Complex Factor Analysis and Extensions

Mouri Sardarabadi, Ahmad; van der Veen, Alle-Jan

DOI
10.1109/TSP.2017.2780047

Publication date
2018

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Signal Processing

Complex Factor Analysis and Extensions

Ahmad Mouri Sardarabadi and Alle-Jan van der Veen

***Abstract*—Many subspace-based array signal processing algorithms assume that the noise is spatially white. In this case the noise covariance matrix is a multiple of the identity and the eigenvectors of the data covariance matrix are not affected by the noise. If the noise covariance is an unknown arbitrary diagonal (e.g., for an uncalibrated array), the eigenvalue decomposition leads to incorrect subspace estimates and it has to be replaced by a more general "Factor Analysis" decomposition (FA), which then reveals all relevant information. We consider this data model and several extensions where the noise covariance matrix has a more general structure, such as banded, sparse, block-diagonal, and cases where we have multiple data covariance matrices that share the same noise covariance matrix. Starting from a nonlinear weighted least squares formulation, we propose new estimation algorithms for both classical FA and its extensions. The optimization is based on Gauss-Newton gradient descent. Generally, this leads to an iteration involving the inversion of a very large matrix. Using the structure of the problem, we show how this can be reduced to the inversion of a matrix with dimension equal to the number of unknown noise covariance parameters. This results in new algorithms that have faster numerical convergence and lower complexity compared to several maximum-likelihood based algorithms that could be considered state-of-the-art. The new algorithms scale well to large dimensions and can replace eigenvalue decompositions in many applications even if the noise can be assumed to be white.**

***Index Terms*—Factor Analysis, covariance matching, subspace estimation, maximum-likelihood**

I. Introduction

Subspace-based techniques for parameter estimation often start with a singular value decomposition (SVD) of a data matrix, or equivalently the eigenvalue decomposition (EVD) of the corresponding data covariance matrix. Without noise, this matrix is considered to be rank-deficient, and its column span is called the signal subspace. With additive noise perturbing the data, an implicit assumption is that this noise is white with covariance matrix $\sigma^2\mathbf{I}$, as adding a scaled identity matrix to the data covariance matrix does not modify the signal subspace. If this is not the case but the noise covariance matrix is known from calibration, whitening techniques can be used as a pre-processing step. However, in many array processing applications this knowledge is not available. A preferable approach is to replace the EVD by techniques that jointly estimate the signal subspace and the noise covariance matrix.

Factor Analysis (FA) is a tool from multivariate statistics that assumes a covariance matrix $\mathbf{R}$ of the data under study (e.g., samples acquired from an array of sensors) can be modeled as

$$\mathbf{R} = \mathbf{A}\mathbf{A}^H + \mathbf{D}, \tag{1}$$

where $\mathbf{A}$ is a "tall" matrix ($\mathbf{A}\mathbf{A}^H$ has low rank), and $\mathbf{D}$ is a positive diagonal matrix. In terms of subspace-based techniques, $\mathbf{A}$ captures the signal subspace while $\mathbf{D}$ can model the noise covariance matrix. Given a sample covariance matrix $\hat{\mathbf{R}}$, the objective of FA is to estimate $\mathbf{A}$ and $\mathbf{D}$.

FA for real-valued matrices was first introduced by Spearman [1] in 1904 to find a quantitative measure for intelligence, given a series of test results. Between 1940 and 1970, Lawley, Anderson, Jöreskog and others developed FA as an established multivariate technique [2]–[6]. Currently, FA is an important and popular tool for latent variable analysis with many applications in various fields of science [7]. However, its application within the signal processing community has been surprisingly limited.

In the context of signal processing, the FA problem and several extensions can be regarded as a specific case of *covariance matching*, studied in detail in [8]. In there, the model (1) is presented more generically in terms of a parametric model $\mathbf{A}(\boldsymbol{\theta})$ and a linear parametric model for the noise covariance (not restricted to diagonal), and maximum likelihood algorithms are presented to estimate the parameters. This relates to the topic of sensor array parameter estimation (e.g., direction of arrival) in the presence of colored noise or spatially correlated noise, under a variety of possible model assumptions such as $\mathbf{D}$ being diagonal, block diagonal, or composed of a linear sum of known matrices [9]–[12].

Generally, algorithms for finding the model parameters in the FA model can be categorized into two groups. "Classical" approaches are based on Maximum Likelihood (ML) or related weighted least squares optimization. This results in large nonlinear optimization problems that are often implemented using Newton-Raphson or more efficient Fletcher-Powell iterations [2], [13], [14]. These algorithms are still very popular and standard toolboxes (Matlab, SPSS) use them. Unfortunately, they are relatively hard to implement and computationally rather complex due to the inversion of a large matrix containing the second-order derivatives, so that approximations are necessary. Alternatively, the ML solution is found using Expectation-Maximization (EM) techniques, first proposed in [15], resulting in algorithms that are simpler to implement but often show slow convergence. The Conditional Maximization (CM) algorithm [16] has quadratic convergence and currently seems most competitive.

A second class of algorithms is inspired by the work of Ledermann in 1940 [17] and gained renewed momentum in recent years due to the popularity of convex optimization. The factors are found using the trace function as a convex

A.M. Sardarabadi was with Fac. Electrical Engineering, Mathematics and Computer Science, TU Delft, The Netherlands. He currently is with The Kapteyn Astronomical Institute, Groningen University, The Netherlands. A.-J. van der Veen is with Fac. Electrical Engineering, Mathematics and Computer Science, TU Delft, The Netherlands. E-mail: a.mourisardarabadi@tudelft.nl, a.j.vanderveen@tudelft.nl. This work was supported in part by NWO under contract 614.00.005.

relaxation of a minimum-rank constraint [18]–[20]. Recently, several new approaches for matrix completion have been proposed that involve low-rank plus sparse matrices [21], [22]. This leads to similar convex optimization algorithms, although not specifically designed with covariance matrices in mind.

In this paper, we aim to present factor analysis as a generic tool to replace EVD in array processing applications. We build upon prior work where we applied FA to calibration and interference detection/filtering in radio astronomy [23]–[26]. These addressed the case where the noise covariance matrix is diagonal with unknown elements. For cases where the noise covariance matrix is no longer diagonal but has a known sparse structure, we propose in this paper the "extended FA" (EFA) model.

We also consider applications where the desired subspace changes rapidly while the noise remains stationary. In this case we can compute a series of short-term covariance matrices or "snapshots" (each of the form (1) but with a common matrix $\mathbf{D}$), requiring an extension toward "joint FA" (JFA). Combined, this leads to "joint extended FA" (JEFA).

In this paper, we focus on the ML-type algorithms, and in particular consider a Weighted Least Squares (WLS) formulation that is minimized using fast-converging Gauss-Newton iterations. Contributions are:

- We extend the FA model to complex data and multi-snapshot observations, and replace the diagonal term with a more general structure (i.e., the JEFA model).
- To avoid large matrix inversions in the computation of the direction of descent, we use the Kronecker structure of these matrices to derive a closed-form expression for the direction of descent of only the noise parameters, without resorting to approximations. This results in a fast algorithm that is scalable to large problem sizes.
- Specializing this approach to the classical FA problem, we arrive at an attractive Alternating WLS algorithm that is easy to implement.
- Simulations show that the proposed algorithms are reliable and outperform many of the currently available algorithms in terms of convergence speed.

The outline of this paper is as follows. In Sec. II we discuss the data and covariance models for classical FA and in Sec. III the proposed extensions to JEFA. Sec. IV gives a brief overview of algorithms used for classical FA. Sec. V presents JEFA as a Nonlinear WLS problem and derives an efficient Gauss-Newton-based algorithm to estimate the parameters. Specializing to the classical FA model leads to an Alternating WLS solution that converges much faster than the existing algorithms. Various model order detection methods are discussed in Sec. VI, and computational complexity in Sec. VII. Finally, in Sec. VIII we use simulations to evaluate the performance of the proposed methods.

Notation

Superscript ^{T} denotes matrix transpose, * denotes complex conjugate, and ^{H} complex conjugate transpose, $\mathrm{vect}(\cdot)$ denotes the stacking of the columns of a matrix in a vector and $\mathrm{unvect}(\cdot)$ is the inverse operation (we assume that the dimensions of the resulting matrix are known). $\mathrm{diag}(\mathbf{a})$ creates a diagonal matrix out of a vector, $\mathrm{vectdiag}(\mathbf{M})$ creates a vector from the diagonal elements of a matrix, $\mathrm{diag}(\mathbf{M}) = \mathrm{diag}(\mathrm{vectdiag}(\mathbf{M}))$, $\mathrm{bdiag}(\{\mathbf{M}_m\}),\ m = 1, \ldots, M$ creates a block-diagonal matrix from the argument matrices. $\mathbf{I}$ is the identity matrix and $\mathbf{1}_M$ is an $M \times 1$ column vector containing only ones.

$\mathcal{E}\{\cdot\}$ is the expectation operator. $\otimes$ denotes the Kronecker product, $\circ$ a Khatri-Rao product (column-wise Kronecker product), and $\odot$ the entrywise multiplication of two matrices of equal size.

For any $P \times Q$ matrix $\mathbf{A}$, we denote by $\mathbf{K}_{P,Q}$ the permutation matrix such that $\mathrm{vect}(\mathbf{A}^T) = \mathbf{K}_{P,Q}\mathrm{vect}(\mathbf{A})$. For any $P \times Q$ matrix $\mathbf{A}$ and $M \times N$ matrix $\mathbf{B}$ we have

$$(\mathbf{A} \otimes \mathbf{B})\mathbf{K}_{Q,N} = \mathbf{K}_{P,M}(\mathbf{B} \otimes \mathbf{A})\,. \tag{2}$$

II. Classical Factor Analysis Model

A. Data Model

To derive the classical FA model, we consider an array of P receiving elements exposed to a mixture of $Q < P$ sources modeled by a complex Gaussian distribution. The array is uncalibrated—each element could have a different gain and noise level. We assume that the noise is a proper complex Gaussian process [27, pp. 39–40] and, for the classical model, uncorrelated between different receiving elements. By stacking the received signals from each receiver, we can model the sampled output of the system as

$$\mathbf{y}[n] = \mathbf{A}_0\mathbf{x}[n] + \mathbf{n}[n], \quad n = 1, \ldots, N \tag{3}$$

where $\mathbf{y}$ is a $P \times 1$ vector of received signals, $\mathbf{A}_0$ is a $P \times Q$ array response matrix, $\mathbf{x}$ is a $Q \times 1$ vector representing the source signals, and $\mathbf{n}$ is a $P \times 1$ vector modeling the noise. N observations are available, and assuming $\mathbf{y}[n]$ is zero mean, we construct the sample covariance matrix as

$$\hat{\mathbf{R}} = \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{y}[n]\mathbf{y}[n]^H.$$

Assuming that the sources and noise contributions are stationary and uncorrelated, the model for $\hat{\mathbf{R}}$ is

$$\mathbf{R} = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbf{A}_0\mathbf{R}_\mathbf{x}\mathbf{A}_0^H + \mathbf{R}_\mathbf{n}\,. \tag{4}$$

$\mathbf{R}_0 = \mathbf{A}_0\mathbf{R}_\mathbf{x}\mathbf{A}_0^H$ is the noise-free covariance matrix, and $\mathbf{R}_\mathbf{n} = \mathcal{E}\{\mathbf{n}\mathbf{n}^H\}$ is the noise covariance matrix. $\mathbf{R}_0$ is of rank Q, and it can be factored as $\mathbf{R}_0 = \mathbf{A}\mathbf{A}^H$ where $\mathbf{A}$ is a $P \times Q$ matrix with the same column span as $\mathbf{A}_0$.

Subspace-based array processing techniques such as MUSIC [28] and ESPRIT [29] have a first step in which the column span of $\mathbf{A}$ is to be estimated. Assuming white noise ($\mathbf{R}_\mathbf{n} = \sigma^2\mathbf{I}$), the eigenvalue decomposition of $\hat{\mathbf{R}}$ is computed. The eigenvectors corresponding to the dominant Q eigenvalues then form an estimate for the column span of $\mathbf{A}$. However, this technique fails if the noise is not white and $\mathbf{R}_\mathbf{n}$ takes another model. Most literature assumes in this case that $\mathbf{R}_\mathbf{n}$ is known so that the data can be prewhitened by $\mathbf{R}_\mathbf{n}^{-1/2}$, reducing it to the previous situation.

Instead, the classical FA model[1] assumes that the additive

[1]Traditionally FA is geared for real-valued data; in this paper we make the straightforward adaptations to complex data.

noise is independent, but not necessarily identical, i.e.,

$$\mathbf{R} = \mathbf{A}\mathbf{A}^H + \mathbf{D}, \tag{5}$$

where $\mathbf{R_n} = \mathbf{D}$ is a diagonal matrix with positive diagonal elements. Given $\hat{\mathbf{R}}$, the objective is to estimate the factors $\mathbf{A}$ and $\mathbf{D}$. In this problem, the number of columns Q of $\mathbf{A}$ (i.e., the number of sources) is assumed to be known. If Q is unknown, it can be estimated by solving the FA problem for several values of Q and employing a Generalized Likelihood Ratio Test. This approach is discussed in Sec. VI.

B. Identifiability and Uniqueness

It is immediately clear that the factors are not uniquely identifiable. E.g., $\mathbf{A}$ is not unique: The columns of $\mathbf{A}$ can be permuted and if $\mathbf{A}$ satisfies the model, then also $\mathbf{A}' = \mathbf{A}\mathbf{Q}$ is valid, for any unitary matrix $\mathbf{Q}$. The column span of $\mathbf{A}$ is invariant under these transformations, and thus these do not harm subspace estimation techniques.

More important is the uniqueness of $\mathbf{D}$. By counting numbers of observations and numbers of unknowns, we see that the number of columns Q of $\mathbf{A}$ cannot be too large, in fact we need $Q < P - \sqrt{P}$ as discussed in Appendix A. Even so, $\mathbf{D}$ is not always unique, as seen from the following example. Consider $\mathbf{R} = \mathbf{A}_1\mathbf{A}_1^H + \mathbf{D}_1$, where

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & 1 \end{bmatrix}^T.$$

Then we also have $\mathbf{R} = \mathbf{A}_2\mathbf{A}_2^H + \mathbf{D}_2$, where

$$\mathbf{A}_2 = \sqrt{2}\begin{bmatrix} 1/2 & 1 & \dots & 1 \end{bmatrix}^T, \quad \mathbf{D}_2 = \mathbf{D}_1 + \tfrac{1}{2}\mathbf{e}_1\mathbf{e}_1^T$$

and $\mathbf{e}_i$ is the ith column of the identity matrix. The problem in this case is caused by a submatrix of $\mathbf{A}_1$ being rank-deficient. This can be considered an uncommon technicality. Appendix A discusses the identifiability conditions in more detail and offers a test, for given $\mathbf{A}$, to establish identifiability of $\mathbf{D}$. Throughout the rest of the paper, we assume that $\mathbf{D}$ can be identified uniquely.

If $\mathbf{D}$ is identifiable, then $\mathbf{A}$ is unique up to a rotation $\mathbf{Q}$. We can make $\mathbf{A}$ unique by adding additional constraints. This essentially amounts to choosing a non-redundant parametrization. Not all algorithms require this, but it may be needed to avoid singularities during the computation of the Cramer-Rao Bound (CRB) or when we use Newton gradient descent techniques. For complex data, Q^2 constraint equations are needed. Common constraints are to force the columns of $\mathbf{A}$ to be orthogonal with respect to a certain weight matrix $\mathbf{W} > 0$, i.e. to require that $\mathbf{A}^H\mathbf{W}\mathbf{A}$ is diagonal.

If we compute a matrix $\mathbf{A}$ without satisfying constraints, the required transformation $\mathbf{Q}$ such that $\mathbf{A}' = \mathbf{A}\mathbf{Q}$ satisfies the constraints is easily determined afterwards. Hence, in most algorithms the constraints do not play a role.

III. Extensions of the Classical Model

We develop two extensions of the classical model: joint and extended factor analysis.

A. Joint Factor Analysis Model

In some applications, the signal subspace (i.e. $\mathbf{A}$) is not stationary, while the noise covariance is stationary. Consider e.g., DOA estimation of moving sources and an uncalibrated array. An available dataset is then partitioned into M short subsets or "snapshots", each containing N samples. This leads to M sample covariance matrices $\hat{\mathbf{R}}_m$, $m = 1, \dots, M$, with model

$$\mathbf{R}_m = \mathbf{A}_m\mathbf{A}_m^H + \mathbf{D}, \quad m = 1, \dots, M\,. \tag{6}$$

$\mathbf{A}_m$ is a low-rank matrix of size $P \times Q_m$ with $Q_m < P$ for all $m = 1, \dots, M$, and $\mathbf{D}$ is a positive real diagonal matrix common among the M models. We call this model Joint Factor Analysis (JFA). The objective is to estimate $\mathbf{D}$ and $\{\mathbf{A}_m\}$ jointly, based on the available sample covariance matrices $\{\hat{\mathbf{R}}_m\}$. In many applications we are just interested in the column span of $\mathbf{A}_m$.

B. Extended and Joint Extended FA Model

Another extension is to consider the noise covariance matrix to be more general than a diagonal matrix, say $\mathbf{R_n} = \mathbf{\Psi}$, where $\mathbf{\Psi}$ has a certain structure, assumed to be known. Here we consider $\mathbf{\Psi}$ of the form

$$\mathbf{\Psi} = \mathbf{M} \odot \mathbf{\Psi},$$

where $\mathbf{M}$ is a symmetric matrix containing only ones and zeros and $\odot$ denotes the Hadamard or entrywise product. We call $\mathbf{M}$ a mask matrix; the main diagonal is assumed to be nonzero. We can model various types of covariance matrices using this approach (for example: block-diagonal matrices, band matrices, sparse matrices, etc.).[2] We assume $\mathbf{M}$ to be known based on the application. The Extended FA (EFA) model then becomes

$$\mathbf{R} = \mathbf{A}\mathbf{A}^H + \mathbf{M} \odot \mathbf{\Psi}\,. \tag{7}$$

Both generalizations can be combined into Joint Extended FA (JEFA), where we have

$$\mathbf{R}_m = \mathbf{A}_m\mathbf{A}_m^H + \mathbf{M} \odot \mathbf{\Psi}\,, \quad m = 1, \dots, M\,. \tag{8}$$

C. Parametrization

All models presented in this section are covariance models, i.e. we can write $\mathbf{R}(\boldsymbol{\theta})$, where the vector $\boldsymbol{\theta}$ represents the unknown parameters in the model. If the parameters are complex, one popular method in signal processing is to represent them using Wirtinger operators and its extensions [30]. Given an unknown parameter θ_i we consider its conjugate θ_i^* as an independent parameter while real parameters are represented only once. Using this method we define the parameter vector as

$$\boldsymbol{\theta} = \left[\boldsymbol{\theta}_{\mathbf{A}_1}^T, \boldsymbol{\theta}_{\mathbf{A}_1^*}^T, \dots, \boldsymbol{\theta}_{\mathbf{A}_M}^T, \boldsymbol{\theta}_{\mathbf{A}_M^*}^T, \boldsymbol{\theta}_{\mathbf{\Psi}}^T\right]^T, \tag{9}$$

where

$$\boldsymbol{\theta}_{\mathbf{A}_m} = \operatorname{vect}(\mathbf{A}_m) \qquad \boldsymbol{\theta}_{\mathbf{\Psi}} = \begin{bmatrix} \boldsymbol{\psi} \\ \boldsymbol{\psi}^* \\ \mathbf{d} \end{bmatrix}.$$
$$\boldsymbol{\theta}_{\mathbf{A}_m^*} = \operatorname{vect}(\mathbf{A}_m^*)$$

[2] A further generalization of this (not considered here) is to model $\mathbf{\Psi}$ as a linear sum of known matrices [8].

Based on the mask $\mathbf{M}$, $\boldsymbol{\psi}$ is a vector consisting of the non-zero elements of the strictly upper triangular part of $\boldsymbol{\Psi}$, while $\mathbf{d} = \text{vectdiag}(\boldsymbol{\Psi})$ represents the diagonal elements of $\boldsymbol{\Psi}$, which are real. Using this parameterization we have

$$\text{vect}(\boldsymbol{\Psi}) = \mathbf{S}_U\boldsymbol{\psi} + \mathbf{S}_L\boldsymbol{\psi}^* + (\mathbf{I}_p \circ \mathbf{I}_p)\mathbf{d},$$

where $\mathbf{S}_U$ and $\mathbf{S}_L$ are selection matrices for the upper and lower triangular part of $\boldsymbol{\Psi}$, based on the mask matrix $\mathbf{M}$, and $\circ$ denotes the Khatri-Rao product (column-wise Kronecker product). We can write this as

$$\text{vect}(\boldsymbol{\Psi}) = \mathbf{J}_{\boldsymbol{\Psi}}\boldsymbol{\theta}_{\boldsymbol{\Psi}}, \tag{10}$$

where

$$\mathbf{J}_{\boldsymbol{\Psi}} = \begin{bmatrix}\mathbf{S}_U & \mathbf{S}_L & \mathbf{I}_P \circ \mathbf{I}_P\end{bmatrix}. \tag{11}$$

Note that $\mathbf{J}_{\boldsymbol{\Psi}}^H\mathbf{J}_{\boldsymbol{\Psi}} = \mathbf{I}$, so that $\mathbf{J}_{\boldsymbol{\Psi}}^H\text{vect}(\boldsymbol{\Psi}) = \boldsymbol{\theta}_{\boldsymbol{\Psi}}$, while $\mathbf{J}_{\boldsymbol{\Psi}}\mathbf{J}_{\boldsymbol{\Psi}}^H$ is a projection that represents the mask: for any $P \times P$ matrix $\mathbf{X}$ with $\mathbf{x} = \text{vect}(\mathbf{X})$ we have

$$\text{vect}(\mathbf{M} \odot \mathbf{X}) = \text{diag}[\text{vect}(\mathbf{M})]\mathbf{x} = \mathbf{J}_{\boldsymbol{\Psi}}\mathbf{J}_{\boldsymbol{\Psi}}^H\mathbf{x}. \tag{12}$$

For classical FA we have $\boldsymbol{\psi} = \mathbf{0}$ and $M = 1$, which leads to a simplified parameterization

$$\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_{\mathbf{A}} \\ \boldsymbol{\theta}_{\mathbf{A}^*} \\ \boldsymbol{\theta}_{\mathbf{D}}\end{bmatrix} = \begin{bmatrix}\text{vect}(\mathbf{A}) \\ \text{vect}(\mathbf{A}^*) \\ \mathbf{d}\end{bmatrix}. \tag{13}$$

Using this parameterization, we discuss in the following sections various methods to find an estimate for $\boldsymbol{\theta}$ given a series of sample covariance matrices $\{\hat{\mathbf{R}}_m\}$. Cramér-Rao Bounds for the presented models have been derived by us before and were presented in [31].

IV. Estimation Algorithms for Classical FA

The classical FA problem was introduced in 1904 [1] and several algorithms were proposed [4], [17], [32], all for real data matrices (although readily extended to the complex case). In this section we briefly review some of these approaches.

A. Ad Hoc Method

The estimation problem can be approached as a two-stage minimization problem [6]. In this approach we minimize the LS cost function

$$\min_{\mathbf{A},\mathbf{D}} \|\hat{\mathbf{R}} - \mathbf{A}\mathbf{A}^H - \mathbf{D}\|_F^2 \tag{14}$$

by an alternating least-squares (ALS) approach, where $\|\cdot\|_F$ is the Frobenius norm. First, for a given $\mathbf{A}$, (14) is minimized with respect to $\mathbf{D}$ and in the next stage, $\mathbf{D}$ is held constant and a new $\mathbf{A}$ is found.

Let the subscript (k) denote the iteration count. The iteration steps are

$$\mathbf{D}_{(k+1)} := \text{diag}(\hat{\mathbf{R}} - \mathbf{A}_{(k)}\mathbf{A}_{(k)}^H) \tag{15}$$

$$\mathbf{U}_{(k+1)}\boldsymbol{\Lambda}_{(k+1)}\mathbf{U}_{(k+1)}^H := \hat{\mathbf{R}} - \mathbf{D}_{(k+1)} \quad \text{[EVD]} \tag{16}$$

$$\mathbf{A}_{(k+1)} := \mathbf{U}_{0,(k+1)}\boldsymbol{\Lambda}_{0,(k+1)}^{1/2}, \tag{17}$$

where $\mathbf{U}_{(k+1)}$ and $\boldsymbol{\Lambda}_{(k+1)}$ follow from an eigenvalue decomposition, and $\mathbf{U}_{0,(k+1)}$ and $\boldsymbol{\Lambda}_{0,(k+1)}$ are the Q dominant eigenvectors and corresponding eigenvalues. A Weighted Least Squares formulation could be considered instead of (14), leading to similar iterations, but involving the EVD of $\boldsymbol{\Psi}^{-1/2}\hat{\mathbf{R}}\boldsymbol{\Psi}^{-1/2}$.

As for most ALS approaches, the rate of convergence is slow (linear). The EVD required at each iteration makes this prohibitive for large problems. Nonetheless, a single iteration of this ad hoc method is often used to initialize other iterative techniques.

B. Maximum Likelihood Estimator

Since the sources and noise are modeled as complex Gaussian, the complex log-likelihood function is given by

$$l(\boldsymbol{\theta}) = N\left[-\log(\pi^P) + \log|\mathbf{R}^{-1}| - \text{tr}(\mathbf{R}^{-1}\hat{\mathbf{R}})\right], \tag{18}$$

where $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{A}\mathbf{A}^H + \mathbf{D}$. The maximum likelihood (ML) approach aims to find $\mathbf{A}$ and $\mathbf{D}$ that maximizes this function. To this end, we find the gradient of the likelihood function (called the Fisher score) and set it equal to zero. The Fisher score for a proper Gaussian distributed signal is given by [27, p.165]

$$\mathbf{g}(\boldsymbol{\theta}) = \begin{bmatrix}\mathbf{g}_{\mathbf{A}} \\ \mathbf{g}_{\mathbf{A}^*} \\ \mathbf{g}_{\mathbf{D}}\end{bmatrix} = N\mathbf{J}^H\left(\mathbf{R}^{-T} \otimes \mathbf{R}^{-1}\right)\text{vect}(\hat{\mathbf{R}} - \mathbf{R}), \tag{19}$$

where the Jacobian $\mathbf{J}(\boldsymbol{\theta})$ is given by

$$\begin{aligned}\mathbf{J} &= \frac{\partial\text{vect}(\mathbf{R})}{\partial\boldsymbol{\theta}^T} = \left[\frac{\partial\text{vect}(\mathbf{R})}{\partial\boldsymbol{\theta}_{\mathbf{A}}^T}, \frac{\partial\text{vect}(\mathbf{R})}{\partial\boldsymbol{\theta}_{\mathbf{A}^*}^T}, \frac{\partial\text{vect}(\mathbf{R})}{\partial\boldsymbol{\theta}_{\mathbf{D}}^T}\right] \\ &= \left[\mathbf{J}_{\mathbf{A}}, \mathbf{J}_{\mathbf{A}^*}, \mathbf{J}_{\mathbf{D}}\right],\end{aligned} \tag{20}$$

which evaluates to (cf. (2))

$$\begin{aligned}&\mathbf{J}_{\mathbf{A}} = \mathbf{A}^* \otimes \mathbf{I}_P, \quad \mathbf{J}_{\mathbf{A}^*} = (\mathbf{I}_P \otimes \mathbf{A})\mathbf{K}_{P,Q}, \\ &\mathbf{J}_{\mathbf{D}} = \mathbf{I}_P \circ \mathbf{I}_P.\end{aligned}$$

From these results and (19), the elements of the Fisher score become

$$\begin{aligned}\mathbf{g}_{\mathbf{A}} &= N(\mathbf{A}^T\mathbf{R}^{-T} \otimes \mathbf{R}^{-1})\text{vect}(\hat{\mathbf{R}} - \mathbf{R}) \\ &= N\text{vect}\left[\mathbf{R}^{-1}(\hat{\mathbf{R}} - \mathbf{R})\mathbf{R}^{-1}\mathbf{A}\right] \qquad (21)\end{aligned}$$

$$\mathbf{g}_{\mathbf{A}^*} = \mathbf{g}_{\mathbf{A}}^* \tag{22}$$

$$\mathbf{g}_{\mathbf{D}} = N\text{vectdiag}\left[\mathbf{R}^{-1}(\hat{\mathbf{R}} - \mathbf{R})\mathbf{R}^{-1}\right]. \tag{23}$$

The ML technique requires us to set (21) and (23) equal to zero, but unfortunately this does not produce a closed-form solution. As a result, different iterative techniques such as the scoring method and EM based approaches have been suggested in the literature.

1) The Scoring Method: Initial algorithms considered the alternating optimization of (21) and (23), and this leads to similar algorithms as the Ad Hoc method [2]. Starting with [13], one line of research has considered Newton-Raphson-like algorithms to numerically compute the ML estimate, as these provide quadratic convergence. In particular, the scoring algorithm is a variant of the Newton-Raphson algorithm where the gradient and Hessian are replaced by the Fisher score and Fisher information matrix, respectively [33]. The Fisher information matrix (FIM) for the Gaussian distribution is given by

$$\mathbf{F} = \mathbf{J}^H(\mathbf{R}^{-T} \otimes \mathbf{R}^{-1})\mathbf{J}, \tag{24}$$

where $\mathbf{J}$ is given by (20). The resulting scoring iterations are

$$\boldsymbol{\theta}_{(k+1)} = \boldsymbol{\theta}_{(k)} + \mu_{(k)}\boldsymbol{\delta}, \tag{25}$$

where $\mu_{(k)}$ is a step size and

$$\boldsymbol{\delta} = \begin{bmatrix}\boldsymbol{\delta}_{\mathbf{A}}^T & \boldsymbol{\delta}_{\mathbf{A}^*}^T & \boldsymbol{\delta}_{\mathbf{D}}^T\end{bmatrix}^T$$

is the direction of descent. The latter follows from solving

$$\mathbf{F}_{(k)}\boldsymbol{\delta} = \mathbf{g}_{(k)}, \tag{26}$$

where $\mathbf{g}_{(k)} = \mathbf{g}(\boldsymbol{\theta}_{(k)})$ is the Fisher score and $\mathbf{F}_{(k)} = \mathbf{F}(\boldsymbol{\theta}_{(k)})$ is the FIM. Since without constraints the parametrization is redundant (see Sec. II-B), the FIM is singular. However, this does not need to cause complications because $\mathbf{g}_{(k)}$ is in the column span of $\mathbf{F}_{(k)}$, so that the system of equations has a solution, and (taking the minimum-norm solution) standard convergence results for the scoring method follow.[3]

A problem with the scoring method is that the matrix $\mathbf{F}$ quickly becomes large, as its dimension is equal to the number of unknown parameters. Solving (26) then becomes unattractive. The literature shows several approximations to reduce the complexity of this step. E.g., the ML method described in [4] is an approximation of the scoring method in which $\mathbf{F}_{(k)}^{\dagger}$ is approximated by $[\mathrm{diag}(\mathbf{F}_{(k)})]^{-1}$, i.e., a Jacobi preconditioner.

2) EM-based Algorithms: Alternatively, the expectation maximization (EM) technique may be used to optimize the likelihood function. For FA, this was first proposed by [15]. Unfortunately, many of the EM algorithms show very slow (linear) convergence. An overview of the original method and several of its derivatives can be found in [16]. In that paper, an alternative Constrained Maximization (CM) algorithm is proposed that is straightforward to implement and shows quadratic convergence. We compare with CM in the simulations.

3) Covariance matching techniques: Factor Analysis can be viewed as a special case of covariance matching, studied in detail in [8]. In there, $\mathbf{A}(\boldsymbol{\theta})$ is modeled parametrically, while the noise covariance $\boldsymbol{\Psi}$ has a linear parametrization as in (10), but for a more general (known) matrix $\mathbf{J}_{\boldsymbol{\Psi}}$. This fits the formalism of what we call Extended Factor Analysis.

In [8], the ML problem is replaced by a Weighted Least Squares (WLS) fitting of the sample covariance, and it is shown that the large sample properties of the estimators are the same. Solving this nonlinear least squares problem using gradient descent techniques is closely connected to the scoring algorithm, and we follow this approach.

For FA, a technique based on WLS was proposed by Jöreskog in 1972 [14] and solved using Newton-Raphson iterations. We compare this method in the simulations in Sec. VIII-A.

This concludes our review of some of the popular estimation techniques for classical FA.

V. Estimation Algorithms for JEFA

In this section we consider the generalization toward the JEFA model (8). Starting from a covariance matching formulation, estimating the parameters for JEFA also leads to a nonlinear weighted least squares problem. As the number of parameters grows quickly, we need to consider scalable approaches. We propose several algorithms.

[3]Alternatively, a non-redundant or constrained parametrization could be used, but it does not seem to offer advantages.

A. Nonlinear Weighted Least Squares

Recall the JEFA data model (8). We start by vectoring and stacking all the covariance matrices to form a single measurement vector

$$\hat{\mathbf{r}} = \begin{bmatrix}\mathrm{vect}^T(\hat{\mathbf{R}}_1), & \dots & ,\mathrm{vect}^T(\hat{\mathbf{R}}_M)\end{bmatrix}^T, \tag{27}$$

and similarly

$$\mathbf{r}(\boldsymbol{\theta}) = \begin{bmatrix}\mathrm{vect}^T(\mathbf{R}_1(\boldsymbol{\theta})), & \dots & ,\mathrm{vect}^T(\mathbf{R}_M(\boldsymbol{\theta}))\end{bmatrix}^T, \tag{28}$$

where $\boldsymbol{\theta}$ is defined by (9). Instead of following the ML formalism, we can estimate the unknown parameters in $\boldsymbol{\theta}$ using nonlinear WLS defined as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{W}^{1/2}[\hat{\mathbf{r}} - \mathbf{r}(\boldsymbol{\theta})]\|_2^2, \tag{29}$$

where $\mathbf{W}$ is a weighting matrix. The optimum weighting matrix is the inverse of the (asymptotic) covariance matrix of the entire dataset, but because we only have access to the sample covariance matrices $\hat{\mathbf{R}}_m$ we use

$$\mathbf{W} = \begin{bmatrix}\hat{\mathbf{R}}_1^{-T} \otimes \hat{\mathbf{R}}_1^{-1} & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \hat{\mathbf{R}}_M^{-T} \otimes \hat{\mathbf{R}}_M^{-1}\end{bmatrix}, \tag{30}$$

which gives an asymptotically optimal solution for a Gaussian distributed data matrix [8].[4]

A very common iterative technique for solving nonlinear optimization problems is the Gauss-Newton algorithm, where the Hessian is replaced by the Gramian of the Jacobians [34]. The updates are similar to the scoring method updates (25):

$$\boldsymbol{\theta}_{(k+1)} = \boldsymbol{\theta}_{(k)} + \mu_{(k)}\boldsymbol{\delta}, \tag{31}$$

where $\boldsymbol{\delta}$ is the direction of descent. To find $\boldsymbol{\delta}$ we need to solve

$$\mathbf{B}(\boldsymbol{\theta}_{(k)})\boldsymbol{\delta} = \mathbf{g}(\boldsymbol{\theta}_{(k)}), \tag{32}$$

where

$$\mathbf{g}(\boldsymbol{\theta}) = \mathbf{J}^H(\boldsymbol{\theta})\mathbf{W}[\hat{\mathbf{r}} - \mathbf{r}(\boldsymbol{\theta})] \tag{33}$$

$$\mathbf{B}(\boldsymbol{\theta}) = \mathbf{J}^H(\boldsymbol{\theta})\mathbf{W}\mathbf{J}(\boldsymbol{\theta}) \tag{34}$$

and the Jacobian $\mathbf{J}(\boldsymbol{\theta})$ is given by

$$\mathbf{J} = \frac{\partial \mathbf{r}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \begin{bmatrix}\mathbf{J}_{\mathbf{A}_1} & \mathbf{J}_{\mathbf{A}_1^*} & \dots & \mathbf{0} & \mathbf{J}_{\boldsymbol{\Psi}} \\ \mathbf{0} & & \dots & \mathbf{0} & \mathbf{J}_{\boldsymbol{\Psi}} \\ \mathbf{0} & \ddots & \ddots & \mathbf{0} & \mathbf{J}_{\boldsymbol{\Psi}} \\ \mathbf{0} & \dots & \mathbf{J}_{\mathbf{A}_M} & \mathbf{J}_{\mathbf{A}_M^*} & \mathbf{J}_{\boldsymbol{\Psi}}\end{bmatrix} \tag{35}$$

$$\mathbf{J}_{\mathbf{A}_m} = (\mathbf{A}_m^* \otimes \mathbf{I}_P), \qquad \mathbf{J}_{\mathbf{A}_m^*} = (\mathbf{I}_P \otimes \mathbf{A}_m)\mathbf{K}_{P,Q_m}$$
$$\mathbf{J}_{\boldsymbol{\Psi}} = [\mathbf{S}_U,\ \mathbf{S}_L,\ \mathbf{I}_P \circ \mathbf{I}_P]\,. \tag{36}$$

The iterations given by (31) are repeated until $\|\mathbf{g}(\boldsymbol{\theta}_{(k)})\|_2 < \epsilon$, where $\epsilon > 0$ depends on the desired accuracy. Clearly, the equations are very similar to the ML equations in Sec. IV-B, except that the covariance matrices in $\mathbf{W}$ (30) have to be inverted only once.

The key step in the Gauss-Newton iteration is solving the linear system (32). For the JEFA model the matrix dimensions can quickly become large. We propose two approaches for solving this system. The first approach (Sec. V-B) is a Krylov-based method directly applied to the system of equations, while the second approach (Sec. V-C) is based on a symbolic

[4]As an aside, we remark that the minimum trace factor analysis discussed by [32] is a special case of the WLS we are considering here.

inversion of $\mathbf{B}$, essentially exploiting the sparse structure of (35).

The optimal step size $\mu_{(k)}$ can also be derived, and this is done in Appendix C. It amounts to solving for the roots of a cubic polynomial, which is computationally simple.

For JFA we can enforce additional constraints such as $\mathbf{D} \geq \epsilon\mathbf{I}$ for some $\epsilon > 0$ using a nonlinear active set approach [34]. The full discussion of this approach is beyond the scope of this paper, but the algorithm presented here can be extended with small modifications.

B. Krylov-Based Method for Direction of Descent

To reduce storage and complexity, we propose to solve (32) using a Krylov subspace-based solver. An overview of such solvers is in [35]. We know that for the FA problem the solution is not unique. This means that the Jacobians and hence $\mathbf{B}$ are singular. One possible Krylov solver that is applicable in this case is the MinresQLP algorithm [36] and for this reason we have chosen this solver for our iterative approach.[5]

MinresQLP is a standard Krylov-subspace iterative solver that requires the availability of a subroutine that performs a matrix-vector multiplications of the form $\mathbf{u} = \mathbf{B}\mathbf{v}$. Other operations in MinresQLP have negligible complexity. We show how we can perform this multiplication efficiently by exploiting the Kronecker structure of $\mathbf{B}(\boldsymbol{\theta})$ and the underlying $\mathbf{J}(\boldsymbol{\theta})$, without needing to store the Jacobians.

We drop the dependency on $\boldsymbol{\theta}$ from the notation and write only $\mathbf{J}$ and $\mathbf{r}$ because $\boldsymbol{\theta}$ does not change while we are solving for $\boldsymbol{\delta}$. To calculate a product $\mathbf{u} = \mathbf{B}\mathbf{v}$ for $\mathbf{B}$ in (34) and arbitrary vectors $\mathbf{u}, \mathbf{v}$ of compatible dimensions, we define the intermediate results

$$\mathbf{z} = \mathbf{J}\mathbf{v}\,, \quad \mathbf{y} = \mathbf{W}\mathbf{z}\,, \quad \mathbf{u} = \mathbf{J}^H\mathbf{y}\,.$$

We partition $\mathbf{u}$ and $\mathbf{v}$ in the same manner as $\boldsymbol{\theta}$ in (9) into

$$\mathbf{v} = \begin{bmatrix} \mathrm{vect}(\mathbf{V}_{\mathbf{A}_1}) \\ \mathrm{vect}(\mathbf{V}_{\mathbf{A}_1^*}) \\ \vdots \\ \mathbf{v}_{\boldsymbol{\Psi}} \end{bmatrix}, \; \mathbf{u} = \begin{bmatrix} \mathrm{vect}(\mathbf{U}_{\mathbf{A}_1}) \\ \mathrm{vect}(\mathbf{U}_{\mathbf{A}_1^*}) \\ \vdots \\ \mathbf{u}_{\boldsymbol{\Psi}} \end{bmatrix}. \tag{37}$$

Likewise we partition $\mathbf{z}$ and $\mathbf{y}$ in the same manner as $\mathbf{r}$ in (28) as

$$\mathbf{z} = \begin{bmatrix} \mathrm{vect}(\mathbf{Z}_1) \\ \vdots \\ \mathrm{vect}(\mathbf{Z}_m) \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathrm{vect}(\mathbf{Y}_1) \\ \vdots \\ \mathrm{vect}(\mathbf{Y}_m) \end{bmatrix}. \tag{38}$$

To find $\mathbf{u}$ we compute $\mathbf{U}_{\mathbf{A}_m}$ $(m = 1, \ldots, M)$ and $\mathbf{u}_{\boldsymbol{\Psi}}$. We assume that $\mathbf{v}$ is such that $\mathbf{V}_{\mathbf{A}_m^*} = \mathbf{V}_{\mathbf{A}_m}^*$, and in that case $\mathbf{U}_{\mathbf{A}_m^*} = \mathbf{U}_{\mathbf{A}_m}^*$. It can be shown that MinresQLP provides vectors $\mathbf{v}$ that have this property, as long as we initialize the iteration with a vector $\mathbf{g}$ with the same property.

The Jacobian for the entire dataset is given by (35). Note that $\mathbf{J}_{\boldsymbol{\Psi}}$ in (36) is identical to $\mathbf{J}_{\boldsymbol{\Psi}}$ as defined by (11), which related the parameter vector $\boldsymbol{\theta}_{\boldsymbol{\Psi}}$ to the $P \times P$ matrix $\boldsymbol{\Psi}$ via $\mathrm{vect}(\boldsymbol{\Psi}) = \mathbf{J}_{\boldsymbol{\Psi}}\boldsymbol{\theta}_{\boldsymbol{\Psi}}$. Similarly, we can define a $P \times P$ matrix $\mathbf{V}_{\boldsymbol{\Psi}}$ as $\mathrm{vect}(\mathbf{V}_{\boldsymbol{\Psi}}) = \mathbf{J}_{\boldsymbol{\Psi}}\mathbf{v}_{\boldsymbol{\Psi}}$ and likewise $\mathrm{vect}(\mathbf{U}_{\boldsymbol{\Psi}}) = \mathbf{J}_{\boldsymbol{\Psi}}\mathbf{u}_{\boldsymbol{\Psi}}$.

Using these relations, we can compute the components of $\mathbf{z} = \mathbf{J}\mathbf{v}$ as

$$\begin{aligned} \mathrm{vect}(\mathbf{Z}_m) &= (\mathbf{A}_m^* \otimes \mathbf{I}_P)\mathrm{vect}(\mathbf{V}_{\mathbf{A}_m}) \\ &\quad + (\mathbf{I}_P \otimes \mathbf{A}_m)\mathbf{K}_{P,Q}\mathrm{vect}(\mathbf{V}_{\mathbf{A}_m^*}) + \mathrm{vect}(\mathbf{V}_{\boldsymbol{\Psi}}) \\ &= \mathrm{vect}\left(\mathbf{V}_{\mathbf{A}_m}\mathbf{A}_m^H + \mathbf{A}_m\mathbf{V}_{\mathbf{A}_m}^H + \mathbf{V}_{\boldsymbol{\Psi}}\right), \end{aligned}$$

where we used $\mathbf{V}_{\mathbf{A}_m^*} = \mathbf{V}_{\mathbf{A}_m}^*$. Unstacking both sides gives

$$\mathbf{Z}_m = \mathbf{V}_{\mathbf{A}_m}\mathbf{A}_m^H + \mathbf{A}_m\mathbf{V}_{\mathbf{A}_m}^H + \mathbf{V}_{\boldsymbol{\Psi}}. \tag{39}$$

Hence, to calculate $\mathbf{z} = \mathbf{J}\mathbf{v}$, we reshape the vector $\mathbf{v}$ into corresponding matrices $\mathbf{V}_{\mathbf{A}_m}$ and $\mathbf{V}_{\boldsymbol{\Psi}}$, and apply (39). The variables $\mathbf{A}_m$ are the current estimates of the unknown parameters and hence require no additional storage.

Next, we compute $\mathbf{y} = \mathbf{W}\mathbf{z}$. Using properties of Kronecker products and the definition of $\mathbf{W}$ in (30), it is straightforward to show that we only need to compute

$$\mathbf{Y}_m = \hat{\mathbf{R}}_m^{-1}\mathbf{Z}_m\hat{\mathbf{R}}_m^{-1}\,, \quad m = 1, \ldots, M. \tag{40}$$

Finally, we calculate $\mathbf{u} = \mathbf{J}^H\mathbf{y}$. From the structure of (35), (36), we find

$$\mathrm{vect}(\mathbf{U}_{\mathbf{A}_m}) = \mathbf{J}_{\mathbf{A}_m}^H\mathrm{vect}(\mathbf{Y}_m) \;\Leftrightarrow\; \mathbf{U}_{\mathbf{A}_m} = \mathbf{Y}_m\mathbf{A}_m \tag{41}$$

while $\mathbf{U}_{\mathbf{A}_m^*} = \mathbf{U}_{\mathbf{A}_m}^*$. The remaining term $\mathbf{U}_{\boldsymbol{\Psi}}$ is given by

$$\mathbf{U}_{\boldsymbol{\Psi}} = \sum_{m=1}^{M} \mathbf{M} \odot \mathbf{Y}_m, \tag{42}$$

where we have used the properties $\mathbf{u}_{\boldsymbol{\Psi}} = \mathbf{J}_{\boldsymbol{\Psi}}^H\mathrm{vect}(\mathbf{U}_{\boldsymbol{\Psi}})$ and, using (12), $\mathbf{J}_{\boldsymbol{\Psi}}^H\mathrm{vect}(\mathbf{X}) = \mathbf{J}_{\boldsymbol{\Psi}}^H\mathbf{J}_{\boldsymbol{\Psi}}\mathbf{J}_{\boldsymbol{\Psi}}^H\mathrm{vect}(\mathbf{X}) = \mathbf{J}_{\boldsymbol{\Psi}}^H\mathrm{vect}(\mathbf{M} \odot \mathbf{X})$.

To summarize, to calculate a matrix-vector product $\mathbf{u} = \mathbf{B}\mathbf{v}$ we reshape $\mathbf{v}$ into $\mathbf{V}_{\mathbf{A}_m}$ and $\mathbf{V}_{\boldsymbol{\Psi}}$ and use (39)–(42) to find the result. The gradient $\mathbf{g}$ in (33) can be calculated in a similar manner by replacing $\mathbf{Z}_m$ in (40) by $\hat{\mathbf{R}}_m - \mathbf{R}_m$ and using (41) and (42) with the result. The procedures that perform these steps are provided to MinresQLP, which then solves $\mathbf{B}\boldsymbol{\delta} = \mathbf{g}$ (32).

C. Direct Method for Direction of Descent

As an alternative technique to Krylov iterations for computing the direction of descent, we now provide a direct approach for solving $\mathbf{B}\boldsymbol{\delta} = \mathbf{g}$. A block LDU (lower-diagonal-upper, or Cholesky) decomposition of the Hermitian matrix $\mathbf{B}$ can be computed symbolically in closed form and leads to the following solution for the descent direction $\boldsymbol{\delta}$.

Define $\mathbf{W}_m = \hat{\mathbf{R}}_m^{-1}$ and the quantities

$$\tilde{\mathbf{W}}_m = \mathbf{W}_m - \mathbf{W}_m\mathbf{A}_m(\mathbf{A}_m^H\mathbf{W}_m\mathbf{A}_m)^{-1}\mathbf{A}_m^H\mathbf{W}_m, \tag{43}$$

$$\tilde{\mathbf{B}}_{\boldsymbol{\Psi}} = \mathbf{J}_{\boldsymbol{\Psi}}^H\left(\sum_m \tilde{\mathbf{W}}_m^T \otimes \tilde{\mathbf{W}}_m\right)\mathbf{J}_{\boldsymbol{\Psi}}, \tag{44}$$

$$\tilde{\mathbf{g}}_{\boldsymbol{\Psi}} = \mathbf{J}_{\boldsymbol{\Psi}}^H\sum_m\left(\tilde{\mathbf{W}}_m^T \otimes \tilde{\mathbf{W}}_m\right)\mathrm{vect}[\hat{\mathbf{R}}_m - \mathbf{R}_m(\boldsymbol{\theta})]\,. \tag{45}$$

As shown in Appendix B, the computation of

$$\boldsymbol{\delta} = \begin{bmatrix} \mathrm{vect}(\boldsymbol{\Delta}_{\mathbf{A}_1}) \\ \mathrm{vect}(\boldsymbol{\Delta}_{\mathbf{A}_1^*}) \\ \vdots \\ \boldsymbol{\delta}_{\boldsymbol{\Psi}} \end{bmatrix}$$

[5] Alternatively, methods such as LSQR or LSMR could be used to solve the equivalent LS problem, $\min_{\boldsymbol{\delta}} \|\mathbf{W}^{1/2}(\mathbf{J}_k\boldsymbol{\delta} - \mathbf{b}_k)\|_2^2$, where $\mathbf{b}_k = \hat{\mathbf{r}} - \mathbf{r}(\boldsymbol{\theta}_k)$. However, working with $\mathbf{W}^{1/2}$ could make these methods computationally less attractive.

reduces to the computation of $\boldsymbol{\delta}_{\boldsymbol{\Psi}}$ from

$$\tilde{\mathbf{B}}_{\boldsymbol{\Psi}}\boldsymbol{\delta}_{\boldsymbol{\Psi}} = \tilde{\mathbf{g}}_{\boldsymbol{\Psi}}\,. \tag{46}$$

Subsequently, we define $\boldsymbol{\Delta}_{\boldsymbol{\Psi}}$ as $\text{vect}(\boldsymbol{\Delta}_{\boldsymbol{\Psi}}) = \mathbf{J}_{\boldsymbol{\Psi}}\boldsymbol{\delta}_{\boldsymbol{\Psi}}$. Closed-form expressions for the $\boldsymbol{\Delta}_{\mathbf{A}_m}$ are

$$\begin{aligned}\boldsymbol{\Delta}_{\mathbf{A}_m} =& \frac{1}{2}(\mathbf{I} + \mathbf{W}_m^{-1}\tilde{\mathbf{W}}_m)(\hat{\mathbf{R}}_m - \mathbf{R}_m(\boldsymbol{\theta}) - \boldsymbol{\Delta}_{\boldsymbol{\Psi}})\cdot \\ &\cdot \mathbf{W}_m\mathbf{A}_m(\mathbf{A}_m^H\mathbf{W}_m\mathbf{A}_m)^{-1}, m = 1,\ldots,M\,, \end{aligned}\tag{47}$$

and $\boldsymbol{\Delta}_{\mathbf{A}_m^*} = \boldsymbol{\Delta}_{\mathbf{A}_m}^*$. Hence, the original matrix inversion problem reduces to solving for $\boldsymbol{\delta}_{\boldsymbol{\Psi}}$ in (46), which has a dimension equal to the number of nonzero entries in the mask $\mathbf{M}$, which is $2P\sum_m Q_m$ fewer unknowns than in $\boldsymbol{\delta}$. In particular, for the JFA model ($\boldsymbol{\Psi}$ diagonal), $\tilde{\mathbf{B}}_{\boldsymbol{\Psi}}$ is just $P \times P$. Since $\boldsymbol{\Psi}$ is well defined if the JEFA model is identifiable, this problem is well-posed.

For large problems, we can also solve (46) using a Krylov-subspace based solver, and the matrix-vector products are similar to the ones presented in the previous section.

D. Alternating WLS method

The approach from Sec. V-C can be developed into a new Alternating WLS method that is similar to the Ad Hoc method discussed in Sec. IV-A, but providing much faster convergence. We consider the update equation for $\boldsymbol{\theta}_{\boldsymbol{\Psi}}$. If we take the step size $\mu_{(k)} = 1$ we have $\boldsymbol{\theta}_{\boldsymbol{\Psi}}^{(k+1)} = \boldsymbol{\theta}_{\boldsymbol{\Psi}}^{(k)} + \boldsymbol{\delta}_{\boldsymbol{\Psi}}$. Starting from (46) and subsequently using $\text{vect}(\boldsymbol{\Psi}^{(k)}) = \mathbf{J}_{\boldsymbol{\Psi}}\boldsymbol{\theta}_{\boldsymbol{\Psi}}^{(k)}$ and the definition of $\tilde{\mathbf{B}}_{\boldsymbol{\Psi}}$ in (44), we obtain

$$\begin{aligned}&\tilde{\mathbf{B}}_{\boldsymbol{\Psi}}\boldsymbol{\theta}_{\boldsymbol{\Psi}}^{(k+1)} \\ &= \tilde{\mathbf{B}}_{\boldsymbol{\Psi}}\boldsymbol{\theta}_{\boldsymbol{\Psi}}^{(k)} + \mathbf{J}_{\boldsymbol{\Psi}}^H\sum_m(\tilde{\mathbf{W}}_m^T \otimes \tilde{\mathbf{W}}_m)\text{vect}(\hat{\mathbf{R}}_m - \mathbf{A}_m\mathbf{A}_m^H - \boldsymbol{\Psi}^{(k)}) \\ &= \mathbf{J}_{\boldsymbol{\Psi}}^H\sum_m\left(\tilde{\mathbf{W}}_m^T \otimes \tilde{\mathbf{W}}_m\right)\text{vect}(\hat{\mathbf{R}}_m - \mathbf{A}_m\mathbf{A}_m^H),\end{aligned}$$

where to simplify the notation we have dropped the dependency on k from $\tilde{\mathbf{B}}_{\boldsymbol{\Psi}}$, $\tilde{\mathbf{W}}_m$ and $\mathbf{A}_m$. Since $\tilde{\mathbf{W}}_m\mathbf{A}_m = \mathbf{0}$ as a result of (43), this reduces to

$$\tilde{\mathbf{B}}_{\boldsymbol{\Psi}}\boldsymbol{\theta}_{\boldsymbol{\Psi}}^{(k+1)} = \mathbf{J}_{\boldsymbol{\Psi}}^H\sum_m\left(\tilde{\mathbf{W}}_m^T \otimes \tilde{\mathbf{W}}_m\right)\text{vect}(\hat{\mathbf{R}}_m). \tag{48}$$

From the definition of $\tilde{\mathbf{B}}_{\boldsymbol{\Psi}}$ in (44), the solution $\boldsymbol{\theta}_{\boldsymbol{\Psi}}^{(k+1)}$ can also be written as the solution to

$$\min_{\boldsymbol{\theta}_{\boldsymbol{\Psi}}}\left\|\tilde{\mathbf{W}}_{(k)}^{1/2}[\hat{\mathbf{r}} - \tilde{\mathbf{J}}_{\boldsymbol{\Psi}}\boldsymbol{\theta}_{\boldsymbol{\Psi}}]\right\|_2^2, \tag{49}$$

where $\tilde{\mathbf{J}}_{\boldsymbol{\Psi}} := [\mathbf{J}_{\boldsymbol{\Psi}}^T,\ldots,\mathbf{J}_{\boldsymbol{\Psi}}^T]^T$ and $\tilde{\mathbf{W}}_{(k)} := \text{bdiag}\{\tilde{\mathbf{W}}_m^T \otimes \tilde{\mathbf{W}}_m\}$. The latter matrix can be interpreted as "projecting out" the contribution of the terms $\mathbf{A}_m\mathbf{A}_m^H$ in $\hat{\mathbf{r}}$ (incorporating an optimal weighting), after which the remaining term $\boldsymbol{\Psi}$ can be estimated. Estimation of $\boldsymbol{\Psi}$ from (49) is computationally efficient, compared to the original problem (32). The problem is convex, and additional constraints such as positivity of $\boldsymbol{\Psi}$ could also be incorporated.

This approach can be formulated as a new Alternating Weighted Least Squares (AWLS) algorithm. Starting from an initial estimate for $\boldsymbol{\Psi}$, in the iteration we estimate the $\mathbf{A}_m$ using the EVD of $\hat{\mathbf{R}}_m - \boldsymbol{\Psi}$ (for $\mathbf{W}_m = \mathbf{I}$) or $\boldsymbol{\Psi}^{-1/2}\hat{\mathbf{R}}_m\boldsymbol{\Psi}^{-1/2}$ (for $\mathbf{W}_m = \hat{\mathbf{R}}_m^{-1}$), similar to Sec. IV-A. Next, we calculate $\tilde{\mathbf{W}}_m$ using (43), which depends only on $\mathbf{A}_m$ and $\mathbf{W}_m$, followed by solving (48) or equivalently (49). For classical FA, this leads to the following iterations (where with abuse of notation we write $\mathbf{W}$ instead of $\mathbf{W}_1$):

$$\begin{aligned}&\mathbf{U}_{(k+1)}\boldsymbol{\Lambda}_{(k+1)}\mathbf{U}_{(k+1)}^H := \mathbf{D}_{(k)}^{-1/2}\hat{\mathbf{R}}\mathbf{D}_{(k)}^{-1/2} \qquad \text{[EVD]} \\ &\mathbf{A}_{(k+1)} := \mathbf{D}_{(k)}^{1/2}\mathbf{U}_{0,(k+1)}(\boldsymbol{\Lambda}_{0,(k+1)} - \mathbf{I})^{1/2} \\ &\tilde{\mathbf{W}} := \mathbf{W} - \mathbf{W}\mathbf{A}_{(k+1)}(\mathbf{A}_{(k+1)}^H\mathbf{W}\mathbf{A}_{(k+1)})^{-1}\mathbf{A}_{(k+1)}^H\mathbf{W} \\ &\mathbf{d}_{(k+1)} := \left[\tilde{\mathbf{W}}^T \odot \tilde{\mathbf{W}}\right]^{-1}\text{vectdiag}(\tilde{\mathbf{W}}\hat{\mathbf{R}}\tilde{\mathbf{W}}) \\ &\mathbf{D}_{(k+1)} := \text{diag}(\mathbf{d}_{(k+1)}).\end{aligned}$$

Note that for $\mathbf{W} = \hat{\mathbf{R}}^{-1}$, we have $\tilde{\mathbf{W}}\hat{\mathbf{R}}\tilde{\mathbf{W}} = \tilde{\mathbf{W}}$. Hence this calculation is not needed.

For $\mathbf{W} = \mathbf{I}$, $\tilde{\mathbf{W}} = \mathbf{P}_{\mathbf{A}}^{\perp}$ is the projection matrix for the orthogonal subspace of $\mathbf{A}$. If we are only interested in the null space of $\mathbf{A}$ for applications such as MUSIC or spatial filtering, we can avoid calculating $\mathbf{A}$ and obtain $\tilde{\mathbf{W}}$ directly from the EVD of $\hat{\mathbf{R}} - \mathbf{D}$.

VI. Goodness of fit and Detection

One of the parameters that needs to be found for the FA model is the factor dimension Q (i.e., $\text{rank}(\mathbf{A})$). In array processing, this relates to detecting the number of sources that the array is exposed to. An extensive literature exists on this topic (an overview can be found in [37, pp. 222-223] [38]). Here we limit the discussion to a general likelihood ratio test (GLRT), which is used to decide whether the FA model fits a given sample covariance matrix. We consider the classical FA model and aim to detect the smallest number of sources for which the model fits the data.

For each Q, two hypotheses are tested against each other. $\mathcal{H}_0$ assumes that there is an FA model underlying the data, while $\mathcal{H}_1$ assumes no structure. For a threshold γ, consider the GLRT

$$\zeta = \frac{L_1^*}{L_0^*} \gtrless \gamma,$$

where L_1^* is the maximum value of the likelihood when $\mathcal{H}_1$ is true, and L_0^* is the maximum value of the likelihood for an FA model. Taking the natural logarithm from both sides we see that the likelihood ratio reduces to

$$\begin{aligned}\lambda &= 2\log(\zeta) \\ &= 2N\left[\text{tr}(\mathbf{R}^{-1}\hat{\mathbf{R}}) - \log|\mathbf{R}^{-1}\hat{\mathbf{R}}| - P\right],\end{aligned}\tag{50}$$

where $\mathbf{R}$ is the best-fitting model with Q sources. From [6, p.267] [39, p.281] we know that λ has an asymptotic χ_s^2 distribution under H_0, where for complex data $s = (P-Q)^2 - P$ is the degree of freedom, as defined by (51) later in Appendix A. We can use this statistic to find a constant false alarm ratio detector. In the special case where $Q = 0$ this test indicates whether there are any sources active during the measurement.

If the GLRT passes for a given estimate Q_0 it also passes for any $Q > Q_0$, and if it fails it also fails for any $Q < Q_0$. Therefore, instead of a linear search for $\hat{Q}$ we propose to use a binary search. In this case the number of needed FA estimates is on average $\log_2(Q_{\max}) + 1$, where $Q_{\max}$ is the maximum number of possible sources for FA given by $Q_{\max} < P - \sqrt{P}$ as shown later in (52).

TABLE I: Complexity of various algorithms per iteration

Model	*Approach*	*Flops per iteration (order)*
FA	Ad Hoc (Sec.IV-A)	P^3
	CM [16]	P^3
	KLD/EM [40]	$P^2Q + Q^3$
	AWLS (Sec.V-D)	P^3
JFA	Krylov NLWLS (Sec.V-A+V-B)	$I_K \sum_m P^2 Q_m$
	Direct NLWLS (Sec.V-A+V-C)	$P^3 + \sum_m P^2 Q_m$
	AWLS (Sec.V-D)	$MP^3 + \sum_m P^2 Q_m$

VII. Computational Complexity

Table I gives an overview of the available and proposed algorithms and shows the complexity for a single iteration of each.

For classical FA, some original algorithms to compare with are the Ad Hoc iterations (Sec. IV-A), the ML approach solved using Conditional Maximization (CM, [16]), or using iterations that minimize the Kullback-Leibler Divergence as a prototype EM algorithm (KLD/EM, [40]). Here we propose to use the new AWLS algorithm presented in Sec. V-D. The main computational complexity is caused by inverting a $P \times P$ matrix and computing the EVD of a $P \times P$ matrix inside the iteration, both with a complexity of order P^3. The number of iterations needed for AWLS is usually very small (see Sec. VIII). Thus, the total complexity of this algorithm for FA is similar to EVD. The Ad Hoc, CM and KLD/EM algorithms have a similar complexity per iteration. However, simulations show that the number of iterations and hence the total complexity of the Ad Hoc and KLD/EM methods is very large. For CM, the number of iterations appear to be two or three times larger than for AWLS, and much more for large Q.

For the JFA model, the available algorithms are based on solving a nonlinear WLS using Gauss-Newton iterations (Sec. V-A), where the key step is solution of $\mathbf{B}\boldsymbol{\delta} = \mathbf{g}$ (32). This could be implemented using a Krylov subspace method (Krylov NLWLS), Sec. V-B. Alternatively, we proposed a direct method (Sec. V-C), where for JFA the main complexity is in the formation of $\tilde{\mathbf{W}}_m$ $(m = 1, \ldots, M)$.

In the table, I_K is the number of iterations needed for the Krylov solver to converge. This number can be chosen to be very small depending on how much improvement is desired with respect to the descent direction provided by the gradient. In the simulations presented next we allow the solver to fully converge based on the default convergence criteria of MinresQLP. For relatively large P (e.g., $P = 100$) I_K is usually less than P, which is a factor $2Q + 1$ smaller than the dimension of the matrix $\mathbf{B}$. This estimate for I_K is based on a final error of $\|\mathbf{B}\boldsymbol{\delta} - \mathbf{g}\|_2 < 10^{-12}$.

In summary, it appears there is no specific computational or storage advantage of Krylov over the direct method. For equal $Q_m = Q$, the computational complexity is of order MP^2Q.

VIII. Simulations

We evaluate the performance of the proposed models and algorithms using a series of simulations. In Sec. VIII-A, we evaluate the convergence speed of the various algorithms, then in Sec. VIII-B, we evaluate the quality of the estimated subspace using classical and Joint FA, and finally in Sec. VIII-D we show that the proposed algorithm for JEFA converges to the CRB as the number of samples becomes large.

A. Convergence Speed

We first evaluate the convergence speed of various algorithms for the classical factor analysis model. An array with $P = 100$ elements is simulated. The matrix $\mathbf{A}$ is chosen randomly with a standard complex Gaussian distribution (i.e. each element is distributed as $\mathcal{CN}(0, 1)$) and $\mathbf{D}$ is chosen randomly with a uniform distribution between 1 and 5.

For $P = 100$, the maximum number of sources is $Q_{\max} = 89$. We show simulation results for $Q = 20$, representative for low-rank cases, and for $Q = 80$ for high-rank cases. Sources and noise are generated using standard unit power complex Gaussian distributions.

The algorithms that we consider are the proposed AWLS (Sec. V-D), the proposed Direct NLWLS (Sec. V-C), and the ML scoring method (Sec. IV-B1) combined with the Krylov solver in Sec. V-B, referred to as Krylov Scoring. We compare with the classical Ad Hoc (Sec. IV-A), the WLS method by Jöreskog [14], and the more recent CM [16] and KLD/EM [40] as a representative of many other EM-type algorithms. We believe that this gives a good range of algorithms indicative of the state-of-the-art (for lack of an agreed standard).[6]

The same initial point is chosen for all the algorithms. As in other literature, we initialize with $\mathbf{D}_{(0)} = [\text{diag}(\hat{\mathbf{R}}^{-1})]^{-1}$.

Fig. 1 shows the convergence rate of the different ML algorithms based on the magnitude of the gradient. In the different panels we vary the number of sources Q and the number of samples N, where $N \to \infty$ represents the case where the covariance data is exactly equal to its model.

We observe that AWLS consistently outperforms all other presented algorithms in terms of the number of iterations needed to reduce the gradient to a given threshold. Typically 10 iterations or less are needed. The Direct NLWLS converges equally fast for infinite data (true $\mathbf{R}$) but it is seen to degrade for finite data size ($N = 1000$), with convergence around 30–40 iterations. Next, Krylov scoring requires consistently around 40–50 iterations. The CM method performs well for smaller Q but not for large Q, where it requires around 100 iterations. The Ad Hoc method is seen to be very sensitive to Q and converges orders of magnitude slower for larger Q, and KLD/EM always converges slowly.

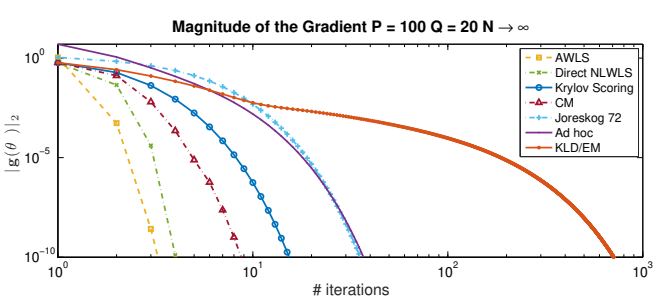
While these results are based on a single realization of the data, we consider the outcome as typical.

B. Subspace Estimation Performance

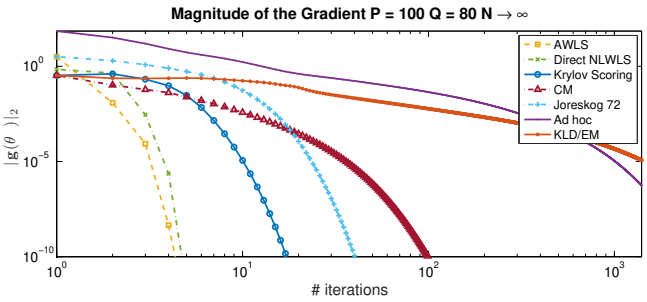
Next, we study the subspace estimation performance of FA and JFA in comparison to EVD for $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$. This gives an indication of the performance penalty if we use FA even if the noise is white and EVD is suitable.

We have chosen $Q_m = 2$, $P = 5$, $M = 5$ and $\sigma = 1$ is the noise power. We study the subspace estimation performance

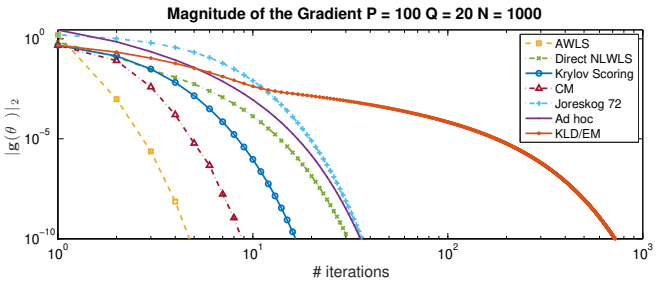
[6] E.g., the Matlab algorithm *factoran* optimizes the ML cost function using a standard optimization toolbox.



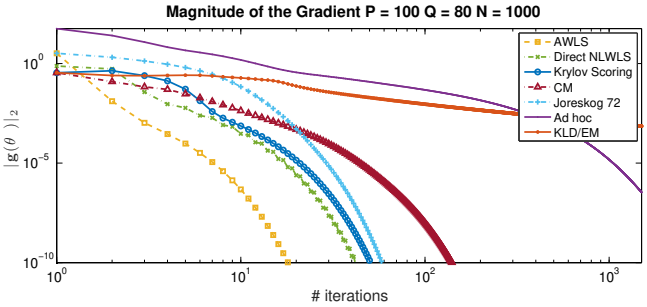
(a) Convergence speed (asymptotic for $N \to \infty$) for $Q = 20$



(b) Convergence speed (asymptotic for $N \to \infty$) for $Q = 80$



(c) Convergence speed for $N = 1000$, $Q = 20$



(d) Convergence speed for $N = 1000$, $Q = 80$

Fig. 1: Convergence for $P = 100$ sensors and varying number of samples N and sources Q.

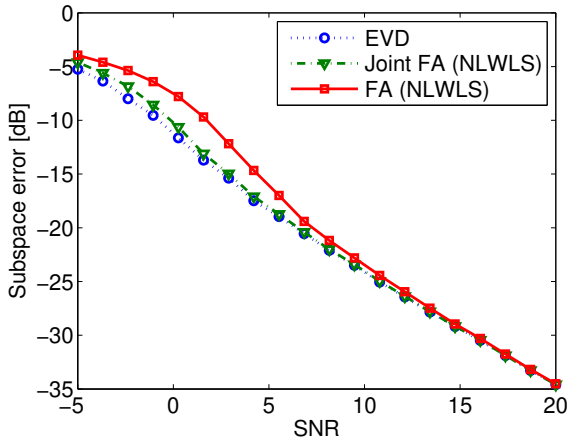


Fig. 2: Subspace error as function of SNR for $\mathbf{\Psi} = \mathbf{I}$ (white noise).

for various signal-to-noise ratios (SNR) ranging from -5 dB to 20 dB per antenna. Each sample covariance matrix is generated using $N = 100$ samples and $\mathbf{A}_m$ is generated as a random complex matrix.

As metric for the accuracy of the estimated subspace, we define a projection matrix $\hat{\mathbf{P}}_m$ onto the null-space of the estimated $\hat{\mathbf{A}}_m$, and measure

$$\text{Subspace error} = \sum_m \frac{\|\hat{\mathbf{P}}_m \mathbf{A}_m \mathbf{A}_m^H \hat{\mathbf{P}}_m\|_F}{\|\mathbf{A}_m \mathbf{A}_m^H\|_F}.$$

Fig. 2 shows the result. FA is the case where the model parameters are estimated separately for each of the $M = 5$ covariance matrices, while JFA shows the effect of jointly processing with a common $\mathbf{\Psi}$. In both cases, the Direct NLWLS algorithm is used. Because FA and JFA have to estimate more parameters, we expect a drop in performance compared to EVD. The simulation shows that for sufficiently high SNR, the algorithms behave the same, while some performance drop occurs for FA at low SNR. JFA exploits the stationarity of the noise component and has a negligible performance penalty with respect to EVD.

We conclude that the use of (J)FA does not result in a significant performance loss, while this model is more general than the white-noise model, making it applicable in many practical situations, e.g., cases where the sensor array is not (yet) accurately calibrated.

C. Convergence to local minima

One of the important points of concern for the proposed iterative methods is the possible convergence to a local minimum. By estimating the distribution of the subspace error we argue that for the NLWLS algorithms proposed in this paper this possibility does not create statistical artifacts in the solutions.

The data set is generated using $P = 100$, $Q = 70$, $M = 1$, $N = 500$ and $\mathbf{\Psi} = \mathbf{I}$. To simulate $\mathbf{A}$, the 100 receivers are randomly spread over an area of 6×6 wavelengths, and 70 sources of equal strength are randomly chosen with minimum angular distance slightly less than 1 degree. This is repeated for each iteration of a Monte-Carlo run. In total 20K runs have been performed.

The results of the Monte-Carlo simulations are used to create the histogram Fig. 3, which shows the distribution of the subspace errors for NLWLS using the direct method and for EVD. We also found that a log-normal distribution fits the histograms quite well.

The smooth behavior of the histogram and its similarity to the behavior of EVD indicates that there are no outliers (large subspace errors) beyond expected deviations of the subspace

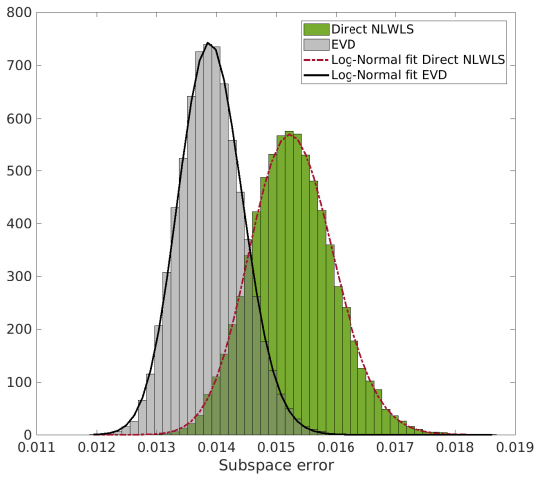


Fig. 3: Distribution of subspace errors for $\mathbf{\Psi} = \mathbf{I}$.

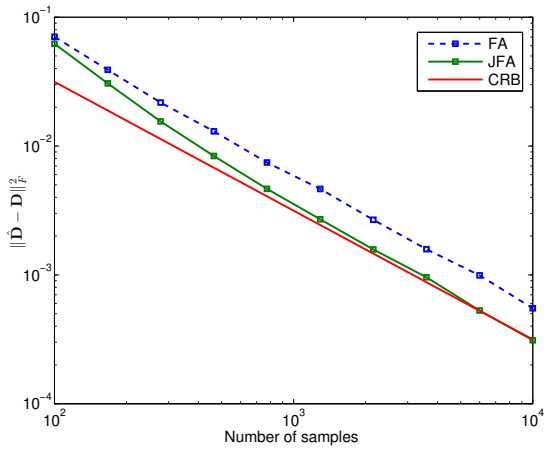


Fig. 4: Performance of the diagonal estimates compared to the CRB

due to finite sample noise. We conclude that the convergence of the algorithm is reliable.

D. Comparison to the Cramér-Rao Bound

In this part we investigate the performance of the proposed Direct NLWLS algorithm using the Cramér-Rao bound in a JFA setting. We use a setup with $P = 5$, $Q_m = 2$, $\mathbf{\Psi} = \mathbf{D}$ with diagonal elements ranging from 0.5 to 1.5. Two different approaches are compared. The first approach is to apply FA separately and then use $\hat{\mathbf{D}} = 1/K \sum_m \hat{\mathbf{D}}_m$. The other approach is to estimate $\hat{\mathbf{D}}$ using JFA.

To measure the performance we use

$$\begin{aligned}\mathcal{E}\{\|\hat{\mathbf{D}} - \mathbf{D}\|_F^2\} &= \mathcal{E}\{\text{vect}(\hat{\mathbf{D}} - \mathbf{D})^H \text{vect}(\hat{\mathbf{D}} - \mathbf{D})\} \\ &= \text{tr}[\mathcal{E}\{\text{vect}(\hat{\mathbf{D}} - \mathbf{D})\text{vect}(\hat{\mathbf{D}} - \mathbf{D})^H\}] \\ &\geq \text{tr}(\mathbf{C}_{\mathbf{\Psi}}),\end{aligned}$$

where $\mathbf{C}_{\mathbf{\Psi}}$ is the sub-matrix of the CRB corresponding to $\mathbf{\Psi}$ that was derived previously in [31]. We estimate $\mathcal{E}\{\|\hat{\mathbf{D}}-\mathbf{D}\|_F^2\}$ using Monte Carlo simulations. Fig. 4 shows the result of this simulation. This figure clearly illustrates that the proposed joint estimation reaches the CRB asymptotically and that applying the estimation separately followed by an averaging results in a sub-optimal estimation with higher variance.

E. Experimental results

The potential of FA and (J)EFA in practical scenarios was demonstrated for spatial filtering of RFI signals present in astronomical data in [31]. Calibration of the Westerbork radio telescope array ($P = 14$ dishes) using the Ad Hoc approach was shown in [23]. Calibration of one station of the LOFAR radio telescope array ($P = 96$ antennas) was reported in [26], [41], [42], and this application is run in daily practice of the array [43]. Using LOFAR data, EFA was demonstrated in [44] to suppress the Milky Way (broadband emission) from a mixture with point sources.

IX. Conclusions

We proposed extensions of the Factor Analysis model to multiple matrices and more general noise covariance structures, and we presented efficient estimation algorithms based on Gauss-Newton iterations. For the classical FA model, we derived a straightforward Alternating WLS algorithm that converges much faster than currently used techniques.

The simulations indicated the reliability and efficiency of the proposed algorithms, showing them feasible for moderately large problem sizes ($P = 100$ sensors).

We consider FA as an extension of the eigenvalue decomposition (EVD) to cases where the noise is not white. The simulations indicated that even if the noise is white, the performance penalty with respect to EVD is minor. Therefore, the more general structure of the extended FA data models enable their application in a wide range of signal processing applications.

Appendix A
Identifiability

One of the challenges with the FA models is the problem of identifiability. As in [45] we call two solutions, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, observationally equivalent if for a set of observations with probability density $p(\mathbf{x}; \boldsymbol{\theta})$, we have $p(\mathbf{x}; \boldsymbol{\theta}_1) = p(\mathbf{x}; \boldsymbol{\theta}_2)$. The problem is called (globally) identifiable if for a solution $\boldsymbol{\theta}$, there are no observationally equivalent solutions on the entire solution space $\mathbf{\Theta}$.

The question we address in this Appendix is: Given a Hermitian matrix $\mathbf{R}$ with decomposition $\mathbf{R} = \mathbf{R}_0 + \mathbf{D}$, with $\mathbf{R}_0 = \mathbf{A}\mathbf{A}^H$ of rank Q and $\mathbf{D}$ diagonal, are $\mathbf{R}_0$ and a non-singular $\mathbf{D}$ identifiable?

Early results on the identification problem were published in [3]. Later work on this subject has been summarized by [46], while [47] gives a more recent overview of important theorems on this subject. However only for $Q = 1$ or $Q = 2$ (very small ranks) do these theorems provide both sufficient and necessary conditions of identifiability. Here we use the

results provided by [45] to formulate necessary and sufficient conditions for identifiability.

A necessary condition for identifiability is that the number of knowns exceeds the number of unknowns. This puts a limit on Q, the number of columns of $\mathbf{A}$. To find this limit we study the degrees of freedom we have for the estimation parameter vector based on a given sample covariance matrix.

For classical FA, the sample covariance matrix consists of $\frac{1}{2}P(P-1)$ complex and P real observations, which are in total P^2 real knowns. The FA model has PQ complex parameters in $\mathbf{A}$ and P real parameters in $\mathbf{D}$, or $2PQ+P$ real parameters in total. We pose Q^2 constraints on $\mathbf{A}$, cf. Sec. II-B. As such the total degrees of freedom becomes[7]

$$\begin{aligned} s &= P^2 - (2PQ + P) + Q^2 \\ &= (P-Q)^2 - P\,. \end{aligned} \tag{51}$$

For the FA model to be identifiable, $s > 0$ is a necessary condition. Solving for Q, we see that the maximum number of sources that could theoretically be detected by classical FA is

$$Q < P - \sqrt{P}\,. \tag{52}$$

Following the same procedure for EFA we find

$$Q < P - \sqrt{\mathrm{tr}(\mathbf{M}^2)}\,, \tag{53}$$

where $\mathrm{tr}(\mathbf{M}^2)$ represents the total number of nonzero entries in the mask $\mathbf{M}$. To find a bound on Q for JEFA we assume for simplicity that Q_m is constant. In this case we find

$$Q < P - \sqrt{\frac{\mathrm{tr}(\mathbf{M}^2)}{M}}. \tag{54}$$

Next, we study identifiability in more detail. Typically, an estimation problem is considered identifiable if the corresponding Fisher Information Matrix (FIM) $\mathbf{F}$ is nonsingular. In our case, some refinements are needed. We know already that the problem has to be complemented with constraints. Further, the FIM depends on the actual parameter values while we would like to say something that relates to the structure of the problem.

To connect to known literature on identifiability, we briefly consider a parameterization of the unknowns in terms of real values. Let $\boldsymbol{\theta}_R$ denote such a parameterization. One way to define $\boldsymbol{\theta}_R$ for classical FA is

$$\boldsymbol{\theta} = \mathbf{T}\boldsymbol{\theta}_R, \tag{55}$$

where

$$\mathbf{T} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_{PQ} & j\mathbf{I}_{PQ} & \mathbf{0} \\ \mathbf{I}_{PQ} & -j\mathbf{I}_{PQ} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sqrt{2}\mathbf{I}_P \end{bmatrix} \tag{56}$$

has size $(2PQ+P) \times (2PQ+P)$. It is straightforward to show that $\mathbf{T}$ is a unitary (and invertible) transformation and hence does not change the number of real unknowns. The Q^2 constraints on $\boldsymbol{\theta}_R$ are written in the form of a function $\mathbf{h}_R(\boldsymbol{\theta}_R) = \mathbf{0}$. The Jacobian of this function is

$$\mathbf{H}_R = \frac{\partial \mathbf{h}_R}{\partial \boldsymbol{\theta}_R^T} \tag{57}$$

of size $Q^2 \times (2PQ+P)$. Let $\mathbf{F}$ be the FIM of the problem, and let $\mathbf{F}_R = \mathbf{T}^H\mathbf{F}\mathbf{T}$ be the real FIM. Define

$$\mathbf{V}_R(\boldsymbol{\theta}_R) = \begin{bmatrix} \mathbf{F}_R \\ \mathbf{H}_R \end{bmatrix}.$$

Reformulating Theorem 2 of [45] gives:

> Suppose $\boldsymbol{\theta}_0$ is in the solution space of $\mathbf{h}(\boldsymbol{\theta}_R) = \mathbf{0}$ and is a regular point of $\mathbf{H}_R(\boldsymbol{\theta}_R)$, and assume $\mathrm{rank}(\mathbf{A}) = Q$. Then $\boldsymbol{\theta}_0$ is locally identifiable if and only if $\mathrm{rank}[\mathbf{V}_R(\boldsymbol{\theta}_0)] = 2PQ + P$.

This means that for a locally identifiable problem $\mathbf{V}_R$ has full column rank. If $\mathrm{rank}(\mathbf{V}_R) < 2PQ + P$, there is another parameterization $\mathbf{R} = \mathbf{R}_0 + \mathbf{D} = \mathbf{R}_1 + \mathbf{D}_1$ such that $\mathrm{rank}(\mathbf{R}_1) \leq \mathrm{rank}(\mathbf{R}_0)$ and $\mathbf{D} \neq \mathbf{D}_1$. In that case the matrix $\mathbf{D}$ cannot be uniquely estimated, and the problem should be complemented with constraints on $\mathbf{D}$ itself. E.g., if in array processing the array signature combined with the noise covariance matrix is unidentifiable then $\mathbf{D}$ also contains part of the signal power and one of the signal subspaces is lost.

In this paper we assume that the signal and noise have proper complex Gaussian distributions. This can be used to simplify the identification criteria. Using Bang's formula we can write the FIM as

$$\mathbf{F}_R = \mathbf{J}_R^H \left(\mathbf{R}^{-T} \otimes \mathbf{R}^{-1}\right) \mathbf{J}_R, \tag{58}$$

where $\mathbf{J}_R = \mathbf{J}\mathbf{T}$. Let $\mathbf{H} = \mathbf{H}_R\mathbf{T}^H$. Considering that $\mathbf{R}^{-T} \otimes \mathbf{R}^{-1}$ is a positive definite matrix and that $\mathbf{H}^H\mathbf{H}$ has the same row space as $\mathbf{H}$, we have

$$\begin{aligned} \mathrm{rank}(\mathbf{V}_R) &= \mathrm{rank}\left(\begin{bmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{V}_R\mathbf{T}^H\right) \\ &= \mathrm{rank}\left(\begin{bmatrix} \mathbf{F} \\ \mathbf{H} \end{bmatrix}\right) \\ &= \mathrm{rank}\left(\begin{bmatrix} \mathbf{J} \\ \mathbf{H} \end{bmatrix}\right). \end{aligned} \tag{59}$$

This means that by studying the rank of the Jacobian we can establish the identifiability of the problem. With Q^2 suitable constraints, $\mathbf{H}$ adds Q^2 independent rows to $\mathbf{J}$, and we require the rank of $\mathbf{J}$ to be (at least) $2PQ + P - Q^2$.

Next, we establish that the rank of $\mathbf{J}$ solely depends on the diagonal structure of $\mathbf{D}$ and on the column span of $\mathbf{A}$, but not on the actual values of $\mathbf{R}$, $\mathbf{D}$ or the power of the sources. Let $\mathbf{A} = \mathbf{U}_0\boldsymbol{\Gamma}^{1/2}\mathbf{Q}^H$ be the (economical) singular value decomposition of $\mathbf{A}$, where $\mathbf{U}_0$ forms an orthonormal basis for the column span of $\mathbf{A}$. We use the structure of $\mathbf{J}$ in (21) to obtain

$$\begin{aligned} \mathbf{J} &= [\mathbf{A}^* \otimes \mathbf{I},\ (\mathbf{I} \otimes \mathbf{A})\mathbf{K},\ \mathbf{I} \circ \mathbf{I}] \\ &= [\mathbf{U}_0^*\boldsymbol{\Gamma}^{1/2}\mathbf{Q}^T \otimes \mathbf{I},\ (\mathbf{I} \otimes \mathbf{U}_0\boldsymbol{\Gamma}^{1/2}\mathbf{Q}^H)\mathbf{K},\ \mathbf{I} \circ \mathbf{I}] \\ &= \tilde{\mathbf{U}} \begin{bmatrix} \boldsymbol{\Gamma}^{1/2}\mathbf{Q}^T \otimes \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{I} \otimes \boldsymbol{\Gamma}^{1/2}\mathbf{Q}^H)\mathbf{K} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}, \end{aligned} \tag{60}$$

where

$$\tilde{\mathbf{U}} = [\mathbf{U}_0^* \otimes \mathbf{I},\ \mathbf{I} \otimes \mathbf{U}_0,\ \mathbf{I} \circ \mathbf{I}], \tag{61}$$

has size $P^2 \times (2PQ+P)$. The latter factor of $\mathbf{J}$ is square and invertible, so that the rank of $\mathbf{J}$ is equal to the rank of $\tilde{\mathbf{U}}$, which only depends on $\mathbf{U}_0$ and the diagonal structure of $\mathbf{D}$ (which is captured by $\mathbf{I} \circ \mathbf{I}$). For the problem to be (locally)

[7] For real-valued data, a similar derivation shows $s = \frac{1}{2}[(P-Q)^2 - (P+Q)]$.

identifiable we need

$$\text{rank}(\tilde{\mathbf{U}}) = 2PQ + P - Q^2. \tag{62}$$

Further, we can show that the submatrix of $\tilde{\mathbf{U}}$ given by the $2PQ$ columns

$$\tilde{\mathbf{U}}_1 = [\mathbf{U}_0^* \otimes \mathbf{I},\ \mathbf{I} \otimes \mathbf{U}_0]$$

has (at least) Q^2 dependent columns. To show this we use the fact that $\tilde{\mathbf{U}}_1\mathbf{Z} = \mathbf{0}$, where

$$\mathbf{Z} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_Q \otimes \mathbf{U}_0 \\ -(\mathbf{U}_0^* \otimes \mathbf{I}_Q) \end{bmatrix}$$

is a unitary basis of size $2PQ \times Q^2$ for the null space of $\tilde{\mathbf{U}}_1$ (i.e. $\mathbf{Z}^H\mathbf{Z} = \mathbf{I}_{Q^2}$).

Thus, identifiability requires that the P columns of $\mathbf{I} \circ \mathbf{I}$ be linearly independent to the columns of $\tilde{\mathbf{U}}_1$.

In summary, we showed that the identifiability of the classical FA problem can be established by examining the rank of $\tilde{\mathbf{U}}$ in (61), which depends only on the column span of $\mathbf{A}$. This result is readily extended to EFA by replacing $\mathbf{I} \circ \mathbf{I}$ in (61) by $\mathbf{J}_{\mathbf{\Psi}}$. The identifiability criteria for EFA becomes

$$\text{rank}(\tilde{\mathbf{U}}) = 2PQ + \text{tr}(\mathbf{M}^2) - Q^2. \tag{63}$$

To conclude, we have used the identifiability problem to find the maximum number of sources that can be modeled using (E)FA. We have also shown that the local identifiability of (E)FA is completely defined by the signal subspace and the structure of the Jacobians with respect to the noise covariance matrix. This structure is completely defined by the masking matrix $\mathbf{M}$ in (7).

Appendix B
Proof of the direct method in Sec. V-C

We prove the expression for $\boldsymbol{\delta}$ given in (46) and (47). Although the result was obtained from executing a symbolic block-LDU decomposition of $\mathbf{D}$, we omit this derivation and only verify the result. We need to prove that $\mathbf{B}\boldsymbol{\delta} = \mathbf{g}$. To simplify the presentation we limit ourselves to the case $M = 1$, and write $\mathbf{A}, \mathbf{W}$ instead of $\mathbf{A}_m, \mathbf{W}_m$. We also write $\mathbf{S} = \mathbf{J}_{\mathbf{\Psi}}$; note that $\mathbf{S}^H\mathbf{S} = \mathbf{I}$ and $\mathbf{S}\mathbf{S}^H\text{vect}(\mathbf{\Delta}_{\mathbf{\Psi}}) = \text{vect}(\mathbf{\Delta}_{\mathbf{\Psi}})$. Then

$$\mathbf{B} = \mathbf{J}^H(\mathbf{W}^T \otimes \mathbf{W})\mathbf{J}$$

$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_{\mathbf{A}} \\ \mathbf{g}_{\mathbf{A}^*} \\ \mathbf{g}_{\mathbf{\Psi}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}))\mathbf{W}\mathbf{A} \\ (\cdots)^* \\ \mathbf{S}^H\text{vect}[\mathbf{W}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}))\mathbf{W}] \end{bmatrix}$$

and

$$\boldsymbol{\delta} = \begin{bmatrix} \frac{1}{2}\text{vect}[(\mathbf{I} + \mathbf{W}^{-1}\tilde{\mathbf{W}})\mathbf{C}] \\ \frac{1}{2}\text{vect}[(\mathbf{I} + \mathbf{W}^{-T}\tilde{\mathbf{W}}^T)\mathbf{C}^*] \\ \mathbf{S}^H\text{vect}(\mathbf{\Delta}_{\mathbf{\Psi}}) \end{bmatrix}$$

where $\mathbf{C} = (\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\mathbf{W}\mathbf{A}(\mathbf{A}^H\mathbf{W}\mathbf{A})^{-1}$, and $\mathbf{\Delta}_{\mathbf{\Psi}}$ satisfies (46), i.e.,

$$\mathbf{S}^H(\tilde{\mathbf{W}}^T \otimes \tilde{\mathbf{W}})\mathbf{S}\boldsymbol{\delta}_{\mathbf{\Psi}} = \mathbf{S}^H(\tilde{\mathbf{W}}^T \otimes \tilde{\mathbf{W}})\text{vect}[\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta})]\,. \tag{64}$$

Also define

$$\mathbf{P} = \mathbf{W}^{1/2}\mathbf{A}(\mathbf{A}^H\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^H\mathbf{W}^{1/2}$$

then $\tilde{\mathbf{W}} = \mathbf{W}^{1/2}(\mathbf{I} - \mathbf{P})\mathbf{W}^{1/2}$. Note that $\mathbf{P}$ is a projection such that $\mathbf{P}(\mathbf{W}^{1/2}\mathbf{A}) = \mathbf{W}^{1/2}\mathbf{A}$, and also $\tilde{\mathbf{W}}\mathbf{A} = \mathbf{0}$. Further, (64) with $\boldsymbol{\delta}_{\mathbf{\Psi}} = \mathbf{S}^H\text{vect}(\mathbf{\Delta}_{\mathbf{\Psi}})$ leads to

$$\tilde{\mathbf{W}}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\tilde{\mathbf{W}} = \mathbf{0}\,. \tag{65}$$

We need to prove that $\mathbf{B}\boldsymbol{\delta} = \mathbf{g}$. The first line of this expression is

$$\begin{aligned}
&\tfrac{1}{2}(\mathbf{A}^T\mathbf{W}^T\mathbf{A}^* \otimes \mathbf{W})\text{vect}[(\mathbf{I} + \mathbf{W}^{-1}\tilde{\mathbf{W}})\mathbf{C}] \\
&+\tfrac{1}{2}(\mathbf{A}^T\mathbf{W}^T \otimes \mathbf{W}\mathbf{A})\mathbf{K}\text{vect}[(\mathbf{I} + \mathbf{W}^{-T}\tilde{\mathbf{W}}^T)\mathbf{C}^*] \\
&+(\mathbf{A}^T\mathbf{W}^T \otimes \mathbf{W})\mathbf{S}\mathbf{S}^H\text{vect}(\mathbf{\Delta}_{\mathbf{\Psi}}) \\
&= \tfrac{1}{2}\text{vect}[\mathbf{W}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\mathbf{W}\mathbf{A} \\
&+\tilde{\mathbf{W}}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\mathbf{W}\mathbf{A} \\
&+\underbrace{\mathbf{W}\mathbf{A}(\mathbf{A}^H\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^H\mathbf{W}}_{=\mathbf{W}-\tilde{\mathbf{W}}}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\mathbf{W}\mathbf{A} \\
&+\mathbf{W}\mathbf{A}(\mathbf{A}^H\mathbf{W}\mathbf{A})\mathbf{A}^H\mathbf{W}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\underbrace{\tilde{\mathbf{W}}\mathbf{W}^{-1}\mathbf{W}\mathbf{A}}_{=\mathbf{0}} \\
&+2\mathbf{W}\mathbf{\Delta}_{\mathbf{\Psi}}\mathbf{W}\mathbf{A}] \\
&= \text{vect}[\mathbf{W}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\mathbf{W}\mathbf{A} + \mathbf{W}\mathbf{\Delta}_{\mathbf{\Psi}}\mathbf{W}\mathbf{A}] \\
&= \text{vect}[\mathbf{W}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}))\mathbf{W}\mathbf{A}] = \mathbf{g}_{\mathbf{A}}
\end{aligned}$$

The second line is the complex conjugate of the first line. The third line is

$$\begin{aligned}
&\tfrac{1}{2}\mathbf{S}^H\text{vect}[\underbrace{\mathbf{W}(\mathbf{I} + \mathbf{W}^{-1}\tilde{\mathbf{W}})}_{=\mathbf{W}+\tilde{\mathbf{W}}}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\cdot \\
&\cdot\underbrace{\mathbf{W}\mathbf{A}(\mathbf{A}^H\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^H\mathbf{W}}_{=\mathbf{W}-\tilde{\mathbf{W}}} \\
&+\underbrace{\mathbf{W}\mathbf{A}(\mathbf{A}^H\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^H\mathbf{W}}_{=\mathbf{W}-\tilde{\mathbf{W}}}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\cdot \\
&\cdot\underbrace{(\mathbf{I} + \tilde{\mathbf{W}}\mathbf{W}^{-1})\mathbf{W}}_{=\mathbf{W}+\tilde{\mathbf{W}}} + 2\mathbf{W}\mathbf{\Delta}_{\mathbf{\Psi}}\mathbf{W}] \\
&= \mathbf{S}^H\text{vect}[\mathbf{W}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}) - \mathbf{\Delta}_{\mathbf{\Psi}})\mathbf{W} + \mathbf{W}\mathbf{\Delta}_{\mathbf{\Psi}}\mathbf{W}] \\
&= \mathbf{S}^H\text{vect}[\mathbf{W}(\hat{\mathbf{R}} - \mathbf{R}(\boldsymbol{\theta}))\mathbf{W}] = \mathbf{g}_{\mathbf{\Psi}},
\end{aligned}$$

where in the first step we used (65).

Appendix C
Step-Size for Weighted Least squares

In Section V-A, we showed parameter estimation using a Weighted Least Squares formulation, solved by a descent algorithm such as GN. Here we discuss the selection of the step-size parameter μ and show that the optimal value can be obtained in closed form. We first investigate how the error term $\mathbf{E} = \hat{\mathbf{R}} - \mathbf{R}$ is updated after each iteration,

$$\begin{aligned}
\mathbf{E}_m^{(k)} &= \hat{\mathbf{R}}_m - \mathbf{A}_m^{(k)}(\mathbf{A}_m^{(k)})^H - \mathbf{\Psi}^{(k)} \\
\mathbf{E}_m^{(k+1)} &= \hat{\mathbf{R}}_m - \mathbf{A}_m^{(k+1)}(\mathbf{A}_m^{(k+1)})^H - \mathbf{\Psi}^{(k+1)} \\
&= \mathbf{E}_m^{(k)} - \mu\mathbf{Y}_m - \mu^2\mathbf{X}_m,
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{Y}_m &= \mathbf{\Delta}_{\mathbf{A}_m}(\mathbf{A}_m^{(k)})^H + \mathbf{A}_m^{(k)}\mathbf{\Delta}_{\mathbf{A}_m}^H + \mathbf{\Delta}_{\mathbf{\Psi}} \\
\mathbf{X}_m &= \mathbf{\Delta}_{\mathbf{A}_m}\mathbf{\Delta}_{\mathbf{A}_m}^H\,.
\end{aligned}$$

Let $\mathbf{e} = \text{vect}(\mathbf{E})$, $\mathbf{x} = \text{vect}(\mathbf{X})$ and $\mathbf{y} = \text{vect}(\mathbf{Y})$. Then the WLS cost function can be written as

$$f(\boldsymbol{\theta}, \boldsymbol{\delta}, \mu) = \sum_m \mathbf{e}_m^H(\mathbf{W}_m^T \otimes \mathbf{W}_m)\mathbf{e}_m$$

and

$$\mathbf{e}_m^{(k+1)} = \mathbf{e}_m^{(k)} - \mu\mathbf{y}_m - \mu^2\mathbf{x}_m.$$

After taking derivatives with respect to μ we need to solve

$$\frac{\partial f}{\partial \mu} = a_1\mu^3 + a_2\mu^2 + a_3\mu + a_4 = 0, \tag{66}$$

where

$$a_1 = 4\sum_m \mathbf{x}_m^H(\mathbf{W}_m^T \otimes \mathbf{W}_m)\mathbf{x}_m, \tag{67}$$

$$a_2 = 6\sum_m^m \Re\{\mathbf{y}_m^H(\mathbf{W}_m^T \otimes \mathbf{W}_m)\mathbf{x}_m\}, \tag{68}$$

$$a_3 = 2\sum_m^m \mathbf{y}_m^H(\mathbf{W}_m^T \otimes \mathbf{W}_m)\mathbf{y}_m \tag{69}$$

$$- 4\sum_m^m \Re\{\mathbf{e}_m^H(\mathbf{W}_m^T \otimes \mathbf{W}_m)\mathbf{x}_m\}, \tag{70}$$

$$a_4 = -2\sum_m \Re\{\mathbf{e}_m^H(\mathbf{W}_m^T \otimes \mathbf{W}_m)\mathbf{y}_m\}. \tag{71}$$

$\Re\{.\}$ is the real part of the argument and we have dropped the dependency on k from the notation.

This is a cubic relation where all the parameters are real, and closed-form solutions exists.

REFERENCES

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

[2] D. N. Lawley, "The estimation of factor loadings by the method of maximum likelihood.," *Proceedings of the Royal Society of Edinburgh*, vol. 60, no. 01, pp. 64–82, 1940.

[3] T. W. Anderson and H. Rubin, "Statistical inference in factor analysis," *In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 5, pp. 111 – 150, 1956.

[4] K. G. Jöreskog, "A general approach to confirmatory maximum likelihood factor analysis," *Psychometrika*, vol. 34, no. 2, pp. 183–202, 1969.

[5] D. N. Lawley and A. Maxwell, *Factor analysis as a statistical method*. 2nd. ed., New York: Am. Elsevier Publ., 1971.

[6] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. Academic Press, 1979.

[7] D. J. Bartholomew, M. Knott, and I. Moustaki, *Latent Variable Models and Factor Analysis: A Unified Approach*. John Wiley and Sons, 2011.

[8] B. Ottersten, P. Stoica, and R. Roy, "Covariance matching estimation techniques for array signal processing applications," *Digital Signal Processing*, vol. 8, no. 3, pp. 185 – 210, 1998.

[9] M. Viberg, P. Stoica, and B. Ottersten, "Array processing in correlated noise fields based on instrumental variables and subspace fitting," *IEEE Trans. Signal Process.*, vol. 43, p. 11871199, Jan. 1995.

[10] V. Nagesha and S. M. Kay, "Maximum likelihood estimation for array processing in colored noise," *IEEE Trans. Signal Process.*, vol. 44, p. 169180, Feb. 1996.

[11] P. Stoica, M. Viberg, K. M. Wong, and Q. Wu, "Maximum-likelihood bearing estimation with partly calibrated arrays in spatially correlated noise fields," *IEEE Trans. Signal Process.*, vol. 44, p. 888899, Apr. 1996.

[12] M. Wax, J. Sheinvald, and A. J. Weiss, "Detection and localization in colored noise via generalized least squares," *IEEE Tr. Signal Process.*, vol. 44, pp. 1734–1743, July 1996.

[13] K. G. Jöreskog, "Some contributions to maximum likelihood factor analysis," *Psychometrika*, vol. 32, no. 4, pp. 433–482, 1967.

[14] K. G. Jöreskog and A. S. Goldberger, "Factor analysis by generalized least squares," *Psychometrika*, vol. 37, pp. 243–260, Sep 1972.

[15] D. Rubin and D. Thayer, "EM algorithms for ML factor analysis," *Psychometrika*, vol. 47, pp. 69–76, 1982. 10.1007/BF02293851.

[16] J.-H. Zhao, P. Yu, and Q. Jiang, "ML estimation for factor analysis: EM or non-EM?," *Statistics and Computing*, vol. 18, pp. 109–123, 2008. 10.1007/s11222-007-9042-y.

[17] W. Ledermann, "On a problem concerning matrices with variable diagonal elements," *Proceedings of the Royal Society of Edinburgh*, vol. 60, pp. 1–17, 1 1940.

[18] A. Shapiro, "Weighted minimum trace factor analysis," *Psychometrika*, vol. 47, no. 3, pp. 243–264, 1982.

[19] A. Shapiro, "Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis," *Psychometrika*, vol. 47, no. 2, pp. 187–199, 1982.

[20] J. Saunderson, V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting," *SIAM Journal on Matrix Analysis and Applications*, vol. 33, no. 4, pp. 1395–1416, 2012.

[21] E. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *arXiv preprint arXiv:0912.3599*, 2009.

[22] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[23] A.-J. Boonstra and A.-J. van der Veen, "Gain calibration methods for radio telescope arrays," *IEEE Tr. Signal Processing*, vol. 51, pp. 25–38, Jan. 2003.

[24] A.-J. van der Veen, A. Leshem, and A.-J. Boonstra, "Array signal processing for radio astronomy," *Experimental Astronomy (EXPA)*, vol. 17, no. 1-3, pp. 231–249, 2004. ISSN 0922-6435.

[25] A.-J. van der Veen, A. Leshem, and A.-J. Boonstra, "Array signal processing for radio astronomy," in *The Square Kilometre Array: An Engineering Perspective* (P. Hall, ed.), pp. 231–249, Dordrecht: Springer, 2005. ISBN 1-4020-3797-x. Reprinted from Experimental Astronomy, 17(1-3),2004.

[26] S. Wijnholds and A.-J. van der Veen, "Multisource self-calibration for sensor arrays," *Signal Processing, IEEE Transactions on*, vol. 57, pp. 3512–3522, Sept 2009.

[27] P. J. Schreier, *Statistical Signal Processing of Complex-Valued Data*. Cambridge University Press, 2010.

[28] R. Schmidt, *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*. PhD thesis, Stanford university, 1981.

[29] R. Roy and T. Kailath, "ESPRIT-Estimation of signals parameters via rotational invariance techniques," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 38, pp. 984–995, July 1989.

[30] A. Hjørungnes, *Complex-Valued Matrix Derivatives with Applications in Signal Processing and Communications*. Cambridge University Press, 2011.

[31] A. Mouri Sardarabadi, A.-J. van der Veen, and A.-J. Boonstra, "Spatial Filtering of RF Interference in Radio Astronomy Using a Reference Antenna Array," *IEEE Trans. Signal Process.*, vol. 64, pp. 432–447, Jan 2016.

[32] S. Y. Lee, "The Gauss-Newton algorithm for the Weighted Least Squares factor analysis," *Journal of the Royal Statistical Society*, vol. 27, June 1978.

[33] S. M. Kay, *Fundamentals of Statistical Signal Processing, Estimation theory*, vol. Volume I. Prentice Hall, 1993.

[34] P. Gill, W. Murray, and M. Wright, *Practical optimization*. London: Academic Press, 1981.

[35] S.-C. T. Choi, *Iterative methods for singular linear equations and least-squares problems*. PhD thesis, Stanford University, 2006.

[36] S. Choi, C. Paige, and M. Saunders, "MINRES-QLP: A Krylov subspace method for indefinite or singular symmetric systems," *SIAM Journal on Scientific Computing*, vol. 33, no. 4, pp. 1810–1836, 2011.

[37] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. Prentice Hall PTR, Jan. 1998.

[38] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *Signal Processing Magazine, IEEE*, vol. 21, pp. 36 – 47, july 2004.

[39] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. Wiley, third edition ed., 2003.

[40] A.-K. Seghouane, "An iterative projections algorithm for ML factor analysis," in *IEEE Workshop on Machine Learning for Signal Processing*, pp. 333–338, Oct. 2008.

[41] S. Wijnholds, S. van der Tol, R. Nijboer, and A.-J. van der Veen, "Calibration challenges for future radio telescopes," *IEEE Signal Processing Magazine*, vol. 27, pp. 30–42, Jan 2010.

[42] A. Mouri Sardarabadi and A.-J. van der Veen, "Application of Krylov based methods in calibration for radio astronomy," in *2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 153–156, June 2014.

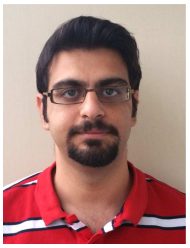
[43] M. P. van Haarlem, M. W. Wise, A. W. Gunst, *et al.*, "LOFAR: The LOw-Frequency ARray," *Astronomy & Applications*, vol. 556, p. A2, 2013.

[44] A. Mouri Sardarabadi and A.-J. van der Veen, "Subspace estimation using factor analysis," in *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 477 –480, June 2012.

[45] T. J. Rothenberg, "Identification in parametric models," *Econometrica*, vol. 39, no. 3, pp. pp. 577–591, 1971.

[46] A. Shapiro, "Identifiability of factor analysis: some results and open problems," *Linear Algebra and its Applications*, vol. 70, no. 0, pp. 1 – 7, 1985.

[47] K. Hayashi and G. A. Marcoulides, "Teacher's corner: Examining identification issues in factor analysis," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 13, no. 4, pp. 631–645, 2006.



Ahmad Mouri Sardarabadi was born in Iran in 1985. He received the Ph.D. degree from TU Delft, The Netherlands, in 2016. He is currently a postdoctoral researcher at Kapteyn Institute of Astronomy, The Netherlands, working on the analysis and development of calibration algorithms for next generation radio telescopes such as the SKA. His research interests are signal processing and multivariate analysis, with applications to radio astronomy.

Alle-Jan van der Veen was born in The Netherlands in 1966. He received the Ph.D. degree (cum laude) from TU Delft, The Netherlands, in 1993. Throughout 1994, he was a postdoctoral scholar at Stanford University, Stanford, CA. He was Chairman of the IEEE SPS Signal Processing for Communications Technical Committee from 2002 to 2004, Editor-in-Chief of the IEEE SIGNAL PROCESSING LETTERS From 2002 to 2005, and Editor-in-Chief of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2006 to 2008. At present, he is a Full Professor in Signal Processing at TU Delft. His research interests are in the general area of system theory applied to signal processing, and in particular algebraic methods for array signal processing, with applications to wireless communications and radio astronomy.