DISSERTATION

# MULTI-CHANNEL FACTOR ANALYSIS: PROPERTIES, EXTENSIONS, AND APPLICATIONS

Submitted by

Gray Stanton

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2024

Doctoral Committee:

Advisor: Haonan Wang
Co-Advisor: Louis Scharf

Piotr Kokoszka
Tianying Wang
Jie Luo

ABSTRACT

MULTI-CHANNEL FACTOR ANALYSIS: PROPERTIES, EXTENSIONS, AND APPLICATIONS

Multi-channel Factor Analysis (MFA) extends factor analysis to the multi-channel or multi-view setting, where latent *common factors* influence all channels while *distinct factors* are specific to individual channels. The within- and across-channel covariance is determined by a *low-rank* matrix, a *block-diagonal* matrix with low-rank blocks, and a *diagonal* matrix, which provides a parsimonious model for both covariances. MFA and related multi-channel methods for data fusion are discussed in Chapter 1. Under conditions on the channel sizes and factor numbers, the results of Chapter 2 show that the generic global identifiability of the aforementioned covariance matrices can be guaranteed *a priori*, and the estimators obtained by maximizing a Gaussian likelihood are shown to be consistent and asymptotically normal even under misspecification. To handle temporal correlation in the latent factors, Chapter 3 introduces Multi-channel Factor Spectral Analysis (MFSA). Results for the identifiability and parameterization properties of the MFSA spectral density model are derived, and a Majorization-Minimization procedure to optimize the Whittle pseudo-likelihood is designed to estimate the MFSA parameters. A simulation study is conducted to explore how temporal correlations in the latent factors affect estimation, and it is demonstrated that MFSA significantly outperforms MFA when the factor series are highly autocorrelated. In Chapter 4, a locally stationary joint multivariate Gaussian process with MFA-type cross-sectional covariance is developed to model multi-vehicle trajectories in a highway environment. A dynamic model-based clustering procedure is designed to partition cohorts of nearby vehicles into *pods* based on the stability of the intra-pod relative vehicle configuration. The performance of this procedure is illustrated by its application to the Next GENeration SIMulation dataset of vehicle trajectories on U.S. Highway 101.

# ACKNOWLEDGEMENTS

Many of the ideas in this dissertation came from the fruitful conversations that Haonan Wang, Louis Scharf, Dongliang Duan, and I have had over the last four years, and I would like to thank them for their invaluable advice. I would also like to thank my officemates Addison Adams and Troy Wixson for their original perspectives and sympathetic ears. Last, my most heartfelt thanks go to my partner, Ryan Filtz, without whom I would surely have foundered.

# DEDICATION

*To those I love who have passed on, and to those I love who are still here.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1

# Introduction and Background

## 1.1 Factor Analysis Background

A foundational method in multivariate statistics, Factor Analysis (FA) has been used to solve problems and obtain insights for over a century and in fields as diverse as radio astronomy (Boonstra and van der Veen, 2003; Van Der Veen et al., 2005; Antman and Leshem, 2020), market research (Stewart, 1981; Howard and O'Sullivan, 2024), machine learning (Chen et al., 2013; Fan et al., 2021; Kilian et al., 2023), array signal processing (Pesavento and Gershman, 2001; Ramírez et al., 2011; Koutrouvelis et al., 2019) and many others. Factor analysis enjoys such a wide-ranging scope of application due to the centrality of its intended purpose, which is to explain the observed correlations between a large number of measured variables in terms of a small number of unobserved *factors* without restrictive assumptions on the scales of measurement. As so much real-world statistical work requires boiling down an undifferentiated mass of variables into a small set of manageable and meaningful quantities, the use of factor analysis will surely persist so long as human understanding is a desired outcome of data exploration.

In its simplest form, the data-generating model which underlies classical factor analysis relates the $n \times 1$ vector of scalar measurements $\mathbf{x}$ to the $r \times 1$ vector $\mathbf{f}$ of latent factors and the $n \times 1$ idiosyncratic noise vector $\mathbf{u}$ as

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{u}, \tag{1.1}$$

where the $n \times r$ matrix $\mathbf{A}$ is called the *loading* matrix, as the $ij$ entry $[\mathbf{A}]_{ij}$ determines how the $i$th scalar measurement $\mathbf{x}_i$ *loads* on the $j$th factor $\mathbf{f}_j$. In (1.1), the observed vector $\mathbf{x}$ is modeled as the linear combination of two parts, namely the *systemic* part $\mathbf{s} \equiv \mathbf{A}\mathbf{f}$ and the error part or *noises* $\mathbf{u}$. The values of the factors $\mathbf{f}$ and the noises $\mathbf{u}$ are not observed, and neither the loading matrix $\mathbf{A}$ nor the noise variances $\phi_i \equiv \text{Var}(\mathbf{u}_i),\ i = 1, \ldots, n$ are known in advance. Estimation of $\mathbf{A}$ and

$\mathbf{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_n)$ on the basis of repeated samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$ is a common goal of classical FA in the exploratory setting.

Factor analysis can be referred to as a *second-order* technique, in that the primary quantity of interest is the second central moment of the measurements, $\mathbf{R}_{\mathbf{xx}} \equiv \mathrm{Var}(\mathbf{x})$. The covariance of the systemic part, $\mathbf{R}_{\mathbf{ss}} \equiv \mathrm{Var}(\mathbf{s})$, is low-rank when the number of factors is less than the number of scalar measurements and the noises $u_i$ and $u_{i'}$ in distinct scalar measurements $i \neq i'$ are assumed to be uncorrelated, so estimating $\mathbf{R}_{\mathbf{ss}}$ produces a succinct description of the second-order relationships between the scalar measurements. This follows as $[\mathbf{R}_{\mathbf{xx}}]_{ii'}$ equals $[\mathbf{R}_{\mathbf{ss}}]_{ii'}$ for $i \neq i'$ and $\mathbf{R}_{\mathbf{ss}}$ is low-rank and therefore "simple". As FA is concerned with the second moment of $\mathbf{x}$, the first moment is typically taken to be zero. In classical factor analysis, models for the higher moments or further distributional assumptions on $\mathbf{f}$ and $\mathbf{u}$ are not an essential component of FA, although joint normality of $(\mathbf{f}, \mathbf{u})$ is commonly used for estimation, to calculate hypothesis tests, or to obtain closed-form expressions for quantities of interest.

In particular, under the orthogonality and scale specification $\mathrm{Var}(\mathbf{f}) = \mathbf{I}_r$ and the specification that the noise $\mathbf{u}$ is uncorrelated with the latent factors, $\mathrm{Cov}(\mathbf{f}, \mathbf{u}) = \mathbf{0}$, estimation of $\mathbf{A}$ and $\mathbf{\Phi}$ allows for partial recovery of the latent factors by the Minimum Mean-Square Error (MMSE) estimates when joint normality is assumed. Under these assumptions, the covariance model of the measurements $\mathbf{x}$ is

$$\mathbf{R}_{\mathbf{xx}} = \mathbf{A}\mathbf{A}^{\mathsf{T}} + \mathbf{\Phi} = \mathbf{R}_{\mathbf{ss}} + \mathbf{R}_{\mathbf{uu}}. \tag{1.2}$$

That is, the measurement covariance is modeled as the sum of two structured summands, namely a rank-$r$ covariance matrix and a diagonal covariance matrix. The properties of such a sum have been studied extensively (see, e.g., Shapiro (1982, 1985); Bekker and ten Berge (1997) and references therein). For the use of classical FA in statistical signal processing, (1.1) can be cast as the multivariate signal model for measurements made at a multi-sensor array where the number of sensors is $n$ and $r$ is the number of *point sources* or total number of transmitting antennae. The factors $\mathbf{f}$ can then be interpreted as the signal transmitted by the point sources, while the error $\mathbf{u}$ represents the sensor noise which is not attributed to the point sources. The matrix $\mathbf{A}$ then is

the *channel* matrix, which is determined by the physical characteristics of signal propagation and accounts for how the transmitted signal is received by the multi-sensor array and is assumed to be constant over a sufficiently small time-scale. In the uncalibrated case, the sensor noise variances $\mathbf{\Phi}$ may be unequal and the channel matrix is unknown, and estimation of $\mathbf{A}$ and $\mathbf{\Phi}$ on the basis of the observed samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$ allows for approximate recovery of the transmitted signal $\mathbf{f}$. From this simple example and its possible extensions, factor analysis is seen to be related to problems of interest within Multiple-Input and Multiple-Output (MIMO) signal processing.

As a statistical method, classical factor analysis was theoretically grounded by the mid-century work of Kendall (1950); Anderson and Rubin (1956); Anderson (1959); Lawley and Maxwell (1962); Jöreskog (1970) and many others. However, the idea of factor analysis and the development of particular forms of FA originated in the field of psychometrics at the turn of the century. To treat the problem of capturing general intelligence, Spearman (1904) devised what is now termed a *one-factor* approach similar to (1.1) with $r = 1$. Continued work by psychologists (such as Thurstone (1935, 1947); Bartlett (1937); Ledermann (1937) ) developed much of the vocabulary and concepts used in the literature of factor analysis. Traces of the utilitarian origin of factor analysis remain, and from a statistical point of view it enjoys neither the theoretical nor computational convenience of other tent-pole multivariate methods such as Principal Component Analysis (PCA) or Canonical Correlation Analysis (CCA) which have similar aims and are closely related to FA (see, e.g., Rao (1955); Browne (1979); Basilevsky (1994)). In particular, the problems of *identifiability* of the covariance model (1.2) and the *computation* of suitable estimates of $\mathbf{A}, \mathbf{\Phi}$ are significantly more challenging than is the case for PCA.

However, it cannot be denied that practitioners have found a vast array of uses for factor analysis, and a staggering number of generalizations, extensions, and derivative methods have been developed which expand the reach and applicability of FA. The foundation of this dissertation is an extension of classical factor analysis to the so-called *multi-view* or *multi-channel* setting, which is the topic of the following section.

## 1.2 Multi-Channel Methods

### 1.2.1 Data Fusion

As the quantification of the world accelerates, the challenge of extracting insights from multiple, diverse datasets continues to increase in importance. Within engineering and applied science, the problem of joint analysis of multiple data sets which contain distinct views of some shared phenomenon has been reified as *data fusion* (Khaleghi et al., 2013; Lahat et al., 2015; Gao et al., 2020). Although the goal of data fusion, namely deriving insights about both commonalities and distinctions within some related collection of datasets, is not new and can be traced back to representative foundational works such as Hotelling (1936); Cattell (1944); Tucker (1966); Carroll and Chang (1970); Kettenring (1971), the quantity and diversity of the datasets now available has led to increased priority on obtaining methods that perform data fusion in a principled fashion. For example, data fusion is particularly relevant to prominent research domains such as brain functionality (where fusing modalities like EEG, fMRI, and DTI can improve spatiotemporal resolution (Levin-Schwartz et al., 2016)), human-machine interaction (where combining visual, auditory, and tactile modalities can improve command-following (Turk, 2014)), and autonomous driving (where fusion of LiDAR, visual, and internal state data can improve navigation (Yeong et al., 2021)).

Within signal processing, data fusion naturally arises in the context of *multi-channel* problems where vectors of measurements collected from multiple sensor arrays are to be jointly analyzed. As each sensor array may possess characteristics unique to the given array (due to, e.g., sensor location or modality) as well as capturing some shared signal, fusion of multiple channels and modalities of data can yield a richer understanding of both the signal of interest and the particularities of the data collection setup. Modern research areas in signal processing such as coordinated multipoint processing (Chu et al., 2013), cooperative relaying in Time-Division Multiple Access (TDMA) systems (Dohler and Li, 2010), and passive radar (Bandiera et al., 2007; Hack et al., 2014) can be seen as possessing a natural multi-channel structure and hence are potentially amenable to analysis

by Multi-channel Factor Analysis (MFA). The central components of MFA are outlined in Sections 1.2.2 and 1.2.3, related multi-channel or multi-view methods of data fusion are reviewed in Section 1.2.4, and toy examples of multi-channel signal processing problems for which MFA is applicable are described in Section 1.3.

### 1.2.2 Multi-Channel Factor Analysis

Multi-channel factor analysis is a method introduced by Ramírez et al. (2020) as an extension of classical or *single-channel* factor analysis to the multi-channel or multi-view setting. The channel-$c$ measurement vector $\mathbf{x}_c$ consists of $n_c$ scalar measurements, where the channel number $c$ ranges from 1 to $C$ for $C \geq 2$, and is modeled as

$$\mathbf{x}_c = \mathbf{A}_c\mathbf{f} + \mathbf{B}_c\mathbf{g}_c + \mathbf{u}_c. \tag{1.3}$$

The latent variables are the *common* factors $\mathbf{f}$ of size $r_0$ which are shared across all channels, the *distinct* factors $\mathbf{g}_c$ of size $r_c$ which are unique to channel $c$, and the channel-$c$ noises $\mathbf{u}_c$ of size $n_c$. In (1.3), the observed channel-$c$ measurement vector $\mathbf{x}_c$ is the linear combination of three unobserved parts, namely the shared systemic part $\mathbf{s}_c \equiv \mathbf{A}_c\mathbf{f}$ in channel $c$ (also referred to as the *signal* in channel $c$), the unshared systemic part $\mathbf{i}_c \equiv \mathbf{B}_c\mathbf{g}_c$ (also referred to as the *interference* in channel $c$), and the channel-$c$ noises $\mathbf{u}_c$.

To jointly analyze the multi-channel measurements $\mathbf{x}_1 \ \ldots \ \mathbf{x}_C$, the vectors are concatenated into the *all-channel* measurement vector $\mathbf{x} \equiv [\mathbf{x}_1^\mathsf{T} \ \ldots \ \mathbf{x}_C^\mathsf{T}]^\mathsf{T}$. From (1.3), the model for $\mathbf{x}$ then is

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{B}\mathbf{g} + \mathbf{u}, \tag{1.4}$$

where $\mathbf{g} = [\mathbf{g}_1^\mathsf{T} \ \ldots \ \mathbf{g}_C^\mathsf{T}]^\mathsf{T}$ and $\mathbf{u} = [\mathbf{u}_1^\mathsf{T} \ \ldots \ \mathbf{u}_C^\mathsf{T}]^\mathsf{T}$ are stacked in the same fashion as $\mathbf{x}$ and the all-channel *common* and *distinct* loading matrices $\mathbf{A}$ and $\mathbf{B}$ are respectively defined as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_C \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 & & \\ & \ddots & \\ & & \mathbf{B}_C \end{bmatrix}. \tag{1.5}$$

If $n \equiv n_1 + \cdots + n_C$ is the total number of scalar measurements across all channels and $r \equiv r_1 + \cdots + r_C$ is the total number of distinct factors across all channels, then $\mathbf{A}$ is of size $n \times r_0$ and $\mathbf{B}$ is of size $n \times r$. We assume that the total number of factors $r_0 + r$ is less than the total number of measurements $n$, and that the number of distinct factors in channel $c$ is less than $n_c$, the number of measurements in channel $c$. In addition, most of the results in this dissertation require that the block matrix $[\mathbf{A}\ \mathbf{B}]$ be full column rank or equivalently that the *signal subspace* $\operatorname{col}\mathbf{A}$ and the *combined interference* subspace $\operatorname{col}\mathbf{B}$ intersect trivially. The channel size vector $\mathbf{n} = [n_1, \ldots, n_C]^\mathsf{T} \in \mathbb{N}_C$ and the factor number vector $\mathbf{r} = [r_0, r_1, \ldots, r_C]^\mathsf{T} \in \mathbb{N}_{C+1}$ are used for economy of notation.

The model (1.4) is the foundation of this dissertation. In Chapter 2, the method of MFA is described in greater detail, conditions which guarantee the generic identifiability of the covariance components are obtained, and the asymptotic properties of the associated estimators are derived. In Chapter 3, (1.3) is extended to discrete time series and the new method of Multi-channel Factor Spectral Analysis (MFSA) is presented. In Chapter 4, the method of MFA underpins the development of a technique to model multi-vehicle trajectory data and categorize vehicles into *pods* on the basis of stability of relative configuration. Finally, Chapter 5 discusses the findings of this dissertation and describes possible directions for future work.

### 1.2.3 MFA Covariance Properties

In the all-channel model (1.4), the latent vectors $\mathbf{f}, \mathbf{g}, \mathbf{u}$ are assumed to be mean-zero and uncorrelated with each other. The latent factors $\mathbf{f}$ and $\mathbf{g}$ can be further assumed to have identity covariance without loss of generality, as discussed in detail in Section 2.2. With these assumptions, the observation covariances $\mathbf{R_{xx}}$ that can be obtained through the data-generating model (1.4) can be written as a sum

$$\mathbf{R_{xx}} = \mathbf{R_{ss}} + \mathbf{R_{ii}} + \mathbf{\Phi} = \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T} + \mathbf{\Phi} \tag{1.6}$$

where $\mathbf{R_{ss}}$ is a covariance matrix of rank $r_0$, $\mathbf{R_{ii}}$ is a block diagonal covariance matrix whose $j$th block is of size $n_c \times n_c$ and rank $r_c$, and $\mathbf{\Phi}$ is a diagonal covariance matrix. The set $\mathcal{R}(\mathbf{n}, \mathbf{r})$ contains

the MFA observation covariance matrices $\mathbf{R_{xx}}$ which can be realized by (1.6) for some $\mathbf{A}, \mathbf{B}, \mathbf{\Phi}$ with the given channel sizes and factor numbers, as detailed in Section 2.2.

To study the properties of the covariance matrices $\mathbf{R_{xx}} \in \mathcal{R}(\mathbf{n}, \mathbf{r})$ for some $\mathbf{n}, \mathbf{r}$, the terms in the decomposition (1.6) can be grouped into $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ and $\mathbf{\Phi}$. Diagonal covariance matrices like $\mathbf{\Phi}$ are well understood, so we concentrate on the properties of those covariance matrices which can be written as the sum $\mathbf{R_{ss}} + \mathbf{R_{ii}}$. If a matrix $\mathbf{K} \in \mathbb{R}_{n\times n}$ can be written as

$$\mathbf{K} \equiv \mathbf{R_{ss}} + \mathbf{R_{ii}} = \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T}$$

for some $\mathbf{A} \in \mathbb{R}_{n\times r_0}$ and $\mathbf{B} = \mathrm{blkdiag}(\mathbf{B}_1, \ldots, \mathbf{B}_C)$ with $\mathbf{B}_c \in \mathbb{R}_{n_c \times r_c},\ c = 1, \ldots, C$, then $\mathbf{K}$ will be referred to as a *noise-free* MFA covariance matrix or as a *block-spiked rank-restricted* covariance with spike sizes $\mathbf{n}$ and rank restrictions $\mathbf{r}$. The set of such $\mathbf{K}$ is denoted by $\mathcal{K}(\mathbf{n}, \mathbf{r})$.

To contextualize the identifiability results of Chapters 2 and 3, the following characterizations of the elements of $\mathcal{K}(\mathbf{n}, \mathbf{r})$ and the random vectors whose variance is in $\mathcal{K}(\mathbf{n}, \mathbf{r})$ will be of use. By constructing the all-channel observation $\mathbf{x}$ by stacking the observations in individual channels, we implicitly decompose the *all-channel observation space* $\mathbb{R}_n$ into the subspace sum $\mathcal{C}_1 \boxplus \cdots \boxplus \mathcal{C}_C$, where the *channel-$c$ observation subspace* $\mathcal{C}_c$ is isomorphic to $\mathbb{R}_{n_c}$. In this perspective, matrices such as $\mathbf{B}$ which are block-diagonal with blocks of the appropriate sizes correspond with the operators which have $\mathcal{C}_c, c = 1, \ldots, C$ as invariant subspaces. As the signal subspace $\mathrm{col}\,\mathbf{A}$ is *not* required to be orthogonal to the combined interference subspace $\mathrm{col}\,\mathbf{B}$, relating the subspace sum $\mathrm{col}\,\mathbf{A} \oplus \mathrm{col}\,\mathbf{B}$ to the channel decomposition $\mathbf{C}_1 \boxplus \cdots \boxplus \mathbf{C}_C$ in the context of the Euclidean inner product is somewhat complicated, and the goal of the following propositions is to provide some clarity. For the first proposition, the maximal rank elements of $\mathcal{K}(\mathbf{n}, \mathbf{r})$ are shown to admit a particular factorization into structured components.

**Proposition 1.** *(Factorization) The positive semi-definite matrix $\mathbf{K} \in \mathbb{R}_{n\times n}$ with rank $r_0 + r$ is in $\mathcal{K}(\mathbf{n}, \mathbf{r})$ if and only if $\mathbf{K} = \mathbf{Q}\mathbf{H}\mathbf{Q}^\mathsf{T}$ for some $\mathbf{Q} \in \mathbb{R}_{n\times(r_0+r)}$ with $\mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{I}_{r_0+r}$ and invertible*

$\mathbf{H} \in \mathbb{R}_{(r_0+r)\times(r_0+r)}$*, where* $\mathbf{Q}$ *and* $\mathbf{H}^{-1}$ *are structured as*

$$\mathbf{Q} = \begin{matrix} \\ n \end{matrix}\!\!\begin{matrix} \begin{matrix} r & r_0 \end{matrix} \\ \begin{bmatrix} \mathbf{W} & \mathbf{U} \end{bmatrix} \end{matrix}, \qquad \mathbf{W} = \mathrm{blkdiag}(\mathbf{W}_1, \ldots, \mathbf{W}_C), \qquad \mathbf{W}_c \in \mathbb{R}_{n_c \times r_c},\ c = 1, \ldots, C,$$

$$\mathbf{H}^{-1} = \begin{matrix} \begin{matrix} r & r_0 \end{matrix} \\ \begin{bmatrix} \mathbf{T} & \mathbf{Y} \\ \mathbf{Y}^\mathsf{T} & \mathbf{S} \end{bmatrix} \end{matrix}\!\!\begin{matrix} r \\ r_0 \end{matrix}, \qquad \mathbf{T} = \mathrm{blkdiag}(\mathbf{T}_1, \ldots, \mathbf{T}_C), \qquad \mathbf{T}_c \in \mathbb{R}_{r_c \times r_c},\ c = 1, \ldots, C.$$

*Proof.* See Section A.1 in Appendix A for proof. □

For the second proposition, if $\mathbf{y}$ is a random vector in $\mathbb{R}_n$ whose covariance is a maximal-rank element $\mathbf{K} \in \mathcal{K}(\mathbf{n}, \mathbf{r})$, then the following proposition relates the channel structure of $\mathbf{K}$ to the geometry of linear MMSE predictors of the channel subvectors $\mathbf{y}_1, \ldots, \mathbf{y}_C$. As defined in Section 1.4, $\mathcal{S}_1 \boxplus \mathcal{S}_2$ is the orthogonal direct sum of subspaces $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{R}_n$, $\mathcal{S}^\perp$ is the orthogonal complement, and $\mathbf{P}_\mathcal{S}$ is the orthogonal projection onto $\mathcal{S}$ in the Euclidean inner product. In addition, for random vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{z}$, the notation $\mathbf{u} \perp \mathbf{v} \mid \mathbf{z}$ indicates that the linear MMSE errors in predicting $\mathbf{u}$ and $\mathbf{v}$ using $\mathbf{z}$ are uncorrelated. That is, $\mathbf{u} \perp \mathbf{v} \mid \mathbf{z}$ is equivalent to $E[(\mathbf{u} - \widehat{\mathbf{u}})(\mathbf{v} - \widehat{\mathbf{v}})^\mathsf{T}] = \mathbf{0}$, where $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ are the LMMSE predictions of $\mathbf{u}$ and $\mathbf{v}$ using $\mathbf{z}$. With these definitions, Proposition 2 shows that the random vectors whose variance is in $\mathcal{K}(\mathbf{n}, \mathbf{r})$ are precisely those whose channel subvectors are conditionally uncorrelated given the orthogonal projection of $\mathbf{y}$ onto a subspace $\mathcal{V}$ of dimension $r_0$ and where the channel prediction errors $\mathbf{y}_c - \hat{\mathbf{y}}_c$ live in a $r_c$-dimensional subspace of $\mathcal{C}_c$.

**Proposition 2.** *(Prediction) Suppose* $\mathbf{y} = [\mathbf{y}_1^\mathsf{T} \ \ldots \ \mathbf{y}_C^\mathsf{T}]^\mathsf{T}$ *is a mean-zero random vector in* $\mathbb{R}_n$ *with subvectors* $\mathbf{y}_c$ *of size* $n_c$ *for* $c = 1, \ldots, C$*. The covariance matrix* $\mathbf{K} \equiv \mathrm{Var}(\mathbf{y})$ *is in* $\mathcal{K}(\mathbf{n}, \mathbf{r})$ *with rank* $r_0 + r$ *if and only if there exists a subspace* $\mathcal{V} \subset \mathbb{R}_n$ *of dimension* $r_0$ *such that*

$$\mathrm{col}\,\mathbf{K} = \mathcal{V} \boxplus (\mathrm{col}\,\mathbf{K} \cap \mathcal{V}^\perp \cap \mathcal{C}_1) \boxplus \cdots \boxplus (\mathrm{col}\,\mathbf{K} \cap \mathcal{V}^\perp \cap \mathcal{C}_C), \tag{1.7}$$

$$\dim(\mathrm{col}\,\mathbf{K} \cap \mathcal{V}^\perp \cap \mathcal{C}_c) = r_c, \quad c = 1, \ldots, C, \tag{1.8}$$

$$\mathbf{y}_c \perp \mathbf{y}_k \mid \mathbf{P}_\mathcal{V}\mathbf{y}, \quad c, k = 1, \ldots, C, \quad c \neq k. \tag{1.9}$$

*Proof.* See Section A.2 in Appendix A for proof. □

If $\mathbf{y} = \mathbf{A}\mathbf{f} + \mathbf{B}\mathbf{g}$ and $\mathbf{Q}_{\mathbf{B},\mathbf{A}}$ is an oblique projection with range $\mathrm{col}\,\mathbf{B}$ and kernel which includes $\mathrm{col}\,\mathbf{A}$, then $\mathbf{Q}_{\mathbf{B},\mathbf{A}}\mathbf{y} = \mathbf{B}\mathbf{g}$. The oblique projection $\mathbf{Q}_{\mathbf{B},\mathbf{A}}$ is adapted to the non-orthogonal subspace decomposition $\mathbb{R}_n = \mathrm{col}\,\mathbf{A} \oplus \mathrm{col}\,\mathbf{B} \oplus \mathcal{Z}$ where $\mathcal{Z}$ is any complementary subspace to $\mathrm{col}\,\mathbf{K}$. As $\mathbf{B}$ is block-diagonal and the channel subvectors $g_1, \ldots, g_C$ are assumed to be uncorrelated, the projection $\mathbf{Q}_{\mathbf{B},\mathbf{A}}$ decorrelates the channels by removing the common factor. This fact is related to but distinct from the result of the above proposition, which decorrelates the channel prediction errors by conditioning on the orthogonal projection of $\mathbf{y}$ onto a certain subspace $\mathcal{V}$. Note that $\mathcal{V}$ will not equal $\mathrm{col}\,\mathbf{A}$ in general unless $\mathrm{col}\,\mathbf{A}$ and $\mathrm{col}\,\mathbf{B}$ are orthogonal subspaces in addition to intersecting trivially.

### 1.2.4 Related Multi-Channel Methods

With the rise of data fusion as a high-level area of research interest, a number of extensions of classic multivariate statistics techniques to the multi-channel setting have been developed. Examples of such multi-view methods include Joint and Individual Variation Explained (JIVE) (Lock et al., 2013), which can be seen as a multi-channel analog of PCA, Shared Independent Component Analysis (ShICA) (Richard et al., 2024) which can bee seen as a multi-channel analog of Independent Component Analysis (ICA), and Joint Independent Subspace Analysis (JISA) (Lahat et al., 2015; Lahat and Jutten, 2019) which is a multi-channel generalization of both Independent Subspace Analysis (ISA) as well as Independent Vector Analysis (IVA) (Silva et al., 2014). Other methods of multi-view data fusion do not directly correspond to a classical multivariate method but instead exploit novel properties of higher-order tensors, such as the essential uniqueness of the Canonical Polyadic Decomposition (CPD) (Carroll and Chang, 1970; Kruskal, 1989). Examples of such methods include Coupled Matrix and Tensor Factorization (CMTF) (Acar et al., 2013) and Generalized Coupled Tensor Factorization (GCTF) (Ermiş et al., 2015) which is its probabilistic analog, as well as Coupled Canonical Polyadic Decomposition (CCPD) (Kuang et al., 2023) and Linked Multi-Way Component Analysis (LMWCA) (Cichocki et al., 2015).

The diversity of methods for multi-channel data fusion represents the substantial freedom which researchers have in specifying how information from the multiple views should be combined, and precisely which aspects of the model are to be shared across views and which are allowed to vary from view to view. Within this abundance of riches, MFA provides value by partnering the intuitive simplicity of FA with a multi-channel model which is sufficiently flexible to cover problems of interest (as seen in Section 1.3) while allowing for important results such as the identifiability guarantees of Chapter 2 to be obtained. Further discussion of MFA identifiability and comparisons to identifiability results of other multi-channel methods is found in Section 2.3. As research into data fusion continues to progress, multi-channel methods such as those discussed above will be streamlined and unified, best practices will be developed, and the empirical performance of these methods on real-world problems will become clear. We believe that MFA as well as the associated results and extensions presented in this dissertation are an important contribution to this progression.

## 1.3 Illustrative Multi-Channel Problems

### 1.3.1 Multi-Channel Blind Source Separation with Interference and Noise

Blind Source Separation (BSS) (Cardoso and Souloumiac, 1994; Cardoso, 1998) is a classic array processing problem where the observed measurement vector $\mathbf{x}$ is assumed to arise as a linear mixture of unknown sources, where the mixture matrix is unknown and the source signals are not observed. The objective of BSS is to estimate the unmixing transformation, allowing the individual transmitted signals to be recovered from the measurements. Distinguishing the signals requires the exploitation of some form of *diversity* between the mixed signals (Sidiropoulos and Bro, 2000), which in the classic ICA approach to BSS consists of the assumption that the sources are independent and non-Gaussian. However, the multi-channel setting allows for additional forms of source diversity (Lahat et al., 2015, Sec. III.C) and can enable source separation beyond what is possible in the single-channel case.

Consider the setting in which a single high-powered source transmits a rank-one signal $f[t]$, which is detected by $C$ spatially separated antenna arrays. The measurement series in each channel

are the vectors $\mathbf{x}_1[t], \ldots, \mathbf{x}_C[t]$. In addition to the high-powered source, the $c$th antenna array also receives interference from $r_c$ weak local sources. The interference from the $r_c$ sources close to the $c$th array is assumed to have negligible effect on the measurements at all other arrays. With the addition of idiosyncratic noises $\mathbf{u}_1[t], \ldots, \mathbf{u}_C[t]$, the measurement series $\mathbf{x}_c[t] \in \mathbb{C}_{n_c}$ can be modeled as

$$\mathbf{x}_c[t] = f[t]\mathbf{a}_c + \sum_{j=1}^{r_c} g_{c,j}[t]\mathbf{b}_{c,j} + \mathbf{u}_c[t].$$

The vector $\mathbf{a}_c \in \mathbb{C}_{n_c}$ controls how the transmitted signal $f[t]$ is received in the $c$th channel, and the $\mathbf{b}_{c,j}$ perform a similar role for the interfering transmissions $g_{c,j}[t]$. With the assumption that the $\mathbf{g}_{c,j}$ are uncorrelated for different channels, MFA can be used to estimate the $\mathbf{a}_c$ and $\mathbf{b}_{c,j}$. Due to rotation invariance, the $g_{c,j}[t]$ cannot be uniquely determined in the Gaussian setting from the second-order statistics (1.6) alone. However, due to the source diversity implied by the multi-channel structure of the problem, the magnitude of the rank-one signal series of interest, $|f[t]|$, can be uniquely separated from the interference and noise up to a scale factor when identifiability conditions on the channel sizes and factor numbers are satisfied, even in the case where all signals are normally distributed. This would not be the case if classical single-channel factor analysis were applied to the all-channel observation vector $\mathbf{x}$, as the class of observation-equivalent source separations would include mixtures of the rank-one signal and the interferences. A simulated experiment related to the application of MFA to multi-channel BSS in the presence of interference and noise is conducted in Section 3.5.2, where modeling the diversity between the temporal correlation of the rank-one signal and that of the interferences in addition to the multi-channel structure is found to improve the recovery of the signal of interest.

### 1.3.2 Target Detection in Passive Radar

As described in Ramírez et al. (2020, Sec. IV.D), MFA with $C = 2$ channels is well-suited to the problem of detecting the presence or absence of a target by a passive radar system. Unlike active radar, a passive radar system is receiver-only and relies on a non-cooperating source of radio emissions. A passive radar system (Hack et al., 2014) has *surveillance* and *reference* antenna

arrays which yield associated measurement vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. The unknown signal $\mathbf{f}$ is transmitted by some strong *opportunistic illuminator* such as a commercial radio station. The reference channel will always receive the signal as $\mathbf{s}_2 = \mathbf{A}_2\mathbf{f}$, where $\mathbf{f}$ will present as $\mathbf{s}_1 = \mathbf{A}_1\mathbf{f}$ in the surveillance channel only when a target is present to reflect the direct path signal from the illuminator and synchronize the channels. Both measurements are potentially corrupted by interferences $\mathbf{i}_1 = \mathbf{B}_1\mathbf{g}_1$ and $\mathbf{i}_2 = \mathbf{B}_2\mathbf{g}_2$, which respectively account for the direct path signal to the surveillance array and the potential multi-path signal to the reference array.

Target detection in a passive radar setup can then be cast as a hypothesis testing problem between the target-present model $\mathcal{H}_1$ the target-absent model $\mathcal{H}_0$,

$$\mathcal{H}_1 : \begin{cases} \mathbf{x}_1 &= \mathbf{A}_1\mathbf{f} + \mathbf{B}_1\mathbf{g}_1 + \mathbf{u}_1 \\ \mathbf{x}_2 &= \mathbf{A}_2\mathbf{f} + \mathbf{B}_2\mathbf{g}_2 + \mathbf{u}_2 \end{cases}, \qquad \mathcal{H}_0 : \begin{cases} \mathbf{x}_1 &= \mathbf{B}_1\mathbf{g}_1 + \mathbf{u}_1 \\ \mathbf{x}_2 &= \mathbf{A}_2\mathbf{f} + \mathbf{B}_2\mathbf{g}_2 + \mathbf{u}_2 \end{cases}.$$

As the experiments of Ramírez et al. (2020) demonstrate, the GLRT of the above problem can be evaluated using MFA and enjoys substantially improved detection performance over a competing coherence-based test, which fails to leverage the assumed low-dimensional nature of the interference terms. In Section 3.5, a similar experiment is conducted where temporal correlation is introduced into the latent factors, and an approximate GLRT detector evaluated with MFSA is shown to outperform the MFA-based detector as the former exploits frequency-varying differences in power spectral densities to distinguish the signal from the interference and noise.

## 1.4 Notation

Matrices and vectors are denoted with bold-faced symbols, and scalars are denoted with light-face symbols. A real matrix of size $n \times m$ is written as $\mathbf{D} \in \mathbb{R}_{n\times m}$, and a column vector of length $n$ is written as $\mathbf{d} \in \mathbb{R}_n$. The zero matrix of dimension $m \times n$ is $\mathbf{0}_{m,n}$ and the $n \times n$ identity matrix is $\mathbf{I}_n$. A zero vector of dimension $n$ is written as $\mathbf{0}_n$. When clear from context, the subscripts may be dropped. The standard basis for $\mathbb{R}_n$ will be written as $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$. Matrix and vector transposes are written as $\mathbf{D}^\mathsf{T}$ and $\mathbf{d}^\mathsf{T}$ respectively. The determinant of $\mathbf{D}$ is written as $\det \mathbf{D}$, and the trace

is written as $\operatorname{tr}\mathbf{D}$. The $(i,j)$th entry of a matrix $\mathbf{D}$ is $[\mathbf{D}]_{ij}$, and similarly for the $i$th entry of a column vector. The matrix obtained by concatenating the columns of $\mathbf{B}$ to the right of the columns of $\mathbf{A}$ is written as $[\mathbf{A}\ \mathbf{B}]$. For row and column index sets $\alpha \subset \{1,\ldots,n\}$ and $\beta \subset \{1,\ldots,m\}$, the submatrix $\mathbf{D}[\alpha,\beta]$ contains the entries $[\mathbf{D}]_{ij}$ with $(i,j) \in \alpha \times \beta$. For $\mathbf{M} \in \mathbb{C}_{n\times m}$, let $\mathbf{M}^{-}$ denote the Moore-Penrose pseudo-inverse of $\mathbf{M}$. The non-negative part of a scalar expression $a \in \mathbb{R}$ is $(a)_+ \equiv \max\{a,0\}$. The normal distribution with mean $\mathbf{m}$ and variance $\mathbf{V}$ is $\mathcal{N}(\mathbf{m},\mathbf{V})$. The Kronecker product of $\mathbf{X}$ and $\mathbf{Y}$ is $\mathbf{X}\otimes\mathbf{Y}$.

The operator $\operatorname{Diag}^{-1}$ applied to a matrix yields the vector containing the diagonal entries. The block-diagonal operator $\operatorname{blkdiag}$ applied to a list of matrices yields the block-diagonal matrix with the listed blocks. The $\operatorname{vec}$ operator vectorizes a matrix by stacking the columns vertically, while $\operatorname{vech}$, applicable to square matrices, vectorizes by extracting only the lower-triangular entries. Denote the spaces of $n\times n$ symmetric, symmetric positive semidefinite, orthogonal, and diagonal matrices as $\operatorname{Sym}(n)$, $\operatorname{PSD}(n)$, $\operatorname{O}(n)$ and $\operatorname{Diag}(n)$ respectively, with $\mathbb{P}_n$ being $\operatorname{PSD}(n)\cap\operatorname{Diag}(n)$. For matrix $\mathbf{A}$ with submatrix $\mathbf{A}_{11}$, the generalized Schur complement of $\mathbf{A}_{11}$ is $\mathbf{A}\setminus\mathbf{A}_{11}$. For symmetric matrices $\mathbf{V},\mathbf{W}$ of the same size $\mathbf{V}\succeq\mathbf{W}$ indicates that $\mathbf{V}-\mathbf{W}$ is positive semi-definite. For vector subspaces $\mathcal{A}$ and $\mathcal{B}$, the subspace intersection is $\mathcal{A}\cap\mathcal{B}$ and the subspace sum and direct sum are respectively denoted by $\mathcal{A}+\mathcal{B}$ and $\mathcal{A}\oplus\mathcal{B}$. If $\mathcal{A}$ and $\mathcal{B}$ are orthogonal, their subspace sum is $\mathcal{A}\boxplus\mathcal{B}$. For a matrix $\mathbf{X}$, $\operatorname{col}\mathbf{X}$ is the linear span of the columns of $\mathbf{X}$. For a linear map $\mathbf{T}$, the image and kernel subspaces are $\operatorname{Im}(\mathbf{T})$ and $\operatorname{Ker}(\mathbf{T})$.

The space of $n\times 1$ column vectors with complex entries is $\mathbb{C}_n$, and the space of $n\times m$ matrices with complex entries is $\mathbb{C}_{n\times m}$. For a scalar $z\in\mathbb{C}$, $\overline{z}$ denotes the complex conjugate of $z$. The argument of a complex number if $\arg z$. The conjugate transposes of $\mathbf{x}$ and $\mathbf{X}$ are written as $\mathbf{x}^{\mathsf{H}}$ and $\mathbf{X}^{\mathsf{H}}$. If $\mathbf{X}\in\mathbb{C}_{n\times n}$ is invertible, then $\mathbf{X}^{-1}$ is its inverse while $\mathbf{X}^{-\mathsf{T}}=(\mathbf{X}^{\mathsf{T}})^{-1}$ and $\mathbf{X}^{-\mathsf{H}}=(\mathbf{X}^{\mathsf{H}})^{-1}$.

# Chapter 2

# Identifiability and Asymptotics of Multi-Channel Factor Analysis

## 2.1 Chapter Overview

In this chapter, the method of multi-channel factor analysis is specified in greater detail, with particular focus on the covariance model (Section 2.2). The problem of identifiability of the covariance components is treated in depth (Section 2.3), with the goal of giving identifiability guarantees which require only the specification of the channel sizes and signal and interference dimensionality. For single-channel factor analysis, practitioners correctly assume identifiability whenever the number of common factors is much smaller than the number of observations (Shapiro, 1985). However, for multi-channel factor analysis, the channel sizes and desired number of common and distinct factors may vary widely, and so the question of identifiability becomes more challenging and less intuitive. This chapter carefully examines the two main sources of non-identifiability of MFA (Section 2.3.1), namely isolation of the idiosyncratic noise variances and separation of signal and interference covariances. Conditions which guarantee the *generic* global and local identifiability of the covariance components are then obtained, providing the desired identifiability results using only what is *a priori* specified, namely the channel sizes and factor numbers. This parallels the advancement of single-channel FA as a statistical method as reviewed in Section 2.3. With this identifiability result, parameter estimates obtained by maximizing a Gaussian likelihood as in Algorithm 1 of Ramírez et al. (2020) are shown to be consistent and asymptotically normal (Section 2.4), even in the case where the true distribution of the latent vectors is non-normal. The work contained in this chapter has been previously published as Stanton et al. (2023a, 2024).

## 2.2 MFA Model

### 2.2.1 Description

An archetypal data collection scheme for which MFA is applicable consists of multiple sensors or observation units, each of which collects a vector of measurements. Often these sensors are homogeneous (such as when all sensors measure voltage), but MFA is also applicable to a heterogeneous collection of sensors. The input from an individual sensor then composes an individual *channel* of observation for some shared signal which is measured by multiple sensors. This channel structure is set by the design of the sensor array, and is known in advance of data collection. The channels are numbered by $c = 1, \ldots, C$ with $n_c$ scalar measurements in channel $c$.

For channel $c$, denote the vector of measurements within that channel as $\mathbf{x}_c$. As in (1.3), the generative model for $\mathbf{x}_c$ is

$$\mathbf{x}_c = \mathbf{A}_c\mathbf{f} + \mathbf{B}_c\mathbf{g}_c + \mathbf{u}_c, \tag{2.1}$$

where $\mathbf{A}_c\mathbf{f}$ is the signal in channel $c$, $\mathbf{B}_c\mathbf{g}_c$ is the channel-$c$ interference that lives within a low-dimensional subspace, and $\mathbf{u}_c$ is the measurement noise. The matrices $\mathbf{A}_c \in \mathbb{R}_{n_c \times r_0}$ and $\mathbf{B}_c \in \mathbb{R}_{n_c \times r_c}$ are the common and distinct factor loadings for channel $c$. The number of common factors $r_0 \leq n$ and distinct factors $r_1, \ldots, r_C$ with $r_c \leq n_c$ determine the flexibility of the model, as the common factor $\mathbf{f}$ is in $\mathbb{R}_{r_0}$ and the distinct factor for channel $c$, $\mathbf{g}_c$, is in $\mathbb{R}_{r_c}$. The remaining portion of each measurement in channel $c$ which is not a result of the influence of the latent factors is contributed by $\mathbf{u}_c \in \mathbb{R}_{n_c}$.

The above data collection scheme and related model (2.1) is appropriate for several signal processing problems, as described in Section 1.3. Another possible application of (2.1) is cooperative relaying in Time-Division Multiple Access (TDMA) systems (Dohler and Li, 2010), where multiple relays each transmit a common signal to a multi-antenna access point in sequential time slots $c = 1, \ldots, C$. The common signal presents in slot $c$ as $\mathbf{A}_c\mathbf{f}$ and is subject to interference $\mathbf{B}_c\mathbf{g}_c$ and measurement noise $\mathbf{u}_c$.

The all-channel observation vector is obtained by stacking the channels as $\mathbf{x} \equiv [\mathbf{x}_1^\mathsf{T} \cdots \mathbf{x}_c^\mathsf{T}]^\mathsf{T}$. The first-order model for the all-channel observations is

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{B}\mathbf{g} + \mathbf{u}, \tag{2.2}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the all-channel loadings,

$$\mathbf{A} \equiv \left[\mathbf{A}_1^\mathsf{T} \ \cdots \mathbf{A}_C^\mathsf{T}\right]^\mathsf{T}, \quad \mathbf{B} \equiv \mathrm{blkdiag}(\mathbf{B}_1, \ldots, \mathbf{B}_C), \tag{2.3}$$

with $\mathbf{g} \equiv [\mathbf{g}_1^\mathsf{T} \cdots \mathbf{g}_C^\mathsf{T}]^\mathsf{T}$ and $\mathbf{u} \equiv [\mathbf{u}_1^\mathsf{T} \cdots \mathbf{u}_C^\mathsf{T}]^\mathsf{T}$. For clarity of notation, continue to let $\mathbf{n} = [n_1, \ldots, n_C]^\mathsf{T}$, and $\mathbf{r} = [r_0, r_1, \ldots, r_C]^\mathsf{T}$. The total number of observations is $n = \sum_{c=1}^{C} n_c$ and the total number of distinct factors is $r = \sum_{c=1}^{C} r_c$. Denote the cumulative sum of the number of observations and distinct factors as $n_{<c} \equiv \sum_{k=1}^{c-1} n_k$ and $r_{<c} \equiv \sum_{k=1}^{c-1} r_k$, respectively. For $c = 1$, $r_{<1}$ and $n_{<1}$ are set to $0$. Similarly, define $n_{>c} \equiv n - n_c - n_{<c}$ and $r_{>c} \equiv r - r_c - r_{<c}$. Table 2.1 summarizes the notation which is frequently used in this chapter.

### 2.2.2 Covariance Specification

In (2.2), the factors $\mathbf{f}, \mathbf{g}$ and the errors $\mathbf{u}$ are latent random quantities, while the factor loadings $\mathbf{A}, \mathbf{B}$ are fixed unknown parameters. The latent factors are assumed to satisfy

$$E[\mathbf{f}] = \mathbf{0}_{r_0}, \qquad E[\mathbf{f}\mathbf{f}^\mathsf{T}] \equiv \mathbf{R}_{\mathbf{f}\mathbf{f}}, \qquad E[\mathbf{g}_c] = \mathbf{0}_{r_c}, \qquad E[\mathbf{g}_c\mathbf{g}_c^\mathsf{T}] \equiv \mathbf{R}_{\mathbf{g}_c\mathbf{g}_c}.$$

Factors of different types are required to be uncorrelated,

$$E[\mathbf{f}\mathbf{g}^\mathsf{T}] = \mathbf{0}_{r_0,r}, \qquad E[\mathbf{g}_c\mathbf{g}_{c'}^\mathsf{T}] = \mathbf{0}_{r_c,r_{c'}}\ c \neq c'.$$

The idiosyncratic errors $\mathbf{u}$ are assumed to satisfy

$$E[\mathbf{u}] = \mathbf{0}_n, \qquad E[\mathbf{u}\mathbf{g}^\mathsf{T}] = \mathbf{0}_{n,r}, \qquad E[\mathbf{u}\mathbf{f}^\mathsf{T}] = \mathbf{0}_{n,r_0}, \qquad E[\mathbf{u}\mathbf{u}^\mathsf{T}] = \mathbf{\Phi},$$

for some covariance matrix $\mathbf{\Phi} \in \mathbb{P}_n$. With this specification, $\mathbf{x}$ has mean zero with covariance matrix

$$\begin{aligned} \mathbf{R}_{\mathbf{xx}} &\equiv \mathbf{A}\mathbf{R}_{\mathbf{ff}}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{R}_{\mathbf{gg}}\mathbf{B}^\mathsf{T} + \mathbf{\Phi}, \\ &= \mathbf{R}_{\mathbf{ss}} + \mathbf{R}_{\mathbf{ii}} + \mathbf{\Phi}, \end{aligned} \tag{2.4}$$

where $\mathbf{R}_{\mathbf{gg}} = \mathrm{blkdiag}(\mathbf{R}_{\mathbf{g}_1\mathbf{g}_1}, \ldots, \mathbf{R}_{\mathbf{g}_C\mathbf{g}_C})$ and the *signal* and *interference* covariances are $\mathbf{R}_{\mathbf{ss}} \equiv \mathbf{A}\mathbf{R}_{\mathbf{ff}}\mathbf{A}^\mathsf{T}$ and $\mathbf{R}_{\mathbf{ii}} \equiv \mathbf{B}\mathbf{R}_{\mathbf{gg}}\mathbf{B}^\mathsf{T}$, respectively. The set $\mathcal{R}(\mathbf{n}, \mathbf{r}) \subset \mathrm{PSD}(n)$ contains all observation covariance matrices realizable by (2.4).

### 2.2.3 Parameterization of MFA Covariance Models

For given channel sizes $\mathbf{n}$ and factor numbers $\mathbf{r}$, the generative model (2.2) for the all-channel observation $\mathbf{x}$ under MFA determines a set of covariance matrices $\mathcal{R}(\mathbf{n}, \mathbf{r}) \subset \mathrm{PSD}(n)$ by (2.4). The set $\mathcal{R}(\mathbf{n}, \mathbf{r})$ can be parameterized in three ways, namely by the triple of structured components $(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi})$, by the loading matrices $\mathbf{A}, \mathbf{B}$ and noise variances $\mathbf{\Phi}$ whose structures are shown in Figure 2.1, or by a vector $\boldsymbol{\eta}$ which captures the degrees of freedom in the $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ parameterization.

**Parameterization by $(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi})$**

In MFA, (2.4) shows that the observation covariance $\mathbf{R}_{\mathbf{xx}}$ is the sum of a low-rank matrix $\mathbf{R}_{\mathbf{ss}}$, a channel-structured block-diagonal matrix $\mathbf{R}_{\mathbf{ii}}$ with low-rank blocks, and a non-negative diagonal matrix $\mathbf{\Phi}$. Any triple $(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi})$ of appropriately structured $n \times n$ matrices determines an element of $\mathcal{R}(\mathbf{n}, \mathbf{r})$ by the second line of (2.4). That is, if $\mathbf{R}_{\mathbf{ss}} \in \mathrm{PSD}(n)$ has rank at most $r_0$, $\mathbf{R}_{\mathbf{ii}} \in \mathrm{PSD}(n)$ is block-diagonal whose $c$th block is $n_c \times n_c$ with rank at most $r_c$, and $\mathbf{\Phi}$ is in $\mathbb{P}_n$, then

$$\mathbf{R}_{\mathbf{xx}}(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi}) \equiv \mathbf{R}_{\mathbf{ss}} + \mathbf{R}_{\mathbf{ii}} + \mathbf{\Phi} \tag{2.5}$$

is in $\mathcal{R}(\mathbf{n}, \mathbf{r})$. This can be seen by taking $\mathbf{R}_{\mathbf{ff}}$ and $\mathbf{R}_{\mathbf{gg}}$ to be identity matrices and obtaining $\mathbf{A}$ and $\mathbf{B}$ from the Cholesky factors of $\mathbf{R}_{\mathbf{ss}}$ and $\mathbf{R}_{\mathbf{ii}}$ respectively.

Recovering $(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi})$ from an estimate of $\mathbf{R}_{\mathbf{xx}}$ is the central goal of MFA, as decomposing $\mathbf{R}_{\mathbf{xx}}$ into the three summands will separate $\mathbf{R}_{\mathbf{ss}}$, which controls the cross-channel covariance,

from $\mathbf{R_{ii}}$, which modifies the within-channel covariance. Both covariance-controlling components are then isolated from the idiosyncratic noise variance for individual inputs. As the summands are separately interpretable and are identifiable from $\mathbf{R_{xx}}$, as will be shown in Section 2.3, the parameterization of $\mathcal{R}(\mathbf{n}, \mathbf{r})$ in terms of $(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$ forms the basis for interpreting the results of MFA.

**Parameterization by** $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$

However, the rank constraints on $\mathbf{R_{ss}}$ and $\mathbf{R_{ii}}$ are inconvenient, as the set of such matrices is not a vector space. It is typical in factor analysis to parameterize in terms of the loading matrices $\mathbf{A}$ and $\mathbf{B}$, so that the rank constraints are automatically satisfied. This increases the complexity of the parameterization map (as it is quadratic rather than linear), but simplifies the domain.

The first line of (2.4) parameterizes $\mathcal{R}(\mathbf{n}, \mathbf{r})$ in terms of $(\mathbf{A}, \mathbf{B}, \mathbf{R_{ff}}, \mathbf{R_{gg}}, \mathbf{\Phi})$. However, without further information about either the loading matrices $\mathbf{A}, \mathbf{B}$ or the factor variances $\mathbf{R_{ff}}, \mathbf{R_{gg}}$, it is clear that the pairs $(\mathbf{A}, \mathbf{R_{ff}})$ and $(\mathbf{B}, \mathbf{R_{gg}})$ are non-identifiable from knowledge of $\mathbf{x}$ alone. As the factors are unobserved, any change of basis on the factor spaces taking $(\mathbf{A}, \mathbf{f})$ to $(\mathbf{A}\mathbf{T}_0, \mathbf{T}_0^{-1}\mathbf{f})$ and $(\mathbf{B}_c, \mathbf{g}_c)$ to $(\mathbf{B}_c\mathbf{T}_c, \mathbf{T}_c^{-1}\mathbf{g}_c)$ leaves the observations unchanged. In the exploratory case where no information beyond the channel structure and the factor space dimensionality is assumed, the invariance of the observations to linear transformations of the factor space is most easily resolved by imposing that the factors be uncorrelated and unit-scale, $\mathbf{R_{ff}} = \mathbf{I}_{r_0}$ and $\mathbf{R_{g_cg_c}} = \mathbf{I}_{r_c}$ for all $c = 1, \ldots, C$. Under this assumption, $\mathcal{R}(\mathbf{n}, \mathbf{r})$ can be parameterized as

$$\mathbf{R_{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}) \equiv \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T} + \mathbf{\Phi}. \tag{2.6}$$

The set of common factor loadings $\mathbf{A}$ is $\mathbb{A} \equiv \mathbb{R}_{n \times r_0}$, while the set of distinct factor loadings is $\mathbb{B} \subset \mathbb{R}_{n \times r}$ containing those $\mathbf{B} \in \mathbb{R}_{n \times r}$ which are block diagonal with $c$th block $\mathbf{B}_c \in \mathbb{R}_{n_c \times r_c}$. The domain of $\mathbf{R_{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ is $\mathbb{A} \times \mathbb{B} \times \mathbb{P}_n$.

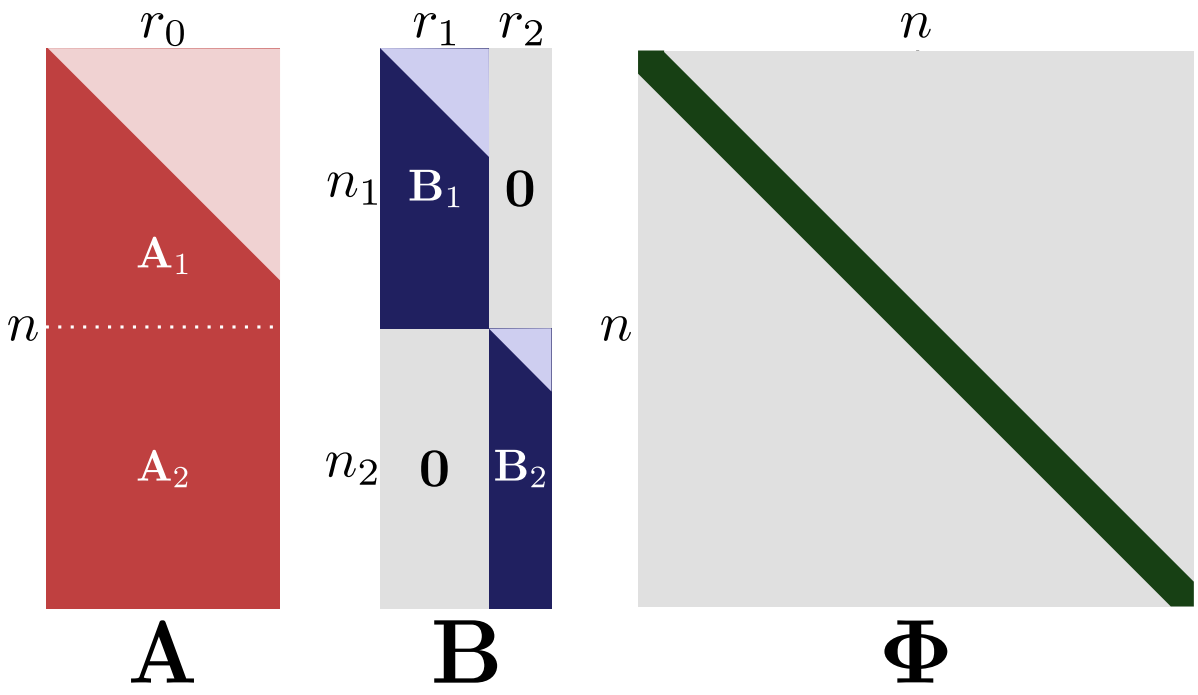



**Figure 2.1:** Depiction of MFA covariance parameters $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ for two channels. Triangles indicate constraints of $\mathbb{A}_L$ and $\mathbb{B}_L$.

**Table 2.1:** Table of commonly used notation and descriptions in Chapter 2

| Quantity | Description | Quantity | Description |
|---|---|---|---|
| $C$ | # of channels | $n_c$ | Chan-$c$ size |
| $r_0$ | Common fac. num. | $r_c$ | Chan-$c$ distinct fac. num. |
| $n$ | Total chan. size | $r$ | Total distinct fac. num. |
| $\mathbf{n} \in \mathbb{N}_C$ | Vector of chan. sizes | $\mathbf{r} \in \mathbb{N}_{C+1}$ | Vector of fac. numbers |
| $\mathbf{A} \in \mathbb{R}_{n \times r_0}$ | Common fac. loadings | $\mathbf{B} \in \mathbb{R}_{n \times r}$ | Distinct fac. loadings |
| $\mathbb{A} \subset \mathbb{R}_{n \times r_0}$ | Set of $\mathbf{A}$s | $\mathbb{B} \subset \mathbb{R}_{n \times r}$ | Set of $\mathbf{B}$s |
| $\mathbb{A}_L, \mathbb{B}_L$ | LT top block subspaces | $\mathbb{A}_L^*, \mathbb{B}_L^*$ | LT with positive diag. |
| $\mathbf{\Phi} \in \mathbb{R}_{n \times n}$ | Diag. noise cov. | $\boldsymbol{\eta} \in \mathbb{R}_L$ | $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ free params |
| $\mathbf{R_{ss}} \in \mathbb{R}_{n \times n}$ | Rank-$r_0$ cov. | $\mathbf{R_{ii}} \in \mathbb{R}_{n \times n}$ | Blkdiag interf. cov. |
| $\mathbf{R_{xx}}$ | MFA observation cov. | $\mathcal{R}(\mathbf{n}, \mathbf{r})$ | Set of $\mathbf{R_{xx}}$s |

**Parameterization by $\boldsymbol{\eta}$**

The above parameterization $\mathbf{R}_{\mathbf{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ by the loading matrices introduces a *rotation invariance*, as $\mathbf{R}_{\mathbf{xx}}(\mathbf{A}\mathbf{Q}_f, \mathbf{B}\mathbf{Q}_g, \mathbf{\Phi})$ equals $\mathbf{R}_{\mathbf{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ for any orthogonal $\mathbf{Q}_f$ and $\mathbf{Q}_g$, where $\mathbf{Q}_g = \text{blkdiag}(\mathbf{Q}_{g,1}, \ldots, \mathbf{Q}_{g,C})$ with $\mathbf{Q}_{g,c} \in \mathbb{R}_{r_c \times r_c},\ c = 1, \ldots, C$. For purposes of estimation and asymptotic analysis, it is desirable to eliminate this invariance by adding artificial restrictions to $\mathbf{A}$ and $\mathbf{B}$, in such a way that the realizable products $\mathbf{A}\mathbf{A}^\mathsf{T}$ and $\mathbf{B}\mathbf{B}^\mathsf{T}$ are not restricted. Analogous restrictions which remove rotation invariance in single-channel factor analysis are well-known, and typically involve orthogonality of loading matrix columns or imposition of structural zeros (Anderson and Rubin, 1956; Jöreskog, 1969).

Here, appropriate restrictions are imposed in the same fashion as in Ramírez et al. (2020). Consider loading matrices $\mathbf{A}$, $\mathbf{B}_1, \ldots, \mathbf{B}_C$ which have lower-triangular (LT) top blocks, $\mathbf{A}_1$ and $\mathbf{B}_{1,1}, \ldots, \mathbf{B}_{C,1}$, of sizes $r_0 \times r_0$ and $r_c \times r_c, c = 1, \ldots, C$ respectively. The remaining rows are unconstrained, and the remaining submatrices are written as $\mathbf{A}_2$ and $\mathbf{B}_{1,2}, \ldots, \mathbf{B}_{C,2}$. Define $\mathbb{A}_L \subset \mathbb{A}$ and $\mathbb{B}_L \subset \mathbb{B}$ as the subspaces of loading matrices which satisfy their respective restrictions. Further, distinguish $\mathbb{A}_L^*$ as the subset where, for each $j = 1, \ldots, r_0$, the main diagonal element $[\mathbf{A}_1]_{jj}$ is either positive or the $j$th column of $\mathbf{A}_1$ is zero. The set $\mathbb{B}_L^*$ is defined similarly.

The non-redundant degrees of freedom in $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ for $\mathbf{A} \in \mathbb{A}_L$ and $\mathbf{B} \in \mathbb{B}_L$ compose the vector $\boldsymbol{\eta} \in \mathbb{R}_L$ as

$$\boldsymbol{\eta} = \left[\text{vech}(\mathbf{A}_1)^\mathsf{T}\ \text{vec}(\mathbf{A}_2)^\mathsf{T}\ \text{vech}(\mathbf{B}_{1,1})^\mathsf{T}\ \text{vec}(\mathbf{B}_{1,2})^\mathsf{T}\ \ldots\ \text{vec}(\mathbf{B}_{C,2})^\mathsf{T}\ \text{Diag}^{-1}(\mathbf{\Phi})^\mathsf{T}\right]^\mathsf{T}, \tag{2.7}$$

where the dimension $L$ is

$$L = nr_0 - \tfrac{1}{2}r_0(r_0 - 1) + \textstyle\sum_{c=1}^{C}\left[n_c r_c - \tfrac{1}{2}r_c(r_c - 1)\right] + n. \tag{2.8}$$

The subset of $\boldsymbol{\eta}$ so obtained is $V \subset \mathbb{R}_L$. The parameterization $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ of $\mathcal{R}(\mathbf{n}, \mathbf{r})$ is obtained by inverting (2.7) for $(\mathbf{A}(\boldsymbol{\eta}), \mathbf{B}(\boldsymbol{\eta}), \mathbf{\Phi}(\boldsymbol{\eta}))$ and taking $\mathbf{R}_{\mathbf{xx}}(\mathbf{A}(\boldsymbol{\eta}), \mathbf{B}(\boldsymbol{\eta}), \mathbf{\Phi}(\boldsymbol{\eta}))$.

## 2.3 Identifiability

### Definition

For multi-channel factor analysis as defined in Section 2.2, we say that an observation covariance matrix $\mathbf{\Sigma_{xx}} \in \mathcal{R}(\mathbf{n}, \mathbf{r})$ is *identified* when it can be uniquely decomposed into a sum of appropriately structured components. That is, in the terms of Section 2.2.3, $\mathbf{\Sigma_{xx}}$ is identified if there is a unique triple $(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$ such that $\mathbf{R_{xx}}(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi}) = \mathbf{\Sigma_{xx}}$. As the covariance matrices $\mathbf{R_{ss}}, \mathbf{R_{ii}}$, and $\mathbf{\Phi}$ contain all information in MFA about the statistical properties of the signal, interference, and noise respectively, any inference must be based on these covariances. However, if there exists a different triple $(\widetilde{\mathbf{R}}_{\mathbf{ss}}, \widetilde{\mathbf{R}}_{\mathbf{ii}}, \widetilde{\mathbf{\Phi}})$ of structured matrices which also sums to $\mathbf{\Sigma_{xx}}$, then any inference based on the individual values for the summands must be suspect. As $\mathbf{\Sigma_{xx}}$ is not known, practical application of MFA requires that the observation covariance model (2.4) be guaranteed to be identified using only what is specified *a priori*, namely the channel sizes and number of common and distinct factors. If the channel sizes $\mathbf{n}$ and factor numbers $\mathbf{r}$ permit such a guarantee, we say that MFA is *identifiable* with those channel sizes and factor numbers.

In this definition, identifiability of MFA is a property of the covariance matrix $\mathbf{\Sigma_{xx}}$ and refers to the uniqueness of the second order decomposition (2.6), *not* uniqueness of the first order generative model (2.1). As discussed in Section 2.2.3, the latent factors themselves are not uniquely identifiable even in the noise-free case, as the bases for the common and distinct factor spaces can be changed without altering the observations. However, if the MFA covariance $\mathbf{\Sigma_{xx}}$ is identified in the above sense and a preferred basis for the factor space is chosen, then the uniqueness of the MFA decomposition allows for linear MMSE estimation of the latent factors $\mathbf{f}$ and $\mathbf{g}_c,\ c = 1, \ldots, C$ in that basis (see (Ramírez et al., 2020, Section IV.B) for a related experiment).

Identification of $(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$ from $\mathbf{\Sigma_{xx}}$ breaks into two subproblems, namely *isolation of the idiosyncratic variances* and *separation of the signal and interference covariances*. That is, the former subproblem refers to whether $\mathbf{\Sigma_{xx}}$ uniquely determines $(\mathbf{R_{ss}} + \mathbf{R_{ii}}, \mathbf{\Phi})$ while the latter subproblem refers to whether $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ uniquely determines $(\mathbf{R_{ss}}, \mathbf{R_{ii}})$.

In Section 2.3.1, conditions on $\mathbf{n}$ and $\mathbf{r}$ which resolve the subproblems and ensure the identifiability of $(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$ from $\mathbf{\Sigma_{xx}}$ are derived. The following Section 2.3.2 investigates the identifiability and associated properties of the parameterization $\mathbf{R_{xx}}(\boldsymbol{\eta})$ of $\mathcal{R}(\mathbf{n}, \mathbf{r})$ in terms of $\boldsymbol{\eta}$, which provides the required technical foundation for the asymptotic analysis of Section 2.4. The relationships between the conditions and results of this section are summarized in Figure 2.2.

**Generic, Global, and Local Identifiability**

To establish the channel sizes and factor numbers for which MFA is identifiable, it is important to realize that certain degenerate $\mathbf{\Sigma_{xx}}$ will never be identified. For example, choose $\mathbf{A}, \mathbf{B}$ such that the block matrix $[\mathbf{A}\ \mathbf{B}]$ has some orthogonal rows and all other rows being zero. The resulting $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ belonging to $\mathcal{K}(\mathbf{n}, \mathbf{r})$ will itself be diagonal and so the noise variances cannot be isolated. Although the subset of $\mathcal{R}(\mathbf{n}, \mathbf{r})$ containing the non-identified MFA observation covariance models is not precisely characterized, it is sufficient for practical applications to find conditions on $\mathbf{n}$ and $\mathbf{r}$ which imply that such non-identified covariance models are atypical. In addition, a distinction can be made between *local* and *global* identifiability. A locally identified triple $(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$ summing to $\mathbf{\Sigma_{xx}}$ is guaranteed to be the unique such triple within some neighborhood, whereas global identifiability extends this guarantee to the entire space.

These types of result are common in the identifiability literature; Shapiro (1985) and Bekker and ten Berge (1997) establish local and global identifiability for single-channel factor analysis in this *generic* sense, while Kiràly and Tomioka (2012) examines generic identifiability in low-rank matrix completion. Formally, we call a subset of a $d$-dimensional real vector space *null* if its image under a linear isomorphism to $\mathbb{R}_d$ has Lebesgue measure zero. A statement is *generically* true if it is true for all elements excepting a null subset.

**Connections to Previous Results in FA**

The question of identifiability in single-channel factor analysis was an area of interest for many years. Based on an equation-counting argument, Ledermann (1937) provided a heuristic for the maximum number of factors (known as the Ledermann bound). Anderson and Rubin (1956) set

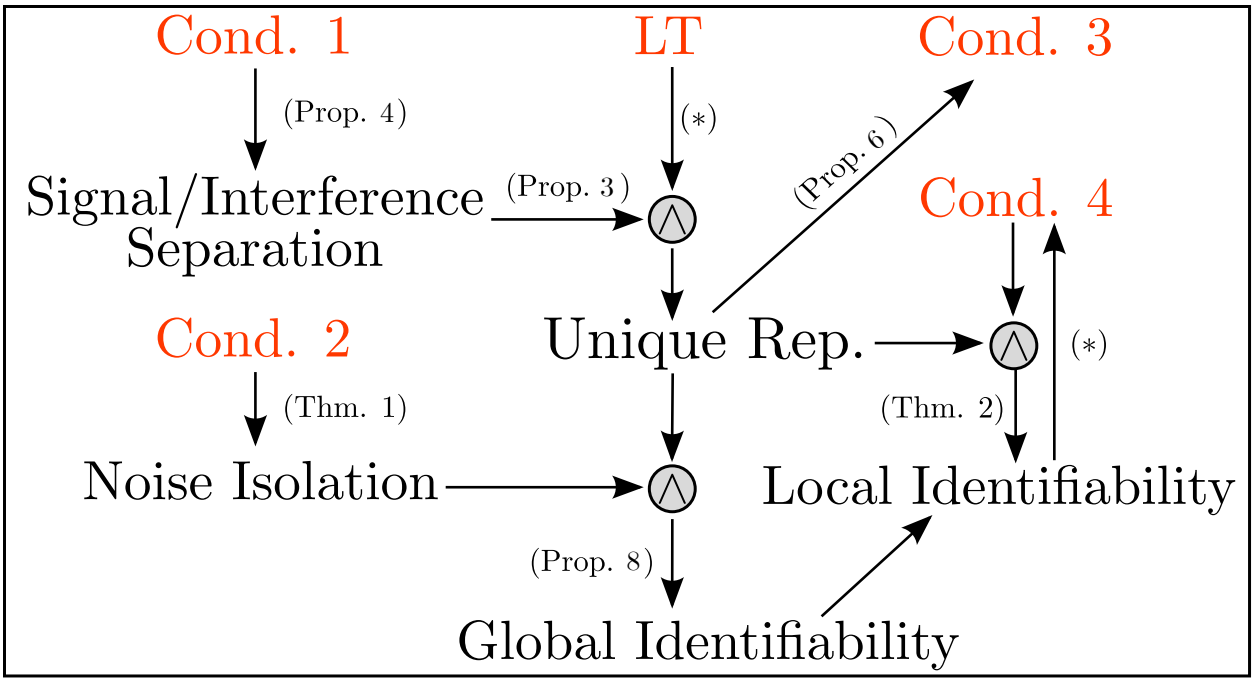



**Figure 2.2:** Diagram indicating the relationships between the conditions imposed on the channel sizes and factor numbers and their implications for the different aspects of MFA identification. Directions marked with asterisks were obtained in Ramírez et al. (2020). "LT" refers to the lower-triangular structure in $\mathbb{A}_L^*$ and $\mathbb{B}_L^*$

out a simple sufficient condition for identifiability by requiring that the loading matrix contain two disjoint full-rank submatrices after removal of a single row, which will be generically satisfied when the number of common factors is less than half the number of observations. Later, Shapiro (1985) demonstrated that the Ledermann bound was generically sufficient for *local* identifiability, providing a maximal number of common factors which approaches $n$ rather than $n/2$. Shapiro also conjectured that this identifiability threshold also held for *global* identifiability, which was later shown to be correct by Bekker and ten Berge (1997).

In this paper, analogous results for multi-channel factor analysis are obtained. The discussion on the identifiability of MFA was opened by Ramírez et al. (2020, Sec. III) and the importance of the problem was recognized. In particular, the authors provided the restrictions which remove rotation invariance used in Proposition 5, and give *necessary* conditions on the factor numbers by an equation-counting argument. The authors conjectured that, as in single-channel FA, the threshold obtained by counting knowns and unknowns should be *sufficient* for identifiability, which here is Condition 4. With the addition of conditions to ensure separability of the signal and interference

covariances, which was not treated by Ramírez et al. (2020), this conjecture is verified for *local* identifiability. For *global* identifiability, we instead require the slightly stronger Condition 2.

**Identifiability in Related Multi-Channel Methods**

Just as classical FA has deep connections with other multivariate statistical methods such as Canonical Correlation Analysis (CCA) (Rao, 1955), MFA can be related to other techniques for multi-channel data analysis. In particular, CCA (both in classical two-view form (Ibrahim and Sidiropoulos, 2019) and in the generalized multi-view form (Sørensen et al., 2021)) has been successfully used to find latent structures shared across multiple channels, which is also an objective of MFA. Other multi-channel techniques such as Joint Independent Subspace Analysis (JISA) (Lahat and Jutten, 2016), Shared Independent Component Analysis (ShICA) (Richard et al., 2024) and Deep CCA (Andrew et al., 2013) also enable discovery of latent structures under differing assumptions on the relations of the shared and unshared aspects to the multi-channel observations. Useful identifiability results for Generalized CCA (Sørensen et al., 2021), Deep CCA (Lyu et al., 2022), JISA (Lahat and Jutten, 2019), and ShICA (Richard et al., 2024) have been obtained through a variety of approaches.

However, the MFA-specific identifiability results of this chapter are not direct consequences of previous results, and differ in two ways. First, Proposition 4 for generic separability of the signal and interference covariances does *not* require that the number of factors $r_0 + r_c$ affecting channel $c$ be less than the channel size, as Condition 1 allows for $r_0 + r_c$ to be greater than $n_c$ for some channels. If $r_0 + r_c > n_c$, then the latent factors $(\mathbf{f}, \mathbf{g}_c)$ cannot be uniquely determined from $\mathbf{x}_c$ alone, even in the noise-free ($\mathbf{\Phi} = \mathbf{0}$) case with $\mathbf{A}_c$ and $\mathbf{B}_c$ known. Second, the presence of noise in the observations substantially alters the identifiability problem, as unique isolation of the noise variance $\mathbf{\Phi}$ neither implies nor is implied by separability of the signal and interference covariances. Specific comparisons to the identifiability results for Deep CCA, Generalized CCA, JISA, and ShICA are discussed in Section B.12 in Appendix B.

### 2.3.1 Identifiability of $\mathbf{R_{xx}}(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$

**Separation of Signal and Interference Covariances**

The first subproblem of MFA identifiability involves the noise-free part of (2.6), namely the combined signal-and-interference covariance $\mathbf{R_{ss}}+\mathbf{R_{ii}}$. To treat the first subproblem, it is convenient to work with the loading parameterization $\mathbf{R_{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$, then transfer to $\mathbf{R_{xx}}(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$.

Define two equivalence relations on $\mathbb{A} \times \mathbb{B}$ by

$$(\mathbf{A}, \mathbf{B}) \sim_1 (\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \iff \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T} = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\mathsf{T} + \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\mathsf{T}$$

$$(\mathbf{A}, \mathbf{B}) \sim_2 (\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \iff \mathbf{A}\mathbf{A}^\mathsf{T} = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\mathsf{T} \text{ and } \mathbf{B}\mathbf{B}^\mathsf{T} = \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\mathsf{T}.$$

Under $\sim_1$, two pairs of loading matrices are equivalent if they correspond to the same sum $\mathbf{R_{ss}}+\mathbf{R_{ii}}$, while under $\sim_2$, two pairs are equivalent if they correspond to the same tuple $(\mathbf{R_{ss}}, \mathbf{R_{ii}})$. It is clear that $\sim_2$ is a *finer* relation than $\sim_1$, and by definition $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ can be uniquely separated into $\mathbf{R_{ss}}$ and $\mathbf{R_{ii}}$ iff the $\sim_2$-equivalence class of $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ contains a single $\sim_1$-equivalence class.

More can be said about the structure of these $\sim_1$ and $\sim_2$ equivalence classes. Application of a well-known result (see, e.g., Anderson and Rubin (1956, Lemma 5.1)) shows that for pairs $(\mathbf{A}, \mathbf{B}), (\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \in \mathbb{A} \times \mathbb{B}$ with $(\mathbf{A}, \mathbf{B}) \sim_1 (\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ we must have

$$\big[\widetilde{\mathbf{A}}\ \widetilde{\mathbf{B}}\big] = \big[\mathbf{A}\ \mathbf{B}\big]\mathbf{Q},$$

for some orthogonal matrix $\mathbf{Q} \in \mathrm{O}(r_0 + r)$. That is, any two $\sim_1$-equivalent pairs are such that the combined loading matrices $[\mathbf{A}\ \mathbf{B}]$ and $[\widetilde{\mathbf{A}}\ \widetilde{\mathbf{B}}]$ represent the same map under different orthonormal bases for the combined factor space of both common and distinct factors. Similarly, if $(\mathbf{A}, \mathbf{B}) \sim_2 (\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ then $\widetilde{\mathbf{A}} = \mathbf{A}\mathbf{Q}_{00}$ for $\mathbf{Q}_{00} \in \mathrm{O}(r_0)$ and $\widetilde{\mathbf{B}}_c = \widetilde{\mathbf{B}}_c\mathbf{Q}_{cc}$ with $\mathbf{Q}_{cc} \in \mathrm{O}(r_c)$ for each

$c = 1, \ldots, C$. Partitioning $\mathbf{Q}$ as,

$$\mathbf{Q} = \begin{array}{c} \begin{matrix} r_0 & r_1 \cdots\cdots\cdots r_C \end{matrix} \\ \begin{bmatrix} \mathbf{Q}_{00} & \mathbf{Q}_{01} & \ldots & \mathbf{Q}_{0C} \\ \mathbf{Q}_{10} & \mathbf{Q}_{11} & \ldots & \mathbf{Q}_{1C} \\ \vdots & \vdots & & \vdots \\ \mathbf{Q}_{C0} & \mathbf{Q}_{C1} & \ldots & \mathbf{Q}_{CC} \end{bmatrix} \end{array} \begin{matrix} r_0 \\ r_1 \\ \vdots \\ r_C \end{matrix} \tag{2.9}$$

then $(\mathbf{A}, \mathbf{B}) \sim_2 (\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ iff $[\widetilde{\mathbf{A}}\ \widetilde{\mathbf{B}}]$ can be obtained from $[\mathbf{A}\ \mathbf{B}]$ by right-multiplication by a *block-diagonal* $\mathbf{Q}$.

One distinction between multi-channel FA and single-channel FA with $r_0 + r$ total factors is that not all products $[\mathbf{A}\ \mathbf{B}]\mathbf{Q}$ will correspond to a valid pair of MFA loading matrices $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \in \mathbb{A} \times \mathbb{B}$. This is due to the structural zeros of $\mathbb{B}$. If $\mathbf{Q}$ is block diagonal, the product $[\mathbf{A}\ \mathbf{B}]\mathbf{Q}$ will preserve the structural zeros in $\mathbf{B}$ and correspond to a valid member of $\mathbb{A} \times \mathbb{B}$. However, the converse is not true without further restrictions on $\mathbf{n}$ and $\mathbf{r}$, as non-block-diagonal $\mathbf{Q}$ which preserve the structural zeros in $\mathbf{B}$ can exist even in non-degenerate cases. A example of this is given in Section B.11.

With the above equivalence relations $\sim_1$ and $\sim_2$, existence of such a non-block-diagonal $\mathbf{Q}$ occurs exactly when $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ cannot be uniquely separated. The following proposition gives sufficient conditions on $(\mathbf{A}, \mathbf{B})$ so that all $\mathbf{Q}$ which preserve the structural zeros of $\mathbf{B}$ are block-diagonal.

**Proposition 3.** *For $\mathbf{A} \in \mathbb{A}$ and $\mathbf{B} \in \mathbb{B}$, suppose that, after possibly renumbering the channels, the submatrices $\mathbf{M}_1, \ldots, \mathbf{M}_C$ of $[\mathbf{A}\ \mathbf{B}]$ have Full Column Rank (FCR), where $\mathbf{M}_c$ is*

$$\mathbf{M}_c = \begin{bmatrix} \mathbf{A}_{<c} & \mathbf{B}_{<c} \\ \mathbf{A}_{>c} & \mathbf{0} \end{bmatrix}, \ \mathbf{M}_1 = [\mathbf{A}_2^\mathsf{T}\ \ldots\ \mathbf{A}_C^\mathsf{T}]^\mathsf{T} \tag{2.10}$$

*with $\mathbf{A}_{<c} = [\mathbf{A}_1^\mathsf{T} \ldots \mathbf{A}_{c-1}^\mathsf{T}]^\mathsf{T}$, $\mathbf{A}_{>c} = [\mathbf{A}_{c+1}^\mathsf{T} \ldots \mathbf{A}_C^\mathsf{T}]^\mathsf{T}$ and $\mathbf{B}_{<c} = \mathrm{blkdiag}(\mathbf{B}_1, \ldots, \mathbf{B}_{c-1})$. Then any $\mathbf{Q} \in \mathrm{O}(r_0 + r)$ such that $[\mathbf{A}\ \mathbf{B}]\mathbf{Q} = [\widetilde{\mathbf{A}}\ \widetilde{\mathbf{B}}]$ for some $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \in \mathbb{A} \times \mathbb{B}$ must have $\mathbf{Q}_{ij} = \mathbf{0}$ for all $i \neq j$ when partitioned as* (2.9)*.*

*Proof.* See Section B.2 in Appendix B for proof. □

If the $\sim_1$-equivalence class of $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ contains any $(\mathbf{A}, \mathbf{B})$ which satisfy the condition of Proposition 3, then all elements in the $\sim_1$-equivalence class belong to the same $\sim_2$-equivalence class and so $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ can be uniquely separated. This can be seen by letting $(\mathbf{A}, \mathbf{B})$ satisfy the above condition, so for any $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \sim_2 (\mathbf{A}, \mathbf{B})$ the pairs must in fact be related by a block-diagonal orthogonal transformation. Hence, the $\sim_2$ and $\sim_1$ equivalence classes collapse by transitivity.

The following condition on $\mathbf{n}$ and $\mathbf{r}$ implies that the hypothesis of Proposition 3 is generically satisfied on $\mathbb{A} \times \mathbb{B}$, and therefore $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ can be uniquely separated into $(\mathbf{R_{ss}}, \mathbf{R_{ii}})$.

**Condition 1.** *The channel sizes $n_1, \ldots, n_C$ and factor numbers $r_0, \ldots, r_C$ satisfy $r_c \leq n_c$ and*

$$r_0 + r_{<c} \leq n - n_c, \tag{2.11}$$

*for all $c = 1, \ldots, C$.*

If $r_0 + r_c \leq n_c$ for all channels, then Condition 1 is satisfied for any ordering of the channels. This follows as $r_0 \leq \min_{c=1,\ldots,C}\{n_c - r_c\}$ implies $r_0 \leq (n_1 - r_1) + \ldots (n_{c-1} - r_{c-1}) + n_{c+1} + \ldots n_C$, and so (2.11) is satisfied for all $c = 1, \ldots, C$. Condition 1 depends on channel ordering, but its use in the following proposition is not order dependent.

**Proposition 4.** (Generic Separability of $\mathbf{R_{ss}} + \mathbf{R_{ii}}$) *If Condition 1 is satisfied for some permutation of the channel numbers, then the subset of $\mathbb{A} \times \mathbb{B}$ which does not satisfy the condition of Proposition 3 is null.*

*Proof.* See Section B.3 in Appendix B for proof. □

**Isolation of Noise Variances**

For single-channel FA identifiability, the main criterion is $\phi$ defined for $n, r, \rho \in \mathbb{N}$ as

$$\phi(n, r, \rho) = \frac{r(r+1)}{2} - \frac{\rho(\rho+1)}{2} - \rho(r - \rho) - n.$$

The threshold for global identifiability of single-channel FA (Bekker and ten Berge, 1997) is then $\phi(n, r, 2r - n) > 0$. For MFA, the analogous criterion function $\psi(\mathbf{n}, \mathbf{r}, \boldsymbol{\rho})$ is

$$\psi(\mathbf{n}, \mathbf{r}, \boldsymbol{\rho}) = n + \phi(n, r_0, \rho_0) + \sum_{c=1}^{C} \phi(n_c, r_c, \rho_c) + r_c(r_0 - \rho_0), \tag{2.12}$$

with non-negative integer vector $\boldsymbol{\rho} = [\rho_0, \rho_1, \ldots, \rho_C]^\mathsf{T}$. The criterion $\psi$ depends on all of the factor numbers $r_0, r_1, \ldots, r_C$ and not a function of the total number of factors alone.

**Condition 2.** *The channel sizes* $\mathbf{n}$ *and factor numbers* $\mathbf{r}$ *satisfy* $r_c \leq n_c$ *and* $r_0 + r \leq n$. *In addition, let* $\psi^*$ *be the smallest criterion value over possible MFA reductions,*

$$\psi^* = \min_{(\mathbf{n}', \mathbf{r}', \boldsymbol{\rho}) \in M} \psi(\mathbf{n}', \mathbf{r}', \boldsymbol{\rho}) \tag{2.13}$$

*where* $M \subset \mathbb{N}_{3C+2}$ *contains all non-negative* $(\mathbf{n}', \mathbf{r}', \boldsymbol{\rho})$ *satisfying*

$$\begin{aligned}
&n'_c \leq n_c,\ c = 1, \ldots, C, \\
&r'_c = (r_c - (n_c - n'_c))_+,\ c = 1, \ldots, C, \\
&r'_0 = [r_0 - \textstyle\sum_{c=1}^{C} (n_c - n'_c - r_c)_+]_+, \\
&\rho_c \leq \min\{r'_c, 2(r'_0 + r'_c) - n'_c\},\ c = 1, \ldots, C, \\
&\rho_0 = \min\{r_0, 2r'_0 + \textstyle\sum_{c=1}^{C} 2r'_c - \rho_c - n'_c\},
\end{aligned} \tag{2.14}$$

*and* $\sum_{c=1}^{C} n'_c > 0$. *Either* $\psi^* > 0$ *or* $M$ *is empty.*

Similarly to Bekker and ten Berge (1997), $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$ is said to have globally identified noise variances if $\mathbf{R}_{\mathbf{xx}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi}) = \mathbf{R}_{\mathbf{xx}}(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\boldsymbol{\Phi}})$ implies that $\boldsymbol{\Phi} = \widetilde{\boldsymbol{\Phi}}$. The following theorem establishes that Condition 2 is sufficient for $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$ to generically have globally identifiable noise variances and hence that the noise variances can be uniquely isolated from $\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T}$.

**Theorem 1.** (Separation of $\boldsymbol{\Phi}$) *If Condition 2 is met,* $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$ *has globally identified noise variances except for a null subset of* $\mathbb{A} \times \mathbb{B} \times \mathbb{P}_n$.

*Proof.* See Section B.7 in Appendix B for proof. □

If the channel sizes and factor numbers meet both Conditions 1 and 2, the results of this section imply that the observation covariance can be uniquely decomposed into the signal, interference, and noise covariances, except for a null set of degenerate cases. Therefore, interpretation of the individual components of MFA is well-founded.

Showing that Condition 2 is sufficient for the unique isolation of the noise variances divides into two cases. The first possibility considered is whether the observation covariance $\mathbf{R_{ss}} + \mathbf{R_{ii}} + \mathbf{\Phi}$ permits a second representation as $\widetilde{\mathbf{R}}_{\mathbf{ss}} + \widetilde{\mathbf{R}}_{\mathbf{ii}} + \widetilde{\mathbf{\Phi}}$ with all noise variances $[\widetilde{\mathbf{\Phi}}]_{ii}$ not equaling $[\mathbf{\Phi}]_{ii},\ i = 1, \ldots, n$. Such a second representation precludes unique isolation of the noise variances, and implies that $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ differs by a diagonal matrix from another noise-free MFA covariance with the same factor numbers. This relationship between two noise-free MFA covariances implies the existence of a symmetric matrix $\mathbf{H}$ of size $r_0 + r$ with appropriate structural zeros and satisfying both an overall rank constraint and rank constraints on the main diagonal blocks, where the constraints are functions of the channel sizes and common and distinct factor numbers. As the overall and block rank constraints interact, the integer vector $\boldsymbol{\rho}$ sets the ranks of the diagonal blocks of $\mathbf{H}$, where the possible values are given in Condition 2 for $n'_c = n_c$, $r'_0 = r_0$ and $r'_c = r_c$ for $c = 1, \ldots, C$. The criterion $\psi$ can be seen (B.8) as the effective number of constraints imposed on $\mathbf{H}$ minus the degrees of freedom in choosing the diagonal difference matrix. If the criterion $\psi$ is positive for all permitted $\boldsymbol{\rho}$, then all block ranks lead to an overdetermined problem, so generically no such second representation exists.

In the second case, the noise variances have $[\widetilde{\mathbf{\Phi}}]_{ii} = [\mathbf{\Phi}]_{ii}$ for $i$ in some index set $\beta$. To resolve the second case, the fact that the difference $[\widetilde{\mathbf{\Phi}}]_{ii} - [\mathbf{\Phi}]_{ii} = 0$ for indices in $\beta$ yields that the associated principal submatrices of $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ and $\widetilde{\mathbf{R}}_{\mathbf{ss}} + \widetilde{\mathbf{R}}_{\mathbf{ii}}$ are equal. Taking the generalized Schur complement in $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ of this submatrix reduces the second case to the first case with smaller channel sizes $n'_c$ and factor numbers $r'_0$ and $r'_c$ for $c = 1, \ldots, C$, where the possible reduced channel sizes and factor numbers (for varying index sets $\beta$) are set out in Condition 2. If $\psi$ is positive for all possible reduced $\mathbf{n}', \mathbf{r}'$ and the associated possible block ranks $\boldsymbol{\rho}$, then $\mathbf{\Phi}$ can be generically be isolated from $\mathbf{R_{ss}} + \mathbf{R_{ii}}$.

### 2.3.2 Identifiability of $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$

The previous section established conditions under which the MFA decomposition of the observation covariance into the signal, interference, and noise covariance matrices is identifiable and thus interpretable. This section provides complementary results for the identifiability of $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$. These results are of technical relevance for the analysis of Section 2.4 as they allow standard parameter estimation theory to be applied.

#### Unique Representative

In constructing the parameterization of $\mathcal{R}(\mathbf{n}, \mathbf{r})$ in terms of $\boldsymbol{\eta}$, the first step is defining the subset $\mathbb{A}_L^* \times \mathbb{B}_L^*$ of $\sim_2$-equivalence class representatives. The following proposition establishes that $\mathbb{A}_L^* \times \mathbb{B}_L^*$ contains a unique representative from each $\sim_2$-equivalence class.

**Proposition 5.** *(LT Uniqueness) For any* $(\mathbf{A}, \mathbf{B}) \in \mathbb{A} \times \mathbb{B}$ *there is a unique* $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \in \mathbb{A}_L^* \times \mathbb{B}_L^*$ *such that* $(\mathbf{A}, \mathbf{B}) \sim_2 (\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$.

*Proof.* See Section B.4 in Appendix B for proof. □

This result parallels the use of LT restrictions to select a unique representative loading matrix in single-channel FA. However, for MFA, a previously unrecognized complication occurs when $\mathbf{R}_{\mathbf{ss}} + \mathbf{R}_{\mathbf{ii}}$ cannot be uniquely separated. In this case, there exist multiple elements of $\mathbb{A}_L^* \times \mathbb{B}_L^*$ which are $\sim_1$ equivalent, and so the LT restrictions do not select a unique representative from each $\sim_1$-equivalence class.

Proposition 6 applies a result for confirmatory factor analysis (Bekker, 1986) to give a *necessary* condition that the channel sizes and factor numbers must satisfy so that the LT restriction will distinguish a unique representative of the $\sim_1$-equivalence class. Condition 1 is *sufficient* for the same result.

**Condition 3.** *The channel sizes* $\mathbf{n}$ *and the factor numbers* $\mathbf{r}$ *satisfy*

$$r_0 r + \sum_{c=1}^{C} r_c r_{<c} \leq \sum_{c=1}^{C} (n - n_c) r_c. \tag{2.15}$$

**Proposition 6.** *If almost all* $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \in \mathbb{A}_L^* \times \mathbb{B}_L^*$ *are the unique representative in* $\mathbb{A}_L^* \times \mathbb{B}_L^*$ *of their* $\sim_1$*-equivalence class, then Condition 3 is satisfied. Conversely, if Condition 1 is satisfied, then almost all* $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ *are the unique representative in* $\mathbb{A}_L^* \times \mathbb{B}_L^*$ *of their* $\sim_1$*-equivalence class.*

*Proof.* See Section B.5 in Appendix B for proof. □

In connection to $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$, if $\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}} \in V$ are obtained by (2.7) for $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi}_0)$ and $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \boldsymbol{\Phi}_0)$ respectively, for $(\mathbf{A}, \mathbf{B})$ and $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ in $\mathbb{A}_L^* \times \mathbb{B}_L^*$, then $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}) = \mathbf{R}_{\mathbf{xx}}(\widetilde{\boldsymbol{\eta}})$ implies $\boldsymbol{\eta} = \widetilde{\boldsymbol{\eta}}$ except on a null subset of $V$. That is, if the noise variances are known, then Condition 1, under which the signal and interference covariances can generically be separated, also yields that $\boldsymbol{\eta}$ is generically globally identifiable. The following two subsections treat the typical case where $\boldsymbol{\Phi}$ is unknown.

### Local Identifiability

For fixed $\boldsymbol{\Sigma}_{\mathbf{xx}} \in \mathcal{R}(\mathbf{n}, \mathbf{r})$, the equation $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}) = \boldsymbol{\Sigma}_{\mathbf{xx}}$ defines a quadratic system of equations in the entries of $\boldsymbol{\eta}$. As this system is nonlinear, simply counting the number of knowns in $\boldsymbol{\Sigma}_{\mathbf{xx}}$ and the number of unknowns in $\boldsymbol{\eta}$ is not sufficient to determine whether a solution is unique. However, linearization of the system by considering the differential $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ allows investigation of *local identification*. Here, local identification at $\boldsymbol{\eta}$ means that there is a neighborhood of $\boldsymbol{\eta}$ on which $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is an invertible map. We say that MFA is *generically locally identifiable* with channel sizes $\mathbf{n}$ and factor numbers $\mathbf{r}$ if almost all $\boldsymbol{\eta} \in V$ are locally identified.

As $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is a smooth map, local identification at $\boldsymbol{\eta}$ follows by the inverse function theorem if $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is injective. To assess this, a key condition follows from the count of knowns and unknowns in MFA with channel sizes $\mathbf{n}$ and factor numbers $\mathbf{r}$, as discussed in Ramírez et al. (2020, Sec. III).

**Condition 4.** *The number of common factors* $r_0$ *satisfies*

$$r_0 \leq \tfrac{1}{2}\left(2n + 1 - \sqrt{8(n+D)+1}\right), \tag{2.16}$$

*where*

$$D = \textstyle\sum_{c=1}^{C} n_c r_c - \tfrac{1}{2} r_c(r_c - 1), \tag{2.17}$$

*and for each channel $c = 1, \ldots, C$, the number of distinct factors in that channel satisfies*

$$r_c \leq \tfrac{1}{2}\big(2n_c + 1 - \sqrt{8n_c + 1}\big). \tag{2.18}$$

The following proposition, which was proven in Ramírez et al. (2020), shows that Condition 4 is *necessary* for $d\mathbf{R}_{\mathbf{xx}}$ to be injective.

**Proposition 7.** *It is* necessary *that the channel sizes* $\mathbf{n}$ *and factor numbers* $\mathbf{r}$ *satisfy Condition 4 for* $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ *to be injective at any* $\boldsymbol{\eta} \in V$.

However, ensuring that $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is *generically* injective is more challenging, as it requires examining the differential itself in addition to the dimensions of the domain and codomain. The following theorem shows that, when combined with the separability result of Proposition 4, Condition 4 is also *sufficient* for local identifiability.

**Theorem 2.** (Local Identifiability) *If the channel sizes* $\mathbf{n}$ *and factor numbers* $\mathbf{r}$ *satisfy Conditions 1 and 4, then the differential* $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ *is generically injective.*

*Proof.* See Section B.8 in Appendix B for proof. □

**Global Identifiability**

Although the local identifiability result of Theorem 2 provides valuable information about the behavior of $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ on small neighborhoods and will be needed in Section 2.4, a stronger global identifiability result for $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is desired. The following proposition combines the results of Section 2.3.2 with Proposition 4 and Theorem 1 to show that $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is an invertible map, excepting a null set of $\boldsymbol{\eta}$.

**Proposition 8.** (Global Identifiability) *If the channel sizes* $\mathbf{n}$ *and factor numbers* $\mathbf{r}$ *satisfy Conditions 1 and 2, then there exists a subset* $\widetilde{V} \subset V$ *such that* $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ *is injective on* $\widetilde{V}$ *and* $V \setminus \widetilde{V}$ *is null.*

*Proof.* See Section B.6 in Appendix B for proof. □

## 2.4 Asymptotics

### 2.4.1 Estimation

Suppose $T$ observation vectors $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are available and are i.i.d. with covariance $\boldsymbol{\Sigma}_{\mathbf{xx}} = \mathbf{R}_{\mathbf{xx}}(\mathring{\boldsymbol{\eta}})$. In this setting, Ramírez et al. (2020) presents an estimation procedure to obtain the value of $\boldsymbol{\eta}$ which maximizes the likelihood of the observations under the assumption that the latent factors and idiosyncratic errors are jointly multivariate normal. Under those distributional assumptions, the implied log density for $\mathbf{x}$ up to constant terms is

$$\log f(\mathbf{x}; \boldsymbol{\eta}) = -\frac{1}{2} \log\det \mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}) - \frac{1}{2} \mathbf{x}^{\mathsf{T}} \mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta}) \mathbf{x}. \tag{2.19}$$

With this density, estimation of $\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi}$ from $\mathbf{x}_1, \ldots \mathbf{x}_T$ is framed as the optimization problem

$$\min_{\boldsymbol{\eta} \in V} \ \log\det \mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}) + \operatorname{tr} \mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta}) \mathbf{S}_T \tag{2.20}$$

where $\mathbf{S}_T$ is the sample covariance, $\mathbf{S}_T = T^{-1} \sum_{j=1}^{T} \mathbf{x}_j \mathbf{x}_j^{\mathsf{T}}$, and the sample objective function $\ell_T$ is

$$\ell_T(\mathbf{S}_T; \boldsymbol{\eta}) \equiv \log\det \mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}) + \operatorname{tr} \mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta}) \mathbf{S}_T. \tag{2.21}$$

To avoid the Heywood cases (Lawley and Maxwell, 1962), we restrict attention to $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$ such that $\min_i [\boldsymbol{\Phi}]_{ii} \geq \epsilon$ for some fixed $\epsilon > 0$. This has the advantage of ensuring that the smallest eigenvalue $\mathbf{R}_{\mathbf{xx}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$ is bounded away from zero. Letting $V' \subset V$ contain all $\boldsymbol{\eta}$ which satisfy this additional requirement, we define the estimators $\widehat{\mathbf{A}}_T$, $\widehat{\mathbf{B}}_T$, $\widehat{\boldsymbol{\Phi}}_T$ as those obtained from the minimizer of $\ell_T(\mathbf{S}_T; \boldsymbol{\eta})$,

$$\widehat{\boldsymbol{\eta}}_T = \operatorname*{argmin}_{\boldsymbol{\eta} \in V'} \ell_T(\mathbf{S}_T; \boldsymbol{\eta}). \tag{2.22}$$

The estimators for the MFA parameters, $\widehat{\mathbf{A}}_T$, $\widehat{\mathbf{B}}_T$, $\widehat{\boldsymbol{\Phi}}_T$, are obtained by inverting (2.7) for $\widehat{\boldsymbol{\eta}}_T$.

As the latent factors and idiosyncratic errors are not observed, the assumption of joint multivariate normality can be difficult to support. Therefore, the asymptotic results of Section 2.4.2 are obtained by treating (2.21) as a *quasi-loglikelihood* (Heyde, 1997) objective function to be optimized, rather than requiring that (2.19) be the true likelihood. The results on the asymptotic

consistency and normality of the estimators do not require the joint normality of the latent factors and errors. These results instead require only mild moment assumptions, so the estimators are asymptotically valid if the latent vectors are non-normal.

### 2.4.2 Asymptotic Properties

In this section it is primarily assumed that the observation vectors $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are independent and identically distributed with mean zero and MFA covariance model (2.6),

$$\text{Var}(\mathbf{x}_1) = \boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \mathbf{R}_{\mathbf{xx}}(\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\boldsymbol{\Phi}}). \tag{2.23}$$

The true values $\mathring{\mathbf{A}}$ and $\mathring{\mathbf{B}} \equiv \text{blkdiag}(\mathring{\mathbf{B}}_1, \ldots \mathring{\mathbf{B}}_C)$ are such that $(\mathring{\mathbf{A}}, \mathring{\mathbf{B}}) \in \mathbb{A}_L^* \times \mathbb{B}_L^*$ and $[\mathring{\boldsymbol{\Phi}}]_{ii} > \epsilon$ for all $i = 1, \ldots, n$. The vectorization (2.7) of $(\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\boldsymbol{\Phi}})$ is $\mathring{\boldsymbol{\eta}} \in V'$. The higher moments of $\mathbf{x}_1$ are not specified, and in particular the observations need not be normally distributed. Theorem 3 also speaks to the misspecified case where $\boldsymbol{\Sigma}_{\mathbf{xx}} \notin \mathcal{R}(\mathbf{n}, \mathbf{r})$.

The next theorem shows that identification of the true $\boldsymbol{\Sigma}_{\mathbf{xx}}$ is enough to ensure that the estimators are consistent for the factor loading parameters $\mathbf{A}$ and $\mathbf{B}$ and the idiosyncratic noise variance $\boldsymbol{\Phi}$. This follows from the fact that $\mathbf{x}$ has finite second moment and the exclusion of singular covariance models in the definition of the parameter space. The objective function $\ell_T$ is then sufficiently well-behaved so that the maximizer $\widehat{\boldsymbol{\eta}}_T$ of $\ell_T$ converges to the maximizer of $\ell_0 \equiv E[\ell_T]$, which is $\mathring{\boldsymbol{\eta}}$. Convergence of $\widehat{\mathbf{A}}_T, \widehat{\mathbf{B}}_T, \widehat{\boldsymbol{\Phi}}_T$ to $\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\boldsymbol{\Phi}}$ then follows.

**Theorem 3.** (Consistency) *Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are an i.i.d. sequence of random vectors with $E[\mathbf{x}_1] = \mathbf{0}$ and positive definite $\text{Var}(\mathbf{x}_1) \equiv \boldsymbol{\Sigma}_{\mathbf{xx}}$. If there exists a unique $\mathring{\mathbf{R}}_{\mathbf{xx}} \in \mathcal{R}(\mathbf{n}, \mathbf{r})$ minimizing $D_{KL}\big(\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{xx}}) || \, \mathcal{N}(\mathbf{0}, \mathring{\mathbf{R}}_{\mathbf{xx}})\big)$ with $\mathring{\mathbf{R}}_{\mathbf{xx}} = \mathbf{R}_{\mathbf{xx}}(\mathring{\boldsymbol{\eta}})$ for $\mathring{\boldsymbol{\eta}} \in V'$ in the interior of the globally identified set $\widetilde{V}$ defined in Proposition 8, then $\widehat{\mathbf{A}}_T, \widehat{\mathbf{B}}_T, \widehat{\boldsymbol{\Phi}}_T$ converge in probability to $\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\boldsymbol{\Phi}}$ respectively.*

*Proof.* See Section B.9 in Appendix B for proof. □

In particular, if the model is correctly specified with $\text{Var}(\mathbf{x}_1) = \mathbf{R}_{\mathbf{xx}}(\mathring{\boldsymbol{\eta}})$ for $\mathring{\boldsymbol{\eta}}$ in the interior of the globally identified set, then the estimators $\widehat{\mathbf{A}}_T, \widehat{\mathbf{B}}_T, \widehat{\boldsymbol{\Phi}}_T$ are consistent.

The following theorem shows that the estimators have a limiting Gaussian distribution when the observation distribution has a finite fourth moment. This is obtained from consistency of the estimators and the nature of $\ell_T$ in a neighborhood of the covariance $\mathbf{R}_{\mathbf{xx}}(\mathring{\boldsymbol{\eta}})$. In particular, Theorem 2 is used to show that the objective generically has a positive second differential, and so the limiting covariance is positive definite. As the limiting distribution is non-degenerate, $\widehat{\boldsymbol{\eta}}_T - \mathring{\boldsymbol{\eta}}$ converges to zero in probability at the standard parametric rate of $T^{-1/2}$.

**Theorem 4.** (Asymptotic Normality) *Assume the conditions of Theorem 3 are satisfied with* $\boldsymbol{\Sigma}_{\mathbf{xx}} = \mathbf{R}_{\mathbf{xx}}(\mathring{\boldsymbol{\eta}})$. *In addition, assume that* $\mathbf{x}_1$ *satisfies* $E[||\mathbf{x}_1||^4] < \infty$. *Under these conditions, the estimated parameters* $\widehat{\boldsymbol{\eta}}_T$ *converges in distribution as*

$$\sqrt{T}(\widehat{\boldsymbol{\eta}}_T - \mathring{\boldsymbol{\eta}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}_L, \mathbf{W}) \tag{2.24}$$

*for positive-definite matrix* $\mathbf{W}$ *from* (B.20) *in Appendix B.*

*Proof.* See Section B.10 in Appendix B for proof. □

## 2.5 Experiments

### 2.5.1 Numeric Comparison of Conditions

In Section 2.3, conditions on the channel sizes and factor numbers and their implications for MFA identifiability are given. To gain intuition for how varying channel sizes and factor numbers affects the satisfaction of these conditions, Figure 2.3 depicts the illustrative case of three equal-size channels. The figure compares Condition 3, which is *necessary* for identifiability, to the hypotheses of Theorem 2 and Proposition 8 which are respectively *sufficient* for generic local and global identifiability. To interpret Figure 2.3, examine channel size $n_1 = 15$ in the middle panel with $r_c = 5$. In this case, Proposition 7 implies that, for $r_0 > 25$, the set of $\sim_2$-representatives $\mathbb{A}_L^* \times \mathbb{B}_L^*$ does not contain unique representatives of almost all $\sim_1$-equivalence classes, preventing the separation of signal and interference. This is indicated by the circle at $n_1 = 15$ and $r_0 = 25$. Further, 19 is the maximum $r_0$ which guarantees local identifiability under Theorem 2 as shown by

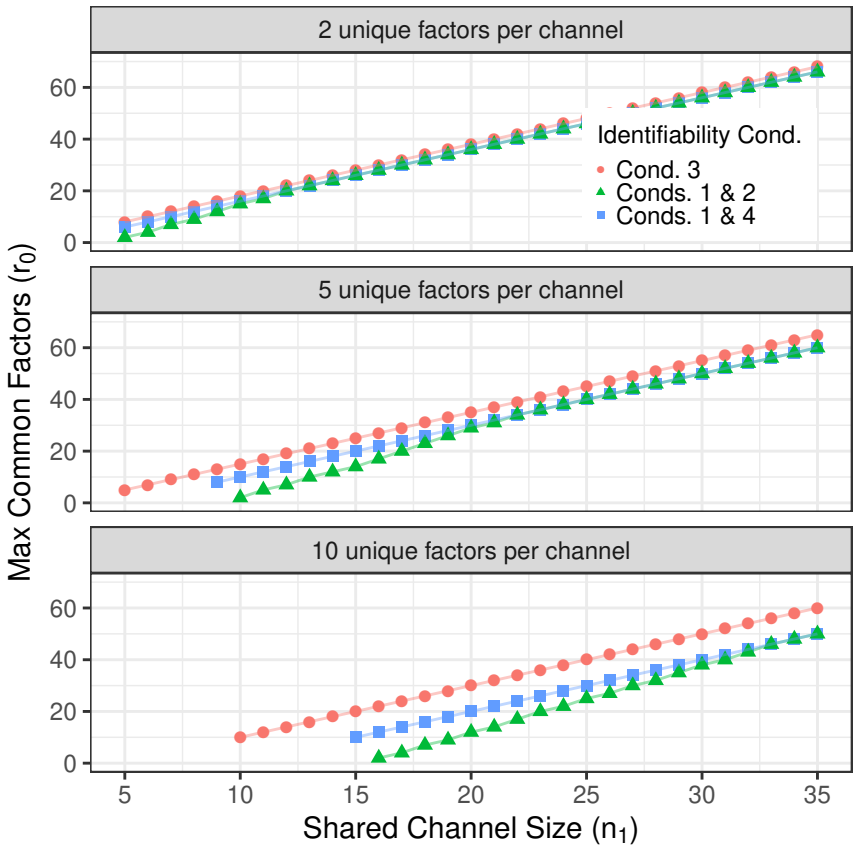



**Figure 2.3:** Comparison of maximum common factor number $r_0$ under three identifiability conditions, for varying channel sizes. Channel structure depicted is three equally-sized channels, with $r_c = 2, 5, 10$ distinct factors for $c = 1, 2, 3$.

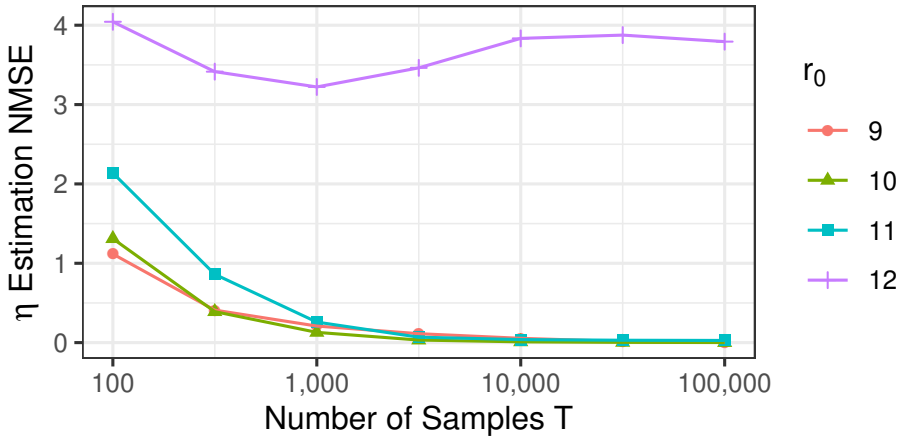



**Figure 2.4:** Experimental validation of Theorem 3 for varying common factor number $r_0$, with $C = 3$ channels of size $n_c = 8$ and $r_c = 2$ distinct factors, $c = 1, 2, 3$. Points indicate average NMSE from 1000 Monte Carlo trials at each setting.

the square at $n_1 = 15$ and $r_0 = 19$. Finally, 14 is the largest $r_0$ which yields global identifiability under Proposition 8, which is indicated by the triangle at $n_1 = 15$ and $r_0 = 14$. The maximum $r_0$ for which generic global and local identifiability can be respectively guaranteed under Proposition 8 and Theorem 2 agree as channel size increases, but the channel size for which the condition agree increases as the distinct factor number increases. In addition, the gap between Condition 3 (which is necessary for identifiability) and Conditions 1 & 4 is constant in the shared channel size and small relative to the total factor number $r_0 + r$. Although the results of this paper give only sufficient conditions for local and global identifiability, the experiment in this section demonstrates that these conditions are close to Condition 3 which is an upper-bound for MFA identifiability. For further discussion and comparisons with unequal channels, see Stanton et al. (2023a).

### 2.5.2 Asymptotic Behavior of Estimators

To verify the consistency of $\widehat{\boldsymbol{\eta}}_T$ resulting from Theorem 3, Figure 2.4 shows the Normalized Mean Square Error (NMSE) $||\widehat{\boldsymbol{\eta}}_T - \mathring{\boldsymbol{\eta}}||^2 / ||\mathring{\boldsymbol{\eta}}||^2$ when the model is correctly specified and the channel sizes and factor numbers satisfy Conditions 1 & 2. For each trial, non-zero entries of the true parameters $\mathring{\mathbf{A}} \in \mathbb{A}_L^*, \mathring{\mathbf{B}} \in \mathbb{B}_L^*$ and $\mathring{\boldsymbol{\Phi}}$ are independent $\mathcal{N}(0, 1)$ samples. For entries constrained to be non-negative, the absolute value is taken. Initial values for the estimation procedure of Ramírez et al. (2020) are independently obtained in the same fashion, and $\widehat{\boldsymbol{\eta}}_T$ is computed from $T$ independent samples with covariance $\boldsymbol{\Sigma}_{\mathbf{xx}}$.

For $C = 3$ channels with $n_c = 8$, Figure 2.3 shows that the largest $r_0$ meeting Conditions 1 & 2 is $r_0 = 9$ while the largest $r_0$ meeting Conditions 1 & 4 is $r_0 = 12$. The decreasing NMSE in Figure 2.4 for $r_0 = 9$ verifies that the parameters $\boldsymbol{\eta}$ can be consistently estimated when global identifiability is guaranteed, while the non-decreasing NMSE for $r_0 = 12$ shows that local identifiability alone is insufficient for consistency. The decreasing NMSE for intermediate cases $9 < r_0 < 12$ may indicate that MFA is globally identifiable for those factor numbers, but this is not given by Proposition 8.

## 2.6 Discussion

This chapter provides a set of theoretical results for multi-channel factor analysis, which justify applying MFA to analyze the second-order structure of multi-channel observations. Conditions on the allowable number of common and distinct factors which guarantee generic uniqueness of the decomposition of the covariance into across-channel, within-channel, and idiosyncratic components are set out in Section 2.3. These identifiability results ensure that conclusions drawn from MFA are meaningful as long as the channel sizes and factor numbers satisfy the appropriate conditions. Further, although the estimation procedure proposed in Ramírez et al. (2020) is obtained by likelihood maximization under the assumption of normality for the latent vectors, the results of Section 2.4 demonstrate that violation of this assumption does not affect the asymptotic validity of the resulting estimators.

When introducing multi-channel factor analysis, Ramírez et al. (2020) discusses the broad potential applicability of the MFA model to diverse problems in signal processing, statistics, and machine learning where channel structure is a relevant feature. The promise of this method comes from the utility of the decomposition of the observation covariance into structured parts corresponding to the latent signal, interference and noise, the uniqueness of which can now be verified.

# Chapter 3

# Multi-Channel Factor Spectral Analysis for Multiple Time Series

## 3.1 Chapter Overview

The previous chapter's discussion of MFA investigates the setting where the structured cross-sectional correlation within and between the observation channels is estimated based on repeated, temporally independent measurements of the channels. However, in many signal processing applications, there is substantial temporal correlation from measurement to measurement. This chapter introduces the method of Multi-channel Factor Spectral Analysis (MFSA) as an extension to MFA (Section 3.2), and the identifiability and parameterization properties are examined. A frequency-domain Gaussian quasi-likelihood (Section 3.3) is used to quantify model fit, and a generalized Majorization-Minimization (MM) procedure is developed for parameter estimation (Section 3.4). A simulation study is conducted (Section 3.5), which demonstrates the improved performance of MFSA over MFA when the measurements are temporally correlated for several problems of interest. Parts of the work in this chapter have previously been published as Stanton et al. (2023b).

## 3.2 MFSA Model

### 3.2.1 Description

The observations of interest are a collection of $C$ discrete time series $\mathbf{x}_1[t], \ldots, \mathbf{x}_C[t]$, $t \in \mathbb{Z}$, where $\mathbf{x}_c[t_0] \in \mathbb{C}_{n_c}$ is a vector of $n_c$ measurements collected at time $t_0$. Each $\mathbf{x}_c[t]$ may be obtained as the series of readings obtained by polling an array of $n_c$ scalar sensors at evenly spaced times that are common to each sensor array. The grouping of the scalar measurements into $C$ distinct MIMO *channels* reflects some structure of *a priori* interest in the data. For example,

the scalar measurements may be grouped by sensor modality, where the scalar measurements in $\mathbf{x}_1[t]$ may have some shared physical meaning (such as all being voltage measurements) while the scalar measurements in $\mathbf{x}_2[t]$ have a distinct meaning (such as all being current measurements). Alternatively, the scalar measurements in $\mathbf{x}_c[t]$ may not have a shared physical meaning but are instead tied together by some detail of the data collection process, such as all being measured at the same location. Grouping the scalar measurements into $C$ channels and jointly analyzing the resulting intra- and cross-channel correlations is at the heart of multi-channel methods, as discussed in Section 1.2.1. This approach is distinct from the two limiting "single-channel" approaches, namely grouping all scalar measurements together into one channel (which ignores the structure of *a priori* interest) or analyzing each $\mathbf{x}_c[t]$ separately (which ignores all potentially interesting cross-channel correlations).

In MFSA, the channel-$c$ observation series $\mathbf{x}_c[t]$ is modeled as the linear combination of the three latent series $\mathbf{s}_c[t], \mathbf{i}_c[t], \mathbf{u}_c[t] \in \mathbb{C}_{n_c}$,

$$\mathbf{x}_c[t] = \mathbf{s}_c[t] + \mathbf{i}_c[t] + \mathbf{u}_c[t]. \tag{3.1}$$

The first series, $\mathbf{s}_c[t]$, represents the manifestation of some unknown *signal* of interest within channel $c$. The second series, $\mathbf{i}_c[t]$, represents the effects of *interference* within channel $c$, while is temporally correlated and affects multiple scalar measurements in channel $c$. The third series, $\mathbf{u}_c[t]$, represents temporally-uncorrelated measurement *noise* in the individual scalar measurements. The distinguishing property between $\mathbf{s}_c[t]$ and $\mathbf{i}_c[t]$ is that the signals in channels $c$ and $c'$, namely $\mathbf{s}_c[t]$ and $\mathbf{s}_{c'}[t]$, are assumed to derive from some shared underlying phenomenon and are therefore cross-sectionally correlated. In contrast, the interferences in channels $c$ and $c'$, namely $\mathbf{i}_c[t]$ and $\mathbf{i}_{c'}[t]$, derive from unrelated phenomena and therefore the $\mathbf{i}_c[t]$ and $\mathbf{i}_{c'}[t]$ are assumed to have no cross-sectional correlation. By the assumption that the underlying phenomenon of interest affects all channels, its effects can be disentangled from those of the interference and noise by examining the commonalities present across all channels, allowing for approximate recovery of $\mathbf{s}_1[t], \ldots, \mathbf{s}_C[t]$ based on $\mathbf{x}_1[t], \ldots, \mathbf{x}_C[t]$ alone.

To jointly analyze the series of measurements obtained from the $C$ channels, it is convenient to combine $\mathbf{x}_1[t], \ldots, \mathbf{x}_C[t]$ into a single *all-channel* measurement series $\mathbf{x}[t] \in \mathbb{C}_n$, as

$$\mathbf{x}[t] = \begin{bmatrix} \mathbf{x}_1^{\mathsf{T}}[t] & \ldots & \mathbf{x}_C^{\mathsf{T}}[t] \end{bmatrix}^{\mathsf{T}}, \; t \in \mathbb{Z},$$

where $n = \sum_{c=1}^{C} n_c$ is the total number of scalar measurements collected across all channels. Corresponding all-channel signal, interference, and noise series are denoted by $\mathbf{s}[t], \mathbf{i}[t]$ and $\mathbf{u}[t]$ respectively and are similarly obtained by vertical concatenation. With these definitions, the $C$ equations contained in (3.1) correspond to the all-channel observation equation,

$$\mathbf{x}[t] = \mathbf{s}[t] + \mathbf{i}[t] + \mathbf{u}[t]. \tag{3.2}$$

For the analysis to follow, it is convenient to introduce the notion of *channel-patterned* vectors and matrices. A vector $\mathbf{v} \in \mathbb{C}_n$ is channel-patterned if it is partitioned into $C$ subvectors of lengths $n_1, \ldots, n_C$ respectively as $\mathbf{v} = [\mathbf{v}_1^{\mathsf{T}} \; \ldots \; \mathbf{v}_C^{\mathsf{T}}]^{\mathsf{T}}$. A Hermitian positive semi-definite matrix $\mathbf{M} \in \mathbb{C}_{n \times n}$ is channel-patterned if it is partitioned into $C^2$ submatrices as

$$\mathbf{M} = \begin{array}{c} \begin{matrix} n_1 & \cdots & \cdots & n_C \end{matrix} \\ \begin{bmatrix} \mathbf{M}_{11} & \ldots & \mathbf{M}_{1C} \\ \vdots & \ddots & \vdots \\ \mathbf{M}_{1C}^{\mathsf{H}} & \ldots & \mathbf{M}_{CC} \end{bmatrix} \end{array} \begin{matrix} n_1 \\ \vdots \\ \\ \\ \\ \vdots \\ n_C \end{matrix}.$$

This patterning reflects that the measurements originate from $C$ distinct channels, which is a nuance that should be preserved in subsequent multi-channel analyses. In the notation of Chapter 1, channel-structured matrices arise from the action of the corresponding operator on the decomposition $\mathcal{C}_1 \boxplus \cdots \boxplus \mathcal{C}_C$ of the all-channel space $\mathbb{C}_n$ into the channel subspaces. The important distinction between intra-channel and cross-channel correlations is represented by differing treatments of on-diagonal and off-diagonal blocks in channel-patterned matrices. For notational convenience, let $\mathbf{n} = [n_1, \ldots, n_C]^{\mathsf{T}} \in \mathbb{N}_C$ be the vector of channel sizes.

To make the expression (3.2) of $\mathbf{x}[t]$ as a sum of three latent series meaningful, additional assumptions must be imposed on $\mathbf{s}[t], \mathbf{i}[t]$ and $\mathbf{u}[t]$ to distinguish the latent series. The assumption

which will be used to distinguish $\mathbf{s}[t]$ from $\mathbf{i}[t]$ is that the received signals in different channels are cross-sectionally correlated while the interferences in different channels are not. To distinguish the *systemic* part $\mathbf{s}[t] + \mathbf{i}[t]$ from the *noise* part $\mathbf{u}[t]$, we assume that $\mathbf{s}[t] + \mathbf{i}[t]$ belongs to some unknown *low-dimensional* subspace which is the same for all $t \in \mathbb{Z}$. Therefore, the part of $\mathbf{x}[t]$ which is orthogonal to this subspace must be due entirely to $\mathbf{u}[t]$. If this subspace and the second moments of $\mathbf{s}[t], \mathbf{i}[t]$ and $\mathbf{u}[t]$ were known, then the systemic part $\mathbf{s}[t] + \mathbf{i}[t]$ can be estimated by an appropriate projection of $\mathbf{x}[t]$ onto the subspace of the systemic part, allowing for $\mathbf{s}[t] + \mathbf{i}[t]$ and $\mathbf{u}[t]$ to be distinguished.

### 3.2.2 Second-Order Properties of the Series

Like MFA, multi-channel factor spectral analysis is a second-order method in that the first moments of the latent series are assumed to be previously handled by centering and removal of deterministic components, while the higher moments of the latent series are not relevant to the method (aside from mild regularity assumptions). Accordingly, we assume that $\mathbf{s}[t]$, $\mathbf{i}[t]$ and $\mathbf{u}[t]$ are wide-sense stationary and completely nondeterministic (i.e. linearly regular), as well as the following first and second moment assumptions. First, all the latent series are mean zero,

$$E[\mathbf{s}[0]] = \mathbf{0}_n \qquad E[\mathbf{u}[0]] = \mathbf{0}_n \qquad E[\mathbf{i}_c[0]] = \mathbf{0}_{n_c} \tag{3.3}$$

for $c = 1, \ldots, C$. Next, we assume that the latent series are uncorrelated across all lags,

$$E[\mathbf{s}[s]\mathbf{i}^{\mathsf{H}}[t]] = \mathbf{0}_{n,n_c}, \quad E[\mathbf{s}[s]\mathbf{u}^{\mathsf{H}}[t]] = \mathbf{0}_{n,n}, \quad E[\mathbf{s}[s]\mathbf{i}^{\mathsf{T}}[t]] = \mathbf{0}_{n,n_c}, \quad E[\mathbf{s}[s]\mathbf{u}^{\mathsf{T}}[t]] = \mathbf{0}_{n,n},$$
$$E[\mathbf{i}_c[s]\mathbf{i}_{c'}^{\mathsf{H}}[t]] = \mathbf{0}_{n_c,n_{c'}} \quad E[\mathbf{i}_c[s]\mathbf{u}^{\mathsf{H}}[t]] = \mathbf{0}_{n_c,n} \quad E[\mathbf{i}_c[s]\mathbf{i}_{c'}^{\mathsf{T}}[t]] = \mathbf{0}_{n_c,n_{c'}} \quad E[\mathbf{i}_c[s]\mathbf{u}^{\mathsf{T}}[t]] = \mathbf{0}_{n_c,n}$$

for all $t, s \in \mathbb{Z}$ and $c, c' = 1, \ldots, C$ with $c \neq c'$. Further, the latent series are assumed to be proper (Neeser and Massey, 1993) and hence have vanishing psuedo-autocovariance

$$E[\mathbf{s}[s]\mathbf{s}^{\mathsf{T}}[t]] = \mathbf{0}_{n,n}, \qquad E[\mathbf{u}[s]\mathbf{u}^{\mathsf{T}}[t]] = \mathbf{0}_{n,n} \qquad E[\mathbf{i}_c[s]\mathbf{i}_c^{\mathsf{T}}[t]] = \mathbf{0}_{n_c,n_c},$$

for all $t, s \in \mathbb{Z}$ and $c = 1, \ldots, C$. Assumptions of the circularity or propriety of complex signals are common but can oversimplify underlying phenomena of interest (Adali et al., 2011). Nevertheless, the assumption of propriety is commonly made for factor analysis in signal processing (Sardarabadi and van der Veen, 2018) and we continue to make that assumption. However, requiring that the latent series are proper is less crucial to MSFA than the assumption that they are pairwise uncorrelated, as the estimates of the measurement PSD remains valid for improper measurement series.

The low-dimensionality assumption on $\mathbf{s}[t]$ and $\mathbf{i}_c[t], c = 1, \ldots, C$ is imposed by assuming that there exist subspaces $\mathcal{S} \subset \mathbb{C}_n$ and $\mathcal{I}_c \subset \mathbb{C}_{n_c}$ such that

$$\mathbf{P}_{\mathcal{S}}\mathbf{s}[t] = \mathbf{s}[t] \qquad \mathbf{P}_{\mathcal{I}_c}\mathbf{i}_c[t] = \mathbf{i}_c[t], \qquad \mathcal{S} \cap (\mathcal{I}_1 \oplus \cdots \oplus \mathcal{I}_C) = \{\mathbf{0}\}, \tag{3.4}$$

for all $t \in \mathbb{Z}$ and $c = 1, \ldots, C$, where $\mathbf{P}_{\mathcal{S}}$ and $\mathbf{P}_{\mathcal{I}_c}$ are projections onto $\mathcal{S}$ and $\mathcal{I}_c$ respectively. The dimensions of the subspaces are taken to be $r_0 \equiv \dim \mathcal{S}$ and $r_c \equiv \dim \mathcal{I}_c$, and for later clarity we define the total interference dimension $r \equiv \sum_{c=1}^{C} r_c$ and the vector of subspace dimensions $\mathbf{r} = [r_0, r_1, \ldots, r_C]^{\mathsf{T}}$. The assumption that the signal subspace $\mathcal{S}$ and interference subspace $\mathcal{I}_1 \oplus \cdots \oplus \mathcal{I}_C$ intersect trivially ensures that if $\mathcal{S}$ and $\mathcal{I}_c,\ c = 1, \ldots, C$ are known, the systemic part $\mathbf{s}[t] + \mathbf{i}[t]$ can be uniquely separated into $\mathbf{s}[t]$ and $\mathbf{i}[t]$ for all values of $\mathbf{s}[t] + \mathbf{i}[t]$. It is satisfied when $r_0 + r \leq n$ and the signal and interference subspaces are in generic position with respect to each other.

The measurement noise $\mathbf{u}[t]$ is not required to live in any subspace, but is required to be white noise with diagonal variance $\mathbf{\Phi}$,

$$E[\mathbf{u}[0]\mathbf{u}^{\mathsf{H}}[0]] \equiv \mathbf{\Phi} \equiv \operatorname{blkdiag}(\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_C),$$

where $\mathbf{\Phi}_c \in \mathbb{C}_{n_c \times n_c}$ is the diagonal variance of $\mathbf{u}_c[t]$.

The above first and second moment assumptions on the latent series allow second-order quantities of interest to be usefully described with the following Wold representations.

**Proposition 9.** *Under the above assumptions on* $\mathbf{s}[t]$, $\mathbf{u}[t]$, *and* $\mathbf{i}_c[t],\ c = 1, \ldots, C$ *the series are uniquely represented as*

$$\begin{aligned}
\mathbf{s}[t] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\omega t}\mathcal{K}_{\mathbf{s}}(\mathrm{e}^{\mathrm{i}\omega})d\mathbf{Z}_0(\omega), \\
\mathbf{i}_c[t] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\omega t}\mathcal{K}_{\mathbf{i}_c}(\mathrm{e}^{\mathrm{i}\omega})d\mathbf{Z}_c(\omega),\ c = 1, \ldots, C, \\
\mathbf{u}[t] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\omega t}\mathbf{\Phi}^{\frac{1}{2}}d\mathbf{W}(\omega),
\end{aligned} \tag{3.5}$$

*where* $\mathcal{K}_{\mathbf{s}}(z) = \sum_{j=0}^{\infty}\mathbf{K}_{\mathbf{s},j}z^{-j}$ *and* $\mathcal{K}_{\mathbf{i}_c}(z) = \sum_{j=0}^{\infty}\mathbf{K}_{\mathbf{i}_c,j}z^{-j}$ *are* $n \times r_0$ *and* $n_c \times r_c$ *matrix transfer functions, while* $\mathbf{Z}_0(\omega)$, $\mathbf{Z}_c(\omega), c = 1, \ldots, C$, *and* $\mathbf{W}(\omega)$ *are orthogonal increments processes with*

$$\begin{aligned}
E[d\mathbf{Z}_0(\omega)d\mathbf{Z}_0^{\mathsf{H}}(\omega)] &= \mathbf{I}_{r_0}d\omega, & E[d\mathbf{Z}_0(\omega)d\mathbf{Z}_0^{\mathsf{T}}(\omega)] &= \mathbf{0}d\omega, \\
E[d\mathbf{Z}_c(\omega)d\mathbf{Z}_c^{\mathsf{H}}(\omega)] &= \mathbf{I}_{r_0}d\omega, & E[d\mathbf{Z}_c(\omega)d\mathbf{Z}_c^{\mathsf{T}}(\omega)] &= \mathbf{0}d\omega, \\
E[d\mathbf{W}(\omega)d\mathbf{W}^{\mathsf{H}}(\omega)] &= \mathbf{I}_{n}d\omega, & E[d\mathbf{W}(\omega)d\mathbf{W}^{\mathsf{T}}(\omega)] &= \mathbf{0}d\omega,
\end{aligned}$$

*for all* $\omega \in (-\pi, \pi]$ *and which are pairwise uncorrelated. The transfer functions have* $\operatorname{col}\mathcal{K}_{\mathbf{s}}(z) \subset \operatorname{col}\mathbf{K}_{\mathbf{s},0}$ *and* $\operatorname{col}\mathcal{K}_{\mathbf{i}_c}(z) \subset \operatorname{col}\mathbf{K}_{\mathbf{i}_c,0}$ *for all* $z \in \mathbb{C}$, *and the top* $r_0 \times r_0$ *block of* $\mathbf{K}_{\mathbf{s},0}$ *and top* $r_c \times r_c$ *block of* $\mathbf{K}_{\mathbf{i}_c,0},\ c = 1, \ldots, C$ *are lower-triangular with positive main diagonal.*

*Proof.* See Section C.1 in Appendix C for proof. □

The Power Spectral Densities (PSDs) of the various latent series are central to MFSA, and are defined using the above proposition for $\omega \in (-\pi, \pi]$ as

$$\mathbf{S}_{\mathbf{ss}}(\omega) = \mathcal{K}_{\mathbf{s}}(\mathrm{e}^{\mathrm{i}\omega})\mathcal{K}_{\mathbf{s}}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega}), \qquad \mathbf{S}_{\mathbf{i}_c\mathbf{i}_c}(\omega) = \mathcal{K}_{\mathbf{i}_c}(\mathrm{e}^{\mathrm{i}\omega})\mathcal{K}_{\mathbf{i}_c}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega}).$$

The assumption (3.4) that the latent series are uncorrelated ensures that the PSD of the all-channel interference is

$$\mathbf{S}_{\mathbf{ii}}(\omega) = \operatorname{blkdiag}\left(\mathbf{S}_{\mathbf{i}_1\mathbf{i}_1}(\omega), \ldots, \mathbf{S}_{\mathbf{i}_C\mathbf{i}_C}(\omega)\right),$$

and so the all-channel measurement PSD is the channel-patterned matrix function,

$$\mathbf{S}_{\mathbf{xx}}(\omega) = \mathbf{S}_{\mathbf{ss}}(\omega) + \mathbf{S}_{\mathbf{ii}}(\omega) + \mathbf{\Phi}. \tag{3.6}$$

Let $\mathcal{D}(\mathbf{n}, \mathbf{r})$ denote the set of MFSA power spectral densities, which can be obtained from (3.6) for some latent series $\mathbf{s}[t], \mathbf{i}[t]$ and $\mathbf{u}[t]$ satisfying the assumptions of Proposition 9.

### 3.2.3 Parameterization

To estimate the power spectral densities $\mathbf{S_{ss}}(\omega), \mathbf{S_{ii}}(\omega)$ and $\mathbf{\Phi}$ of the latent series using measurements from all channels at a finite number of times, it is convenient to define a finite dimensional parameterization for some subset of all $\mathbf{S_{xx}}(\omega)$ obtained by (3.6).

To do so, let $\mathbb{L}_{n\times m} \subset \mathbb{C}_{n\times m}$ for $n \geq m$ be the subspace of $n \times m$ complex matrices whose top $m \times m$ block is lower-triangular. Further define the affine subspace $\mathbb{L}^1_{n\times m} \subset \mathbb{L}_{n\times m}$ as the subset where the top $m \times m$ block is unit lower-triangular (that is, the entries on the main diagonal are $1$), and the subspace $\mathbb{L}^0_{n\times m}$ as the subset where the top $m \times m$ block is strictly lower-triangular (that is, the entries on the main diagonal are $0$).

In a slight abuse of notation, let $\mathbb{L}_{n\times m\times p}$ be the space of $n \times pm$ complex matrices which are obtained by horizontally stacking $p$ elements in $\mathbb{L}_{n\times m}$. That is, if $\mathbf{L} \in \mathbb{L}_{n\times m\times p}$ then

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_1 & \dots & \mathbf{L}_p \end{bmatrix}$$

where $\mathbf{L}_j \in \mathbb{L}_{n\times m}$ for $j = 1, \dots, p$. The subset $\mathbb{L}^1_{n\times m\times p}$ is the subset of $\mathbb{L}_{n\times m\times p}$ where each block $\mathbf{L}_j$ belongs to $\mathbb{L}^1_{n\times m}$, and $\mathbb{L}^0_{n\times m\times p}$ is defined similarly. Each element in $\mathbb{L}_{n\times m\times p}$ is then associated with a $n \times m$ matrix of polynomials in $z^{-1}$ of maximum degree $p$ using $\mathbf{u}_p(z) = [z^{-1}, \dots, z^{-p}]^\mathsf{T}$ as

$$\mathbf{L}(z) \equiv \begin{bmatrix} \mathbf{I}_m \\ \mathbf{0}_{(n-m),m} \end{bmatrix} + \mathbf{L}(\mathbf{u}_p(z) \otimes \mathbf{I}_n). \tag{3.7}$$

Using the associated polynomials, a subset $\mathbb{L}^{<1}_{n\times m\times p} \subset \mathbb{L}_{n\times m\times p}$ can be defined by requiring that the top-block main diagonal polynomials $[\mathbf{L}(z)]_{jj},\ j = 1, \dots, m$ be Schur-stable. That is, $\mathbf{L} \in \mathbb{L}^{<1}_{n\times m\times p} \subset \mathbb{L}_{n\times m\times p}$ if the roots of $[\mathbf{L}(z)]_{jj}$ have absolute value $< 1$ for all $j = 1, \dots, m$. These component polynomials are of particular interest, as if $z$ is not a root for any $[\mathbf{L}(z)]_{jj},\ j = 1, \dots, C$, then $\mathbf{L}(z)$ is full rank. Similarly, define $\mathbb{L}^{SE<1}_{n\times m\times p} \subset \mathbb{L}_{n\times m\times p}$ such that only the south-east top-block polynomial $[\mathbf{L}(z)]_{mm}$ is required to be Schur-stable and the other polynomials are unrestricted.

Computing the measurement spectral density requires inversion of a matrix of the form $\mathbf{L}(\mathrm{e}^{\mathrm{i}\omega})$, and restricting attention to some subset $\mathbb{L}^{\leq\kappa,B}_{m\times m\times p}$ of $\mathbb{L}^{<1}_{m\times m\times p}$ can ensure that this matrix is sufficiently well-conditioned for all $\omega$. To define this subset, choose $0 \leq \kappa < 1$ and $0 \leq B < \infty$, and write $\mathbf{L} = \mathbf{L}' + \mathbf{L}''$ where $\mathbf{L}' \in \mathbb{L}^{0}_{m\times m\times p}$ contains the strictly lower-triangular part of each coefficient matrix and $\mathbf{L}'' = [\mathbf{L}''_1 \ \ldots \ \mathbf{L}''_p]$ contains the remaining diagonal part of each coefficient matrix. Then $\mathbf{L} \in \mathbb{L}^{<1}_{m\times m\times p}$ is in $\mathbb{L}^{\kappa,B}_{m\times m\times p}$ if the roots of the main diagonal polynomials $[\mathbf{L}''(z)]_{jj}$ have absolute value $\leq \kappa$ for $j = 1, \ldots, m$ and $||\mathbf{L}'||_{2,\infty} \leq B$, where $||\mathbf{X}||_{2,\infty}$ is the maximum $\ell_2$ norm of the rows of $\mathbf{X}$. If $\mathbf{L} \in \mathbb{L}^{\kappa,B}_{m\times m\times p}$, then the condition number $\mathrm{cond}\left(\mathbf{L}(\mathrm{e}^{\mathrm{i}\omega})\right)$ is bounded above (Guggenheimer et al., 1995) as

$$\mathrm{cond}\left(\mathbf{L}(\mathrm{e}^{\mathrm{i}\omega})\right) \leq \frac{2((1+\kappa)^p + mpB)^{m/2}}{(1-\kappa)^{mp}} \tag{3.8}$$

where $v_{jk},\ k = 1, \ldots p$ are the zeros of $[\mathbf{L}(z)]_{jj}$ and

$$\begin{aligned} |\det\left(\mathbf{L}(\mathrm{e}^{\mathrm{i}\omega})\right)| &\geq \textstyle\prod_{j=1}^{m}\prod_{k=1}^{p}|(1 - v_{jk}\mathrm{e}^{-\mathrm{i}\omega})| \\ &\geq (1_\kappa)^{mp} \\ ||\mathbf{L}(\mathrm{e}^{\mathrm{i}\omega})||_F^2 &\leq \textstyle\sum_{j=1}^{m}\left|[\mathbf{L}(\mathrm{e}^{\mathrm{i}\omega})]_{jj}\right|^2 + m^2p||\mathbf{L}''||_{2,\infty}^2 \\ &\leq m(1+\kappa)^p + m^2pB \end{aligned}$$

as $||\mathbf{u}_p(\mathrm{e}^{\mathrm{i}\omega}) \otimes \mathbf{I}_m||_F$ is $\sqrt{mp}$, so $||\mathbf{L}''(\mathbf{u}_p(z) \otimes \mathbf{I}_m)||_F \leq m\sqrt{p}||\mathbf{L}''||_{2,\infty}$. In a similar vein, define the subset $\mathbb{L}^{SE\leq\kappa}_{m\times m\times p} \subset \mathbb{L}^{SE<1}_{m\times m\times p}$ as the set where the south-east polynomial $[\mathbf{L}(z)]_{mm}$ has roots with magnitude at most $\kappa$, and $\mathbb{L}^{SE=1}_{m\times m\times p}$ where the south-east coefficient $[\mathbf{L}_j]_{mm}$ equals 1 for each coefficient block $\mathbf{L}_j,\ j = 1, \ldots, p$.

Next, define $\mathbb{P}_n \subset \mathbb{C}_{n\times n}$ to be the set of $n \times n$ diagonal matrices with positive diagonal entries. To avoid numeric underflow in the estimation procedure, we will work within the further subset $\mathbb{P}_n^{\geq\epsilon} \subset \mathbb{P}_n$, where the diagonal entries are $\geq \epsilon$ for some $\epsilon > 0$.

Let $\mathbf{r}, \mathbf{p}, \mathbf{q}$ be three vectors in $\mathbb{N}_{C+1}$, namely $\mathbf{r} = [r_0, r_1, \ldots, r_C]^\mathsf{T}$, $\mathbf{p} = [p_0, p_1, \ldots, p_C]^\mathsf{T}$ and $\mathbf{q} = [q_0, q_1, \ldots, q_C]^\mathsf{T}$, which contain the subspace dimensions, numerator orders, and denominator

orders respectively. The parameter space $\boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ is then defined as the Cartesian product,

$$\begin{aligned}\boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q}) =& \mathbb{P}_n \times \mathbb{L}^1_{n\times r_0} \times \mathbb{L}^{SE<1}_{r_0\times r_0\times p_0} \times \mathbb{L}^{<1}_{r_0\times r_0\times q_0} \times \mathbb{P}_{r_0} \\ & \times \mathbb{L}^1_{n_1\times r_1} \times \mathbb{L}^{SE<1}_{r_1\times r_1\times p_1} \times \mathbb{L}^{<1}_{r_1\times r_1\times q_1} \times \mathbb{P}_{r_1} \\ & \vdots \\ & \times \mathbb{L}^1_{n_C\times r_C} \times \mathbb{L}^{SE<1}_{r_C\times r_C\times p_C} \times \mathbb{L}^{<1}_{r_C\times r_C\times q_C} \times \mathbb{P}_{r_C}.\end{aligned}$$

The subset $\boldsymbol{\Theta}^*(\mathbf{r}, \mathbf{p}, \mathbf{q})$ is obtained by replacing all $\mathbb{P}_\cdot$ with $\mathbb{P}^{\geq\epsilon}_\cdot$, $\mathbb{L}^{<1}_{\cdot\times\cdot\times\cdot}$ with $\mathbb{L}^{\leq\kappa,B}_{\cdot\times\cdot\times\cdot}$, and $\mathbb{L}^{SE<1}_{\cdot\times\cdot\times\cdot}$ with $\mathbb{L}^{SE\leq\kappa}_{\cdot\times\cdot\times\cdot}$ of the appropriate dimensions for some specified $0 < \epsilon < 1$, $0 \leq \kappa < 1$ and $0 \leq B < \infty$. The components of a parameter tuple $\boldsymbol{\theta} \in \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ are labeled

$$\boldsymbol{\theta} = (\boldsymbol{\Phi}, \mathbf{A}, \mathbf{M}, \mathbf{C}, \boldsymbol{\Psi}, \mathbf{B}_1, \mathbf{N}_1, \mathbf{D}_1, \boldsymbol{\Upsilon}_1, \ldots, \mathbf{B}_C, \mathbf{N}_C, \mathbf{D}_C, \boldsymbol{\Upsilon}_C). \tag{3.9}$$

As the columns of $\mathbf{A}$ are channel-patterned, it will also be convenient to define $\mathbf{A} = [\mathbf{A}_1^{\mathsf{H}} \ \ldots \ \mathbf{A}_C^{\mathsf{H}}]^{\mathsf{H}}$ where $\mathbf{A}_c \in \mathbb{C}_{n_c\times r_0}$ is the block of $\mathbf{A}$ that lies in channel $c$ for $c = 1, \ldots, C$. Additionally, the combined matrix $\mathbf{B} = \mathrm{blkdiag}(\mathbf{B}_1, \ldots, \mathbf{B}_C)$ contains the bases for each of the interference subspaces. The parameter space $\boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ corresponds to a manifold with boundary of real dimension,

$$\begin{aligned}\dim \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q}) = & \ n + 2nr_0 - r_0(r_0+1) + (p_0+q_0)r_0(r_0+1) + r_0 \\ & + \textstyle\sum_{c=1}^{C} 2n_c r_c - r_c(r_c+1) + (p_c+q_c)r_c(r_c+1) + r_c.\end{aligned}$$

The parameter tuple $\boldsymbol{\theta}$ then defines the matrix transfer functions,

$$\mathcal{K}_{\mathbf{s}}(z;\boldsymbol{\theta}) \equiv \mathbf{A}\mathbf{M}(z)\mathbf{C}^{-1}(z)\boldsymbol{\Psi}^{\frac{1}{2}}, \qquad \mathcal{K}_{\mathbf{i}_c}(z;\boldsymbol{\theta}) \equiv \mathbf{B}_c\mathbf{N}_c(z)\mathbf{D}_c^{-1}(z)\boldsymbol{\Upsilon}_c^{\frac{1}{2}}, \tag{3.10}$$

where $\boldsymbol{\Psi}^{\frac{1}{2}}$ and $\boldsymbol{\Upsilon}_c^{\frac{1}{2}}$ are the positive-definite square roots. The associated spectral densities are

$$\mathbf{S}_{\mathbf{ss}}(\omega;\boldsymbol{\theta}) \equiv \mathcal{K}_{\mathbf{s}}(\mathrm{e}^{\mathrm{i}\omega};\boldsymbol{\theta})\mathcal{K}_{\mathbf{s}}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega};\boldsymbol{\theta}), \qquad \mathbf{S}_{\mathbf{i}_c\mathbf{i}_c}(\omega;\boldsymbol{\theta}) \equiv \mathcal{K}_{\mathbf{i}_c}(\mathrm{e}^{\mathrm{i}\omega};\boldsymbol{\theta})\mathcal{K}_{\mathbf{i}_c}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega};\boldsymbol{\theta})$$

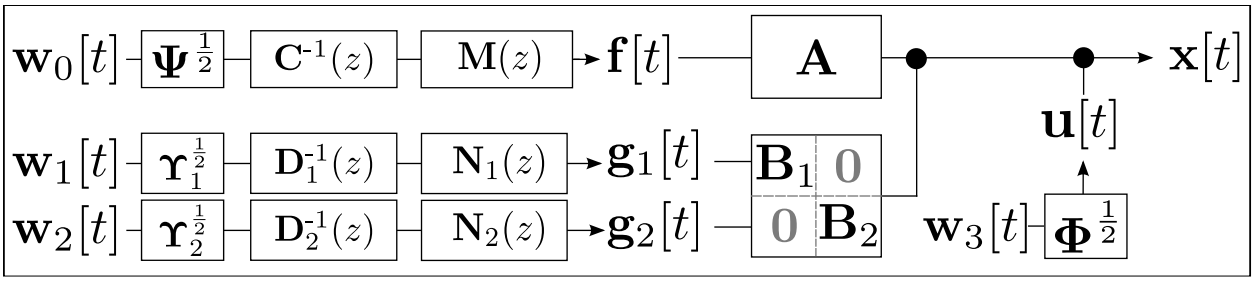



**Figure 3.1:** Two-channel synthesis diagram for the all-channel observation series $\mathbf{x}[t]$ under MFSA parameterization $\boldsymbol{\theta} \in \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$. The series $\mathbf{w}_0[t], \mathbf{w}_1[t]$ and $\mathbf{w}_2[t]$ are pairwise-uncorrelated white noise series with identity variance. All boxed quantities are structured parameters to be estimated on the basis of the measurements $\mathbf{x}[0], \ldots, \mathbf{x}[T-1]$.

with $\mathbf{S}_{\mathbf{ii}}\big(\omega; \boldsymbol{\theta}\big) = \mathrm{blkdiag}\big(\mathbf{S}_{\mathbf{i}_1\mathbf{i}_1}(\omega; \boldsymbol{\theta}), \ldots, \mathbf{S}_{\mathbf{i}_C\mathbf{i}_C}(\omega; \boldsymbol{\theta})\big)$. The power spectral density for the all-channel measurements with $\theta \in \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ then is

$$\mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta}) = \mathbf{S}_{\mathbf{ss}}(\omega; \boldsymbol{\theta}) + \mathbf{S}_{\mathbf{ii}}(\omega; \boldsymbol{\theta}) + \boldsymbol{\Phi}. \tag{3.11}$$

In the above parameterization, $\mathbf{A} \in \mathbb{L}_{n \times r_0}$ is a basis for the signal subspace $\mathcal{S}$, while $\mathbf{B}_c \in \mathbb{L}_{n_c \times r_c}$ is a basis for the channel-c inference subspace $\mathcal{I}_c$. The $r_0 \times r_0$ blocks of $\mathbf{M} \in \mathbb{L}_{r_0 \times r_0 \times p_0}$ and $\mathbf{D} \in \mathbb{L}^{<1}_{r_0 \times r_0 \times q_0}$ are the lower-triangular coefficient matrices in the numerator and denominator respectively of the rational matrix function $\mathbf{M}(z)\mathbf{D}^{-1}(z)$, which controls how the columns of $\mathcal{K}_{\mathbf{s}}(z)$ are obtained as frequency-dependent combinations of the columns of $\mathbf{A}$. The $r_c \times r_c$ blocks of $\mathbf{N}_c \in \mathbb{L}^{<1}_{r_c \times r_c \times p_c}$ and $\mathbf{D}_c \in \mathbb{L}^{<1}_{r_c \times r_c \times q_c}$ perform an analogous role for $\mathcal{K}_{\mathbf{i}_c}(z)$ and $\mathbf{B}_c$. Last, $\Psi$ and $\boldsymbol{\Upsilon}_1, \ldots \boldsymbol{\Upsilon}_C$ set the variance of the input white noise for the associated latent series. This synthesis process is depicted in Figure 3.1.

Note that the Schur-stability of the main diagonal polynomials of $\mathbf{C}(z)$ and $\mathbf{D}_c(z),\ c = 1, \ldots, C$ ensure that all entries of $\mathcal{K}_{\mathbf{s}}(z; \boldsymbol{\theta})$ and $\mathcal{K}_{\mathbf{i}_c}(z; \boldsymbol{\theta}),\ c = 1, \ldots, C$ are analytic for $|z| \geq 1$ and so they represent stable systems for all $\boldsymbol{\theta}$. Hence $\mathcal{K}_{\mathbf{s}}(z)$ has a stable, causal power series representation $\sum_{j=0}^{\infty} \mathbf{K}_{\mathbf{s},j}(\boldsymbol{\theta}) z^{-j}$ for $|z| \geq 1$, whose constant term is

$$\mathbf{K}_{\mathbf{s},0}(\boldsymbol{\theta}) = \lim_{z \to \infty} \mathcal{K}_{\mathbf{s}}(z; \boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\Phi}^{\frac{1}{2}},$$

which is lower triangular with positive main diagonal entries in the top $r_0 \times r_0$ block. The constant terms $\mathbf{K}_{\mathbf{i}_c,0}(\boldsymbol{\theta}),\ c = 1 \ldots, C$ similarly have the appropriate structure. This ensures that $\mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta})$ does in fact belong to the set of PSDs obtained by (3.6) by way of Proposition 9. However, the location of the zeroes of the main diagonal polynomials of $\mathbf{M}(z)$ and $\mathbf{N}_c(z), c = 1 \ldots, C$ are unconstrained, and so $\mathcal{K}_{\mathbf{s}}(z; \boldsymbol{\theta})$ and $\mathcal{K}_{\mathbf{i}_c}(z; \boldsymbol{\theta}), c = 1, \ldots, C$ may represent systems which do not have stable left-inverses.

### 3.2.4 Identifiability

In MFSA, the first-order model for the measurements is (3.1), which posits that the measurements in the individual channels are in fact a sum of a low-dimensional *signal* $\mathbf{s}_c[t]$ which has cross-channel correlations, a low-dimensional *interference* $\mathbf{i}_c[t]$ which is uncorrelated across channels, and measurement *noise* $\mathbf{u}_c[t]$. As the latent series $\mathbf{s}[t], \mathbf{i}[t]$ and $\mathbf{u}[t]$ are not directly observed, we must determine what statistical properties of the latent series can be uniquely determined based on sufficiently many channel measurements. The purpose of MFSA then is to estimate these quantities, which form a basis for drawing conclusions about the measurement series and allow for a decomposition of the measurements into signal, interference, and noise parts.

For multi-channel factor analysis, Chapter 2 demonstrates that the covariances of the signal, interference, and noise parts can generically be uniquely identified from the all-channel measurement covariance when conditions on the channel sizes and signal and interference dimensionality are satisfied. Given sufficiently many channel measurements, the covariances of the unobserved parts can be estimated to any desired degree of precision. This allows for the approximate recovery of signal, interference, and noise parts using MMSE estimates from the estimated covariances of the unobserved parts.

Building on that work, the following proposition shows that the PSDs of the latent series can be uniquely identified from the PSD of the all-channel measurement series as long as the values of the PSDs of the latent series can be uniquely identified at a single (non-degenerate) frequency.

**Proposition 10.** (Single-Freq. Identification) *If* $\mathbf{S_{xx}}(\omega_0)$ *uniquely determines* $(\mathbf{S_{ss}}(\omega_0), \mathbf{S_{ii}}(\omega_0), \mathbf{\Phi})$ *for some* $\omega_0 \in (-\pi, \pi]$ *and*

$$\operatorname{rank} \begin{bmatrix} \mathbf{S_{ss}}(\omega_0) & \mathbf{S_{ii}}(\omega_0) \end{bmatrix} = r_0 + r,$$

*then* $\mathbf{S_{xx}}(\omega)$ *uniquely determines* $(\mathbf{S_{ss}}(\omega), \mathbf{S_{ii}}(\omega), \mathbf{\Phi})$ *for almost all* $\omega \in (-\pi, \pi]$.

*Proof.* See Section C.2 in Appendix C for proof. □

Previous results for MFA identifiability provides conditions on the channel sizes $\mathbf{n}$ and subspace dimensions $\mathbf{r}$ which suffice for the generic global identifiability of $\mathbf{S_{ss}}$, $\mathbf{S_{ii}}$ and $\mathbf{\Phi}$ from $\mathbf{S_{xx}}$ at some frequency. Proposition 10 can then be used to obtain the generic global identifiability of the MFSA decomposition (3.6) at all frequencies. To see this, let $\omega_0$ be some frequency such that (10) is satisfied, and let $\mathbf{S}^o$ be the real $2n \times 2n$ matrix,

$$\mathbf{S}^o = \begin{bmatrix} \operatorname{Re} \mathbf{S_{xx}}(\omega_0) & \operatorname{Im} \mathbf{S_{xx}}(\omega_0) \\ -\operatorname{Im} \mathbf{S_{xx}}(\omega_0) & \operatorname{Re} \mathbf{S_{xx}}(\omega_0) \end{bmatrix}.$$

By hypothesis, $\mathbf{S_{xx}}(\omega_0) = \mathbf{A}\mathbf{A}^{\mathsf{H}} + \mathbf{B}\mathbf{B}^{\mathsf{H}} + \mathbf{\Phi}$ for some full rank $\mathbf{A} \in \mathbb{C}_{n \times r_0}$ and $\mathbf{B}$ equalling $\operatorname{blkdiag}(\mathbf{B}_1, \ldots, \mathbf{B}_C)$ with full rank $\mathbf{B}_c \in \mathbb{C}_{n_c \times r_c}$. Hence,

$$\begin{aligned} \operatorname{Re} \mathbf{S_{ss}}(\omega_0) &= [\operatorname{Re} \mathbf{A} \ \operatorname{Im} \mathbf{A}][\operatorname{Re} \mathbf{A} \ \operatorname{Im} \mathbf{A}]^{\mathsf{T}}, \\ \operatorname{Re} \mathbf{S_{i_c i_c}}(\omega_0) &= [\operatorname{Re} \mathbf{B}_c \ \operatorname{Im} \mathbf{B}_c][\operatorname{Re} \mathbf{B}_c \ \operatorname{Im} \mathbf{B}_c]^{\mathsf{T}}, c = 1, \ldots, C \\ \operatorname{Re} \mathbf{S_{xx}}(\omega_0) &= \operatorname{Re} \mathbf{S_{ss}}(\omega_0) + \operatorname{Re} \mathbf{S_{ii}}(\omega_0) + \mathbf{\Phi}, \end{aligned}$$

and so $\operatorname{Re} \mathbf{S_{xx}}(\omega_0)$ is the sum of a real $n \times n$ covariance matrix of rank at most $2r_0$, a block-diagonal covariance matrix with blocks of size $n_c \times n_c$ with rank at most $2r_c$ and a diagonal covariance matrix. This is the setting of Theorem 1, and so if Condition 2 is satisfied with channel sizes $\mathbf{n}$ and factor numbers $2\mathbf{r}$, then generically $\mathbf{\Phi}$ can be uniquely isolated from $\operatorname{Re} \mathbf{S_{ss}}(\omega_0) + \operatorname{Re} \mathbf{S_{ii}}(\omega_0)$. Defining the matrices $\mathbf{A}^o$ and $\mathbf{B}_c^o$ for $c = 1, \ldots, C$ as

$$\mathbf{A}^o = \begin{bmatrix} \operatorname{Re} \mathbf{A} & \operatorname{Im} \mathbf{A} \\ -\operatorname{Im} \mathbf{A} & \operatorname{Re} \mathbf{A} \end{bmatrix} \quad \mathbf{B}_c^o = \begin{bmatrix} \operatorname{Re} \mathbf{B}_c & \operatorname{Im} \mathbf{B}_c \\ -\operatorname{Im} \mathbf{B}_c & \operatorname{Re} \mathbf{B}_c \end{bmatrix},$$

with $\mathbf{B}^o = \text{blkdiag}(\mathbf{B}_1^o, \ldots, \mathbf{B}_C^o)$ and $\mathbf{\Phi}^o = \text{blkdiag}(\mathbf{\Phi}, \mathbf{\Phi})$, we further have

$$\mathbf{S}^o - \mathbf{\Phi}^o = [\mathbf{A}^o \; \mathbf{B}^o][\mathbf{A}^o \; \mathbf{B}^o]^\mathsf{T}.$$

If the submatrices of $[\mathbf{A}^o \; \mathbf{B}^o]$ satisfy the condition of Proposition 3, then $\mathbf{A}^o\mathbf{A}^{o\mathsf{T}}$ and $\mathbf{B}^o\mathbf{B}^{o\mathsf{T}}$ will be uniquely determined from $\mathbf{S}^o - \mathbf{\Phi}^o$. This will generically occur when Condition 1 is satisfied with channel sizes $2\mathbf{n}$ and subspace dimensions $2\mathbf{r}$. As $\mathbf{A}^o\mathbf{A}^{o\mathsf{T}}$ determines $\mathbf{S}_{\mathbf{ss}}(\omega_0)$ and similarly for $\mathbf{B}^o\mathbf{B}^{o\mathsf{T}}$ and $\mathbf{S}_{\mathbf{ii}}(\omega_0)$, the measurement PSD $\mathbf{S}_{\mathbf{xx}}(\omega_0)$ will uniquely determine $(\mathbf{S}_{\mathbf{ss}}(\omega_0), \mathbf{S}_{\mathbf{ii}}(\omega_0), \mathbf{\Phi})$ generically when the above conditions on $\mathbf{n}$ and $\mathbf{r}$ are satisfied.

### 3.2.5 Parameterization Coverage and Uniqueness

The parameterization outlined in Section 3.2.3 defines a map from $\boldsymbol{\theta} \in \mathbf{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ into $\mathcal{D}(\mathbf{n}, \mathbf{r})$, the set of MFSA power spectral densities. For the parameterization to be useful, the union of the parameterization images for varying $\mathbf{r}, \mathbf{p}, \mathbf{q}$ should cover a sufficiently large subset of $\mathcal{D}(\mathbf{n}, \mathbf{r})$. As the following proposition shows, the union of parameterization images includes all $\mathbf{S}_{\mathbf{xx}}(\omega)$ obtained from rational transfer functions which are full rank and have no poles on the unit circle.

**Proposition 11.** (Coverage) *Suppose* $\mathbf{S}_{\mathbf{xx}}(\omega)$ *is obtained from* (3.6) *for some* $\mathcal{K}_{\mathbf{s}}(z)$, $\mathcal{K}_{\mathbf{i}_c}(z),\ c = 1, \ldots, C$ *and* $\mathbf{\Phi}$ *as defined in Proposition 9. Further suppose that* $\mathcal{K}_{\mathbf{s}}(z)$ *and* $\mathcal{K}_{\mathbf{i}_c}(z),\ c = 1, \ldots, C$ *are full rank when* $|z| = 1$ *and that all components of* $\mathcal{K}_{\mathbf{s}}(z)$ *and* $\mathcal{K}_{\mathbf{i}_c}(z),\ c = 1, \ldots, C$ *are rational functions of* $z^{-1}$ *taking finite values when* $|z| = 1$. *Then there exist* $\mathbf{r}, \mathbf{p}, \mathbf{q} \in \mathbb{N}_{C+1}$ *and* $\boldsymbol{\theta} \in \mathbf{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ *such that* $\mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta}) = \mathbf{S}_{\mathbf{xx}}(\omega)$ *for almost all* $\omega \in (-\pi, \pi]$.

*Proof.* See Section C.3 in Appendix C for proof. □

In addition to being sufficiently expressive, a desirable property of a parameterization is that the map from the parameter space into the set of MFSA observation spectral densities $\mathcal{D}(\mathbf{n}, \mathbf{r})$ be one-to-one. As the all-channel measurements $\mathbf{x}[t], t = 0, \ldots, T-1$ are used to obtain a best-fitting MFSA spectral density for the chosen $\mathbf{r}, \mathbf{p}, \mathbf{q}$, it is required for numeric convergence that those best-fitting power spectral densities corresponds to a unique $\boldsymbol{\theta} \in \mathbf{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$. As the following proposition shows, the map from $\mathbf{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ into $\mathcal{D}(\mathbf{n}, \mathbf{r})$ is in fact one-to-one except on a null

subset of the parameter space. Hence, the estimated $\mathbf{S}_{\mathbf{xx}}(\omega;\boldsymbol{\theta})$ will generically correspond to a unique tuple of parameter values.

**Proposition 12.** (Uniqueness) *If the channel sizes* $\mathbf{n}$ *and subspace dimensions* $\mathbf{r}$ *are such that the conclusion of Proposition 10 holds for* $\mathbf{S}_{\mathbf{xx}}(\omega;\boldsymbol{\theta})$ *except on a null subset of* $\boldsymbol{\Theta}(\mathbf{r},\mathbf{p},\mathbf{q})$*, then there exists a null subset* $N \subset \boldsymbol{\Theta}(\mathbf{r},\mathbf{p},\mathbf{q})$ *such that the map* $\boldsymbol{\theta} \mapsto \mathbf{S}_{\mathbf{xx}}(\omega;\boldsymbol{\theta})$ *is injective on* $\boldsymbol{\Theta}(\mathbf{r},\mathbf{p},\mathbf{q}) \setminus N$.

*Proof.* See Section C.4 in Appendix C for proof. □

The preceding two propositions concerning the parameter space $\boldsymbol{\Theta}(\mathbf{r},\mathbf{p},\mathbf{q})$ and the parameterization (3.11) show that the class of parametric models considered is flexible enough to represent a broad set of MFSA spectral densities while still retaining the desired unique connection between the parameter tuples and the represented spectral density. As the next section shows, this parameterization is also convenient for estimation.

## 3.3 Estimation

We assume that measurements $\mathbf{x}_c[t]$ from all channels $c = 1,\ldots,C$ are collected at $T$ equally-spaced times $t = 0,\ldots,T-1$. The primary goal of MFSA is to obtain a decomposition of the all-channel measurement PSD $\mathbf{S}_{\mathbf{xx}}(\omega)$ into the power spectral densities of the latent series as (3.6). To this end, the all-channel *periodogram* $\mathbf{I}(\omega)$ will be used as an empirical analog of the unknown $\mathbf{S}_{\mathbf{xx}}(\omega)$ for the purpose of estimating the terms in the decomposition (3.6). In particular, the tapered DTFT

$$\mathbf{x}(\omega) = \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} h\left(\frac{t}{T}\right) \mathrm{e}^{\mathrm{i}\omega} \mathbf{x}[t]$$

and the tapered periodogram $\mathbf{I}(\omega) = \mathbf{x}(\omega)\mathbf{x}^{\mathsf{H}}(\omega)$ are the inputs to Algorithm 1, where the taper function $h(y) : [0,1] \to \mathbb{R}$ is normalized to $1 = \sum_{t=0}^{T-1} h(t/T)$. Use of taper functions or similar methods of smoothing the periodogram are known (Dahlhaus, 1988) to improve finite-sample performance of spectral estimation techniques, as they reduce the edge effect created by using a finite number of observations. The all-channel periodogram $\mathbf{I}(\omega) \in \mathbb{C}_{n\times n}$ is a channel-patterned matrix whose diagonal blocks $\mathbf{I}_{cc}(\omega)$ are the intra-channel periodograms obtained from $\mathbf{x}_c[t]$ while

the off-diagonal blocks $\mathbf{I}_{cc'}(\omega),\ c \neq c'$ are the cross-channel periodograms obtained from $\mathbf{x}_c[t]$ and $\mathbf{x}_{c'}[t]$.

For the purpose of estimation, we consider the subspace dimensions $\mathbf{r}$ and the numerator and denominator orders $\mathbf{p}$ and $\mathbf{q}$ to be specified *a priori*. To ensure that unique minima exist for the relevant sub-problems, we further require that $T > \max\{||\mathbf{r}||_\infty, ||\mathbf{p}||_\infty, ||\mathbf{q}||_\infty, n\}$. The procedure described in the remainder of this section is designed to obtain a value $\boldsymbol{\theta} \in \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ which solves the following optimization problem.

### 3.3.1 Optimization Problem

As a measure of goodness of fit, we use the finite-sample Whittle pseudo-likelihood,

$$\ell(\boldsymbol{\theta}) = \frac{1}{T}\sum_{k=0}^{T-1} \log\det \mathbf{S}_{\mathbf{xx}}(\omega_k; \boldsymbol{\theta}) + \operatorname{tr} \mathbf{S}_{\mathbf{xx}}^{-1}(\omega_k; \boldsymbol{\theta})\mathbf{I}(\omega_k),$$

for the Fourier frequencies $\omega_k = 2\pi k/T$ for $k = 0, \ldots, T-1$. The Whittle pseudo-likelihood, both in the form of an integral and in the form of a Riemann sum over the fundamental frequencies, has a long history of use in time series analysis (Whittle, 1953; Dahlhaus, 1988; Rao and Yang, 2020) and signal processing (Itakura and Saito, 1968; Ramírez et al., 2011). As a measure of model fit, it can be motivated as minimizing the Itakura-Saito spectral divergence of the periodogram from the target PSD (Taniguchi, 1987; Takabatake and Yano, 2023), or as the asymptotic log-likelihood of the periodogram at a fixed set of frequencies (Brillinger, 1975; Choudhuri, 2004).

The optimization problem by which the parameters of MFSA are estimated then is,

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}^*(\mathbf{r},\mathbf{p},\mathbf{q})} \ell(\boldsymbol{\theta}), \tag{3.12}$$

where, for numeric stability, the set of parameter values optimized over is $\boldsymbol{\Theta}^*(\mathbf{r}, \mathbf{p}, \mathbf{q})$ which includes the constraints that the noise and innovation variances are bounded away from zero by a positive constant and that the magnitude of the zeros of the appropriate entries of the numerator and denominator polynomials are bounded away from 1 by some positive constant. A useful aspect of these constraints is that they ensure that the matrices $\mathbf{C}^{-1}(\mathrm{e}^{\mathrm{i}\omega})\boldsymbol{\Psi}\mathbf{C}^{-\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})$ and

$\mathbf{D}_c^{-1}(\mathrm{e}^{\mathrm{i}\omega})\boldsymbol{\Upsilon}_c\mathbf{D}_c^{-\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega}),\ c=1,\ldots,C$ are finite and positive definite for all $\omega \in (-\pi,\pi]$. This will allow certain applications of the matrix determinant and inverse lemmas to $\ell(\boldsymbol{\theta})$ so as to transform it into a form which is convenient for optimization and exploits the dimensionality assumptions of MFSA.

The frequency-flat component of the objective is defined as,

$$\begin{aligned}\epsilon(\boldsymbol{\Phi},\boldsymbol{\Psi},\boldsymbol{\Upsilon}_1,\ldots,\boldsymbol{\Upsilon}_C) = \textstyle\sum_{i=1}^{n}\log[\boldsymbol{\Phi}]_{ii} + [T^{-1}\sum_{k=0}^{T-1}\mathbf{I}(\omega_k)]_{ii}[\boldsymbol{\Phi}^{-1}]_{ii} \\ + \textstyle\sum_{j=1}^{r_0}\log[\boldsymbol{\Psi}]_{jj} + \sum_{c=1}^{C}\sum_{j=1}^{r_c}\log[\boldsymbol{\Upsilon}_c]_{jj}\end{aligned} \tag{3.13}$$

in which the data enter through the main diagonal of the sample covariance matrix for the tapered all-channel measurements.

As the systemic part $\mathbf{s}[t]+\mathbf{i}[t]$ of the all-channel measurement $\mathbf{x}[t]$ is constrained to live in an $r_0+r$-dimensional subspace, many aspects of the optimization problem (3.12) can be defined in terms of $(r_0+r)\times(r_0+r)$ matrices with the appropriate structure. As the total subspace dimension $r_0+r$ is typically much less than the total channel size $n$, working with $r_0+r\times r_0+r$ matrices instead of $n\times n$ matrices can produce significant computational savings. Accordingly, a Hermitian positive semi-definite matrix $\mathbf{Y}\in\mathbb{C}_{r_0+r\times r_0+r}$ is *factor-patterned* if it is partitioned into $(C+1)^2$ blocks as

$$\mathbf{Y} = \begin{array}{c} \begin{matrix} r_0 & r_1 & \cdots & r_C \end{matrix} \\ \begin{bmatrix} \mathbf{Y}_{00} & \mathbf{Y}_{01} & \ldots & \mathbf{Y}_{0C} \\ \mathbf{Y}_{01}^{\mathsf{H}} & \mathbf{Y}_{11} & \ldots & \mathbf{Y}_{1C} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Y}_{0C}^{\mathsf{H}} & \mathbf{Y}_{1C}^{\mathsf{H}} & \ldots & \mathbf{Y}_{CC} \end{bmatrix} \end{array} \begin{matrix} r_0 \\ r_1 \\ \vdots \\ r_C \end{matrix}.$$

As the signal subspace $\mathcal{S}$ and interference subspaces $\mathcal{I}_1,\ldots,\mathcal{I}_C$ are of central importance to MFSA, we can describe the diagonal blocks $\mathbf{Y}_{\ell\ell},\ \ell=0,\ldots,C$ as intra-subspace quantities and the off-diagonal blocks $\mathbf{Y}_{\ell\ell'},\ \ell\neq\ell'$ as cross-subspace quantities.

With this notion, the transformed periodogram $\mathbf{J}(\omega)$ can be defined as the factor-patterned matrix whose blocks are given by,

$$\begin{aligned}
\mathbf{J}_{00}(\omega) &\equiv \mathbf{M}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{A}^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{I}(\omega)\mathbf{\Phi}^{-1}\mathbf{A}\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega}), \\
\mathbf{J}_{0c}(\omega) &\equiv \mathbf{M}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{A}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{I}_{cc}(\omega)\mathbf{\Phi}_c^{-1}\mathbf{B}_c\mathbf{N}_c(\mathrm{e}^{\mathrm{i}\omega}), \\
\mathbf{J}_{cc'}(\omega) &\equiv \mathbf{N}_c^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{B}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{I}_{cc'}(\omega)\mathbf{\Phi}_{c'}^{-1}\mathbf{B}_{c'}\mathbf{N}_{c'}(\mathrm{e}^{\mathrm{i}\omega}),
\end{aligned} \tag{3.14}$$

for all $\omega \in (-\pi, \pi]$, with $c, c' = 1, \ldots, C$ and $c \neq c'$. Note that $\mathbf{J}(\omega)$ depends both on the data through $\mathbf{I}(\omega)$ and the parameter tuple $\boldsymbol{\theta}$, but that dependence is suppressed for clarity of notation. Additionally define the factor-patterned block-diagonal matrix $\mathbf{E}(\omega)$ as

$$\begin{aligned}
\mathbf{E}_{00}(\omega) &\equiv \mathbf{C}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{\Psi}^{-1}\mathbf{C}(\mathrm{e}^{\mathrm{i}\omega}), \\
\mathbf{E}_{cc}(\omega) &\equiv \mathbf{D}_c^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{\Upsilon}_c^{-1}\mathbf{D}_c(\mathrm{e}^{\mathrm{i}\omega}),\ c = 1, \ldots, C,
\end{aligned} \tag{3.15}$$

and let $\mathbf{F}(\omega) \in \mathbb{C}_{n \times r_0 + r}$ be the frequency-varying combined loading matrix,

$$\mathbf{F}(\omega) \equiv \begin{bmatrix} \mathbf{A}\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega}) & \mathbf{B}\mathbf{N}(\mathrm{e}^{\mathrm{i}\omega}) \end{bmatrix}. \tag{3.16}$$

With the above definitions, we can define the factor-patterned spectral density $\mathbf{T}(\omega)$ as

$$\mathbf{T}(\omega) = \left(\mathbf{E}(\omega) + \mathbf{F}^{\mathsf{H}}(\omega)\mathbf{\Phi}^{-1}\mathbf{F}(\omega)\right)^{-1}.$$

As previously described, the restriction that $\boldsymbol{\theta}$ is in $\mathbf{\Theta}^*(\mathbf{r}, \mathbf{p}, \mathbf{q})$ ensures that $\mathbf{E}(\omega)$ is finite and positive definite for all $\omega$, hence $\mathbf{E}(\omega) + \mathbf{F}^{\mathsf{H}}(\omega)\mathbf{\Phi}^{-1}\mathbf{F}(\omega)$ is itself positive definite and so $\mathbf{T}(\omega)$ exists for all $\omega$. Using the matrix determinant and inverse lemmas, the objective $\ell(\boldsymbol{\theta})$ can be shown to equal

$$\begin{aligned}
\ell(\boldsymbol{\theta}) = -T^{-1} \textstyle\sum_{k=0}^{T-1} \left[ \log\det \mathbf{T}(\omega_k) + \operatorname{tr} \mathbf{T}(\omega_k)\mathbf{J}(\omega_k) \right] \\
+ \epsilon(\mathbf{\Phi}, \mathbf{\Psi}, \mathbf{\Upsilon}_1, \ldots, \mathbf{\Upsilon}_C) + \gamma(\mathbf{C}) + \textstyle\sum_{c=1}^{C} \delta_c(\mathbf{D}_c),
\end{aligned} \tag{3.17}$$

for all $\boldsymbol{\theta} \in \mathbf{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$, where $\gamma(\mathbf{C})$ and $\delta_c(\mathbf{D}_c)$ are

$$\begin{aligned}
\gamma(\mathbf{C}) &\equiv -T^{-1} \textstyle\sum_{k=0}^{T-1} \log\det \mathbf{C}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k})\mathbf{C}(\mathrm{e}^{\mathrm{i}\omega_k}), \\
\delta_c(\mathbf{D}_c) &\equiv -T^{-1} \textstyle\sum_{k=0}^{T-1} \log\det \mathbf{D}_c^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k})\mathbf{D}_c(\mathrm{e}^{\mathrm{i}\omega_k}).
\end{aligned}$$

The expression (3.17) of $\ell(\boldsymbol{\theta})$ demonstrates an important aspect of the optimization problem (3.12). The parameterization of $\mathbf{S}_{\mathbf{xx}}(\omega;\boldsymbol{\theta})$ is naturally described in terms of signal-related parameter blocks $(\mathbf{A},\mathbf{M},\mathbf{C},\boldsymbol{\Psi})$ and the channel-$c$ interference-related parameter blocks $(\mathbf{B}_c,\mathbf{N}_c,\mathbf{D}_c,\boldsymbol{\Upsilon}_c)$ for $c=1,\ldots,C$. These parameter blocks are intermingled within $\ell(\boldsymbol{\theta})$ and so they must be optimized jointly rather than separately. However, when examining (3.17) it can be seen that if $\mathbf{T}(\omega)$ were block-diagonal for each $\omega$, then the optimization subproblems for the signal-related and interference-related parameter blocks would be separable as the trace of the product of a matrix with a block-diagonal matrix is the sum of the traces of the products of the diagonal blocks. The quantity which prevents $\mathbf{T}(\omega)$ from being block-diagonal is the factor-patterned matrix,

$$\mathbf{R} \equiv \begin{bmatrix} \mathbf{A}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{A} & \mathbf{A}_1^{\mathsf{H}}\boldsymbol{\Phi}_1^{-1}\mathbf{B}_1 & \ldots & \mathbf{A}_C^{\mathsf{H}}\boldsymbol{\Phi}_C^{-1}\mathbf{B}_C \\ \mathbf{B}_1^{\mathsf{H}}\boldsymbol{\Phi}_1^{-1}\mathbf{A}_1 & \mathbf{B}_1^{\mathsf{H}}\boldsymbol{\Phi}_1^{-1}\mathbf{B}_1 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_C^{\mathsf{H}}\boldsymbol{\Phi}_C^{-1}\mathbf{A}_C & \mathbf{0} & \ldots & \mathbf{B}_C^{\mathsf{H}}\boldsymbol{\Phi}_C^{-1}\mathbf{B}_C \end{bmatrix}.$$

In terms of the inner product on $\mathbb{C}_{n\times(r_0+r)}$ defined by $\boldsymbol{\Phi}$ (namely $\mathrm{tr}(\mathbf{X}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{Y})$ for $\mathbf{X},\mathbf{Y}\in\mathbb{C}_{n\times(r_0+r)}$), $\mathbf{R}$ is the Gram matrix for the columns of $[\mathbf{A}\ \mathbf{B}]$. So, the above matrix will be block-diagonal if and only if the signal subspace $\mathcal{S}$, for which the columns of $\mathbf{A}$ are a basis, is $\boldsymbol{\Phi}$-orthogonal to the combined interference subspace $\mathcal{I}_1\oplus\cdots\oplus\mathcal{I}_C$. In MFSA, we assume that the signal and combined interference subspaces intersect trivially (which is generically true when the total subspace dimension $r_0+r$ is less than the total channel size $n$), but $\boldsymbol{\Phi}$-orthogonal signal and combined interference subspaces would imply a particular relationship between $\mathbf{s}[t]$, $\mathbf{i}[t]$ and $\mathbf{u}[t]$ which may not hold for many problems. Nevertheless, a surrogate function is designed for $\ell(\boldsymbol{\theta})$ in the following sections which leverages this geometric perspective and allows for the signal-related and interference-related parameter blocks to be optimized separately.

### 3.3.2 Optimizing $\ell(\boldsymbol{\theta})$ by Majorization-Minimization

To find a $\boldsymbol{\theta}\in\boldsymbol{\Theta}(\mathbf{r},\mathbf{p},\mathbf{q})$ that optimizes (3.12), a Majorization-Minimization (MM) approach is used. MM algorithms can be seen as extensions of the Expectation-Maximization (EM) approach

(Dempster et al., 1977; Moon, 1996) to estimation that relax the "incomplete data" formulation of EM and instead directly obtain a surrogate function with the needed properties (Sun et al., 2017). Benefits of MM estimation procedures are that the sequence of objective values obtained is monotonically decreasing, which leads to improved stability and can reduce sensitivity to initialization, and that parameter constraints can more easily be incorporated when minimizing the surrogate function than is the case for gradient-based methods.

For MM algorithms, a surrogate function $g(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}})$ for the objective $\ell(\boldsymbol{\theta})$ is constructed, which is a global upper bound $g(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}) \geq \ell(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}$ and has the equality $g(\boldsymbol{\theta}|\boldsymbol{\theta}) = \ell(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. So, an update to the parameter tuple taking $\widetilde{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}'$ which decreases the surrogate as $g(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}}) \leq g(\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{\theta}})$ must also decrease the objective, $\ell(\boldsymbol{\theta}') \leq \ell(\widetilde{\boldsymbol{\theta}})$. The surrogate function can be designed to have certain desirable properties (such as separation of parameter blocks and closed form updates for the parameter blocks) which make it easier to find an $\boldsymbol{\theta}'$ which decreases the surrogate rather than directly optimize the objective.

To obtain a convenient surrogate function, the following elementary majorizing inequalities will be used. The ability to obtain a surrogate by combining matrix majorizing inequalities which have clear geometric interpretations is an advantage of optimizing in the spectral domain, as corresponding time-domain inequalities would be much more difficult to handle. In the following expressions, $\mathbf{X}, \widetilde{X} \in \mathbb{C}_{n\times m}$ are arbitrary matrices, $\mathbf{S}, \widetilde{\mathbf{S}} \in \mathbb{C}_{m\times m}$ are positive definite, and $\mathbf{V}, \mathbf{W}, \mathbf{M}, \widetilde{\mathbf{V}} \in \mathbb{C}_{m\times m}$, $\mathbf{D} \in \mathbb{C}_{n\times n}$ are positive semi-definite with $\mathbf{M} \succeq \mathbf{W}$,

$$\log\det \mathbf{S} \leq \log\det \widetilde{\mathbf{S}} + \operatorname{tr} \widetilde{\mathbf{S}}^{-1}(\mathbf{S} - \widetilde{\mathbf{S}}), \tag{3.18}$$

$$-\operatorname{tr}(\mathbf{S}^{-1}\mathbf{V}) \leq \operatorname{tr} \widetilde{\mathbf{S}}^{-1}\widetilde{\mathbf{V}}\widetilde{\mathbf{S}}^{-1}\mathbf{S} - 2\operatorname{Re}\operatorname{tr} \widetilde{\mathbf{S}}^{-1}\widetilde{\mathbf{V}}^{\frac{1}{2}}(\mathbf{V}^{\frac{1}{2}})^{\mathsf{H}}, \tag{3.19}$$

$$\operatorname{tr} \mathbf{X}^{\mathsf{H}}\mathbf{D}\mathbf{X}\mathbf{W} \leq \operatorname{tr} \mathbf{X}^{\mathsf{H}}\mathbf{D}\mathbf{X}\mathbf{M} - 2\operatorname{Re}\operatorname{tr} \widetilde{\mathbf{X}}^{\mathsf{H}}\mathbf{D}\mathbf{X}(\mathbf{M} - \mathbf{W}) + \operatorname{tr} \widetilde{\mathbf{X}}^{\mathsf{H}}\mathbf{D}\widetilde{\mathbf{X}}(\mathbf{M} - \mathbf{W}). \tag{3.20}$$

Expressions of the form $\mathbf{V}^{\frac{1}{2}}$ indicate any matrix such that $\mathbf{V} = \mathbf{V}^{\frac{1}{2}}(\mathbf{V}^{\frac{1}{2}})^{\mathsf{H}}$. That is, $\mathbf{V}^{\frac{1}{2}}$ is *any* matrix square root of $\mathbf{V}$, not necessarily the unique positive semi-definite square root. The first inequality (3.18) follows from the concavity of the log-determinant, as it is globally upperbounded by its differential at any $\widetilde{\mathbf{S}}$. The second inequality (3.19) follows from the non-negativity of

$||(\mathbf{S}^{\frac{1}{2}})^{\mathsf{H}}(\mathbf{S}^{-1}\mathbf{V}^{\frac{1}{2}} - \widetilde{\mathbf{S}}^{-1}\widetilde{\mathbf{V}}^{\frac{1}{2}})||_F^2$. The third inequality (3.20) follows from the non-negativity of $||(\mathbf{X}\mathbf{D}^{\frac{1}{2}} - \widetilde{\mathbf{X}}\mathbf{D}^{\frac{1}{2}})^{\mathsf{H}}(\mathbf{M}-\mathbf{W})^{\frac{1}{2}}||_F^2$, as a square root of $\mathbf{M}-\mathbf{W}$ exists by $\mathbf{M} \succeq \mathbf{W}$. If $\mathbf{S} = \widetilde{\mathbf{S}}$, $\mathbf{X} = \widetilde{\mathbf{X}}$ and $\mathbf{V}^{\frac{1}{2}} = \widetilde{\mathbf{V}}^{\frac{1}{2}}$ (hence $\mathbf{V} = \widetilde{\mathbf{V}}$) then all three hold with equality.

Before applying the above majorizing inequalities to $\ell(\boldsymbol{\theta})$ in the form (3.17), needed geometric and algebraic relationships between various relevant quantities will be established in the following sections.

### 3.3.3 Majorization of $\gamma(\mathbf{C})$ and $\delta_c(\mathbf{D}_c)$

As $\mathbf{C}(\mathrm{e}^{\mathrm{i}\omega})$ and $\mathbf{D}_c(\mathrm{e}^{\mathrm{i}\omega})$ are lower-triangular, $\log\det\mathbf{C}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{C}(\mathrm{e}^{\mathrm{i}\omega})$ and $\log\det\mathbf{D}_c^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{D}_c(\mathrm{e}^{\mathrm{i}\omega})$ reduce to the sum of the log squared magnitudes of the main-diagonal polynomials. As the zeros of these polynomials are constrained to have magnitude at most $\kappa < 1$, it will be convenient to work with their respective zeros $v_{jm}$ and $w_{c,jm}$, which have

$$[\mathbf{C}(z)]_{jj} = \textstyle\prod_{m=1}^{q_0}(1 - v_{jm}z^{-1}),\ j = 1,\ldots,r_0,$$
$$[\mathbf{D}_c(z)]_{jj} = \textstyle\prod_{m=1}^{q_c}(1 - w_{c,jm}z^{-1}),\ j = 1,\ldots,r_c.$$

As the main-diagonal polynomials have constant term 1, the vector of coefficients and the set of zeros are in one-to-one correspondence. As the update step for $\mathbf{C}$ and $\mathbf{D}_c,\ c = 1,\ldots,C$ defined in Section 3.4.3 updates $v_{jm}$ and $w_{c,jm}$ directly, the zeros for the $[\mathbf{C}(z)]_{jj}$ and $[\mathbf{D}_c(z)]_{jj}$ can be stored rather than their coefficients. This also avoids ambiguity in the indexing of the zeros, as they can be numbered arbitrarily at the initialization step and the indexing preserved subsequently.

As $\ell(\boldsymbol{\theta})$ depends on the spectral density at the Fourier frequencies, we use the identity

$$-\frac{1}{T}\sum_{k=0}^{T-1}\log|1 - v\mathrm{e}^{\mathrm{i}\omega_k}|^2 = \tfrac{1}{T}\sum_{j=1}^{\infty}\sum_{k=0}^{T-1} j^{-1}(v^j\mathrm{e}^{-\mathrm{i}\omega_k j} + \overline{v}^j\mathrm{e}^{\mathrm{i}\omega_k j}) = -\frac{2}{T}\log|1 - v^T|,$$

which holds for all $|v| < 1$ by the Mercator series expansion for $-\log(1-z)$ and that $\sum_{k=0}^{T-1}\mathrm{e}^{\mathrm{i}\omega_k j}$ is $T$ if $j$ is a multiple of $T$ and zero otherwise. The first and second Wirtinger derivatives of $f(v) \equiv -\frac{2}{T}\log|1 - v^T|$ are

$$\frac{\partial f}{\partial z} = \frac{v^{T-1}}{1 - v^T}, \qquad \frac{\partial f}{\partial \overline{z}} = \frac{\overline{v}^{T-1}}{1 - \overline{v}^T},$$

$$\frac{\partial^2 f}{\partial z^2} = \frac{v^{T-2}(v^T + T - 1)}{(1 - v^T)^2}, \qquad \frac{\partial^2 f}{\partial \overline{z}^2} = \frac{\overline{v}^{T-2}(\overline{v}^T + T - 1)}{(1 - \overline{v}^T)^2},$$

and $\frac{\partial^2 f}{\partial z \partial \overline{z}} = 0$. If $f(v)$ is expressed as a function of the real and imaginary parts, the largest eigenvalue of the associated Hessian $\mathbf{D}^2 f(x, y)$ satisfies

$$\begin{aligned}\lambda_{\max}(\mathbf{D}^2 f)(x, y) &= \frac{2|v|^{T-2}|(v^T + T - 1)|}{|1 - v^T|^2} \\ &\leq \frac{2\kappa^{T-1}(\kappa^T + T - 1)}{(1 - \kappa^T)^2} \equiv \eta(\kappa, T),\end{aligned}$$

for all $v = x + \mathrm{i}y$ with $|v| \leq \kappa$, by the relation of the real Hessian to the second Wirtinger derivatives given by van den Bos (1994). As the real Hessian satisfies $\mathbf{D}^2 f(x, y) \preceq \eta(\kappa, T)\mathbf{I}_2$ for all $(x, y)$ with radius at most $\kappa$, the quadratic majorization (see, e.g., Sun et al. (2017)) via Taylor expansion for twice-differentiable real functions can be employed. The resulting inequality, when expressed in terms of $v$, is

$$f(v) \leq f(\widetilde{v}) + 2\,\mathrm{Re}\,\frac{\widetilde{v}^{T-1}}{1 - \widetilde{v}^T}(v - \widetilde{v}) + \frac{\eta(\kappa, T)}{2}|v - \widetilde{v}|^2,$$

which holds for all $v, \widetilde{v}$ with magnitude at most $\kappa$ and holds with equality when $v = \widetilde{v}$. Applying this to all zeros, we obtain

$$\begin{aligned}\gamma(\mathbf{C}) &\leq \textstyle\sum_{j=1}^{r_0} \sum_{m=1}^{q_0} f(\widetilde{v}_{jm}) + \frac{\eta(\kappa,T)}{2}|v_{jm} - \widetilde{v}_{jm}|^2 \\ &\quad + 2\,\mathrm{Re}\,\widetilde{v}_{jm}^{T-1}(1 - \widetilde{v}_{jm}^T)^{-1}(v_{jm} - \widetilde{v}_{jm}), \\ \delta_c(\mathbf{D}_c) &\leq \textstyle\sum_{j=1}^{r_c} \sum_{m=1}^{q_c} f(\widetilde{w}_{c,jm}) + \frac{\eta(\kappa,T)}{2}|w_{c,jm} - \widetilde{w}_{c,jm}|^2 \\ &\quad + 2\,\mathrm{Re}\,\widetilde{w}_{c,jm}^{T-1}(1 - \widetilde{w}_{c,jm}^T)^{-1}(w_{c,jm} - \widetilde{w}_{c,jm}),\end{aligned} \tag{3.21}$$

for all $\mathbf{C}, \widetilde{\mathbf{C}} \in \mathbb{L}^{\leq\kappa,B}_{r_0 \times r_0 \times q_0}$ and $\mathbf{D}_c, \widetilde{\mathbf{D}}_c \in \mathbb{L}^{\leq\kappa,B}_{r_c \times r_c \times q_c}$. These inequalities hold with equality when $\mathbf{C} = \widetilde{\mathbf{C}}$ and $\mathbf{D}_c = \widetilde{\mathbf{D}}_c$, as this implies that all main-diagonal zeros are equal, $v_{jm} = \widetilde{v}_{jm}$ and $w_{c,jm} = \widetilde{w}_{c,jm}$ for all $c, j, m$ up to permutation. To use these inequalities in the construction of the

surrogate function, define $\gamma_*(\mathbf{C}|\widetilde{\boldsymbol{\theta}})$ and $\delta_{c,*}(\mathbf{D}_c|\widetilde{\boldsymbol{\theta}})$ to be

$$\gamma_*(\mathbf{C}|\widetilde{\boldsymbol{\theta}}) \equiv \sum_{j=1}^{r_0}\sum_{m=1}^{q_0} \frac{\eta(\kappa,T)}{2}|v_{jm}-\widetilde{v}_{jm}|^2 + 2\,\mathrm{Re}\,\widetilde{v}_{jm}^{T-1}(1-\widetilde{v}_{jm}^{T})^{-1}(v_{jm}-\widetilde{v}_{jm}),$$

$$\delta_{c,*}(\mathbf{D}_c|\widetilde{\boldsymbol{\theta}}) \equiv \sum_{j=1}^{r_c}\sum_{m=1}^{q_c} \frac{\eta(\kappa,T)}{2}|w_{c,jm}-\widetilde{w}_{c,jm}|^2 + 2\,\mathrm{Re}\,\widetilde{w}_{c,jm}^{T-1}(1-\widetilde{w}_{c,jm}^{T})^{-1}(w_{c,jm}-\widetilde{w}_{c,jm}).$$

### 3.3.4 Block-Diagonal Bounds in the Loewner Order

As mentioned previously, separation of the signal-related parameter blocks and the $C$ channel-specific parameter blocks is prevented by the fact that $\mathbf{T}(\omega)$ and related matrices are not generally block-diagonal. However, the results of this section allow us to sidestep this problem through the use of majorizing inequality (3.20) with a constructed block-diagonal matrix which is the needed upper bound in the Loewner order (Bhatia, 2007). This block-diagonal matrix is constructed in three steps, first by finding a block-diagonal lower bound for the Gram matrix $\mathbf{R}$, finding a block-diagonal upper bound for the transformed periodogram $\mathbf{J}(\omega)$, and combining these to obtain an upper bound for $\mathbf{T}(\omega) + \mathbf{T}(\omega)\mathbf{J}(\omega)\mathbf{T}(\omega)$.

The first quantity for which we will obtain a block-diagonal lower bound is $\mathbf{R}$, the Gram matrix for the columns of $[\mathbf{A}\ \mathbf{B}]$ in the inner product induced by $\boldsymbol{\Phi}$. To do so, let $\sigma$ be the smallest principal angle between $\operatorname{col}\mathbf{A}$ and $\operatorname{col}\mathbf{B}_1 \oplus \cdots \oplus \operatorname{col}\mathbf{B}_C$,

$$\sigma = \min_{\mathbf{w},\mathbf{y}} \left\{ \arccos \frac{|\mathbf{w}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{y}|}{(\mathbf{w}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{w})^{\frac{1}{2}}(\mathbf{y}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{y})^{\frac{1}{2}}} \right\} \tag{3.22}$$

for $\mathbf{w} \in \operatorname{col} A$ and $\mathbf{y} \in \operatorname{col}\mathbf{B}$. Note that $0 \le \sigma \le \pi/2$ and $\sigma \neq 0$ if $\operatorname{col}\mathbf{A}$ and $\operatorname{col}\mathbf{B}_1 \oplus \cdots \oplus \operatorname{col}\mathbf{B}_C$ intersect trivially. Recalling that $\mathbf{A}_c \in \mathbb{C}_{n_c \times r_c}$ is the block of $\mathbf{A}$ which corresponds to channel $c$, let $\rho_c$ for $c = 1, \ldots, C$ be the smallest principal angle between $\operatorname{col}\mathbf{A}_c$ and $\operatorname{col}\mathbf{B}_c$ respectively,

$$\rho_c = \min_{\mathbf{w}_c,\mathbf{y}_c} \left\{ \arccos \frac{|\mathbf{w}_c^{\mathsf{H}}\boldsymbol{\Phi}_c^{-1}\mathbf{y}_c|}{(\mathbf{w}_c^{\mathsf{H}}\boldsymbol{\Phi}_c^{-1}\mathbf{w}_c)^{\frac{1}{2}}(\mathbf{y}_c^{\mathsf{H}}\boldsymbol{\Phi}_c^{-1}\mathbf{y}_c)^{\frac{1}{2}}} \right\}, \tag{3.23}$$

for $\mathbf{w}_c \in \operatorname{col}\mathbf{A}_c$ and $\mathbf{y}_c \in \operatorname{col}\mathbf{B}_c$. If $\sigma$ is large, the signal and combined interference subspaces are close to orthogonal, while if $\rho_c$ is large, the channel-c interference subspace and the projection of the signal subspace onto channel $c$ are nearly orthogonal. So, $\sigma$ and $\rho_1, \ldots, \rho_C$ capture distinct elements

of the geometric relationships between the signal and interference subspaces. The following proposition uses the two distinct perspectives to find block-diagonal lower bounds.

**Proposition 13.** *If* $\mathbf{R}_{*,1}$ *and* $\mathbf{R}_{*,2}$ *are the factor-patterned matrices*

$$\mathbf{R}_{*,1} = \text{blkdiag}\left((1-\cos\sigma)\mathbf{R}_{00}, \ldots, (1-\cos\sigma)\mathbf{R}_{CC}\right)$$

$$\mathbf{R}_{*,2} = \text{blkdiag}\left(\textstyle\sum_{c=1}^{C}(1-\cos\rho_c)\mathbf{A}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{A}_c, (1-\cos\rho_1)\mathbf{R}_{11}, \ldots, (1-\cos\rho_C)\mathbf{R}_{CC}\right),$$

*then* $\mathbf{R}_{*,1} \preceq \mathbf{R}$ *and* $\mathbf{R}_{*,2} \preceq \mathbf{R}$.

*Proof.* See Section C.5 in Appendix C for proof. □

If $\sigma \neq 0$, then $\mathbf{R}_{*,1}$ is positive definite. In contrast, $\mathbf{R}_{*,2}$ will fail to be positive definite if any $\rho_c, c = 1, \ldots, C$ is zero, which may occur if $r_0 + r_c > n_c$. If $\sigma$ is close to zero yet $\rho_c$ is large, the $c+1$th diagonal block in $\mathbf{R}_{*,2}$ can be much larger than the same block in $\mathbf{R}_{*,1}$, and so appropriately combining the lower bounds will improve the tightness.

As both $\mathbf{R}_{*,1}$ and $\mathbf{R}_{*,2}$ are block-diagonal lower bounds for $\mathbf{R}$, so to is any convex combination $\lambda\mathbf{R}_{*,1} + (1-\lambda)\mathbf{R}_{*,2}$. To select a combination of $\mathbf{R}_{*.1}$ and $\mathbf{R}_{*,2}$, a reasonable choice is to choose the combination which is closest to $\mathbf{R}$ in the Frobenius norm by minimizing $||\mathbf{R} - \lambda\mathbf{R}_{*,1} - (1-\lambda)\mathbf{R}_{*.2}||_F^2$ for $\lambda \in [0,1]$. As $\mathbf{R}_{*,1}$ and $\mathbf{R}_{*,2}$ are block-diagonal, some algebra shows that the unconstrained global minimizer is

$$\frac{\sum_{c=1}^{C}(\cos\rho_c - \cos\sigma)\cos\rho_c(||\mathbf{A}_c\mathbf{\Phi}_c^{-1}\mathbf{A}_c||_F^2 + ||\mathbf{B}_c\mathbf{\Phi}_c^{-1}\mathbf{B}_c||_F^2)}{\sum_{c=1}^{C}(\cos\rho_c - \cos\sigma)^2(||\mathbf{A}_c\mathbf{\Phi}_c^{-1}\mathbf{A}_c||_F^2 + ||\mathbf{B}_c\mathbf{\Phi}_c^{-1}\mathbf{B}_c||_F^2)}. \tag{3.24}$$

The denominator is zero iff $\mathbf{R}_{*,1} = \mathbf{R}_{*,2}$ as $||\mathbf{B}_c\mathbf{\Phi}_c^{-1}\mathbf{B}_c||_F^2$ is positive for all $c$. If $\mathbf{R}_{*,1} \neq \mathbf{R}_{*,2}$, $||\mathbf{R} - \lambda\mathbf{R}_{*,1} - (1-\lambda)\mathbf{R}_{*.2}||_F^2$ is a non-negative upward-opening quadratic, so the constrained global minimizer $\lambda_*$ is equal to (3.24) if it is within $[0,1]$ or the nearest endpoint if it is outside the interval. Define the linear combination,

$$\mathbf{R}_* = \lambda_*\mathbf{R}_{*,1} + (1-\lambda_*)\mathbf{R}_{*,2}, \tag{3.25}$$

which satisfies $\mathbf{R}_* \preceq \mathbf{R}$ and is block-diagonal. The lower bound $\mathbf{R}_*$ then yields a block-diagonal upper bound $\mathbf{T}_*(\omega) \succeq \mathbf{T}(\omega)$ for each $\omega$ by

$$\mathbf{T}_*(\omega) = \big(\mathbf{E}(\omega) + \mathbf{L}^{\mathsf{H}}(\omega)\mathbf{R}_*\mathbf{L}(\omega)\big)^{-1}, \tag{3.26}$$

where $\mathbf{L}(\omega) = \operatorname{blkdiag}\big(\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega}), \mathbf{N}_1(\mathrm{e}^{\mathrm{i}\omega}), \ldots, \mathbf{N}_C(\mathrm{e}^{\mathrm{i}\omega})\big)$. That $\mathbf{T}_*(\omega)$ is an upper bound for $\mathbf{T}(\omega)$ follows from

$$\mathbf{F}^{\mathsf{H}}(\omega)\mathbf{\Phi}^{-1}\mathbf{F}(\omega) = \mathbf{L}^{\mathsf{H}}(\omega)\mathbf{R}\mathbf{L}(\omega) \succeq \mathbf{L}^{\mathsf{H}}(\omega)\mathbf{R}_*\mathbf{L}(\omega).$$

Hence $\mathbf{T}^{-1}(\omega) - \mathbf{T}_*^{-1}(\omega) \succeq \mathbf{0}$, so $\mathbf{T}_*(\omega) \succeq \mathbf{T}(\omega)$ for all $\omega$.

Next, we require a block-diagonal upper bound for the transformed periodogram $\mathbf{J}(\omega)$. To obtain this, let $\mathbf{P}_0$ be the projection onto $\operatorname{col}\mathbf{A}$ which is orthogonal in the $\mathbf{\Phi}$ inner product, that is, $\mathbf{P}_0 = \mathbf{A}(\mathbf{A}^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{H}}\mathbf{\Phi}^{-1}$. Similarly, let $\mathbf{P}_c = \mathbf{B}_c(\mathbf{B}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{B}_c)^{-1}\mathbf{B}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}$ be the appropriate projection onto $\operatorname{col}\mathbf{B}_C$ for $c = 1, \ldots, C$. Let $m_0(\omega) \in \mathbb{R}$ be the relative magnitude of the periodogram within $\operatorname{col}\mathbf{A}$, and similarly let $m_c(\omega),\ c = 1, \ldots, C$ be the relative magnitude of the channel-$c$ periodogram within $\operatorname{col}\mathbf{B}_c$. If $\mathbf{x}(\omega)$ is not $\mathbf{\Phi}$-orthogonal to $\operatorname{col}\mathbf{A} \oplus \operatorname{col}\mathbf{B}_1 \oplus \ldots \operatorname{col}\mathbf{B}_C$, then $m_0(\omega)$ is

$$m_0(\omega) \equiv \frac{\operatorname{tr}\mathbf{\Phi}^{-1}\mathbf{P}_0\mathbf{I}(\omega)\mathbf{P}_0^{\mathsf{H}}}{\operatorname{tr}\mathbf{\Phi}^{-1}\mathbf{P}_0\mathbf{I}(\omega)\mathbf{P}_0^{\mathsf{H}} + \sum_{c=1}^{C}\operatorname{tr}\mathbf{\Phi}_c^{-1}\mathbf{P}_c\mathbf{I}_{cc}(\omega)\mathbf{P}_c^{\mathsf{H}}}.$$

If the denominator were zero, then $\mathbf{P}_0\mathbf{x}(\omega)$ and $\mathbf{P}_c\mathbf{x}_c(\omega)$ are zero for all $c$, which occurs only if $\mathbf{x}(\omega)$ is $\mathbf{\Phi}$-orthogonal to both $\operatorname{col}\mathbf{A}$ and $\operatorname{col}\mathbf{B}_1 \oplus \ldots \operatorname{col}\mathbf{B}_C$. Similarly, let $m_c$ be

$$m_c(\omega) \equiv \frac{\operatorname{tr}\mathbf{\Phi}_c^{-1}\mathbf{P}_c\mathbf{I}_{cc}(\omega)\mathbf{P}_c^{\mathsf{H}}}{\operatorname{tr}\mathbf{\Phi}^{-1}\mathbf{P}_0\mathbf{I}(\omega)\mathbf{P}_0^{\mathsf{H}} + \sum_{c=1}^{C}\operatorname{tr}\mathbf{\Phi}_c^{-1}\mathbf{P}_c\mathbf{I}_{cc}(\omega)\mathbf{P}_c^{\mathsf{H}}}$$

for $c = 1, \ldots, C$. The following proposition uses these magnitudes to obtain a block-diagonal upper bound to $\mathbf{J}(\omega)$. Note that, if $m_0(\omega) = 0$ for some $\omega$, then $\mathbf{P}_0\mathbf{x}(\omega) = \mathbf{0}$ and so $\mathbf{x}(\omega)$ is $\mathbf{\Phi}$-orthogonal to $\operatorname{col}\mathbf{A}$. Therefore, $\mathbf{A}^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{x}(\omega) = \mathbf{0}$ and so $\mathbf{J}_{00}(\omega) = \mathbf{0}$. Taking $m_0^{-1}(\omega)\mathbf{0} = \mathbf{0}$ when $m_0(\omega) = 0$ and similarly for all $m_c(\omega)$, the upper bound $\mathbf{J}_*(\omega)$ is well-defined.

**Proposition 14.** *If $\mathbf{J}_*(\omega)$ is the factor-patterned matrix*

$$\mathbf{J}_*(\omega) = \operatorname{blkdiag}(m_0^{-1}(\omega)\mathbf{J}_{00}(\omega), \ldots, m_C^{-1}(\omega)\mathbf{J}_{CC}(\omega)),$$

*then* $\mathbf{J}_*(\omega) \succeq \mathbf{J}(\omega)$ *for all* $\omega \in (-\pi, \pi]$.

*Proof.* See Section C.6 in Appendix C for proof. □

Last, it is necessary to obtain a block-diagonal upper bound for the quantity

$$\mathbf{G}(\omega) \equiv \mathbf{T}(\omega) + \mathbf{T}(\omega)\mathbf{J}(\omega)\mathbf{T}(\omega). \tag{3.27}$$

From the previous steps, we have the block-diagonal upper bounds $\mathbf{T}(\omega) \preceq \mathbf{T}_*(\omega)$ and $\mathbf{J}(\omega) \preceq \mathbf{J}_*(\omega)$. However, it does not follow that $\mathbf{T}_*(\omega) + \mathbf{T}_*(\omega)\mathbf{J}_*(\omega)\mathbf{T}_*(\omega)$ is an upper bound for $\mathbf{G}(\omega)$. Instead, to use $\mathbf{T}_*(\omega)$ and $\mathbf{J}_*(\omega)$ to find a block-diagonal bound, note that $\mathbf{G}(\omega)$ is the sum of a positive definite matrix plus a rank-one positive semi-definite matrix. Use of the following lemma is the central step to obtain a bound for $\mathbf{G}(\omega)$.

**Lemma 1.** (Tilted SM) *For non-zero* $\mathbf{x} \in \mathbb{C}_m$ *and* $a, b \in \mathbb{R}$ *with* $b \neq -\mathbf{x}^{\mathsf{H}}\mathbf{x}$*, the matrix*

$$a\mathbf{I}_m - \frac{b\mathbf{x}\mathbf{x}^{\mathsf{H}}}{b + \mathbf{x}^{\mathsf{H}}\mathbf{x}} \tag{3.28}$$

*is positive definite iff* $a > 0$ *and*

$$\max\{\frac{a}{\mathbf{x}^{\mathsf{H}}\mathbf{x}} + \frac{a}{b}, \frac{\mathbf{x}^{\mathsf{H}}\mathbf{x}}{b + \mathbf{x}^{\mathsf{H}}\mathbf{x}}\} > 1, \tag{3.29}$$

*in which case the inverse is*

$$a^{-1}\mathbf{I}_m + \frac{(a^{-1}b)\mathbf{x}\mathbf{x}^{\mathsf{H}}}{ab + (a-b)\mathbf{x}^{\mathsf{H}}\mathbf{x}}.$$

*Proof.* See Section C.7 in Appendix C for proof. □

If $a = b = 1$, the lemma reduces to the usual Sherman-Morrison formula for a rank-one update to the identity.

**Proposition 15.** *If* $\mathbf{G}_*(\omega) = \mathbf{T}_*(\omega)$ *for* $\mathbf{J}(\omega) = \mathbf{0}$ *and*

$$\mathbf{G}_*(\omega) = m(\omega)\lambda(\omega)\big(\mathbf{T}_*(\omega) + h^{-1}(\omega)\mathbf{T}_*(\omega)\mathbf{J}_*(\omega)\mathbf{T}_*(\omega)\big)$$

*for* $\mathbf{J}(\omega) \neq \mathbf{0}$*, where*

$$
\begin{aligned}
m(\omega) &= \max\{1, \operatorname{tr} \mathbf{T}_*(\omega)\mathbf{J}(\omega)\}, \\
\lambda(\omega) &= \lambda_{\max}\big(\mathbf{T}_*^{-\frac{1}{2}}(\omega)\mathbf{T}(\omega)\mathbf{T}_*^{-\frac{1}{2}}(\omega)\big), \\
h(\omega) &= (\lambda^{-1}(\omega) - 1) \operatorname{tr} \mathbf{T}_*(\omega)\mathbf{J}(\omega) + \lambda^{-1}(\omega) \operatorname{tr} \mathbf{T}(\omega)\mathbf{J}(\omega),
\end{aligned}
$$

*then* $\mathbf{G}_*(\omega)$ *is block-diagonal with* $\mathbf{G}_*(\omega) \succeq \mathbf{G}(\omega)$.

*Proof.* See Section C.7 in Appendix C for proof. □

As $\mathbf{0} \prec \mathbf{T}(\omega) \preceq \mathbf{T}_*(\omega)$ implies $0 < \lambda(\omega) \leq 1$ and $\operatorname{tr} \mathbf{T}(\omega)\mathbf{J}(\omega) > 0$ when $\mathbf{J}(\omega) \neq \mathbf{0}$ by $\mathbf{T}(\omega)$ positive definite, $\lambda^{-1}(\omega)$ and $h^{-1}(\omega)$ exist and $\mathbf{G}_*(\omega)$ is well-defined. The block-diagonal upper bound $\mathbf{G}_*(\omega)$ allows for construction of a surrogate function which separates the signal and interference related parameter blocks, as described in the following section.

### 3.3.5 Surrogate Function $\mathrm{g}(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}})$

The majorizing inequalities (3.18) - (3.20) can be combined with the block-diagonal upper bound to obtain a surrogate function which separates the signal- and interference-related parameter blocks and enables closed form updates to the various parameters. To do so, let $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ be two parameter tuples, where the components of $\boldsymbol{\theta}$ are denoted as (3.9) and the components of $\widetilde{\boldsymbol{\theta}}$ are denoted as $(\widetilde{\boldsymbol{\Phi}}, \ldots, \widetilde{\boldsymbol{\Upsilon}}_C)$. Quantities such as $\mathbf{T}(\omega)$ dependent on the parameter values in $\widetilde{\boldsymbol{\theta}}$ are similarly denoted as $\widetilde{\mathbf{T}}(\omega)$.

Applying (3.18) with $\mathbf{S} = \mathbf{T}^{-1}(\omega)$ and $\widetilde{\mathbf{S}} = \widetilde{\mathbf{T}}^{-1}(\omega)$ yields the majorizing inequality

$$
-\log\det \mathbf{T}(\omega) \leq -\log\det \widetilde{\mathbf{T}}(\omega) - (r_0 + r) + \operatorname{tr} \widetilde{\mathbf{T}}(\omega)\mathbf{E}(\omega) + \operatorname{tr} \widetilde{\mathbf{T}}(\omega)\mathbf{F}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}^{-1}\mathbf{F}(\omega),
$$

for all $\omega$, with equality at $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$. Applying (3.19) with $\mathbf{V} = \mathbf{J}(\omega)$, $\widetilde{\mathbf{V}} = \widetilde{\mathbf{J}}(\omega)$, $\mathbf{S} = \mathbf{T}^{-1}(\omega)$ and $\widetilde{\mathbf{S}} = \widetilde{\mathbf{T}}^{-1}(\omega)$ yields the majorizing inequality

$$
-\operatorname{tr} \mathbf{T}(\omega)\mathbf{J}(\omega) \leq \operatorname{tr} \widetilde{\mathbf{T}}(\omega)\widetilde{\mathbf{J}}(\omega)\widetilde{\mathbf{T}}(\omega)\big(\mathbf{E}(\omega) + \mathbf{F}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}^{-1}\mathbf{F}(\omega)\big) - 2\operatorname{Re}\operatorname{tr} \widetilde{\mathbf{T}}(\omega)\widetilde{\mathbf{J}}^{\frac{1}{2}}(\omega)(\mathbf{J}^{\frac{1}{2}}(\omega))^{\mathsf{H}},
$$

where $\mathbf{J}^{\frac{1}{2}}(\omega)$ and $\widetilde{\mathbf{J}}^{\frac{1}{2}}(\omega)$ are respective matrix square roots of $\mathbf{J}(\omega)$ and $\widetilde{\mathbf{J}}(\omega)$ where $\boldsymbol{\theta}$ equalling $\widetilde{\boldsymbol{\theta}}$ implies $\mathbf{J}^{\frac{1}{2}}(\omega)$ equals $\widetilde{\mathbf{J}}^{\frac{1}{2}}(\omega)$. In particular, we can take

$$\mathbf{J}^{\frac{1}{2}}(\omega) = \mathbf{F}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}^{-1}\mathbf{x}(\omega), \qquad \widetilde{\mathbf{J}}^{\frac{1}{2}}(\omega) = \widetilde{\mathbf{F}}^{\mathsf{H}}(\omega)\widetilde{\boldsymbol{\Phi}}^{-1}\mathbf{x}(\omega).$$

Combining these inequalities with the negative log majorization of Section 3.3.3 and applying them to (3.17) yields

$$\begin{aligned}\ell(\boldsymbol{\theta}) \leq & -\text{const} + \epsilon(\boldsymbol{\theta}) + \gamma_*(\mathbf{C}|\widetilde{\boldsymbol{\theta}}) + \textstyle\sum_{c=1}^{C} \delta_{c,*}(\mathbf{D}_c|\widetilde{\boldsymbol{\theta}}) \\ & + T^{-1} \textstyle\sum_{k=0}^{T-1} \operatorname{tr} \widetilde{\mathbf{G}}(\omega_k)\mathbf{E}(\omega_k) \\ & + T^{-1} \textstyle\sum_{k=0}^{T-1} \operatorname{tr} \widetilde{\mathbf{G}}(\omega_k)\mathbf{F}^{\mathsf{H}}(\omega_k)\boldsymbol{\Phi}^{-1}\mathbf{F}(\omega_k) \\ & - 2T^{-1} \textstyle\sum_{k=0}^{T-1} \operatorname{Re}\operatorname{tr} \widetilde{\mathbf{T}}(\omega_k)\widetilde{\mathbf{F}}^{\mathsf{H}}(\omega_k)\widetilde{\boldsymbol{\Phi}}^{-1}\mathbf{I}(\omega_k)\boldsymbol{\Phi}^{-1}\mathbf{F}(\omega_k)\end{aligned} \tag{3.30}$$

where $\widetilde{\mathbf{G}}(\omega)$ is $\widetilde{\mathbf{T}}(\omega) + \widetilde{\mathbf{T}}(\omega)\widetilde{\mathbf{J}}(\omega)\widetilde{\mathbf{T}}(\omega)$ and const is

$$\begin{aligned}\text{const} = & \, T^{-1} \textstyle\sum_{k=0}^{T-1} \log\det \widetilde{\mathbf{C}}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k})\widetilde{\mathbf{C}}(\mathrm{e}^{\mathrm{i}\omega_k}) + T^{-1} \textstyle\sum_{c=1}^{C} \sum_{k=0}^{T-1} \log\det \widetilde{\mathbf{D}}_c^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k})\widetilde{\mathbf{D}}_c(\mathrm{e}^{\mathrm{i}\omega_k}) \\ & + (r_0 + r) + T^{-1} \textstyle\sum_{k=0}^{T-1} \log\det \widetilde{\mathbf{T}}(\omega_k),\end{aligned}$$

which does not depend on $\boldsymbol{\theta}$.

In the right-hand side of (3.30), multiplicative interaction between signal and interference parameter blocks in $\boldsymbol{\theta}$ occurs only in the terms of the form $\operatorname{tr} \widetilde{\mathbf{G}}(\omega)\mathbf{F}(\omega)\boldsymbol{\Phi}^{-1}\mathbf{F}^{\mathsf{H}}(\omega)$, as $\mathbf{E}(\omega)$ is block-diagonal. By using the block-diagonal upper bound for $\widetilde{\mathbf{G}}(\omega)$ found in Proposition 15 in conjunction with the inequality (3.20), this interaction can be eliminated. Taking $\mathbf{X} = \mathbf{F}(\omega)$ and $\widetilde{\mathbf{X}} = \widetilde{\mathbf{F}}(\omega)$ while $\mathbf{D} = \boldsymbol{\Phi}^{-1}$, $\mathbf{W} = \tilde{\mathbf{G}}(\omega)$ and $\mathbf{M} = \mathbf{G}_*(\omega)$ in (3.20), we further obtain

$$\begin{aligned}\operatorname{tr} \widetilde{\mathbf{G}}(\omega)\mathbf{F}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}^{-1}\mathbf{F}(\omega) \leq & \operatorname{tr} \widetilde{\mathbf{G}}_*(\omega)\mathbf{F}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}^{-1}\mathbf{F}(\omega) - 2\operatorname{Re}\operatorname{tr}\big(\widetilde{\mathbf{G}}_*(\omega) - \widetilde{\mathbf{G}}(\omega)\big)\widetilde{\mathbf{F}}(\omega)^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{F}(\omega) \\ & + \operatorname{tr}\big(\widetilde{\mathbf{G}}_*(\omega) - \widetilde{\mathbf{G}}(\omega)\big)\widetilde{\mathbf{F}}(\omega)^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\widetilde{\mathbf{F}}(\omega).\end{aligned}$$

To combine the above majorizations into a surrogate function, construct the coefficient $n \times (r_0 + r)$ matrix $\mathbf{K}(\omega)$ as

$$\widetilde{\mathbf{K}}(\omega) = \widetilde{\mathbf{F}}(\omega)\big(\widetilde{\mathbf{G}}_*(\omega) - \widetilde{\mathbf{G}}(\omega)\big) + \mathbf{I}(\omega)\widetilde{\boldsymbol{\Phi}}^{-1}\widetilde{\mathbf{F}}(\omega)\widetilde{\mathbf{T}}(\omega).$$

The matrix $\widetilde{\mathbf{K}}(\omega)$ can be partitioned in a fashion similar to $[\mathbf{A}\ \mathbf{B}]$,

$$\widetilde{\mathbf{K}}(\omega) = \begin{array}{c} \begin{matrix} r_0 & r_1 & \dots & r_C \end{matrix} \\ \begin{bmatrix} \widetilde{\mathbf{K}}_{10}(\omega) & \widetilde{\mathbf{K}}_{11}(\omega) & \cdots & \widetilde{\mathbf{K}}_{1C}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathbf{K}}_{C0}(\omega) & \widetilde{\mathbf{K}}_{C1}(\omega) & \dots & \widetilde{\mathbf{K}}_{CC}.(\omega) \end{bmatrix} \end{array} \begin{matrix} n_1 \\ \vdots \\ \\ \\ \\ \\ n_C \end{matrix},$$

and let $\widetilde{\mathbf{K}}_0(\omega) \equiv [\widetilde{\mathbf{K}}_{10}^{\mathsf{T}}(\omega), \dots, \widetilde{\mathbf{K}}_{C0}^{\mathsf{T}}(\omega)]^{\mathsf{T}}$. In terms of the parameter blocks in $\boldsymbol{\theta}$ as well as the quantities $\widetilde{\mathbf{G}}_*(\omega), \widetilde{\mathbf{G}}(\omega), \widetilde{\mathbf{F}}(\omega)$ and $\widetilde{\mathbf{K}}(\omega)$ depending on $\widetilde{\boldsymbol{\theta}}$, the surrogate function can be constructed as a sum of the following quantities. Define the surrogate parts $a$ and $b_c$ as

$$\begin{aligned} a(\mathbf{A}, \mathbf{M}, \boldsymbol{\Phi}|\widetilde{\boldsymbol{\theta}}) \equiv \frac{1}{T}\sum_{k=0}^{T-1} \Big[ & \operatorname{tr} \widetilde{\mathbf{G}}_{*,00}(\omega_k)\mathbf{M}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k})\mathbf{A}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{A}\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega_k}) \\ & - 2\operatorname{Re}\operatorname{tr} \widetilde{\mathbf{K}}_0^{\mathsf{H}}(\omega_k)\boldsymbol{\Phi}^{-1}\mathbf{A}\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega_k})\Big], \\ b_c(\mathbf{B}_c, \mathbf{N}_c, \boldsymbol{\Phi}_c|\widetilde{\boldsymbol{\theta}}) \equiv \frac{1}{T}\sum_{k=0}^{T-1}\Big[ & \operatorname{tr} \widetilde{\mathbf{G}}_{*,cc}(\omega_k)\mathbf{N}_c^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k})\mathbf{B}_c^{\mathsf{H}}\boldsymbol{\Phi}_c^{-1}\mathbf{B}_c\mathbf{N}_c(\mathrm{e}^{\mathrm{i}\omega_k}) \\ & - 2\operatorname{Re}\operatorname{tr} \widetilde{\mathbf{K}}_{cc}^{\mathsf{H}}(\omega_k)\boldsymbol{\Phi}_c^{-1}\mathbf{B}_c\mathbf{N}_c(\mathrm{e}^{\mathrm{i}\omega_k})\Big], \end{aligned} \tag{3.31}$$

and define the surrogate parts $c, d_c$ and $e$ as

$$\begin{aligned} c(\mathbf{C}, \boldsymbol{\Psi}|\widetilde{\boldsymbol{\theta}}) &\equiv \frac{1}{T}\sum_{k=0}^{T-1} \operatorname{tr} \widetilde{\mathbf{G}}_{00}(\omega)\mathbf{C}^{\mathsf{H}}(\mathrm{e}^{-\mathrm{i}\omega_k})\boldsymbol{\Psi}^{-1}\mathbf{C}(\mathrm{e}^{\mathrm{i}\omega_k}) + \gamma_*(\mathbf{C}|\widetilde{\boldsymbol{\theta}}), \\ d_c(\mathbf{D}_c, \boldsymbol{\Upsilon}_c|\widetilde{\boldsymbol{\theta}}) &\equiv \frac{1}{T}\sum_{k=0}^{T-1} \operatorname{tr} \widetilde{\mathbf{G}}_{cc}(\omega)\mathbf{D}_c^{\mathsf{H}}(\mathrm{e}^{-\mathrm{i}\omega})\boldsymbol{\Upsilon}_c^{-1}\mathbf{D}_c(\mathrm{e}^{\mathrm{i}\omega}) + \delta_{c,*}(\mathbf{D}_c|\widetilde{\boldsymbol{\theta}}), \\ e(\boldsymbol{\Phi}|\widetilde{\boldsymbol{\theta}}) &\equiv \frac{1}{T}\sum_{k=0}^{T-1} \operatorname{tr} \boldsymbol{\Phi}^{-1}\widetilde{\mathbf{F}}(\omega_k)(\widetilde{\mathbf{G}_*}(\omega_k) - \widetilde{\mathbf{G}}(\omega_k))\widetilde{\mathbf{F}}^{\mathsf{H}}(\omega_k). \end{aligned} \tag{3.32}$$

Combining the above parts, we obtain the surrogate function,

$$\begin{aligned} g(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}) = -\mathrm{const} &+ a(\mathbf{A}, \mathbf{M}, \boldsymbol{\Phi}|\widetilde{\boldsymbol{\theta}}) + c(\mathbf{C}, \boldsymbol{\Psi}|\widetilde{\boldsymbol{\theta}}) + e(\boldsymbol{\Phi}|\widetilde{\boldsymbol{\theta}}) \\ &+ \sum_{c=1}^{C} b_c(\mathbf{B}_c, \mathbf{N}_c, \boldsymbol{\Phi}_c|\widetilde{\boldsymbol{\theta}}) + d_c(\mathbf{D}_c.\boldsymbol{\Upsilon}_c|\widetilde{\boldsymbol{\theta}}), \end{aligned} \tag{3.33}$$

which possesses the desired upper-bound properties as it is iteratively constructed through application of majorizing inequalities which possess the needed properties.

## 3.4 Algorithm

Through the manipulations of Section 3.3.2, a surrogate function $g(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}})$ is obtained which separates the parameters into the (overlapping) blocks $(\mathbf{A}, \mathbf{M}, \boldsymbol{\Phi}), (\mathbf{B}_1, \mathbf{N}_1, \boldsymbol{\Phi}_1), \ldots, (\mathbf{B}_C, \mathbf{N}_C, \boldsymbol{\Phi}_C)$ and the non-overlapping blocks $(\mathbf{C}, \boldsymbol{\Psi}), (\mathbf{D}_1, \boldsymbol{\Upsilon}_1), \ldots (\mathbf{D}_C, \boldsymbol{\Upsilon}_C)$. The first collection of blocks are overlapping only by the mutual dependence on the noise variances $\boldsymbol{\Phi}$.

If the parameter values at step $s$ are denoted by $\boldsymbol{\theta}^{(s)}$, new parameter values $\boldsymbol{\theta}^{(s+1)}$ which satisfy $g(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\theta}^{(s)}) \leq g(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s)}) = \ell(\boldsymbol{\theta}^{(s)})$ can be sequentially obtained through block coordinate descent applied to $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$. In particular, after first updating $\boldsymbol{\Phi}^{(s)}$ to $\boldsymbol{\Phi}^{(s+1)}$, the remainder of the block updates can be performed in parallel as the optimization subproblems are decoupled when $\boldsymbol{\Phi}$ is fixed at $\boldsymbol{\Phi}^{(s+1)}$. Within each block, updating the block components in an appropriate sequence allows for each update to be obtained in closed-form. The outline of the estimation procedure is given by Algorithm 1, and the details of the update calculations are given in Subsections 3.4.1 through 3.4.4. In Algorithm 1, a default initialization for $\boldsymbol{\theta}^{(s)}$ is provided; alternatively, MFA can be used to provide initial values for $\mathbf{A}, \mathbf{B}, \boldsymbol{\Psi}, \boldsymbol{\Upsilon}$ and $\boldsymbol{\Phi}$.

By application of a general result for global convergence of Majorization-Minimization procedures (Razaviyayn et al., 2013) and the generic unique representative result of Proposition 12, the sequence of parameter values $\{\boldsymbol{\theta}^{(s)}\}_{s=0}^{\infty}$ can be guaranteed to converge to a stationary point of $\ell(\boldsymbol{\theta})$ in most cases of interest. This convergence result is given in the following theorem.

**Theorem 5.** (Convergence) *Suppose the distribution of the random vector* $\left[\mathbf{x}^{\mathsf{T}}[0], \ldots, \mathbf{x}^{\mathsf{T}}[T-1]\right]^{\mathsf{T}}$ *is absolutely continuous with respect to the Lebesgue measure, the channel sizes* $\mathbf{n}$ *and factor numbers* $\mathbf{r}$ *are such that MFSA is generically globally identifiable in the sense of Section 3.2.4, and* $T > \max\{||\mathbf{r}||_\infty, ||\mathbf{p}||_\infty, ||\mathbf{q}||_\infty, n\}$. *If these assumptions are met, the sequence of iterates* $\{\boldsymbol{\theta}^{(s)}\}_{s=0}^{\infty}$ *converges to a stationary point of* $\ell(\boldsymbol{\theta})$ *almost surely.*

*Proof.* See Section C.8 in Appendix C for proof. □

**Algorithm 1:** MFSA Estimation Procedure

**Data:** Periodogram $\mathbf{I}(\omega_k)$, factor numbers $\mathbf{r}$, numerator orders $\mathbf{p}$, denominator orders $\mathbf{q}$
**Result:** Parameter values $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*(\mathbf{r}, \mathbf{p}, \mathbf{q})$ solving (3.12).
Set $s \leftarrow 0$;
Initialize loadings as $\mathbf{A}^{(0)} \leftarrow [\mathbf{I}_{r_0}\ \mathbf{0}]^\mathsf{T}, \mathbf{B}_c^{(0)} \leftarrow [\mathbf{I}_{r_c}\ \mathbf{0}]^\mathsf{T}$;
Initialize idiosyncratic noise variances $\boldsymbol{\Phi}^{(0)} \leftarrow \mathbf{I}_n$;
Initialize input variances $\boldsymbol{\Psi}^{(0)} \leftarrow \mathbf{I}_{r_0}$, $\boldsymbol{\Upsilon}_c^{(0)} \leftarrow \mathbf{I}_{r_c}$;
Initialize $\mathbf{M}^{(0)} \leftarrow \mathbf{0}_{r_0, p_0 r_0}$ and $\mathbf{N}_c^{(0)} \leftarrow \mathbf{0}_{r_c, p_c r_c}$;
Initialize $\mathbf{C}^{(0)} \leftarrow \mathbf{0}_{r_0, q_0 r_0}$ and $\mathbf{D}_c^{(0)} \leftarrow \mathbf{0}_{r_c, q_c r_c}$;
**while** $\boldsymbol{\theta}^{(s)}$ *not converged* **do**
  Update to $\boldsymbol{\Phi}^{(s+1)}$ as (3.35) ;
  Update to $\boldsymbol{\Psi}^{(s+1)}$ and $\boldsymbol{\Upsilon}_c^{(s+1)}$ as (3.37) ;
  Update to $\mathbf{A}^{(s+1)}$ and $\mathbf{B}_c^{(s+1)}$ as (3.40), (3.42);
  Update off-diagonal $\mathbf{C}'^{(s+1)}, \mathbf{D}_c'^{(s+1)}$ as (3.47), (3.54) ;
  Update diagonal $\mathbf{C}''^{(s+1)}, \mathbf{D}_c''^{(s+1)}$ as (3.52), (3.56);
  Update non-SE $\mathbf{M}'^{(s+1)}, \mathbf{N}_c'^{(s+1)}$ as (3.59), (3.64);
  Update SE part of $\mathbf{M}^{(s+1)}, \mathbf{N}_c^{(s+1)}$ as (3.62), (3.65);
  $s \leftarrow s + 1$
Return $\boldsymbol{\theta}^{(s)}$;

### 3.4.1 Updating $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Upsilon}_c$

The parameter most intertwined with other parameters in the surrogate function $g$ is the noise variance $\boldsymbol{\Phi}$ as it occurs in terms $a, b$ and $e$. Holding all the other parameters fixed at their values in step $s$, minimizing $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ for $\boldsymbol{\Phi} \in \mathbb{P}_n^{\geq \epsilon}$ then reduces to

$$\min_{\boldsymbol{\Phi} \in \mathbb{P}_n^{\geq \epsilon}} a(\mathbf{A}^{(s)}, \mathbf{M}^{(s)}, \boldsymbol{\Phi}|\boldsymbol{\theta}^{(s)}) + e(\boldsymbol{\Phi}|\boldsymbol{\theta}^{(s)}) + \sum_{c=1}^{C} b(\mathbf{B}_c^{(s)}, \mathbf{N}_c^{(s)}, \boldsymbol{\Phi}_c|\boldsymbol{\theta}^{(s)}) \tag{3.34}$$

This optimization problem has a simple closed-form solution. To see this, first recall that the majorizing equality (3.20) applied to the surrogate function holds with equality when $\mathbf{F}(\omega) = \widetilde{\mathbf{F}}(\omega)$. However, $\mathbf{F}(\omega)$ is a function of $(\mathbf{A}, \mathbf{M})$ and $(\mathbf{B}_c, \mathbf{N}_c), c = 1, \ldots, C$ which still equal their values in $\boldsymbol{\theta}^{(s)}$. Therefore, the objective in (3.34) in fact equals

$$\log\det \boldsymbol{\Phi} + T^{-1} \textstyle\sum_{k=0}^{T-1} \operatorname{tr} \mathbf{I}(\omega_k) \boldsymbol{\Phi}^{-1} + T^{-1} \sum_{k=0}^{T-1} \mathbf{F}^{(s)}(\omega_k) \mathbf{G}^{(s)}(\omega_k) \mathbf{F}^{(s)\mathsf{H}}(\omega_k) \boldsymbol{\Phi}^{-1}$$

$$-\tfrac{2}{T}\sum_{k=0}^{T-1}\operatorname{Re}\operatorname{tr}\mathbf{F}^{(s)}(\omega_k)\mathbf{T}^{(s)}(\omega_k)\mathbf{F}^{(s)\mathsf{H}}(\omega_k)(\mathbf{\Phi}^{(s)})^{-1}\mathbf{I}(\omega_k)\mathbf{\Phi}^{-1}.$$

To simplify the above expression, define $\mathbf{y}(\omega) \in \mathbb{C}_n$ as

$$\mathbf{y}^{(s)}(\omega) = \Big(\mathbf{I}_n - \mathbf{F}^{(s)}(\omega)\mathbf{T}^{(s)}(\omega)\mathbf{F}^{(s)\mathsf{H}}(\omega)(\mathbf{\Phi}^{(s)})^{-1}\Big)\,\mathbf{x}(\omega).$$

If $\mathbf{u}(\omega)$ is the tapered DTFT of $\mathbf{u}[t]$, then $\mathbf{y}^{(s)}(\omega)$ is the LMMSE estimate of $\mathbf{u}(\omega)$ from $\mathbf{x}(\omega)$ at $\boldsymbol{\theta}^{(s)}$ when the mean squared error is measured in the $\mathbf{\Phi}$ inner product. Expanding $\mathbf{G}^{(s)}(\omega)$ and $\mathbf{J}^{(s)}(\omega)$, the outer product $\mathbf{y}(\omega)\mathbf{y}^{(s)\mathsf{H}}(\omega)$ equals

$$\begin{aligned}\mathbf{y}^{(s)}(\omega)\mathbf{y}^{(s)\mathsf{H}}(\omega) = \mathbf{I}(\omega) &+ \mathbf{F}^{(s)}(\omega)\big(\mathbf{G}^{(s)}(\omega) - \mathbf{T}^{(s)}(\omega)\big)\mathbf{F}^{(s)\mathsf{H}} \\ &- 2\operatorname{Re}\mathbf{F}^{(s)}(\omega)\mathbf{T}^{(s)}(\omega)\mathbf{F}^{(s)\mathsf{H}}(\omega)(\mathbf{\Phi}^{(s)})^{-1}\mathbf{I}(\omega).\end{aligned}$$

Let $\mathbf{V}^{(s)}(\omega)$ be the $n \times n$ positive semi-definite matrix,

$$\mathbf{V}^{(s)}(\omega) = \mathbf{y}^{(s)}(\omega)\mathbf{y}^{(s)\mathsf{H}}(\omega) + \mathbf{F}^{(s)}(\omega)\mathbf{T}^{(s)}(\omega)\mathbf{F}^{(s)\mathsf{H}}(\omega),$$

then the expansion of $\mathbf{y}^{(s)}(\omega)\mathbf{y}^{(s)\mathsf{H}}(\omega)$ gives that the objective in (3.34) equals

$$\log\det\mathbf{\Phi} + \tfrac{1}{T}\sum_{k=0}^{K-1}\operatorname{tr}\mathbf{V}^{(s)}(\omega_k)\mathbf{\Phi}^{-1}.$$

As the noise PSD $\mathbf{\Phi}$ is frequency-flat, the minimizer of (3.34) is

$$\mathbf{\Phi}^{(s+1)} = \operatorname*{argmin}_{\mathbf{\Phi}\in\mathbb{P}_n^{\geq\epsilon}} \log[\mathbf{\Phi}]_{ii} + \frac{[\mathbf{V}^{(s)}]_{ii}}{[\mathbf{\Phi}]_{ii}},$$

where $\mathbf{V}^{(s)}$ is the frequency-average $T^{-1}\sum_{k=0}^{T-1}\mathbf{V}^{(s)}(\omega_k)$. This matrix $\mathbf{V}^{(s)}$ will be almost surely positive definite under the assumption that $T > \max\{||\mathbf{r}||_\infty, ||\mathbf{p}||_\infty, ||\mathbf{q}||_\infty, n\}$ and the joint density of the $\mathbf{x}[t]$ is absolutely continuous with respect to the Lebesgue measure as $\mathbf{y}^{(s)}(\omega)$ is obtained from $\mathbf{x}(\omega)$ by a continuous map. So, (3.34) has a unique minimizer over $\mathbb{P}_n^{\geq\epsilon}$ given by

$$[\mathbf{\Phi}^{(s+1)}]_{ii} = \max\big\{[\mathbf{V}^{(s)}]_{ii}, \epsilon\big\},\ i = 1, \ldots, n \tag{3.35}$$

as $\log x + \frac{a}{x}$ is unimodal on $(0, \infty)$.

Minimizing $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ for $\boldsymbol{\Psi}$ and $\boldsymbol{\Upsilon}_c,\ c = 1, \ldots, C$ while holding $\mathbf{C}$ and $\mathbf{D}_c,\ c = 1, \ldots, C$ fixed at $\mathbf{C}^{(s)}$ and $\mathbf{D}_c^{(s)}$ respectively reduces to similar optimization problems,

$$\begin{aligned}\boldsymbol{\Psi}^{(s+1)} &= \underset{\boldsymbol{\Psi} \in \mathbb{P}_{r_0}^{\geq \epsilon}}{\operatorname{argmin}} \sum_{j=1}^{r_0} \log[\boldsymbol{\Psi}]_{jj} + \frac{[\mathbf{W}_0]_{jj}}{[\boldsymbol{\Psi}]_{jj}} \\ \boldsymbol{\Upsilon}_c^{(s+1)} &= \underset{\boldsymbol{\Upsilon}_c \in \mathbb{P}_{r_c}^{\geq \epsilon}}{\operatorname{argmin}} \sum_{j=1}^{r_c} \log[\boldsymbol{\Psi}]_{jj} + \frac{[\mathbf{W}_c]_{jj}}{[\boldsymbol{\Upsilon}_c]_{jj}},\ c = 1, \ldots, C\end{aligned} \tag{3.36}$$

where $\mathbf{W}^{(s)}$ and $\mathbf{W}_c^{(s)},\ c = 1, \ldots, C$ are

$$\begin{aligned}\mathbf{W}_0^{(s)} &= T^{-1} \textstyle\sum_{k=0}^{T-1} \mathbf{C}^{(s)}(\mathrm{e}^{\mathrm{i}\omega_k}) \mathbf{G}_{00}^{(s)}(\omega_k) \mathbf{C}^{(s)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k}), \\ \mathbf{W}_c^{(s)} &= T^{-1} \textstyle\sum_{k=0}^{T-1} \mathbf{D}_c^{(s)}(\mathrm{e}^{\mathrm{i}\omega_k}) \mathbf{G}_{cc}^{(s)}(\omega_k) \mathbf{D}_c^{(s)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k}).\end{aligned}$$

These problems have also have the unique minimizers,

$$\begin{aligned}[\boldsymbol{\Psi}^{(s+1)}]_{jj} &= \max\left\{[\mathbf{W}_0^{(s)}]_{jj}, \epsilon\right\},\ j = 1, \ldots, r_0, \\ [\boldsymbol{\Upsilon}_c^{(s+1)}]_{jj} &= \max\left\{[\mathbf{W}_c^{(s)}]_{jj}, \epsilon\right\},\ j = 1, \ldots, r_c,\end{aligned} \tag{3.37}$$

for $c = 1, \ldots, C$.

### 3.4.2 Updating $\mathbf{A}$ and $\mathbf{B}_c$

Having updated $\boldsymbol{\Phi}$ to $\boldsymbol{\Phi}^{(s+1)}$ and holding $\mathbf{M}$ fixed at $\mathbf{M}^{(s)}$, the profile optimization for $\mathbf{A}$ is

$$\min_{\mathbf{A} \in \mathbb{L}_{n \times r_0}^1} a(\mathbf{A}, \mathbf{M}^{(s)}, \boldsymbol{\Phi}^{(s+1)}|\boldsymbol{\theta}^{(s)}). \tag{3.38}$$

To solve this problem, set

$$\mathbf{X}_0^{(s)}(\omega) = \mathbf{M}^{(s)}(\mathrm{e}^{\mathrm{i}\omega}) \mathbf{G}_{*,00}^{(s)}(\omega) \mathbf{M}^{(s)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega}), \qquad \mathbf{Y}_0^{(s)}(\omega) = \mathbf{K}_0^{(s)}(\omega) \mathbf{M}^{(s)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega}),$$

and let $\mathbf{X}^{(s)}$ and $\mathbf{Y}^{(s)}$ be the frequency averages $T^{-1} \sum_{k=0}^{T-1} \mathbf{X}^{(s)}(\omega_k)$ and $T^{-1} \sum_{k=0}^{T-1} \mathbf{Y}^{(s)}(\omega_k)$ respectively. The objective in (3.38) then equals

$$\operatorname{tr} \mathbf{X}^{(s)} \mathbf{A}^{\mathsf{H}} (\boldsymbol{\Phi}^{(s+1)})^{-1} \mathbf{A} - 2 \operatorname{Re} \operatorname{tr} \mathbf{Y}^{(s)\mathsf{H}} (\boldsymbol{\Phi}^{(s+1)})^{-1} \mathbf{A}.$$

The matrix $\mathbf{X}^{(s)}$ can be seen to be positive definite. First, note that

$$\tfrac{1}{T}\sum_{k=0}^{T-1}\mathbf{M}^{(s)}(\mathrm{e}^{\mathrm{i}\omega_k}) = \mathbf{I}_{r_0} + \tfrac{1}{T}\sum_{k=0}^{T-1}\sum_{\ell=1}^{p_0}\mathbf{M}_j^{(s)}\mathrm{e}^{\mathrm{i}\omega_k\ell} = \mathbf{I}_{r_0}, \tag{3.39}$$

where the second equality follows from $p_0 < T$. Next, $\mathbf{G}^{(s)}(\omega_k)$ positive definite for each $\omega_k$ implies $\tau = \min_k\{\lambda_{\min}\big(\mathbf{G}^{(s)}(\omega_k)\big)\}$ is positive. Combining the previous facts with the definition (3.7) of $\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega})$, we have

$$\mathbf{X}^{(s)} \succeq \tau T^{-1}\sum_{k=0}^{T-1}\mathbf{M}^{(s)}(\mathrm{e}^{\mathrm{i}\omega_k})\mathbf{M}^{(s)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k}) = \tau\mathbf{I} + T^{-1}\sum_{\ell=1}^{p_0}\mathbf{M}_\ell^{(s)}\mathbf{M}_\ell^{(s)\mathsf{H}} \succ \mathbf{0}.$$

As $\mathbf{X}^{(s)}$ is positive definite, (3.38) can be put into the equivalent least squares form,

$$\min_{\mathbf{A}\in\mathbb{L}^1_{n\times r_0}} ||(\mathbf{\Phi}^{(s+1)})^{-\frac{1}{2}}\mathbf{A}\mathbf{L}_0^{(s)} - (\mathbf{\Phi}^{(s+1)})^{-\frac{1}{2}}\mathbf{Y}_0^{(s)}(\mathbf{L}_0^{(s)})^{-\mathsf{H}}||_F^2,$$

where $\mathbf{L}_0^{(s)}$ is the lower-triangular Cholesky factor of $\mathbf{X}_0^{(s)}$. As $\mathbf{A} \in \mathbb{L}^1_{n\times r_0}$, there is a $\mathbf{A}' \in \mathbb{L}^0_{n\times r_0}$ such that

$$\mathbf{A} = \begin{bmatrix}\mathbf{I}_{r_0}\\ \mathbf{0}\end{bmatrix} + \mathbf{A}'.$$

As $\mathbf{L}_0^{(s)}$ and $(\mathbf{\Phi}^{(s+1)})^{\frac{1}{2}}$ are lower-triangular, the quantity $(\mathbf{\Phi}^{(s+1)})^{-\frac{1}{2}}\mathbf{A}'\mathbf{L}_0^{(s)}$ is also in $\mathbb{L}^0_{n\times r_0}$. So, the optimal $\mathbf{A}'$ in $\mathbb{L}^0_{n\times r_0}$ can be found by orthogonal projection onto $\mathbb{L}^0_{n\times r_0}$ followed by left and right multiplication by $(\mathbf{\Phi}^{(s+1)})^{\frac{1}{2}}$ and $(\mathbf{L}_0^{(s)})^{-1}$ respectively. The unique minimizing $\mathbf{A} \in \mathbb{L}^1_{n\times r_0}$ is then

$$\mathbf{A}^{(s+1)} = \begin{bmatrix}\mathbf{I}_{r_0}\\ \mathbf{0}\end{bmatrix} + (\mathbf{\Phi}^{(s+1)})^{\frac{1}{2}}\mathbf{H}_0^{(s)}(\mathbf{L}_0^{(s)})^{-1}, \tag{3.40}$$

where $\mathbf{H}_0^{(s)}$ is the projection onto $\mathbb{L}^0_{n\times r_0}$ whose entries are

$$[\mathbf{H}_0^{(s)}]_{ij} = \Big[(\mathbf{\Phi}^{(s+1)})^{-\frac{1}{2}}\big(\mathbf{Y}_0^{(s)}(\mathbf{L}_0^{(s)})^{-1} - [\mathbf{L}_0^{(s)\mathsf{T}}\ \mathbf{0}^{\mathsf{T}}]^{\mathsf{T}}\big)\Big]_{ij},$$

for $1 \le j \le r_0$ and $j < i \le n$ and $[\mathbf{H}_0^{(s)}]_{ij} = 0$ for $i \ge j$.

The profile optimization problem for $\mathbf{B}_c$ is similar. To update $\mathbf{B}_c$ while holding $\mathbf{N}_c$ fixed at $\mathbf{N}_c^{(s)}$ and $\mathbf{\Phi}$ at $\mathbf{\Phi}^{(s+1)}$ results in the profile optimization problem

$$\min_{\mathbf{B}_c \in \mathbb{L}_{n_c \times r_c}^1} b_c(\mathbf{B}_c, \mathbf{N}_c^{(s)}, \mathbf{\Phi}_c^{(s+1)} | \boldsymbol{\theta}^{(s)}). \tag{3.41}$$

Define the relevant quantities,

$$\mathbf{X}_c^{(s)}(\omega) = \mathbf{N}_c^{(s)}(\mathrm{e}^{\mathrm{i}\omega}) \mathbf{G}_{*,cc}^{(s)}(\omega) \mathbf{N}_c^{(s)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega}), \qquad \mathbf{Y}_c^{(s)}(\omega) = \mathbf{K}_{cc}^{(s)}(\omega) \mathbf{N}_c^{(s)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega}),$$

with $\mathbf{X}_c^{(s)} = T^{-1} \sum_{k=0}^{T-1} \mathbf{X}_c^{(s)}(\omega_k)$ and $\mathbf{Y}_c^{(s)} = T^{-1} \sum_{k=0}^{T-1} \mathbf{Y}_c^{(s)}(\omega_k)$. A similar argument to that above yields that the unique solution to (3.41) is given by

$$\mathbf{B}_c^{(s+1)} = \begin{bmatrix} \mathbf{I}_{r_c} \\ \mathbf{0} \end{bmatrix} + (\mathbf{\Phi}_c^{(s+1)})^{\frac{1}{2}} \mathbf{H}_c^{(s)} (\mathbf{L}_c^{(s)})^{-1} \tag{3.42}$$

where $\mathbf{L}_c^{(s)}$ is the Cholesky factor of $\mathbf{X}_c^{(s)}$ and $\mathbf{H}_c^{(s)}$ has the entries

$$[\mathbf{H}_c^{(s)}]_{ij} = \left[ (\mathbf{\Phi}_c^{(s+1)})^{-\frac{1}{2}} \left( \mathbf{Y}_c^{(s)} (\mathbf{L}_c^{(s)})^{-1} + [\mathbf{L}_c^{(s)\mathsf{T}} \; \mathbf{0}^{\mathsf{T}}]^{\mathsf{T}} \right) \right]_{ij},$$

for $1 \leq j \leq r_c$ and $j < i \leq n_c$ and $[\mathbf{H}_c^{(s)}]_{ij} = 0$ for $i \geq j$.

### 3.4.3 Updating $\mathbf{C}$ and $\mathbf{D}_c$

The terms in $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ which depends on $\mathbf{C}$ are $c(\mathbf{C}, \mathbf{\Psi}|\boldsymbol{\theta}^{(s)})$ and $\gamma_*(\mathbf{C}|\boldsymbol{\theta}^{(s)})$. Obtaining a new value for $\mathbf{C}$ which reduces the surrogate function is complicated by the constraint that the magnitude of the zeros of the main diagonal polynomials of $\mathbf{C}(z)$ are at most $\kappa$. However, if the main-diagonal coefficients are held constant at their values in $\boldsymbol{\theta}^{(s)}$, the minimization problem to obtain the coefficients of the sub-diagonal polynomials is unconstrained. Dividing $\mathbf{C}(z)$ as

$$\mathbf{C}(z) = \mathbf{C}'(z) + \mathbf{C}''(z)$$

with $\mathbf{C}'(z)$ the strictly lower-triangular part and $\mathbf{C}''(z)$ the diagonal part,

$$\mathbf{C}'(z) \equiv \mathbf{C}'(\mathbf{u}_{q_0}(z) \otimes \mathbf{I}_{r_0}), \qquad \mathbf{C}''(z) \equiv \mathbf{I}_{r_0} + \mathbf{C}''(\mathbf{u}_{q_0}(z) \otimes \mathbf{I}_{r_0}),$$

where $\mathbf{C}'_j$ and $\mathbf{C}''_j$, $j = 1, \ldots, q_0$ are respectively the strictly lower-triangular part and diagonal part of the $j$th coefficient block $\mathbf{C}_j$, and $\mathbf{C}' \equiv [\mathbf{C}'^{\mathsf{T}}_1 \ \ldots \ \mathbf{C}'^{\mathsf{T}}_{q_0}]^{\mathsf{T}}$ while $\mathbf{C}'' \equiv [\mathbf{C}''^{\mathsf{T}}_1 \ \ldots \ \mathbf{C}''^{\mathsf{T}}_{q_0}]^{\mathsf{T}}$.

With this division, the optimization problem for $\mathbf{C}'$ is

$$\min_{\mathbf{C}' \in \mathbb{L}^{0,B}_{r_0 \times r_0 \times q_0}} c(\mathbf{C}', \mathbf{C}''^{(s)}, \mathbf{\Psi}^{(s+1)} \,|\boldsymbol{\theta}^{(s)}), \tag{3.43}$$

as $\gamma_*(\mathbf{C}|\boldsymbol{\theta}^{(s)})$ is constant in $\mathbf{C}'$. Setting $\mathbf{U}_0(\omega) = \mathbf{u}_{q_0}(\mathrm{e}^{\mathrm{i}\omega}) \otimes \mathbf{I}_{r_0}$, the objective in (3.43) equals

$$\begin{aligned} &T^{-1} \textstyle\sum_{k=0}^{T-1} \operatorname{tr} \mathbf{U}_0(\omega_k) \mathbf{G}^{(s)}_{00}(\omega_k) \mathbf{U}^{\mathsf{H}}_0(\omega_k) \mathbf{C}'^{\mathsf{H}} (\mathbf{\Psi}^{(s+1)})^{-1} \mathbf{C}' \\ &- 2T^{-1} \textstyle\sum_{k=0}^{T-1} \operatorname{Re} \operatorname{tr} \mathbf{U}_0(\omega_k) \mathbf{G}^{(s)}_{00}(\omega_k) (\mathbf{C}''^{(s)})^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k}) (\mathbf{\Psi}^{(s+1)})^{-1} \mathbf{C}' \end{aligned}$$

plus terms that are constant in $\mathbf{C}'$. Defining the needed frequency-average matrices

$$\mathbf{O}^{(s)}_0 = \tfrac{1}{T} \textstyle\sum_{k=0}^{T-1} \mathbf{U}(\omega_k) \mathbf{G}^{(s)}_{00}(\omega_k) \mathbf{U}^{\mathsf{H}}(\omega_k), \qquad \mathbf{Q}^{(s)}_0 = \tfrac{1}{T} \textstyle\sum_{k=0}^{T-1} \mathbf{C}''^{(s)}(\mathrm{e}^{\mathrm{i}\omega_k}) \mathbf{G}^{(s)}_{00}(\omega_k) \mathbf{U}^{\mathsf{H}}(\omega_k),$$

then as $\mathbf{G}(\omega_k)$ is positive definite and the set of Fourier basis vectors $\{\mathbf{u}_{p_0}(\omega_k)\}_{k=0}^{T-1}$ is linearly independent by $T > \max\{q_0, r_0\}$, the average $\mathbf{O}^{(s)}_0$ is positive definite. With these definitions, the objective in (3.43) is

$$\operatorname{tr}(\mathbf{\Psi}^{(s+1)})^{-1} \mathbf{C}' \mathbf{O}^{(s)}_0 \mathbf{C}'^{\mathsf{H}} + 2 \operatorname{Re} \operatorname{tr} (\mathbf{\Psi}^{(s+1)})^{-1} (\mathbf{C}'^{\mathsf{H}})^{\mathsf{H}} \mathbf{Q}^{(s)\mathsf{H}}_0 \tag{3.44}$$

plus terms constant in $\mathbf{C}'$. Let $\mathbf{c}'_j$ and $\mathbf{q}_{0,j}$ for $j = 2, \ldots, r_0$ be the row-vectors containing the $j$th row of $\mathbf{C}'$ and $\mathbf{Q}^{(s)}_0$. Then as

$$\operatorname{vec}\big(\mathbf{O}^{\frac{1}{2}}_0 \mathbf{C}'^{\mathsf{H}} \mathbf{\Psi}^{-\frac{1}{2}}\big) = \big(\mathbf{\Psi}^{-\frac{1}{2}} \otimes \mathbf{O}^{\frac{1}{2}}_0\big) \operatorname{vec}(\mathbf{C}'^{\mathsf{H}}),$$

where $\big(\mathbf{\Psi}^{-\frac{1}{2}} \otimes \mathbf{O}^{\frac{1}{2}}_0\big)$ is block-diagonal, some algebra shows that (3.44) has equivalent vector form,

$$\textstyle\sum_{j=2}^{r_0} \frac{1}{[\mathbf{\Psi}^{(s+1)}]_{jj}} \mathbf{c}'_j \mathbf{O}^{(s)}_0 \mathbf{c}'^{\mathsf{H}}_j - \frac{2}{[\mathbf{\Psi}^{(s+1)}]_{jj}} \operatorname{Re} \mathbf{c}'_j \mathbf{q}^{\mathsf{H}}_{0,j}. \tag{3.45}$$

The constraint that $\mathbf{C}'$ be in $\mathbb{L}^{0,B}_{r_0 \times r_0 \times q_0}$ yields that $\mathbf{c}'_j$ has $q_0 j(j+1)/2$ components constrained to be zero with remaining components obeying the norm constraint $||\mathbf{C}'||_{2,\infty} < B$. To use this fact,

define the matrices $\mathbf{Z}_{m,j,p}$ for $1 \leq j \leq m$ as

$$\mathbf{Z}_{m,j,p} = \text{blkdiag}\left(\begin{bmatrix}\mathbf{I}_j \\ \mathbf{0}_{m-j,j}\end{bmatrix}, \overset{p-2}{\ldots}, \begin{bmatrix}\mathbf{I}_j \\ \mathbf{0}_{m-j,j}.\end{bmatrix}\right) \tag{3.46}$$

Then $\mathbf{Z}_{m,j,p}$ is the $pm \times pj$ matrix which transforms a vector $\mathbf{v}$, patterned as $[\mathbf{v}_1^\mathsf{T} \ \ldots \ \mathbf{v}_p^\mathsf{T}]^\mathsf{T}$ with each $\mathbf{v}_k,\ k = 1, \ldots, p$ being length $j$, as

$$\mathbf{Z}_{m,j,p}\mathbf{v} = \begin{bmatrix}\mathbf{v}_1 \\ \mathbf{0}_{m-j} \\ \vdots \\ \mathbf{v}_p \\ \mathbf{0}_{m-j}\end{bmatrix}.$$

Reparameterizing the vectorized form (3.45) with $\mathbf{c}_j'^\mathsf{H} = \mathbf{Z}_{r_0,j-1,q_0}\mathbf{v}_j$ for some $\mathbf{v}_j \in \mathbb{C}_{r_0(j-1)}$, the problem (3.43) is equivalent to minimizing the sum of quadratic forms in distinct variables, where the $j$th quadratic form is minimized over $\mathbb{C}_{q_0(j-1)}$ without constraints. As $\mathbf{Z}_{r_0,j,p_0}$ is full column rank and $\mathbf{O}_0^{(s)}$ is positive definite, so too is $\mathbf{Z}_{r_0,j,q_0}^\mathsf{H}\mathbf{O}_0^{(s)}\mathbf{Z}_{r_0,j,q_0}$. Therefore, the new value of $\mathbf{C}'$ which uniquely minimizes (3.43) can be found by projecting $\mathbf{Q}_0^{(s)}$ onto the convex set $\mathbb{L}_{r_0\times r_0\times q_0}^{0,B}$ which is denoted by $\mathbf{Q}_0'^{(s)}$. For $B$ sufficiently large, the unconstrained projection would be such that the $j$th row $\mathbf{q}_{0,j}'$ of $\mathbf{Q}_0'^{(s)}$ for $j = 2, \ldots, r_0$ is

$$\mathbf{q}_{0,j}' = \mathbf{q}_{0,j}\mathbf{Z}_{r_0,j-1,p_0}(\mathbf{Z}_{r_0,j-1,q_0}^\mathsf{H}\mathbf{O}_0^{(s)}\mathbf{Z}_{r_0,j-1,q_0})^{-1}\mathbf{Z}_{r_0,j-1,q_0}^\mathsf{H}.$$

If $||\mathbf{q}_{0,j}'||_2 \leq B$, then the constraint on the maximum row norm of $\mathbf{Q}_0'^{(s)}$ does not bind $\mathbf{q}_{0,j}'$, which is given by the above expression. If $||\mathbf{q}_{0,j}'||_2 > B$ as computed above, then the revised row $\mathbf{q}_{0,j}'$ minimizing (3.45) and satisfying the norm constraint can be found as an orthogonal projection onto an ellipsoid, which does not have closed form in general but can be computed efficiently (Jia et al., 2017). In either case, the global minimizer of (3.43) is unique and the update is

$$\mathbf{C}'^{(s+1)} = \mathbf{Q}_0'^{(s)}. \tag{3.47}$$

After updating $\mathbf{C}'$, it remains to update $\mathbf{C}''$. Holding $\mathbf{C}'$ fixed at $\mathbf{C}'^{(s+1)}$, the profile optimization problem is

$$\min_{\mathbf{C}'' \in \mathbb{L}_{r_0 \times r_0 \times q_0}^{\leq \kappa, 0}} c(\mathbf{C}'^{(s+1)}, \mathbf{C}'', \mathbf{\Psi}^{(s+1)} | \boldsymbol{\theta}^{(s)}), \tag{3.48}$$

where the objective in (3.48) can be written as

$$\begin{aligned} &\gamma_*(\mathbf{C}|\boldsymbol{\theta}^{(s)}) + T^{-1} \sum_{j=1}^{r_0} \sum_{k=0}^{T-1} \frac{[\mathbf{G}_{00}^{(s)}(\omega_k)]_{jj}}{[\mathbf{\Psi}^{(s+1)}]_{jj}} \left| [\mathbf{C}''(\mathrm{e}^{\mathrm{i}\omega_k})]_{jj} \right|^2 \\ &- \frac{2}{T} \sum_{j=1}^{r_0} \sum_{k=0}^{T-1} \mathrm{Re}\, \frac{[\mathbf{G}_{00}^{(s)}(\omega_k) \mathbf{C}'^{(s+1)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega_k})]_{jj}}{[\mathbf{\Psi}^{(s+1)}]_{jj}} [\mathbf{C}''(\mathrm{e}^{\mathrm{i}\omega_k})]_{jj} \end{aligned} \tag{3.49}$$

plus terms constant in $\mathbf{C}''$ as $\mathbf{C}''(z)$ and $\mathbf{\Psi}^{(s+1)}$ are diagonal. To ensure the constraint that $\mathbf{C}$ belongs to $\mathbb{L}_{r_0 \times r_0 \times p_0}^{\leq \kappa, B}$ is satisfied, $[\mathbf{C}''(z)]_{jj}$ for $j = 1, \ldots, r_0$ is parameterized by its zeros $v_{jm}$ as

$$[\mathbf{C}''(z)]_{jj} = \textstyle\prod_{m=1}^{q_0} (1 - v_{jm} z^{-1}).$$

Note that if $[\mathbf{C}''(z)]_{jj}$ is stored and updated in factored form, it is not necessary to explicitly determine the zeros of a polynomial from its coefficients at any step.

The Wirtinger derivative of (3.48) with respect to $v_{jm}$ is

$$\begin{aligned} &\frac{\eta}{2} (\overline{v}_{jm} - \overline{v}_{jm}^{(s)}) + (v_{jm}^{(s)})^{T-1} \big(1 - (v_{jm}^{(s)})^T\big)^{-1} \\ &+ T^{-1} \textstyle\sum_{k=0}^{T-1} a_{jm}(\omega_k) (\overline{v}_{jm} - \mathrm{e}^{-\mathrm{i}\omega_k}) - b_{jm}(\omega_k) e^{-\mathrm{i}\omega_k} \end{aligned}$$

where $a_{jm}(\omega)$ and $b_{jm}(\omega)$ are

$$\begin{aligned} a_{jm}(\omega) &= \frac{[\mathbf{G}_{00}^{(s)}(\omega)]_{jj}}{[\mathbf{\Psi}^{(s+1)}]_{jj}} \prod_{m' \neq m} |1 - v_{jm'} \mathrm{e}^{-\mathrm{i}\omega}|^2, \\ b_{jm}(\omega) &= \frac{[\mathbf{G}_{00}^{(s)}(\omega) \mathbf{C}'^{(s+1)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})]_{jj}}{[\mathbf{\Psi}^{(s+1)}]_{jj}} \prod_{m' \neq m} (1 - v_{jm'} \mathrm{e}^{-\mathrm{i}\omega}). \end{aligned} \tag{3.50}$$

Note that $|v_{jm'}| < 1$ for all $m' \neq m$ implies $a_{jm}(\omega)$ is real and positive for all $\omega$, as $\mathbf{G}_{00}^{(s)}(\omega)$ is positive definite. As (3.48) is real, the unique stationary point $v_{jm}^* \in \mathbb{C}$ is

$$v_{jm}^* = \frac{\frac{\eta}{2} v_{jm}^{(s)} - \frac{(\overline{v}_{jm}^{(s)})^{T-1}}{1 - (\overline{v}_{jm}^{(s)})^T} + T^{-1} \sum_{k=0}^{T-1} \mathrm{e}^{\mathrm{i}\omega_k} \big(a_{jm}(\omega_k) + \overline{b_{jm}(\omega_k)}\big)}{\frac{\eta}{2} + T^{-1} \sum_{k=0}^{T-1} a_{jm}(\omega_k)}, \tag{3.51}$$

which can be shown to be the unconstrained global minimizer using $a_{jm}(\omega) > 0$ and $\eta > 0$. If $|\mathbf{v}^*_{jm}| > \kappa$, then the constrained minimum occurs on the boundary. If $\kappa \mathrm{e}^{\mathrm{i}\theta_{mj}}$ is substituted for $v_{jm}$ in (3.48), the associated optimization problem reduces to minimizing an objective of the form $-r' \cos(\theta_{mj} + \theta') + \text{const}$, which has unique global minimum at $\theta_{mj} = -\theta' = \operatorname{Arg} v^*_{jm}$.

If $v^{(s)}_{jm},\ m = 1, \ldots, q_0$ are the zeros of $[\mathbf{C}''^{(s)}(z)]_{jj}$ counted with multiplicity, the new zero locations can be computed iteratively for $m = 1$ through $m = q_0$ as

$$v^{(s+1)}_{jm} = \min\{1, \kappa/|v^{*(s)}_{jm}|\} \cdot v^{*(s)}_{jm}, \tag{3.52}$$

where $a^{(s)}_{jm}(\omega)$ and $b^{(s)}_{jm}(\omega)$ are computed as (3.50) using $v^{(s+1)}_{j1}, \ldots, v^{(s+1)}_{j(m-1)}, v^{(s)}_{j(m+1)}, \ldots, v^{(s)}_{jm}$, and $v^{*(s)}_{jm}$ is obtained as (3.51) using $a^{(s)}_{jm}(\omega)$ and $b^{(s)}_{jm}(\omega)$. This update step is performed for all $j = 1, \ldots, r_0$ to update $\mathbf{C}''$, and, when combined with the update step for $\mathbf{C}'$, yields an updated $\mathbf{C}^{(s+1)}$ within $\mathbb{L}^{\kappa,B}_{r_0 \times r_0 \times q_0}$ which decreases the surrogate function.

The update steps for $\mathbf{D}_c,\ c = 1, \ldots, C$ is analogous to the update step for $\mathbf{C}$, as $\mathbf{D}_c(z)$ can similarly be partitioned into $\mathbf{D}'_c(z) + \mathbf{D}''_c(z)$ with corresponding off-diagonal and diagonal coefficient blocks $\mathbf{D}'_c$ and $\mathbf{D}''_c$. The optimization problem for $\mathbf{D}'_c$ is

$$\min_{\mathbf{D}'_c \in \mathbb{L}^{0,B}_{r_c \times r_c \times q_c}} d_c(\mathbf{D}'_c, \mathbf{D}''^{(s)}_c, \boldsymbol{\Upsilon}^{(s+1)} | \boldsymbol{\theta}^{(s)}) \tag{3.53}$$

as $\delta_{c,*}(\mathbf{D}_c|\boldsymbol{\theta}^{(s)})$ is constant in $\mathbf{D}''_c$. The needed quantities to update $\mathbf{D}'_c$ are

$$\mathbf{O}^{(s)}_c = \tfrac{1}{T} \textstyle\sum_{k=0}^{T-1} \mathbf{U}_c(\omega_k) \mathbf{G}^{(s)}_{cc}(\omega_k) \mathbf{U}^{\mathsf{H}}_c(\omega_k),$$
$$\mathbf{Q}^{(s)}_c = \tfrac{1}{T} \textstyle\sum_{k=0}^{T-1} \mathbf{D}''^{(s)}_c(\mathrm{e}^{\mathrm{i}\omega_k}) \mathbf{G}^{(s)}_{cc}(\omega_k) \mathbf{U}^{\mathsf{H}}_c(\omega_k),$$

with $\mathbf{U}_c(\omega) = \mathbf{u}_{q_c}(\mathrm{e}^{\mathrm{i}\omega}) \otimes \mathbf{I}_{r_c}$. If $\mathbf{Q}'^{(s)}_c \in \mathbb{L}^{0,B}_{r_c \times r_c \times q_c}$ is the projection of $\mathbf{Q}^{(s)}_c$ onto $\mathbb{L}^{0,B}_{r_c \times r_c \times q_c}$ which has $j$th row

$$\mathbf{q}'_{c,j} = \mathbf{q}_{c,j} \mathbf{Z}_{r_c, j-1, q_c} (\mathbf{Z}^{\mathsf{H}}_{r_c, j-1, q_c} \mathbf{O}^{(s)}_c \mathbf{Z}_{r_c, j-1, q_c})^{-1} \mathbf{Z}^{\mathsf{H}}_{r_c, j-1, q_c},$$

where $\mathbf{q}_{c,j}$ is the $j$th row of $\mathbf{Q}^{(s)}_c$ for $j = 2, \ldots, r_c$, then the profile objective minimum for $\mathbf{D}'_c$ is

$$\mathbf{D}'^{(s+1)}_c = \mathbf{Q}'^{(s)}_c. \tag{3.54}$$

With the obtained values for $\mathbf{D}_c'^{(s+1)}$, the profile optimization problem $\mathbf{D}_c''$ is

$$\min_{D_c'' \in \mathbb{L}_{r_c \times r_c \times q_c}^{\leq \kappa, 0}} d_c(\mathbf{D}_c'^{(s+1)}, \mathbf{D}_c'', \mathbf{\Upsilon}_c^{(s+1)} | \boldsymbol{\theta}^{(s+1)}) + \delta_{c,*}(\mathbf{D}_c'^{(s+1)}, \mathbf{D}_c'' | \boldsymbol{\theta}^{(s+1)}), \tag{3.55}$$

which can be updated in the same fashion as $\mathbf{C}''$. To update the on-diagonal polynomials $\mathbf{D}_c''(z)$ with zeros $w_{c,jm}$ for $j = 1, \ldots, r_c$ and $m = 1, \ldots, q_c$ requires

$$f_{c,jm}(\omega) = \frac{[\mathbf{G}_{cc}^{(s)}(\omega)]_{jj}}{[\mathbf{\Upsilon}_c^{(s+1)}]_{jj}} \prod_{m' \neq m} |1 - w_{c,jm'} \mathrm{e}^{-\mathrm{i}\omega}|^2,$$

$$h_{c,jm}(\omega) = \frac{[\mathbf{G}_{cc}^{(s)}(\omega) \mathbf{D}_c'^{(s+1)\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})]_{jj}}{[\mathbf{\Upsilon}_c^{(s+1)}]_{jj}} \prod_{m' \neq m} (1 - w_{c,jm'} \mathrm{e}^{-\mathrm{i}\omega}),$$

which yield the unconstrained minimum

$$w_{c,jm}^* = \frac{\frac{\eta}{2} w_{c,jm}^{(s)} - \frac{(\overline{w}_{c,jm}^{(s)})^{T-1}}{1 - (\overline{w}_{c,jm}^{(s)})^T} + \sum_{k=0}^{T-1} \mathrm{e}^{\mathrm{i}\omega_k} \left( f_{c,jm}(\omega_k) + \overline{h_{c,jm}(\omega_k)} \right)}{\frac{\eta}{2} + \sum_{k=0}^{T-1} f_{c,jm}(\omega_k)}. \tag{3.56}$$

Iteratively computing $f_{c,jm}^{(s)}(\omega)$, $h_{c,jm}^{(s)}(\omega)$ and $w_{jm}^{*(s)}$ using the previously computed zeros in the same fashion as $a_{jm}^{(s)}(\omega)$, $b_{jm}^{(s)}(\omega)$ and $v_{jm}^{*(s)}$ respectively yields the update,

$$w_{c,jm}^{(s+1)} = \min\{1, \kappa / |w_{c,jm}^{*(s)}|\} \cdot w_{c,jm}^{*(s)}.$$

### 3.4.4 Updating $\mathrm{M}$ and $\mathrm{N}_c$

To update $\mathrm{M}$ holding $\mathbf{A}$ and $\mathbf{\Phi}$ fixed at the new values $\mathbf{A}^{(s+1)}$ and $\mathbf{\Phi}^{(s+1)}$, the relevant term in the concentrated surrogate function is $a(\mathbf{A}^{(s+1)}, \mathbf{M}, \mathbf{\Phi}^{(s+1)} | \boldsymbol{\theta}^{(s)})$. As for the update of $\mathbf{C}$, the constraint that the zeros of the south-east polynomial $[\mathbf{M}(z)]_{r_0 r_0}$ have magnitude at most $\kappa$ adds an additional complication to the minimization problem. Updating the values for the other coefficients in $\mathbf{M}$ prior to updating the zeros in $[\mathbf{M}(z)]_{r_0 r_0}$ will ensure that this constraint is satisfied.

To separate the two types of coefficients in $\mathbf{M}$, first take an $LDL^{\mathsf{H}}$ decomposition as

$$\widehat{\mathbf{S}}_0^{(s+1)} \widehat{\mathbf{\Gamma}}_0^{(s+1)} \widehat{\mathbf{S}}_0^{(s+1)\mathsf{H}} = \mathbf{\Pi} \mathbf{A}^{(s+1)\mathsf{H}} (\mathbf{\Phi}^{(s+1)})^{-1} \mathbf{A}^{(s+1)} \mathbf{\Pi},$$

where $\mathbf{\Pi}$ is the anti-diagonal permutation matrix (that is, the order-reversing permutation). Let the reversed $\mathbf{S}_0^{(s+1)}$ and $\mathbf{\Gamma}_0^{(s+1)}$ be

$$\mathbf{S}_0^{(s+1)} = \mathbf{\Pi}\widehat{\mathbf{S}}_0^{(s+1)\mathsf{H}}\mathbf{\Pi}, \qquad\qquad \mathbf{\Gamma}_0^{(s+1)} = \mathbf{\Pi}\widehat{\mathbf{\Gamma}}_0{}^{(s+1)}\mathbf{\Pi}.$$

Then $\mathbf{S}_0^{(s+1)} \in \mathbb{L}^1_{r_0\times r_0}$ is unit lower-triangular and $\mathbf{\Gamma}^{(s+1)} \in \mathbb{P}_{r_0}$ is diagonal with real positive diagonal components. If we further set $\widehat{\mathbf{M}} = \mathbf{S}_0^{(s+1)}\mathbf{M}$, then $\widehat{\mathbf{M}}$ is in $\mathbb{L}_{r_0\times r_0\times p_0}$ where the diagonal entries of the coefficient blocks $\widehat{\mathbf{M}}_j$ equal those of $\mathbf{M}_j$ (in particular, $[\widehat{\mathbf{M}}_j]_{r_0r_0} = [\mathbf{M}_j]_{r_0r_0}$). With this transformation, the quantity

$$\mathbf{M}^{\mathsf{H}}(z)\mathbf{A}^{(s+1)\mathsf{H}}(\mathbf{\Phi}^{(s+1)})^{-1}\mathbf{A}^{(s+1)}\mathbf{M}(z)$$

equals, for all $\mathbf{M} \in \mathbb{L}_{r_0\times r_0\times p_0}$,

$$\Big(\mathbf{S}_0^{(s+1)} + \widehat{\mathbf{M}}\big(\mathbf{u}_{p_0}(z)\otimes\mathbf{I}_{r_0}\big)\Big)^{\mathsf{H}}\mathbf{\Gamma}_0^{(s+1)}\Big(\mathbf{S}_0^{(s+1)} + \widehat{\mathbf{M}}\big(\mathbf{u}_{p_0}(z)\otimes\mathbf{I}_{r_0}\big)\Big).$$

Next, for each coefficient block $\widehat{\mathbf{M}}_j,\ j = 1,\ldots,p_0$ in $\mathbf{M}$, let $\widehat{\mathbf{M}}'_j \in \mathbb{L}_{r_0\times r_0}$ be the matrix with south-east entry $[\widehat{\mathbf{M}}'_j]_{r_0r_0} = 1$ and all other entries equal to those of $\widehat{\mathbf{M}}_j$. If $\boldsymbol{\xi}_0 \in \mathbb{C}_{p_0r_0\times p_0r_0}$ is the diagonal matrix

$$\boldsymbol{\xi}_0 = \mathrm{blkdiag}(\mathbf{I}_{r_0-1}, [\mathbf{M}_1]_{r_0r_0}, \ldots, \mathbf{I}_{r_0-1}, [\mathbf{M}_{p_0}]_{r_0r_0})$$

for $j = 1,\ldots,p_0$, then $\widehat{\mathbf{M}} = \widehat{\mathbf{M}}'\boldsymbol{\xi}_0$ where $\widehat{\mathbf{M}}' = [\widehat{\mathbf{M}}'_1\ \ldots\ \widehat{\mathbf{M}}'_{p_0}]$. If $\boldsymbol{\xi}_0^{(s)}$ is constructed as previously described using $\mathbf{M}^{(s)}$, then the optimization problem for $\widehat{\mathbf{M}}'$ is

$$\min_{\widehat{\mathbf{M}}'\in\mathbb{L}^{SE=1}_{r_0\times r_0\times p_0}} a(\mathbf{A}^{(s+1)}, \widehat{\mathbf{M}}', \boldsymbol{\xi}_0^{(s)}, \mathbf{\Phi}^{(s+1)}|\boldsymbol{\theta}^{(s)}). \tag{3.57}$$

Defining the quantities,

$$\begin{aligned}
\mathbf{P}_0^{(s)}(\omega) &\equiv \boldsymbol{\xi}_0^{(s)}\mathbf{U}_0(\omega)\mathbf{G}^{(s)}_{*,00}(\omega)\mathbf{U}_0^{\mathsf{H}}(\omega)\boldsymbol{\xi}_0^{(s)\mathsf{H}},\\
\mathbf{\Delta}_0^{(s)}(\omega) &\equiv \mathbf{K}_0^{(s)\mathsf{H}}(\omega)(\mathbf{\Phi}^{(s+1)})^{-1}\mathbf{A}^{(s+1)}(\mathbf{S}_0^{(s+1)})^{-1},\\
\mathbf{L}_0^{(s)}(\omega) &\equiv \big[\mathbf{\Delta}_0^{(s)\mathsf{H}}(\omega) - \mathbf{\Gamma}_0^{(s+1)}\mathbf{S}_0^{(s+1)}\mathbf{G}^{(s)}_{*,00}(\omega)\big]\mathbf{U}_0^{\mathsf{H}}(\omega)\boldsymbol{\xi}_0^{(s)\mathsf{H}}
\end{aligned}$$

with $\mathbf{P}_0^{(s)} = T^{-1}\sum_{k=0}^{T-1}\mathbf{P}_0^{(s)}(\omega_k)$ and $\mathbf{L}_0^{(s)} = T^{-1}\sum_{k=0}^{T-1}\mathbf{L}_0^{(s)}(\omega_k)$. Note that $\mathbf{P}_0^{(s)}$ is positive definite for the same reason as $\mathbf{O}_0^{(s)}$ when $T > \max\{p_0, r_0\}$ as long as $[\mathbf{M}_j^{(s)}]_{r_0 r_0} \neq 0$ for all $j = 1, \ldots, p_0$ which occurs almost surely when the distribution of $\mathbf{x}$ is absolutely continuous. Then the objective in (3.57) is equivalent to

$$\operatorname{tr}\boldsymbol{\Gamma}^{(s+1)}\widehat{\mathbf{M}}'\mathbf{P}_0^{(s)}\widehat{\mathbf{M}}'^{\mathsf{H}} - 2\operatorname{Re}\operatorname{tr}(\widehat{\mathbf{M}}'^{\mathsf{H}})\mathbf{L}_0^{(s)\mathsf{H}} \tag{3.58}$$

plus terms which are constant in $\widehat{\mathbf{M}}'$. A similar argument to that used in Section 3.4.3 yields that (3.58) has equivalent vectorized form

$$\sum_{j=1}^{r_0}[\boldsymbol{\Gamma}^{(s+1)}]_{jj}\widehat{\mathbf{m}}_j'\mathbf{P}_0^{(s)}\widehat{\mathbf{m}}_j'^{\mathsf{H}} - 2\operatorname{Re}\widehat{\mathbf{m}}_j'\boldsymbol{\ell}_{0,j}^{\mathsf{H}}$$

where $\widehat{\mathbf{m}}_j'$ and $\boldsymbol{\ell}_{0,j}$ are the row-vectors containing the $j$th row of $\widehat{\mathbf{M}}'$ and $\mathbf{L}_0^{(s)}$ respectively. For $j = 1, \ldots, r_0 - 1$, let $\widehat{\mathbf{L}}_0^{(s)}$ be the orthogonal projection of $\mathbf{L}_0^{(s)}$ onto the affine space $\mathbb{L}_{r_0\times r_0\times p_0}^{SE=1}$, which has $j$th row $\widehat{\boldsymbol{\ell}}_{0,j}$ for $j = 1, \ldots, r_0 - 1$ given by

$$\widehat{\boldsymbol{\ell}}_{0,j} = [\boldsymbol{\Gamma}_0^{(s+1)}]_{jj}^{-1}\boldsymbol{\ell}_{0,j}\mathbf{Z}_{r_0,j,p_0}(\mathbf{Z}_{r_0,j,p_0}^{\mathsf{H}}\mathbf{P}_0^{(s)}\mathbf{Z}_{r_0,j,p_0})^{-1}\mathbf{Z}_{r_0,j,p_0}^{\mathsf{H}},$$

and for $j = r_0$, the displacement of the affine space $\mathbb{L}_{r_0\times r_0\times p_0}^{SE=1}$ yields the projected row,

$$\widehat{\boldsymbol{\ell}}_{0,r_0} = \mathbf{w}_0 + ([\boldsymbol{\Gamma}_0^{(s+1)}]_{r_0 r_0}^{-1}\boldsymbol{\ell}_{0,r_0} - \mathbf{w}_0\mathbf{P}_0^{(s)})\mathbf{Z}_{r_0,r_0-1,p_0}(\mathbf{Z}_{r_0,j,p_0}^{\mathsf{H}}\mathbf{P}_0^{(s)}\mathbf{Z}_{r_0,j,p_0})^{-1}\mathbf{Z}_{r_0,r_0-1,p_0}^{\mathsf{H}},$$

where $\mathbf{w}_0^{\mathsf{T}} \in \mathbb{C}_{p_0 r_0}$ has entries numbered $k \cdot r_0$ for $k = 1, \ldots, p_0$ equaling 1 and all others zero. Then $\widehat{\mathbf{M}}' = \widehat{\mathbf{L}}_0^{(s)}$ is the unique minimizer of (3.57), so the update $\mathbf{M}'^{(s+1)}$ then is

$$\mathbf{M}'^{(s+1)} = (\mathbf{S}_0^{(s+1)})^{-1}\widehat{\mathbf{L}}_0^{(s)}. \tag{3.59}$$

To update zeros in the south-east polynomial, suppose $\mathbf{S}_0^{(s+1)} + \widehat{\mathbf{M}}(\mathbf{u}_{p_0}(z) \otimes \mathbf{I}_{r_0})$ is patterned as

$$\mathbf{S}_0^{(s+1)} + \widehat{\mathbf{M}}(\mathbf{u}_{p_0}(z) \otimes \mathbf{I}_{r_0}) \equiv \begin{bmatrix} \widehat{\mathbf{M}}_{11}(z) & \mathbf{0} \\ \widehat{\mathbf{m}}_{21}(z) & \widehat{m}_{22}(z) \end{bmatrix}. \tag{3.60}$$

Then $\widehat{\mathbf{M}}_{11}(z)$ and $\widehat{\mathbf{m}}_{21}(z)$ are functions of $\widehat{\mathbf{M}}'$ alone while $m_{22}(z) = [\mathbf{M}(z)]_{r_0 r_0}$. If $\mathbf{M}'$ is held constant at $\mathbf{M}'^{(s+1)}$, the profile objective for $[\mathbf{M}(z)]_{r_0 r_0}$ is

$$\begin{aligned} &\tfrac{1}{T}\textstyle\sum_{k=0}^{T-1}[\mathbf{\Gamma}_0^{(s+1)}\mathbf{G}_{*,00}^{(s)}(\omega_k)]_{r_0 r_0}\,|[\mathbf{M}(z)]_{r_0 r_0}|^2 \\ &+ \tfrac{2}{T}\textstyle\sum_{k=0}^{T-1}\mathrm{Re}\left(\mathbf{g}_0^{(s)}(\omega)\widehat{\mathbf{m}}_{21}^{(s+1)\mathsf{H}}(\omega) - [\mathbf{\Delta}_0(\omega_k)]_{r_0 r_0}\right)[\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega_k})]_{r_0 r_0}, \end{aligned} \tag{3.61}$$

plus terms which are constant in $[\mathbf{M}(z)]_{r_0 r_0}$. Here, $\widehat{\mathbf{m}}_{21}^{(s+1)}$ is the row-vector of length $r_0 - 1$ from (3.60) using $\widehat{\mathbf{M}}'^{(s+1)}$, while $\mathbf{g}_0^{(s)}(\omega)$ is the first $r_0 - 1$ entries in the bottom row of $\mathbf{G}_{*,00}^{(s)}(\omega)$. As in Section 3.4.3, the polynomial $[\mathbf{M}(z)]_{r_0 r_0}$ can be parameterized in terms of its zeros $x_j,\ j = 1, \ldots, p_0$ as

$$[\mathbf{M}(z)]_{r_0 r_0} = \textstyle\prod_{j=1}^{p_0}(1 - x_j z^{-1}).$$

Defining the quantities

$$\begin{aligned} p_j(\omega) &= [\mathbf{\Gamma}_0^{(s+1)}\mathbf{G}_{*,00}^{(s)}(\omega)]_{r_0 r_0}\textstyle\prod_{j' \neq j}^{p_0}(1 - x_{j'}\mathrm{e}^{-\mathrm{i}\omega}), \\ q_j(\omega) &= \left([\mathbf{\Delta}_0^{(s)}(\omega)]_{r_0 r_0} - \left(\mathbf{g}_0^{(s)}(\omega)\widehat{\mathbf{m}}_{21}^{(s+1)\mathsf{H}}(\omega)\right)\right)\textstyle\prod_{j' \neq j}^{p_0}(1 - x_{j'}\mathrm{e}^{-\mathrm{i}\omega}), \end{aligned}$$

then as in Section 3.4.3, the minimizing $x_i^{(s+1)}$ can be found as

$$\begin{aligned} x_j^{*(s)} &= \frac{\sum_{k=0}^{T-1}\mathrm{e}^{\mathrm{i}\omega_k}\left(p_j^{(s)}(\omega_k) + \overline{q_j^{(s)}(\omega_k)}\right)}{\sum_{k=0}^{T-1}p_j^{(s)}(\omega_k)}, \\ x_j^{(s+1)} &= \min\{1, \kappa/|x_j^{*(s)}|\} \cdot x_j^{*(s)}, \end{aligned} \tag{3.62}$$

where $p_j^{(s)}(\omega)$ and $q_j^{(s)}(\omega)$ are computed using $x_1^{(s+1)}$ through $x_{j-1}^{(s+1)}$ and $x_{j+1}^{(s)}$ through $x_{p_0}^{(s)}$.

The updates for $\mathbf{N}_c,\ c = 1, \ldots, C$ are analogous, as the relevant term in the surrogate function is $b_c(\mathbf{B}_c^{(s+1)}, \mathbf{N}_c, \mathbf{\Upsilon}_c^{(s+1)}|\boldsymbol{\theta}^{(s)})$. To update the entries in $\mathbf{N}_c$ prior to updating the location of the zeros of $[\mathbf{N}_c(z)]_{r_c r_c}$, let $\widehat{\mathbf{S}}_c^{(s+1)}\widehat{\mathbf{\Gamma}}_c^{(s+1)}\widehat{\mathbf{S}}_c^{(s+1)\mathsf{H}}$ be the $LDL^{\mathsf{H}}$ decomposition of $\mathbf{\Pi}\mathbf{B}_c^{(s+1)\mathsf{H}}(\mathbf{\Phi}_c)^{-1}\mathbf{B}_c^{(s+1)}\mathbf{\Pi}$ where here $\mathbf{\Pi}$ is the order-reversing permutation matrix of the appropriate size. The reversed $\mathbf{S}_c^{(s+1)}$ and $\mathbf{\Gamma}_c^{(s+1)}$ are obtained in the same fashion as $\mathbf{S}_0^{(s+1)}$ and $\mathbf{\Gamma}_0^{(s+1)}$. Let $\widehat{\mathbf{N}}_c = \mathbf{S}_c^{(s+1)}\mathbf{N}_c$ be the transformed coefficient matrix, and let $\widehat{\mathbf{N}}'_c = [\widehat{\mathbf{N}}'_{c,1}\ \ldots\ \widehat{\mathbf{N}}'_{c,p_c}]$ be the matrix in $\mathbb{L}_{r_c \times r_c \times p_c}^{SE=1}$ where $[\widehat{\mathbf{N}}'_{c,j}]_{r_c r_c} = 1$ and all other entries in $\widehat{\mathbf{N}}'_{c,j}$ are the same as in the corresponding coefficient matrix

$\widehat{\mathbf{N}}_{c,j}$ for $j = 1, \ldots, p_c$. Taking

$$\boldsymbol{\xi}_c^{(s)} = \text{blkdiag}(\mathbf{I}_{r_c-1}, [\mathbf{N}_{c,1}^{(s)}]_{r_c r_c}, \ldots, \mathbf{I}_{r_c-1}, [\mathbf{N}_{c,p_c}^{(s)}]_{r_c r_c}),$$

we have $\widehat{\mathbf{N}}_c = \widehat{\mathbf{N}}_c' \boldsymbol{\xi}_c$, and the profile optimization problem for $\widehat{\mathbf{N}}_c'$ is

$$\min_{\widehat{\mathbf{N}}_c' \in \mathbb{L}_{r_c \times r_c \times p_c}^{SE=1}} b_c(\mathbf{B}_c^{(s+1)}, \widehat{\mathbf{N}}_c', \boldsymbol{\xi}_c^{(s)}, \boldsymbol{\Phi}_c^{(s+1)} | \boldsymbol{\theta}^{(s)}). \tag{3.63}$$

To find the update for $\mathbf{N}_c'$, the needed quantities are

$$\begin{aligned}
\mathbf{P}_c^{(s)}(\omega) &\equiv \boldsymbol{\xi}_c^{(s)} \mathbf{U}_c(\omega) \mathbf{G}_{*,cc}^{(s)}(\omega) \mathbf{U}_c^{\mathsf{H}}(\omega) \boldsymbol{\xi}_c^{(s)\mathsf{H}}, \\
\boldsymbol{\Delta}_c^{(s)}(\omega) &\equiv \mathbf{K}_c^{(s)\mathsf{H}}(\omega) (\boldsymbol{\Phi}_c^{(s+1)})^{-1} \mathbf{B}_c^{(s+1)} (\mathbf{S}_c^{(s+1)})^{-1}, \\
\mathbf{L}_c^{(s)}(\omega) &\equiv \left[ \boldsymbol{\Delta}_c^{(s)\mathsf{H}}(\omega) - \boldsymbol{\Gamma}_c^{(s+1)} \mathbf{S}_c^{(s+1)} \mathbf{G}_{*,cc}^{(s)}(\omega) \right] \mathbf{U}_c^{\mathsf{H}}(\omega) \boldsymbol{\xi}_c^{(s)\mathsf{H}},
\end{aligned}$$

and set $\mathbf{P}_c^{(s)} = T^{-1} \sum_{k=0}^{T-1} \mathbf{P}_c^{(s)}(\omega_k)$ and $\mathbf{L}_c^{(s)} = T^{-1} \sum_{k=0}^{T-1} \mathbf{L}_c^{(s)}(\omega_k)$. Let $\widehat{\mathbf{L}}_c^{(s)}$ be the projection of $\mathbf{L}_c^{(s)}$ whose $j$th row $\widehat{\boldsymbol{\ell}}_{c,j}$ for $j = 1, \ldots, r_c - 1$ is given by

$$\widehat{\boldsymbol{\ell}}_{c,j} = [\boldsymbol{\Gamma}_c^{(s+1)}]_{jj}^{-1} \boldsymbol{\ell}_{c,j} \mathbf{Z}_{r_c,j,p_c} (\mathbf{Z}_{r_c,j,p_c}^{\mathsf{H}} \mathbf{P}_c^{(s)} \mathbf{Z}_{r_c,j,p_c})^{-1} \mathbf{Z}_{r_c,j,p_c}^{\mathsf{H}},$$

and the $r_c$th row is

$$\widehat{\boldsymbol{\ell}}_{c,r_c} = \mathbf{w}_c + ([\boldsymbol{\Gamma}_c^{(s+1)}]_{r_c r_c}^{-1} \boldsymbol{\ell}_{c,r_c} - \mathbf{w}_c \mathbf{P}_c^{(s)}) \mathbf{Z}_{r_c,r_c-1,p_c} (\mathbf{Z}_{r_c,j,p_c}^{\mathsf{H}} \mathbf{P}_c^{(s)} \mathbf{Z}_{r_c,j,p_c})^{-1} \mathbf{Z}_{r_c,r_c-1,p_c}^{\mathsf{H}}.$$

where $\mathbf{w}_c^{\mathsf{T}} \in \mathbb{C}_{p_c r_c}$ has every $r_c$th entry being 1 and all other entries zero. The update $\mathbf{N}_c^{\prime(s+1)}$ is obtained as

$$\mathbf{N}_c^{\prime(s+1)} = (\mathbf{S}_c^{(s+1)})^{-1} \widehat{\mathbf{L}}_c^{(s)}. \tag{3.64}$$

For the south-east coefficients, if the zeros of $[\mathbf{N}_c(z)]_{r_c r_c}$ are denoted as $y_{c,j}$ for $j = 1, \ldots, p_c$ so that

$$[\mathbf{N}_c(z)]_{r_c r_c} = \textstyle\prod_{j=1}^{p_0} (1 - y_{c,j} z^{-1}),$$

then the needed quantities to update $y_{c,j}$ are

$$t_{c,j}(\omega) = [\mathbf{\Gamma}_c^{(s+1)} \mathbf{G}_{*,cc}^{(s)}(\omega)]_{r_c r_c} \prod_{j' \neq j}^{p_c} (1 - y_{c,j'} \mathrm{e}^{-\mathrm{i}\omega})$$
$$u_{c,j}(\omega) = \left( [\mathbf{\Delta}_c^{(s)}(\omega)]_{r_c r_c} - \left( \mathbf{g}_c^{(s)}(\omega) \widehat{\mathbf{n}}_{c,21}^{(s+1)\mathsf{H}}(\omega) \right) \right) \prod_{j' \neq j}^{p_c} (1 - y_{c,j'} \mathrm{e}^{-\mathrm{i}\omega}),$$

where $\mathbf{g}_c^{(s)}(\omega)$ is the first $r_c - 1$ entries in the bottom row of $\mathbf{G}_{*,cc}^{(s)}(\omega)$ and $\widehat{\mathbf{n}}_{c,21}^{(s+1)\mathsf{H}}(\omega)$ is the first $r_c - 1$ entries of the bottom row of $\mathbf{S}_c^{(s)+1)} + \widehat{\mathbf{N}}_c^{(s+1)} (\mathbf{u}_{q_c} \otimes \mathbf{I}_{r_c})$. As above, the minimizing $\mathbf{y}_{c,j}^{(s+1)}$ can be found as

$$y_{c,j}^{*(s)} = \frac{\sum_{k=0}^{T-1} \mathrm{e}^{\mathrm{i}\omega_k} \left( t_{c,j}^{(s)}(\omega_k) + \overline{u_{c,j}^{(s)}(\omega_k)} \right)}{\sum_{k=0}^{T-1} t_{c,j}^{(s)}(\omega_k)},$$
$$y_{c,j}^{(s+1)} = \min\{1, \kappa / |y_{c,j}^{*(s)}|\} \cdot y_{c,j}^{*(s)}, \tag{3.65}$$

where $t_{c,j}^{(s)}(\omega)$ and $u_{c,j}^{(s)}(\omega)$ are computed using $y_{c,1}^{(s+1)}, \ldots, y_{c,j-1}^{(s+1)}$ and $y_{c,j+1}^{(s)}, \ldots, y_{c,p_c}^{(s)}$.

## 3.5 Experiments

In this section, a simulation study is conducted in order to examine how modeling both the cross-sectional and temporal correlations of the observations (as in MFSA) can lead to improved performance relative to MFA, which models only the cross-sectional correlations. For the three problems of cross-sectional covariance estimation, transmitted signal recovery, and target detection in passive radar, the following experiments demonstrate that MFSA has improved performance characteristics when the observations have significant temporal correlations.

In the following experiments, the relative contribution of the signal, interference, and noise series to the channel-$c$ observation are quantified by the channel-$c$ average power ratios, defined as

$$\eta_{\mathbf{s}_c} = \frac{1}{2\pi P_c} \int_{-\pi}^{\pi} \operatorname{tr} \mathbf{S}_{\mathbf{s}_c \mathbf{s}_c}(\omega) d\omega, \qquad \eta_{\mathbf{i}_c} = \frac{1}{2\pi P_c} \int_{-\pi}^{\pi} \operatorname{tr} \mathbf{S}_{\mathbf{i}_c \mathbf{i}_c}(\omega) d\omega, \qquad \eta_{\mathbf{u}_c} = \frac{\operatorname{tr} \mathbf{\Phi}_c}{P_c}, \tag{3.66}$$

where $P_c$ is the average power in channel $c$,

$$P_c = \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{tr} \mathbf{S}_{\mathbf{x}_c \mathbf{x}_c}(\omega) d\omega,$$

and $\mathbf{S}_{\mathbf{s}_c}(\omega), \mathbf{S}_{\mathbf{i}_c \mathbf{i}_c}(\omega)$ and $\mathbf{\Phi}_c$ are the power spectral densities of $\mathbf{s}_c, \mathbf{i}_c$ and $\mathbf{u}_c$ respectively.

### 3.5.1 Cross-Sectional Correlation Estimation

In multi-channel problems, the cross-sectional correlations within and between channels are often of primary interest. To understand how the presence of non-trivial temporal correlations in the latent series affect estimation of zero-lag cross-sectional correlation matrix $\mathbf{R}_{\mathbf{xx}} \equiv E\big[\mathbf{x}[0]\mathbf{x}^{\mathsf{H}}[0]\big]$, we simulate $\mathbf{x}[t]$ using temporally correlated latent factor series. MFA, which exploits the low-dimensional channel-structured observation model (2.2) but the assumes that the observations are temporally *independent*, still allows for consistent estimation of $\mathbf{R}_{\mathbf{xx}}$ even when the observations are temporally dependent, under mild moment conditions. Therefore, the estimates $\widehat{\mathbf{R}}_{\mathrm{MFA}}$ obtained using the method of Ramírez et al. (2020) can be compared to the estimated $\widehat{\mathbf{R}}_{\mathrm{MFSA}} \equiv (2\pi)^{-1}\int_{-\pi}^{\pi}\mathbf{S}_{\mathbf{xx}}(\omega;\widehat{\boldsymbol{\theta}})d\omega$ obtained as a result of Algorithm 1. In addition, we compare to cross-sectional covariance estimates which do not follow the low-dimensional channel-structured observation model, namely $\widehat{\mathbf{R}}_{\mathrm{ARMA}}$ and $\widehat{\mathbf{R}}_{\mathrm{Samp}}$, which are respectively obtained by fitting a multivariate ARMA model directly to the all-channel observation series $\mathbf{x}[t]$ and by using the sample covariance matrix $\mathbf{R}_{\mathrm{Samp}} = T^{-1}\sum_{t=0}^{T-1}\mathbf{x}[t]\mathbf{x}^{\mathsf{H}}[t]$ as a nonparametric estimator.

To make this comparison, we perform an experiment analogous to that of Experiment 3 in Ramírez et al. (2020, Sec. IV.c), with $C = 3$ channels of sizes $n_1 = n_2 = n_3 = 7$ influenced by $r_0 = 2$ common factors and $r_c = 1$ distinct factors for each channel, with the added feature that the common and distinct factors exhibit significant temporal correlation. The average power ratios are the same across the three channels,

$$\eta_{\mathbf{s},c} = 0.1, \qquad \eta_{\mathbf{i},c} = 0.5, \qquad \eta_{\mathbf{u},c} = 0.4, \qquad c = 1,2,3.$$

In the temporally correlated setting, the latent factors are independent realizations of multivariate Gaussian $AR(3)$ processes, where the diagonal poles of greatest magnitude , $v_{j1}, j = 1,\ldots,r_0$ and $w_{c,j1}, c = 1,2,3, j = 1,\ldots,r_c$, are set to $0.95$ and the remaining poles are uniformly distributed with $|v_{j,m}| < 0.95$ for $j = 1,\ldots,r_0, m = 2,3$ and similarly $|w_{c,jm}| < 0.95$ for $c = 1,2,3, j = 1,\ldots,r_c, m = 2,3$. In the temporally uncorrelated setting, the latent factors are independent standard normal random variables. In both settings, the remaining unconstrained parameters are

sampled as independent standard normal variables, and positive-constrained parameters are sampled as the absolute value of independent standard normal random variables. The channel-$c$ loading parameters $\mathbf{A}_c$ and $\mathbf{B}_c$ are then scaled so as to obtain the target power ratios.

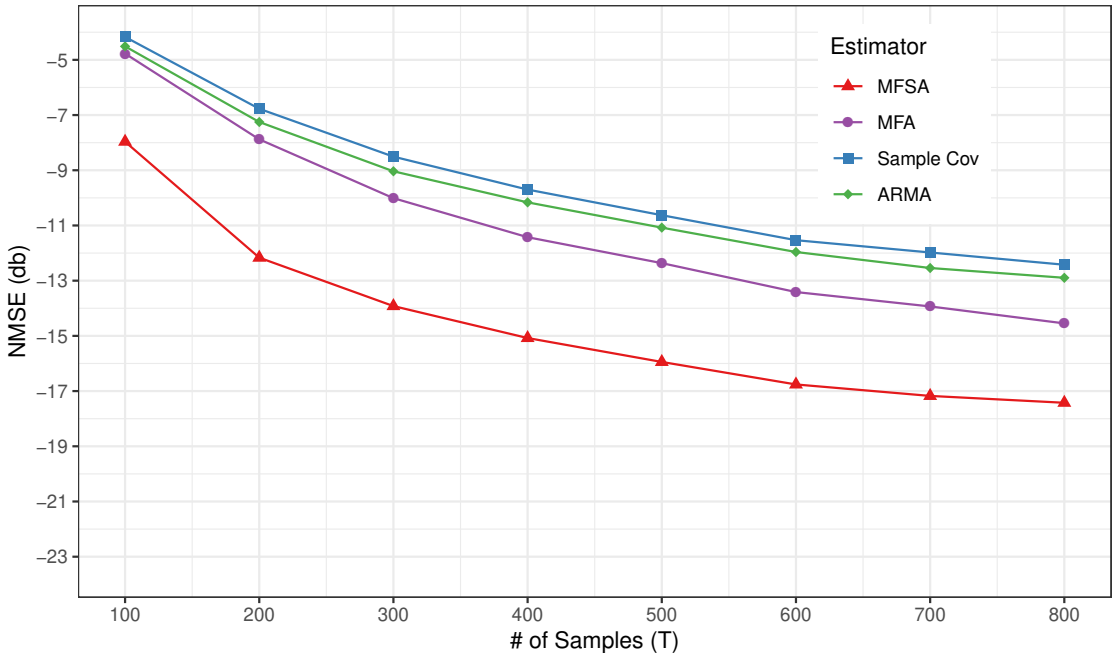
In this experiment, the measure of cross-sectional covariance estimation quality is the Normalized Mean-Squared Error (NMSE), defined as

$$NMSE = E\left[\frac{||\mathbf{R}_{\mathbf{xx}} - \widehat{\mathbf{R}}||_F^2}{||\mathbf{R}_{\mathbf{xx}}||_F^2}\right],$$

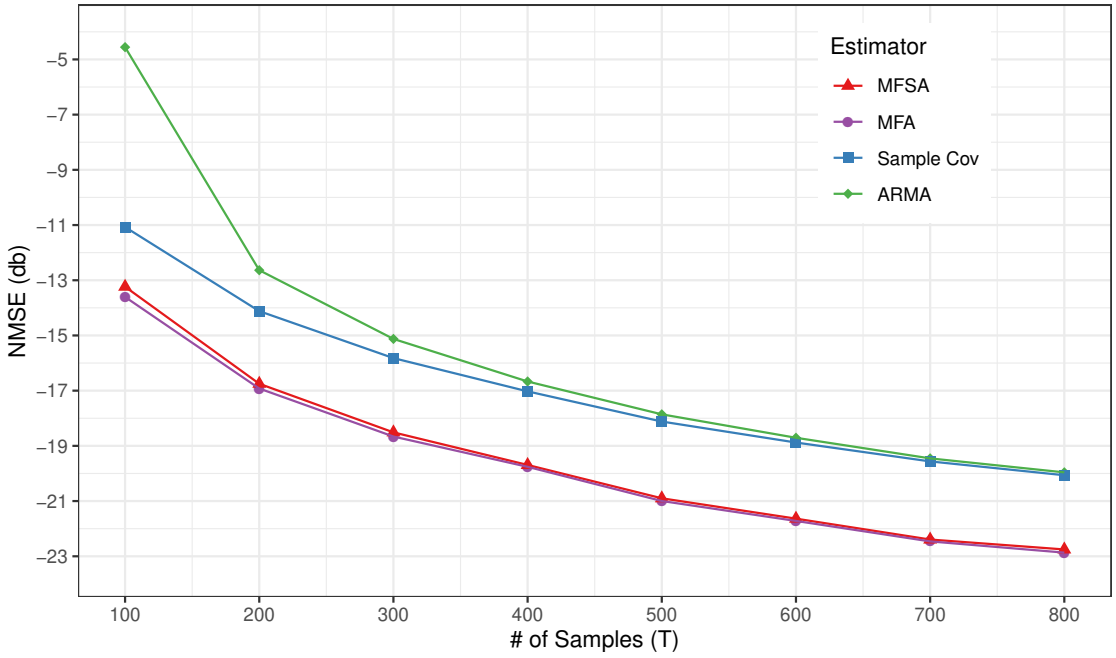
for $\widehat{\mathbf{R}} = \widehat{\mathbf{R}}_{\text{MFSA}}, \widehat{\mathbf{R}}_{\text{MFA}}, \widehat{\mathbf{R}}_{\text{ARMA}}, \widehat{\mathbf{R}}_{\text{Samp}}$, which is estimated by averaging $250$ Monte-Carlo trials at each combination of simulation settings.

In Figures 3.2a and 3.2b, we compare the achieved NMSE for varying number of samples $T$ when the latent factor exhibit temporal correlations versus the setting where the latent factors are uncorrelated across time. Figure 3.2a demonstrates that $\widehat{\mathbf{R}}_{\text{MFSA}}$ exhibits improved performance for estimation of the cross-sectional covariance when non-zero temporal correlation in the latent factors is present, and that the performance gap between $\widehat{\mathbf{R}}_{\text{MFSA}}$ and $\widehat{\mathbf{R}}_{\text{MFA}}$ increases as the number of samples increases. Conversely, when the latent factors are temporally uncorrelated, Figure 3.2b shows that $\widehat{\mathbf{R}}_{\text{MFA}}$ and $\widehat{\mathbf{R}}_{\text{MFSA}}$ perform similarly except when the sample size is small.

In both settings, the estimates $\widehat{\mathbf{R}}_{\text{MFSA}}$ and $\widehat{\mathbf{R}}_{\text{MFA}}$ that leverage the fact that the signal and interference series are belong to the subspace $\mathcal{S} \oplus (\mathcal{I}_1 \oplus \mathcal{I}_2 \oplus \mathcal{I}_3)$ which is of dimension $r_0 + r = 5$ significantly outperform the estimates $\widehat{\mathbf{R}}_{\text{ARMA}}$ and $\widehat{\mathbf{R}}_{\text{Samp}}$ which work directly in the observation space of dimension $n = 21$. As $\widehat{\mathbf{R}}_{\text{MFA}}$ outperforms $\widehat{\mathbf{R}}_{\text{ARMA}}$ in both settings, this experiment demonstrates that leveraging the low-dimensional and channel-structured aspects of the observations while ignoring temporal correlations can be more useful than the reverse, when the factor sizes are sufficiently small relative to the observation sizes and even when the factors are temporally correlated. However, the fact that $\widehat{\mathbf{R}}_{\text{MFSA}}$ is the best-performing estimator in the temporally-correlated setting demonstrates that leveraging both the low-dimensional and channel-structured aspects and modeling the temporal correlations of the latent factors an substantially improve estimation of the cross-sectional correlation matrix.




**(a)** Factor series are temporally correlated




**(b)** Factors are temporally uncorrelated

**Figure 3.2:** Normalized mean-squared error achieved by estimators of zero-lag cross-sectional covariance $\mathbf{R}_{\mathbf{xx}}$ with varying $T$ at different temporal correlations.

### 3.5.2 Transmitted Signal Recovery

In the previous experiment, the cross-sectional correlation was of primary interest while modeling the temporal correlation served only to improve the estimates of $\mathbf{R}_{\mathbf{xx}}$. However, modeling both cross-sectional and temporal correlation directly is useful for many problems of interest. In the multi-channel blind source separation problem of Section 1.3, estimation of the transmitted signal $\mathbf{f}[t]$ on the basis of the observed data is a crucial task. When the latent series exhibit temporal correlations, the information about the transmitted signal at time $t$ is contained in both the measurements at time $t$ as well as at other times. Leveraging the information contained in $\mathbf{x}[s], s \neq t$ to predict $\mathbf{f}[t]$ requires estimation of the entire covariance function, which can be accomplished with MFSA.

To understand the utility of modeling the temporal correlation of the latent factors, this section reports an experiment with $T = 400$ samples, $C = 2$ channels of sizes $n_1 = n_2 = 5$, with $r_0 = 1$ common factors and $r_1 = r_2 = 1$ distinct factors per channel. The average power ratios are equal across channels and set to $\eta_{\mathbf{s},c} = 0.3, \eta_{\mathbf{i},c} = 0.4$ and $\eta_{\mathbf{u},c} = 0.3$. The latent factors are sampled as independent Gaussian ARMA$(2, 1)$ processes where the degree of temporal correlation varies for a fixed sample size. To accomplish this, the largest diagonal poles $v_{j1}$ for $j = 1, \ldots, r_0$ and $w_{c,j1}, C = 1, 2, j = 1, \ldots, r_c$ is set to a value in the interval $[0, 0.95]$ and the magnitude of the other diagonal poles and zeros are uniformly distributed with magnitude less than the magnitude of the largest poles. The remaining parameters are sampled in the same fashion as in Section 3.5.1. The loading parameters are scaled to obtain the desired average power ratios.

For MFSA, the estimates of the transmitted signal $\widehat{\mathbf{f}}_{\text{MFSA}}[t]$ are obtained as the MMSE estimates using all observations $\mathbf{x}[0], \ldots, \mathbf{x}[T-1]$,

$$\widehat{\mathbf{f}}[t] = \widehat{\boldsymbol{\gamma}}[t](\mathbf{I}_T \otimes \widehat{\mathbf{A}}^{\mathsf{H}})\widehat{\boldsymbol{\Gamma}}_{\mathbf{xx}}^{-1}\mathbf{x}, \tag{3.67}$$

where $\mathbf{x} = [\mathbf{x}^{\mathsf{T}}[0], \ldots, \mathbf{x}^{\mathsf{T}}[T-1]]^{\mathsf{T}}$ is the $nT$ vector of stacked observations, $\widehat{\boldsymbol{\Gamma}}_{\mathbf{xx}}$ is the block Topeliz matrix with $sh$-th $n \times n$ block being

$$\widehat{\boldsymbol{\Gamma}}_{\mathbf{xx}}[s-h] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}(s-h)}\mathbf{S}_{\mathbf{xx}}(\omega; \widehat{\boldsymbol{\theta}})d\omega,$$

and $\widehat{\boldsymbol{\gamma}}[t]$ is the $r_0 \times r_0 T$ block matrix whose $s$th block is $\widehat{\boldsymbol{\Gamma}}_{\mathbf{xx}}[t-s]$. These predictions can be compared to the MMSE estimates obtained under the model that the factors are independent,

$$\widehat{\mathbf{f}}_{\mathrm{MFA}}[t] = \mathbf{R}_{\mathbf{ff}}^{1/2} \widehat{\mathbf{A}}^{\mathsf{H}} \widehat{\mathbf{R}}_{\mathrm{MFA}}^{-1} \mathbf{x}[t], \tag{3.68}$$

and to the oracle MMSE estimates of $\mathbf{f}[t]$ obtained as in (3.67) by replacing $\mathbf{S}_{\mathbf{xx}}(\omega; \widehat{\boldsymbol{\theta}})$ with the true power spectral density. In (3.68), the factor $\mathbf{R}_{\mathbf{ff}}^{1/2}$ is included to scale the estimated factors $\widehat{\mathbf{f}}_{\mathrm{MFA}}[t]$ appropriately, as MFA uses the scaling convention where the latent factors have identity covariance. The measure of transmitted signal recovery is the NMSE, which here is

$$NMSE = E\left[\frac{\sum_{t=0}^{T-1} ||\mathbf{f}[t] - \widehat{\mathbf{f}}[t]||^2}{\sum_{t=0}^{T-1} ||\mathbf{f}[t]||^2}\right],$$

and is estimated by $250$ Monte-Carlo trials at each setting.

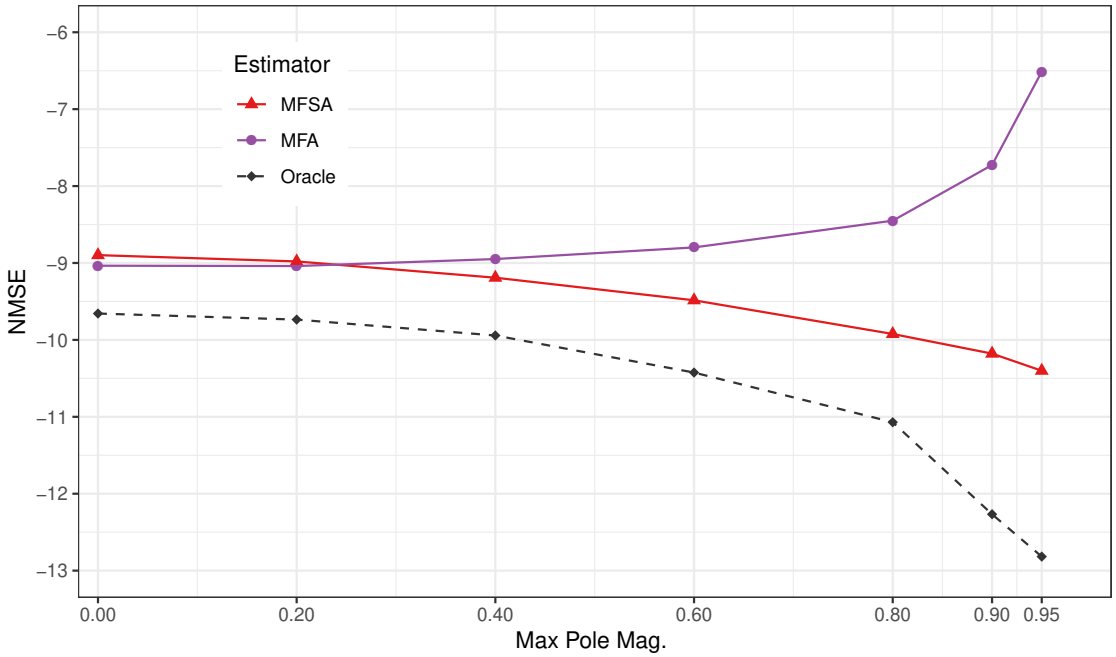
In Figure 3.3a, it can be seen that the estimates $\widehat{\mathbf{f}}_{\mathrm{MFSA}}[t]$ obtained by estimating both temporal and cross-sectional correlations significantly outperform the estimates $\widehat{\mathbf{f}}_{\mathrm{MFA}}[t]$ (which only use the cross-sectional correlations) when the latent series have substantial temporal correlations. Conversely, when the temporal correlation is low, most of the information about $\mathbf{f}[t]$ is contained in $\mathbf{x}_t$ and so the estimates $\widehat{\mathbf{f}}_{\mathrm{MFA}}[t]$ are preferable.

As a further demonstration, Figures 3.3b and 3.3c compare the magnitudes of the factor estimates $||\widehat{\mathbf{f}}_{\mathrm{MFSA}}[t]||$ and $||\widehat{\mathbf{f}}_{\mathrm{MFA}}[t]||$ to the true factor magnitude $||\mathbf{f}[t]||$. The MFSA estimates track the true factor substantially better than the MFA estimates, as the MFSA estimates are able to borrow information about the factors from nearby times $s \neq t$ while MFA uses only the same-time observations $\mathbf{x}[t]$ to estimate $\mathbf{f}[t]$.

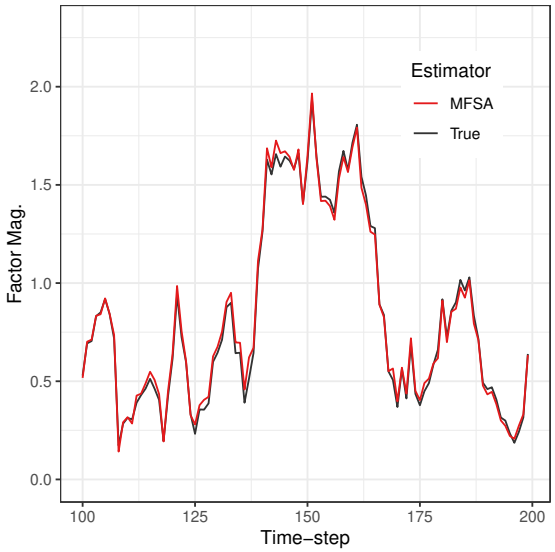
### 3.5.3 Passive Radar Application

As described in Section 1.3, the target detection problem for passive radar can be framed as a two-channel hypothesis testing problem,

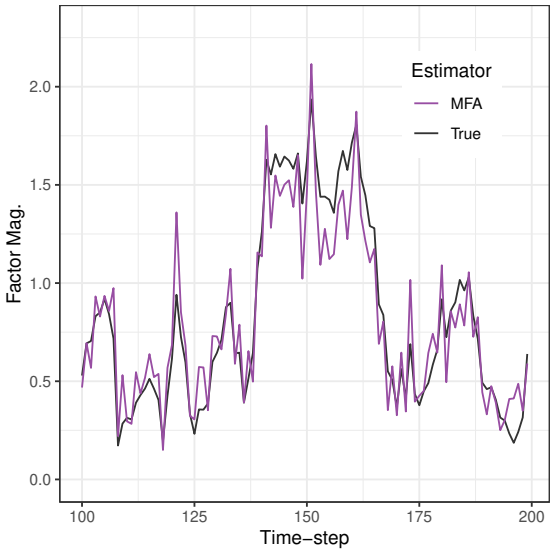
$$\mathcal{H}_1 : \begin{cases} \mathbf{x}_1[t] &= \mathbf{A}_1\mathbf{f}[t] + \mathbf{B}_1\mathbf{g}_1[t] + \mathbf{u}_1[t] \\ \mathbf{x}_2[t] &= \mathbf{A}_2\mathbf{f}[t] + \mathbf{B}_2\mathbf{g}_2[t] + \mathbf{u}_2[t] \end{cases}, \quad \mathcal{H}_0 : \begin{cases} \mathbf{x}_1[t] &= \mathbf{B}_1\mathbf{g}_1[t] + \mathbf{u}_1[t] \\ \mathbf{x}_2[t] &= \mathbf{A}_2\mathbf{f}[t] + \mathbf{B}_2\mathbf{g}_2[t] + \mathbf{u}_2[t] \end{cases}.$$




**(a)** NMSE achieved by $\mathbf{f}[t]$ estimators, for varying degrees of temporal correlation in the latent factors.




**(b)** Comparison of MFSA estimates to true factors




**(c)** Comparison of MFA estimates to true factors

**Figure 3.3:** Comparison of factor estimator performance when latent factors are temporally correlated.

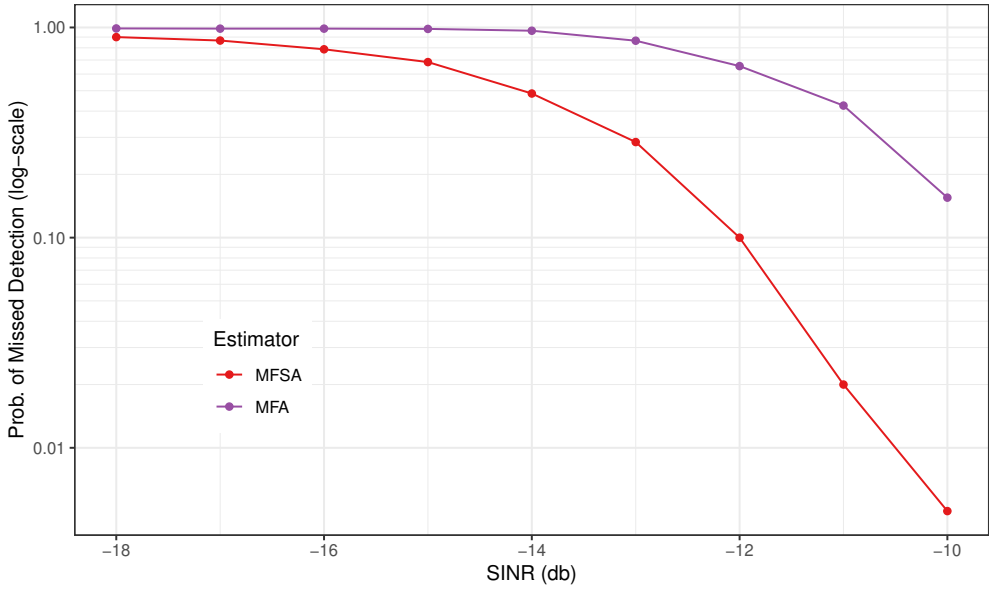



**Figure 3.4:** Probability of missed detection for approximate GLRT $\mathcal{S}\big(\mathbf{I}(\omega)\big)$ using temporal correlations (MFSA) and GLRT $\mathcal{G}\big(\mathbf{x}[0],\ldots,\mathbf{x}[T-1]\big)$ if no temporal correlations were present (MFA), for varying SINR and fixed false-alarm probability $p_{fa} = 10^{-2}$.

Channels $1$ and $2$ are respectively the surveillance and reference channel, and $\mathcal{H}_1$ is the target-present condition while $\mathcal{H}_0$ is the target-absent condition. To test if $\mathbf{x}[t]$ was synthesized by $\mathcal{H}_1$ versus $\mathcal{H}_0$, a test $\mathcal{G}(\mathbf{x}[0],\ldots,\mathbf{x}[T-1])$, which is GLRT when the observations are normal and temporally independent, is proposed in Ramírez et al. (2020, Eq. 75).

To examine how temporal correlation effects the distinguishability of $\mathcal{H}_1$ from $\mathcal{H}_0$, we repeat the passive radar experiment with the introduction of temporally correlated factors. The channels contain $n_1 = n_2 = 10$ scalar measurements each, with $r_0 = 1$ common factor and $r_1 = r_2 = 1$ distinct factors. The total number of time-points is $T = 400$, and the latent series are independent realizations of Gaussian ARMA(1,1) processes where the largest diagonal poles are set at $0.95$. The remainder of the sampling scheme is as in the previous subsection, with the exception that the average SINR, defined as,

$$\mathrm{SINR}_c(db) = 10\log_{10}\left(\frac{\eta_{\mathbf{s}_c}}{\eta_{\mathbf{i}_c} + \eta_{\mathbf{u}_c}}\right),$$

is varied over the range $-18$ through $-10$. The channel-$c$ common factor loading matrix $\mathbf{A}_c$ for $c = 1, 2$ is scaled to obtain the target SINR.

As the estimator proposed in Section 3.3 are obtained by maximizing the Whittle psuedo-likelihood, the test

$$\mathcal{S}\big(\mathbf{I}(\omega)\big) = -2\Big[\max_{\theta\in\boldsymbol{\Theta}_1} \ell\big(\mathbf{I}(\omega);\boldsymbol{\theta}\big) - \max_{\theta\in\boldsymbol{\Theta}_0} \ell\big(\mathbf{I}(\omega);\boldsymbol{\theta}\big)\Big] \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\lessgtr}} \tau \tag{3.69}$$

can be defined, which will be an (approximate) GLRT (Ramírez et al., 2023, Ch. 8.3) when the latent vectors are jointly normal for an appropriate threshold $\tau$, which is estimated by Monte-Carlo simulation under $\mathcal{H}_0$.

Figure 3.4 compares the performance of this test against the temporally-independent GLRT $\mathcal{G}(\mathbf{x}[0],\ldots,\mathbf{x}[T-1])$. It can be seen that the approximate GLRT $\mathcal{S}$, which is a function of the PSD over all frequencies, has improved detection performance compared to $\mathcal{G}$, which is a function of the sample cross-sectional covariance alone.

## 3.6 Discussion

In this chapter, the method of Multi-channel Factor Spectral Analysis (MFSA) is introduced as an extension of MFA to allow for temporal correlations within the latent factors. As many problems within statistics and signal processing to which MFA may be applied exhibit temporal correlation in addition to the within- and across-channel cross-sectional covariance modeled by MFA, this extension can be of value for such problems. As demonstrated by the experiments of Section 3.5, explicitly modeling the temporal correlations can yield dividends on core multi-channel problems like signal recovery and target detection. The theoretical results on MFSA identifiability and the properties of the parameterization found in Sections 3.2.4 and 3.2.5 characterize the model's properties and can be useful building blocks for further results, as is the case for Theorem 5. The MM estimation procedure designed in Section 3.4 to optimize the Whittle psuedo-likelihood demonstrates that the chosen parameterization is tractable despite its complexity. In total, the development of MFSA is a useful addition to the class of second-order methods for analyzing multivariate time series across multiple channels.

# Chapter 4

# Application of MFA to Multi-Vehicle Dynamics in a Highway Environment

## 4.1 Chapter Overview

In this chapter, a model for the trajectories of multiple vehicles in a highway environment is developed. The cross-sectional correlation of these trajectories is jointly modeled by the MFA decomposition (Section 4.4.2), which isolates the across-vehicle covariance from the low-rank within-vehicle covariance and parsimoniously models the common influence of the highway conditions. The real-world data which is analyzed consists of location and kinematic data for many vehicles on a tracked highway segment, where the membership on the tracked segment changes (Section 4.4.1). The resulting partially overlapping trajectories adds complexity to analysis of cross-vehicle correlation, and so the data are modeled as a partially observed joint multivariate Gaussian process (Section 4.4) whose cross-sectional covariance is of MFA form. As the cross-vehicle covariances may change, the MFA decomposition is extended (Section 4.4.3) to allow for time-varying parameter curves. These parameter curves are estimated by the method of moving segments (Section 4.6.2), and the resulting estimators are justified through the theory of locally stationary processes. The goal of clustering vehicles into *pods* whose relative configuration is close to static is introduced and characterized geometrically (Section 4.5). The podding problem is connected to the MFA decomposition, and a dynamic model-based podding procedure (Section 4.6.3) is developed. The results of applying this procedure to *cohorts* of spatiotemporally-proximate vehicles within the Next GENeration SIMulation (NGSIM) highway dataset are investigated (Section 4.7).

## 4.2 Motivation

Understanding vehicle behavior on highways is a fundamental problem for traffic researchers and designers of autonomous driving systems. The highway traffic stream is marked by both substantial regularity in vehicle motion, as well as deviations from the local average motion by individual vehicles. Decomposing the highway traffic stream into more interpretable components is a well-studied problem (Avila and Mezić, 2020), with continued application to decision-making for traffic agencies. However, most existing methods are tailored to identify patterns in aggregate vehicle density and so have reduced capability to describe the motion of individual vehicles or small groups of vehicles. In this Chapter, we present a model-based approach to dynamically segregating cohorts of vehicles into *pods*, which can provide insights at the level of individual vehicles.

The defining characteristic of a pod of vehicles is the stability of the relationships between vehicles within the same pod, when compared to vehicles in other pods. The spatial homogeneity of highways suggests that the relative positions and velocities of pairs of vehicles can be a proxy for the relationship between said vehicles, so the relative configuration of the vehicles within a pod will be more stable than the overall traffic stream. Motivating examples of pods include small groups of vehicles which remain nearby due to safety or familiarity, despite overtaking opportunities. Other collections which meet our pod definition include single outlying vehicles whose motion is potentially erratic and thus very different from that of nearby vehicles, as well as the lane-specific groups which arise when motion is constrained due to congestion.

This pod decomposition can provide a novel lens onto the emergence of regularities within micro-scale traffic data, and tapping into the heightened stability of nearby pods can represent a pro-social strategy for autonomous driving systems. Previous work has shown that cooperative or altruistic autonomous driving systems can yield globally improved outcomes for all participants when the relationships of other vehicles are appropriately accounted for (Toghi et al., 2022, 2021). Additionally, assessing the riskiness of traffic conditions can depend on the stability of the ego vehicle's relation to nearby vehicles, which is directly related to the pod decomposition.

A hurdle to the analysis of traffic data at the level of individual vehicles is the changing identity of the vehicles on a given stretch of highway. Previous work has successfully modeled individual trajectories as multivariate Gaussian processes, which incorporate dependencies between the multiple features associated with a given vehicle as well as temporal dependence. Splitting a cohort into pods requires examining multiple contemporaneous trajectories, so a *joint* multivariate Gaussian process is used to model time-varying inter-vehicle relationships.

The approach taken for dynamic pod assignment relies on developing a model for the relationships between the trajectories of spatiotemporally-proximate vehicles. The modeling strategy used leverages the high degree of coordination and regularity present in vehicle trajectory data via latent factor methods. In light of the multivariate nature of vehicle trajectory observations, we employ multi-channel factor analysis to model the cross-sectional covariance, which distinguishes between the tight, physically-required covariance for features corresponding to the same vehicle and the looser cross-vehicle covariances which are of central interest. This method of analysis is extended to a factor decomposition for the joint multivariate Gaussian process, where factor loading estimation is performed for time-varying parameters on a collection of vehicles whose composition may also vary with time.

## 4.3 Related Work

### 4.3.1 Trajectory Modeling

Modeling individual vehicle trajectories is an area of intense research interest, with competing approaches ranging from classic techniques like the Kalman filter (Mercat et al., 2019), sophisticated deep-learning algorithms (see Leon and Gavrilescu (2021) for a recent survey), and probabilistic models such as multivariate Gaussian processes (Goli et al., 2018; Patterson et al., 2020). The primary application of these methods is for prediction, where neural-network-based methods typically represent the state-of-the-art. However, experiments such as those by Zhu et al. (2019) have found that regression approaches remain competitive with deep-learning methods. These probabilistic models have the advantages of parameter interpretability, uncertainty quantification,

and the ability to be fit with relatively less training data. These advantages underlie the successful use of multivariate Gaussian process regression as a technique for trajectory modeling in both the domain of vehicle traffic (Yoon et al., 2021), and beyond (Eerland et al., 2016; Buelta et al., 2021).

The definition of a pod and its basis in the stability of geometric configuration of a group of a vehicles has interesting connections to vehicular *platoons* (El-Zaher et al., 2012; Maiti et al., 2017). A vehicular platoon is a multi-vehicle system of reactive agents which seek to maintain a predefined relative configuration as they travel together down the highway, without any physical coupling. Smart design of platoon agents can enable reduced fuel usage and increase highway safety and efficiency (Alam et al., 2010). The self-organization of the agents obtained by their individual actions responding to local perceptions allows for the maintenance of the desired geometry in the presence of external shocks (El-Zaher et al., 2011). As the identified pods are likewise defined by having a stable relative configuration, they can be viewed as naturally occurring examples of transient platoons among human drivers who are similarly motivated by personal safety and efficiency.

### 4.3.2 Dynamic Clustering

Measures of similarity between objects can be used to break down a collection of objects into subgroups via cluster analysis. Clustering methods are tightly connected with mixture models, which assume the observed objects have unobserved labels that identify a particular mixture component from which that object is sampled. For mixture models, the problem of choosing the number of mixture components can be side-stepped by moving from a finite mixture model to a infinite mixture model. A standard model which enables this is the Dirichlet process mixture model, which uses a Bayesian nonparametric approach to specify the mixing distribution and allows for an unbounded number of mixture components (Antoniak, 1974). However, posterior inference for Dirichlet process mixtures can be challenging. A tractable alternative was given by Kulis and Jordan (2012), whose DP-means algorithm derives a straightforward $k$-means-like iterative approach from the small-variance limit of the Dirichlet process mixture model with Gaussian kernel.

Clustering methods have been applied to vehicle trajectory data for a variety of goals. These include driving style identification (de Zepeda et al., 2021; Wang et al., 2020, 2017), vehicle state segmentation (Yi et al., 2016), risk assessment (Guo and Fang, 2013; Shi et al., 2022), maneuver detection and execution (Pang, 2019; Mahajan et al., 2020), lane detection (Hota et al., 2009), and traffic shockwave detection (Lu and Skabardonis, 2007), among others.

Static cluster analysis assigns a fixed label to a specific object. For the problem of identifying vehicle pods, we require the flexibility that the pod assignments and composition can change over time. This problem of time-changing cluster assignments is known as *dynamic clustering*, and it provides a sequence of labels for a given time series which capture the evolution of similarities between objects as time passes. Dynamic clustering methods for time-series data is an active area of research (Bouchachia, 2012; Sartório and Fonseca, 2020; Aghabozorgi et al., 2015). Recent work in dynamic driving style identification by de Zepeda et al. (2021) has underscored the importance of directly incorporating the nonstationarity of driver behavior to obtain deeper insights.

A particular application of dynamic clustering to vehicles which has received a great deal of interest is the development of vehicular ad-hoc networks (VANETs). For these networks, identification of groups of vehicles which are spatially close and likely to remain so for some time is necessary to reduce the costs of network construction and maintenance. Accordingly, a number of clustering methods are based on problem-specific distances such as Link Expiration Time (Benslimane et al., 2011) or heuristics like highest-degree and lowest-ID (Lin and Gerla, 1997). More sophisticated algorithms such as WCA (Chatterjee et al., 2002), VWCA (Daeinabi et al., 2011), or QMM-VANET (Fatemidokht and Rafsanjani, 2020) weigh combinations of time-varying factors such as distance, average speed, trust, and network degree to assign time-varying clusters. These methods differ substantially from the dynamic clustering algorithm proposed in this Chapter, as they are specifically tailored towards mobile network construction.

## 4.4 Trajectory Model

### 4.4.1 Multi-Vehicle Trajectories

Development of a dynamic podding method requires simultaneous analysis of the motion of many vehicles in order to identify commonalities. For this purpose, complete trajectory datasets (Section 4.4.1) provide the necessary trajectory data for many vehicles. Previously successful models of trajectory data with multivariate Gaussian processes can be extended to *joint* multivariate Gaussian processes, which provides the desired framing for the discovery of commonalities in vehicle motion.

#### Complete Vehicle Trajectory Datasets

A useful class of datasets for traffic researchers and ADS engineers are *complete vehicle trajectory* datasets (Seo et al., 2020). A complete vehicle trajectory dataset consists of positions for all vehicle within a delineated road section, observed at high frequency over some substantial duration. From these high-frequency observations, trajectories can be interpolated for all vehicles on the road during the collection interval. A classic example of a complete vehicle trajectory dataset is the Next GENeration SIMulation (NGSIM) project (U.S Department of Transportation, Federal Highway Administration, 2016), which captured vehicle trajectories on 3 road segments recorded for 30-45 minute durations at 10Hz. More modern examples include HighD (Krajewski et al., 2018), Zen Traffic Dataset (Hanshin Exp. Co. Ltd., 2024), and others.

For complete vehicle trajectory datasets, the structure of the resulting recordings poses several analytic challenges. As the typical travel-time down the road segment is much shorter that the total recording duration, the captured trajectories are non- or partially-overlapping in time. That is, many pairs of vehicles do not ever occupy the tracked road segment at the same time, and those pairs which do simultaneously occupy the tracked segment may only do so for a short time and may be spatially distant. These complications make joint trajectory modeling difficult, and many analyses develop marginal models for the trajectories of individual vehicles without directly modeling the

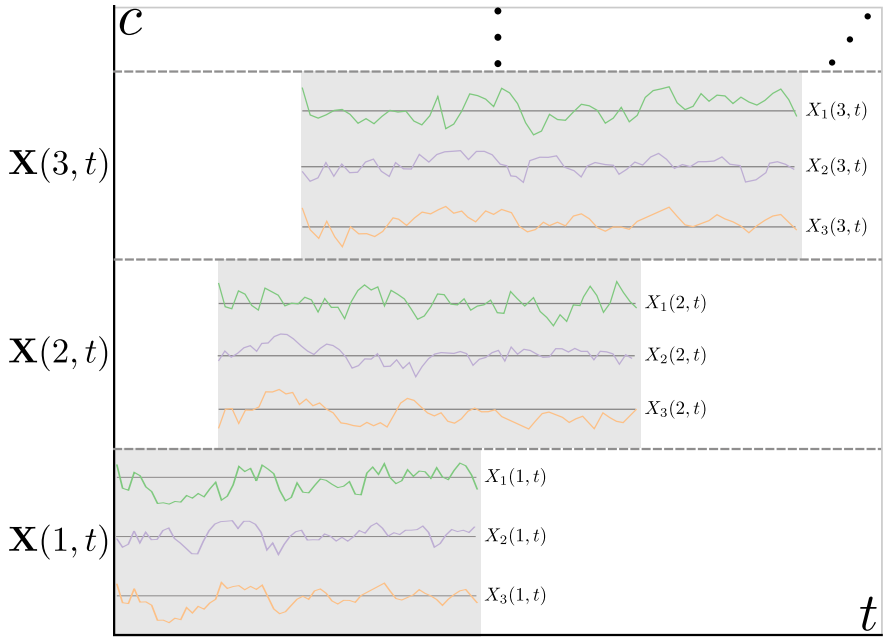



**Figure 4.1:** A schematic diagram of multivariate trajectories for vehicles $c = 1, 2, 3, \ldots$ with $n = 3$ features. The time-index $t$ of the trajectory is the global recording time, leading to partially-overlapping trajectories of differing lengths. White spaces indicate potential $(c, t)$ pairs which are not observed, due to pair occurring before the vehicle enters the tracked road segment or after it leaves. Between its entry and exit times, a vehicle's trajectory is completely observed.

interactions between spatially and temporally proximate vehicles. As the interactions between vehicles are of central interest, we instead frame the data as a partially-observed joint multivariate Gaussian process, where the structure of the partially-observed data is depicted in Figure 4.1.

### Joint Multivariate Gaussian Processes

A Gaussian Process (GP) on an index set $I$ is a flexible and widely-used method of specifying a random function $x : I \to \mathbb{R}$, via the constraint that for any finite set of indices $t_1, \ldots t_q$, the random vector $[x(t_1), \ldots, x(t_q)]^\mathsf{T}$ will have a multivariate Gaussian distribution. A GP can be fully specified by a mean function $\mu : I \to \mathbb{R}$ and a symmetric, positive definite covariance kernel $\kappa : I \times I \to \mathbb{R}$. The GP model can be extended to the $n$-dimensional *multivariate* Gaussian process (MV-GP) (Crépey and Dixon, 2020; Chen et al., 2023), which in turn models a random function $\mathbf{x}(t) : I \to \mathbb{R}_n$ via the mean function $\boldsymbol{\mu}(t) : I \to \mathbb{R}_n$, covariance kernel $\kappa(t_1, t_2) : I \times I \to \mathbb{R}$ and column-covariance matrix $\mathbf{R} \in \mathbb{R}_{n \times n}$ by requiring that the marginal distribution for any vector of

indices $\mathbf{t} = [t_1, \ldots, t_T]^\mathsf{T}$ satisfy

$$\mathbf{X}(\mathbf{t}) \sim \mathcal{MN}(\mathbf{M}(\mathbf{t}), \mathbf{R}, \mathbf{U}(\mathbf{t})),$$

where the matrices $\mathbf{X}(t), \mathbf{M}(t) \in \mathbb{R}_{n\times T}$ and $\mathbf{U}(\mathbf{t}) \in \mathbb{R}_{T\times T}$ are

$$\mathbf{X}(\mathbf{t})=\begin{bmatrix}\mathbf{x}(t_1) & \ldots & \mathbf{x}(t_T)\end{bmatrix},\ \mathbf{M}(\mathbf{t})=\begin{bmatrix}\boldsymbol{\mu}(t_1) & \ldots & \boldsymbol{\mu}(t_T)\end{bmatrix},\ \mathbf{U}(\mathbf{t})=\begin{bmatrix}\kappa(t_1,t_1) & \ldots & \kappa(t_1,t_T)\\ \vdots & \ddots & \vdots \\ \kappa(t_T,t_1) & \ldots & \kappa(t_T,t_T)\end{bmatrix}.$$

As the matrix normal distribution $\mathcal{MN}(\mathbf{M}, \mathbf{R}, \mathbf{U})$ for random matrix $\mathbf{Y} \in \mathbb{R}_{n\times T}$ is equivalent to requiring that $\mathrm{vec}(\mathbf{Y})$ have multivariate normal distribution with mean $\mathrm{vec}(\mathbf{M})$ and covariance $\mathbf{R} \otimes \mathbf{U}$, the covariance of the MV-GP model is *separable* into the index covariance kernel $\kappa(t_1, t_2)$ and the column-covariance matrix $\mathbf{R}$. The multivariate Gaussian process (MV-GP) method has been successfully been used as a model for both continuous-time ($I \subset \mathbb{R}$) and discrete-time ($I \subset \mathbb{Z}$) trajectory data (Patterson et al., 2020; Eerland et al., 2016).

An objective of the current work is to examine the relationship between the trajectories of nearby vehicles, and hence the joint distribution of multiple trajectories is to be specified. To build on the utility of the MV-GP model for individual trajectories, the MV-GP model can be extended to a multivariate stochastic process with domain $\mathbb{N} \times I$ and range $\mathbb{R}_n$. This $n$-dimensional *joint* multivariate Gaussian Process (JMV-GP) gives a probabilistic model for an unbounded number of trajectories, such that the marginal distribution of any fixed finite cohort is an MV-GP. Elementary properties of JMV-GPs are provided in Section D.1 in Appendix D.

The mean function $\boldsymbol{\mu}(c, t) : \mathbb{N} \times I \to \mathbb{R}_n$ and covariance kernel $\boldsymbol{\kappa}(c_1, t_1, c_2, t_2) : (\mathbb{N} \times I) \times (\mathbb{N} \times I) \to \mathbb{R}_{n\times n}$ serve to fully specify the JMV-GP. Two potential properties of these functions will be used in model development, namely *stationarity* and *separability*. A JMV-GP is temporally *stationary* if there exist functions $\boldsymbol{\mu}_0(c)$ and $\boldsymbol{\kappa}_0(c_1, c_2, \tau)$ such that $\boldsymbol{\mu}(c, t) = \boldsymbol{\mu}_0(c)$ and $\boldsymbol{\kappa}(c_1, t_1, c_2, t_2) = \boldsymbol{\kappa}_0(c_1, c_2, |t_1 - t_2|)$. Similarly, a JMV-GP has *separable covariance* if $\boldsymbol{\kappa}$ factors as

$$\boldsymbol{\kappa}(c_1, t_1, c_2, t_2) = \kappa_{\text{time}}(t_1, t_2)\boldsymbol{\kappa}_{\text{id}}(c_1, c_2) \tag{4.1}$$

where $\kappa_{\text{time}} : I \times I \to \mathbb{R}$ is a scalar covariance kernel, and $\boldsymbol{\kappa}_{\text{id}} : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_{n\times n}$ is a matrix covariance kernel.

For JMV-GPs with separable covariance, the marginal distribution for *rectangular* index sets has simple form. An set of indices is rectangular if it can be written as the Cartesian product $\{c_1, \ldots, c_C\} \times \{t_1, \ldots, t_T\}$, with associated index vectors $\mathbf{c} = [c_1, \ldots, c_C]^\mathsf{T}$ and $\mathbf{t} = [t_1, \ldots, t_T]^\mathsf{T}$. The index set then consists of the same set of trajectories observed at the same set of times, and hence the measurements can be arranged into a rectangular $Cn \times T$ matrix:

$$\mathbf{X}(\mathbf{c}, \mathbf{t}) = \begin{bmatrix} \mathbf{x}(c_1, t_1) & \ldots & \mathbf{x}(c_1, t_T) \\ \vdots & \ddots & \vdots \\ \mathbf{x}(c_C, t_1) & \ldots & \mathbf{x}(c_C, t_T) \end{bmatrix}. \tag{4.2}$$

As shown in Section D.1, the separability condition implies that $\mathbf{X}(\mathbf{c}, \mathbf{t})$ has matrix normal distribution,

$$\mathbf{X}(\mathbf{c}, \mathbf{t}) \sim \mathcal{MN}\big(\mathbf{M}(\mathbf{c}, \mathbf{t}), \mathbf{R}(\mathbf{c}), \mathbf{U}(\mathbf{t})\big) \tag{4.3}$$

with the following block-structured matrices for $i = 1, \ldots, C$ and $j = 1, \ldots, T$,

$$\begin{aligned} \mathbf{M}(\mathbf{c}, \mathbf{t}) &= [\boldsymbol{\mu}(c_i, t_j)]_{ij} \quad (n \times 1 \text{ blocks}), \\ \mathbf{U}(\mathbf{t}) &= [\kappa_{\text{time}}(t_i, t_j)]_{ij} \quad (1 \times 1 \text{ blocks}), \\ \mathbf{R}(\mathbf{c}) &= [\boldsymbol{\kappa}_{\text{id}}(c_i, c_i)]_{ij} \quad (n \times n \text{ blocks}). \end{aligned} \tag{4.4}$$

This motivates the description of the defined process as a *joint* multivariate Gaussian process as, for any fixed index vector $\mathbf{c} = [c_1, \ldots, c_C]^\mathsf{T}$ and separable covariance kernel $\boldsymbol{\kappa}(c_1, t_1, c_2, t_2) = \kappa_{\text{time}}(t_1, t_2)\boldsymbol{\kappa}_{\text{id}}(c_1, c_2)$, the stochastic process $\mathbf{x}(\mathbf{c}, t) = [\mathbf{x}^\mathsf{T}(c_1, t) \ \ldots \ \mathbf{x}^\mathsf{T}(c_C, t)]^\mathsf{T}$ is an MV-GP with covariance kernel $\kappa_{\text{time}}(t_1, t_2)$ and column-covariance $\mathbf{R}(\mathbf{c})$.

### 4.4.2 Stationary Separable JVM-GPs with MFA Decomposition

For the purpose of joint trajectory analysis, the main modeling choice is picking the class of cross-sectional covariance kernels $\boldsymbol{\kappa}_{\text{id}}$ under consideration. Unlike for spatial or temporal Gaussian processes, we may have no reason to assume that $c_1$ being close to $c_2$ should imply that they are

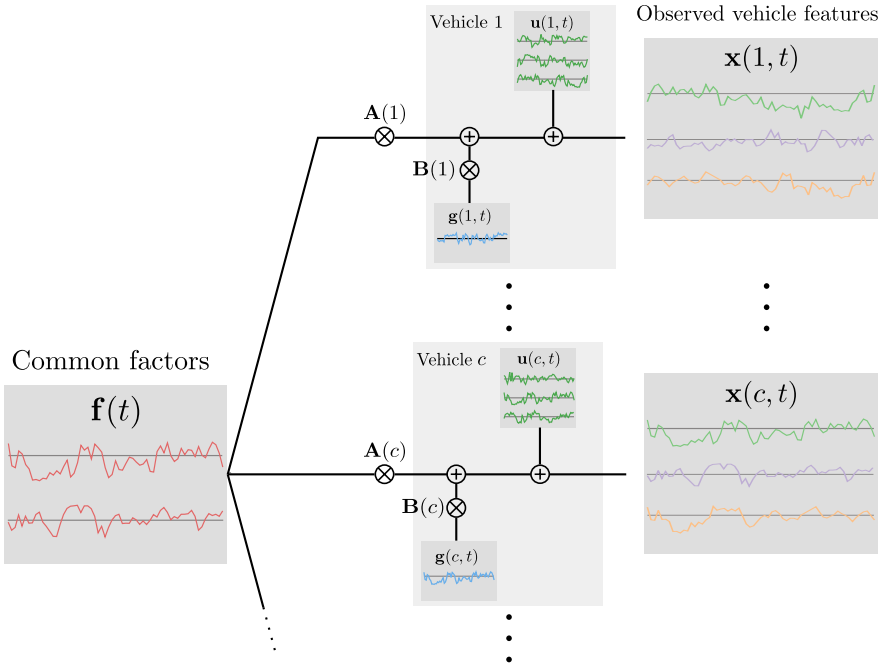



**Figure 4.2:** A synthesis diagram for a JMV-GP with stationary MFA decomposition as defined in (4.5). The $r_0 = 2$ common factors form a latent MV-GP which is multiplied by vehicle-specific loadings $\mathbf{A}(c)$ and provide all cross-vehicle correlation in the model. Each vehicle has a distinct factor as $r_1 = 1$, which models additional within-vehicle correlation between features, modified by $\mathbf{B}(c)$. The idiosyncratic error process $\mathbf{u}(c,t) = \mathbf{\Phi}^{\frac{1}{2}}\mathbf{w}(c,t)$ explains remaining variability in the vehicles features. The linear combination of these three sources produces the observed series $\mathbf{x}(c,t)$ for vehicle $c$.

more correlated than if $c_1$ is far from $c_2$. Instead, we use the MFA covariance model (1.6) as the cross-sectional covariance, wherein the observed across-vehicle commonalities are due to the influence of a small number of latent across-vehicle factors. Further, the covariances within the same vehicle are affected by within-vehicle factors, and remaining variation is attributed to idiosyncratic noise. By partitioning the covariance in this way, we isolate the across-vehicle covariances (which are of primary interest), and model them parsimoniously. For clarity of presentation, we assume that the number of scalar measurements associated with each vehicle is $n$ and does not vary across vehicles. Similarly, we assume that the number of distinct factors per vehicle is $r_1$ for each vehicle.

To define the stationary separable MFA model for the observation JMV-GP $\mathbf{x}(c,t)$, select some stationary temporal covariance kernel $\kappa_{\text{time}}(|t_1 - t_2|; \boldsymbol{\theta})$ which is known up to a finite dimensional vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ of temporal covariance parameters and where the associated covariance matrices $\mathbf{U}(\mathbf{t}; \boldsymbol{\theta})$ are positive definite for all $\mathbf{t}, \boldsymbol{\theta} \in \boldsymbol{\Theta}$. For identification, assume that $\kappa_{\text{time}}$ is scaled such that $\kappa_{\text{time}}(0; \boldsymbol{\theta}) = 1$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let the latent across-vehicle factors $\mathbf{f}(t)$ be a $r_0$-dimensional zero-mean MV-GP with covariance kernel $\kappa_{\text{time}}(|t_1 - t_2|; \boldsymbol{\theta})$ and column-covariance matrix $\mathbf{I}_{r_0}$. The within-vehicle factors $\mathbf{g}(c,t)$ are an $r_1$-dimensional JMV-GP and the noise $\mathbf{w}(c,t)$ is an $n$-dimensional JMV-GP. Both $\mathbf{g}(c,t)$ and $\mathbf{w}(c,t)$ are zero mean with stationary covariance kernels which is separable into $\kappa_{\text{time}}(|t_1 - t_2|; \boldsymbol{\theta})$ and $\boldsymbol{\kappa}_{\text{id}}(c_1, c_2)$ being the appropriately sized identity matrix if $c_1 = c_2$ or the zero matrix if $c_1 \neq c_2$. The latent JMV-GPs $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{w}$ are are assumed to be independent from one another. With these assumptions, the observation process $\mathbf{x}(c,t)$, defined as

$$\mathbf{x}(c,t) = \mathbf{A}(c)\mathbf{f}(t) + \mathbf{B}(c)\mathbf{g}(c,t) + \boldsymbol{\Phi}^{\frac{1}{2}}(c)\mathbf{w}(c,t) + \boldsymbol{\mu}(c) \tag{4.5}$$

for fixed mean function $\boldsymbol{\mu} : \mathbb{N} \to \mathbb{R}_n$, is a stationary separable JMV-GP with multi-channel factor decomposition. The cross-sectional covariance parameters are the common factor loadings $\mathbf{A} : \mathbb{N} \to \mathbb{R}_{n \times r_0}$, distinct factor loadings $\mathbf{B} : \mathbb{N} \to \mathbb{R}_{n \times r_1}$, and noise variances $\boldsymbol{\Phi} : \mathbb{N} \to \mathbb{P}_n$. The covariance function for the observation process then has two forms. The within-vehicle covariance

$\boldsymbol{\kappa}_{\mathbf{x}}(c, t_1, c, t_2)$ for vehicle $c$ is

$$\boldsymbol{\kappa}_{\mathbf{x}}(c, t_1, c, t_2) = \kappa_{\text{time}}(|t_1 - t_2|; \boldsymbol{\theta}) \left(\mathbf{A}(c)\mathbf{A}^{\mathsf{T}}(c) + \mathbf{B}(c)\mathbf{B}^{\mathsf{T}}(c) + \boldsymbol{\Phi}(c)\right) \tag{4.6}$$

while the across-vehicle covariance $\boldsymbol{\kappa}_{\mathbf{x}}(c_1, t_1, c_2, t_2)$ between $c_1 \neq c_2$ is

$$\boldsymbol{\kappa}_{\mathbf{x}}(c_1, t_1, c_2, t_2) = \kappa_{\text{time}}(|t_1 - t_2|; \boldsymbol{\theta})\mathbf{A}(c_1)\mathbf{A}^{\mathsf{T}}(c_2). \tag{4.7}$$

For a rectangular index set with index vectors $\mathbf{c}$ and $\mathbf{t}$, the marginal distribution of $\mathbf{X}(\mathbf{c}, \mathbf{t})$ is

$$\mathbf{X}(\mathbf{c}, \mathbf{t}) \sim \mathcal{MN}\big(\mathbf{M}(\mathbf{c}), \mathbf{A}(\mathbf{c})\mathbf{A}^{\mathsf{T}}(\mathbf{c}) + \mathbf{B}(\mathbf{c})\mathbf{B}^{\mathsf{T}}(\mathbf{c}) + \boldsymbol{\Phi}(\mathbf{c}), \mathbf{U}(\mathbf{t}; \boldsymbol{\theta})\big) \tag{4.8}$$

where

$$\mathbf{A}(\mathbf{c}) = \begin{bmatrix} \mathbf{A}(c_1) \\ \vdots \\ \mathbf{A}(c_C) \end{bmatrix} \quad \mathbf{B}(\mathbf{c}) = \begin{bmatrix} \mathbf{B}(c_1) & & \\ & \ddots & \\ & & \mathbf{B}(c_C) \end{bmatrix} \tag{4.9}$$

and $\boldsymbol{\Phi}(\mathbf{c}) = \text{diag}(\boldsymbol{\Phi}(c_1), \ldots, \boldsymbol{\Phi}(c_C))$. The mean matrix $\mathbf{M}(\mathbf{c})$ and temporal covariance matrix $\mathbf{U}(\mathbf{t}; \boldsymbol{\theta})$ are as defined in (4.4).

### 4.4.3 JMV-GPs with Time-Varying MFA Decomposition

The assumption of a stationary covariance for $\mathbf{x}(c, t)$ implies that the covariance between same-time observations of two series is constant. In the context of vehicle trajectories, this may not be tenable over longer periods of observation. Instead, it is desirable to allow for a changing relationship between vehicles, which can be accomplished by allowing the loadings to vary slowly over time. Extending the observation process (4.5) to include time-varying parameter curves $\mathbf{A}(c, t)$, $\mathbf{B}(c, t)$, $\boldsymbol{\Phi}(c, t)$ and $\boldsymbol{\mu}(c, t)$ gives the time-varying decomposition,

$$\mathbf{x}(c, t) = \mathbf{A}(c, t)\mathbf{f}(t) + \mathbf{B}(c, t)\mathbf{g}(c, t) + \boldsymbol{\Phi}^{\frac{1}{2}}(c, t)\mathbf{w}(c, t) + \boldsymbol{\mu}(c, t). \tag{4.10}$$

Then $\mathbf{x}(c,t)$ is a non-separable JMV-GP whose within-vehicle ($c_1 = c_2$) covariance is

$$\begin{aligned}\boldsymbol{\kappa}_{\mathbf{x}}(c_1,t_1,c_1,t_2) = \kappa_{\text{time}}(t_1,t_2;\boldsymbol{\theta})\Big(&\mathbf{A}(c_1,t_1)\mathbf{A}^{\mathsf{T}}(c_1,t_2) + \mathbf{B}(c_1,t_1)\mathbf{B}^{\mathsf{T}}(c_1,t_2) \\ &+ \boldsymbol{\Phi}^{\frac{1}{2}}(c_1,t_1)\boldsymbol{\Phi}^{\frac{1}{2}}(c_1,t_2)\Big)\end{aligned} \tag{4.11}$$

and across-vehicle ($c_1 \neq c_2$) covariance is

$$\boldsymbol{\kappa}_{\mathbf{x}}(c_1,t_1,c_2,t_2) = \kappa_{\text{time}}(t_1,t_2;\boldsymbol{\theta})\big(\mathbf{A}(c_1,t_1)\mathbf{A}(c_2,t_2)^{\mathsf{T}}\big). \tag{4.12}$$

Estimation of the time-varying factor decomposition can be approximately reduced to estimation of the stationary separable factor decomposition, if the time-indices fall within some sufficiently small window and the parameter curves are slowly varying. Formally relating the time-varying decomposition to stationary approximations at each time point can be accomplished in the framework of locally stationary processes (Dahlhaus, 1996b, 2012), which provides justification for the approach presented in Section 4.6.2. Locally stationary process theory obtains properties of estimators for time-varying parameters under *infill* asymptotics, wherein the number of observations used to estimate the individual values of the parameter curve is allowed to increase while the "total duration" of observation is held constant. A locally stationary process can defined as being close to an appropriate collection of stationary processes (Vogt, 2012), and stationary estimation procedures based on moving windows of the observations can be justified via infill asymptotic considerations (Chiann and Morettin, 2005).

For the present problem, let $\mathcal{T}$ indicates the total number of time-points observed and define a sequence of JMV-GPs $\{\mathbf{x}_{\mathcal{T}}(c,t)\}_{\mathcal{T}=1}^{\infty}$ on $\mathbb{N} \times \{1 \dots \mathcal{T}\}$. For consideration of infill asymptotic behavior, underlying continuous parameter curves are instead defined on the *rescaled* time interval $[0,1]$ via the mapping $t \mapsto t/\mathcal{T}$. That is, the parameter curves of the form $\mathbf{A}(c,t)$ are related to an underlying time-rescaled parameter curve $\mathbf{A}'(\mathbf{c},u) : \mathbb{N} \times [0,1] \to \mathbb{R}_{n \times r_0}$ by

$$\mathbf{A}(c,t) = \mathbf{A}'(c,t/\mathcal{T}).$$

Given the time-rescaled cross-sectional parameter curves $\mathbf{A}'(c,u), \mathbf{B}'(c,u)$ and $\mathbf{\Phi}'(c,u)$ whose domains are $\mathbb{N} \times [0,1]$ as well as a time-rescaled temporal covariance curve $\boldsymbol{\theta}'(u) : [0,1] \to \mathbf{\Theta}$ and $\boldsymbol{\mu}'(c,u) : \mathbb{N} \times [0,1] \to \mathbb{R}_n$, the sequence of JMV-GPs $\{\mathbf{x}_\mathcal{T}(c,t)\}_{\mathcal{T}=1}^\infty$ on $\mathbb{N} \times \{1,\ldots,\mathcal{T}\}$ is

$$\mathbf{x}_\mathcal{T}(c,t) \equiv \mathbf{A}'(c,t/\mathcal{T})\mathbf{f}_\mathcal{T}(t) + \mathbf{B}'(c,t/\mathcal{T})\mathbf{g}_\mathcal{T}(c,t) + \mathbf{\Phi}'^{\frac{1}{2}}(c,t/\mathcal{T})\mathbf{w}_\mathcal{T}(c,t) + \boldsymbol{\mu}'(c,t/\mathcal{T}). \tag{4.13}$$

Here, $\{\mathbf{f}_\mathcal{T}(t)\}_{\mathcal{T}=1}^\infty$ is a sequence of zero-mean MV-GPs on $\{1,\ldots,\mathcal{T}\}$ with non-stationary covariance kernels $\{\kappa_{\mathcal{T},\text{time}}\big(t_1,t_2;\boldsymbol{\theta}'(u)\big)\}_{\mathcal{T}=1}^\infty$ specified by a temporal parameter curve $\boldsymbol{\theta}'(u)$ and column-covariance $\mathbf{I}_{r_0}$. Similarly, $\{\mathbf{g}_\mathcal{T}(c,t)\}_{\mathcal{T}=1}^\infty$ and $\{\mathbf{w}(c,t)\}_{\mathcal{T}=1}^\infty$ are sequences of non-stationary separable zero-mean JMV-GPs on $\mathbb{N}\times\{1,\ldots,\mathcal{T}\}$ with temporal kernels $\{\kappa_{\mathcal{T},\text{time}}\big(t_1,t_2;\boldsymbol{\theta}'(u)\big)\}_{\mathcal{T}=1}^\infty$ and $\boldsymbol{\kappa}_\text{id}(c_1,c_2)$ being the appropriately sized identity matrix if $c_1 = c_2$ or zero if $c_1 \neq c_2$.

**Local Stationarity of $\mathbf{x}_\mathcal{T}(c,t)$**

As suggested by the name, locally stationary processes can be defined (Vogt, 2012) as those processes which are, at any time, close in probability to some (time-dependent) member of a collection of stationary processes on the rescaled time interval. That is, a sequence of vector-valued processes $\{\mathbf{y}_\mathcal{T}(t)\}_{\mathcal{T}=1}^\infty$ with domain $\{1,\ldots,T\}$ is *locally stationary* if, for each rescaled time-point $u \in [0,1]$, there exists a strictly stationary process $\widetilde{\mathbf{y}}_t(u)$ such that

$$||\mathbf{y}_\mathcal{T}(t) - \widetilde{\mathbf{y}}_t(u)|| \leq \left(|t/\mathcal{T} - u| + \mathcal{T}^{-1}\right) U_{t,\mathcal{T}}(u) \quad \text{a.s,} \tag{4.14}$$

where $U_{t,\mathcal{T}}(u)$ is a positive process with a finite $\rho$th moment for $\rho > 0$ not dependent on $u$ or $\mathcal{T}$. This ensures $U_{t,\mathcal{T}}(u)$ is bounded in probability, so $||\mathbf{y}_\mathcal{T}(t) - \widetilde{\mathbf{y}}_t(u)||$ is $O_p\left(|t/\mathcal{T} - u| + \mathcal{T}^{-1}\right)$. Motivated by this definition, we can define a locally stationary sequence of JMV-GPs $\{\mathbf{x}_\mathcal{T}(c,t)\}_{\mathcal{T}=1}^\infty$ by requiring that, for any fixed finite cohort $\mathbf{c} = [c_1,\ldots,c_C]^\mathsf{T}$, the associated sequence of stacked processes $\{\mathbf{x}_\mathcal{T}(t)\}_{\mathcal{T}=1}^\infty$ with $\mathbf{x}_\mathcal{T}(t) = [\mathbf{x}^\mathsf{T}(c_1,t) \ \ldots \ \mathbf{x}^\mathsf{T}(c_C,t)]^\mathsf{T}$ is locally stationary as defined in (4.14) for $\rho$ not depending on $\mathbf{c}, u$ or $\mathcal{T}$.

With this definition, the following proposition shows that local stationarity of the latent processes and smooth and bounded cross-sectional parameter curves ensures that $\{\mathbf{x}_\mathcal{T}(c,t)\}_{\mathcal{T}=1}^\infty$ is a locally stationary sequence of JMV-GPs. The assumptions on the cross-sectional parameter curves used for

this proposition and for Theorem 6 are contained in the following assumption. Here, $\mathbf{c}_{\mathcal{T}}(t)$ and $\mathbf{c}'(u)$ are time-varying cohort vectors in original and rescaled time respectively, with $\mathbf{c}_{\mathcal{T}}(t) = \mathbf{c}'(t/\mathcal{T})$. The respective lengths $C_{\mathcal{T}}(t)$ and $C'(u)$ are such that $\mathbf{c}_{\mathcal{T}}(t) \in \mathbb{N}^{C_{\mathcal{T}}(t)}$ and $\mathbf{c}'(u) \in \mathbb{N}^{C'(u)}$.

**Assumption 1.** *(Cross-Sectional) The time-rescaled cross-sectional parameter curves satisfy:*

*(A1) For all $c$, $\mathbf{A}'(c,u)$, $\mathbf{B}'(c,u)$, $\mathbf{\Phi}'^{\frac{1}{2}}(c,u)$ and $\boldsymbol{\mu}'(c,u)$ are Lipschitz continuous in $u$ for some Lipschitz constants that do not depend on $c$.*

*(A2) The norms $||\mathbf{A}'(c,u)||_F$, $||\mathbf{B}'(c,u)||_F$, and $||\mathbf{\Phi}'(c,u)||_F$ are uniformly bounded above by constants that do not depend on $(c,u)$.*

*(A3) The noise covariance is bounded away from zero,*

$$\min_{i\in\{1,\ldots,n\}} [\mathbf{\Phi}'(c,u)]_{ii} \geq \epsilon > 0,$$

*for some constant $\epsilon$ which does not depend on $(c,u)$,*

*(A4) The cross-sectional covariance parameter curve $\boldsymbol{\eta}\big(\mathbf{c}'(u),u\big)$ is in the globally identified subset for all $u$.*

Note that the curve $\boldsymbol{\eta}\big(\mathbf{c}'(u),u\big)$ is obtained by vectorizing the loadings $\mathbf{A}\big(\mathbf{c}'(u)\big), \mathbf{B}(\mathbf{c}'(u))$ and $\mathbf{\Phi}\big(\mathbf{c}'(u)\big)$ as in (2.7) and the globally identified subset is as defined in Proposition 8. As the shared channel size and factor numbers are held constant at $n, r_0$ and $r_1$ respectively, Assumption (A4) implies a lower bound on the cohort size $C'(u) \geq C_{\min}$ for all $u$ for some $C_{\min}$ which is a function of $n, r_0$ and $r_1$. With the first two of these assumptions, the following proposition can be proven.

**Proposition 16.** *Suppose $\{\mathbf{f}_{\mathcal{T}}(t)\}_{\mathcal{T}=1}^{\infty}$ is a locally stationary sequence of MV-GPs on $\{1,\ldots,\mathcal{T}\}$, and $\{\mathbf{g}_{\mathcal{T}}(c,t)\}_{\mathcal{T}=1}^{\infty}$, $\{\mathbf{w}_{\mathcal{T}}(c,t)\}_{\mathcal{T}=1}^{\infty}$ are locally stationary sequences of JMV-GPs on $\mathbb{N}\times\{1,\ldots,\mathcal{T}\}$. If $\{\mathbf{x}_{\mathcal{T}}(c,t)\}_{\mathcal{T}=1}^{\infty}$ is the sequence of JMV-GPs defined by* (4.13) *for $\mathbf{A}'(c,u), \mathbf{B}'(c,u)$ and $\mathbf{\Phi}'(c,u)$ satisfying (A1) and (A2), then $\{\mathbf{x}_{\mathcal{T}}(c,t)\}_{\mathcal{T}=1}^{\infty}$ is locally stationary.*

*Proof.* See Section D.2 in Appendix D for proof. □

The above proof requires that the latent processes be locally stationary, which depends on the properties of the covariance kernels $\{\kappa_{\mathcal{T},\text{time}}\big(t_1,t_2;\boldsymbol{\theta}'(u)\big)\}_{\mathcal{T}=1}^{\infty}$. If the $\{\kappa_{\mathcal{T},\text{time}}\big(t_1,t_2;\boldsymbol{\theta}'(u)\big)\}_{\mathcal{T}=1}^{\infty}$ arise from a model which is known to be locally stationary, such as the Gaussian tvARMA$(p,q)$ process (Dahlhaus, 2012, Ex. 4.2), then the latent processes will also be locally stationary.

Further, if $\mathbf{c}'(u)$ is a time-varying cohort vector with maximum size $C_{\max}$, then the max cohort vector $\mathbf{c}_{\max} \in \mathbb{N}^{C_{\max}}$ can be defined as the ascending list of indices where $c$ is listed in $\mathbf{c}_{\max}$ if there exists some $u$ such that $c$ is listed in $\mathbf{c}'(u)$. Define the selector matrix $\mathbf{Z}(u)$ as the $C'(u) \times C_{\max}$ block matrix associated with $\mathbf{c}'(u)$ by

$$\mathbf{D}_{ij}(u) = \begin{cases} \mathbf{I}_n & [\mathbf{c}'(u)]_i = [\mathbf{c}_{\max}]_j \\ \mathbf{0}_{n,n} & \text{o.w.} \end{cases},$$

$$[\mathbf{Z}(u)]_{ij} = \mathbf{D}_{ij}(u) \text{ for } i = 1,\ldots,C'(u),\ j = 1,\ldots,C_{\max}.$$

If $\mathbf{x}_{\mathcal{T}}(t)$ and $\widetilde{\mathbf{x}}_t(u)$ is the stacked process associated with the constant cohort $\mathbf{c}_{\max}$ and its stationary local approximations as obtained in the proof of Proposition 16, then $||\mathbf{Z}(t/\mathcal{T})\mathbf{x}_{\mathcal{T}}(t) - \mathbf{Z}(t/\mathcal{T})\widetilde{\mathbf{x}}_t(u)|| \leq ||\mathbf{x}_{\mathcal{T}}(t) - \widetilde{\mathbf{x}}_t(u)||$, so $||\mathbf{Z}(t/\mathcal{T})\mathbf{x}_{\mathcal{T}}(t) - \mathbf{Z}(t/\mathcal{T})\widetilde{\mathbf{x}}_t(u)||$ is $O_p\left(|t/\mathcal{T} - u| + \mathcal{T}^{-1}\right)$. For a time-varying cohort $\mathbf{c}'(u)$ with bounded size, the vector $\mathbf{Z}(t/\mathcal{T})\widetilde{\mathbf{x}}_t(u)$ is the stationary local approximation to $\mathbf{Z}(t/\mathcal{T})\mathbf{x}_{\mathcal{T}}(t)$. This is necessary because the difference $\mathbf{Z}(t/\mathcal{T})\mathbf{x}_{\mathcal{T}} - \mathbf{Z}(u)\widetilde{\mathbf{x}}_t(u)$ is only well-defined if $\mathbf{c}'(t/\mathcal{T}) = \mathbf{c}'(u)$.

## 4.5 Podding via Relative Configuration

To identify stable sub-groups of vehicles as they travel down the highway, we focus on the *relative configuration* of the vehicles. That is, only the motion of vehicles which alters their relative position and velocity are considered, and continued constant-speed down-highway movement is ignored. Given a cohort of vehicles with complete observations over a given window, the space of relative configurations is identified with a subspace of the trajectory space. Similarly, partitioning the cohort into pods creates a collection of subspaces which capture the relative configuration of vehicles within the pod. A loss function for poddings which measures the instability of the

within-pod relative configurations is constructed. Under the stationary separable covariance model of Section 4.4.2, a simple expression for the expected podding loss is obtained.

### 4.5.1 Overall Relative Configuration

Suppose a cohort of vehicles with index vector $\mathbf{c} = [c_1, \ldots, c_C]$ are observed at times $\mathbf{t} = [t_1, \ldots, t_T]$, with $n$ features measured for each vehicle. The cumulative changes in relative configuration can be obtained from this vector-valued time series. The space of relative configurations of the cohort can be identified as follows. On the trajectory space $\mathbb{R}_{Cn\times T}$ equipped with the Frobenius inner product, we can define the action of a all-vehicle translation $T_{\mathbf{V}}$ as

$$T_{\mathbf{V}}(\mathbf{X}\big(\mathbf{c},\mathbf{t}\big)) = \mathbf{X}(\mathbf{c},\mathbf{t}) + (\mathbf{1}_C \otimes \mathbf{V}) = \begin{bmatrix} \mathbf{x}(c_1,t_1)+\mathbf{v}_1 & \ldots & \mathbf{x}(c_1,t_T)+\mathbf{v}_T \\ \vdots & \ddots & \vdots \\ \mathbf{x}(c_C,t_1)+\mathbf{v}_1 & \ldots & \mathbf{x}(c_C,t_T)+\mathbf{v}_T \end{bmatrix} \tag{4.15}$$

where $\mathbf{V} \in \mathbb{R}_{n\times T}$ has columns $[\mathbf{v}_1 \ \ldots \ \mathbf{v}_T]$ and $\mathbf{1}_m \in \mathbb{R}_m$ is the vector of ones. Configurations of the same cohort are equivalent if they differ by some all-vehicle translation $T_{\mathbf{V}}$. The space of relative configurations of the cohort is defined as the quotient space under this equivalence relation, which can be identified with a particular subspace of $\mathbb{R}_{Cn\times T}$. Therefore, projecting onto the space of relative configurations removes the effect of the down-highway motion $\mathbf{V}$ which is applied uniformly to each vehicle.

**Proposition 17.** *The space of relative configurations of the cohort* $\mathbf{c}$ *at times* $\mathbf{t}$*, defined as the quotient space of* $\mathbb{R}_{Cn\times T}$ *under the above equivalence relation, is isomorphic to the* $(C-1)Tn$ *dimensional subspace* $C^{\perp} \subset \mathbb{R}_{Cn\times T}$*, where*

$$C = \mathrm{Span}\{\mathbf{1}_C \otimes \mathbf{e}_i \otimes \widetilde{\mathbf{e}}_j \mid i = 1,\ldots,n,\ j = 1,\ldots,T\}.$$

*Here,* $\mathbf{e}_i \in \mathbb{R}_n$ *and* $\widetilde{\mathbf{e}}_j \in \mathbb{R}_T$ *are the standard basis vectors of the appropriate sizes. The orthogonal projection* $P_{\mathcal{C}^{\perp}}(\mathbf{Y}) = \mathbf{Y} - \mathbf{1}_C \otimes \big(C^{-1}(\mathbf{1}_C^{\mathsf{T}} \otimes \mathbf{I}_n)\mathbf{Y}\big)$ *for* $\mathbf{Y} \in \mathbb{R}_{Cn\times T}$ *is equivalent to the quotient space projection.*

*Proof.* See proof of Proposition 18 with $\mathbf{m}_B = \mathbf{1}_C$. □

Given a subset $B$ of the cohort $\{c_1, \ldots, c_C\}$, the space of relative configurations of the vehicles in $B$ can be identified with an $(|B|-1)Tn$-dimensional subspace of $\mathbb{R}_{Cn\times T}$ in the same fashion. To see this, define the binary membership vector $\mathbf{m}_B \in \mathbb{R}_C$ which has a 1 in the $j$th component if $c_j \in B$ and zero otherwise. The space of all configurations of the vehicles in $B$ can be identified with the $|B|Tn$ dimensional subspace $S_B \subset \mathbb{R}_{Cn\times T}$ defined as the image of $\mathbb{R}_{Cn\times T}$ under left-multiplication by $\mathbf{P}_{S_B} = \mathrm{diag}(\mathbf{m}_B) \otimes \mathbf{I}_n$. As for the full cohort, the space of relative configurations of the *pod* whose members are given by $B$ is defined as the quotient space of $\mathbb{R}_{Cn\times T}$ under the equivalence relation arising from the transformation $\mathbf{T}_{\mathbf{V},B}\big(\mathbf{X}(\mathbf{c},\mathbf{t})\big) = \mathbf{P}_{S_B}\mathbf{X}(\mathbf{c},\mathbf{t}) + (\mathbf{m}_B \otimes \mathbf{V})$.

**Proposition 18.** *For any $B \subset \{c_1, \ldots, c_C\}$ at times $\mathbf{t}$, the space of relative configurations of the pod with members in $B$ can be identified with the $(|B|-1)Tn$ dimensional subspace $C_B^{\perp} \cap S_B$, where*

$$C_B = \mathrm{Span}\{\mathbf{m}_B \otimes \mathbf{e}_{i:n} \otimes \mathbf{e}_{j:T} \mid i = 1, \ldots, n,\ j = 1, \ldots, T\}. \tag{4.16}$$

*The orthogonal projection $P_B$ onto $C_B^{\perp} \cap S_B$ is $P_B(\mathbf{Y}) \equiv \mathbf{P}_{S_B}\mathbf{Y} - \mathbf{m}_B \otimes \big(|B|^{-1}(\mathbf{m}_B^{\mathsf{T}} \otimes \mathbf{I}_n)\mathbf{Y}\big)$.*

*Proof.* See Section D.3 in Appendix D for proof. □

If B is a singleton, then $C_B^{\perp} \cap S_B = \{\mathbf{0}\}$ and all trajectories yield the same trivial relative configuration for the vehicle in pod $B$.

### 4.5.2 Podding Loss

A podding $\mathcal{P}$ can be defined as a partition of the cohort $\mathbf{c}$ into the subsets $B_1, \ldots B_K$, with each $B_k$ designating a pod of vehicles. For a stable podding, the within-pod relative configurations should be close to constant. To measure this, first define the time-average configuration $\overline{\mathbf{x}}(\mathbf{c},\mathbf{t})$ of the vehicles in $\mathbf{c}$ over the times in $\mathbf{t}$ as

$$\overline{\mathbf{x}}(\mathbf{c},\mathbf{t}) = T^{-1}\mathbf{X}(\mathbf{c},\mathbf{t})\mathbf{1}_T. \tag{4.17}$$

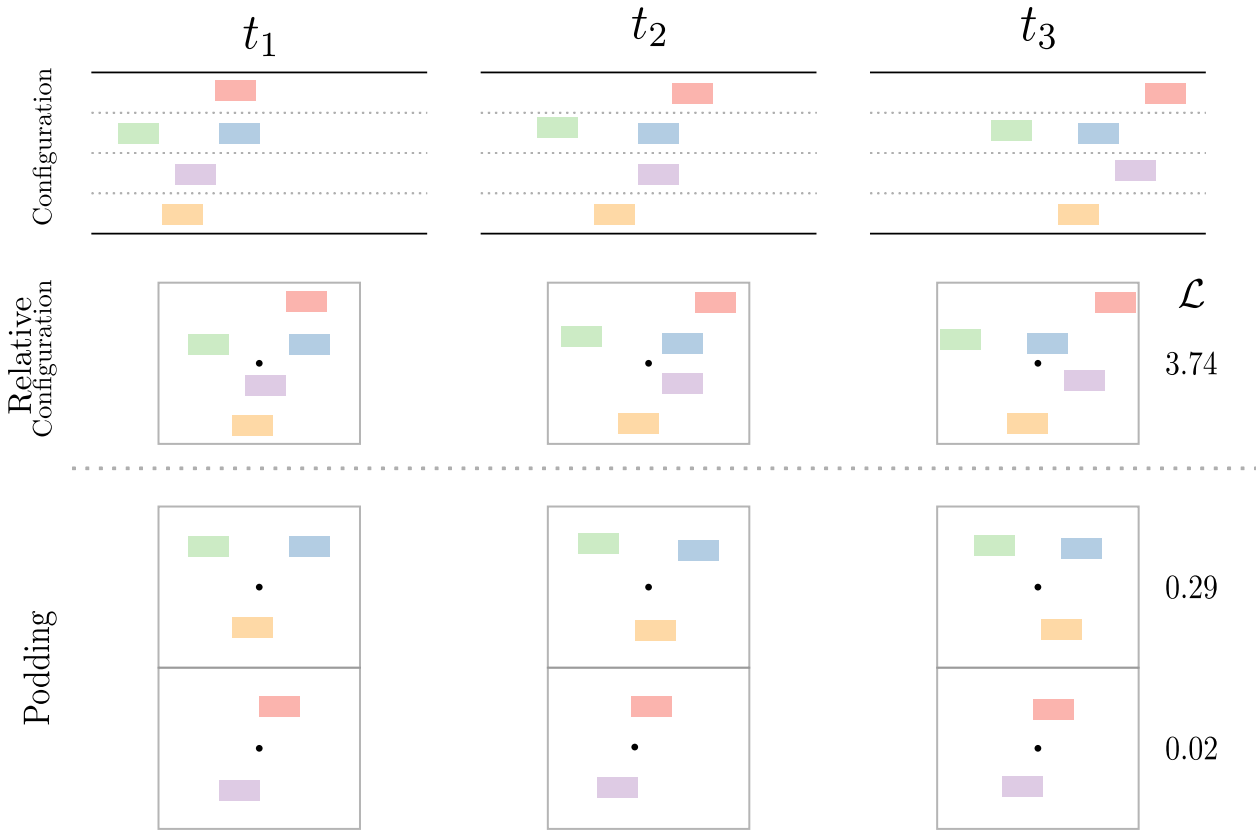



**Figure 4.3:** Example of podding decomposition for $C = 5$ vehicles and $T = 3$ time points with $n = 2$ features (Along- and across-highway position). Top row depicts absolute configuration for the vehicle positions on highway segment at $t_1, t_2, t_3$. The second row shows relative configuration of the overall cohort at each of the three time-points, while the third row shows the within-pod relative configurations for the optimal 2-part podding. The central black dot in each relative configuration box is the origin of the configuration space. The podding loss $\mathcal{L}$ from (4.18) corresponding to each pod is given in the third column.

The deviation from the time-average configuration when measured on the associated relative configuration subspaces captures the appropriate notion of within-pod stability. For the measurement matrix $\mathbf{X}(\mathbf{c},\mathbf{t})$ and a podding $\mathcal{P}$ for the cohort of vehicles $\mathbf{c}$ measured at times $\mathbf{t}$, define the *podding loss* $\mathcal{L}\big(\mathcal{P},\mathbf{X}(\mathbf{c},\mathbf{t})\big)$ as the (scaled) squared magnitude of the deviation from the time-average configuration, $\mathbf{X}(\mathbf{c},\mathbf{t}) - (\overline{\mathbf{x}}(\mathbf{c},\mathbf{t}) \otimes \mathbf{1}_T^\mathsf{T})$ when projected onto the subspace $(C_{B_1}^\perp \cap S_{B_1}) \boxplus \cdots \boxplus (C_{B_K}^\perp \cap S_{B_K})$ of the relative configurations of the pods,

$$\mathcal{L}\big(\mathcal{P},\mathbf{X}(\mathbf{c},\mathbf{t})\big) = \frac{1}{T-1}\sum_{B\in\mathcal{P}} ||P_B\big(\mathbf{X}(\mathbf{c},\mathbf{t}) - \overline{\mathbf{x}}(\mathbf{c},\mathbf{t}) \otimes \mathbf{1}_T^\mathsf{T}\big)||_F^2. \tag{4.18}$$

An example of the relation between a vehicle podding and the podding loss is shown in Figure 4.3.

The loss function $\mathcal{L}$ measures the per-time squared average deviation from a static within-pod relative configuration. A loss of zero is attained only if the podding breaks the cohort $\mathcal{C}$ into pods where the relative configuration of the vehicles is unchanging during the times $\{t_1,\ldots,t_T\}$, which is a "perfect" podding. The above podding loss is non-increasing as partition fineness increases, and the trivial podding $\mathcal{P}_{\min}$ which assigns each vehicle its own pod will always achieve a loss of zero. To avoid this, we can impose a penalty for finer poddings. In the same setting as $\mathcal{L}$ above, define the penalized podding loss $\mathcal{L}_\lambda\big(\mathcal{P},\mathbf{X}(\mathbf{c},\mathbf{t})\big)$ with penalty weight $\lambda \geq 0$ as

$$\mathcal{L}_\lambda\big(\mathcal{P},\mathbf{X}(\mathbf{c},\mathbf{t})\big) = \mathcal{L}\big(\mathcal{P},\mathbf{X}(\mathbf{c},\mathbf{t})\big) + \lambda|\mathcal{P}|. \tag{4.19}$$

If $\mathbf{S}(\mathbf{c},\mathbf{t})$ is the sample lag-zero covariance matrix for the cohort $\mathbf{c}$ at times $\mathbf{t}$,

$$\mathbf{S}(\mathbf{c},\mathbf{t}) = (T-1)^{-1}(\mathbf{X}(\mathbf{c},\mathbf{t}) - \overline{\mathbf{x}} \otimes \mathbf{1}_T^\mathsf{T})(\mathbf{X}(\mathbf{c},\mathbf{t}) - \overline{\mathbf{x}} \otimes \mathbf{1}_T^\mathsf{T})^\mathsf{T},$$

then some algebra shows that $\mathcal{L}\big(\mathcal{P},\mathbf{X}(\mathbf{c},\mathbf{t})\big)$ can be written in terms of $\mathbf{S}(\mathbf{c},\mathbf{t})$ as

$$\begin{aligned}\mathcal{L}\big(\mathcal{P},\mathbf{S}(\mathbf{c},\mathbf{t})\big) = \sum_{B\in\mathcal{P}} \operatorname{tr}\mathbf{P}_{S_B}\mathbf{S}(\mathbf{c},\mathbf{t})\mathbf{P}_{S_B} - \frac{2}{|B|}\operatorname{tr}\left(\mathbf{m}_B^\mathsf{T} \otimes \left(\mathbf{P}_{S_B}\mathbf{S}(\mathbf{c},\mathbf{t})\mathbf{P}_{S_B}(\mathbf{m}_B \otimes \mathbf{I}_n)\right)\right) \\ + \frac{1}{|B|}\operatorname{tr}\mathbf{P}_{S_B}\mathbf{S}(\mathbf{c},\mathbf{t})\mathbf{P}_{S_B}(\mathbf{m}_B\mathbf{m}_B^\mathsf{T} \otimes \mathbf{I}_n).\end{aligned}$$

If $\mathbf{S}(\mathbf{c},\mathbf{t})$ is partitioned into the $n \times n$ blocks corresponding to each pair of vehicles,

$$\mathbf{S}(\mathbf{c},\mathbf{t}) = \begin{bmatrix} \mathbf{S}_{11}(\mathbf{c},\mathbf{t}) & \dots & \mathbf{S}_{C1}(\mathbf{c},\mathbf{t}) \\ \vdots & \ddots & \vdots \\ \mathbf{S}_{C1}^{\mathsf{T}}(\mathbf{c},\mathbf{t}) & \dots & \mathbf{S}_{CC}(\mathbf{c},\mathbf{t}), \end{bmatrix}$$

then the above expression can be simplified into

$$\mathcal{L}\big(\mathcal{P},\mathbf{S}(\mathbf{c},\mathbf{t})\big) = \sum_{B\in\mathcal{P}} \left[ \sum_{c\in B} \frac{|B|-1}{|B|} \operatorname{tr} \mathbf{S}_{cc}(\mathbf{c},\mathbf{t}) - \frac{1}{|B|} \sum_{\substack{c,c'\in B \\ c\neq c'}} \operatorname{tr} \mathbf{S}_{cc'}(\mathbf{c},\mathbf{t}) \right]. \tag{4.20}$$

Under the stationary MFA model (4.5), the cross-sectional covariance is channel-structured and so the expected podding loss $E[\mathcal{L}\big(\mathcal{P},\mathbf{S}(\mathbf{c},\mathbf{t})\big)]$ for $\mathbf{X}(\mathbf{c},\mathbf{t})$ distributed as (4.3) is

$$E[\mathcal{L}\big(\mathcal{P},\mathbf{S}(\mathbf{c},\mathbf{t})\big)] = \sum_{B\in\mathcal{P}} \left[ \frac{1}{2|B|} \sum_{c\in B} \sum_{c'\in B} ||\mathbf{A}(c) - \mathbf{A}(c')||_F^2 + \frac{|B|-1}{|B|} \sum_{c\in B} ||\mathbf{B}(c)||_F^2 + ||\mathbf{\Phi}(c)||_F \right]. \tag{4.21}$$

In Section 4.6.3, the above expression for the expected podding loss for cohort $\mathbf{c}$ under the stationary model (4.5) is used to construct a dynamic podding rule for the locally stationary model using a related kernel distance measure.

## 4.6 Estimation and Podding

Given a time-varying cohort of vehicles and their associated partially-overlapping trajectories, our goal is to dynamically cluster the cohort into pods. The model-based dynamic podding procedure outlined in Section 4.5 requires estimating the vehicle-specific loading parameters on the basis of the trajectories of this time-varying cohort $\mathbf{c}_{\mathcal{T}}(t)$, while allowing for these loadings to slowly vary. This is accomplished by taking small windows of length $W$ over which the cohort is static and completely observed and applying a stationary psuedo-likelihood maximization algorithm (Section 4.6.1). These estimates are stitched together into an estimate of the time-varying loading curves (Section 4.6.2) and this is shown to be a consistent estimation procedure under infill asymptotics. Using these curves, a dynamic podding estimator is obtained by selecting the optimal podding for

the stationary approximation to the time-varying factor decomposition (Section 4.6.3) at each time $t$.

### 4.6.1 Estimation of Stationary Approximations

For vehicles $\mathbf{c} = [c_1, \ldots, c_C]$ with $n$ features each, observed at times $\mathbf{s} \equiv [s_1, \ldots, s_W]$, the $Cn \times W$ windowed data $\mathbf{X}(\mathbf{c}, \mathbf{s})$ is modeled as a sample from a separable stationary JMV-GP with MFA decomposition, (4.5). The mean matrix $\mathbf{M}(c)$ is estimated by the time-average configuration $\overline{\mathbf{x}}(\mathbf{c}, \mathbf{s}) \otimes \mathbf{1}_W^\mathsf{T}$ as in Section (4.5). The remaining parameters to be estimated are the cross-sectional covariance parameters $\mathbf{A}(\mathbf{c}), \mathbf{B}(\mathbf{c}), \mathbf{\Phi}(\mathbf{c})$ and the temporal covariance parameters $\boldsymbol{\theta}$. Estimates of these parameters are obtained by maximizing the stationary pseudo-log-likelihood,

$$\begin{aligned}\widetilde{\ell}\big(\mathbf{X}(\mathbf{c}, \mathbf{s}); \mathbf{A}, \mathbf{B}, \mathbf{\Phi}, \boldsymbol{\theta}\big) = &- \operatorname{tr} \mathbf{U}^{-1}(\boldsymbol{\theta}) \mathbf{Z}^\mathsf{T} \mathbf{R}^{-1}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}) \mathbf{Z} \\ &- W \log\det \mathbf{R}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}) - Cn \log\det \mathbf{U}(\boldsymbol{\theta}),\end{aligned} \tag{4.22}$$

plus terms which do not depend on the unknown parameters, where

$$\begin{aligned}\mathbf{Z} &= \mathbf{X}(\mathbf{c}, \mathbf{s}) - \overline{\mathbf{x}}(\mathbf{c}, \mathbf{s}) \otimes \mathbf{1}_W^\mathsf{T}, \\ \mathbf{R}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}) &= \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T} + \mathbf{\Phi}.\end{aligned}$$

For the cross-sectional parameter spaces $\mathbb{A}_L^*, \mathbb{B}_L^*$ and $\mathbb{P}_n$ are as defined in Chapter 2, then the optimization problem on the given window is

$$\max_{(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}, \boldsymbol{\theta}) \in \mathbb{A}_L^* \times \mathbb{B}_L^* \times \mathbb{P}_n \times \mathbf{\Theta}} \widetilde{\ell}\big(\mathbf{X}(\mathbf{c}, \mathbf{s}); \mathbf{A}, \mathbf{B}, \mathbf{\Phi}, \boldsymbol{\theta}\big). \tag{4.23}$$

Estimation of the scale matrices of the matrix normal likelihood can be accomplished via a *flip-flop* algorithm (Srivastava et al., 2007). Numeric experiments by Werner et al. (2008) have demonstrated that flip-flop methods for estimating covariance matrices with Kronecker structure tend to converge to a global optimum more rapidly than Newton-Raphson methods and have improved asymptotic efficiency over projection of the sample covariance onto the space of covariance matrices with Kronecker product structure. Algorithm 2 solves the optimization problem (4.23) by sequentially updating the parameter blocks $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ and $\boldsymbol{\theta}$ using the existing MFA block

majorization-minimization approach (Ramírez et al., 2020, Alg. 1) and off-the-shelf methods for $\boldsymbol{\theta}$ respectively. The updates for $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$ and $\boldsymbol{\theta}$ are detailed in the following two subsections. The procedure is outlined in Algorithm 2.

**Algorithm 2:** Stationary estimation of parameters on single window

**Input:** Centered observation matrix $\mathbf{Z}$ for vehicles $c_1, \ldots, c_C$ at times $s_1 \ldots s_W$
**Output:** Estimated loading parameters $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\Phi}}$, estimated temporal covariance parameters $\widehat{\boldsymbol{\theta}}$

$s \leftarrow 0$
Initialize $\widehat{\mathbf{A}}^{(0)}, \widehat{\mathbf{B}}^{(0)}, \widehat{\boldsymbol{\Phi}}^{(0)}, \widehat{\boldsymbol{\theta}}^{(0)}$
**repeat**
  $\mathbf{S}(\widehat{\boldsymbol{\theta}}^{(s)}) \leftarrow \frac{1}{W-1}\left(\mathbf{Z}\mathbf{U}(\widehat{\boldsymbol{\theta}}^{(s)})^{-\frac{1}{2}}\right)\left(\mathbf{Z}\mathbf{U}(\widehat{\boldsymbol{\theta}}^{(s)})^{-\frac{1}{2}}\right)^{\mathsf{T}}$
  Update to $\widehat{\mathbf{A}}^{(s+1)}$ as (4.27) using $\mathbf{S}(\widehat{\boldsymbol{\theta}}^{(s)}), \widehat{\mathbf{B}}^{(s)}, \widehat{\boldsymbol{\Phi}}^{(s)}$
  Update to $\widehat{\mathbf{B}}^{(s+1)}$ as (4.29) using $\mathbf{S}(\widehat{\boldsymbol{\theta}}^{(s)}), \widehat{\mathbf{A}}^{(s+1)}, \widehat{\boldsymbol{\Phi}}^{(s)}$
  Update to $\widehat{\boldsymbol{\Phi}}^{(s+1)}$ as (4.31) using $\mathbf{S}(\widehat{\boldsymbol{\theta}}^{(s)}), \widehat{\mathbf{A}}^{(s+1)}, \widehat{\mathbf{B}}^{(s+1)}$
  $\widehat{\mathbf{R}}^{(s+1)} \leftarrow \widehat{\mathbf{A}}^{(s+1)}\widehat{\mathbf{A}}^{(s+1)\mathsf{T}} + \widehat{\mathbf{B}}^{(s+1)}\widehat{\mathbf{B}}^{(s+1)\mathsf{T}} + \widehat{\boldsymbol{\Phi}}^{(s+1)}$
  $\widetilde{\mathbf{Z}}(\widehat{\mathbf{R}}^{(s+1)}) \leftarrow (\widehat{\mathbf{R}}^{(s+1)})^{-\frac{1}{2}}\mathbf{Z}$
  Update to $\widehat{\boldsymbol{\theta}}^{(s+1)}$ as in 4.6.1 using $\widetilde{\mathbf{Z}}(\widehat{\mathbf{R}}^{(s+1)})$
**until** $\widehat{\mathbf{A}}^{(s)}, \widehat{\mathbf{B}}^{(s)}, \widehat{\boldsymbol{\Phi}}^{(s)}$ *and* $\widehat{\boldsymbol{\theta}}^{(s)}$ *converged*

**Updating** $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$

Holding the estimated temporal covariance parameters $\boldsymbol{\theta}$ fixed at $\widehat{\boldsymbol{\theta}}^{(s)}$, the profile optimization problem for $\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi}$ is

$$\max_{(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})\in\mathbb{A}_L^*\times\mathbb{B}_L^*\times\mathbb{P}_n} -\operatorname{tr}\left(\mathbf{U}(\widehat{\boldsymbol{\theta}}^{(s)})^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})\mathbf{Z}\right) - W\log\det\mathbf{R}(\mathbf{A},\mathbf{B},\boldsymbol{\Phi}). \tag{4.24}$$

If the temporally-whitened covariance $\mathbf{S}(\boldsymbol{\theta})$ is defined to be

$$\mathbf{S}(\boldsymbol{\theta}) = \frac{1}{W-1}\left(\mathbf{Z}\mathbf{U}(\boldsymbol{\theta})^{-1/2}\right)\left(\mathbf{Z}\mathbf{U}(\boldsymbol{\theta})^{-1/2}\right)^{\mathsf{T}}, \tag{4.25}$$

then maximizing the above profile likelihood is equivalent to solving the problem

$$\max_{(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})\in\mathbb{A}_L^*\times\mathbb{B}_L^*\times\mathbb{P}_*} -\log\det\mathbf{R}(\mathbf{A},\mathbf{B},\boldsymbol{\Phi}) - \operatorname{tr}\left(\mathbf{R}^{-1}\mathbf{S}(\widehat{\boldsymbol{\theta}}^{(s)})\right). \tag{4.26}$$

under constraints for $\mathbf{A}, \mathbf{B}, \mathbf{\Phi}$ which are equivalent to that of Ramírez et al. (2020, Alg. 1). Those authors provide a three-step block-minimization-majorization approach to identifying the maximizing $\mathbf{A}, \mathbf{B}, \mathbf{\Phi}$. For completeness, those steps are summarized below.

**Common factor loading update**

Holding $\mathbf{B}$ and $\mathbf{\Phi}$ fixed at $\widehat{\mathbf{B}}^{(s)}$ and $\widehat{\mathbf{\Phi}}^{(s)}$, the maximizing $\widehat{\mathbf{A}}^{(s+1)}$ has the closed form

$$\widehat{\mathbf{A}}^{(s+1)} = (\mathbf{E}^{(s)})^{\frac{1}{2}} \mathbf{C}_0^{(s)} (\mathbf{D}_0^{(s)})^{\frac{1}{2}} \mathbf{Q}_0^{(s)}, \tag{4.27}$$

where $\mathbf{E}^{(s)} = \widehat{\mathbf{B}}^{(s)} \widehat{\mathbf{B}}^{(s)\mathsf{T}} + \widehat{\mathbf{\Phi}}^{(s)} \widehat{\mathbf{\Phi}}^{(s)\mathsf{T}}$ is the estimated error covariance, which is used to isolate the influence of the common factor as

$$\mathbf{E}^{-\frac{1}{2}} \mathbf{S}(\widehat{\boldsymbol{\theta}}^{(s)}) (\mathbf{E}^{(s)})^{-\frac{1}{2}} = \mathbf{C}_0^{(s)} \mathbf{\Lambda}_0^{(s)} \mathbf{C}_0^{(s)\mathsf{T}}. \tag{4.28}$$

The right-hand side is the eigendecomposition of the left, with $\mathbf{\Lambda}_0^{(s)} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{Cn})$ sorted in descending order and $\mathbf{C}_0^{(s)}$ orthogonal. The matrix $\mathbf{D}_0^{(s)} = \operatorname{diag}(d_1, \ldots, d_r, 0, \ldots, 0)$ with $d_i = \max(\lambda_i - 1, 0)$ is the truncated and thresholded matrix of eigenvalues. The orthogonal matrix $\mathbf{Q}_0^{(s)}$ is chosen to satisfy the constraints of $\mathbb{A}_L^*$.

**Vehicle-specific factor loading update**

Holding $\mathbf{A}$ and $\mathbf{\Phi}$ fixed at $\widehat{\mathbf{A}}^{(s+1)}$ and $\widehat{\mathbf{\Phi}}^{(s)}$, the updated estimate for $\widehat{\mathbf{B}}$ can be obtained by maximization of a minorizing function, which is equivalent to $C$ classical factor analyses applied separately to the individual vehicles. For vehicle $c_i$, the maximizing $\widehat{\mathbf{B}}_i^{(s+1)}$ is obtained in the same fashion as in step 1, with

$$\widehat{\mathbf{B}}_i^{(s+1)} = \left(\mathbf{\Phi}_i^{(s+1)}\right)^{-\frac{1}{2}} \mathbf{C}_i^{(s)} \left(\mathbf{D}_i^{(s)}\right)^{-\frac{1}{2}} \mathbf{Q}_i^{(s)}, \tag{4.29}$$

where the eigendecomposition of the appropriate whitened vehicle covariance block is

$$\left(\mathbf{\Phi}_i^{(s)}\right)^{-\frac{1}{2}} \left(\mathbf{S}_{ii}(\widehat{\boldsymbol{\theta}}^{(s)}) - \widehat{\mathbf{A}}_i^{(s+1)} \widehat{\mathbf{A}}_i^{(s+1)\mathsf{T}}\right) \left(\mathbf{\Phi}_i^{(s)}\right)^{-\frac{1}{2}} = \mathbf{C}_i^{(s)} \mathbf{\Lambda}_i^{(s)} \mathbf{C}_i^{(s)\mathsf{T}}, \tag{4.30}$$

and $\mathbf{S}_{ii}(\widehat{\boldsymbol{\theta}}^{(s)})$ is the $i$th $n \times n$ block on the diagonal of $\mathbf{S}(\widehat{\boldsymbol{\theta}}^{(s)})$. Blocks $\mathbf{A}_i, \mathbf{B}_i, \boldsymbol{\Phi}_i$ are blocks of their respective matrices for vehicle $c_i$. The matrix $\mathbf{D}_i^{(s)}$ is obtained from thresholding of the eigenvalues as in the step above, and $\mathbf{Q}_i^{(s)}$ enforces the constraints of $\mathbb{B}_L^*$.

**Idiosyncratic variance update**

Holding $\mathbf{A}$ and $\mathbf{B}$ fixed at $\widehat{\mathbf{A}}^{(s+1)}$ and $\widehat{\mathbf{B}}^{(s+1)}$, the updated estimate for $\boldsymbol{\Phi}$ is obtained by assigning the residual sample variance to $\boldsymbol{\Phi}$, via

$$\widehat{\boldsymbol{\Phi}}^{(s+1)} = \operatorname{diag}\left(\mathbf{S}(\widehat{\boldsymbol{\theta}}^{(s)}) - \widehat{\mathbf{A}}^{(s+1)}\widehat{\mathbf{A}}^{(s+1)\mathsf{T}} - \widehat{\mathbf{B}}^{(s+1)}\widehat{\mathbf{B}}^{(s+1)\mathsf{T}}\right). \tag{4.31}$$

The updated estimates $\widehat{\mathbf{A}}^{(s+1)}$ and $\widehat{\mathbf{B}}^{(s+1)}$ ensure that the right-hand side has exclusively non-negative entries.

**Updating $\boldsymbol{\theta}$**

Holding the cross-sectional covariance parameters $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$ fixed at $(\widehat{\mathbf{A}}^{(s)}, \widehat{\mathbf{B}}^{(s)}, \widehat{\boldsymbol{\Phi}}^{(s)})$ with associated cross-sectional covariance matrix $\widehat{\mathbf{R}}^{(s)} = \widehat{\mathbf{A}}^{(s)}\widehat{\mathbf{A}}^{(s)\mathsf{T}} + \widehat{\mathbf{B}}^{(s)}\widehat{\mathbf{B}}^{(s)\mathsf{T}} + \widehat{\boldsymbol{\Phi}}^{(s)}$, the profile optimization problem for $\boldsymbol{\theta}$ is

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} -\operatorname{tr}\left(\mathbf{U}(\boldsymbol{\theta})^{-1}\mathbf{Z}^{\mathsf{T}}(\widehat{\mathbf{R}}^{(s)})^{-1}\mathbf{Z}\right) - Cn\log\det\mathbf{U}(\boldsymbol{\theta}). \tag{4.32}$$

Let the spatially-whitened data $\widetilde{\mathbf{Z}}(\mathbf{R})$ be defined as

$$\widetilde{\mathbf{Z}}(\mathbf{R}) = \mathbf{R}^{-\frac{1}{2}}\mathbf{Z}, \tag{4.33}$$

and let $\widetilde{\mathbf{z}}_i$ be the $i$th row of $\widetilde{\mathbf{Z}}\big(\widehat{\mathbf{R}}^{(s)}\big)$. The profile optimization problem then is

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \quad -\sum_{i=1}^{Cn}\widetilde{\mathbf{r}}_i^{\mathsf{T}}\mathbf{U}^{-1}(\boldsymbol{\theta})\widetilde{\mathbf{r}}_i - Cn\log\det\mathbf{U}(\boldsymbol{\theta}).$$

This optimization problem is the same as that solved by a maximum likelihood estimation routine for $Cn$ independent replicates of a zero-mean Gaussian time series with covariance model $\mathbf{U}(\boldsymbol{\theta})$. Depending on the choice of temporal correlation model $\mathbf{U}(\boldsymbol{\theta})$, off-the-shelf methods for Gaussian maximum likelihood estimation of $\boldsymbol{\theta}$ may exist. For example, if $\kappa_{\text{time}}(|t_1 - t_2|; \boldsymbol{\theta})$ is the autoco-

variance function of a ARMA model, the parameter vector can be estimated by the procedure of Brockwell and Davis (1991, Ch. 8.7).

### 4.6.2 Time-Varying Parameter Estimation via Moving Segments

An elementary and intuitive method for handling time-varying parameters $\alpha(t)$ in locally stationary models is the method of moving segments (Allen and Rabiner, 1977; Dahlhaus, 1996a; Chiann and Morettin, 2005), where small windows of the data are taken around each $t$. A method for estimating the parameter of interest for the stationary time series is then applied separately to each window, and the resulting estimate $\widehat{\boldsymbol{\alpha}}_{\mathcal{T}}(t/\mathcal{T})$ is assigned to the midpoint of the window. Interpolating these estimates in rescaled-time produces a curve $\widehat{\boldsymbol{\alpha}}_{\mathcal{T}}(u)$ which serves as a estimator of the underlying time-varying parameter curve $\boldsymbol{\alpha}(u)$. If the parameter curve $\boldsymbol{\alpha}(u)$ is sufficiently smooth and consistency of the stationary estimators $\widehat{\boldsymbol{\alpha}}_{\mathcal{T}}(t/\mathcal{T})$ holds as the window size increases, the moving-window estimator $\widehat{\boldsymbol{\alpha}}_{\mathcal{T}}(u)$ will typically converge to $\boldsymbol{\alpha}(u)$ under infill asymptotics as $\mathcal{T} \to \infty$.

This approach can be applied to estimation of the rescaled-time parameter curves $\mathbf{A}'(c,u)$, $\mathbf{B}'(c,u)$ and $\boldsymbol{\Phi}'(c,u)$ as in (4.10). Let $\mathbf{c}_{\mathcal{T}}(t)$ be a potentially time-varying vector of vehicle indices with $\max_{t\in\{1,\ldots,\mathcal{T}\}} \mathbf{c}_{\mathcal{T}}(t) = C_{\max} < \infty$, which is assumed to be obtained from a rescaled-time parameter curve $\mathbf{c}'(u)$ by $\mathbf{c}_{\mathcal{T}}(t) = \mathbf{c}'(t/\mathcal{T})$. For each time window centered at $t$, let $\underline{s}_{\mathcal{T}}(t)$ be the smallest time in $\{1,\ldots,\mathcal{T}\}$ such that $\mathbf{c}_{\mathcal{T}}(s)$ equals $\mathbf{c}_{\mathcal{T}}(t)$ for all $s \in \{\underline{s}_{\mathcal{T}}(t),\ldots,t\}$. Similarly define $\overline{s}_{\mathcal{T}}(t)$ as the largest time in $\{1,\ldots,\mathcal{T}\}$ such that $\mathbf{c}_{\mathcal{T}}(s)$ equals $\mathbf{c}_{\mathcal{T}}(t)$ for all $s \in \{t,\ldots,\overline{s}(t)\}$. With these definitions, let the vector of windowed times $\mathbf{s}_{\mathcal{T}}(t)$ be $[\min\{\underline{s}(t),(t-\frac{W}{2}+1)\},\ldots,\max\{\overline{s}(t),(t+\frac{W}{2})\}]^{\mathsf{T}}$. As $\mathbf{c}_{\mathcal{T}}(t)$ is constant on the times $\mathbf{s}_{\mathcal{T}}(t)$, the cohort $\mathbf{c}_{\mathcal{T}}(s)$ equals $\mathbf{c}_{\mathcal{T}}(t)$ for all times $s$ in $\mathbf{s}_{\mathcal{T}}(t)$ and so yields a rectangular index set, allowing the construction of the windowed dataset $\mathbf{X}\big(\mathbf{c}_{\mathcal{T}}(t),\mathbf{s}_{\mathcal{T}}(t)\big)$.

On each windowed dataset, the estimation procedure (2) based on likelihood maximization for the stationary approximation can be applied to obtain the estimates $\widehat{\mathbf{A}}(c),\widehat{\mathbf{B}}(c),\widehat{\boldsymbol{\Phi}}(c)$ and $\boldsymbol{\theta}$ for $c$ in $\mathbf{c}_{\mathcal{T}}(t)$. The window estimates $\widehat{\mathbf{A}}(c),\widehat{\mathbf{B}}(c),\widehat{\boldsymbol{\Phi}}(c)$ and $\widehat{\boldsymbol{\theta}}$ for $c$ in $\mathbf{c}_{\mathcal{T}}(t)$ obtained using the windowed data $\mathbf{X}\big(\mathbf{c}_{\mathcal{T}}(t),\mathbf{s}_{mathcalT}(t)\big)$ centered at time $t$ are assigned to $\widehat{\mathbf{A}}'(c,t/\mathcal{T}),\widehat{\mathbf{B}}'(c,t/\mathcal{T}),\widehat{\boldsymbol{\Phi}}'(c,t/\mathcal{T})$

and $\widehat{\boldsymbol{\theta}}'(t/\mathcal{T})$ respectively. The overall estimated parameter curves $\widehat{\mathbf{A}}'(c,u), \widehat{\mathbf{B}}'(c,u), \widehat{\boldsymbol{\Phi}}'(c,u)$ and $\widehat{\boldsymbol{\theta}}'(u)$ are obtained by interpolating these window estimates, which are obtained the on the grid $\frac{1}{\mathcal{T}}, \ldots, \frac{\mathcal{T}-1}{\mathcal{T}}, 1$ in the rescaled-time interval $[0,1]$. That is, $\widehat{\mathbf{A}}'(c,u)$ equals $\widehat{\mathbf{A}}'(c, \lceil \mathcal{T}u \rceil/\mathcal{T})$ for $u > 0$ and $\widehat{\mathbf{A}}'(c,0)$ is set to $\widehat{\mathbf{A}}'(c, 1/\mathcal{T})$. The other curves are interpolated in the same fashion.

The following theorem demonstrates that, under appropriate conditions on the cross-sectional and temporal covariance parameter curves, the moving-window estimation procedure will be consistent for the parameter curve segments of a specified time-varying collection of vehicles. The needed assumptions on the temporal covariance are specified in the following Assumption.

**Assumption 2.** *(Temporal) The covariance kernels* $\{\kappa_{\mathcal{T},\text{time}}\big(t_1, t_2; \boldsymbol{\theta}'(u)\big)\}_{\mathcal{T}=1}^{\infty}$*, parameter curve* $\boldsymbol{\theta}'(u)$*, and parameter space* $\boldsymbol{\Theta}$ *satisfy:*

*(B1) The temporal covariance parameter curve* $\boldsymbol{\theta}'(u)$ *is Lipschitz continuous.*

*(B2) If* $\boldsymbol{\theta}'(u)$ *is the constant value* $\boldsymbol{\theta}_0$*, then* $\kappa_{\mathcal{T},\text{time}}\big(t_1, t_2; \boldsymbol{\theta}'(u)\big)$ *equals* $k(|t_1 - t_2|; \boldsymbol{\theta}_0)$ *for all* $t_1, t_2$ *and* $\mathcal{T}$*, where* $k(|j|; \boldsymbol{\theta})$ *is some absolutely summable autocovariance function that is continuous in* $\boldsymbol{\theta}$ *and has unit variance* $k(0; \boldsymbol{\theta}) = 1$*.*

*(B3) For any* $t \in \{1, \ldots, \mathcal{T}\}$ *and all* $-t + 1 \leq j \leq \mathcal{T} - t$*,*

$$\kappa_{\mathcal{T},\text{time}}\big(t, t+j; \boldsymbol{\theta}'(u)\big) = k(|j|; \boldsymbol{\theta}_{t,\mathcal{T}}) + O(\mathcal{T}^{-1}),$$

*uniformly in* $t$ *and* $j$*, where* $\boldsymbol{\theta}_{t,\mathcal{T}} = \boldsymbol{\theta}'(t/\mathcal{T})$*.*

*(B4) The parameter space and temporal covariance kernels are such that, for any distinct times* $t_1, \ldots, t_T$ *in* $\{1, \ldots, \mathcal{T}\}$*, the matrix*

$$\mathbf{K}_{\mathcal{T}}\big(\mathbf{t}; \boldsymbol{\theta}'(u)\big) = \begin{bmatrix} \kappa_{\mathcal{T},\text{time}}\big(t_1, t_t; \boldsymbol{\theta}'(u)\big) & \ldots & \kappa_{\mathcal{T},\text{time}}\big(t_1, t_T; \boldsymbol{\theta}'(u)\big) \\ \vdots & \ddots & \vdots \\ \kappa_{\mathcal{T},\text{time}}\big(t_T, t_t; \boldsymbol{\theta}'(u)\big) & \ldots & \kappa_{\mathcal{T},\text{time}}\big(t_T, t_T; \boldsymbol{\theta}'(u)\big) \end{bmatrix},$$

*has* $\lambda_{\min}\Big(\mathbf{K}_{\mathcal{T}}(\mathbf{t}; \boldsymbol{\theta}'(u))\Big) \geq \epsilon$ *for some* $\epsilon > 0$ *that does not depend on* $\mathcal{T}$ *or* $\boldsymbol{\theta}'(u)$*.*

*(B5) The parameter space $\boldsymbol{\Theta}$ is compact and the estimator $\widehat{\boldsymbol{\theta}}$ obtained by minimizing the stationary log-likelihood* (4.22) *for known $\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi}$ is consistent as $W \to \infty$.*

The central local stationarity assumption (B3) is shown to hold in Dahlhaus (2012, Sec. 4) for a class of linear locally stationary processes. Under Assumptions 1 and 2 on the cross-sectional and temporal parameter curves respectively, application of a general theorem (Dahlhaus, 2012, Thm. 3.2) for locally stationary processes shows that the parameter curves estimated by the method of moving segments are consistent estimators of the relevant portions of the true curves.

**Theorem 6.** *(Consistency) Let $\{\mathbf{x}_{\mathcal{T}}(c,t)\}_{\mathcal{T}=1}^{\infty}$ be the sequence of JMV-GPs on $\mathbb{N} \times \{1, \ldots, \mathcal{T}\}$ defined by* (4.13) *for $\mathbf{A}'(c,u), \mathbf{B}'(c,u), \boldsymbol{\Phi}'(c,u)$ satisfying Assumptions (A1)-(A4) using the MV-GP sequence $\{\mathbf{f}_{\mathcal{T}}(t)\}_{\mathcal{T}=1}^{\infty}$ and JVM-GP sequences $\{\mathbf{g}_{\mathcal{T}}(c,t)\}_{\mathcal{T}=1}^{\infty}$ and $\{\mathbf{w}_{\mathcal{T}}(c,t)\}_{\mathcal{T}=1}^{\infty}$ as defined in Section 4.4.3 with temporal covariances $\{\kappa_{\mathcal{T},\text{time}}\big(t_1, t_2; \boldsymbol{\theta}'(u)\big)\}_{\mathcal{T}=1}^{\infty}$ and $\boldsymbol{\theta}'(u)$ satisfying (B1) - (B5). If the estimated parameter curves $\widehat{\mathbf{A}}'(c,u)$, $\widehat{\mathbf{B}}'(c,u)$, $\widehat{\boldsymbol{\Phi}}'(c,u)$ and $\widehat{\boldsymbol{\theta}}'(u)$ are obtained by the method of movement segments, as specified in Section 4.6.2, for windows $\mathbf{s}_{\mathcal{T}}(t)$ whose size $W(\mathcal{T})$ is such that $W(\mathcal{T}) \to \infty$ and $W(\mathcal{T})/\mathcal{T} \to 0$ as $\mathcal{T} \to \infty$ and for a cohort curve $\mathbf{c}'(u)$ with bounded size and which is right-continuous with a finite number of points of discontinuity in $(0,1)$, then for any $u_0 \in (0,1)$ the estimates $\widehat{\mathbf{A}}'(c,u_0)$, $\widehat{\mathbf{B}}'(c,u_0)$, $\widehat{\boldsymbol{\Phi}}'(c,u_0)$ and $\widehat{\boldsymbol{\theta}}'(u_0)$ converge in probability to the true values $\mathbf{A}'(c,u_0), \mathbf{B}'(c,u_0), \boldsymbol{\Phi}'(c,u_0)$ and $\boldsymbol{\theta}'(u_0)$ for any $c$ in $\mathbf{c}'(u_0)$.*

*Proof.* See Section D.4 in Appendix D for proof. □

### 4.6.3 Dynamic Pod Estimation

As illustrated in Section 4.5, within the framework of the podding losses accrued for non-static relative configurations, a JMV-GP $\mathbf{x}(c,t)$ with stationary factor decomposition (4.5) has the expected podding loss (4.21) for cohort $\mathbf{c}$. The optimal podding $\mathcal{P}_k^*$ of $\mathbf{c}$ into $k$ pods for the stationary $\mathbf{x}(\mathbf{c}, \mathbf{t})$ is

$$\begin{aligned}\widetilde{\mathcal{P}}_k^*(\mathbf{c}) = \underset{|\mathcal{P}|=k}{\operatorname{argmin}} \sum_{B \in \mathcal{P}} \frac{1}{2|B|} \sum_{c_i, c_j \in B} ||\mathbf{A}(c_i) - \mathbf{A}(c_j)||^2 \\ + \sum_{B \in \mathcal{P}} \frac{|B|-1}{|B|} \sum_{c_i \in B} \left( ||\mathbf{B}(c_i)||^2 + ||\boldsymbol{\Phi}(c_i)|| \right).\end{aligned} \tag{4.34}$$

For any cohort of vehicles $\mathbf{c}$ and fixed pod number $k$, finding the optimal podding can be framed as kernel $k$-means problem under the kernel described below. This connects the podding problem with the large literature on clustering, and provides a springboard for efficient estimation by enabling the model-based podding procedure where the parameters $\mathbf{A}(c), \mathbf{B}(c)$ and $\mathbf{\Phi}(c)$ are first estimated using $\mathbf{X}(\mathbf{c}, \mathbf{t})$.

This approach extends to the time-varying factor decomposition, as within each window the estimated parameter curves $\widehat{\mathbf{A}}'(c,u), \widehat{\mathbf{B}}'(c,u)$ and $\widehat{\mathbf{\Phi}}'(c,u)$ are obtained by likelihood maximization for the stationary local approximation. We can define the dynamic podding rule $\mathcal{P}_k^*(\mathbf{c}_\mathcal{T}(t), t)$ by using the optimal podding rule for the local approximation at each time:

$$\mathcal{P}_k^*(\mathbf{c}_\mathcal{T}(t), t) = \widetilde{\mathcal{P}}_k^*\big(\mathbf{c}_\mathcal{T}(t)\big).$$

This completes the dynamic podding method, in that the time-varying factor decomposition is locally approximated by a stationary factor decomposition, which has an optimal podding under the expected podding loss. As the locally stationary approximation varies, so too does the optimal podding, leading to the above dynamic podding rule.

To estimate $\mathcal{P}_k^*(\mathbf{c}_\mathcal{T}(t), t)$, first note that the stationary podding problem (4.34) is equivalent to a kernel k-means problem with scalar kernel $k : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$,

$$k(c, c') = \begin{cases} \operatorname{tr} \mathbf{A}^\mathsf{T}(c)\mathbf{A}(c') & \text{if } c \neq c' \\ ||\mathbf{A}(c)||_F^2 + ||\mathbf{B}(c)||^2 + ||\mathbf{\Phi}(c)||_F & \text{if } c = c'. \end{cases}$$

This is easily seen to be a positive semi-definite kernel, as for any vector of vehicle indices $\mathbf{c} = [c_1, \ldots, c_C]^\mathsf{T}$, the associated $C \times C$ kernel matrix is the sum of the Gram matrix $\mathbf{V}^\mathsf{T}\mathbf{V}$ with $\mathbf{V} = [\operatorname{vec}(\mathbf{A}(c_1)) \ \ldots \ \operatorname{vec}(\mathbf{A}(c_C))]$ plus a non-negative diagonal matrix and is therefore positive semi-definite. In this formulation, $\widetilde{\mathcal{P}}_k^*(\mathbf{c})$ can be computed efficiently using the kernel k-means algorithm (Dhillon et al., 2004, Alg. 1). In the time-varying parameter setting, let $\widehat{k}(c, c'; t)$ be the above kernel computed using $\widehat{\mathbf{A}}'(c, t/\mathcal{T}), \widehat{\mathbf{B}}'(c, t/\mathcal{T})$ and $\widehat{\mathbf{\Phi}}'(c, t/\mathcal{T})$.

For the penalized loss (4.19), the estimation of the optimal partition $\mathcal{P}^*_\lambda(\mathbf{c}_\mathcal{T}(t), t)$ with the addition of a penalization term for the number of pods and penalty weight $\lambda$ can be accomplished via the kernel DP-means algorithm (Kulis and Jordan, 2012). The resulting optimization problem is minimization of the within-cluster pairwise discrepancies with the addition of a penalty of $\lambda$ times the number of clusters, over the space of all partitions. The resulting clustering algorithm is similar to $k$-means, with the addition that new cluster formation will occur when a given observation is more than $\lambda$ units from all other cluster centers. This flexibility is well suited to the dynamic podding problem, as an *a priori* fixed number of pods would be unrealistic. This dynamic podding with varying observation identities is obtained via Algorithm 3.

**Algorithm 3:** Model-based dynamic podding using DP-means

**Data:** Estimated loading parameters curves $\widehat{\mathbf{A}}'(c,u), \widehat{\mathbf{B}}'(c,u), \widehat{\mathbf{\Phi}}'(c,u)$, time-varying vehicles index vector $\mathbf{c}_\mathcal{T}(t)$, start and end times $t_{\text{start}}$ and $t_{\text{end}}$

**Input:** Pod count penalty $\lambda$

**Output:** $\widehat{\mathcal{P}}_\lambda(\mathbf{c}_\mathcal{T}(t), t)$, the dynamic podding of $\mathbf{c}_\mathcal{T}(t)$

**for** $t \leftarrow t_{\text{start}} \dots t_{\text{end}}$ **do**
- **if** $t = t_{\text{start}}$ **then**
  - $\mathcal{P} \leftarrow \{\mathbf{c}_\mathcal{T}(t)\}$ `// initialize to trivial podding`
- **else**
  - $\mathcal{P} \leftarrow \widehat{\mathcal{P}}_\lambda(\mathbf{c}_\mathcal{T}(t-1), t-1)$ `// start at previous step podding`
- **repeat**
  - **for** $i \leftarrow 1, \dots, \text{len}(\mathbf{c}(t))$ **do**
    - Compute $d_{i,B}$ for all pods $B \in \mathcal{P}$ as

$$d_{i,B} = \widehat{k}(c_i, c_i; t) + |B|^{-1} \sum_{c_j \in B} \widehat{k}(c_j, c_j; t) - 2\widehat{k}(c_i, c_j; t)$$

    - **if** $\min_B d_{i,B} > \lambda$ **then**
      - In $\mathcal{P}$, move $c_i$ to a new pod
    - **else**
      - In $\mathcal{P}$, move $c_i$ to pod $\text{argmin}_{B \in \mathcal{P}} d_{i,B}$
  - **until** $\mathcal{P}$ *does not change*
- $\widehat{\mathcal{P}}_\lambda(\mathbf{c}(t), t) \leftarrow \mathcal{P}$

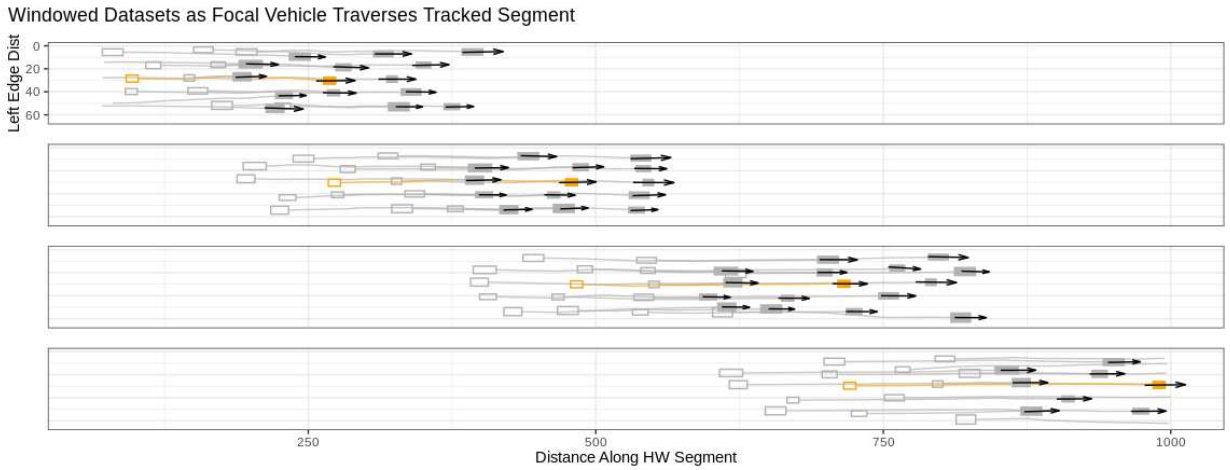



**Figure 4.4:** Example of how the trajectories of the cohort of the ego vehicle (in orange) can be divided into windows. Data are partitioned into five-second $(W = 50)$ windows with inclusion radius $\rho = 60$. The non-filled rectangle is initial vehicle position with filled being final in window, arrow magnitude and direction give velocity.

In the following section, we apply the above dynamic podding procedure to cohorts of spatially proximate vehicles centered around an *ego vehicle* and discuss the resulting podding decomposition.

## 4.7 Results

### 4.7.1 Data

For the purposes of illustrating the performance of the dynamic podding procedure based on the time-varying MFA model (4.10) as well as exploring the qualitative aspects of the resulting dynamic podding estimates, we use the NGSIM US-101 complete trajectory dataset as our primary data source (U.S Department of Transportation, Federal Highway Administration, 2016). The dataset consists of location and kinematic data for thousands of vehicles as they travelled along the US 101 highway, polled with a high sampling rate of $10\text{Hz}$. The trajectories were recorded along a $2100\text{ft}$ section of the highway during three $15$ minute periods, and the data encompass a variety of traffic conditions from free-flowing to heavily-congested. Vehicles occupy five primary southbound lanes and one auxiliary southbound lane between on- and off-ramps.

Measurement of vehicle position and bounding box was recorded by 8 overhead cameras. For the purposes of this analysis, $n = 4$ vehicle features were investigated: Vehicle position (in feet) relative to the left edge of the highway, Vehicle position (in feet) relative to the beginning of the tracked segment, and vehicle velocity components along those axis (in feet/second). The latter two features were derived via centered numerical differentiation.

The full US-101 dataset consists of nearly five million vehicle-timepoint observations, and as discussed in Section 4.4.1, the trajectories are predominantly non- or partially-overlapping. Accordingly, the moving-window method developed in Section 4.6 will be applied to sequences of rectangular index sets where identity of the comprising vehicles are allowed to change. With the goal of applicability towards autonomous driving, an interpretable class of sequences of rectangular index sets can be defined around a given *ego* vehicle, an example of which is depicted in Figure 4.4. These index sets will include all nearby vehicles for a chosen ego vehicle, and therefore capture the experience of the ego vehicle during its journey down the tracked segment and also represent a cohort of co-travelling vehicles.

Explicitly, a sequence of rectangular index sets is extracted as follows. First, choose some ego vehicle $c_0$, whose first and last recorded positions were at times $t_{\text{start}}$ and $t_{\text{end}}$. Fix a time window length $W$, and a spatial bound $\rho > 0$ for distance from the ego vehicle. For times $t \in \{t_{\text{start}}, \ldots, t_{\text{end}}\}$, define the associated vehicles $\mathbf{c}_0(t)$ as all $c \in \mathbb{N}$ such that there exists some $s \in (t - W/2 + 1) \ldots (t + W/2)$ where $d_{\text{space}}(\mathbf{x}(c, s), \mathbf{x}(c_0, s)) < \rho$, when $d_{\text{space}}$ is the Euclidean distance applied to the two position features of $\mathbf{x}(c, t)$. The time-varying cohort $\mathbf{c}_0(t)$ then consists of all vehicles who were within $\rho$ feet of the ego vehicle $h$ at any point during the window $(t - W/2 + 1) \ldots (t + W/2)$. The ego vehicle itself will always be in $\mathbf{c}_0(t)$, and the identity of the other vehicles is allowed to change as they fall behind or move ahead of the ego vehicle.

This defines the sequence of rectangular index sets with for cohort $\mathbf{c}_0(t)$ at times $\mathbf{s}_(t) = [(t - W/2 + 1), \ldots, (t + W/2)]$ as $t$ ranges from $t_{\text{start}}$ to $t_{\text{end}}$. The estimation procedure of Section 4.6.2 and for the time-varying factor decomposition can be applied, which yields estimates of the parameter curves $\widehat{\mathbf{A}}(c, t), \widehat{\mathbf{B}}(c, t), \widehat{\mathbf{\Phi}}(c, t)$ for all $(c, t)$ pairs such that $c \in \mathbf{c}_0(t)$. These

estimated curve segments enable the use of the Algorithm 3 to $\mathbf{c}_0(t)$ and so a dynamic podding $\widehat{\mathcal{P}}_\lambda(\mathbf{c}_0(t), t)$ of the surrounding vehicles is obtained. Computing the dynamic podding $\widehat{\mathcal{P}}_\lambda(\mathbf{c}_0(t), t)$ requires selecting three tuning parameters, namely the number of common factors $r_0$, the number of vehicle-specific factors $r_1$, and the pod creating parameter $\lambda$. These tuning parameters are selected empirically by a double cross-validation procedure (Zeng et al., 2019), as discussed in Section D.5 in Appendix D.

### 4.7.2 Podding Evaluation

The utility of the proposed method for dynamic podding can be assessed both by examining the qualitative aspects of the time-varying pods, and by comparing the resulting pods to those arising from simpler methods.

Two example dynamic poddings of the same cohort at $4$ times are shown in Figure 4.5, where the vehicles are grouped into time-varying pods based on Algorithm 3 with $W = 50, r_0 = 2, r_1 = 1$ and $\lambda = 40, 60$. The differing $\lambda$ values are seen to decompose the cohort at different levels of granularity. For the larger value $\lambda = 60$, the differing speeds between the slower leftmost two lanes and the faster three rightmost lanes is the dominant influence in the podding. For the smaller value $\lambda = 40$, additional patterns can be discerned. For example, vehicle $1$ in the middle lane is assigned its own pod in three out of the four timesteps, which is due to the fact that it is the only vehicle with no leading vehicle in its lane within the cohort and so is able to (and chooses to) overtake the majority of the cohort. This is in contrast to vehicle $10$ in lane 2, which, despite having no within-cohort leading vehicle, maintains a steady configuration with vehicles $12$ and $13$ and so is assigned to the same pod as them in the second and third times.

For a more quantitative comparison, two other dynamic podding methods are investigated in addition to the proposed dynamic podding method ("MCFM" in Figure 4.6). For the first additional method, vehicle-specific factors are removed by setting $r_1 = 0$, which reduces the multi-channel factor model to a classical single-channel factor model ("SCFM" in Figure 4.6). Comparison of the proposed method to the single-channel model assesses the merit of inclusion of vehicle-specific

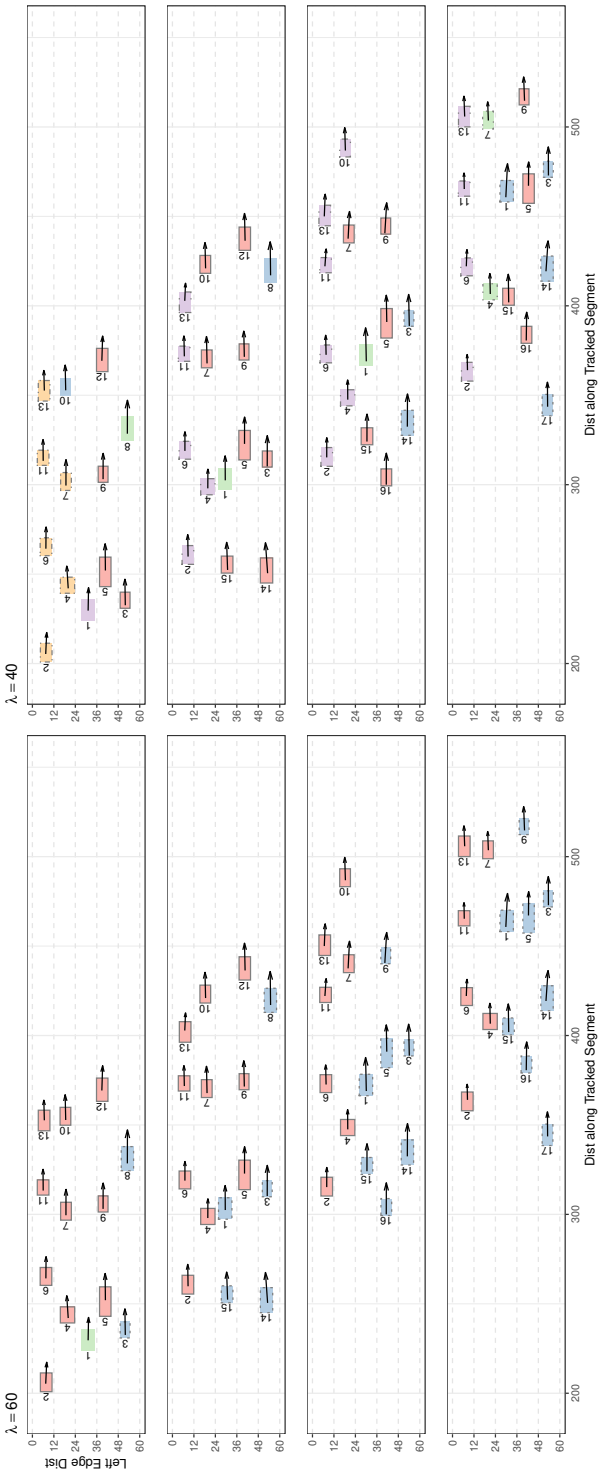



**Figure 4.5:** Example dynamic podding obtained by application of Algorithm 3 with $\lambda = 40, 60$, $r_0 = 2, r_1 = 1, W = 50$. Vehicles are colored according to pod membership, 4 total timepoints are depicted, with the timepoints spaced by 20 timesteps each. Vehicle bounding box is to-scale, and line-type of bounding box outline also indicates pod membership. Vehicles with no bounding box outline belong to a single-vehicle pod.

factors. The second comparison method is a model-free method which directly minimizes the penalized podding loss (4.19) using the formulation (4.20) applied to the window sample covariance matrix (“Trajectory" in Figure 4.6).

For each method, we can evaluate the quality of the resulting pods in a fashion similar to the blocked cross-validation approach outlined in Section D.5. For a subsample of $500$ cohorts, windows of length $W + H$ are extracted and the optimal podding on the first $W$ timesteps is computed for each method over a range of $\lambda$. The resulting within-window podding loss is shown in Figure 4.6, top. Using the same podding assignments, the penalized podding loss (4.19) is computed on the $H$ held-out timesteps, and the out-of-window podding loss is shown in Figure 4.6, mid. The overall average number of pods chosen is shown in figure 4.6, bottom.

Examining Figure 4.6, it is unsurprising that the method based around directly clustering the trajectories minimizes the within-window loss, as the podding produced is also the loss-minimizing podding. However, this performance fails to generalize outside of the training window, as the out-of-window loss is substantially greater for the trajectory-based podding than for both the single-channel and multi-channel factor-based poddings over the range of $\lambda$ determined in Section D.5. The reason for this discrepancy is shown in bottom Figure 4.6, which indicates that the trajectory-based podding tends to produce a greater number of pods than the factor-based approaches. Out-of-window, these additional pods fail to sufficiently reduce the within-pod change in relative configuration to offset the inclusion penalty, and so are effectively spurious.

## 4.8 Discussion

This chapter employs the idea behind multi-channel factor analysis to develop a model-based decomposition of cohorts of vehicles into pods, where the within-pod relative configuration of vehicles is more stable than the overall traffic stream.

A generative model for the joint motion of cohorts of vehicles is developed, where the cross-sectional covariance is of MFA form (1.6). This multi-channel factor model for vehicle position and velocity separates the modeled correlations in motion across vehicles from the modeled correlation of

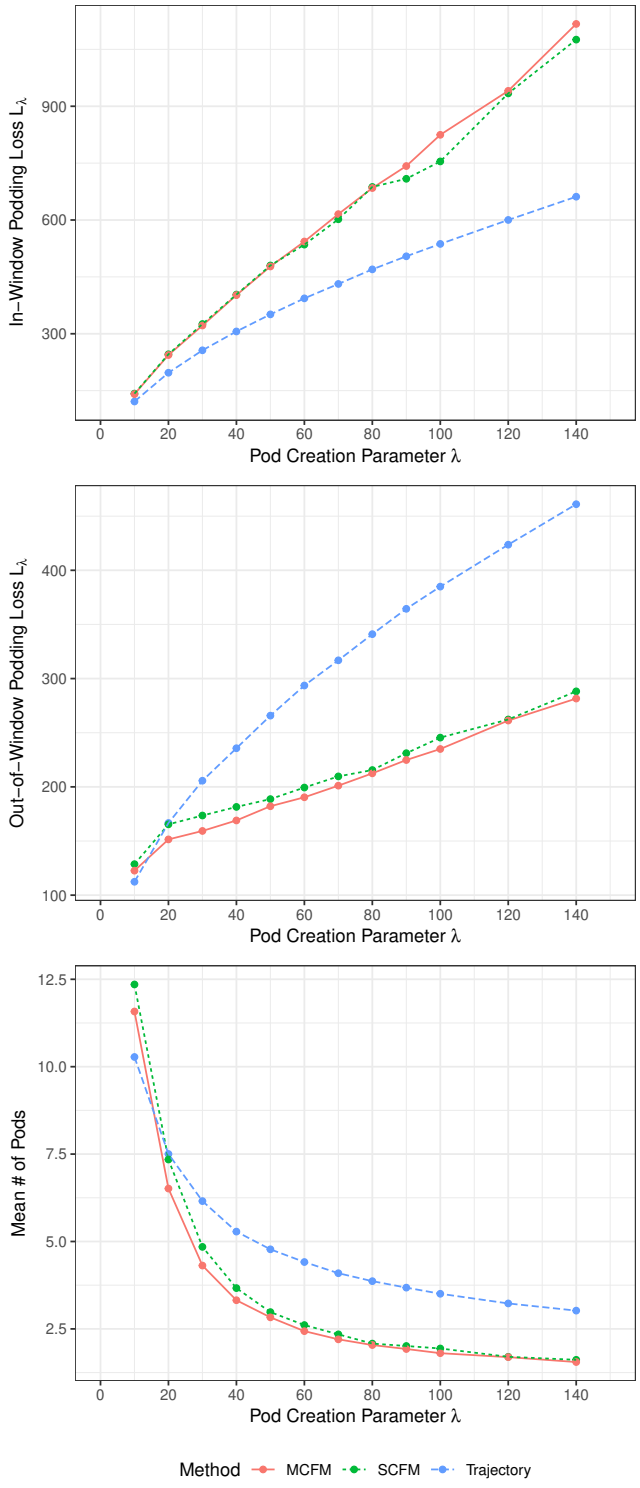



**Figure 4.6:** Comparison of dynamic podding methods. *Top*: The within-window penalized podding loss (4.19) for all three methods, over a range of $\lambda$. *Mid*: The resulting penalized podding loss on the held-out observations when the podding obtained on the training window is propagated forward, as detailed in Section 4.7.2. *Bottom*: The average number of pods per window, for each method.

the kinematic variables for a single vehicle. The across-vehicle correlations are modeled by a small number of factor loading parameters, which isolate and summarize the observed motion coordination between vehicles. In order to model the joint motion of cohorts of vehicles within complete highway trajectory datasets, it is necessary to allow for change over time in the relationships between vehicles as well as change in which vehicles compose the cohort. The time-varying model (4.10) allows for change over time to the factor loading parameters, and a moving-window estimation procedure is constructed.

These loading parameter curves summarize how the motion of a given vehicle relates to the others within its time-varying cohort. A geometric measure for the stability of the relative configuration of a collection of vehicles is developed. It is demonstrated that assigning vehicles to pods to minimize expected relative configuration loss under the generative model produces a more robust podding method than a model-free trajectory-based approach. This model-based podding is fit into the framework of algorithmic clustering methods with a particular kernel function, and application of these methods to the loading parameter curves provides a dynamic podding method which is optimal under the generative model.

Both the locally stationary multi-channel factor model for vehicle trajectories and the resulting dynamic podding method can be applied to help answer questions of interest to researchers. Driver actions which differ significantly those of nearby drivers may be questionable, and so identifying drivers with a pattern of taking such actions is of interest (Dinh et al., 2021). Within the framework of this Chapter, these outlying drivers will be frequently assigned to a single-vehicle pod if their relative position to surrounding vehicles is not sufficiently stable. So, examining the proportion of the trip down the tracked highway section which is spent in a single-vehicle pod can allow for identification of outlying drivers. This method does not depend on defining pre-sepecified dangerous maneuvers or desired vehicle configurations, but is instead dependant solely on the across-vehicle relationships.

Frameworks for driver decision-making (Frédéric et al., 2006) propose that drivers balance individual efficiency and societal norms when planning their vehicle trajectory. Proposed trajectories

which maintain more stable relationships relative to other vehicles while still accomplishing the desired goal are likely to be more socially acceptable. Incorporation of the observed trajectories of the current or target pod into a trajectory planning algorithms can therefore improve the acceptability of proposed trajectories, as these pods are defined based on the stability of their relative configuration. Socially-aware trajectory planning approaches can then improve the performance of autonomous vehicles in mixed traffic environments.

For cohorts of vehicles traveling down the highway, the regularity of their joint motion allows for parsimonious modeling of their trajectories. The time-varying relationships between nearby vehicles can be captured in a parametric model and used to create a dynamic decomposition into pods of vehicles who maintain a stable relative configuration. The existence of these pods is an interesting feature of highway traffic and is potentially informative about how human drivers understand their relationship to other vehicles in the context of societal driving.

# Chapter 5

# Conclusion and Future Work

## 5.1 Discussion

Multi-channel factor analysis maintains much of the applicability and interpretability of classical factor analysis, and many of the properties and uses of MFA can be seen as extensions from the single-channel case to the multi-channel case. For example, the generic global identifiability result of Chapter 2 was obtained by first examining the corresponding result for classic FA and uncovering what missing pieces were needed for the multi-channel version. However, the multi-channel setting does result in some new properties, such as the differing threshold for generic local and global identifiability. The interplay between the three structured components in the MFA covariance model increases the complexity which must be managed to obtain the desired results, but the usefulness of MFA is sufficient to make the effort worthwhile.

In a similar fashion, the extension from MFA to MFSA adds another layer of complexity to balance. However, as the construction of the surrogate function on the basis of block-diagonal upper bound in Section 3.3.4 exemplifies, the complexity can be reduced by adopting the correct perspective on the problem at hand. Further, as the experiments in Section 3.5 demonstrate, the complications created by the addition of an explicit model for the temporal correlations can be well worth the cost.

Like the use of MFA detailed in Chapter 4, many problems can be framed in such a way that MFA is naturally applicable, even when the problem does not appear to have a channel structure at first glance. Just as classical factor analysis works to model the covariances between distinct variables in a simple way and without restricting the variances, the simultaneous and parsimonious modeling of both the within-channel and across-channel covariances is the core advantage of MFA. As multi-channel methods proliferate, the straightforward appeal of multi-channel factor analysis will continue to interest practitioners.

## 5.2 Future Research Directions

The results of this dissertation are obtained assuming that the signal and interference dimensions are prespecified. Many applications of interest would require estimating these dimensions from the observations, which is a challenging order selection problem as the MFA covariance sets $\mathcal{S}(\mathbf{n}, \mathbf{r})$ for varying factor numbers $\mathbf{r}$ are nested in a partial order, rather than a total order. In single-channel FA, techniques such as maximization of an information criterion (Akaike, 1987), bi-cross-validation (Owen and Wang, 2016), or eigenvalue analysis (Ahn et al., 2009) can be used to estimate the number of common factors. Adapting these techniques or inventing new methods of order selection in MFA is a very important direction for future work.

An additional interesting direction for future work is the high-dimensional setting. Many modern techniques in factor analysis derive their validity from asymptotic regimes where, in addition to the number of samples, the measurement dimensionality also increases. This can allow for the asymptotically exact recovery of the latent factors, and can be used to justify simple, computationally efficient estimation procedures which perform well as long as the measurement dimensionality is sufficiently large. Applying the same idea to MFA is made more complicated by the question of determining what the high-dimensional asymptotic regime should be. Options include increasing the number of channels, fixing the number of channels and increasing channel size, or allowing both number of channels and channel size to go to infinity while controlling the ratio of the largest channel size to the smallest channel size. These differing asymptotic regimes may have diverse implications for the identification of the latent variables and future work should explore what form *high-dimensional* MFA could take.

For MFSA, the power spectral densities in the model developed in Chapter 3 are formulated in terms of rational transfer functions. The resulting method necessarily has all the advantages and disadvantages of classic multivariate time series models, and it could be of interest to instead utilize a nonparametric or semiparametric approach to modeling the power spectral density of the latent series.

Finally, it could be of interest to apply the joint trajectory model of Chapter 4 to sensor data collected from autonomous vehicles with rich suites of sensors. Data fusion is a topic of great interest in the autonomous driving literature, and combining the data collected from several sensors across multiple vehicles could potentially yield interesting findings about road state and pro-social driving behaviors, such as socially-aware trajectory planning as discussed in Section 4.8.

# Bibliography


Evrim Acar, Morten Arendt Rasmussen, Francesco Savorani, Tormod Næs, and Rasmus Bro. Understanding data fusion within the framework of coupled matrix and tensor factorizations. *Chemometrics and Intelligent Laboratory Systems*, 129:53–63, 2013.

Tülay Adali, Peter J. Schreier, and Louis L. Scharf. Complex-valued signal processing: The proper way to deal with impropriety. *IEEE Transactions on Signal Processing*, 59(11):5101–5125, 2011.

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering - a decade review. *Inf. Syst.*, 53(C):16–38, 2015.

Seung Ahn, W Carey, and Alex Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81, 2009.

Hirotugu Akaike. Factor analysis and AIC. *Psychometrika*, 52:317–332, 1987.

Assad Al Alam, Ather Gattami, and Karl Henrik Johansson. An experimental study on the fuel reduction potential of heavy duty vehicle platooning. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 306–311, 2010.

Jont B. Allen and Lawrence R. Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65:1558–1564, 1977.

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 1959.

Theodore W. Anderson and Herman Rubin. Statistical inference in factor analysis. *University of California Press*, pages 111–150, 1956.

Galen Andrew, Raman Arora, Jeff Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Intl. Conference on Machine Learning*, 2013.

Avishay Antman and Amir Leshem. Radio transient detection in radio astronomical arrays. *IEEE Transactions on Signal Processing*, 68:5648–5663, 2020.

Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152 – 1174, 1974.

Allan M. Avila and Igor Mezić. Data-driven analysis and forecasting of highway traffic dynamics. *Nature Communications*, 11, 2020.

Francesco Bandiera, Olivier Besson, Danilo Orlando, Giuseppe Ricci, and Louis L. Scharf. GLRT-based direction detectors in homogeneous noise and subspace interference. *IEEE Transactions on Signal Processing*, 55(6):2386–2394, 2007.

Maurice Bartlett. The statistical concept of mental factors. *British Journal of Psychology. General Section*, 28:97 – 104, 1937.

Alexander Basilevsky. *Statistical Factor Analysis and Related Methods*. John Wiley & Sons, New York, 1994.

Paul A. Bekker. A note on the identification of restricted factor loading matrices. *Psychometrika*, 51:607–611, 1986.

Paul A. Bekker and Jos M.F. ten Berge. Generic global indentification in factor analysis. *Linear Algebra and its Applications*, 264:255–263, 1997.

Abderrahim Benslimane, Tarik Taleb, and Rajarajan Sivaraj. Dynamic clustering-based adaptive mobile gateway management in integrated VANET — 3G heterogeneous wireless networks. *IEEE Journal on Selected Areas in Communications*, 29(3):559–570, 2011.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, Princeton, 2007.

Albert J. Boonstra and Alle-Jan van der Veen. Gain calibration methods for radio telescope arrays. *IEEE Transactions on Signal Processing*, 51(1):25–38, 2003.

Hamid Bouchachia. Dynamic clustering. *Evolving Systems*, 3, 2012.

David R. Brillinger. *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston, New York, 1975.

Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer New York, New York, 1991.

M. W. Browne. The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology*, 32(1):75–86, 1979.

Almudena Buelta, Alberto Olivares, Ernesto Staffetti, Waqas Aftab, and Lyudmila Mihaylova. A Gaussian process iterative learning control for aircraft trajectory tracking. *IEEE Transactions on Aerospace and Electronic Systems*, 57(6):3962–3973, 2021.

Jean-Francois Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86 (10):2009–2025, 1998.

Jean-Francois Cardoso and Antoine Souloumiac. Blind beamforming for non Gaussian signals. *Radar and Signal Processing, IEE Proceedings F*, 140:362 – 370, 1994.

J. Douglas Carroll and Jih Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrika*, 35:283–319, 1970.

Raymond Bernard Cattell. Parallel proportional profiles and other principles for determining the choice of factors by rotation. *Psychometrika*, 9:267–283, 1944.

Mainak Chatterjee, Sajal K Das, and Damla Turgut. WCA: A weighted clustering algorithm for mobile ad hoc networks. *Cluster computing*, 5(2):193–204, 2002.

Bo Chen, Gungor Polatkan, Guillermo Sapiro, David Blei, David Dunson, and Lawrence Carin. Deep learning with hierarchical convolutional factor analysis. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1887–1901, 2013.

Zexun Chen, Jun Fan, and Kuo Wang. Multivariate Gaussian processes: Definitions, examples and applications. *METRON*, 81(2):181–191, 2023.

Chang Chiann and Pedro Alberto Morettin. Time-domain estimation of time-varying linear systems. *Journal of Nonparametric Statistics*, 17:365 – 383, 2005.

Nidhan Choudhuri. Contiguity of the Whittle measure for a Gaussian time series. *Biometrika*, 91 (1):211–218, 2004.

X. Chu, D. L. Perez, Y. Yang, and F. Gunnarsson. *Heterogeneous Cellular Networks: Theory, Simulation and Deployment*. Cambridge University Press, 2013.

Andrzej Cichocki, Danilo Mandic, Lieven De Lathauwer, Guoxu Zhou, Qibin Zhao, Cesar Caiafa, and Huy Anh Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2):145–163, 2015.

Stéphane Crépey and Matthew F. Dixon. Gaussian process regression for derivative portfolio modeling and application to credit valuation adjustment computations. *The Journal of Computational Finance*, 24(1), 2020.

Ameneh Daeinabi, Akbar Ghaffar Pour Rahbar, and Ahmad Khademzadeh. VWCA: An efficient clustering algorithm in vehicular ad hoc networks. *Journal of Network and Computer Applications*, 34(1):207–222, 2011.

Rainer Dahlhaus. Small Sample Effects in Time Series Analysis: A New Asymptotic Theory and a New Estimate. *The Annals of Statistics*, 16(2):808 – 841, 1988.

Rainer Dahlhaus. On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Processes and their Applications*, 62(1):139–168, 1996a.

Rainer Dahlhaus. Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In *Athens Conference on Applied Probability and Time Series Analysis*, pages 145–159, New York, NY, 1996b. Springer New York.

Rainer Dahlhaus. Locally stationary processes. In Tata Subba Rao, Suhasini Subba Rao, and C.R. Rao, editors, *Time Series Analysis: Methods and Applications*, volume 30 of *Handbook of Statistics*, pages 351–413. Elsevier, 2012.

Maria Valentina Niño de Zepeda, Fanlin Meng, Jinya Su, Xiao-Jun Zeng, and Qian Wang. Dynamic clustering analysis for driving styles identification. *Engineering Applications of Artificial Intelligence*, 97:104096, 2021.

Arthur. P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 551 – 556, New York, 2004.

Thinh Hoang Dinh, Vincent Martinez, and Daniel Delahaye. Recognition of outlying driving behaviors: A data-driven perspective with applications to V2X collective perception. In *2021 IEEE Vehicular Networking Conference (VNC)*, pages 52–59, 2021.

Mischa Dohler and Yonghui Li. *Cooperative Communications: Hardware, Channel & PHY*. John Wiley & Sons, New York, 2010.

Willem Eerland, Simon Box, and Andras Sobester. Modeling the dispersion of aircraft trajectories using Gaussian processes. *Journal of Guidance, Control, and Dynamics*, 39, 2016.

Madeleine El-Zaher, Franck Gechter, Pablo Gruer, and Mohammad Hajjar. A new linear platoon model based on reactive multi-agent systems. In *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, pages 898–899, 2011.

Madeleine El-Zaher, Baudouin Dafflon, Franck Gechter, and Jean-Michel Contet. Vehicle platoon control with multi-configuration ability. *Procedia Computer Science*, 9:1503–1512, 2012.

Lasha Ephremidze, Gennady Mishuris, and Ilya Spitkovsky. On the exact spectral factorization of rational matrix functions with applications to paraunitary filter banks. *Journal of Fourier Analysis and Applications*, 30, 2024.

Beyza Ermiş, Evrim Acar, and A. Taylan Cemgil. Link prediction in heterogeneous data via generalized coupled tensor factorization. *Data Min. Knowl. Discov.*, 29(1):203–236, 2015.

Jianqing Fan, Kaizheng Wang, Yiqiao Zhong, and Ziwei Zhu. Robust high dimensional factor models with applications to statistical machine learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(2):303, 2021.

Hamideh Fatemidokht and Marjan Kuchaki Rafsanjani. QMM-VANET: An efficient clustering algorithm based on QoS and monitoring of malicious vehicles in vehicular ad hoc networks. *Journal of Systems and Software*, 165:110561, 2020.

Vanderhaegen Frédéric, Sebastien Chalme, Francoise Anceaux, and Patrick Millot. Principles of cooperation and competition: Application to car driver behavior analysis. *Cognition, Technology & Work*, 8:183–192, 2006.

Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation*, 32(5):829–864, 2020.

James E. Gentle. *Numerical Linear Algebra for Applications in Statistics*. Springer New York, New York, 1998.

Sepideh Afkhami Goli, Behrouz H. Far, and Abraham O. Fapojuwo. Vehicle trajectory prediction with Gaussian process regression in connected vehicle environment $\star$. In *IEEE Intelligent Vehicles Symposium*, pages 550–555, 2018.

H. Guggenheimer, Alan Edelman, and Charles R. Johnson. A simple estimate of the condition number of a linear system. *College Mathematics Journal*, 26:2–5, 1995.

Feng Guo and Youjia Fang. Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention*, 61:3–9, 2013.

Daniel E. Hack, Lee K. Patton, Braham Himed, and Michael A. Saville. Detection in passive MIMO radar networks. *IEEE Transactions on Signal Processing*, 62(11):2999–3012, 2014.

Hanshin Exp. Co. Ltd. Zen traffic data. https://zen-traffic-data.net/english/outline/, 2024. Accessed: 2021-09-23.

Christopher C. Heyde. *Quasi-Likelihood And Its Application*. Springer New York, New York, 1997.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

Rudra Hota, Shahanaz Syed, Subhadip Bandyopadhyay, and Pisipati Radha Krishna. A simple and efficient lane detection using clustering and weighted regression. In *Proceedings of the 15th International Conference on Management of Data*, pages 1–8, 2009.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936.

Matt C. Howard and Rachel O'Sullivan. A systematic review of exploratory factor analysis in marketing: providing recommended guidelines and evaluating current practices. *Journal of Marketing Theory and Practice*, pages 1–22, 2024.

Mohamed Salah Ibrahim and Nicholas D. Sidiropoulos. Cell-edge interferometry: Reliable detection of unknown cell-edge users via canonical correlation analysis. In *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, 2019.

F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proceedings of the 6th of the International Congress on Acoustics*, pages C17–C20, 1968.

Gigla Janashia, Edem Lagvilava, and Lasha Ephremidze. A new method of matrix spectral factorization. *IEEE Transactions on Information Theory*, 57(4):2318–2326, 2011.

Zehui Jia, Xingju Cai, and Deren Han. Comparison of several fast algorithms for projection onto an ellipsoid. *Journal of Computational and Applied Mathematics*, 319:320–337, 2017.

Karl G. Jöreskog. A general method for analysis of covariance structures. *Biometrika*, 57:239–251, 1970.

Karl G. Jöreskog. A general approach to confirmatory factor analysis. *Psychometrika*, 34:183–202, 1969.

Thomas Kailath. *Linear Systems*. Prentice-Hall, Eaglewood Cliffs, N.J., 1980.

Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 2009.

Maurice G. Kendall. Factor analysis as a statistical technique. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12(1):60–73, 1950.

Jon R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

Bahador Khaleghi, Alaa Khamis, Fakhreddine O. Karray, and Saiedeh N. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, 2013.

Pascal Kilian, Daniel Leyhr, Christopher J. Urban, Oliver Höner, and Augustin Kelava. A deep learning factor analysis model based on importance-weighted variational inference and normalizing flow priors: Evaluation within a set of multidimensional performance assessments in youth elite soccer players. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 16(5): 474–487, 2023.

Franz Kiràly and Ryota Tomioka. A combinatorial algebraic approach for the identifiability of low-rank matrix completion. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 1, 2012.

Tono Kollo and Rosen Dietrich. *Advanced Multivariate Statistics with Matrices*. Springer Dordrecht, Dordrecht, Netherlands, 2005.

Andreas I. Koutrouvelis, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. Estimation of sensor array signal model parameters using factor analysis. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125, 2018.

Joseph B. Kruskal. *Rank, decomposition, and uniqueness for 3-way and n-way arrays*, page 7–18. North-Holland Publishing Co., NLD, 1989. ISBN 0444874100.

Li-Dan Kuang, Zhi-Ming He, Jianming Zhang, and Feng Li. Coupled canonical polyadic decomposition of multi-group fMRI data with spatial reference and orthonormality constraints. *Biomedical Signal Processing and Control*, 80:104232, 2023.

Brian Kulis and Michael I Jordan. Revisiting k-means: new algorithms via Bayesian nonparametrics. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1131–1138, 2012.

Dana Lahat and Christian Jutten. Joint independent subspace analysis using second-order statistics. *IEEE Transactions on Signal Processing*, 64(18):4891–4904, 2016.

Dana Lahat and Christian Jutten. Joint independent subspace analysis: Uniqueness and identifiability. *IEEE Transactions on Signal Processing*, 67(3):684–699, 2019.

Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

Kenneth Lange. *Optimization*. Springer New York, New York, NY, 2013.

D. N. Lawley and A. E. Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229, 1962.

Walter Ledermann. On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika*, 2:85–93, 1937.

John M. Lee. *Introduction to Smooth Manifolds*. Springer New York, New York, 2002.

Florin Leon and Marius Gavrilescu. A review of tracking and trajectory prediction methods for autonomous driving. *Mathematics*, 9(6):660, 2021.

Yuri Levin-Schwartz, Yang Song, Peter J. Schreier, V.D. Calhoun, and Tülay Adali. Sample-poor estimation of order and common signal subspace with application to fusion of medical imaging data. *NeuroImage*, 134:486–493, 2016.

Chun-hung R. Lin and Mario Gerla. Adaptive clustering for mobile wireless networks. *IEEE Journal on Selected Areas in Communications*, 15(7):1265–1275, 1997.

Eric Lock, Katherine Hoadley, J.S. Marron, and Andrew Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics*, 7:523–542, 2013.

Xiao-Yun Lu and Alexander Skabardonis. Freeway traffic shockwave analysis: exploring the NGSIM trajectory data. In *86th Annual Meeting of the Transportation Research Board, Washington, DC*. Citeseer, 2007.

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*, 2022.

Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, New York, 1988.

Vishal Mahajan, Christos Katrakazas, and Constantinos Antoniou. Prediction of lane-changing maneuvers with automatic labeling and deep learning. *Transportation Research Record*, 2674(7): 336–347, 2020.

Santa Maiti, Stephan Winter, and Lars Kulik. A conceptualization of vehicle platoons and platoon operations. *Transportation Research Part C: Emerging Technologies*, 80:1–19, 2017.

Jean Pierre Mercat, Nicole El Zoghby, Guillaume Sandou, Dominique Beauvois, and Guillermo Pita Gil. Inertial single vehicle trajectory prediction baselines and applications with the ngsim dataset, 2019. Accessed: 09-29-2021.

Todd K. Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13 (6):47–60, 1996.

Freddy D. Neeser and J.L. Massey. Proper complex random processes with applications to information theory. *IEEE Transactions on Information Theory*, 39(4):1293–1302, 1993.

Art B. Owen and Jingshu Wang. Bi-Cross-Validation for factor analysis. *Statistical Science*, 31(1): 119 – 139, 2016.

Meng-Yuan Pang. Trajectory data based clustering and feature analysis of vehicle lane-changing behavior. In *2019 4th International Conference on Electromechanical Control Technology and Transportation (ICECTT)*, pages 229–233, 2019.

Andrew Patterson, Aditya Gahlawat, and Naira Hovakimyan. Learning probabilistic intersection traffic models for trajectory prediction, 2020. Accessed 09-29-2021.

Marius Pesavento and A.B. Gershman. Maximum-likelihood direction-of-arrival estimation in the presence of unknown nonuniform noise. *IEEE Transactions on Signal Processing*, 49(7): 1310–1324, 2001.

David Ramírez, Gonzalo Vazquez-Vilar, Roberto Lopez-Valcarce, Javier Via, and Ignacio Santamaria. Detection of rank- $p$ signals in cognitive radio networks with uncalibrated multiple antennas. *IEEE Transactions on Signal Processing*, 59(8):3764–3774, 2011.

David Ramírez, Ignacio Santamaria, Louis L. Scharf, and Steven Van Vaerenbergh. Multi-channel factor analysis with common and unique factors. *IEEE Transactions on Signal Processing*, 68: 113–126, 2020.

David Ramírez, Ignacio Santamaría, and Louis L. Scharf. *Coherence in Signal Processing and Machine Learning*. Springer Cham, 2023.

C. Radhakrishna Rao. Estimation and tests of significance in factor analysis. *Psychometrika*, pages 93–111, 1955.

Suhasini Subba Rao and Junho Yang. Reconciling the Gaussian and Whittle likelihood with an application to estimation in the frequency domain. *The Annals of Statistics*, 2020.

Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

Hugo Richard, Pierre Ablin, Bertrand Thirion, Alexandre Gramfort, and Aapo Hyvärinen. Shared independent component analysis for multi-subject neuroimaging. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, 2024.

Yu. A. Rozanov. Spectral properties of multivariate stationary processes and boundary properties of analytic matrices. *Theory of Probability & Its Applications*, 5(4):362–376, 1960.

Yu. A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, CA, 1967.

Ahmad Mouri Sardarabadi and Alle-Jan van der Veen. Complex factor analysis and extensions. *IEEE Transactions on Signal Processing*, 66(4):954–967, 2018.

Victhor S. Sartório and Thaís C. O. Fonseca. Dynamic clustering of time series data, 2020. Accessed: 09-29-2021.

Toru Seo, Yusuke Tago, and Norihito Shinkai. Evaluation of large-scale complete vehicle trajectories dataset on two kilometers highway segment for one hour duration: Zen traffic data. Accessed: 2021-09-23, 2020.

Alexander Shapiro. Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, 47:187–199, 1982.

Alexander Shapiro. Identifiability of factor analysis: some results and open problems. *Linear Algebra and its Applications*, 70:1–7, 1985.

Xiupeng Shi, Yiik Diew Wong, Chen Chai, Michael Zhi-Feng Li, Tianyi Chen, and Zeng Zeng. Automatic clustering for unsupervised risk diagnosis of vehicle driving for smart road. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):17451–17465, 2022.

Nikolaos D. Sidiropoulos and R. Bro. On communication diversity for blind identifiability and the uniqueness of low-rank decomposition of N-way arrays. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 5, pages 2449–2452 vol.5, 2000.

Rogers F. Silva, Sergey M. Plis, Tülay Adali, and Vince D. Calhoun. Multidataset independent subspace analysis extends independent vector analysis. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 2864–2868, 2014.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Muni S. Srivastava, Tatjana von Rosen, and Dietrich von Rosen. Models with a kronecker product covariance structure: Estimation and testing. *Mathematical Methods of Statistics*, 17:357–370, 2007.

Gray Stanton, David Ramírez, Ignacio Santamaria, Louis L. Scharf, and Haonan Wang. Identifiability in multi-channel factor analysis. In *57th Asilomar Conference on Signals, Systems, and Computers*, pages 1344–1349, 2023a.

Gray Stanton, Haonan Wang, Dongliang Duan, and Louis L. Scharf. Multi-channel factor analysis for temporally and spatially correlated time series. In *57th Asilomar Conference on Signals, Systems, and Computers*, pages 1339–1343, 2023b.

Gray Stanton, David Ramírez, Ignacio Santamaria, Louis Scharf, and Haonan Wang. Multi-channel factor analysis: Identifiability and asymptotics. *IEEE Transactions on Signal Processing*, 72: 3562–3577, 2024.

David W. Stewart. The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, 18(1):51–62, 1981.

Ying Sun, Prabhu Babu, and Daniel Pérez Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65:794–816, 2017.

Mikael Sørensen, Charilaos I. Kanatsoulis, and Nicholas D. Sidiropoulos. Generalized canonical correlation analysis: A subspace intersection approach. *IEEE Transactions on Signal Processing*, 69:2452–2467, 2021.

Tetsuya Takabatake and Keisuke Yano. Towards a robust frequency-domain analysis: Spectral Rényi divergence revisited, 2023.

Masanobu Taniguchi. Minimum contrast estimation for spectral densities of stationary processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(3):315–325, 1987.

Terence Tao. *An Introduction to Measure Theory*. Amer Mathematical Society, Providence, RI, 2011.

Louis L. Thurstone. *The Vectors of Mind*. University of Chicago Press, Chicago, 1935.

Louis L. Thurstone. *Multiple-Factor Analysis*. University of Chicago Press, Chicago, 1947.

Behrad Toghi, Rodolfo Valiente, Dorsa Sadigh, Ramtin Pedarsani, and Yaser P. Fallah. Cooperative autonomous vehicles that sympathize with human drivers. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4517–4524, 2021.

Behrad Toghi, Rodolfo Valiente, Dorsa Sadigh, Ramtin Pedarsani, and Yaser P. Fallah. Social coordination and altruism in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):24791–24804, 2022.

Ledyard Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311, 1966.

Matthew Turk. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195, 2014.

U.S Department of Transportation, Federal Highway Administration. Next generation simulation (NGSIM) vehicle trajectories and supporting data. Provided by ITS DataHub through Data.transportation.gov, 2016. Accessed: 2021-09-23.

Adrianus W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

Adriaan van den Bos. Complex gradient and Hessian. *IEE Proceedings - Vision, Image and Signal Processing*, 141:380–382(2), 1994.

Alle-Jan Van Der Veen, Amir Leshem, and Albert-Jan Boonstra. *Array Signal Processing for Radio Astronomy*, pages 231–249. Springer Netherlands, Dordrecht, 2005.

Michael Vogt. Nonparametric regression for locally stationary time series. *The Annals of Statistics*, 40(5):2601 – 2633, 2012.

Lin Wang, Qing-Feng Lin, Zhen-Yu Wu, and Bin Yu. A data-driven estimation of driving style using deep clustering. In *COTA International Conference of Transportation Professionals*, pages 4183–4194. American Society of Civil Engineers, 2020.

Wenshuo Wang, Junqiang Xi, Alexandre Chong, and Lin Li. Driving style classification using a semisupervised support vector machine. *IEEE Transactions on Human-Machine Systems*, 47: 650–660, 2017.

Karl Werner, Magnus Jansson, and Petre Stoica. On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56:478 – 491, 2008.

Peter Whittle. The analysis of multiple stationary time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(1):125–139, 1953.

De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, and Joseph Walsh. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 21(6), 2021.

Dewei Yi, Jinya Su, Cunjia Liu, and Wen-Hua Chen. Data-driven situation awareness algorithm for vehicle lane change. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 998–1003, 2016.

Youngmin Yoon, Changhee Kim, Jongmin Lee, and Kyongsu Yi. Interaction-aware probabilistic trajectory prediction of cut-in vehicles using Gaussian process for proactive control of autonomous vehicles. *IEEE Access*, 9:63440–63455, 2021.

Xianli Zeng, Yingcun Xia, and Linjun Zhang. Double cross validation for the number of factors in approximate factor models, 2019. Accessed: 09-30-21.

Fuzhen Zhang, editor. *The Schur Complement and its Applications*. Springer New York, New York, 2005.

Jiacheng Zhu, Shenghao Qin, Wenshuo Wang, and Ding Zhao. Probabilistic trajectory prediction for autonomous vehicles with attentive recurrent neural process, 2019. Accessed: 09-29-2021.

# Appendices

# Appendix A

# Supplementary Materials for Chapter 1

## A.1 Proof of Proposition 1

*Proof.* First, suppose $\mathbf{K} \in \mathcal{K}(\mathbf{n}, \mathbf{r})$ with rank $r_0 + r$. By the definition of $\mathcal{K}(\mathbf{n}, \mathbf{r})$, there exist some $\mathbf{A} \in \mathbb{R}_{n \times r_0}$ and $\mathbf{B} = \mathrm{blkdiag}(\mathbf{B}_1, \ldots, \mathbf{B}_C)$ with $\mathbf{B}_c \in \mathbb{R}_{n_c \times r_c}$ for $c = 1, \ldots, C$ such that $[\mathbf{B}\ \mathbf{A}]$ has rank $r_0 + r$ and $\mathbf{K} = [\mathbf{B}\ \mathbf{A}][\mathbf{B}\ \mathbf{A}]^\mathsf{T} = \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T}$. Let $\mathbf{QR}$ be the thin QR decomposition of $[\mathbf{B}\ \mathbf{A}]$, so $\mathbf{Q} \in \mathbb{R}_{n \times (r_0 + r)}$ has orthogonal columns and $\mathbf{R}$ is upper triangular. The matrices are partitioned as

$$\mathbf{Q} = \begin{matrix} \begin{matrix} r_1 & \ldots & r_C & r_0 \end{matrix} \\ \begin{bmatrix} \mathbf{Q}_1 & \ldots & \mathbf{Q}_C & \mathbf{Q}_0 \end{bmatrix} \end{matrix} n \qquad \mathbf{R} = \begin{matrix} \begin{matrix} r_1 & \ldots, & r_C & r_0 \end{matrix} \\ \begin{bmatrix} \mathbf{R}_{11} & \ldots & \mathbf{R}_{1C} & \mathbf{R}_{10} \\ & \ddots & \vdots & \vdots \\ & & \mathbf{R}_{CC} & \mathbf{R}_{C0} \\ & & & \mathbf{R}_{00} \end{bmatrix} \end{matrix} \begin{matrix} r_1 \\ \vdots \\ r_C \\ r_0 \end{matrix}. \tag{A.1}$$

As $\mathbf{B}$ is block diagonal, the equation $[\mathbf{B}\ \mathbf{A}] = \mathbf{QR}$ implies

$$\begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{0} \end{bmatrix} = \mathbf{Q}_1 \mathbf{R}_{11},$$

and as $[\mathbf{B}\ \mathbf{A}]$ is full column rank, the diagonal entries of $\mathbf{R}_{11}$ are positive and hence $\mathbf{R}_{11}$ is invertible. Multiplying the above display by $\mathbf{R}_{11}^{-1}$ on the right, we obtain that $\mathbf{Q}$ in fact equals $[\mathbf{Q}_{11}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \ldots, \mathbf{0}^\mathsf{T}]^\mathsf{T}$ where $\mathbf{Q}_{11} = \mathbf{B}_1 \mathbf{R}_{11}^{-1}$, so $\mathrm{col}\, \mathbf{Q}_1 \subset \mathcal{C}_1$, and $\mathbf{Q}_{11} \in \mathbb{R}_{n_1 \times r_1}$ has orthogonal columns. Then as the top-right $n_1 \times (r - r_1)$ submatrix of $\mathbf{B}$ is zero as $\mathbf{B}$ is block diagonal, the equation $\mathbf{Q}^\mathsf{T}[\mathbf{B}\ \mathbf{A}] = \mathbf{R}$ implies $\mathbf{R}_{1k} = \mathbf{0}_{r_1, r_k}$ for $k = 2, \ldots, C$. As an induction hypothesis, suppose $\mathrm{col}\, \mathbf{Q}_k \subset \mathcal{C}_k$ for $k = 1, \ldots, c-1$ and $\mathbf{R}_{kj} = \mathbf{0}_{r_k, r_j}$ for $k = 1, \ldots, c-1$ and $j = k+1, \ldots, C$.

In the same fashion as for $c = 1$, the equation $[\mathbf{B}\ \mathbf{A}] = \mathbf{QR}$ yields

$$\begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{B}_c \\ \vdots \\ \mathbf{0} \end{bmatrix} = \mathbf{Q}_1\mathbf{R}_{1c} + \cdots + \mathbf{Q}_{c-1}\mathbf{R}_{(c-1)c} + \mathbf{Q}_c\mathbf{R}_{cc} = \mathbf{Q}_c\mathbf{R}_{cc}$$

where the second inequality follows from the induction hypothesis. As $\mathbf{R}_{cc}$ is invertible, we similarly obtain $\operatorname{col}\mathbf{Q}_c \subset \mathcal{C}_c$. Similarly, as the submatrix of $\mathbf{B}$ consisting of the $n_c$ rows containing the block $\mathbf{B}_c$ and the $r - (r_1 + \cdots + r_c)$ columns to the right of $\mathbf{B}_c$ are zero, the equation $\mathbf{Q}^\mathsf{T}[\mathbf{B}\ \mathbf{A}] = \mathbf{R}$ then implies that $\mathbf{R}_{cj} = \mathbf{0}$ for $j = c+1, \ldots, C$ and so the induction step is verified. We obtain that $\mathbf{Q}$ and $\mathbf{R}$ are in fact structured as

$$\mathbf{Q} = \begin{array}{c} \begin{matrix} r_1 & \ldots & r_C & r_0 \end{matrix} \\ \begin{bmatrix} \mathbf{Q}_{11} & \ldots & \mathbf{0} & \mathbf{Q}_{10} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \ldots & \mathbf{Q}_{CC} & \mathbf{Q}_{C0} \end{bmatrix} \begin{matrix} n_1 \\ \vdots \\ n_C \end{matrix} \end{array} \qquad \mathbf{R} = \begin{array}{c} \begin{matrix} r_1 & \ldots & r_C & r_0 \end{matrix} \\ \begin{bmatrix} \mathbf{R}_{11} & \ldots & \mathbf{0} & \mathbf{R}_{10} \\ & \ddots & \vdots & \vdots \\ & & \mathbf{R}_{CC} & \mathbf{R}_{C0} \\ & & & \mathbf{R}_{00} \end{bmatrix} \begin{matrix} r_1 \\ \vdots \\ r_C \\ r_0 \end{matrix} \end{array}. \tag{A.2}$$

As $\mathbf{K} = [\mathbf{B}\ \mathbf{A}][\mathbf{B}\ \mathbf{A}]^\mathsf{T}$, this implies that $\mathbf{K} = \mathbf{Q}(\mathbf{R}\mathbf{R}^\mathsf{T})\mathbf{Q}^\mathsf{T}$. The inverse of the middle term is $\mathbf{R}^{-\mathsf{T}}\mathbf{R}^{-1}$, and the structure of $\mathbf{R}$ and the formula for the inverse of a block upper triangular matrix imply that $\mathbf{R}^{-1}$ has the same structure as $\mathbf{R}$,

$$\mathbf{R}^{-1} = \begin{bmatrix} \mathbf{R}_{11}^{-1} & \ldots & \mathbf{0} & \mathbf{Z}_{10} \\ & \ddots & \vdots & \vdots \\ & & \mathbf{R}_{CC}^{-1} & \mathbf{Z}_{C0} \\ & & & \mathbf{R}_{00}^{-1} \end{bmatrix}$$

where $\mathbf{Z} \in \mathbb{R}_{r \times r_0}$ is

$$\mathbf{Z} = -\begin{bmatrix} \mathbf{R}_{11}^{-1} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{R}_{CC}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{10} \\ \vdots \\ \mathbf{R}_{C0} \end{bmatrix} \mathbf{R}_{00}^{-1} = \begin{bmatrix} \mathbf{Z}_{10} \\ \vdots \\ \mathbf{Z}_{C0} \end{bmatrix} \begin{matrix} r_1 \\ \vdots \\ r_C \end{matrix}.$$

Writing $\mathbf{R}_B = \mathrm{blkdiag}(\mathbf{R}_{11}, \dots, \mathbf{R}_{CC})$, it can be seen that

$$\mathbf{H}^{-1} \equiv= \mathbf{R}^{-\mathsf{T}}\mathbf{R}^{-1} = \begin{bmatrix} \mathbf{R}_B^{-\mathsf{T}}\mathbf{R}_B^{-1} & \mathbf{R}_B^{-\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{R}_B^{-1} & \mathbf{R}_{00}^{-\mathsf{T}}\mathbf{R}_{00}^{-1} \end{bmatrix}$$

and so $\mathbf{QHQ}^{\mathsf{T}}$ has the desired properties.

For the other direction, suppose $\mathbf{K} = \mathbf{QHQ}^{\mathsf{T}}$ for appropriately structured $\mathbf{Q}$ and $\mathbf{H}^{-1}$. As $\mathbf{H}$ is full rank and $\mathbf{Q}$ has orthogonal columns, we necessarily have that $\mathbf{K}$ is rank $r_0 + r$. Let $\mathbf{LL}^{\mathsf{T}}$ be the Cholesky decomposition of $\mathbf{H}^{-1}$, so that

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{T} & \mathbf{Y} \\ \mathbf{Y}^{\mathsf{T}} & \mathbf{S} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_{11} & \mathbf{0} \\ \mathbf{L}_{10} & \mathbf{L}_{00} \end{bmatrix} \begin{bmatrix} \mathbf{L}_{11}^{\mathsf{T}} & \mathbf{L}_{10}^{\mathsf{T}} \\ \mathbf{0} & \mathbf{L}_{00}^{\mathsf{T}} \end{bmatrix}$$

and the formula for the Cholesky decomposition of a $2 \times 2$ block matrix yields that $\mathbf{T} = \mathbf{L}_{11}\mathbf{L}_{11}^{\mathsf{T}}$. As $\mathbf{T}$ is block-diagonal, so too is its Cholesky factor $\mathbf{L}_{11}$. We then have that

$$\mathbf{K} = \mathbf{QHQ}^{\mathsf{T}} = [\mathbf{W}\ \mathbf{U}] \begin{bmatrix} \mathbf{L}_{11}^{-\mathsf{T}} & \mathbf{Z}^{\mathsf{T}} \\ \mathbf{0} & \mathbf{L}_{00}^{-\mathsf{T}} \end{bmatrix} \begin{bmatrix} \mathbf{L}_{11}^{-1} & \mathbf{0} \\ \mathbf{Z} & \mathbf{L}_{00}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{W}^{\mathsf{T}} \\ \mathbf{U}^{\mathsf{T}} \end{bmatrix},$$

where $\mathbf{Z} = -\mathbf{L}_{00}^{-1}\mathbf{L}_{10}\mathbf{L}_{11}^{-1}$. Taking $\mathbf{B} = \mathbf{W}\mathbf{L}_{11}^{-\mathsf{T}}$ and $\mathbf{A} = \mathbf{W}\mathbf{Z}^{\mathsf{T}} + \mathbf{U}\mathbf{L}_{00}^{-\mathsf{T}}$, we have that $\mathbf{B}$ is block diagonal with the appropriately sized blocks, and $\mathbf{K} = \mathbf{AA}^{\mathsf{T}} + \mathbf{BB}^{\mathsf{T}}$. So, $\mathbf{K} \in \mathcal{K}(\mathbf{n}, \mathbf{r})$ as desired. □

## A.2 Proof of Proposition 2

*Proof.* First, assume that $\mathbf{K}$ is in $\mathcal{K}(\mathbf{n}, \mathbf{r})$ with rank $r_0 + r$. Then $\mathbf{K} = [\mathbf{B}\ \mathbf{A}][\mathbf{B}\ \mathbf{A}]^{\mathsf{T}}$ for some appropriately structured $\mathbf{B}$ and $\mathbf{A}$. As in the proof of Proposition 1, let $\mathbf{QR}$ be the thin QR decomposition of $[\mathbf{B}\ \mathbf{A}]$, and $\mathbf{Q}$ and $\mathbf{R}$ have the structure of (A.1) and (A.2). By Proposition 1,

$\mathbf{K} = \mathbf{Q}\mathbf{H}\mathbf{Q}^\mathsf{T}$ and so $\operatorname{col}\mathbf{K} = \operatorname{col}\mathbf{Q}$. Let $\mathcal{V} = \operatorname{col}\mathbf{Q}_0$, which has dimension $r_0$ as $\mathbf{Q}_0$ is $n \times r_0$ with orthonormal columns. Then the columns of $\mathbf{Q}_c$ are a basis for $\operatorname{col}\mathbf{K} \cap \mathcal{V}^\perp \cap \mathcal{C}_c$. This follows from the facts that $\operatorname{col}\mathbf{Q}_c \subset \operatorname{col}\mathbf{K}$, $\operatorname{col}\mathbf{Q}_c \subset (\operatorname{col}\mathbf{Q}_0)^\perp$ as $\mathbf{Q}$ has orthogonal columns, and $\operatorname{col}\mathbf{Q}_c \subset \mathcal{C}_c$ by the structure of (A.2). So, $\operatorname{col}\mathbf{Q}_c \subset \operatorname{col}\mathbf{K} \cap \mathcal{V}^\perp \cap \mathcal{C}_c$, while if $\mathbf{z} \in \operatorname{col}\mathbf{K} \cap \mathcal{V}^\perp \cap \mathcal{C}_c$, then

$$\mathbf{z} = \mathbf{Q}\mathbf{w} = \begin{bmatrix} \mathbf{Q}_{11} & \dots & \mathbf{0} & \mathbf{Q}_{10} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{Q}_{CC} & \mathbf{Q}_{C0} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_C \\ \mathbf{w}_0 \end{bmatrix}$$

by $\mathbf{z} \in \operatorname{col}\mathbf{K}$ for some $\mathbf{w} \in \mathbb{R}_{r_0+r}$. As $\mathbf{z} \in \mathcal{V}^\perp$, $\mathbf{Q}_0^\mathsf{T}\mathbf{z} = [\mathbf{0}, \dots, \mathbf{0}, \mathbf{w}_0] = \mathbf{0}$ and so $\mathbf{w}_0 = \mathbf{0}$. As $\mathcal{C}_{c'}$ for $c' \neq c$ is orthogonal to $\mathcal{C}_c$, we also have $\mathbf{Q}_{c'}^\mathsf{T}\mathbf{z} = \mathbf{0}$ and so we similarly have $\mathbf{w}_{c'} = \mathbf{0}$. Therefore, $z$ in fact equals $\mathbf{Q}_c\mathbf{w}_c$ and so $z \in \operatorname{col}\mathbf{Q}_c$ which implies $\operatorname{col}\mathbf{Q}_c = \mathbf{K} \cap \mathcal{V}^\perp \cap \mathcal{C}_c$. As $\mathbf{Q}_c$ is $n \times r_c$ with orthonormal columns, we have established (1.8) while (1.7) follows from the fact that $\mathbf{Q}$ has orthonormal columns.

To show (1.9), first note that $\mathbf{P}_\mathcal{V}$ equals $\mathbf{Q}_0\mathbf{Q}_0^\mathsf{T}$. If $\mathbf{H}$ is partitioned as

$$\mathbf{H} = \begin{array}{c} \begin{matrix} r_1 & \dots & r_C & r_0 \end{matrix} \\ \begin{bmatrix} \mathbf{H}_{11} & \dots & \mathbf{H}_{1C} & \mathbf{H}_{10} \\ & \ddots & & \\ \mathbf{H}_{1C}^\mathsf{T} & \dots & \mathbf{H}_{CC} & \mathbf{H}_{C0} \\ \mathbf{H}_{10}^\mathsf{T} & \dots & \mathbf{H}_{C0}^\mathsf{T} & \mathbf{H}_{00} \end{bmatrix} \end{array} \begin{matrix} r_1 \\ \vdots \\ r_C \\ r_0, \end{matrix}$$

then $\operatorname{Var}(\mathbf{P}_\mathcal{V}\mathbf{y}) = \mathbf{Q}_0\mathbf{Q}_0^\mathsf{T}\mathbf{Q}\mathbf{H}\mathbf{Q}^\mathsf{T}\mathbf{Q}_0\mathbf{Q}_0^\mathsf{T} = \mathbf{Q}_0\mathbf{H}_{00}\mathbf{Q}_0^\mathsf{T}$. Writing $\mathbf{H}_{\cdot 0}$ for $[\mathbf{H}_{10}^\mathsf{T} \ \dots \ \mathbf{H}_{C0}^\mathsf{T} \ \mathbf{H}_{00}^\mathsf{T}]^\mathsf{T}$, we similarly have $\operatorname{Cov}(\mathbf{y}, \mathbf{P}_\mathcal{V}\mathbf{y}) = \mathbf{Q}\mathbf{H}_{\cdot 0}\mathbf{Q}_0^\mathsf{T}$. With these results, we have that the LMMSE estimator $\widehat{\mathbf{y}}$ of $\mathbf{y}$ given the projection $\mathbf{P}_\mathcal{V}\mathbf{y}$ is $\widehat{\mathbf{y}} = \mathbf{Q}\mathbf{H}_{\cdot 0}\mathbf{H}_{00}^{-1}\mathbf{Q}_0^\mathsf{T}\mathbf{y}$.

Partitioning $\widehat{\mathbf{y}}$ into $[\widehat{\mathbf{y}}_1^\mathsf{T} \ \dots \ \widehat{\mathbf{y}}_C^\mathsf{T}]^\mathsf{T}$ and using the structure of $\mathbf{Q}$, we obtain that

$$\widehat{\mathbf{y}}_c = (\mathbf{Q}_{cc}\mathbf{H}_{c0}\mathbf{H}_{00}^{-1} + \mathbf{Q}_{c0})\mathbf{Q}_0^\mathsf{T}\mathbf{y}, \quad c = 1, \dots, C.$$

For notational simplicity, let $\mathbf{G}_c$ denote the be the $n_c \times n$ regression matrix $(\mathbf{Q}_{cc}\mathbf{H}_{c0}\mathbf{H}_{00}^{-1} + \mathbf{Q}_{c0})\mathbf{Q}_0^\mathsf{T}$. To show (1.9), we need that, for $1 \leq c < k \leq C$,

$$\mathbf{0} = E[(\mathbf{y}_c - \widehat{\mathbf{y}}_c)(\mathbf{y}_k - \widehat{\mathbf{y}}_k)^\mathsf{T}] = \mathbf{K}_{ck} - \mathbf{G}_c\mathbf{K}_{\cdot k} - \mathbf{K}_{\cdot c}^\mathsf{T}\mathbf{G}_k^\mathsf{T} + \mathbf{G}_c\mathbf{K}\mathbf{G}_k^\mathsf{T},$$

where $\mathbf{K}_{ck} \equiv E[\mathbf{y}_c\mathbf{y}_k^\mathsf{T}]$ and $\mathbf{K}_{\cdot c} = E[\mathbf{y}\mathbf{y}_c^\mathsf{T}]$ are the appropriate submatrices of $\mathbf{K}$. The quantity $\mathbf{G}_c\mathbf{K}_{\cdot k}$ can be expressed in terms of $\mathbf{Q}$ and $\mathbf{H}$ as

$$\mathbf{G}_c\mathbf{K}_{\cdot k} = \mathbf{Q}_{cc}^\mathsf{T}\mathbf{H}_{c0}\mathbf{H}_{00}^{-1}\mathbf{H}_{k0}^\mathsf{T}\mathbf{Q}_{kk}^\mathsf{T} + \mathbf{Q}_{c0}\mathbf{H}_{k0}^\mathsf{T}\mathbf{Q}_{k0}^\mathsf{T} + \mathbf{Q}_{cc}\mathbf{H}_{c0}\mathbf{Q}_{k0}^\mathsf{T} + \mathbf{Q}_{c0}\mathbf{H}_{00}\mathbf{Q}_{k0}^\mathsf{T},$$

while $\mathbf{G}_c\mathbf{K}\mathbf{G}_k^\mathsf{T} = \mathbf{G}_c\mathbf{K}_{\cdot k}$ can be verified using $\mathbf{Q}_0^\mathsf{T}\mathbf{K}\mathbf{Q}_0 = \mathbf{H}_{00}$. As the structure of $\mathbf{Q}$ implies

$$\mathbf{K}_{ck} = \mathbf{Q}_{cc}\mathbf{H}_{ck}\mathbf{Q}_{kk}^\mathsf{T} + \mathbf{Q}_{cc}\mathbf{H}_{c0}\mathbf{Q}_{k0}^\mathsf{T} + \mathbf{Q}_{c0}\mathbf{H}_{k0}^\mathsf{T}\mathbf{Q}_{k0}^\mathsf{T} + \mathbf{Q}_{c0}\mathbf{H}_{00}\mathbf{Q}_{k0}^\mathsf{T},$$

combining the above results yields

$$\begin{aligned} E[(\mathbf{y}_c - \widehat{\mathbf{y}}_c)(\mathbf{y}_k - \widehat{\mathbf{y}}_k)^\mathsf{T}] &= \mathbf{K}_{ck} - \mathbf{K}_{\cdot c}^\mathsf{T}\mathbf{G}_k^\mathsf{T} \\ &= \mathbf{Q}_{cc}(\mathbf{H}_{ck} - \mathbf{H}_{c0}\mathbf{H}_{00}^{-1}\mathbf{H}_{k0}^\mathsf{T})\mathbf{Q}_{kk}^\mathsf{T}. \end{aligned}$$

However, the quantity $\mathbf{H}_{ck} - \mathbf{H}_{c0}\mathbf{H}_{00}^{-1}\mathbf{H}_{k0}^\mathsf{T}$ equals zero due to the structure of $\mathbf{H}^{-1}$ implied by Proposition 1. Recalling the expression of the inverse of a $2 \times 2$ block matrix in terms of the Schur complement of the south-east block, we have that $\mathbf{T}$, the north-west block of $\mathbf{H}^{-1}$, equals $(\mathbf{H}/\mathbf{H}_{00})^{-1}$. As $\mathbf{K} \in \mathcal{K}(\mathbf{n}, \mathbf{r})$ implies $\mathbf{T}$ is block-diagonal by Proposition 1, we obtain that $(\mathbf{H}/\mathbf{H}_{00})^{-1}$ and hence $\mathbf{H}/\mathbf{H}_{00}$ is similarly block diagonal. Noting that the quantity $\mathbf{H}_{ck} - \mathbf{H}_{c0}\mathbf{H}_{00}^{-1}\mathbf{H}_{k0}^\mathsf{T}$ is the value of the off-diagonal block in $\mathbf{H}/\mathbf{H}_{00}$ of size $n_c \times n_k$, we obtain that $\mathbf{H}_{ck} - \mathbf{H}_{c0}\mathbf{H}_{00}^{-1}\mathbf{H}_{k0}^\mathsf{T}$ for all $1 \leq c < k \leq C$, and so (1.9) is satisfied.

In the reverse direction, suppose (1.7), (1.8), and (1.9) are satisfied for some covariance $\mathbf{K} = \mathrm{Var}(\mathbf{y})$ and subspace $\mathcal{V}$ of dimension $r_0$. As (1.8) implies that $\operatorname{col}\mathbf{K}$ has dimension $r_0 + r$, we have $\operatorname{rank}\mathbf{K} = r_0 + r$. Let the columns of $\mathbf{Q}_0 \in \mathbb{R}_{n\times r_0}$ be an orthonormal basis of $\mathcal{V}$, and similarly let the columns of $\mathbf{Q}_c \in \mathbb{R}_{n\times r_c}$ be an orthonormal basis for $\operatorname{col}\mathbf{K} \cap \mathcal{V}^\perp \cap \mathcal{C}_c$ for $c = 1, \ldots, C$. Let $\mathbf{Z} \in \mathbb{R}_{n\times(n-r_0-r)}$ be an orthonormal basis for $(\operatorname{col}\mathbf{K})^\perp$. By (1.7), the matrix

$\widetilde{\mathbf{Q}} = [\mathbf{Q}_1 \ \dots \ \mathbf{Q}_C \ \mathbf{Q}_0 \ \mathbf{Z}]$ is orthogonal, and as $\operatorname{col} \mathbf{Q}_c \subset \mathbf{C}_c$ for $c = 1, \dots, C$ we also have that $[\mathbf{Q}_1, \dots, \mathbf{Q}_C] = \operatorname{blkdiag}(\mathbf{Q}_{11}, \dots, \mathbf{Q}_{CC})$ for $\mathbf{Q}_{cc} \in \mathbb{R}_{n_c \times r_c}$. The product $\widetilde{\mathbf{Q}}^\mathsf{T}\mathbf{K}\widetilde{\mathbf{Q}}$ can be partitioned as

$$\begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \equiv \widetilde{\mathbf{Q}}^\mathsf{T}\mathbf{K}\widetilde{\mathbf{Q}} = \begin{array}{c} \begin{matrix} r_1 & \dots & r_C & r_0 & n - r_0 - r \end{matrix} \\ \begin{bmatrix} \mathbf{S}_{11} & \dots & \mathbf{S}_{1C} & \mathbf{S}_{10} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{S}_{1C}^\mathsf{T} & \dots & \mathbf{S}_{CC} & \mathbf{S}_{C0} & \mathbf{0} \\ \mathbf{S}_{10}^\mathsf{T} & \dots & \mathbf{S}_{C0}^\mathsf{T} & \mathbf{S}_{00} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{matrix} r_1 \\ \vdots \\ r_C \\ r_0 \\ n - r_0 - r \end{matrix} \end{array}$$

as $\mathbf{Z}^\mathsf{T}\mathbf{K} = \mathbf{0}$ by construction. As $\mathbf{P}_\mathcal{V} = \mathbf{Q}_0\mathbf{Q}_0^\mathsf{T}$ by construction of $\mathbf{Q}_0$, we have $\operatorname{Var}(\mathbf{P}_\mathcal{V}\mathbf{y}) = \mathbf{Q}_0\mathbf{Q}_0^\mathsf{T}\mathbf{K}\mathbf{Q}_0\mathbf{Q}_0^\mathsf{T} = \mathbf{Q}_0\mathbf{S}_{00}\mathbf{Q}_0^\mathsf{T}$ and $\operatorname{Cov}(\mathbf{y}, \mathbf{P}_\mathcal{V}\mathbf{y}) = \mathbf{K}\mathbf{Q}_0\mathbf{Q}_0^\mathsf{T} = \widetilde{\mathbf{Q}}\mathbf{S}\widetilde{\mathbf{Q}}^\mathsf{T}\mathbf{Q}_0\mathbf{Q}_0^\mathsf{T} = \widetilde{\mathbf{Q}}\mathbf{S}_{\cdot 0}\mathbf{Q}_0^\mathsf{T}$, where $\mathbf{S}_{\cdot 0} = [\mathbf{S}_{10}^\mathsf{T} \ \dots \ \mathbf{S}_{C0}^\mathsf{T} \ \mathbf{S}_{00}^\mathsf{T} \ \mathbf{0}^\mathsf{T}]^\mathsf{T}$. Combining these formula and using the block-diagonal structure of the first $r$ columns of $\widetilde{\mathbf{Q}}$, we obtain

$$\hat{\mathbf{y}}_c = (\mathbf{Q}_{cc}\mathbf{S}_{c0} + \mathbf{Q}_{c0}\mathbf{S}_{00})\mathbf{S}_{00}^{-1}\mathbf{Q}_0^\mathsf{T}\mathbf{y}, \quad c = 1, \dots, C.$$

Therefore, $\mathbf{Q}_{cc}^\mathsf{T}\hat{\mathbf{y}}_c = \mathbf{S}_{c0}\mathbf{S}_{00}^{-1}\mathbf{Q}_0^\mathsf{T}\mathbf{y}$ as the columns of $\mathbf{Q}_{cc}$ and $\mathbf{Q}_{c0}$ are orthogonal. By (1.9) for all $1 \le c < k \le C$,

$$\begin{aligned} \mathbf{0} &= \mathbf{Q}_{cc}^\mathsf{T} E[(\mathbf{y}_c - \widehat{\mathbf{y}}_c)(\mathbf{y} - \widehat{\mathbf{y}}_k)^\mathsf{T}]\mathbf{Q}_{kk} = E[(\mathbf{Q}_{cc}^\mathsf{T}\mathbf{y} - \mathbf{Q}_{cc}^\mathsf{T}\widehat{\mathbf{y}}_c)(\mathbf{Q}_{kk}^\mathsf{T}\mathbf{y}_k - \mathbf{Q}_{kk}^\mathsf{T}\widehat{\mathbf{y}}_k)^\mathsf{T}] \\ &= \mathbf{Q}_{cc}^\mathsf{T} E[\mathbf{y}_c\mathbf{y}_k^\mathsf{T}]\mathbf{Q}_{kk}^\mathsf{T} - \mathbf{S}_{c0}\mathbf{S}_{00}^{-1}\mathbf{Q}_0^\mathsf{T} E[\mathbf{y}\mathbf{y}_k^\mathsf{T}]\mathbf{Q}_{kk}^\mathsf{T} - \mathbf{Q}_{cc}^\mathsf{T} E[\mathbf{y}_c\mathbf{y}^\mathsf{T}]\mathbf{S}_{00}^{-1}\mathbf{S}_{k0}^\mathsf{T} + \mathbf{S}_{c0}\mathbf{S}_{00}^{-1}\mathbf{Q}_0^\mathsf{T}\mathbf{K}\mathbf{Q}_0\mathbf{S}_{00}^{-1}\mathbf{S}_{k0}^\mathsf{T} \\ &= \mathbf{S}_{ck} - \mathbf{S}_{c0}\mathbf{S}_{00}^{-1}\mathbf{S}_{k0}^\mathsf{T} \end{aligned}$$

and so all the off-diagonal blocks of $\mathbf{S}/\mathbf{S}_{00}$ are zero. As this implies that $(\mathbf{S}/\mathbf{S}_{00})^{-1}$ is similarly block-diagonal, we obtain that $\mathbf{K} = \mathbf{Q}\mathbf{S}\mathbf{Q}^\mathsf{T}$ where $\mathbf{Q} = [\mathbf{Q}_1 \ \dots \ \mathbf{Q}_C \ \mathbf{Q}_0]$ and $\mathbf{S}^{-1}$ have the structure required in Proposition 1, and we conclude that $\mathbf{K} \in \mathcal{K}(\mathbf{n}, \mathbf{r})$ as desired. □

# Appendix B

# Supplementary Materials for Chapter 2

## B.1 Lemmas

For ease of notation, let $\mathbb{A} \oplus \mathbb{B} \subset \mathbb{R}_{n\times(r_0+r)}$ be the subspace of matrices that can be written as $[\mathbf{A}\ \mathbf{B}]$ for some $\mathbf{A} \in \mathbb{A}, \mathbf{B} \in \mathbb{B}$. The spaces $\mathbb{A} \oplus \mathbb{B}$ and $\mathbb{A} \times \mathbb{B}$ are trivially isomorphic. Further, let $(\mathbb{A} \oplus \mathbb{B})^*$ contain all FCR elements of $\mathbb{A} \oplus \mathbb{B}$. As long as $r_0 + r \leq n$ and $r_c \leq n_c$ for each $c$, $(\mathbb{A} \oplus \mathbb{B})^*$ is an open submanifold. The set $\mathbb{A}_L \oplus \mathbb{B}_L$ is defined similarly.

As many of the propositions proved here involve null sets, we distinguish between null subsets of the unrestricted loadings $\mathbb{A} \times \mathbb{B}$ and null subsets of the equivalence class representatives $\mathbb{A}^*_L \times \mathbb{B}^*_L$. A null subset of $\mathbb{A}\times\mathbb{B}$ need not correspond to a null subset of representatives. For example, $\mathbb{A}^*_L \times \mathbb{B}^*_L$ is itself null in $\mathbb{A} \times \mathbb{B}$. Lemma 2 connects the two notions.

**Lemma 2.** *If $\mathcal{C} \subset \mathbb{A} \times \mathbb{B}$ is a null set that is a union of $\sim_2$-equivalence classes, then the set of representatives $\tilde{\mathcal{C}} \subset \mathbb{A}^*_L \times \mathbb{B}^*_L$ is null in $\mathbb{A}_L \times \mathbb{B}_L$.*

**Lemma 3.** *Let $\mathbf{H} = [\mathbf{H}_1^\intercal\ \mathbf{H}_2^\intercal]^\intercal$ and $\mathbf{Z} = [\mathbf{Z}_1^\intercal\ \mathbf{Z}_2^\intercal]^\intercal$ be $m \times p$ matrices with $\mathbf{H}_1, \mathbf{Z}_1 \in \mathbb{R}_{p\times p}$. If $\mathbf{H}_1$ is invertible, $\mathbf{H}_1$ and $\mathbf{Z}_1$ are LT, then $\mathbf{H}\mathbf{Z}^\intercal + \mathbf{Z}\mathbf{H}^\intercal = \mathbf{0}$ implies $\mathbf{Z} = \mathbf{0}$.*

**Lemma 4.** *(Maximal Rank Sub-Loadings) Let $(\mathbb{A} \oplus \mathbb{B})^{**}$ be the subset of $(\mathbb{A} \oplus \mathbb{B})^*$ containing FCR $[\mathbf{A}\ \mathbf{B}]$ where, for all $c = 1, \ldots, C$, the rank of all submatrices of the channel $c$ loadings $[\mathbf{A}_c\ \mathbf{B}_c]$ are maximal. That is, if $\mathbf{D}$ is any $s \times t$ submatrix of $[\mathbf{A}_c\ \mathbf{B}_c]$, then $\operatorname{rank}(\mathbf{D}) = \min\{s, t\}$. The complement $\mathbb{A} \oplus \mathbb{B} \setminus (\mathbb{A} \oplus \mathbb{B})^{**}$ is null.*

*Proof of Lemma 2.* Let $(\mathbf{A}, \mathbf{B})$ be a random element in $\mathbb{A} \times \mathbb{B}$ with all non-zero entries being independent standard normal random variables. As $\mathcal{C}$ is null, $P\big((\mathbf{A}, \mathbf{B}) \in \mathcal{C}\big) = 0$ by the equivalence of the Gaussian and Lebesgue measures. If $\mathbf{A}_1$ is the first $r_0$ rows of $\mathbf{A}$ and $\mathbf{A}_2$ the remaining rows, then $\mathbf{A}_1$ is full rank almost surely. So, the Cholesky decomposition $\mathbf{L}\mathbf{L}^\intercal$ of $\mathbf{A}_1\mathbf{A}_1^\intercal$ has positive diagonal almost surely. As $\mathbf{A}_1\mathbf{A}_1^\intercal$ is Wishart with $r_0$ degrees of freedom and scale $\mathbf{I}_{r_0}$, the Bartlett

decomposition implies that $\mathbf{L}$ has independent entries with $[\mathbf{L}_0]_{ii} \sim \chi^2_{r_0-i+1}$ and $[\mathbf{L}_0]_{ij} \sim \mathcal{N}(0,1)$ for $i < j$. Further, we have that $\mathbf{A}_1\mathbf{Q}_0^\mathsf{T} = \mathbf{L}$ for some random orthogonal $\mathbf{Q}_0$. Although $\mathbf{Q}_0$ depends on $\mathbf{A}_1$, independence of $\mathbf{Q}_0$ and $\mathbf{A}_2$ and the sphericity of $\mathbf{A}_2$ implies that $\mathbf{A}_2\mathbf{Q}_0$ still has independent $\mathcal{N}(0,1)$ entries and is independent of $\mathbf{L}_0$. Arguing similarly for $\mathbf{B}_1, \ldots, \mathbf{B}_2$ to obtain $\mathbf{L}_1, \ldots, \mathbf{L}_C$ and $\mathbf{Q}_1, \ldots, \mathbf{Q}_C$, we have that $[\mathbf{A}\ \mathbf{B}] \stackrel{d}{=} [\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]\mathbf{Q}$ where $\tilde{\mathbf{A}} = [\mathbf{L}_0^\mathsf{T}\ \mathbf{A}_2^\mathsf{T}]^\mathsf{T}$, $\tilde{\mathbf{B}}_c = [\mathbf{L}_c^\mathsf{T}\ \mathbf{B}_{c,2}^\mathsf{T}]^\mathsf{T}$, $\tilde{\mathbf{B}} = \mathrm{blkdiag}(\tilde{\mathbf{B}}_1, \ldots, \tilde{\mathbf{B}_C})$, and $\mathbf{Q} = \mathrm{blkdiag}(\mathbf{Q}_0, \ldots, \mathbf{Q}_C)$. Therefore,

$$P\big((\mathbf{A}, \mathbf{B}) \in \mathcal{C}\big) = 0 = P\big((\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \tilde{\mathcal{C}}\big)$$

by the closure of $\mathcal{C}$ under right-multiplication by block-diagonal orthogonal matrices. From the result $P\big((\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \tilde{\mathcal{C}}\big) = 0$ and $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ has independent entries whose distributions dominate the Lebesgue measure on $\mathbb{A}_L \times \mathbb{B}_L$, we obtain that $\tilde{\mathcal{C}}$ is null in $\mathbb{A}_L \times \mathbb{B}_L$. □

*Proof of Lemma 3.* The lemma will follow from the square-matrix case with $m = p$, where $\mathbf{H}$ is a square invertible LT matrix of size $p \times p$. For any $p \times p$ LT matrix $\mathbf{Z}$, if

$$\mathbf{H}\mathbf{Z}^\mathsf{T} + \mathbf{Z}\mathbf{H}^\mathsf{T} = \mathbf{0} \tag{B.1}$$

we show that $\mathbf{Z} = \mathbf{0}$.

For $p = 1$, (B.1) is $2[\mathbf{H}]_{11}[\mathbf{Z}]_{11} = 0$, and $[\mathbf{H}]_{11} \neq 0$ as $\mathbf{H}$ is invertible, so $[\mathbf{Z}]_{11} = 0$. Suppose the lemma is true for all $1 \leq k < p$. Note that (B.1) implies that $\mathbf{H}\mathbf{Z}^\mathsf{T}$ is skew-symmetric and hence has zero diagonal. The top left entry of $\mathbf{H}\mathbf{Z}^\mathsf{T}$ is $[\mathbf{H}]_{11}[\mathbf{Z}]_{11}$, and so, as $[\mathbf{H}]_{11}$ is non-zero as $\mathbf{H}$ is invertible and LT, $[\mathbf{Z}]_{11}$ is zero. Skew-symmetry yields that $[\mathbf{H}]_{11}[\mathbf{Z}]_{i1} = -[\mathbf{H}]_{i1}[\mathbf{Z}]_{11}$, and so $[\mathbf{Z}]_{i1} = 0$ for all $i = 1, \ldots, p$. The leading columns of $\mathbf{Z}, \mathbf{H}$ can then be dropped, and the square-matrix case of the lemma follows from the induction hypothesis. To extend to the non-square case, note that the hypothesis of Lemma 2 implies $\mathbf{H}_1\mathbf{Z}_1^\mathsf{T} + \mathbf{Z}_1\mathbf{H}_1^\mathsf{T} = \mathbf{0}$, and so $\mathbf{Z}_1 = \mathbf{0}$ by the above proof for the square case. With the given partition of $\mathbf{H}$ and $\mathbf{Z}$, the upper right block of the equation $\mathbf{H}\mathbf{Z}^\mathsf{T} + \mathbf{Z}\mathbf{H}^\mathsf{T} = \mathbf{0}$ reduces to $\mathbf{H}_1\mathbf{Z}_2^\mathsf{T} = \mathbf{0}$ when $\mathbf{Z}_1 = \mathbf{0}$. As $\mathbf{H}_1$ is invertible, we obtain $\mathbf{Z}_2 = \mathbf{0}$ as well and the lemma follows. □

*Proof of Lemma 4.* Fix $c$ and index sets $\alpha \subset \{1, \ldots, n_c\}$ and $\beta \subset \{1, \ldots, r_0 + r_c\}$, and let $\det_{1,c,\alpha,\beta} : (\mathbb{A} \oplus \mathbb{B}) \to \mathbb{R}$ be the determinant of $([\mathbf{A}_c\ \mathbf{B}_c][\alpha, \beta])^{\mathsf{T}}([\mathbf{A}_c\ \mathbf{B}_c][\alpha, \beta])$. Similarly, let $\det_{2,c,\alpha,\beta}$ be the determinant of $([\mathbf{A}_c\ \mathbf{B}_c][\alpha, \beta])([\mathbf{A}_c\ \mathbf{B}_c][\alpha, \beta])^{\mathsf{T}}$. Then $\operatorname{rank}([\mathbf{A}_c\ \mathbf{B}_c][\alpha, \beta]) < \min\{|\alpha|, |\beta|\}$ only if $[\mathbf{A}_c\ \mathbf{B}_c]$ is in $\det^{-1}_{1,c,\alpha,\beta}(0) \cup \det^{-1}_{2,c,\alpha,\beta}(0)$. As $\det_{1,c,\alpha,\beta}$ and $\det_{2,c,\alpha,\beta}$ are non-zero polynomials in the matrix entries, their zero sets are null. Taking the finite union over all $c, \alpha, \beta$ proves the lemma. □

## B.2 Proof of Proposition 3

*Proof.* First, set $c = C$, so $\mathbf{M}_C$ is the submatrix of $[\mathbf{A}\ \mathbf{B}]$ obtained by dropping the rows and columns for the last channel. Partition $\mathbf{Q}$ as

$$\mathbf{Q} = \begin{array}{c} \begin{matrix} r_{<C} & r_C \end{matrix} \\ \begin{bmatrix} \mathbf{Q}_{\widehat{C}\widehat{C}} & \mathbf{Q}_{\widehat{C}C} \\ \mathbf{Q}_{C\widehat{C}} & \mathbf{Q}_{CC} \end{bmatrix} \begin{matrix} r_{<C} \\ r_C \end{matrix} \end{array}.$$

The top $n - n_C$ rows of the last $r_C$ columns of $\tilde{\mathbf{B}}$ are zero as $\tilde{\mathbf{B}} \in \mathbb{B}$, so $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}] = [\mathbf{A}\ \mathbf{B}]\mathbf{Q}$ yields

$$\mathbf{0}_{(n-n_C),r_C} = [\mathbf{M}_C\ \mathbf{0}_{(n-n_C),r_C}][\mathbf{Q}^{\mathsf{T}}_{\widehat{C}C}\ \mathbf{Q}^{\mathsf{T}}_{CC}]^{\mathsf{T}}.$$

Thus, $\mathbf{0} = \mathbf{M}_C\mathbf{Q}_{\widehat{C}C}$ and, as $\mathbf{M}_C$ is full column rank, $\mathbf{Q}_{\widehat{C}C}$ is zero. A standard result for $2 \times 2$ block orthogonal matrices (see, e.g., Horn and Johnson (1990, Lem. 2.1.10)) then implies that $\mathbf{Q}_{C\widehat{C}}$ is zero as well. So, in fact $\mathbf{Q} = \operatorname{blkdiag}(\mathbf{Q}_{\widehat{C}\widehat{C}}, \mathbf{Q}_{CC})$.

Next, for $1 \leq c < C$ suppose that $\mathbf{Q}$ is structured as

$$\mathbf{Q} = \begin{array}{c} \begin{matrix} r_{<c} & r_c & r_{>c} \end{matrix} \\ \begin{bmatrix} \mathbf{Q}_{\widehat{c}\widehat{c}} & \mathbf{Q}_{\widehat{c}c} & \mathbf{0} \\ \mathbf{Q}_{c\widehat{c}} & \mathbf{Q}_{cc} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}_{>c} \end{bmatrix} \begin{matrix} r_{<c} \\ r_c \\ r_{>c} \end{matrix} \end{array},$$

where $\mathbf{Q}_{>c} = \mathrm{blkdiag}(\mathbf{Q}_{(c+1)(c+1)}, \ldots, \mathbf{Q}_{CC})$. The $r_c$ columns of $\tilde{\mathbf{B}}$ for the channel-$c$ loadings are zero outside of the $n_c$ rows within channel $c$, so the structure for $\mathbf{Q}$ implies

$$\mathbf{0}_{(n-n_c),r_c} = [\mathbf{M}_c\ \mathbf{0}_{n-n_c,r_c}\ \mathbf{Z}][\mathbf{Q}_{\widehat{cc}}^{\mathsf{T}}\ \mathbf{Q}_{cc}^{\mathsf{T}}\ \mathbf{0}^{\mathsf{T}}]^{\mathsf{T}}$$

for $\mathbf{Z}$ the appropriate submatrix of $\mathbf{B}$. The equation above implies $\mathbf{0}_{(n-n_c),r_c} = \mathbf{M}_c\mathbf{Q}_{\widehat{cc}}$ and, as $\mathbf{M}_c$ has full column rank by assumption, we obtain that $\mathbf{Q}_{\widehat{cc}} = \mathbf{0}$. The same reasoning as in the $c = C$ case yields that $\mathbf{Q}_{c\widehat{c}} = \mathbf{0}$ by the orthogonality of $\mathbf{Q}$. Therefore, $\mathbf{Q}$ must in fact be $\mathrm{blkdiag}(\mathbf{Q}_{\widehat{cc}}, \mathbf{Q}_{cc}, \mathbf{Q}_{>c})$. Induction on $c$ yields the result. □

## B.3 Proof of Proposition 4

*Proof.* Let $\mathcal{C}$ be the subset of $\mathbb{A} \times \mathbb{B}$ where the $C$ submatrix-rank conditions of Proposition 3 are not all satisfied. First, consider $c = 1$, where $\mathbf{M}_1$ is an unstructured $(n - n_1) \times r_0$ matrix, which is full column rank except on the closed and null set $\mathcal{C}_1$ as long as $n - n_1 \geq r_0$. Next, consider $1 < c \leq C$. As $\mathbf{B}_1, \ldots, \mathbf{B}_c$ are unstructured with at least as many rows as columns, they are each full column rank except on a closed null set. So, except on a closed null set, we can form the matrix $\mathbf{G} \equiv \mathrm{blkdiag}(\mathbf{G}_1, \ldots, \mathbf{G}_{c-1})$ with

$$\mathbf{G}_k = \begin{bmatrix} \mathbf{B}_{k,1}^{-1} & \mathbf{0} \\ -\mathbf{B}_{k,2}\mathbf{B}_{k,1}^{-1} & \mathbf{I}_{n_c - r_c} \end{bmatrix},$$

for $k = 1, \ldots, (c-1)$ and $\mathbf{B}_{k,1}$ and $\mathbf{B}_{k,2}$ being the top $r_k$ rows and bottom $n_k - r_k$ rows of $\mathbf{B}_k$ respectively. Taking $\mathbf{M}'_c \equiv \mathrm{blkdiag}(\mathbf{G}, \mathbf{I}_{r_{>c}})\mathbf{M}_c$, we have $\mathrm{rank}(\mathbf{M}'_c) = \mathrm{rank}(\mathbf{M}_c)$ as $\mathbf{G}$ is non-singular. Partitioning $\mathbf{M}'_c$ as a $2 \times 2$ block matrix in the same pattern as $\mathbf{M}_c$, the upper right block of $\mathbf{M}'_c$ is

$$\mathbf{G}\mathbf{B}_{<c} = \mathrm{blkdiag}\left( \begin{bmatrix} \mathbf{I}_{r_1} \\ \mathbf{0}_{(n_1 - r_1), r_1} \end{bmatrix}, \ldots, \begin{bmatrix} \mathbf{I}_{r_c} \\ \mathbf{0}_{(n_{c-1} - r_{c-1}), r_{c-1}} \end{bmatrix} \right).$$

The upper-left block, $\mathbf{G}\mathbf{A}_{<c}$, continues to be an unstructured matrix, as we can take $\mathbf{A}'_{<c} = \mathbf{G}^{-1}\mathbf{A}_{<c}$. Let the size-$r_{<c}$ index set $\beta \subset \{1, \ldots, (n - n_c)\}$ be

$$\beta = \bigcup_{k=1}^{c-1} \{(n_{<k} + 1), \ldots, (n_{<k} + r_k)\},$$

which selects the rows corresponding to the non-zero parts of $\mathbf{G}\mathbf{B}_{<c}$ in $\mathbf{M}'_c$. Denote the complement as $\beta^c = \{1, \ldots, (n - n_c)\} \setminus \beta$. Then $\mathbf{M}'_c[\beta^c, \{1, \ldots, r_0\}]$ is an unstructured $(n - n_c - r_{<c}) \times r_0$ matrix. As $r_0 \leq n - n_c - r_{<c}$ by assumption, generically there is some index set $\alpha \subset \beta^c$ of size $r_0$ where $\mathbf{M}'_c[\alpha, \{1, \ldots, r_0\}]$ is non-singular. Then the $r_0 + r_{<c}$ rows of $\mathbf{M}'_c$ selected by $\alpha \cup \beta$ are linearly independent, as those selected by $\alpha$ have zeros in the last $r_{<c}$ columns while the rows selected by $\beta$ have a non-zero entry in distinct columns. Therefore, $\mathbf{M}_c$ is full column rank except on the closed null set $\mathcal{C}_c$.

As the subsets $\mathcal{C}_1, \ldots, \mathcal{C}_C$ excluded by the above argument are each closed and null, so too is their finite union $\mathcal{C}$. For all $c$, if $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}] = [\mathbf{A}\ \mathbf{B}]\mathbf{Q}$ for block-diagonal $\mathbf{Q}$, the corresponding submatrix $\tilde{\mathbf{M}}_c$ of $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]$ is $\mathbf{M}_c\mathbf{Q}_{<c}$ where $\mathbf{Q}_{<c}$ is the northwest $(r_0 + r_{<c}) \times (r_0 + r_{<c})$ submatrix of $\mathbf{Q}$. As $\mathbf{Q}_{<c}$ is invertible, this implies that $\operatorname{rank}(\tilde{\mathbf{M}}_c) = \operatorname{rank}(\mathbf{M}_c)$ for each $c$. This shows that $\mathcal{C}$ is a union of $\sim_2$-equivalence classes, and so Lemma 2 yields that the set of $\sim_2$ representatives $\tilde{\mathcal{C}}$ is null in $\mathbb{A}_L \times \mathbb{B}_L$. Note that $\tilde{\mathcal{C}} = \mathbb{A}_L^* \times \mathbb{B}_L^* \cap \mathcal{C}$, and so $\tilde{\mathcal{C}}$ is relatively closed in $\mathbb{A}_L^* \times \mathbb{B}_L^*$. □

## B.4 Proof of Proposition 5

*Proof.* For the top blocks $\mathbf{A}_1, \mathbf{B}_{1,1}, \ldots, \mathbf{B}_{C,1}$, let $\mathbf{L}_0\mathbf{L}_0^\mathsf{T}, \mathbf{L}_1\mathbf{L}_1^\mathsf{T}, \ldots, \mathbf{L}_C\mathbf{L}_C^\mathsf{T}$ be a Cholesky factorization of $\mathbf{A}_1\mathbf{A}_1^\mathsf{T}, \mathbf{B}_{1,1}\mathbf{B}_{1,1}^\mathsf{T}, \ldots, \mathbf{B}_{C,1}\mathbf{B}_{C,1}^\mathsf{T}$ respectively. By Gentle (1998, Ch. 3.2.2), the lower-triangular factors $\mathbf{L}_0, \ldots, \mathbf{L}_C$ with either positive diagonal entries or zero columns is unique, even in the rank-deficient case. Taking the SVD of the top blocks and lower-triangular factors respectively, there exist orthogonal $\mathbf{Q}_0, \mathbf{Q}_1, \ldots, \mathbf{Q}_C$ of the appropriate sizes such that $\mathbf{A}_1 = \mathbf{L}_0\mathbf{Q}_0$ and $\mathbf{B}_{c,1} = \mathbf{L}_c\mathbf{Q}_c$ for $c = 1, \ldots, C$. Taking $\tilde{\mathbf{A}} = [\mathbf{L}_0^\mathsf{T}\ \mathbf{Q}_0\mathbf{A}_2^\mathsf{T}]^\mathsf{T}$ and $\tilde{\mathbf{B}}_c = [\mathbf{L}_c^\mathsf{T}\ \mathbf{Q}_c\mathbf{B}_{c,2}^\mathsf{T}]^\mathsf{T}$ provides the desired $\sim_2$-equivalent representative. □

## B.5 Proof of Proposition 6

*Proof.* First, it is shown that Condition 3 is necessary. The combined matrix $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]$ in $\mathbb{A}_L \oplus \mathbb{B}_L$ can be thought of as a loading matrix for confirmatory factor analysis with certain structural zeros. In particular, $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]$ has exactly

$$K \equiv \frac{r_0(r_0-1)}{2} + \sum_{c=1}^{C} \frac{r_c(r_c-1)}{2} + \sum_{c=1}^{C} r_c(n-n_c)$$

entries constrained to be zero, and so the subspace $\mathbb{A}_L \oplus \mathbb{B}_L$ is defined by $K$ linear restrictions. By application of the result of Bekker (1986), if $K > (r_0+1)(r_0+r+1)/2$ then the equation $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}][\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]^\mathsf{T} = [\mathbf{A}\ \mathbf{B}][\mathbf{A}\ \mathbf{B}]^\mathsf{T}$ has multiple solutions with $[\mathbf{A}\ \mathbf{B}]$ in $\mathbb{A}_L \oplus \mathbb{B}_L$ for almost all values of $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]$. As $K \leq (r_0+1)(r_0+r+1)/2$ is equivalent to the condition (2.15), if almost all $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ are to be unique representatives of their $\sim_1$-equivalence class, it is required that (2.15) be satisfied.

To see that Condition 1 is sufficient, note that Propositions 4 and 3 and Lemma 2 imply that the $\sim_1$-equivalence class of $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ equals the $\sim_2$-equivalence class of $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ except on null subset of $\mathbb{A}_L^* \times \mathbb{B}_L^*$. Application of standard results on lower-triangular constraints in single-channel FA (e.g. Anderson and Rubin (1956, Sec. 5)) then yields the result. □

## B.6 Proof of Proposition 8

*Proof.* Let $\mathcal{U}$ be as defined in the proof of Theorem 1, and recall that $\mathbb{A} \oplus \mathbb{B}$ and $\mathbb{A} \times \mathbb{B}$ are isomorphic. Let $\mathcal{C}$ be the subset of $\mathbb{A} \times \mathbb{B}$ where the full-rank submatrices condition of Proposition 3 is violated. As the definition of $\mathcal{U}$ (B.2) depends only on $\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T}$, it is clear that $\mathcal{U}$ is a union of $\sim_2$-equivalence classes. As $\mathcal{C}$ and $\mathcal{U}$ are null by Proposition 3 and Theorem 1 respectively and both are unions of $\sim_2$ equivalence classes, then their intersection $\mathcal{C} \cap \mathcal{U}$ is a null union of $\sim_2$ equivalence classes. By Lemma 2, the set of representatives corresponding to $\mathcal{C} \cap \mathcal{U}$ is null in $\mathbb{A}_L^* \times \mathbb{B}_L^*$. As $V$ and $\mathbb{A}_L^* \times \mathbb{B}_L^* \times \mathbb{P}_n$ are isomorphic, the set of $\boldsymbol{\eta} \in V$ with $(\mathbf{A}(\boldsymbol{\eta}), \mathbf{B}(\boldsymbol{\eta})) \in \mathcal{C} \cap \mathcal{U}$ is null as well. If $\tilde{V}$ is the complement of said set of $\boldsymbol{\eta}$, $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is injective on $\tilde{V}$. □

## B.7 Proof of Theorem 1

*Proof.* Define $\mathcal{B} \subset \mathbb{A} \times \mathbb{B} \times \mathbb{P}_n$ as the subset where $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ does not have globally identified noise variances. For any $\mathbf{\Phi}' \succeq \mathbf{\Phi}$, if $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ is in $\mathcal{B}$ then so too is $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}')$ as $\mathbf{R}_{\mathbf{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}) = \mathbf{R}_{\mathbf{xx}}(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{\Phi}})$ implies $\mathbf{R}_{\mathbf{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi} + [\mathbf{\Phi}' - \mathbf{\Phi}]) = \mathbf{R}_{\mathbf{xx}}(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{\Phi}} + [\mathbf{\Phi}' - \mathbf{\Phi}])$ with $\mathbf{\Phi}' \neq \tilde{\mathbf{\Phi}} + [\mathbf{\Phi}' - \mathbf{\Phi}]$. As the set of $\mathbf{\Phi}'$ greater than $\mathbf{\Phi}$ has positive Lebesgue measure in $\mathrm{Diag}(n)$, $\mathcal{B}$ is null in $\mathbb{A} \times \mathbb{B} \times \mathbb{P}_n$ iff its projection onto $\mathcal{A} \times \mathcal{B}$ is null. That this projection is null is shown in the remainder of the proof.

Let $\mathcal{U} \subset \mathbb{A} \oplus \mathbb{B}$ contain those $[\mathbf{A}\ \mathbf{B}]$ such that there exists some other $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]$ and a diagonal $\boldsymbol{\varphi} = \mathbf{\Phi} - \tilde{\mathbf{\Phi}}$ with

$$\mathbf{A}\mathbf{A}^{\mathsf{T}} + \mathbf{B}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\varphi} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\mathsf{T}} + \tilde{\mathbf{B}}\tilde{\mathbf{B}}^{\mathsf{T}}, \qquad \boldsymbol{\varphi} \neq \mathbf{0}. \tag{B.2}$$

If $[\mathbf{A}\ \mathbf{B}] \in \mathcal{U}$ for $\boldsymbol{\varphi} = \mathbf{\Phi} - \tilde{\mathbf{\Phi}} \neq 0$, then $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ is in $\mathcal{B}$, and so showing $\mathcal{U}$ to be null will imply that $\mathcal{B}$ is null.

To show $\mathcal{U}$ is null under the conditions of Theorem 1, $\mathcal{U}$ is partitioned into a number of cases. Let $\mathcal{U}^*$ contain the elements of $\mathcal{U}$ satisfying the maximal rank condition of Lemma 4, $\mathcal{U}^* = (\mathbb{A} \oplus \mathbb{B})^{**} \cap \mathcal{U}$. Next, note that $\mathcal{U}^*$ can be written as the finite union of $\mathcal{U}^*_\beta$ where $\beta \subset \{1, \ldots, n\}$ is a proper subset of the possible indices. The subset $\mathcal{U}_\beta$ is obtained by adding the restriction that $[\boldsymbol{\varphi}]_{ii} = 0$ for all $i \in \beta$ and $[\boldsymbol{\varphi}]_{jj} \neq 0$ for $j \in \beta^c$ to (B.2). The proof proceeds in two steps. In the first step, $\beta$ is the empty set and so $\boldsymbol{\varphi}$ is non-singular. Results from differential geometry will imply that $\mathcal{U}^*_\emptyset$ is null. In the second step, the diagonal of $\boldsymbol{\varphi}$ has zeros, which enables reduction to the invertible case with smaller $\mathbf{n}$ and $\mathbf{r}$.

**Case 1: $\varphi$ non-singular**

In the primary case, $\beta = \emptyset$ and so $\boldsymbol{\varphi}$ is non-singular. To eliminate the quantification over $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ in the definition of $\mathcal{U}$, construct the block matrices

$$\mathbf{M} = \begin{bmatrix} \boldsymbol{\varphi} & \mathbf{A} & \mathbf{B} \\ \mathbf{A}^\mathsf{T} & -\mathbf{I}_{r_0} & \mathbf{0} \\ \mathbf{B}^\mathsf{T} & \mathbf{0} & -\mathbf{I}_r \end{bmatrix}, \quad \mathbf{M}_c = \begin{bmatrix} \boldsymbol{\varphi}_c & \mathbf{A}_c & \mathbf{B}_c \\ \mathbf{A}_c^\mathsf{T} & -\mathbf{I}_{r_0} & \mathbf{0} \\ \mathbf{B}_c^\mathsf{T} & \mathbf{0} & -\mathbf{I}_{r_c} \end{bmatrix},$$

for $c = 1, \ldots, C$, where $\boldsymbol{\varphi}_c$ is the submatrix of $\boldsymbol{\varphi}$ in the $c$th channel. Additivity of rank with respect to the Schur complement implies

$$\begin{aligned} r_0 + r + \operatorname{rank}(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T} + \tilde{\mathbf{B}}\tilde{\mathbf{B}}^\mathsf{T}) &= n + \operatorname{rank}(\mathbf{H}), \\ r_0 + r_c + \operatorname{rank}(\tilde{\mathbf{A}}_c\tilde{\mathbf{A}}_c^\mathsf{T} + \tilde{\mathbf{B}}_c\tilde{\mathbf{B}}_c^\mathsf{T}) &= n_c + \operatorname{rank}(\mathbf{H}_c), \end{aligned} \tag{B.3}$$

for $c = 1, \ldots, C$, where $\mathbf{H}$ is

$$\mathbf{H}(\mathbf{A}, \mathbf{B}, \boldsymbol{\varphi}) = \begin{bmatrix} \mathbf{I}_{r_0} + \mathbf{A}^\mathsf{T}\boldsymbol{\varphi}^{-1}\mathbf{A} & \mathbf{A}^\mathsf{T}\boldsymbol{\varphi}^{-1}\mathbf{B} \\ \mathbf{B}^\mathsf{T}\boldsymbol{\varphi}^{-1}\mathbf{A}^\mathsf{T} & \mathbf{I}_r + \mathbf{B}^\mathsf{T}\boldsymbol{\varphi}^{-1}\mathbf{B} \end{bmatrix}, \tag{B.4}$$

and similarly $\mathbf{H}_c$ is

$$\mathbf{H}_c(\mathbf{A}_c, \mathbf{B}_c, \boldsymbol{\varphi}_c) = \mathbf{I}_{r_0 + r_c} + [\mathbf{A}_c\ \mathbf{B}_c]^\mathsf{T}\boldsymbol{\varphi}_c^{-1}[\mathbf{A}_c\ \mathbf{B}_c].$$

Further, the block-diagonal structure of $\mathbf{B}$ yields that the lower-right part of $\mathbf{H}$ is block-diagonal with blocks of size $r_c \times r_c, c = 1, \ldots, C$. The lower-right part of $\mathbf{H}_c$ equals the $c$th block of the lower-right part of $\mathbf{H}$. Combining the bounds

$$\begin{aligned} \operatorname{rank}(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T} + \tilde{\mathbf{B}}\tilde{\mathbf{B}}^\mathsf{T}) &\leq r_0 + r, \\ \operatorname{rank}(\tilde{\mathbf{A}}_c\tilde{\mathbf{A}}_c^\mathsf{T} + \tilde{\mathbf{B}}_c\tilde{\mathbf{B}}_c^\mathsf{T}) &\leq \min(n_c, r_0 + r_c), \end{aligned}$$

with (B.3) yields that

$$\begin{aligned} \operatorname{rank}\left(\mathbf{H}(\mathbf{A}, \mathbf{B}, \boldsymbol{\varphi})\right) &\leq 2(r_0 + r) - n, \\ \operatorname{rank}(\mathbf{I}_{r_c} + \mathbf{B}_c^\mathsf{T}\boldsymbol{\varphi}_c^{-1}\mathbf{B}_c) &\leq \min\{r_c, 2(r_0 + r_c) - n_c\}, \end{aligned} \tag{B.5}$$

for $c = 1, \ldots, C$. To establish the second line of (B.5), we combine (B.3) with the above bounds to yield that $\operatorname{rank}(\mathbf{H}_c) \leq \min\{r_0 + r_c, 2(r_0 + r_c) - n_c\}$. As $\mathbf{I}_{r_c} + \mathbf{B}_c^\mathsf{T} \boldsymbol{\varphi}_c^{-1} \mathbf{B}_c$ is the lower-right block of $\mathbf{H}_c$ of size $r_c \times r_c$, its rank is bounded above by the minimum of the block size and rank of the whole matrix as $\min\{r_c, \min\{r_0 + r_c, 2(r_0 + r_c) - n_c)\}$. This expression then equals the RHS of the second line of (B.5).

Therefore, showing that the set of $[\mathbf{A}\ \mathbf{B}]$ where $\mathbf{H}(\mathbf{A}, \mathbf{B}, \boldsymbol{\varphi})$ satisfies (B.5) for some invertible $\boldsymbol{\varphi}$ is null implies that $\mathcal{U}_\emptyset^*$ is null as well. If either $2(r_0+r)-n < 0$ or there is a $c$ with $2(r_0+r_c)-n_c < 0$, a bound in (B.5) is negative and so $\mathcal{U}_\emptyset$ is empty. For the remainder of this case, assume that $2(r_0 + r) \geq n$ and $2(r_0 + r_c) \geq n_c$ for all $c$.

For the codomain of $\mathbf{H}$, let $\mathcal{S} \subset \operatorname{Sym}(r_0 + r)$ contain the vector space of all symmetric matrices whose $r \times r$ lower-right part is block-diagonal with blocks of sizes $r_c \times r_c$ for $c = 1, \ldots, C$, which has dimension

$$\dim \mathcal{S} = \tfrac{r_0(r_0+1)}{2} + r_0 r + \textstyle\sum_{c=1}^{C} \tfrac{r_c(r_c+1)}{2}.$$

Recall that both $(\mathbb{A} \oplus \mathbb{B})^{**}$, which contains $[\mathbf{A}\ \mathbf{B}]$ satisfying the maximal rank submatrix condition of Lemma 4, and the non-singular diagonal matrices $\mathbb{P}_n^{>0}$ are open submanifolds of their respective vector spaces. So, $\mathbf{H}$ is a smooth map from the product manifold $(\mathbb{A} \oplus \mathbb{B})^{**} \times \mathbb{P}_n^{>0}$ to $\mathcal{S}$. The differential $d\mathbf{H}$ of this map at $(\mathbf{A}, \mathbf{B}, \boldsymbol{\varphi})$ is

$$\begin{aligned} d\mathbf{H} = {} & [\mathbf{A}\ \mathbf{B}]^\mathsf{T} \boldsymbol{\varphi}^{-1} [d\mathbf{A}\ d\mathbf{B}] + [d\mathbf{A}\ d\mathbf{B}]^\mathsf{T} \boldsymbol{\varphi}^{-1} [\mathbf{A}\ \mathbf{B}] \\ & + [\mathbf{A}\ \mathbf{B}]^\mathsf{T} \boldsymbol{\varphi}^{-1} d\boldsymbol{\varphi} \boldsymbol{\varphi}^{-1} [\mathbf{A}\ \mathbf{B}]. \end{aligned}$$

The differential is surjective. To see this, consider the subspace of tangent vectors $(d\mathbf{A}, d\mathbf{B}, \mathbf{0})$ with $[d\mathbf{A}\ d\mathbf{B}] = \boldsymbol{\varphi}^{-1}[\mathbf{A}\ \mathbf{B}]\mathbf{L}$, for $\mathbf{L} \in \mathbb{R}_{(r_0+r)\times(r_0+r)}$ lower-triangular and partitioned as $\mathbf{L} = [\mathbf{L}_{11}\ \mathbf{0}; \mathbf{L}_{12}\ \mathbf{L}_{22}]$, where $\mathbf{L}_{22}$ is block-diagonal with $c$th block of size $r_c \times r_c$. The space of such $\mathbf{L}$ has equal dimension to $\mathcal{S}$. It can be verified that $\boldsymbol{\varphi}^{-1}[\mathbf{A}\ \mathbf{B}]\mathbf{L}$ has structural zeros in the appropriate places, and so is a valid choice for $[d\mathbf{A}\ d\mathbf{B}]$. On this subspace, $d\mathbf{H}$ is injective. Suppose that, for some $\mathbf{L}$ with the above structure,

$$d\mathbf{H} = \mathbf{0} = [\mathbf{A}\ \mathbf{B}]^\mathsf{T} \boldsymbol{\varphi}^{-2} [\mathbf{A}\ \mathbf{B}]\mathbf{L} + \mathbf{L}^\mathsf{T} [\mathbf{A}\ \mathbf{B}]^\mathsf{T} \boldsymbol{\varphi}^{-2} [\mathbf{A}\ \mathbf{B}].$$

The matrix $[\mathbf{A}\ \mathbf{B}]^\mathsf{T}\boldsymbol{\varphi}^{-2}[\mathbf{A}\ \mathbf{B}]$ is positive-definite as $\boldsymbol{\varphi}^{-2}$ is positive and $[\mathbf{A}\ \mathbf{B}]$ is FCR. Therefore, the matrix $[\mathbf{A}\ \mathbf{B}]^\mathsf{T}\boldsymbol{\varphi}^{-2}[\mathbf{A}\ \mathbf{B}]$ has Cholesky-type decomposition $\mathbf{U}\mathbf{U}^\mathsf{T}$ with $\mathbf{U}$ upper-triangular and non-singular. This is obtained by taking the usual Cholesky decomposition of the original matrix with the order of rows and columns reversed. So, the above equation can be manipulated to yield $\mathbf{U}^\mathsf{T}\mathbf{L}(\mathbf{U}^{-1})^\mathsf{T} = -\mathbf{U}^{-1}\mathbf{L}^\mathsf{T}\mathbf{U}$ and so $\mathbf{U}^\mathsf{T}\mathbf{L}(\mathbf{U}^{-1})^\mathsf{T}$ is skew-symmetric. However, as the LHS is lower-triangular while the right is upper-triangular, we must also have that $\mathbf{U}^\mathsf{T}\mathbf{L}(\mathbf{U}^{-1})^\mathsf{T}$ diagonal. Diagonal skew-symmetric matrices must be zero, and $\mathbf{U}^\mathsf{T}\mathbf{L}(\mathbf{U}^{-1})^\mathsf{T}$ implies $\mathbf{L} = \mathbf{0}$ as $\mathbf{U}$ is invertible. So, $d\mathbf{H}$ is injective on a subspace with the same dimension as the codomain, and so $d\mathbf{H}$ is surjective. As $(\mathbf{A}, \mathbf{B}, \boldsymbol{\varphi}) \in (\mathbb{A} \oplus \mathbb{B})^{**} \times \mathbb{P}_n^{>0}$ was arbitrary, $\mathbf{H}$ is a smooth submersion.

Next, we will show that the subset of $\mathcal{S}$ where the rank conditions (B.5) are satisfied can be written as a union of embedded submanifolds. For any index set $\alpha \subset \{1, \ldots, (r_0 + r)\}$, relate $\alpha$ to the block-structure of $\mathbf{H}$ by defining $\rho_0 \equiv |\alpha \cap \{1, \ldots, r_0\}|$ and $\rho_c \equiv |\alpha \cap \{(r_{<c}+1), \ldots, (r_{<c}+r_c)\}|$ for each $c$. Then for those $\alpha$ where $\boldsymbol{\rho} \equiv [\rho_0, \rho_1, \ldots, \rho_C]$ satisfies

$$\begin{aligned} \rho_0 + \textstyle\sum_{c=1}^{C} \rho_c &\leq 2(r_0 + r) - n, \\ \rho_c &\leq 2(r_0 + r_c) - n_c, \end{aligned} \tag{B.6}$$

let $\mathcal{S}_\alpha$ be the subset of $\mathbf{S} \in \mathcal{S}$ where the principal submatrix $\mathbf{S}[\alpha, \alpha]$ is non-singular and $\operatorname{rank}(\mathbf{S}) = |\alpha|$. If $\mathbf{H}(\mathbf{A}, \mathbf{B}, \boldsymbol{\varphi})$ satisfies (B.5), then $\mathbf{H}(\mathbf{A}, \mathbf{B}, \boldsymbol{\varphi}) \in \mathcal{S}_\alpha$ for some $\alpha$ satisfying (B.6). For $\mathbf{S} \in \mathcal{S}_\alpha$, symmetry and invertibility of $\mathbf{S}[\alpha, \alpha]$ implies the complementary submatrix, $\mathbf{S}[\alpha^c, \alpha^c]$, is a smooth function of $\mathbf{S}[\alpha, \alpha]$ and $\mathbf{S}[\alpha, \alpha^c]$. This fact is equivalent to the well-known matrix completion result that, for $\mathbf{C} \in \operatorname{Sym}(n)$ partitioned as $\mathbf{C} = [\mathbf{C}_{11}\ \mathbf{C}_{12};\ \mathbf{C}_{12}^\mathsf{T}\ \mathbf{C}_{22}]$ with $\mathbf{C}_{11}$ invertible and $\operatorname{rank}(\mathbf{C}_{11}) = \operatorname{rank}(\mathbf{C})$, we have $\mathbf{C}_{22} = \mathbf{C}_{12}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}^\mathsf{T}$.

For all $\mathbf{S} \in \mathcal{S}_\alpha$, the submatrices $\mathbf{S}[\alpha, \alpha]$ and $\mathbf{S}[\alpha, \alpha^c]$ inherit structural zeros from $\mathcal{S}$ which are determined by $\alpha$. Define the vector space $\mathcal{W}_\alpha \subset \operatorname{Sym}(\rho_0 + \rho)$ containing all symmetric matrices with structural zeros in locations which match those of $\mathbf{S}[\alpha, \alpha]$, and similarly let $\mathcal{Y}_\alpha \subset \mathbb{R}_{\rho \times (r_0 + r - \rho)}$ be the vector space containing all matrices with structural zeros matching those of $\mathbf{S}[\alpha, \alpha^c]$. As $\mathbf{I}_{|\alpha|}$ is in $\mathcal{W}_\alpha$, the subset $\mathcal{W}_\alpha^*$ containing only non-singular matrices is the preimage of $\det_\alpha^{-1}(\mathbb{R} \setminus \{0\})$

and so is a non-empty open submanifold of the same dimension as $\mathcal{W}_\alpha$. Here $\det_\alpha : \mathcal{W}_\alpha \to \mathbb{R}$ is the usual determinant with domain restricted to $\mathcal{W}_\alpha$. The dimensions of these spaces are

$$\begin{aligned} \dim(\mathcal{W}_\alpha) &= \tfrac{\rho_0(\rho_0+1)}{2} + \rho_0\rho + \textstyle\sum_{c=1}^{C} \tfrac{\rho_c(\rho_c+1)}{2}, \\ \dim(\mathcal{Y}_\alpha) &= \rho_0(r_0 + r - \rho_0) - \rho_0\rho + \textstyle\sum_{c=1}^{C} \rho_c(r_c - \rho_c). \end{aligned} \tag{B.7}$$

Then there is the obvious embedding from the product manifold $\mathcal{W}_\alpha^* \times \mathcal{Y}_\alpha$ into $\mathcal{S}$ obtained by setting $\mathbf{S}[\alpha, \alpha]$ equal to the first component, $\mathbf{S}[\alpha, \alpha^c]$ and $\mathbf{S}[\alpha^c, \alpha]$ to the second component and its transpose respectively, then smoothly obtaining $\mathbf{S}[\alpha^c, \alpha^c]$ as the unique low-rank matrix completion. So, $\mathcal{S}_\alpha$ can be treated as an embedded submanifold of dimension $\dim(\mathcal{W}_\alpha) + \dim(\mathcal{Y}_\alpha)$.

As $\mathcal{S}_\alpha$ is an embedded submanifold, $\mathbf{H}$ is automatically transverse for $\mathcal{S}_\alpha$ by virtue of being a submersion. So, a standard application of Sard's Theorem (see, e.g., Lee (2002, Thm. 6.30)) ensures that $\mathbf{H}^{-1}(\mathcal{S}_\alpha)$ is an embedded submanifold of $(\mathbb{A} \oplus \mathbb{B})^{**} \times \mathbb{P}_n^{>0}$ with codimension equal to the codimension of $\mathcal{S}_\alpha$ in $\mathcal{S}$, namely

$$\begin{aligned} \operatorname{codim} \mathcal{S}_\alpha = {} & \tfrac{r_0(r_0+1) - \rho_0(\rho_0+1)}{2} + \textstyle\sum_{c=1}^{C} \tfrac{r_c(r_c+1) - \rho_c(\rho_c+1)}{2} \\ & + (r_0 - \rho_0)r - \rho_0(r_0 - \rho_0) - \textstyle\sum_{c=1}^{C} \rho_c(r_c - \rho_c). \end{aligned}$$

Let $\pi$ be the projection map from $(\mathbb{A} \oplus \mathbb{B})^{**} \times \mathbb{P}_n^{>0}$ to $(\mathbb{A} \oplus \mathbb{B})^{**}$. Dimensional considerations (Lee, 2002, p 131) then imply that $\pi\left(\mathbf{H}^{-1}(\mathcal{S}_\alpha)\right)$ is null if

$$\dim(\mathbb{A} \oplus \mathbb{B})^* + n - \operatorname{codim} \mathcal{S}_\alpha < \dim(\mathbb{A} \oplus \mathbb{B})^*. \tag{B.8}$$

If the above inequality is satisfied for all $\alpha$ with $\boldsymbol{\rho}$ meeting (B.6), then $\mathcal{U}_\beta^*$ is null for $\beta = \emptyset$. This follows as for any $[\mathbf{A}\ \mathbf{B}]$ in $\mathcal{U}_\emptyset$, $\mathbf{H}(\mathbf{A}, \mathbf{B}, \boldsymbol{\varphi})$ is in $\mathcal{S}_\alpha$ for some $\boldsymbol{\varphi}$ and some $\alpha$ satisfying (B.6). Hence $\mathcal{U}_\emptyset^*$ is a subset of $\bigcup_\alpha \pi\left(\mathbf{H}^{-1}(\mathcal{S}_\alpha)\right)$, and the latter is a finite union of null sets.

**Case 2: $\boldsymbol{\varphi}$ singular**

We will show that $\mathcal{U}_\beta$ with $|\beta| > 0$ is also null by reducing to the non-singular case with smaller channel sizes and factor numbers. To do so, let $j_c \equiv |\beta \cap \{(r_{<c} + 1), \ldots, (r_{<c} + r_c)\}|$ be the

number of zeros in the $c$th channel on the diagonal of $\boldsymbol{\varphi}$ and let the channels be numbered such that $j_1 \geq j_2 \geq \cdots \geq j_C$.

For the first channel, we can take $\boldsymbol{\varphi} = \mathrm{Diag}(\mathbf{0}_{j_1}, \boldsymbol{\varphi}')$ without loss of generality by permuting $\mathbf{x}_1$. Continuing to let $\mathbf{A}_1$ and $\mathbf{B}_1$ be the common and distinct factor loadings for channel 1 respectively, define the submatrices $\mathbf{A}_{11}$ and $\mathbf{B}_{11}$ containing the first $j_1$ rows of $\mathbf{A}_1$ and $\mathbf{B}_1$ respectively. Similarly define $\tilde{\mathbf{A}}_{11}$ and $\tilde{\mathbf{B}}_{11}$ with respect to $\tilde{\mathbf{A}}_1$ and $\tilde{\mathbf{B}}_1$. The remaining submatrices $\mathbf{A}_{12}, \mathbf{B}_{12}$ and $\tilde{\mathbf{A}}_{12}, \tilde{\mathbf{B}}_{12}$ contain the last $n_1 - j_1$ rows of $\mathbf{A}_1, \mathbf{B}_1$ and $\tilde{\mathbf{A}}_1, \tilde{\mathbf{B}}_1$ respectively. Finally, let $\mathbf{A}' = [\mathbf{A}_{12}^\mathsf{T}\ \mathbf{A}_2^\mathsf{T}\ \ldots\ \mathbf{A}_C]^\mathsf{T}$ and $\mathbf{B}' = \mathrm{blkdiag}(\mathbf{B}_{12}, \mathbf{B}_2, \ldots, \mathbf{B}_C)$ be the loadings after exclusion of the top $j_1$ rows, with $\tilde{\mathbf{A}}'$ and $\tilde{\mathbf{B}}'$ being similar.

With these definitions, the zeros of $\boldsymbol{\varphi}$ show that (B.2) implies

$$\mathbf{A}_{11}\mathbf{A}_{11}^\mathsf{T} + \mathbf{B}_{11}\mathbf{B}_{11}^\mathsf{T} = \tilde{\mathbf{A}}_{11}\tilde{\mathbf{A}}_{11}^\mathsf{T} + \tilde{\mathbf{B}}_{11}\tilde{\mathbf{B}}_{11}^\mathsf{T}. \tag{B.9}$$

As $\boldsymbol{\varphi}$ is diagonal, the off-diagonal blocks are also equal,

$$[\mathbf{A}_{11}\ \mathbf{B}_{11}\ \mathbf{0}_{r_{>1}}][\mathbf{A}'\ \mathbf{B}']^\mathsf{T} = [\tilde{\mathbf{A}}_{11}\ \tilde{\mathbf{B}}_{11}\ \mathbf{0}_{r_{>1}}][\tilde{\mathbf{A}}'\ \tilde{\mathbf{B}}']^\mathsf{T}. \tag{B.10}$$

Next, recall that the generalized Schur complement of $\mathbf{A}_{11}\mathbf{A}_{11}^\mathsf{T} + \mathbf{B}_{11}\mathbf{B}_{11}^\mathsf{T}$ in $\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T}$ is

$$\mathbf{A}'\mathbf{A}' + \mathbf{B}'\mathbf{B}' - [\mathbf{A}'\ \mathbf{B}']\mathbf{W}[\mathbf{A}'\ \mathbf{B}']^\mathsf{T}, \tag{B.11}$$

where $\mathbf{W}$ is defined as

$$\mathbf{W} = [\mathbf{A}_{11}\mathbf{B}_{11}\ \mathbf{0}]^\mathsf{T}(\mathbf{A}_{11}\mathbf{A}_{11}^\mathsf{T} + \mathbf{B}_{11}\mathbf{B}_{11}^\mathsf{T})^{-}[\mathbf{A}_{11}\mathbf{B}_{11}\ \mathbf{0}],$$

with $(\cdot)^-$ being the Moore-Penrose psuedo-inverse. As $\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T}$ is positive semi-definite, the generalized Schur complement is uniquely defined (Zhang, 2005, Ch. 6), and so (B.11) equals

$$[\mathbf{A}'\ \mathbf{B}'] \begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r_{>1}} \end{bmatrix} \begin{bmatrix} \mathbf{A}'^\mathsf{T} \\ \mathbf{B}'^\mathsf{T} \end{bmatrix}, \tag{B.12}$$

where $\mathbf{P}$ is the orthogonal projection onto $\mathrm{Ker}([\mathbf{A}_{11}\ \mathbf{B}_{11}])$, which has dimension $(r_0 + r_1 - j_1)_+$ by maximal rank submatrix condition. To represent $\mathbf{P}$, note that $\dim \mathrm{Ker}(\mathbf{B}_{11})$ is $r_1' \equiv (r_1 - j_c)_+$ as

$\mathbf{B}_{11}$ is a $j_1 \times r_1$ submatrix of $[\mathbf{A}_1\ \mathbf{B}_1]$. Choosing $\mathbf{v}'_1, \ldots, \mathbf{v}'_{r'_1}$ as an orthogonal basis for $\mathrm{Ker}(\mathbf{B}_{11})$, the vectors $\mathbf{v}_i = [\mathbf{0}_{r_0}^\mathsf{T}\ \mathbf{v}_i]^\mathsf{T}$ are also in $\mathrm{Ker}([\mathbf{A}_{11}\ \mathbf{B}_{11}])$. To the list $\mathbf{v}_1, \ldots, \mathbf{v}_{r'_1}$, we can extend to an orthogonal basis of $\mathrm{Ker}([\mathbf{A}_{11}\ \mathbf{B}_{11}])$ by adding $\mathbf{w}_1, \ldots, \mathbf{w}_{r'_0}$ where

$$r'_0 \equiv (r_0 + r_1 - j_1)_+ - (r_1 - j_1)_+ = [r_0 - (j_1 - r_1)_+]_+.$$

If $\mathbf{W} = [\mathbf{w}_1\ \ldots\ \mathbf{w}_{r'_0}]$ and $\mathbf{V}' = [\mathbf{v}'_1\ \ldots\ \mathbf{v}_{r'_1}]$, then if

$$\mathbf{D} \equiv \begin{array}{c} \begin{array}{cccc} r'_0 & r_0 - r'_0 & r'_1 & r_1 - r'_1 \end{array} \\ \begin{bmatrix} \mathbf{W}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_1 & \mathbf{0} & \mathbf{V}' & \mathbf{0} \end{bmatrix} \end{array} \begin{array}{c} r_0 \\ r_1 \end{array},$$

the projection $\mathbf{P}$ equals $\mathbf{D}\mathbf{D}^\mathsf{T}$, where $\mathbf{W}_0$ and $\mathbf{W}_1$ contain the first $r_0$ rows and the remaining $r_1$ rows of $\mathbf{W}$ respectively.

So, the inner term in (B.12) factors into $\mathbf{R}\mathbf{R}^\mathsf{T}$ where $\mathbf{R} = [\mathbf{D}\ \mathbf{0};\ \mathbf{0}\ \mathbf{I}_{r_{>1}}]$. As $\mathbf{R}$ has a block lower-triangular form, it can be verified that $[\mathbf{A}'\ \mathbf{B}']\mathbf{R}$ continues to have the appropriate channel structure. The generalized Schur complement can then be represented as $\mathbf{A}_o\mathbf{A}_o^\mathsf{T} + \mathbf{B}_o\mathbf{B}_o^\mathsf{T}$ where $[\mathbf{A}_o\ \mathbf{B}_o]$ is obtained from $[\mathbf{A}'\ \mathbf{B}']\mathbf{R}$ by dropping the zero columns.

Using this representation, we can take the lower $(n - j_1) \times (n - j_1)$ block of (B.2) and subtract $[\mathbf{A}'\ \mathbf{B}']\mathbf{W}[\mathbf{A}'\ \mathbf{B}']$ from both sides. By the equalities (B.9) and (B.10), this implies the relation between the generalized Schur complements,

$$\mathbf{K} \setminus \mathbf{K}_{11} = \tilde{\mathbf{K}} \setminus \tilde{\mathbf{K}}_{11} + \boldsymbol{\varphi}'$$

where $\mathbf{K} \equiv \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T}$ and $\mathbf{K}_{11} \equiv \mathbf{A}_{11}\mathbf{A}_{11}^\mathsf{T} + \mathbf{B}_{11}\mathbf{B}_{11}^\mathsf{T}$, and $\tilde{\mathbf{K}}, \tilde{\mathbf{K}}_{11}$ are defined similarly using $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}$. As discussed above, this implies there are $\mathbf{A}^o, \mathbf{B}^o$ and $\tilde{\mathbf{A}}^o, \tilde{\mathbf{B}}^o$ such that

$$\mathbf{A}^o\mathbf{A}^{o\mathsf{T}} + \mathbf{B}^o\mathbf{B}^{o\mathsf{T}} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^{o\mathsf{T}} + \tilde{\mathbf{B}}^o\tilde{\mathbf{B}}^{o\mathsf{T}} + \boldsymbol{\varphi}'.$$

In the case with $j_c = 0$ for $c > 1$, The above procedure exhibits a reduction from $\mathcal{U}^*_\beta$ into $\mathcal{U}'^*_\emptyset$ where $\mathcal{U}'$ is the set of loadings satisfying (B.2) for channel sizes $n_1 - j_1, n_2, \ldots, n_C$ and factor numbers

$r'_0, r'_1, r_2, \ldots, r_C$. As the orthogonal projection $\mathbf{P}$ varies smoothly with $[\mathbf{A}_{11}\ \mathbf{B}_{11}]$, the matrix $\mathbf{D}$ can be chosen to smoothly vary in $[\mathbf{A}_{11}\ \mathbf{B}_{11}]$ and so the reduction is smooth.

In other cases, the above procedure can be iterated to remove $j_c$ zeros from $\boldsymbol{\varphi}_c$ each time, yielding a smooth reduction from $\mathcal{U}^*_\beta$ to $\mathcal{U}'^*_\emptyset$. Showing $\mathcal{U}^*_\beta$ to be null reduces to showing $\mathcal{U}'^*_\emptyset$ to be null, where $\mathcal{U}'$ is the non-separable set (B.2) with channel sizes $n_1 - j_1, \ldots, n_C - j_C$ and factor sizes $r'_0 = (r_0 - \sum_{c=1}^{C}(j_c - r_c)_+)_+$, $r'_c = (r_c - j_c)_+$ for $c = 1, \ldots, C$.

**Verification of Condition 2**

To demonstrate that Condition 2 is sufficient to imply that $\mathcal{U}$ is null, first assume that $M$ as defined in (2.14) is non-empty. For $\mathcal{U}^*_\emptyset$, if the subset of $M$ containing $(\mathbf{n}', \mathbf{r}', \boldsymbol{\rho})$ with $\mathbf{n}' = \mathbf{n}, \mathbf{r}' = \mathbf{r}$ is empty, then (B.6) cannot be satisfied with $\boldsymbol{\rho}$ non-negative and so either $2(r_0 + r) - n < 0$ or there is a $c$ with $2(r_0 + r_c) - n_c < 0$. That is, $\mathbf{H}(\mathbf{A}, \mathbf{B}, \boldsymbol{\varphi})$ cannot satisfy (B.5) for any $(\mathbf{A}, \mathbf{B})$ and so $\mathcal{U}^*_\emptyset$ is empty. If the aforementioned subset of $M$ is non-empty, then at least some non-negative $\boldsymbol{\rho}$ satisfying (B.6) is possible. Then the dimension condition (B.8) is equivalent to $\psi(\mathbf{n}, \mathbf{r}, \boldsymbol{\rho}) > 0$ with $\psi$ defined in (2.12). As a function of $\rho_0$ alone, $\psi$ is decreasing on $[0, r_0]$, so $\psi$ being positive when $\rho_0$ is at its maximum feasible value implies $\psi$ is positive for all smaller values of $\rho_0$. So, $\min_\rho \psi(\mathbf{n}, \mathbf{r}, \rho) > 0$ ensures that (B.8) is satisfied for all valid $\alpha$. Therefore, $\mathcal{U}^*_\emptyset$ is null.

For any index set $\beta$ with $j_c, c = 1, \ldots, C$ zeros in the $c$th channel on the diagonal of $\boldsymbol{\varphi}$, taking $n'_c = n_c - j_c$ and $r'_0, r'_1, \ldots r'_C$ as in (2.14) effects the reduction of $\mathcal{U}^*_\beta$ to $\mathcal{U}'^*_\emptyset$ as discussed in the previous subsection. The above argument can be applied to $\mathcal{U}'^*_\emptyset$, showing that $\psi^* > 0$ implies $\mathcal{U}'^*_\emptyset$ is null for all possible reductions. Therefore, $\mathcal{U}^*_\beta$ is null for all $\beta$. The remaining part of $\mathcal{U}$ is a subset of $(\mathbb{A} \oplus \mathbb{B}) \setminus (\mathbb{A} \oplus \mathbb{B})^{**}$, which is null by Lemma 4. So, $\mathcal{U}$ is a subset of the finite union of the null sets $\mathcal{U}^*_\emptyset, \mathcal{U}^*_\beta$ for all $\beta$, and $(\mathbb{A} \oplus \mathbb{B}) \setminus (\mathbb{A} \oplus \mathbb{B})^{**}$.

Finally, if $M$ is empty, then no feasible value for $\boldsymbol{\rho}$ exists for any possible reduction and so $\mathcal{U}^*$ is empty. Then $\mathcal{U} \subset \mathbb{A} \oplus \mathbb{B} \setminus (\mathbb{A} \oplus \mathbb{B})^{**}$ and so is null. As $\mathbb{A} \oplus \mathbb{B}$ is isomorphic to $\mathbb{A} \times \mathbb{B}$, the result of Theorem 1 follows. □

## B.8 Proof of Theorem 2

*Proof.* First, note that $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is equivalent to $d\mathbf{R}_{\mathbf{xx}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$ when $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi}) \in \mathbb{A}_L^* \times \mathbb{B}_L^* \times \mathbb{P}_n$ is obtained from $\boldsymbol{\eta}$ by inverting (2.7). For $(d\mathbf{A}, d\mathbf{B}) \in \mathbb{A}_L \times \mathbb{B}_L$ and $d\boldsymbol{\Phi} \in \mathrm{Diag}(n)$, the differential

$$d\mathbf{R}_{\mathbf{xx}} = \mathbf{A}d\mathbf{A}^\mathsf{T} + d\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}d\mathbf{B}^\mathsf{T} + d\mathbf{B}\mathbf{B}^\mathsf{T} + d\boldsymbol{\Phi} \tag{B.13}$$

is a linear map in $(d\mathbf{A}, d\mathbf{B}, d\boldsymbol{\Phi})$ from a vector space of dimension $L$ as defined in (2.8) to a space of dimension $n(n+1)/2$. Necessary conditions for injectivity can be obtained by dimensionality considerations. First, the dimension of the domain must be no greater than the codomain for the map to be injective, so $L \leq n(n+1)/2$. The previous inequality is equivalent to the condition

$$r_0 \leq \frac{1}{2}\left(2n + 1 - \sqrt{8(n+D)+1}\right),$$

where $D$ is defined in (2.17). Similarly, setting $d\mathbf{A}, d\boldsymbol{\Phi}$ and $d\mathbf{B}_2, \ldots, d\mathbf{B}_C$ to zero (of the appropriate dimensions), the restricted $d\mathbf{R}_{\mathbf{xx}}$ is a linear map from a vector space of dimension $n_1 r_1 - \frac{r_1(r_1-1)}{2}$ into the subspace of symmetric matrices with only the top $n_1 \times n_1$ block being non-zero. Again, the dimension of the restricted domain must be no greater than the codomain, meaning $n_1 r_1 - r_1(r_1-1)/2 \leq n_1(n_1+1)/2$. This is equivalent to $r_1 \leq \frac{1}{2}\left(2n_1 + 1 - \sqrt{8n_1+1}\right)$. The same considerations for other blocks yield the conditions that, for all $c = 1, \ldots, C$, $r_c \leq \frac{1}{2}\left(2n_c + 1 - \sqrt{8n_c+1}\right)$. Combined, this is Condition 4 and so Proposition 7 is proven.

Next, we will show that the Condition 4 combined with the separability Condition 1 is sufficient for $d\mathbf{R}_{\mathbf{xx}}$ to be generically injective in $(d\mathbf{A}, d\mathbf{B}, d\boldsymbol{\Phi})$. Assume that the combined matrix $[\mathbf{A}\ \mathbf{B}]$ is FCR, which excludes a null subset of $\mathbb{A}_L^* \times \mathbb{B}_L^*$. The proof proceeds in three steps: first showing that $d\mathbf{A} \mapsto \mathbf{A}d\mathbf{A}^\mathsf{T} + d\mathbf{A}\mathbf{A}^\mathsf{T}$ and $d\mathbf{B} \mapsto \mathbf{B}d\mathbf{B}^\mathsf{T} + d\mathbf{B}\mathbf{B}^\mathsf{T}$ are separately injective, then showing the sum of the two is generically injective, then finally showing that the map $(d\mathbf{A}, d\mathbf{B}, d\boldsymbol{\Phi}) \mapsto d\mathbf{R}_{\mathbf{xx}}$ is generically injective and therefore $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is generically injective.

Define the linear maps $\mathbf{F_A} : \mathbb{A}_L \to \mathrm{Sym}(n)$ and $\mathbf{F_B} : \mathbb{B}_L \to \mathrm{Sym}(n)$ as $\mathbf{F_A}(\mathbf{X}) \equiv \mathbf{A}\mathbf{X}^\mathsf{T} + \mathbf{X}\mathbf{A}^\mathsf{T}$ and $\mathbf{F_B}(\mathbf{Y}) \equiv \mathbf{B}\mathbf{Y}^\mathsf{T} + \mathbf{Y}\mathbf{B}^\mathsf{T}$. Similarly, $\mathbf{F_{A,B}} : \mathbb{A}_L \times \mathbb{B}_L \to \mathrm{Sym}(n)$ is

$$\begin{aligned}\mathbf{F_{A,B}}(\mathbf{X}, \mathbf{Y}) &\equiv \mathbf{F_A}(\mathbf{X}) + \mathbf{F_B}(\mathbf{Y}) \\ &= [\mathbf{A}\ \mathbf{B}][\mathbf{X}\ \mathbf{Y}]^\mathsf{T} + [\mathbf{X}\ \mathbf{Y}][\mathbf{A}\ \mathbf{B}]^\mathsf{T}.\end{aligned} \tag{B.14}$$

For $\mathbf{F_A}$, an application of Lemma 3 implies that $\mathbf{F_A}$ is injective, as $\mathbf{F_A}(\mathbf{X}) = \mathbf{0}$ implies $\mathbf{X} = \mathbf{0}$ by the structure of $\mathbb{A}_L$ and the FCR assumption. Arguing channel-wise for $\mathbf{F_B}$, $\mathbf{F_B}$ is injective in a similar fashion.

Second, define the image subspaces as $\mathcal{A} \equiv \mathrm{Im}(\mathbf{F_A})$ and $\mathcal{B} \equiv \mathrm{Im}(\mathbf{F_B})$. Injectivity of $\mathbf{F_A}$ and $\mathbf{F_B}$ implies

$$\begin{aligned}\dim(\mathcal{A}) &= nr_0 - \tfrac{r_0(r_0-1)}{2} \\ \dim(\mathcal{B}) &= \textstyle\sum_{c=1}^{C} n_c r_c - \tfrac{r_c(r_c-1)}{2}.\end{aligned}$$

The sum map $\mathbf{F_{A,B}}$ will be injective iff $\mathcal{A} \cap \mathcal{B} = \{\mathbf{0}\}$, as existence of $(\mathbf{X}, \mathbf{Y})$ with $\mathbf{F_{A,B}}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$ implies that $\mathbf{F_A}(\mathbf{X}) = -\mathbf{F_B}(\mathbf{Y})$ and so $\mathbf{F_B}(\mathbf{Y})$ is in $\mathcal{A} \cap \mathcal{B}$. If $\mathcal{A} \cap \mathcal{B} = \{\mathbf{0}\}$, then injectivity of $\mathbf{F_A}$ and $\mathbf{F_B}$ separately implies $(\mathbf{X}, \mathbf{Y}) = (\mathbf{0}, \mathbf{0})$. The converse direction follows as if $\mathbf{G} \in \mathcal{A} \cap \mathcal{B}$ and $\mathbf{G} \neq \mathbf{0}$, then exist non-zero $\mathbf{X}$ and $\mathbf{Y}$ such that $\mathbf{F_A}(\mathbf{X}) = \mathbf{G} = \mathbf{F_B}(\mathbf{Y})$ and so $\mathbf{F_{A,B}}(\mathbf{X}, -\mathbf{Y}) = \mathbf{0}$.

To show $\mathcal{A} \cap \mathcal{B} = \{\mathbf{0}\}$, suppose $\mathbf{F_{A,B}}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$ and so

$$\mathbf{A}\mathbf{X}^\mathsf{T} + \mathbf{X}\mathbf{A}^\mathsf{T} = -\mathbf{B}\mathbf{Y}^\mathsf{T} - \mathbf{Y}\mathbf{B}^\mathsf{T}. \tag{B.15}$$

Next, suppose there exists some $\mathbf{v} \in \mathbb{R}_n$ with $\mathbf{v} \in \mathrm{Ker}(\mathbf{A}^\mathsf{T}) \cap \mathrm{Ker}(\mathbf{B}^\mathsf{T})$ but at least one of $\mathbf{X}^\mathsf{T}\mathbf{v}$ or $\mathbf{Y}^\mathsf{T}\mathbf{v}$ is non-zero. Then applying the transformations in (B.15) to $\mathbf{v}$, we obtain

$$\mathbf{A}\mathbf{X}^\mathsf{T}\mathbf{v} = \mathbf{B}(-\mathbf{Y})^\mathsf{T}\mathbf{v}.$$

The LHS is then in $\mathrm{Im}(\mathbf{A})$ and the RHS is in $\mathrm{Im}(\mathbf{B})$. As $\mathrm{Im}(\mathbf{A}) \cap \mathrm{Im}(\mathbf{B}) = \{\mathbf{0}\}$ is implied by the FCR assumption on $[\mathbf{A}\ \mathbf{B}]$, we must have that both $\mathbf{A}\mathbf{X}^\mathsf{T}\mathbf{v}$ and $\mathbf{B}(-\mathbf{Y})^\mathsf{T}\mathbf{v}$ are zero. However, $\mathbf{A}$ and $\mathbf{B}$ both being full rank implies that $\mathrm{Ker}(\mathbf{A}) = \{\mathbf{0}\}$ and $\mathrm{Ker}(\mathbf{B}) = \{\mathbf{0}\}$. As at least one of

$\mathbf{X}^\mathsf{T}\mathbf{v}$ and $(-\mathbf{Y})^\mathsf{T}\mathbf{v}$ is non-zero by assumption, at least one of $\mathbf{A}\mathbf{X}^\mathsf{T}\mathbf{v}$ and $\mathbf{B}(-\mathbf{Y})^\mathsf{T}\mathbf{v}$ is non-zero, yielding a contradiction. So, for all $\mathbf{v} \in \mathrm{Ker}(\mathbf{A}^\mathsf{T}) \cap \mathrm{Ker}(\mathbf{B}^\mathsf{T})$, $\mathbf{X}^\mathsf{T}\mathbf{v}$ and $\mathbf{Y}^\mathsf{T}\mathbf{v}$ are zero.

Hence, we have that $\mathrm{Ker}(\mathbf{X}^\mathsf{T}) \supset \mathrm{Ker}(\mathbf{A}^\mathsf{T}) \cap \mathrm{Ker}(\mathbf{B}^\mathsf{T})$ while $\mathrm{Ker}(\mathbf{Y}^\mathsf{T}) \supset \mathrm{Ker}(\mathbf{A}^\mathsf{T}) \cap \mathrm{Ker}(\mathbf{B}^\mathsf{T})$, which is equivalent to $\mathrm{Im}(\mathbf{X}) \subset \mathrm{Im}(\mathbf{A}) \oplus \mathrm{Im}(\mathbf{B})$ and $\mathrm{Im}(\mathbf{Y}) \subset \mathrm{Im}(\mathbf{A}) \oplus \mathrm{Im}(\mathbf{B})$. So, all columns of $[\mathbf{X}\ \mathbf{Y}]$ can be written as a linear combination of the columns of $[\mathbf{A}\ \mathbf{B}]$ and there exists some $\mathbf{W} \in \mathbb{R}_{(r_0+r)\times(r_0+r)}$ such that

$$[\mathbf{X}\ \mathbf{Y}] = [\mathbf{A}\ \mathbf{B}]\mathbf{W}. \tag{B.16}$$

Combining (B.16) and (B.14), $\mathbf{F}_{\mathbf{A},\mathbf{B}}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$ can be written as $[\mathbf{A}\ \mathbf{B}](\mathbf{W}^\mathsf{T} + \mathbf{W})[\mathbf{A}\ \mathbf{B}]^\mathsf{T} = \mathbf{0}$. As $[\mathbf{A}\ \mathbf{B}]$ is FCR, left and right multiplying by $([\mathbf{A}\ \mathbf{B}]^\mathsf{T}[\mathbf{A}\ \mathbf{B}])^{-1}[\mathbf{A}\ \mathbf{B}]^\mathsf{T}$ and its transpose yields that $(\mathbf{W} + \mathbf{W}^\mathsf{T}) = \mathbf{0}$, so $\mathbf{W}$ is skew-symmetric.

However, Proposition 4 implies that $\mathbf{W}$ must be zero unless $(\mathbf{A}, \mathbf{B})$ belong to a null subset of $\mathbb{A}_L^* \times \mathbb{B}_L^*$. To see this, recall that the Cayley transform gives that there exists a $\mathbf{Q} \in \mathrm{O}(r_0 + r)$ such that $\mathbf{W} = (\mathbf{Q} - \mathbf{I})(\mathbf{Q} + \mathbf{I})^{-1}$. So, (B.16) is equivalent to

$$[\mathbf{X}\ \mathbf{Y}] + [\mathbf{A}\ \mathbf{B}] = ([\mathbf{A}\ \mathbf{B}] - [\mathbf{X}\ \mathbf{Y}])\,\mathbf{Q}.$$

The above implies that $(\mathbf{X}+\mathbf{A}, \mathbf{Y}+\mathbf{B}) \sim_1 (\mathbf{A}-\mathbf{X}, \mathbf{B}-\mathbf{Y})$, where both pairs are in $\mathbb{A}_L \times \mathbb{B}_L$. Note that $\mathbf{F}_{\mathbf{A},\mathbf{B}}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$ implies that $\mathbf{F}_{\mathbf{A},\mathbf{B}}(\epsilon\mathbf{X}, \epsilon\mathbf{Y}) = \mathbf{0}$ for any $\epsilon$, meaning that $(\mathbf{A} - \mathbf{X}, \mathbf{B} - \mathbf{Y})$ can be assumed to belong to an arbitrary neighborhood of $(\mathbf{A}, \mathbf{B})$ in $\mathbb{A}_L^* \times \mathbb{B}_L^*$. By the proof of Proposition 4, the subset $\tilde{\mathcal{C}} \subset \mathbb{A}_L^* \times \mathbb{B}_L^*$ where the full rank submatrix condition of Proposition 3 is not satisfied is closed and null. Generically, $(\mathbf{A}, \mathbf{B})$ is in the complement $\mathbb{A}_L^* \times \mathbb{B}_L^* \setminus \tilde{\mathcal{C}}$ and so $(\mathbf{A} - \mathbf{X}, \mathbf{B} - \mathbf{Y})$ also belongs to the complement for small enough $(\mathbf{X}, \mathbf{Y})$ as $[\mathbf{A}\ \mathbf{B}]$ is FCR. So, Proposition 3 applies to $(\mathbf{A} - \mathbf{X}, \mathbf{B} - \mathbf{Y})$, meaning that $\mathbf{Q}$ must be block-diagonal and so $(\mathbf{X} + \mathbf{A}, \mathbf{Y} + \mathbf{B}) \sim_2 (\mathbf{A} - \mathbf{X}, \mathbf{B} - \mathbf{Y})$. By Proposition 5, the $\sim_2$-representatives in $\mathbb{A}_L^* \times \mathbb{B}_L^*$ are unique, hence $(\mathbf{X}, \mathbf{Y}) = (\mathbf{0}, \mathbf{0})$. Thus, $\mathbf{F}_{\mathbf{A},\mathbf{B}}$ is generically injective.

For the third step, we complete the proof by showing that

$$(\mathcal{A} \oplus \mathcal{B}) \cap \mathrm{Diag}(n) = \{\mathbf{0}\}, \tag{B.17}$$

which is equivalent to showing that the differential (B.13) is injective. The show (B.17), we examine the orthogonal complement $((\mathcal{A} \oplus \mathcal{B}) \cap \mathrm{Diag}(n))^{\perp}$. Standard properties of the subspace lattice (Kollo and Dietrich, 2005, Ch. 1) show that $((\mathcal{A} \oplus \mathcal{B}) \cap \mathrm{Diag}(n))^{\perp}$ equals $(\mathcal{A}^{\perp} \cap \mathcal{B}^{\perp}) \cap (\mathcal{A}^{\perp} \cap \mathcal{B}^{\perp} \cap \mathrm{Diag}(n)^{\perp})^{\perp} \oplus \mathrm{Diag}(n)^{\perp}$. As $\dim(\mathrm{Diag}(n)^{\perp})$ is $n(n-1)/2$, it suffices to show

$$n = \dim\big((\mathcal{A}^{\perp} \cap \mathcal{B}^{\perp}) \cap (\mathcal{A}^{\perp} \cap \mathcal{B}^{\perp} \cap \mathrm{Diag}(n)^{\perp})^{\perp}\big). \tag{B.18}$$

Expanding $\mathcal{B}^{\perp}$ using that $\mathbf{F}_{\mathbf{B}}(\mathbf{Y})$ is channel-structured block-diagonal, we see that any matrix with zeros in the block-diagonal is within $\mathcal{B}^{\perp}$. Let $\mathcal{Y}$ be the subspace of such matrices,

$$\mathcal{Y} = \mathrm{Span}\big(\{\mathbf{e}_i\mathbf{e}_j^{\mathsf{T}} + \mathbf{e}_j\mathbf{e}_i^{\mathsf{T}} : \exists\, c \text{ s.t. } i \leq r_{<c} + r_c < j\}\big).$$

The structure of $\mathcal{Y}$ then implies $\mathcal{Y} \subset \mathcal{B}^{\perp}$ and therefore $\mathcal{B}^{\perp} = \mathcal{Y} \oplus (\mathcal{B}^{\perp} \cap \mathcal{Y}^{\perp})$. The subspace $\mathcal{B}^{\perp} \cap \mathcal{Y}^{\perp}$ contains all channel-structured block-diagonal matrices where for all $c$, the $c$th block is orthogonal to all matrices of the form $\mathbf{B}_c\mathbf{Y}_c^{\mathsf{T}} + \mathbf{Y}_c\mathbf{B}_c^{\mathsf{T}}$. Hence, the subspace $\mathcal{B}^{\perp} \cap \mathcal{Y}^{\perp}$ be can be written as $\mathcal{W}_1 \oplus \cdots \oplus \mathcal{W}_C$, where $\mathcal{W}_c \subset \mathrm{Sym}(n)$ contains the matrices in $\mathcal{B}^{\perp} \cap \mathcal{Y}^{\perp}$ whose entries for blocks other than $c$ are all zero.

The intersection $\mathcal{A}^{\perp} \cap \mathcal{B}^{\perp} \cap \mathrm{Diag}(n)^{\perp}$ equals $\mathcal{A}^{\perp} \cap (\mathcal{W}_1 \oplus \cdots \oplus \mathcal{W}_c \oplus \mathcal{Y}) \cap \mathrm{Diag}(n)^{\perp}$, which in turn equals $\mathcal{A}^{\perp} \cap (((\mathcal{W}_1 \oplus \cdots \oplus \mathcal{W}_c) \cap \mathrm{Diag}(n)^{\perp}) \oplus \mathcal{Y})$ as $\mathcal{Y} \subset \mathrm{Diag}(n)^{\perp}$. The subspace $(\mathcal{W}_1 \oplus \cdots \oplus \mathcal{W}_c) \cap \mathrm{Diag}(n)^{\perp}$ consists of all $\mathbf{C} \in \mathcal{B}^{\perp} \cap \mathcal{Y}^{\perp}$ with zero diagonal. This space equals $\widetilde{\mathcal{W}}_1 \oplus \cdots \oplus \widetilde{\mathcal{W}}_C$, where $\widetilde{\mathcal{W}}_c$ is the subspace of all elements of $\mathcal{W}_c$ with zero diagonal. So, $\dim(\mathcal{A}^{\perp} \cap \mathcal{B}^{\perp} \cap \mathrm{Diag}(n)^{\perp})$ is

$$\dim(\mathcal{A}) + \textstyle\sum_{c=1}^{C} \dim(\widetilde{\mathcal{W}}_c) + \dim(\mathcal{Y}) - \dim(\mathcal{A} + \sum_{c=1}^{C} \widetilde{\mathcal{W}}_c + \mathcal{Y}).$$

The dimension of $\mathcal{W}_c$ can be written as

$$\begin{aligned}\dim(\widetilde{\mathcal{W}}_c) &= \dim(\mathcal{W}_c) + \tfrac{n_c(n_c-1)}{2} - (\tfrac{n_c(n_c+1)}{2} - J_c) \\ &= \dim(\mathcal{W}_c) - (n_c - J_c),\end{aligned}$$

where $J_c$ is the dimension of the intersection between the subspace of matrices in $\mathrm{Sym}(n_c)$ that are realized as $\mathbf{B}_c\mathbf{Y}_c^{\mathsf{T}} + \mathbf{Y}_c\mathbf{B}_c^{\mathsf{T}}$ and the subspace of diagonal matrices. This will be zero iff the

individual channel is locally identifiable as a single channel factor model, which will generically occur when (2.18) is satisfied, as shown by Shapiro (1985, Thm. 3.2). If the above is satisfied, $\mathcal{B} \cap \mathrm{Diag}(n) = \{\mathbf{0}\}$. From this, we obtain that

$$\begin{aligned} \mathcal{A}^\perp + \mathcal{B}^\perp = (\mathcal{A}^\perp + \mathcal{B}^\perp) \cap (\mathcal{A}^\perp + (\mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp)) \\ \oplus (\mathcal{A}^\perp + \mathcal{B}^\perp) \cap (\mathcal{A} \cap (\mathcal{B} \oplus \mathrm{Diag}(n))). \end{aligned} \tag{B.19}$$

However, the term $\mathcal{A} \cap (\mathcal{B} \oplus \mathrm{Diag}(n))$ can be shown to be $\{\mathbf{0}\}$ by the results from Shapiro (1982) for single-channel FA when (2.16) is satisfied. In particular, if $\mathbf{C} \in \mathcal{A} \cap (\mathcal{B} \oplus \mathrm{Diag}(n))$, then as $\mathbf{C} \in \mathcal{A}$, Shapiro (1982, Lemma 2.1) implies that

$$\mathbf{v}_i^\mathsf{T} \mathbf{C} \mathbf{v}_j = 0, \quad 1 \leq i \leq j \leq n - r_0,$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_{n-r_0} \in \mathbb{R}_n$ is a basis for $\mathrm{Ker}(\mathbf{A}^\mathsf{T})$. As the structure of $\mathbb{A}$ does not restrict $\mathrm{Ker}(\mathbf{A}^\mathsf{T})$, all the above linear constraints on $\mathbf{C}$ are generically independent. There are $\binom{n-r_0}{2}$ constraints, and $\mathbf{C} \in (\mathcal{B} \oplus \mathrm{Diag}(n))$ has $\dim(\mathcal{B} \oplus \mathrm{Diag}(n))$ degrees of freedom. Hence, $\mathcal{A} \cap (\mathcal{B} \oplus \mathrm{Diag}(n))$ is $\{\mathbf{0}\}$ iff

$$\tfrac{(n-r_0)(n-r_0+1)}{2} \geq \dim(\mathcal{B}) + \dim(\mathrm{Diag}(n)),$$

as the RHS is $\dim(\mathcal{B} \oplus \mathrm{Diag}(n))$. This is equivalent to the condition (2.16), and therefore $\mathcal{A} \cap (\mathcal{B} \oplus \mathrm{Diag}(n)) = \{\mathbf{0}\}$. This implies that the RHS of (B.19) equals $(\mathcal{A}^\perp + (\mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp))$. Finally, (B.18) can be expanded as,

$$\begin{aligned} &\dim\left((\mathcal{A}^\perp \cap \mathcal{B}^\perp) \cap (\mathcal{A}^\perp \cap \mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp)^\perp\right) \\ &= \dim(\mathcal{A}^\perp) + \dim(\mathcal{B}^\perp) - \dim(\mathcal{A}^\perp + \mathcal{B}^\perp) \\ &\quad - \dim(\mathcal{A}^\perp \cap \mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp) \\ &= \dim(\mathcal{A}^\perp) + \textstyle\sum_{c=1}^C \dim(\mathcal{W}_c) + \dim(\mathcal{Y}) \\ &\quad - \dim(\mathcal{A}^\perp + \mathcal{B}^\perp) - \dim(\mathcal{A}^\perp) - \textstyle\sum_{c=1}^C \dim(\widetilde{W}_c) \\ &\quad + \dim(\mathcal{A}^\perp + (\mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp)). \end{aligned}$$

By the previous results, this simplifies to

$$n - \big[\dim(\mathcal{A}^\perp + \mathcal{B}^\perp) - \dim(\mathcal{A}^\perp + (\mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp))\big] - \textstyle\sum_{c=1}^{C} J_c,$$

which equals $n$ iff all $J_c = 0$ and

$$\dim(\mathcal{A}^\perp + (\mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp)) = \dim(\mathcal{A}^\perp + \mathcal{B}^\perp).$$

This occurs generically when the criteria in Proposition 2 are satisfied, and so the differential $d\mathbf{R}_{\mathbf{xx}}$ will be injective at almost all $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}) \in \mathbb{A}_L^* \times \mathbb{B}_L^* \times \mathrm{Diag}(n)$. This space is isomorphic to $V$, so $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is generically injective. □

## B.9 Proof of Theorem 3

*Proof.* Define the function $\ell_0$ over $\boldsymbol{\eta} \in V$,

$$\begin{aligned}\ell_0(\boldsymbol{\eta}) \equiv E[\ell_T(\boldsymbol{\eta})] &= \log\det \mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}) + \mathrm{tr}\, \mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})\mathbf{\Sigma}_{\mathbf{xx}} \\ &= 2D_{KL}\big(\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{xx}}) || \mathcal{N}(\mathbf{0}, \mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}))\big) + \log\det \mathbf{\Sigma}_{\mathbf{xx}}.\end{aligned}$$

By hypothesis, $\ell_0$ has a unique minimum over $V'$ at $\mathring{\boldsymbol{\eta}}$ as $\mathring{\boldsymbol{\eta}}$ belongs to the globally identified set. Further, this minimum is well-separated. To see this, first note that the sublevel sets of $\ell_0(\boldsymbol{\eta})$ are compact by the proof of Ramírez et al. (2020, Thm. 1) with $\mathbf{S} = \mathbf{\Sigma}_{\mathbf{xx}}$ and the fact that the map $\boldsymbol{\eta} \mapsto (\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ is continuous. Let $m = \ell_0(\mathring{\boldsymbol{\eta}})$, which is finite as $\mathbf{\Sigma}_{\mathbf{xx}}$ and $\mathring{\mathbf{R}}_{\mathbf{xx}}$ are positive definite and so yield a finite KL-divergence. For any $h > 0$ and all $\epsilon > 0$, the closed set $V'$ is partitioned into the three sets $B_\epsilon \cap L_h$, $B_\epsilon^c \cap L_h$ and $L_h^c$, where

$$\begin{aligned}B_\epsilon &= \{\boldsymbol{\eta} \in V' \,;\, ||\boldsymbol{\eta} - \mathring{\boldsymbol{\eta}}|| < \epsilon\}, \\ L_h &= \{\boldsymbol{\eta} \in V' \,;\, \ell_0(\boldsymbol{\eta}) \leq m + h\}.\end{aligned}$$

The set $L_h \cap B_\epsilon^c$ is the intersection of a closed set and a compact set, so it is compact. Therefore, the infimum of the continuous function $\ell_0(\boldsymbol{\eta})$ over $L_h \cap B_\epsilon^c$ is achieved at some $\boldsymbol{\eta}'$. By the assumption of a unique minimum, $\ell_0(\boldsymbol{\eta}') > m$ for all $\epsilon > 0$. Additionally, as $h > 0$, the infimum of $\ell_0$ over $L_h^c$ is strictly greater than $m$. Therefore, $\inf_{\boldsymbol{\eta} \in B_\epsilon^c} \ell(\boldsymbol{\eta}) > \ell(\mathring{\boldsymbol{\eta}})$ and so $\mathring{\boldsymbol{\eta}}$ is a well-separated minimum.

The deviation of $\ell_T$ from $\ell_0$ is controlled as,

$$\begin{aligned}\sup_{\boldsymbol{\eta}\in V'}|\ell_T(\boldsymbol{\eta})-\ell_0(\boldsymbol{\eta})| &= \sup_{\boldsymbol{\eta}\in V'}|\operatorname{tr}\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})(\mathbf{S}_T-\boldsymbol{\Sigma}_{\mathbf{xx}})| \\ &\leq \sup_{\boldsymbol{\eta}\in V'}||\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})||_F||\mathbf{S}_T-\boldsymbol{\Sigma}_{\mathbf{xx}}||_F \\ &\leq \epsilon^{-1}\sqrt{n}||\mathbf{S}_T-\boldsymbol{\Sigma}_{\mathbf{xx}}||_F.\end{aligned}$$

The third line follows from the definition of $V'$, which imposes that $\lambda_{\min}(\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}))\geq\epsilon$. As the second moment of $\mathbf{x}$ exists, $\mathbf{S}_T$ is consistent so $||\mathbf{S}_T-\boldsymbol{\Sigma}_{\mathbf{xx}}||_F\xrightarrow{p}0$. So, $\ell_T(\boldsymbol{\eta})$ converges uniformly in probability to the limiting function $\ell_0(\boldsymbol{\eta})$. As $\widehat{\boldsymbol{\eta}}_T$ minimizes $\ell_T$, standard results for $M$-estimators (Vaart, 1998, p. 45) imply that $\widehat{\boldsymbol{\eta}}_T\xrightarrow{p}\mathring{\boldsymbol{\eta}}$, so $(\widehat{\mathbf{A}}_T,\widehat{\mathbf{B}}_T,\widehat{\boldsymbol{\Phi}}_T)\xrightarrow{p}(\mathring{\mathbf{A}},\mathring{\mathbf{B}},\mathring{\boldsymbol{\Phi}})$ as well. Note that if $\boldsymbol{\Sigma}_{\mathbf{xx}}\in\mathcal{R}(\mathbf{n},\mathbf{r})$, then the minimizing $\mathring{\mathbf{R}}_{\mathbf{xx}}$ is $\boldsymbol{\Sigma}_{\mathbf{xx}}$, which is the unique minimum by the Gibbs inequality. □

## B.10 Proof of Theorem 4

*Proof.* For any $r>0$, define the set of $\boldsymbol{\eta}\in V'$ with $||\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})-\boldsymbol{\Sigma}_{\mathbf{xx}}||_F<r$ and $||\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})-\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}||_F<r$. As $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}),\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})$ are continuous and $\mathring{\boldsymbol{\eta}}$ is in both sets, their intersection is a non-empty open neighborhood of $\mathring{\boldsymbol{\eta}}$. In this neighborhood, the objective function $\ell_T(\boldsymbol{\eta})$ is Lipschitz, as for any $\boldsymbol{\eta}_1,\boldsymbol{\eta}_2$ in the neighborhood with $\mathbf{R}_1\equiv\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}_1),\mathbf{R}_2\equiv\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}_2)$, the difference $|\ell_T(\boldsymbol{\eta}_1)-\ell_T(\boldsymbol{\eta}_2)|$ is bounded above,

$$\begin{aligned}|\ell_T(\boldsymbol{\eta}_1)-\ell_T(\boldsymbol{\eta}_2)| &\leq |\log\det\mathbf{R}_1\mathbf{R}_2^{-1}|+|\operatorname{tr}\mathbf{S}_T[\mathbf{R}_1^{-1}-\mathbf{R}_2^{-1}]| \\ &\leq |\operatorname{tr}\mathbf{R}_1[\mathbf{R}_2^{-1}-\mathbf{R}_1^{-1}]|+|\operatorname{tr}\mathbf{S}_T[\mathbf{R}_2^{-1}-\mathbf{R}_1^{-1}]| \\ &\leq (||\mathbf{R}_1-\boldsymbol{\Sigma}_{\mathbf{xx}}||_F+||\boldsymbol{\Sigma}_{\mathbf{xx}}||_F+||\mathbf{S}_T||_F)\times \\ &\qquad ||\mathbf{R}_1^{-1}||_F||\mathbf{R}_2^{-1}||_F||\mathbf{R}_1-\mathbf{R}_2||_F \\ &\leq m(r,\boldsymbol{\Sigma}_{\mathbf{xx}},\mathbf{S})||\mathbf{R}_1-\mathbf{R}_2||_F,\end{aligned}$$

with $m(r,\boldsymbol{\Sigma}_{\mathbf{xx}},\mathbf{S})=(r+||\boldsymbol{\Sigma}_{\mathbf{xx}}||_F+||\mathbf{S}||_F)(r+||\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}||_F)^2$. Since by assumption $E[||\mathbf{x}_1||^4]<\infty$, it is implied that $E[||\mathbf{S}_T||_F^2]<\infty$ and therefore $E[m(r,\boldsymbol{\Sigma}_{\mathbf{xx}},\mathbf{S})^2]<\infty$. As $||\mathbf{R}||_F$ is bounded within

the neighborhood, it is similarly true that the associated $||\mathbf{A}||_F, ||\mathbf{B}||_F$ are bounded. Therefore,

$$\begin{aligned}||\mathbf{R}_1 - \mathbf{R}_2||_F &\leq 2||\mathbf{A}_1||_F||\mathbf{A}_1 - \mathbf{A}_2||_F + ||\mathbf{\Phi}_1 - \mathbf{\Phi}_2||_F \\ &\quad + 2||\mathbf{B}_1||_F||\mathbf{B}_1 - \mathbf{B}_2||_F \\ &\leq C(r, \mathbf{\Sigma_{xx}})||\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2||_2\end{aligned}$$

for some non-random $C(r, \mathbf{\Sigma_{xx}})$. Hence, $\ell_T(\boldsymbol{\eta})$ is Lipschitz within some neighborhood of $\mathring{\boldsymbol{\eta}}$ with first differential

$$\begin{aligned}d\ell_T(\boldsymbol{\eta}, d\boldsymbol{\eta}) = &\operatorname{tr}(\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}, d\boldsymbol{\eta})) \\ &- \operatorname{tr}(\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}, d\boldsymbol{\eta})\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})\mathbf{S}_T).\end{aligned}$$

Both $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}, d\boldsymbol{\eta})$ and $\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})$ exist for all $\boldsymbol{\eta} \in V$, so $d\ell_T(\boldsymbol{\eta}, d\boldsymbol{\eta})$ is well-defined. The second differential expands as

$$\begin{aligned}d^2\ell_T(\boldsymbol{\eta}, d\boldsymbol{\eta}) = &\, 2\operatorname{tr}\left([\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}, d\boldsymbol{\eta})]^2\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})\mathbf{S}_T\right) \\ &- \operatorname{tr}\left([\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}, d\boldsymbol{\eta})]^2\right).\end{aligned}$$

Taking expectation and evaluating at $\mathring{\boldsymbol{\eta}}$, the above display equals $||\mathbf{\Sigma}_{\mathbf{xx}}^{-1/2}d\mathbf{R}_{\mathbf{xx}}(\mathring{\boldsymbol{\eta}}, d\boldsymbol{\eta})\mathbf{\Sigma}_{\mathbf{xx}}^{-1/2}||_F^2$ for $d\mathbf{R}_{\mathbf{xx}}$ as in (B.13), evaluated at $(\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\mathbf{\Phi}})$ with tangent vector $(d\mathbf{A}, d\mathbf{B}, d\mathbf{\Phi})$. This norm is zero only if $d\mathbf{R}_{\mathbf{xx}}(\mathring{\boldsymbol{\eta}}, d\boldsymbol{\eta})$ is zero. However, as in the proof of Theorem 3, identifiability of $\mathring{\boldsymbol{\eta}}$ implies that $d\mathbf{R}_{\mathbf{xx}}$ is non-zero for all non-zero $(d\mathbf{A}, d\mathbf{B}, d\mathbf{\Phi})$. Let $\mathbf{V}_0$ be the Hessian matrix of $\ell_0$ at $\mathring{\boldsymbol{\eta}}$. As the second differential is positive, the defining property of the second differential (Magnus and Neudecker, 1988, p. 119) shows that the quadratic form $d\boldsymbol{\eta}^\mathsf{T}\mathbf{V}_0 d\boldsymbol{\eta}$ is positive for all $d\boldsymbol{\eta} \neq \mathbf{0}$, so $\mathbf{V}_0$ is positive definite. Standard results for M-estimators (e.g. Vaart (1998, p. 53)) then imply that $\sqrt{T}(\widehat{\boldsymbol{\eta}}_T - \mathring{\boldsymbol{\eta}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{W})$, where

$$\mathbf{W} = \mathbf{V}_0^{-1}E\Big[\frac{\partial\ell_T(\mathring{\boldsymbol{\eta}})}{\partial\boldsymbol{\eta}}\frac{\partial\ell_T(\mathring{\boldsymbol{\eta}})^\mathsf{T}}{\partial\boldsymbol{\eta}}\Big]\mathbf{V}_0^{-1}. \tag{B.20}$$

□

## B.11 Example of Non-BD Q Preserving Channel Structure

The following example with $C = 2, n_1 = 3 = n_2$ and $r_0 = 3, r_1 = 1, r_2 = 1$ shows that non-block-diagonal $\mathbf{Q}$ can preserve the block structure of $\mathbf{B}$,

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ -1 & 2 & 0 \\ 2 & 1 & 1 \\ -3 & 4 & 1 \\ -6 & 5 & 1 \\ 5 & -4 & 3 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 2 & 0 \\ -1 & 0 \\ 3 & 0 \\ 0 & 1 \\ 0 & -3 \\ 0 & 2 \end{bmatrix},$$

$$\mathbf{Q} = \begin{bmatrix} 0.740 & 0.352 & -0.221 & 0.483 & -0.214 \\ -0.579 & 0.683 & 0.0553 & 0.439 & 0.053 \\ 0.069 & 0.073 & 0.801 & -0.052 & -0.588 \\ -0.022 & -0.528 & 0.332 & 0.712 & 0.321 \\ 0.334 & 0.354 & 0.443 & -0.253 & 0.709 \end{bmatrix},$$

as it can be verified that $[\mathbf{A}\ \mathbf{B}]\mathbf{Q}$ produces a $\tilde{\mathbf{B}}$ which is appropriately structured yet $\mathbf{Q}$ is not block-diagonal. The corresponding $\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T}$ provides an example of a non-trivial noise-free MFA covariance matrix which cannot be uniquely separated into signal and interference parts.

## B.12 Identifiability Result Comparison

The generative model (Lyu et al., 2022, Equation 5) of Deep CCA can be related to the generative model of MFA in the two-channel, noise-free ($\mathbf{\Phi} = \mathbf{0}$) case by setting $\mathbf{z} = \mathbf{f}$, $\mathbf{c}^{(q)} = \mathbf{g}_q$, and

$$g^{(q)}\left(\begin{bmatrix} \mathbf{z} \\ \mathbf{c}^{(q)} \end{bmatrix}\right) = \begin{bmatrix} \mathbf{A}_q & \mathbf{B}_q \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{g}_q \end{bmatrix},$$

for $q = 1, 2$. However, Lyu et al. (2022) further assumes that $\mathbf{g}^{(q)}$ is an invertible transformation for each $q$, and that assumption is a crucial component of the proof of the authors' main identifiability

result (Lyu et al., 2022, Theorem 1). In the noise-free case, our primary identifiability condition is Condition 1 which enables the signal and interference covariances to generically be uniquely separable. Condition 1 allows for $r_0 + r_c > n_c$ for some (but not all) of the channels, which makes $[\mathbf{A}_c\ \mathbf{B}_c]$ a wide matrix. For example, $n_1 = 10, n_2 = 5$ and $r_0 = 4, r_1 = 3, r_2 = 2$ are channel sizes and factor numbers which satisfy Condition 1 and $r_0 + r_2 > n_2$. In this case $[\mathbf{A}_c\ \mathbf{B}_c]$ has no left inverse and so the identifiability result of Lyu et al. (2022) is not applicable. In the case of multiple smaller channels or substantially different channel sizes, requiring that $r_0 + r_c \leq n_c$ for all $c = 1, \ldots, C$ will be significantly more restrictive than Condition 1.

A similar point also holds for the relation of the identifiability results of Sørensen et al. (2021) for Generalized CCA. To translate the noise-free case of MFA into the notation of Sørensen et al. (2021), set the number of views as $N = C$, the channel sizes as $J_n = r_n$ for $n = 1, \ldots, N$, the number of observations as $I = T$, the common factor dimension as $R = r_0$, and the distinct factor dimensions as $L_n = r_n$ for $n = 1, \ldots, N$. If $\mathbf{X}_n$ is the $I \times J_n$ matrix whose $m$th row is the transpose of the $m$th sample of our channel-$n$ observation vector $\mathbf{x}_n$, $\mathbf{F}$ is the $I \times R$ matrix whose $m$th row is the transpose of the $m$th value for our common factors $\mathbf{f}$, and $\mathbf{G}_n$ is the $I \times L_n$ matrix whose $m$th row is the transpose of the $m$th value for our channel-$n$ distinct factors $\mathbf{g}_n$, then the observation equation (9) in Sørensen et al. (2021) becomes

$$\begin{aligned}\mathbf{X}_n &= \mathbf{F}\mathbf{A}_n^\mathsf{T} + \mathbf{G}_n\mathbf{B}_n^\mathsf{T} \\ &= [\mathbf{F}\ \mathbf{G}_n]\mathbf{S}_n^\mathsf{T}\end{aligned}$$

where $\mathbf{A}_n$ and $\mathbf{B}_n$ are our common and distinct factor loading matrices for the $c$th channel and $\mathbf{S}_n = [\mathbf{A}_n\ \mathbf{B}_n]$. The identifiability results of Sørensen et al. (2021, Section IV) require that $\mathbf{S}_n$ be full column rank (and so $R + L_n \leq J_n$), while our Condition 1 allows for $R + L_n > J_n$ for some channels.

Similarly, the identifiability results of MFA in this chapter differ from those for JISA obtained by Lahat and Jutten (2019). To translate the noise-free ($\mathbf{\Phi} = \mathbf{0}$) case of MFA to that of JISA's generative model (Lahat and Jutten, 2019, Equation (2)), set the channel sizes as $R = C$, source number $R = 2$, common source $\mathbf{s}_1^{[k]} = \mathbf{s}_1^{[1]} = \mathbf{f}$ for all $k = 1, \ldots, K$ and channel-specific source

$\mathbf{s}_2^{[k]} = \mathbf{g}_k$. Then the channel-$k$ generative model in JISA is

$$\begin{aligned}\mathbf{x}^{[k]} &= \mathbf{A}_1^{[k]}\mathbf{s}_1^{[k]} + \mathbf{A}_2^{[k]}\mathbf{s}_2^{[k]} \\ &= [\mathbf{A}_1^{[k]}\ \mathbf{A}_2^{[k]}] \begin{bmatrix} \mathbf{s}_1^{[k]} \\ \mathbf{s}_2^{[k]}, \end{bmatrix}\end{aligned}$$

where $\mathbf{A}_1^{[k]}$ is the common factor loading matrix in channel $k$ (which is $\mathbf{A}_k$ in the current chapter), and $\mathbf{A}_2^{[k]}$ is the distinct factor loading matrix in channel $k$ (which is $\mathbf{B}_k$ in the current chapter). The mixing matrix $\mathbf{A}^{[k]}$ equals $[\mathbf{A}_1^{[k]}\ \mathbf{A}_2^{[k]}]$, which is $[\mathbf{A}_k\ \mathbf{B}_k]$ in our notation. With these definitions, the MFA in the noise-free scenario is indeed a special case of the general JISA model outlined in Lahat and Jutten (2019, Section II.A). However, the identifiability results obtained in Lahat and Jutten (2019, Section V) are obtained under a number of additional assumptions. In particular, Assumption (A4) in Lahat and Jutten (2019) is that the mixing matrix $\mathbf{A}^{[k]}$ is non-singular for all channels $k = 1, \ldots, K$, which differs from the the primary condition for unique separability of the signal and interference. The second identifiability subproblem treated in the current chapter, namely unique isolation of the noise variances, is discussed in Lahat and Jutten (2019, Section VIII.D). In that section, those authors note that the inclusion of noise will increase the complexity of the model identifiability problem, and they mention that the additional analysis needed for the noisy case is outside the scope of that paper. Therefore, although the noise-free case of MFA can be seen to be a special case of the general JISA model, our results on MFA identifiability are substantially different from the identifiability results of Lahat and Jutten (2019).

The method of Shared Independent Component Analysis (ShICA) represents a different approach to modeling multi-channel observations from that used in MFA. In the ShICA generative model (Richard et al., 2024, Equation 1), the channel-$i$ observation for $i = 1, \ldots, m$ and $m \geq 3$ is of size $p$ and it is generated by adding $p$-dimensional channel-specific noise $\mathbf{n}_i$ to a shared latent component $\mathbf{s}$, also of size $p$, and passing the sum through a non-singular $p \times p$ mixing matrix $\mathbf{A}_i$. In the ShICA model, the all-channel observations $\mathbf{x} = [\mathbf{x}_1^\mathsf{T}\ \ldots\ \mathbf{x}_m^\mathsf{T}]^\mathsf{T}$ live in a subspace of dimension $p$ within the all-channel observation space $\mathbb{R}_{mp}$. The effect of the channel specific noise $\mathbf{n}_i$ on the channel-$i$

observation is $\mathbf{A}_i\mathbf{n}_i$, while the effect of the shared component is $\mathbf{A}_i\mathbf{s}$ and so the effects of the shared component and noise components do not belong to distinct low-dimensional subspaces. The identifiability results of Richard et al. (2024, Theorem 1) utilize the assumption that the diagonal covariances of the signal $\mathbf{s}$ and noise components $\mathbf{n}_i$ have sufficient diversity to allow their effects on the multi-channel observations to be distinguished. This is a different approach than used for separating the signal and interference covariances in MFA, where our Proposition 1 distinguishes signal and interference using the properties of the combined loading matrix $[\mathbf{A}\ \mathbf{B}]$ and relies on the fact that the shared component $\mathbf{Af}$ and unshared component $\mathbf{Bg}$ in $\mathbf{x}$ live in different subspaces. Therefore, identifiability of MFA is a different problem from identifiability of ShICA, even in the noise-free case of MFA.

# Appendix C

# Supplementary Materials for Chapter 3

## C.1 Proof of Proposition 9

*Proof.* The proof is similar to a standard proof of the existence of the Wold decomposition (see, e.g. (Rozanov, 1967, Ch. 2)) with some modifications to account for the assumptions on the series at hand. We assume that the components of $\mathbf{s}[t], \mathbf{i}[t], \mathbf{u}[t], t \in \mathbb{Z}$ are defined on the same probability space, and let $\mathcal{L}^2$ be the Hilbert space of square-integrable complex random variables defined on that probability space. Next define the subspaces $\mathcal{H}_{\mathbf{s}}[t]$ as the closed span of $\left\{ [\mathbf{s}[k]]_i \, ; k \leq t, i = 1, \ldots, n \right\}$, and define $\mathcal{H}_{\mathbf{i}_c}[t],\ c = 1 \ldots, C$ and $\mathcal{H}_{\mathbf{u}}[t]$ similarly, and let $\mathcal{H}_{\mathbf{s}}[\infty], \mathcal{H}_{\mathbf{i}_c}[\infty], \mathcal{H}_{\mathbf{u}}[\infty]$ be the closed span of all values of their respective processes. The innovation subspaces $\mathcal{D}_{\mathbf{s}}[t]$ are $\mathcal{H}_{\mathbf{s}}[t] \cap (\mathcal{H}_{\mathbf{s}}[t-1])^{\perp}$, and by the hypothesis that $\mathbf{s}[t]$ lives in the $r_0$-dimensional subspace $\mathcal{S}$ for all $t$, we have that $n - r_0$ of the components $\left[\mathbf{s}[t]\right]_i$ can be written as a linear combination of the remaining $r_0$. Without loss of generality, suppose the components are numbered so that the $[\mathbf{s}[t]]_{i'}$ for $i' = r_0 + 1, \ldots n$ are linear combinations of the $[\mathbf{s}[t]]_{i''}$ for $i'' = 1, \ldots, r_0$,

$$[\mathbf{s}[t]]_{i'} = \textstyle\sum_{i=1}^{r_0} \alpha_{i',i''} \, [\mathbf{s}[t]]_{i''} \, ,$$

for some coefficients $\alpha_{i',i''} \in \mathbb{C}$. As $\mathcal{D}_{\mathbf{s}}[t]$ is generated by $\{(\mathbf{I} - \mathbf{P}_{\mathcal{H}_{\mathbf{s}}[t-1]}) \, [\mathbf{s}[t]]_i \, , i = 1, \ldots, n\}$, where $\mathbf{P}_{\mathcal{H}_{\mathbf{s}}[t-1]}$ is the orthogonal projection onto $\mathcal{H}_{\mathbf{s}}[t-1]$, if we let $\epsilon_{i''}[t] = (\mathbf{I} - \mathbf{P}_{\mathcal{H}_{\mathbf{s}}[t-1]}) \, [\mathbf{s}[t]]_{i''}$, we obtain

$$(\mathbf{I} - \mathbf{P}_{\mathcal{H}_{\mathbf{s}}[t-1]}) \, [\mathbf{s}[t]]_{i'} = \textstyle\sum_{i=1}^{r_0} \alpha_{i',i''} \epsilon_{i''}[t],$$

and so $\dim \mathbf{D}_{\mathbf{s}}[t] \leq r_0$ for all $t$. A similar definition and argument shows that $\dim \mathbf{D}_{\mathbf{i}_c}[t] \leq r_c$ for all $t$.

Next, let $\mathbf{U}_{\mathbf{s}}$ be the unitary operator on $\mathcal{H}_{\mathbf{s}}[\infty]$ defined by $\mathbf{U}_{\mathbf{s}} \, [\mathbf{s}[t]]_i = [\mathbf{s}[t+1]]_i$ for all $t \in \mathbb{Z}, i = 1, \ldots, n$, which is guaranteed to exist as $\mathbf{s}[t]$ is WSS (see, e.g., Rozanov (1967, Ch. 1)). As $\mathbf{s}[t]$ is completely nondeterministic, $\bigcap_{t \in \mathbb{Z}} \mathcal{H}_{\mathbf{s}}[t]$ is the trivial vector space and so $\mathcal{H}_{\mathbf{s}}[\infty]$ equals the

orthogonal direct sum $\bigoplus_{t\in\mathbb{Z}} \mathcal{D}_{\mathbf{s}}[t]$. Without loss of generality, we can assume that $\dim \mathcal{D}_{\mathbf{s}}[0] = r_0$, as if $\dim \mathcal{D}_{\mathbf{s}}[t] < r_0$ for all $t$, then $\mathbf{s}[t]$ in fact belongs to a proper subspace of $\mathcal{S}$ and the preceding argument can instead be run for that subspace. Then, let $[\boldsymbol{\varepsilon}_{\mathbf{s}}[0]]_k$ for $k = 1, \ldots, r_0$ be an orthonormal basis for $\mathcal{D}_{\mathbf{s}}[0]$ and set $[\boldsymbol{\varepsilon}_{\mathbf{s}}[t]]_k = \mathbf{U}_{\mathbf{s}}^t [\boldsymbol{\varepsilon}_{\mathbf{s}}[0]]_k$. As $\mathcal{H}_{\mathbf{s}}[t] = \bigoplus_{s\leq t} \mathcal{D}_{\mathbf{s}}[s]$, $\mathbf{s}[t]$ is obtained by application of a causal linear time-invariant filter of $\boldsymbol{\varepsilon}[t]$, hence

$$\mathbf{s}[t] = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\omega t} \mathcal{K}_{\mathbf{s}}(\mathrm{e}^{\mathrm{i}\omega}) d\mathbf{Z}_0(\omega)$$

where $\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\omega} d\mathbf{Z}(\omega)$ is the spectral representation of $\boldsymbol{\varepsilon}_{\mathbf{s}}[t]$, and $\mathcal{K}_{\mathbf{s}}(z) = \sum_{j=0}^{\infty} \mathbf{K}_{\mathbf{s},j} z^{-j}$ is the matrix transfer function of the linear filter. As $\boldsymbol{\varepsilon}_{\mathbf{s}}[t]$ is white noise with $E[\boldsymbol{\varepsilon}_{\mathbf{s}}[0]\boldsymbol{\varepsilon}_{\mathbf{s}}^{\mathsf{H}}[0]] = \mathbf{I}_{r_0}$, $\mathbf{Z}_0(\omega)$ is an orthogonal increments process with $E[d\mathbf{Z}(\omega) d\mathbf{Z}^{\mathsf{H}}(\omega)] = \mathbf{I}_{r_0}$ for all $\omega \in (-\pi, \pi]$. Further, as $\mathcal{H}_{\mathbf{s}}[t] = \bigoplus_{s\leq t} \mathcal{D}_{\mathbf{s}}[s]$, the innovations $\boldsymbol{\epsilon}_{\mathbf{s}}[t]$ are obtained by application of a causal linear time-invariant filter of $\mathbf{s}[t]$. By assumption, $\mathbf{s}[t]$ is proper, and hence so is $\epsilon_{\mathbf{s}}[t]$ (Neeser and Massey, 1993), which in turn implies that $\mathbf{E}[\mathbf{Z}_0(\omega)\mathbf{Z}_0^{\mathsf{T}}(0)] = \mathbf{0} d\omega$ as the pseudo-autocovariance function of $\boldsymbol{\epsilon}_{\mathbf{s}}$ is zero.

As $\mathbf{K}_{\mathbf{s},0} = E[\mathbf{s}[0]\boldsymbol{\varepsilon}^{\mathsf{H}}[0]]$, $\mathbf{K}_{\mathbf{s},0}$ must be full rank or $\dim \mathcal{D}_{\mathbf{s}}[t]$ would be strictly less than $r_0$ for all $t$. Hence, we we can choose the orthonormal basis for $\mathcal{D}_{\mathbf{s}}[0]$, which is $\boldsymbol{\varepsilon}_{\mathbf{s}}[0]$, so that $\mathbf{K}_{\mathbf{s},0}$ has lower-triangular top block with positive diagonal. Finally, $\operatorname{col} \mathcal{K}_{\mathbf{s}}(z) \subset \mathbf{K}_{\mathbf{s},0}$, as

$$\mathbf{P}_{\mathcal{S}}\mathbf{s}[t] = \sum_{j=0}^{\infty} \mathbf{P}_{\mathcal{S}} \mathbf{K}_{\mathbf{s},j} \varepsilon_{\mathbf{s}}[t-j] = \mathbf{s}[t]$$

and so the uniqueness of the linear filter coefficients yields $\mathbf{K}_{\mathbf{s},j} = \mathbf{P}_{\mathcal{S}}\mathbf{K}_{\mathbf{s},j}$ for all $j$. Hence $\mathbf{P}_{\mathcal{S}}\mathcal{K}_{\mathbf{s}}(z) = \mathcal{K}_{\mathbf{s}}(z)$ for all $z$ and we have $\operatorname{col} \mathcal{K}_{\mathbf{s}}(z) \subset \operatorname{col} \mathbf{K}_{\mathbf{s},0}$ as $\mathbf{K}_{\mathbf{s},0}$ is full rank.

A similar argument yields $\boldsymbol{\varepsilon}_{\mathbf{i}_c}[t], \mathcal{K}_c(z)$ and $\mathbf{Z}_c(\omega)$ for $c = 1, \ldots, C$ with the desired properties. As $\mathbf{u}[t]$ is white noise with $E[\mathbf{u}[0]\mathbf{u}^{\mathsf{H}}[0]] = \boldsymbol{\Phi}$, its immediately follows that it has spectral representation $\int_{\pi}^{\pi} \mathrm{e}^{\mathrm{i}\omega} \boldsymbol{\Phi}^{\frac{1}{2}} d\mathbf{W}(\omega)$ for an orthogonal increments process $\mathbf{W}(\omega)$ with the desired properties. Finally, by the assumption (3.4), it follows that the subspaces $\mathcal{H}_{\mathbf{s}}[\infty]$, $\mathcal{H}_{\mathbf{u}}[\infty]$ and $\mathcal{H}_{\mathbf{i}_c}[\infty],\ c = 1, \ldots, C$ are pairwise orthogonal, and therefore

$$E[\boldsymbol{\varepsilon}_{\mathbf{s}}[s]\boldsymbol{\varepsilon}_{\mathbf{i}_c}^{\mathsf{H}}[t]] = \mathbf{0}_{r_0,r_c}, \quad E[\boldsymbol{\varepsilon}_{\mathbf{s}}[s]\mathbf{u}^{\mathsf{H}}[t]] = \mathbf{0}_{r_0,n}, \quad E[\boldsymbol{\varepsilon}_{\mathbf{i}_c}[s]\boldsymbol{\varepsilon}_{\mathbf{i}_{c'}}^{\mathsf{H}}[t]] = \mathbf{0}_{r_c,r_{c'}}, \quad E[\boldsymbol{\varepsilon}_{\mathbf{i}_c}[s]\mathbf{u}^{\mathsf{H}}[t]] = \mathbf{0}_{r_c,n},$$

for all $s, t \in \mathbb{Z}$. As the cross-covariances of the form $E[d\mathbf{Z}_0(\omega)d\mathbf{Z}_c^{\mathsf{H}}(\omega)]$ between the orthogonal increments processes are the Fourier transforms of the above covariance functions, it follows that the orthogonal increments processes are uncorrelated and the proposition follows. □

## C.2 Proof of Proposition 10

*Proof.* Let $(\mathbf{S}_{\mathbf{ss}}(\omega), \mathbf{S}_{\mathbf{ii}}(\omega), \mathbf{\Phi})$ and $(\widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega), \widetilde{\mathbf{S}}_{\mathbf{ii}}(\omega), \widetilde{\mathbf{\Phi}})$ be two triples obtained as described in Section 3.2.2 such that

$$\mathbf{S}_{\mathbf{ss}}(\omega) + \mathbf{S}_{\mathbf{ii}}(\omega) + \mathbf{\Phi} = \mathbf{S}_{\mathbf{xx}}(\omega) = \widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega) + \widetilde{\mathbf{S}}_{\mathbf{ii}}(\omega) + \widetilde{\mathbf{\Phi}}.$$

By hypothesis, we have $(\mathbf{S}_{\mathbf{ss}}(\omega_0), \mathbf{S}_{\mathbf{ii}}(\omega_0), \mathbf{\Phi})$ equals $(\widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega_0), \widetilde{\mathbf{S}}_{\mathbf{ii}}(\omega_0), \widetilde{\mathbf{\Phi}})$ for some $\omega_0$. Hence $\mathbf{\Phi} = \widetilde{\mathbf{\Phi}}$ and so

$$\mathbf{S}_{\mathbf{ss}}(\omega) + \mathbf{S}_{\mathbf{ii}}(\omega) = \mathbf{S}_{\mathbf{xx}}(\omega) - \mathbf{\Phi} = \widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega) + \widetilde{\mathbf{S}}_{\mathbf{ii}}(\omega) \tag{C.1}$$

for all $\omega$. Next, let $\mathbf{P}_{\mathcal{S}}$ be an oblique projection onto $\operatorname{col} \mathbf{S}_{\mathbf{ss}}(\omega_0)$ whose kernel contains $\operatorname{col} \mathbf{S}_{\mathbf{ii}}(\omega_0)$. Such an oblique projection must exist, as $\mathbf{S}_{\mathbf{ss}}(\omega) = \mathcal{K}_{\mathbf{s}}(\mathrm{e}^{\mathrm{i}\omega})\mathcal{K}_{\mathbf{s}}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})$ with $\mathcal{K}_{\mathbf{s}}(\mathrm{e}^{\mathrm{i}\omega})$ an $n \times r_0$ matrix implies that $\operatorname{rank} \mathbf{S}_{\mathbf{ss}}(\omega) \leq r_0$ for all $\omega$. Similarly, $\operatorname{rank} \mathbf{S}_{\mathbf{ii}}(\omega) \leq r$ for all $\omega$. Hence (10) implies that $\operatorname{col} \mathbf{S}_{\mathbf{ss}}(\omega_0) \cap \operatorname{col} \mathbf{S}_{\mathbf{ii}}(\omega_0) = \{\mathbf{0}\}$, and so a projection $\mathbf{P}_{\mathcal{S}}$ with the desired image and kernel exists.

Next, as $\operatorname{col} \mathbf{K}_{\mathbf{s}}(z) \subset \operatorname{col} \mathbf{K}_{\mathbf{s},0}$ by Proposition 9, we have $\operatorname{col} \mathbf{S}_{\mathbf{ss}}(\omega_0) \subset \operatorname{col} \mathbf{K}_{\mathbf{s},0}$. In fact, we have $\operatorname{col} \mathbf{S}_{\mathbf{ss}}(\omega_0) = \operatorname{col} \mathbf{K}_{\mathbf{s},0}$ as both subspaces are dimension $r_0$. So, $\mathbf{P}_{\mathcal{S}}\mathbf{K}_{\mathbf{s},0} = \mathbf{K}_{\mathbf{s},0}$ and therefore $\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{ss}}(\omega) = \mathbf{S}_{\mathbf{ss}}(\omega)$. A similar argument shows that $\operatorname{col} \mathbf{S}_{\mathbf{ii}}(\omega) \subset \operatorname{col} \mathbf{S}_{\mathbf{ii}}(\omega_0)$ for all $\omega$. Thus, $\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{ii}}(\omega) = \mathbf{0}_{n,n}$ for all $\omega$, as we have $\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{ii}}(\omega_0) = \mathbf{0}_{n,n}$ by construction. Left-multiplying (C.1) by $\mathbf{P}_{\mathcal{S}}$, we obtain

$$\mathbf{P}_{\mathcal{S}}(\mathbf{S}_{\mathbf{xx}}(\omega) - \mathbf{\Phi}) = \mathbf{S}_{\mathbf{ss}}(\omega) = \mathbf{P}_{\mathcal{S}}\widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega) + \mathbf{P}_{\mathcal{S}}\widetilde{\mathbf{S}}_{\mathbf{ii}}(\omega).$$

However, a similar argument to the above shows that $\operatorname{col} \widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega) \subset \operatorname{col} \widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega_0) = \operatorname{col} \mathbf{S}_{\mathbf{ss}}(\omega_0)$, while $\operatorname{col} \widetilde{\mathbf{S}}_{\mathbf{ii}}(\omega) \subset \operatorname{col} \widetilde{\mathbf{S}}_{\mathbf{ii}}(\omega_0)$. Thus, $\mathbf{P}_{\mathcal{S}}\widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega) = \widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega)$ and $\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{ii}}(\omega) = \mathbf{0}_{n,n}$ for all $\omega$, and therefore $\mathbf{S}_{\mathbf{ss}}(\omega) = \widetilde{\mathbf{S}}_{\mathbf{ss}}(\omega)$. Subtracting $\mathbf{S}_{\mathbf{ss}}(\omega)$ from both sides of (C.1) yields $\mathbf{S}_{\mathbf{ii}}(\omega) = \widetilde{\mathbf{S}}_{\mathbf{ii}}(\omega)$ for all $\omega$.

□

## C.3 Proof of Proposition 11

*Proof.* First, define the coordinate $z$-spectrum $\mathbf{S}_{\mathbf{ff}}(z) \in \mathbb{C}_{r_0 \times r_0}$ as

$$\mathbf{S}_{\mathbf{ff}}(z) = \mathbf{K}_{\mathbf{s},0}^{-} \mathcal{K}_{\mathbf{s}}(z) \mathcal{K}_{\mathbf{s}}^{\mathsf{H}}(1/\overline{z}) (\mathbf{K}_{\mathbf{s},0}^{-})^{\mathsf{H}},$$

where $\mathbf{K}_{\mathbf{s},0}^{-}$ is the Moore-Penrose pseudo-inverse of $\mathbf{K}_{\mathbf{s},0}$. Then $\mathbf{S}_{\mathbf{ff}}(z)$ has the Cholesky-like decomposition $\mathbf{S}_{\mathbf{ff}}(z) = \mathbf{L}_{+}(z)\mathbf{L}_{+}^{\mathsf{H}}(1/\overline{z})$ with $\mathbf{L}_{+}(z)$ being a lower-triangular matrix of rational functions whose diagonal entries are the corresponding canonical scalar spectral factors (see Ephremidze et al. (2024); Janashia et al. (2011) for details of this construction) and so all zeroes and poles of the rational functions on the main diagonal have absolute value $\leq 1$. The off-diagonal entries of $\mathbf{L}_{+}(z)$ may have poles for $|z| \geq 1$. However, the assumption that all entries of $\mathcal{K}_{\mathbf{s}}(z)$ are finite if $|z| = 1$ implies that all entries of $\mathbf{S}_{\mathbf{ff}}(z)$ are finite when $|z| = 1$. This in turn ensures that all entries of $\mathbf{L}_{+}(z)$ are finite when $|z| = 1$. To see this, first consider the base case where $\mathbf{S}_{\mathbf{ff}}(z)$ is $1 \times 1$, in which case $|\mathbf{L}_{+}(z)\overline{\mathbf{L}_{+}(z)}| = |\mathbf{S}_{\mathbf{ff}}(z)| < \infty$ implies $|\mathbf{L}_{+}(z)| < \infty$ as desired. Returning to the original case, note that the upper-left entry $[\mathbf{S}_{\mathbf{ff}}(z)]_{11}$ is $[\mathbf{L}_{+}(z)]_{11}\overline{[\mathbf{L}_{+}(z)]_{11}}$, hence $[\mathbf{L}_{+}(z)]_{11}$ is finite when $|z| = 1$. Unless $[\mathbf{L}_{+}(z)]_{11}$ is identically zero, the remaining entries in the first column of $\mathbf{L}_{+}(z)$ are

$$[\mathbf{L}_{+}(z)]_{i1} = \frac{1}{[\mathbf{L}_{+}(z)]_{11}} [\mathbf{S}_{\mathbf{ff}}(z)]_{i1}.$$

By hypothesis, $\mathcal{K}_{\mathbf{s}}(z)$ is full rank on the unit circle and so $\mathbf{S}_{\mathbf{ff}}(z)$ is positive definite when $|z| = 1$. Hence $[\mathbf{L}_{+}(z)]_{11} \neq 0$ on the unit circle, so $|[\mathbf{L}_{+}(z)]_{i1}| < \infty$ when $|z| = 1$. For the rest of the columns of $\mathbf{L}_{+}(z)$, note that they are the obtained by factorizing the Schur complement of $[\mathbf{L}_{+}(z)]_{11}$ within $\mathbf{L}_{+}(z)$, which is an $(r_0 - 1) \times (r_0 - 1)$ matrix of rational functions of $z^{-1}$ which is positive definite with finite entries for all $|z| = 1$. By induction, we then have that no entry of $\mathbf{L}_{+}(z)$ has a pole on the unit circle.

Next, suppose the $ij$th entry of $\mathbf{L}_{+}(z)$ has a pole $p$ with $|p| > 1$. Right-multiplying $\mathbf{L}_0(z)$ by the diagonal paraunitary transformation $\mathbf{U}_p(z)$ whose $jj$th entry is $(z^{-1} - p)/(1 - \overline{p}z^{-1})$ and all other

diagonal entries are 1 reflects the pole $p$ in the $ij$th entry to $1/\overline{p}$, which is within the unit circle. For other entries in the $j$th column, the right-multiplication introduces a zero at $p$ and a pole at $1/\overline{p}$. This procedure can be iterated until all poles outside the unit circle have been reflected, and we can define $\mathbf{L}'(z)$ as the resulting rational lower triangular matrix. For the south-east entry of $\mathbf{L}_+(z)$, it already has all poles and zeros within the unit circle and so is unchanged.

Next, note that $\det \mathbf{L}'(z)$ is the product of $\det \mathbf{L}_+(z)$ and a finite number of the terms $\det \mathbf{U}_p(z)$. As $\det \mathbf{L}_+(z)$ is the product of the canonical spectral factors on the diagonal, $\lim_{z\to\infty} \det \mathbf{L}_+(z)$ is finite and non-zero as the spectral factors have all poles and zeros within the unit circle. Further, $\lim_{z\to\infty} \det \mathbf{U}_p(z) = -p$ which is non-zero as $|p| > 1$. Hence $\lim_{z\to\infty} \det \mathbf{L}'(z)$ is finite and non-zero, so $\mathbf{L}'_0 \equiv \lim_{z\to\infty} \mathbf{L}'(z)$ is full-rank. Let $\mathbf{L}(z) = \mathbf{L}'(z)\mathbf{U}$, where $\mathbf{U}$ is the constant diagonal unitary unitary matrix that takes the diagonal elements of $\mathbf{L}'_0$ to be positive. As $\mathbf{L}(z)$ is obtained from $\mathbf{L}_+(z)$ by right-multiplication with paraunitary matrices, the factorization $\mathbf{L}(z)\mathbf{L}^{\mathsf{H}}(1/\overline{z}) = \mathbf{S}_{\mathbf{ff}}(z)$ continues to hold. As the constant term of $[\mathbf{L}_+]_{r_0 r_0}$ is already 1, we have $[\mathbf{L}(z)]_{r_0 r_0} = [\mathbf{L}_+(z)]_{r_0 r_0}$.

Further, define $\mathbf{L}_0 \equiv \mathbf{L}'_0\mathbf{U} \equiv \mathbf{T}\mathbf{\Psi}^{\frac{1}{2}}$ where $\mathbf{\Psi}^{\frac{1}{2}}$ is diagonal with $[\mathbf{\Psi}^{\frac{1}{2}}]_{jj} = [\mathbf{K}_{\mathbf{s},0}]_{jj}[\mathbf{L}'_0\mathbf{U}]_{jj}$ for $j = 1, \ldots, r_0$, which is positive by the construction of $\mathbf{K}_{\mathbf{s},0}$ and $\mathbf{L}'_0\mathbf{U}$. The matrix $\mathbf{T}$ is then lower-triangular with $[\mathbf{T}]_{jj} = 1/[\mathbf{K}_{\mathbf{s},0}]_{jj},\ j = 1, \ldots, r_0$ and so $\mathbf{K}_{\mathbf{s},0}\mathbf{T}$ has unit-triangular top $r_0 \times r_0$ block. If we define $\mathbf{R}(z) = \mathbf{T}^{-1}\mathbf{L}(z)\mathbf{\Psi}^{-\frac{1}{2}}$, then $\lim_{z\to\infty} \mathbf{R}(z) = \mathbf{I}_{r_0}$.

To show that $\mathbf{R}(z)$ has the appropriately-structured factorization, let $d(z)$ be the monic least common multiple of the denominator polynomials for all entries in $\mathbf{L}(z)$, which has $d(z) \neq 0$ for all $|z| \geq 1$ by the construction of $\mathbf{L}(z)$. Then $(\mathbf{T}^{-1}d(z)\mathbf{L}(z), d(z)\mathbf{\Psi}^{-\frac{1}{2}})$ is a right matrix fractional decomposition of $\mathbf{R}(z)$ as $d(z)\mathbf{L}(z)$ is a matrix of polynomials in $z^{-1}$. Next, let $(\mathbf{M}(z), \mathbf{C}(z))$ be an irreducible right matrix fractional decomposition of $\mathbf{R}(z)$, with $\mathbf{M}(z)$ and $\mathbf{C}(z)$ being lower-triangular matrices of polynomials in $z^{-1}$ whose constant terms are $\mathbf{I}_{r_0}$ and $[\mathbf{C}(z)]_{jj} \neq 0$ for all $|z| \geq 1$, $j = 1, \ldots, r_0$. Such an MFD must exist, as if $(\mathbf{M}''(z), \mathbf{C}''(z))$ is any irreducible right MFD of $\mathbf{R}(z)$, then $\mathbf{C}''(z)$ can be made lower-triangular through right-multiplication by a unimodular matrix $\mathbf{B}(z)$ which takes $\mathbf{C}''(z)$ to row-Hermite form. So, $(\mathbf{M}'(z), \mathbf{C}'(z)) \equiv (\mathbf{M}''(z)\mathbf{B}(z), \mathbf{C}''(z)\mathbf{B}(z))$ is another irreducible right MFD with $\mathbf{C}'(z)$ lower-triangular. As $\mathbf{M}'(z)$ equals $\mathbf{R}(z)\mathbf{C}'(z)$, this also

implies $\mathbf{M}'(z)$ is lower-triangular. So,

$$\mathbf{I}_{r_0} = \lim_{z\to\infty} \mathbf{R}(z) = \lim_{z\to\infty} \mathbf{M}'(z)\mathbf{C}'^{-1}(z)$$

and therefore the constant terms $\mathbf{M}'_0$ and $\mathbf{C}'_0$ must be full-rank and equal. Letting $(\mathbf{M}(z), \mathbf{C}(z)) \equiv (\mathbf{M}'(z)\mathbf{M}'^{-1}_0, \mathbf{C}'(z)\mathbf{C}'^{-1}_0)$, we have an irreducible right MFD with the numerator and denominator polynomials being lower-triangular with constant term $\mathbf{I}_{r_0}$. As $(\mathbf{T}^{-1}d(z)\mathbf{L}(z), d(z)\mathbf{\Psi}^{-\frac{1}{2}})$ is a right MFD and $(\mathbf{M}(z), \mathbf{C}(z))$ is an irreducible right MFD of $\mathbf{R}(z)$, there exists a matrix exists a matrix $\mathbf{W}(z)$ which is polynomial in $z^{-1}$ such that $d(z)\mathbf{\Phi}^{-\frac{1}{2}} = \mathbf{C}(z)\mathbf{W}(z)$ (Kailath, 1980, Lemma 6.5-5). Hence,

$$d(z) = \det \mathbf{\Phi}^{\frac{1}{2}} \det \mathbf{C}(z) \det \mathbf{W}(z),$$

and as the right-hand side is a polynomial in $z^{-1}$, we can conclude that $\det \mathbf{C}(z) \neq 0$ for $|z| \geq 1$. As $\mathbf{C}(z)$ is lower-triangular, this implies that the main-diagonal polynomials $[\mathbf{C}(z)]_{jj} \neq 0$ for $|z| \geq 1$, $j = 1, \ldots, r_0$, as desired. Finally, as $[\mathbf{L}_0]_{r_0r_0} = 1$, the south-east entry $[\mathbf{R}(z)]_{r_0r_0}$ equals $[\mathbf{L}_+(z)]_{r_0r_0}$, so $[\mathbf{M}(z)]_{r_0r_0}$ has all zeros within the unit circle by construction.

Combining the above definitions,

$$\begin{aligned}\mathbf{S}_{\mathbf{ff}}(z) &= \mathbf{L}(z)\mathbf{L}^{\mathsf{H}}(1/\overline{z}) = \mathbf{T}\mathbf{R}(z)\mathbf{\Psi}\mathbf{R}^{\mathsf{H}}(1/\overline{z})\mathbf{T}^{\mathsf{H}} \\ &= \mathbf{T}\mathbf{M}(z)\mathbf{C}^{-1}(z)\mathbf{\Psi}\mathbf{C}^{-\mathsf{H}}(1/\overline{z})\mathbf{M}^{\mathsf{H}}(1/\overline{z})\mathbf{T}^{\mathsf{H}}.\end{aligned}$$

As $\operatorname{col}\mathcal{K}_{\mathbf{s}}(z) \subset \operatorname{col}\mathbf{K}_{\mathbf{s},0}$ for all $z$, $\mathbf{K}_{\mathbf{s},\mathbf{0}}\mathbf{K}^{-}_{\mathbf{s},0}\mathcal{K}_{\mathbf{s},0}(z) = \mathcal{K}_{\mathbf{s}}(z)$. Letting $\mathbf{A} = \mathbf{K}_{s,0}\mathbf{T}$, whose top $r_0 \times r_0$ block is unit lower-triangular by the construction of $\mathbf{T}$, we have

$$\mathbf{S}_{\mathbf{ss}}(\omega) = \mathbf{A}\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{C}^{-1}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{\Psi}\mathbf{C}^{-\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{M}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{A}^{\mathsf{H}}.$$

Defining $\mathbf{S}_{\mathbf{g}_c\mathbf{g}_c}(z) = \mathbf{K}^{-}_{\mathbf{i}_c,0}\mathcal{K}_{\mathbf{i}_c}(z)\mathcal{K}^{\mathsf{H}}(1/\overline{z})(\mathbf{K}^{-}_{\mathbf{i}_c,0})^{\mathsf{H}}$, an analogous argument to that above provides

$$\mathbf{S}_{\mathbf{i}_c\mathbf{i}_c}(\omega) = \mathbf{B}_c\mathbf{N}_c(\mathrm{e}^{\mathrm{i}\omega})\mathbf{D}_c^{-1}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{\Upsilon}_c\mathbf{D}_c^{-\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{N}_c^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{B}_c^{\mathsf{H}},$$

with the appropriately-structured factors. Setting $r_0 = \dim\operatorname{col}\mathbf{K}_{\mathbf{s},0}$ and $r_c = \dim\operatorname{col}\mathbf{K}_{\mathbf{i}_c,0},\ c = 1, \ldots, C$ and $p_0, p_1, \ldots p_C$ and $q_0, q_1, \ldots, q_C$ as the maximum polynomial degree of the entries of

$\mathbf{M}(z)$,$\mathbf{N}_1(z), \ldots, \mathbf{N}_C(z)$ and $\mathbf{C}(z)$,$\mathbf{D}_1(z), \ldots, \mathbf{D}_C(z)$ respectively, the above construction implies that there is a parameter tuple $\boldsymbol{\theta} \in \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ such that

$$\mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta}) = \mathbf{S}_{\mathbf{xx}}(\omega)$$

for all $\omega \in (-\pi, \pi]$. □

## C.4 Proof of Proposition 12

*Proof.* By hypothesis, there is a null subset $N' \subset \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ such that the $\mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta})$ uniquely determines $(\mathbf{S}_{\mathbf{ss}}(\omega; \boldsymbol{\theta}), \mathbf{S}_{\mathbf{ii}}(\omega; \boldsymbol{\theta}), \boldsymbol{\Phi})$ on $\boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q}) \setminus N'$. Suppose that $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ are such that

$$\mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta}) = \mathbf{S}_{\mathbf{xx}}(\omega; \widetilde{\boldsymbol{\theta}}) \tag{C.2}$$

for all $\omega \in (-\pi, \pi]$. For notational convenience, let $\mathcal{K}_{\mathbf{s}}(z), \mathcal{K}_{\mathbf{i}_c}(z)$ and $\widetilde{\mathcal{K}}_{\mathbf{s}}(z),\ \widetilde{\mathcal{K}}_{\mathbf{i}_c}(z)$ denote the transfer functions (3.10) $\mathcal{K}_{\mathbf{s}}(z; \boldsymbol{\theta}), \mathcal{K}_{\mathbf{i}_c}(z; \boldsymbol{\theta})$ and $\mathcal{K}_{\mathbf{s}}(z; \widetilde{\boldsymbol{\theta}}), \mathcal{K}_{\mathbf{i}_c}(z; \widetilde{\boldsymbol{\theta}})$ respectively, for $c = 1, \ldots, C$. Similarly, denote the components of $\widetilde{\boldsymbol{\theta}}$ as $\widetilde{\boldsymbol{\Phi}}$, $\widetilde{\mathbf{A}}$, and so on. Next, let $D_0^c \subset \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ contain parameter tuples such that the sets of zeros of $\det \mathbf{M}(z)$ and $\det \mathbf{C}(z)$ are disjoint and the determinants have maximal degree (that is, $\deg \det \mathbf{M}(z) = r_0 p_0$ and $\deg \det \mathbf{C}(z) = r_0 q_0$), and define similarly $D_c^c$ with respect to $\det \mathbf{N}_c(z)$ and $\det \mathbf{D}_c(z)$ for $c = 1, \ldots, C$. Finally, let $Z_0 \subset \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ contain parameter tuples such that there exists a $z_0 \neq 0$ where one of the initial $r_0 - 1$ columns of $\mathcal{K}_{\mathbf{s}}(z)$ vanishes identically at $z_0$ ( that is, $[\mathcal{K}_{\mathbf{s}}(z_0)]_{ij} = 0$ for some $1 \leq j < r_0$ and all $i = 1, \ldots n$). Similarly define $Z_c$ with respect to the initial $r_c - 1$ columns of $\mathcal{K}_{\mathbf{i}_c}(z)$ for $c = 1, \ldots, C$.

To see that the above sets are null, first consider $D_0$. If $\boldsymbol{\theta} \in D_0$, then either some leading coefficient $[\mathbf{M}_{p_0}]_{jj}$ or $[\mathbf{C}_{q_0}]_{jj}$ is zero for some $j = 1, \ldots, r_0$ or the tuple $(x_1, \ldots, x_{r_0 p_0}, y_1, \ldots, y_{r_0 q_0})$ has some repeated element where $x_1, \ldots, x_{r_0 p_0}$ are the roots of $\det \mathbf{M}(z)$ and $y_1, \ldots, y_{r_0 q_0}$ are the roots of $\det \mathbf{C}(z)$, counted with multiplicity. As the coefficients of $[\mathbf{M}(z)]_{jj}$ and $[\mathbf{C}(z)]_{jj}$ live in an open subset of $\mathbb{C}_{p_0}$ and $\mathbb{C}_{q_0}$ respectively, the first case corresponds to a Lebesgue null subset of $\boldsymbol{\Theta}$. For the second case, note that each root belongs to an open subset of $\mathbb{C}$ and so the set of possible tuples is an open subset of $\mathbb{C}_{r_0(p_0+q_0)}$. It can then be shown that the set of tuples with non-distinct

elements is Lebesgue null and hence the corresponding set of coefficients for $[\mathbf{M}(z)]_{jj}$ and $[\mathbf{C}(z)]_{jj}$, $j = 1, \ldots, r_0$ is also Lebesgue null.

Next, we will show that $Z_0 \cap D_0^c$ is null, and so $Z_0$ is null as $D_0$ is null by the above argument. To see this, note the $j$th column of $\mathbf{K_s}(z_0)$ vanishing is equivalent to the $j$th column of $\mathbf{M}(z_0)\mathbf{C}^{-1}(z_0)$ vanishing as $\mathbf{A}$ is full-rank and $\mathbf{\Phi}$ is full-rank and diagonal. For $j < r_0$, the main-diagonal and sub-diagonal entries of $\mathbf{M}(z)\mathbf{C}^{-1}(z)$ respectively are

$$r_j(z) \equiv \frac{[\mathbf{M}(z)]_{jj}}{[\mathbf{C}(z)]_{jj}},$$
$$s_j(z) \equiv [\mathbf{M}(z)]_{(j+1)j} - \frac{[\mathbf{M}(z)]_{(j+1)(j+1)}[\mathbf{C}(z)]_{(j+1)j}}{[\mathbf{C}(z)]_{jj}}.$$

Either $[\mathbf{M}(z)]_{(j+1)j}$ and $[\mathbf{C}(z)]_{(j+1)j}$ are zero for all $j$ or $s_j(z)$ is a non-constant rational function with at most $p_0 q_0$ zeros. The former case is clearly Lebesgue null in the set of all coefficients, so assume that $s_j(z)$ is not always zero. As $s_j(z)$ does not depend on the coefficients of $[\mathbf{M}(z)]_{jj}$, it can be seen that for any tuple of zeroes $(x_1, \ldots, x_{p_0})$ of $[\mathbf{M}(z)]_{jj}$, the set of coefficients of $[\mathbf{M}(z)]_{(j+1)j}$, $[\mathbf{M}(z)]_{(j+1)(j+1)}$, $[\mathbf{C}(z)]_{(j+1)j}$ and $[\mathbf{C}(z)]_{jj}$ such that $s_j(x_k) = 0$ for some $k = 1, \ldots, p_0$ is null in the open set of all possible coefficients.Then as $\boldsymbol{\theta} \in Z_0$ implies that there is some $j < r_0$ and some $z_0$ such that $t_j(z_0) = 0 = s_j(z_0)$, the above argument yields that $Z_0$ is Lebesgue null. A similar argument shows that $D_c$ and $Z_c$ are Lebesgue-null, for $c = 1, \ldots, C$.

Let $N$ be $N' \cup D_0 \cup Z_0 \cup \cdots \cup D_C \cup Z_C$, which is a finite union of null subsets and so is null. Assume that $\theta \in \mathbf{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q}) \setminus N$. As $\theta \notin N'$, $\mathbf{S_{xx}}(\omega; \boldsymbol{\theta})$ uniquely determines $(\mathbf{S_{ss}}(\omega; \boldsymbol{\theta}), \mathbf{S_{ii}}(\omega; \boldsymbol{\theta}), \mathbf{\Phi})$ for all $\omega$. We must then have that $\mathbf{S_{ss}}(\omega; \boldsymbol{\theta}) = \mathbf{S_{ss}}(\omega; \widetilde{\boldsymbol{\theta}})$, $\mathbf{S_{ii}}(\omega; \boldsymbol{\theta}) = \mathbf{S_{ii}}(\omega; \widetilde{\boldsymbol{\theta}})$ and $\mathbf{\Phi} = \widetilde{\mathbf{\Phi}}$, as otherwise $\mathbf{S_{ss}}(\omega; \widetilde{\boldsymbol{\theta}}) + \mathbf{S_{ii}}(\omega; \widetilde{\boldsymbol{\theta}}) + \widetilde{\mathbf{\Phi}}$ would provide a distinct decomposition of $\mathbf{S_{xx}}(\omega; \boldsymbol{\theta})$, violating the assumption that $\boldsymbol{\theta} \in N'$.

As $\mathcal{K}_{\mathbf{s}}(z), \mathcal{K}_{\mathbf{i}_c}(z)$ and $\widetilde{\mathcal{K}}_{\mathbf{s}}(z), \widetilde{\mathcal{K}}_{\mathbf{i}_c}(z)$ are analytic for $|z| \geq 1$, typical results in rational spectral factorization (Rozanov, 1960) show that $\mathbf{S_{ss}}(\omega; \boldsymbol{\theta}) = \mathbf{S_{ss}}(\omega; \widetilde{\boldsymbol{\theta}})$ for all $\omega$ implies

$$\mathcal{K}_{\mathbf{s}}(z) = \widetilde{\mathcal{K}}_{\mathbf{s}}(z)\mathbf{U}(z) \tag{C.3}$$

for all $z$, where $\mathbf{U}(z)$ is a paraunitary matrix of rational functions of $z^{-1}$. As the top $r_0 \times r_0$ blocks of $\mathcal{K}_\mathbf{s}(z)$ and $\widetilde{\mathcal{K}}_\mathbf{s}(z)$ are lower-triangular and $\mathbf{U}(z)$ is unitary when $|z| = 1$, we further have that $\mathbf{U}(z)$ is diagonal. So,

$$[\mathcal{K}_\mathbf{s}(z)]_{ij} = [\widetilde{\mathcal{K}}_\mathbf{s}(z)]_{ij}[\mathbf{U}(z)]_{ii} \tag{C.4}$$

for all $j = 1, \ldots, r_0, i = j, \ldots, n$.

For the south-east entry where $j = r_0$, both $[\mathbf{M}(z)]_{r_0 r_0}/[\mathbf{C}(z)]_{r_0 r_0}$ and $[\widetilde{\mathbf{M}}(z)]_{r_0 r_0}/[\widetilde{\mathbf{C}}(z)]_{r_0 r_0}$ have all poles and zeros within the unit circle. By the uniqueness of the canonical spectral factorization in the scalar case, this implies that $[\mathbf{U}(z)]_{r_0 r_0}$ is a unit-magnitude constant. As the constant term in $\mathbf{M}(z), \widetilde{\mathbf{M}}(z)$ and $\mathbf{C}(z), \widetilde{\mathbf{C}}(z)$ is $\mathbf{I}_{r_0}$, we obtain that $[\mathbf{U}(z)]_{r_0 r_0} = 1$.

Next, consider the case with $j < r_0$ and $i = j, \ldots, n$. By construction, $[\mathcal{K}_\mathbf{s}(z)]_{ij}$ and $[\widetilde{\mathcal{K}}_\mathbf{s}(z)]_{ij}$ have no poles for $|z| \geq 1$, and so $[\mathbf{U}(z)]_{jj} = 0$ implies $[\mathcal{K}_\mathbf{s}(z)]_{ij} = 0$ for any $|z| \geq 1$. As $\boldsymbol{\theta} \notin Z_0$ implies that the $j$th column of $\mathbf{K}_\mathbf{s}(z)$ does not vanish identically, this implies that $[\mathbf{U}(z)]_{jj} \neq 0$ for all $|z| \geq 1$ and $j = 1, \ldots, r_0 - 1$.

As $[\mathbf{U}(z)]_{jj}$ is a scalar paraunitary transformation, the fact that $[\mathbf{U}(z)]_{jj}$ has no zeros for $|z| \geq 1$ implies that $[\mathbf{U}(z)]_{jj}$ has no poles inside the unit circle, for $j = 1, \ldots, r_0 - 1$. For the main diagonal with $i = j$, (C.4) becomes

$$\frac{[\mathbf{M}(z)]_{jj}[\mathbf{\Phi}]_{jj}}{[\mathbf{C}(z)]_{jj}} = \frac{[\widetilde{\mathbf{M}}(z)]_{jj}[\widetilde{\mathbf{\Phi}}]_{jj}}{[\widetilde{\mathbf{C}}(z)]_{jj}}[\mathbf{U}(\mathbf{z})]_{jj}.$$

As $[\mathbf{U}(z)]_{jj}$ has no poles in the unit circle, if $p$ with $|p| < 1$ is a zero of $[\mathbf{C}(z)]_{jj}$ with multiplicity $m$, then $p$ must also be a zero of $[\widetilde{\mathbf{C}}(z)]_{jj}$ with multiplicity $\widetilde{m} \geq m$. However, as $\deg[\widetilde{\mathbf{C}}(z)]_{jj} \leq \deg[\mathbf{C}(z)]_{jj}$ by $\boldsymbol{\theta} \in D_0^c$, we in fact have $\widetilde{m} = m$. As both $[\widetilde{\mathbf{C}}(z)]_{jj}$ and $[\mathbf{C}(z)]_{jj}$ have all zeros within the unit circle and constant term 1, this implies $[\mathbf{C}(z)]_{jj} = [\widetilde{\mathbf{C}}(z)]_{jj}$. So, we in fact have

$$[\mathbf{M}(z)]_{jj}[\mathbf{\Phi}]_{jj} = [\widetilde{\mathbf{M}}(z)]_{jj}[\widetilde{\mathbf{\Phi}}]_{jj}[\mathbf{U}(\mathbf{z})]_{jj}. \tag{C.5}$$

Next, note that $(\det \mathbf{C}(z))\mathbf{C}^{-1}(z)$ and $(\det \mathbf{C}(z))\widetilde{\mathbf{C}}^{-1}(z)$ are matrices of polynomials in $z^{-1}$. This follows as the application of back-substitution to compute $\mathbf{C}^{-1}(z)$ and $\widetilde{\mathbf{C}}^{-1}(z)$ shows that all the denominator polynomials in $\mathbf{C}^{-1}(z)$ and $\widetilde{\mathbf{C}}^{-1}(z)$ divide $\det \mathbf{C}(z)$ and $\det \widetilde{\mathbf{C}}(z)$ respectively.

Then as $\det \widetilde{\mathbf{C}}(z) = \prod_{j=1}^{r_0} [\widetilde{\mathbf{C}}(z)]_{jj} = \prod_{j=1}^{r_0} [\mathbf{C}(z)]_{jj} = \det \mathbf{C}(z)$ by the previous step, we have that all entries in $(\det \mathbf{C}(z))\mathbf{C}^{-1}(z)$ and $(\det \mathbf{C}(z))\widetilde{\mathbf{C}}^{-1}(z)$ are polynomials in $z^{-1}$. So,

$$\frac{\det \mathbf{C}(z)[\mathcal{K}_{\mathbf{s}}(z)]_{ij}}{\det \mathbf{C}(z)[\widetilde{\mathcal{K}}_{\mathbf{s}}(z)]_{ij}} = \frac{[\mathbf{M}(z)\big((\det \mathbf{C}(z))\mathbf{C}^{-1}(z)\big)]_{ij}}{[\widetilde{\mathbf{M}}(z)\big((\det \mathbf{C}(z))\widetilde{\mathbf{C}}^{-1}(z)\big)]_{ij}}$$

is a ratio of polynomials in $z^{-1}$ which equals $[\mathbf{U}(\mathbf{z})]_{jj}$ by (C.4). Hence if $[\mathbf{U}(\mathbf{z})]_{jj}$ equals zero for $z \neq 0$, the numerator $\det \mathbf{C}(z)[\mathcal{K}_{\mathbf{s}}(z)]_{ij}$ must be 0 for all $i = j, \ldots, n$. However, (C.5) implies that, if $[\mathbf{U}(z)]_{jj} = 0$ then $[\mathbf{M}(z)]_{jj} = 0$. As the zeros of the main-diagonal polynomials of $\mathbf{M}(z)$ and $\mathbf{C}(z)$ are disjoint by $\boldsymbol{\theta} \in D_0^c$, $[\mathbf{M}(z)]_{jj} = 0$ implies $\det \mathbf{C}(z) \neq 0$, hence we in fact have $[\mathcal{K}_{\mathbf{s}}(z)]_{ij} = 0$ for all $i = j, \ldots, n$, which contradicts $\boldsymbol{\theta} \notin Z_0$. So, $[\mathbf{U}(z)]_{jj}$ is a paraunitary transformation which is non-zero for all $z \in \mathbb{C}$, and therefore it equals $\mathrm{e}^{\mathrm{i}\varphi}$ for some $\varphi \in \mathbb{R}$. Finally, taking the limit of (C.5) as $z \to \infty$, we obtain

$$[\boldsymbol{\Phi}^{\frac{1}{2}}]_{jj} = [\widetilde{\boldsymbol{\Phi}}^{\frac{1}{2}}]_{jj} \mathrm{e}^{\mathrm{i}\varphi}.$$

As $[\boldsymbol{\Phi}]_{jj}$ and $[\widetilde{\boldsymbol{\Phi}}]_{jj}$ are positive, we conclude $[\mathbf{U}(z)]_{jj} = 1$. Thus, $\mathbf{U}(z) = \mathbf{I}_{r_0}$ and so the transfer functions are equal, $\mathcal{K}_{\mathbf{s}}(z) = \widetilde{\mathcal{K}}_{\mathbf{s}}(z)$.

It remains to show that $\mathbf{A} = \widetilde{\mathbf{A}}$, $\mathbf{M} = \widetilde{\mathbf{M}}$, $\mathbf{C} = \widetilde{\mathbf{C}}$, and $\boldsymbol{\Phi} = \widetilde{\boldsymbol{\Phi}}$. Taking the limit as $z \to \infty$ of $\mathcal{K}_{\mathbf{s}}(z) = \widetilde{\mathcal{K}}_{\mathbf{s}}(z)$, we obtain

$$\mathbf{A}\boldsymbol{\Phi}^{\frac{1}{2}} = \widetilde{\mathbf{A}}\widetilde{\boldsymbol{\Phi}}^{\frac{1}{2}},$$

and as $\mathbf{A}, \widetilde{\mathbf{A}}$ are unit-triangular and $\boldsymbol{\Phi}, \widetilde{\boldsymbol{\Phi}}$ are positive and diagonal, we obtain $\mathbf{A} = \widetilde{\mathbf{A}}$ and $\boldsymbol{\Phi} = \widetilde{\boldsymbol{\Phi}}$. As the top $r_0 \times r_0$ block of $\mathbf{A}$ and $\boldsymbol{\Phi}^{\frac{1}{2}}$ are both invertible, $\mathcal{K}_{\mathbf{s}}(z) = \widetilde{\mathcal{K}}_{\mathbf{s}}(z)$ implies

$$\mathbf{M}(z)\mathbf{C}^{-1}(z) = \widetilde{\mathbf{M}}(z)\widetilde{\mathbf{C}}^{-1}(z).$$

As $(\mathbf{M}(z), \mathbf{C}(z))$ and $(\widetilde{\mathbf{M}}(z), \widetilde{\mathbf{C}}(z))$ are both right MFDs of the same transfer function, we have $\widetilde{\mathbf{C}}(z) = \mathbf{C}(z)\mathbf{R}(z)$ for some matrix $\mathbf{R}(z)$ of polynomials in $z^{-1}$, which must be lower-triangular. As shown above, the diagonal elements of $\mathbf{C}(z)$ and $\widetilde{\mathbf{C}}(z)$ are equal, hence $\mathrm{diag}(\mathbf{R}(z)) = \mathbf{1}_{r_0}$. Taking limits as $z \to \infty$ of $\mathbf{C}(z)$ and $\widetilde{\mathbf{C}}(z)\mathbf{R}(z)$, we also obtain that the constant term of $\mathbf{R}(z)$ is

$\mathbf{I}_{r_0}$. For the first sub-diagonal of $\widetilde{\mathbf{C}}(z)$, we then have

$$[\widetilde{\mathbf{C}}(z)]_{(j+1)j} = [\mathbf{C}(z)]_{(j+1)j} + [\mathbf{C}(z)]_{jj}[\mathbf{R}(z)]_{(j+1)j}.$$

As $\deg[\mathbf{C}(z)]_{jj} = q_0$ by $\boldsymbol{\theta} \in D_0^c$ while $\deg[\mathbf{C}(z)]_{(j+1)j}$ and $\deg[\widetilde{\mathbf{C}}(z)]_{(j+1)j}$ are at most $q_0$, we must have $\deg[\mathbf{R}(z)]_{(j+1)j} = 0$. As the constant term of $\mathbf{R}(z)$ is $\mathbf{I}_{r_0}$, this then implies $[\mathbf{R}(z)]_{(j+1)j} = 0$ for all $z$, $j = 1, \ldots, r_0 - 1$ and so the first subdiagonal of $\mathbf{R}(z)$ is zero. Iterating this argument then shows that all subdiagonals of $\mathbf{R}(z)$ are zero, so $\mathbf{R}(z) = \mathbf{I}_{r_0}$. This then implies that $\mathbf{M}(z) = \widetilde{\mathbf{M}}(z)$ and $\mathbf{C}(z) = \widetilde{\mathbf{C}}(z)$.

A similar argument to the above shows that $\mathbf{B}_c = \widetilde{\mathbf{B}}_c$, $\mathbf{N}_c(z) = \widetilde{\mathbf{N}}_c(z)$, $\mathbf{D}_c(z) = \widetilde{\mathbf{D}}(z)$ and $\boldsymbol{\Upsilon}_c = \widetilde{\boldsymbol{\Upsilon}}_c$ for all $c = 1, \ldots, C$. Hence we can conclude $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$, and so $\boldsymbol{\theta} \mapsto \mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta})$ is injective on $\boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q}) \setminus N$.

□

## C.5 Proof of Proposition 13

*Proof.* If $\sigma = 0$, then $\mathbf{R}_{*,1} = \mathbf{0}$ and so $\mathbf{R}_{*,1} \preceq \mathbf{R}$ trivially as $\mathbf{R}$ is a Gram matrix. Then $\mathbf{R} - \mathbf{R}_{*,1} \succeq \mathbf{0}$ is equivalent to requiring

$$(\cos^2 \sigma)\mathbf{A}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{A} - \mathbf{A}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{B}(\mathbf{B}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{B})^{-1}\mathbf{B}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{A} \tag{C.6}$$

be positive semi-definite for $\sigma > 0$. This can be shown by taking the Schur complement of the southeast $r \times r$ block of $\mathbf{R} - \mathbf{R}_{*,1}$ and multiplying by $\cos\sigma$. The matrix $\mathbf{B}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{B} = \sum_{c=1}^{C} \mathbf{B}_c^{\mathsf{H}}\boldsymbol{\Phi}_c^{-1}\mathbf{B}_c$ is positive definite as $\mathbf{B}_c \in \mathbb{L}^1_{n_c \times r_c}$. To verify the above display, first note that $\boldsymbol{\Phi}^{-\frac{1}{2}}\mathbf{B}(\mathbf{B}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{B})^{-1}\mathbf{B}^{\mathsf{H}}\boldsymbol{\Phi}^{-\frac{1}{2}}$ is the projection onto $\operatorname{col}\boldsymbol{\Phi}^{-\frac{1}{2}}\mathbf{B}$ which is orthogonal in the Euclidean inner product. Hence the projection can be written as $\mathbf{Q}\mathbf{Q}^{\mathsf{H}}$ for unitary $\mathbf{Q}$, and thus $\mathbf{B}(\mathbf{B}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{B})^{-1}\mathbf{B}^{\mathsf{H}}$ equals $\mathbf{Y}\mathbf{Y}^{\mathsf{H}}$ with $\mathbf{Y} = \boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{Q}$ and $\mathbf{Y}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{Y} = \mathbf{I}_{r_c}$. Similarly, if $\mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{H}}$ is the compact SVD of $\boldsymbol{\Phi}^{-\frac{1}{2}}\mathbf{A}$, then $\mathbf{A} = \mathbf{W}\mathbf{D}\mathbf{V}^{\mathsf{H}}$ with $\mathbf{W} = \boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{U}$ and $\mathbf{W}^{\mathsf{H}}\boldsymbol{\Phi}^{-1}\mathbf{W} = \mathbf{I}_{r_0}$. With these

definitions, the matrix (C.6) equals

$$(\mathbf{V}\mathbf{D}^{\mathsf{H}})(\cos^2\sigma\mathbf{I}_n - \mathbf{W}^{\mathsf{H}}\mathbf{Y}\mathbf{Y}^{\mathsf{H}}\mathbf{W})(\mathbf{V}\mathbf{D}^{\mathsf{H}})^{\mathsf{H}}.$$

The inner term in the above expression is positive semi-definite, as the largest singular value $s_1$ of $\mathbf{W}^{\mathsf{H}}\mathbf{Y}$ satisfies

$$s_1 = \max_{\mathbf{a},\mathbf{b}} |\mathbf{a}^{\mathsf{H}}\mathbf{W}^{\mathsf{H}}\mathbf{Y}\mathbf{b}| \leq \cos(\sigma),$$

for $\mathbf{a} \in \mathbb{C}_{r_0}$ and $\mathbf{b} \in \mathbb{C}_n$ with $\mathbf{a}^{\mathsf{H}}\mathbf{a} = 1 = \mathbf{b}^{\mathsf{H}}\mathbf{b}$. The inequality follows from $\mathbf{W}\mathbf{a} \in \operatorname{col}\mathbf{A}$ and $\mathbf{Y}\mathbf{b} \in \operatorname{col}\mathbf{B}$ with $\mathbf{a}^{\mathsf{H}}\mathbf{W}^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{W}\mathbf{a} = 1 = \mathbf{b}^{\mathsf{H}}\mathbf{Y}^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{Y}\mathbf{b}$. Hence if $s_1 > \cos(\sigma)$, we could take $\mathbf{w} = \mathbf{W}\mathbf{a}$ and $\mathbf{y} = \mathbf{Y}\mathbf{b}$ in (3.22) and obtain a contradiction. Therefore, the largest eigenvalue of $\mathbf{W}^{\mathsf{H}}\mathbf{Y}\mathbf{Y}^{\mathsf{H}}\mathbf{W}$ is bounded above by $\sigma^2$, which in turn implies that (C.6) is positive semi-definite. So, $\mathbf{R}_{*,1} \preceq \mathbf{R}$.

To prove $\mathbf{R}_{*,2} \preceq \mathbf{R}$, let $\mathbf{S}^{(c)}$ for $c = 1, \ldots, C$ be the factor-patterned matrices whose blocks are

$$\begin{aligned}
\mathbf{S}^{(c)}_{00} &= \cos\rho_c \mathbf{A}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{A}_c, \\
\mathbf{S}^{(c)}_{0k} &= \mathbf{A}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{B}_c, \\
\mathbf{S}^{(c)}_{kk} &= \cos\rho_c \mathbf{B}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{B}_c,
\end{aligned}$$

for $k = 1, \ldots C$ and $\mathbf{S}^{(c)}_{kk'} = \mathbf{0}$ for $k \neq k'$. It can be seen that $\sum_{c=1}^{C} \mathbf{S}^{(c)} = \mathbf{R} - \mathbf{R}_{*,2}$.

The requirement that $\mathbf{S}^{(c)}$ be positive semi-definite is equivalent to requiring that

$$(\cos^2\rho_c)\mathbf{A}^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{A}_c - \mathbf{A}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{B}_c(\mathbf{B}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{B}_c)^{-1}\mathbf{B}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{A}_c$$

be positive semi-definite, which can be demonstrated by a similar argument to that above. As $\mathbf{S}^{(c)}$ is positive semi-definite for each $c$, we have $\mathbf{R} \succeq \mathbf{R}_{*,2}$. □

## C.6 Proof of Proposition 14

*Proof.* Let $\mathbf{U}_0\mathbf{D}_0\mathbf{V}_0^{\mathsf{H}}$ be the compact SVD of $\mathbf{A}$, and define $\mathbf{W}_0 = \mathbf{\Phi}^{\frac{1}{2}}\mathbf{U}_0$ which is $\mathbf{\Phi}$-orthogonal as $\mathbf{W}_0^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{W}_0 = \mathbf{I}_{r_0}$. Similarly, let $\mathbf{B}_c = \mathbf{W}_c\mathbf{D}_c\mathbf{V}_c^{\mathsf{H}}$ with $\mathbf{W}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{W}_c = \mathbf{I}_{r_c}$. Then define $\mathbf{q}_0(\omega) = \mathbf{W}_0^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{x}(\omega)$ and similarly $\mathbf{q}_c(\omega) = \mathbf{W}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{x}_c(\omega)$. Without loss of generality, assume

that $\mathbf{q}_0$ and all $\mathbf{q}_c$ are non-zero at $\omega$. This is not problematic, as $\mathbf{q}_0(\omega) = \mathbf{0}$ only if $\mathbf{A}^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{x}(\omega) = \mathbf{0}$, in which case $\mathbf{J}_{00}(\omega) = \mathbf{0}$ for all $k = 0, \ldots, C$. A similar result holds when $\mathbf{q}_c(\omega) = \mathbf{0}$ for any $c = 1, \ldots, C$. For $\omega$ where some subset of the $\mathbf{q}$ are zero, the associated blocks of $\mathbf{J}(\omega)$ and $\mathbf{J}_*(\omega)$ are also zero, while the following argument can be applied to the remaining non-zero $\mathbf{q}$ and associated submatrices.

If $\mathbf{q}(\omega) = [\mathbf{q}_0^{\mathsf{T}}(\omega) \ \ldots \ \mathbf{q}_C^{\mathsf{T}}(\omega)]^{\mathsf{T}}$ and $\mathbf{Y} = \mathrm{blkdiag}(\mathbf{V}_0\mathbf{D}_0^{\mathsf{H}}, \ldots, \mathbf{V}_C\mathbf{D}_C^{\mathsf{H}})$, then

$$\begin{bmatrix} \mathbf{A}^{\mathsf{H}} \\ \mathbf{B}^{\mathsf{H}} \end{bmatrix} \mathbf{\Phi}^{-1}\mathbf{I}(\omega)\mathbf{\Phi}^{-1} \begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix} = \mathbf{Y}\mathbf{q}(\omega)\mathbf{q}^{\mathsf{H}}(\omega)\mathbf{Y}^{\mathsf{H}}. \tag{C.7}$$

Recall that for projections $\mathbf{Q}_1$ and $\mathbf{Q}_2$ which are orthogonal in the Euclidean inner product, $\mathbf{Q}_1 \preceq \mathbf{Q}_2$ iff $\mathrm{col}\,\mathbf{Q}_1 \subset \mathrm{col}\,\mathbf{Q}_2$. Let $\mathbf{Q}(\omega) = ||\mathbf{q}(\omega)||^{-2}\mathbf{q}(\omega)\mathbf{q}^{\mathsf{H}}(\omega)$ be the rank-1 orthogonal projection associated with $\mathbf{q}(\omega)$, and similarly define $\mathbf{Q}_k(\omega) = ||\mathbf{q}_k(\omega)||^{-2}\mathbf{q}_k(\omega)\mathbf{q}_k^{\mathsf{H}}(\omega)$ for $k = 0, \ldots, C$. Then,

$$\mathbf{Q}(\omega) \preceq \mathrm{blkdiag}\left(\mathbf{Q}_0(\omega), \ldots, \mathbf{Q}_C(\omega)\right) \tag{C.8}$$

as the right hand side is the orthogonal projection onto

$$\mathrm{col}\begin{bmatrix} \mathbf{q}_0(\omega) & \ldots & \mathbf{0}_{r_0} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{r_C} & \ldots & \mathbf{q}_C(\omega) \end{bmatrix}$$

which has $\mathrm{col}\,\mathbf{Q}(\omega)$ as a subspace. Next, note that $\mathbf{W}_0\mathbf{W}_0^{\mathsf{H}}\mathbf{\Phi}^{-1}$ equals $\mathbf{P}_0$ as $\mathbf{W}_0\mathbf{W}_0^{\mathsf{H}}\mathbf{\Phi}^{-1}$ is idempotent, $\mathbf{W}_0$ is full-rank with the same range as $\mathbf{A}$, and $\mathbf{W}_0\mathbf{W}_0^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{v} = \mathbf{0}$ iff $\mathbf{v}$ is $\mathbf{\Phi}$-orthogonal to $\mathrm{col}\,\mathbf{A}$. So,

$$\begin{aligned} ||\mathbf{q}_0(\omega)||^2 &= \mathbf{x}^{\mathsf{H}}(\omega)\mathbf{\Phi}^{-1}\mathbf{W}_0(\mathbf{W}_0^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{W}_0)\mathbf{W}_0^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{x}(\omega) \\ &= (\mathbf{P}_0\mathbf{x}(\omega))^{\mathsf{H}}\mathbf{\Phi}^{-1}(\mathbf{P}_0\mathbf{x}(\omega)) = \mathrm{tr}\,\mathbf{\Phi}^{-1}\mathbf{P}_0\mathbf{I}(\omega)\mathbf{P}_0^{\mathsf{H}}, \end{aligned}$$

and similarly $||\mathbf{q}_c(\omega)||^2 = \mathrm{tr}\,\mathbf{\Phi}_c^{-1}\mathbf{P}_c\mathbf{I}_{cc}(\omega)\mathbf{P}_c^{\mathsf{H}}$. Therefore, $||\mathbf{q}(\omega)||^2/||\mathbf{q}_0(\omega)||^2 = m_0^{-1}(\omega)$ and similarly for $\mathbf{m}_c(\omega),\ c = 1, \ldots, C$. Multiplying both sides of (C.8) by $||\mathbf{q}(\omega)||^2$, which equals

$||\mathbf{q}_0(\omega)||^2 + \cdots + ||\mathbf{q}_C(\omega)||^2$, and plugging into (C.7), we obtain

$$\begin{bmatrix} \mathbf{A}^{\mathsf{H}} \\ \mathbf{B}^{\mathsf{H}} \end{bmatrix} \mathbf{\Phi}^{-1}\mathbf{I}(\omega)\mathbf{\Phi}^{-1} \begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix} \preceq \mathbf{K}_*(\omega)$$

where $\mathbf{K}_*$ is the factor-patterned matrix with diagonal blocks

$$\mathbf{K}_{*,00}(\omega) = m_0^{-1}(\omega)\mathbf{A}^{\mathsf{H}}\mathbf{\Phi}^{-1}\mathbf{I}(\omega)\mathbf{\Phi}^{-1}\mathbf{A},$$
$$\mathbf{K}_{*,cc}(\omega) = m_c^{-1}(\omega)\mathbf{B}_c^{\mathsf{H}}\mathbf{\Phi}_c^{-1}\mathbf{I}_{cc}(\omega)\mathbf{\Phi}_c^{-1}\mathbf{B}_c,\ c = 1, \ldots, C$$

and zero off-diagonal blocks. If $\mathbf{L}(\omega) = \mathrm{blkdiag}(\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega}), \mathbf{N}_1(\mathrm{e}^{\mathrm{i}\omega}), \ldots, \mathbf{N}_C(\mathrm{e}^{\mathrm{i}\omega}))$, then left and right-multiplying the above inequality by $\mathbf{L}(\omega)$ and $\mathbf{L}^{\mathsf{H}}(\omega)$ respectively yields

$$\mathbf{J}(\omega) \preceq \mathbf{J}_*(\omega),$$

as desired.

□

## C.7 Proofs of Proposition 15 and Lemma 1

### Proof of Lemma 1

*Proof.* Let $\mathbf{M}$ be the matrix in (3.28) and let $\mathbf{y}_1, \ldots, \mathbf{y}_{m-1} \in \mathbb{C}_m$ be a set of orthonormal vectors with $\mathbf{y}_j \perp \mathbf{x}$. Then $\mathbf{M}\mathbf{y}_j = a\mathbf{y}_j$ for each $\mathbf{y}_j$, and so $\mathbf{M}$ is positive definite only if $a > 0$. The vector $\mathbf{x}$ is an eigenvector of $\mathbf{M}$ with eigenvalue

$$\lambda = a - (b\mathbf{x}^{\mathsf{H}}\mathbf{x})/(b + \mathbf{x}^{\mathsf{H}}\mathbf{x})$$

This divides into 3 cases when $a > 0$. If $-\mathbf{x}^{\mathsf{H}}\mathbf{x} < b < 0$, then the quantity subtracted is negative and so $\lambda > 0$. The second term in (3.29) is greater than one iff $0 > b > -\mathbf{x}^{\mathsf{H}}\mathbf{x}$, which accounts for this case. If $b \geq 0$ or $b < -\mathbf{x}^{\mathsf{H}}\mathbf{x}$ then $\lambda > 0$ is equivalent to $a/\mathbf{x}^{\mathsf{H}}\mathbf{x} + a/b > 1$, which can be seen by multiplying $\lambda > 0$ by the positive quantity $(b + \mathbf{x}^{\mathsf{H}}\mathbf{x})/(b\mathbf{x}^{\mathsf{H}}\mathbf{x})$. Hence when (3.29) is satisfied, $(\mathbf{y}_1, \ldots, \mathbf{y}_{n-1}, \mathbf{x})$ is an eigenbasis for $\mathbf{M}$ whose corresponding eigenvalues are positive,

so $\mathbf{M}$ is positive definite. Conversely, if $a \leq 0$ or (3.28) is not satisfied, then $\mathbf{M}$ has a non-positive eigenvalue. The expression for the inverse can be verified by direct calculation. □

## Proof of Proposition 15

*Proof.* If $\mathbf{J}(\omega) = \mathbf{0}$, then the conclusion reduces to $\mathbf{T}_*(\omega) \succeq \mathbf{T}(\omega)$, which has previously been shown. For $\mathbf{J}(\omega) \neq \mathbf{0}$, instead of finding an upper bound for $\mathbf{G}(\omega)$ (which is positive definite for all $\omega$ as $\mathbf{T}(\omega)$ is positive definite), we will find a lower bound for $\mathbf{G}^{-1}(\omega) = \mathbf{T}^{-1}(\omega)\big(\mathbf{T}^{-1}(\omega) + \mathbf{J}(\omega)\big)^{-1}\mathbf{T}^{-1}(\omega)$. To do so, let $\mathbf{W}(\omega)$ and $\mathbf{K}(\omega)$ be

$$\mathbf{W}(\omega) = \mathbf{T}_*^{\frac{1}{2}}(\omega)\mathbf{T}^{-1}(\omega)\mathbf{T}_*^{\frac{1}{2}}(\omega),$$
$$\mathbf{K}(\omega) = \mathbf{T}_*^{\frac{1}{2}}(\omega)\mathbf{J}(\omega)\mathbf{T}_*^{\frac{1}{2}}(\omega),$$

and note that $\mathbf{W}(\omega) \succeq \mathbf{I}$ for all $\omega$ as $\mathbf{T}_*(\omega) \succeq \mathbf{T}(\omega)$, while $\mathbf{K}(\omega)$ is rank-1 positive semi-definite. The inverse then is,

$$\mathbf{G}^{-1}(\omega) = \mathbf{T}_*^{-\frac{1}{2}}(\omega)\mathbf{W}(\omega)\big(\mathbf{W}(\omega) + \mathbf{K}(\omega)\big)^{-1}\mathbf{W}(\omega)\mathbf{T}_*^{-\frac{1}{2}}(\omega).$$

Next, note that

$$\mathbf{W}(\omega) + \mathbf{K}(\omega) \preceq m(\omega)\big(\mathbf{W}(\omega) + (\operatorname{tr}\mathbf{K}(\omega))^{-1}\mathbf{K}(\omega)\big)$$

as either $\operatorname{tr}\mathbf{K}(\omega) \leq 1$, so $m(\omega) = 1$ and $\big(\operatorname{tr}\mathbf{K}(\omega)\big)^{-1}\mathbf{K}(\omega) \succeq \mathbf{K}(\omega)$ or $\operatorname{tr}\mathbf{K}(\omega) > 1$, in which case $m(\omega) = \operatorname{tr}\mathbf{K}(\omega)$ and $\big(\operatorname{tr}\mathbf{K}(\omega)\big)\mathbf{W}(\omega) \succeq \mathbf{W}(\omega)$. As $\mathbf{J}(\omega) \neq 0$ and $\mathbf{T}_*(\omega)$ is positive definite, we have $\operatorname{tr}\mathbf{K}(\omega) > 0$. Using this bound and applying the Sherman-Morrison formula yields

$$\begin{aligned} &m^{-1}(\omega)\mathbf{W}(\omega)\big(\mathbf{W}(\omega) + \big(\operatorname{tr}\mathbf{K}(\omega)\big)^{-1}\mathbf{K}(\omega)\big)^{-1}\mathbf{W}(\omega) = \\ &m^{-1}(\omega)\big(\mathbf{W}(\omega) - (\operatorname{tr}\mathbf{K}(\omega) + \operatorname{tr}\mathbf{W}^{-1}(\omega)\mathbf{K}(\omega))^{-1}\mathbf{K}(\omega)\big) \\ &\quad\preceq \mathbf{W}(\omega)\big(\mathbf{W}(\omega) + \mathbf{K}(\omega)\big)^{-1}\mathbf{W}(\omega). \end{aligned} \tag{C.9}$$

Next, note that $\lambda(\omega) > 0$ as $\mathbf{T}(\omega)$ and $\mathbf{T}_*(\omega)$ are positive definite, hence $\lambda^{-1}(\omega) = \lambda_{\min}\big(\mathbf{W}(\omega)\big)$ exists. Letting

$$\mathbf{V}(\omega) \equiv \lambda^{-1}(\omega)\mathbf{I} - \big(\operatorname{tr}\mathbf{K}(\omega) + \operatorname{tr}\mathbf{W}^{-1}(\omega)\mathbf{K}(\omega)\big)^{-1}\mathbf{K}(\omega), \tag{C.10}$$

we have

$$\mathbf{V}(\omega) \preceq \mathbf{W}(\omega) - \big(\operatorname{tr}\mathbf{K}(\omega) + \operatorname{tr}\mathbf{W}^{-1}(\omega)\mathbf{K}(\omega)\big)^{-1}\mathbf{K}(\omega),$$

as $\lambda_{\min}\big(\mathbf{W}(\omega)\big)\mathbf{I} \preceq \mathbf{W}(\omega)$. If $\operatorname{tr}\mathbf{K}(\omega) + \operatorname{tr}\mathbf{W}^{-1}(\omega)\mathbf{K}(\omega) \neq 1$, we can apply Lemma 1 to $\mathbf{V}(\omega)$ with $a \equiv \lambda^{-1}(\omega)$, $\mathbf{x}\mathbf{x}^{\mathsf{H}} \equiv \mathbf{K}(\omega)$ and

$$b \equiv \frac{\operatorname{tr}\mathbf{K}(\omega)}{\operatorname{tr}\mathbf{K}(\omega) + \operatorname{tr}\mathbf{W}^{-1}(\omega)\mathbf{K}(\omega) - 1}.$$

Then $b \neq -\operatorname{tr}\mathbf{K}(\omega)$ as $\operatorname{tr}\mathbf{K}(\omega) \geq \operatorname{tr}\mathbf{W}^{-1}(\omega)\mathbf{K}(\omega) > 0$. With these values, the first term in the criterion (3.28) for Lemma 1 has

$$\lambda^{-1}(\omega)\left(1 + \frac{\operatorname{tr}\mathbf{W}^{-1}(\omega)\mathbf{K}(\omega)}{\operatorname{tr}\mathbf{K}(\omega)}\right) > 1.$$

as $\lambda^{-1}(\omega) \geq 1$ by $\mathbf{W}(\omega) \succeq \mathbf{I}$ and $\operatorname{tr}\mathbf{W}^{-1}(\omega)\mathbf{K}(\omega) > 0$ by $\mathbf{W}(\omega)$ positive definite. Hence $\mathbf{V}(\omega)$ is positive definite with inverse

$$\mathbf{V}^{-1}(\omega) = \lambda(\omega)\left(\mathbf{I} + h^{-1}(\omega)\mathbf{K}(\omega)\right) \tag{C.11}$$

as $\operatorname{tr}\mathbf{K} = \operatorname{tr}\mathbf{T}_*(\omega)\mathbf{J}(\omega)$ and $\operatorname{tr}\mathbf{W}^{-\mathbf{1}}(\omega)\mathbf{K}(\omega) = \operatorname{tr}\mathbf{T}(\omega)\mathbf{J}(\omega)$ implies

$$\begin{aligned}\frac{b}{ab + (b-a)\mathbf{x}^{\mathsf{H}}\mathbf{x}} &= \frac{1}{(\lambda^{-1}(\omega)-1)\operatorname{tr}\mathbf{K}(\omega) + \lambda^{-1}(\omega)\operatorname{tr}\mathbf{W}^{-\mathbf{1}}\mathbf{K}(\omega)} \\ &= h^{-1}(\omega).\end{aligned}$$

In the case where $\operatorname{tr}\mathbf{K}(\omega) + \operatorname{tr}\mathbf{W}^{-1}\mathbf{K}(\omega) = 1$, $\mathbf{V}(\omega)$ is still a positive definite lower bound with inverse given by (C.11). As $\operatorname{tr}\mathbf{K} > 0$ and $\operatorname{tr}\mathbf{W}^{-1}(\omega)\mathbf{K}(\omega) > 0$, this case has $\operatorname{tr}\mathbf{K}(\omega) < 1$. Then as $\lambda^{-1}(\omega) \geq 1$, constructing an eigenbasis for $\lambda^{-1}(\omega)\mathbf{I} - \mathbf{K}(\omega)$ shows

$$\mathbf{W}(\omega) - \mathbf{K}(\omega) \succeq \mathbf{V}(\omega) = \lambda^{-1}(\omega)\mathbf{I} - \mathbf{K}(\omega) \succ \mathbf{0}.$$

Letting $\mathbf{x}\mathbf{x}^{\mathsf{H}} = \mathbf{K}(\omega)$, applying the Sherman-Morrison formula to obtain the inverse of$\lambda^{-1}(\omega) + (-\mathbf{x})\mathbf{x}^{\mathsf{H}}$ shows that $\mathbf{V}^{-1}(\omega)$ is given by (C.11) as well, as $h(\omega) = \lambda^{-1}(\omega) - \operatorname{tr}\mathbf{K}(\omega)$ in this case.

Combining (C.10) with (C.9) in the expression for $\mathbf{G}^{-1}(\omega)$ yields

$$m^{-1}(\omega)\mathbf{T}_*^{-\frac{1}{2}}\mathbf{V}(\omega)\mathbf{T}_*^{-\frac{1}{2}} \preceq \mathbf{G}^{-1}(\omega)$$

where both terms are positive definite. Inverting and using (C.11) shows

$$m(\omega)\lambda(\omega)\big(\mathbf{T}_*(\omega) + h^{-1}(\omega)\mathbf{T}_*(\omega)\mathbf{J}(\omega)\mathbf{T}_*(\omega)\big) \succeq \mathbf{G}(\omega).$$

Replacing $\mathbf{J}(\omega)$ with its block-diagonal upper bound $\mathbf{J}_*(\omega)$ in the left hand side gives $\mathbf{G}_*(\omega) \succeq \mathbf{G}(\omega)$, as desired. □

## C.8 Proof of Theorem 5

*Proof.* As the upper bounds on the magnitudes of the roots of the main diagonal polynomials of $\mathbf{C}(z)$ and $\mathbf{D}_c(z),\ c = 1, \ldots, C$ as well as on the south-east polynomials of $\mathbf{M}(z)$ and $\mathbf{N}_c,\ c = 1, \ldots, C$ do not lead to convex sets in the space of coefficients, we can reparameterize $\mathbb{L}^{\kappa,B}_{m\times m\times p}$ as $\mathbb{L}^{0,B}_{m\times m\times p} \times \mathbb{D}^{mp}_{\leq\kappa}$ where $\mathbb{L}^{0,B}_{m\times m\times p}$ is the subset of off-diagonal coefficient matrices $\mathbf{L} \in \mathbb{L}^{0}_{m\times m\times p}$ satisfying $||\mathbf{L}||_{2,\infty} \leq B$ and $\mathbb{D}_{\leq\kappa}$ is the set of $z \in \mathbb{C}$ with $|z| \leq \kappa$. As both these subsets are compact and convex, so too is their product. Similarly, $\mathbb{L}^{SE\leq\kappa}_{m\times m\times p}$ can be reparameterized as $\mathbb{L}^{SE=1}_{m\times m\times p} \times \mathbb{D}^{p}_{\leq\kappa}$, which is then convex (but $\mathbb{L}^{SE=1}_{m\times m\times p}$ is closed and unbounded). Further, it will be convenient to parameterize by the precision matrices $\mathbf{\Phi}^{-1}$, $\mathbf{\Psi}^{-1}$ and $\mathbf{\Upsilon}_c^{-1},\ c = 1, \ldots, C$ rather than the covariances. The space of corresponding precision matrices is then $\mathbb{P}^{\leq\epsilon^{-1}}_{m}$ contains all $m \times m$ diagonal positive definite matrices with all diagonal entries bounded above by $\epsilon^{-1}$.

Substituting these new sets into the definition of the parameter space $\mathbf{\Theta}^*(\mathbf{r}, \mathbf{p}, \mathbf{q})$ given in Section 3.2.3 defines the space $\mathbf{\Theta}^{**}(\mathbf{r}, \mathbf{p}, \mathbf{q})$, which as a Cartesian product of closed and convex sets is itself closed and convex. As there is a smooth bijection between the coefficients of a polynomial with constant term 1 and the roots of the polynomial as well as a smooth bijection between a positive definite matrix and its inverse, there is a smooth bijection between the parameter spaces $\mathbf{\Theta}^*(\mathbf{r}, \mathbf{p}, \mathbf{q})$ and $\mathbf{\Theta}^{**}(\mathbf{r}, \mathbf{p}, \mathbf{q})$. Accordingly, we will prove convergence of the sequence of iterates $\boldsymbol{\theta}^{(s)\prime} \in \mathbf{\Theta}^{**}(\mathbf{r}, \mathbf{p}, \mathbf{q})$ corresponding to $\boldsymbol{\theta}^{(s)} \in \mathbf{\Theta}^{*}(\mathbf{r}, \mathbf{p}, \mathbf{q})$. In a small abuse of notation, $\ell(\boldsymbol{\theta}')$ and $g(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}}')$ will denote the objective and surrogate function applied after application of the smooth bijection taking $\boldsymbol{\theta}' \mapsto \boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}' \mapsto \widetilde{\boldsymbol{\theta}}$.

First, it must be shown that the sublevel sets $\{\boldsymbol{\theta}' \in \boldsymbol{\Theta}^{**} \mid \ell(\boldsymbol{\theta}') \leq \alpha\}$ are compact for all $\alpha \in \mathbb{R}$. To do so, note that

$$\ell(\boldsymbol{\theta}') \geq \frac{1}{T}\sum_{k=0}^{T-1} \log\det \mathbf{S}_{\mathbf{xx}}(\omega_k;\boldsymbol{\theta}') + \epsilon \operatorname{tr} \mathbf{I}(\omega).$$

as $\boldsymbol{\Phi} \succeq \epsilon \mathbf{I}_n$. Then as $\mathbf{S}_{\mathbf{xx}}(\omega;\boldsymbol{\theta}') \succeq \boldsymbol{\Phi}$ for all $\omega$, we have $\log\det \mathbf{S}_{\mathbf{xx}}(\omega;\boldsymbol{\theta}') \geq \log\det \boldsymbol{\Phi} \geq \log(\epsilon^{n-1}||\boldsymbol{\Phi}||_2)$. So, if any entry in $\boldsymbol{\Phi}$ diverges (and so the corresponding entry in $\boldsymbol{\Phi}^{-1}$ goes to zero), then $\ell(\boldsymbol{\theta}')$ goes to $+\infty$ regardless of the other parameter values.

Next, for $j = 1, \ldots r_0$, the diagonal entry of $\mathbf{S}_{\mathbf{xx}}(\omega;\boldsymbol{\theta})$ is bounded below by

$$[\mathbf{S}_{\mathbf{xx}}(\omega;\boldsymbol{\theta}')]_{jj} \geq [\mathbf{S}_{\mathbf{ss}}(\omega;\boldsymbol{\theta}')]_{jj} \geq \frac{|[\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega})]_{jj}|^2}{|[\mathbf{C}(\mathrm{e}^{\mathrm{i}\omega})]_{jj}|^2}[\boldsymbol{\Psi}]_{jj},$$

as $[\mathbf{S}_{\mathbf{ss}}(\omega;\boldsymbol{\theta})]_{jj}$ is the squared norm of the $j$th row of $\mathbf{A}\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{C}^{-1}(\mathrm{e}^{\mathrm{i}\omega})\boldsymbol{\Psi}^{\frac{1}{2}}$. The denominator is bounded above as $|[\mathbf{C}(\mathrm{e}^{\mathrm{i}\omega})]_{jj}|^2 \leq (1+\kappa)^{2q_0}$, while by the discussion around (3.39) shows that the numerator satisfies

$$1 = T^{-1}\textstyle\sum_{k=0}^{T-1}[\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega_k})]_{jj} \leq T^{-1}\sum_{k=0}^{T-1}|[\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega_k})]_{jj}|,$$

for all $j = 1, \ldots, r_0$. Hence for all $j$, there is some $k_0$ such that $|[\mathbf{M}(\mathrm{e}^{\mathrm{i}\omega_{k_0}})]_{jj}|^2 \geq 1$ and we obtain

$$||\mathbf{S}_{\mathbf{xx}}(\omega_{k_0};\boldsymbol{\theta}')||_2 \geq [\mathbf{S}_{\mathbf{xx}}(\omega_{k_0})]_{jj} \geq \frac{1}{(1+\kappa)^{2q_0}}[\boldsymbol{\Psi}]_{jj}.$$

As $\mathbf{S}_{\mathbf{xx}}(\omega;\boldsymbol{\theta}') \succeq \epsilon\mathbf{I}_n$, we similarly have $\log\det \mathbf{S}_{\mathbf{xx}}(\omega_{k_0};\boldsymbol{\theta}') \geq \log(\epsilon^{n-1}||\mathbf{S}_{\mathbf{xx}}(\omega_{k_0};\boldsymbol{\theta}')||_2)$. So, if any entry of $\boldsymbol{\Psi}$ diverges, then $\ell(\boldsymbol{\theta}')$ goes to $+\infty$ regardless of the other parameter values. A similar argument applied to the $(\sum_{d=1}^{c-1} n_d + j)$th diagonal entry of $\mathbf{S}_{\mathbf{xx}}$ for $c = 1, \ldots, C$ and $1 \leq j \leq r_c$ shows that the same conclusion holds for $\boldsymbol{\Upsilon}_c,\ c = 1, \ldots, C$. Hence all entries of $\boldsymbol{\Phi}, \boldsymbol{\Psi}$ and $\boldsymbol{\Upsilon}_c,\ c = 1, \ldots, C$ are bounded above by some $K_1 < \infty$ if $\ell(\boldsymbol{\theta}') \leq \alpha$.

Next, note that

$$\begin{aligned}
||\mathbf{D}(\mathrm{e}^{\mathrm{i}\omega})\boldsymbol{\Psi}^{-1}\mathbf{D}^{\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega})||_2 &\leq \epsilon^{-1}||\mathbf{D}(\mathrm{e}^{\mathrm{i}\omega})||_2^2 \\
&\leq \epsilon^{-1}\sum_{j=1}^{r_0}|[\mathbf{D}(\mathrm{e}^{\mathrm{i}\omega})]_{jj}|^2 + \epsilon^{-1}||\mathbf{D}'(\mathrm{e}^{\mathrm{i}\omega})||_F^2 \\
&\leq \epsilon^{-1}(1+\kappa)^{2r_0q_0} + \epsilon^{-1}r_0B^2,
\end{aligned}$$

and so $\mathbf{E}^{-1}(\omega) = \mathbf{D}^{-1}(\mathrm{e}^{\mathrm{i}\omega})\mathbf{\Psi}\mathbf{D}^{-\mathsf{H}}(\mathrm{e}^{\mathrm{i}\omega}) \succeq K_2\mathbf{I}_{r_0}$ for $K_2 = (\epsilon^{-1}(1+\kappa)^{2r_0p_0} + \epsilon^{-1}r_0B^2)^{-1} > 0$. Let $\widetilde{\mathbf{U}}(\omega) = [1, \mathrm{e}^{\mathrm{i}\omega}, \ldots, \mathrm{e}^{\mathrm{i}\omega p_0}]^{\mathsf{T}} \otimes \mathbf{I}_{r_0}$ and $\widetilde{\mathbf{M}} = [\mathbf{I}_{r_0}\ \mathbf{M}]$, and so $\mathbf{M}(z) = \widetilde{\mathbf{M}}\mathbf{U}(z)$. Define matrices $\mathbf{S}$, $\mathbf{E}$ and $\mathbf{U}$ as

$$\mathbf{S} = \mathrm{blkdiag}\left(\mathbf{S}(\omega_0;\boldsymbol{\theta}'), \ldots, \mathbf{S}(\omega_{T-1};\boldsymbol{\theta}')\right),$$

$$\mathbf{E} = \mathrm{blkdiag}\left(\mathbf{E}(\omega_0), \ldots, \mathbf{E}(\omega_{T-1})\right),$$

$$\widetilde{\mathbf{U}} = \mathrm{blkdiag}\left(\widetilde{\mathbf{U}}(\omega_0), \ldots, \widetilde{\mathbf{U}}(\omega_{T-1})\right).$$

The $nT \times nT$ matrix $\mathbf{S}$ then has

$$\mathbf{S} = (\mathbf{I}_T \otimes \mathbf{A}\widetilde{\mathbf{M}})\mathbf{U}\mathbf{E}^{-1}\mathbf{U}^{\mathsf{H}}(\mathbf{I}_T \otimes \mathbf{A}\widetilde{\mathbf{M}})^{\mathsf{H}} \succeq K_2(\mathbf{I}_T \otimes \mathbf{A}\widetilde{\mathbf{M}})\mathbf{U}\mathbf{U}^{\mathsf{H}}(\mathbf{I}_T \otimes \mathbf{A}\widetilde{\mathbf{M}})^{\mathsf{H}},$$

and by taking the trace we obtain

$$nT||\mathbf{S}||_2 \geq \mathrm{tr}\,\mathbf{S} \geq K_2\,\mathrm{tr}\,\mathbf{A}\widetilde{\mathbf{M}}\big(\textstyle\sum_{k=0}^{T-1}\widetilde{\mathbf{U}}(\omega_k)\widetilde{\mathbf{U}}^{\mathsf{H}}(\omega_k)\big)\widetilde{\mathbf{M}}^{\mathsf{H}}\mathbf{A}^{\mathsf{H}}.$$

However, $p_0 < T$ implies

$$\textstyle\sum_{k=0}^{T-1}\widetilde{\mathbf{U}}(\omega_k)\widetilde{\mathbf{U}}^{\mathsf{H}}(\omega_k) = T\mathbf{I}_{(p_0+1)r_0},$$

as all off-diagonal entries sum to zero. Therefore,

$$\tfrac{n}{K_2}||\mathbf{S}||_2 \geq ||\mathbf{A}\widetilde{\mathbf{M}}||_F^2 = ||\mathbf{A}||_F^2 + \textstyle\sum_{j=1}^{p_0}||\mathbf{A}\mathbf{M}_j||_F^2, \tag{C.12}$$

and so if any entry of $\mathbf{A}$ diverges, so too does $||\mathbf{S}||_2$. As

$$\textstyle\sum_{k=0}^{T-1}\log\det\mathbf{S}_{\mathbf{xx}}(\omega_k;\boldsymbol{\theta}') = \log\det\mathbf{S} \geq \log(\epsilon^{nT-1}||\mathbf{S}||_2),$$

this then implies that $\ell(\boldsymbol{\theta}')$ goes to positive infinity if any entry of $\mathbf{A}$ diverges, regardless of the other parameter values. Hence $\ell(\boldsymbol{\theta}') \leq \alpha$ implies $||\mathbf{A}||_F \leq K_3$ for some $K_3$, and so the largest singular value of $\mathbf{A}$ is also $\leq K_3$. If $\mathbf{A}_1$ is the unit lower-triangular matrix containing the top $r_0$ rows of $\mathbf{A}$, then both $||\mathbf{A}_1||_2$ bounded above by $K_3$ and

$$\lambda_{\min}(\mathbf{A}_1\mathbf{A}_1^{\mathsf{H}}) \leq \sigma_{\min}(\mathbf{A})$$

follow by the Cauchy interlacing theorem. As $\det(\mathbf{A}_1\mathbf{A}_1^{\mathsf{H}}) = 1$ by $\mathbf{A}_1$ unit lower triangular, then $\lambda_{\min}(\mathbf{A}_1\mathbf{A}_1^{\mathsf{H}}) \geq K_3^{-(r_0-1)} > 0$ when $\ell(\boldsymbol{\theta}') \leq \alpha$. But then (C.12) implies

$$\tfrac{n}{K_2}||\mathbf{S}||_2 \geq K_3^{-r_0-1}||\mathbf{M}||_F^2$$

and so if any entry of $\mathbf{M}$ diverges, $\ell(\boldsymbol{\theta}')$ will go to $+\infty$. A similar argument to the above can be applied to $\mathbf{B}_c, \mathbf{N}_c$ and $\mathbf{D}_c$ to show that any entry of those matrices diverging implies $\ell(\boldsymbol{\theta}')$ diverges. Therefore, $\ell(\boldsymbol{\theta}') \leq \alpha$ implies that all the non-diagonal parameters live in a bounded subset of their respective spaces, each of which is closed. Also, $\ell(\boldsymbol{\theta}') \leq \alpha$ implies the diagonal entries of $\boldsymbol{\Phi}^{-1}, \boldsymbol{\Psi}^{-1}$ and $\boldsymbol{\Upsilon}_c^{-1},\ c = 1, \ldots, C$ live in a closed interval $[K_1^{-1}, \epsilon^{-1}]$. As $\alpha$ was arbitrary, combining these results implies that the sublevel sets of $\ell(\boldsymbol{\theta}')$ are compact.

Next, we verify that the objective function $\ell(\boldsymbol{\theta}')$ and its surrogate $g(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}}')$ satisfy (A1)-(A4) of Razaviyayn et al. (2013, Assumption 1). Assumptions (A1) and (A2) are satisfied as $g$ is a surrogate for $\ell$ and hence by its construction in Section 3.3.5 we have $g(\boldsymbol{\theta}'|\boldsymbol{\theta}') = \ell(\boldsymbol{\theta}')$ and $g(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}}') \leq \ell(\boldsymbol{\theta}')$. Assumption (A4) requires that $g(\boldsymbol{\theta}'|\widetilde{\boldsymbol{\theta}}')$ be continuous in $(\boldsymbol{\theta}', \widetilde{\boldsymbol{\theta}}')$, which is equivalent to continuity in $(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})$ of $g(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}})$ as the map $\boldsymbol{\theta}' \mapsto \boldsymbol{\theta}$ is smooth. For fixed $\widetilde{\boldsymbol{\theta}}$ (and therefore fixed $\widetilde{\mathbf{G}}(\omega), \widetilde{\mathbf{G}}_*(\omega)$ and $\widetilde{\mathbf{K}}(\omega)$) the continuity of $g(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}})$ can be immediately verified by inspecting the definitions (3.31), (3.32) and (3.33). Recalling that the eigenvalues and singular values of a matrix are Lipschitz continuous functions of its entries, the matrices $\widetilde{\mathbf{G}}(\omega_k)$, $\widetilde{\mathbf{G}}_*(\omega_k)$ and $\widetilde{\mathbf{K}}(\omega_k)$ for $k = 0, \ldots, T-1$ are also continuous functions of $\widetilde{\omega}$. Hence (A4) is satisfied. Finally, the definition of $\boldsymbol{\Theta}^*(\mathbf{r}, \mathbf{p}, \mathbf{q})$ depends on chosen values $\epsilon > 0$, $0 < \kappa < 1$, and $0 < B < \infty$ to ensure that the spectral density $\mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta})$ is well-behaved for $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*(\mathbf{r}, \mathbf{p}, \mathbf{q})$. Relaxing these values to e.g. $\epsilon/2$, $(1-\kappa)/2 + \kappa/2$ and $2B$ and considering the interior of the resulting parameter space provides an open set containing $\boldsymbol{\Theta}^*(\mathbf{r}, \mathbf{p}, \mathbf{q})$ on which $g(\theta, \widetilde{\boldsymbol{\theta}})$ continues to satisfy (A1) and (A2). As both $\ell(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}})$ are differentiable in $\boldsymbol{\theta}$ in this open set, the result of Razaviyayn et al. (2013, Proposition 1) implies that (A3) is satisfied.

As Razaviyayn et al. (2013, Assumption 1) is satisfied, Razaviyayn et al. (2013, Theorem 1) can be applied to show that every limit point of the sequence of iterates $\boldsymbol{\theta}^{(s)}$ is a stationary point of

$\ell(\boldsymbol{\theta})$. As stated, Razaviyayn et al. (2013, Theorem 1) applies to iterate sequences obtained where $\boldsymbol{\theta}^{(s+1)}$ is the global minimum of $g(\theta|\boldsymbol{\theta}^{(s+1)})$. However, inspection of the proof shows that only the property that $g(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\theta}^{(s)}) \leq g(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s)})$ is needed, which is satisfied for the sequence of iterates obtained from Algorithm 1 as each step is obtained by minimizing $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ over some block of parameters in $\boldsymbol{\theta}$ while holding the others fixed, and each step also has a unique minimum under the assumptions of Theorem 5. Combining this result with the fact that the sublevel sets are compact, we obtain that the sequence of iterates $\boldsymbol{\theta}^{(s)}$ converges in norm to the set of stationary points of $\ell(\boldsymbol{\theta})$.

The above result does not preclude the possibility of cycling within the set of stationary points of $\ell(\boldsymbol{\theta})$. However, the properties of the optimization subproblems as well as the generic uniqueness of Proposition 12 will ensure convergence to a single point almost surely. To show this, we first show that each update in Algorithm 1 is obtained by minimizing a strictly convex function, and as the change in magnitude of the profile surrogate function decreases to zero by the previous results, we must also have that the magnitude of the change in the iterates $||\boldsymbol{\theta}^{(s+1)} - \boldsymbol{\theta}^{(s)}||$ goes to zero as well. Note that the optimization subproblems solved in the steps of Algorithm 1 are all of two forms. First, for the positive-constrained parameter blocks $\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Upsilon}$, the optimization problems 3.34 and 3.37 are equivalent to minimizing the strictly convex function $-\log y + ay$ on $(0, \epsilon^{-1})$. As the minima are unique and the change in magnitude of the function values $|\ell(\boldsymbol{\theta}^{(s+1)}) - \ell(\boldsymbol{\theta}^{(s)})|$ goes to zero by the previous results, the properties of strictly convex functions imply that $||\boldsymbol{\Phi}^{(s+1)} - \Phi^{(s)}||^2 \to 0$ and similarly for $\boldsymbol{\Psi}$ and $\boldsymbol{\Upsilon}$. Second, the updates for the other parameter blocks are each obtained by minimizing a quadratic form (see profile optimization problems (3.38), (3.41), (3.43), (3.48), (3.53), (3.55), (3.57), (3.61), and (3.63)). Under the assumptions of Theorem 5, these quadratic forms are positive definite with probability one for each $s$ and so are strictly convex. By the same argument as above, monotonic convergence of the objective values implies that $||\mathbf{A}^{(s+1)} - \mathbf{A}^{(s)}|| \to 0$ and similarly for the remaining parameter blocks. Combined, we obtain that the total magnitude of the change in the iterates goes to zero, $||\boldsymbol{\theta}^{(s+1)} - \boldsymbol{\theta}^{(s)}|| \to 0$. Last, if the set of stationary points of $\ell(\boldsymbol{\theta})$ is discrete, then convergence of the sequence of iterates to the set of stationary points and the magnitude of the step-wise change going to zero suffices to show that the sequence of iterates

converges to a single point which is a stationary point of $\ell(\boldsymbol{\theta})$ (Lange, 2013). By the assumed generic global identifiability, the result of Proposition 12, and the assumed absolute continuity of the data distribution, we have that all elements in the sequence of iterates almost surely lie within the unique-representative subset $\boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q}) \setminus N$ on which the map $\boldsymbol{\theta} \mapsto \mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta})$ is injective, and so there does not exist a connected non-isolated set of $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q}) \setminus N$ mapping to the same $\mathbf{S}_{\mathbf{xx}}(\omega; \boldsymbol{\theta})$. Therefore, the set of stationary points of $\ell(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}(\mathbf{r}, \mathbf{p}, \mathbf{q}) \setminus N$ is discrete, and the sequence of iterates converges to a single stationary point, as desired. □

# Appendix D

# Supplementary Materials for Chapter 4

## D.1 Properties of Joint Multivariate Gaussian Processes

### Positive semi-definite matrix kernel

Recall that a symmetric function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a PSD kernel on the set $\mathcal{X}$ if for any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ and $w_1, \ldots w_n \in \mathbb{R}$,

$$\sum_{i=1} \sum_{j=1} w_i w_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0. \tag{D.1}$$

For a PSD matrix kernel, this generalizes if $\boldsymbol{\kappa} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}$ is a function such that for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have

$$\boldsymbol{\kappa}(\mathbf{x}_1, \mathbf{x}_2) = \boldsymbol{\kappa}^{\mathsf{T}}(\mathbf{x}_2, \mathbf{x}_1) \tag{D.2}$$

and for any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ and $\mathbf{w}_1, \ldots \mathbf{w}_n \in \mathbb{R}^m$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{w}_i^{\mathsf{T}} \boldsymbol{\kappa}(\mathbf{x}_i, \mathbf{x}_j) \mathbf{w}_j \geq 0. \tag{D.3}$$

Then $\boldsymbol{\kappa}(\cdot, \cdot)$ is a PSD matrix kernel of dimension $m$.

In matrix form, the second requirement is that

$$\begin{bmatrix} \mathbf{w}_1^{\mathsf{T}} & \ldots & \mathbf{w}_n^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\kappa}(\mathbf{x}_1, \mathbf{x}_1) & \ldots & \boldsymbol{\kappa}(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\kappa}(\mathbf{x}_n, \mathbf{x}_1) & \ldots & \boldsymbol{\kappa}(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_n \end{bmatrix} \geq 0 \tag{D.4}$$

and so the PSD kernels can also be characterized as giving rise to symmetric, positive-semi-definite matrices with a block structure such that the $ij$th $m \times m$ block is $\boldsymbol{\kappa}(\mathbf{x}_i, \mathbf{x}_j)$. This in turn entails the requirement that $\boldsymbol{\kappa}(\mathbf{x}, \mathbf{x})$ is PSD for all $\mathbf{x}$.

## Definition and Existence

Let $\mathbf{x}(c,t)$ be a multivariate stochastic process with index set $\mathbb{N} \times I$ and range $\mathbb{R}^m$. For every finite list of indices $L = \{(c_1,t_1)\ldots,(c_p,t_p)\}$, let $\mathbf{x}(L)$ be the $mp \times 1$ random vector,

$$\mathbf{x}(L) = \begin{bmatrix} \mathbf{x}(c_1,t_1) \\ \vdots \\ \mathbf{x}(c_p,t_p) \end{bmatrix}.$$

If for every $L$ and for every list of vectors $\mathbf{w}_1,\ldots,\mathbf{w}_p \in \mathbb{R}^m$ with $\mathbf{w} = [\mathbf{w}_1^{\mathsf{T}} \ \ldots \ \mathbf{w}_p^{\mathsf{T}}]^{\mathsf{T}}$, the characteristic function of $\mathbf{x}(L)$ satisfies

$$E\left[\exp\left(\mathrm{i}\mathbf{w}^{\mathsf{T}}\mathbf{x}(L)\right)\right] = \exp\left(-\frac{1}{2}\sum_{j=1}^{p}\sum_{j'=1}^{p}\mathbf{w}_j^{\mathsf{T}}\boldsymbol{\Sigma}_{(c_j,t_j)(c_{j'},t_{j'})}\mathbf{w}_{j'} + \mathrm{i}\sum_{j=1}^{q}\mathbf{w}_j^{\mathsf{T}}\boldsymbol{\mu}_{(c_j,t_j)}\right), \tag{D.5}$$

for some parameters $\{\mu_{(c,t)}\}$ and $\{\Sigma_{(c_1,t_1)(c_2,t_2)}\}$, then $\mathbf{x}(c,t)$ is an $m$-dimensional JMV-GP. The parameters are the mean and covariance of the JMV-GP. The covariance can be specified in terms of a PSD matrix kernel $\boldsymbol{\kappa}(c_1,t_1,c_2,t_2)$ which takes values in $\mathbb{R}^{m\times m}$ such that

$$\mathrm{Cov}\left(\mathbf{x}(c_1,t_1),\mathbf{x}(c_2,t_2)\right) = \boldsymbol{\kappa}(c_1,t_1,c_2,t_2).$$

For any index set $I$, any mean function $\boldsymbol{\mu} : \mathbb{N}\times I \to \mathbb{R}^m$ and any PSD matrix kernel function $\boldsymbol{\kappa} : \mathbb{N}\times I\times\mathbb{N}\times I \to \mathbb{R}^{m\times m}$, there exists a probability triple $(\Omega,\mathcal{F},P)$ and $m$-dimensional JMV-GP $\mathbf{x}(c,t)$ such that, for any finite list of indices $L = \{(c_1,t_1)\ldots(c_p,t_p)\}$ , $\mathbf{x}(L)$ satisfies

$$\begin{aligned} E\left[\mathbf{x}(L)\right] &= \mathbf{m}(L), \\ E\left[\left(\mathbf{x}(L)-\mathbf{m}(L)\right)\left(\mathbf{x}(L)-\mathbf{m}(L)\right)^{\mathsf{T}}\right] &= \mathbf{R}(L), \end{aligned} \tag{D.6}$$

with

$$\mathbf{m}(L) = \begin{bmatrix} \boldsymbol{\mu}(c_1,t_1) \\ \vdots \\ \boldsymbol{\mu}(c_p,t_p) \end{bmatrix}, \qquad \mathbf{R}(L) = \begin{bmatrix} \boldsymbol{\kappa}(c_1,t_1,c_1,t_1) & \ldots & \boldsymbol{\kappa}(c_1,t_1,c_p,t_p) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\kappa}^{\mathsf{T}}(c_1,t_1,c_p,t_p) & \ldots & \boldsymbol{\kappa}(c_p,t_p,c_p,t_p) \end{bmatrix}.$$

This follows from a standard application of the Kolmogorov extension theorem (Tao, 2011, Thm. 2.4.3). The characteristic function definition (D.5) with

$$\begin{aligned} \Sigma_{(c_j,t_j)(c_{j'},t_{j'})} &= [\mathbf{R}(L)]_{jj'}, \\ \boldsymbol{\mu}_{c_j,t_j} &= [\mathbf{m}(L)]_j, \end{aligned} \tag{D.7}$$

requires that the finite-dimensional distributions satisfy

$$\mathbf{x}(L) \sim \mathcal{N}(\mathbf{m}(L), \mathbf{R}(L)), \tag{D.8}$$

for every finite $L$. These finite-dimensional distributions satisfy the consistency condition of the Kolmogorov extension theorem, as for any sublist of indices $J = \{(c_{j_1}, t_{j_1}) \ldots, (c_{j_q}, t_{j_q})\}$ of $L$, the marginal distribution of $\mathbf{x}(J)$ satisfies

$$\mathbf{x}(J) \overset{d}{=} \mathbf{P}_J \mathbf{X}(L), \tag{D.9}$$

where

$$\mathbf{x}(J) \sim \mathcal{N}\left( \begin{bmatrix} \boldsymbol{\mu}(c_{j_1}, t_{j_1}) \\ \vdots \\ \boldsymbol{\mu}(c_{j_q}, t_{j_q}) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\kappa}(c_{j_1}, t_{j_1}, c_{j_1}, t_{j_1}) & \ldots & \boldsymbol{\kappa}(c_{j_1}, t_{j_1}, c_{j_q}, t_{j_q}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\kappa}^{\mathsf{T}}(c_{j_1}, t_{j_1}, c_{j_q}, t_{j_q}) & \ldots & \boldsymbol{\kappa}(c_{j_q}, t_{j_q}, c_{j_q}, t_{j_q}) \end{bmatrix} \right),$$

by the specification of the finite-dimensional distributions. This follows as the multivariate normal distribution transforms appropriately under projections, hence the $pm \times qm$ projection matrix defined as

$$\mathbf{D}_j = \begin{cases} \mathbf{I}_m & \text{if } (c_j, t_j) \in J \\ \mathbf{0}_{m,m} & \text{o.w.} \end{cases},$$

$$\mathbf{P}_J = \text{blkdiag}(\mathbf{D}_1, \ldots, \mathbf{D}_p),$$

provides the appropriate connection as

$$\mathbf{P}_J \mathbf{x}(L) \overset{d}{=} \mathcal{N}\left(\mathbf{P}_J \mathbf{m}(L), \mathbf{P}_J \mathbf{R}(L) \mathbf{P}_J\right) \overset{d}{=} \mathbf{x}(J). \tag{D.10}$$

As the covariance kernel $\boldsymbol{\kappa}$ is assumed to be a PSD matrix kernel, the marginal covariance matrix $\mathbf{P}_J\mathbf{K}\mathbf{P}_J$ is also PSD and so the marginal multivariate normal distribution for $J$ is well-defined. So, the finite-dimensional distributions are consistent, and the Kolmogorov extension theorem gives the existence of a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ supporting a stochastic process $\mathbf{x}$ on the index set $\mathbb{N} \times I$ with the desired finite-dimensional distributions. This process $\mathbf{x}$ satisfies the defining characteristic function identity, and so can be taken as an appropriate JMV-GP with mean function $\boldsymbol{\mu}$ and covariance kernel $\boldsymbol{\kappa}$.

If $\mathbf{x}(c,t)$ is a JMV-GP with separable kernel $\boldsymbol{\kappa}(c_1,t_1,c_2,t_2) = \kappa_{\text{time}}(t_1,t_2)\boldsymbol{\kappa}_{\text{id}}(c_1,c_2)$, then for any rectangular set of indices $K = \{c_1,\ldots,c_C\} \times \{t_1,\ldots,t_T\}$ and any $mCT \times 1$ vector

$$\mathbf{w} \equiv [\mathbf{w}_{1,1}^\mathsf{T} \ \ldots \ \mathbf{w}_{1,T}^\mathsf{T} \ \ldots \ \mathbf{w}_{C,T}^\mathsf{T}]^\mathsf{T},$$

the associated characteristic function of $\mathbf{x}(K) = \text{vec}(\mathbf{X}(\mathbf{c},\mathbf{t}))$ is

$$\begin{aligned}
&E\left[\exp(\mathrm{i}\textstyle\sum_{i=1}^{C}\sum_{j=1}^{T}\mathbf{w}_{i,j}^\mathsf{T}\mathbf{x}(c_i,t_j)\right] \\
&= \exp\left(\sum_{j=1}^{T}\sum_{j'=1}^{T} -\frac{1}{2}\sum_{i=1}^{C}\sum_{i'=1}^{C}\mathbf{w}_{i,j}^\mathsf{T}\kappa_{\text{time}}(t_j,t_{j'})\boldsymbol{\kappa}_{\text{id}}(c_i,c_{i'})\mathbf{w}_{i',j'} + \mathrm{i}\sum_{j=1}^{T}\sum_{i=1}^{C}\mathbf{w}_{i,j}^\mathsf{T}\boldsymbol{\mu}(c_i,t_j)\right) \\
&= \exp\left(-\frac{1}{2}\sum_{j=1}^{T}\sum_{j'=1}^{T}\kappa_{\text{time}}(t_j,t_{j'})\sum_{i=1}^{C}\sum_{i'=1}^{C}\mathbf{w}_{i,j}^\mathsf{T}\boldsymbol{\kappa}_{\text{id}}(c_i,c_{i'})\mathbf{w}_{i',j'} + \mathrm{i}\sum_{j=1}^{T}\sum_{i=1}^{C}\mathbf{w}_{i,j}^\mathsf{T}\boldsymbol{\mu}(c_i,t_j)\right) \\
&= \exp\left(-\frac{1}{2}\mathbf{w}^\mathsf{T}(\mathbf{U}(\mathbf{t})\otimes\mathbf{R}(\mathbf{c}))\mathbf{w} + \mathrm{i}\mathbf{w}^\mathsf{T}\text{vec}(\mathbf{M}(\mathbf{c},\mathbf{t}))\right),
\end{aligned}$$

where $\mathbf{U}, \mathbf{R}, \mathbf{M}$ are as defined in (4.4). The last line is the characteristic function of a multivariate normal distribution with mean $\text{vec}\big(\mathbf{M}(\mathbf{c},\mathbf{t})\big)$ and covariance $\mathbf{U}(\mathbf{t})\otimes\mathbf{R}(\mathbf{c})$, which is in turn equivalent to the matrix-normal distribution $\mathcal{MN}(\mathbf{M}(\mathbf{c},\mathbf{t}), \mathbf{R}(\mathbf{c}), \mathbf{U}(\mathbf{t}))$.

**Additivity and Linear Transformations**

As the JMV-GP has Gaussian marginals, it is not surprising that linear transformations of JMV-GPs and sums of independent JMV-GPs will also be JMV-GPs. In particular, let $\mathbf{x}(c,t)$ and $\mathbf{y}(c,t)$ be two $m$-dimensional JMV-GPs defined on the same probability triple, with means $\boldsymbol{\mu}_\mathbf{x}(c,t), \boldsymbol{\mu}_\mathbf{y}(c,t)$ and covariance kernels $\boldsymbol{\kappa}_\mathbf{x}(c_1,t_1,c_2,t_2), \boldsymbol{\kappa}_\mathbf{y}(c_1,t_1,c_2,t_2)$ respectively. Assume

that $\mathbf{x}, \mathbf{y}$ are independent as for finite index sets

$$L_1 = \{(c_1, t_1), \ldots (c_p, t_p)\},$$

$$L_2 = \{(c_{p+1}, t_{p+1}), \ldots, (c_{p+q}, t_{p+q})\},$$

the random vectors $\mathbf{x}(L_1)$ and $\mathbf{y}(L_2)$ are independent. The sum process $\mathbf{z}(c,t) = \mathbf{x}(c,t) + \mathbf{y}(c,t)$ is also an $m$-dimensional JMV-GP with

$$\boldsymbol{\mu}_{\mathbf{z}}(c,t) = \boldsymbol{\mu}_{\mathbf{x}}(c,t) + \boldsymbol{\mu}_{\mathbf{y}}(c,t),$$

$$\boldsymbol{\kappa}_{\mathbf{z}}(c,t) = \boldsymbol{\kappa}_{\mathbf{x}}(c_1,t_1,c_2,t_2) + \boldsymbol{\kappa}_{\mathbf{y}}(c_1,t_1,c_2,t_2).$$

This follows as the characteristic function of $\mathbf{z}(L)$ for any $mp \times 1$ vector $\mathbf{w} = [\mathbf{w}_1^{\mathsf{T}} \ \ldots \ \mathbf{w}_p^{\mathsf{T}}]^{\mathsf{T}}$ is,

$$\begin{aligned} E\left[\mathrm{i}\mathbf{w}^{\mathsf{T}}\mathbf{z}(L)\right] &= E\left[\mathrm{i}\sum_{j=1}^{p}\mathbf{w}_j^{\mathsf{T}}\mathbf{x}(c_j,t_j)\right] E\left[\mathrm{i}\sum_{j=1}^{p}\mathbf{w}_j^{\mathsf{T}}\mathbf{y}(c_j,t_j)\right] \\ &= \exp\Bigg(-\frac{1}{2}\sum_{j=1}^{p}\sum_{j'=1}^{p}\mathbf{w}_j^{\mathsf{T}}\left(\boldsymbol{\kappa}_{\mathbf{x}}(c_j,t_j,c_{j'},t_{j'}) + \boldsymbol{\kappa}_{\mathbf{y}}(c_j,t_j,c_{j'},t_{j'})\right)\mathbf{w}_{j'} \\ &\qquad + \mathrm{i}\sum_{j=1}^{q}\mathbf{w}_j^{\mathsf{T}}\left(\boldsymbol{\mu}_{\mathbf{x}}(c_j,t_j) + \boldsymbol{\mu}_{y}(c_j,t_j)\right)\Bigg), \end{aligned}$$

by the assumption of independence.

Last, if $\mathbf{A}(c,t) : \mathbb{N} \times I : \mathbb{R}^{d\times m}$ is a matrix-valued function, the transformed process

$$\mathbf{v}(c,t) = \mathbf{A}(c,t)\mathbf{x}(c,t),$$

will be a $d$-dimensional JMV-GP. This follows as the characteristic function of $\mathbf{v}(L)$ is

$$E\left[\mathrm{i}\sum_{j=1}^{p}\mathbf{w}_j^{T}\mathbf{v}(c_j,t_j)\right] = E\left[\mathrm{i}\sum_{j=1}^{q}(\mathbf{A}^{\mathsf{T}}(c_j,t_j)\mathbf{w}_j)^{\mathsf{T}}\mathbf{x}(c_j,t_j)\right]$$

$$= \exp\left(-\tfrac{1}{2}\textstyle\sum_{j=1}^{p}\sum_{j'=1}^{p}\mathbf{w}_j^{\mathsf{T}}\mathbf{A}(c_j,t_j)\boldsymbol{\kappa}_x(c_j,t_j,c_{j'},t_{j'})\mathbf{A}^{\mathsf{T}}(c_{j'},t_{j'})\mathbf{w}_{j'} + \mathrm{i}\textstyle\sum_{j=1}^{q}\mathbf{w}_j^{\mathsf{T}}\mathbf{A}(c_j,t_j)\boldsymbol{\mu}_x(c_j,t_j)\right).$$

Hence $\mathbf{v}(c,t)$ is a $d$-dimensional JMV-GP with mean function $\boldsymbol{\mu}_{\mathbf{v}}(c,t) = \mathbf{A}(c,t)\boldsymbol{\mu}_{\mathbf{x}}(c,t)$ and covariance kernel $\boldsymbol{\kappa}_{\mathbf{v}}(c_1,t_1,c_2,t_2) = \mathbf{A}(c_1,t_1)\boldsymbol{\kappa}_{\mathbf{x}}(c_1,t_1,c_2,t_2)\mathbf{A}^{\mathsf{T}}(c_2,t_2)$.

## D.2 Proof of Proposition 16

*Proof.* As $\{\mathbf{f}_{\mathcal{T}}(t)\}$ is locally stationary, let $\widetilde{\mathbf{f}}_t(u)$ and $U^{(f)}_{t,\mathcal{T}}(u)$ be the associated stationary approximation satisfying (4.14). For a cohort $\mathbf{c} = [c_1, \ldots, c_C]^{\mathsf{T}}$, let $\mathbf{g}_{\mathcal{T}}(t)$ and $\mathbf{w}_{\mathcal{T}}(t)$ be the stacked vectors,

$$\mathbf{g}_{\mathcal{T}}(t) = [\mathbf{g}^{\mathsf{T}}_{\mathcal{T}}(c_1, t) \ \ldots \ \mathbf{g}^{\mathsf{T}}_{\mathcal{T}}(c_C, t)]^{\mathsf{T}},$$

$$\mathbf{w}_{\mathcal{T}}(t) = [\mathbf{w}^{\mathsf{T}}_{\mathcal{T}}(c_1, t) \ \ldots \ \mathbf{w}^{\mathsf{T}}_{\mathcal{T}}(c_C, t)]^{\mathsf{T}}.$$

By assumption, these processes have associated stationary approximations $\widetilde{\mathbf{g}}_t(u)$ $\widetilde{\mathbf{w}}_t(u)$ satisfying (4.14) for some positive processes $U^{(g)}_{t,\mathcal{T}}(u)$ and $U^{(w)}_{t,\mathcal{T}}(u)$. From (A1), the positive constants $k_1, k_2, k_3, k_4$ are the Lipschitz constants for $\mathbf{A}'(c,u)$, $\mathbf{B}'(c,u)$ $\mathbf{\Phi}'(c,u)$, and $\boldsymbol{\mu}'(c,u)$ respectively, while the positive constants $K_1, K_2, K_3$ are the respective bounds on $||\mathbf{A}(c,u)||_F, ||\mathbf{B}(c,u)||_F$ and $||\mathbf{\Phi}(c,u)||_F$ guaranteed by (A2). Define the approximation

$$\widetilde{\mathbf{x}}_t(u) = \mathbf{A}'(u)\widetilde{\mathbf{f}}_t(u) + \mathbf{B}'(u)\widetilde{\mathbf{g}}_t(u) + \mathbf{\Phi}'^{\frac{1}{2}}(u)\widetilde{\mathbf{w}}_t(u) + \boldsymbol{\mu}'(u), \tag{D.11}$$

with $\mathbf{A}'(u) = \left[\mathbf{A}'^{\mathsf{T}}(c_1, u) \ \ldots \ \mathbf{A}'^{\mathsf{T}}(c_C, u)\right]^{\mathsf{T}}$, while $\mathbf{B}'(u) = \text{blkdiag}\left(\mathbf{B}'(c_1, u), \ldots, \mathbf{B}'(c_C, u)\right)$ and $\mathbf{\Phi}'(u)$ is defined similarly. The mean $\boldsymbol{\mu}'(u)$ equals $\left[\boldsymbol{\mu}'^{\mathsf{T}}(c_1, u) \ \ldots \ \boldsymbol{\mu}'^{\mathsf{T}}(c_C, u)\right]^{\mathsf{T}}$. Combining (D.11) with (4.13) yields

$$\begin{aligned}
||\mathbf{x}_{\mathcal{T}}(t) - \widetilde{\mathbf{x}}_t(u)|| \leq\ & ||\mathbf{A}'(t/\mathcal{T}) - \mathbf{A}'(u)||_F||\mathbf{f}_{\mathcal{T}}(t)|| + ||\mathbf{A}'(u)||_F||\mathbf{f}_{\mathcal{T}}(t) - \widetilde{\mathbf{f}}_t(u)|| \\
&+ ||\mathbf{B}'(t/\mathcal{T}) - \mathbf{B}'(u)||_F||\mathbf{g}_{\mathcal{T}}(t)|| + ||\mathbf{B}'(u)||_F||\mathbf{g}_{\mathcal{T}}(t) - \widetilde{\mathbf{g}}_t(u)|| \\
&+ ||\mathbf{\Phi}'^{\frac{1}{2}}(t/\mathcal{T}) - \mathbf{\Phi}'^{\frac{1}{2}}(u)||_F||\mathbf{w}_{\mathcal{T}}(t)||_F + ||\mathbf{\Phi}'(u)||^{\frac{1}{2}}_F||\mathbf{w}_{\mathcal{T}}(t) - \widetilde{\mathbf{w}}_t(u)|| \\
&+ ||\boldsymbol{\mu}'_{\mathcal{T}}(t/\mathcal{T}) - \boldsymbol{\mu}'(u)|| \\
\leq\ & (k_1||\mathbf{f}_{\mathcal{T}}(t)|| + k_2||\mathbf{g}_{\mathcal{T}}(t)|| + k_3||\mathbf{w}_{\mathcal{T}}(t)|| + k_4)|t/\mathcal{T} - u| \\
&+ \left(K_1U^{(f)}_{t,\mathcal{T}} + K_2U^{(g)}_{t,\mathcal{T}} + K_3^{\frac{1}{2}}U^{(w)}_{t,\mathcal{T}}\right)(|t/\mathcal{T} - u| + \mathcal{T}^{-1}).
\end{aligned}$$

Therefore, the requirement (4.14) is satisfied with

$$U^{(x)}_{t,\mathcal{T}} = K_1U^{(f)}_{t,\mathcal{T}} + K_2U^{(g)}_{t,\mathcal{T}} + K_3^{\frac{1}{2}}U^{(w)}_{t,\mathcal{T}} + k_1||\mathbf{f}_{\mathcal{T}}(t)|| + k_2||\mathbf{g}_{\mathcal{T}}(t)|| + k_3||\mathbf{w}_{\mathcal{T}}(t)|| + k_4,$$

as $\mathbf{f}_{\mathcal{T}}(t), \mathbf{g}_{\mathcal{T}}(t)$ and $\mathbf{w}_{\mathcal{T}}(t)$ have identity covariance and so $U^{(x)}_{t,\mathcal{T}}$ will have a finite moment. □

## D.3 Proof of Propositions 2 and 3

*Proof.* Suppose $\mathbf{Y}$ is such that $\mathbf{T}_{\mathbf{V},B}(\mathbf{Y}) = \mathbf{0}$ for some $\mathbf{V} \in \mathbb{R}^{n\times T}$. This implies that the projection onto $S_B$, $\mathbf{P}_{S_B}\mathbf{Y}$, equals $-\mathbf{m}_B \otimes \mathbf{V}$ while the projection onto $S_B^{\perp}$ is unconstrained. So, if $N \subset \mathbb{R}^{Cn\times T}$ is the subspace of such $\mathbf{Y}$, then counting the degrees of freedom in $\mathbf{V}$ and $S_B^{\perp}$ yields $\dim N = Tn + (C - |B|)Tn$. Also, $P_B(\mathbf{Y}) = \mathbf{0}$ iff $\mathbf{Y} - \mathbf{m}_B \otimes (|B|^{-1}(\mathbf{m}_B^{\mathsf{T}} \otimes \mathbf{I}_n)\mathbf{Y})$ is in $S_B^{\perp}$, which is equivalent to $\mathbf{P}_{S_B}\mathbf{Y} = -\mathbf{m}_B \otimes \mathbf{V}$ for some $\mathbf{V}$. Hence, $\operatorname{Ker} P_B = N$. By the first isomorphism theorem, the quotient space $\mathbb{R}^{Cn\times p}/N$ is isomorphic to the subspace $\operatorname{col} P_B$, and the rank-nullity theorem gives $\dim \operatorname{range} P_B = (|B| - 1)Tn$. To verify that the map $P_B(\mathbf{Y})$ is the orthogonal projection on $N$, it can be seen that

$$\mathbf{m}_B \otimes \left(|B|^{-1}(\mathbf{m}_B^{\mathsf{T}} \otimes \mathbf{I}_n)(\mathbf{m}_B \otimes \mathbf{V})\right) = \mathbf{m}_B \otimes \mathbf{V},$$

and so

$$P_B(P_B(\mathbf{Y})) = \mathbf{P}_{S_B}\mathbf{Y} - \mathbf{P}_{S_B}\mathbf{m}_B \otimes (|B|^{-1}(\mathbf{m}_B^{\mathsf{T}} \otimes \mathbf{I}_n)\mathbf{Y}) = P_B(\mathbf{Y})$$

by the multilinearity of the Kronecker product and the fact that $(\mathbf{m}_B^{\mathsf{T}} \otimes \mathbf{I}_n)\mathbf{P}_{S_B} = (\mathbf{m}_B^{\mathsf{T}} \otimes \mathbf{I}_n)$. By above, all elements in $\operatorname{Ker} \mathbf{P}_B$ are of the form $\mathbf{m}_B \otimes \mathbf{V}$, and as

$$\begin{aligned}
&\operatorname{tr}\left((\mathbf{m}_B \otimes \mathbf{V})^{\mathsf{T}}(\mathbf{P}_{S_B}\mathbf{Y} - \mathbf{m}_B \otimes (|B|^{-1}(\mathbf{m}_B^{\mathsf{T}} \otimes \mathbf{I}_n)\mathbf{Y}))\right) \\
&= \operatorname{tr} \mathbf{Y}^{\mathsf{T}}\mathbf{P}_{S_B}(\mathbf{m}_B \otimes \mathbf{V}) - \operatorname{tr}(\mathbf{m}_B^{\mathsf{T}}\mathbf{m}_B \otimes |B|^{-1}\mathbf{V}^{\mathsf{T}}(\mathbf{m}_B^{\mathsf{T}} \otimes \mathbf{I}_n)\mathbf{Y}) \\
&= \operatorname{tr} \mathbf{Y}^{\mathsf{T}}(\mathbf{P}_{S_B}(\mathbf{m}_B \otimes \mathbf{V}) - (\mathbf{m}_B \otimes \mathbf{I}_n)\mathbf{V}) = 0
\end{aligned}$$

as $\mathbf{m}_B \otimes \mathbf{V}$ is in $S_B$ and $(\mathbf{m}_B \otimes \mathbf{I}_n)\mathbf{V} = (\mathbf{m}_B \otimes \mathbf{I}_n)(1 \otimes \mathbf{V}) = (\mathbf{m}_B \otimes \mathbf{V})$. Hence $P_B$ is idempotent and the kernel and image are orthogonal in the Frobenius inner product, so $P_B$ is the orthogonal projection onto its image, which is isomorphic to the quotient space $\mathbb{R}^{Cn\times p}/N$. □

## D.4 Proof for Theorem 6

*Proof.* Let $C'(u)$ be the size of the cohort curve $\mathbf{c}'(u)$ at each $u$, which has $\max_{u\in[0,1]} C'(u) < \infty$ by assumption. As $\mathbf{c}'(u)$ has a finite number of points of discontinuity, let $0 < v_1 < \cdots < v_D < 1$ denote them. As the codomain of $\mathbf{c}'(u)$ is a discrete space and $\mathbf{c}'(u)$ is right-continuous, $\mathbf{c}'(u)$ is constant on the intervals $\mathcal{I}_0 = [0, v_1), \mathcal{I}_1 = [v_1, v_2), \ldots, \mathcal{I}_D = [v_D, 1)$. Let $\delta > 0$ be the minimum width of these intervals. Fix some $u_0 \in (0, 1)$, and note that we have two cases, namely where $u_0$ equals $v_j$ for some $j = 1, \ldots, D$ or $u_0$ is a continuity point of $\mathbf{c}'(u)$. We treat the second case, as the first differs only in that the window will be one sided. Let $\mathcal{I}_{u_0}$ be the interval within whose interior $u_0$ belongs, and let the time $t_0$ be $\lceil \mathcal{T} u_0 \rceil$, so $\lceil \mathcal{T} u_0 \rceil / \mathcal{T}$ is the next grid point. As $u_0$ is a continuity point of $\mathbf{c}'(u)$ so too is $t_0$ for $\mathcal{T}$ sufficiently large. Similarly, $\mathbf{c}'(t_0) = \mathbf{c}'(u_0)$ as long as $\lceil u_0 \mathcal{T} \rceil / \mathcal{T} < \sup \mathcal{I}_{u_0}$, which will be the case for $\mathcal{T}$ sufficiently large. Hence we can assume that $\mathbf{c}'(t_0/\mathcal{T}) = \mathbf{c}'(u_0)$ and so the parameter curve estimates at $u_0$ will be obtained using the windowed data $\mathbf{X}\big(\mathbf{c}'(u_0), \mathbf{s}_\mathcal{T}(t_0)\big)$ with $\mathbf{s}_\mathcal{T}(t_0)$ of length $W(\mathcal{T})$.

Define the $C'(u_0)nW(\mathcal{T}) \times 1$ vector $\mathbf{z}$ as,

$$\mathbf{z} = \operatorname{vec}\left(\mathbf{X}\big(\mathbf{c}'(u_0), \mathbf{s}_\mathcal{T}(t_0)\big) - \overline{\mathbf{x}}\big(\mathbf{c}'(u_0), \mathbf{s}_\mathcal{T}(t_0)\big) \otimes \mathbf{1}_{W(\mathcal{T})}\right),$$

so that the stationary approximation pseudo-log-likelihood $\widetilde{\ell}_{t_0,\mathcal{T}}$ can be written in terms of $\mathbf{z}$ as

$$\widetilde{\ell}_{t_0,\mathcal{T}}(\mathbf{z}; \mathbf{A}, \mathbf{B}, \mathbf{\Phi}, \boldsymbol{\theta}) = -\mathbf{z}^\mathsf{T}\big(\mathbf{U}_\mathcal{T}(t_0; \boldsymbol{\theta}) \otimes \mathbf{R}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})\big)^{-1}\mathbf{z} - \log\det\big(\mathbf{U}_\mathcal{T}(t_0; \boldsymbol{\theta}) \otimes \mathbf{R}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})\big),$$

where $\mathbf{A}$ is $C'(u_0)n \times r_0$, $\mathbf{B}$ is $C'(u_0)n \times C'(u_0)r_1$ and $\mathbf{\Phi}$ is $C'(u_0)n \times C'(u_0)n$. Then $\mathbf{U}_\mathcal{T}(t_0; \boldsymbol{\theta})$ is

$$\mathbf{U}_\mathcal{T}(t_0; \boldsymbol{\theta}) = \begin{bmatrix} 1 & \ldots & k(|s_1 - s_{W(\mathcal{T})}|; \boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ k(|s_{W(\mathcal{T})} - s_1|; \boldsymbol{\theta}) & \ldots & 1 \end{bmatrix},$$

where $\mathbf{s}_\mathcal{T}(t_0) = [s_1, \ldots, s_{W(\mathcal{T})}]^\mathsf{T}$ and $s_j = t_0 - W(\mathcal{T})/2 + j$.

Under the time-varying model (4.13), the vector $\mathbf{z}$ is distributed $\mathcal{N}\big(\mathbf{0}, \mathbf{V}_{\mathcal{T}}(t_0)\big)$, where $\mathbf{V}_{\mathcal{T}}(t_0)$ is the $C'(u_0)nW(\mathcal{T}) \times C'(u_0)nW(\mathcal{T})$ matrix with blocks of size $C'(u_0)n \times C'(u_0)n$ defined by

$$\begin{aligned}
\mathbf{R}_{\mathcal{T},ij}(t_0) &\equiv \mathbf{A}'(\mathbf{c}'(u_0), s_i/\mathcal{T})\mathbf{A}'^{\mathsf{T}}(\mathbf{c}'(u_0), s_j/\mathcal{T}) + \mathbf{B}'(\mathbf{c}'(u_0), s_i/\mathcal{T})\mathbf{B}'^{\mathsf{T}}(\mathbf{c}'(u_0), s_j/\mathcal{T}) \\
&\quad + \mathbf{\Phi}'^{\frac{1}{2}}(\mathbf{c}'(u_0), s_i/\mathcal{T})\mathbf{\Phi}'^{\frac{1}{2}}(\mathbf{c}'(u_0), s_j/\mathcal{T}), \\
[\mathbf{V}_{\mathcal{T}}(t_0)]_{ij} &\equiv \kappa_{\mathcal{T},\text{time}}\big(s_i, s_j; \boldsymbol{\theta}(u)\big)\mathbf{R}_{\mathcal{T},ij}(t_0),
\end{aligned}$$

for $i, j = 1, \ldots, W(\mathcal{T})$. The cross-sectional covariance matrices are

$$\begin{aligned}
\mathbf{A}'(\mathbf{c}'(u_0), u) &= \big[\mathbf{A}'^{\mathsf{T}}([\mathbf{c}'(u_0)]_1, u) \ \ldots \ \mathbf{A}'^{\mathsf{T}}([\mathbf{c}'(u_0)]_{C'(u_0)}, u)\big]^{\mathsf{T}}, \\
\mathbf{B}'(\mathbf{c}'(u_0), u) &= \text{blkdiag}\left(\mathbf{B}'([\mathbf{c}'(u_0)]_1, u), \ldots, \mathbf{B}'([\mathbf{c}'(u_0)]_{C'(u_0)}, u)\right), \\
\mathbf{\Phi}'(\mathbf{c}'(u_0), u) &= \text{blkdiag}\left(\mathbf{\Phi}'([\mathbf{c}'(u_0)]_1, u), \ldots, \mathbf{\Phi}'([\mathbf{c}'(u_0)]_{C'(u_0)}, u)\right).
\end{aligned}$$

With these definitions, the log-likelihood for $\mathbf{z}$ under the time-varying model is

$$\ell_{t_0,\mathcal{T}}\big(\mathbf{z}; \mathbf{A}'(c,u), \mathbf{B}'(c,u), \mathbf{\Phi}'(c,u), \boldsymbol{\theta}'(u)\big) = -\mathbf{z}^{\mathsf{T}}\mathbf{V}_{\mathcal{T}}^{-1}(t_0)\mathbf{z} - \log\det\mathbf{V}_{\mathcal{T}}(t_0),$$

plus constant terms. The difference between the time-varying and stationary log-likelihoods is,

$$\begin{aligned}
&d_{t_0,\mathcal{T}}(\mathbf{z}) = \ell_{t_0,\mathcal{T}}(\mathbf{z}; \mathbf{A}'(c,u), \mathbf{B}'(c,u), \mathbf{\Phi}'(c,u), \boldsymbol{\theta}'(u)) - \widetilde{\ell}_{t_0,\mathcal{T}}(\mathbf{z}; \mathbf{A}, \mathbf{B}, \mathbf{\Phi}, \boldsymbol{\theta}) \\
&= -\mathbf{z}^{\mathsf{T}}\big(\mathbf{V}_{\mathcal{T}}^{-1}(t_0) - (\mathbf{U}_{\mathcal{T}}^{-1}(t_0;\boldsymbol{\theta}) \otimes \mathbf{R}^{-1}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})\big)\mathbf{z} - \log\det\mathbf{V}_{\mathcal{T}}(t_0)\big(\mathbf{U}_{\mathcal{T}}^{-1}(t_0;\boldsymbol{\theta}) \otimes \mathbf{R}^{-1}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})\big).
\end{aligned}$$

To control $|d_{t_0,\mathcal{T}}(\mathbf{z})|$, the first needed bound is

$$\begin{aligned}
&\left|\mathbf{z}^{\mathsf{T}}\big(\mathbf{V}_{\mathcal{T}}^{-1}(t_0) - \mathbf{U}_{\mathcal{T}}^{-1}(t_0;\boldsymbol{\theta}) \otimes \mathbf{R}^{-1}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})\big)\mathbf{z}\right| \leq ||\mathbf{z}||^2 ||\mathbf{V}_{\mathcal{T}}^{-1}(t_0) - \mathbf{U}_{\mathcal{T}}^{-1}(t_0;\boldsymbol{\theta}) \otimes \mathbf{R}^{-1}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})||_2 \\
&\leq ||\mathbf{z}||^2 ||\mathbf{V}_{\mathcal{T}}^{-1}(t_0)||_2 ||\mathbf{U}_{\mathcal{T}}^{-1}(t_0;\boldsymbol{\theta}) \otimes \mathbf{R}^{-1}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})||_2 ||\mathbf{V}_{\mathcal{T}}(t_0) - (\mathbf{U}_{\mathcal{T}}(t_0;\boldsymbol{\theta}) \otimes \mathbf{R}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}))||_F
\end{aligned} \tag{D.12}$$

where $||\cdot||_2$ is the $\ell_2$ operator norm. By Assumptions (A3) and (B4), the smallest eigenvalue of $\mathbf{V}_{\mathcal{T}}(t_0)$ is be bounded below. To see this, define

$$\mathbf{u} = \begin{bmatrix} \mathbf{\Phi}'^{\frac{1}{2}}(\mathbf{c}'(u_0), s_1/\mathcal{T})\mathbf{w}_{\mathcal{T}}(s_1) \\ \vdots \\ \mathbf{\Phi}'^{\frac{1}{2}}(\mathbf{c}'(u_0), s_{W(\mathcal{T})}/\mathcal{T})\mathbf{w}_{\mathcal{T}}(s_{W(\mathcal{T})}) \end{bmatrix},$$

and note that $\text{Var}(\mathbf{u}) \preceq \text{Var}(\mathbf{z}) = \mathbf{V}_{\mathcal{T}}(t_0)$ as, by (4.13), $\mathbf{z}$ can be written as a sum of independent random vectors with $\mathbf{u}$ as one of the vectors. The variance of $\mathbf{u}$ is,

$$\begin{aligned}\mathbf{P} &= \text{blkdiag}(\boldsymbol{\Phi}'^{\frac{1}{2}}(\mathbf{c}'(t_0), s_1/\mathcal{T}), \dots \boldsymbol{\Phi}'^{\frac{1}{2}}(\mathbf{c}'(t_0), s_1/\mathcal{T})),\\ \text{Var}(\mathbf{u}) &= \mathbf{P}^{\frac{1}{2}}\big(\mathbf{K}(\mathbf{w}_{\mathcal{T}}(t_0); \boldsymbol{\theta}'(u)) \otimes \mathbf{I}_{C'(u_0)n}\big)\mathbf{P}^{\frac{1}{2}},\end{aligned}$$

for $\mathbf{K}$ as defined in (B4), and so $\lambda_{\min}\big(\text{Var}(\mathbf{u})\big) \geq \epsilon^2$ from (A3) and (B4). Hence $||\mathbf{V}_{\mathcal{T}}^{-1}(t_0)||_2$ is bounded above by $\epsilon^{-2}$. Similarly, if $\boldsymbol{\theta}_{t_0,\mathcal{T}} = \boldsymbol{\theta}'(t_0/\mathcal{T})$, $\mathbf{A}_{t_0,\mathcal{T}} = \mathbf{A}'\big(\mathbf{c}'(u_0), t_0/\mathcal{T}\big)$ and similarly for $\mathbf{B}_{t_0,\mathcal{T}}$ and $\boldsymbol{\Phi}_{t_0,\mathcal{T}}$, then $||\mathbf{U}_{\mathcal{T}}^{-1}(t_0; \boldsymbol{\theta}_{t_0,\mathcal{T}}) \otimes \mathbf{R}^{-1}(\mathbf{A}_{t_0,\mathcal{T}}, \mathbf{B}_{t_0,\mathcal{T}}, \boldsymbol{\Phi}_{t_0,\mathcal{T}})||_2$ is bounded above by $\epsilon^{-2}$ as the smallest eigenvalue of a Kronecker product is the product of the factors' smallest eigenvalues.

On the other side, by the boundedness assumption (A2) and the fact that the local covariance kernels $k(|j|; \boldsymbol{\theta})$ are scaled to have unit variance, the norm of each block $[\mathbf{V}_{\mathcal{T}}(t_0)]_{ij}$ is bounded above by a constant. Hence $||\mathbf{V}_{\mathcal{T}}(t_0)||_F = O(W(\mathcal{T}))$ as there are $C'(u_0)^2 W(\mathcal{T})^2$ blocks.

Finally, the term $||\mathbf{V}_{\mathcal{T}}(t_0) - (\mathbf{U}_{\mathcal{T}}(t_0; \boldsymbol{\theta}_{t_0,\mathcal{T}}) \otimes \mathbf{R}(\mathbf{A}_{t_0,\mathcal{T}}, \mathbf{B}_{t_0,\mathcal{T}}, \boldsymbol{\Phi}_{t_0,\mathcal{T}}))||_F$ is $O(W(\mathcal{T})^2\mathcal{T}^{-1})$. To see this, consider the $ij$th block of $\mathbf{V}_{\mathcal{T}}(t_0) - (\mathbf{U}_{\mathcal{T}}(t_0; \boldsymbol{\theta}_{t_0,\mathcal{T}}) \otimes \mathbf{R}(\mathbf{A}_{t_0,\mathcal{T}}, \mathbf{B}_{t_0,\mathcal{T}}, \boldsymbol{\Phi}_{t_0,\mathcal{T}}))$, which is

$$\kappa_{\mathcal{T},\text{time}}\big(s_i, s_j; \boldsymbol{\theta}'(u)\big)\mathbf{R}_{\mathcal{T},ij}(t_0) - k(|s_i - s_j|; \boldsymbol{\theta}_{t_0,T})\mathbf{R}(\mathbf{A}_{t_0,\mathcal{T}}, \mathbf{B}_{t_0,\mathcal{T}}, \boldsymbol{\Phi}_{t_0,\mathcal{T}}).$$

The Frobenius norm of this block is bounded above by

$$\begin{aligned}&|\kappa_{\mathcal{T},\text{time}}\big(s_i, s_j; \boldsymbol{\theta}(u)\big) - k(|s_i - s_j|; \boldsymbol{\theta}_{t,\mathcal{T}})| \cdot ||\mathbf{R}_{\mathcal{T},ij}(t_0)||_F\\ &\quad + |k(|s_i - s_j|; \boldsymbol{\theta}_{t,\mathcal{T}})| \cdot ||\mathbf{R}_{\mathcal{T},ij}(t_0) - \mathbf{R}(\mathbf{A}_{t_0,\mathcal{T}}, \mathbf{B}_{t_0,\mathcal{T}}, \boldsymbol{\Phi}_{t_0,\mathcal{T}})||_F\\ &= O(\mathcal{T}^{-1})O(1) + O(1)O(W(\mathcal{T})\mathcal{T}^{-1}) = O(W(\mathcal{T})\mathcal{T}^{-1})\end{aligned}$$

by assumptions (B3), (A2), (A1) and the unit variance of $k(|j|; \boldsymbol{\theta})$. As there are $O(W(\mathcal{T})^2)$ blocks in $\mathbf{V}_{\mathcal{T}}(t_0)$, $||\mathbf{V}_{\mathcal{T}}(t_0) - (\mathbf{U}_{\mathcal{T}}(t_0; \boldsymbol{\theta}_{t_0,\mathcal{T}}) \otimes \mathbf{R}(\mathbf{A}_{t_0,\mathcal{T}}, \mathbf{B}_{t_0,\mathcal{T}}, \boldsymbol{\Phi}_{t_0,\mathcal{T}}))||_F$ is $O(W(\mathcal{T})^2\mathcal{T}^{-1})$. Combining the above bounds with the fact that $||\mathbf{z}||^2/\operatorname{tr}\mathbf{V}_{\mathcal{T}}(t_0) = O_p(1)$ and so $||\mathbf{z}||^2$ is $O_p(W(\mathcal{T}))$ by the boundedness assumptions and number of blocks in $\mathbf{V}_{\mathcal{T}}(t_0)$, we have that

$$\big|\mathbf{z}^{\mathsf{T}}\big(\mathbf{V}_{\mathcal{T}}^{-1}(t_0) - \mathbf{U}_{\mathcal{T}}^{-1}(t_0; \boldsymbol{\theta}) \otimes \mathbf{R}^{-1}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})\big)\mathbf{z}\big| = O_p(W(\mathcal{T})^2\mathcal{T}^{-1}).$$

The other term in $d_{t_0,\mathcal{T}}(\mathbf{z})$ can be bounded using the inequality

$$\log\det \mathbf{V}_{\mathcal{T}}(t_0)\big(\mathbf{U}_{\mathcal{T}}^{-1}(t_0;\boldsymbol{\theta})\otimes\mathbf{R}^{-1}(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})\big) \leq \operatorname{tr}\left(\mathbf{V}_{\mathcal{T}}(t_0)\big(\mathbf{U}_{\mathcal{T}}^{-1}(t_0;\boldsymbol{\theta})\otimes\mathbf{R}^{-1}(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})\big)-\mathbf{I}\right)$$
$$\leq ||\mathbf{V}_{\mathcal{T}}(t_0)-\mathbf{U}_{\mathcal{T}}(t_0;\boldsymbol{\theta})\otimes\mathbf{R}(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})||_F||\mathbf{U}_{\mathcal{T}}^{-1}(t_0;\boldsymbol{\theta})\otimes\mathbf{R}^{-1}(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})||_F,$$

where the first inequality is obtained from the concavity of $\log\det$ and the second is obtained by Cauchy-Schwarz. For $\mathbf{A}_{t_0,\mathcal{T}}, \mathbf{B}_{t_0,\mathcal{T}}, \boldsymbol{\Phi}_{t_0,\mathcal{T}}$ and $\boldsymbol{\theta}_{t_0,\mathcal{T}}$, the previous results similarly imply the rate

$$\log\det \mathbf{V}_{\mathcal{T}}(t_0)\big(\mathbf{U}_{\mathcal{T}}^{-1}(t_0;\boldsymbol{\theta}_{t_0,\mathcal{T}})\otimes\mathbf{R}^{-1}(\mathbf{A}_{t_0,\mathcal{T}},\mathbf{B}_{t_0,\mathcal{T}},\boldsymbol{\Phi}_{t_0,\mathcal{T}})\big) = O(W(\mathcal{T})^2\mathcal{T}^{-1}).$$

Combined with the bound for the other term, the likelihood difference average is

$$\frac{1}{W(\mathcal{T})}|d_{t_0,\mathcal{T}}(\mathbf{z})| = O_p(W(\mathcal{T})\mathcal{T}^{-1}),$$

for $\mathbf{A}=\mathbf{A}_{t_0,\mathcal{T}}, \mathbf{B}=\mathbf{B}_{t_0,\mathcal{T}}, \boldsymbol{\Phi}=\boldsymbol{\Phi}_{t_0,\mathcal{T}}$ and $\boldsymbol{\theta}=\boldsymbol{\theta}_{t_0,\mathcal{T}}$. As $W(\mathcal{T})/\mathcal{T}\to 0$, we have that $\frac{1}{W(\mathcal{T})}|d_{t_0,\mathcal{T}}(\mathbf{z})|$ converges in probability to zero.

Next, define the expected likelihood $\ell^0_{t,\mathcal{T}}(\mathbf{A}'(c,u),\mathbf{B}'(c,u),\boldsymbol{\Phi}'(c,u),\boldsymbol{\theta}'(u))$, which is

$$\begin{aligned}\ell^0_{t,\mathcal{T}}\big(\mathbf{A}'(c,u),\mathbf{B}'(c,u),\boldsymbol{\Phi}'(c,u),\boldsymbol{\theta}'(u)\big) &= E\Big[\ell_{t_0,\mathcal{T}}\big(\mathbf{z};\mathbf{A}'(c,u),\mathbf{B}'(c,u),\boldsymbol{\Phi}'(c,u),\boldsymbol{\theta}'(u)\big)\Big]\\ &= -C'(u_0)nW(\mathcal{T})-\log\det\mathbf{V}_{\mathcal{T}}(t_0).\end{aligned}$$

By the covariance assumptions (B2) and (B3), the compactness of the parameter space assumed in (B5) and $(A2)$, and the smoothness of the parameter curves (B1) and (A1), we have that the difference $\left|\frac{\ell^0_{t,\mathcal{T}}}{W(\mathcal{T})}-\frac{\ell_{t,\mathcal{T}}}{W(\mathcal{T})}\right|$ converges in probability to zero uniformly in the parameter curves as $\mathcal{T}, W(\mathcal{T})\to\infty$. Also, $|d_{t_0,\mathcal{T}}(\mathbf{z})|/W(\mathcal{T})$ converges in probability to zero by $W(\mathcal{T})/\mathcal{T}\to 0$, the consistency of the stationary component estimators from (B5) and Theorem 3, the Kronecker structure of the stationary log-likelihood, and the smoothness of the parameter curves as $t_0/\mathcal{T}\to u_0$ when $\mathcal{T}\to\infty$. With these results, we can invoke Dahlhaus (2012, Thm. 3.2) and so obtain $\widehat{\mathbf{A}}'(c,u_0),\widehat{\mathbf{B}}'(c,u_0),\widehat{\boldsymbol{\Phi}}'(c,u_0)$ and $\widehat{\boldsymbol{\theta}}'(u_0)$ converge in probability to $\mathbf{A}'(c,u_0),\mathbf{B}'(c,u_0),\boldsymbol{\Phi}'(c,u_0)$ and $\boldsymbol{\theta}'(u_0)$ respectively for all $c$ belonging to $\mathbf{c}'(u_0)$, as desired. □

## D.5 Tuning Parameter Selection

### Factor Structure Selection

Execution of Algorithm 3 for estimation of the factor loading curves requires selection of 3 tuning parameters: the number of common factors $r_0$, the number of vehicle-specific factors $r_1$, and the window size $W$. The use of the aforementioned $n = 4$ features under the MFA identifiability Condition 3 imposes $r_1 \leq 2$, so an exhaustive comparison is feasible. Simultaneous selection of $r_0$ and $W$ can be accomplished under the following scheme, which is reminiscent of Double Cross-Validation (DCV) for selection of the number of common factors in approximate factor models. Recent work by Zeng et al. (2019), has demonstrated that DCV can serve as a robust alternative to spectral methods of selecting the number of common factors.

To find optimal values for $r_0$ and $W$ while accounting for temporal dependence, a blocked cross-validation method is used. Windows of length $W + H$ are extracted, the stationary approximation (Algorithm 2) with $r_0$ common factors is fit to the first $W$ observations, then the fit quality is evaluated on the remaining $H$ held-out observations. This evaluation is performed by extracting the fitted observation values $\widehat{\mathbf{x}}(c_i, s)$ for vehicle $c_i$ on the held-out timesteps. To accomplish this, the estimated common factor loadings $\widehat{\mathbf{A}}(\mathbf{c})$ are obtained from the first $W$ observations, and the common factor $\widehat{\mathbf{f}}_{(i)}(s)$ on the held-out timesteps are estimated by regressing the observations for all other vehicles on their estimated loadings. The resulting $H$ fitted values $\widehat{\mathbf{x}}(c_i, s) = \widehat{\mathbf{A}}(c_i)\widehat{\mathbf{f}}_{(i)}(s)$ are then obtained without use of the observed values for vehicle $c_i$ on the held-out timesteps, so direct comparison of the fitted and actual values is valid.

This procedure is repeated for each vehicle in turn, and squared error is averaged over the number of vehicles, features, and timesteps to give the following $DCV$ criterion for the given window of $W + H$ timesteps:

$$DCV = \frac{1}{HCn} \sum_{i=1}^{C} \sum_{s=W+1}^{W+H} ||\mathbf{x}(c_i, s) - \overline{\mathbf{x}}(c_i) - \widehat{\mathbf{A}}(c_i)\widehat{\mathbf{f}}_{(i)}(s)||_F^2,$$

where $\overline{\mathbf{x}}(c_i)$ is the temporal average for vehicle $c_i$ on the held-out observations. A low value for $DCV$ implies that the estimated loadings $\widehat{\mathbf{A}}$ continue to capture the relevant between-vehicle covariance during the following $H$ timesteps, so that knowledge of the motion of the other vehicles allows for prediction of the motion of the held-out vehicle.

As it is necessary for the choice of $W$ and $r_0$ to serve for a variety of traffic conditions, $500$ ego vehicles and associated temporally-separated windows were subsampled from the full NGSIM data. The resulting median $DCV$ criterion for window lengths of $W = 30, 50, 70, 100$ and $r_0 = 1 \ldots 10$ are shown in Figure D.1. For all window lengths $W$, the held-out sample length was fixed at $H = 10$. The global minimum of the criterion occurs at $W = 50, r_0 = 2, r_1 = 1$, which will be used as the factor structure for the remainder of the results section.

## Podding Granularity

Selection of the factor structure parameters as detailed in the previous section enables the estimation of the time-varying loading curves $\mathbf{A}'(c,u), \mathbf{B}'(c,u), \mathbf{\Phi}'(c,u)$ for all vehicles within the cohort of interest. Converting these curves into a dynamic podding is accomplished by Algorithm 3, but this podding is dependent on the choice of penalty weight $\lambda$. The penalty weight $\lambda$ controls how different a vehicle must be from other vehicles before it is assigned its own pod, and so governs the granularity of the podding. The number of pods then adapts to the observed similarity of the vehicle trajectories.

The $\lambda$ range for which the estimated dynamic poddings are non-trivial is explored in Figures D.2a and D.2b. Figure D.2a shows the relationship between the out-of-sample non-penalized podding loss (4.18) and the number of pods, averaged across all timesteps for the cohort-window subsamples. For values of $\lambda < 40$, the podding loss rapidly decreases as the number of pods grows, indicating that the introduction of potentially-spurious pods. On the other hand, for values of $\lambda > 120$, the out-of-sample podding loss grows more rapidly. Hence for values of $\lambda$ between $40$ and $120$, interesting and potentially useful poddings structures at differing levels of granularity can be extracted.

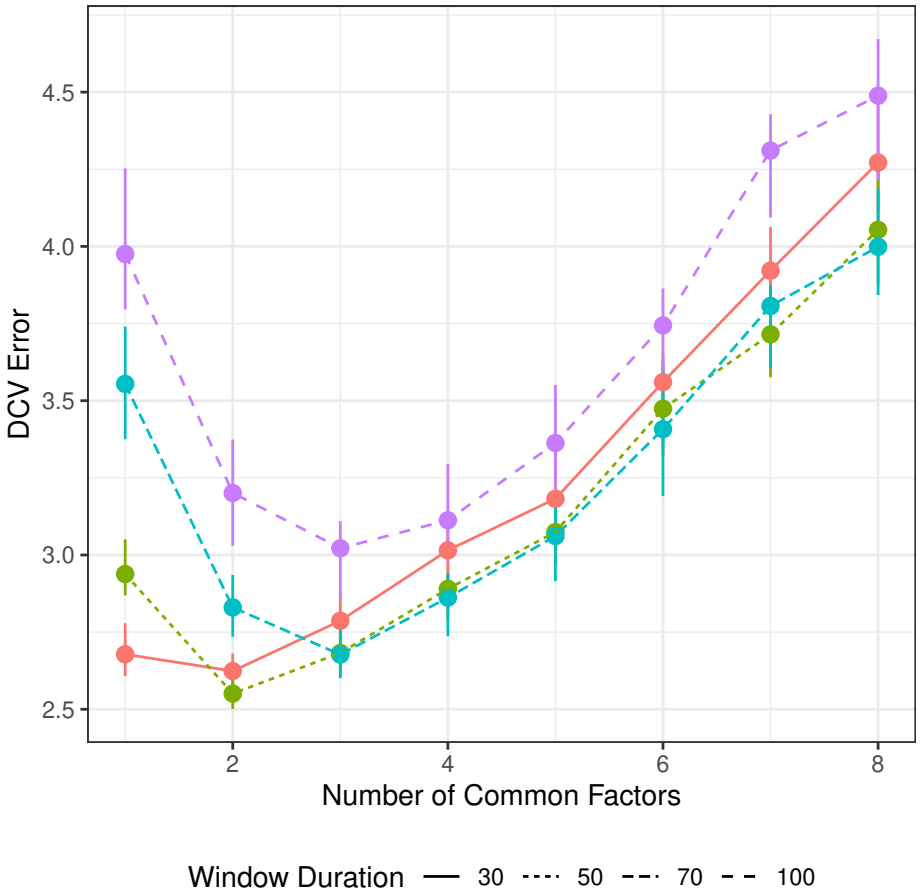
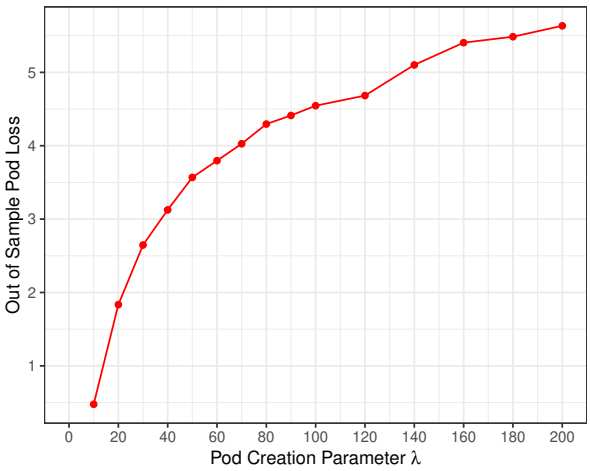
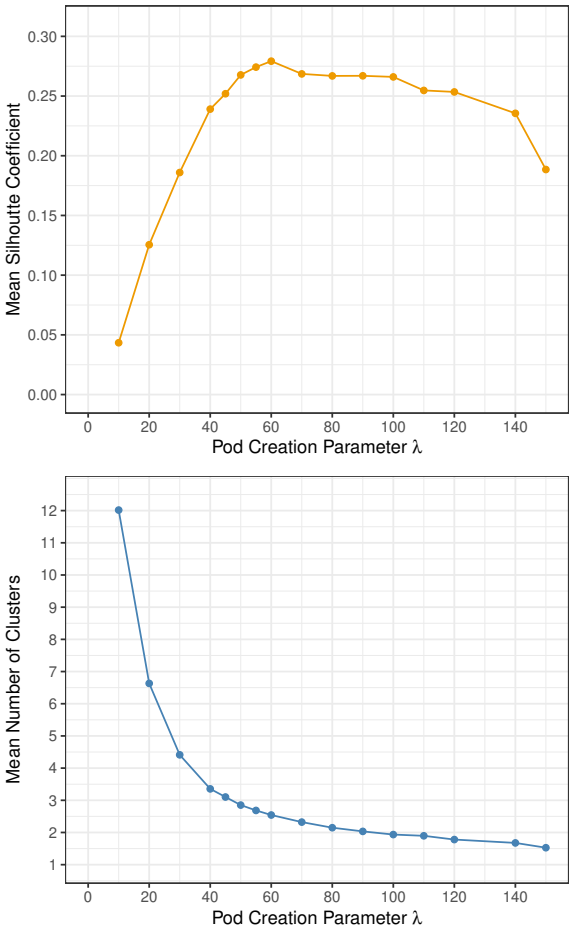



**Figure D.1:** An plot of the out-of-sample median fit quality for various number of common factors and window duration $W$. All models include one vehicle-specific factor per vehicle, $r_1 = 1$. The vertical axis plots the DCV squared error per vehicle. The minimum of this criterion occurs at $r_0 = 2$, $W = 50$. Error bars are $80\%$ bootstrap confidence intervals for the median, based on the subsampling method.

This range is further confirmed by examination of Figure D.2b, which shows the average Silhouette Coefficient (SC) for varying $\lambda$. The SC is a common measure for internal validation of a clustering assignment (Kaufman and Rousseeuw, 2009), based around the comparison of the average distance to observations in the same cluster to those in the next closest cluster. The discrepancy between vehicles used here that of Algorithm 3. As can be seen in Figure D.2b, the SC peaks at $\lambda = 60$, decays rapidly for $\lambda < 40$ and decays more slowly for $\lambda > 120$.




**(a)** Mean out-of-sample podding loss (4.19) evaluated on $H$ held-out timesteps.




**(b)** Top: Mean silhouette coefficient obtained for varying lambda. Bottom: Mean number of pods per window for varying lambda.

**Figure D.2:** Characteristics of resulting pods from Algorithm 3, for varying pod creation parameter $\lambda$.