ 0 & 0 & 0 & \dots & 0 & P_{2q} \end{bmatrix} L - \begin{bmatrix} 0 & 0 & \dots & 0 & -P_0 \\ I & 0 & \dots & 0 & -P_1 \\ 0 & I & \ddots & 0 & -P_2 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & I & -P_{2q-1} \end{bmatrix}, \tag{10a}$$

$$R = \begin{bmatrix} 0 & 0 & \dots & I \end{bmatrix} \in \mathbb{C}^{d \times 2qd}, \tag{10b}$$

where all blocks $d \times d$. Here, $\tilde{C}(L)$ is a *matrix pencil*, i.e., a linear matrix polynomial. The pair $(R, \tilde{C}(L))$ is a *right pencil pair* for the matrix polynomial $P(L)$. Informally, this means that the pencil $\tilde{C}(L)$ has the same eigenvalues and multiplicities as $P(L)$, and that the pair $(R, \tilde{C}(L))$ can be used to recover its right eigenvectors. See (Gohberg et al., 1988, Section I.2) for a rigorous definition and in particular Proposition 2.2 therein for a proof of this statement.

A right pencil pair $(X, LE - A)$ is called *strictly equivalent* to $(R, \tilde{C}(L))$ if there are two invertible matrices $F_1, F_2 \in \mathbb{C}^{dn \times dn}$ such that $LE - A = F_1(\tilde{C}(L))F_2$ and $X = RF_2$ (Gohberg et al., 1988, Section I.4). Note that this implies that the eigenvalues of $LE - A$ are the same as those of $\tilde{C}(L)$ and $P(L)$, since multiplying by invertible matrices on both sides preserves the eigenvalues of the pencil.

We introduce now another result in numerical linear algebra, the *generalized Schur decomposition*, also known as *QZ decomposition* (Golub and Van Loan (2013)).

Theorem 4 ((Golub and Van Loan, 2013, Theorem 7.7.1), Kågström (1993))**.** *Given a matrix pencil $LE - A \in \mathbb{C}^{m \times m}[L]$, there are unitary matrices $Q, Z \in \mathbb{C}^{m \times m}$ such that $S := QEZ$ and $T := QAZ$ are upper triangular matrices. In particular, if $LE - A$ is regular, its eigenvalues are given by T_{ii}/S_{ii}, $i = 1, 2, \dots, m$. Moreover, one can find such a factorization in which the ratios T_{ii}/S_{ii} come in any prescribed order.*

We first prove the following lemma.

Lemma 5. *Given a right pencil pair $(X, LE - A)$ with no eigenvalues on the unit circle, let*

$$Q(LE - A)Z = L\begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix} - \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \tag{11}$$

be its generalized Schur decomposition, with the eigenvalues ordered so that those of $LS_{11} - T_{11}$ are inside the unit circle and those of $LS_{22} - T_{22}$ are outside, and let $XZ = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ be partitioned with the same row block sizes. Then, there exists a matrix $\hat{X}_2$ such that $(X, LE - A)$ is strictly equivalent to

$$\left(\begin{bmatrix} X_1 & \hat{X}_2 \end{bmatrix}, \begin{bmatrix} LI - T_1 & 0 \\ 0 & LT_2 - I \end{bmatrix} \right), \tag{12}$$

with $T_1 := S_{11}^{-1} T_{11}$, $T_2 := T_{22}^{-1} S_{22}$.

Note that proofs of lemmas and theorems are relegated in the Appendix. Now, let us start from $P(L) = L^q \Gamma(L) = \Theta(L)\Sigma_{\mathbf{v}} H(L)$, of degree $2q$, as above; we construct $\tilde{C}(L)$ as in (10a), and compute its ordered Schur decomposition $Q\tilde{C}(L)Z$, decomposed as in (11), and $XZ = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$. The pencil (12), in the language of Gohberg et al. (1988), is a Γ-*decomposed right pencil pair*, with Γ equal to the unit circle, and $H(L)$ is a Γ-*spectral right divisor*. Hence we may apply the "only if" part of (Gohberg et al., 1988, Theorem 3.2), and conclude that the two blocks in (12) have both size qd, and that the matrix

$$Y = \begin{bmatrix} X_1 \\ X_1 T_1 \\ X_1 T_1^2 \\ \vdots \\ X_1 T_1^{q-1} \end{bmatrix} \tag{13}$$

is nonsingular.

The final part of (Gohberg et al., 1988, Theorem 3.2) provides a formula to reconstruct the polynomial $H(L)$ from X_1 and T_1, although in a different form that does not yield the coefficients explicitly; however, by comparing with (Gohberg et al., 1982, Theorem 2.4), we see that we can use (7) and (8) instead, as in the previous section but with D replaced by T_1.

To summarize the results of this section, an improved algorithm to recover Θ_i from the ACGF $\Gamma(L)$ of a VMA(q) of dimension d is presented as Algorithm 2.

Remark 6. Software packages such as Matlab, Mathematica and R contain functions to compute the generalized Schur decomposition needed here. The computation is numerically robust even in the case of repeated or clustered eigenvalues.

3. Covariance estimation: the META approach

Having good estimates of the autocovariances Γ_k is crucial for the accuracy of methods based on spectral factorization. Sample autocovariances $\hat{\Gamma}_k := \frac{1}{n-k}\sum_{t=k}^{n} \mathbf{y}_t \mathbf{y}_{t-k}^T$

Algorithm 2: Spectral factorization of a polynomial function (in a more rigorous and stable way than Algorithm (1)).

Input: degree q and coefficients $\Gamma_0 = \Gamma_0^T, \Gamma_1, \dots, \Gamma_q$ of an ACGF satisfying Assumptions P and Q.

Output: coefficients $\Sigma_{\hat{\mathbf{v}}}$ and Θ_i, $i = 1, 2, \dots, q$ of its factorization (3).

1 Construct the right pencil pair $(R, \tilde{C}(L))$ in (10a) for the matrix polynomial $P(L) = L^q \Gamma(L)$;

2 compute a generalized Schur decomposition of $\tilde{C}(L)$, ordered so that the eigenvalues inside the unit circle are in the first entries, and partition the resulting matrices as in (11);

3 let X_1 be the first block of $X = RZ = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$, and $T_1 = S_{11}^{-1} T_{11}$;

4 invert the matrix Y in (13), and denote by V_i its blocks, as in Theorem 3;

5 compute Θ_i, for $i = 1, 2, \dots, q$, using (8), and $\Sigma_{\hat{\mathbf{v}}}$ using (9).

are notoriously slow to converge to their asymptotic values. In this section, we suggest a method to extract these moment estimates from the ML estimates of several (univariate) aggregated processes, replacing one large-dimensional optimization problem with many small-dimensional ones. This procedure generalizes and extends the approach proposed by Poloni and Sbrana (2015b) and Poloni and Sbrana (2015a), which works only for VMA(q) processes with symmetric transfer functions. The method is based on the following result, which is an easy consequence of the existence of the canonical factorization (3).

Lemma 7. *Let $\mathbf{y} = (\mathbf{y}_t)$ be a stationary linear process with degree-q ACGF $\Gamma(L)$ satisfying Assumptions P and Q. Let $\mathbf{w}(L) \in \mathbb{R}[L]^{1\times d}$, $\mathbf{w}(L) \neq 0$, be a vector polynomial of degree r, and consider the process $x^{(\mathbf{w})} = \mathbf{w}(L)\mathbf{y}$. Then,*

1. *The ACGF of $x^{(\mathbf{w})}$, which is the palindromic Laurent polynomial*

$$\gamma^{(\mathbf{w})}(L) = \gamma_{q+r}^{(\mathbf{w})} L^{-q-r} + \cdots + \gamma_1^{(\mathbf{w})} L^{-1} + \gamma_0^{(\mathbf{w})} + \gamma_1^{(\mathbf{w})} L^{1} + \cdots + \gamma_{q+r}^{(\mathbf{w})} L^{q+r},$$

satisfies the equation

$$\mathbf{w}(L)\Gamma(L)\mathbf{w}(L)^\star = \gamma^{(\mathbf{w})}(L). \tag{14}$$

2. $\gamma^{(\mathbf{w})}(e^{i\lambda}) > 0$ *for each* $\lambda \in [0, 2\pi]$*, i.e.,* $\gamma^{(\mathbf{w})}(L)$ *satisfies Assumption P.*
3. $x^{(\mathbf{w})}$ *can be reparametrized as a (univariate) MA($q + r$)*

$$x^{(\mathbf{w})} = \theta(L)u^{(\mathbf{w})}, \quad \theta^{(\mathbf{w})}(L) = 1 + \sum_{i=1}^{q+r} \theta_i^{(\mathbf{w})} L^i, \quad \mathbb{E}\left[\left(u_t^{(\mathbf{w})}\right)^2\right] = \omega^{(\mathbf{w})} > 0. \tag{15}$$

We obtain a realization of the aggregated process $x_t^{(\mathbf{w})}$ using the formula $x_t^{(\mathbf{w})} = \mathbf{w}^{(0)}\mathbf{y}_t + \mathbf{w}^{(1)}\mathbf{y}_{t-1} + \cdots + \mathbf{w}^{(r)}\mathbf{y}_{t-r}$, where the $\mathbf{w}^{(i)}$ are the coefficients of $\mathbf{w}(L)$, i.e., $\mathbf{w}(L) = \mathbf{w}^{(0)} + \mathbf{w}^{(1)}L + \cdots + \mathbf{w}^{(r)}L^r$. Using maximum likelihood, we can get an estimator $\hat{\beta}_{\mathbf{w}}$ for the vector of parameters

$$\beta_{\mathbf{w}} = \begin{bmatrix} \omega^{(\mathbf{w})} \\ \theta_1^{(\mathbf{w})} \\ \theta_2^{(\mathbf{w})} \\ \vdots \\ \theta_{q+r}^{(\mathbf{w})} \end{bmatrix} \in \mathbb{R}^{q+r+1}. \tag{16}$$

We gather in a vector

$$\gamma_{\mathbf{w}} = \begin{bmatrix} \gamma_0^{(\mathbf{w})} \\ \gamma_1^{(\mathbf{w})} \\ \vdots \\ \gamma_{q+r}^{(\mathbf{w})} \end{bmatrix} \in \mathbb{R}^{q+r+1} \tag{17}$$

the coefficients of $\gamma^{(w)}(L)$ (autocovariances). Closed-form expressions for them as a function of $\beta_{\mathbf{w}}$ are simple to obtain by expanding the expression $\gamma^{(\mathbf{w})}(L) = \theta^{(\mathbf{w})}(L)\omega^{(\mathbf{w})}\theta^{(\mathbf{w})}(L^{-1})$ and equating coefficients. For instance, if $q = r = 1$, one has

$$\gamma_{\mathbf{w}} = \begin{bmatrix} \gamma_0^{(\mathbf{w})} \\ \gamma_1^{(\mathbf{w})} \\ \gamma_2^{(\mathbf{w})} \end{bmatrix} = \begin{bmatrix} \omega^{(\mathbf{w})}(1 + (\theta_1^{(\mathbf{w})})^2 + (\theta_2^{(\mathbf{w})})^2) \\ \omega^{(\mathbf{w})}(\theta_1^{(\mathbf{w})} + \theta_1^{(\mathbf{w})}\theta_2^{(\mathbf{w})}) \\ \omega^{(\mathbf{w})}\theta_2^{(\mathbf{w})} \end{bmatrix}. \tag{18}$$

We use these expressions, adding hats to each variable, to compute an estimator $\hat{\gamma}_{\mathbf{w}}$ from $\hat{\beta}_{\mathbf{w}}$. We can interpret these estimates as giving us partial information on a (yet to determine) estimator $\hat{\Gamma}(L)$ of $\Gamma(L)$, according to the relation

$$\mathbf{w}(L)\hat{\Gamma}(L)\mathbf{w}(L)^{\star} = \hat{\gamma}^{(\mathbf{w})}(L). \tag{19}$$

Indeed, if we gather the unknown entries of the ACGF in the parameter vector

$$\hat{\mathbf{z}} = \begin{bmatrix} \operatorname{vech}(\hat{\Gamma}_0) \\ \operatorname{vec}(\hat{\Gamma}_1) \\ \vdots \\ \operatorname{vec}(\hat{\Gamma}_q) \end{bmatrix} \in \mathbb{R}^{m\times 1}, \quad m = \frac{d(d+1)}{2} + qd^2, \tag{20}$$

expanding both sides of (19) as Laurent polynomials and equating coefficients gives $q + r + 1$ linear equations involving the entries of $\hat{\mathbf{z}}$.

We repeat this process for a sufficient number of different vectors $\mathbf{w}(L)$, until we have enough equations to determine the entries of $\hat{\mathbf{z}}$.

3.1. An example

Consider the following bivariate VMA(1) process

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix}.$$

Let us label the coefficients of its autocovariances as

$$\Gamma_0 = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad \Gamma_1 = \begin{bmatrix} d & e \\ f & g \end{bmatrix}.$$

We first choose a constant vector polynomial $\mathbf{w}(L) = \begin{bmatrix} 1 & 0 \end{bmatrix}$. The left-hand side of Equation (14) is

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \left(\begin{bmatrix} d & f \\ e & g \end{bmatrix} L^{-1} \begin{bmatrix} a & b \\ b & c \end{bmatrix} + \begin{bmatrix} d & e \\ f & g \end{bmatrix} L \right) \begin{bmatrix} 1 \\ 0 \end{bmatrix} = dL^{-1} + a + dL, \tag{21}$$

while its right-hand side is

$$(1 + \theta_1 L)\omega_1(1 + \theta_1 L^{-1}) = \theta_1 \omega_1 L^{-1} + (1 + \theta_1^2)\omega_1 + \theta_1 \omega_1 L, \tag{22}$$

where $x_t = u_t + \theta_1 u_{t-1}$, $\mathbb{E}\left[u_t^2\right] = \omega_1$ is the MA(1) representation of the process $x_t = \mathbf{w}(L)\mathbf{y}_t = y_{1,t}$. We obtain estimates $\hat{\theta}_1$, $\hat{\omega}_1$ using a scalar maximum-likelihood estimator on the time series $y_{1,t}$, and equate the coefficients of (21) and (22), getting the two equations

$$\hat{a} = (1 + \hat{\theta}_1^2)\hat{\omega}_1, \quad \text{and} \tag{23a}$$

$$\hat{d} = \hat{\theta}_1 \hat{\omega}_1. \tag{23b}$$

We repeat the procedure on the vector $\mathbf{w}_2(L) = \begin{bmatrix} 0 & 1 \end{bmatrix}$, obtaining

$$\hat{c} = (1 + \hat{\theta}_2^2)\hat{\omega}_2, \quad \text{and} \tag{23c}$$

$$\hat{g} = \hat{\theta}_2 \hat{\omega}_2, \tag{23d}$$

where $\hat{\theta}_2$, $\hat{\omega}_2$ are estimates for the parameters of the MA(1) representation $y_{2,t} = u_t + \theta_2 u_{t-1}$, $\mathbb{E}\left[u_t^2\right] = \omega_2$.

We now use $\mathbf{w}_3(L) = \begin{bmatrix} 1 & L \end{bmatrix}$: this time, the process $\mathbf{w}_3(L)\mathbf{y}_t = y_{1,t} + y_{2,t-1}$ is a scalar MA(2) with ACGF

$$\mathbf{w}_3(L)\Gamma(L)\mathbf{w}_3(L)^\star = \hat{f} L^{-2} + (\hat{d} + \hat{b} + \hat{g})L^{-1} + \hat{a} + \hat{c} + 2\hat{e} + (\hat{d} + \hat{b} + \hat{g})L + \hat{f} L^2,$$

so we get equations

$$\hat{a} + \hat{c} + 2\hat{e} = \hat{\omega}\left[1 + (\hat{\theta})^2 + (\hat{\psi})^2\right], \tag{23e}$$

$$\hat{d} + \hat{b} + \hat{g} = \hat{\theta}\hat{\omega} + \hat{\theta}\hat{\omega}\hat{\psi}, \quad \text{and} \tag{23f}$$

$$\hat{f} = \hat{\psi}\hat{\omega}. \tag{23g}$$

where $\hat{\theta}$, $\hat{\psi}$, $\hat{\omega}$ are obtained via a scalar maximum-likelihood estimator on the MA representation $y_{1,t} + y_{2,t-1} = u_t + \theta u_{t-1} + \psi u_{t-2}$, $\mathbb{E}\left[u_t^2\right] = \omega$.

Equations (23a)–(23g) form a nonsingular system of 7 equations in 7 unknowns, from which we can recover all the entries of Γ_0 and Γ_1.

Remark 8. If one uses only aggregation vectors $\mathbf{w}(L)$ which do not depend explicitly on L, the resulting equations depend on e and f only through the quantity $e+f$. Hence using nonconstant values of $\mathbf{w}(L)$ is necessary to get a nonsingular system.

Remark 9. There is no guarantee that the estimated ACGF $\hat{\Gamma}(L) = \hat{\Gamma}_1^T L^{-1} + \hat{\Gamma}_0 + \hat{\Gamma}_1 L$ satisfies Assumption P. We present in Section 3.3 several possible solutions to continue the estimation process even in the case in which an indefinite spectral density matrix is returned.

Remark 10. An equivalent of Lemma 7 does not hold for more generic VARMA processes. For example, the contemporaneous aggregation of a d-dimensional VAR(1) is, in general, an ARMA(d, $d-1$), not an AR(1) (see the discussion in Granger and Morris (1976) and Hamilton (1994)). Although this does not exclude the possibility to recover the ACGF from the aggregation of scalar processes, it is not immediate to generalize this estimation strategy to a VARMA(p, q).

3.2. Estimating autocovariances using weighted least-squares

In the example in Section 3.1, we have a square linear system with 7 equations in 7 unknowns. We present in this section a more general approach that allows one to make use of a larger number of aggregation vectors. The framework is the theory of overdetermined linear systems and linear least squares (see e.g. (Golub and Van Loan, 2013, Sections 5 and 6)).

For a given aggregation polynomial $\mathbf{w}(L)$ of degree $r^{(\mathbf{w})}$, we can write the equations obtained from (19) as

$$X_{\mathbf{w}} \hat{\mathbf{z}} = \hat{\gamma}_{\mathbf{w}}, \tag{24}$$

where $X_{\mathbf{w}} \in \mathbb{R}^{(q+r^{(\mathbf{w})}+1)\times m}$ is a fixed coefficient matrix, which depends only on the choice of $\mathbf{w}(L)$. (An explicit formula for this matrix $X_{\mathbf{w}}$ is given in Appendix Appendix B.)

We repeat the aggregation with k different sets of weights $\mathbf{w}_1(L), \mathbf{w}_2(L), \dots, \mathbf{w}_k(L)$, and combine these equations to get a larger linear system

$$X\hat{\mathbf{z}} = \hat{\gamma}, \quad X = \begin{bmatrix} X_{\mathbf{w}_1} \\ X_{\mathbf{w}_2} \\ \vdots \\ X_{\mathbf{w}_k} \end{bmatrix}, \quad \hat{\gamma} = \begin{bmatrix} \hat{\gamma}_{\mathbf{w}_1} \\ \hat{\gamma}_{\mathbf{w}_2} \\ \vdots \\ \hat{\gamma}_{\mathbf{w}_k} \end{bmatrix}. \tag{25}$$

In the general case, this system is overdetermined, but we can obtain a solution in the least squares sense as

$$\hat{\mathbf{z}} = \arg\min \left\| \hat{W}^{1/2}(X\hat{\mathbf{z}} - \hat{\gamma}) \right\|, \tag{26}$$

for any given positive definite weighting matrix $\hat{W}$. Standard theory leads to the closed form

$$\hat{\mathbf{z}} = (X^T \hat{W} X)^{-1} (X^T \hat{W} \hat{\gamma}), \tag{27}$$

Setting $\hat{W} = I$, for instance, corresponds to ordinary least squares.

Note that we are not in the usual setting in which least squares are used in statistics: we work with a fixed number k of aggregation vectors, and we are not interested in the behavior when $k \to \infty$, but rather when the number of observations of the time series n tends to infinity and $\hat{\gamma}$ converges to its exact asymptotic value. Also, the errors in the entries of $\hat{\gamma}$ are not i.i.d., in the typical case.

Nevertheless, we can use some statistical insight to make a more effective choice of $\hat{W}$. We expect the ideal value of $\hat{W}$ to be $W = V^{-1}$, where V is the asymptotic covariance matrix of $\hat{\gamma}$. Partitioning V conformably with (25), we have

$$V = \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1k} \\ V_{21} & V_{22} & \cdots & V_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ V_{k1} & V_{k2} & \cdots & V_{kk} \end{bmatrix}.$$

Each diagonal block V_{ii}, for $i = 1, 2, \ldots, k$, contains the asymptotic covariance $V_{\mathbf{w}_i}$ of $\hat{\gamma}_{\mathbf{w}_i}$.

Under the assumption of Gaussian noises, a closed-form expression for the asymptotic covariance $\Omega_{\mathbf{w}}$ of the ML estimator $\hat{\beta}_{\mathbf{w}}$ of a scalar MA is classical (see (Box and Jenkins, 1976, Section 7.2.6) and (Brockwell and Davis, 2009, Example 8.8.2)). Then, by propagation of uncertainty, the asymptotic covariance of $\gamma_{\mathbf{w}}$ is

$$V_{\mathbf{w}} = J_{\mathbf{w}} \Omega_{\mathbf{w}} J_{\mathbf{w}}^T, \quad \text{with } J_{\mathbf{w}} = \frac{\partial \gamma_{\mathbf{w}}}{\partial \beta_{\mathbf{w}}}.$$

For instance, if $q + r = 2$, we have

$$\Omega_{\mathbf{w}} = \begin{bmatrix} 2(\omega^{(\mathbf{w})})^2 & 0 & 0 \\ 0 & 1 - (\theta_2^{(\mathbf{w})})^2 & \theta_1^{(\mathbf{w})}(1 - \theta_2^{(\mathbf{w})}) \\ 0 & \theta_1^{(\mathbf{w})}(1 - \theta_2^{(\mathbf{w})}) & 1 - (\theta_2^{(\mathbf{w})})^2 \end{bmatrix}$$

and, differentiating (18),

$$J_{\mathbf{w}} = \begin{bmatrix} 1 + (\theta_1^{(\mathbf{w})})^2 + (\theta_2^{(\mathbf{w})})^2 & 2\theta_1^{(\mathbf{w})}\omega^{(\mathbf{w})} & 2\theta_2^{(\mathbf{w})}\omega^{(\mathbf{w})} \\ \theta_1^{(\mathbf{w})} + \theta_1^{(\mathbf{w})}\theta_2^{(\mathbf{w})} & (1 + \theta_2^{(\mathbf{w})})\omega^{(\mathbf{w})} & \theta_1^{(\mathbf{w})}\omega^{(\mathbf{w})} \\ \theta_2^{(\mathbf{w})} & 0 & \omega^{(\mathbf{w})} \end{bmatrix}. \tag{28}$$

Replacing everywhere $\theta_i^{(\mathbf{w})}$ and $\omega^{(\mathbf{w})}$ with their estimated values, we obtain estimates $\hat{V}_{ii}$ for each diagonal blocks V_{ii}, $i = 1, 2, \ldots, k$. The elements of V outside its block diagonal are more complicated to estimate, since the aggregated processes $x^{(\mathbf{w})}$ are not independent from each other, even in the case of Gaussian noises. We give an explicit (although long) expression to compute the full covariance matrix in the proof of Theorem 19. However, in practice, we recommend replacing the off-diagonal blocks with zeros and set $\hat{W} = \operatorname{diag}(\hat{V}_{11}, \hat{V}_{22}, \ldots, \hat{V}_{kk})^{-1}$. This is just a crude approximation, but it allows one to keep into account at least the different variances of each entry of γ, and obtain a more accurate estimation than unweighted least squares.

As an additional benefit, with this block diagonal form of $\hat{W}$ we get the formulas

$$X^T \hat{W} X = \sum_{i=1}^{k} X_{\mathbf{w}_i}^T \hat{V}_{ii}^{-1} X_{\mathbf{w}_i}, \qquad \text{and } X^T \hat{W} \hat{\gamma} = \sum_{i=1}^{k} X_{\mathbf{w}_i}^T \hat{V}_{ii}^{-1} \hat{\gamma}_{\mathbf{w}_i}, \tag{29}$$

which are simpler and more computationally effective than the general version, as they do not require assembling the matrices X and $\hat{W}$ and inverting the latter.

Remark 11. The asymptotic properties of the estimator proved in Section 5 hold irrespective of the choice of $\hat{W}$.

Remark 12. The least-squares problem (26) is uniquely solvable if and only if the matrix X has full column rank. A necessary restriction for this to happen is that one chooses at least as many equations as unknowns; i.e.,

$$\sum_{\mathbf{w} \in \mathscr{W}} (q + r^{(\mathbf{w})} + 1) \geq qd^2 + \frac{d(d+1)}{2}. \tag{30}$$

As argued in Section 3.1, some vectors with $r^{(\mathbf{w})} \geq 1$ are necessary, because using degree-0 vectors only yields equations that depend on Γ_i only through $(\Gamma_i + \Gamma_i^T)$, for each $i \geq 1$. In our experiments, choosing random weights and a sufficient number of degree-1 vectors gives a matrix X with full column rank whenever (30) holds; we haven't investigated further theoretical conditions to ensure full rank of X.

Remark 13. For (30) to hold, a number $k \approx d^2$ of weight vectors is sufficient. Hence the computational cost of this estimation procedure is $O(nd^2)$ operations, where n is the number of samples. Estimators which work directly on the multivariate problem performing operations on d-dimensional matrices and vectors (such as multivariate maximum likelihood) also require $O(nd^2)$ or $O(nd^3)$ operations.

3.3. Positiveness of the spectral density matrix

As highlighted in Remark 9, this procedure is not guaranteed to produce an ACGF $\hat{\Gamma}(L)$ which satisfies Assumption P (positive definiteness of the spectral density matrix for each frequency λ). We show an example in Figures 1–4: the first two figures show a case in which the estimation procedure produces an estimated ACGF satisfying the assumption; the next two show a rarer one in which this assumption is not satisfied.

There are three possible ways to solve this problem.

1. Repeat the estimation procedure with a different set of weights $\mathscr{W}$.
2. Use the method presented in Brüll et al. (2013) to compute the Laurent polynomial $\tilde{\Gamma}(L)$ closest to $\hat{\Gamma}(L)$ which satisfies Assumption P.
3. Replace $\hat{\Gamma}(L)$ with $\hat{\Gamma}(L) + tI$, for a suitable value of $t > 0$. In practice, we try several values of t, starting from $t = 0.001\|\hat{\Gamma}_0\|$ and increasing it iteratively until the assumption is satisfied. In terms of the plots in Figures 3–4, adding a multiple of the identity corresponds to translating each of the dashed lines up by an amount t; so, at least in this example, one can see that the amount t needed to move them above the x axis is negligible with respect to the error already performed by the estimation procedure. This procedure is inspired by the well-known ideas of Tikhonov regularization and shrinkage estimation.

Figure 1: Eigenvalues of the true and estimated spectral density matrices for a simulated example of Model 2 with $n = 300$. The eigenvalues of $\Gamma(e^{i\lambda})$ for $\lambda \in [-\pi, \pi]$ are in gray; in black dashed the eigenvalues of $\hat{\Gamma}(e^{i\lambda})$ for an instance of $\hat{\Gamma}(L)$ generated by our estimation procedure.

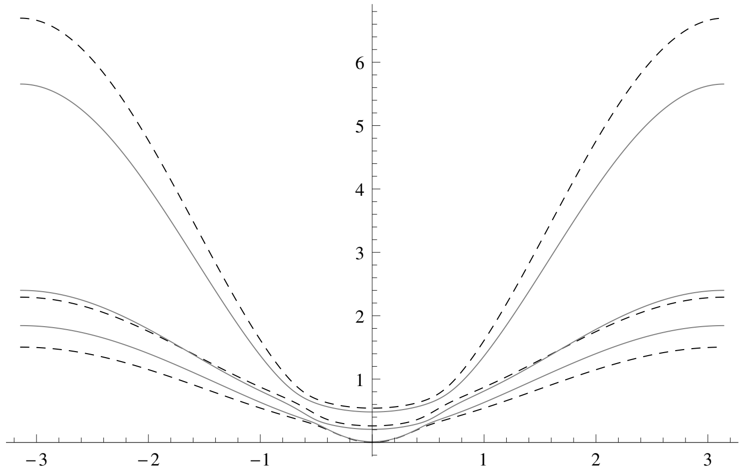


Figure 2: Zoom of Figure 1

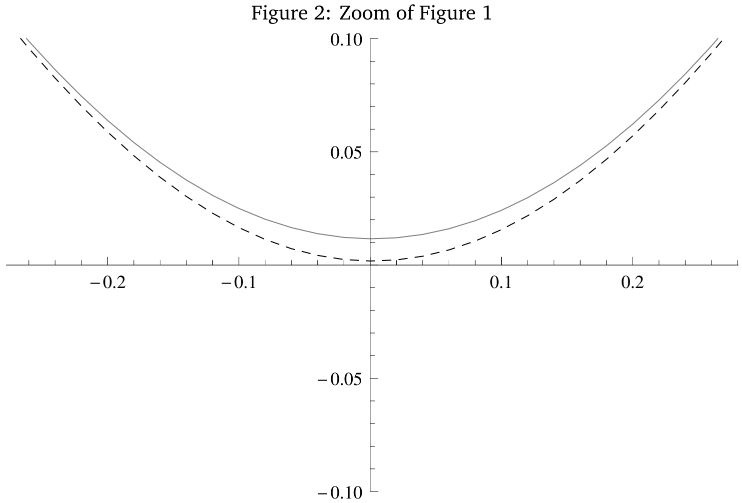


Figure 3: Eigenvalues of the true and estimated spectral density matrices for another example of Model 2 with $n = 300$. Unlike the case in Figure 1, here $\hat{\Gamma}(L)$ does not satisfy Assumption P.

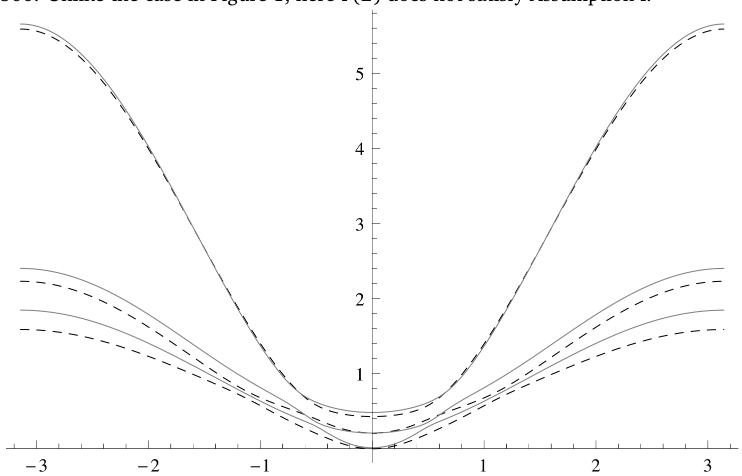
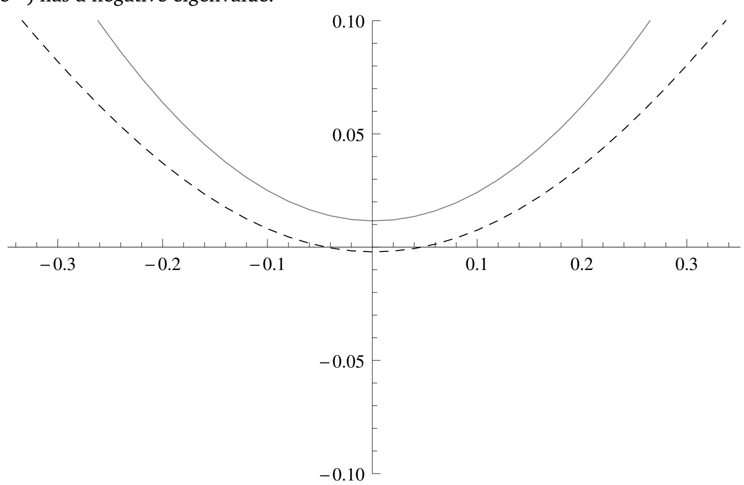


Figure 4: Zoom of Figure 3. One of the dashed lines crosses the x axis, so for $\lambda \in [-0.04, 0.04]$ the matrix $\hat{\Gamma}(e^{i\lambda})$ has a negative eigenvalue.



Among them, we decided to adopt the latter method. Although simple, it seems to work well in practice.

Remark 14. Since $\hat{\Gamma}(L)$ is a regular matrix polynomial,

$$\hat{t} = \min\{z : z \text{ is an eigenvalue of } \hat{\Gamma}(\mathrm{e}^{i\lambda}) \text{ for some } \lambda \in [0, 2\pi]\}$$

exists finite, so this procedure always succeeds, because increasing t iteratively at some point we get $t > \hat{t}$ and hence $\hat{\Gamma}(\mathrm{e}^{i\lambda}) + tI$ is positive definite for each λ.

To sum up, our covariance estimation algorithm is presented as Algorithm 3. The

---

Algorithm 3: META algorithm for the estimation of a VMA(q) process.

Input: Degree q and observed values $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ of a linear model whose ACGF $\Gamma(L)$ satisfies Assumptions P and Q.

Output: Estimates $\hat{\Theta}_i$ for the coefficients of its invertible VMA(q) representation.

Choose a set of k aggregation weight polynomials $\mathscr{W} = \{\mathbf{w}_1(L), \mathbf{w}_2(L), \dots, \mathbf{w}_k(L)\}$ such that the matrix X has full column rank.

foreach $\mathbf{w} \in \mathscr{W}$ **do**

  compute the observations of the aggregated MA($q + r$) process $x = \mathbf{w}(L)\mathbf{y}$;

  estimate its parameters $\hat{\boldsymbol{\beta}}_{\mathbf{w}}$ using a univariate ML estimator;

  compute estimated moments $\hat{\gamma}_{\mathbf{w}}$ from $\hat{\boldsymbol{\beta}}_{\mathbf{w}}$, using formulas such as (18);

  construct the matrix $\hat{V}_{\mathbf{w}} = \hat{J}_{\mathbf{w}} \hat{\Omega}_{\mathbf{w}} \hat{J}_{\mathbf{w}}^T$, where $\hat{J}_{\mathbf{w}}$ and $\hat{\Omega}_{\mathbf{w}}$ are defined like $J_{\mathbf{w}}$ and $\Omega_{\mathbf{w}}$, but replacing the (unknown) exact values $\theta_i^{(\mathbf{w})}$ and $\omega^{(\mathbf{w})}$ with their estimates $\hat{\boldsymbol{\beta}}_{\mathbf{w}}$;

end

compute $X^T \hat{W} X$ and $X^T \hat{W} \hat{\gamma}$ using (29);

compute $\hat{\mathbf{z}} = (X^T \hat{W} X)^{-1} (X^T \hat{W} \hat{\gamma})$;

repeat

  try computing Θ_i, $i = 1, 2, \dots, q$, using Algorithm (1) or (2);

  replace $\hat{\Gamma}_0$ with $\hat{\Gamma}_0 + 0.001 \|\hat{\Gamma}_0\| I$.

until *the algorithm (1 or 2) succeeds (detecting qd eigenvalues inside the unit circle and qd outside)*;

---

most computationally intensive part of the algorithm are the k scalar ML estimations, which can be performed in parallel.

Remark 15. We do not have to worry about the properties of this regularization procedure when we present the asymptotic theory, because if the method is consistent then $\hat{\Gamma}(L) \xrightarrow{\text{a.s.}} \Gamma(L)$ and hence it is positive definite almost surely.

Remark 16. Suppose that the estimation method is applied to a model with roots on the unit circle, i.e., $\det G(e^{i\lambda}) = 0$ for a finite number of values $\lambda \in \mathbb{R}$ (and hence, in particular, Assumption P is not satisfied). In terms of the spectral plots of Figures 3–4, this means that the gray line (exact eigenvalues) is tangent to the x-axis. Note that in this case the aggregate process $\gamma^{(\mathbf{w})}(L)$ still satisfies Assumption P, except from the unlikely case in which the aggregation weight vector $\mathbf{w}(L)$ is chosen such that $\mathbf{w}(e^{i\lambda})G(e^{i\lambda}) = 0$, i.e., it matches exactly a left eigenvector of $G(L)$. Hence for almost all choices of the aggregation weights the asymptotic properties described in Appendix Appendix C hold, and the method produces a consistent estimated ACGF $\hat{\Gamma}(L)$ (whose spectral plot may either touch the x-axis or lie entirely above it). The regularization procedure described in this section will then produce a nearby $\tilde{\Gamma}(L)$ which satisfies Assumption P, and hence an estimated $\hat{\Theta}(L)$ which is always invertible and converges (assuming that the parameter t is chosen in a way such that $t \to 0$ as $n \to \infty$) to the factor $\Theta(L)$ of a factorization (3) of the process with $\det \Theta(z) \neq 0$ for $|z| < 1$.

4. Comparing estimation methods

In this section we investigate the small sample properties of the META approach by comparing it with two generally employed estimation methods. These are the (long) autoregressive approach based on the least squares estimation and the conditional maximum likelihood approach. The indirect estimation of moving average models through autoregressive processes dates back to Durbin (1959)(see also variants such as Hannan and Rissanen (1982); Kapetanios (2003); Koreisha and Pukkila (1990)). A similar estimator for multivariate models has been considered by Spliid (1983) and Galbraith et al. (2002). Its main attractive lies in the fact that the estimation of the AR part can be performed in a simple way using least squares.

The conditional maximum likelihood estimator (CML) has been studied in Wilson (1973); Dunsmuir and Hannan (1976); Hannan and Deistler (2012) and Harvey (1990). It is possible to evaluate the multivariate likelihood function exactly, but the maximization procedure requires the use of multivariate and high-dimensional optimization techniques. We refer to Kascha (2012) for a general discussion and comparison of different estimation approaches.

Our empirical comparison has been carried out using two different software packages, each with its strength and weaknesses:

- Wolfram Mathematica 8 with the Time Series 1.4 package (Wolfram Research, 2007). Mathematica has excellent symbolic computation capabilities, which make it easy to deal with Laurent polynomials in L in full generality; moreover, it includes a CML estimator for multivariate VARMA models. On the minus side, the language is intrinsically extremely slow, and this shows especially in the performance of MA estimators, which we need in abundance. As it is an interpreted language, its speed is extremely dependent on the specific way in which a function is coded and executed, and hence it is less suitable to perform time comparisons.

- EViews 8 (IHS Global, 2013). EViews is a standard package in econometrics; its capabilities are more limited when it comes to abstraction and symbolic computation, and it does not include a CML estimator for multivariate time

Model n.	# of parameters	# of vectors $\mathbf{w}_i(L)$ with degree $r=0$	# of vectors $\mathbf{w}_i(L)$ with degree $r=1$
Model 1	7	4	16
Model 2	15	6	30
Model 3	11	8	25
Model 4	24	10	40

Table 1: Number of vectors of degree 0 and 1 used in each experiment.

series. Hence the only estimator against which we can compare is the long-AR method. On the other hand, the included scalar MA estimator is much faster than the one in Mathematica, and this is crucial to properly assess the speed of our method.

The experiments have been performed using an Intel(R) Core(TM) i5.3210M CPU @ 2.5 GHz.

We consider two cases: the VMA(1) and the VMA(2). For each of them we take a bivariate and a trivariate model. The chosen coefficient matrices are

$$\text{Model 1:}\quad \Theta_1 = \begin{bmatrix} -0.5 & -0.3 \\ -0.1 & -0.7 \end{bmatrix} \; \Sigma_{\mathbf{v}} = \begin{bmatrix} 1 & .2 \\ .2 & 1.3 \end{bmatrix}$$

$$\text{Model 2:}\quad \Theta_1 = \begin{bmatrix} -.6 & -.1 & -.2 \\ -.1 & -.7 & -.2 \\ -.1 & -.2 & -.5 \end{bmatrix} \; \Sigma_{\mathbf{v}} = \begin{bmatrix} 1 & .1 & .2 \\ .1 & 1.2 & .2 \\ .2 & .2 & 1.4 \end{bmatrix}$$

$$\text{Model 3:}\quad \Theta_1 = \begin{bmatrix} -0.6 & -0.4 \\ -0.2 & -0.7 \end{bmatrix} \; \Theta_2 = \begin{bmatrix} 0.5 & 0.4 \\ 0.2 & 0.4 \end{bmatrix} \; \Sigma_{\mathbf{v}} = \begin{bmatrix} 1 & .1 \\ .1 & 1.2 \end{bmatrix}$$

$$\text{Model 4:}\quad \Theta_1 = \begin{bmatrix} -.6 & -.3 & -.3 \\ -.2 & -.7 & -.2 \\ -.2 & -.2 & -.7 \end{bmatrix} \; \Theta_2 = \begin{bmatrix} .3 & .2 & .2 \\ .1 & .5 & .1 \\ .2 & .2 & .4 \end{bmatrix} \; \Sigma_{\mathbf{v}} = \begin{bmatrix} 1 & .2 & .2 \\ .2 & 1.3 & .2 \\ .2 & .2 & 1.1 \end{bmatrix} .$$

For each model, we considered two different sample sizes, $n = 300$ and $n = 800$, and used Gaussian random-generated noises $\mathbf{v}_t$. Each experiment has been repeated 1000 times, each time with different random numbers.

We used the least-squares approach described in Section 3.2, with some weight vectors of the form $\mathbf{w}(L) = \mathbf{w}^{(0)}$, of degree $r = 0$, and some of the form $\mathbf{w}(L) = \mathbf{w}^{(0)} + \mathbf{w}^{(1)}L$, of degree $r = 1$. The entries of these vectors $\mathbf{w}^{(0)}$ (and of $\mathbf{w}^{(1)}$, when it is present) are drawn independently from a normal distribution $N(0,1)$. The number of weight vectors used in each experiment is detailed in Table 1. As error measure, we used the relative error in the Frobenius norm (root mean squared error of the matrix entries)

$$RMSE = \frac{\|\hat{\Theta}_i - \Theta_i\|_F}{\|\Theta_i\|_F} \quad \text{for } i = 1,2; \tag{31}$$

Figure 5: Accuracy comparison of several estimation methods: RMSE(Θ_1) for Model 1 with $n = 300$ and $n = 800$

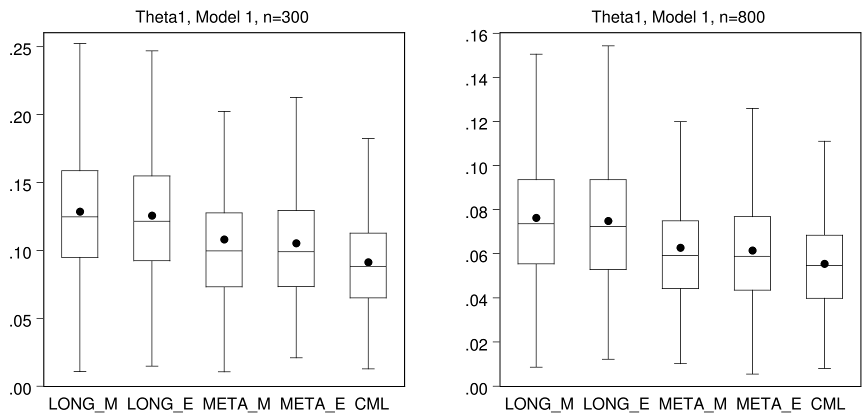


Figure 6: Accuracy comparison of several estimation methods: RMSE(Θ_1) for Model 2 with $n = 300$ and $n = 800$

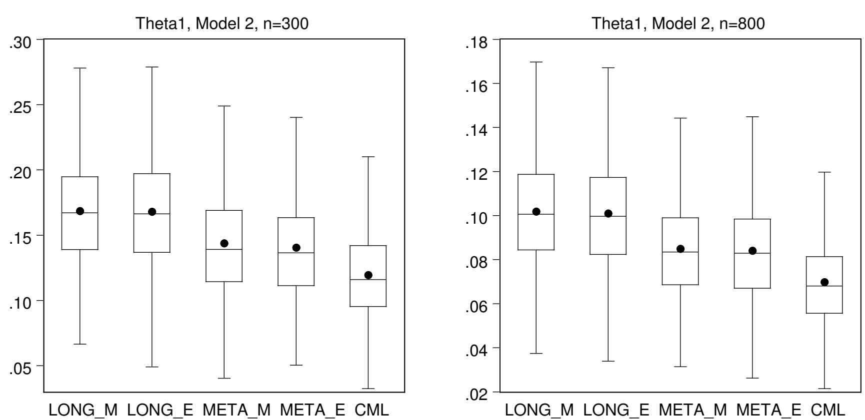


Figure 7: Accuracy comparison of several estimation methods: RMSE(Θ_1) and RMSE(Θ_2) for Model 3 with $n = 300$

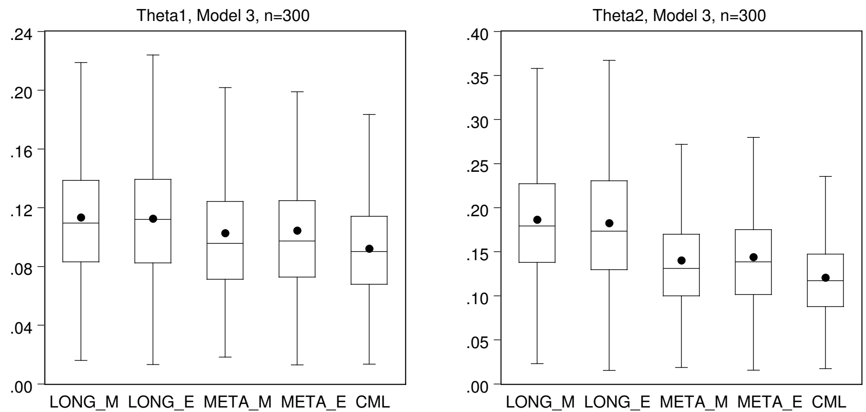


Figure 8: Accuracy comparison of several estimation methods: RMSE(Θ_1) and RMSE(Θ_2) for Model 3 with $n = 800$

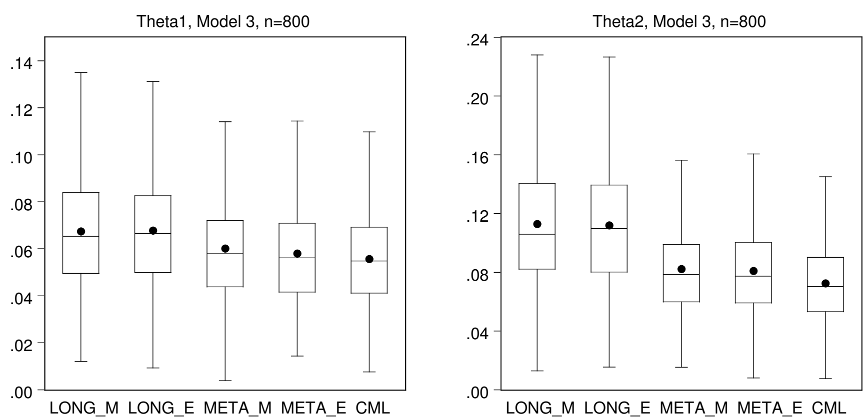


Figure 9: Accuracy comparison of several estimation methods: RMSE(Θ_1) and RMSE(Θ_2) for Model 4 with $n = 300$

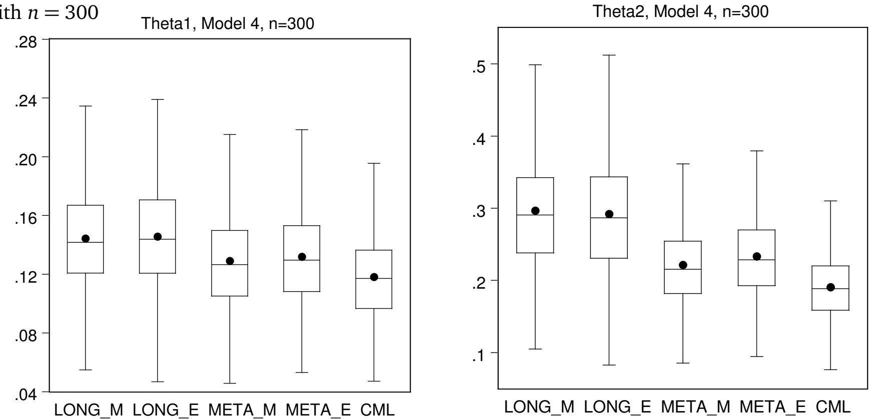


Figure 10: Accuracy comparison of several estimation methods: RMSE(Θ_1) and RMSE(Θ_2) for Model 4 with $n = 800$

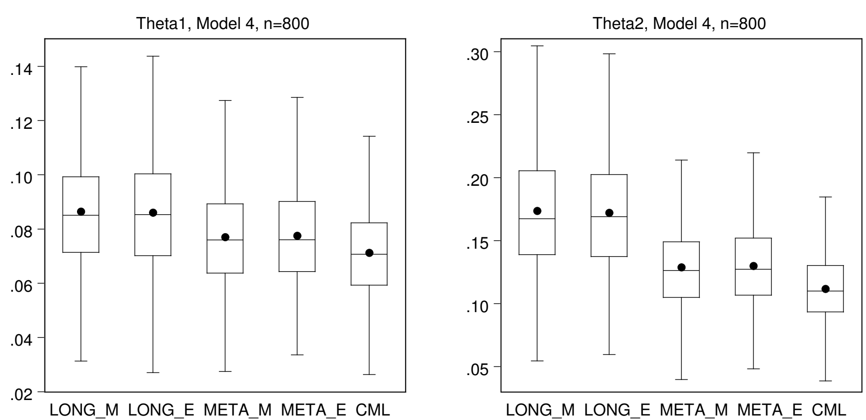


Figure 11: Average computational time needed for the estimation by each software (in seconds)

Software	Method	Model1 N=300	Model2 N=300	Model3 N=300	Model4 N=300
Mathematica	LONG	.003	.003	.003	.003
Mathematica	META	3.81	7.61	24.69	41.73
Mathematica	CML	.118	.176	.177	.258
Eviews	LONG	.001	.002	.001	.002
Eviews	META	.023	.048	.053	.088
		Model1 N=800	Model2 N=800	Model3 N=800	Model4 N=800
Mathematica	LONG	.004	.005	.004	.005
Mathematica	META	6.61	12.86	29.89	49.71
Mathematica	CML	.229	.312	.341	.456
Eviews	LONG	.002	.002	.001	.002
Eviews	META	.059	.105	.128	.193

The empirical results (RMSE) are shown in Figure 5 to 10 using box plots. Each chart compares the performance of the long autoregression (with 10 lags, i.e., we estimate first a VAR(10) model), the META and the CML. For all methods but the CML, the subscripts $_M$ and $_E$ denote the software employed (Mathematica or Eviews). The time required by each estimator is in Figure 11.

One of the main empirical finding is that the META outperforms in terms of accuracy the LONGAR but not the CML for all the considered models. In addition, For Model 1 and Model 3, The performance of the META tends to get closer to that of the CML estimator.

Remark 17. A common addition to the LONGAR setup is truncating the model using an information criterion (AIC or BIC) and refining the estimate (Hannan-Rissanen estimator). These additions did not improve the accuracy of the results, in our experiments. Another common estimation strategy is the multivariate innovations algorithm (Brockwell and Davis, 2009, Proposition 11.4.2). This estimation method also did not improve the accuracy of the LONGAR setup. Finally, we also considered the sample estimation of the ACGF but it returned significantly worse results than the LONGAR method, so it has not been included in the comparison.

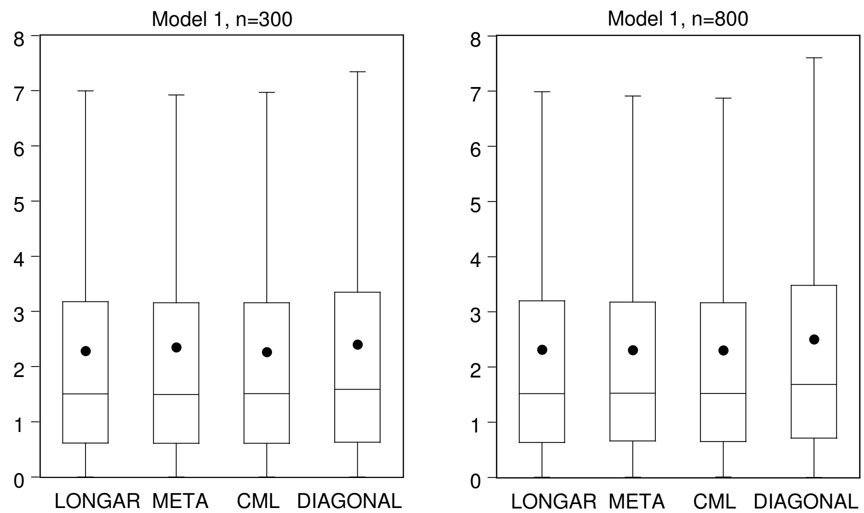
The implementation of our method in Mathematica is extremely slow, but this is an artifact of the Mathematica implementation; the experiments performed with EViews reveal a time which is in line with the other estimation methods.

4.1. Forecasting and the effectiveness of VMA models

The optimal 1-step-ahead forecast for a VMA(q) is given by

$$\hat{\mathbf{y}}_{n+1} = \sum_{i=1}^{q} \Theta_i \mathbf{v}_{n-i-1}.$$

Figure 12: Forecasting accuracy comparison of several estimation methods: $\|y_{n+1} - \hat{y}_{n+1}\|^2$ for Model 1 with $n = 300$ and $n = 800$



Hence accurate estimates the coefficients Θ_i directly imply better forecasting accuracy. However, due to the large impact of the new noise vector $\mathbf{v}_{n+1}$, it might be complicated to reveal this accuracy empirically in simulated experiments. We have computed the forecast error $\|\mathbf{y}_{n+1} - \hat{\mathbf{y}}_{n+1}\|^2$ obtained using the estimates $\hat{\Theta}_i$ from the different methods. As an additional competitor, we have compared against estimating a MA(q) on each component of the time series separately (labelled "diagonal" in the graphs). Clearly, this method ignores completely the cross-correlation between the processes.

The results are reported in Figures 12 to 15. Result show that there is a forecast gain when using estimation methods for VMA models, rather than relying on forecasting equation-by-equation. This would be probably more visible by increasing the number of Monte Carlo experiments that might reveal the difference in forecasting accuracy which we expect from the increased accuracy in the estimates $\hat{\Theta}_i$.

5. Asymptotic properties

We describe in this section the asymptotic consistency and normality properties of META when the estimator used for the underlying aggregated processes x_t is a quasi-maximum likelihood estimator. We speak about *quasi*-maximum likelihood because we use the expression for the Gaussian likelihood, although we do not assume that the noise in the model (1) is Gaussian.

The derivation of the asymptotic properties is complicated by the fact that the noise processes $u_t^{(\mathbf{w})}$ of the various aggregations $x_t^{(\mathbf{w})} = \mathbf{w}(L)\mathbf{y}_t$ are neither mutually independent nor uncorrelated.

Figure 13: Forecasting accuracy comparison of several estimation methods: $\|y_{n+1} - \hat{y}_{n+1}\|^2$ for Model 2 with $n = 300$ and $n = 800$

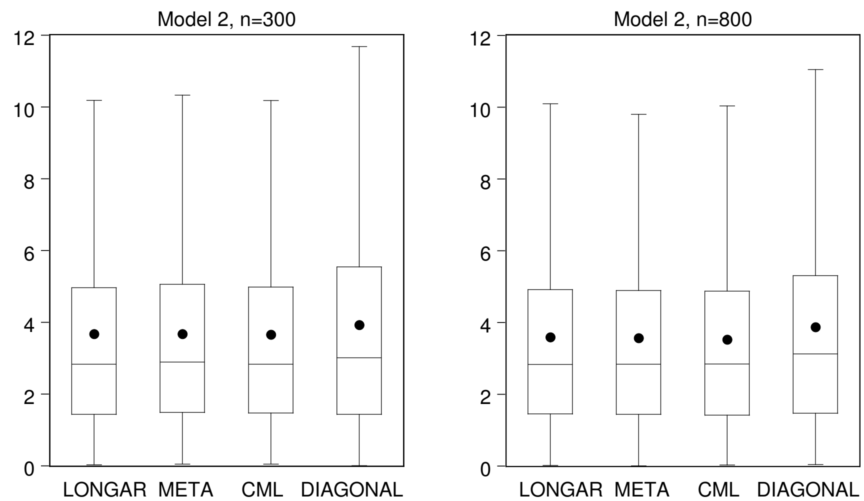


Figure 14: Forecasting accuracy comparison of several estimation methods: $\|y_{n+1} - \hat{y}_{n+1}\|^2$ for Model 3 with $n = 300$ and $n = 800$

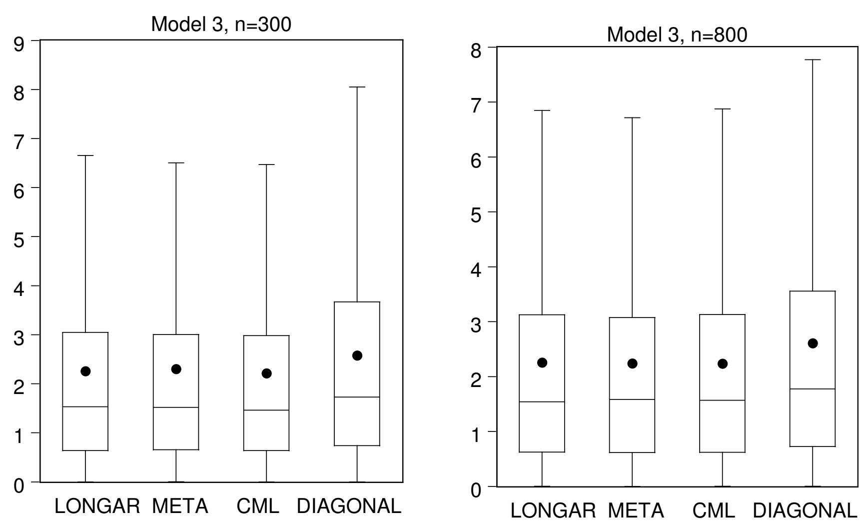
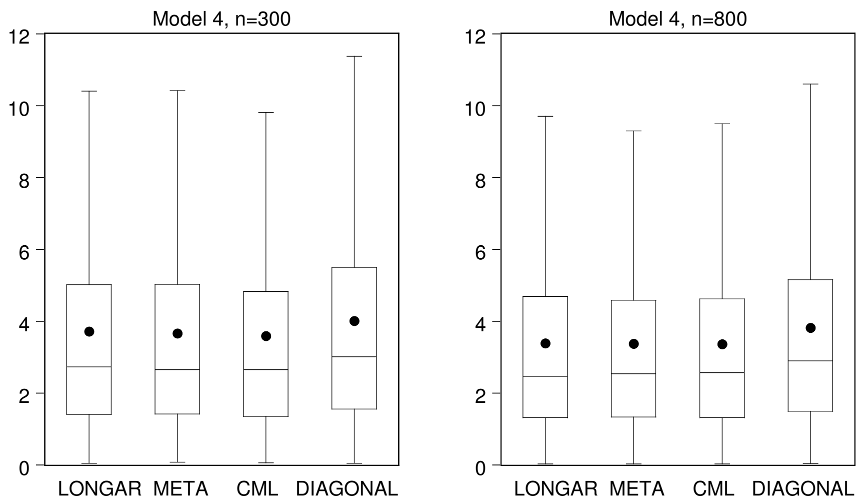


Figure 15: Forecasting accuracy comparison of several estimation methods: $\|y_{n+1} - \hat{y}_{n+1}\|^2$ for Model 4 with $n = 300$ and $n = 800$



Our proofs, presented in the Appendix, follow the ones of Poloni and Sbrana (2015b), extending their framework to this more general model.

Let

$$\boldsymbol{\alpha} = \begin{bmatrix} \operatorname{vech}(\Sigma_{\hat{\mathbf{v}}}) \\ \operatorname{vec}(\Theta_1) \\ \operatorname{vec}(\Theta_2) \\ \vdots \\ \operatorname{vec}(\Theta_q) \end{bmatrix} \in \mathbb{R}^{m \times 1}.$$

be the vector of parameters of the true VMA representation of $\mathbf{y}_t$, and $\hat{\boldsymbol{\alpha}}$ its estimate produced by Algorithm 3. Then, the following result holds.

Theorem 18. *Let $\mathbf{y}_t$ be an ergodic stationary linear process whose ACGF satisfies Assumptions P and Q. Suppose that the aggregation weights $\mathscr{W}$ are chosen so that the matrix X in (25) has full column rank. Then, $\hat{\boldsymbol{\alpha}} \xrightarrow{a.s.} \boldsymbol{\alpha}$ when $n \to \infty$. If, moreover, the fourth moments of $\mathbf{v}_t$ are finite, then $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{law} N(0, \Psi)$ for a suitable matrix Ψ.*

An expression for the matrix Ψ is given in terms of several quantities that are explicitly computable (although in practice rather unwieldy).

Theorem 19. *Under the hypotheses of Theorem 18,*

$$\Psi = A^{-1}(X^T W X)^{-1} X^T W J \mathscr{I}^{-1} \Xi \mathscr{I}^{-1} J^T W^T X (X^T W X)^{-1} (A^T)^{-1},$$

where the matrices $A, W, J, \mathscr{I}, \Xi$ are defined in the Appendix.

6. Conclusions

The estimation of a vector moving average (VMA) process represents a challenging task since the multivariate likelihood estimator is extremely slow to converge. In this paper we provide an alternative estimation method (META approach) based on two steps: we first compute several aggregations of the variables of the system and apply likelihood estimators to the resulting univariate processes; we then recover the VMA parameters using linear algebra tools. We show that the suggested estimator is consistent and asymptotically normal. In addition, some numerical experiments show the good performance of this estimator in small samples compared with standard methods.

The practical advantage of the suggested approach is that in this way we work with ML estimates of univariate processes only, therefore avoiding the complexity of maximizing the multivariate likelihood function directly. Another benefit is that the required univariate estimations can be performed in parallel for different values of the weight vectors, on a computer architecture that supports it. In contrast, estimation with the multivariate likelihood method is an intrinsically serial task, difficult to parallelize.

The suggested method not only is fast but it can also be implemented by standard statistical/econometric packages requiring only the estimation of univariate processes.

Some open issues need further investigation. Indeed, even if the estimator seems to work well in practice with random choices of the aggregation vectors $\mathbf{w}(L)$, a natural question is whether it is possible to find the optimal choice for these weight vectors. Moreover, it might be worth investigating empirical strategies to approximate the asymptotic covariance matrix using, for instance, bootstrap techniques (see Kotchoni (2014)). This might improve the performance of the suggested estimator. Moreover, a generalization from VMA(q) processes to VARMA(p, q) would give a more powerful and general estimator. We believe that this might be feasible, however, one should consider the complications arising with the contemporaneous aggregation of ARMA processes, as described in Remark 10.

7. Acknowledgments

We wish to thank Oliver Linton, for helping us improving the original version of this paper. We are grateful to one of the anonymous referees of Brüll et al. (2013), who has brought to our attention the idea of shrinkage estimation. F. Poloni acknowledges the support of INdAM/GNCS (Istituto Nazionale di Alta Matematica) and of the PRA projects of the University of Pisa.

Bart, H., Gohberg, I., Kaashoek, M. A., Ran, A. C. M., 2010. A state space approach to canonical factorization with applications. Vol. 200 of Operator Theory: Advances and Applications. Birkhäuser Verlag, Basel; Birkhäuser Verlag, Basel, linear

Operators and Linear Systems.
URL http://dx.doi.org/10.1007/978-3-7643-8753-2

Box, G. E., Jenkins, G. M., 1976. Time series analysis: forecasting and control, revised ed. Holden-Day.

Brockwell, P. J., Davis, R. A., 2009. Time series: theory and methods. Springer Science & Business Media.

Brüll, T., Poloni, F., Sbrana, G., Schröder, C., 2013. Enforcing solvability of a nonlinear matrix equation and estimation of multivariate ARMA time series. Tech. Rep. Matheon Preprint #1027, DFG-Forschungszentrum Matheon.
URL http://nbn-resolving.de/urn:nbn:de:0296-matheon-12409

Chu, E. K.-W., Huang, T.-M., Lin, W.-W., Wu, C.-T., 2010. Palindromic eigenvalue problems: a brief survey. Taiwanese J. Math. 14 (3A), 743–779.

Dufour, J.-M., Pelletier, D., 2005. Practical methods for modelling weak VARMA processes: identification, estimation and specification with a macroeconomic application. Preprint.

Dunsmuir, W., Hannan, E. J., 1976. Vector linear time series models. Advances in Appl. Probability 8 (2), 339–364.

Durbin, J., 1959. Efficient estimation of parameters in moving-average models. Biometrika, 306–316.

Durbin, J., 1960. The fitting of time-series models. Revue de l'Institut International de Statistique, 233–244.

Durbin, J., Koopman, S. J., 2012. Time series analysis by state space methods. No. 38. Oxford University Press.

Ephremidze, L., 2010. An elementary proof of the polynomial matrix spectral factorization theorem.

Fuller, W. A., 1996. Introduction to statistical time series, 2nd Edition. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, a Wiley-Interscience Publication.

Galbraith, J. W., Ullah, A., Zinde-Walsh, V., 2002. Estimation of the vector moving average model by vector autoregression. Econometric Reviews 21 (2), 205–219.

Gohberg, I., Kaashoek, M. A., Lancaster, P., 1988. General theory of regular matrix polynomials and band Toeplitz operators. Integral Equations Operator Theory 11 (6), 776–882.
URL `http://dx.doi.org/10.1007/BF01208810`

Gohberg, I., Lancaster, P., Rodman, L., 1982. Matrix polynomials. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, computer Science and Applied Mathematics.

Golub, G. H., Van Loan, C. F., 2013. Matrix computations, 4th Edition. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD.

Gourieroux, C., Monfort, A., Renault, E., 1993. Indirect inference. Journal of Applied Econometrics 8, S85–S118.
URL `http://www.jstor.org/stable/2285076`

Granger, C. W., Morris, M. J., 1976. Time series modelling and interpretation. Journal of the Royal Statistical Society. Series A (General), 246–257.

Hamilton, J. D., 1994. Time series analysis. Princeton University Press, Princeton, NJ.

Hannan, E. J., Deistler, M., 2012. The statistical theory of linear systems. Vol. 70 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, reprint of the 1988 original.
URL `http://dx.doi.org/10.1137/1.9781611972191.ch1`

Hannan, E. J., Kavalieris, L., 1984. A method for autoregressive-moving average estimation. Biometrika 71 (2), 273–280.

Hannan, E. J., Rissanen, J., 1982. Recursive estimation of mixed autoregressive-moving average order. Biometrika 69 (1), 81–94.

Harvey, A. C., 1990. Forecasting, structural time series models and the Kalman filter. Cambridge university press.

Hunt, K. J. (Ed.), 1993. Polynomial methods in optimal control and filtering. Vol. 49 of IEE Control Engineering Series. Peter Peregrinus, Ltd., London; on behalf of the Institution of Electrical Engineers (IEE), London.
URL `http://dx.doi.org/10.1049/PBCE049E`

Hyndman, R., Koehler, A. B., Ord, J. K., Snyder, R. D., 2008. Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media.

IHS Global, 2013. Eviews 8. `http://www.eviews.com/EViews8/ev8main.html`.

Isserlis, L., 1918. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. Biometrika 12, 134–139.

Kågström, B., 1993. A direct method for reordering eigenvalues in the generalized real Schur form of a regular matrix pair (A, B). In: Linear algebra for large scale and real-time applications (Leuven, 1992). Vol. 232 of NATO Adv. Sci. Inst. Ser. E Appl. Sci. Kluwer Acad. Publ., Dordrecht, pp. 195–218.

Kapetanios, G., 2003. A note on an iterative least-squares estimation method for ARMA and VARMA models. Economics Letters 79 (3), 305–312.

Kascha, C., 2012. A comparison of estimation methods for vector autoregressive moving-average models. Econometric Reviews 31 (3), 297–324.

Koreisha, S., Pukkila, T., 1990. A generalized least-squares approach for estimation of autoregressive moving-average models. Journal of Time Series Analysis 11 (2), 139–151.

Kotchoni, R., 2014. The indirect continuous-GMM estimation. Comput. Statist. Data Anal. 76, 464–488.
URL `https://doi.org/10.1016/j.csda.2013.09.023`

Ling, S., McAleer, M., 2010. A general asymptotic theory for time-series models. Statistica Neerlandica 64 (1), 97–111.

Lütkepohl, H., 2005. New introduction to multiple time series analysis. Springer-Verlag, Berlin.
URL `http://dx.doi.org/10.1007/978-3-540-27752-1`

Lütkepohl, H., 2012. Forecasting aggregated vector ARMA processes. Vol. 284. Springer Science & Business Media.

McLeod, A. I., 1999. Necessary and sufficient condition for nonsingular Fisher information matrix in ARMA and fractional ARIMA models. Amer. Statist. 53 (1), 71–72.
URL `http://dx.doi.org/10.2307/2685656`

Peligrad, M., Utev, S., 2006. Central limit theorem for stationary linear processes. The Annals of Probability 34 (4), 1608–1622.

Plagborg-Møller, M., et al., 2015. Bayesian inference on structural impulse response functions. Tech. rep.

Poloni, F., Sbrana, G., 2015a. Multivariate trend-cycle extraction with the Hodrick-Prescott filter. (forthcoming) Macroeconomic Dynamics.

Poloni, F., Sbrana, G., 2015b. A note on forecasting demand using the multivariate exponential smoothing framework. International Journal of Production Economics 162 (0), 143 – 150.

URL http://www.sciencedirect.com/science/article/pii/S0925527315000298

Ravenna, F., 2007. Vector autoregressions and reduced form representations of DSGE models. Journal of monetary economics 54 (7), 2048–2064.

Reinsel, G. C., 2003. Elements of multivariate time series analysis. Springer Science & Business Media.

Rozanov, Y. A., 1967. Stationary random processes, vi+211.

Sbrana, G., Silvestrini, A., Venditti, F., Jun. 2015. Short term inflation forecasting: the M.E.T.A. approach. Temi di discussione (Economic working papers) 1016, Bank of Italy, Economic Research and International Relations Area.
URL http://ideas.repec.org/p/bdi/wptemi/td_1016_15.html

Spliid, H., 1983. A fast estimation method for the vector autoregressive moving average model with exogenous variables. Journal of the American Statistical Association 78 (384), 843–849.

Stewart, G. W., Sun, J. G., 1990. Matrix perturbation theory. Computer Science and Scientific Computing. Academic Press Inc., Boston, MA.

Wilson, G. T., 1973. The estimation of parameters in multivariate time series models. Journal of the Royal Statistical Society. Series B (Methodological), 76–85.

Wolfram Research, 2007. Mathematica time series version 1.4. Package for Mathematica 7 and 8. https://www.wolfram.com/products/applications/timeseries/.

Appendix A. Proof of the lemmas in the first sections

Proof of Lemma 5. One can show (Stewart and Sun, 1990, Section VI.1.2) that if $LS_{11} - T_{11}$ and $LS_{22} - T_{22}$ have no common eigenvalues (as is the case here, since the former are outside the unit circle and the latter are inside), then there exist matrices Z_1 and Z_2 such that

$$\begin{bmatrix} I & Z_1 \\ 0 & I \end{bmatrix}\begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix}\begin{bmatrix} I & Z_2 \\ 0 & I \end{bmatrix} = \begin{bmatrix} S_{11} & 0 \\ 0 & S_{22} \end{bmatrix}, \quad \text{and}$$
$$\begin{bmatrix} I & Z_1 \\ 0 & I \end{bmatrix}\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}\begin{bmatrix} I & Z_2 \\ 0 & I \end{bmatrix} = \begin{bmatrix} T_{11} & 0 \\ 0 & T_{22} \end{bmatrix}.$$

Choosing

$$F_1 = \begin{bmatrix} S_{11}^{-1} & 0 \\ 0 & T_{22}^{-1} \end{bmatrix}\begin{bmatrix} I & Z_1 \\ 0 & I \end{bmatrix} Q, \quad \text{and } F_2 = Z\begin{bmatrix} I & Z_2 \\ 0 & I \end{bmatrix},$$

we get exactly

$$F_1(LE - A)F_2 = \begin{bmatrix} LI - S_{11}^{-1}T_{11} & 0 \\ 0 & LT_{22}^{-1}S_{22} - I \end{bmatrix}, \text{ and } XF_2 = \begin{bmatrix} X_1 & X_2 + X_1 Z_2 \end{bmatrix}.$$

□

Proof of Lemma 7. The first point is clear. For each $\lambda \in [0, 2\pi]$, the scalar quantity $\gamma^{(\mathbf{w})}(e^{i\lambda}) = \mathbf{w}(e^{i\lambda})\Gamma(e^{i\lambda})\mathbf{w}(e^{i\lambda})^\star$ is positive since the matrix $\Gamma(e^{i\lambda})$ is positive definite. The third point follows from the existence of the factorization (3). □

Appendix B. Explicit form of $X_\mathbf{w}$

The matrix $X_\mathbf{w}$ in (24) contains the coefficients of (19), when written down explicitly as a system of linear equations with unknowns the coefficients $\hat{\mathbf{z}}$ of the estimated ACGF, as in (20), and right-hand side $\gamma_\mathbf{w}$. For instance, in the system (23e)–(23g), the unknowns are

$$\begin{bmatrix} \hat{a} \\ \hat{b} \\ \vdots \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \operatorname{vech} \hat{\Gamma}_0 \\ \operatorname{vec} \hat{\Gamma}_1 \end{bmatrix},$$

and the matrix $X_\mathbf{w}$ is

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

One can obtain an explicit expression for this matrix in the general case. If we set $\mathbf{w}(L) = \mathbf{w}^{(0)} + \mathbf{w}^{(1)}L + \cdots + \mathbf{w}^{(r)}L^r$, then expanding (19) gives

$$\hat{\gamma}_\ell^{(\mathbf{w})} = \sum_{h_1 + k - h_2 = \ell} \mathbf{w}^{(h_1)}\hat{\Gamma}_k(\mathbf{w}^{(h_2)})^T,$$

hence the coefficient of the unknown $(\hat{\Gamma}_k)_{ij}$ (keeping into account that this unknown appears also in $\hat{\Gamma}_{-k} = \hat{\Gamma}_k^T$) is

$$\sum_{h_1-h_2=\ell-k} \mathbf{w}_i^{(h_1)}(\mathbf{w}_j^{(h_2)})^T + \sum_{h_1-h_2=\ell+k} \mathbf{w}_j^{(h_1)}(\mathbf{w}_i^{(h_2)})^T.$$

Putting each coefficient into its place inside the matrix, one obtains the following closed form for X.

$$X_{\mathbf{w}} = \begin{bmatrix} \mathbf{x}_{0,0}^T & \mathbf{x}_{0,1}^T & \cdots & \mathbf{x}_{0,q}^T \\ \mathbf{x}_{1,0}^T & \mathbf{x}_{1,1}^T & \cdots & \mathbf{x}_{1,q}^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{q+r,0}^T & \mathbf{x}_{q+r,1}^T & \cdots & \mathbf{x}_{q+r,q}^T \end{bmatrix},$$

where the vectors $\mathbf{x}_{\ell,k}$ are given by

$$\mathbf{x}_{\ell,k} = \sum_{h_1-h_2=\ell-k} \operatorname{vec}\left(\mathbf{w}^{(h_1)}(\mathbf{w}^{(h_2)})^T\right) + \sum_{h_1-h_2=\ell+k} \operatorname{vec}\left(\mathbf{w}^{(h_2)}(\mathbf{w}^{(h_1)})^T\right), \quad \text{for } k > 0,$$

and

$$\mathbf{x}_{\ell,0} = \sum_{h_1-h_2=\ell} \operatorname{vech}\left(\mathbf{w}^{(h_1)}(\mathbf{w}^{(h_2)})^T + \mathbf{w}^{(h_2)}(\mathbf{w}^{(h_1)})^T - \operatorname{diag}(\mathbf{w}^{(h_1)})\operatorname{diag}(\mathbf{w}^{(h_2)})\right).$$

The last term comes from the fact that the diagonal elements with $i = j$ are summed twice instead of once in the previous sum of two half-vectorizations.

Appendix C. Asymptotic theory

We start this section with the asymptotic theory of the QML estimator of an aggregate univariate MA process, which we use in our procedure. Note that the univariate reparametrized noise $u_t^{(\mathbf{w})}$ is uncorrelated, but is not in general independent. Hence the standard textbook results which assume independent noise do not hold in our case, and we have to adapt some of the proofs.

Appendix C.1. (Quasi-)maximum likelihood of univariate processes

In this Section Appendix C.1, we refer to a single aggregate MA process $x^{(\mathbf{w})}$, and drop the sub- or superscript $\mathbf{w}$ for ease of notation.

The expression for the Gaussian negative log-likelihood function of a univariate MA$(q+r)$ process $x_t = \theta(L)u_t$ is classical, see e.g. Box and Jenkins (1976):

$$\mathscr{L}(\tilde{\beta}) = \sum_{t=1}^{n} \ell_t(\tilde{\beta}), \quad \text{with } \ell_t(\tilde{\beta}) = \frac{1}{2}\log\tilde{\omega} + \frac{\tilde{u}_t^2}{2\tilde{\omega}}, \quad \tilde{u}_t = \tilde{\theta}(L)^{-1}x_t, \quad \tilde{\beta} = \begin{bmatrix} \tilde{\omega} \\ \tilde{\theta}_1 \\ \tilde{\theta}_2 \\ \vdots \\ \tilde{\theta}_{q+r} \end{bmatrix}.$$

The maximum likelihood estimator is given by

$$\hat{\beta} = \arg\min \mathscr{L}(\tilde{\beta}). \tag{C.1}$$

We assume here that the parameter space for the $\tilde{\theta}_i$ is restricted to a set of invertible processes for which the roots of the associated polynomial $\tilde{\theta}(L)$ satisfy $|\lambda| < \delta$ for a fixed $\delta < 1$, so that the power series $\tilde{\theta}(L)^{-1} = \sum \tilde{\psi}_i L^i$ has coefficients that decay as $\tilde{\psi}_i = O(\delta^i)$ for each $\tilde{\theta}$ in our admissible parameter space. Similarly, we assume that $\tilde{\omega}$ is bounded away from 0 and ∞.

The partial derivatives of the asymptotic likelihood function $\ell_t(\tilde{\beta})$ with respect to the parameters $\tilde{\theta}_i$ and $\tilde{\omega}$, evaluated in β, are

$$\xi_t^{(0)} = \left.\frac{\partial \ell_t(\tilde{\beta})}{\partial \tilde{\omega}}\right|_{\beta} = \frac{\omega - u_t^2}{2\omega^2}, \quad \text{and} \tag{C.2a}$$

$$\xi_t^{(h)} = \left.\frac{\partial \ell_t(\tilde{\beta})}{\partial \tilde{\theta}_h}\right|_{\beta} = \frac{1}{\omega}u_t^{(h)}u_t, \quad \text{with } u_t^{(h)} = -\frac{L^h}{\theta(L)^2}x_t = -\frac{L^h}{\theta(L)}u_t,\ h = 1,2,\ldots,q+r. \tag{C.2b}$$

Since the noise process u_t is not independent, but only uncorrelated, it may not be obvious that the true parameters β maximize the asymptotic likelihood; we prove it in the next lemma.

Lemma 20. *Let* $\mathbf{v}$ *be an i.i.d. process with finite variance,* $x = \mathbf{w}(L)\Theta(L)\mathbf{w}$*, and* $\theta(L), \omega$ *as in* (15) *and* (16)*. Then,* $\beta = \arg\min \mathbb{E}\left[\ell_t(\tilde{\beta})\right]$.

Proof. By taking a derivative in $\tilde{\omega}$, it is easy to show that the minimum of

$$\mathbb{E}\left[\ell_t(\tilde{\beta})\right] = \frac{1}{2}\log\tilde{\omega} + \frac{\mathbb{E}\left[\tilde{u}_t^2\right]}{2\tilde{\omega}}$$

occurs when $\tilde{\omega} = \mathbb{E}\left[\tilde{u}_t^2\right]$. In that case the function reduces to $\frac{1}{2}\left(\log \mathbb{E}\left[\tilde{u}_t^2\right] + 1\right)$, which is increasing in $\mathbb{E}\left[\tilde{u}_t^2\right]$. So we only need to prove that the minimum of $\mathbb{E}\left[\tilde{u}_t^2\right]$ is achieved by u_t.

The variance of $\tilde{u} = \frac{1}{\tilde{\theta}(L)}\mathbf{w}(L)G(L)\mathbf{v}$ is the constant term in its ACGF

$$\frac{1}{\tilde{\theta}(L)}\mathbf{w}(L)G(L)\Sigma_{\mathbf{v}}G(L)^\star\mathbf{w}(L)^\star\frac{1}{\tilde{\theta}(L^{-1})} = \frac{1}{\tilde{\theta}(L)}\theta(L)\omega\theta(L^{-1})\frac{1}{\tilde{\theta}(L^{-1})} = \omega a(L)a(L^{-1}),$$

where $a(L)$ is the power series

$$a(L) = \tilde{\theta}(L)^{-1}\theta(L) = 1 + a_1 L + a_2 L^2 + \ldots.$$

The constant term of $a(L)a(L^{-1})$ is $1 + a_1^2 + a_2^2 + \cdots \geq 1$, and equality holds if and only if $a(L) = 1$, i.e., $\tilde{\theta}(L) = \theta(L)$. □

Moreover, in the following we shall need the fact that the Fisher information matrix $\mathscr{I} = \mathbb{E}\left[\nabla^2 \ell_t(\beta)\right]$ is nonsingular. This is proved, for instance, in McLeod (1999).

Note that $\mathscr{I} = \Omega_{\mathbf{w}}^{-1}$, with $\Omega_{\mathbf{w}}$ as in Section 3.2, as proved for instance in (Box and Jenkins, 1976, Section 7.1).

Appendix C.2. Consistency and normality of the aggregate estimates

We now turn to prove that the quasi-maximum likelihood estimator $\hat{\beta}$ of

$$\beta = \begin{bmatrix} \beta_{\mathbf{w}_1} \\ \beta_{\mathbf{w}_2} \\ \vdots \\ \beta_{\mathbf{w}_k} \end{bmatrix},$$

where the $\beta_{\mathbf{w}_i}$ are defined as in (16), is consistent and jointly normal. For the former property, it is enough to prove that each of them is consistent when considered alone.

Theorem 21. *Let $\mathbf{y} = G(L)\mathbf{v}$ be a stationary ergodic linear model whose ACGF satisfies Assumptions P and Q. Then, the quasi-maximum likelihood estimator $\hat{\beta}_{\mathbf{w}}$ in* (C.1) *for the process $x^{(\mathbf{w})} = \mathbf{w}(L)G(L)\mathbf{v}$ is consistent, i.e., $\hat{\beta}_{\mathbf{w}} \xrightarrow{a.s.} \beta_{\mathbf{w}}$ when $n \to \infty$.*

Proof. We rely on (Ling and McAleer, 2010, Theorem 1). We have proved that the likelihood has a maximum in the exact values in Lemma 20, and due to our choice of the parameter space its expectation is bounded. Since the coefficients of $\frac{1}{\tilde{\theta}(L)}\mathbf{w}(L)G(L)$ are bounded uniformly by $O(\delta^i)$, Assumption 2(i) in Ling and McAleer (2010) is satisfied as well, hence asymptotic consistency hold. □

Establishing joint normality is more involved: since the $x_t^{(\mathbf{w})}$ are neither independent nor uncorrelated from each other, we cannot rely on the classical central limit results. We use instead a central limit result for weakly dependent sequences from Peligrad and Utev (2006), which we summarize and report as follows.

Theorem 22. *For an i.i.d. sequence of random variables $(\mathbf{v}_t)_{t=\dots,-1,0,1,\dots}$, denote by $\mathscr{F}_a^b$ the σ-field generated by $\mathbf{v}_t$ with $a \le t \le b$ and define $\xi_t = f(\mathbf{v}_t, \mathbf{v}_{t-1}, \dots)$, $t \in \mathbb{Z}$. Assume that $\mathbb{E}[\xi_0] = 0$, $\mathbb{E}\left[\xi_0^2\right] = \omega < \infty$, and*

$$\sum_{t=1}^{\infty} \frac{1}{\sqrt{t}} \|\xi_0 - \mathbb{E}\left[\xi_0 \mid \mathscr{F}_{-t}^0\right]\|_{\mathbb{L}_2} < \infty, \tag{C.3}$$

where $\|X\|_{\mathbb{L}_2} := \mathbb{E}\left[X^2\right]^{1/2}$. Then,

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \xi_t \xrightarrow{law} N(0, \omega). \tag{C.4}$$

Proof. Peligrad and Utev (2006) contains a stronger result on triangular sequences (Corollary 5); the statement (C.4) is a special case that can be obtained by setting

$$a_i = \begin{cases} 1 & i = 0 \\ 0 & \text{otherwise} \end{cases}$$

in the thesis of their Theorem 1, so that $b_n = \sqrt{n}$. □

We start from a lemma.

Lemma 23. *Let $(\mathbf{v}_t)_{t=\dots,-1,0,1,\dots}$ be a sequence of vector-valued i.i.d. random variables ($\mathbf{v}_i \in \mathbb{R}^d$ for each i), and $C_i, D_i \in \mathbb{R}^{1\times d}$, for $i = 1, 2, \dots$, be such that $\|C_i\| = O(\delta^i)$, $\|D_i\| = O(\delta^i)$ for some $\delta < 1$. Define the univariate linear processes $c = C(L)\mathbf{v}$ and $d = D(L)\mathbf{v}$ as*

$$C(L) = \sum_{i=0}^{\infty} C_i L^i, \quad D(L) = \sum_{i=0}^{\infty} D_i L^i.$$

Let $M = \mathbb{E}[c_t d_t]$ and $\xi_t = c_t d_t - M$. Then,

$$\|\xi_0 - \mathbb{E}\left[\xi_0 \mid \mathscr{F}_{-t}^0\right]\|_{\mathbb{L}_2} = O(\delta^t) \quad \textit{when } t \to \infty. \tag{C.5}$$

Proof. We decompose c_0 and d_0 into

$$c_0 = \underbrace{\sum_{i=0}^{t} C_i \mathbf{v}_{-i}}_{:=p_t} + \underbrace{\sum_{i>t} C_i \mathbf{v}_{-i}}_{:=q_t}, \quad \text{and } d_0 = \underbrace{\sum_{i=0}^{t} D_i \mathbf{v}_{-i}}_{:=r_t} + \underbrace{\sum_{i>t} D_i \mathbf{v}_{-i}}_{:=s_t}, \tag{C.6}$$

where p_t and r_t are functions in the σ-field $\mathscr{F}_{-t}^0$ and q_t and s_t are independent from it. One has

$$\begin{aligned}
\mathbb{E}\left[\xi_0 \mid \mathscr{F}_{-t}^0\right] &= \mathbb{E}\left[c_0 d_0 - M \mid \mathscr{F}_{-t}^0\right] = \mathbb{E}\left[(p_t + q_t)(r_t + s_t) \mid \mathscr{F}_{-t}^0\right] - M \\
&= p_t r_t + \underbrace{\mathbb{E}\left[q_t \mid \mathscr{F}_{-t}^0\right]}_{=0} r_t + p_t \underbrace{\mathbb{E}\left[s_t \mid \mathscr{F}_{-t}^0\right]}_{=0} + \mathbb{E}\left[q_t s_t \mid \mathscr{F}_{-t}^0\right] - M \\
&= p_t r_t + \mathbb{E}\left[q_t s_t \mid \mathscr{F}_{-t}^0\right] - M,
\end{aligned}$$

thus

$$\begin{aligned}
\left\|\xi_0 - \mathbb{E}\left[\xi_0 \mid \mathscr{F}_{-t}^0\right]\right\|_{\mathbb{L}_2} &= \|q_t r_t + p_t s_t + q_t s_t - \mathbb{E}\left[q_t s_t \mid \mathscr{F}_{-t}^0\right]\|_{\mathbb{L}_2} \\
&\leq \|q_t\|_{\mathbb{L}_2} \|r_t\|_{\mathbb{L}_2} + \|p_t\|_{\mathbb{L}_2} \|s_t\|_{\mathbb{L}_2} + 2\|q_t\|_{\mathbb{L}_2} \|s_t\|_{\mathbb{L}_2}.
\end{aligned} \tag{C.7}$$

Take a constant $K > 0$ such that $\|C_i\| < K\delta^i$ and $\|D_i\| < K\delta^i$; since the $\mathbf{v}_t$ are independent, we have

$$\begin{aligned}
\|p_t\|_{\mathbb{L}_2}^2 &\leq \sum_{i=0}^{t} \|C_i\|^2 \|\mathbf{v}_i\|_{\mathbb{L}_2}^2 \leq \frac{K^2 \|\Sigma_{\mathbf{v}}\|}{1 - \delta^2} = O(1), \quad \text{and} \\
\|q_t\|_{\mathbb{L}_2}^2 &\leq \sum_{i>t} \|C_i\|^2 \|\mathbf{v}_i\|_{\mathbb{L}_2}^2 \leq \frac{K^2 \|\Sigma_{\mathbf{v}}\| \delta^{2t+2}}{1 - \delta^2} = O(\delta^{2t}),
\end{aligned}$$

and analogously $\|r_t\|_{\mathbb{L}_2} = O(1)$, $\|s_t\|_{\mathbb{L}_2} = O(\delta^t)$. Plugging these estimates into (C.7), we get the required bound. □

Next, we consider the vector

$$\boldsymbol{\xi}_t = \begin{bmatrix} \nabla_{\mathbf{w}_1} \ell_t^{(\mathbf{w}_1)}(\beta_{\mathbf{w}_1}) \\ \vdots \\ \nabla_{\mathbf{w}_k} \ell_t^{(\mathbf{w}_k)}(\beta_{\mathbf{w}_k}) \end{bmatrix},$$

(where $\nabla_{\mathbf{w}}$ denotes the gradient with respect to the parameters $\tilde{\beta}_{\mathbf{w}}$), and prove a central limit result for it.

Lemma 24. *Let $\mathbf{y} = G(L)\mathbf{v}$ be a stationary ergodic linear model whose ACGF satisfies Assumptions P and Q, and $\mathscr{W}$ be a finite set of aggregation weights $\mathbf{w}(L) \in \mathbb{R}[L]^{1\times d}$. If $\mathbf{v}$ has finite fourth moments, then*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{\xi}_t \xrightarrow{law} N(0, \Xi) \quad \textit{for } n \to \infty,$$

where $\Xi = \mathbb{E}\left[\boldsymbol{\xi}_t \boldsymbol{\xi}_t^T\right]$.

Proof. By the Cramer-Wold device (Brockwell and Davis, 2009, Proposition 6.3.1), it is sufficient to prove that the (scalar) central limit theorem holds for a generic linear combination of its entries

$$\xi_t = \sum_i a_i (\xi^{(i)})_t,$$

where $a_i \in \mathbb{R}$ and $(\xi^{(i)})_t$ is in the form (C.2a) or (C.2b) for the aggregated process associated to some $\mathbf{w} \in \mathscr{W}$. Each of these forms for $\xi^{(i)}$ satisfies the hypotheses of Lemma 23: indeed, in the case (C.2a), take

$$c_t = d_t = \frac{u_t}{\sqrt{2}\omega}, \quad \text{and } M = \frac{\mathbb{E}\left[u_t^2\right]}{2\omega^2} = \frac{1}{2\omega},$$

and in the case (C.2b), take

$$c_t = u_t^{(i)}, \quad d_t = \frac{1}{\omega} u_t, \quad \text{and } M = 0.$$

Thus (C.5) holds with ξ_0 replaced by $\xi_0^{(i)}$. Using linearity of the expectation and the triangle inequality, one can obtain

$$\|\xi_0 - \mathbb{E}\left[\xi_0 \mid \mathscr{F}_{-t}^0\right]\|_{\mathbb{L}_2} = O(\delta^t),$$

where δ is the maximum of the decay rates of the processes $\xi^{(i)}$. Hence Condition (C.3) holds. Moreover, $\mathbb{E}[\xi_0] = 0$ and $\mathbb{E}\left[\xi_0^2\right] < \infty$ (this follows from the fact that $\mathbf{v}$ has finite fourth moments). Hence, by Theorem 22, the CLT holds. □

We are now ready to state and prove the main normality result.

Theorem 25. *Let* $\mathbf{y} = G(L)\mathbf{v}$ *be a stationary ergodic linear model whose ACGF satisfies Assumptions P and Q, and* $\mathscr{W}$ *be a finite set of aggregation weights* $\mathbf{w}(L) \in \mathbb{R}[L]^{1\times d}$*. If* $\mathbf{v}$ *has finite fourth moments, then*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{law} N(0, \mathscr{I}^{-1}\Xi\mathscr{I}^{-1}),$$

where

$$\mathscr{I} = \operatorname{diag}(\mathscr{I}_{\mathbf{w}_1}, \mathscr{I}_{\mathbf{w}_2}, \ldots, \mathscr{I}_{\mathbf{w}_k}),\ \mathscr{I}_{\mathbf{w}_i} = \mathbb{E}\left[\nabla^2_{\mathbf{w}_i} \ell_t^{(\mathbf{w}_i)}(\beta_{\mathbf{w}_i})\right],\ i = 1, 2, \ldots, k. \tag{C.8}$$

Proof. The first-order optimality conditions for the ML estimates state that $0 = \frac{1}{n}\sum_{t=1}^{n} \nabla_{\mathbf{w}} \ell_t(\hat{\theta}^{(\mathbf{w})}(L), \hat{\omega}^{(\mathbf{w})})$. Using a multivariate Taylor expansion around $\beta_{\mathbf{w}}$, we get

$$0 = \frac{1}{n}\sum_{t=1}^{n} \nabla_{\mathbf{w}} \ell_t(\beta_{\mathbf{w}}) + \left(\frac{1}{n}\sum_{t=1}^{n} \nabla^2_{\mathbf{w}} \ell_t(\tilde{\beta}_{\mathbf{w}})\right)(\hat{\beta}_{\mathbf{w}} - \beta_{\mathbf{w}}), \tag{C.9}$$

for a suitable vector $\tilde{\beta}_{\mathbf{w}}$ on the segment that joins $\hat{\beta}_{\mathbf{w}}$ and $\beta_{\mathbf{w}}$. If $\hat{\beta}_{\mathbf{w}}$ and $\beta_{\mathbf{w}}$ are close enough (which happens almost surely for large enough n, thanks to Theorem 21), then by continuity the Hessian matrix is invertible, thus we can rewrite (C.9) as

$$\hat{\beta}_{\mathbf{w}} - \beta_{\mathbf{w}} = -\left(\frac{1}{n}\sum_{t=1}^{n} \nabla^2_{\mathbf{w}} \ell_t(\tilde{\beta}_{\mathbf{w}})\right)^{-1} \left(\frac{1}{n}\sum_{t=1}^{n} \nabla_{\mathbf{w}} \ell_t(\beta_{\mathbf{w}})\right). \tag{C.10}$$

This expansion (C.10) is valid for every $\mathbf{w} \in \mathscr{W}$.

Stacking these expansions one above the other and multiplying by $\sqrt{n}$ we get

$$\sqrt{n}(\hat{\beta} - \beta) = -M(\tilde{\beta})^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{\xi}_t, \tag{C.11}$$

with $M(\tilde{\beta})$ the block diagonal matrix containing $\frac{1}{n}\sum \nabla^2_{\mathbf{w}_i} \ell_t^{(\mathbf{w}_i)}(\tilde{\beta}_{\mathbf{w}_i})$ in its diagonal blocks. By consistency, each of these blocks converges almost surely to its asymptotic value $\mathscr{I}_{\mathbf{w}_i}$; hence $M(\tilde{\beta}) \xrightarrow{\text{a.s.}} \mathscr{I}$. By Lemma (24), $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \boldsymbol{\xi}_t \xrightarrow{\text{law}} N(0, \Xi)$. Thus the thesis follows by Slutsky's theorem. □

We are now ready to give a proof of Theorem 18.

Proof of Theorem 18. Independently of the technique used to solve the linear systems (exact solution or least squares), the estimator that we have described is an analytical function of $\hat{\beta}$. This is made explicit, for instance, by the integral representation in (Gohberg et al., 1988, Theorem III.3.2) for the right divisor $\Theta(L)^\star$ of $\Gamma(L)$. Since $\hat{\beta}$ is consistent and asymptotically normal by Theorems 21 and 25, consistency and normality of our estimator follows from the Delta method. □

Appendix C.3. Computation of Ξ

In principle, each entry of the matrix Ξ can be computed, in terms of the coefficients of $G(L)$, the aggregation vectors $\mathscr{W}$, and the moments of the noise vector $\mathbf{v}_t$.

Indeed, each entry of Ξ is in the form $\mathbb{E}\left[\xi_t^{(h)}\hat{\xi}_t^{(\ell)}\right]$, with $\xi_t^{(h)}$ and $\hat{\xi}_t^{(\ell)}$ two of the quantities defined in (C.2a) or (C.2b), where we take two (possibly distinct) choices of the aggregation vector $\mathbf{w}(L)$ and $\hat{\mathbf{w}}(L)$, and denote with a hat all quantities computed starting from $\hat{\mathbf{w}}(L)$ instead of $\mathbf{w}(L)$. For each choice of the aggregation vector $\mathbf{w}(L)$ we have

$$u_t = \theta(L)^{-1}\mathbf{w}(L)G(L)\mathbf{v}_t = C_0\mathbf{v}_t + C_1\mathbf{v}_{t-1} + C_2\mathbf{v}_{t-2} + \dots,$$

where $C_i \in \mathbb{R}^{1\times d}$ are the coefficients of the power series expansion in L of the rational function $\theta(L)^{-1}\mathbf{w}(L)G(L)$, and a similar expansion holds for $u_t^{(h)}$ in (C.2b). Hence there are four sets of power series expansions coefficients $C_i, D_j, E_k, F_m \in \mathbb{R}^{1\times d}$, all explicitly computable, such that

$$\begin{aligned}
\mathbb{E}\left[\xi_t^{(h)}\hat{\xi}_t^{(\ell)}\right] &= \frac{1}{\omega\hat{\omega}}\mathbb{E}\left[u_t^{(h)}u_t\hat{u}_t^{(\ell)}\hat{u}_t\right] \\
&= \frac{1}{\omega\hat{\omega}}\mathbb{E}\left[\left(\sum_{i=0}^{\infty}C_i\mathbf{v}_{t-i}\right)\left(\sum_{j=0}^{\infty}D_j\mathbf{v}_{t-j}\right)\left(\sum_{k=0}^{\infty}E_k\mathbf{v}_{t-k}\right)\left(\sum_{m=0}^{\infty}F_m\mathbf{v}_{t-m}\right)\right] \\
&= \frac{1}{\omega\hat{\omega}}\sum_{i,j,k,m=0}^{\infty}\mathbb{E}\left[(C_i\mathbf{v}_{t-i})(D_j\mathbf{v}_{t-j})(E_k\mathbf{v}_{t-k})(F_m\mathbf{v}_{t-m})\right] \\
&= \frac{1}{\omega\hat{\omega}}\sum_{i,j,k,m=0}^{\infty}(C_i\otimes D_j\otimes E_k\otimes F_m)\mathbb{E}\left[\mathbf{v}_{t-i}\otimes\mathbf{v}_{t-j}\otimes\mathbf{v}_{t-k}\otimes\mathbf{v}_{t-m}\right]
\end{aligned}$$

(if $h, \ell > 0$, or a slightly more complicated expression involving also second moments $\mathbb{E}\left[\mathbf{v}_{t-i}\otimes\mathbf{v}_{t-j}^T\right]$ if one of them is zero, because of the constant term in (C.2a)).

Since $\mathbf{v}_t$ is assumed to be an independent noise, $\mathbb{E}\left[\mathbf{v}_{t-i} \otimes \mathbf{v}_{t-j} \otimes \mathbf{v}_{t-k} \otimes \mathbf{v}_{t-m}\right]$ is nonzero only when $i = j = k = m$, $i = j \neq k = m$, $i = k \neq j = m$ or $i = m \neq j = k$, but in general we need all the second and fourth moments of $\mathbf{v}_t$ in the computation.

In the case where the noise $\mathbf{v}_t$ is Gaussian, we can find a more explicit expression. We need the following two results. The first is a special case of Isserlis' theorem (Isserlis (1918)).

Lemma 26. *Let X_1, X_2, X_3, X_4 be four zero-mean scalar Gaussian random variables (not necessarily uncorrelated, and possibly coinciding). Then,*

$$\mathbb{E}\left[X_1 X_2 X_3 X_4\right] = \mathbb{E}\left[X_1 X_2\right]\mathbb{E}\left[X_3 X_4\right] + \mathbb{E}\left[X_1 X_3\right]\mathbb{E}\left[X_2 X_4\right] + \mathbb{E}\left[X_1 X_4\right]\mathbb{E}\left[X_2 X_3\right].$$

The second is a generalization of Lemma 1.

Lemma 27. *Let $\mathbf{y}_t = C(L)\mathbf{v}_t$ and $\mathbf{z}_t = D(L)\mathbf{v}_t$ be two stationary linear models constructed from the same i.i.d. process $\mathbf{v}_t$ (with $\mathbb{E}\left[\mathbf{v}_s \mathbf{v}_t^T\right] = \Sigma_\mathbf{v} \delta_{st}$). Then, $\mathbb{E}\left[\mathbf{y}_t \mathbf{z}_t^T\right]$ is equal to the coefficient of e^0 in the Fourier series of $C(\mathrm{e}^{i\lambda})\Sigma_v D(\mathrm{e}^{-i\lambda})^T$*

Proof. Expand $C(L)$ and $D(L)$ in a power series convergent on the unit disc, i.e., $C(L)\mathbf{v}_t = C_0\mathbf{v}_t + C_1\mathbf{v}_{t-1} + C_2\mathbf{v}_{t-2} + \dots$ and $D(L)\mathbf{v}_t = D_0\mathbf{v}_t + D_1\mathbf{v}_{t-1} + D_2\mathbf{v}_{t-2} + \dots$. Since $\mathbb{E}\left[\mathbf{v}_t \mathbf{v}_s^T = 0\right]$ whenever $s \neq t$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbf{y}_t \mathbf{z}_t^T\right] &= \mathbb{E}\left[(C_0\mathbf{v}_t + C_1\mathbf{v}_{t-1} + C_2\mathbf{v}_{t-2} + \dots)(D_0\mathbf{v}_t + D_1\mathbf{v}_{t-1} + D_2\mathbf{v}_{t-2} + \dots)^T\right] \\
&= \mathbb{E}\left[C_0\mathbf{v}_t\mathbf{v}_t^T D_0^T + C_1\mathbf{v}_{t-1}\mathbf{v}_{t-1}^T D_1^T + C_2\mathbf{v}_{t-2}\mathbf{v}_{t-2}^T D_2^T + \dots\right] \\
&= C_0\Sigma_\mathbf{v} D_0^T + C_1\Sigma_\mathbf{v} D_1^T + C_2\Sigma_\mathbf{v} D_2^T + \dots,
\end{aligned}$$

which is the coefficient of L^0 in the (unique) Laurent series expansion of $C(L)\Sigma_\mathbf{v} D(L)^\star$ that is convergent on the unit circle. This power series expansion is the Fourier series of $C(\mathrm{e}^{i\lambda})\Sigma_v D(\mathrm{e}^{-i\lambda})^T$. □

Recall that the aggregate processes u_t are constructed as

$$u_t = \frac{1}{\theta(L)} x_t = \frac{1}{\theta(L)} \mathbf{w}(L)\mathbf{y}_t = \frac{1}{\theta(L)} \mathbf{w}(L) G(L)\mathbf{v}_t, \tag{C.12}$$

and their derivatives $u^{(h)}$ needed in (C.2b) are

$$u_t^{(h)} = \frac{-L^h}{(\theta(L))^2} x_t = \frac{-L^h}{(\theta(L))^2} \mathbf{w}(L) G(L)\mathbf{v}_t. \tag{C.13}$$

With (C.12), (C.13) and Lemma 27, we can compute $\mathbb{E}[u_t \hat{u}_t]$ as the coefficient of e^0 in the Fourier series of

$$\frac{1}{\theta(L)}\mathbf{w}(L)G(L)\Sigma_{\mathbf{v}}G(L)^\star\hat{\mathbf{w}}(L)^\star\frac{1}{\hat{\theta}(L^{-1})} = \frac{1}{\theta(L)}\mathbf{w}(L)\Gamma(L)\hat{\mathbf{w}}(L)^\star\frac{1}{\hat{\theta}(L^{-1})},$$

and analogously $\mathbb{E}\left[u_t^{(h)}\hat{u}_t^{(l)}\right]$ for each h, l as the coefficient of e^0 in the Fourier series of

$$\frac{L^h}{(\theta(L))^2}\mathbf{w}(L)\Gamma(L)\hat{\mathbf{w}}(L)^\star\frac{L^{-l}}{(\hat{\theta}(L^{-1}))^2}.$$

826 A similar formula holds for $\mathbb{E}\left[u_t\hat{u}_t^{(l)}\right]$ for each l.

With these expressions and Lemma 26, one can compute the entries of Ξ. We have for the case $h, l > 0$

$$\begin{aligned}\mathbb{E}\left[\xi_t^{(h)}\hat{\xi}_t^{(l)}\right] &= \mathbb{E}\left[\frac{1}{\omega}u_t^{(h)}u_t\frac{1}{\hat{\omega}}\hat{u}_t^{(l)}\hat{u}_t\right]\\ &= \frac{1}{\omega\hat{\omega}}\left(\underbrace{\mathbb{E}\left[u_t^{(h)}u_t\right]}_{=0}\underbrace{\mathbb{E}\left[\hat{u}_t^{(l)}\hat{u}_t\right]}_{=0}+\mathbb{E}\left[u_t^{(h)}\hat{u}_t^{(l)}\right]\mathbb{E}[u_t\hat{u}_t]+\mathbb{E}\left[u_t^{(h)}\hat{u}_t\right]\mathbb{E}\left[u_t\hat{u}_t^{(l)}\right]\right)\end{aligned}$$

and for the special cases when one or both indices h, l are zero

$$\begin{aligned}\mathbb{E}\left[\xi_t^{(0)}\hat{\xi}_t^{(l)}\right] &= \mathbb{E}\left[\frac{\omega-u_t^2}{2\omega^2}\frac{1}{\hat{\omega}}\hat{u}_t^{(l)}\hat{u}_t\right] = \frac{1}{2\omega^2\hat{\omega}}\left(\omega\underbrace{\mathbb{E}\left[\hat{u}_t^{(l)}\hat{u}_t\right]}_{=0}-\mathbb{E}\left[u_t^2\hat{u}_t^{(l)}\hat{u}_t\right]\right)\\ &= \frac{-1}{2\omega^2\hat{\omega}}\left(\mathbb{E}\left[u_t^2\right]\underbrace{\mathbb{E}\left[\hat{u}_t^{(l)}\hat{u}_t\right]}_{=0}+\mathbb{E}\left[u_t\hat{u}_t^{(l)}\right]\mathbb{E}[u_t\hat{u}_t]+\mathbb{E}[u_t\hat{u}_t]\mathbb{E}\left[u_t\hat{u}_t^{(l)}\right]\right)\\ &= \frac{-1}{\omega^2\hat{\omega}}\mathbb{E}\left[u_t\hat{u}_t^{(l)}\right]\mathbb{E}[u_t\hat{u}_t],\\ \mathbb{E}\left[\xi_t^{(0)}\hat{\xi}_t^{(0)}\right] &= \mathbb{E}\left[\frac{\omega-u_t^2}{2\omega^2}\frac{\hat{\omega}-\hat{u}_t^2}{2\hat{\omega}^2}\right]\\ &= \frac{1}{4\omega^2\hat{\omega}^2}\left(\omega\hat{\omega}-\omega\underbrace{\mathbb{E}\left[\hat{u}_t^2\right]}_{=\hat{\omega}}-\underbrace{\mathbb{E}\left[u_t^2\right]}_{=\omega}\hat{\omega}+\mathbb{E}\left[u_t^2\hat{u}_t^2\right]\right)\\ &= \frac{1}{4\omega^2\hat{\omega}^2}\left(-\omega\hat{\omega}+\underbrace{\mathbb{E}\left[u_t^2\right]}_{=\omega}\underbrace{\mathbb{E}\left[\hat{u}_t^2\right]}_{=\hat{\omega}}+\mathbb{E}[u_t\hat{u}_t]\mathbb{E}[u_t\hat{u}_t]+\mathbb{E}[u_t\hat{u}_t]\mathbb{E}[u_t\hat{u}_t]\right)\\ &= \frac{1}{2\omega^2\hat{\omega}^2}\mathbb{E}[u_t\hat{u}_t]\mathbb{E}[u_t\hat{u}_t].\end{aligned}$$

Appendix C.4. Computation of A

In this section, we give an explicit expression for the matrix $A = \frac{\partial \mathbf{z}}{\partial \boldsymbol{\alpha}}$, which is needed in the computation of the asymptotic covariance of the estimator.

The vector $\mathbf{z}$ can be written as a simple function of $\boldsymbol{\alpha}$, by expanding the relation $\Gamma(L) = \Theta(L)\Sigma_{\mathbf{v}}\Sigma(L)^{\star}$; in particular, also its Jacobian matrix A admits an explicit expression that can be obtained with some bookkeeping. Denoting by C, D, E the commutation, duplication and elimination matrices (Lütkepohl, 2005, Appendix A.12), it can be written as

$$A = \left[\begin{array}{c|ccc} A_{00} & A_{01} & \cdots & A_{0q} \\ \hline A_{10} & & & \\ \vdots & & A_{**} & \\ A_{q0} & & & \end{array}\right]$$

with

$$\begin{aligned}
A_{00} &= E\left(I + \sum_{i=1}^{q} \Theta_i \otimes \Theta_i\right)D, \\
A_{0i} &= E\left(\Theta_i \Sigma \otimes I + (I \otimes (\Theta_i \Sigma))C\right), \\
A_{i0} &= \left(I \otimes \Theta_i + \sum_{j=1}^{q-i} \Theta_j \otimes \Theta_{j+i}\right)D, \quad \text{and} \\
A_{**} &= \begin{bmatrix} \Sigma \otimes I & \Theta_1 \Sigma \otimes I & \Theta_2 \Sigma \otimes I & \cdots & \Theta_{q-1}\Sigma \otimes I \\ 0 & \Sigma \otimes I & \Theta_1 \Sigma \otimes I & \cdots & \Theta_{q-2}\Sigma \otimes I \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \Sigma \otimes I & \Theta_1 \Sigma \otimes I \\ 0 & 0 & \cdots & 0 & \Sigma \otimes I \end{bmatrix} \\
&\quad + \begin{bmatrix} (I \otimes \Theta_2 \Sigma)C & (I \otimes \Theta_3 \Sigma)C & \cdots & (I \otimes \Theta_q \Sigma)C & 0 \\ (I \otimes \Theta_3 \Sigma)C & (I \otimes \Theta_4 \Sigma)C & \iddots & 0 & 0 \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ (I \otimes \Theta_q \Sigma)C & 0 & \iddots & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix},
\end{aligned}$$

where the latter expression is the sum of a triangular matrix and an anti-triangular one with 0 on the main anti-diagonal. For instance, for $q = 2$, one has

$$A = \begin{bmatrix} E(I + \Theta_1 \otimes \Theta_1 + \Theta_2 \otimes \Theta_2)D & E(\Theta_1 \Sigma \otimes I + (I \otimes \Theta_1 \Sigma)C) & E(\Theta_2 \Sigma \otimes I + (I \otimes \Theta_2 \Sigma)C) \\ (I \otimes \Theta_1 + \Theta_1 \otimes \Theta_2)D & \Sigma \otimes I + (I \otimes \Theta_2 \Sigma)C & \Theta_1 \Sigma \otimes I \\ (I \otimes \Theta_2)D & 0 & \Sigma \otimes I \end{bmatrix}.$$

Note that, in the case of a scalar MA, A reduces to the matrix $J_{\mathbf{w}}$ in (28).

Appendix C.5. Proof of Theorem 19

Proof of Theorem 19. The asymptotic covariance of the estimator $\hat{\gamma}$ in (25) is $J\mathscr{I}^{-1}\Xi\mathscr{I}^{-1}J^T$, with $J = \mathrm{diag}(J_{\mathbf{w}_1}, J_{\mathbf{w}_2}, \ldots, J_{\mathbf{w}_k})$. Since $\hat{\mathbf{z}}$ is computed in our estimator by solving (27), and

$$\hat{W} \xrightarrow{\text{a.s.}} W = \mathrm{diag}(V_{(\mathbf{w}_1)}, V_{(\mathbf{w}_2)}, \ldots, V_{(\mathbf{w}_k)})^{-1}, \tag{C.14}$$

its asymptotic covariance is

$$(X^T W X)^{-1} X^T W J \mathscr{I}^{-1} \Xi \mathscr{I}^{-1} J^T W^T X (X^T W X)^{-1}.$$

The final step of the computation is determining $\hat{\boldsymbol{\alpha}}$ from the estimated covariances in $\hat{\mathbf{z}}$. The Jacobian matrix of the function that maps $\mathbf{z}$ to $\boldsymbol{\alpha}$ is A^{-1}, hence, putting everything together, the asymptotic variance of the estimator $\hat{\boldsymbol{\alpha}}$ is

$$\Psi = A^{-1}(X^T W X)^{-1} X^T W J \mathscr{I}^{-1} \Xi \mathscr{I}^{-1} J^T W^T X (X^T W X)^{-1} (A^T)^{-1}.$$

□