Multi-Channel Factor Analysis: Identifiability and Asymptotics

Gray Stanton, *Student Member, IEEE,* David Ramírez, *Senior Member, IEEE,* Ignacio Santamaria, *Senior Member, IEEE,* Louis Scharf, *Life Fellow, IEEE,* and Haonan Wang

***Abstract*— Recent work by Ramírez et al. [2] has introduced Multi-Channel Factor Analysis (MFA) as an extension of factor analysis to multi-channel data that allows for latent factors common to all channels as well as factors specific to each channel. This paper validates the MFA covariance model and analyzes the statistical properties of the MFA estimators. In particular, a thorough investigation of model identifiability under varying latent factor structures is conducted, and sufficient conditions for generic global identifiability of MFA are obtained. The development of these identifiability conditions enables asymptotic analysis of estimators obtained by maximizing a Gaussian likelihood, which are shown to be consistent and asymptotically normal even under misspecification of the latent factor distribution.**

***Index Terms*—Asymptotic normality, consistency, factor analysis (FA), identifiability, multi-channel factor analysis (MFA)**

I. Introduction

Factor analysis (FA) is a statistical technique for modeling second-order structure within a collection of measurements. The method explains an observed vector through an unobserved systemic part (which is typically of scientific or engineering interest) and an unobserved noise part. Classical or *single-channel* factor analysis was originally developed within the field of psychometrics by Spearman [3] as a method to identify a small number of unobserved random *factors* which explain the between-individual variation in psychometric scores. In signal processing, FA and its extensions [4]–[6] are employed in the uncalibrated setting where the noise variance is anisotropic and unknown [7]–[9].

A recent extension to factor analysis by Ramírez et al. [2] is of central interest to this paper. These authors developed *multi-channel* factor analysis (MFA), which enables joint factor analysis of observations collected from *multiple channels*. Many problems and associated methods possess a natural channel structure [10]–[13], such as grouping multi-sensor data by sensor modality [14], [15]. In MFA, some factors, termed *common* factors, may influence all channels. In addition to common factors, in MFA each channel may also possess *distinct* factors influencing that channel alone. MFA then decomposes the vector of observations into latent vectors that can be described as a signal that influences all channels, within-channel interference, and idiosyncratic noise. This decomposition is of value, as detecting a weak signal that presents across all channels in the presence of channel-specific interference and noise is a goal in domains such as passive radar [16], [17], speech recognition [18], [19], and astronomy [20], [21].

Previous work on MFA has provided an estimation procedure for the parameters of the model based on likelihood-maximization under normality assumptions. However, for the output of the procedure to be meaningful, it is crucial that MFA be guaranteed to be identifiable using only what is known *a priori*, namely the channel sizes and the dimensions of the signal and interference subspaces. For single-channel factor analysis, practitioners correctly assume identifiability whenever the number of common factors is much smaller than the number of observations [22]. 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. In [2], the question of identifiability is recognized and some necessary conditions on the maximum number of common and distinct factors are discussed. This paper extends that discussion by carefully examining the two main sources of non-identifiability of MFA, namely isolation of the idiosyncratic noise variances and separation of signal and interference covariances.

The purpose of this paper is to give identifiability guarantees requiring only the specification of the channel sizes and signal and interference dimensionality. The asymptotic properties of the MFA estimators are then derived, which provides the previously-missing theoretical underpinnings for interpretation of the MFA parameter estimates. This parallels the advancement of single-channel FA as a statistical method as reviewed in Section III. The main contributions of this paper to MFA are

1) Necessary and sufficient conditions for separation of signal and interference covariances.
2) Sufficient conditions on the number of common and distinct factors for generic global identifiability.
3) Proof of the asymptotic consistency and normality of estimators derived from Algorithm 1 in [2].

The sufficient conditions for generic global identifiability of the MFA covariance model ensure that, for reasonable numbers of common and distinct factors, the decomposition of the observation covariance into common, distinct, and idiosyncratic

This paper is the journal version of [1].

Gray Stanton and Haonan Wang are with the Department of Statistics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: gstanton@colostate.edu; wanghn@stat.colostate.edu).

David Ramírez is with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid 28903, Spain, and also with Gregorio Marañón Health Research Institute, Madrid 28007, Spain (e-mail: david.ramirez@uc3m.es).

Ignacio Santamaria is with the Department of Communications Engineering, Universidad de Cantabria, 39005 Santander, Spain (e-mail: i.santamaria@unican.es).

Louis Scharf is with the Department of Mathematics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: louis.scharf@colostate.edu).

parts will be unique for almost all population covariance matrices. With this identifiability result, parameter estimates obtained by maximizing a Gaussian likelihood are shown to be consistent and asymptotically normal, even in the case where the true distribution of the latent vectors is non-normal.

A. 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 ith 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 $(i, j) \in \alpha\times\beta$. 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 operator 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 vec operator vectorizes a matrix by stacking the columns vertically, while 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 $\operatorname{Diag}_{\geq 0}(n)$ 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 $\mathcal{A} \oplus \mathcal{B}$. For a linear map $\mathbf{T}$, the image and kernel subspaces are $\operatorname{Im}(\mathbf{T})$ and $\operatorname{Ker}(\mathbf{T})$.

II. Model

A. 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 $\mathbf{x}_c$. The generative model for $\mathbf{x}_c$ is

$$\mathbf{x}_c = \mathbf{A}_c\mathbf{f} + \mathbf{B}_c\mathbf{g}_c + \mathbf{u}_c, \tag{1}$$

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 (1) is appropriate for several signal processing problems. In passive radar [23], the observations are collected from two multi-sensor arrays which make up the *reference* and *surveillance* channels. The common signal $\mathbf{f}$ affects both channels as $\mathbf{A}_1\mathbf{f}$ and $\mathbf{A}_2\mathbf{f}$ when a target is reflected by an opportunistic illuminator. As the multi-sensor arrays are spatially separated, the interferences can be modeled as the uncorrelated terms $\mathbf{B}_1\mathbf{g}_1$ and $\mathbf{B}_2\mathbf{g}_2$. Finally, the measurements are contaminated by uncorrelated noises whose variances are unknown in the absence of an accurate calibration. Another possible application of (1) is cooperative relaying in Time-Division Multiple Access (TDMA) systems [24], 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

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{B}\mathbf{g} + \mathbf{u}, \tag{2}$$

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

$$\mathbf{A} \equiv \begin{bmatrix}\mathbf{A}_1^\mathsf{T} & \cdots \mathbf{A}_C^\mathsf{T}\end{bmatrix}^\mathsf{T}, \quad \mathbf{B} \equiv \operatorname{blkdiag}(\mathbf{B}_1, \ldots, \mathbf{B}_C), \tag{3}$$

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, let $\mathbf{n} \equiv [n_1, \ldots, n_C]$, and $\mathbf{r} \equiv [r_0, r_1, \ldots, r_C]$. The total number of observations is $n \equiv \sum_{c=1}^C n_c$ and the total number of distinct factors is $r \equiv \sum_{c=1}^C r_c$. 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 I summarizes the commonly used notation.

B. Covariance Specification

In (2), the factors $\mathbf{f}, \mathbf{g}$ and the errors $\mathbf{u}$ are unobserved random quantities while the factor loadings $\mathbf{A}, \mathbf{B}$ are fixed unknown parameters. The latent factors are assumed to satisfy

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

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

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

for some covariance matrix $\mathbf{\Phi} \in \mathrm{Diag}_{\geq 0}(n)$. Under the above specification, $\mathbf{x}$ has mean zero with covariance matrix

$$\begin{aligned}\mathbf{R_{xx}} &\equiv \mathbf{A}\mathbf{R_{ff}}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{R_{gg}}\mathbf{B}^\mathsf{T} + \mathbf{\Phi}, \\ &= \mathbf{R_{ss}} + \mathbf{R_{ii}} + \mathbf{\Phi},\end{aligned} \tag{4}$$

where $\mathbf{R_{gg}} = \mathrm{blkdiag}(\mathbf{R}_{\mathbf{g}_1\mathbf{g}_1}, \ldots, \mathbf{R}_{\mathbf{g}_C\mathbf{g}_C})$ and the *signal* and *interference* covariances are $\mathbf{R_{ss}} \equiv \mathbf{A}\mathbf{R_{ff}}\mathbf{A}^\mathsf{T}$ and $\mathbf{R_{ii}} \equiv \mathbf{B}\mathbf{R_{gg}}\mathbf{B}^\mathsf{T}$, respectively. The set $\mathcal{R}(\mathbf{n},\mathbf{r}) \subset \mathrm{PSD}(n)$ contains all observation covariance matrices realizable by (4).

C. Parameterization of MFA Covariance Models

For given channel sizes $\mathbf{n}$ and factor numbers $\mathbf{r}$, the generative model (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 (4). The set $\mathcal{R}(\mathbf{n},\mathbf{r})$ can be parameterized in three ways, namely by the triple of structured components $(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$, by the loading matrices $\mathbf{A}, \mathbf{B}$ and noise variances $\mathbf{\Phi}$ whose structures are shown in Figure 1, or by a vector $\boldsymbol{\eta}$ which captures the degrees of freedom in the $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ parameterization.

1) Parameterization by $(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$: In MFA, (4) shows that the observation covariance $\mathbf{R_{xx}}$ is the sum of a low-rank matrix $\mathbf{R_{ss}}$, a channel-structured block-diagonal matrix $\mathbf{R_{ii}}$ with low-rank blocks, and a non-negative diagonal matrix $\mathbf{\Phi}$. Any triple $(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$ of appropriately structured $n \times n$ matrices determines an element of $\mathcal{R}(\mathbf{n},\mathbf{r})$ by the second line of (4). That is, if $\mathbf{R_{ss}} \in \mathrm{PSD}(n)$ has rank at most r_0, $\mathbf{R_{ii}} \in \mathrm{PSD}(n)$ is block-diagonal whose cth block is $n_c \times n_c$ with rank at most r_c, and $\mathbf{\Phi}$ is in $\mathrm{Diag}_{\geq 0}(n)$, then

$$\mathbf{R_{xx}}(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi}) \equiv \mathbf{R_{ss}} + \mathbf{R_{ii}} + \mathbf{\Phi} \tag{5}$$

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

Recovering $(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$ from an estimate of $\mathbf{R_{xx}}$ is the central goal of MFA, as decomposing $\mathbf{R_{xx}}$ into the three summands will separate $\mathbf{R_{ss}}$, 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 III, 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.

2) 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.

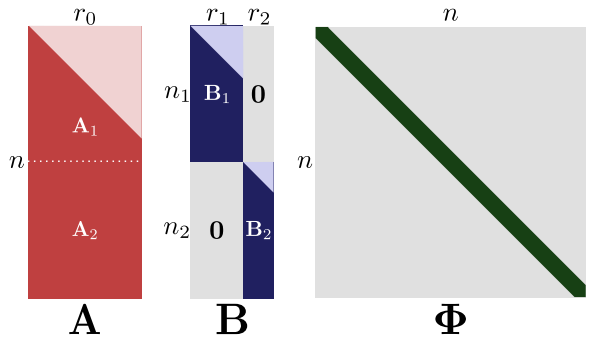

Figure 1: Depiction of MFA covariance parameters $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ for two channels. Triangles indicate constraints of $\mathbb{A}_L$ and $\mathbb{B}_L$.

Table I: Table of commonly used notation and descriptions.

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

The first line of (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 $\mathbf{R}_{\mathbf{g}_c\mathbf{g}_c} = \mathbf{I}_{r_c}$ for all $c = 1, \ldots, C$.

Under this assumption, $\mathcal{R}(\mathbf{n},\mathbf{r})$ can be parameterized as

$$\mathbf{R_{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi}) \equiv \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T} + \mathbf{\Phi}. \tag{6}$$

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 cth 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 $\mathbb{A} \times \mathbb{B} \times \mathrm{Diag}_{\geq 0}(n)$.

3) Parameterization by $\boldsymbol{\eta}$: The above parameterization $\mathbf{R_{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ by the loading matrices introduces a *rotation invariance*, as $\mathbf{R_{xx}}(\mathbf{A}\mathbf{Q}_f, \mathbf{B}\mathbf{Q}_g, \mathbf{\Phi})$ equals $\mathbf{R_{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ for any orthogonal $\mathbf{Q}_f$ and $\mathbf{Q}_g$, where $\mathbf{Q}_g = \mathrm{blkdiag}(\mathbf{Q}_{g,1}, \ldots, \mathbf{Q}_{g,C})$ 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 [25] [26].

Here, appropriate restrictions are imposed in the same fashion

as in [2]. 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 jth 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

$$\begin{aligned}\boldsymbol{\eta} = \big[&\operatorname{vech}(\mathbf{A}_1)^\mathsf{T} \ \operatorname{vec}(\mathbf{A}_2)^\mathsf{T} \ \operatorname{vech}(\mathbf{B}_{1,1})^\mathsf{T} \\ &\operatorname{vec}(\mathbf{B}_{1,2})^\mathsf{T} \ \ldots \ \operatorname{vec}(\mathbf{B}_{C,2})^\mathsf{T} \ \operatorname{Diag}^{-1}(\mathbf{\Phi})^\mathsf{T}\big]^\mathsf{T},\end{aligned} \tag{7}$$

where the dimension L is

$$L = nr_0 - \tfrac{1}{2}r_0(r_0-1) + \textstyle\sum_{c=1}^{C}\left[n_c r_c - \tfrac{1}{2}r_c(r_c-1)\right] + n. \tag{8}$$

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 (7) for $(\mathbf{A}(\boldsymbol{\eta}), \mathbf{B}(\boldsymbol{\eta}), \mathbf{\Phi}(\boldsymbol{\eta}))$ and taking $\mathbf{R}_{\mathbf{xx}}(\mathbf{A}(\boldsymbol{\eta}), \mathbf{B}(\boldsymbol{\eta}), \mathbf{\Phi}(\boldsymbol{\eta}))$.

III. Identifiability

1) Definition: For multi-channel factor analysis as defined in Section II, we say that an observation covariance matrix $\mathbf{\Sigma}_{\mathbf{xx}} \in \mathcal{R}(\mathbf{n}, \mathbf{r})$ is *identified* when it can be uniquely decomposed into a sum of appropriately structured components. That is, in the terms of Section II-C, $\mathbf{\Sigma}_{\mathbf{xx}}$ is identified if there is a unique triple $(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi})$ such that $\mathbf{R}_{\mathbf{xx}}(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi}) = \mathbf{\Sigma}_{\mathbf{xx}}$. As the covariance matrices $\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}$, 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 $(\tilde{\mathbf{R}}_{\mathbf{ss}}, \tilde{\mathbf{R}}_{\mathbf{ii}}, \tilde{\mathbf{\Phi}})$ of structured matrices which also sums to $\mathbf{\Sigma}_{\mathbf{xx}}$, then any inference based on the individual values for the summands must be suspect. As $\mathbf{\Sigma}_{\mathbf{xx}}$ is not known, practical application of MFA requires that the observation covariance model (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 $\mathbf{\Sigma}_{\mathbf{xx}}$ and refers to the uniqueness of the second order decomposition (6), *not* uniqueness of the first order generative model (1). As discussed in Section (II-C2), 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 $\mathbf{\Sigma}_{\mathbf{xx}}$ 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 [2, Section IV.B] for a related experiment).

Identification of $(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi})$ from $\mathbf{\Sigma}_{\mathbf{xx}}$ breaks into two subproblems, namely *isolation of the idiosyncratic variances* and *separation of the signal and interference covariances*.

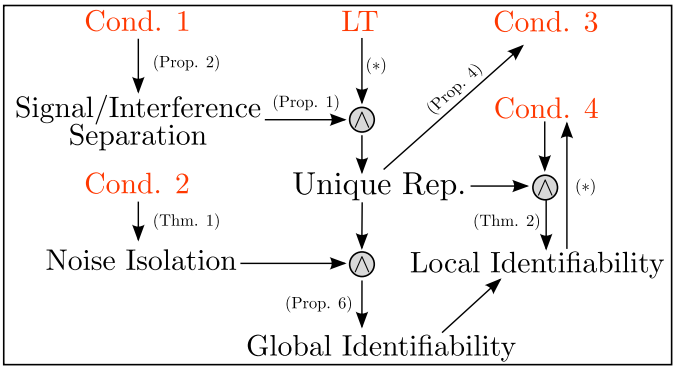

Figure 2: Diagram indicating the relationships between the conditions imposed on the channel sizes and factor numbers and their implications for the different aspects of MFA identification. Directions marked with asterisks were obtained in [2]. "LT" refers to the lower-triangular structure in $\mathbb{A}_L^*$ and $\mathbb{B}_L^*$

That is, the former subproblem refers to whether $\mathbf{\Sigma}_{\mathbf{xx}}$ uniquely determines $(\mathbf{R}_{\mathbf{ss}} + \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi})$ while the latter subproblem refers to whether $\mathbf{R}_{\mathbf{ss}} + \mathbf{R}_{\mathbf{ii}}$ uniquely determines $(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}})$.

In Section III-A, conditions on $\mathbf{n}$ and $\mathbf{r}$ which resolve the subproblems and ensure the identifiability of $(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi})$ from $\mathbf{\Sigma}_{\mathbf{xx}}$ are derived. The following Section III-B investigates the identifiability and associated properties of the parameterization $\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ of $\mathcal{R}(\mathbf{n}, \mathbf{r})$ in terms of $\boldsymbol{\eta}$, which provides the required technical foundation for the asymptotic analysis of Section IV. 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 $\mathbf{\Sigma}_{\mathbf{xx}}$ 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 $\mathbf{R}_{\mathbf{ss}} + \mathbf{R}_{\mathbf{ii}}$ 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 $(\mathbf{R}_{\mathbf{ss}}, \mathbf{R}_{\mathbf{ii}}, \mathbf{\Phi})$ summing to $\mathbf{\Sigma}_{\mathbf{xx}}$ 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; [22] and [27] establish local and global identifiability for single-channel factor analysis in this *generic* sense, while [28] 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.

3) 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 [29] provided a heuristic for

the maximum number of factors (known as the Ledermann bound). Anderson [25] set 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 [22] 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 [27].

In this paper, analogous results for multi-channel factor analysis are obtained. The discussion on the identifiability of MFA was opened by [2, Sec. III] and the importance of the problem was recognized. In particular, the authors provided the restrictions which remove rotation invariance used in Proposition 3, 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 in [2], this conjecture is verified for *local* identifiability. For *global* identifiability, we instead require the slightly stronger Condition 2.

4) Identifiability in Related Multi-Channel Methods: Just as classical FA has deep connections with other multivariate statistical methods such as Canonical Correlation Analysis (CCA) [30], MFA can be related to other techniques for multi-channel data analysis. In particular, CCA (both in classical two-view form [31] and in the generalized multi-view form [32]) 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) [33], Shared Independent Component Analysis (ShICA) [34] and Deep CCA [35] 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 [32], Deep CCA [36], JISA [37], and ShICA [34] have been obtained through a variety of approaches.

However, the MFA-specific identifiability results obtained in this paper are not direct consequences of previous results, and differ in two ways. First, Proposition 2 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.

A. Identifiability of $\mathbf{R_{xx}}(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$

1) Separation of Signal and Interference Covariances: The first subproblem of MFA identifiability involves the noise-free part of (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 relate back to $\mathbf{R_{xx}}(\mathbf{R_{ss}}, \mathbf{R_{ii}}, \mathbf{\Phi})$.

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

$$(\mathbf{A}, \mathbf{B}) \sim_1 (\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \iff \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T} + \tilde{\mathbf{B}}\tilde{\mathbf{B}}^\mathsf{T}$$
$$(\mathbf{A}, \mathbf{B}) \sim_2 (\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \iff \mathbf{A}\mathbf{A}^\mathsf{T} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T} \text{ and } \mathbf{B}\mathbf{B}^\mathsf{T} = \tilde{\mathbf{B}}\tilde{\mathbf{B}}^\mathsf{T}.$$

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 associated with $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ 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., [38, Lemma 5.1]) shows that for pairs $(\mathbf{A}, \mathbf{B}), (\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \mathbb{A} \times \mathbb{B}$ with $(\mathbf{A}, \mathbf{B}) \sim_1 (\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ we must have

$$\begin{bmatrix}\tilde{\mathbf{A}} & \tilde{\mathbf{B}}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{B}\end{bmatrix}\mathbf{Q},$$

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 $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]$ represent the same map under different orthonormal bases for the combined factor space of both common and distinct factors. Similarly, if $(\mathbf{A}, \mathbf{B}) \sim_2 (\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ then $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{Q}_{00}$ for $\mathbf{Q}_{00} \in \mathrm{O}(r_0)$ and $\tilde{\mathbf{B}}_c = \mathbf{B}_c\mathbf{Q}_{cc}$ with $\mathbf{Q}_{cc} \in \mathrm{O}(r_c)$ for each $c = 1, \ldots, C$. Partitioning $\mathbf{Q}$ as,

$$\mathbf{Q} = \begin{array}{c} \begin{matrix} r_0 & r_1 \cdots\cdots\cdots r_C \end{matrix} \\ \begin{bmatrix} \mathbf{Q}_{00} & \mathbf{Q}_{01} & \cdots & \mathbf{Q}_{0C} \\ \mathbf{Q}_{10} & \mathbf{Q}_{11} & \cdots & \mathbf{Q}_{1C} \\ \vdots & \vdots & & \vdots \\ \mathbf{Q}_{C0} & \mathbf{Q}_{C1} & \cdots & \mathbf{Q}_{CC} \end{bmatrix} \begin{matrix} r_0 \\ r_1 \\ \vdots \\ r_C \end{matrix} \end{array} \tag{9}$$

then $(\mathbf{A}, \mathbf{B}) \sim_2 (\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ iff $[\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]$ 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 $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \mathbb{A} \times \mathbb{B}$. 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 the Supplementary Materials.

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 1. *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*

$$\mathbf{M}_c = \begin{bmatrix} \mathbf{A}_{<c} & \mathbf{B}_{<c} \\ \mathbf{A}_{>c} & \mathbf{0} \end{bmatrix},\ \mathbf{M}_1 = [\mathbf{A}_2^\mathsf{T}\ \ldots\ \mathbf{A}_C^\mathsf{T}]^\mathsf{T} \tag{10}$$

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* $\mathbf{B}_{<c} = \text{blkdiag}(\mathbf{B}_1, \ldots, \mathbf{B}_{c-1})$. *Then any* $\mathbf{Q} \in \text{O}(r_0 + r)$ *such that* $[\mathbf{A}\ \mathbf{B}]\mathbf{Q} = [\tilde{\mathbf{A}}\ \tilde{\mathbf{B}}]$ *for some* $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \mathbb{A} \times \mathbb{B}$ *must have* $\mathbf{Q}_{ij} = \mathbf{0}$ *for all* $i \neq j$ *when partitioned as* (9)*.*

Proof: See Supplementary Materials 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 1, 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 $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \sim_2 (\mathbf{A}, \mathbf{B})$ 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 1 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*

$$r_0 + r_{<c} \leq n - n_c, \tag{11}$$

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 (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 2. (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 1 is null.*

Proof: See Supplementary Materials for proof. ∎

2) Isolation of Noise Variances: For single-channel FA identifiability, the main criterion is ϕ 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 [27] is then $\phi(n, r, 2r - n) > 0$. For MFA, the analogous criterion function $\psi(\mathbf{n}, \mathbf{r}, \boldsymbol{\rho})$ is

$$\psi(\mathbf{n}, \mathbf{r}, \boldsymbol{\rho}) = n + \phi(n, r_0, \rho_0) + \sum_{c=1}^{C} \phi(n_c, r_c, \rho_c) + r_c(r_0 - \rho_0), \tag{12}$$

with non-negative integer vector $\boldsymbol{\rho} = [\rho_0, \rho_1, \ldots, \rho_C]$. The criterion ψ 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* ψ^* *be the smallest criterion value over possible MFA reductions,*

$$\psi^* = \min_{(\mathbf{n}', \mathbf{r}', \boldsymbol{\rho}) \in M} \psi(\mathbf{n}', \mathbf{r}', \boldsymbol{\rho}) \tag{13}$$

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

$$\begin{aligned} &n'_c \leq n_c,\ c = 1, \ldots, C, \\ &r'_c = (r_c - (n_c - n'_c))_+,\ c = 1, \ldots, C, \\ &r'_0 = [r_0 - \textstyle\sum_{c=1}^{C}(n_c - n'_c - r_c)_+]_+, \\ &\rho_c \leq \min\{r'_c, 2(r'_0 + r'_c) - n'_c\},\ c = 1, \ldots, C, \\ &\rho_0 = \min\{r_0, 2r'_0 + \textstyle\sum_{c=1}^{C} 2r'_c - \rho_c - n'_c\}, \end{aligned} \tag{14}$$

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

Similarly to [27], $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi})$ is said to have globally identified noise variances if $\mathbf{R_{xx}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi}) = \mathbf{R_{xx}}(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\boldsymbol{\Phi}})$ implies that $\boldsymbol{\Phi} = \tilde{\boldsymbol{\Phi}}$. 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* $\mathbb{A} \times \mathbb{B} \times \text{Diag}_{\geq 0}(n)$.

Proof: See Appendix 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 $\mathbf{R_{ss}}+\mathbf{R_{ii}}+\boldsymbol{\Phi}$ permits a second representation as $\widetilde{\mathbf{R}}_{\mathbf{ss}}+\widetilde{\mathbf{R}}_{\mathbf{ii}}+\widetilde{\boldsymbol{\Phi}}$ with all noise variances $[\widetilde{\boldsymbol{\Phi}}]_{ii}$ not equaling $[\boldsymbol{\Phi}]_{ii},\ i = 1, \ldots, n$. 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 ψ can be seen (31) as the effective number of constraints imposed on $\mathbf{H}$ minus the degrees of freedom in choosing the diagonal difference matrix. If the criterion ψ 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 $[\widetilde{\boldsymbol{\Phi}}]_{ii} = [\boldsymbol{\Phi}]_{ii}$ for i in some index set β. To resolve the second case, the fact that the difference $[\widetilde{\boldsymbol{\Phi}}]_{ii} - [\boldsymbol{\Phi}]_{ii} = 0$ for indices in β 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 $c = 1, \dots, C$, where the possible reduced channel sizes and factor numbers (for varying index sets β) are set out in Condition 2. If ψ is positive for all possible reduced $\mathbf{n}', \mathbf{r}'$ and the associated possible block ranks $\boldsymbol{\rho}$, then $\boldsymbol{\Phi}$ can be generically be isolated from $\mathbf{R_{ss}} + \mathbf{R_{ii}}$.

B. Identifiability of $\mathbf{R_{xx}}(\boldsymbol{\eta})$

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 $\mathbf{R_{xx}}(\boldsymbol{\eta})$. These results are of technical relevance for the analysis of Section IV as they allow standard parameter estimation theory to be applied.

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

Proof: See Supplementary Materials for proof. ∎

This result parallels the use of LT restrictions to select a unique representative loading matrix in single-channe FA. However, for MFA, a previously unrecognized complication occurs when $\mathbf{R_{ss}} + \mathbf{R_{ii}}$ 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 4 applies a result for confirmatory factor analysis [39] 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*

$$r_0 r + \textstyle\sum_{c=1}^{C} r_c r_{<c} \leq \sum_{c=1}^{C} (n - n_c) r_c. \tag{15}$$

Proposition 4. *If almost all $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \mathbb{A}_L^* \times \mathbb{B}_L^*$ 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 $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ are the unique representative in $\mathbb{A}_L^* \times \mathbb{B}_L^*$ of their $\sim_1$-equivalence class.*

Proof: See Supplementary Materials for proof. ∎

In connection to $\mathbf{R_{xx}}(\boldsymbol{\eta})$, if $\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}} \in V$ are obtained by (7) applied to $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi}_0)$ and $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \boldsymbol{\Phi}_0)$ respectively, for $(\mathbf{A}, \mathbf{B})$ and $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ in $\mathbb{A}_L^* \times \mathbb{B}_L^*$, then $\mathbf{R_{xx}}(\boldsymbol{\eta}) = \mathbf{R_{xx}}(\tilde{\boldsymbol{\eta}})$ implies $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$ 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.

2) Local Identifiability: For fixed $\boldsymbol{\Sigma_{xx}} \in \mathcal{R}(\mathbf{n}, \mathbf{r})$, the equation $\mathbf{R_{xx}}(\boldsymbol{\eta}) = \boldsymbol{\Sigma_{xx}}$ defines a quadratic system of equations in the entries of $\boldsymbol{\eta}$. As this system is nonlinear, simply counting the number of knowns in $\boldsymbol{\Sigma_{xx}}$ 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 $d\mathbf{R_{xx}}(\boldsymbol{\eta})$ allows investigation of *local identification*. Here, local identification at $\boldsymbol{\eta}$ means that there is a neighborhood of $\boldsymbol{\eta}$ on which $\mathbf{R_{xx}}(\boldsymbol{\eta})$ 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 $\mathbf{R_{xx}}(\boldsymbol{\eta})$ is a smooth map, local identification at $\boldsymbol{\eta}$ follows by the inverse function theorem if $d\mathbf{R_{xx}}(\boldsymbol{\eta})$ is injective. To assess this, a key condition follows from the tabulation of knowns and unknowns in MFA with channel sizes $\mathbf{n}$ and factor numbers $\mathbf{r}$, as discussed in [2, Sec. III].

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

$$r_0 \leq \tfrac{1}{2}\left(2n + 1 - \sqrt{8(n + D) + 1}\right), \tag{16}$$

where

$$D = \textstyle\sum_{c=1}^{C} n_c r_c - \frac{1}{2} r_c (r_c - 1) \tag{17}$$

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

$$r_c \leq \tfrac{1}{2}\big(2n_c + 1 - \sqrt{8n_c + 1}\big). \tag{18}$$

The following proposition, which was proven in [2], shows that Condition 4 is *necessary* for $d\mathbf{R_{xx}}$ to be injective.

Proposition 5. *It is* necessary *that the channel sizes* $\mathbf{n}$ *and factor numbers* $\mathbf{r}$ *satisfy Condition 4 for $d\mathbf{R_{xx}}(\boldsymbol{\eta})$ to be injective at any $\boldsymbol{\eta} \in V$.*

However, ensuring that $d\mathbf{R_{xx}}(\boldsymbol{\eta})$ 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 2, 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 $d\mathbf{R_{xx}}(\boldsymbol{\eta})$ is generically injective.*

Proof: See Appendix for proof. ∎

3) Global Identifiability: Although the local identifiability result of Theorem 2 provides valuable information about the behavior of $\mathbf{R_{xx}}(\boldsymbol{\eta})$ on small neighborhoods and will be needed in Section IV, a stronger global identifiability result for $\mathbf{R_{xx}}(\boldsymbol{\eta})$ is desired. The following proposition combines the results of Section III-B1 with Proposition 2 and Theorem 1 to show that $\mathbf{R_{xx}}(\boldsymbol{\eta})$ is an invertible map, excepting a null set of $\boldsymbol{\eta}$.

Proposition 6. (Global Identifiability) *If the channel sizes* $\mathbf{n}$ *and factor numbers* $\mathbf{r}$ *satisfy Conditions 1 and 2, then there exists a subset $\tilde{V} \subset V$ such that $\mathbf{R_{xx}}(\boldsymbol{\eta})$ is injective on $\tilde{V}$ and $V \setminus \tilde{V}$ is null.*

Proof: See Supplementary Materials for proof. ∎

IV. Asymptotics

A. Estimation

Suppose T observation vectors $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are available and are i.i.d. with covariance $\mathbf{\Sigma_{xx}} = \mathbf{R_{xx}}(\mathring{\boldsymbol{\eta}})$. In this setting, [2] 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 $\mathbf{x}$ is

$$\log f(\mathbf{x}; \boldsymbol{\eta}) = -\frac{1}{2}\log\det \mathbf{R_{xx}}(\boldsymbol{\eta}) - \frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})\mathbf{x} + K. \tag{19}$$

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

$$\min_{\boldsymbol{\eta}\in V} \ \log\det \mathbf{R_{xx}}(\boldsymbol{\eta}) + \operatorname{tr} \mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})\mathbf{S}_T \tag{20}$$

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 ℓ_T is

$$\ell_T(\mathbf{S}_T; \boldsymbol{\eta}) \equiv \log\det \mathbf{R_{xx}}(\boldsymbol{\eta}) + \operatorname{tr} \mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})\mathbf{S}_T. \tag{21}$$

To avoid the Heywood cases [40], we will restrict attention to $(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ such that $\min_i[\mathbf{\Phi}]_{ii} \geq \epsilon$ for some fixed $\epsilon > 0$. This has the advantage of ensuring that the smallest eigenvalue $\mathbf{R_{xx}}(\mathbf{A}, \mathbf{B}, \mathbf{\Phi})$ is bounded away from zero. Letting $V' \subset V$ contain all $\boldsymbol{\eta}$ which satisfy this additional requirement, we define the estimators $\hat{\mathbf{A}}_T$, $\hat{\mathbf{B}}_T$, $\hat{\mathbf{\Phi}}_T$ as those obtained from the minimizer of $\ell_T(\mathbf{S}_T; \boldsymbol{\eta})$,

$$\hat{\boldsymbol{\eta}}_T = \operatorname*{argmin}_{\boldsymbol{\eta}\in V'} \ell_T(\mathbf{S}_T; \boldsymbol{\eta}). \tag{22}$$

The estimators for the MFA parameters, $\hat{\mathbf{A}}_T$, $\hat{\mathbf{B}}_T$, $\hat{\mathbf{\Phi}}_T$, are obtained by inverting (7) for $\hat{\boldsymbol{\eta}}_T$.

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 IV-B are obtained by treating (21) as a *quasi-loglikelihood* [41] objective function to be optimized, rather than requiring that (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.

B. 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 (6),

$$\operatorname{Var}(\mathbf{x}_1) = \mathbf{\Sigma_{xx}} \equiv \mathbf{R_{xx}}(\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\mathbf{\Phi}}). \tag{23}$$

The true values $\mathring{\mathbf{A}}$ and $\mathring{\mathbf{B}} \equiv \operatorname{blkdiag}(\mathring{\mathbf{B}}_1, \ldots \mathring{\mathbf{B}}_C)$ are such that $(\mathring{\mathbf{A}}, \mathring{\mathbf{B}}) \in \mathbb{A}_L^* \times \mathbb{B}_L^*$ and $[\mathring{\mathbf{\Phi}}]_{ii} > \epsilon$ for all $i = 1, \ldots, n$. The vectorization (7) of $(\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\mathbf{\Phi}})$ 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 $\mathbf{\Sigma_{xx}} \notin \mathcal{R}(\mathbf{n}, \mathbf{r})$.

The next theorem shows that identification of the true $\mathbf{\Sigma_{xx}}$ is enough to ensure that the estimators are consistent for the factor loading parameters $\mathbf{A}$ and $\mathbf{B}$ and the idiosyncratic noise variance $\mathbf{\Phi}$. 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 ℓ_T is then sufficiently well-behaved so that the maximizer $\hat{\boldsymbol{\eta}}_T$ of ℓ_T converges to the maximizer of $\ell_0 \equiv E[\ell_T]$, which is $\mathring{\boldsymbol{\eta}}$. Convergence of $\hat{\mathbf{A}}_T, \hat{\mathbf{B}}_T, \hat{\mathbf{\Phi}}_T$ to $\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\mathbf{\Phi}}$ 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 $\operatorname{Var}(\mathbf{x}_1) \equiv \mathbf{\Sigma_{xx}}$. If there exists a unique $\mathring{\mathbf{R}}_{\mathbf{xx}} \in \mathcal{R}(\mathbf{n}, \mathbf{r})$ minimizing $D_{KL}\big(\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_{xx}}) || \ \mathcal{N}(\mathbf{0}, \mathring{\mathbf{R}}_{\mathbf{xx}})\big)$ with $\mathring{\mathbf{R}}_{\mathbf{xx}} = \mathbf{R_{xx}}(\mathring{\boldsymbol{\eta}})$ for $\mathring{\boldsymbol{\eta}} \in V'$ in the interior of the globally identified set $\tilde{V}$ defined in Proposition 6, then $\hat{\mathbf{A}}_T, \hat{\mathbf{B}}_T, \hat{\mathbf{\Phi}}_T$ converge in probability to $\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\mathbf{\Phi}}$ respectively.*

Proof: See Appendix for proof. ∎

In particular, if the model is correctly specified with $\operatorname{Var}(\mathbf{x}_1) = \mathbf{R_{xx}}(\mathring{\boldsymbol{\eta}})$ for $\mathring{\boldsymbol{\eta}}$ in the interior of the globally identified set, then the estimators $\hat{\mathbf{A}}_T, \hat{\mathbf{B}}_T, \hat{\mathbf{\Phi}}_T$ 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 ℓ_T in a neighborhood of the covariance $\mathbf{R_{xx}}(\mathring{\boldsymbol{\eta}})$. 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, $\hat{\boldsymbol{\eta}}_T - \mathring{\boldsymbol{\eta}}$ 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 $\mathbf{\Sigma_{xx}} = \mathbf{R_{xx}}(\mathring{\boldsymbol{\eta}})$. In addition, assume that $\mathbf{x}_1$ satisfies $E[||\mathbf{x}_1||^4] < \infty$. Under these conditions, the estimated parameters $\hat{\boldsymbol{\eta}}_T$ converges in distribution as*

$$\sqrt{T}(\hat{\boldsymbol{\eta}}_T - \mathring{\boldsymbol{\eta}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}_L, \mathbf{W}) \tag{24}$$

for positive-definite matrix $\mathbf{W}$ in the Appendix, (43).

Proof: See Appendix for proof. ∎

V. Experiments

A. Numeric Comparison of Conditions

In Section III, 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 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 6 which are respectively *sufficient* for generic local and global identifiability. To interpret Figure 3, examine channel size $n_1 = 15$ in the middle panel with $r_c = 5$. In this case, Proposition 5 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

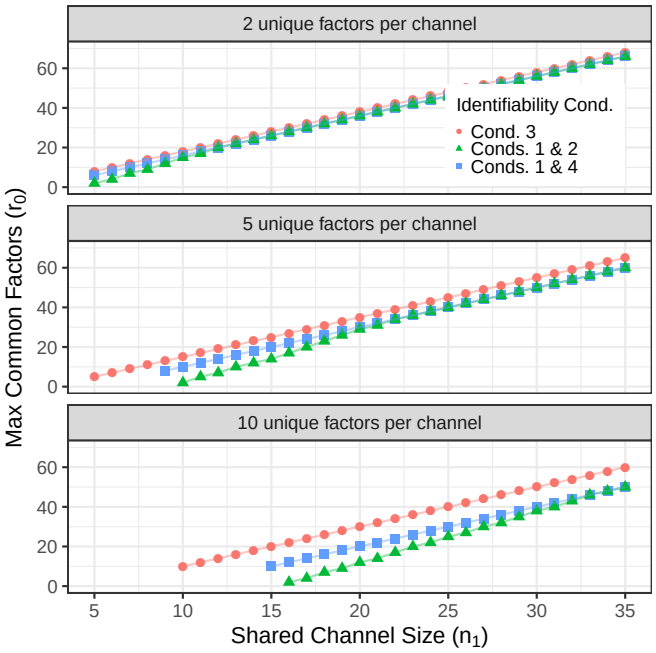

Figure 3: Comparison of maximum common factor number r_0 under three identifiability conditions, for varying channel sizes. Channel structure depicted is three equally-sized channels, with $r_c = 2, 5, 10$ distinct factors for $c = 1, 2, 3$.

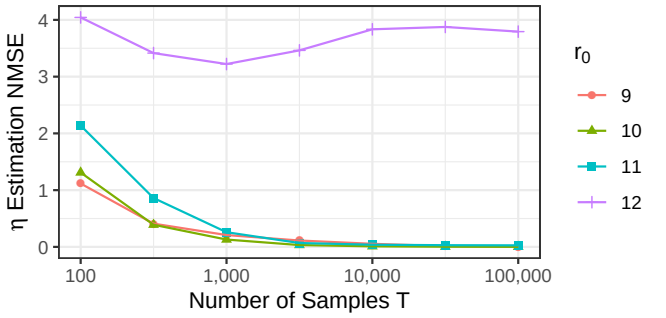

Figure 4: Experimental validation of Theorem 3 for varying common factor number r_0, with $C = 3$ channels of size $n_c = 8$ and $r_c = 2$ distinct factors, $c = 1, 2, 3$. Points indicate average NMSE from 1000 Monte Carlo trials at each setting.

which guarantees local identifiability under Theorem 2 as shown by the square at $n_1 = 15$ and $r_0 = 19$. Finally, 14 is the largest r_0 which yields global identifiability under Proposition 6, 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 6 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 [1].

B. Asymptotic Behavior of Estimators

To verify the consistency of $\hat{\boldsymbol{\eta}}_T$ resulting from Theorem 3, Figure 4 shows the Normalized Mean Square Error (NMSE) $||\hat{\boldsymbol{\eta}}_T - \mathring{\boldsymbol{\eta}}||^2 / ||\mathring{\boldsymbol{\eta}}||^2$ 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 [2] are independently obtained in the same fashion, and $\hat{\boldsymbol{\eta}}_T$ is computed from T independent samples with covariance $\boldsymbol{\Sigma}_{\mathbf{xx}}$.

For $C = 3$ channels with $n_c = 8$, Figure 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 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 6.

VI. Discussion

This paper 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 III. 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 [2] is obtained by likelihood maximization under the assumption of normality for the latent vectors, the results of Section IV demonstrate that violation of this assumption does not affect the asymptotic validity of the resulting estimators.

When introducing multi-channel factor analysis, [2] 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. The identifiability results of this paper 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. In single-channel FA, techniques such as maximization of an information criterion [42], bi-cross-validation [43], or eigenvalue analysis [44] can be used to estimate the number of

common factors. Adapting these techniques for order selection in MFA is an important direction for future work.

ACKNOWLEDGMENT

The authors thank the reviewers for their constructive comments. This work was supported in part by National Science Foundation grants DMS-1923142, CNS-1932413, and DMS-2123761. The work of I. Santamaria was funded by AEI /10.13039/501100011033 and FEDER UE under grant PID2022-137099NB-C43 (MADDIE). The work of D. Ramírez was partially supported by MICIU/AEI/10.13039/501100011033/FEDER, UE, under grant PID2021-123182OB-I00 (EPiCENTER), by the Office of Naval Research (ONR) Global under contract N62909-23-1-2002, and by the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D SORUS project.

REFERENCES

[1] G. Stanton, D. Ramírez, I. Santamaria, L. L. Scharf, and H. Wang, "Identifiability in multi-channel factor analysis," in *Asilomar Conference on Signals, Systems, and Computers*, pp. 1344–1349, 2023.

[2] D. Ramírez, I. Santamaria, L. L. Scharf, and S. Van Vaerenbergh, "Multi-channel factor analysis with common and unique factors," *IEEE Transactions on Signal Processing*, vol. 68, pp. 113–126, 2020.

[3] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904. Place: US Publisher: Univ of Illinois Press.

[4] A. Klami, S. Virtanen, E. Leppäaho, and S. Kaski, "Group factor analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 2136–2147, 2015.

[5] A. M. Sardarabadi and A.-J. van der Veen, "Complex factor analysis and extensions," *IEEE Transactions on Signal Processing*, vol. 66, no. 4, pp. 954–967, 2018.

[6] D. Ramírez, I. Santamaría, and L. L. Scharf, *Coherence in Signal Processing and Machine Learning*. Springer Cham, 2023.

[7] M. Pesavento and A. Gershman, "Maximum-likelihood direction-of-arrival estimation in the presence of unknown nonuniform noise," *IEEE Transactions on Signal Processing*, vol. 49, no. 7, pp. 1310–1324, 2001.

[8] D. Ramírez, G. Vazquez-Vilar, R. Lopez-Valcarce, J. Via, and I. Santamaria, "Detection of rank-P signals in cognitive radio networks with uncalibrated multiple antennas," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3764–3774, 2011.

[9] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Estimation of sensor array signal model parameters using factor analysis," in *27th European Signal Processing Conference (EUSIPCO)*, 2019.

[10] D. Cochran and H. Gish, "Multiple-channel detection using generalized coherence," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 2883–2886 vol.5, 1990.

[11] D. Cochran, H. Gish, and D. Sinno, "A geometric approach to multiple-channel signal detection," *IEEE Transactions on Signal Processing*, vol. 43, no. 9, pp. 2049–2057, 1995.

[12] D. Ramírez, J. Via, I. Santamaría, and L. L. Scharf, "Detection of spatially correlated Gaussian time series," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5006–5015, 2010.

[13] I. Santamaria, L. L. Scharf, J. Via, H. Wang, and Y. Wang, "Passive detection of correlated subspace signals in two MIMO channels," *IEEE Transactions on Signal Processing*, vol. 65, no. 20, pp. 5266–5280, 2017.

[14] N. M. Correa, T. Adalı, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE Signal Processing Magazine*, vol. 27, pp. 39–50, 2010.

[15] S. Bhinge, Q. Long, Y. Levin-Schwartz, Z. Boukouvalas, V. D. Calhoun, and T. Adalı, "Non-orthogonal constrained independent vector analysis: Application to data fusion," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2666–2670, 2017.

[16] F. Bandiera, O. Besson, D. Orlando, G. Ricci, and L. L. Scharf, "GLRT-based direction detectors in homogeneous noise and subspace interference," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2386–2394, 2007.

[17] D. E. Hack, L. K. Patton, B. Himed, and M. A. Saville, "Detection in passive MIMO radar networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2999–3012, 2014.

[18] E. Zwyssig, M. Ravanelli, P. Svaizer, and M. Omologo, "A multi-channel corpus for distant-speech interaction in presence of known interferences," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4480–4484, 2015.

[19] W. Zhang, B. Lin, L. Ma, A. Zhou, and G. Wu, "Temporal-frequency-spatial features fusion for multi-channel informed target speech separation," in *5th International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 168–174, 2022.

[20] A. Leshem and A.-J. van der Veen, "Multichannel detection and spatial signature estimation with uncalibrated receivers," in *IEEE Signal Processing Workshop on Statistical Signal Processing*, pp. 190–193, 2001.

[21] A. Antman and A. Leshem, "Radio transient detection in radio astronomical arrays," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5648–5663, 2020.

[22] A. Shapiro, "Identifiability of factor analysis: Some results and open problems," *Linear Algebra and its Applications*, vol. 70, pp. 1–7, 1985.

[23] D. E. Hack, L. K. Patton, B. Himed, and M. A. Saville, "Detection in passive MIMO radar networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2999–3012, 2014.

[24] M. Dohler and Y. Li, *Cooperative Communications: Hardware, Channel & PHY*. John Wiley & Sons, Ltd, 2010.

[25] T. W. Anderson and H. Rubin, "Statistical inference in factor analysis," *University of California Press*, pp. 111–150, 1956.

[26] K. Jöreskog, "A general approach to confirmatory factor analysis," *Psychometrika*, vol. 34, pp. 183–202, 1969.

[27] P. A. Bekker and J. M. ten Berge, "Generic global identification in factor analysis," *Linear Algebra and its Applications*, vol. 264, pp. 255–263, 1997. Sixth Special Issue on Linear Algebra and Statistics.

[28] F. Király and R. Tomioka, "A combinatorial algebraic approach for the identifiability of low-rank matrix completion," *29th International Conference on Machine Learning (ICML)*, 2012.

[29] W. Ledermann, "On the rank of the reduced correlational matrix in multiple-factor analysis," *Psychometrika*, vol. 2, pp. 85–93, 1937.

[30] C. R. Rao, "Estimation and tests of significance in factor analysis," *Psychometrika*, vol. 20, pp. 93–111, 1955.

[31] M. S. Ibrahim and N. 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)*, pp. 1–5, 2019.

[32] M. Sørensen, C. I. Kanatsoulis, and N. D. Sidiropoulos, "Generalized canonical correlation analysis: A subspace intersection approach," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2452–2467, 2021.

[33] D. Lahat and C. Jutten, "Joint independent subspace analysis using second-order statistics," *IEEE Transactions on Signal Processing*, vol. 64, no. 18, pp. 4891–4904, 2016.

[34] H. Richard, P. Ablin, B. Thirion, A. Gramfort, and A. Hyvarinen, "Shared independent component analysis for multi-subject neuroimaging," in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.

[35] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Intl. Conference on Machine Learning*, 2013.

[36] Q. Lyu, X. Fu, W. Wang, and S. Lu, "Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective," in *International Conference on Learning Representations*, 2022.

[37] D. Lahat and C. Jutten, "Joint independent subspace analysis: Uniqueness and identifiability," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 684–699, 2019.

[38] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. Wiley, 1959.

[39] P. A. Bekker, "A note on the identification of restricted factor loading matrices.," *Psychometrika*, vol. 51, pp. 607–611, 1986.

[40] D. N. Lawley and A. E. Maxwell, "Factor analysis as a statistical method," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 12, no. 3, pp. 209–229, 1962.

[41] C. C. Heyde, *Quasi-Likelihood and its Application*. Springer New York, 1997.

[42] H. Akaike, "Factor analysis and AIC," *Psychometrika*, vol. 52, pp. 317–332, 1987.

[43] A. B. Owen and J. Wang, "Bi-Cross-Validation for factor analysis," *Statistical Science*, vol. 31, no. 1, pp. 119 – 139, 2016.

[44] S. Ahn, W. Carey, and A. Horenstein, "Eigenvalue ratio test for the number of factors," *Econometrica*, vol. 81, 2009.

[45] J. M. Lee, *Introduction to Smooth Manifolds*. New York: Springer, 2002.

[46] F. Zhang, ed., *The Schur Complement and its Applications*. Springer New York, 2005.
[47] A. Shapiro, "Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis," *Psychometrika*, vol. 47, pp. 187–199, 1982.
[48] A. W. v. d. Vaart, *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.

APPENDIX
PROOFS OF THEOREMS $1-4$

For ease of notation, let $\mathbb{A}\oplus\mathbb{B}\subset\mathbb{R}^{n\times(r_0+r)}$ be the subspace of matrices which 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 1 connects the two notions. Proofs of the following lemmas can be found in the Supplementary Materials for this paper.

Lemma 1. *If $\mathcal{C}\subset\mathbb{A}\times\mathbb{B}$ is a null set which is a union of $\sim_2$-equivalence classes, then the set of representatives $\tilde{C}\subset\mathbb{A}_L^*\times\mathbb{B}_L^*$ is null in $\mathbb{A}_L\times\mathbb{B}_L$.*

Lemma 2. *Let $\mathbf{H}=[\mathbf{H}_1^\mathsf{T}\ \mathbf{H}_2^\mathsf{T}]^\mathsf{T}$ and $\mathbf{Z}=[\mathbf{Z}_1^\mathsf{T}\ \mathbf{Z}_2^\mathsf{T}]^\mathsf{T}$ 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 $\mathbf{H}\mathbf{Z}^\mathsf{T}+\mathbf{Z}\mathbf{H}^\mathsf{T}=\mathbf{0}$ implies $\mathbf{Z}=\mathbf{0}$.*

Lemma 3. *(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 Theorem 1: Define $\mathcal{B}\subset\mathbb{A}\times\mathbb{B}\times\operatorname{Diag}_{\geq 0}(n)$ as the subset where $(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})$ does not have globally identified noise variances. For any $\boldsymbol{\Phi}'\succeq\boldsymbol{\Phi}$, if $(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})$ is in $\mathcal{B}$ then so too is $(\mathbf{A},\mathbf{B},\boldsymbol{\Phi}')$ as $\mathbf{R}_{\mathbf{xx}}(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})=\mathbf{R}_{\mathbf{xx}}(\tilde{\mathbf{A}},\tilde{\mathbf{B}},\tilde{\boldsymbol{\Phi}})$ implies $\mathbf{R}_{\mathbf{xx}}(\mathbf{A},\mathbf{B},\boldsymbol{\Phi}+[\boldsymbol{\Phi}'-\boldsymbol{\Phi}])=\mathbf{R}_{\mathbf{xx}}(\tilde{\mathbf{A}},\tilde{\mathbf{B}},\tilde{\boldsymbol{\Phi}}+[\boldsymbol{\Phi}'-\boldsymbol{\Phi}])$ with $\boldsymbol{\Phi}'\neq\tilde{\boldsymbol{\Phi}}+[\boldsymbol{\Phi}'-\boldsymbol{\Phi}]$. As the set of $\boldsymbol{\Phi}'$ greater than $\boldsymbol{\Phi}$ has positive Lebesgue measure in $\operatorname{Diag}(n)$, $\mathcal{B}$ is null in $\mathbb{A}\times\mathbb{B}\times\operatorname{Diag}_{\geq 0}(n)$ 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 $\boldsymbol{\varphi}=\boldsymbol{\Phi}-\tilde{\boldsymbol{\Phi}}$ with

$$\mathbf{A}\mathbf{A}^\mathsf{T}+\mathbf{B}\mathbf{B}^\mathsf{T}+\boldsymbol{\varphi}=\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T}+\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\mathsf{T},\qquad\boldsymbol{\varphi}\neq\mathbf{0}. \tag{25}$$

If $[\mathbf{A}\ \mathbf{B}]\in\mathcal{U}$ for $\boldsymbol{\varphi}=\boldsymbol{\Phi}-\tilde{\boldsymbol{\Phi}}\neq 0$, then $(\mathbf{A},\mathbf{B},\boldsymbol{\Phi})$ 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 3, $\mathcal{U}^*=(\mathbb{A}\oplus\mathbb{B})^{**}\cap\mathcal{U}$. 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 (25). The proof proceeds in two steps. In the first step, β 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: φ 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 cth channel. Additivity of rank with respect to the Schur complement implies

$$\begin{aligned}r_0+r+\operatorname{rank}(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T}+\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\mathsf{T})&=n+\operatorname{rank}(\mathbf{H}),\\ r_0+r_c+\operatorname{rank}(\tilde{\mathbf{A}}_c\tilde{\mathbf{A}}_c^\mathsf{T}+\tilde{\mathbf{B}}_c\tilde{\mathbf{B}}_c^\mathsf{T})&=n_c+\operatorname{rank}(\mathbf{H}_c),\end{aligned} \tag{26}$$

for $c=1,\ldots,C$, where $\mathbf{H}$ is

$$\mathbf{H}(\mathbf{A},\mathbf{B},\boldsymbol{\varphi})=\begin{bmatrix}\mathbf{I}_{r_0}+\mathbf{A}^\mathsf{T}\boldsymbol{\varphi}^{-1}\mathbf{A}&\mathbf{A}^\mathsf{T}\boldsymbol{\varphi}^{-1}\mathbf{B}\\ \mathbf{B}^\mathsf{T}\boldsymbol{\varphi}^{-1}\mathbf{A}^\mathsf{T}&\mathbf{I}_r+\mathbf{B}^\mathsf{T}\boldsymbol{\varphi}^{-1}\mathbf{B}\end{bmatrix}, \tag{27}$$

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 cth 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 (26) yields that

$$\begin{aligned}\operatorname{rank}\big(\mathbf{H}(\mathbf{A},\mathbf{B},\boldsymbol{\varphi})\big)&\leq 2(r_0+r)-n,\\ \operatorname{rank}(\mathbf{I}_{r_c}+\mathbf{B}_c^\mathsf{T}\boldsymbol{\varphi}_c^{-1}\mathbf{B}_c)&\leq\min\{r_c,2(r_0+r_c)-n_c\},\end{aligned} \tag{28}$$

for $c=1,\ldots,C$. To establish the second line of (28), we combine (26) 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 (28).

Therefore, showing that the set of $[\mathbf{A}\ \mathbf{B}]$ where $\mathbf{H}(\mathbf{A},\mathbf{B},\boldsymbol{\varphi})$ satisfies (28) 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 (28) 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

$$\dim\mathcal{S}=\tfrac{r_0(r_0+1)}{2}+r_0r+\textstyle\sum_{c=1}^C\tfrac{r_c(r_c+1)}{2}.$$

Recall that both $(\mathbb{A}\oplus\mathbb{B})^{**}$, which contains $[\mathbf{A}\ \mathbf{B}]$ satisfying the maximal rank submatrix condition of Lemma 3, and the non-singular diagonal matrices $\mathrm{Diag}^*(n)$ are open submanifolds of their respective vector spaces. So, $\mathbf{H}$ is a smooth map from the product manifold $(\mathbb{A}\oplus\mathbb{B})^{**}\times\mathrm{Diag}^*(n)$ 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 cth 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. So, $[\mathbf{A}\ \mathbf{B}]^{\mathsf{T}}\boldsymbol{\varphi}^{-2}[\mathbf{A}\ \mathbf{B}]$ 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 $(\mathbf{A},\mathbf{B},\boldsymbol{\varphi})\in(\mathbb{A}\oplus\mathbb{B})^{**}\times\mathrm{Diag}^*(n)$ was arbitrary, $\mathbf{H}$ is a smooth submersion.

Next, we will show that the subset of $\mathcal{S}$ where the rank conditions (28) are satisfied can be written as a union of embedded submanifolds. For any index set $\alpha\subset\{1,\dots,(r_0+r)\}$, relate α to the block-structure of $\mathbf{H}$ by defining $\rho_0 \equiv |\alpha\cap\{1,\dots,r_0\}|$ and $\rho_c \equiv |\alpha\cap\{(r_{<c}+1),\dots,(r_{<c}+r_c)\}|$ for each c. Then for those α where $\boldsymbol{\rho}\equiv[\rho_0,\rho_1,\dots,\rho_C]$ satisfies

$$\begin{aligned}\rho_0 + \textstyle\sum_{c=1}^{C}\rho_c \le 2(r_0+r) - n,\\ \rho_c \le 2(r_0+r_c) - n_c,\end{aligned} \tag{29}$$

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 $\mathrm{rank}(\mathbf{S}) = |\alpha|$. If $\mathbf{H}(\mathbf{A},\mathbf{B},\boldsymbol{\varphi})$ satisfies (28), then $\mathbf{H}(\mathbf{A},\mathbf{B},\boldsymbol{\varphi})\in\mathcal{S}_\alpha$ for some α satisfying (29). 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 $\mathbf{C}\in\mathrm{Sym}(n)$ partitioned as $\mathbf{C} = [\mathbf{C}_{11}\ \mathbf{C}_{12};\ \mathbf{C}_{12}^{\mathsf{T}}\ \mathbf{C}_{22}]$ with $\mathbf{C}_{11}$ invertible and $\mathrm{rank}(\mathbf{C}_{11}) = \mathrm{rank}(\mathbf{C})$, 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 α. Define the vector space $\mathcal{W}_\alpha\subset\mathrm{Sym}(\rho_0+\rho)$ 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

$$\begin{aligned}\dim(\mathcal{W}_\alpha) = \tfrac{\rho_0(\rho_0+1)}{2} + \rho_0\rho + \textstyle\sum_{c=1}^{C}\tfrac{\rho_c(\rho_c+1)}{2},\\ \dim(\mathcal{Y}_\alpha) = \rho_0(r_0+r-\rho_0) - \rho_0\rho + \textstyle\sum_{c=1}^{C}\rho_c(r_c-\rho_c).\end{aligned} \tag{30}$$

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., [45, Thm. 6.30]) ensures that $\mathbf{H}^{-1}(\mathcal{S}_\alpha)$ is an embedded submanifold of $(\mathbb{A}\oplus\mathbb{B})^{**}\times\mathrm{Diag}^*(n)$ with codimension equal to the codimension of $\mathcal{S}_\alpha$ in $\mathcal{S}$, namely

$$\begin{aligned}\mathrm{codim}\,\mathcal{S}_\alpha = \tfrac{r_0(r_0+1)-\rho_0(\rho_0+1)}{2} + \textstyle\sum_{c=1}^{C}\tfrac{r_c(r_c+1)-\rho_c(\rho_c+1)}{2}\\ + (r_0-\rho_0)r - \rho_0(r_0-\rho_0) - \textstyle\sum_{c=1}^{C}\rho_c(r_c-\rho_c).\end{aligned}$$

Let π be the projection map from $(\mathbb{A}\oplus\mathbb{B})^{**}\times\mathrm{Diag}^*(n)$ to $(\mathbb{A}\oplus\mathbb{B})^{**}$. Dimensional considerations [45, p 131] then imply that $\pi\left(\mathbf{H}^{-1}(\mathcal{S}_\alpha)\right)$ is null if

$$\dim(\mathbb{A}\oplus\mathbb{B})^* + n - \mathrm{codim}\,\mathcal{S}_\alpha < \dim(\mathbb{A}\oplus\mathbb{B})^*. \tag{31}$$

If the above inequality is satisfied for all α with $\boldsymbol{\rho}$ meeting (29), 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 α satisfying (29). 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 $j_c \equiv |\beta\cap\{(r_{<c}+1),\dots,(r_{<c}+r_c)\}|$ be the number of zeros in the cth channel on the diagonal of $\boldsymbol{\varphi}$ and let the channels be numbered such that $j_1\ge j_2\ge\cdots\ge j_C$.

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 $\mathbf{A}' = [\mathbf{A}_{12}^{\mathsf{T}}\ \mathbf{A}_2^{\mathsf{T}}\ \dots\ \mathbf{A}_C]^{\mathsf{T}}$ and $\mathbf{B}' = \mathrm{blkdiag}(\mathbf{B}_{12},\mathbf{B}_2,\dots,\mathbf{B}_C)$ 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 (25) implies

$$\mathbf{A}_{11}\mathbf{A}_{11}^{\mathsf{T}} + \mathbf{B}_{11}\mathbf{B}_{11}^{\mathsf{T}} = \tilde{\mathbf{A}}_{11}\tilde{\mathbf{A}}_{11}^{\mathsf{T}} + \tilde{\mathbf{B}}_{11}\tilde{\mathbf{B}}_{11}^{\mathsf{T}}. \tag{32}$$

As $\boldsymbol{\varphi}$ is diagonal, the off-diagonal blocks are also equal,

$$[\mathbf{A}_{11}\ \mathbf{B}_{11}\ \mathbf{0}_{r_{>1}}][\mathbf{A}'\ \mathbf{B}']^{\mathsf{T}} = [\tilde{\mathbf{A}}_{11}\ \tilde{\mathbf{B}}_{11}\ \mathbf{0}_{r_{>1}}][\tilde{\mathbf{A}}'\ \tilde{\mathbf{B}}']^{\mathsf{T}}. \tag{33}$$

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

$$\mathbf{A}'\mathbf{A}' + \mathbf{B}'\mathbf{B}' - [\mathbf{A}'\ \mathbf{B}']\mathbf{W}[\mathbf{A}'\ \mathbf{B}']^\mathsf{T}, \tag{34}$$

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 [46, Ch. 6], and so (34) equals

$$[\mathbf{A}'\ \mathbf{B}']\begin{bmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{r_{>1}}\end{bmatrix}\begin{bmatrix}\mathbf{A}'^\mathsf{T}\\ \mathbf{B}'^\mathsf{T}\end{bmatrix}, \tag{35}$$

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

$$\mathbf{D} \equiv \begin{array}{c}\begin{matrix} r_0' & r_0 - r_0' & r_1' & r_1 - r_1'\end{matrix}\\ \begin{bmatrix}\mathbf{W}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{W}_1 & \mathbf{0} & \mathbf{V}' & \mathbf{0}\end{bmatrix}\end{array}\begin{matrix} r_0\\ r_1\end{matrix},$$

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 (35) 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 (25) and subtract $[\mathbf{A}'\ \mathbf{B}']\mathbf{W}[\mathbf{A}'\ \mathbf{B}']$ from both sides. By the equalities (32) and (33), 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 (25) 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 (25) 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 (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 (29) 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 (28) 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 (29) is possible. Then the dimension condition (31) is equivalent to $\psi(\mathbf{n}, \mathbf{r}, \boldsymbol{\rho}) > 0$ with ψ defined in (12). As a function of ρ_0 alone, ψ is decreasing on $[0, r_0]$, so ψ being positive when ρ_0 is at its maximum feasible value implies ψ is positive for all smaller values of ρ_0. So, $\min_\rho \psi(\mathbf{n}, \mathbf{r}, \rho) > 0$ ensures that (31) is satisfied for all valid α. Therefore, $\mathcal{U}_\emptyset^*$ is null.

For any index set β with $j_c, c = 1, \ldots, C$ zeros in the cth channel on the diagonal of $\boldsymbol{\varphi}$, taking $n_c' = n_c - j_c$ and $r_0', r_1', \ldots r_C'$ as in (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 β. 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 3. So, $\mathcal{U}$ is a subset of the finite union of the null sets $\mathcal{U}_\emptyset^*, \mathcal{U}_\beta^*$ for all β, 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. ∎

Proof of Theorem 2: 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 $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Phi}) \in \mathbb{A}_L^* \times \mathbb{B}_L^* \times \mathrm{Diag}_{\geq 0}(n)$ is obtained from $\boldsymbol{\eta}$ by inverting (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

$$d\mathbf{R}_{\mathbf{xx}} = \mathbf{A}d\mathbf{A}^\mathsf{T} + d\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{B}d\mathbf{B}^\mathsf{T} + d\mathbf{B}\mathbf{B}^\mathsf{T} + d\boldsymbol{\Phi} \tag{36}$$

is a linear map in $(d\mathbf{A}, d\mathbf{B}, d\boldsymbol{\Phi})$ from a vector space of dimension L as defined in (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 (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 the condition $r_1 \leq \frac{1}{2}\left(2n_1 + 1 - \sqrt{8n_1 + 1}\right)$. 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 5 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 $(d\mathbf{A}, d\mathbf{B}, d\mathbf{\Phi})$. 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 $(d\mathbf{A}, d\mathbf{B}, d\mathbf{\Phi}) \mapsto d\mathbf{R}_{\mathbf{xx}}$ is generically injective and therefore $d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})$ is generically injective.

Define the linear maps $\mathbf{F}_\mathbf{A} : \mathbb{A}_L \to \mathrm{Sym}(n)$ and $\mathbf{F}_\mathbf{B} : \mathbb{B}_L \to \mathrm{Sym}(n)$ as $\mathbf{F}_\mathbf{A}(\mathbf{X}) \equiv \mathbf{A}\mathbf{X}^\mathsf{T} + \mathbf{X}\mathbf{A}^\mathsf{T}$ and $\mathbf{F}_\mathbf{B}(\mathbf{Y}) \equiv \mathbf{B}\mathbf{Y}^\mathsf{T} + \mathbf{Y}\mathbf{B}^\mathsf{T}$. Similarly, $\mathbf{F}_{\mathbf{A},\mathbf{B}} : \mathbb{A}_L \times \mathbb{B}_L \to \mathrm{Sym}(n)$ is

$$\begin{aligned}\mathbf{F}_{\mathbf{A},\mathbf{B}}(\mathbf{X},\mathbf{Y}) &\equiv \mathbf{F}_\mathbf{A}(\mathbf{X}) + \mathbf{F}_\mathbf{B}(\mathbf{Y}) \\ &= [\mathbf{A}\ \mathbf{B}][\mathbf{X}\ \mathbf{Y}]^\mathsf{T} + [\mathbf{X}\ \mathbf{Y}][\mathbf{A}\ \mathbf{B}]^\mathsf{T}.\end{aligned} \tag{37}$$

For $\mathbf{F}_\mathbf{A}$, an application of Lemma 2 implies that $\mathbf{F}_\mathbf{A}$ is injective, as $\mathbf{F}_\mathbf{A}(\mathbf{X}) = \mathbf{0}$ implies $\mathbf{X} = \mathbf{0}$ by the structure of $\mathbb{A}_L$ and the FCR assumption. Arguing channel-wise for $\mathbf{F}_\mathbf{B}$, $\mathbf{F}_\mathbf{B}$ is injective in a similar fashion.

Second, define the image subspaces as $\mathcal{A} \equiv \mathrm{Im}(\mathbf{F}_\mathbf{A})$ and $\mathcal{B} \equiv \mathrm{Im}(\mathbf{F}_\mathbf{B})$. Injectivity of $\mathbf{F}_\mathbf{A}$ and $\mathbf{F}_\mathbf{B}$ 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 $\mathbf{F}_{\mathbf{A},\mathbf{B}}$ will be injective iff $\mathcal{A} \cap \mathcal{B} = \{\mathbf{0}\}$, as existence of $(\mathbf{X},\mathbf{Y})$ with $\mathbf{F}_{\mathbf{A},\mathbf{B}}(\mathbf{X},\mathbf{Y}) = \mathbf{0}$ implies that $\mathbf{F}_\mathbf{A}(\mathbf{X}) = -\mathbf{F}_\mathbf{B}(\mathbf{Y})$ and so $\mathbf{F}_\mathbf{B}(\mathbf{Y})$ is in $\mathcal{A} \cap \mathcal{B}$. If $\mathcal{A} \cap \mathcal{B} = \{\mathbf{0}\}$, then injectivity of $\mathbf{F}_\mathbf{A}$ and $\mathbf{F}_\mathbf{B}$ 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 $\mathbf{F}_\mathbf{A}(\mathbf{X}) = \mathbf{G} = \mathbf{F}_\mathbf{B}(\mathbf{Y})$ and so $\mathbf{F}_{\mathbf{A},\mathbf{B}}(\mathbf{X},-\mathbf{Y}) = \mathbf{0}$.

To show $\mathcal{A} \cap \mathcal{B} = \{\mathbf{0}\}$, suppose $\mathbf{F}_{\mathbf{A},\mathbf{B}}(\mathbf{X},\mathbf{Y}) = \mathbf{0}$ and so

$$\mathbf{A}\mathbf{X}^\mathsf{T} + \mathbf{X}\mathbf{A}^\mathsf{T} = -\mathbf{B}\mathbf{Y}^\mathsf{T} - \mathbf{Y}\mathbf{B}^\mathsf{T}. \tag{38}$$

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 (38) 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

$$[\mathbf{X}\ \mathbf{Y}] = [\mathbf{A}\ \mathbf{B}]\mathbf{W}. \tag{39}$$

Combining (39) and (37), $\mathbf{F}_{\mathbf{A},\mathbf{B}}(\mathbf{X},\mathbf{Y}) = \mathbf{0}$ 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 2 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, (39) is equivalent to

$$[\mathbf{X}\ \mathbf{Y}] + [\mathbf{A}\ \mathbf{B}] = \big([\mathbf{A}\ \mathbf{B}] - [\mathbf{X}\ \mathbf{Y}]\big)\,\mathbf{Q}.$$

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 ϵ, 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 2, the subset $\tilde{\mathcal{C}} \subset \mathbb{A}_L^* \times \mathbb{B}_L^*$ where the full rank submatrix condition of Proposition 1 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 1 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 3, 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

$$(\mathcal{A} \oplus \mathcal{B}) \cap \mathrm{Diag}(n) = \{\mathbf{0}\}, \tag{40}$$

which is equivalent to showing that the differential (36) is injective. To show (40), we examine the orthogonal complement $\big((\mathcal{A} \oplus \mathcal{B}) \cap \mathrm{Diag}(n)\big)^\perp$. Standard properties of the subspace lattice show that $\big((\mathcal{A} \oplus \mathcal{B}) \cap \mathrm{Diag}(n)\big)^\perp$ 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

$$n = \dim\big((\mathcal{A}^\perp \cap \mathcal{B}^\perp) \cap (\mathcal{A}^\perp \cap \mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp)^\perp\big). \tag{41}$$

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,

$$\mathcal{Y} = \mathrm{Span}\big(\{\mathbf{e}_i\mathbf{e}_j^\mathsf{T} + \mathbf{e}_j\mathbf{e}_i^\mathsf{T} :\ \exists\ c \text{ s.t. } i \le r_{<c} + r_c < j\}\big).$$

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 cth 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 (18) is satisfied, as shown in [22, Thm. 3.2]. If the above is satisfied, $\mathcal{B} \cap \mathrm{Diag}(n) = \{\mathbf{0}\}$. From this, we obtain that

$$\begin{aligned}\mathcal{A}^\perp + \mathcal{B}^\perp = (\mathcal{A}^\perp + \mathcal{B}^\perp) \cap (\mathcal{A}^\perp + (\mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp)) \\ \oplus (\mathcal{A}^\perp + \mathcal{B}^\perp) \cap (\mathcal{A} \cap (\mathcal{B} \oplus \mathrm{Diag}(n))).\end{aligned} \tag{42}$$

However, the term $\mathcal{A} \cap (\mathcal{B} \oplus \mathrm{Diag}(n))$ can be shown to be $\{\mathbf{0}\}$ by the results from [47] for single-channel FA when (16) is satisfied. In particular, if $\mathbf{C} \in \mathcal{A} \cap (\mathcal{B} \oplus \mathrm{Diag}(n))$, then as $\mathbf{C} \in \mathcal{A}$, [47, Lemma 2.1] implies that

$$\mathbf{v}_i^\mathsf{T}\mathbf{C}\mathbf{v}_j = 0, \quad 1 \le i \le j \le n - r_0,$$

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

$$\tfrac{(n-r_0)(n-r_0+1)}{2} \ge \dim(\mathcal{B}) + \dim(\mathrm{Diag}(n)),$$

as the RHS is $\dim(\mathcal{B} \oplus \mathrm{Diag}(n))$. This is equivalent to the condition (16), and therefore $\mathcal{A} \cap (\mathcal{B} \oplus \mathrm{Diag}(n)) = \{\mathbf{0}\}$. This implies that the RHS of (42) equals $(\mathcal{A}^\perp + (\mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp))$. Finally, (41) 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

$$n - \left[\dim(\mathcal{A}^\perp + \mathcal{B}^\perp) - \dim(\mathcal{A}^\perp + (\mathcal{B}^\perp \cap \mathrm{Diag}(n)^\perp))\right] - \textstyle\sum_{c=1}^C J_c,$$

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. ∎

Proof of Theorem 3: Define the function ℓ_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, ℓ_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 [2, 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

$$\begin{aligned}B_\epsilon &= \{\boldsymbol{\eta} \in V' \;;\; ||\boldsymbol{\eta} - \mathring{\boldsymbol{\eta}}|| < \epsilon\}, \\ L_h &= \{\boldsymbol{\eta} \in V' \;;\; \ell_0(\boldsymbol{\eta}) \le m + h\}.\end{aligned}$$

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 ℓ_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 ℓ_T from ℓ_0 is controlled as,

$$\begin{aligned}\sup_{\boldsymbol{\eta} \in V'} |\ell_T(\boldsymbol{\eta}) - \ell_0(\boldsymbol{\eta})| &= \sup_{\boldsymbol{\eta} \in V'} |\mathrm{tr}\,\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})(\mathbf{S}_T - \mathbf{\Sigma}_{\mathbf{xx}})| \\ &\le \sup_{\boldsymbol{\eta} \in V'} ||\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})||_F ||\mathbf{S}_T - \mathbf{\Sigma}_{\mathbf{xx}}||_F \\ &\le \epsilon^{-1}\sqrt{n}||\mathbf{S}_T - \mathbf{\Sigma}_{\mathbf{xx}}||_F.\end{aligned}$$

The third line follows from the definition of V', which imposes that $\lambda_{\min}(\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta})) \ge \epsilon$. As the second moment of $\mathbf{x}$ exists, $\mathbf{S}_T$ is consistent so $||\mathbf{S}_T - \mathbf{\Sigma}_{\mathbf{xx}}||_F \xrightarrow{p} 0$. So, $\ell_T(\boldsymbol{\eta})$ converges uniformly in probability to the limiting function $\ell_0(\boldsymbol{\eta})$. As $\hat{\boldsymbol{\eta}}_T$ minimizes ℓ_T, standard results for M-estimators [48, p. 45] imply that $\hat{\boldsymbol{\eta}}_T \xrightarrow{p} \mathring{\boldsymbol{\eta}}$, so $(\hat{\mathbf{A}}_T, \hat{\mathbf{B}}_T, \hat{\mathbf{\Phi}}_T) \xrightarrow{p} (\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\mathbf{\Phi}})$ as well. Note that if $\mathbf{\Sigma}_{\mathbf{xx}} \in \mathcal{R}(\mathbf{n}, \mathbf{r})$, then the minimizing $\mathring{\mathbf{R}}_{\mathbf{xx}}$ is $\mathbf{\Sigma}_{\mathbf{xx}}$, which is the unique minimum by the Gibbs inequality. ∎

Proof of Theorem 4: For any $r > 0$, define the set of $\boldsymbol{\eta} \in V'$ with $||\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}) - \mathbf{\Sigma}_{\mathbf{xx}}||_F < r$ and $||\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta}) - \mathbf{\Sigma}_{\mathbf{xx}}^{-1}||_F < r$. 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,

$$\begin{aligned}|\ell_T(\boldsymbol{\eta}_1) - \ell_T(\boldsymbol{\eta}_2)| &\le |\log\det\mathbf{R}_1\mathbf{R}_2^{-1}| + |\mathrm{tr}\,\mathbf{S}_T[\mathbf{R}_1^{-1} - \mathbf{R}_2^{-1}]| \\ &\le |\,\mathrm{tr}\,\mathbf{R}_1[\mathbf{R}_2^{-1} - \mathbf{R}_1^{-1}]| + |\,\mathrm{tr}\,\mathbf{S}_T[\mathbf{R}_2^{-1} - \mathbf{R}_1^{-1}]| \\ &\le \big(||\mathbf{R}_1 - \mathbf{\Sigma}_{\mathbf{xx}}||_F + ||\mathbf{\Sigma}_{\mathbf{xx}}||_F + ||\mathbf{S}_T||_F\big) \times \\ &\qquad ||\mathbf{R}_1^{-1}||_F ||\mathbf{R}_2^{-1}||_F ||\mathbf{R}_1 - \mathbf{R}_2||_F \\ &\le m(r, \mathbf{\Sigma}_{\mathbf{xx}}, \mathbf{S})||\mathbf{R}_1 - \mathbf{R}_2||_F,\end{aligned}$$

with $m(r, \mathbf{\Sigma}_{\mathbf{xx}}, \mathbf{S}) = (r + ||\mathbf{\Sigma}_{\mathbf{xx}}||_F + ||\mathbf{S}||_F)(r + ||\mathbf{\Sigma}_{\mathbf{xx}}^{-1}||_F)^2$. Since by assumption $E[||\mathbf{x}_1||^4] < \infty$, it is implied that $E[||\mathbf{S}_T||_F^2] < \infty$ and therefore $E[m(r, \mathbf{\Sigma}_{\mathbf{xx}}, \mathbf{S})^2] < \infty$. 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,

$$\begin{aligned}||\mathbf{R}_1 - \mathbf{R}_2||_F &\le 2||\mathbf{A}_1||_F||\mathbf{A}_1 - \mathbf{A}_2||_F + ||\mathbf{\Phi}_1 - \mathbf{\Phi}_2||_F \\ &\quad + 2||\mathbf{B}_1||_F||\mathbf{B}_1 - \mathbf{B}_2||_F \\ &\le C(r, \mathbf{\Sigma}_{\mathbf{xx}})||\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2||_2\end{aligned}$$

for some non-random $C(r, \mathbf{\Sigma}_{\mathbf{xx}})$. Hence, $\ell_T(\boldsymbol{\eta})$ is Lipschitz within some neighborhood of $\mathring{\boldsymbol{\eta}}$ with first differential

$$\begin{aligned}d\ell_T(\boldsymbol{\eta}, d\boldsymbol{\eta}) = &\,\mathrm{tr}(\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}, d\boldsymbol{\eta})) \\ &- \mathrm{tr}(\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}, d\boldsymbol{\eta})\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})\mathbf{S}_T).\end{aligned}$$

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

$$d^2\ell_T(\boldsymbol{\eta}, d\boldsymbol{\eta}) = 2\,\mathrm{tr}\left([\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}, d\boldsymbol{\eta})]^2\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})\mathbf{S}_T\right) - \mathrm{tr}\left([\mathbf{R}_{\mathbf{xx}}^{-1}(\boldsymbol{\eta})d\mathbf{R}_{\mathbf{xx}}(\boldsymbol{\eta}, d\boldsymbol{\eta})]^2\right).$$

Taking expectation and evaluating at $\mathring{\boldsymbol{\eta}}$, the above display equals $||\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1/2}d\mathbf{R}_{\mathbf{xx}}(\mathring{\boldsymbol{\eta}}, d\boldsymbol{\eta})\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1/2}||_F^2$ for $d\mathbf{R}_{\mathbf{xx}}$ as in (36), evaluated at $(\mathring{\mathbf{A}}, \mathring{\mathbf{B}}, \mathring{\boldsymbol{\Phi}})$ with tangent vector $(d\mathbf{A}, d\mathbf{B}, d\boldsymbol{\Phi})$. 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 $(d\mathbf{A}, d\mathbf{B}, d\boldsymbol{\Phi})$. Let $\mathbf{V}_0$ be the Hessian matrix of ℓ_0 at $\mathring{\boldsymbol{\eta}}$. As the second differential is positive, 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. [48, p. 53]) then imply that $\sqrt{T}(\hat{\boldsymbol{\eta}}_T - \mathring{\boldsymbol{\eta}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{W})$, where

$$\mathbf{W} = \mathbf{V}_0^{-1}E\Big[\frac{\partial\ell_T(\mathring{\boldsymbol{\eta}})}{\partial\boldsymbol{\eta}}\frac{\partial\ell_T(\mathring{\boldsymbol{\eta}})^{\mathsf{T}}}{\partial\boldsymbol{\eta}}\Big]\mathbf{V}_0^{-1}. \tag{43}$$

■