Blind Identification of Mixtures of Quasi-stationary Sources

LEE, Ka Kit

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Philosophy
in
Electronic Engineering

The Chinese University of Hong Kong
July 2012

Abstract of thesis entitled:

Blind Identification of Mixtures of Quasi-stationary Sources

Submitted by LEE, Ka Kit

for the degree of Master of Philosophy

at The Chinese University of Hong Kong in July 2012

Blind identification of linear instantaneous mixtures of quasi-stationary sources (BI-QSS) has received great research interest over the past few decades, motivated by its application in blind speech separation. In this problem, we identify the unknown mixing system coefficients by exploiting the time-varying characteristics of quasi-stationary sources. Traditional BI-QSS methods fall into two main categories: i) Parallel Factor Analysis (PARAFAC), which is based on tensor decomposition; ii) Joint Diagonalization (JD), which is based on approximate joint diagonalization of multiple matrices. In both PARAFAC and JD, the joint-source formulation is used in general; i.e., the algorithms are designed to identify the whole mixing system simultaneously.

In this thesis, I devise a novel blind identification framework using a Khatri-Rao (KR) subspace formulation. The proposed formulation is different from the traditional formulations in that it decomposes the blind identification problem into a number of per-source, structurally less complex subproblems. For the overdetermined mixing models, a specialized alternating projections algorithm is proposed for the KR subspace formulation. The resulting algorithm is not only empirically found to be very competitive, but also has a theoretically neat convergence guarantee. Even better, the proposed algorithm can be applied to the underdetermined mixing models in a straightforward manner. Rank minimization heuristics are proposed to speed up the algorithm for the underdetermined mixing model. The advantages on employing the rank minimization heuristics are demonstrated by simulations.

摘要

由於在盲語音分離的應用，線性準平穩源訊號混合的盲識別獲得了巨大的研究興趣。在這個問題上，我們利用準穩態源訊號的時變特性來識別未知的混合系統系數。傳統的方法有二：i) 基於張量分解的平行因子分析 (PARAFAC)；ii) 基於對多個矩陣的聯合對角化的聯合對角化算法 (JD)。一般來說，PARAFAC和JD都採用了源聯合的提取方法；即是說，對應所有訊號源的系統係數在算法上是同時進行識別的。

在這篇論文中，我利用Khati-Rao(KR)子空間來設計一種新的盲識別算法。在我設計的算法中提出一種與傳統的方法不同的提法。在我設計的算法中，盲識別問題被分解成數個結構上相對簡單的子問題,分別對應不同的源。在超定混合模型，我們提出了一個專門的交替投影算法 (AP)。由此產生的算法，不但能從經驗發現是非常有競爭力的，而且更有理論上的利落收斂保證。另外，作為一個有趣的延伸，該算法可循一個簡單的方式應用於欠定混合模型。對於欠定混合模型，我們提出啟發式的秩最小化算法從而提高算法的速度。

Acknowledgement

On my master study in the Chinese University of Hong Kong, I would like to thank my supervisor Prof. Wing-Kin Ma for providing me an excellent research environment. Prof. Ma teaches me that we should always focus on important insights, instead of mathematical details. Also, Prof. Ma urges me to always ask myself "what I am doing?", which helped me a great deal. Thanks to Prof. Ma, I discovered my strength and weakness. I would like to express my gratitude and appreciation to Prof. Tan Lee, Prof. Thierry Blu and Prof. Jose Bioucas for the valuable comments and discussions during my MPhil oral defense. I would like to thank Wai-Hoi To as well, for his kind help in both my research and studies. More importantly, the course teachings could not be running that smooth without his help. Thanks Fu Xiao for his help, particularly on this work. Thanks Pan Jiaxian, Li Qiang, Wu Xiaoxiao, Joseph, Wang Haipeng, David, Gary, Wilson, Liu Yun, Cai Ran for their kind helps and supports. Last but not least, I have to thank my family and friends for their encouragement and supports toward my studies. I am truly indebted and thankful.

This thesis is dedicated to my family and friends

Contents

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

Blind identification of mixtures refers to the procedure on identifying the system (or, more specifically, the mixing matrix) mixing multiple sources, using only the system output. Blind identification of mixtures is closely related to another interesting problem known as blind source separation (BSS). In fact, many BSS algorithms employ a two-steps approach: i) estimate the mixing matrix (i.e. blind identification of mixtures); and then ii) separate the sources using the estimated mixing matrix [9, 20, 28, 29]. Blind identification using sensor arrays has received great research interest over the past few decades owing to its wide applicability. Important applications include direction of arrival (DOA) estimation in sensor array processing [2, 10, 30], artifact removal in biomedical signal analysis [33], blind equalization [32] and blind beamforming [31] in wireless communication, to mention but a few. The main interest of this thesis is blind identification of speech mixtures, motivated by the well-known cocktail party problem. In the cocktail party problem, a number of people are speaking simultaneously in a room, and one is trying to follow one of the discussions.

In the context of blind identification of mixtures, properties of the source signals and/or mixing matrix are exploited. In particular, utilizing the quasi-stationarity of the source signals has become one of the most promising directions. There are two major classes of methods utilizing the quasi-stationarity. The first class of methods is known as Parallel

Factor Analysis (PARAFAC, also known as CANDECOMP) , which is based on three-way array data fitting [10–15, 22, 23]. It has strong connection to an attractive subject in mathematics known as tensor decomposition [35–37]. Making use of some powerful results in the context of tensor decomposition, a surprising identifiability result has been shown: blind identification of mixtures of quasi-stationary sources (BI-QSS) using PARAFAC is possible even when the number of sensors is roughly half of the number of sources [37]. Existing algorithms such as Trilinear Alternating Least Squares (TALS) [10,11] and Alternating-Columns Diagonal-Centers (ACDC) [12] take advantage of the aforementioned identifiability result. The second class of methods is known as Joint Diagonalization (JD), where the BI-QSS problem is formulated as a problem of jointly diagonalizating multiple matrices [4–9, 13]. Joint diagonalization of matrices is a fundamental problem in matrix analysis and has attracted much interest. Elegant results are provided in this context; see [54] and the reference therein. JD-based methods such as UWEDGE [9] and FF-DIAG [4] are well-known for their high efficiency. Another famous JD algorithm developed by Pham is equivalent to maximum likelihood estimation under the Gaussian source assumption [5, 6].

In this work, we propose a subspace approach to handle BI-QSS. Specifically, we first devise a blind identification criterion based on the quasi-stationarity of the source signals. As we will see, the criterion involves the subspace formed by the self Khatri-Rao (KR) product of the mixing matrix; therefore, we call the criterion KR subspace criterion [2]. One salient feature of the KR subspace criterion is that it suggests per-source identification, as opposed to the joint-source nature of PARAFAC and JD.

In overdetermined mixing models (i.e. we have more sensors than sources in a given system), prewhitening is commonly performed before the BI-QSS procedures [4,7]. Rather unexpectedly, the prewhitening procedure dramatically improves the convergence behavior of the proposed algorithm. With the employment of prewhitening, it can be shown that the proposed algorithm converges globally to a true mixing matrix column in one iteration with probability one. Simulation results indicate that the proposed algorithm exhibits both excellent blind identification performance and good runtime performance.

An even more challenging problem is BI-QSS in underdetermined mixing models [20, 21]. As mentioned previously, PARAFAC-based algorithms are able to handle the underdetermined mixing models [22, 23]. Apart from that, most algorithms handling the underdetermined mixing models rely on further exploiting properties of source signals, such as sparsity in time-frequency domain [26] and conditional independence [18, 19]. Based on the identifiability analysis, the KR subspace criterion devised in this work can deal with the underdetermined mixing models, without further exploiting any property of source signals. Although prewhitening is no longer possible in the underdetermined mixing models, we propose rank-minimization heuristics [40, 41, 46, 47] to speed up the identification procedure. There is an interesting connection between the proposed method and the Huber loss function known in robust statistics [55] and smooth optimization [48, 49]. Simulation results suggest that the proposed rank minimization heuristics significantly reduce the number of iterations, subsequently improving the runtime performance.

□ **End of chapter.**

Chapter 2

Settings of Quasi-Stationary Signals based Blind Identification

In this chapter, we will introduce the signal model and assumptions used in this work. Specifically, we will focus on blind identification of mixtures of quasi-stationary sources (BI-QSS). By the quasi-stationary assumption of the source signals, a specialized local covariance model will be introduced. After that, we will discuss two important and commonly employed preprocessing procedures in BI-QSS.

2.1 Signal Model

Consider a system with N sensors receiving K source signals. Denote the received signal of the nth sensor by $x_n(t)$, and the source signal emitted from the kth source by $s_k(t)$. In this thesis, we assume linear instantaneous mixing model; i.e., the received signal vector $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^T \in \mathbb{C}^N$ is represented as:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{v}(t),\ t = 0, 1, 2, \ldots, \tag{2.1}$$

where t is the time index, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K] \in \mathbb{C}^{N \times K}$ is the mixing matrix, $\mathbf{s}(t) = [s_1(t), \ldots, s_K(t)]^T \in \mathbb{C}^K$ is the source signal vector and $\mathbf{v}(t)$ is the noise in the received signal vector. The scenario is depicted in Fig. 2.1. There are K speakers speaking simultaneously and there are N microphones recording the speech mixtures. By the observations $\mathbf{x}(t)$

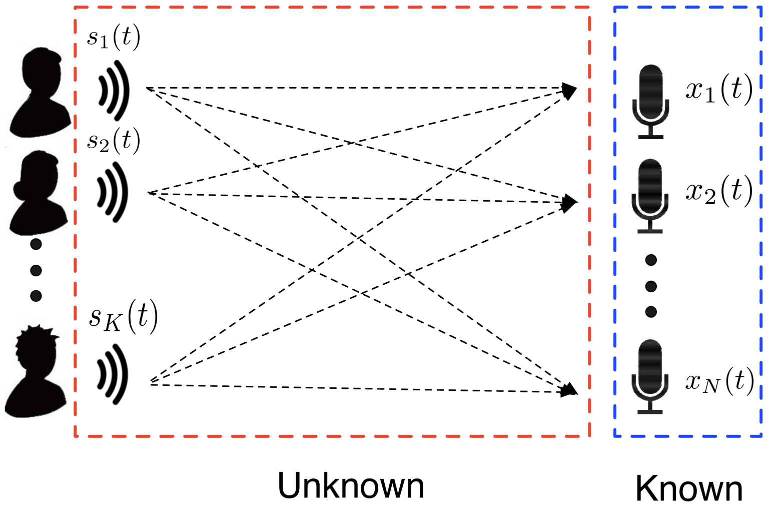

Figure 2.1: Blind identification scenario.

obtained from the microphone array, our goal is to retrieve the mixing matrix $\mathbf{A}$.

Remark: it has been noticed that a more realistic model for blind speech separation should be the linear convolutive mixing model; i.e.,

$$\mathbf{x}(t) = \mathbf{A} \star \mathbf{s}(t) + \mathbf{v}(t), \; t = 0, 1, 2, \ldots, \tag{2.2}$$

where $\star$ is the linear convolution operator. Yet, it is known that a blind identification problem with linear convolutive mixing model can be transformed into a set of blind identification problems with linear instantaneous mixing model using discrete Fourier Transform (DFT), for example [14, 23] . In other words, efficient algorithms developed based on the linear instantaneous mixing model can be extended to handle the convolutive mixing model. For the sake of self-containedness, the aforementioned transformation will be introduced briefly in Appendix A.

2.2 Assumptions

Strictly speaking, it is extremely difficult to identify the mixing matrix $\mathbf{A}$ in a completely blind fashion (i.e. no prior information at all), if not impossible. In most cases, properties of the source signals and/or the mixing matrix have to be exploited. In speech applications, there are many properties one may exploit, e.g. sparsity in the time-frequency domain [24–26], local dominance [16], conditional independence [18, 19] and non-Gaussianity [28, 29]. In this thesis, we exploit quasi-stationarity of the source signals, which is defined as:

Definition 2.1 *A random process $s(t)$ with mean*

$$\mu(t) = \mathbb{E}\{s(t)\}$$

and autocorrelation function

$$R(t, t') = \mathbb{E}\{s(t)s(t')\}$$

is said to be wide-sense quasi-stationary if $s(t)$ is wide-sense stationary for each time window $[(m-1)L+1, mL]$, with L denoting the frame length; i.e.,

$$\mu(t) = \mu_m, \ \textit{for } t \in [(m-1)L+1, mL], \tag{2.3}$$

and

$$R(t_1, t_1+\tau) = R(t_2, t_2+\tau), \ \textit{for } t_1, t_2 \in [(m-1)L+1, mL-\tau], \tau \in [0, L-1]. \tag{2.4}$$

As an illustration, a speech segment is given in Fig. 2.2. Although the speech signal is non-stationary, it is fair to say that it is stationary within short time windows (segmented by the red dotted line), i.e., quasi-stationary according to Definition 2.1.

In BI-QSS, we make the following basic assumptions:

(A1) The source signals $s_k(t), k = 1, \ldots, K$, are mutually independent, with zero mean.

(A2) The noise vector $\mathbf{v}(t)$ is wide-sense stationary with zero mean and covariance matrix $\sigma^2\mathbf{I}$, and is statistically independent of the source signals.

(A3) The source signals are wide-sense quasi-stationary with frame length L; specifically, the power of the source k in the mth frame is:

$$\mathbb{E}\{|s_k(t)|^2\} = d_{mk} \geq 0, \ \text{for } t \in [(m-1)L+1, mL].$$

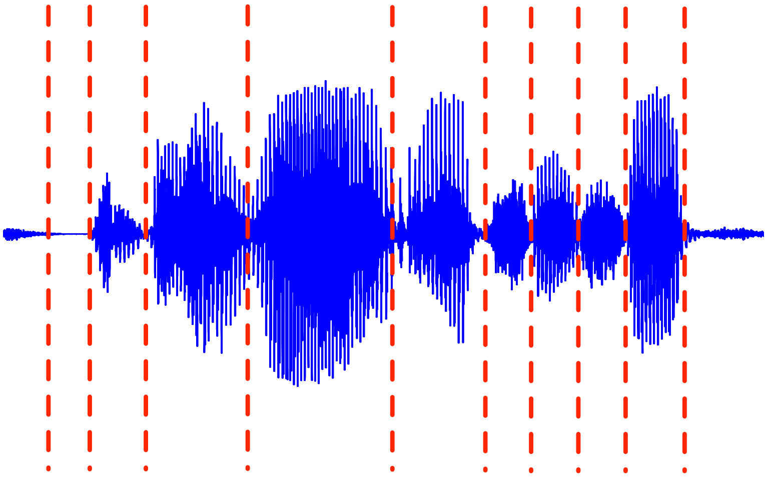

Figure 2.2: A speech segment.

Remark: it is also worthwhile to note that the structure of the mixing matrix $\mathbf{A}$ can also be exploited to aid the blind identification procedure. For example, if the sensor array has a uniform linear array structure, then under some mild assumptions $\mathbf{A}$ has a Vandemonde structure. The Vandemonde structure of $\mathbf{A}$ has been utilized to devise an efficient direction-of-arrival (DOA) estimation algorithm [2].

2.3 Local Covariance Model

With the quasi-stationary assumption (A3), we can derive one covariance matrix of $\mathbf{x}(t)$ per frame:

$$\mathbf{R}_m = \mathbb{E}\{\mathbf{x}(t)\mathbf{x}(t)^H\} \in \mathbb{C}^{N\times N}, \text{ for } t \in [(m-1)L+1, mL], \tag{2.5}$$

where m denotes the frame index. We call $\mathbf{R}_m$ the local covariance matrix of the mth frame. According to assumptions (A1), (A2) and (A3), $\mathbf{R}_m$ can be equivalently written as

$$\mathbf{R}_m = \mathbf{A}\mathbf{D}_m\mathbf{A}^H + \sigma^2\mathbf{I}, \tag{2.6}$$

where $\mathbf{D}_m = \text{Diag}(d_{m1}, \ldots, d_{mK}) \in \mathbb{R}^{K\times K}$ is the source local covariance matrix of the mth frame. In practice, $\mathbf{R}_m$ can be estimated by local time average:

$$\mathbf{R}_m = \frac{1}{L}\sum_{t=(m-1)L+1}^{mL} \mathbf{x}(t)\mathbf{x}(t)^H. \tag{2.7}$$

Now, suppose that the local covariance matrices $\{\mathbf{R}_1, \ldots, \mathbf{R}_M\}$ are available, where M is the total number of frames. The goal of BI-QSS is to retrieve $\mathbf{A}$ from $\{\mathbf{R}_1, \ldots, \mathbf{R}_M\}$ without prior knowledge on $\{\mathbf{D}_1, \ldots, \mathbf{D}_M\}$ and σ^2.

Before we introduce some existing BI-QSS algorithms, we will discuss two important preprocessing procedures in BI-QSS, which are noise covariance removal and prewhitening.

Remark: for blind speech separation utilizing the second-order statistics (SOSs), it would be interesting to investigate some typical categories of speech sounds and their impact. The discussion is given in Appendix D.

2.4 Noise Covariance Removal

It is known that the noise covariance matrix $\sigma^2\mathbf{I}$ in Eq. (2.6) can be removed before BI-QSS [1]. For overdetermined mixing models (i.e. $N > K$), the eigenvalue decomposition (EVD) of $\mathbf{R}_m$ is given by

$$\begin{aligned}\mathbf{R}_m &= \mathbf{U}_m \boldsymbol{\Lambda}_m \mathbf{U}_m^H + \sigma^2 \mathbf{I} \\ &= \mathbf{U}_m \begin{bmatrix} \lambda_{m1}+\sigma^2 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \lambda_{m2}+\sigma^2 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{mK}+\sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \mathbf{U}_m^H. \end{aligned} \tag{2.8}$$

where $\mathbf{U}_m \in \mathbb{C}^{N\times N}$ is the unitary matrix whose columns are the eigenvectors of $\mathbf{R}_m$, and $\boldsymbol{\Lambda} = \text{Diag}(\lambda_{m1}, \ldots, \lambda_{mK}) \in \mathbb{R}^{N\times N}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues.

Hence, the noise power can be estimated by

$$\hat{\sigma}^2 = \min_{m=1,\ldots,M} \lambda_{\min}(\mathbf{R}_m), \tag{2.9}$$

where $\lambda_{\min}(\mathbf{X})$ represents the smallest eigenvalue of $\mathbf{X}$. Therefore, the noise covariance matrix can be removed by

$$\mathbf{R}_m := \mathbf{R}_m - \hat{\sigma}^2\mathbf{I},\ m = 1, \ldots, M. \tag{2.10}$$

For underdetermined mixing models (i.e. $K > N$), we define the frame averaged covariance matrix $\overline{\mathbf{R}}$ as

$$\overline{\mathbf{R}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{R}_m. \tag{2.11}$$

According to our local covariance model (cf. Eq. (2.6)), we have

$$\overline{\mathbf{R}} = \mathbf{A}\overline{\mathbf{D}}\mathbf{A}^H + \sigma^2\mathbf{I}, \tag{2.12}$$

where $\overline{\mathbf{D}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{D}_m$. By this observation, the noise covariance can be removed by

$$\mathbf{R}_m := \mathbf{R}_m - \overline{\mathbf{R}} = \mathbf{A}(\mathbf{D}_m - \overline{\mathbf{D}})\mathbf{A}^H,\ m = 1, \ldots, M. \tag{2.13}$$

Therefore, we will focus on the following noise-free local covariance model in the sequel:

$$\mathbf{R}_m = \mathbf{A}\mathbf{D}_m\mathbf{A}^H,\ m = 1, \ldots, M. \tag{2.14}$$

2.5 Prewhitening

In many BI-QSS algorithms (mainly joint-diagonalization based), a preprocessing procedure called prewhitening is employed [1,4,7,12,29]. The purpose of prewhitening is to constrain the solution in order to avoid trivial solution and at the same time reduce the computational complexity. Although some authors suggested that prewhitening will cause damage to the problem structure and avoid the algorithms from getting good solutions [4, 12], empirically we find that the estimation accuracy with prewhitening can be quite satisfactory. Furthermore, as we will see, the employment of prewhitening makes a significant difference to our proposed algorithm.

According to the noise-free local covariance model, the frame averaged covariance matrix is (cf. Eq (2.12))

$$\overline{\mathbf{R}} = \mathbf{A}\overline{\mathbf{D}}\mathbf{A}^H. \tag{2.15}$$

Since $\overline{\mathbf{R}}$ is positive semi-definite, we can perform the square root factorization on $\overline{\mathbf{R}}$; i.e.,

$$\overline{\mathbf{R}} = \mathbf{B}\mathbf{B}^H, \tag{2.16}$$

where $\mathbf{B} \in \mathbb{C}^{N \times K}$. Then, the prewhitening procedure is:

$$\tilde{\mathbf{R}}_m = \mathbf{B}^\dagger \mathbf{R}_m (\mathbf{B}^\dagger)^H, \ m = 1, \ldots, M, \tag{2.17}$$

where $\mathbf{B}^\dagger$ is the Moore-Penrose pesudoinverse of $\mathbf{B}$. It can be verified that $\tilde{\mathbf{R}}_m$ can be equivalently written as

$$\tilde{\mathbf{R}}_m = \tilde{\mathbf{A}}\tilde{\mathbf{D}}_m\tilde{\mathbf{A}}^H, \ m = 1, \ldots, M, \tag{2.18}$$

where

$$\tilde{\mathbf{D}}_m = \overline{\mathbf{D}}^{-1}\mathbf{D}_m, \ \tilde{\mathbf{A}} = \mathbf{B}^\dagger \mathbf{A}\overline{\mathbf{D}}^{1/2}. \tag{2.19}$$

Note that $\tilde{\mathbf{A}}$ is an unitary matrix. In particular, Eq. (2.18) is equivalent to the noise-free local covariance model (cf. Eq. (2.14)) with unitary mixing matrix $\tilde{\mathbf{A}}$. Once the equivalent mixing matrix $\tilde{\mathbf{A}}$ is identified, the original mixing matrix $\mathbf{A}$ can be retrieved by $\mathbf{B}\tilde{\mathbf{A}}$, according to Eq. (2.19).

2.6 Summary

In this chapter, we have introduced the system model used throughout this thesis. A few basic assumptions were discussed. Special attention has been put on the crucial assumption: the quasi-stationarity of the source signals. Afterwards, the local covariance model, an important building block of BI-QSS, has been introduced. Based on that, the objective of BI-QSS was then stated explicitly. In the next two chapters, we will introduce blind identification algorithms specialized for the local covariance model.

□ **End of chapter.**

Chapter 3

Review on Some Existing BI-QSS Algorithms

In this chapter, we review some existing BI-QSS algorithms. Firstly, we will introduce joint diagonalization (JD), in which BI-QSS is handled by joint diagonalization of multiple matrices. Here, we will introduce two classical JD-based algorithms: Fast Frobenius Diagnlaization (FFDIAG) and Pham's JD. Then, we will introduce parallel factor analysis (PARAFAC), in which a three-way array data fitting approach is employed to handle BI-QSS. A highly related topic known as tensor decomposition will first be introduced briefly. The attractive identifiability result will then be discussed in detail. Following that, we will describe two popular PARAFAC-based algorithms, namely, Alternating-Columns Diagonal-Centers (ACDC) and Trilinear Alternating Least-Squares (TALS).

3.1 Joint Diagonalization

In Joint Diagonalization (JD)-based algorithms, we make one additional assumption:

(A4 for JD) The mixing matrix $\mathbf{A}$ is invertible.

In other words,

$$\mathbf{W} \triangleq \mathbf{A}^{-1} \tag{3.1}$$

exists. For the case where $N \geq K$, the invertibility can be ensured using dimension reduction methods such as prewhitening. According to the noise-free local covariance model (cf. Section 2.4), we have

$$\begin{aligned}\mathbf{W}\mathbf{R}_m\mathbf{W}^H &= (\mathbf{A}^{-1}\mathbf{A})\mathbf{D}_m(\mathbf{A}^H\mathbf{A}^{-H}) \\ &= \mathbf{D}_m,\ m = 1, \ldots, M,\end{aligned} \tag{3.2}$$

where $\mathbf{A}^{-H} \triangleq (\mathbf{A}^{-1})^H$. Simple as it is, Eq. (3.2) suggests a way to find the mixing matrix: we want to find a matrix $\mathbf{W}$ that jointly diagonalizes all the local covariance matrices $\{\mathbf{R}_1, \ldots, \mathbf{R}_M\}$. Let us start with discussing Fast Frobenius Diagonalization (FFDIAG).

3.1.1 Fast Frobenius Diagonalization [4]

Fast Frobenius Diagonalization (FFDIAG) is a method of finding a solution of the following optimization problem:

$$\min_{\mathbf{W}} \sum_{m=1}^{M} \text{off}(\mathbf{W}\mathbf{R}_m\mathbf{W}^H), \tag{3.3}$$

where off$(\cdot)$ is the off-diagonal operator defined as

$$\text{off}(\mathbf{X}) = \sum_{i \neq j} X_{ij}^2. \tag{3.4}$$

In essence, FFDIAG attempts to find a matrix $\mathbf{W}$ such that the squared sum of all off-diagonal elements of $\{\mathbf{W}\mathbf{R}_1\mathbf{W}^H, \ldots, \mathbf{W}\mathbf{R}_M\mathbf{W}^H\}$ are minimized. However, by a careful look at Problem (3.3), the solution is trivial: $\mathbf{W} = \mathbf{0}$. Obviously, it is not desirable. In FFDIAG, the trivial solution is avoided by enforcing the invertibility of $\mathbf{W}$. The demixing matrix $\mathbf{W}$ is updated multiplicatively:

$$\mathbf{W}^{(k+1)} = (\mathbf{I} + \mathbf{V}^{(k)})\mathbf{W}^{(k)}, \tag{3.5}$$

where $\mathbf{V}^{(k)}$ is constrained to be all zeros on its main diagonal and k is the iteration index. According to the update rule (3.5), to ensure the invertibility of $\mathbf{W}$, it suffices to ensure the invertibility of $\mathbf{I} + \mathbf{V}$. The matrix invertibility can be guaranteed according to the following theorem:

Theorem 3.1 *[4] If $\mathbf{X} \in \mathbb{C}^{n\times n}$ is strictly diagonally-dominant, i.e.*

$$|x_{ii}| > \sum_{j\neq i} |x_{ij}|,\ \forall\ i = 1,\ldots,n, \tag{3.6}$$

then $\mathbf{X}$ is invertible.

Owing to the special structure of $\mathbf{I}+\mathbf{V}$ and Theorem 3.1, it suffices to ensure that

$$\max_i \sum_{j\neq i} |V_{ij}| = \|\mathbf{V}\|_\infty < 1 \tag{3.7}$$

where $\|\cdot\|_\infty$ is the matrix infinity norm, defined as the maximum row sum of a matrix. Therefore, the invertibility of $\mathbf{W}$ can be ensured by dividing $\mathbf{V}$ by its infinity norm whenever it exceeds some fixed number $\theta < 1$. An even stricter condition can be imposed using the Frobenius norm:

$$\mathbf{V}^{(k)} := \frac{\theta}{\|\mathbf{V}^{(k)}\|_F}\mathbf{V}^{(k)}, \tag{3.8}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Subsequently, we want to know how to compute $\mathbf{V}$ such that the objective function of Problem (3.3) is minimized. Let us consider the update of the local covariance matrices:

$$\mathbf{R}_m^{(k+1)} = (\mathbf{I}+\mathbf{V}^{(k)})\mathbf{R}_m^{(k)}(\mathbf{I}+\mathbf{V}^{(k)})^H,\ m = 1,\ldots,M. \tag{3.9}$$

We separate the diagonal and off-diagonal parts of $\mathbf{R}_m^{(k)}$, i.e.

$$\mathbf{R}_m^{(k)} = \mathcal{D}_m^{(k)} + \mathcal{E}_m^{(k)},\ m = 1,\ldots,M, \tag{3.10}$$

where $\mathcal{D}_m^{(k)}$ and $\mathcal{E}_m^{(k)}$ denote the diagonal and off-diagonal parts of $\mathbf{R}_m^{(k)}$, respectively. In [4], the authors made an important assumption:

(A5 for FFDIAG) $\|\mathbf{V}^{(k)}\|$ and $\{\|\mathcal{E}_1^{(k)}\|,\ldots,\|\mathcal{E}_M^{(k)}\|\}$ are small.

It can be justified as follows: when the local covariance matrices $\{\mathbf{R}_1^{(k)},\ldots,\mathbf{R}_M^{(k)}\}$ are almost diagonalized (i.e. $\{\|\mathcal{E}_1^{(k)}\|,\ldots,\|\mathcal{E}_M^{(k)}\|\}$ are small), the update $\mathbf{V}^{(k)}$ should be close to zero and hence $\|\mathbf{V}^{(k)}\|$ is small. Eventually, according to Eq. (3.9), we have

$$\begin{aligned}\mathbf{R}_m^{(k+1)} &= (\mathbf{I}+\mathbf{V}^{(k)})(\mathcal{D}_m^{(k)}+\mathcal{E}_m^{(k)})(\mathbf{I}+\mathbf{V}^{(k)})^H \\ &\approx \mathcal{D}_m^{(k)} + \mathbf{V}^{(k)}\mathcal{D}_m^{(k)} + \mathcal{D}_m^{(k)}(\mathbf{V}^{(k)})^H + \mathcal{E}_m^{(k)},\end{aligned} \tag{3.11}$$

where quadratic terms are ignored because of the assumption (A5). We can see that the local covariance model is highly simplified. Furthermore, by ignoring the already diagonal term $\mathcal{D}_m^{(k)}$, $\mathbf{V}^{(k+1)}$ can be found by considering the following problem (cf. Problem (3.3)):

$$\mathbf{V}^{(k+1)} = \arg\min_{\mathbf{V}} \sum_{m=1}^{M} \text{off}(\mathbf{V}\mathcal{D}_m^{(k)} + \mathcal{D}_m^{(k)}\mathbf{V}^H + \mathcal{E}_m^{(k)}), \tag{3.12}$$

which can be solved efficiently (the exact implementation details are skipped here).

3.1.2 Pham's JD [5, 6]

Pham's JD is a classical BI-QSS algorithm making use of the principle of maximum likelihood (ML) estimation.

Firstly, assume the source signal vector $\mathbf{s}(t)$ is a K-variate Gaussian random process with zero mean and covariance matrix $\mathbf{\Sigma}(t)$. As the sources are mutually independent, $\mathbf{\Sigma}(t)$ is a diagonal matrix. According to the linear instantaneous mixing model in the absence of noise (cf. Section 2.1):

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{3.13}$$

we know that $\mathbf{x}(t)$ is also a K-variate Gaussian random process with zero mean and covariance matrix $\mathbf{A}\mathbf{\Sigma}(t)\mathbf{A}^H$. Hence, the joint probability density function (p.d.f.) of $\mathbf{x}(t)$ is:

$$p(\mathbf{x}(t)) = \frac{1}{(2\pi)^{K/2}|\det(\mathbf{A}\mathbf{\Sigma}(t)\mathbf{A}^H)|^{1/2}} e^{-\frac{1}{2}\mathbf{x}(t)^H(\mathbf{A}\mathbf{\Sigma}(t)\mathbf{A}^H)^{-1}\mathbf{x}(t)}. \tag{3.14}$$

Recall that we have assumed the source signals are quasi-stationary (cf. Section 2.2). Now, consider the time window $[(m-1)L+1, mL]$; specifically,

$$\mathbf{\Sigma}(t) = \mathbf{D}_m, \text{ for } t \in [(m-1)L+1, mL].$$

Define a function $f(\mathbf{A}, \mathbf{D}_m)$ by

$$f(\mathbf{A}, \mathbf{D}_m) \triangleq \frac{1}{L} \sum_{t=(m-1)L+1}^{mL} -\log p(\mathbf{x}(t)). \tag{3.15}$$

Then, we have (ignoring the terms independent of $\mathbf{A}$ and $\mathbf{D}_m$)

$$\begin{aligned}
f(\mathbf{A}, \mathbf{D}_m) &= \frac{1}{L} \sum_{t=(m-1)L+1}^{mL} \frac{1}{2} \left(\mathbf{x}(t)^H (\mathbf{A}\boldsymbol{\Sigma}(t)\mathbf{A}^H)^{-1} \mathbf{x}(t) + \log |\det(\mathbf{A}\boldsymbol{\Sigma}(t)\mathbf{A}^H)| \right) \\
&= \frac{1}{L} \sum_{t=(m-1)L+1}^{mL} \frac{1}{2} \left(\mathrm{Tr}(\boldsymbol{\Sigma}(t)^{-1} \mathbf{A}^{-1} \mathbf{x}(t)\mathbf{x}(t)^H \mathbf{A}^{-H}) + \log |\det(\mathbf{A}\boldsymbol{\Sigma}(t)\mathbf{A}^H)| \right) \\
&= \frac{1}{2} \left(\mathrm{Tr}(\mathbf{D}_m^{-1} \underbrace{\mathbf{A}^{-1}\hat{\mathbf{R}}_m\mathbf{A}^{-H}}_{\triangleq \mathbf{U}}) + \log |\det(\mathbf{A}\mathbf{D}_m\mathbf{A}^H)| \right) \\
&= \frac{1}{2} \left(\sum_{k=1}^{K} d_{mk}^{-1} u_{kk} + 2 \log |\det(\mathbf{A})| + \sum_{k=1}^{K} \log d_{mk} \right)
\end{aligned} \tag{3.16}$$

where d_{mk} is the (k,k)th element of $\mathbf{D}_m$ and

$$\hat{\mathbf{R}}_m \triangleq \frac{1}{L} \sum_{(m-1)L+1}^{mL} \mathbf{x}(t)\mathbf{x}(t)^H$$

is the local covariance matrix of the mth frame estimated using local time average.

In order to maximize the likelihood function, we want to minimize $f(\mathbf{A}, \mathbf{D}_m)$. Differentiating $f(\mathbf{A}, \mathbf{D}_m)$ w.r.t. d_{mk} yields:

$$\begin{aligned}
\frac{\partial f(\mathbf{A}, \{d_{m1}, \ldots, d_{mK}\})}{\partial d_{mk}} &= \frac{1}{d_{mk}} - d_{mk}^{-2} u_{kk} \\
&= \frac{d_{mk} - u_{kk}}{d_{mk}^2}.
\end{aligned} \tag{3.17}$$

By setting Eq. (3.17) to zero, the local minimizer is $d_{mk}^{\star} = u_{kk}$, $\forall\ k = 1, \ldots, K$, as $d_{mk} > 0, \forall\ m, k$.

By substituting the solution of $\mathbf{D}_m$ back to $f(\mathbf{A}, \mathbf{D}_m)$, we have (again, the terms independent of $\mathbf{A}$ are ignored)

$$\begin{aligned}
f(\mathbf{A}) \triangleq& \inf_{\mathbf{D}_m} f(\mathbf{A}, \mathbf{D}_m) \\
=& \frac{1}{2} \left(\mathrm{Tr}(\mathrm{Diag}(\mathbf{U})^{-1}\mathbf{U}) + 2 \log |\det(\mathbf{A})| + \log |\det(\mathrm{Diag}(\mathbf{U}))| \right) \\
=& \frac{1}{2} (\log |\det(\mathbf{A}\mathbf{A}^H)| + \log |\det(\hat{\mathbf{R}}_m^{-1})| - \\
& \log |\det(\hat{\mathbf{R}}_m^{-1})| + \log |\det(\mathrm{Diag}(\mathbf{A}^{-1}\hat{\mathbf{R}}_m\mathbf{A}^{-H}))|) \\
=& \frac{1}{2} \left(- \log |\det(\mathbf{A}^{-1}\hat{\mathbf{R}}_m\mathbf{A}^{-H})| + \log |\det(\mathrm{Diag}(\mathbf{A}^{-1}\hat{\mathbf{R}}_m\mathbf{A}^{-H}))| \right),
\end{aligned} \tag{3.18}$$

where Diag$(\cdot)$ forces its argument to diagonal by setting all its off-diagonal elements to zero. Furthermore, by the Hadamard's inequality, we know that for $\mathbf{X} \in \mathbb{C}^{K\times K}$

$$|\det(\mathbf{X})| \leq \prod_{i=1}^{K} x_{ii} = |\det(\mathrm{Diag}(\mathbf{X}))|. \tag{3.19}$$

Therefore, we have

$$-\log|\det(\mathbf{X})| \geq -\log|\det(\mathrm{Diag}(\mathbf{X}))|. \tag{3.20}$$

Eventually, we conclude that $f(\mathbf{A})$ is minimized when $\mathbf{A}^{-1}\hat{\mathbf{R}}_m\mathbf{A}^{-H}$ is a diagonal matrix.

For BI-QSS, we want to minimize the following function w.r.t. $\mathbf{A}, \{\mathbf{D}_1, \ldots, \mathbf{D}_M\}$:

$$g(\mathbf{A}, \mathbf{D}_1, \ldots, \mathbf{D}_M) = \sum_{m=1}^{M} \frac{1}{2}\left(\mathrm{Tr}(\mathbf{D}_m^{-1}\mathbf{A}^{-1}\hat{\mathbf{R}}_m\mathbf{A}^{-H}) + \log|\det(\mathbf{A}\mathbf{D}_m\mathbf{A}^H)|\right). \tag{3.21}$$

With similar arguments as above, we conclude that

$$\begin{aligned} g(\mathbf{A}) &= \inf_{\mathbf{D}_1,\ldots,\mathbf{D}_M} g(\mathbf{A}, \mathbf{D}_1, \ldots, \mathbf{D}_M) \\ &= \sum_{m=1}^{M} \frac{1}{2}\left(-\log|\det(\mathbf{A}^{-1}\hat{\mathbf{R}}_m\mathbf{A}^{-H})| + \log|\det(\mathrm{Diag}(\mathbf{A}^{-1}\hat{\mathbf{R}}_m\mathbf{A}^{-H}))|\right). \end{aligned} \tag{3.22}$$

Then, $g(\mathbf{A})$ is minimized if $\{\mathbf{A}^{-1}\hat{\mathbf{R}}_1\mathbf{A}^{-H}, \ldots, \mathbf{A}^{-1}\hat{\mathbf{R}}_M\mathbf{A}^{-H}\}$ are all diagonal matrices.

Remark 1: a subtle requirement by the Pham's JD is that all the local covariance matrices must be positive-definite, as the log-determinant function has been introduced.

Remark 2: fundamentally, JD-based methods are not applicable to $K > N$, as there is be no way to reduce the dimension of $\mathbf{A}$ to square and at the same time ensuring the invertibility.

3.2 Parallel Factor Analysis

Parallel Factor Analysis (PARAFAC) is a data analysis tool firstly introduced in psychometrics. Recall the noise-free local covariance model (cf. Section 2.4):

$$\mathbf{R}_m = \mathbf{A}\mathbf{D}_m\mathbf{A}^H,\ m = 1, \ldots, M. \tag{3.23}$$

For the PARAFAC-based algorithms, the blind identification criterion is:

$$\min_{\mathbf{A},\{\mathbf{D}_1,\ldots,\mathbf{D}_M\}:\text{diagonal}} \sum_{m=1}^{M} \|\mathbf{R}_m - \mathbf{A}\mathbf{D}_m\mathbf{A}^H\|_F^2, \tag{3.24}$$

which is a data fitting problem. In other words, PARAFAC finds the matrices $\{\mathbf{A}, \mathbf{D}_1, \ldots, \mathbf{D}_M\}$ that best fit the data $\mathbf{R}_m$ in a least-squares sense. From the above formulation, we can already see some advantages against JD: i) it does not have the risk of yielding the trivial solution $\mathbf{A} = \mathbf{0}$; ii) it does not require $\mathbf{A}^{-1}$ to exist. Moreover, as we will discuss, PARAFAC is applicable to underdetermined mixing models (i.e. $K > N$).

In this section, we will first discuss tensor decomposition. In particular, the identifiability result for PARAFAC-based algorithms will be discussed. Then, we will introduce the Alternating-Columns Diagonal-Centers (ACDC) and Trilinear Alternating Least-Squares (TALS) algorithms, in which Problem (3.24) is handled using alternating optimization.

3.2.1 Tensor Decomposition [37]

Tensor decomposition is an attractive topic in the context of multilinear algebra, in which the methods used in linear algebra are extended. Important applications can be found in signal processing, numerical linear algebra, computer vision, to mention but a few. For a comprehensive review on tensor decomposition, please refer to [56] and the references therein.

Consider a matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$ with rank equals 3. Then, the rank-3 decomposition of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{a}_1\mathbf{b}_1^T + \mathbf{a}_2\mathbf{b}_2^T + \mathbf{a}_3\mathbf{b}_3^T. \tag{3.25}$$

for some vectors $\mathbf{a}_i \in \mathbb{R}^I, \mathbf{b}_i \in \mathbb{R}^J, i = 1, 2, 3$. Equivalently, we write $\mathbf{X} = \mathbf{A}\mathbf{B}^T$, where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] \in \mathbb{R}^{I \times 3}$ and $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3] \in \mathbb{R}^{J \times 3}$. It is important to note that the outer product form in (3.25) is not unique, as

$$\mathbf{X} = \mathbf{A}\mathbf{B}^T = \mathbf{A}\mathbf{T}\mathbf{T}^{-1}\mathbf{B}^T,$$

for any invertible matrix $\mathbf{T}$.

Now, consider a three dimensional tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$. Denote x_{ijk} by the (i, j, k)th element of $\mathcal{X}$. The F-component trilinear decomposition of $\mathcal{X}$ is defined as

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf}, \ \forall \ i, j, k. \tag{3.26}$$

In (3.26), $\mathcal{X}$ is expressed as a sum of F rank-1 tensors, where the tensor rank is defined as the minimum number of rank-1 components required in decomposing a given tensor. A surprising result in tensor decomposition lies in its uniqueness, which is missing in matrix rank-r decomposition. Define the matrices $\mathbf{A} \in \mathbb{R}^{I \times F}$, $\mathbf{B} \in \mathbb{R}^{J \times F}$ and $\mathbf{C} \in \mathbb{R}^{K \times F}$ with

$$A_{if} = a_{if}, \ B_{jf} = b_{jf}, \ C_{kf} = c_{kf}. \tag{3.27}$$

Under some mild conditions, the tensor decomposition is unique up to scaling and permutation ambiguities; that is, given $\mathcal{X}$, $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are unique up to the inherently unresolvable ambiguities.

Before we state the uniqueness theorem of three dimensional tensor decomposition, an important concept called Kruskal rank (krank) has to be introduced.

Definition 3.1 *Given a matrix $\mathbf{X} \in \mathbb{R}^{I \times F}$. We have*

$$\text{krank}(\mathbf{X}) = r$$

if and only if every r columns of $\mathbf{X}$ are linearly independent, and there exists one set of $r+1$ columns of $\mathbf{X}$ are linearly dependent.

Then, we state the famous uniqueness result shown by Kruskal:

Theorem 3.2 *[37] Consider the F-component trilinear decomposition*

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf}, \ \forall i, j, k. \tag{3.28}$$

Given $\mathcal{X}$, the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ (defined in Eq. (3.27)) are unique up to scaling and permutation ambiguities if

$$\text{krank}(\mathbf{A}) + \text{krank}(\mathbf{B}) + \text{krank}(\mathbf{C}) \geq 2F + 2. \tag{3.29}$$

Before we further discuss the tensor decomposition, let us recall the noise-free local covariance model (cf. Section 2.4):

$$\mathbf{R}_m = \mathbf{A}\mathbf{D}_m\mathbf{A}^H,\ m = 1, 2, \ldots, M. \tag{3.30}$$

Define a matrix $\mathbf{\Psi} \in \mathbb{R}^{M\times K}$ as

$$\mathbf{\Psi} = \begin{bmatrix} \text{diag}(\mathbf{D}_1)^T \\ \text{diag}(\mathbf{D}_2)^T \\ \vdots \\ \text{diag}(\mathbf{D}_M)^T \end{bmatrix}, \tag{3.31}$$

where $\text{diag}(\mathbf{D}_m)$ takes the main diagonal of $\mathbf{D}_m$ and stacks it into a column vector. Then, the local covariance model in (3.30) can be rewritten as

$$\mathbf{R}_m = \mathbf{A}\mathcal{D}_m(\mathbf{\Psi})\mathbf{A}^H, m = 1, 2, \ldots, M, \tag{3.32}$$

where $\mathcal{D}_m(\mathbf{\Psi})$ takes the mth row of $\mathbf{\Psi}$ and stacks it into a diagonal matrix.

By writting out the (i, j)th element of $\mathbf{R}_m$ and Theorem 3.2, it can be shown that the K-component trilinear decomposition is unique if

$$\text{krank}(\mathbf{A}) + \text{krank}(\mathbf{A}^*) + \text{krank}(\mathbf{\Psi}) \geq 2K + 2. \tag{3.33}$$

In BI-QSS, it is reasonable to make the following assumptions:

(A4 for PARAFAC) The mixing matrix $\mathbf{A} \in \mathbb{C}^{N\times K}$ has full Kruskal rank.

(A5 for PARAFAC) $\mathbf{\Psi} \in \mathbb{R}^{M\times K}$ has full Kruskal rank.

Assumption (A4) means that for $N \geq K$, $\mathbf{A}$ has full column rank; for $N < K$, any K columns of $\mathbf{A}$ are linearly independent. For source signals coming from significantly different paths, assumption (A4) is well-justified. Physically, the ith column of $\mathbf{\Psi}$ describes the power distribution of the ith source. Hence, for sources having significantly different power distributions, assumption (A5) is also well-justified. Now, let us investigate the uniqueness results case by case:

Case 1: $N \geq K, M \geq K$. According to assumption (A4), condition (3.33) becomes

$$2K + \text{krank}(\mathbf{\Psi}) \geq 2K + 2. \tag{3.34}$$

According to assumption (A5), krank($\mathbf{\Psi}$) $= K$. Therefore, the tensor decomposition is unique when there are more than or equal to 2 sources (i.e. $K \geq 2$) in a given system.

Case 2: $N \geq K$ and $M < K$. Here, krank($\mathbf{\Psi}$) $= M$ according to (A5). Then, if we partition the received signals into no less than 2 frames (i.e. $M \geq 2$), the tensor decomposition is unique.

Case 3: $N < K$ and $M \geq K$. According to assumption (A4), we have krank($\mathbf{A}$)=N. Then, condition (3.33) becomes

$$2N + \text{krank}(\mathbf{\Psi}) \geq 2K + 2.$$

Also, as krank($\mathbf{\Psi}$) = K, the tensor decomposition is unique when

$$K \leq 2N - 2. \tag{3.35}$$

Case 4: $N < K$ and $M < K$. In this case, we have krank($\mathbf{\Psi}$) $= M$. Therefore, the tensor decomposition is unique when

$$K \leq N - 1 + \frac{M}{2}. \tag{3.36}$$

Remark 1: Eq. (3.35) and (3.36) shows the applicability of PARAFAC-based algorithms in underdetermined mixing models.

Remark 2: in BI-QSS, we usually have $M \gg K$ in order to capture the non-stationarity of the source signals. That is, the uniqueness result of PARAFAC used in BI-QSS is stated in Case 1 and Case 3.

Remark 3: the uniqueness result of tensor decomposition is very powerful. For a general tensor decomposition problem, we can shuffle $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ in Eq. (3.28) in order to get the best identifiability. Moreover, there exists some even more powerful probabilistic uniqueness results of tensor decomposition [11].

Another important property of tensor is the matrix unfolding. It is a generalization of vectorization in matrix case. Vectorization is a procedure in converting a matrix into a vector, while preserving all the elements; matrix unfolding is to convert a tensor into multiple matrices. A three dimensional tensor gets three different matrix unfoldings. Consider a three dimensional tensor as a cuboid, as shown in Fig. 3.1. There are three different ways to cut the cuboid into slices (which are matrices), corresponding to three different matrix unfoldings. Then, matrix unfoldings of a three dimensional tensor are obtained by stacking the slices into big matrices. Matrix unfolding is an important procedure used in TALS, as we will see shortly.

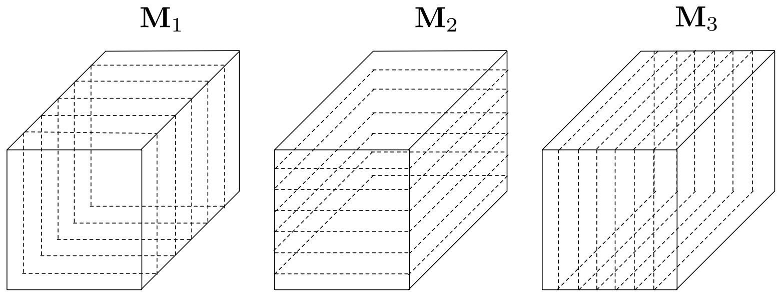

Figure 3.1: Matrix unfoldings.

3.2.2 Alternating-Columns Diagonal-Centers [12]

Recall the blind identification criterion of PARAFAC:

$$\min_{\mathbf{A},\{\mathbf{D}_1,\ldots,\mathbf{D}_M\}:\text{diagonal}} \sum_{m=1}^{M} \|\mathbf{R}_m - \mathbf{A}\mathbf{D}_m\mathbf{A}^H\|_F^2. \tag{3.37}$$

The idea of Alternating-Columns Diagonal-Centers (ACDC) is to use alternating optimization to handle Problem (3.37). Specifically, we first update $\mathbf{A}$ with fixed $\{\mathbf{D}_1, \ldots, \mathbf{D}_M\}$; then we update $\{\mathbf{D}_1, \ldots, \mathbf{D}_M\}$ with fixed $\mathbf{A}$. The author of ACDC [12] called the first phase the "alternating columns" phase and the second phase the "diagonal centers" phase. We will discuss two phases separately.

Alternating Columns

The first phase is called "alternating columns" (AC). In this phase, the diagonal matrices $\{\mathbf{D}_1, \ldots \mathbf{D}_M\}$ are all fixed. The mixing matrix $\mathbf{A}$ is updated in a column-by-column fashion.

The objective function of (3.37) can be rewritten as

$$C_{LS}(\mathbf{A}, \mathbf{D}_1, \ldots, \mathbf{D}_M) = \sum_{m=1}^{M} \|\mathbf{R}_m - \mathbf{A}\mathbf{D}_m\mathbf{A}\|_F^2 = \sum_{m=1}^{M} \left\| \mathbf{R}_m - \sum_{k=1}^{K} d_{mk}\mathbf{a}_k\mathbf{a}_k^H \right\|_F^2, \tag{3.38}$$

where $\mathbf{a}_k$ is the kth column of $\mathbf{A}$ and d_{mk} is the (k,k)th element of $\mathbf{D}_m$. Define $\grave{\mathbf{R}}_m$ by

$$\grave{\mathbf{R}}_m = \mathbf{R}_m - \sum_{k \neq l} d_{mk}\mathbf{a}_k\mathbf{a}_k^H. \tag{3.39}$$

Here, we want to update the lth column of $\mathbf{A}$ (i.e. $\mathbf{a}_l$). Consider C_{LS} for fixed $\mathbf{A}$, $\{\mathbf{D}_1, \dots \mathbf{D}_M\}$ except $\mathbf{a}_l$:

$$\begin{aligned} C_{LS}(\mathbf{a}_l) =& \sum_{m=1}^{M} \left\| \mathbf{R}_m - \sum_{k=1}^{K} d_{mk}\mathbf{a}_k\mathbf{a}_k^H \right\|_F^2 \\ =& \sum_{m=1}^{M} \left\| \grave{\mathbf{R}}_m - d_{ml}\mathbf{a}_l\mathbf{a}_l^H \right\|_F^2 \\ =& \sum_{m=1}^{M} \mathrm{Tr}\left([\grave{\mathbf{R}}_m - d_{ml}\mathbf{a}_l\mathbf{a}_l^H]^H [\grave{\mathbf{R}}_m - d_{ml}\mathbf{a}_l\mathbf{a}_l^H] \right) \\ =& -\mathbf{a}_l^H \left[\sum_{m=1}^{M} d_{ml} \left[\grave{\mathbf{R}}_m^H + \grave{\mathbf{R}}_m \right] \right] \mathbf{a}_l + (\mathbf{a}_l^H\mathbf{a}_l)^2 \sum_{m=1}^{M} (d_{ml})^2 + \text{Constant}. \end{aligned} \tag{3.40}$$

Decomposing $\mathbf{a}_l$ into

$$\mathbf{a}_l = c\boldsymbol{\alpha}, \text{where } \boldsymbol{\alpha}^H\boldsymbol{\alpha} = 1, \tag{3.41}$$

the objective function (3.40) reduces to

$$C_{LS}(c, \boldsymbol{\alpha}) = -2c^2\boldsymbol{\alpha}^H\mathbf{P}\boldsymbol{\alpha} + c^4 p + \text{Constant}, \tag{3.42}$$

where $\mathbf{P} = \sum_{m=1}^{M} d_{ml}\grave{\mathbf{R}}_m$ and $p = \sum_{m=1}^{M}(d_{ml})^2$. Differentiating Eq. (3.42) w.r.t. c and equating the result to zero yields $c = 0$ or

$$c^2 = \frac{1}{p}\boldsymbol{\alpha}^H\mathbf{P}\boldsymbol{\alpha}. \tag{3.43}$$

As $\mathbf{P}$ is positive semi-definite, $c = \sqrt{\frac{1}{p}\boldsymbol{\alpha}^H\mathbf{P}\boldsymbol{\alpha}}$. By substituting the result back to $C_{LS}(c, \boldsymbol{\alpha})$, the partial minimization problem w.r.t. $\boldsymbol{\alpha}$ is:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & -\frac{1}{p}(\boldsymbol{\alpha}^H\mathbf{P}\boldsymbol{\alpha})^2 \\ \text{s.t. } & \boldsymbol{\alpha}^H\boldsymbol{\alpha} = 1, \end{aligned} \tag{3.44}$$

in which the solution is given by a principle eigenvector of $\mathbf{P}$. The AC phase stops when all K columns of $\mathbf{A}$ are updated once.

Diagonal Centers

The second phase is called "diagonal centers" (DC). In this phase, the mixing matrix $\mathbf{A}$ is fixed and one diagonal matrix $\mathbf{D}_m$ is updated at one time. The objective function C_{LS} for fixed $\mathbf{A}$ and $\{\mathbf{D}_1, \ldots, \mathbf{D}_{m-1}, \mathbf{D}_{m+1}, \ldots, \mathbf{D}_M\}$ is

$$C_{LS}(\mathbf{D}_m) = \|\mathbf{R}_m - \mathbf{A}\mathbf{D}_m\mathbf{A}^H\|_F^2. \tag{3.45}$$

Define $\mathbf{y}_m = \text{vec}(\mathbf{R}_m)$ and $\mathbf{d}_m = \text{diag}(\mathbf{D}_m)$, where $\text{vec}(\cdot)$ is the vectorization operator and $\text{diag}(\cdot)$ is the diagonalization operator. Then, (3.45) becomes

$$C_{LS}(\mathbf{d}_m) = [\mathbf{y}_m - (\mathbf{A}^* \odot \mathbf{A})\mathbf{d}_m]^H[\mathbf{y}_m - (\mathbf{A}^* \odot \mathbf{A})\mathbf{d}_m], \tag{3.46}$$

where $\odot$ denotes the Khatri-Rao product (also known as column-wise Kronecker product). To minimize $\mathbf{C}_{LS}$ w.r.t. $\mathbf{d}_m$, the solution is given by

$$\mathbf{d}_m = \left((\mathbf{A}^* \odot \mathbf{A})^H(\mathbf{A}^* \odot \mathbf{A})\right)^{-1}(\mathbf{A}^* \odot \mathbf{A})^H\mathbf{y}_m. \tag{3.47}$$

The DC phase stops when $\{\mathbf{D}_1, \ldots, \mathbf{D}_M\}$ are all updated once. ACDC alternates the AC phase and DC phase, until a stopping criterion is met.

3.2.3 Trilinear Alternating Least-Squares [10, 11]

Similar to ACDC, Trilinear Alternating Least-Squares (TALS) algorithm handles Problem (3.24) using alternating optimization. One of the distinguishing features of TALS is that Problem (3.24) is rewritten as three data fitting problems employing matrix unfolding of tensors. By properly arranging those three problems, Problem (3.24) can be handled efficiently.

Now, define the tensor $\mathcal{R} \in \mathbb{C}^{N\times N\times M}$ by[1]

$$\mathcal{R}_{:,:,m} = \mathbf{R}_m = \mathbf{A}\mathbf{D}_m\mathbf{A}^H, m = 1, \ldots, M. \tag{3.48}$$

Here, we are interested in matrix unfoldings of $\mathcal{R}$. There are many ways in defining matrix unfoldings and we adopt the convention used in [11][2].

[1] We use the notations used in MATLAB for convenience.

[2] There are some typos in this paper. I have verified the expressions used here and confirmed they are consistent with the MATLAB code provided.

Define the matrices $\mathbf{R}_a \in \mathbb{R}^{MN\times N}$, $\mathbf{R}_b \in \mathbb{R}^{MN\times N}$ and $\mathbf{R}_c \in \mathbb{R}^{N^2\times M}$ by

$$\mathbf{R}_a = \begin{bmatrix} \mathcal{R}_{:,:,1} \\ \mathcal{R}_{:,:,2} \\ \vdots \\ \mathcal{R}_{:,:,M} \end{bmatrix}, \ \mathbf{R}_b = \begin{bmatrix} \mathcal{R}^T_{:,1,:} \\ \mathcal{R}^T_{:,2,:} \\ \vdots \\ \mathcal{R}^T_{:,N,:} \end{bmatrix}, \mathbf{R}_c = \begin{bmatrix} \mathcal{R}_{1,:,:} \\ \mathcal{R}_{2,:,:} \\ \vdots \\ \mathcal{R}_{N,:,:} \end{bmatrix}. \tag{3.49}$$

It can be verified that the matrix unfoldings have the following forms:

$$\mathbf{R}_a = (\mathbf{\Psi} \odot \mathbf{A})\mathbf{B},\ \mathbf{R}_b = (\mathbf{B}^T \odot \mathbf{\Psi})\mathbf{A}^T,\ \mathbf{R}_c = (\mathbf{A} \odot \mathbf{B}^T)\mathbf{\Psi}^T, \tag{3.50}$$

where $\mathbf{\Psi}$ is defined in (3.31) and[3] $\mathbf{B} = \mathbf{A}^H$.

Then, we consider the following three data fitting problems:

$$\min \|\mathbf{R}_a - (\mathbf{\Psi} \odot \mathbf{A})\mathbf{B}\|_F^2, \tag{3.51}$$

$$\min \|\mathbf{R}_b - (\mathbf{B}^T \odot \mathbf{\Psi})\mathbf{A}^T\|_F^2, \tag{3.52}$$

$$\min \|\mathbf{R}_c - (\mathbf{A} \odot \mathbf{B}^T)\mathbf{\Psi}^T\|_F^2. \tag{3.53}$$

Note that the variables of the above minimization problems are not stated. The variables $\mathbf{A}, \mathbf{B}, \mathbf{\Psi}$ must be updated once in these minimization problems, in order to find the best fit. An important observation is that

$$\min_{\mathbf{B}} \|\mathbf{R}_a - (\mathbf{\Phi} \odot \mathbf{A})\mathbf{B}\|_F^2, \tag{3.54}$$

$$\min_{\mathbf{A}} \|\mathbf{R}_b - (\mathbf{B}^T \odot \mathbf{\Psi})\mathbf{A}^T\|_F^2, \tag{3.55}$$

$$\min_{\mathbf{\Psi}} \|\mathbf{R}_c - (\mathbf{A} \odot \mathbf{B}^T)\mathbf{\Psi}^T\|_F^2, \tag{3.56}$$

are linear least-squares fitting problems, and the solutions lead us to the following simple closed form updates:

$$\mathbf{B}^{(k+1)} := (\mathbf{\Phi}^{(k)} \odot \mathbf{A}^{(k)})^\dagger \mathbf{R}_a, \tag{3.57}$$

$$\mathbf{A}^{(k+1)} := \left(\left((\mathbf{B}^T)^{(k+1)} \odot \mathbf{\Phi}^{(k)} \right)^\dagger \mathbf{R}_b \right)^T, \tag{3.58}$$

$$\mathbf{\Phi}^{(k+1)} := \left(\left(\mathbf{A}^{(k+1)} \odot (\mathbf{B}^T)^{(k+1)} \right)^\dagger \mathbf{R}_c \right)^T. \tag{3.59}$$

Upon convergence, the mixing matrix $\mathbf{A}$ is given by

$$\hat{\mathbf{A}} = \frac{\mathbf{A} + \mathbf{B}^H}{2}. \tag{3.60}$$

[3] We have this transformation because TALS was actually designed to handle the unsymmetric cases, i.e. $\mathbf{R}_m = \mathbf{A}\mathbf{D}_m\mathbf{B}$.

3.3 Summary

In this chapter, we have briefly introduced four BI-QSS algorithms including FFDIAG, Pham's JD, TALS and ACDC. Extra assumptions made by those algorithms are stated explicitly.

For FFDIAG, the introduction of the demixing matrix $\mathbf{W}$ simplifies the problem structure. By imposing some implicit invertibility constraint on $\mathbf{W}$, trivial solution can be avoided. The problem structure is further simplified by making one extra assumption (cf. (A5 for FFDIAG)). For Pham's JD, the maximum likelihood (ML) estimation is employed in handling BI-QSS. It gives an alternative interpretation of joint diagonalization.

PARAFAC is highly related to tensor decomposition. In this chapter, we have first provided a short discussion on tensor decomposition, with emphasis put on its uniqueness result. It reveals the powerful identifiability of PARAFAC-based algorithms. Then, the matrix unfolding of tensor was introduced. Afterwards, we discussed ACDC which handles BI-QSS using alternating optimization. TALS, making use of the matrix unfoldings of tensor, rewrites the BI-QSS problem into three least-squares data fitting problems which can be handled efficiently.

An important observation is that a joint-source identification approach is used for both JD and PARAFAC; i.e., they aim at identifying the whole mixing matrix $\mathbf{A}$ simultaneously. In the next chapter, we will propose a blind identification criterion which inherently suggests per-source identification.

□ **End of chapter.**

Chapter 4

Proposed Algorithms

In this chapter, we will first devise a blind identification criterion based on the subspace characteristic of the observations. The proposed criterion is structurally different from that of PARAFAC and JD. Specifically, it inherently suggests a per-source identification procedure for identifying the mixing matrix $\mathbf{A}$ [1, 2]. In addition, the proposed criterion provides strong identifiability.

Then, we will devise BI-QSS algorithms utilizing the proposed criterion. A simple technique called Alternating Projections (AP) [3] will be used to handle the problem. The resulting BI-QSS algorithm contains only simple closed form updates. An all-column identification procedure will also be introduced to complement the per-source identification procedure suggested by the proposed criterion.

Afterwards, we will focus on overdetermined mixing models, in which prewhitening is possible. With the employment of prewhitening, the convergence behavior of AP is dramatically improved [1]. We will show that the proposed algorithm converges to the true mixing matrix columns in one iteration with probability one. As a side benefit, the all-column identification procedure can be significantly simplified in overdetermined mixing models.

We will then consider the more challenging underdetermined mixing models. As prewhitening is no longer possible, AP may be inefficient. We will propose rank minimization heuristics to speed up the algorithm.

Recently, the rank minimization heuristic has received great research interest in many areas like image processing and video processing [40–42,44]. Furthermore, in order to improve the runtime performance, ideas from the augmented Lagrangian method [41,48,52] is used. Interestingly, the resulting algorithm possesses a smooth optimization interpretation [43]. Specifically, we will show the connection between the proposed algorithm and the Huber loss function widely used in robust statistics [55].

At the end of this chapter, we consider a practical situation in which corrupted local covariance matrices exist. To the best of my knowledge, there are few works that explicitly consider blind identification with robustness against corrupted data. In this thesis, we will consider robust subspace extraction, in which the corrupted local covariance matrices are detected and dropped in the subspace extraction procedure.

4.1 KR Subspace Criterion

Let us first devise the proposed blind identification criterion. Recall the noise-free local covariance model (cf. Section 2.3):

$$\mathbf{R}_m = \mathbf{A}\mathbf{D}_m\mathbf{A}^H, \ m = 1, \ldots, M. \tag{4.1}$$

Define $\mathbf{y}_m \in \mathbb{C}^{N^2}$ by:

$$\begin{aligned}\mathbf{y}_m \triangleq \text{vec}(\mathbf{R}_m) &= \text{vec}(\sum_{k=1}^{K} d_{mk}\mathbf{a}_k\mathbf{a}_k^H) \\ &= \sum_{k=1}^{K} d_{mk}(\mathbf{a}_k^* \otimes \mathbf{a}_k) \qquad (4.2) \\ &= (\mathbf{A}^* \odot \mathbf{A})\mathbf{d}_m \qquad (4.3)\end{aligned}$$

where vec(·) is the vectorization operator, $\odot$ is the Khatri-Rao (KR) product and $\mathbf{d}_m = [d_{m1}, \ldots, d_{mK}]^T$. The equality in (4.2) is due to a property of Kronecker product[1]. Furthermore, stacking all $\mathbf{y}_m$ yields

$$\mathbf{Y} \triangleq [\mathbf{y}_1, \ldots, \mathbf{y}_M] = (\mathbf{A}^* \odot \mathbf{A})\boldsymbol{\Psi}^T \in \mathbb{C}^{N^2 \times M}, \tag{4.4}$$

[1] We have $\mathbf{a} \otimes \mathbf{b} = \text{vec}(\mathbf{b}\mathbf{a}^T)$.

where $\mathbf{\Psi}^T = [\mathbf{d}_1, \ldots, \mathbf{d}_M]$ (cf. Eq. (3.31)). We are interested in the subspace characteristics of $\mathbf{Y}$. Here, we make the following two additional assumptions:

(A4) The mixing matrix $\mathbf{A} \in \mathbb{C}^{N\times K}$ has full Kruskal rank;

(A5) $\mathbf{\Psi}$ has full column rank for $M > K$;

where the justifications are already given in Section 3.2.1.

According to assumption (A4) , it can be shown that

$$\mathbf{A}^* \odot \mathbf{A} \in \mathbb{C}^{N^2\times K}$$

has full column rank for $K \leq 2N-1$. Therefore, it is easy to verify that

$$\mathcal{R}(\mathbf{Y}) = \mathcal{R}(\mathbf{A}^* \odot \mathbf{A}), \tag{4.5}$$

where $\mathcal{R}(\mathbf{X})$ is the range space of $\mathbf{X}$. An important implication of (4.5) is that

$$\mathbf{a}_k^* \otimes \mathbf{a}_k \in \mathcal{R}(\mathbf{Y}),\ k = 1, \ldots, K,$$

where $\mathcal{R}(\mathbf{Y})$ can be extracted from the observations. Specifically, according to the full column rank condition of $\mathbf{A}^* \odot \mathbf{A}$ and $\mathbf{\Psi}$, $\mathbf{Y}$ admits the following compact singular value decomposition (SVD):

$$\mathbf{Y} = \mathbf{U}_s \mathbf{\Sigma}_s \mathbf{V}_s^H, \tag{4.6}$$

where $\mathbf{\Sigma}_s \in \mathbb{R}^{K\times K}$ is the non-zero singular value matrix and $\mathbf{U}_s \in \mathbb{C}^{N^2\times K}$ and $\mathbf{V}_s \in \mathbb{C}^{M\times K}$ are the associated left and right singular matrices, respectively. We also have

$$\mathcal{R}(\mathbf{U}_s) = \mathbf{R}(\mathbf{Y}). \tag{4.7}$$

Based on Eq. (4.5) and (4.7), we know that the mixing matrix columns satisfy

$$\mathbf{a}_k^* \otimes \mathbf{a}_k \in \mathcal{R}(\mathbf{U}_s). \tag{4.8}$$

Therefore, the mixing matrix columns can be identified by the following blind identification criterion:

$$\begin{aligned} &\text{find } \mathbf{a} \\ &\text{s.t. } \mathbf{a}^* \otimes \mathbf{a} \in \mathcal{R}(\mathbf{U}_s). \end{aligned} \tag{4.9}$$

As the above identification criterion involved the subspace of the self KR product of $\mathbf{A}$, we will call it the *KR Subspace Criterion* in the sequel. One distinguishing feature of the KR subspace criterion is that it suggests a per-source identification procedure in identifying $\mathbf{A}$.

Before seeking a way to find a solution of the KR subspace criterion, let us investigate its identifiability; i.e. under what conditions the mixing matrix columns can be uniquely identified using the KR subspace criterion, up to a scaling factor. For Vandemonde mixing matrix $\mathbf{A}$, the identifiability of the KR subspace criterion has been investigated in [2]. In this thesis, we generalize the result to consider $\mathbf{A}$ with full Kruskal rank.

Theorem 4.1 *Under* (2.14)*, (A4), (A5). Then, $K \leq 2N-2$ is a necessary and sufficient condition for*

$$\mathbf{a} = c\mathbf{a}_k \iff \mathbf{a}^* \otimes \mathbf{a} \in \mathcal{R}(\mathbf{U}_s), \tag{4.10}$$

for any $k = 1, \ldots, K$ and for any non-zero constant c.

The proof of Theorem 4.1 is given in Appendix B.1. Theorem 4.1 implies that the mixing matrix columns can be unambiguously identified using the KR subspace criterion up to a scaling factor. In addition, it reveals that the KR subspace criterion can deal with underdetermined mixing models (i.e. $K > N$).

4.2 Blind Identification using Alternating Projections

In this section, we will focus on per-source blind identification suggested by the KR subspace criterion. Motivated by the KR subspace criterion, we consider the following problem:

$$\begin{aligned} &\min_{\alpha,\mathbf{a},\mathbf{h}} \|\alpha\mathbf{a}^* \otimes \mathbf{a} - \mathbf{h}\|^2 \\ &\text{ s.t. } |\alpha| = 1, \|\mathbf{a}\|^2 = 1, \mathbf{h} \in \mathcal{R}(\mathbf{U}_s), \end{aligned} \tag{4.11}$$

where $\|\cdot\|$ is the Euclidean norm. In essence, we want to minimize the difference between a vector having a self Kronecker product structure and a vector in the KR subspace $\mathcal{R}(\mathbf{U}_s)$.

We employ alternating projections (AP) to handle Problem (4.11). AP is a simple technique for finding an intersection point of some given sets [3]. For Problem (4.11), AP splits the problem into two partial minimization problems, one minimizing (4.11) w.r.t. $\mathbf{h}$ only while the other minimizing (4.11) w.r.t. $(\alpha, \mathbf{a})$ only. The partial minimization problem of Problem (4.11) w.r.t. $\mathbf{h}$ with $(\alpha, \mathbf{a})$ fixed is:

$$\begin{aligned} &\min_{\mathbf{h}} \ \|\alpha \mathbf{a}^* \otimes \mathbf{a} - \mathbf{h}\|^2 \\ &\text{s.t. } \mathbf{h} \in \mathcal{R}(\mathbf{U}_s), \end{aligned} \tag{4.12}$$

which is a linear projection problem. The solution of Problem (4.12) is given by

$$\mathbf{h} = \mathbf{U}_s \mathbf{U}_s^H (\alpha \mathbf{a}^* \otimes \mathbf{a}). \tag{4.13}$$

In addition, the partial minimization problem of Problem (4.11) w.r.t. $(\alpha, \mathbf{a})$ with $\mathbf{h}$ fixed is:

$$\begin{aligned} &\min_{\alpha, \mathbf{a}} \ \|\alpha \mathbf{a}^* \otimes \mathbf{a} - \mathbf{h}\|^2 \\ &\text{s.t. } |\alpha| = 1, \|\mathbf{a}\|^2 = 1. \end{aligned} \tag{4.14}$$

Problem (4.14) also admits closed form solutions. To see this, consider the objective function of Problem (4.14):

$$\begin{aligned} \|\alpha \mathbf{a}^* \otimes \mathbf{a} - \mathbf{h}\|^2 &= |\alpha|^2 \|\mathbf{a}^* \otimes \mathbf{a}\|^2 - 2\text{Re}\{\alpha^* (\mathbf{a}^* \otimes \mathbf{a})^H \mathbf{h}\} + \|\mathbf{h}\|^2 \\ &= 1 - 2\text{Re}\{\alpha^* \mathbf{a}^H \text{vec}^{-1}(\mathbf{h}) \mathbf{a}\} + \|\mathbf{h}\|^2 && (4.15) \\ &= 1 - 2\text{Re}\{\alpha^* \mathbf{a}^H \tilde{\mathbf{H}} \mathbf{a}\} + \|\mathbf{h}\|^2 && (4.16) \\ &\geq 1 - 2|\mathbf{a}^H \tilde{\mathbf{H}} \mathbf{a}| + \|\mathbf{h}\|^2, && (4.17) \end{aligned}$$

where $\tilde{\mathbf{H}} \triangleq \frac{1}{2}\left(\text{vec}^{-1}(\mathbf{h}) + (\text{vec}^{-1}(\mathbf{h}))^H\right)$ and $\text{vec}^{-1}(\cdot)$ is the devectorization operator. Again, a property of Kronecker product[2] has been used in obtaining (4.15). To minimize (4.17) w.r.t. $\mathbf{a}$, the solution is given by

$$\mathbf{a} = \mathbf{q}_{\max}(\tilde{\mathbf{H}}), \tag{4.18}$$

[2] We have $(\mathbf{b}^* \otimes \mathbf{a})^H \text{vec}(\mathbf{C}) = \mathbf{a}^H \mathbf{C} \mathbf{b}$

where $\mathbf{q}_{\max}(\mathbf{X})$ denotes a unit-norm eigenvector of $\mathbf{X}$ associated with an eigenvalue with the largest absolute value $\lambda_{\max}(\mathbf{X})$. Moreover, it is easy to verify that the equality in (4.17) holds when

$$\alpha = \lambda_{\max}(\tilde{\mathbf{H}})/|\lambda_{\max}(\tilde{\mathbf{H}})|. \tag{4.19}$$

The alternating projections algorithm iteratively updates $(\alpha, \mathbf{a}, \mathbf{h})$ until some stopping criterion is satisfied. We end this section by providing the AP algorithm in **Algorithm 1**.

Algorithm 1 Alternating Projections Algorithm for Problem (4.11).

```
Input: $\mathbf{U}_s$: KR subspace; $\mathbf{h} \in \mathcal{R}(\mathbf{U}_s)$: a randomly generated initial point
Output: $\mathbf{a}$: a mixing matrix column;
repeat
    $\tilde{\mathbf{H}} = \frac{1}{2}\left(\mathrm{vec}^{-1}(\mathbf{h}) + (\mathrm{vec}^{-1}(\mathbf{h}))^H\right)$;
    $\mathbf{a} = \mathbf{q}_{\max}(\tilde{\mathbf{H}})$;
    $\alpha = \lambda_{\max}(\tilde{\mathbf{H}})/|\lambda_{\max}(\tilde{\mathbf{H}})|$;
    $\mathbf{h} = \mathbf{U}_s\mathbf{U}_s^H(\alpha\mathbf{a}^* \otimes \mathbf{a})$;
until a stopping criterion is satisfied.
```

4.2.1 All-Columns Identification

As the KR subspace criterion suggests a per-column identification approach, a way to identify all mixing matrix columns must be found. Here, we propose a simple all-column identification procedure: we randomly generate a large number of initializations $\mathbf{h} \in \mathcal{R}(\mathbf{U}_s)$. Every time a solution is obtained from AP, we check the cross-correlation with all previously identified columns. We accept a newly identified column if the cross-correlations with the previously identified columns are less than certain predefined constant $0 < \delta < 1$. The exact algorithm is listed in **Algorithm 2**.

Algorithm 2 All-column Identification using Alternating Projections.

```
Input: $\mathbf{U}_s$: KR subspace; $\delta$: a small constant;
Output: $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$: mixing matrix;
initialize $i = 1$; randomly generate an initial point $\mathbf{h} \in \mathcal{R}(\mathbf{U}_s)$;
use the AP algorithm in Algorithm 1 with KR subspace $\mathbf{U}_s$ and initialization $\mathbf{h}$ to obtain $\mathbf{a}$;
if $i = 1$ then
    goto step 11;
end if
while $i < K$ do
    repeat
        randomly generate an initial point $\mathbf{h} \in \mathcal{R}(\mathbf{U}_s)$;
        use the AP algorithm in Algorithm 1 with KR subspace $\mathbf{U}_s$ and initialization $\mathbf{h}$ to obtain $\mathbf{a}$;
    until $\frac{|\mathbf{a}_j^H \mathbf{a}|}{\|\mathbf{a}_j\|\|\mathbf{a}\|} < \delta, \ \forall\, j = 1, \ldots, i-1$.
    $\mathbf{a}_i = \mathbf{a}$; $i = i + 1$;
end while
```

4.3 Overdetermined Mixing Models ($N > K$): Prewhitened Alternating Projection Algorithm (PAPA)

Although we have shown that the operation on AP consists of only simple closed form updates, it has been noticed that AP may require a large number of iterations to converge in general [3]. More importantly, the all-column identification procedure in **Algorithm 2** is inefficient. Rather unexpectedly, both the convergence behavior of AP and the way to identify all columns become dramatically different when the local covariance matrices are prewhitened before the BI-QSS (cf. Section 2.5). In this section, we will focus on the overdetermined mixing models in which prewhitening is possible.

To facilitate our discussion, let us first introduce the following Theorem:

Theorem 4.2 *Suppose that* $\mathbf{A}$ *is unitary and the model error is absent* *(i.e. the local covariance model in Eq.* (2.14) *holds ideally). If the initialization of the AP algorithm is randomly generated by* $\mathbf{h} = \mathbf{U}_s \boldsymbol{\xi}$ *where* $\boldsymbol{\xi} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$*, then, the AP algorithm converges to any one of the true mixing matrix columns up to a scaling factor in one iteration with probability one.*

The proof of Theorem 4.2 is given in Appendix B.2. Recall that the equivalent mixing matrix $\tilde{\mathbf{A}}$ after prewhitening the local covariance matrices $\{\mathbf{R}_1, \ldots, \mathbf{R}_M\}$ is unitary (cf. Section 2.5). Hence, for the overdetermined mixing models, we can first employ prewhitening, followed by applying AP. In Fig. 4.1, the projection error

$$\|(\mathbf{I} - \mathbf{U}_s \mathbf{U}_s^H)\mathbf{a}^* \otimes \mathbf{a}\|$$

against iteration is shown to demonstrate this desirable behavior. The simulation is based on a perfect data scenario; i.e. the local covariance matrices

$$\mathbf{R}_m = \mathbf{A}\mathbf{D}_m\mathbf{A}^H, \; m = 1, \ldots, M$$

are synthetically generated, where $\mathbf{A} \in \mathbb{R}^{6\times 5}$. We can see that the number of iterations required by AP can be significantly reduced with the employment of prewhitening.

Even better, with the employment of prewhitening, we can systematically identify all columns of the mixing matrix by exploiting the column orthogonality of $\tilde{\mathbf{A}}$ (cf. Section 2.5). Suppose that we have already identified $\tilde{\mathbf{a}}_r$. It can be shown that

$$\mathcal{R}(\mathbf{P}^{\perp}_{\tilde{\mathbf{a}}_r^* \otimes \tilde{\mathbf{a}}_r} \mathbf{U}_s) = \mathcal{R}(\mathbf{P}^{\perp}_{\tilde{\mathbf{a}}_r^* \otimes \tilde{\mathbf{a}}_r} \tilde{\mathbf{A}}^* \odot \tilde{\mathbf{A}}) = \mathcal{R}(\tilde{\mathbf{A}}_{-r} \odot \tilde{\mathbf{A}}_{-r}), \tag{4.20}$$

where $\tilde{\mathbf{A}}_{-r} \triangleq [\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_{r-1}, \tilde{\mathbf{a}}_{r+1}, \ldots, \tilde{\mathbf{a}}_K]$ and $\mathbf{P}^{\perp}_{\tilde{\mathbf{a}}_r^* \otimes \tilde{\mathbf{a}}_r}$ is the orthogonal complement projector of $\tilde{\mathbf{a}}_r^* \otimes \tilde{\mathbf{a}}_r$. Therefore, by extracting the basis matrix $\mathbf{Q}_s \in \mathbb{C}^{N^2 \times (K-1)}$ from $\mathbf{P}^{\perp}_{\tilde{\mathbf{a}}_r^* \otimes \tilde{\mathbf{a}}_r} \mathbf{U}_s$ using SVD and updating the KR subspace

$$\mathbf{U}_s := \mathbf{Q}_s,$$

the identified column $\tilde{\mathbf{a}}_r$ will be completely eliminated from the KR subspace. For convenience, we call the proposed algorithm the *Prewhitened*

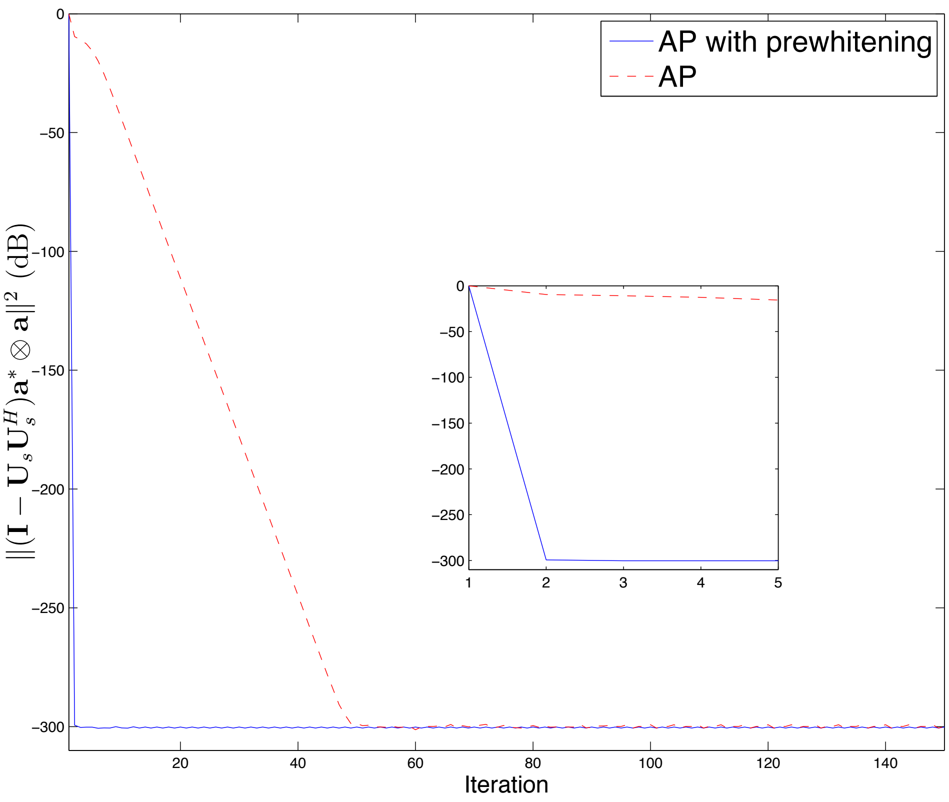

Figure 4.1: Projection error against iteration

Alternating Projections Algorithm (PAPA) where the complete pseudocode is provided in **Algorithm 3**.

4.4 Underdetermined Mixing Models ($N < K$)

According to Theorem 4.1, BI-QSS algorithms devised from the KR subspace criterion (4.9) will also be applicable for $K \leq 2N - 2$. However, prewhitening is no longer possible when $K > N$. Therefore, AP may exhibit slow convergence. In this section, we propose rank minimization heuristics to speed up the AP convergence.

4.4.1 Rank Minimization Heuristic

Without prewhitening, the alternating projections algorithm in **Algorithm 1** may be slow in terms of number of iterations required [3]. Yet, there is one interesting observation on the AP algorithm:

Algorithm 3 Prewhitened Alternating Projection Algorithms.

```
Input: local covariance matices $\{\mathbf{R}_1, \ldots, \mathbf{R}_M\}$;
Output: $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K] = \mathbf{B}\tilde{\mathbf{A}}$: mixing matrix;
1: perform the noise covariance removal procedure describe in Section 2.4;
2: compute $\overline{\mathbf{R}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{R}_m$ and perform a square-root factorization on $\overline{\mathbf{R}}$;
   i.e., $\overline{\mathbf{R}} = \mathbf{B}\mathbf{B}^H$;
3: compute $\tilde{\mathbf{R}}_m = \mathbf{B}^\dagger \mathbf{R}_m (\mathbf{B}^\dagger)^H, \ m = 1, \ldots, M$;
4: compute the compact SVD of $\mathbf{Y} = [\mathrm{vec}(\tilde{\mathbf{R}}_1), \ldots, \mathrm{vec}(\tilde{\mathbf{R}}_M)]$; i.e., $\mathbf{Y} =$
   $\mathbf{U}_s \Sigma_s \mathbf{V}_s^H$, and set $i = 1$;
5: use the AP algorithm in Algorithm 1 with KR subspace $\mathbf{U}_s$ and initialization $\mathbf{h} = \mathbf{U}_s \boldsymbol{\xi}$ for $\boldsymbol{\xi} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ to obtain $\tilde{\mathbf{a}}_i$;
6: compute $\mathbf{P}^{\perp}_{\tilde{\mathbf{a}}_i^* \otimes \tilde{\mathbf{a}}_i} = \mathbf{I} - (\tilde{\mathbf{a}}_i^* \otimes \tilde{\mathbf{a}}_i)(\tilde{\mathbf{a}}_i^* \otimes \tilde{\mathbf{a}}_i)^H$ and obtain the basis matrix
   $\mathbf{Q}_s \in \mathbb{C}^{K^2 \times (K-i)}$of $\mathbf{P}^{\perp}_{\tilde{\mathbf{a}}_i^* \otimes \tilde{\mathbf{a}}_i} \mathbf{U}_s$ using SVD;
7: update the KR subspace $\mathbf{U}_s := \mathbf{Q}_s$, set $i := i + 1$ and goto step 5 until
   $i > K$;
8: $\tilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_K]$.
```

Observation 4.1 *If the matrix* $\tilde{\mathbf{H}}$ *in* ***Algorithm 1*** *is of rank-1, then* $\mathbf{a}$ *satisfies the KR subspace criterion in its next iteration.*

The proof of Observation 4.1 is given in Appendix B.3. The desirable property of $\tilde{\mathbf{H}}$ can be incorporated by introducing a regularization term. In essence, we consider the following rank regularized minimization problem (the objective function of Problem (4.11) is devectorized for notational convenience):

$$\begin{aligned} \min_{\alpha, \mathbf{a}, \mathbf{H}} \ & \|\alpha \mathbf{a}\mathbf{a}^H - \mathbf{H}\|_F^2 + \gamma \mathrm{rank}(\mathbf{H}) \\ \text{s.t.} \ & |\alpha| = 1, \|\mathbf{a}\|^2 = 1, \mathrm{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s), \end{aligned} \tag{4.21}$$

where $\mathbf{H} \triangleq \mathrm{vec}^{-1}(\mathbf{h})$ and γ is a properly chosen regularization constant. In Problem (4.21), we put incentive to lower the rank alongside with finding a vector $\mathbf{a}^* \otimes \mathbf{a}$ in the KR subspace $\mathcal{R}(\mathbf{U}_s)$.

However, Problem (4.21) is challenging to be solved in general, owing to the combinatorial nature of the rank function [40]. All is not lost, nonetheless, as it has been known that nuclear norm is a simple yet efficient heuristic in replacing the rank function [45–47]. Nuclear norm is defined as

$$\|\mathbf{X}\|_* = \sigma_1 + \ldots + \sigma_r, \tag{4.22}$$

where σ_i is the ith non-zero singular value of $\mathbf{X}$ and r is the rank of $\mathbf{X}$. Therefore, instead of attempting to solve Problem (4.21), we consider the following nuclear norm regularization problem:

$$\begin{aligned} \min_{\alpha,\mathbf{a},\mathbf{H}} \ & \|\alpha\mathbf{a}\mathbf{a}^H - \mathbf{H}\|_F^2 + \gamma\|\mathbf{H}\|_* \\ \text{s.t. } & |\alpha| = 1, \|\mathbf{a}\|^2 = 1, \text{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s). \end{aligned} \tag{4.23}$$

Interestingly, the above formulation is similar to ℓ_1-ℓ_2 optimization used in basis pursuit (BP) [43], a very powerful technique used in compressive sensing [39, 47]. In BP, we consider the following problem:

$$\min_{\mathbf{z}} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|^2 + \gamma\|\mathbf{z}\|_1, \tag{4.24}$$

where $\mathbf{y}$ is the observation, $\mathbf{A}$ is the dictionary, $\mathbf{z}$ is the variable known to be sparse and γ is a properly chosen regularization constant. In BP, the ℓ_2-norm term is for the solution accuracy and the ℓ_1-norm term is to promote the solution sparsity. The nuclear norm can be considered as a matrix extension of the ℓ_1-norm in promoting low-rankness of the solution.

Again, we use AP to handle Problem (4.23). For fixed $\mathbf{h}$, the partial minimization problem of Problem (4.23) w.r.t. $(\alpha, \mathbf{a})$ is essentially the same as Problem (4.14). Therefore, the closed form solutions of α and $\mathbf{a}$ in **Algorithm 1** can be applied. For fixed $(\alpha, \mathbf{a})$, the partial minimization problem of Problem (4.23) w.r.t. $\mathbf{h}$ is:

$$\begin{aligned} \min_{\mathbf{H}} \ & \|\alpha\mathbf{a}\mathbf{a}^H - \mathbf{H}\|_F^2 + \gamma\|\mathbf{H}\|_* \\ \text{s.t. } & \text{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s), \end{aligned} \tag{4.25}$$

which is a convex optimization problem. Thus, convex optimization algorithms such as the Interior Point Method (IPM) [59] can be used to handle Problem (4.25). Here, we use `CVX` [62, 63] to handle Problem (4.25). The AP algorithm for Problem (4.23) is provided in **Algorithm 4**.

Algorithm 4 Alternating Projections Algorithm for Problem (4.23).

```
Input: $\mathbf{U}_s$: KR subspace; $\mathbf{h} \in \mathcal{R}(\mathbf{U}_s)$: a randomly generated initial point;
Output: $\mathbf{a}$: a mixing matrix column;
1: repeat
2:     $\tilde{\mathbf{H}} = \frac{1}{2}\left(\mathbf{H} + \mathbf{H}^H\right)$;
3:     $\mathbf{a} = \mathbf{q}_{\max}(\tilde{\mathbf{H}})$;
4:     $\alpha = \lambda_{\max}(\tilde{\mathbf{H}})/|\lambda_{\max}(\tilde{\mathbf{H}})|$;
5:     compute $\mathbf{H}$ by handling Problem (4.25) using CVX;
6: until a stopping criterion is satisfied.
```

4.4.2 Alternating Projections Algorithm with Huber Function Regularization

Solving Problem (4.25) using general purpose toolboxes (e.g. `CVX`) can be slow in general, in terms of runtime. For alternating projections, simple closed form solutions are highly desirable (cf. **Algorithm 1**). Therefore, we are motivated to find a computational efficient way to update $\mathbf{H}$.

In Problem (4.25), the subspace constraint $\text{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s)$ prevents one from getting closed form solution. In fact, in the absence of the subspace constraint, $\mathbf{H}$ can be updated in closed form using the singular value thresholding (SVT) [40, 44]. Specifically, the SVT is defined as:

$$\text{SVT}(\mathbf{X}, \mu) = \mathbf{U}(\mathbf{\Sigma} - \mu\mathbf{I})_+\mathbf{V}^H, \tag{4.26}$$

where $\mathbf{X} = \mathbf{U\Sigma V}^H$ and $(\cdot)_+$ is the element-wise thresholding operator defined as:

$$\left[(\mathbf{X})_+\right]_{ij} = \begin{cases} x_{ij} - \mu & \text{if } x_{ij} \geq \mu, \\ 0 & \text{if } x_{ij} < \mu. \end{cases} \tag{4.27}$$

Let us consider the following equivalent formulation of Problem (4.25):

$$\begin{aligned} \min_{\mathbf{H},\mathbf{G}} \ & \|\alpha\mathbf{a}\mathbf{a}^H - \mathbf{H}\|_F^2 + \gamma\|\mathbf{G}\|_* \\ \text{s.t.} \ & \text{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s), \mathbf{H} = \mathbf{G}, \end{aligned} \tag{4.28}$$

where $\mathbf{G}$ is a splitting variable. Although Problem (4.28) is equivalent to (4.25), the variables in the Frobenius norm and the nuclear norm are now split up in the objective function. Now, consider the following approximation:

$$\begin{aligned} \min_{\mathbf{H},\mathbf{G}} \ & \|\alpha\mathbf{a}\mathbf{a}^H - \mathbf{H}\|_F^2 + \rho\|\mathbf{H}-\mathbf{G}\|_F^2 + \gamma\|\mathbf{G}\|_* \\ \text{s.t. } & \text{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s), \end{aligned} \tag{4.29}$$

where an augmented term is added for the constraint $\mathbf{H} = \mathbf{G}$ in Problem (4.28) and ρ is a properly chosen regularization constant for the augmented term. With a properly chosen ρ, the solution of Problem (4.29) should be close to that of Problem (4.28). It is important to note that the subspace constraint and nuclear norm are now split to two different variables $\mathbf{H}$ and $\mathbf{G}$.

Again, we use alternating projections to handle Problem (4.29). As mentioned previously (cf. Section 4.4.1), the partial minimization problem of Problem (4.29) w.r.t α and $\mathbf{a}$ both have closed form solutions (cf. **Algorithm 1**). Furthermore, the partial minimization problem of Problem (4.29) w.r.t both $\mathbf{H}$ and $\mathbf{G}$ also admit closed form solutions. Firstly, let us consider the partial minimization problem of Problem (4.29) w.r.t. $\mathbf{H}$:

$$\begin{aligned} \min_{\mathbf{H}} \ & \|\alpha\mathbf{a}\mathbf{a}^H - \mathbf{H}\|_F^2 + \rho\|\mathbf{H}-\mathbf{G}\|_F^2 \\ \text{s.t. } & \text{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s), \end{aligned} \tag{4.30}$$

which is a linear projection problem. The solution of Problem (4.30) is given by

$$\mathbf{H} = \text{vec}^{-1}\left(\frac{1}{1+\rho}\mathbf{U}_s\mathbf{U}_s^H\left(\alpha\mathbf{a}^* \otimes \mathbf{a} + \rho\text{vec}(\mathbf{G})\right)\right). \tag{4.31}$$

Then, we consider the partial minimization problem of Problem (4.29) w.r.t. $\mathbf{G}$:

$$\min_{\mathbf{G}} \ \rho\|\mathbf{H}-\mathbf{G}\|_F^2 + \gamma\|\mathbf{G}\|_*, \tag{4.32}$$

whose solution is given by the SVT of $\mathbf{H}$ (for the sake of self-containedness, the derivation of SVT is provided in Appendix C):

$$\mathbf{G} = \text{SVT}\left(\mathbf{H}, \frac{\gamma}{2\rho}\right). \tag{4.33}$$

Therefore, as both $\alpha, \mathbf{a}, \mathbf{H}$ and $\mathbf{G}$ admit closed form updates, the algorithm is expected to be efficient. More importantly, the above formulation possesses a smooth optimization interpretation. Minimizing a nonsmooth function (e.g. ℓ_1 norm) is computationally undesirable, in the

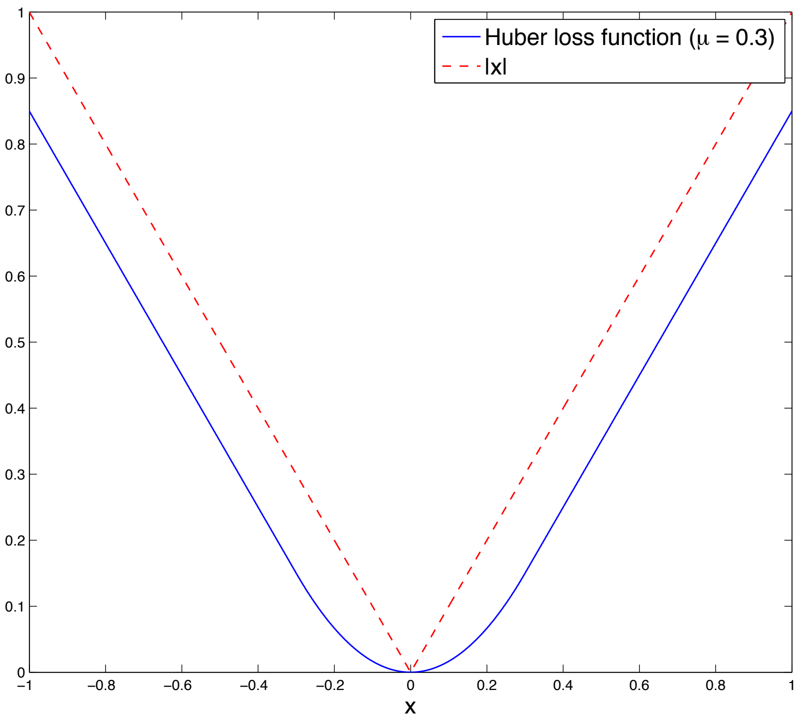

Figure 4.2: Huber loss function.

sense that non-descent iterative method such as subgradient method [61] has to be used. In the compressive sensing society, the ℓ_1-norm $\|\mathbf{x}\|_1$ is usually substituted by a general penalty function $\boldsymbol{\varphi}_s(\mathbf{x}) = \sum_i \varphi_s(x_i)$ [43]. In order to accelerate convergence, naturally we would choose a convex and smooth φ_s. A popular choice is the Huber loss function defined as

$$\varphi_s(x_i) = \begin{cases} \frac{x_i^2}{2\mu}, & 0 \leq |x_i| \leq \mu, \\ |x_i| - \frac{\mu}{2}, & \text{otherwise}, \end{cases} \tag{4.34}$$

which is depicted in Fig. 4.3. Huber loss function makes compromise between the ℓ_1 and ℓ_2-norm. It has stronger incentive in penalizing small values, and is essentially the same as ℓ_1-norm when the value is greater than certain threshold. Now, consider the following proposition:

Proposition 4.1 *In Problem* (4.32)*, the optimal objective value is equivalent to*

$$\min_{\mathbf{G}} \rho\|\mathbf{H} - \mathbf{G}\|_F^2 + \gamma\|\mathbf{G}\|_* = \sum_{i=1}^{r} \varphi_s(\sigma_i), \tag{4.35}$$

where r is the rank of $\mathbf{H}$, σ_i is the ith singular value of $\mathbf{H}$ and

$$\varphi_s(\sigma_i) = \begin{cases} \gamma\sigma_i - \frac{\gamma^2}{4\rho}, & \text{if } \sigma_i \geq \frac{\gamma}{2\rho}, \\ \rho\sigma_i^2, & \text{if } \sigma_i < \frac{\gamma}{2\rho}. \end{cases} \tag{4.36}$$

The proof of Proposition 4.1 is given in Appendix B.4. From Proposition 4.1, Problem (4.29) can be rewritten as

$$\begin{aligned} &\min_{\mathbf{H},\mathbf{G}} \ \|\alpha\mathbf{a}\mathbf{a}^H - \mathbf{H}\|_F^2 + \rho\|\mathbf{H} - \mathbf{G}\|_F^2 + \gamma\|\mathbf{G}\|_* \\ &\quad \text{s.t. } \operatorname{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s), \\ =&\min_{\mathbf{H}} \ \left\{\|\alpha\mathbf{a}\mathbf{a}^H - \mathbf{H}\|_F^2 + \min_{\mathbf{G}} \rho\|\mathbf{H} - \mathbf{G}\|_F^2 + \gamma\|\mathbf{G}\|_*\right\} \\ &\quad \text{s.t. } \operatorname{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s), \\ =&\min_{\mathbf{H}} \ \|\alpha\mathbf{a}\mathbf{a}^H - \mathbf{H}\|_F^2 + \sum_{i=1}^{r} \varphi_s(\sigma_i) \\ &\quad \text{s.t. } \operatorname{vec}(\mathbf{H}) \in \mathcal{R}(\mathbf{U}_s), \end{aligned} \tag{4.37}$$

where σ_i is the ith singular value of $\mathbf{H}$. Specifically, a smooth approximation of the nuclear norm is employed. For convenience, we call the algorithm *KR Huber*, in which the algorithm is given in **Algorithm 5**.

To illustrate the efficacy of nuclear norm regularization and Huber regularization, let us check the convergence behavior of AP. The projection error $\|(\mathbf{I} - \mathbf{U}_s\mathbf{U}_s^H)\mathbf{a}^* \otimes \mathbf{a}\|^2$ is shown in Fig. 4.2. Here, we consider the perfect data scenario (cf. Section 4.3) where $\mathbf{A} \in \mathbb{R}^{5\times 7}$. The regularization constants are $(\lambda, \rho) = (0.5, 1)$. We can see that the nuclear norm regularization dramatically reduces the number of iterations required by AP. The Huber regularization reduces the number of iterations as well, although the reduction is traded off by the computational efficiency.

4.5 Robust KR Subspace Extraction

In our framework, a very important step is to extract the KR subspace $\mathbf{U}_s$. The KR subspace is extracted by computing the compact SVD of $\mathbf{Y}$ (cf. Section 4.2):

$$\mathbf{Y} = [\operatorname{vec}(\mathbf{R}_1), \ldots, \operatorname{vec}(\mathbf{R}_M)].$$

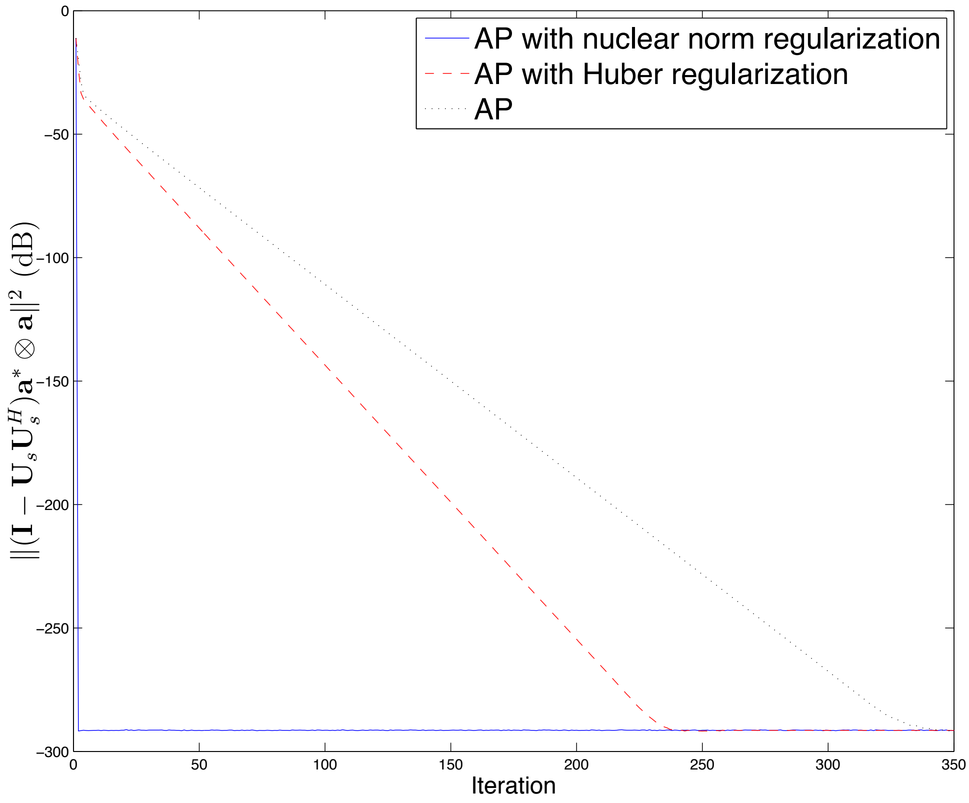

Figure 4.3: Projection error against iteration

If there is no modeling error on the local covariance matrices $\{\mathbf{R}_1,\ldots,\mathbf{R}_M\}$, using SVD is satisfactory. In practice, however, the estimated local covariance matrices may suffer from severe modeling errors; i.e.,

$$\mathbf{R}_m=\mathbf{A}\mathbf{D}_m\mathbf{A}^H+\mathbf{E}_m,\ m=1,\ldots,M, \tag{4.38}$$

where $\{\mathbf{E}_1,\ldots,\mathbf{E}_M\}$ represent modeling errors. In BI-QSS, the local covariance model is based on the assumption that the sources are mutually independent; nevertheless, in some cases, this assumption can be severely violated. In other words, the source covariance matrices $\{\mathbf{D}_1,\ldots,\mathbf{D}_M\}$ may no longer be diagonal. In Fig. 4.4, the correlation between two speech sources against the frame number is shown. The correlation of two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ is defined as

$$\frac{\mathbf{x}_1^H\mathbf{x}_2}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|}.$$

We can see that the correlation is small in general. However, in some frames, the source correlation can be quite high. The local covariance

Algorithm 5 KR Huber.

```
Input: $\mathbf{U}_s$: KR subspace; $\mathbf{h} \in \mathcal{R}(\mathbf{U}_s), \mathbf{G}$: randomly generated initial points.
Output: $\mathbf{a}$: a mixing matrix column;
$\mathbf{H} = \text{vec}^{-1}(\mathbf{h})$;
repeat
    $\tilde{\mathbf{H}} = \frac{1}{2}\left(\mathbf{H} + \mathbf{H}^H\right)$;
    $\mathbf{a} = \mathbf{q}_{\max}(\tilde{\mathbf{H}})$;
    $\alpha = \lambda_{\max}(\tilde{\mathbf{H}})/|\lambda_{\max}(\tilde{\mathbf{H}})|$;
    repeat
        $\mathbf{G} = \text{SVT}\left(\mathbf{H}, \frac{\gamma}{2\rho}\right)$;
        $\mathbf{H} = \text{vec}^{-1}\left(\frac{1}{1+\rho}\mathbf{U}_s\mathbf{U}_s^H(\alpha\mathbf{a}^* \otimes \mathbf{a} + \rho\text{vec}(\mathbf{G}))\right)$;
    until a stopping criterion is satisfied.
until a stopping criterion is satisfied.
```

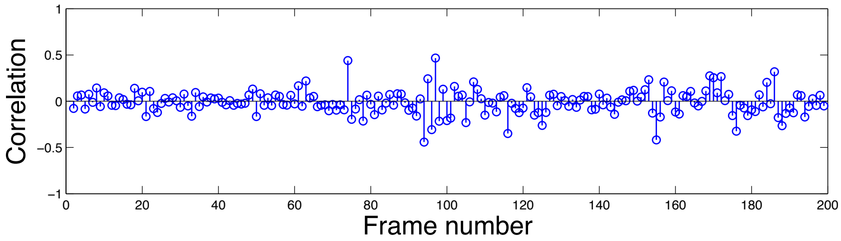

Figure 4.4: Source correlation.

matrices corresponding to those frames will be corrupted. Eventually, the KR subspace obtained from $\mathbf{Y}$ will be corrupted as well, i.e.

$$\mathcal{R}(\mathbf{U}_s) \neq \mathcal{R}(\mathbf{A}^* \odot \mathbf{A}). \tag{4.39}$$

Therefore, identifying mixing matrix columns from the corrupted KR subspace will be problematic.

In this section, we propose a way to detect and drop highly corrupted local covariance matrices together with the KR subspace extraction procedure. To facilitate our discussion, let us first review the procedure in extracting the KR subspace using SVD. The SVD of $\mathbf{Y}$ may be described

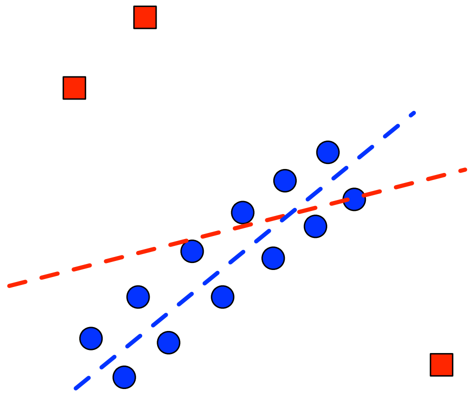

Figure 4.5: Linear regression with outliers.

by the following optimization problem:

$$
\begin{aligned}
&\min_{\mathbf{K},\mathbf{U}_s} \ \|\mathbf{Y}-\mathbf{U}_s\mathbf{K}\|_F^2 \\
&\text{s.t. } \mathbf{U}_s^H\mathbf{U}_s = \mathbf{I} \\
= &\min_{\{\mathbf{k}_1,\ldots,\mathbf{k}_M\},\mathbf{U}_s} \ \sum_{m=1}^{M} \|\mathbf{y}_m-\mathbf{U}_s\mathbf{k}_m\|^2 \\
&\text{s.t. } \mathbf{U}_s^H\mathbf{U}_s = \mathbf{I},
\end{aligned}
\tag{4.40}
$$

where $\mathbf{K} = [\mathbf{k_1},\ldots,\mathbf{k}_M] = \mathbf{\Sigma}_s\mathbf{V}_s^H$ According to Problem (4.40), SVD finds a basis matrix $\mathbf{U}_s$ and a set of coefficient vectors $\{\mathbf{k}_1,\ldots,\mathbf{k}_M\}$ that best fit the observations $\{\mathbf{y}_1,\ldots,\mathbf{y}_M\}$ in a least-squares sense. If there are some highly corrupted data in the set $\{\mathbf{y}_1,\ldots,\mathbf{y}_M\}$, finding a best least-squares fit might result in corrupted subspace. Consider a linear regression problem in the presence of outliers depicted in Fig. 4.5. In Fig. 4.5, the normal data (the blue dots) show a clear linear pattern; yet the linear regression model may give us a wrong model, e.g. the red dotted line, because of the existence of outliers (the red squares). Here, if we can detect and remove outliers before finding the linear regression model, hopefully a more representative model can be obtained (i.e. the blue dotted line).

Here, we incorporate a recently proposed method to drop highly corrupted frames [38]. The rationale behind the method is that if the fitting error

$$\|\mathbf{y_m}-\mathbf{U}_s\mathbf{k}_m\|^2 \tag{4.41}$$

is large, $\mathbf{y}_m$ is likely to be corrupted. Therefore, we want to eliminate those highly corrupted local covariance matrices during the extraction of KR subspace. Consider the following problem:

$$\min_{\text{num}\{\mathbf{z}_1,\ldots,\mathbf{z}_M\}\leq Z} \left\{ \min_{\{\mathbf{k}_1,\ldots\mathbf{k}_M\},\mathbf{U}_s:\mathbf{U}_s^H\mathbf{U}_s=\mathbf{I}} \sum_{m=1}^{M} \|\mathbf{y}_m - \mathbf{U}_s\mathbf{k}_m - \mathbf{z}_m\|^2 \right\}, \tag{4.42}$$

where $\text{num}\{\mathbf{z}_1,\ldots,\mathbf{z}_M\}$ denotes the number of non-zero vectors in $\{\mathbf{z}_1,\ldots,\mathbf{z}_M\}$ and Z is the predicted number of highly corrupted frames. Here, the additional variables $\{\mathbf{z}_1,\ldots,\mathbf{z}_M\}$ is introduced to cancel out the highly corrupted terms, determined by the fitting error $\|\mathbf{y}_m - \mathbf{U}_s\mathbf{k}_m\|^2$.

Problem (4.42) can be handled using alternating optimization. Specifically, the partial minimization problem of Problem (4.42) w.r.t. $\{\mathbf{U}_s,\mathbf{k}_1,\ldots,\mathbf{k}_M\}$ with $\{\mathbf{z}_1,\ldots,\mathbf{z}_M\}$ fixed is:

$$\begin{aligned} \min_{\{\mathbf{k}_1,\ldots,\mathbf{k}_M\},\mathbf{U}_s} \ & \sum_{m=1}^{M} \|(\mathbf{y}_m - \mathbf{z}_m) - \mathbf{U}_s\mathbf{k}_m\|^2 \\ \text{s.t. } & \mathbf{U}_s^H\mathbf{U}_s = \mathbf{I}, \end{aligned} \tag{4.43}$$

which can be handled by SVD (cf. Problem (4.40)); moreover, the partial minimization problem w.r.t. $\{\mathbf{z}_1,\ldots,\mathbf{z}_M\}$ with $\{\mathbf{U}_s,\mathbf{k}_1,\ldots,\mathbf{k}_M\}$ fixed is

$$\min_{\text{num}\{\mathbf{z}_1,\ldots,\mathbf{z}_M\}\leq Z} \sum_{m=1}^{M} \|(\mathbf{y}_m - \mathbf{U}_s\mathbf{k}_m) - \mathbf{z}_m\|^2, \tag{4.44}$$

whose solution is given by

$$\mathbf{z}_m = \begin{cases} \mathbf{y}_m - \mathbf{U}_s\mathbf{k}_m, & \text{if } m \in \mathcal{I} = \{i_1, i_2, \ldots, i_Z\}, \\ \mathbf{0}, & \text{if } m \notin \mathcal{I}, \end{cases} \tag{4.45}$$

where i_k is the index of the kth largest value in $\{\|\mathbf{y}_1-\mathbf{U}_s\mathbf{k}_1\|^2,\ldots,\|\mathbf{y}_M-\mathbf{U}_s\mathbf{k}_M\|^2\}$. From the above procedure, we can see that the most corrupted Z frames will be discarded in the next subspace extraction procedure, while the good frames remain intact. The robust KR subspace extraction procedure is given in **Algorithm 6**.

4.6 Summary

In this chapter, we have introduced the KR subspace criterion based on the local covariance model. The criterion is significantly different from

Algorithm 6 Robust KR Subspace Extraction.

```
Input: $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$: vectorized covariance matrices;
Output: $\mathbf{U}_s$: robust KR subspace;
initialize $\{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$.
repeat
    compute the compact SVD of $\tilde{\mathbf{Y}} = [\mathbf{y}_1 - \mathbf{z}_1, \ldots, \mathbf{y}_M - \mathbf{z}_M]$;
    i.e., $\tilde{\mathbf{Y}} = \mathbf{U}_s\mathbf{K}$;
    compute the index set $\mathcal{I}$ by checking the value of
    $\{\|\mathbf{y}_1 - \mathbf{U}_s\mathbf{k}_1\|^2, \ldots, \|\mathbf{y}_M - \mathbf{U}_s\mathbf{k}_M\|^2\}$;
    $\mathbf{z}_m = \begin{cases} \mathbf{y}_m - \mathbf{U}_s\mathbf{k}_m, & \text{if } m \in \mathcal{I} = \{i_1, i_2, \ldots, i_Z\}, \\ \mathbf{0}, & \text{if } m \notin \mathcal{I}, \end{cases}$
until a stopping criterion is satisfied.
```

that of PARAFAC and JD, in the sense that per-column identification procedure is suggested. We have shown that $K \leq 2N - 2$ is an sufficient condition for the KR subspace criterion to uniquely identify the mixing matrix columns, up to scaling factor.

Then, an algorithm handling BI-QSS utilizing the KR subspace criterion was devised. Specifically, alternating projections (AP) is used to handle BI-QSS whose solution can be computed in simple closed forms. An all-column identification procedure based on multiple randomized initializations is used to complement the per-column identification nature of the KR subspace criterion.

The subsequent development was divided into two parts. First of all, we considered overdetermined mixing models ($N > K$). In that case, we have shown that employing prewhitening before BI-QSS is beneficial to our proposed algorithm. With the employment of prewhitening, the proposed algorithm possesses provably rapid convergence. Moreover, we proposed a specialized all-column identification procedure by exploiting the unitarity of the mixing matrix after prewhitening.

In underdetermined mixing models ($K > N$), prewhitening is no longer possible. In order to speed up the convergence, rank minimization heuristics were proposed. Specifically, we considered a nuclear norm regularized problem. We have demonstrated that the nuclear regularization works well. Then, we proposed an efficient way to handle the nuclear norm regularized problem which shows connections with the Huber loss function used in smooth optimization and robust estimation.

Finally, we considered a practical situation where corrupted local covariance matrices are present. A robust subspace extraction procedure was proposed to mitigate the effect of corrupted data.

□ **End of chapter.**

Chapter 5

Simulation Results

In this chapter, we will use simulations to demonstrate the efficacy of PAPA and KR Huber. We will first introduce the general settings used throughout this chapter. Then, the discussion will be divided into two parts. In the first part, we will show the performance of PAPA in overdetermined mixing models. We will compare PAPA with FFDIAG [4], UWEDGE [9], Pham's JD [5] and TALS [11]. In the second part, the performance of KR Huber in underdetermined mixing models will be examined. We will compare KR Huber with TALS [11] and ACDC [12].

5.1 General Settings

The general settings used in both overdetermined and underdetermined mixing models are listed below:

(S1) The mixing matrix $\mathbf{A} \in \mathbb{R}^{N \times K}$ is randomly generated at each trial, where the elements of $\mathbf{A}$ are i.i.d. Gaussian distributed with zero mean and unit variance. The columns of $\mathbf{A}$ are then normalized to unit norm.

(S2) The source signals are speech recordings. A database consisting of 23 speech signals, each with length 6 seconds, is used. The speech signals are normalized to zero mean and unit power. The sampling rate is equal to 16kHz.

(S3) The received signals $\mathbf{x}(t)$ are partitioned into M frames, each with

L samples. Furthermore, in the estimation of local covariance matrices, frames with 50% overlapping is employed; i.e.

$$\mathbf{R}_m = \frac{1}{L} \sum_{t=-0.5(m-1)L+1}^{0.5(m-1)L+L} \mathbf{x}(t)\mathbf{x}^H(t). \tag{5.1}$$

(S4) To simulate a noisy environment, noise is added to the observations. The signal-to-noise ratio (SNR) is defined as

$$\text{SNR} = \frac{1}{T} \sum_{t=0}^{T-1} \frac{\text{E}\{\|\mathbf{A}\mathbf{s}(t)\|^2\}}{\text{E}\{\|\mathbf{v}(t)\|^2\}},$$

where $T = M \times L$ is the length of source signals.

(S5) The noise removal procedure discussed in Section 2.4 was applied to the estimated local covariance matrices $\{\mathbf{R}_1, \ldots, \mathbf{R}_M\}$.

(S6) The robust subspace extraction procedure provided in **Algorithm 6** is used. It stops when the relative change in objective value is lower than 10^{-3}, or the number of iterations exceed 15. We assume 15% of the frames (after frame overlapping) are corrupted.

(S7) The proposed algorithms will be stopped when its relative change in objective value is lower than 10^{-6}; i.e.,

$$\frac{|f^{(n+1)} - f^{(n)}|}{|f^{(n)}|} < 10^{-6},$$

where $f^{(n)}$ is the objective value of the algorithm at the nth iteration; we stop other BI-QSS algorithms using similar stopping criteria for benchmarking.

(S8) The maximum allowable number of iterations is 2000.

(S9) The performance measure employed here is the average mean square error (MSE), defined as

$$\text{MSE} = \min_{\substack{\boldsymbol{\pi} \in \boldsymbol{\Pi}, \\ c_1, \ldots, c_K \in \{\pm 1\}}} \frac{1}{K} \sum_{k=1}^{K} \left\| \frac{\mathbf{a}_k}{\|\mathbf{a}_k\|} - c_k \frac{\hat{\mathbf{a}}_{\pi(k)}}{\|\hat{\mathbf{a}}_{\pi(k)}\|} \right\|^2,$$

where $\boldsymbol{\Pi}$ is the set of all bijections $\boldsymbol{\pi} : \{1, 2, \ldots, K\} \rightarrow \{1, 2, \ldots, K\}$; $\mathbf{A}$ and $\hat{\mathbf{A}}$ are the true and estimated mixing matrix, respectively. In essence, the MSE is calculated after fixing the scaling and permutation ambiguities.

(S10) All the algorithms are run on a Desktop PC with Core 2 Duo 3GHz CPU and 4GB RAM, and the codes are all written in MATLAB. One thousand independent trials are performed.

5.2 Overdetermined Mixing Models

In this section, we consider overdetermined mixing models. Specifically, we will compare PAPA with FFDIAG [4], UWEDGE [9], Pham's JD[1] [5], and TALS [11]. Except for TALS, the noise covariance removed local covariance matrices are prewhitened.

5.2.1 Simulation 1 - Performance w.r.t. SNR

Let us first examine the performance of BI-QSS w.r.t. SNR. We fix $(N, K) = (6, 5)$ and $(M, L) = (399, 200)$. Note that the number of available frames M is calculated after frame overlapping. The average MSEs of the various algorithms w.r.t. SNR are shown in Fig. 5.1. We can see that TALS performs well in low SNR regime. For SNR $\geq$ 5dB, PAPA yields the best estimation accuracy.

Table 5.1 shows the average runtimes of the various algorithms w.r.t. SNR. For PAPA, we also listed the average number of iterations required for the AP to converge. PAPA clearly demonstrated its high efficiency. It is at least four times faster than UWEDGE. It is also worth noticing that even in face of modeling error, PAPA converges with small number of iterations. It matches the rapid convergence behavior shown in Theorem 4.2.

5.2.2 Simulation 2 - Performance w.r.t. the Number of Available Frames M

Next, let us examine the effect of the number of available frames M on BI-QSS. We fix $(N, K) = (6, 5)$ and $L = 200$. The SNR is fixed at 10dB. The average MSEs of the various algorithms w.r.t. M are shown in Fig.

[1] Diagonal loading $\delta\mathbf{I}$ with $\delta = 1e^{-6}$ is added to the local covariance matrices to ensure positive definiteness.

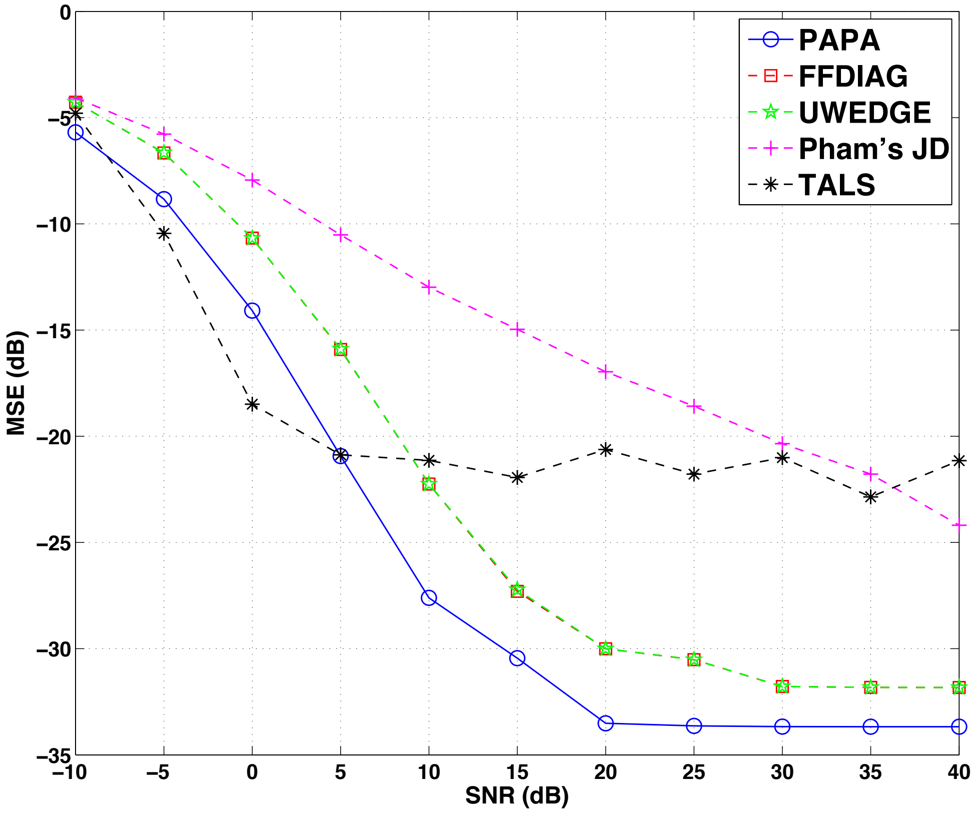

Figure 5.1: The average MSEs of the various algorithms w.r.t. SNR.

5.2. When the number of frames is not enough, BI-QSS algorithms are not able to capture the non-stationarity of the source signals. Therefore, estimation accuracy of the various algorithms are poor. We can see when M is sufficiently large, PAPA yields the best estimation accuracy.

Table 5.2 shows the average runtimes of the various algorithms w.r.t. M. When the number of frames is not sufficient, the runtime of PAPA is less competitive. However, we can see that the runtime of PAPA grows slowly against M, comparing with other algorithms. It is due to the fact that after extracting the robust KR subspace, the number of frames M has no effect to PAPA.

5.2.3 Simulation 3 - Performance w.r.t. the Number of Sources K

We also test the proposed algorithm with different number of sources. Here, the number of sensors is equal to $K + 1$ and $(M, L) = (399, 200)$.

Table 5.1: The average runtimes of the various algorithms w.r.t. SNR.

		SNR-10dB	0dB	10dB	20dB	30dB	40dB
PAPA	Time (sec.)	**0.00722**	**0.00798**	**0.00556**	**0.00537**	**0.00536**	**0.00539**
	Iteration	13.22	6.972	4.336	3.614	3.504	3.479
FFDIAG	Time (sec.)	0.167	0.0685	0.0391	0.034	0.0334	0.0329
UWEDGE	Time (sec.)	0.186	0.0554	0.03	0.0246	0.0234	0.0236
Pham's JD	Time (sec.)	5.935	1.952	1.401	1.119	1.045	1.04
TALS	Time (sec.)	2.484	1.252	1.327	1.307	1.335	1.348

Table 5.2: The average runtimes of various algorithms w.r.t. M.

		M=39	159	319	479	639	799	959
PAPA	Time (sec.)	0.00849	**0.00588**	**0.00568**	**0.00625**	**0.00699**	**0.00762**	**0.0083**
	Iteration	15.0702	5.153	4.394	4.289	4.374	4.39	4.277
FFDIAG	Time (sec.)	0.00729	0.00172	0.0333	0.0479	0.0624	0.0796	0.0956
UWEDGE	Time (sec.)	**0.0062**	0.0112	0.0211	0.0306	0.0388	0.0490	0.0586
Pham's JD	Time (sec.)	0.0575	0.3501	1.097	2.29	3.789	6.208	8.959
TALS	Time (sec.)	0.386	0.594	1.068	1.996	3.543	5.553	7.402

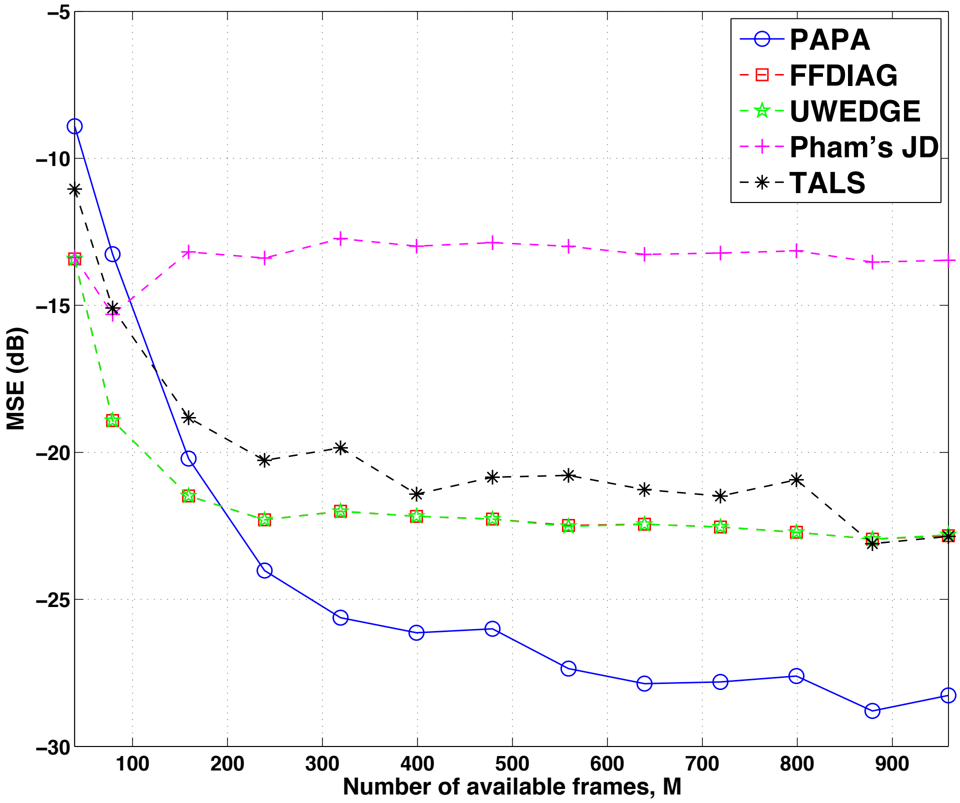

Figure 5.2: The average MSEs of the various algorithms w.r.t. M.

The SNR is fixed at 10dB. The performance is given in Table 5.3. We can see that PAPA gives satisfactory estimation accuracy and at the same time quite efficient.

5.3 Underdetermined Mixing Models

In this section, we consider underdetermined mixing models. We will compare KR Huber with TALS [11] and ACDC [12]. Throughout this section, we fix $(N, K) = (5, 7)$ and $(M, L) = (399, 400)$. Moreover, it has been noticed that in underdetermined mixing models, good conditioning of $\mathbf{A}$ is crucial [20–22]. Here, we constrain the columns of $\mathbf{A}$ to have cross-correlation less than 0.8; i.e.,

$$|\mathbf{a}_i^H \mathbf{a}_j| < 0.8, \forall\ i \neq j.$$

Table 5.3: Identification performance of various algorithms w.r.t. K.

		K =6	7	8	9	10
PAPA	MSE (dB)	**-25.68**	**-23.49**	**-22.65**	**-22.29**	**-21.07**
	Time (sec.)	**0.0104**	**0.0176**	**0.0268**	**0.0426**	**0.0616**
	Iteration	4.626	4.794	4.918	5.008	5.092
FFDIAG	MSE (dB)	-20.48	-19.71	-18.64	-18.78	-17.91
	Time (sec.)	0.0491	0.0628	0.0737	0.0946	0.103
UWEDGE	MSE (dB)	-20.48	-19.71	-18.64	-18.78	-17.91
	Time (sec.)	0.0464	0.0581	0.0691	0.0872	0.0977

For KR Huber (cf. **Algorithm 4**), we set $(\gamma, \rho) = (0.5, 1)$. Also, the inner alternating projections (Problem (4.29)) terminates when

$$\frac{|g^{(n+1)} - g^{(n)}|}{|g^{(n)}|} < 10^{-3},$$

where $g^{(n)}$ is the objective value of Problem (4.29) at the nth iteration.

5.3.1 Simulation 1 - Success Rate of KR Huber

In underdetermined mixing models, all-column identification of KR Huber relies on randomized initializations (cf. **Algorithm 2**). In this section, let us begin with examining the efficacy of **Algorithm 2**. Here, we fix the maximum number of randomized initializations to $20 \times K = 140$. **Algorithm 2** will be terminated when K columns with cross-correlations

$$|\hat{\mathbf{a}}_i^H \hat{\mathbf{a}}_j| < 0.8$$

are obtained.

The performance of **Algorithm 2** applying to KR Huber is given in Table 5.4. Specifically, **Algorithm 2** is considered to be “successful” if K columns with cross-correlation less than 0.8 are obtained. For low SNR (SNR=-10dB), **Algorithm 2** performs badly. When the SNR is so low, the modeling error is severe. Even with robust KR subspace extraction, the extracted subspace can be severely corrupted. Fortunately, for reasonable SNR, promising success rate can be seen.

Table 5.4: Performance on the all-column identification heuristic.

	Success rate					
SNR (dB)	-10	0	10	20	30	40
KR Huber (%)	6.2	79.7	87.7	87.5	88.1	87.6

Table 5.5: The average runtimes of the various algorithms w.r.t. SNR.

		SNR-10dB	0dB	10dB	20dB	30dB	40dB
KR Huber	Time (sec.)	**2.6**	**0.54**	**0.304**	**0.31**	**0.291**	**0.29**
TALS	Time (sec.)	2.654	1.82	1.782	1.778	1.772	1.849
ACDC	Time (sec.)	19.3	14.05	13.88	14.43	14.92	14.32

5.3.2 Simulation 2 - Performance w.r.t. SNR

Subsequently, let us compare KR Huber with TALS and ACDC. For the case where **Algorithm 2** fails, the remaining unidentified columns are randomly generated. In those instances, the estimation is completed by putting the estimated mixing matrix $\hat{\mathbf{A}}$ to TALS as an initialization. The average MSEs of the various algorithms w.r.t SNR are shown in Fig. 5.3. In the low SNR regime, PARAFAC-based algorithms are competitive. On the other hand, for SNR $\geq$5dB, KR Huber starts to work better. KR Huber yields significantly better estimation accuracy comparing to TALS and ACDC when SNR $\geq$15dB.

The average runtimes of the various algorithms w.r.t. SNR are listed in Table 5.5. In the low SNR regime, as KR Huber is almost impossible to get all K columns using **Algorithm 2**. As it relies heavily on TALS, the runtime performance is similar to TALS. In the high SNR regime, however, significantly better runtime performance of KR Huber can be seen.

5.3.3 Simulation 3 - Performance w.r.t. M

Let us examine the effect of the number of available frames M on BI-QSS. We fix $(N, K) = (5, 7)$ and $L = 400$. The SNR is fixed at 10dB. The average MSEs of the various algorithms w.r.t. M are shown in Fig. 5.4.

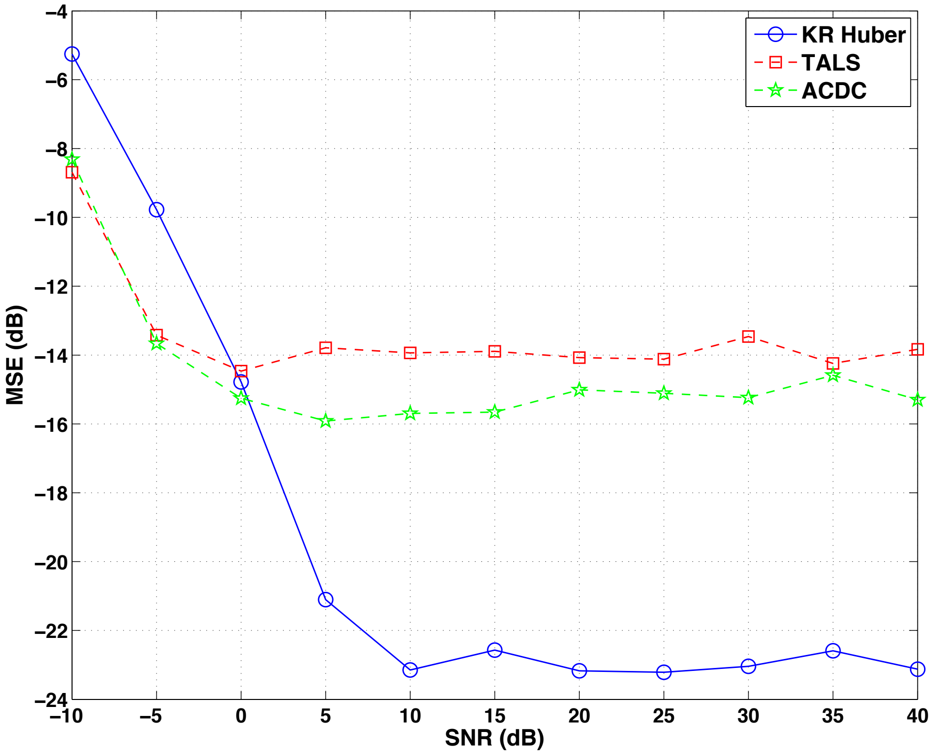

Figure 5.3: The average MSEs of the various algorithms w.r.t SNR.

Table 5.6: The average runtimes of the various algorithms w.r.t. M.

		M=39	119	199	279	359	439	479
KR Huber	Time (sec.)	0.943	0.49	**0.329**	**0.318**	**0.303**	**0.308**	**0.318**
TALS	Time (sec.)	**0.201**	**0.314**	0.48	0.729	0.983	1.268	1.48
ACDC	Time (sec.)	4.1	6.67	8.28	10.6	12.84	14.79	16.58

KR Huber performs substantially better than both TALS and ACDC, and the performance gap is even wider when we have more frames.

Table 5.6 shows the average runtimes of the various algorithms w.r.t. M. To KR Huber, insufficiency of available frames has huge impact. However, when M is getter greater, the average runtime of KR Huber is almost invariant w.r.t. M. On the other hand, the computational efficiency of TALS and ACDC are getting worse when M increases.

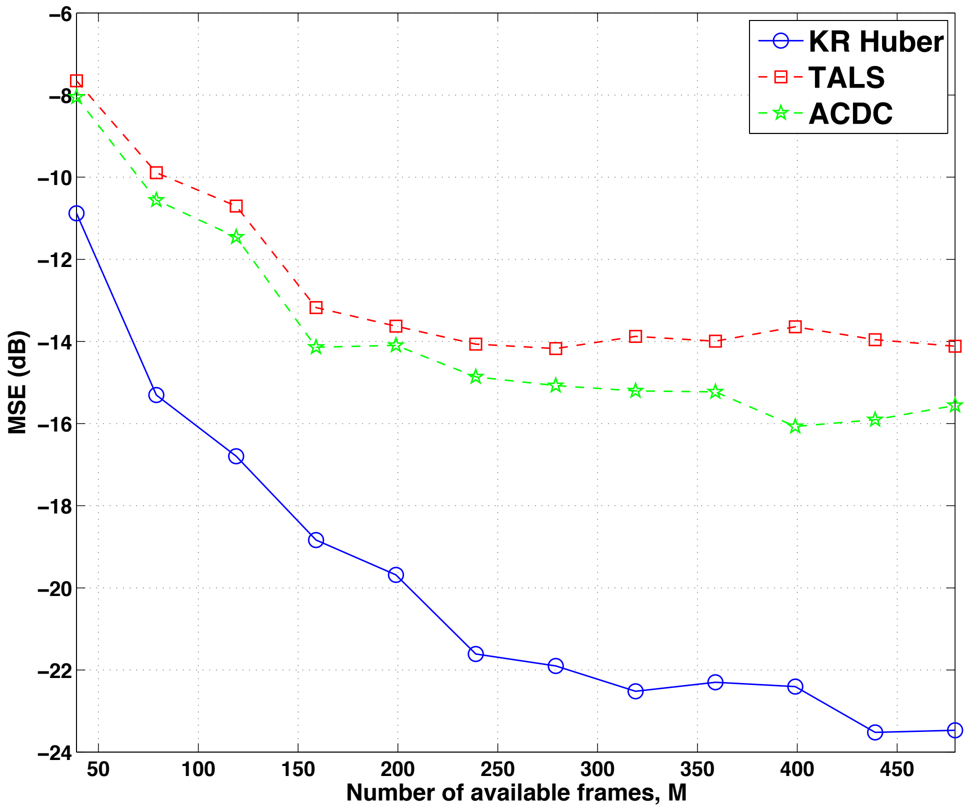

Figure 5.4: The average MSE of the various algorithms w.r.t. M.

5.3.4 Simulation 4 - Performance w.r.t. N

Finally, we test KR Huber with different number of sensors. Here, the number of sources is equal to $2N-3$ and $(M, L) = (399, 200)$. The SNR is fixed at 10dB. The performance is given in Table 5.7. We can see that KR Huber estimates the mixing matrix $\mathbf{A}$ quite accurately even when we have so many sources. In terms of runtime performance, KR Huber is also attractive.

5.4 Summary

In this chapter, we have demonstrated the efficacy of PAPA and KR Huber by extensive simulations. The two algorithms yield competitive estimation accuracy and computational efficiency. Hence, we conclude that the proposed BI-QSS algorithms are efficient and serve as good BI-QSS alternatives in practice.

Table 5.7: Performance of various algorithms against N.

		N =5	6	7	8	9
KR Huber	MSE (dB)	**-22.95**	**-21.08**	**-19.69**	**-17.96**	**-17.52**
	Time (sec.)	**0.293**	**0.58**	**0.948**	**1.52**	**2.281**
TALS	MSE (dB)	-14.06	-14.17	-13.89	-14.12	-13.73
	Time (sec.)	1.468	2.247	3.364	4.927	6.277
ACDC	MSE (dB)	-15.43	-15.55	-15.75	-15.65	-15.39
	Time (sec.)	13.77	18.3	22.14	21.7	25.36

□ **End of chapter.**

Chapter 6

Conclusion and Future Works

In this thesis, blind identification of mixtures of quasi-stationary sources (BI-QSS) has been studied. Specifically, a second-order statistics (SOSs) based per-source blind identification criterion has been developed based on exploiting the time-varying characteristics of quasi-stationary signals. We have shown that this blind identification criterion uniquely determines (up to scaling ambiguity) the mixing matrix columns for $K \leq 2N - 2$, where N and K denote the number of sensors and sources in a given system, respectively. Moreover, for the overdetermined mixing models, a specialized alternating projections algorithm has been proposed, in which the mixing matrix column can be uniquely determined in one iteration with probability one under some mild conditions. For the more challenging underdetermined mixing models, rank-minimization heuristics were proposed to speed up the alternating projections algorithm. In practical situations, the existence of corrupted data is inevitable. A specialized robust subspace extraction procedure was employed to mitigate the damage caused by corrupted data. Extensive simulation results illustrate that KR subspace based algorithms are competitive in both computational complexity and estimation accuracy.

As a future direction, it would be interesting to seek a systematic way in performing all-column identification in underdetermined mixing models. We have already seen the competitive performance of KR Huber when the all-column identification heuristic works well. Also, with the ability to systematically cancel the estimated columns, the runtime performance can be significantly improved. Besides, it would be interesting to see how the joint BSS [17] work can be incorporated in our proposed algorithms in efficiently handling convolutive mixtures. Owing to the high efficiency of PAPA, real-time blind identification would also be an interesting direction to pursue.

□ **End of chapter.**

Appendix A

Convolutive Mixing Model

Recall the linear convolutive mixing model:

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{A} \star \mathbf{s}(t) + \mathbf{v}(t) \\ &= \sum_{\tau=0}^{L} \mathbf{A}(\tau)\mathbf{s}(t-\tau) + \mathbf{v}(t),\ t = 0, 1, 2, \ldots, \end{aligned} \tag{A.1}$$

where L is the length of the convolutional sum. This linear convolutive mixing model is considered to be a realistic model because the reverberation of the recording environment is taken into account. A typical reverberative environment is illustrated in Fig. A.1. In a reverberative environment, in addition to the directed sound, there are also reflected sounds from the surfaces. The reflected sounds are differently delayed and mixed to form the received signal vector $\mathbf{x}(t)$.

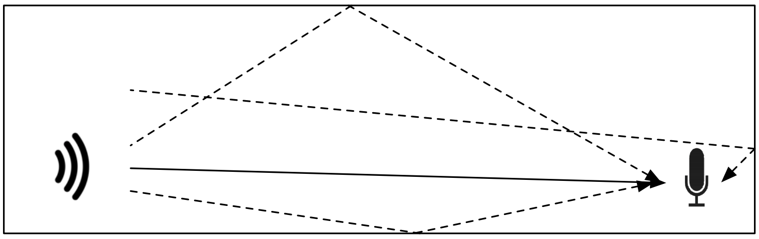

Figure A.1: Reverberation.

To handle the linear convolutive mixing model, one may transform it into the frequency domain and solve a set of blind identification problems with linear instantaneous mixing model corresponding to different frequencies. To do so, we make use of the discrete Fourier Transform

(DFT). Recall an important property of DFT:

$$\mathcal{F}(x(t) \circ y(t)) = \mathcal{F}(x(t))\mathcal{F}(y(t)), \tag{A.2}$$

where $\mathcal{F}(\cdot)$ is the DFT operator and $\circ$ is the circular convolution operator. It is known that linear convolution can be approximated by a circular convolution if the size of DFT is much larger than L [14]. Hence, by applying DFT to both sides of Eq. (A.1), we get

$$\mathbf{x}(f,t) \approx \mathbf{A}(f)\mathbf{s}(f,t) + \mathbf{v}(f,t), \text{ for } T \gg L, \tag{A.3}$$

where $\mathbf{x}(f,t) \triangleq \sum_{\tau=0}^{T-1} \mathbf{x}(t+\tau)e^{-\frac{j2\pi f\tau}{T}}$ is the DFT of $[\mathbf{x}(t), \ldots, \mathbf{x}(t+T-1)]$. The expressions of $\mathbf{A}(f)$ and $\mathbf{s}(f,t)$ carry in the same fashion. With the model in Eq. (A.3), the local covariance model can be derived. The local covariance matrix in the mth frame at frequency f is defined as

$$\mathbf{R}_m(f) = \mathbb{E}\{\mathbf{x}(f,t)\mathbf{x}(f,t)^H\}, \text{ for } t \in [(m-1)L+1, mL]. \tag{A.4}$$

Since we have assumed that the sources are mutually independent (cf. Assumption (A1)), we have

$$\mathbf{R}_m(f) = \mathbf{A}(f)\mathbf{D}_m(f)\mathbf{A}(f)^H + \sigma^2(f)\mathbf{I}. \tag{A.5}$$

where $\mathbf{D}_m(f)$ is the source local covariance matrix in the mth frame at frequency f and $\sigma^2(f)$ is the variance of $\mathbf{v}(f,t)$ at frequency f. One important observation is that Eq. (A.5) is exactly the same as the local covariance model with linear instantaneous mixing model (cf. Eq. (2.6)). Therefore, for fixed f, technique developed based on the linear instantaneous mixing model (e.g. PAPA and KR Huber) can be applied to identify $\mathbf{A}(f)$.

However, there is another practical and even more challenging issue that one must be careful of. It is well known that $\mathbf{A}(f)$ can only be specified up to permutation and scaling; i.e.,

$$\hat{\mathbf{A}}(f) = \mathbf{D}(f)\mathbf{\Pi}(f)\mathbf{A}(f), \ f = 0, \ldots, T-1. \tag{A.6}$$

With instantaneous mixing model, these ambiguities are immaterial. However, it poses a serious problem in convolutive mixing model. Specifically, consistent permutations and scalings across frequencies are necessary for correctly reconstructing the sources $\mathbf{s}(t)$ [14,23]. In fact, resolving the frequency-dependent scaling and permutation ambiguities arising in convolutive mixing model is a very challenging topic and active research are still on-going.

Appendix B

Proofs

B.1 Proof of Theorem 4.1

Let us begin with the sufficiency. For $K \leq 2N-2$, we have $\mathcal{R}(\mathbf{U}_s) = \mathcal{R}(\mathbf{A}^* \odot \mathbf{A})$. Then, it is trivial that

$$\mathbf{a} = c\mathbf{a}_k \Longrightarrow \mathbf{a}^* \otimes \mathbf{a} \in \mathcal{R}(\mathbf{U}_s).$$

For the converse, we first write $\mathbf{a}^* \otimes \mathbf{a} = (\mathbf{A}^* \odot \mathbf{A})\boldsymbol{\beta}$, where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_K]^T \in \mathbb{C}^K$. Equivalently, by devectorization, we have

$$\mathbf{a}\mathbf{a}^H = \mathbf{A}\mathbf{D}\mathbf{A}^H, \tag{B.1}$$

where $\mathbf{D} = \mathrm{Diag}(\boldsymbol{\beta})$. Assume without loss that $\beta_1, \ldots, \beta_J > 0$, where $J \leq K$. Let us examine the ranges of J:

Case 1: when $J \leq N$, we have

$$\mathbf{a}\mathbf{a}^H = \check{\mathbf{A}}\check{\mathbf{D}}\check{\mathbf{A}}^H, \tag{B.2}$$

where $\check{\mathbf{D}} = [\beta_1, \ldots, \beta_J] \in \mathbb{C}^{J \times J}$ and $\check{\mathbf{A}} = [\mathbf{a}_1, \ldots, \mathbf{a}_J] \in \mathbb{C}^{N \times J}$. Since $\check{\mathbf{A}}$ has full column rank (according to (A4)) and L.H.S. of (B.2) is of rank 1, there can only be one non-zero element in $\check{\mathbf{D}}$, which implies $\mathbf{a} = c\mathbf{a}_k$.

Case 2: when $N < J \leq K$, we will show it by contradiction. We first observe that $\mathrm{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_N\} = \mathbb{C}^N$. Then, assume

$$\mathbf{a} = \mathbf{A}_1\boldsymbol{\eta}$$

without loss of generality where $\mathbf{A}_1 = [\mathbf{a}_1, \ldots, \mathbf{a}_N] \in \mathbb{C}^{N \times N}$ and $\boldsymbol{\eta} =$

$[\eta_1, \ldots, \eta_N]$. By (B.1), we have

$$\mathbf{a}\mathbf{a}^H = \mathbf{A}_1\boldsymbol{\eta}\boldsymbol{\eta}^H\mathbf{A}_1^H = \mathbf{A}_1\mathbf{D}_1\mathbf{A}_1^H + \mathbf{A}_2\mathbf{D}_2\mathbf{A}_2^H, \tag{B.3}$$

where $\mathbf{D}_1 = \mathrm{Diag}(\beta_1, \ldots, \beta_N) \in \mathbb{C}^{N\times N}$, $\mathbf{D}_2 = \mathrm{Diag}(\beta_{N+1}, \ldots, \beta_J, 0, \ldots, 0) \in \mathbb{C}^{K-N\times K-N}$ and $\mathbf{A}_2 = [\mathbf{a}_{N+1}, \ldots, \mathbf{a}_K] \in \mathbb{C}^{N\times K-N}$. According to Eq. (B.3), we have

$$\mathbf{A}_1\boldsymbol{\eta}\boldsymbol{\eta}^H\mathbf{A}_1^H - \mathbf{A}_1\mathbf{D}_1\mathbf{A}_1^H = \mathbf{A}_2\mathbf{D}_2\mathbf{A}_2^H. \tag{B.4}$$

According to the full column rank condition of $\mathbf{A}_1$ and $\mathbf{A}_2$, we have

$$\mathrm{rank}(\mathbf{A}_1\boldsymbol{\eta}\boldsymbol{\eta}^H\mathbf{A}_1^H - \mathbf{A}_1\mathbf{D}_1\mathbf{A}_1^H) = \mathrm{rank}(\boldsymbol{\eta}\boldsymbol{\eta}^H - \mathbf{D}_1) \geq N - 1$$

and

$$\mathrm{rank}(\mathbf{A}_2\mathbf{D}_2\mathbf{A}_2^H) = \mathrm{rank}(\mathbf{D}_2) = J - N \leq K - N.$$

As a consequence, we arrive at $K \geq 2N - 1$, which contradict to the assumption $K \leq 2N - 2$. It means that case 2 can never happen and the sufficiency is completed.

For the necessity, we proof it by contradiction. Consider a Vandermonde $\mathbf{A}$; i.e., $\mathbf{a}_k = [1, e^{j\theta_k}, \ldots, e^{j(N-1)\theta_k}]$, for which $\theta_k \in [0, 2\pi)$ and $\theta_i \neq \theta_j$, $\forall\ k \neq j$. This $\mathbf{A}$ has full Kruskal rank [58], which satisfies (A4). Suppose that

$$\mathbf{a} = [1, e^{j\psi}, \ldots, e^{j(N-1)\psi}]$$

for some ψ and $K = 2N - 1$ (thus $K > 2N - 2$). Then, by choosing

$$\theta_k = \begin{cases} \frac{2\pi(k-1)}{N}, & k = 1, \ldots, N, \\ \frac{2\pi(k-N-1)}{N} + \frac{\pi}{N}, & k = N+1, \ldots, 2N-1, \end{cases}$$

and $\psi = \frac{2\pi(N-1)}{N} + \frac{\pi}{N}$. It can be verified that $\mathbf{A}_1$ and $[\mathbf{A}_2, \mathbf{a}]$ are both unitary. By choosing $\mathbf{D}_1 = \mathbf{I}$ and $\mathbf{D}_2 = -\mathbf{I}$, both sides of Eq. (B.3) will be equal to $\mathbf{I}$. It implies $\mathbf{a}^* \otimes \mathbf{a} \in \mathcal{R}(\mathbf{U}_s)$ but $\mathbf{a} \neq c\mathbf{a}_k$, $\forall\ k$ and the proof is completed. ■

B.2 Proof of Theorem 4.2

It can be verified that for a unitary $\mathbf{A}$, $\mathbf{A}^* \odot \mathbf{A}$ is semi-unitary (i.e. $(\mathbf{A}^* \odot \mathbf{A})^H(\mathbf{A}^* \odot \mathbf{A}) = \mathbf{I}$). With this property, it can be shown that

$$\mathbf{U}_s = (\mathbf{A}^* \odot \mathbf{A})\boldsymbol{\Gamma} \tag{B.5}$$

for some unitary $\boldsymbol{\Gamma} \in \mathbb{C}^{K\times K}$. By the unitarity of $\boldsymbol{\Gamma}$, we have $\boldsymbol{\eta} \triangleq \boldsymbol{\Gamma}\boldsymbol{\xi} \sim \mathcal{CN}(0, \mathbf{I})$. Subsequently, the initialization can be expressed as

$$\mathbf{h} = \mathbf{U}_s\boldsymbol{\xi} = (\mathbf{A}^* \odot \mathbf{A})\boldsymbol{\eta}. \tag{B.6}$$

Let us consider the devectorization of $\mathbf{h}$. Devectorizing both sides of (B.6) yields

$$\text{vec}^{-1}(\mathbf{h}) = \mathbf{A}\text{Diag}(\boldsymbol{\eta})\mathbf{A}^H. \tag{B.7}$$

Therefore, we have

$$\begin{aligned} \tilde{\mathbf{H}} &= \frac{1}{2}\left(\text{vec}^{-1}(\mathbf{h}) + \text{vec}^{-1}(\mathbf{h})^H\right) \\ &= \mathbf{A}\left(\frac{\text{Diag}(\boldsymbol{\eta} + \boldsymbol{\eta}^*)}{2}\right)\mathbf{A}^H \\ &= \mathbf{A}\text{Diag}(\text{Re}\{\boldsymbol{\eta}\})\mathbf{A}^H. \end{aligned} \tag{B.8}$$

Since $\mathbf{A}$ is unitary, the right hand side of Eq. (B.8) is in fact an eigenvalue decomposition (EVD) of $\tilde{\mathbf{H}}$. The remaining question is whether (B.8) is the unique EVD. It is known that if the eigenvalues $\text{Re}\{\eta_1\}, \ldots, \text{Re}\{\eta_K\}$ are distinct, then the correspecting EVD is unique. As $\boldsymbol{\eta}$ is a continuous random vector, $\text{Re}\{\eta_i\} = \text{Re}\{\eta_j\}$ holds with probability zero for any $i \neq j$ and the proof is completed. ■

B.3 Proof of Observation 4.1

Denote $(\alpha^{(k)}, \mathbf{a}^{(k)}, \mathbf{h}^{(k)}, \tilde{\mathbf{H}}^{(k)})$ by the solutions of AP after the kth iteration according to **Algorithm 1** (cf. Section 4.2). Suppose that $\tilde{\mathbf{H}}^{(k)}$ is of rank-1, i.e.

$$\tilde{\mathbf{H}}^{(k)} = \beta\mathbf{b}\mathbf{b}^H,$$

for some $\beta \in \mathbb{C}, \mathbf{b} \in \mathbb{C}^N$. Then, in the next AP iteration (cf. **Algorithm 1**), we have

$$\alpha^{(k+1)}\mathbf{a}^{(k+1)}(\mathbf{a}^{(k+1)})^H = \beta\mathbf{b}\mathbf{b}^H = \tilde{\mathbf{H}}^{(k)}. \tag{B.9}$$

We also have

$$\begin{aligned}\mathbf{h}^{(k+1)} &= \mathbf{U}_s\mathbf{U}_s^H\left(\alpha^{(k+1)}\left(\mathbf{a}^{(k+1)}\right)^* \otimes \mathbf{a}^{(k+1)}\right)\\ &= \mathbf{U}_s\mathbf{U}_s^H\text{vec}(\tilde{\mathbf{H}}^{(k)}).\end{aligned} \tag{B.10}$$

Therefore, the objective value of (4.11) after the $(k+1)$th iteration is

$$\begin{aligned}&\|\alpha^{(k+1)}\left(\left(\mathbf{a}^{(k+1)}\right)^* \otimes \mathbf{a}^{(k+1)}\right) - \mathbf{h}^{(k+1)}\|^2\\ =&\|(\mathbf{I}-\mathbf{U}_s\mathbf{U}_s^H)\text{vec}(\tilde{\mathbf{H}}^{(k)})\|^2\\ =&0,\end{aligned} \tag{B.11}$$

where the last equality is due to the fact that $\text{vec}(\tilde{\mathbf{H}}) \in \mathcal{R}(\mathbf{U}_s)$ as $\mathbf{h} \in \mathcal{R}(\mathbf{U}_s)$. It means that $(\mathbf{a}^{(k+1)})^* \otimes \mathbf{a}^{(k+1)} \in \mathcal{R}(\mathbf{U}_s)$, in which the KR subspace criterion is satisfied. Thus, the AP terminates as it has already reached an intersection of two given sets and the proof is completed. ■

B.4 Proof of Proposition 4.1

Let us assume $\mathbf{H}$ is of rank-1 for simplicity. Denote the non-zero singular value of $\mathbf{H}$ by σ. Therefore, for $\sigma \geq \frac{\gamma}{2\rho}$, we have

$$\begin{aligned}\rho\|\mathbf{H}-\mathbf{G}^\star\|_F^2 + \gamma\|\mathbf{G}^\star\|_* &= \rho\left(\sigma - (\sigma - \frac{\gamma}{2\rho})\right)^2 + \lambda(\sigma - \frac{\gamma}{2\rho})\\ &= \gamma\sigma - \frac{\gamma^2}{4\rho}.\end{aligned} \tag{B.12}$$

And for $\sigma < \frac{\gamma}{2\rho}$, we have

$$\rho\|\mathbf{H}-\mathbf{G}^\star\|_F^2 + \gamma\|\mathbf{G}^\star\|_* = \rho\sigma^2. \tag{B.13}$$

To summarize, we have

$$\rho\|\mathbf{H}-\mathbf{G}^\star\|_F^2 + \gamma\|\mathbf{G}^\star\|_* = \begin{cases}\gamma\sigma - \frac{\gamma^2}{4\rho}, & \text{if } \sigma \geq \frac{\gamma}{2\rho},\\ \rho\sigma^2, & \text{if } \sigma < \frac{\gamma}{2\rho},\end{cases} \tag{B.14}$$

which is equivalent to a Huber loss function mapping the singular value of $\mathbf{H}$. The above derivation can be generalized to $\mathbf{H}$ of rank r in a straightforward manner and the proof is completed. ■

Appendix C

Singular Value Thresholding

Consider the following problem:

$$\min_{\mathbf{G}} \; \rho\|\mathbf{H}-\mathbf{G}\|_F^2+\gamma\|\mathbf{G}\|_*. \tag{C.1}$$

The above problem is a proximal minimization problem in which the solution is given by:

$$\mathbf{G}^\star=\mathrm{SVT}(\mathbf{H},\frac{\gamma}{2\rho}), \tag{C.2}$$

where $\mathrm{SVT}(\cdot)$ is the singular value thresholding operator (cf. Eq. (4.26) and (4.27)).

To see this, first note that the objective function of Problem (C.1) is strictly convex. Therefore, there exists a unique minimizer to Problem (C.1). Thus, we want to show that the unique minimizer is equal to $\mathrm{SVT}(\mathbf{H},\frac{\lambda}{2\rho})$.

For convenience, define

$$h(\mathbf{G})=\rho\|\mathbf{H}-\mathbf{G}\|_F^2+\gamma\|\mathbf{G}\|_*. \tag{C.3}$$

It is known that [61]

$$h(\mathbf{G}^\star)=\inf_{\mathbf{G}} h(\mathbf{G}) \Longleftrightarrow \mathbf{0}\in\partial h(\mathbf{G}^\star), \tag{C.4}$$

where $\partial h(\mathbf{G}^\star)$ is the subdifferential of $h(\mathbf{G})$ at $\mathbf{G}^\star$. In other words, we have

$$\mathbf{0}\in\mathbf{H}-\mathbf{G}^\star+\frac{\gamma}{2\rho}\partial\|\mathbf{G}^\star\|_*. \tag{C.5}$$

It is known that [44]

$$\partial\|\mathbf{X}\|_*=\{\mathbf{U}\mathbf{V}^H+\mathbf{W}|\mathbf{U}^H\mathbf{W}=\mathbf{0},\mathbf{W}\mathbf{V}=\mathbf{0},\|\mathbf{W}\|_2\leq 1\}, \tag{C.6}$$

where $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$ and $\|\cdot\|_2$ is the dual norm of nuclear norm, i.e. the spectral norm [54].

We set $\hat{\mathbf{G}} = \mathrm{SVT}(\mathbf{H}, \frac{\gamma}{2\rho})$. Now, we want to show that $\hat{\mathbf{G}}$ obeys Eq. (C.5). Firstly, we decompose $\mathbf{H}$ as

$$\mathbf{H} = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^H + \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^H, \tag{C.7}$$

where $\mathbf{U}_0$, $\mathbf{V}_0$ (resp. $\mathbf{U}_1$, $\mathbf{V}_1$) are the left and right singular matrices associated with singular values greater than or equal to (resp. smaller than) $\frac{\gamma}{2\rho}$. By this specialized decomposition, we have

$$\hat{\mathbf{G}} = \mathbf{U}_0(\mathbf{\Sigma}_0 - \frac{\gamma}{2\rho}\mathbf{I})\mathbf{V}_0^H, \tag{C.8}$$

and therefore

$$\begin{aligned}\hat{\mathbf{G}} - \mathbf{H} &= -\frac{\gamma}{2\rho}\mathbf{U}_0\mathbf{V}_0^H - \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^H \\ &= -\frac{\gamma}{2\rho}(\mathbf{U}_0\mathbf{V}_0^H + \mathbf{W}),\end{aligned} \tag{C.9}$$

where $\mathbf{W} \triangleq \frac{2\rho}{\gamma}\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^H$. By the unitarity of $[\mathbf{U}_0, \mathbf{U}_1]$ and $[\mathbf{V}_0, \mathbf{V}_1]$, we have

$$\mathbf{U}_0^H\mathbf{W} = \mathbf{0},\ \ \mathbf{W}\mathbf{V}_0 = \mathbf{0}. \tag{C.10}$$

Moreover, by the decomposition structure of $\mathbf{H}$, diagonal elements of $\mathbf{\Sigma}_1$ are all less than $\frac{\gamma}{2\rho}$; i.e.,

$$\|\mathbf{W}\|_2 \leq 1. \tag{C.11}$$

Hence, we have

$$\mathbf{U}_0\mathbf{V}_0^H + \mathbf{W} \in \partial\|\hat{\mathbf{G}}\|_*, \tag{C.12}$$

which implies that

$$\mathbf{0} \in \mathbf{H} - \hat{\mathbf{G}} + \frac{\gamma}{2\rho}\partial\|\hat{\mathbf{G}}\|_*. \tag{C.13}$$

According to Eq. (C.5) and the strict convexity of Problem (C.1), we have

$$\hat{\mathbf{G}} = \mathbf{G}^\star = \mathrm{SVT}(\mathbf{H}, \frac{\gamma}{2\rho}), \tag{C.14}$$

and the derivation is completed.

Appendix D

Categories of Speech Sounds and Their Impact on SOSs-based BI-QSS Algorithms

Although methods devised in this thesis are applicable for many different kinds of quasi-stationary signals such as EEG signals [33], speech signals are of particular interest motivated by the cocktail party problem. Hence, it would be interesting to analyze the second-order statistics (SOSs) based BI-QSS methods under different typical categories of speech sounds. In spoken language, a speech signal is composed of three basic elements: vowels, consonants and silent pauses.

D.1 Vowels

Vowels are referred to sounds pronounced with an open vocal tract. It has been noticed that voiced vowels are typically quasi-stationary over 40–80ms time windows [34].

D.2 Consonants

In contrast to vowels, consonants are produced by severely restricting or even completely stopping the flow of air by controlling the vocal tract. Typically, the quasi-stationary assumption for consonants is only valid when the analyze window is very short (less than 20ms [34]).

To sum up, in order to capture a stationary frame, the window length should not be too long. For example, in automatic speech recognition (ASR), windows with length 20–30ms is used to capture the quasi-stationarity of speech signals [34]. For frames significantly violating the quasi-stationary assumption, those frames are corrupted. The proposed robust subspace extraction procedure (cf. Section 4.5) will automatically detect and eliminate those corrupted frames.

D.3 Silent Pauses

Simply put, silent pause means that the speakers momentarily stops their speech. In order to simplify the discussion, we consider a scenario in which two speakers are talking simultaneously and more than two microphones are recording the speech mixtures. The effect of silent pauses can be best understood by considering the local covariance model:

$$\begin{aligned}\mathbf{R}_m &= \mathbb{E}\{\mathbf{x}(t)\mathbf{x}(t)^H\}, \text{ for } t \in [(m-1)L+1, mL] \\ &= \mathbf{A}\mathbf{D}_m\mathbf{A}^H + \sigma^2\mathbf{I}, \; m = 1, \ldots, M.\end{aligned}$$

Case 1: If two speech segments are both in silent pauses, which means that there is no sources being active in this frame. In noise-free situation, $\mathbf{R}_m$ will be equal to zero. However, in the presence of noise, if there is a prior information on the existence of such frame, the noise covariance removal (Section 2.4) can be performed effectively. In essence, we can first perform a scanning, and then estimate the noise power in that particular frame.

Case 2: Suppose that only source one is active in frame m. Then, we have

$$\mathbf{R}_m = d_{m1}\mathbf{a}_1\mathbf{a}_1^H.$$

The above rank one structure of the local covariance matrix is favorable. In fact, if we know a prior that there exists two rank one frames, one with source one being active and the other frame with source two being active, the so-called local dominance condition is satisfied. By exploiting this local dominance structure of $\{\mathbf{R}_1, \ldots, \mathbf{R}_M\}$, very simple algorithm to identify the mixing matrix $\mathbf{A}$ has been devised. For detailed discussion on the local dominance and the interesting insights it brings us, we refer readers to [16].

□ **End of chapter.**

Bibliography

[1] K.-K. Lee, W.-K. Ma, Y.-L. Chiou, T.-H. Chan, and C.-Y. Chi, " Blind identification of mixtures of quasi-stationary sources using a Khatri-Rao subspace approach," *Asilomar Conf. 2011*, Pacific Grove, Nov. 6-9, 2011.

[2] W.-K. Ma, T.-H. Hsieh, and C.-Y. Chi, "DOA estimation of quasi-stationary signals with less sensors than sources and unknown spatial noise covariance: A Khatri-Rao subspace approach," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2168–2180, Apr. 2010.

[3] S. Boyd and J. Dattoro, *Alternating Projections*, EE392o: Optimization Projects, Stanford University, 2003, available online: `http://www.stanford.edu/class/ee392o/alt_proj.pdf`

[4] A. Ziehe, P. Laskov, and G. Nolte, "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation," *J. Mach. Learn. Res.*, vol. 5, pp. 801–818, 2004.

[5] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non-stationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[6] D.-T. Pham, "Joint approximate diagonalization of positive definite Hermitian matrices."

[7] R. Vollgraf and K. Obermayer, "Quadratic optimization for simultaneous matrix diagonalization," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3270 –3278, Sep. 2006.

[8] J.-F. Cardoso, and A. Souloumiac "Jacobi angles for simultaneous diagonalization," *SIAM J. Mat. Anal. Appl*, vol. 17, pp. 161–164 1996

[9] P. Tichavský, and A. Yeredor "Fast approximate joint diagonalization incorporating weight matrices," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 878 – 891, Mar. 2009.

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

[11] Y. Rong, S.-A. Vorobyov, A.-B. Gershman, and N.-D. Sidiropoulos, "Blind spatial signature estimation via time-varying user power loading and parallel factor analysis," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1697 – 1710, May. 2005.

[12] A. Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1545 –1553, Jul. 2002.

[13] X. Luciani, and L. Albera "Semi-algebraic canonical decomposition of multi-way arrays and joint eigenvalue decomposition," *ICASSP*, Prague, Czech Republic, May. 22-27, 2011.

[14] K.-N. Mokios, N.-D. Sidiropoulos, and A. Potamianos "Blind speech separation using PARAFAC analysis and integer least squares," *ICASSP*, Toulouse, France, May 14-19, 2006.

[15] F. Roemer, and M. Haardt "A closed-form solution for parallel factor (PARAFAC) analysis," *ICASSP*, Caesars Palace, Las Vegas, Nevada, USA, Mar. 30- Apr. 4, 2008.

[16] X. Fu, and W.-K. Ma, "A simple closed-form solution for overdetermined blind separation of locally sparse quasi-stationary sources," *ICASSP, accepted*, Kyoto, Japan, Mar. 25-30, 2012.

[17] J. Vía, M. Anderson, X.-L. Li, and T ulay Adali, "Joint blind source separation from second-order statistics: necessary and sufficient identifiability conditions," *ICASSP*, Prague, Czech Republic, May. 22-27, 2011.

[18] M. Cooke, J.-R. Hershey and S.-J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech Language*, vol. 24, no. 1, pp. 1-15, Elsevier Ltd, 2010.

[19] Y.-T. Yeung, T. Lee, and C.-C. Leung, "Integrating multiple observations for model-based single-microphone speech separation with conditional random fields," *ICASSP, accepted*, Kyoto, Japan, Mar. 25-30, 2012.

[20] P. Tichavský, and Z. Koldovský, " Weight adjusted tensor method for blind separation of underdetermined mixtures of nonstationary sources," *IEEE. Trans. Signal Process.*, vol. 59, no. 3, pp. 1037–1047, Mar. 2011.

[21] X. Luciani, A.-L.-.F. de Almeida, and P. Comon "Blind identification of underdetermined mixtures based on the characteristic function: the complex case," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 540–553, Feb. 2011.

[22] L.-D. Lathauwer, and J. Castaing "Blind identification of underdetermined mixtures by simultaneous matrix diagonalization," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1096–1105, Mar. 2008.

[23] D. Nion, K.-N. Mokios, N.-D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 6, pp. 1193–1207, Aug. 2010.

[24] K. Rahbar, and J.-P. Reilly "A frequency domain method for blind source separation of convolutive audio mixtures," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 832–844, Sep. 2005.

[25] L. Parra, and C. Spence "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech and Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[26] D. Peng, and Y. Xiang "Underdetermined blind source separation based on relaxed sparsity condition of sources," *IEEE. Trans. Signal Process.*, vol. 57, no. 2, pp. 809–814, Feb. 2009.

[27] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 8, pp. 434 - 444, Feb. 1997.

[28] A. Hyvärinen, J. Krhunen, and E. Oja "Independent component analysis," Wiley-Interscience, 2001.

[29] P. Comon, and C. Jutten "Handbook of blind source separation: independent component analysis and applications," Academic Press, 2010.

[30] X. Guo, S. Miron, D. Brie, S. Zhu, and X. Liao "A CANDECOMP/PARAFAC perspective on uniqueness of DOA estimation using a vector sensor array," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3475–3481, Jul. 2011.

[31] A.-J. van der Veen, and A. Leshem "Constant modulus beamforming," *Robust Adaptive beamforming*, Wiley Series in Telecommnication and Signal Processing, Chapter 6, 2005.

[32] M.-K. Tsatanis, and Z. Xu "Constrained Optimization Methods for Direct Blind Equalization" *IEEE Journal on Selected Areas in Comm.*, vol. 17, no. 3, pp. 424–433, Mar. 1999.

[33] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M.-J. Mckeown, V. Iragui, and T.-J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology, 37 (2000)*, 163-178. Cambridge University Press.

[34] V. Tyagi, H. Bourlard, and C. Wellekens "On variable-scale piecewise stationary spectral analysis of speech signals for ASR," *Speech Communication,* vol. 48, no. 9, pp. 1182–1191, Sep. 2006.

[35] L.-D. Lathauwer, B.-D. Moor, and J. Vandewalle "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.* , vol. 21, no. 4, pp. 1253–1278, Jan. 1999.

[36] G. Bergqvist, and E.-G. Larsson "The higher-order singular value decomposition: theory and an application," *IEEE Signal Process. Magazine*, vol. 27, no. 3, pp. 151–154, May 2010.

[37] N.-D. Sidiropoulos, R. Bro "On the uniqueness of multilinear decomposition of N-way arrays," *J. Chemometrics*, vol. 14, pp. 229 – 239, May. 2005.

[38] H.-E. Huang, T.-H. Chan, A. Ambikapathi, W.-K. Ma, and C.-Y. Chi, "Outlier-robust dimension reduction and its impact on hyperspectral endmember extraction," *to appear in WHISPERS,* Shanghai, China, June 5-7, 2012.

[39] M.-F. Duarte, and Y.-C. Eldar "Structured Compressed Sensing: from theory to applications," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053–4085, Sep. 2011.

[40] Z. Wen, W. Yin, and Y. Zhang "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm."

[41] Y. Shen, Z. Wen, and Y. Zhang "Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization," 2011

[42] Q. Ling, Y. Xu, W. Yin, and Z. Wen "Decentralized low-rank matrix completion," *ICASSP, accepted*, Kyoto, Japan, Mar. 25-30, 2012.

[43] M. Zibulevsky, M. Elad, "L1-L2 optimization in signal and image processing," *IEEE Signal Process. Magazine,* vol 27, no. 3, pp 76-88, May. 2010.

[44] J.-F. Cai, E.-J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. on Optimization,* vol. 20, no. 4, pp 1956-1982, Mar. 2010.

[45] M. Fazel, H. Hindi, and S. Boyd "Log-Det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices," *Proceed. American Control Conference,* no. 3, pp. 2156 – 2162, Jun. 2003.

[46] M. Fazel, H. Hindi, and S. Boyd "A rank minimization heuristic with application to minimum order system approximation," *Proceedings American Control Conference*, vol. 6, pp. 4734-4739, Jun. 2001.

[47] E.-J. Candès, X. Li, Y. Ma, and J. Wright "Robust Principal Component Analysis?," *Journal of the ACM*, vol. 58, no. 3, May 2011.

[48] J.-M. Bioucas-Dias, M.-A.-T. Figueiredo, and J.-P. Oliveira "Total variation-based image deconvolution: a majorization-minimization approach," *ICASSP* , Toulouse, France, May 14-19, 2006.

[49] D.-R. Hunter, K. Lange "A tutorial on MM algorithms," *Amer. Statist* , pp 30–37, 2004.

[50] Q. Ke, and T. Kanade "Robust subspace computation using L1 norm," Aug. 2003.

[51] M. Rajin, P. Comon, and R.-A. Harshman, "Enhanced line search: a novel method to accelerate PARAFAC," SIAM J. Matrix Analy. Appl., vol. 30, no. 3, Sep. 2008.

[52] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning.*, vol. 3, no. 1, pp. 1–122, 2011.

[53] K.-B. Petersen, and M.-S. Pedersen "The Matrix Cookbook," MIT, Feb. 2006.

[54] R.-A. Horn, and C.-R. Johnson "Matrix Analysis," Cambridge University Press, 1990.

[55] P.-J. Huber "Robust regression: asymptotics, conjectures and Monte Carlo," *Ann. Statist.*, vol. 1, no. 5, pp 799 – 821 1973.

[56] T. G. Kolda, and B. W. Bader "Tensor decompositions and applications," *SIAM Review*, June, 2008.

[57] M. Wax, and J. Sheinvald "A least-squares approach to joint diagonalization," *IEEE Signal Processing Letters*, vol. 4, no. 2, Feb. 1999.

[58] N.-D. Sidiropoulos, and X. Liu, "Identifiability results for blind beamforming in incoherent multipath with small delay spread," *IEEE Trans. Signal Processing*, vol. 49, no. 1, pp. 228 – 236, Jan. 2001.

[59] S. Boyd, and L. Vandenberghe "Convex Optimization," Cambridge University Press, 2004.

[60] T.-M. Cover, and J.-A. Thomas, "Elements of Information Theory," Wiley-Interscience, 1991.

[61] S. Boyd, *Subgradients*, EE364bo: Lecture Slides and Notes, Stanford University, 2010, available online: `http://www.stanford.edu/class/ee364b/lectures/subgradients_slides.pdf`

[62] M. Grant, and S. Boyd "CVX: Matlab software for disciplined convex programming, version 1.21," Apr. 2011.

[63] M. Grant, and S. Boyd "Graph implementations for nonsmooth convex programs," *Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences*, Springer-Verlag Limited, page 95–110, 2008.

[64] B. Chen, S. He, Z. Li, and S. Zhang "Maximum block improvement and polynomial optimization".