Estimation of a harmonic component and banded covariance matrix in a multivariate time series

Zhanna Andrushchenko

Research Report
Centre of Biostochastics

Swedish University of Agricultural Sciences

Report 2010:05
ISSN 1651-8543

Estimation of a harmonic component and banded covariance matrix in a multivariate time series

Zhanna Andrushchenko[1]

Department of Energy and Technology and Centre of Biostochastics,
Swedish University of Agricultural Sciences
P.O. Box 7032, SE-750 07 Uppsala, Sweden

Abstract

A multivariate time series that is asymptotically stationary to second order and contains a harmonic component is considered. The special case where the marginal univariate time series are independent and the marginal covariance matrix of errors is banded of order m is investigated in detail. Explicit analytical estimators for the harmonic component and the covariance matrix are presented. Properties of estimators are also given.

Keywords: Banded covariance matrix; Covariance matrix estimation; Harmonic regression; Multivariate normal distribution; Multivariate time series

AMS classification: 62F12, 62F30, 62H12, 62M10, 62M15

[1]E-mail address to the correspondence author: Zhanna.Andrushchenko@et.slu.se

1 Introduction

Periodic phenomena are frequently observed in nature, with examples to be found in climatology (Bloomfield et al., 1994), economics (Doran and Quilkey, 1972), geosciences (MacDonald, 1989), etc. In many biomedical experiments the measurements of a physiological variable taken at different time points show a certain periodic pattern (Tong, 1976).

There are two distinct yet broadly equivalent modes of time series analysis. On the one hand are the time-domain methods, which deal mainly with the autocovariance function and the cross-covariance functions of the series, and lead towards the construction of structural or parametric models. On the other hand are the frequency-domain methods of spectral analysis. These are based on an extension of the methods of Fourier analysis which originate in the idea that over a finite interval, any analytical function can be approximated, to whatever degree of accuracy desired, by taking a weighted sum of sine and cosine functions of different frequencies.

The origin of spectral analysis can be traced to the pioneering work of Sir Arthur Schuster (Schuster, 1898) who introduced a numerical method for spectrum analysis, the periodogram. Early works for testing the presence of hidden periodicities in which the noise is Gaussian white noise include Schuster (1898), Fisher (1929) and Hartley (1949). Later research focused more on the detection of signal in the presence of colored noise. In this case, the fundamental work was done by Whittle (1952) and Priestley (1962a,b). Walker (1971) and Hannan (1973) were among other contributors to the theory of periodogram estimation. For a review of a variety of spectral analysis methods available, see e.g. Stoica and Moses (1997).

Most existing time series techniques are applicable to time series that are stationary in some sense. A sequence of random variables $\{X_i\},\ (i = 1, 2, \ldots)$ is said to be (a) *stationary to second order* if, for every integer k, the expectations $E(X_i),\ E(X_iX_{i+k})\ (i > 0, i + k > 0)$ exist finitely and are independent of i; (b) *asymptotically stationary to second order* if, for every integer k, the expectations $E(X_i),\ E(X_iX_{i+k})\ (i > 0, i+k > 0)$ exist finitely and have finite limits as $i \to \infty$; (c) *stationary* if the distribution function of $(X_{i+1}, \ldots, X_{i+k})$ is independent of i, for every positive integer k (Diananda, 1954). Although in practise many signals show non-stationary behavior, most of non-stationary time series can be transformed to stationary series.

In many signal processing problems involving multivariate random data from a population the covariance must be estimated. If one has a priori knowledge of structure in the true underlying covariance of the random process, this information can be exploited in covariance estimation procedures. In some

applications, the structure of the problem suggests that the underlying, true covariance matrix is the Kronecker product of two valid covariance matrices. If such a separability between matrices can be assumed, this dramatically reduces the number of parameters to be estimated. For example, let us consider a process that varies in both time and space, and is measured on p locations and n times. Then the covariance matrix has $np(np+1)/2$ parameters. One way to dramatically reduce the number of parameters is to assume space-time separability. This model has only $n(n+1)/2 + p(p+1)/2 - 1$ parameters.

While in many statistical problems the observations are independent, in time series successive observations may be dependent, and the dependence may depend on the position in the sequence. A sequence of r.v.s $(X_1,\ X_2,\ \ldots,\ X_n)$ is said to be m-dependent if two sets $(X_1,\ X_2\ \ldots,\ X_r),\ \ (X_s,\ X_{s+1}\ \ldots,\ X_n)$ are independent, provided $s - r > m$ (Hoeffding and Robbins, 1948). This implies a banded covariance structure obtained by setting all covariances more than m steps apart equal to zero. Such a structure was studied by Hoeffding and Robbins (1948), Diananda (1954), Cocke (1972), Berk (1973), Ferguson (1996), and, recently, by Andrushchenko et al. (2008) and Ohlson et al. (2009). The assumption of such a covariance matrix is in many cases a natural assumption and is usually much more natural than the presumption of independence between observations, which is frequently applied. It should be noted, however, that Andrushchenko et al. (2008) and Ohlson et al. (2009) allowed unequal variances and correlation coefficients. Such a situation differs from standard time series assumptions and results in more parameters to be estimated than in conventional time series.

This paper deals with a multivariate time series that is asymptotically stationary to second order (or can be transformed to such) and that contains a harmonic component.

The model is based on a spectral approach, under the assumption that the observations are multivariate normally distributed. The case where the univariate time series are independent and the covariance matrix of errors is banded of order m is investigated in detail. Each univariate time series is analyzed with a spectral method, and estimators are calculated via a modified least-squares method. Finally, the banded covariance structure is explicitly estimated via the maximum likelihood approach.

Basically, the algorithm attempts to: (1) identify the spectral peak in an empirical spectrum of the time series data; (2) assign a harmonic component to the spectral peak; (3) estimate a covariance matrix. The main aim is to find reasonable explicit estimators for both for the mean (harmonic structure) and the covariance matrix.

The paper is organized as follows. Section 2 describes the model, while Section 3 presents an estimation procedure. Section 3.1 deals with the harmonic part, and a theorem about consistency of estimators is presented and proved. Section 3.2 provides the algorithm for estimation of the banded covariance matrix. Here the proposed algorithm consists of maximizing a modified likelihood function via inserting the estimated parameters from previous steps. Finally, Section 4 summarizes the paper. Some useful collections of formulae are presented in the Appendices: classical harmonic regression and periodogram analysis are reviewed in Appendix A; different Fourier transforms are described in Appendix B; and Appendix C contains the main formulae related to matrix normal distribution, some further definitions, and the main inequalities used throughout the paper.

2 Model

Unless otherwise stated, scalars and matrix elements are denoted by ordinary letters, vectors by small bold letters, and matrices by capital bold letters.

For any square matrix $\mathbf{A}\colon n\times n$, $\mathbf{A}_k$ $(k<n)$ represents the $k\times k$ submatrix located in the top left-hand corner of $\mathbf{A}$; $\mathbf{A}'$ means the transpose of $\mathbf{A}$, $\mathbf{A}^{-1}$ means the inverse of $\mathbf{A}$, and $|\mathbf{A}|$ stands for the determinant of the matrix.

Let a matrix $\mathbf{A}\colon p\times n$ be:

$$\mathbf{A}=(a_{ij})=\begin{pmatrix} a_{11} & \dots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{p1} & \dots & a_{pn}\end{pmatrix}=\begin{pmatrix}\mathbf{a}_1:\ldots:\mathbf{a}_n\end{pmatrix},\ \mathbf{a}_j=\begin{pmatrix} a_{1j}\\ \vdots\\ a_{pj}\end{pmatrix}. \quad (2.1)$$

To make the formulae more compact, we also use the following notation:

$$(\mathbf{a}_i:\ldots:\mathbf{a}_j)=\mathbf{A}_{i:j},\ \ \mathbf{a}_{k|i:j}=\mathbf{a}_k|\mathbf{A}_{i:j}. \quad (2.2)$$

For any reasonable matrices, $(\mathbf{A}-c\mathbf{B})'(\ldots)$ means $(\mathbf{A}-c\mathbf{B})'(\mathbf{A}-c\mathbf{B})$.

Other notations are conventional: "p.d.f." for probability density function, "i.i.d" for independent and identically distributed, "LS" for least squares, "ML" for maximum likelihood, "RHS" for right-hand side.

As mentioned in the Introduction, the multivariate time series under consideration is asymptotically stationary to second order or can be transformed to stationary. To analyze such a time series that contains a harmonic component, one can use a spectral approach. Classical harmonic regression together with use of the periodogram as an estimator are presented in detail in Appendix A.

We assume that the data form a matrix $\mathbf{X} = (x_{ij})\colon p \times n$, where p is the number of univariate time series and n is the number of observations in each univariate time series. We also assume that $\mathbf{X}$ follows the matrix normal distribution (see Appendix C.1) $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi})$, where $\mathbf{M} = (m_{ij})\colon p \times n$ describes the harmonic structure, $\mathbf{\Sigma} = (\sigma_{ij})\colon p \times p$ describes the covariance between rows, and $\mathbf{\Psi} = (\psi_{ij})\colon n \times n$ describes the covariance between columns. As usual, we assume that both the covariance matrices are positive definite. It is worth noting that the role of $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ can be interchanged by considering the transpose of $\mathbf{X}$: if $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi})$, then $\mathbf{X}' \sim N_{n,p}(\mathbf{M}', \mathbf{\Psi}, \mathbf{\Sigma})$.

The joint p.d.f. of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\mathbf{X}) = (2\pi)^{-np/2}\, |\mathbf{\Sigma}|^{-n/2}\, |\mathbf{\Psi}|^{-p/2} \times \exp\Big(-\frac{1}{2}\, \mathrm{tr}\big(\mathbf{\Sigma}^{-1}\, (\mathbf{X} - \mathbf{M})\, \mathbf{\Psi}^{-1}\, (\mathbf{X} - \mathbf{M})'\big)\Big). \tag{2.3}$$

The joint p.d.f. of $\mathbf{X}$, considered as a function of the parameters $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ (for fixed observed $\mathbf{X}$) will serve as the likelihood function. Some useful formulae related to the matrix normal distribution are presented in Appendix C.1, where we adopt the notation of Kollo and von Rosen (2005, Chapter 2.2).

We start from a model that consists of a sum of a finite number of sinusoids and a noise component. We allow the amplitudes of all fundamental oscillations and the fundamental frequencies to be different for different univariate time series. If the sampling times are taken to be equally spaced so that $t = 1, \ldots, n$, the model can be written as

$$x_{it} = \sum_{k=1}^{q} \big(A_{ik} \cos(\omega_{ik} t) + B_{ik} \sin(\omega_{ik} t)\big) + \varepsilon_{it}, \quad i = 1, \ldots, p, \quad t = 1, \ldots, n. \tag{2.4}$$

Here A_{ik} and B_{ik} are unknown parameters, and ω_{ik} are unknown different angular frequencies measured in radians per unit time. It is assumed here that there are few (for example q) fundamental frequencies that are responsible for a harmonic structure of time series. It is also assumed that errors, ε_{it}, are normally distributed, but are not necessarily white noise.

Limiting the consideration to harmonic structure with $q = 1$, then:

$$x_{it} = A_i \cos(\omega_i t) + B_i \sin(\omega_i t) + \varepsilon_{it}, \quad i = 1, \ldots, p, \quad t = 1, \ldots, n. \tag{2.5}$$

We now have a set of p regression equations. By means of notation, the model (2.5) can be transformed to:

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\beta}_i + \mathbf{u}_i, \quad i = 1, \ldots, p, \tag{2.6}$$

with

$$\mathbf{y}_i = (x_{it})' = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix} : \ n \times 1; \quad \mathbf{u}_i = (\varepsilon_{it})' = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in} \end{pmatrix} : \ n \times 1;$$

$$\mathbf{Z}_i = \begin{pmatrix} \cos(\omega_i) & \sin(\omega_i) \\ \cos(2\omega_i) & \sin(2\omega_i) \\ \vdots & \vdots \\ \cos(n\omega_i) & \sin(n\omega_i) \end{pmatrix} : \ n \times 2; \quad \boldsymbol{\beta}_i = \begin{pmatrix} A_i \\ B_i \end{pmatrix} : \ 2 \times 1. \tag{2.7}$$

If ω_i were known this model would coincide with classical linear seemingly unrelated regression (SUR) model (Zellner, 1962). A SUR model comprises several individual relationships that are linked by the fact that their error terms are correlated. Zellner (1962) has shown that by stacking the observations either in the t dimension or for each i, the model (2.6) can be further transformed to a single-equation regression model. The estimation then yields regression estimators that are at least asymptotically more efficient than single-equation least squares (LS) estimators. However, if any equation is misspecified, then the entire $\boldsymbol{\beta}$ will be inconsistently estimated by this method. In this sense, equation-by-equation LS provides some degree of robustness since it is not affected by misspecification in other equations of the system. Therefore we will limit ourselves to equation-by-equation estimation.

The model (2.5) can be rewritten in matrix form as:

$$\mathbf{X} = \mathbf{M} + \mathbf{E}, \tag{2.8}$$

where

$$\mathbf{M} = \boldsymbol{\beta}\mathbf{Z}, \ \ \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1' & 0 & \dots & 0 \\ 0 & \boldsymbol{\beta}_2' & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \boldsymbol{\beta}_p' \end{pmatrix}, \ \ \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1' \\ \mathbf{Z}_2' \\ \vdots \\ \mathbf{Z}_p' \end{pmatrix}, \tag{2.9}$$

with $\boldsymbol{\beta}$: $p \times 2p$ and $\mathbf{Z}$: $2p \times n$, $\boldsymbol{\beta}_i$ and $\mathbf{Z}_i$ given by (2.7), and $\mathbf{E} = (\varepsilon_{it})$: $p \times n$. Thus, $\mathbf{M}$ can be presented as

$$\mathbf{M} = (\mathbf{m}_1 : \dots : \mathbf{m}_n), \ \ \mathbf{m}_j = \begin{pmatrix} A_1 \cos(j\omega_1) + B_1 \sin(j\omega_1) \\ \vdots \\ A_p \cos(j\omega_p) + B_p \sin(j\omega_p) \end{pmatrix}, \ \ j = 1, \dots, n. \tag{2.10}$$

Since $A_i,\ B_i,\ \omega_i$ are all unknown, the model (2.8) is nonlinear in the parameters. To estimate the harmonic component of the model, a modified LS method can be used (see e.g. Appendix A for details).

So far, the covariance structure in the model has not been specified. This is intentional. As far as the calculation of LS estimators is concerned, random errors need not to be white noise.

The estimation of the covariance matrices is an important problem by itself and is a crucial part of many signal processing algorithms. Two limiting cases are worth mentioning. First, in multivariate statistics, the observations in each row are frequently considered as i.i.d. Therefore $\mathbf{\Psi}$ is assumed to be $\mathbf{I}$, the identity matrix, and the estimation of $\mathbf{\Sigma}$ is of primary interest. For the case when $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{I})$, with $\mathbf{\Sigma}$ being a banded of order m, the explicit estimators for the mean and for the covariance matrix were found in Andrushchenko et al. (2008) (for $m = 1$) and Ohlson et al. (2009) (for arbitrary m). It was also shown that estimation of the mean is not affected by the covariance structure, whereas estimator of the mean itself is the part of estimator of the covariance matrix.

Another limiting case corresponds to time series, where nearby observations are correlated, and is considered below. Here, it is reasonable to assume that $\mathbf{\Psi}$ has a banded covariance structure of order m, i.e. all covariances more than m steps apart the main diagonal are equal to zero, $\psi_{ij} = 0$ for $|i-j| > m$. Moreover, we allow unequal variances and correlation coefficients:

$$\mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \dots & \psi_{1,m+1} & 0 & 0 & \dots & 0 \\ \psi_{21} & \dots & \psi_{2,m+1} & \psi_{2,m+2} & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \psi_{m+1,1} & \dots & \dots & \psi_{m+1,2m+1} & 0 & \dots & 0 \\ 0 & \psi_{m+2,2} & \dots & \dots & \psi_{m+2,2m+2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \psi_{n-1,n-m-1} & \psi_{n-1,n-m} & \dots & \psi_{n-1,n} \\ 0 & \dots & 0 & 0 & \psi_{n,n-m} & \dots & \psi_{n,n} \end{pmatrix}. \tag{2.11}$$

Such a situation differs from standard time series cases and results in more parameters to be estimated. Next, we assume that each row or each univariate time series is independent of the others, therefore $\mathbf{\Sigma}$ is $\mathbf{I}$. With such assumptions about the covariance matrices, $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{I}, \mathbf{\Psi})$.

The main aim is to find reasonable explicit estimators for both the mean (harmonic structure) and the covariance matrix. As far as we limit the consid-

eration to a harmonic structure with $q = 1$, we need to estimate $3p$ parameters for the mean $(A_i, B_i, \omega_i,\ i = 1, \ldots, p)$ and $\frac{1}{2}(2n - m)(m + 1)$ nonzero parameters for the banded covariance matrix. As we have $p \times n$ observations, with both p and n being large enough, we are able to estimate all the parameters for the harmonic structure and the covariance matrix explicitly.

3 Estimation

3.1 Estimation of the harmonic structure

We have shown in Section 2 that if each row in model (2.8) can be treated independently from other rows, regression techniques can be applied to estimate parameters in each row.

A natural estimation procedure for each row is the following (with the index indicating the row number omitted for convenience):

1. The Fast Fourier transform (see Appendix B for details) is used to calculate the spectrum. Hidden periodicities are detected and estimated by periodogram analysis. Namely, $\hat{\omega}$ is determined by maximizing the periodogram

$$\hat{\omega} = \arg\max_{\omega} I(\omega), \tag{3.1}$$

with the periodogram $I(\omega)$ given by

$$I(\omega) = \frac{2}{n}\Big|\sum_{t=1}^{n} x_t \exp(i\omega t)\Big|^2. \tag{3.2}$$

Here i is the imaginary unit, and for a complex variable z, $|z|^2 = zz^*$, where * means conjugate.

2. The estimators of A and B are obtained by minimizing a modified residual sum of squares, $U(A, B, \omega)$, see Appendix A, (A.13) and (A.15), for details:

$$\hat{A}(\hat{\omega}) = \frac{2}{n}\sum_{t=1}^{n} x_t \cos(\hat{\omega}t), \quad \hat{B}(\hat{\omega}) = \frac{2}{n}\sum_{t=1}^{n} x_t \sin(\hat{\omega}t). \tag{3.3}$$

3. After that, the matrix $\mathbf{M}$ is estimated via (2.10) as $\hat{\mathbf{M}} = \mathbf{M}(\hat{A}, \hat{B}, \hat{\omega})$.

It might be expected that this procedure could be improved, by computing a weighted regression using the known or estimated covariance matrix or residuals. However, Hannan (1973) has shown that asymptotically this is not so, basically because the regression component is of nature of a signal sent at one frequency, so that ultimately only the noise at that frequency matters. Of course for finite sample sizes and a very irregular spectrum of $\mathbf{X}$ near $\hat{\boldsymbol{\omega}}$ the influence of the covariance structure might be significant, even though its influence vanishes asymptotically. However, the covariance structure of $\mathbf{X}$ is unknown and has to be estimated. For simplicity we confine ourselves to unweighted procedures.

For the case where errors are i.i.d. with $E(\varepsilon_t^2) = v < \infty$, Walker (1971) has shown that the estimators $\hat{A}, \hat{B}, \hat{\omega}$, and $\hat{v}$ are all consistent as $n \to \infty$. Here we prove the consistency of $\hat{A}, \hat{B}$, and $\hat{\omega}$ when errors are m-dependent. Estimation of the covariance structure is given in Section 3.2.

Theorem 3.1. *For each row in* $\mathbf{X}$*, let*

$$x_t = A_0 \cos(\omega_0 t) + B_0 \sin(\omega_0 t) + \varepsilon_t \quad (0 < \omega_0 < \pi, \quad t = 1, \ldots, n), \tag{3.4}$$

where the index 0 *is used to indicate true value,* ε_t *are* m*-dependent normally distributed with* $E(\varepsilon_t) = 0$ *and* $E(\varepsilon_t^2) = v < \infty$*, whereas* $E(\varepsilon_t \varepsilon_k) < v < \infty$ *for* $|t-k| \le m$ *and* $E(\varepsilon_t \varepsilon_k) = 0$ *for* $|t-k| > m$*. Then the estimators* $\hat{A}$*,* $\hat{B}$*,* $\hat{\omega}$ *are all consistent as* $n \to \infty$*.*

It is worth remembering that the index, indicating the row number has been omitted for convenience, and each row is treated separately. The generic notation "v" is different for each row i and, formally, $v \equiv \psi_{ii}$.

Proof. First, we show that $\hat{\omega}$ is a consistent estimator of ω_0, namely we show that $\hat{\omega} - \omega_0 = o_p(n^{-1})$ as $n \to \infty$. We then use this to prove the consistency of $\hat{A}$ and $\hat{B}$.

We start from the definition of the periodogram, and use

$$A_0 \cos(\omega_0 t) + B_0 \sin(\omega_0 t) = D_0 \exp(i\omega_0 t) + D_0^* \exp(-i\omega_0 t), \tag{3.5}$$

where

$$D_0 = \frac{1}{2}(A_0 - iB_0), \quad D_0^* = \frac{1}{2}(A_0 + iB_0). \tag{3.6}$$

Next, we introduce a function

$$h(u) = \sum_{t=1}^{n} \exp(iut) = \begin{cases} \frac{\sin(nu/2)}{\sin(u/2)} \exp\left(\frac{i(n+1)u}{2}\right) & (0 < u < 2\pi), \\ n & (u = 0 \text{ or } 2\pi). \end{cases} \tag{3.7}$$

Then, from definition (3.2) of $I(\omega)$, with x_t being given by (3.4),

$$\begin{aligned}\frac{n}{2}I(\omega) &= \Big|\sum_{t=1}^{n} x_t \exp(i\omega t)\Big|^2 \\ &= \Big|\sum_{t=1}^{n} \exp(i\omega t)\Big(D_0 \exp(i\omega_0 t) + D_0^* \exp(-i\omega_0 t) + \varepsilon_t\Big)\Big|^2 \\ &= \Big|\sum_{t=1}^{n} \exp(i\omega t)\Big(D_0 \exp(i\omega_0 t) + D_0^* \exp(-i\omega_0 t)\Big)\Big|^2 + \Big|\sum_{t=1}^{n} \varepsilon_t \exp(i\omega t)\Big|^2 \\ &+ 2\Re\Big\{\sum_{t=1}^{n} \varepsilon_t \exp(-i\omega t)\Big(D_0\ h(\omega+\omega_0) + D_0^*\ h(\omega-\omega_0)\Big)\Big\}, \qquad (3.8)\end{aligned}$$

where $\Re$ means the real part of the expression.

Using (3.7), the first term on the RHS of (3.8) can be transformed to:

$$\begin{aligned}&\Big|\sum_{t=1}^{n} \exp(i\omega t)\Big(D_0 \exp(i\omega_0 t) + D_0^* \exp(-i\omega_0 t)\Big)\Big|^2 \\ &\quad= \Big|D_0 h(\omega+\omega_0) + D_0^* h(\omega-\omega_0)\Big|^2 \\ &\quad= \big|D_0\big|^2\Big(|h(\omega-\omega_0)|^2 + |h(\omega+\omega_0)|^2\Big) \\ &\quad+ D_0^2\ h(\omega+\omega_0)h^*(\omega-\omega_0) + D_0^{*2}\ h(\omega-\omega_0)h^*(\omega+\omega_0). \qquad (3.9)\end{aligned}$$

Since $0 < \omega_0 < \pi$, (3.7) implies that:

$$\max_{0\le\omega\le\pi} |h(\omega-\omega_0)| = n, \quad \max_{0\le\omega\le\pi} |h(\omega+\omega_0)| = O(1). \qquad (3.10)$$

Therefore the RHS of (3.9) is dominated by $\big|D_0\big|^2|h(\omega-\omega_0)|^2$, resulting in

$$\Big|\sum_{t=1}^{n} \exp(i\omega t)\Big(D_0 \exp(i\omega_0 t) + D_0^* \exp(-i\omega_0 t)\Big)\Big|^2 = \big|D_0\big|^2|h(\omega-\omega_0)|^2 + O(n). \qquad (3.11)$$

The second term on the RHS of (3.8) can be estimated as:

$$\Big|\sum_{t=1}^{n}\varepsilon_t\exp(i\omega t)\Big|^2=\sum_{t=1}^{n}\varepsilon_t\exp(i\omega t)\sum_{\tau=1}^{n}\varepsilon_\tau^*\exp(-i\omega\tau)$$
$$=\sum_{t=1}^{n}\sum_{\tau=1}^{n}\varepsilon_t\varepsilon_\tau\exp(i\omega(t-\tau))=\sum_{t=1}^{n}\varepsilon_t^2+2\sum_{t=1}^{n-s}\sum_{s=1}^{n-1}\varepsilon_t\varepsilon_{t+s}\exp(-i\omega s)$$
$$\leq\sum_{t=1}^{n}\varepsilon_t^2+2\sum_{s=1}^{n-1}\Big|\sum_{t=1}^{n-s}\varepsilon_t\varepsilon_{t+s}\Big|, \tag{3.12}$$

with the expectation being:

$$E\Big|\sum_{t=1}^{n}\varepsilon_t\exp(i\omega t)\Big|^2\leq E\Big(\sum_{t=1}^{n}\varepsilon_t^2\Big)+2\sum_{s=1}^{n-1}E\Big|\sum_{t=1}^{n-s}\varepsilon_t\varepsilon_{t+s}\Big|$$
$$\overset{C-S}{\leq}E\Big(\sum_{t=1}^{n}\varepsilon_t^2\Big)+2\sum_{s=1}^{n-1}\Big(E\Big(\sum_{t=1}^{n-s}\varepsilon_t\varepsilon_{t+s}\Big)^2\Big)^{1/2}$$
$$\overset{m-dep.}{=}E\Big(\sum_{t=1}^{n}\varepsilon_t^2\Big)+2\sum_{s=1}^{m}\Big(E\sum_{t,\tau=1}^{n-s}\varepsilon_t\varepsilon_{t+s}\varepsilon_\tau\varepsilon_{\tau+s}\Big)^{1/2}$$
$$\leq nv+2\sum_{s=1}^{m}\Big((n-s)v^2\Big)^{1/2}=v\Big(n+2\sum_{r=n-m}^{n-1}r^{1/2}\Big)$$
$$\leq v\Big(n+2\int_{r=n-m}^{n}r^{1/2}dr\Big)=v\Big(n+\frac{4}{3}\big(n^{3/2}-(n-m)^{3/2}\big)\Big). \tag{3.13}$$

Here, $\overset{C-S}{}$ means "by the Cauchy-Schwarz inequality", see formula (C.13) in Appendix C, and $\overset{m-dep.}{}$ means "taking into account m-dependence". Assuming large n and reasonable m and applying the Markov inequality (C.14) to (3.13), with $Z=\Big|\sum_{t=1}^{n}\varepsilon_t\exp(i\omega t)\Big|^2$, leads to the following:

$$vn\geq EZ\overset{Markov}{\geq}C_\epsilon n\,P\big(Z\geq C_\epsilon n\big), \tag{3.14}$$

where $\overset{Markov}{}$ means "by the Markov inequality", see formula (C.14), and C_ϵ is any given constant. This means that for all $\epsilon>0$ there is a $C_\epsilon=v/\epsilon$, such that $P\big(Z\leq C_\epsilon n\big)\geq 1-\epsilon$. According to (C.10) this implies that

$$\Big|\sum_{t=1}^{n}\varepsilon_t\exp(i\omega t)\Big|^2=O_p(n). \tag{3.15}$$

Note, that Walker (1971) obtained $O_p(n^{3/2})$ here, because instead of (3.13) he used an upper bound which was $O(n^{3/2})$.

To estimate the third term of (3.8), we start with the following:

$$E\Big|\sum_{t=1}^{n}\varepsilon_t\exp(-i\omega t)\Big| \le E\Big|\sum_{t=1}^{n}\varepsilon_t\Big| \overset{C-S}{\le} \Big(E\Big(\sum_{t=1}^{n}\varepsilon_t\Big)^2\Big)^{1/2}$$
$$\overset{triangle}{\le} \Big(E\Big(\sum_{t=1}^{n}\varepsilon_t^2\Big)\Big)^{1/2} \le \big(nv\big)^{1/2}. \tag{3.16}$$

Here, $\overset{triangle}{}$ means "by the triangle inequality", see (C.15). Applying the Markov inequality again, the last term on the RHS of (3.8) can be estimated as:

$$2\Re\Big\{\sum_{t=1}^{n}\varepsilon_t\exp(-i\omega t)\Big(D_0\ h(\omega+\omega_0)+D_0^*\ h(\omega-\omega_0)\Big)\Big\} = O_p(n^{3/2}). \tag{3.17}$$

Combining all these estimators and taking into account (3.10) gives:

$$\max_{0\le\omega\le\pi}\Big(\frac{n}{2}I(\omega)-|D_0|^2\ |h(\omega-\omega_0)|^2\Big) = O(n)+O_p(n)+O_p(n^{3/2}), \tag{3.18}$$

and, after using (3.6),

$$\max_{0\le\omega\le\pi}\Big(I(\omega)-\frac{1}{2n}\big(A_0^2+B_0^2\big)\ |h(\omega-\omega_0)|^2\Big) = O_p(n^{1/2}). \tag{3.19}$$

The function $|h(u)|^2 = \sin^2(nu/2)/\sin^2(u/2)$ $(0 < u < 2\pi)$ decreases monotonically from its absolute maximum of n^2 at $u = 0$ to 0 at $u = 2\pi/n$, and then oscillates between local maxima and minima. Thus, for any δ which is sufficiently small,

$$\max_{|\omega-\omega_0|\ge\delta/n}|h(\omega-\omega_0)|^2 \le \frac{\sin^2\big(\frac{\delta}{2}\big)}{\sin^2\big(\frac{\delta}{2n}\big)}, \tag{3.20}$$

and therefore, for large n,

$$\max_{|\omega-\omega_0|\ge\delta/n} I(\omega) \le \frac{1}{2n}\big(A_0^2+B_0^2\big)\frac{\sin^2\big(\frac{\delta}{2}\big)}{\sin^2\big(\frac{\delta}{2n}\big)} + O_p(n^{1/2}). \tag{3.21}$$

Thus, with probability tending to 1 as $n\to\infty$,

$$\lim_{n\to\infty}\Big(n^{-1}\max_{|\omega-\omega_0|\ge\delta/n} I(\omega)\Big) \le \frac{1}{2}\big(A_0^2+B_0^2\big)\lim_{n\to\infty}\left(\frac{\sin^2\big(\frac{\delta}{2}\big)}{n^2\sin^2\big(\frac{\delta}{2n}\big)}\right) \le \frac{1}{2}\big(A_0^2+B_0^2\big). \tag{3.22}$$

On the other hand, from (3.10) and (3.19) it follows that

$$\lim_{n\to\infty}\big(n^{-1}I(\omega_0)\big) \xrightarrow{p} \frac{1}{2}\big(A_0^2+B_0^2\big). \tag{3.23}$$

This implies that:

$$P\big(\max_{|\omega-\omega_0|\geq\delta/n} I(\omega)\leq I(\omega_0)\big)\to 1 \;\; as \;\; n\to\infty. \tag{3.24}$$

Suppose $\hat{\omega}\in|\omega-\omega_0|\geq\delta/n$. Then $\max_{|\omega-\omega_0|\geq\delta/n} I(\omega)=I(\hat{\omega})$, and

$$P\big(\max_{|\omega-\omega_0|\geq\delta/n} I(\omega_0)>I(\omega_0)\big)=P\big(I(\hat{\omega})>I(\omega_0)\big)=1 \tag{3.25}$$

which contradicts (3.24). Thus, $\hat{\omega}\in|\omega-\omega_0|<\delta/n$ and

$$P\big(n|\hat{\omega}-\omega_0|<\delta\big)\to 1 \;\; as \;\; n\to\infty, \tag{3.26}$$

Since δ can be arbitrarily small, the last expression is equivalent to

$$\hat{\omega}-\omega_0=o_p(n^{-1}) \;\; as \;\; n\to\infty. \tag{3.27}$$

and, therefore $\hat{\omega}$ is a consistent estimator of ω_0.

Now, we show that $\hat{A}$ is a consistent estimator of A_0 and $\hat{B}$ is a consistent estimator of B_0. Using (3.3)-(3.7),

$$\begin{aligned}\hat{A}+i\hat{B} &= \frac{2}{n}\sum_{t=1}^{n} x_t\exp(i\hat{\omega}t)\\ &= \frac{2}{n}\sum_{t=1}^{n}\Big(D_0\exp(i\omega_0 t)+D_0^*\exp(-i\omega_0 t)+\varepsilon_t\Big)\exp(i\hat{\omega}t)\\ &= \frac{2}{n}\Big(D_0h(\hat{\omega}+\omega_0)+D_0^*h(\hat{\omega}-\omega_0)\Big)+\frac{2}{n}\sum_{t=1}^{n}\varepsilon_t\exp(i\hat{\omega}t).\end{aligned} \tag{3.28}$$

Thus:

$$\begin{aligned}|\hat{A}-A_0+i(\hat{B}-B_0)| &= |\hat{A}+i\hat{B}-2D_0^*|\\ &\leq \frac{2|D_0|}{n}\Big(|h(\hat{\omega}+\omega_0)|+|h(\hat{\omega}-\omega_0)-n|\Big)+\frac{2}{n}\Big|\sum_{t=1}^{n}\varepsilon_t\exp(i\hat{\omega}t)\Big|.\end{aligned} \tag{3.29}$$

(3.10) implies that:

$$\frac{1}{n}|h(\hat{\omega}+\omega_0)| \xrightarrow{p} 0 \;\; as \;\; n \to \infty. \tag{3.30}$$

As far as $h(u)$ is continuous and differentiable on $(0, u)$, with $h(0) = n$, then one can apply the mean value theorem to $h(u)$ to get:

$$\frac{|h(\hat{\omega}-\omega_0) - h(0)|}{|(\hat{\omega}-\omega_0)|} = |h'(u)|. \tag{3.31}$$

On the other side,

$$|h'(u)| = \Big|\Big(\sum_{t=1}^{n} \exp(iut)\Big)'\Big| = \Big|\sum_{t=1}^{n} t\exp(iut)\Big| \leq \sum_{t=1}^{n} t < n^2, \tag{3.32}$$

and therefore,

$$\Big|\frac{1}{n}\big(h(\hat{\omega}-\omega_0) - n\big)\Big| < n|\hat{\omega}-\omega_0|. \tag{3.33}$$

Together with consistency of $\hat{\omega}$, (3.27), this implies that:

$$\Big|\frac{1}{n}\big(h(\hat{\omega}-\omega_0) - n\big)\Big| \xrightarrow{p} 0 \;\; as \;\; n \to \infty. \tag{3.34}$$

From (3.15) it follows that the last term in (3.29) is $O_p(n^{-1/2})$, and therefore:

$$\Big|\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t \exp(i\hat{\omega}t)\Big| \xrightarrow{p} 0 \;\; as \;\; n \to \infty. \tag{3.35}$$

Finally,

$$|\hat{A} - A_0 + i(\hat{B} - B_0)| \xrightarrow{p} 0 \;\; as \;\; n \to \infty. \tag{3.36}$$

A complex random variable converges to 0 if and only if its real and imaginary parts converge to 0. Therefore:

$$\hat{A} \xrightarrow{p} A_0 \;\; as \;\; n \to \infty, \;\; \hat{B} \xrightarrow{p} B_0 \;\; as \;\; n \to \infty, \tag{3.37}$$

i.e., $\hat{A}$ is a consistent estimator of A_0 and $\hat{B}$ is a consistent estimator of B_0. $\square$

3.2 Estimation of the covariance matrix $\mathbf{\Psi}$

According to Section 2, $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{I}, \mathbf{\Psi})$, with $\mathbf{\Psi}$ being a banded matrix of order m. The joint p.d.f. of $\mathbf{X}$ with $\mathbf{M}$ replaced by its estimator $\hat{\mathbf{M}}$ is given by:

$$f_{\mathbf{X}}(\mathbf{X}) = c\,|\mathbf{\Psi}|^{-p/2} \exp\Big(-\frac{1}{2}\,\mathrm{tr}\big(\mathbf{\Psi}^{-1}\,(\mathbf{X}-\hat{\mathbf{M}})'\,(\mathbf{X}-\hat{\mathbf{M}})\big)\Big). \tag{3.38}$$

Here $c = (2\pi)^{-np/2}$ and $\hat{\mathbf{M}} = (\hat{\mathbf{m}}_{ij})$, with $A,\ B$ and ω being replaced by their estimators $\hat{A},\ \hat{B}$ and $\hat{\omega}$ (see (2.10), (3.3) and (3.1)). Below, the symbol $\hat{}$ over $\mathbf{M}$ and $\mathbf{m}$ is omitted to make formulae easier to read.

To estimate the covariance matrix, we start with the fact that if $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{I}, \mathbf{\Psi})$, then $\mathbf{X}' \sim N_{n,p}(\mathbf{M}', \mathbf{\Psi}, \mathbf{I})$, and use the approach presented in Andrushchenko et al. (2008) and Ohlson et al. (2009). We start from the likelihood function as the product of marginal and conditional distributions (see e.g. (C.7)). First, we maximize the first factor in the likelihood function (marginal distribution with unstructured covariance matrix) and estimate the parameters. Those parameters which also appear in the next factor of the likelihood function (conditional part) are replaced by the estimators from the previous part. The estimation proceeds in a similar manner until the parameters of the last factor have been obtained.

If the covariance matrix $\mathbf{\Psi}\colon n \times n$ is banded of order m, then the submatrix $\mathbf{\Psi}_{m+1}\colon (m+1) \times (m+1)$ is unstructured and the marginal distribution is:

$$\mathbf{X}'_{1:m+1} \sim N_{m+1,p}(\mathbf{M}'_{1:m+1}, \mathbf{\Psi}_{m+1}, \mathbf{I}_p). \tag{3.39}$$

The estimators of the elements of such a covariance matrix coincide with the usual MLEs ($k, j \le m+1$):

$$\hat{\psi}_{k,j} = \frac{1}{p}\,(\mathbf{x}_k - \mathbf{m}_k)'(\mathbf{x}_j - \mathbf{m}_j), \quad \hat{\psi}_{k,k} = \frac{1}{p}\,(\mathbf{x}_k - \mathbf{m}_k)'(\mathbf{x}_k - \mathbf{m}_k). \tag{3.40}$$

For $k = m+2, \ldots, n$, let us partition the covariance matrix as follows:

$$\mathbf{\Psi}_k = \begin{pmatrix} \mathbf{\Psi}_{k-1} & \boldsymbol{\psi}_k \\ \boldsymbol{\psi}'_k & \psi_{kk} \end{pmatrix}, \tag{3.41}$$

where

$$\boldsymbol{\psi}_k = (\underbrace{0, \ldots, 0}_{k-m-1}, \underbrace{\psi_{k,k-m}, \ldots, \psi_{k,k-1}}_{m})'. \tag{3.42}$$

Then the conditional distribution can be written as:

$$\mathbf{x}'_{k|1:k-1} \sim N_{1,p}(\mathbf{m}'_{k|1:k-1}, \psi_{k|1:k-1}, \mathbf{I}_p), \tag{3.43}$$

with

$$\begin{aligned}\mathbf{m}_{k|1:k-1} &= \mathbf{m}_k + \boldsymbol{\psi}'_k \boldsymbol{\Psi}_{k-1}^{-1}(\mathbf{X}_{1:k-1} - \mathbf{M}_{1:k-1}),\\ \psi_{k|1:k-1} &= \psi_{kk} - \boldsymbol{\psi}'_k \boldsymbol{\Psi}_{k-1}^{-1} \boldsymbol{\psi}_k.\end{aligned} \tag{3.44}$$

It can be simplified using a formula for the elements of the inverse of a matrix and a definition of the determinant:

$$\begin{aligned}\left(\boldsymbol{\Psi}_k^{-1}\right)_{ij} &= \left((-1)^{i+j}\frac{C_{ij}^{(k)}}{|\boldsymbol{\Psi}_k|}\right)_{ij},\\ |\boldsymbol{\Psi}_k| &= \sum_{i=1}^{k}(-1)^{i+k}\psi_{ki}\, C_{ki}^{(k)} = \psi_{kk}\, C_{kk}^{(k)} + \sum_{i=1}^{k-1}(-1)^{i+k}\psi_{ki}\, C_{ki}^{(k)}\\ &= \psi_{kk}|\boldsymbol{\Psi}_{k-1}| - \sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(-1)^{i+j}\psi_{ki}\, C_{ij}^{(k-1)}\, \psi_{jk}.\end{aligned} \tag{3.45}$$

Here $C_{ij}^{(k)}$ are the minors of the matrix $\boldsymbol{\Psi}_k$ (the determinant of the matrix formed by removing the i^{th} row and the j^{th} column of the matrix $\boldsymbol{\Psi}_k$). It should be noted that (3.45) is valid for any square matrix. Applying these formulae to the banded covariance matrix of order m results in:

$$\begin{aligned}\mathbf{m}_{k|1:k-1} &= \mathbf{m}_k + \sum_{i=1}^{k-1}\sum_{j=k-m}^{k-1}(\mathbf{x}_i - \mathbf{m}_i)\,(-1)^{i+j}\,\frac{C_{ij}^{(k-1)}}{|\boldsymbol{\Psi}_{k-1}|}\,\psi_{jk}\\ &= \mathbf{m}_k + \sum_{i=1}^{k-1}\sum_{j=k-m}^{k-1}(-1)^{i+j}\,\frac{C_{ij}^{(k-1)}}{|\boldsymbol{\Psi}_{k-2}|}\,(\mathbf{x}_i - \mathbf{m}_i)\,\psi_{jk}\,\frac{|\boldsymbol{\Psi}_{k-2}|}{|\boldsymbol{\Psi}_{k-1}|}\\ &= \mathbf{m}_k + (\tilde{\mathbf{X}}_{k-1} - \tilde{\mathbf{M}}_{k-1})\,\boldsymbol{\beta}^{(k-1)} = \tilde{\mathbf{M}}_k + \tilde{\mathbf{X}}_{k-1}\,\boldsymbol{\beta}^{(k-1)},\end{aligned}$$

$$\begin{aligned}\psi_{k|1:k-1} &= \psi_{kk} - \sum_{j=k-m}^{k-1}\sum_{i=k-m}^{k-1}\psi_{ki}\,(-1)^{i+j}\,\frac{C_{ij}^{(k-1)}}{|\boldsymbol{\Psi}_{k-1}|}\,\psi_{jk}\\ &= \frac{1}{|\boldsymbol{\Psi}_{k-1}|}\Big(\psi_{kk}|\boldsymbol{\Psi}_{k-1}| - \sum_{j=k-m}^{k-1}\sum_{i=k-m}^{k-1}\psi_{ki}\,(-1)^{i+j}\,C_{ij}^{(k-1)}\,\psi_{jk}\Big)\\ &= \frac{|\boldsymbol{\Psi}_k|}{|\boldsymbol{\Psi}_{k-1}|},\end{aligned} \tag{3.46}$$

where

$$\tilde{\mathbf{X}}_{k-1} = (\tilde{\mathbf{x}}_{k-m}^{(k-1)} \colon \ldots \colon \tilde{\mathbf{x}}_{k-1}^{(k-1)}), \quad \tilde{\mathbf{x}}_j^{(k-1)} = \sum_{i=1}^{k-1} (-1)^{i+j} \frac{C_{ij}^{(k-1)}}{|\boldsymbol{\Psi}_{k-2}|} \mathbf{x}_i,$$

$$\tilde{\mathbf{M}}_{k-1} = (\tilde{\mathbf{m}}_{k-m}^{(k-1)} \colon \ldots \colon \tilde{\mathbf{m}}_{k-1}^{(k-1)}), \quad \tilde{\mathbf{m}}_j^{(k-1)} = \sum_{i=1}^{k-1} (-1)^{i+j} \frac{C_{ij}^{(k-1)}}{|\boldsymbol{\Psi}_{k-2}|} \mathbf{m}_i,$$

$$\boldsymbol{\beta}^{(k-1)} = (\beta_{k-m}^{(k-1)}, \ \ldots, \ \beta_{k-1}^{(k-1)})', \quad \beta_j^{(k-1)} = \psi_{jk} \frac{|\boldsymbol{\Psi}_{k-2}|}{|\boldsymbol{\Psi}_{k-1}|}, \tag{3.47}$$

and $j = k-m, \ldots, k-1$.

For $k = m+2, \ldots, n$, the likelihood function is given by

$$\begin{aligned} L_k &= L_{m+1} \prod_{i=m+2}^{k} L_{i|1:i-1} \\ &= L_{m+1} \prod_{i=m+2}^{k} c \left(\psi_{i|1:i-1}\right)^{-p/2} \exp\Big(- \frac{1}{2\psi_{i|1:i-1}} (\mathbf{x}_i - \mathbf{m}_{i|1:i-1})' (\ldots) \Big), \end{aligned} \tag{3.48}$$

and the likelihood has to be maximized with respect to $(2k-m)(m+1)/2$ unknown parameters, namely, $(m+1)(m+2)/2$ parameters for unstructured covariance submatrix, and $(k-m-1)(m+1)$ parameters in $\boldsymbol{\psi}_k$ and ψ_{kk}.

Instead of doing this directly, we use the sequential approach suggested above. We start with $k = m+2$ and represent the likelihood as (3.48). The estimation starts with the unstructured covariance matrix, where the estimators are given by (3.40), and proceeds with the next factor, inserting estimators from the previous factor. For $k = m+3, \ldots, n$, the likelihood (3.48) can be presented as

$$L_k = L_{k-1} c \left(\psi_{k|1:k-1}\right)^{-p/2} \exp\Big(- \frac{1}{2\psi_{k|1:k-1}} (\mathbf{x}_k - \mathbf{m}_{k|1:k-1})' (\ldots) \Big), \tag{3.49}$$

and the parameters in L_{k-1} are already estimated. The estimation proceeds with the next factor, each time inserting estimators from the previous factor until the parameters of the last factor have been obtained. This means that for any $k = m+2, \ldots, n$, we need to estimate only $m+1$ unknown parameters ($\beta_j^{(k-1)}$, $j = k-m, \ldots, k-1$, and $\psi_{k|1:k-1}$). This results in:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(k-1)} &= \big((\widehat{\mathbf{X}}_{k-1} - \widehat{\mathbf{M}}_{k-1})' (\widehat{\mathbf{X}}_{k-1} - \widehat{\mathbf{M}}_{k-1})\big)^{-1} (\widehat{\mathbf{X}}_{k-1} - \widehat{\mathbf{M}}_{k-1})' (\mathbf{x}_k - \mathbf{m}_k), \\ \hat{\psi}_{k|1:k-1} &= \frac{1}{p} \left(\mathbf{x}_k - \mathbf{m}_k - (\widehat{\mathbf{X}}_k - \widehat{\mathbf{M}}_k)\, \hat{\boldsymbol{\beta}}^{(k-1)}\right)' (\ldots), \end{aligned} \tag{3.50}$$

where

$$\widehat{\mathbf{X}}_{k-1} = (\widehat{\mathbf{x}}_{k-m}^{(k-1)} : \ldots : \widehat{\mathbf{x}}_{k-1}^{(k-1)}), \quad \widehat{\mathbf{x}}_j^{(k-1)} = \sum_{i=1}^{k-1} (-1)^{i+j} \frac{\hat{C}_{ij}^{(k-1)}}{|\hat{\mathbf{\Psi}}_{k-2}|} \mathbf{x}_i,$$
$$\widehat{\mathbf{M}}_{k-1} = (\widehat{\mathbf{m}}_{k-m}^{(k-1)} : \ldots : \widehat{\mathbf{m}}_{k-1}^{(k-1)}), \quad \widehat{\mathbf{m}}_j^{(k-1)} = \sum_{i=1}^{k-1} (-1)^{i+j} \frac{\hat{C}_{ij}^{(k-1)}}{|\hat{\mathbf{\Psi}}_{k-2}|} \mathbf{m}_i, \tag{3.51}$$

The estimators of initial parameters can be carried out from here. In general, they can be written as:

$$\hat{\psi}_{k,j} = \hat{\beta}_j^{(k-1)} \, \hat{\psi}_{k-1|1:k-2}, \tag{3.52}$$
$$\hat{\psi}_{kk} = \hat{\psi}_{k|1:k-1} + \hat{\boldsymbol{\psi}}_k' \hat{\mathbf{\Psi}}_{k-1}^{-1} \hat{\boldsymbol{\psi}}_k, \tag{3.53}$$

with $\hat{\boldsymbol{\psi}}_k$ being given by (3.42) with $\psi_{k,j}$ replaced by their estimators $\hat{\psi}_{k,j}$, (3.52).

4 Conclusions

In this paper, we have presented a way of dealing with a multivariate time series that contains a harmonic component, under the condition that covariance structure is banded of order m. An algorithm consists of first estimating the harmonic structure and then using these estimates in estimating the banded covariance matrix. Our analysis of the properties of estimators of the harmonic part revealed that the estimators of parameters describing the harmonic structure are all consistent. Explicit analytical estimators for the elements of the banded covariance matrix are presented.

The main contribution of this article is finding the reasonable explicit estimators for both the mean (harmonic structure) and the covariance matrix.

Acknowledgements

The work of Zhanna Andrushchenko was supported by the Swedish Research Council, VR 621-2002-5578.

The author is grateful to Prof. Dietrich von Rosen for discussions and suggestions during the work with this manuscript.

References

Andrushchenko, Z., Ohlson, M., and von Rosen, D. (2008). Estimation of banded covariance matrices in a multivariate normal distribution. Centre of Biostochastics, Swedish University of Agricultural Sciences. Research Report 2008:2.

Berk, K. (1973). A central limit theorem for m-dependent random variables with unbounded m. *The Annals of Probability*, 1:352–354.

Bloomfield, P., Hurd, H. L., and Lund, R. B. (1994). Periodic correlation in stratospheric ozone data. *Journal of Time Series Analysis*, 15:127–150.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer, New York, 2nd edition.

Cocke, W. J. (1972). Central limit theorems for sums of dependent vector variables. *The Annals of Mathematical Statistics*, 43:968–976.

Cox, D. and Hinkley, D. (1974). *Theoretical statistics*. Chapman and Hall, London.

Diananda, P. H. (1954). The central limit theorems for m-dependent variables asymptotically stationary to second order. *Proc. Cambridge Philos. Soc.*, 50:287–292.

Doran, H. E. and Quilkey, J. J. (1972). Harmonic analysis of seasonal data: some important properties. *Amer. J. of Agricultural Economics*, 54:646–651.

Ferguson, T. S. (1996). *A course in large sample theory*. Chapman and Hall, London.

Fisher, R. A. (1929). Tests of significance in harmonic analysis. *Proc of the Royal Society of London, Series A*, 125:54–59.

Hannan, E. J. (1973). The estimation of frequency. *J. Appl. Prob.*, 10:510–519.

Hartley, H. O. (1949). Tests of significance in harmonic analysis. *Biometrika*, 36:194–201.

Hoeffding, W. and Robbins, H. (1948). The central limit theorem for dependent random variables. *Duke Math. J.*, 15:773–780.

Kollo, T. and von Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices*. Springer, Dordrecht.

MacDonald, G. J. (1989). Spectral analysis of time series generated by nonlinear processes. *Reviews of Geophysics*, 27:449–469.

Ohlson, M., Andrushchenko, Z., and von Rosen, D. (2009). Explicit estimators under m-dependence for a multivariate normal distribution. *Annals of the Institute of Statistical Mathematics*. (*accepted*).

Priestley, M. B. (1962a). The analysis of stationary processes with mixed spectra - l. *Journal of the Royal Statistical Society, Series B*, 24:215–233.

Priestley, M. B. (1962b). The analysis of stationary processes with mixed spectra - ll. *Journal of the Royal Statistical Society, Series B*, 24:511–529.

Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26-day period of meteorological phenomena. *Terrestrial Magnetism and Atmospheric Electricity*, 3:13–41.

Shannon, C. E. (1949). Communication in the presence of noise. *Proc. Institute of Radio Engineers*, 37:10–21. Reprint as classic paper in: Proc. IEEE, Vol. 86, No. 2, Feb 1998.

Stoica, P. and Moses, R. (1997). *Introduction to Spectral Analysis*. Prentice Hall.

Tong, Y. L. (1976). Parameter estimation in studying circadian rhythms. *Biometrics*, 32:85–94.

Walker, A. M. (1971). On the estimation of a harmonic component in a time series with stationary independent residuals. *Biometrika*, 58:21–36.

Whittle, P. (1952). The simulttaneous estimation of a time series harmonic component and covariance structure. *Trabajos de Estadistica*, 3:43–57.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57:348–368.

A Classical harmonic regression and periodogram

In the statistical analysis of time series, Fourier methods are used to discover and analyze the regularity or periodicity in data. The most popular way of performing frequency analysis has been to apply the fast Fourier transform (FFT), see Appendix B, to the data. FFT allows the separation of various rhythms and an estimation of their frequencies independently of each other, a difficult task to perform visually if several rhythmic activities occur simultaneously.

The basis of frequency domain (Fourier) analysis for time series is the spectral representation theorem for stationary processes (see Brockwell and Davis (1991, Sect.4) for details). Roughly, the theorem says that one may think of a stationary time series as being formed by a general mean μ, a harmonic component consisting of the sum of finite number of sinusoids with angular frequencies ω_k and a noise component ε. If the sampling times are taken to be equally spaced, so that $t = 1, \ldots, n$ and $x(t) = x_t$, the model becomes:

$$x_t = \mu + \sum_{k=1}^{q} \big(A_k \cos(\omega_k t) + B_k \sin(\omega_k t)\big) + \varepsilon, \;\; 0 \leq \omega_k \leq \pi, \tag{A.1}$$

where $\omega_k, \;\; k = 1, ..., q$ are different angular frequencies, and A_k and B_k are mutually uncorrelated, mean-zero random variables. Ideally, q is small and the ω_k are well separated. The reason for restricting ω_k to the range $(0, \pi)$ is that for a discrete process measured at unit intervals of time, the variation at frequencies higher than π cannot be distinguished from variation at a corresponding frequency in $(0, \pi)$. The frequency $\omega = \pi$ is called the Nyquist frequency. For a discrete process measured at intervals of time of length Δt, the Nyquist frequency is $\pi/\Delta t$. In general, the Nyquist frequency is half the sampling frequency of a discrete signal processing system. The Nyquist-Shannon sampling theorem (Shannon, 1949) states that perfect reconstruction of a signal is possible when the Nyquist frequency exceeds the highest frequency of the signal being sampled. If lower sampling rates are used, the original signal may not be completely recoverable from the sampled signal.

In the case when all the frequencies are known, the model (A.1) becomes linear in parameters $\boldsymbol{\theta} = (\mu, \;\; A_1, \;\; B_1, \ldots, \;\; A_q, \;\; B_q)'$. Thus, the harmonic model is a special case of the classical linear regression model. Therefore the method of ordinary LS produces asymptotically best linear unbiased estimators, regardless of the covariance structure of $\boldsymbol{\varepsilon}$. When the maximum period is known and data cover few replicates of this fundamental period, then LS

estimators (LSEs) become

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x_t, \;\; \hat{A}_k = \frac{2}{n}\sum_{t=1}^{n} x_t \cos(\omega_k t), \;\; \hat{B}_k = \frac{2}{n}\sum_{t=1}^{n} x_t \sin(\omega_k t), \;\; k = 1, \ldots, q. \tag{A.2}$$

If the underlying frequencies are unknown, Schuster's periodogram can be used to estimate them. Schuster's periodogram is based on a model of a mean-zero process x_t, which has a harmonic component consisting of the sum of finite number (say, q) of sinusoids with unknown angular frequencies ω_k and a noise component ε:

$$x_t = \sum_{k=1}^{q} \big(A_k \cos(\omega_k t) + B_k \sin(\omega_k t)\big) + \varepsilon, \tag{A.3}$$

where ε is a purely random process that has a vanishing mean and constant variance, $E[\varepsilon] = 0$, $E[\varepsilon^2] = v < \infty$. Given n observations of x_t, the problem is to determine the set of unknown parameters q, A_k, B_k, ω_k, v $(k = 1, \ldots, q)$.

Schuster's periodogram, $I(\omega)$, is defined as:

$$I(\omega) = \frac{2}{n}\left(\Big(\sum x_t \cos(\omega t)\Big)^2 + \Big(\sum x_t \sin(\omega t)\Big)^2\right) = \frac{2}{n}\Big|\sum_{t=1}^{n} x_t \exp(i\omega t)\Big|^2. \tag{A.4}$$

If the observed data are indeed periodic, the plot of the periodogram function exhibits large positive values (the squares of the amplitudes associated with the frequencies) at the true values of the underlying frequencies present in the data, and at all the other points it is close to zero.

It is clear that the model (A.3) is nonlinear. It can be solved numerically using the nonlinear least squares (NLS) method (see e.g. Stoica and Moses (1997)).

The problem of estimating all the parameters in (A.3) was first dealt with by Whittle (1952). He used a method of estimation that was approximately equivalent to the LS. His argument is illustrated by considering a process with a single sinusoid, $q = 1$,

$$x_t = A\cos(\omega t) + B\sin(\omega t) + \varepsilon. \tag{A.5}$$

One forms the regression sum of squares, $S(A, B, \omega)$, as if ω were known:

$$S(A, B, \omega) = \sum_{t=1}^{n} \big(x_t - A\cos(\omega t) - B\sin(\omega t)\big)^2, \tag{A.6}$$

Then the "LS" estimators of A and B are obtained by the equations:

$$\begin{aligned}\sum x_t \cos(\omega t) - \tilde{A}\sum \cos^2(\omega t) - \tilde{B}\sum \sin(\omega t)\cos(\omega t) = 0,\\ \sum x_t \sin(\omega t) - \tilde{A}\sum \sin(\omega t)\cos(\omega t) - \tilde{B}\sum \sin^2(\omega t) = 0,\end{aligned} \tag{A.7}$$

and, since $\sum \sin(\omega t)\cos(\omega t) \approx 0$ and $\sum \cos^2(\omega t) \approx \sum \sin^2(\omega t) \approx n/2$, the solutions become:

$$\tilde{A} = \frac{2}{n}\sum_{t=1}^{n} x_t \cos(\omega t), \quad \tilde{B} = \frac{2}{n}\sum_{t=1}^{n} x_t \sin(\omega t). \tag{A.8}$$

Now, the angular frequency, ω, is also an unknown parameter the least squares estimate of which, $\hat{\omega}$, is yielded by the equation

$$\hat{\omega} = \arg\min_{\omega} \big(S(\tilde{A}, \tilde{B}, \omega)\big). \tag{A.9}$$

With some rearrangement, it is found that $S(\tilde{A}, \tilde{B}, \omega) = \sum x_t^2 - \frac{n}{2}(\tilde{A}^2 + \tilde{B}^2)$, so that the estimation equation for ω is in effect:

$$\hat{\omega} = \arg\max_{\omega} \Big(\frac{n}{2}\big(\tilde{A}^2 + \tilde{B}^2\big)\Big) = \arg\max_{\omega} \Big(\big(\sum x_t \cos(\omega t)\big)^2 + \big(\sum x_t \sin(\omega t)\big)^2\Big). \tag{A.10}$$

Thus, $\hat{\omega}$ corresponds to the greatest ordinate of the periodogram (A.4), and the magnitude of this ordinate provides the least squares estimate of $A^2 + B^2$, the amplitude of the component. Whittle (1952) argued that the periodogram gives the best estimates of the frequencies in the LS sense. On practise, the most popular way to estimate the frequency ω is to apply the (radix-2) FFT algorithm directly to the data (or to the data with zero padding).

Walker (1971) gives an rigorous proof of Whittle's statements concerning the asymptotic distribution of the estimators for the case of independent errors. He used a method of estimation that was approximately equivalent to the LS, becoming approximately the method of ML estimation when ε has a normal distribution so that x_t becomes a normal or Gaussian process.

If the errors are normally distributed, the log-likelihood function of the observation $x_1, x_2, \ldots, x_n$ is:

$$L(A, B, \omega, \sigma^2) = -\frac{n}{2}\log(2\pi v) - \frac{1}{2v}\, S(A, B, \omega), \tag{A.11}$$

where S is the regression sum of squares:

$$S(A, B, \omega) = \sum_{t=1}^{n} \big(x_t - A\cos(\omega t) - B\sin(\omega t)\big)^2, \tag{A.12}$$

The estimators of A, B and ω are thus obtained by minimizing the residual sum of squares S, and the estimator of v is equal to the minimum sum of squares divided by n.

Calculation of the parameters minimizing S is simplified if the sampling times are taken to be equally spaced, so that $t = 1, 2, \ldots, n$. It is also easier to proceed in this case by introducing a new function:

$$U(A, B, \omega) = \sum_{t=1}^{n} x_t^2 - 2\sum_{t=1}^{n} x_t\big(A\cos(\omega t) + B\sin(\omega t)\big) + \frac{n}{2}(A^2 + B^2). \quad \text{(A.13)}$$

Since

$$S(A, B, \omega) - U(A, B, \omega) = \frac{1}{2}\sum_{t=1}^{n}\Big((A^2 - B^2)\cos(2\omega t) + 2AB\sin(2\omega t)\Big), \quad \text{(A.14)}$$

which is $O(1)$ as $n \to \infty$ (provided $\omega \neq 0, \pi$), one can expect the effect on the estimators to be negligible for large n provided that ω_0, the true value of ω, is not equal to 0 or π; so one restricts $0 < \omega_0 < \pi$. Under these conditions, minimization of U is equivalent to minimizing S.

If the estimates for A, B, and ω are denoted by $\hat{A}, \hat{B}$, and $\hat{\omega}$, minimizing U yields

$$\hat{A} = \frac{2}{n}\sum_{t=1}^{n} x_t\cos(\hat{\omega}t), \quad \hat{B} = \frac{2}{n}\sum_{t=1}^{n} x_t\sin(\hat{\omega}t), \quad \text{(A.15)}$$

provided the estimated frequency $\hat{\omega}$ is such that:

$$I(\hat{\omega}) = \max\big(I(\omega)\big), \quad \text{(A.16)}$$

where $I(\omega)$ is Schuster's periodogram (A.4). Also,

$$\hat{v} = \frac{1}{n}\Big(\sum_{t=1}^{n} x_t^2 - I(\hat{\omega})\Big). \quad \text{(A.17)}$$

Walker (1971) showed that the estimators $\hat{A}, \hat{B}, \hat{\omega}$ and $\hat{v}$ are all consistent as $n \to \infty$.

The case in which the signal consists of a number of sinusoids is more complex. For the case of q sinusoids, one gets the unsurprising result that minimization drives all the $\hat{\omega}_k$, $k = 1, \ldots, q$, equal to the frequency at which $I(\omega)$ obtains its maximum value. To prevent this, some condition such as:

$$\min(|\hat{\omega}_k - \hat{\omega}_l|) \geq 2\pi/n \quad \text{(A.18)}$$

should be placed on the frequencies.

A possible method is to proceed sequentially, first calculating the maximum value of the periodogram, using e.g. FFT. Alternatively, one could use a coarse FFT and follow that by a Newton-Raphson search for a maximum. Once $\hat{\omega}_1$ is identified, estimates of the coefficients $\hat{A}_1$ and $\hat{B}_1$ are obtained by:

$$\hat{A}_1 = \frac{2}{n}\sum_{t=1}^{n} x_t \cos(\hat{\omega}_1 t), \quad \hat{B}_1 = \frac{2}{n}\sum_{t=1}^{n} x_t \sin(\hat{\omega}_1 t). \tag{A.19}$$

Term-by-term subtraction forms a new series x'_t,

$$x'_t = x_t - \hat{A}_1 \cos(\hat{\omega}_1 t) - \hat{B}_1 \sin(\hat{\omega}_1 t). \tag{A.20}$$

The process is repeated successively, using subsequent estimates of $\hat{\omega}_k$, $k = 2, \ldots$.

B Fourier transforms

The Fourier Transform (FT) is a mathematical procedure used to transform a signal defined in the time domain into one defined in the frequency domain.

The **integral Fourier transform** operates on continuous functions. Given a function $x(t)$ of a real variable t, the FT of $x(t)$ is defined as:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) \exp(-i\omega t) dt, \tag{B.1}$$

provided the integral exists for every real ω. A sufficient condition for $X(\omega)$ to exist is $\int_{-\infty}^{\infty} |x(t)| dt < \infty$. If (B.1) is regarded as an integral equation for $x(t)$ given $X(\omega)$, then:

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega) \exp(i\omega t) d\omega, \tag{B.2}$$

and $x(t)$ is called the inverse FT of $X(\omega)$. The two functions $x(t)$ and $X(\omega)$ are commonly called a FT pair.

In practise, the FT cannot be applied directly if measurements are taken only at discrete times. A modified form of the FT, known as the **Discrete Fourier Transform** (DFT), is used in the case of sampled (discrete) signals. When $x(t)$ is only defined for integer values of $t = 0, \ldots, N-1$, then:

$$X(\omega) = \sum_{t} x(t) \exp(-i\omega t), \quad -\pi \leq \omega \leq \pi \tag{B.3}$$

is the DFT of $x(t)$. The inverse transform is:

$$x(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) \exp(i\omega t) d\omega, \quad t = 0, \ldots, N-1. \tag{B.4}$$

When the DFT is applied to a discrete signal, the result is a set of sine and cosine coefficients. When sine and cosine waves of appropriate frequencies are multiplied by these coefficients and then added together, the original signal waveform is exactly reconstructed. The procedure by which the sine and cosine coefficients are calculated is straightforward in principle, although in practise it requires a great deal of computation.

The **Fast Fourier Transform** (FFT) is a class of special algorithms which implement the DFT with considerable savings in computational time. Functionally, the FFT decomposes the set of data to be transformed into a series of smaller data sets. Then, it decomposes those smaller sets into even smaller sets. At each stage of processing, the results of the previous stage are combined in special way. Finally, it calculates the DFT of each small data set.

By far the most common FFT is the Cooley-Tukey algorithm. The most well-known use of it is to divide the set of size N into two pieces of size $N/2$ at each step, and is therefore limited to power-of-two sizes, but any factorization can be used in general. These are called the radix-2 and mixed-radix cases, respectively. While it is possible to develop FFT algorithms that work with any number of points, maximum efficiency of computation is obtained by constraining the number of time points to be an integer power of 2.

The ratio between a DFT computation and a FFT one for the same N is proportional to $N/\log_2(N)$. In many applications, N is not a power of 2 and hence the previously described radix-2 FFT algorithm cannot be applied directly. However, one can increase the length of the given sequence by means of zero padding, $\{x(1), \ldots, x(N), 0, 0, \ldots\}$ until the length of the so-obtained sequence is a power of 2.

Other Fourier transforms and series:

The **finite Fourier transform** converts one finite sequence of coefficients into another sequence of the same length. The transform and its inverse are:

$$X_k = \sum_{j=0}^{N-1} x_j \exp\Big(- \frac{i2\pi}{N} \, kj \Big), \quad k = 0, \ldots, N-1, \tag{B.5}$$

$$x_j = \frac{1}{N} \sum_{k=0}^{N-1} X_k \exp\Big(\frac{i2\pi}{N} \, kj \Big), \quad j = 0, \ldots, N-1. \tag{B.6}$$

A **Fourier series** converts a periodic function $f(x)$, with a period L, into

an infinite sequence of Fourier coefficients c_n:

$$c_n = \frac{1}{L}\int_{-L/2}^{L/2} f(x)\exp(-inx)dx, \quad n = 0, 1, \ldots. \tag{B.7}$$

With these coefficients, the complex form of the Fourier series is:

$$f(x) = \sum_{n=0}^{\infty} c_n \exp(inx). \tag{B.8}$$

The integral FT involves only integrals. The finite FT involves only finite sums of coefficients. Fourier series and the DFT involve both integrals and sequences. It is possible to turn any of the transforms into any of the others by taking limits or restricting domains.

C Auxiliary results

C.1 Matrix normal distribution

Let the observation matrix be $\mathbf{X} : p \times n$.

Definition C.1 (**Definition 2.2.1, Kollo and von Rosen (2005)**). *Let $\mathbf{\Sigma} = \boldsymbol{\sigma}\boldsymbol{\sigma}'$ and $\mathbf{\Psi} = \boldsymbol{\psi}\boldsymbol{\psi}'$, where $\boldsymbol{\sigma} : p \times r$ and $\boldsymbol{\psi} : n \times s$. A matrix $\mathbf{X} : p \times n$ is said to be matrix normally distributed with parameters $\mathbf{M}$, $\mathbf{\Sigma}_p$, $\mathbf{\Psi}_n$, if it has the same distribution as $\mathbf{M} + \boldsymbol{\sigma}\mathbf{U}\boldsymbol{\psi}'$, where $\mathbf{M} : p \times n$ is non-random and $\mathbf{U} : r \times s$ consists of s i.i.d. $N_r(\mathbf{0}, \mathbf{I}_n)$ vectors $\mathbf{U}_i$, $i = 1, 2, \ldots, s$. If $\mathbf{X} : p \times n$ is matrix normally distributed, this will be denoted $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{\Sigma}_p, \mathbf{\Psi}_n)$.*

We note that $\mathbf{\Sigma}$ describes a covariance between rows ("spatial" covariance structure), and $\mathbf{\Psi}$ describes a covariance between columns ("temporal" covariance structure). Then, if $\mathbf{\Psi} = \mathbf{I}_n$, the columns of $\mathbf{X}$ are independently distributed; if $\mathbf{\Sigma} = \mathbf{I}_p$, the rows of $\mathbf{X}$ are independent. As usual, we assume that both the covariance matrices are positive definite.

The joint probability density function of $\mathbf{X}$ is given by:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{X}) = {} & (2\pi)^{-np/2}\,|\mathbf{\Sigma}_p|^{-n/2}\,|\mathbf{\Psi}_n|^{-p/2} \\ & \times \exp\Big(-\frac{1}{2}\operatorname{tr}\big(\mathbf{\Sigma}_p^{-1}\,(\mathbf{X}-\mathbf{M})\,\mathbf{\Psi}_n^{-1}\,(\mathbf{X}-\mathbf{M})'\big)\Big). \end{aligned} \tag{C.1}$$

Let $\mathbf{X}$, $\mathbf{M}$, $\mathbf{\Sigma}_p$, $\mathbf{\Psi}$ be partitioned as:

$$\begin{aligned}
\mathbf{X} &= \begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{pmatrix} : \begin{pmatrix} r \times s & r \times (n-s) \\ (p-r) \times s & (p-r) \times (n-s) \end{pmatrix}, \\
\mathbf{M} &= \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix} : \begin{pmatrix} r \times s & r \times (n-s) \\ (p-r) \times s & (p-r) \times (n-s) \end{pmatrix}, \\
\mathbf{\Sigma}_p &= \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} : \begin{pmatrix} r \times r & r \times (p-r) \\ (p-r) \times r & (p-r) \times (p-r) \end{pmatrix}, \\
\mathbf{\Psi}_n &= \begin{pmatrix} \mathbf{\Psi}_{11} & \mathbf{\Psi}_{12} \\ \mathbf{\Psi}_{21} & \mathbf{\Psi}_{22} \end{pmatrix} : \begin{pmatrix} s \times s & s \times (n-s) \\ (n-s) \times s & (n-s) \times (n-s) \end{pmatrix},
\end{aligned} \tag{C.2}$$

and

$$\begin{aligned}
\mathbf{X}_{\bullet 1} &= \begin{pmatrix} \mathbf{X}_{11} \\ \mathbf{X}_{21} \end{pmatrix}, \ \mathbf{X}_{\bullet 2} = \begin{pmatrix} \mathbf{X}_{12} \\ \mathbf{X}_{22} \end{pmatrix}, \\
\mathbf{X}_{1\bullet} &= \begin{pmatrix} \mathbf{X}_{11} : \mathbf{X}_{12} \end{pmatrix}, \ \mathbf{X}_{2\bullet} = \begin{pmatrix} \mathbf{X}_{21} : \mathbf{X}_{22} \end{pmatrix}, \\
\mathbf{M}_{\bullet 1} &= \begin{pmatrix} \mathbf{M}_{11} \\ \mathbf{M}_{21} \end{pmatrix}, \ \mathbf{M}_{\bullet 2} = \begin{pmatrix} \mathbf{M}_{12} \\ \mathbf{M}_{22} \end{pmatrix}, \\
\mathbf{M}_{1\bullet} &= \begin{pmatrix} \mathbf{M}_{11} : \mathbf{M}_{12} \end{pmatrix}, \ \mathbf{M}_{2\bullet} = \begin{pmatrix} \mathbf{M}_{21} : \mathbf{M}_{22} \end{pmatrix}.
\end{aligned} \tag{C.3}$$

Lemma C.1 (**Theorem 2.2.5, Kollo and von Rosen (2005)**). *Suppose that* $\mathbf{\Sigma}_{22}^{-1}$ *and* $\mathbf{\Psi}_{22}^{-1}$ *exist. Then we have the conditional distributions:*

$$\begin{aligned}
\mathbf{X}_{\bullet 1} | \mathbf{X}_{\bullet 2} &\sim N_{p,s}\left(\mathbf{M}_{\bullet 1} + (\mathbf{X}_{\bullet 2} - \mathbf{M}_{\bullet 2})\mathbf{\Psi}_{22}^{-1}\mathbf{\Psi}_{21},\ \mathbf{\Sigma},\ \mathbf{\Psi}_{1|2}\right), \\
\mathbf{X}_{1\bullet} | \mathbf{X}_{2\bullet} &\sim N_{r,n}\left(\mathbf{M}_{1\bullet} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{X}_{2\bullet} - \mathbf{M}_{2\bullet}),\ \ \mathbf{\Sigma}_{1|2},\ \mathbf{\Psi}\right).
\end{aligned} \tag{C.4}$$

Obviously, if $\mathbf{\Sigma}_{11}^{-1}$ and $\mathbf{\Psi}_{11}^{-1}$ exist, then

$$\begin{aligned}
\mathbf{X}_{\bullet 2} | \mathbf{X}_{\bullet 1} &\sim N_{p,n-s}\left(\mathbf{M}_{\bullet 2} + (\mathbf{X}_{\bullet 1} - \mathbf{M}_{\bullet 1})\mathbf{\Psi}_{11}^{-1}\mathbf{\Psi}_{12},\ \mathbf{\Sigma},\ \mathbf{\Psi}_{2|1}\right), \\
\mathbf{X}_{2\bullet} | \mathbf{X}_{1\bullet} &\sim N_{p-r,n}\left(\mathbf{M}_{2\bullet} + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}(\mathbf{X}_{1\bullet} - \mathbf{M}_{1\bullet}),\ \ \mathbf{\Sigma}_{2|1},\ \mathbf{\Psi}\right).
\end{aligned} \tag{C.5}$$

Here

$$\begin{aligned}
\mathbf{\Sigma}_{1|2} &= \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}, \ \ \mathbf{\Psi}_{1|2} = \mathbf{\Psi}_{11} - \mathbf{\Psi}_{12}\mathbf{\Psi}_{22}^{-1}\mathbf{\Psi}_{21}, \\
\mathbf{\Sigma}_{2|1} &= \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}, \ \ \mathbf{\Psi}_{2|1} = \mathbf{\Psi}_{22} - \mathbf{\Psi}_{21}\mathbf{\Psi}_{11}^{-1}\mathbf{\Psi}_{12},
\end{aligned} \tag{C.6}$$

and we adopt the notation of Kollo and von Rosen (2005, Chapter 2.2).

Now the joint probability density function of $\mathbf{X}$ can be written as the product of conditional and marginal distributions:

$$f_{\mathbf{X}}(\mathbf{X}) = f(\mathbf{X}_{\bullet 1})\ f(\mathbf{X}_{\bullet 2} \mid \mathbf{X}_{\bullet 1}) = f(\mathbf{X}_{1\bullet})\ f(\mathbf{X}_{2\bullet} \mid \mathbf{X}_{1\bullet}). \tag{C.7}$$

The probability density function considered as a function of the parameters $\mathbf{M}$, $\mathbf{\Sigma}_p$ and $\mathbf{\Psi}_n$ (for fixed observed $\mathbf{X}$) will serve as the likelihood function.

C.2 $O(\ldots)$ and $o(\ldots)$

The notations $O(\ldots)$ and $o(\ldots)$ are used to describe the limiting behavior of a function for very large arguments. Suppose $f(x)$ and $g(x)$ are two functions defined on some subset of the real numbers. One says $f(x) = O(g(x)),\ \ as\ \ x \to \infty$, if there exist a positive real number C and a real number x_0 such that

$$|f(x)| \leq C|(g(x)|\ \ for\ \ all\ \ x > x_0. \tag{C.8}$$

The relation $f(x) \in\ o(g(x))$ means that $f(x)$ becomes insignificant relative to $g(x)$ as x approaches infinity. Formally, it states that:

$$\lim \frac{f(x)}{g(x)} = 0,\ \ as\ \ x \to \infty. \tag{C.9}$$

Similarly, the notations $O_p(\ldots)$ and $o_p(\ldots)$ are used to describe the limiting behavior of an r.v. as the number of observations tends to infinity ($n \to \infty$): $Z_n = O_p(n^\alpha)$ if for all $\epsilon > 0$ there exist $C_\epsilon > 0$ such that:

$$P\Big(|Z_n| \leq C_\epsilon n^\alpha\Big) \geq 1 - \epsilon\ \ for\ \ all\ \ n; \tag{C.10}$$

$Z_n = o_p(n^\alpha)$ means an r.v. such that for any $\epsilon > 0$:

$$P\Big(n^{-\alpha}|Z_n| > \epsilon\Big) = 0\ \ as\ \ n \to \infty. \tag{C.11}$$

The notation $o_p(1)$ is useful in showing the consistency of estimators (see e.g. Cox and Hinkley (1974, Ch.9)): an estimator $\hat{\theta}$ is consistent for θ if:

$$\hat{\theta} = \theta + o_p(1) \iff P\big(|\hat{\theta} - \theta| > \epsilon\big) \to 0\ \ as\ \ n \to \infty \iff \hat{\theta} \xrightarrow{p} \theta. \tag{C.12}$$

C.3 Some important inequalities

Inequality C.1 (**The Cauchy-Schwarz inequality**). *Suppose that X and Y have finite variances, and $\|X\|_r = \big(EX^r\big)^{1/r}$, $r > 0$, is a norm. Then:*

$$|E\,XY| \leq E|XY| \leq \|X\|_2 \cdot \|Y\|_2 = \big(EX^2EY^2\big)^{1/2}. \tag{C.13}$$

Inequality C.2 (**The Markov inequality**). *Suppose that $E|X|^r < \infty$ for some $r > 0$, and let $x > 0$. Then:*

$$P(|X| \geq x) \leq \frac{E|X|^r}{x^r}. \tag{C.14}$$

Inequality C.3 (**The triangle inequality**). *For x, $y \in \mathbb{R}^n$, with $\mathbb{R}^n$ being a real number space, the standard triangle inequality states that:*

$$|x + y| \leq |x| + |y|. \tag{C.15}$$