Multi-Channel Factor Analysis with Common and Unique Factors

David Ramírez, *Senior Member, IEEE,* Ignacio Santamaria, *Senior Member, IEEE,*
Louis L. Scharf, *Life Fellow, IEEE,* and Steven Van Vaerenbergh, *Senior Member, IEEE*

Abstract— This work presents a generalization of classical factor analysis (FA). Each of M channels carries measurements that share factors with all other channels, but also contains factors that are unique to the channel. Furthermore, each channel carries an additive noise whose covariance is diagonal, as is usual in factor analysis, but is otherwise unknown. This leads to a problem of multi-channel factor analysis with a specially structured covariance model consisting of shared low-rank components, unique low-rank components, and diagonal components. Under a multivariate normal model for the factors and the noises, a maximum likelihood (ML) method is presented for identifying the covariance model, thereby recovering the loading matrices and factors for the shared and unique components in each of the M multiple-input multiple-output (MIMO) channels. The method consists of a three-step cyclic alternating optimization, which can be framed as a block minorization-maximization (BMM) algorithm. Interestingly, the three steps have closed-form solutions and the convergence of the algorithm to a stationary point is ensured. Numerical results demonstrate the performance of the proposed algorithm and its application to passive radar.

Index Terms— Block minorization-maximization (BMM) algorithms, expectation-maximization (EM) algorithms, maximum likelihood (ML) estimation, multi-channel factor analysis (MFA), multiple-input multiple-output (MIMO) channels, passive radar.

I. INTRODUCTION

Classical factor analysis (FA) was pioneered by Spearman in his seminal paper [1]. Spearman and others applied FA to problems in psychology, and especially to the analysis of the correlation of children's scores across different academic subjects. Later, with the work of Lawley, Anderson, and

D. Ramírez is with the Department of Signal Theory and Communications, University Carlos III of Madrid, Leganés, Spain and with the Gregorio Marañón Health Research Institute, Madrid, Spain (e-mail: david.ramirez@uc3m.es).

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

S. Van Vaerenbergh is with the Department of Mathematics, Statistics and Computing, University of Cantabria, Santander, Spain (e-mail: steven.vanvaerenbergh@unican.es).

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

others [2]–[4], a more rigorous approach was developed, which made FA a well-established technique in multivariate statistics. FA now finds many applications in science and engineering. For instance, in the field of array signal processing, FA has been applied to radio astronomy [5], [6], cognitive radio [7], direction-of-arrival estimation [8]–[10], modal analysis [11], [12], and detection or source enumeration [13]–[16].

In the classical FA model, measurements in a single MIMO channel are modeled as a set of unknown factors scaling the modes of an unknown factor loading matrix, plus a multivariate normal noise of unknown, but diagonal, covariance. The factors are typically treated as multivariate normal, with identity covariance, so that the net effect is to posit a multivariate normal measurement with a structured covariance consisting of an unknown low-rank component to account for the factor loadings plus an unknown diagonal matrix to account for the additive noise. Thus, in second-order FA, the problem is to estimate a low-rank plus diagonal covariance matrix from several multivariate observations. This model has been recently extended in [17] to consider more complicated covariance structures, i.e., not diagonal, but this structure needs to be sparse and known. Even another extension was developed in [18], where the precision matrix, the inverse of the covariance matrix, is assumed to be composed by a low-rank component plus a sparse one. The sparsity pattern in this model enforces relations of conditional independence between observed variables, whereas the low-rank component favors models explained by a reduced number of latent hidden factors.

Common estimation approaches for the FA model are based on the maximum likelihood (ML) criterion. Unfortunately, even under the Gaussian assumption, the maximization problem has no closed-form solution and numerical methods must be employed. A convergent numerical procedure for obtaining the maximum likelihood estimates was first given by Joreskog [19], [20] (cf. Chapter 9 in [3]). Other optimization approaches have been investigated for this problem, ranging from steepest descent [4] and alternating optimization methods [7], [21], [22], to Expectation-Maximization (EM) algorithms [23], [24]. Most of the proposed techniques assumed known the number of factors, i.e., the dimension of the low-rank component. When this is not the case, it needs to be estimated [25]. Moreover, if we abandon the ML criterion, there are other alternatives. For instance, the work in [26] derives a robust technique based on the low-rank plus sparse factorization of the precision matrix that also provides an estimate for the number of factors as a result of the optimization procedure.

The technique of FA has been extended to multiple channels of multivariate observations. To the best of our knowledge, the first generalization was developed by Tucker [27], where he proposed the so-called *inter-battery* FA. In this work, the observations of two channels are composed by linear combinations of *common* factors and independent noises without a particular covariance structure. Additionally, he derived an estimation algorithm based on the least squares (LS) criterion, which was later related to canonical correlation analysis (CCA) by Browne [28]. The extension of the inter-battery FA model to more than two channels was developed in [29], [30].

The recent work in [17] also proposes a different generalization of FA to several channels, which assumes that the factors at each channel are independent, but the noise covariance matrix is common. This work also allows for a number of factors in each channel that may be different. A different generalization is presented in [31], and termed group factor analysis (GFA). In GFA, the factors may be common to all channels or to a subset of them. Other work related to multi-channel factor analysis includes the parallel factor (PARAFAC) analysis model [32] and independent vector analysis (IVA) [33]. Our model for the channel covariance differs significantly from the channel covariance in the PARAFAC model [32]. The multiple channels (i.e., the third dimension in the three-way array) of PARAFAC are obtained from displaced but otherwise identical subarrays, which induces a shift-invariant structure in the loading matrices of the common factors. Further, the noise covariance model in [32] is white. Our model is different is several aspects, namely, we never have rotational invariance, we have both common and unique (or channel-specific) factors, and our model for the noise covariance is diagonal with unknown variances. The standard model in IVA accounts for the dependence of a set of common sources or factors observed through several mixing matrices, but it does not consider channel-specific factors or noises whose variances are unequal [33].

This paper extends the inter-battery FA model to more than two channels and to noise covariance structures that account for additive noise and unique channel factors, which are missing in the original inter-battery analysis of Tucker. This multi-channel FA (MFA) model has many applications in signal processing, machine learning, and communications. In multi-view learning [34], for example, common factors would model information that is shared among all views and unique factors would account for effects that are specific to each view. Similarly, MFA could be used to fuse different modalities of brain imaging data (EEG, fMRI, and sMRI) [35], [36], where common factors account for information contained in all modalities, and unique factors are used to model information specific to each modality. As another example, the MFA model may have application in cellular networks that apply coordinated multipoint (CoMP) processing, where mobile users at the edge of a cell could be connected to several base stations (BS) thus playing the role of common factors in an MFA model. Each BS can also be receiving signals from a few specific users, either in the corresponding macro-cell or from a nearby small-cell, which would be the unique factors in each BS channel. This and similar multi-tier signal models are commonly used in heterogeneous cellular networks (HetNets) [37]. Finally, the proposed model is particularly relevant for passive radar since it accounts for leakage of a reference channel transmission into the surveillance channel [22]. We will describe with more detail the application of our MFA model to passive radar in Section IV.

The iterative procedure to obtain the maximum likelihood estimates of the multi-channel FA model developed in this paper bears resemblance to ML estimation in the FA model, where there also does not exist a closed-form solution. The iterative procedure consists of three steps and was derived following the block minorization-maximization approach [38], [39]. In the first step, a closed-form solution for the loading matrices of the common factors is found by maximizing the log-likelihood. In the second step, the estimate of the loading matrices for the uncommon factors is obtained by maximizing a global lower bound of the log-likelihood, similarly to the EM algorithm. The third step also returns a closed-form solution for the estimate of the diagonal noise covariance matrices by maximizing another minimizer obtained by linearizing the aforementioned EM-based lower bound. We prove that this algorithm converges to a stationary point of the log-likelihood and demonstrate its performance on several illustrative problems.

A. Outline

The outline of this paper is as follows: Section II summarizes the classical FA model, as well as an ML estimation procedure based on an alternating optimization approach. A brief introduction to inter-battery FA and the proposed generalization are presented in Section III. This section also describes the ML estimation of the unknown parameters. The alternating optimization ML algorithm is derived in Section III-B. Finally, in Section IV the performance of the proposed method is illustrated by means of numerical simulations, and the main conclusions are summarized in Section V.

B. Notation

In this paper, matrices are denoted by bold-faced upper case letters, bold-faced lower case letters denote column vectors, and scalars are denoted by light-face lower case letters. A real matrix of dimension $M \times N$ is denoted $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $\mathbf{x} \in \mathbb{R}^M$ indicates that $\mathbf{x}$ is a real vector of dimension M. The superscript $(\cdot)^T$ denotes transpose, and the determinant, Frobenius norm and trace of a matrix $\mathbf{A}$ are denoted $\det(\mathbf{A})$, $\|\mathbf{A}\|_F$ and $\mathrm{tr}(\mathbf{A})$, respectively. The notation $\mathbf{x} \sim \mathcal{N}_M(\boldsymbol{\mu}, \mathbf{R})$ indicates that $\mathbf{x}$ is an M-dimensional Gaussian random vector of mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$ and $E[\cdot]$ represents the expectation operator. The identity matrix of size $L \times L$ is $\mathbf{I}_L$, $\mathbf{0}_{M \times N}$ denotes the zero matrix of the dimension $M \times N$. We use $\mathbf{A}^{1/2}$ to denote the symmetric square root matrix of the symmetric matrix $\mathbf{A}$. Finally, $\mathrm{diag}(\mathbf{A})$ constructs a diagonal matrix from the diagonal of $\mathbf{A}$, the operator $\mathrm{blkdiag}$ denotes block-diagonal concatenation of matrices, and $\delta[n]$ denotes the Kronecker delta.

II. CLASSICAL FA

In single-channel (or classical) factor analysis (FA) [2]–[4], the real-valued observations $\mathbf{x} \in \mathbb{R}^L$ are modeled as[1]

$$\mathbf{x} = \mathbf{H}\mathbf{f} + \mathbf{e}, \tag{1}$$

where $\mathbf{f} \in \mathbb{R}^p$ contains the p factors, and $\mathbf{H} \in \mathbb{R}^{L \times p}$ is the factor loading matrix; p is usually much smaller than L. The L-dimensional noise vector $\mathbf{e}$ is typically assumed zero mean, Gaussian distributed and its components are independent, i.e., $\mathbf{e} \sim \mathcal{N}_L(\mathbf{0}, \mathbf{\Sigma})$, where the covariance matrix $\mathbf{\Sigma}$ is diagonal with positive elements. In classical FA, the factors $\mathbf{f}$ are assumed to be zero-mean Gaussian with identity covariance. As a consequence, the measurements $\mathbf{x}$ are zero-mean Gaussian with an $L \times L$ covariance matrix

$$\mathbf{R} = \mathbf{H}\mathbf{H}^T + \mathbf{\Sigma}. \tag{2}$$

That is, the covariance matrix is a positive semi-definite rank-p matrix plus a diagonal covariance $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_L^2)$, with $\sigma_i^2 > 0$. The FA model implies that, conditioned on the factors, the observations are uncorrelated, and hence the common factors explain all the dependence structure among the observations.

The invariances of this model determine the identifiability of this second-order model. The covariance model of (1) is invariant to the transformation $\mathbf{H} \longrightarrow \mathbf{H}\mathbf{Q}, \mathbf{f} \longrightarrow \mathbf{Q}^T\mathbf{f}$, where $\mathbf{Q}$ is any orthogonal matrix. The model is therefore unique only up to equivalence class of frames $\mathbf{H}$, denoted by the subspace $\langle \mathbf{H} \rangle$, which is a point on a Grassmann manifold of dimension p. As a unique representative of this class of equivalence, we take a loading matrix $\mathbf{H}$ such that its $p \times p$ upper block is lower triangular with positive ordered diagonal values. Clearly, this particular representative achieves the same log-likelihood in the second-order FA model as any other point in $\langle \mathbf{H} \rangle$. As we shall see, the choice of this unique representative is motivated to ensure the convergence of the proposed algorithm to a stationary point, but is otherwise irrelevant. Moreover, under the aforementioned assumption, any estimation procedure will provide a unique solution only when the number of factors p is sufficiently small in comparison to the dimension of the ambient space. A model is said to be generically identified if we can find a unique FA factorization as in (2) for almost every pair of matrices $\{\mathbf{H}, \mathbf{\Sigma}\}$ viewed as points in a parameter space of dimension $(Lp + L)$ [40]. The non-identifiable models therefore should live in a set of zero Lebesgue measure. According to this definition, it was proven in [41] that a necessary and sufficient condition for a FA model to be generically identified is

$$(L - p)^2 - (L + p) > 0. \tag{3}$$

Other definitions of identifiability are possible. In [17], a model is considered identifiable if the corresponding Fisher information matrix is nonsingular. Using this definition, it is shown in [17] that (3) is a necessary (but not sufficient) condition for the uniqueness of the solution.

[1]To simplify the exposition, the case of real-valued channels is considered throughout this work, but its extension to the complex-valued case is straightforward.

A. ML estimation in the FA model

Maximum likelihood is the most common principle for estimation in factor analysis. However, since it is not possible to find the ML estimates of $\{\mathbf{H}, \mathbf{\Sigma}\}$ in closed-form, solutions based on iterative procedures have been typically proposed. These include numerical procedures by Joreskog based on first-order or second-order derivatives [19], [20], alternating optimization methods [7], [21], [22], and EM-type algorithms [21], [23], [24].

In our experience, alternating optimization methods are preferable for moderate-size problems. For instance, the alternating optimization approach in [22] operates as follows. It starts with the likelihood function for N observations of $\mathbf{x}$, $\mathbf{x}[1], \ldots, \mathbf{x}[N]$, which is to be maximized with respect to the factor loading matrix $\mathbf{H}$ and the diagonal noise covariance matrix $\mathbf{\Sigma}$.

The likelihood of the observations is

$$l(\mathbf{H}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{LN/2} \det^{N/2}(\mathbf{R})} \exp\left[-\frac{N}{2}\text{tr}(\mathbf{R}^{-1}\mathbf{S})\right], \tag{4}$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}[n]\mathbf{x}^T[n], \tag{5}$$

is the sample covariance matrix. The ML estimation problem can be re-formulated as

$$\underset{\mathbf{H} \in \mathbb{H}, \mathbf{\Sigma}}{\text{maximize}} \quad \mathcal{L}(\mathbf{H}, \mathbf{\Sigma}), \tag{6}$$

where $\mathbb{H}$ denotes the set of structured $n \times p$ matrices such that its $p \times p$ upper block is lower triangular with positive ordered diagonal values and the objective function is

$$\mathcal{L}(\mathbf{H}, \mathbf{\Sigma}) = -\log\det(\mathbf{H}\mathbf{H}^T + \mathbf{\Sigma}) - \text{tr}\left[(\mathbf{H}\mathbf{H}^T + \mathbf{\Sigma})^{-1}\mathbf{S}\right]. \tag{7}$$

There is no closed-form solution to the problem (6), but it is possible to find a local maximum of the likelihood by applying an alternating optimization approach. Concretely, [22] proposed an algorithm to find the estimate of the precision matrix, which has an equivalence in terms of the parameters of the covariance matrix. Then, defining the noise-whitened sample covariance matrix

$$\tilde{\mathbf{S}} = \hat{\mathbf{\Sigma}}^{-1/2}\mathbf{S}\hat{\mathbf{\Sigma}}^{-1/2}, \tag{8}$$

and its eigenvalue decomposition (EVD)

$$\tilde{\mathbf{S}} = \tilde{\mathbf{W}}\text{diag}\left(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_L\right)\tilde{\mathbf{W}}^T, \tag{9}$$

with $\tilde{\lambda}_i \geq \tilde{\lambda}_{i+1}$, the estimate of $\mathbf{H}$ is

$$\hat{\mathbf{H}} = \hat{\mathbf{\Sigma}}^{1/2}\tilde{\mathbf{W}}\tilde{\mathbf{D}}^{1/2}\mathbf{Q}, \tag{10}$$

where $\hat{\mathbf{\Sigma}}$ is the previous estimate of $\mathbf{\Sigma}$, $\tilde{\mathbf{D}} = \text{diag}\left(\tilde{d}_1, \ldots, \tilde{d}_p, 0, \ldots, 0\right)$, with $\tilde{d}_i = \max(\tilde{\lambda}_i - 1, 0)$, and $\mathbf{Q}$ is the unique orthogonal matrix that imposes the structure of the set $\mathbb{H}$. Concretely, let $\mathbf{B}$ be the $p \times p$ upper block of $\hat{\mathbf{\Sigma}}^{1/2}\tilde{\mathbf{W}}\tilde{\mathbf{D}}^{1/2}$ and define its LQ decomposition as $\mathbf{L}\mathbf{Q}_{\mathbf{H}} = \mathbf{B}$ with the diagonal elements of $\mathbf{L}$ ordered in decreasing absolute value, then $\mathbf{Q}$ is given by

$$\mathbf{Q} = \mathbf{Q}_{\mathbf{H}}^T\text{sign}\left[\text{diag}(\mathbf{L})\right]. \tag{11}$$
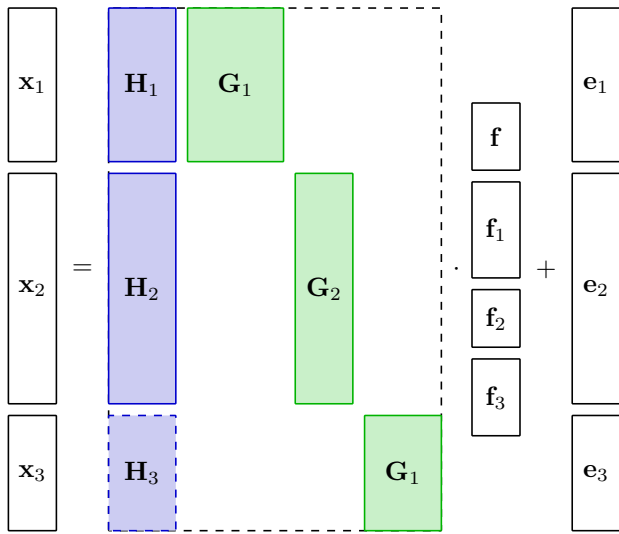
Fig. 1: Diagram of the proposed multi-channel factor analysis model for $M = 3$ channels. The observations are represented by $\mathbf{x}_i$; the loading matrices for the common factors $\mathbf{f}$ and for the unique factors $\mathbf{f}_i$ are depicted by $\mathbf{H}_i$ (blue) and $\mathbf{G}_i$ (green), respectively; $\mathbf{e}_i$ is the channel noise, for $i = 1, \ldots, M$.

Now, given $\hat{\mathbf{H}}$, the estimate of $\mathbf{\Sigma}$ is

$$\hat{\mathbf{\Sigma}} = \mathrm{diag}\left(\mathbf{S} - \hat{\mathbf{H}}\hat{\mathbf{H}}^T\right). \tag{12}$$

Since each step of the above procedure obtains the unique minimum of the the cost function, this alternating algorithm is ensured to attain a stationary point of the log-likelihood [42].

III. Multi-channel Factor Analysis

The first generalization of FA to more than one channel was introduced by Tucker in the fifties [27]. This generalization, known as inter-battery FA, aims at extracting factors $\mathbf{f}$ common to two sets of variables (or batteries), and is based on the generative model

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{H}_1\mathbf{f} + \mathbf{e}_1, \\ \mathbf{x}_2 &= \mathbf{H}_2\mathbf{f} + \mathbf{e}_2, \end{aligned} \tag{13}$$

where the covariance matrices of the noise vectors does not have any further structure besides being arbitrary positive definite matrices. In the work of Tucker, a solution is proposed based on a least squares (LS) criterion, which results in the singular value decomposition (SVD) of the sample cross-covariance matrix between the two data sets. Interestingly, a few decades later, Browne presented a connection between the inter-battery FA and canonical correlation analysis (CCA) in [28]. The extension of the inter-battery FA model to more than two channels was developed in [29], [30].

We propose the following generalization of inter-battery FA analysis. We consider $M \geq 2$ channels with noise covariance matrices that have further structure to account for the existence in each channel of a factor component that is unique to the channel. The generative model is

$$\mathbf{x}_i = \mathbf{H}_i\mathbf{f} + \mathbf{G}_i\mathbf{f}_i + \mathbf{e}_i, \quad i = 1, \ldots, M, \tag{14}$$

where $\mathbf{H}_i \in \mathbb{R}^{L_i \times p}$ is the loading matrix in channel i for the common factors $\mathbf{f}$ and $\mathbf{G}_i \in \mathbb{R}^{L_i \times p_i}$ is the loading matrix in channel i for the unique factors $\mathbf{f}_i$; $\mathbf{e}_i \sim \mathcal{N}_{L_i}(\mathbf{0}, \mathbf{\Sigma}_i)$ is the noise in channel i. This model is illustrated in Fig. 1 for $M = 3$. The noise covariance matrices, $\mathbf{\Sigma}_i$, are diagonal and invertible, and the noises at different channels are uncorrelated: $E[\mathbf{e}_i\mathbf{e}_j^T] = \mathbf{\Sigma}_i\delta[i - j]$. Moreover, common and specific factors are uncorrelated: $E[\mathbf{f}\,\mathbf{f}_i^T] = \mathbf{0}_{p \times p_i}, \forall i$, and $E[\mathbf{f}_i\mathbf{f}_j^T] = \mathbf{0}_{p_i \times p_j}$, for $i \neq j$. In this multi-channel generative model, the common factors explain the inter-channel dependence structure, whereas the unique factors explain the intra-channel dependence structure. Further, conditioned on both the common and unique factors, the multi-channel observations are uncorrelated. This structure makes our model different from other multi-channel models assumed in PARAFAC [32], IVA [33] or multiset CCA [43], [44].

As with single-channel FA, only the subspaces $\langle\mathbf{H}\rangle$ and $\langle\mathbf{G}_i\rangle$ can be identified. Thus, without loss of generality, we consider the factors to be normalized as follows: $E[\mathbf{f}\,\mathbf{f}^T] = \mathbf{I}_p$ and $E[\mathbf{f}_i\mathbf{f}_i^T] = \mathbf{I}_{p_i}$.

Assuming a multivariate normal model for common and uncommon factors, the composite vector $\mathbf{x} = [\mathbf{x}_1^T \cdots \mathbf{x}_M^T]^T$ is distributed as $\mathcal{N}_L(\mathbf{0}, \mathbf{R})$ with a structured covariance matrix that is

$$\mathbf{R} = \mathbf{H}\mathbf{H}^T + \mathbf{E}. \tag{15}$$

Here, the composite loading matrix is $\mathbf{H} = [\mathbf{H}_1^T \cdots \mathbf{H}_M^T]^T \in \mathbb{R}^{L \times p}$, with $L = \sum_{i=1}^M L_i$. The unique-factors-plus-noise covariance matrix is

$$\mathbf{E} = \mathrm{blkdiag}[\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_M] = \mathbf{G}\mathbf{G}^T + \mathbf{\Sigma}, \tag{16}$$

where the composite loading matrix for the uncommon factors and the composite noise covariance matrix are, respectively,

$$\mathbf{G} = \mathrm{blkdiag}[\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_M], \tag{17}$$

and

$$\mathbf{\Sigma} = \mathrm{blkdiag}[\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \ldots, \mathbf{\Sigma}_M]. \tag{18}$$

Moreover, $\mathbf{E}_i = \mathbf{G}_i\mathbf{G}_i^T + \mathbf{\Sigma}_i$, with $\mathbf{G}_i \in \mathbb{R}^{L_i \times p_i}$ and $\mathbf{\Sigma}_i \in \mathbb{R}^{L_i \times L_i}$.

The identification of a second-order MFA model is not unique due to the problem invariances. However, as explained in Sec. II, we can choose unique representatives for the loading matrices for the common and unique factors as follows. The unique solution for $\mathbf{H}$ is such that its $p \times p$ upper block is lower triangular with positive ordered diagonal elements and, similarly, the $p_i \times p_i$ upper block of $\mathbf{G}_i$ is also lower triangular with positive ordered diagonal elements. We will denote the sets of matrices with this structure by $\mathbb{H}$ and $\mathbb{G}_i$, respectively, and $\mathbb{G}$ denotes the set of matrices $\mathbf{G} = \mathrm{blkdiag}[\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_M]$, with $\mathbf{G}_i \in \mathbb{G}_i$. These constraints yield unique solutions for the loading matrices, and are necessary to ensure the convergence of the alternating optimization algorithm to a stationary point. Moreover, as in single-channel FA, the MFA model is not identifiable without a constraint on the number of parameters to be identified. In the following, we present a necessary condition on the largest number of common and specific factors that yield a unique solution for the covariance

matrix. To do so, we need to count how many knowns and unknowns the model has. The number of knowns is given by the number of different elements of the sample covariance matrix, which are $L(L+1)/2$. The number of unknowns is slightly more involved to compute. Let us start with the number of unknowns in $\mathbf{E}_i$, which are those of the classical FA model, i.e., $L_i + L_i p_i - p_i(p_i - 1)/2$. Finally, since the number of unknowns in $\mathbf{H}\mathbf{H}^T$ is $Lp - p(p-1)/2$, for the solution to be unique it is required that

$$\frac{L(L+1)}{2} - Lp + \frac{p(p-1)}{2}$$
$$- \sum_{i=1}^{M} \left(L_i + L_i p_i - \frac{p_i(p_i - 1)}{2} \right) > 0. \quad (19)$$

Additionally, after removing the common factors each single-channel FA model must be also identifiable, which adds the following conditions

$$(L_i - p_i)^2 - (L_i + p_i) > 0, \qquad i = 1, \dots, M. \quad (20)$$

A. ML estimation in the MFA model

In this section, we present the ML estimation of the unknown parameters in the MFA model. In particular, assuming that N observations of each channel are available, $\mathbf{x}_i[1], \dots, \mathbf{x}_i[N], i = 1, \dots, M$, the goal is to estimate the composite common-factor loading matrix $\mathbf{H}$, the composite channel-specific loading matrix $\mathbf{G}$, and the composite diagonal noise covariance matrix $\mathbf{\Sigma}$ that maximize the log-likelihood. Hence, the ML estimates of $\mathbf{H}, \mathbf{G}$, and $\mathbf{\Sigma}$ are obtained by solving the maximization problem

$$\underset{\mathbf{H} \in \mathbb{H}, \mathbf{G} \in \mathbb{G}, \mathbf{\Sigma}}{\text{maximize}} \quad \mathcal{L}(\mathbf{H}, \mathbf{G}, \mathbf{\Sigma}), \quad (21)$$

where the objective function is

$$\mathcal{L}(\mathbf{H}, \mathbf{G}, \mathbf{\Sigma}) = -\log \det(\mathbf{R}) - \text{tr}(\mathbf{R}^{-1}\mathbf{S}), \quad (22)$$

with $\mathbf{R}$ given in (15) and the sample covariance matrix given in (5) with the vector of multi-channel observations $\mathbf{x}[n] = [\mathbf{x}_1^T[n] \cdots \mathbf{x}_M^T[n]]^T$.

The maximization problem in (21) does not have a closed-form solution. In this work, we propose therefore to use an alternating optimization approach, as described in the following section.

B. Alternating Optimization Algorithm

We propose a cyclic alternating-optimization algorithm for maximizing the log-likelihood function in (22) subject to the constraints that ensure the uniqueness of the loading matrices for common and unique factors. The procedure applies three steps in a cyclic fashion, which are derived using the block minorization-maximization (BMM) framework [38], [39]. At each of the three steps, a subset of variables is optimized by maximizing a global lower bound of the cost function, while the remaining variables are fixed at previously estimated values. The fixed parameters at the $(k + 1)$th iteration are denoted by $\hat{\mathbf{H}}^{(k)}$, $\hat{\mathbf{G}}^{(k)}$, and $\hat{\mathbf{\Sigma}}^{(k)}$, whereas the parameters to be optimized are denoted without a hat. That is, $\mathcal{L}(\mathbf{H}, \hat{\mathbf{G}}^{(k)}, \hat{\mathbf{\Sigma}}^{(k)})$ is the objective function at the $(k + 1)$th iteration for fixed values of $\mathbf{G}$ and $\mathbf{\Sigma}$.

a) Step 1: Estimation of $\mathbf{H}$*:* The first step of the proposed method consists in estimating $\mathbf{H}$, assuming that $\mathbf{G}$ and $\mathbf{\Sigma}$ are fixed. Thus, the optimization problem at the $(k+1)$th iteration is

$$(\mathcal{P}_1) \quad \underset{\mathbf{H} \in \mathbb{H}}{\text{maximize}} \quad \mathcal{L}(\mathbf{H}, \hat{\mathbf{G}}^{(k)}, \hat{\mathbf{\Sigma}}^{(k)}). \quad (23)$$

In this case, it is possible to obtain a closed-form solution by maximizing the log-likelihood directly, that is, no lower bound is necessary. To do so, we may define

$$\hat{\mathbf{E}} = \hat{\mathbf{G}}^{(k)}\hat{\mathbf{G}}^{(k)T} + \hat{\mathbf{\Sigma}}^{(k)}, \quad (24)$$

which is fixed since $\hat{\mathbf{\Sigma}}^{(k)}$ and $\hat{\mathbf{G}}^{(k)}$ are fixed at their values of the kth iteration. The whitened sample covariance matrix and its eigenvalue decomposition are

$$\tilde{\mathbf{S}} = \hat{\mathbf{E}}^{-1/2}\mathbf{S}\hat{\mathbf{E}}^{-1/2} = \tilde{\mathbf{W}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{W}}^T \quad (25)$$

where $\tilde{\mathbf{\Lambda}} = \text{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_L)$ with $\tilde{\lambda}_i \geq \tilde{\lambda}_{i+1}$.

From the original result of Anderson [45], the solution to (23) is

$$\hat{\mathbf{H}}^{(k+1)} = \hat{\mathbf{E}}^{1/2}\tilde{\mathbf{W}}\tilde{\mathbf{D}}^{1/2}\mathbf{Q}, \quad (26)$$

where $d_i = \max(\tilde{\lambda}_i - 1, 0)$, p is the number of common factors, and $\mathbf{Q}$ is an orthogonal matrix selected to impose the structure of the set $\mathbb{H}$, which is obtained as in Section II-A. A consequence of this result is that the value of $\mathbf{H}\mathbf{H}^T$ that maximizes the log-likelihood is

$$\hat{\mathbf{H}}^{(k+1)}\hat{\mathbf{H}}^{(k+1)T} = \hat{\mathbf{E}}^{1/2}\tilde{\mathbf{W}}\tilde{\mathbf{D}}\tilde{\mathbf{W}}^T\hat{\mathbf{E}}^{1/2}.$$

b) Step 2: Estimation of $\mathbf{G}$*:* In this step, the channel-specific loading matrices, $\mathbf{G}_i$, for fixed $\mathbf{H}$ and $\mathbf{\Sigma}$, are estimated. The optimization problem at the $(k + 1)$th iteration is

$$(\mathcal{P}_2) \quad \underset{\mathbf{G} \in \mathbb{G}}{\text{maximize}} \quad \mathcal{L}(\hat{\mathbf{H}}^{(k+1)}, \mathbf{G}, \hat{\mathbf{\Sigma}}^{(k)}), \quad (27)$$

which has no closed-form solution. Following the BMM framework, we propose to find a global lower-bound based on the EM approach and maximize this bound. Interestingly, we will show that this step amounts to removing the loading matrix for the common factors $\hat{\mathbf{H}}^{(k+1)}$ from the corresponding block in the diagonal of the sample covariance. Then, apply Anderson's result [45] and select the unique loading matrix with the required structure.

Here is the idea. After marginalization with respect to the Gaussian factors $\mathbf{f}$, the model for the measurement $\mathbf{x}$ is $\mathbf{x} \sim \mathcal{N}_L(\mathbf{0}, \mathbf{H}\mathbf{H}^T + \mathbf{G}\mathbf{G}^T + \mathbf{\Sigma})$. The problem is to find joint ML estimates for $\{\mathbf{H}, \mathbf{G}, \mathbf{\Sigma}\}$ in this model. We might say we started with the joint distribution for $\{\mathbf{x}, \mathbf{f}\}$, with $\mathbf{x}$ normally distributed, conditioned on $\mathbf{f}$, and $\mathbf{f}$ normal. The distribution of $\mathbf{f}$ is conjugate with respect to the conditional distribution of $\mathbf{x}$, so the marginalization of the joint distribution of $\{\mathbf{x}, \mathbf{f}\}$ is easy, producing the second-order normal distribution $\mathcal{N}_L(\mathbf{0}, \mathbf{H}\mathbf{H}^T + \mathbf{G}\mathbf{G}^T + \mathbf{\Sigma})$. But the maximization of the likelihood in this second-order model with respect to $\mathbf{G}$ is intractable, even in an alternating maximization with $\mathbf{H}$ and $\mathbf{\Sigma}$ fixed at their most recent estimates.

So, we replace the model $\mathcal{N}_L(\mathbf{0}, \mathbf{H}\mathbf{H}^T + \mathbf{G}\mathbf{G}^T + \mathbf{\Sigma})$ with the conditional first-order model $\mathcal{N}_L(\mathbf{H}\mathbf{f}, \mathbf{G}\mathbf{G}^T + \mathbf{\Sigma})$, treating $\mathbf{f}$ as an unmeasured random effect, and proceed with an EM-based

lower bound in this first-order model, which is maximized to actually find a $\hat{\mathbf{G}}$ that increases the log-likelihood in the second-order model $\mathbf{x} \sim \mathcal{N}_L(\mathbf{0}, \mathbf{HH}^T + \mathbf{GG}^T + \boldsymbol{\Sigma})$, for fixed $\{\mathbf{H}, \boldsymbol{\Sigma}\}$.

Now, we will compute the lower bound by only assuming that $\mathbf{H}$ is fixed. We do not consider fixed $\boldsymbol{\Sigma}$ as the bound derived here will be also useful in the next step. Then, let us rewrite the log-likelihood for fixed $\mathbf{H}$ as follows

$$\mathcal{L}\left(\hat{\mathbf{H}}^{(k+1)}, \mathbf{G}, \boldsymbol{\Sigma}\right) = \sum_{n=1}^{N} \log l\left(\mathbf{x}[n]|\mathbf{G}, \boldsymbol{\Sigma}\right) \qquad (28)$$

where $l\left(\mathbf{x}[n]|\mathbf{G}, \boldsymbol{\Sigma}\right)$ is the likelihood of $\mathbf{x}[n]$ for $\mathbf{H}$ fixed to $\hat{\mathbf{H}}^{(k+1)}$, and the equality is up to constant terms that do not modify the optimization problem. The factors $\mathbf{f}[n]$, with $\mathbf{f}[n]$ denoting the nth realization of $\mathbf{f}$, are considered unmeasured random effects, or hidden data, so the augmented data is $\{\mathbf{x}[n], \mathbf{f}[n]\}$ [21], [23], [24]. For this choice of the augmented data set, the lower-bound on $\mathcal{L}\left(\hat{\mathbf{H}}^{(k+1)}, \mathbf{G}, \boldsymbol{\Sigma}\right)$ is

$$\mathcal{L}\left(\hat{\mathbf{H}}^{(k+1)}, \mathbf{G}, \boldsymbol{\Sigma}\right) \geq Q\left(\mathbf{G}, \boldsymbol{\Sigma}\right) + C, \qquad (29)$$

where C is a constant and

$$Q\left(\mathbf{G}, \boldsymbol{\Sigma}\right) = \\ \sum_{n=1}^{N} E\left[\log l\left(\mathbf{x}[n]|\mathbf{f}[n], \mathbf{G}, \boldsymbol{\Sigma}, \hat{\mathbf{H}}^{(k+1)}, \hat{\mathbf{G}}^{(k)}, \hat{\boldsymbol{\Sigma}}^{(k)}\right)\right]. \qquad (30)$$

The conditional distribution of $\mathbf{x}[n]$ given $\mathbf{f}[n]$ and the parameters of the previous iteration is

$$\mathbf{x}[n]|\mathbf{f}[n] \sim \mathcal{N}_L\left(\hat{\mathbf{H}}^{(k+1)}\mathbf{f}[n], \mathbf{E}\right), \qquad (31)$$

where $\mathbf{E} = \mathbf{GG}^T + \boldsymbol{\Sigma}$. Then, the expectation in (30) becomes (up to constant terms)

$$E\left[\log l\left(\mathbf{x}[n]|\mathbf{f}[n], \mathbf{G}, \boldsymbol{\Sigma}, \hat{\mathbf{H}}^{(k+1)}, \hat{\mathbf{G}}^{(k)}, \hat{\boldsymbol{\Sigma}}^{(k)}\right)\right] = \\ -\frac{1}{2}\mathbf{x}[n]^T\mathbf{E}^{-1}\mathbf{x}[n] + \mathbf{x}[n]^T\mathbf{E}^{-1}\hat{\mathbf{H}}^{(k+1)}\hat{\mathbf{f}}[n] \\ -\frac{1}{2}\log\det\left(\mathbf{E}\right) - \frac{1}{2}\text{tr}\left(\hat{\mathbf{H}}^{(k+1)T}\mathbf{E}^{-1}\hat{\mathbf{H}}^{(k+1)}\mathbf{C}[n]\right), \qquad (32)$$

where

$$\hat{\mathbf{f}}[n] = E\left[\mathbf{f}[n]|\mathbf{x}[n]\right] = \mathbf{Wx}[n], \qquad (33)$$

is the expected value of the factors, which is the minimum mean squared estimator (MMSE) of $\mathbf{f}[n]$ given $\mathbf{x}[n]$, and

$$\mathbf{C}[n] = E\left[\mathbf{f}[n]\mathbf{f}[n]^T|\mathbf{x}[n]\right] = \mathbf{Wx}[n]\mathbf{x}^T[n]\mathbf{W}^T \\ + \left(\mathbf{I} + \hat{\mathbf{H}}^{(k+1)T}\hat{\mathbf{E}}^{-1}\hat{\mathbf{H}}^{(k+1)}\right)^{-1}, \qquad (34)$$

is the second order moment of the factors given the observations. In (33) and (34), the MMSE matrix $\mathbf{W}$ is

$$\mathbf{W} = \hat{\mathbf{H}}^{(k+1)T}\left(\hat{\mathbf{H}}^{(k+1)}\hat{\mathbf{H}}^{(k+1)T} + \hat{\mathbf{E}}\right)^{-1}, \qquad (35)$$

where $\hat{\mathbf{E}}$ is defined in (24). Plugging now (33) and (34) into (30) yields

$$Q\left(\mathbf{G}, \boldsymbol{\Sigma}\right) = -\frac{N}{2}\log\det\left(\mathbf{E}\right) \\ -\frac{N}{2}\text{tr}\left[\mathbf{E}^{-1}\left(\left(\mathbf{I} - 2\hat{\mathbf{H}}^{(k+1)}\mathbf{W}\right)\mathbf{S} + \hat{\mathbf{H}}^{(k+1)}\bar{\mathbf{C}}\hat{\mathbf{H}}^{(k+1)T}\right)\right], \qquad (36)$$

where

$$\bar{\mathbf{C}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{C}[n] = \left(\mathbf{I} + \hat{\mathbf{H}}^{(k+1)T}\hat{\mathbf{E}}^{-1}\hat{\mathbf{H}}^{(k+1)}\right)^{-1} + \mathbf{WSW}^T. \qquad (37)$$

Exploiting the block-diagonal structure of $\mathbf{E}$, $Q\left(\mathbf{G}, \boldsymbol{\Sigma}\right)$ can be written as

$$Q\left(\mathbf{G}, \boldsymbol{\Sigma}\right) = -\frac{N}{2}\sum_{i=1}^{M}\left[\log\det\left(\mathbf{G}_i\mathbf{G}_i^T + \boldsymbol{\Sigma}_i\right) \\ + \text{tr}\left(\left(\mathbf{G}_i\mathbf{G}_i^T + \boldsymbol{\Sigma}_i\right)^{-1}\mathbf{T}_i\right)\right], \qquad (38)$$

where $\mathbf{T}_i$ is the ith block of the appropriate dimensions in the diagonal of

$$\mathbf{T} = \left(\mathbf{I} - 2\hat{\mathbf{H}}^{(k+1)}\mathbf{W}\right)\mathbf{S} + \hat{\mathbf{H}}^{(k+1)}\bar{\mathbf{C}}\hat{\mathbf{H}}^{(k+1)T}. \qquad (39)$$

The following lemma allows a further simplification of the expected log-likelihood function.

Lemma 1: The matrices $\mathbf{T}$ and $\mathbf{P} = \mathbf{S} - \hat{\mathbf{H}}^{(k+1)}\hat{\mathbf{H}}^{(k+1)T}$ are identical.

Proof: Using (26), $\mathbf{T}$ can be written as

$$\mathbf{T} = \hat{\mathbf{E}}^{1/2}\hat{\mathbf{W}}\boldsymbol{\Delta}\hat{\mathbf{W}}^T\hat{\mathbf{E}}^{1/2}, \qquad (40)$$

where

$$\boldsymbol{\Delta} = \text{diag}\left(\delta_1, \ldots, \delta_p, \tilde{\lambda}_{p+1}, \ldots, \tilde{\lambda}_L\right), \qquad (41)$$

with $\delta_j = \min(\tilde{\lambda}_j, 1)$. On the other hand, substituting $\hat{\mathbf{H}}^{(k+1)}\hat{\mathbf{H}}^{(k+1)T} = \hat{\mathbf{E}}^{1/2}\tilde{\mathbf{W}}\tilde{\mathbf{D}}\tilde{\mathbf{W}}^T\hat{\mathbf{E}}^{1/2}$ into $\mathbf{P}$, we obtain

$$\mathbf{P} = \mathbf{S} - \hat{\mathbf{H}}^{(k+1)}\hat{\mathbf{H}}^{(k+1)T} \\ = \hat{\mathbf{E}}^{1/2}\hat{\mathbf{W}}\left(\tilde{\boldsymbol{\Lambda}} - \tilde{\mathbf{D}}\right)\hat{\mathbf{W}}^T\hat{\mathbf{E}}^{1/2} \\ = \hat{\mathbf{E}}^{1/2}\hat{\mathbf{W}}\boldsymbol{\Delta}\hat{\mathbf{W}}^T\hat{\mathbf{E}}^{1/2} = \mathbf{T}, \qquad (42)$$

which proves the lemma.

□

From this result, we finally find that the expected log-likelihood function can be written as

$$Q\left(\mathbf{G}, \boldsymbol{\Sigma}\right) = -\frac{N}{2}\sum_{i=1}^{M}\left[\log\det\left(\mathbf{G}_i\mathbf{G}_i^T + \boldsymbol{\Sigma}_i\right) \\ + \text{tr}\left(\left(\mathbf{G}_i\mathbf{G}_i^T + \boldsymbol{\Sigma}_i\right)^{-1}\mathbf{P}_i\right)\right], \qquad (43)$$

where

$$\mathbf{P}_i = \mathbf{S}_i - \hat{\mathbf{H}}_i^{(k+1)}\hat{\mathbf{H}}_i^{(k+1)T} \qquad (44)$$

and $\mathbf{S}_i$ is the sample covariance matrix of the ith channel. The interesting point from this derivation is that maximizing the

lower bound can be decoupled into M standard FA problems, that is,

$$Q\left(\mathbf{G}, \boldsymbol{\Sigma}\right) = \sum_{i=1}^{M} Q_i\left(\mathbf{G}_i, \boldsymbol{\Sigma}_i\right) \qquad (45)$$

where

$$Q_i\left(\mathbf{G}_i, \boldsymbol{\Sigma}_i\right) = -\frac{N}{2} \log \det\left(\mathbf{G}_i\mathbf{G}_i^T + \boldsymbol{\Sigma}_i\right)$$
$$+ \operatorname{tr}\left(\left(\mathbf{G}_i\mathbf{G}_i^T + \boldsymbol{\Sigma}_i\right)^{-1}\mathbf{P}_i\right), \qquad (46)$$

is identical to (7). To obtain the low-rank component $\mathbf{G}_i$ that models the loading matrix for the unique factors, we have to maximize $Q_i\left(\mathbf{G}_i, \hat{\boldsymbol{\Sigma}}_i^{(k)}\right)$ for fixed $\hat{\boldsymbol{\Sigma}}_i^{(k)}$. To this end, we proceed as before. Defining the whitened version of $\mathbf{P}_i$ and its EVD as

$$\tilde{\mathbf{P}}_i = \left[\hat{\boldsymbol{\Sigma}}_i^{(k)}\right]^{-1/2} \mathbf{P}_i \left[\hat{\boldsymbol{\Sigma}}_i^{(k)}\right]^{-1/2} = \tilde{\mathbf{W}}_i \tilde{\boldsymbol{\Lambda}}_i \tilde{\mathbf{W}}_i^T \qquad (47)$$

where $\tilde{\boldsymbol{\Lambda}}_i = \operatorname{diag}(\tilde{\lambda}_{i,1}, \ldots, \tilde{\lambda}_{i,L_i})$ with $\tilde{\lambda}_{i,j} \geq \tilde{\lambda}_{i,j+1}$, the value of $\mathbf{G}_i$ that maximizes the lower bound is found from the fundamental result of Anderson [45]:

$$\hat{\mathbf{G}}_i^{(k+1)} = \left[\hat{\boldsymbol{\Sigma}}_i^{(k)}\right]^{1/2} \tilde{\mathbf{W}}_i \tilde{\mathbf{D}}_i^{1/2} \mathbf{Q}_i. \qquad (48)$$

Here, $\tilde{\mathbf{D}}_i = \operatorname{diag}(d_{i,1}, \ldots, d_{i,p_i}, 0, \ldots, 0)$, with $d_{i,j} = \max(\tilde{\lambda}_{i,j} - 1, 0)$ and p_i is the number of specific factors in the ith channel. In addition, the orthogonal matrix $\mathbf{Q}_i$ is computed like in Step 1 to ensure that $\hat{\mathbf{G}}_i^{(k+1)} \in \mathbb{G}_i$.

c) Step 3: Estimation of $\boldsymbol{\Sigma}$: The last step of the proposed algorithm is to estimate $\boldsymbol{\Sigma}$ as the solution to the optimization problem

$$(\mathcal{P}_3) \quad \underset{\boldsymbol{\Sigma}}{\operatorname{maximize}} \quad \mathcal{L}\left(\hat{\mathbf{H}}^{(k+1)}, \hat{\mathbf{G}}^{(k+1)}, \boldsymbol{\Sigma}\right). \qquad (49)$$

The problem $(\mathcal{P}_3)$ does not admit a closed-form solution, and neither does the maximization of the lower bound $Q_i\left(\hat{\mathbf{G}}_i^{(k+1)}, \boldsymbol{\Sigma}_i\right)$, derived in the previous step. Nevertheless, it is possible to find yet another lower bound on $Q_i\left(\hat{\mathbf{G}}_i^{(k+1)}, \boldsymbol{\Sigma}_i\right)$, which admits a closed-form maximizer, as follows.

Defining $\boldsymbol{\Phi}_i = \boldsymbol{\Sigma}_i^{-1}$, we may rewrite $Q_i\left(\hat{\mathbf{G}}_i^{(k+1)}, \boldsymbol{\Sigma}_i\right)$ as

$$Q_i\left(\hat{\mathbf{G}}_i^{(k+1)}, \boldsymbol{\Phi}_i^{-1}\right) = R_1(\boldsymbol{\Phi}_i) - R_2(\boldsymbol{\Phi}_i), \qquad (50)$$

where

$$R_1(\boldsymbol{\Phi}_i) = \operatorname{tr}(\log \boldsymbol{\Phi}_i) - \operatorname{tr}(\mathbf{P}_i \boldsymbol{\Phi}_i) \qquad (51)$$

and

$$R_2(\boldsymbol{\Phi}_i) = \sum_{j=1}^{p_i} \log(\max(\tilde{\beta}_{i,j}, 1)) - \max(\tilde{\beta}_{i,j}, 1) + 1 \qquad (52)$$

with $\log$ denoting the matrix logarithm and $\tilde{\beta}_{i,j}$ being the jth eigenvalue of $\boldsymbol{\Phi}_i^{1/2} \mathbf{P}_i \boldsymbol{\Phi}_i^{1/2}$. Clearly, $R_1(\boldsymbol{\Phi}_i)$ and $R_2(\boldsymbol{\Phi}_i)$ are both concave functions, but their difference is not concave. Khamaru and Mazumder in [46] propose a global lower-bound

of $Q_i\left(\hat{\mathbf{G}}_i^{(k+1)}, \boldsymbol{\Phi}_i^{-1}\right)$ based on the linearization of $R_2(\boldsymbol{\Phi}_i)$ using the subgradient at $\hat{\boldsymbol{\Sigma}}_i^{(k)}$, which yields

$$\bar{Q}_i\left(\hat{\mathbf{G}}_i^{(k+1)}, \boldsymbol{\Phi}_i^{-1}\right) = \operatorname{tr}(\log \boldsymbol{\Phi}_i) - \operatorname{tr}(\mathbf{P}_i \boldsymbol{\Phi}_i) + \operatorname{tr}(\mathbf{A}_i \boldsymbol{\Phi}_i), \qquad (53)$$

where the subgradient with respect to $\boldsymbol{\Phi}_i$ is

$$\mathbf{A}_i = \operatorname{diag}\left(\left[\hat{\boldsymbol{\Sigma}}_i^{(k)}\right]^{1/2} \tilde{\mathbf{W}}_i \tilde{\boldsymbol{\Xi}}_i \tilde{\mathbf{W}}_i^T \left[\hat{\boldsymbol{\Sigma}}_i^{(k)}\right]^{-1/2} \mathbf{P}_i\right), \qquad (54)$$

with

$$\boldsymbol{\Xi}_i = \operatorname{diag}(\xi_{i,1}, \ldots, \xi_{i,p_i}, 0, \ldots, 0), \qquad (55)$$

and $\xi_{i,j} = \max(1 - 1/\tilde{\lambda}_{i,j}, 0)$. Thus, we have that

$$\mathcal{L}\left(\hat{\mathbf{H}}^{(k+1)}, \hat{\mathbf{G}}^{(k+1)}, \boldsymbol{\Sigma}\right)$$
$$\geq Q\left(\hat{\mathbf{G}}^{(k+1)}, \boldsymbol{\Sigma}\right) + C \geq \bar{Q}\left(\hat{\mathbf{G}}^{(k+1)}, \boldsymbol{\Sigma}\right) + C, \qquad (56)$$

where

$$\bar{Q}\left(\mathbf{G}^{(k+1)}, \boldsymbol{\Sigma}\right) = \sum_{i=1}^{M} \bar{Q}_i\left(\hat{\mathbf{G}}_i^{(k+1)}, \boldsymbol{\Phi}_i^{-1}\right). \qquad (57)$$

Finally, the maximizer of (53) is given by

$$\hat{\boldsymbol{\Phi}}_i^{-1} = \operatorname{diag}\left(\mathbf{P}_i - \mathbf{A}_i\right). \qquad (58)$$

After some straightforward manipulations, we may rewrite the subgradient in (54) as

$$\mathbf{A}_i = \operatorname{diag}\left(\hat{\mathbf{G}}_i^{(k+1)} \hat{\mathbf{G}}_i^{(k+1)T}\right), \qquad (59)$$

which yields

$$\hat{\boldsymbol{\Sigma}}_i^{(k+1)} = \operatorname{diag}\left(\mathbf{P}_i - \hat{\mathbf{G}}_i^{(k+1)} \hat{\mathbf{G}}_i^{(k+1)T}\right). \qquad (60)$$

or, equivalently,

$$\hat{\boldsymbol{\Sigma}}^{(k+1)} = \operatorname{diag}\left(\mathbf{S} - \hat{\mathbf{H}}^{(k+1)} \hat{\mathbf{H}}^{(k+1)T} - \hat{\mathbf{G}}^{(k+1)} \hat{\mathbf{G}}^{(k+1)T}\right). \qquad (61)$$

The non-negativity of the elements in the diagonal of $\boldsymbol{\Sigma}$ has not been imposed. However, taking into account (48), it is easy to show that the elements of $\hat{\boldsymbol{\Sigma}}_i$ are indeed non-negative.

C. Initialization and convergence

The algorithm for ML multi-channel factor analysis, or ML-MFA, is initialized at $\hat{\boldsymbol{\Sigma}}^{(0)} = \mathbf{I}_L$ and $\hat{\mathbf{G}}_i^{(0)} = \mathbf{0}_{L_i \times p_i}$, respectively. A smarter initialization of $\boldsymbol{\Sigma}$, which could achieve faster convergence for small signal-to-noise ratios [4], is $\hat{\boldsymbol{\Sigma}}^{(0)} = \operatorname{diag}(\mathbf{S})$. A summary of the ML-MFA algorithm is presented in Algorithm 1.

The following theorem proves the convergence of the ML-MFA algorithm to a stationary point of (21).

Theorem 1: Denote by

$$\{\hat{\boldsymbol{\Theta}}^{(k)}\} = \{\hat{\mathbf{H}}^{(k)}, \hat{\mathbf{G}}^{(k)}, \hat{\boldsymbol{\Sigma}}^{(k)}\} \qquad (62)$$

the sequence of iterates generated by Algorithm 1. Then, assuming that the problem is identifiable and that the solution at each iteration has positive noise variances (i.e., the so-called Heywood cases [4] are excluded), the sequence $\{\hat{\boldsymbol{\Theta}}^{(k)}\}$ converges to a stationary point $\hat{\boldsymbol{\Theta}}^*$ of (21).

Proof: See Appendix I. $\qquad\square$

Algorithm 1 ML-MFA algorithm.

1: **Initialize**: $k = 0, \hat{\boldsymbol{\Sigma}}^{(0)} = \mathbf{I}_L$ and $\hat{\mathbf{G}}_i^{(0)} = \mathbf{0}_{L_i \times p_i}$
2: **repeat**
3:     Step 1: Estimate $\hat{\mathbf{H}}^{(k+1)}$ according to (26)
4:     Cancel out the effect of the common factors using (44)
5:     Step 2: Estimate $\hat{\mathbf{G}}_i^{(k+1)}$ following (48)
6:     Step 3: Estimate $\hat{\boldsymbol{\Sigma}}_i^{(k+1)}$ as in (60)
7:     $k = k + 1$
8: **until** convergence
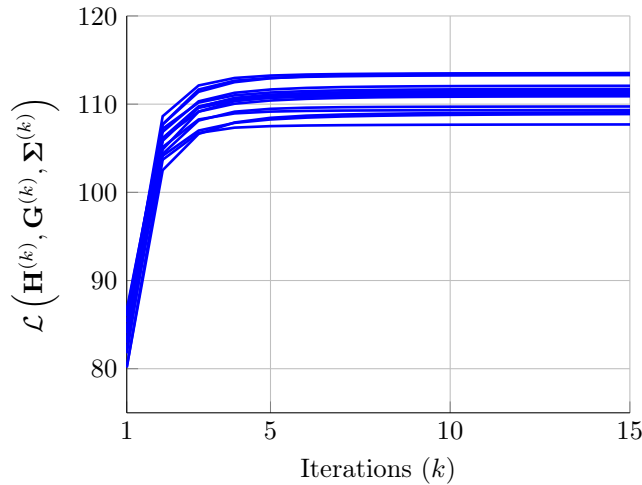
Fig. 2: Convergence of the ML-MFA algorithm.

IV. NUMERICAL RESULTS

A. Demonstrating convergence

In the first example, we study the convergence of Algorithm 1 by considering $M = 3$ channels of dimensions $L_1 = 20, L_2 = 15$, and $L_3 = 10$. The number of observations is $N = 100$, and the number of common and unique factors are, respectively, $p = 2$ and $p_1 = 4, p_2 = 3$, and $p_3 = 2$. Moreover, the power ratio explained by the common, unique, and noise components for the ith channel with respect to the total power are given by

$$\eta_c = \text{tr}(\mathbf{H}_i \mathbf{H}_i^T)/\text{tr}(\mathbf{R}_i) = 0.3, \tag{63}$$

$$\eta_s = \text{tr}(\mathbf{G}_i \mathbf{G}_i^T)/\text{tr}(\mathbf{R}_i) = 0.3, \tag{64}$$

$$\eta_n = \text{tr}(\boldsymbol{\Sigma}_i)/\text{tr}(\mathbf{R}_i) = 0.4, \tag{65}$$

where $\mathbf{R}_i$ is the covariance matrix of the ith channel:

$$\mathbf{R}_i = \mathbf{H}_i \mathbf{H}_i^T + \mathbf{G}_i \mathbf{G}_i^T + \boldsymbol{\Sigma}_i. \tag{66}$$

Note that, for simplicity, the power ratios for all channels are identical, although it is straightforward to extend the model to unequal power ratios.

The results for this example are shown in Fig. 2, where the convergence curves for 15 runs of the proposed method are plotted. The loading and covariance matrices are randomly generated. That is, the model is different in each run. Consequently the value of achieved log-likelihood varies from run-to-run.
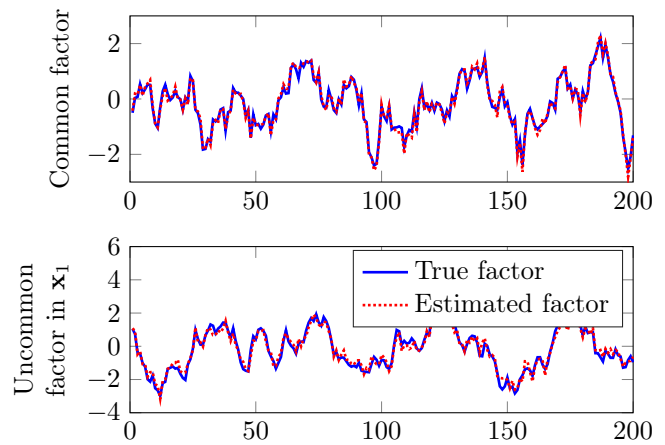

Fig. 3: Estimated and true factors.

B. Estimating the common and unique factors

In the second example, the identification of the composite covariance matrix for all channels is used in uncoupled MMSE estimates of common and unique factors:

$$\hat{\mathbf{f}}[n] = \hat{\mathbf{H}}^T \hat{\mathbf{R}}^{-1} \mathbf{x}[n], \tag{67}$$

$$\hat{\mathbf{f}}_i[n] = \hat{\mathbf{G}}_i^T \hat{\mathbf{R}}_i^{-1} \mathbf{x}_i[n], \tag{68}$$

where $\hat{\mathbf{R}}_i$ is the ith block in the diagonal of $\hat{\mathbf{R}}$. The results are shown in Fig. 3 for an experiment with $p = p_i = 1$ factors, which are now AR(1) signals,[2] and $N = 1000$. The remaining parameters of the measurement model are those in the previous example. As can be seen in Fig. 3, the estimated factors in this scenario are nearly identical to the true factors.

C. Mean-squared error of the estimated covariance model

The next example compares the performance of the MFA method, a naive method that applies a single-channel FA algorithm whose solution for $\mathbf{H}$ is projected afterwards onto the set of matrices with the structure in Fig. 1, and a two-step method that consists on first applying MAXVAR-CCA [43] to estimate the common factors, then cancel out their effect using (44), and finally applying single-channel FA to each $\mathbf{P}_i$. This combined CCA+FA approach may be interpreted as a variation of the proposed ML-MFA method, where only one iteration of Step 1 is taken. Moreover, we also include in the comparison the group factor analysis (GFA) model proposed in [31]. GFA learns a structured sparse FA model so that the factor loading matrix is group-wise sparse. Sparsity in GFA is enforced by assuming independent gamma distributions as the precision parameter of the prior distribution for the elements of the loading matrix, and approximate inference is performed using the mean-field variational approximation. A final point to mention is that, while MFA, CCA+FA, and the naive method need an estimate of the number of common and unique factors, GFA only needs to know the total number of factors, K, and the variational optimization procedure finds

---

[2]The temporal correlation induced by the AR(1) model is only used for visualization purposes and not exploited in the estimation algorithm, which still considers independent and identically distributed observations.
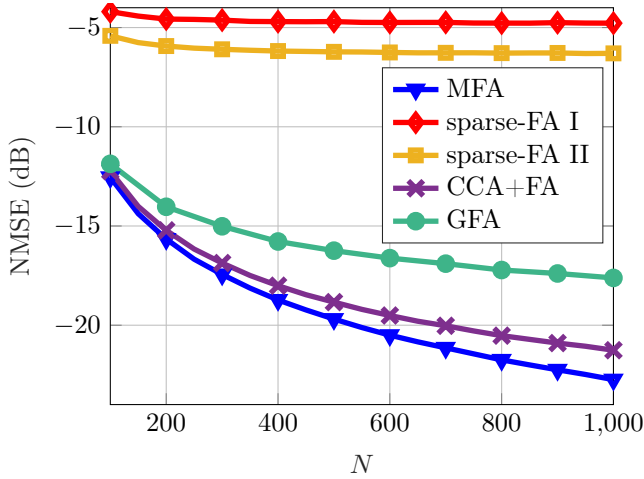
Fig. 4: NMSE in the estimate of $\mathbf{R}$ for the MFA, block-sparse FA, CCA+FA, and GFA models.

the most adequate group-wise sparse structure for the multi-channel loading matrix. As a figure of merit, we use the normalized mean-squared error in the estimate of $\mathbf{R}$, which is defined as

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

estimated by averaging 1000 Monte Carlo trials for each value of N.

We generate data according to the proposed MFA model with $M = 4$ channels of dimensions $L_1 = 6, L_2 = 8, L_3 = 10$, and $L_4 = 12$, $p = 2$ common factors and $p_i = 1$ specific factor in each channel, and the proportion of variance explained by the common factors, the specific factors, and the noise are, respectively,

$$\eta_c = 0.1, \qquad \eta_s = 0.5, \qquad \eta_n = 0.4. \qquad (69)$$

Fig. 4 shows the NMSE for the MFA, CCA+FA, and the GFA models as a function of the sample size N, as well as for two versions of the proposed naive method. In the first one, labeled as *sparse-FA I*, the projection is performed after convergence of the FA algorithm, whereas in the second one, labeled as *sparse-FA II*, the projection is performed after each iteration. For the MFA, block-sparse FA, and CCA+FA, we use the correct number of common and unique factors, while for the GFA we use the correct number of total factors $K = p + p_1 + p_2 + p_3 + p_4 = 6$. As Fig. 4 shows, the gain obtained by properly enforcing the right sparsity structure in the composite loading matrix (cf. Fig. 1) increases with the number of samples. Moreover, imposing this structure using the naive approach results in a very poor performance. Thus, we will discard this method in the next experiment.
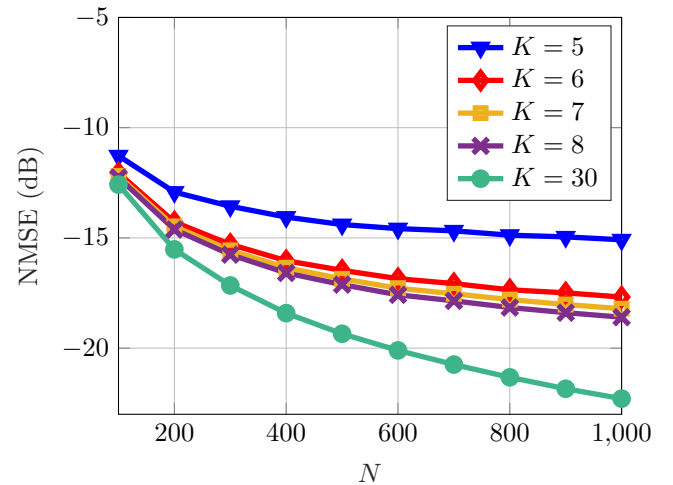
This is admittedly a rigged experiment, as the measurements are generated from a model matched to the MFA structure assuming that the exact number of common factors, p, and the exact number of unique factors, p_i, are known. In the next example we evaluate the NMSE performance of MFA,
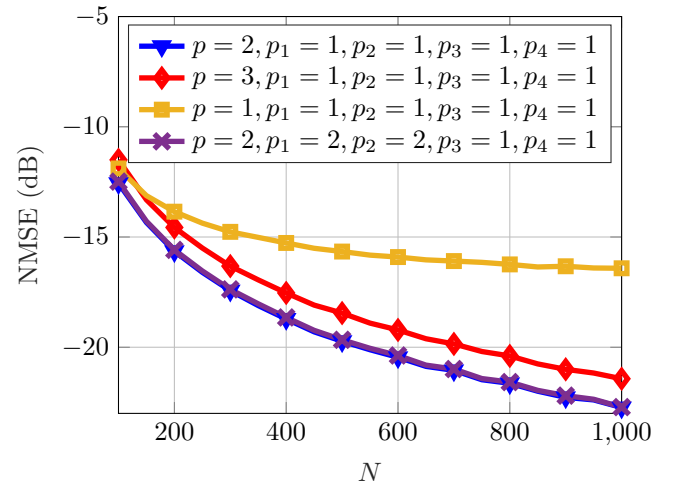
(a) NMSE (dB) for MFA
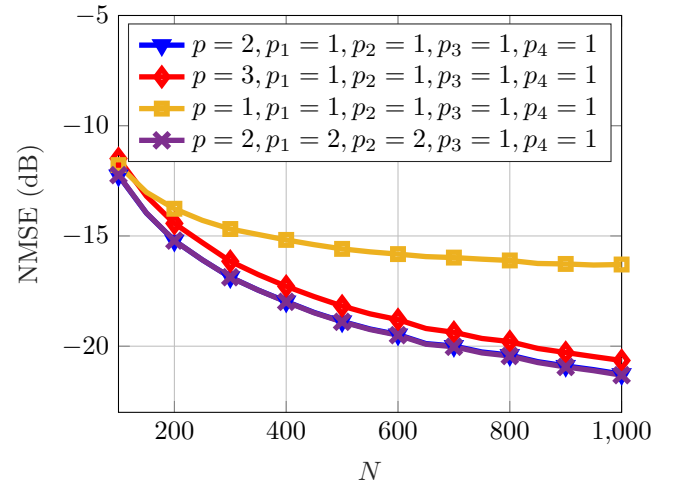
(b) NMSE (dB) for CCA+FA

(c) NMSE (dB) for GFA

Fig. 5: Robustness of MFA, CCA+FA and GFA against mismatched models.

CCA+FA, and GFA against mismatched models using the parameters of the last experiment. Let us recall that the true number of common factors is $p = 2$, and the unique factors for the 4 channels are $p_i = 1$. Fig. 5a shows that the performance of MFA is rather insensitive to an overestimation of either the number of common or unique factors. However, MFA is sensitive to underestimation of the number of unique factors. The same behavior can be observed for the CCA+FA model as can be seen in Fig. 5b. Finally, the GFA model experiences also a similar behavior, as Fig. 5c shows: it is robust against an overestimation of the total number of factors present in the true model, but sensitive to under-estimation of the number of factors. In fact, this example suggests that GFA is more sensitive to model order underestimation than is MFA or CCA+FA.

So far, we have not commented on the computational complexity of the MFA, CCA+FA, and GFA models. Here, the computational complexity is measured only by the most expensive operations (matrix factorizations and inverses). First, we will compute the computational complexity of the MFA algorithm per iteration. Step 1 of the algorithm requires the computation of $\hat{\mathbf{E}}^{1/2}$ and $\hat{\mathbf{E}}^{-1/2}$, which can be computed using the EVD with a cost of $\sum_{i=1}^{M} \mathcal{O}(n_i^3)$, where $\mathcal{O}(\cdot)$ is the Landau's big O notation. Moreover, in Step 1, the EVD of $\tilde{\mathbf{S}}$ is computed, which has a complexity of $\mathcal{O}(n^3)$, as well as the LQ decomposition of $\mathbf{B}$, which has a complexity of $\mathcal{O}(p^3)$. In Step 2, after we have removed the effect of the common factor (with a negligible complexity), it is necessary to compute the EVDs of $\tilde{\mathbf{P}}_i, i = 1, \ldots, M$, which amounts to a complexity of $\sum_{i=1}^{M} \mathcal{O}(n_i^3)$ and the LQ decompositions of the corresponding blocks with a complexity of $\sum_{i=1}^{M} \mathcal{O}(p_i^3)$. Moreover, the complexity of Step 3 can be neglected. To sum up, the complexity per iteration of the MFA is

$$\text{Comp.}_{\text{MFA}} = \mathcal{O}(n^3) + \mathcal{O}(p^3) + \sum_{i=1}^{M} \mathcal{O}(n_i^3) + \sum_{i=1}^{M} \mathcal{O}(p_i^3). \quad (70)$$

The complexity of the CCA+FA approach can be obtained similarly. Keep in mind that, as we have pointed out before, this method can be seen as a specialization of our algorithm where only one iteration of Step 1 is taken. Thus, the computation complexity is $\mathcal{O}(n^3) + \sum_{i=1}^{M} \mathcal{O}(n_i^3)$ due to CCA and then that of FA, which is $\sum_{i=1}^{M} \mathcal{O}(n_i^3) + \sum_{i=1}^{M} \mathcal{O}(p_i^3)$ per iteration. Admittedly, the complexity of the MFA algorithm is higher due to the multiple iterations of Step 1. However, as the results have shown, this higher complexity can be worth to improve the performance.

The most expensive operations performed by the GFA model in each of its iterations are the updates of its latent variables and projection matrices. These require the inversion of $M + 1$ matrices of size $K \times K$, where M is the number of channels and $K = p + \sum_i p_i$ is the total number of factors. Thus, the complexity per iteration is $(M + 1)\mathcal{O}(K^3)$. Since the involved matrices are fairly small, these operations are rather inexpensive and performing a single iterate is relatively fast. Nevertheless, the convergence of GFA in Figs. 4 and 5c required several hundreds up to thousands of iterations, which is up to two orders of magnitude more than MFA.

The GFA implementation of [31] also includes a variational approximation scheme to solve the rotational disambiguity of the solution, which adds an important computational burden to each iteration. Since this operation is not required to estimate the covariance matrix, it was not performed in our experiments.

D. Application to Passive Radar

A passive radar is equipped with both surveillance and reference antenna arrays [47]. The detection problem is to test $\mathcal{H}_1$: target present vs $\mathcal{H}_0$: target absent:

$$
\begin{aligned}
\mathcal{H}_1 : &\begin{cases} \mathbf{x}_1[n] = \mathbf{H}_1\mathbf{f}[n] + \mathbf{G}_1\mathbf{f}_1[n] + \mathbf{e}_1[n], \\ \mathbf{x}_2[n] = \mathbf{H}_2\mathbf{f}[n] + \mathbf{G}_2\mathbf{f}_2[n] + \mathbf{e}_2[n], \end{cases} \\
\mathcal{H}_0 : &\begin{cases} \mathbf{x}_1[n] = \mathbf{G}_1\mathbf{f}_1[n] + \mathbf{e}_1[n], \\ \mathbf{x}_2[n] = \mathbf{H}_2\mathbf{f}[n] + \mathbf{G}_2\mathbf{f}_2[n] + \mathbf{e}_2[n]. \end{cases}
\end{aligned}
\quad (71)
$$

Here, $\mathbf{x}_1[n]$ and $\mathbf{x}_2[n]$ are respectively the surveillance and reference observations, $\mathbf{f}[n]$ is the unknown signal transmitted by the opportunistic illuminators, and $\mathbf{H}_1$ and $\mathbf{H}_2$ correspond to the channels between the illuminators and the surveillance and reference antennas. The factor $\mathbf{f}[n]$ is common when there is a target present to reflect the direct path signal, and the scanning surveillance channel comes into synchrony with the reference channel. The factors $\mathbf{f}_1[n]$ and $\mathbf{f}_2[n]$, and their channels $\mathbf{G}_1$ and $\mathbf{G}_2$, model the local interference at the surveillance and reference antenna arrays. Local interference in the surveillance channel models the direct path signal to the surveillance channel, and local interference in the reference channel allows for the modeling of multipath from the transmitter. We assume that the number of common and unique factors is known, which is not unrealistic for this application.

In [22], the model in (71) has been studied under different assumptions on the composite covariance matrix for the surveillance and direct channels. One of these assumptions is that there is no channel specific interference in the surveillance and reference channels, and that the covariances for the noises $\mathbf{e}_1$ and $\mathbf{e}_2$ are positive definite, but not diagonal. In this case, [22] derived the generalized likelihood ratio test (GLRT):

$$T(\mathbf{x}[1], \ldots, \mathbf{x}[N]) = \prod_{i=1}^{p} \frac{1}{1 - k_i^2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta, \quad (72)$$

where η is a properly selected threshold and k_i is the ith canonical correlation between the surveillance and reference channels. That is, k_i is the ith singular value of $\mathbf{C} = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}$, with $\mathbf{S}_{ij}$ the ijth block of $\mathbf{S}$. The statistic may be termed a coherence statistic, as $\mathbf{C}$ is a coherence matrix. The statistic $1 - 1/T(\mathbf{x}[1], \ldots, \mathbf{x}[N])$ makes the coherence interpretation more clear:

$$1 - \frac{1}{T(\mathbf{x}[1], \ldots, \mathbf{x}[N])} = 1 - \prod_{i=1}^{p}(1 - k_i^2). \quad (73)$$

Let us compare the GLRT in (72) with the GLRT statistic for the problem (71)

$$\mathscr{G}(\mathbf{x}[1], \ldots, \mathbf{x}[N]) = \frac{\underset{\mathbf{H}_1,\mathbf{H}_2,\mathbf{G},\boldsymbol{\Sigma}}{\max} l(\mathbf{H}_1, \mathbf{H}_2, \mathbf{G}, \boldsymbol{\Sigma})}{\underset{\mathbf{H}_2,\mathbf{G},\boldsymbol{\Sigma}}{\max} l(\mathbf{H}_2, \mathbf{G}, \boldsymbol{\Sigma})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta. \quad (74)$$
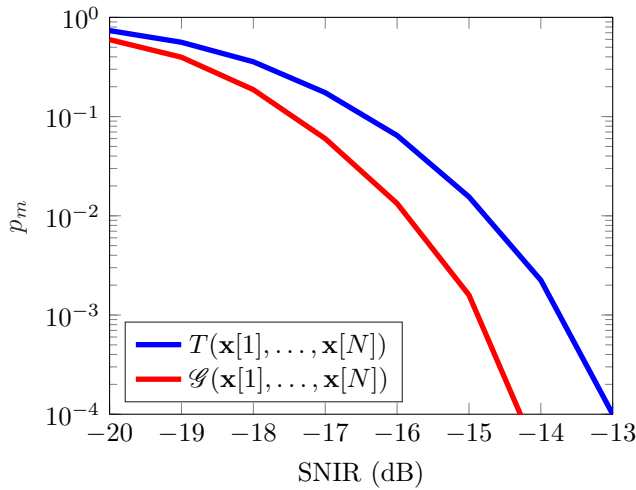
Fig. 6: Probability of missed detection (p_m) for varying SINR and fixed probability of false alarm $p_{fa} = 10^{-3}$.

Here $l(\cdot)$ is the Gaussian likelihood function. The maximization in the numerator is identical to that in Section III and we can therefore solve it using the ML-MFA algorithm. Under $\mathcal{H}_0$, the measurement model is

$$
\begin{aligned}
\mathbf{x}_1[n] &= \mathbf{G}_1 \mathbf{f}_1[n] + \mathbf{e}_1[n], \\
\mathbf{x}_2[n] &= \mathbf{H}_2 \mathbf{f}[n] + \mathbf{G}_2 \mathbf{f}_2[n] + \mathbf{e}_2[n],
\end{aligned}
\tag{75}
$$

which is equivalent to two (independent) FA problems. Thus, the computation of the compressed likelihood under the null hypothesis $\mathcal{H}_0$ may be carried out by solving two FA problems as in Section II.

To evaluate the performance of the statistics $\mathscr{G}(\mathbf{x}[1], \ldots, \mathbf{x}[N])$ and $T(\mathbf{x}[1], \ldots, \mathbf{x}[N])$, let us construct the following experiment. The noise covariance matrices are generated as $\boldsymbol{\Sigma}_i = \mathrm{diag}(\sigma_{i,1}^2, \ldots, \sigma_{i,L_i}^2)$ with $\sigma_{i,j}^2$ uniformly distributed between 0 and 1, and the elements of the common and uncommon loading matrices are generated as independent complex normals with zero mean and unit variance; the common loading matrices are scaled to achieve the desired signal-to-interference-plus-noise ratio (SINR). The surveillance and reference channels are both equipped with $L_i = 10$ antennas, the number of antennas at the illuminator is $p = 3$, the interferers have both $p_i = 1$ antenna, and the number of available samples is $N = 200$. The results for this scenario are shown in Figure 6. This figure shows the probability of missed detection (p_m) for fixed probability of false alarm $p_{fa} = 10^{-3}$ and varying SINR, which is defined as

$$
\mathrm{SINR \ (dB)} = 10 \log_{10} \left(\frac{\mathrm{tr}(\mathbf{H}_i \mathbf{H}_i^H)}{\mathrm{tr}(\mathbf{G}_i \mathbf{G}_i^H + \boldsymbol{\Sigma}_i)} \right).
\tag{76}
$$

As we can see, the proposed detector in (74) outperforms the detector $T(\mathbf{x}[1], \ldots, \mathbf{x}[N])$ in (72) because it exploits the additional structure induced by the low-rank interferers, which is to say the statistic $\mathscr{G}(\mathbf{x}[1], \ldots, \mathbf{x}[N])$ is matched to the measurement model and the statistic $T(\mathbf{x}[1], \ldots, \mathbf{x}[N])$ is mismatched.

V. CONCLUSIONS

This paper reports an extension to factor analysis (FA) for several MIMO channels that share factors and also contain unique factors. One important application of these results is to the problem of target detection in passive radar. Compared to other multi-channel generalizations of FA, such as inter-battery FA, the model proposed in this paper allows for shared *and* unique factors in each channel. The net of this model is to produce a multivariate Gaussian distribution for the set of MIMO channels in which a composite covariance matrix is structured in a very special way. The maximum likelihood problem is to identify this structured covariance matrix from a sequence of multi-channel measurements. Since there is no closed-form solution, we report an iterative algorithm, consisting of a sequence of three steps, which are derived using the block minorization-maximization framework. We prove the convergence of the algorithm to a stationary point and demonstrate its performance with numerical experiments on illustrative problems. In the theory developed here, the number of factors must be known for each channel, suggesting further refinements for order determinations in each channel.

APPENDIX I
PROOF OF THEOREM 1

The proposed algorithm is a block minorization-maximization (BMM) algorithm [38], also known as block successive minimization algorithm [39], with a cyclic selection rule. BMM algorithms are a generalization of the well-known block coordinate ascend methods [42], where instead of maximizing the objective function with respect to each block of variables, global lower bounds are maximized. This allows for more flexible algorithms that still guarantee convergence to a stationary point. Concretely, Razaviyayn, Hong, and Luo established in [39] the conditions under which BMM algorithms converge to a stationary points. Hence, the objective of this appendix is to show that these conditions are met for our particular problem. The conditions to ensure convergence of BMM algorithms are [39]:

(C1) Each block of variables belongs to a convex set.
(C2) The maximizer of the global lower bounds is unique for at least 2 blocks.
(C3) The global lower bounds satisfy the regularity conditions given by [39, Assumption 2].
(C4) The level set, defined as

$$
\mathcal{X}^{(0)} = \{\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma} \,|\, \mathcal{L}(\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}) \geq \mathcal{L}(\mathbf{H}^{(0)}, \mathbf{G}^{(0)}, \boldsymbol{\Sigma}^{(0)})\},
\tag{77}
$$

is compact.
(C5) The log-likelihood is regular at any point in $\mathcal{X}^{(0)}$.

Before proceeding, let us note that Algorithm 1 is a special case of a BMM algorithm since the estimate of $\mathbf{H}$ is obtained by directly maximizing the log-likelihood (i.e., the original cost function) and no lower-bound is required. On the other hand, the estimates for $\mathbf{G}$ and $\boldsymbol{\Sigma}$ are obtained by maximizing the global lower-bounds $Q\left(\mathbf{G}, \hat{\boldsymbol{\Sigma}}^{(k)}\right) + C$ and $\bar{Q}\left(\hat{\mathbf{G}}^{(k+1)}, \boldsymbol{\Sigma}\right) + C$, respectively.

The matrices $\mathbf{H} \in \mathbb{H}$, $\mathbf{G} = \text{blkdiag}[\mathbf{G}_1, \ldots, \mathbf{G}_M]$, where $\mathbf{G}_i \in \mathbb{G}_i$, and $\boldsymbol{\Sigma} = \text{blkdiag}[\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M]$, where $\boldsymbol{\Sigma}_i$ is a diagonal matrix with positive elements, clearly belong to convex sets. Therefore, the condition (C1) is satisfied. Moreover, as we have seen in Section III, the maximizer for $\boldsymbol{\Sigma}$ of the lower bound is unique, and the maximizers for $\mathbf{H}$ and $\mathbf{G}$ of the lower bounds are also unique when the lower triangular structure in the upper blocks of $\mathbf{H}$ and $\mathbf{G}$ is imposed, which can be safely done due to the invariances of factor analysis. Thus, condition (C2) is also met since all the maximizers of the lower bounds are unique.

Since the Gaussian log-likelihood is a smooth function, using [39, Proposition 2], it is easy to check that the regularity conditions of (C3) are satisfied for $Q\left(\mathbf{G}, \hat{\boldsymbol{\Sigma}}^{(k)}\right) + C$. Similarly, and taking into account that $\bar{Q}\left(\hat{\mathbf{G}}^{(k+1)}, \boldsymbol{\Sigma}\right) + C$ was obtained by linearizing $Q\left(\mathbf{G}, \hat{\boldsymbol{\Sigma}}^{(k)}\right) + C$, the regularity conditions are also satisfied for this lower bound.

To prove that the level set is compact, we shall use the Heine-Borel theorem, which states that a subset $\mathcal{S}$ of $\mathbb{R}^s$ (with the usual metric) is compact if and only if it is closed and bounded. In our problem, the solutions $\{\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}\}$ belong to a subset of $\mathbb{R}^{Lp + \sum_i L_i p_i + L}$, and therefore the Heine-Borel theorem applies. First, we study the closedness of the level set. Any point in $\mathcal{X}^{(0)}$ satisfies

$$\mathcal{L}(\mathbf{H}^{(0)}, \mathbf{G}^{(0)}, \boldsymbol{\Sigma}^{(0)})\} \leq \mathcal{L}(\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}) \leq \mathcal{L}(\mathbf{H}^\star, \mathbf{G}^\star, \boldsymbol{\Sigma}^\star), \tag{78}$$

where $\{\mathbf{H}^\star, \mathbf{G}^\star, \boldsymbol{\Sigma}^\star\}$ is the global maximum. It is easy to show that the log-likelihood at the global maximum is bounded above by the log-likelihood for the unstructured estimate of $\mathbf{R}$, i.e., $\hat{\mathbf{R}} = \mathbf{S}$, given by

$$-\log \det(\mathbf{S}) - L, \tag{79}$$

which is finite with probability one for $N \geq L$, implying that the set defined in (78) is closed. Then, since the log-likelihood is continuous for proper solutions (solutions with positive noise variances), the inverse image of the set in (78) must be closed. That is, the level set is closed.

To study the boundedness of $\mathcal{X}^{(0)}$, we shall decompose the covariance matrix as

$$\mathbf{R} = \mathbf{H}\mathbf{H}^T + \mathbf{G}\mathbf{G}^T + \boldsymbol{\Sigma} = a\tilde{\mathbf{R}}, \tag{80}$$

with a positive and $\text{tr}(\tilde{\mathbf{R}}) = \text{tr}(\mathbf{R}^{(0)})$. Here,

$$\tilde{\mathbf{R}} = \tilde{\mathbf{H}}\tilde{\mathbf{H}}^T + \tilde{\mathbf{G}}\tilde{\mathbf{G}}^T + \tilde{\boldsymbol{\Sigma}}, \tag{81}$$

with

$$\text{tr}(\tilde{\mathbf{R}}) = \|\tilde{\mathbf{H}}\|^2 + \sum_{i=1}^{M} \|\tilde{\mathbf{G}}_i\|^2 + \text{tr}(\tilde{\boldsymbol{\Sigma}}) = \text{tr}(\mathbf{R}^{(0)}). \tag{82}$$

Therefore, the set

$$\left\{ \tilde{\mathbf{H}}, \tilde{\mathbf{G}}, \tilde{\boldsymbol{\Sigma}} \,\middle|\, \|\tilde{\mathbf{H}}\|^2 + \sum_{i=1}^{M} \|\tilde{\mathbf{G}}_i\|^2 + \text{tr}(\tilde{\boldsymbol{\Sigma}}) = \text{tr}(\mathbf{R}^{(0)}) \right\}, \tag{83}$$

is bounded since $\text{tr}(\mathbf{R}^{(0)})$ is finite, which implies that $\mathcal{X}^{(0)}$ is bounded if a is bounded. Thus, we need to show that the values of a fulfilling

$$-\log \det(\tilde{\mathbf{R}}) - n \log a - \frac{1}{a}\text{tr}(\tilde{\mathbf{R}}^{-1}\mathbf{S})$$
$$\geq -\log \det(\mathbf{R}^{(0)}) - \text{tr}([\mathbf{R}^{(0)}]^{-1}\mathbf{S}), \tag{84}$$

are finite. Since we consider identifiable systems and proper solutions, we have that $\tilde{\boldsymbol{\Sigma}} \succ \mathbf{0}$ and therefore $\tilde{\mathbf{R}} \succ \mathbf{0}$. Now, if we can find a finite a that fulfills

$$\log a + \frac{\text{tr}(\mathbf{S})}{n\tilde{\rho}_{\min}a} \leq \frac{\log \det(\mathbf{R}^{(0)}) + \text{tr}([\mathbf{R}^{(0)}]^{-1}\mathbf{S}) - n \log \tilde{\rho}_{\min}}{n}, \tag{85}$$

where $\tilde{\rho}_{\min} > 0$ is the smallest eigenvalue of $\tilde{\mathbf{R}}$, the values of a fulfilling (84) are also finite. For $a > \text{tr}(\mathbf{S})/n\tilde{\rho}_{\min}$, the function on the left-hand side of (85) is increasing, and there must therefore exist a finite a_0 such that

$$\log a_0 + \frac{C}{na_0} = \frac{B}{n}, \tag{86}$$

which proves that a is bounded for any point in $\mathcal{X}^{(0)}$. Hence, the level set is bounded. Then, since the level set is closed and bounded, it is compact.

Finally, we study the regularity of the log-likelihood function according to the definition in [39]. Concretely, the log-likelihood is regular at a point $\{\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}\}$ in its domain if

$$\ell(\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}, \bar{\mathbf{H}}, \bar{\mathbf{G}}, \bar{\boldsymbol{\Sigma}}) \leq 0, \tag{87}$$

such that $\ell(\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}, \bar{\mathbf{H}}, \mathbf{0}, \mathbf{0}) \leq 0$, $\ell(\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}, \mathbf{0}, \bar{\mathbf{G}}, \mathbf{0}) \leq 0$, and $\ell(\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}, \mathbf{0}, \mathbf{0}, \bar{\boldsymbol{\Sigma}}) \leq 0$. Here, the directional derivative in direction $\{\bar{\mathbf{H}}, \bar{\mathbf{G}}, \bar{\boldsymbol{\Sigma}}\}$ is defined as

$$\ell(\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}, \bar{\mathbf{H}}, \bar{\mathbf{G}}, \bar{\boldsymbol{\Sigma}}) =$$
$$\liminf_{\tau \to 0} \frac{\mathcal{L}(\mathbf{H} + \tau\bar{\mathbf{H}}, \mathbf{G} + \tau\bar{\mathbf{G}}, \boldsymbol{\Sigma} + \tau\bar{\boldsymbol{\Sigma}}) - \mathcal{L}(\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma})}{\tau}. \tag{88}$$

Defining

$$\mathbf{R} = \mathbf{H}\mathbf{H}^T + \mathbf{G}\mathbf{G}^T + \boldsymbol{\Sigma}, \tag{89}$$

and

$$\bar{\mathbf{R}} = \tau^2\bar{\mathbf{H}}\bar{\mathbf{H}}^T + \tau^2\bar{\mathbf{G}}\bar{\mathbf{G}}^T + \tau\mathbf{H}\bar{\mathbf{H}}^T + \tau\bar{\mathbf{H}}\mathbf{H}^T$$
$$+ \tau\mathbf{G}\bar{\mathbf{G}}^T + \tau\bar{\mathbf{G}}\mathbf{G}^T + \tau\bar{\boldsymbol{\Sigma}}, \tag{90}$$

the first term in the numerator of (88) may be rewritten as

$$\mathcal{L}(\mathbf{H} + \tau\bar{\mathbf{H}}, \mathbf{G} + \tau\bar{\mathbf{G}}, \boldsymbol{\Sigma} + \tau\bar{\boldsymbol{\Sigma}}) =$$
$$-\log \det(\mathbf{R} + \bar{\mathbf{R}}) - \text{tr}((\mathbf{R} + \bar{\mathbf{R}})^{-1}\mathbf{S}). \tag{91}$$

Since $\tau \to 0$, which implies $\bar{\mathbf{R}} \to \mathbf{0}$, we may substitute the functions in (91) by their first-order approximations, which yields

$$\ell(\mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}, \bar{\mathbf{H}}, \bar{\mathbf{G}}, \bar{\boldsymbol{\Sigma}}) = \liminf_{\tau \to 0} \frac{\text{tr}((\mathbf{R}^{-1}\mathbf{S}\mathbf{R}^{-1} - \mathbf{R}^{-1})\bar{\mathbf{R}})}{\tau}. \tag{92}$$

Now, taking the limit and noticing that the trace is a continuous function we find that

$$
\begin{aligned}
\ell(\mathbf{H}, \mathbf{G}, \mathbf{\Sigma}, \bar{\mathbf{H}}, \bar{\mathbf{G}}, \bar{\mathbf{\Sigma}}) =\ & 2\text{tr}\left(\bar{\mathbf{H}}^T(\mathbf{R}^{-1}\mathbf{S}\mathbf{R}^{-1} - \mathbf{R}^{-1})\mathbf{H}\right) \\
& + 2\text{tr}\left(\bar{\mathbf{G}}^T(\mathbf{R}^{-1}\mathbf{S}\mathbf{R}^{-1} - \mathbf{R}^{-1})\mathbf{G}\right) \\
& + \text{tr}\left(\bar{\mathbf{\Sigma}}(\mathbf{R}^{-1}\mathbf{S}\mathbf{R}^{-1} - \mathbf{R}^{-1})\right).
\end{aligned} \quad (93)
$$

Each of the three terms in the right-hand side of (93) corresponds to $\ell(\mathbf{H}, \mathbf{G}, \mathbf{\Sigma}, \bar{\mathbf{H}}, \mathbf{0}, \mathbf{0})$, $\ell(\mathbf{H}, \mathbf{G}, \mathbf{\Sigma}, \mathbf{0}, \bar{\mathbf{G}}, \mathbf{0})$, and $\ell(\mathbf{H}, \mathbf{G}, \mathbf{\Sigma}, \mathbf{0}, \mathbf{0}, \bar{\mathbf{\Sigma}})$, respectively. Thus, if all three terms are negative, (87) is fulfilled and, therefore, the log-likelihood is regular.

To conclude, the proof follows since conditions (C1)-(C5) are satisfied.

REFERENCES

[1] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, pp. 72–101, 1904.

[2] B. S. Everitt, *An Introduction to Latent Variable Models*. Chapman and Hall, London, 1984.

[3] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. New York: Academic, 1979.

[4] D. N. Lawley and A. E. Maxwell, *Factor Analysis as a Statistical Method*. American Elsevier Pub. Co., 1971.

[5] A.-J. Boonstra and A.-J. Van der Veen, "Gain calibration methods for radio telescope arrays," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 25–38, 2003.

[6] A. J. van der Veen, A. Leshem, and A. J. Boonstra, "Array signal processing for radio astronomy," *Experimental Astronomy*, vol. 17, no. 1, pp. 231–249, Jun. 2004.

[7] D. Ramírez, G. Vázquez-Vilar, R. López-Valcarce, J. Vía, and I. Santamaría, "Detection of rank-P signals in cognitive radio networks with uncalibrated multiple antennas," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3764–3774, Aug. 2011.

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

[9] B. Friedlander and A. J. Weiss, "Direction finding using noise covariance modeling," *IEEE Trans. Signal Process.*, vol. 43, no. 7, pp. 1557–1567, Jul. 1995.

[10] Q. Wu and K. M. Wong, "UN-MUSIC and UN-CLE: An application of generakized correlation analysis to the estimation of the directionsl of arrival in unknown correlated noise," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2331–2343, Sep. 1994.

[11] T. Z. Kou and A. A. L. Beex, "Parameter identification of exponential signals in colored environments," in *IFAC Symp. Identification and System Parameters Identification*, Beijing, China, 1988, pp. 1023–1028.

[12] B. Peeters and G. D. Roeck, "Reference-based stochastic subspace identification for output-only modal analysis," *Mech. Systems and Signal Process.*, vol. 13, no. 6, pp. 855–878, 1999.

[13] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. on Acoustics, Speech and Signal Process.*, vol. ASSP-33, no. 2, p. 387, Apr. 1985.

[14] P. Stoica and M. Cedervall, "Detection tests for array processing in unknown correlated noise fields," *IEEE Trans. Signal Process.*, vol. 45, no. 9, pp. 2351–2362, Sep. 1997.

[15] A. Leshem and A.-J. Van der Veen, "Multichannel detection of Gaussian signals with uncalibrated receivers," *IEEE Signal Process. Lett.*, vol. 8, no. 4, pp. 120–122, Apr. 2001.

[16] E. Byrne and P. Schniter, "Adaptive detection of structured signals in low-rank interference," Tech. Rep., 2018.

[17] A. M. Sardarabadi and A. J. van der Veen, "Complex factor analysis and extensions," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 954–967, Feb. 2018.

[18] V. Ciccone, A. Ferrante, and M. Zorzi, "Factor models with real data: A robust estimation of the number of factors," *IEEE Trans. Auto. Control*, vol. 64, no. 6, pp. 2412–2425, 2018.

[19] K. G. Joreskog, "Testing a simple structure hypotheses in factor analysis," *Psychometrica*, vol. 31, pp. 165–178, 1966.

[20] ——, "Some contributions to maximum likelihood factor analysis," vol. 32, pp. 443–482, 1967.

[21] J.-H. Zhao, P. L. H. Yu, and Q. Jiang, "ML estimation for factor analysis: EM or non-EM?" *Stat. Comp.*, vol. 18, no. 2, pp. 109–123, 2008.

[22] I. Santamaría, L. Scharf, J. Via, Y. Wang, and H. Wang, "Passive detection of correlated subspace signals in two MIMO channels," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1752–1764, Mar. 2017.

[23] D. Rubin and D. Thayer, "EM algorithms for ML factor analysis," vol. 47, no. 1, pp. 69–76, 1982.

[24] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," Tech. Rep., 1996.

[25] Y. Song, P. Schreier, D. Ramírez, and T. Hasija, "Canonical correlation analysis of high-dimensional data with very small sample support," *Signal Process.*, vol. 128, pp. 449–458, Nov. 2016.

[26] V. Ciccone, A. Ferrante, and M. Zorzi, "Robust identification of "sparse plus low-rank" graphical models: An optimization approach," in *IEEE Conf. on Decision and Control*, 2019, pp. 2241–2246.

[27] L. R. Tucker, "An inter-battery method of factor analysis," vol. 23, no. 2, pp. 111–135, 1958.

[28] M. W. Browne, "The maximum likelihood solution in inter-battery factor analysis," *Bristish Journal Math. Stat. Psyc.*, vol. 32, pp. 75–86, 1979.

[29] R. P. McDonald, "Three common factor models for groups of variables," vol. 35, no. 1, pp. 111–128, Mar. 1970.

[30] M. W. Browne, "Factor analysis of multiple batteries by maximum likelihood," *Bristish Journal Math. Stat. Psyc.*, vol. 33, no. 2, pp. 184–199, Nov. 1980.

[31] A. Klami, S. Virtanen, E. Leppäaho, and S. Kaski, "Group factor analysis," *IEEE Trans. Neural Netw. and Learning Syst.*, vol. 26, no. 9, pp. 2136–2147, Sep. 2015.

[32] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 2377–2388, 2000.

[33] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adali, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4399–4410, 2014.

[34] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *Neural Inf. Process. Systems*, 2010, pp. 982–990.

[35] T. Adali, Y. Levin-Schwartz, and V. D. Calhoun, "Multimodal data fusion using source separation: Two effective models based on ica and iva and their properties," *Proc. IEEE*, vol. 103, no. 9, pp. 1478–1493, Sep. 2015.

[36] Y. Levin-Schwartz, Y. Song, P. J. Schreier, V. D. Calhoun, and T. Adali, "Sample-poor estimation of order and common signal subspace with application to fusion of medical imaging data," *NeuroImage*, vol. 134, pp. 486–493, Jul. 2016.

[37] X. Chu, D. L. Perez, Y. Yang, and F. Gunnarsson, *Heterogeneous Cellular Networks*. Cambridge University Press, 2013.

[38] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[39] M. Razaviyayn, M. Hong, and Z. Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Opt.*, vol. 23, no. 2, pp. 1126–1153, 2013.

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

[41] P. A. Bekker and J. M. F. ten Berge, "Generic global identification in factor analysis," *Linear Algebra Appl.*, vol. 264, pp. 255–263, 1997.

[42] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.

[43] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[44] J. Vía, I. Santamaría, and J. Pérez, "A learning algorithm for adaptive canonical correlation analysis of several data sets," *Neural Networks*, vol. 20, no. 1, pp. 139–152, Jan. 2007.

[45] T. W. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 122–148, Mar. 1963.

[46] K. Khamaru and R. Mazumder, "Computation of the maximum likelihood estimator in low-rank factor analysis," *Mathematical Programming*, vol. 176, no. 1, pp. 279–310, Jul. 2019.

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

David Ramírez (M'12-SM'16) received the Telecommunication Engineer degree and the Ph.D. degree in Electrical Engineering from the University of Cantabria, Spain, in 2006 and 2011, respectively. From 2006 to 2011, he was with the Communications Engineering Department, University of Cantabria, Spain. In 2011, he joined as a Research Associate the University of Paderborn, Germany, and later on, he became Assistant Professor (Akademischer Rat). He is now Associate Professor with the University Carlos III of Madrid. Dr. Ramírez has been a visiting researcher at the University of Newcastle, Australia and at the University College London. His current research interests include signal processing for wireless communications, statistical signal processing, change-point management, and signal processing over graphs. Dr. Ramírez has been involved in several national and international research projects on these topics. He was the recipient of the 2012 IEEE Signal Processing Society Young Author Best Paper Award and the 2013 extraordinary Ph.D. award of the University of Cantabria. Moreover, he currently serves as associate editor for the IEEE Transactions on Signal Processing. Finally, Dr. Ramírez is a member of the IEEE Technical Committee on Signal Processing Theory and Methods and was Publications Chair of the 2018 IEEE Workshop on Statistical Signal Processing.

Louis Scharf (LF'07) is Research Professor of Mathematics and Emeritus Professor of Electrical and Computer Engineering at Colorado State University, Fort Collins, CO. His research interests are in statistical signal processing and machine learning, as it applies to adaptive array processing for radar, sonar, and communication; modal analysis for electric power monitoring; and image processing for classification. He has made original contributions to matched and adaptive subspace detection, invariance theories for signal processing, and reduced-rank signal processing. He has co-authored the books, L.L. Scharf, "Statistical Signal Processing: Detection, Estimation, and Time Series Analysis," Addison-Wesley, 1991, and P.J. Schreier and L.L. Scharf, "Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals," Cambridge University Press, 2010. Professor Scharf has received several awards for his professional service and his contributions to statistical signal processing, including the Technical Achievement and Society Awards from the IEEE Signal Processing Society (SPS); the Donald W. Tufts Award for Underwater Acoustic Signal Processing, the Diamond Award from the University of Washington, and the 2016 IEEE Jack S. Kilby Medal for Signal Processing. He is a Life Fellow of IEEE.

Ignacio Santamaria (M'96-SM'05) received the Telecommunication Engineer degree and the Ph.D. degree in electrical engineering from the Universidad Politecnica de Madrid (UPM), Spain, in 1991 and 1995, respectively. In 1992, he joined the Department of Communications Engineering, University of Cantabria, Spain, where he is currently Full Professor. He has co-authored more than 200 publications in refereed journals and international conference papers, and holds two patents. His current research interests include signal processing algorithms and information-theoretic aspects of multiuser multiantenna wireless communication systems, multivariate statistical techniques and machine learning theories. He has been involved in numerous national and international research projects on these topics.

He has been a visiting researcher at the University of Florida (in 2000 and 2004), at the University of Texas at Austin (in 2009), and at the Colorado State University (in 2015). Prof. Santamaria was Technical Co-Chair of the 2nd International ICST Conference on Mobile Lightweight Wireless Systems (MOBILIGHT 2010), Special Sessions Co-Chair of the 2011 European Signal Processing Conference (EUSIPCO 2011), and General Co-Chair of the 2012 IEEE Workshop on Machine Learning for Signal Processing (MLSP 2012). From 2009 to 2014, he was a member of the IEEE Machine Learning for Signal Processing Technical Committee. He served as Associate Editor and Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING (2011-2015). He was a co-recipient of the 2008 EEEfCOM Innovation Award, as well as coauthor of a paper that received the 2012 IEEE Signal Processing Society Young Author Best Paper Award.

Steven Van Vaerenbergh (M'11-SM'15) received the M.Sc. degree in Electrical Engineering from Ghent University, Belgium, in 2003, and the Ph.D. degree in Telecommunications Engineering from the University of Cantabria, Spain, in 2010. He was a visiting researcher at the Computational NeuroEngineering Laboratory, University of Florida, in 2008. Until 2018 he was a Postdoctoral Associate with the Department of Telecommunications Engineering, University of Cantabria, where his research covered theory and algorithms for pattern recognition, time-series prediction, system identification, and online machine learning. He is currently Assistant Professor at the Department of Mathematics, Statistics and Computing of the University of Cantabria. His current research interests include statistical signal processing and machine learning efficiency.