A Convergence and Asymptotic Analysis of Nonlinear Separation Model

Lu Wang* and Tomoaki Ohtsuki†
* Graduate School of Science and Technology, Keio University, Yokohama, Japan
† Department of Information and Computer Science, Keio University, Yokohama, Japan
E-mail: wanglu@ohtsuki.ics.keio.ac.jp, ohtsuki@ics.keio.ac.jp

***Abstract*—Nonlinear blind source separation is the process of estimating either the original signals or mixture functions from the degraded signals, without any prior information about original sources. The key idea is to recover the sources by estimating an approximation function so as to approximate the inverse of mixing function. However, in practice, the approximation function is derived from some estimation algorithm with finite sample size, which leads to the performance loss. In this paper, we work on the convergence and asymptotic analysis of the separation approach, which uses the flexible approximation to extract the nonlinearity of mixture function so that to make the problem linearly separable. The analysis stems from the performance of a mismatched estimator that accesses the finite sample size. By providing a closed-form expression of normalized mean squared error (NMSE), we can present a novel algebraic formalization that leads to the upper bound on the estimation error. The simulation results show that if the flexible approximation can extract the nonlinearity of mixing functions, the minimized NMSE can be achieved as the sample size tends to be infinity. This implies that the algorithm is feasible to separate the distortion of the nonlinear mixture.**

I. Introduction

The purpose of independent component analysis (ICA) and blind source separation (BSS) [1], [2], [3], is to extract m mutually independent elements from n observed mixtures. Let us consider the following linear instantaneous mixing system with m inputs and n outputs as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \quad t = 1, 2, \cdots, T, \tag{1}$$

where $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_m(t)]^\top$ are the signals with m channels, $s_i(t)$ denotes the sample of the i-th source at time index t. $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_n(t)]^\top$ denotes the observed mixtures with n channels, which is assumed to be generated by $n \times m$ mixing matrix $\mathbf{A}$ and source signals $\mathbf{s}(t)$.

Commonly, the separation process of ICA is conducted on the assumption that the sources are statistical independent. For a linear mixing model, if the number of sources equals to the number of channels ($m = n$), the demixng matrix $\mathbf{W}$ can be defined as $\mathbf{W} = \mathbf{A}^{-1}$. The recovered signals are represented as $\hat{\mathbf{s}}(t) = \mathbf{W}\mathbf{x}(t)$. The linear BSS aims at estimating $\mathbf{W}$ and recovered signals $\hat{\mathbf{s}}(t)$ using only the observed signals $\mathbf{x}(t)$.

An obvious extension for the task of BSS is that the observed signals are assumed to be generated from a set of sources by a nonlinear, instantaneous and invertible function $\mathcal{F}$, i.e., $\mathbf{x}(t) = \mathcal{F}(\mathbf{s}(t))$ for all $t = 1, \cdots, T$. Roughly, the nonlinear blind source separation seeks to find the mixing function (or its inverse function $\mathcal{G} = \mathcal{F}^{-1}$), solely based on the assumption that the sources are statistically independent. However, the indeterminacies imposed by the nonlinear model are difficult to handle [4], [5]. The obstacle for the nonlinear BSS problem is that solutions are non-unique without extra constraints [6]. The recovery inconsistency has been tackled by adding further prior information directly in the model or as a regularization term in the optimization processing procedure.

Most nonlinear algorithms utilize single approximation to extract the nonlinearity, such as the multi-layer perceptron (MLP) in the neural network [7], [8], which is employed for estimating the nonlinear separation function. By restricting the smoothness of the target transforming, MLP provides the regularized solutions to ensure that nonlinear ICA leads to the sources separable. However, the example presented in [9] shows that the smoothness property is not a sufficient condition for this purpose. Hyvärinen and Pajunen [6] show conformal mapping may helpful. Nonlinear ICA is able to estimate a separation mapping up to the rotation when the mapping functions are restricted to the set of conformal mapping. Unfortunately, the angle preservation conditions seem very restrictive [10]. In particular, it is not realistic in the framework of the nonlinear mappings associated to the nonlinear sensor array.

A novel approach named as Vanishing Ideal based Non-Linear SEparation Model (ViNLisem) was proposed in [11], which relies on a novel mathematical construction with multi-layer architecture. By considering a situation where a set of flexible approximations are utilized to extract the nonlinearity, the approach breaks a nonlinear distortion down into the version of the linear case in the feature space.

Nevertheless, the approximation function are generated adaptively depend solely on the input data, then the function and its empirical counterpart that is assumed to be derived by the estimation algorithm with the finite sample size could differ, which is said to be mismatched or misspecified. Experience with real data often exposes the limitations of any assumed model, since modeling errors at some level are always present. Therefore, understanding the possible performance loss that the separation algorithm subject to model misspecification is of practical interest and critical. In this paper, we work on the convergence and asymptotic analysis of an approximation function, so that propose an analytical expression of performance measure.

A. Our Contribution

This paper provides a theoretical analysis to ViNLisem algorithm [11], which includes the closed-form expressions on normalized mean squared error (NMSE), as well as proposing a new algebraic formalization that leads to the upper bound on the performance loss. The analysis stems from the performance of a mismatched estimator that accesses the finite sample size, which is explored by two parts. One is to derive an iterative expression of the coefficient matrix. Another part aims to establish a closed-form expression of discrepancy between the truth model and its counterpart.

Using Maximum Likelihood (ML) estimation, the coefficient matrix is modeled as deterministic but depend on the data. By maximizing the likelihood function, the convergence point of the maximum likelihood estimator could be interpreted as the stationary point that minimizes Kullback-Leibler (KL) divergence between the truth model and the approximated expression. Then, the natural gradient of likelihood function is utilized to obtain an optimal solution, which is propagated to yield the component of NMSE estimation.

Then, we establish a closed-form expression of discrepancy between the truth covariance and its counterpart. A major focus is on the derivation of the estimation of covariance matrices, which can be treated exactly or approximately as an estimation of a finite or infinite sample size. Thus, the spectral norm utilized to obtain the upper bound under a range of matrix operator norm and divergence losses, as well as solving the non-parametric function mis-specification problems with finite sample size.

The rest of the paper is organized as follows. In Section 2, we introduce the separation model, which is denoted as ViNLisem. Then the problem formulation is given mathematically. An iterative expression of the coefficient matrix is presented in Section 3. Section 4 aims to establish a close-form expression of discrepancy between the truth covariance and its counterpart. Some simulation experiments are carried out to corroborate the theoretical results in Section 5. We conclude the paper in Section 6.

II. Model and Problem Formulation

The nonlinear BSS problem is formally described as follows. The observed mixture $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_n(t)]^\top$ is assumed to be generated from a set of statistically independent sources $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_n(t)]^\top$ by a nonlinear, instantaneous and invertible function as

$$\mathbf{x}(t) = \mathcal{F}(\mathbf{s}(t)), \quad t = 1, \cdots, T, \tag{2}$$

where $\{\cdot\}^\top$ denotes the transpose, and t is the sample (time) index. This process can be described on the left-hand side of Fig. 1, which is denoted as a mixing-separating system.

However, without any extra constraints for mixing function, the solutions are non-unique [6]. The approach in [11] was proposed to tackle the ill-posedness with a few assumptions. By utilizing a flexible approximation to extract the nonlinearity, the distortion of mixing functions can be transformed into the version of the linear case in the feature space. This process described on the right-hand side of Fig. 1, which is denoted as a demixing system.

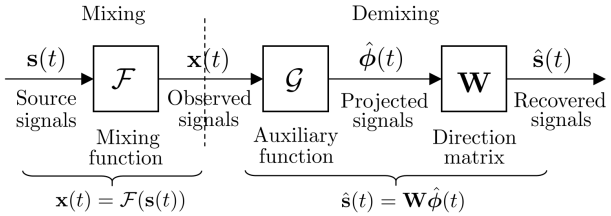


Fig. 1: The mixing-separating system of the nonlinear blind source separation. The block $\mathcal{F}$ are generic nonlinear functions that lead to a mixture process. The observed signals are $\mathbf{x}(t)$, which are assumed to be generated from source signals by a nonlinear mixing function. The $\mathcal{G}$ block in the demixing process, implementing a flexible approximation, as the auxiliary function is used to extract the nonlinearity of mixing functions. Thus, the projected signals $\hat{\boldsymbol{\phi}}(t)$ can make the problem linearly separable. The block $\mathbf{W}$ is a coefficient matrix, performing a linear operator that derive the estimator of original signals from the projected signals.

Given a set of auxiliary functions that allowed us to construct the nonlinear variants by some vanishing polynomials, such as $g_1(\mathbf{x}(t)), \cdots, g_k(\mathbf{x}(t)) \in \mathcal{G}$, where $g_i(\mathbf{x}(t))$ is i-th vanishing polynomial that the observed signals $\mathbf{x}(t)$ are mapped implicitly into some feature space $\boldsymbol{\Phi} : \mathbb{R}^n \to \mathbb{R}^k$, i.e., $\hat{\phi}_i(\mathbf{x}(t))$ represents the projected value from polynomial $\hat{\phi}_i(\mathbf{x}(t)) = g_i(\mathbf{x}(t))$. The feature space is spanned from such polynomial functions that enable us to work on $\mathcal{G}$. Thus, the projected data in the feature space lead to a linear combination for the demixing process

$$\hat{\mathbf{s}}_j(t) = \sum_i W_{ji} \hat{\phi}_i(\mathbf{x}(t)), \tag{3}$$

where W_{ji} denotes the (j,i)-th element of the coefficient matrix $\mathbf{W}$. $\hat{\phi}_i(\mathbf{x}(t))$ is the model assumed to derive the estimation under the finite samples size, denote as $\hat{\phi}_i(t)$ for short.

In this paper, we work on a theoretical analysis of the proposed separation model [11] as described in Fig. 1, so that measure the quantities of the recovered signals. In other words, the problem consists in estimating $\hat{\mathbf{s}}(t)$ to given a closed-form expression for normalized mean squared error (NMSE), as well as proposing a new algebraic formalization that leads to the upper bound.

First, normalized mean squared error (NMSE) as the figure of merit is used to recovered signals $\hat{\mathbf{s}}(t)$ by

$$\begin{aligned} \widehat{\text{NMSE}} &\triangleq \frac{1}{T} \sum_{t=1}^{T} \|\hat{\mathbf{s}}(t) - \mathbf{s}(t)\|_F^2 \\ &= \frac{1}{T} \sum_{t=1}^{T} \left\| \mathbf{W}\hat{\boldsymbol{\phi}}(t) - \mathbf{W}\boldsymbol{\phi}(t) \right\|_F^2, \end{aligned} \tag{4}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm, i.e. $\|\mathbf{A}\|_F^2 = \mathrm{tr}\{\mathbf{A}\mathbf{A}^\top\}$.

Since we assume that the coefficient matrix $\mathbf{W}$ is fixed, which only depend on the observed mixture. The theoretical analysis only considers the discrepancy from such polynomials $\hat{\boldsymbol{\phi}}(t)$ that derived from the finite data points. We thus begin by defining a convenient error term. In order to focus on well-defined quantities, we consider the error

$$\boldsymbol{\delta}_\phi(t) = \hat{\boldsymbol{\phi}}(t) - \boldsymbol{\phi}(t), \tag{5}$$

where $\boldsymbol{\phi}(t)$ is true data model and $\hat{\boldsymbol{\phi}}(t)$ is the model assumed to derive the estimation under the finite samples size. Thus NMSE can be rewritten as

$$\begin{aligned}\widehat{\mathrm{NMSE}} &= \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{W}\boldsymbol{\delta}_\phi(t)\|_F^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathrm{tr}\left\{\mathbf{W}\boldsymbol{\delta}_\phi(t)\boldsymbol{\delta}_\phi(t)^\top\mathbf{W}^\top\right\} \\ &= \mathrm{tr}\left\{\mathbf{W}\bar{\boldsymbol{\Sigma}}_{\delta_\phi}\mathbf{W}^\top\right\}. \end{aligned} \tag{6}$$

The second equality used the definition of Frobenius norm. In the third equality, the results derived from the definition of empirical counterpart $\bar{\boldsymbol{\Sigma}}_{\delta_\phi} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\delta}_\phi(t)\boldsymbol{\delta}_\phi(t)^\top$.

Without loss of generality, we assume that the original source $\mathbf{s}(t)$ is independent of $\mathbf{s}(t')$ for all $t \neq t'$. For any t, $\mathbf{s}(t)$ has the zero vector mean and the covariance matrix is $\boldsymbol{\Sigma}_s$. For projected signals $\boldsymbol{\phi}(t)$, the linear model (3) implies the relation that

$$\boldsymbol{\Sigma}_s = \mathbf{W}\boldsymbol{\Sigma}_\phi\mathbf{W}^\top, \tag{7}$$

where $\boldsymbol{\Sigma}_\phi$ is the covariance matrix of $\boldsymbol{\phi}(t)$. The corresponding empirical counterpart is defined as

$$\bar{\boldsymbol{\Sigma}}_s \triangleq \frac{1}{T}\sum_{t=1}^{T}\mathbf{s}(t)\mathbf{s}(t)^\top. \tag{8}$$

The objective of BSS is to recover $\mathbf{s}(t)$ from the mixture $\mathbf{x}(t)$ so that its correlation satisfies $\boldsymbol{\Sigma}_{s_i,s_j} \triangleq \mathbb{E}[\bar{\boldsymbol{\Sigma}}_{s_i,s_j}] = \mathbf{0}_{n\times n}$ for $i \neq j$. However, due to the finite sample size, it does not hold for its empirical counterpart, i.e. $\bar{\boldsymbol{\Sigma}}_{s_i,s_j} \neq \mathbf{0}_{n\times n}$. In this paper, the coefficient matrix $\mathbf{W}$ is assumed to be given, thus the linear model (3) implies the relation of covariance matrices $\boldsymbol{\Sigma}_\phi = \mathbf{W}^-\boldsymbol{\Sigma}_s\mathbf{W}^{-\top}$ and their empirical counterpart $\bar{\boldsymbol{\Sigma}}_\phi = \mathbf{W}^-\bar{\boldsymbol{\Sigma}}_s\mathbf{W}^{-\top}$. The notation is denoted as $\mathbf{W}^- = \mathbf{W}^{-1}$ for simple expression using the following content.

In practical, only the samples of finite size is available that lead to NMSE of (6) under an empirical counterpart. The corresponding representation under the infinite size can be obtained by approximating the mathematical expectations with the sample means, which is given by

$$\begin{aligned}\mathrm{NMSE} &\triangleq \mathbb{E}\{\widehat{\mathrm{NMSE}}\} \\ &= \mathrm{tr}\left\{\boldsymbol{\Sigma}_{\delta_\phi}\mathrm{Cov}(\mathbf{W}^\top)\right\} + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \end{aligned} \tag{9}$$

The detail derivation can be found in Appendix A. The objective of this paper is to access the performance of the mismatched estimator, as well as proposing a formalization to the performance loss. This can be formulated by a contrast function.

Problem 1. Given a set of projected data $\boldsymbol{\phi}(t)$ that is a fixed point from the theoretical model. If its empirical counterpart has an almost surely foxed point $\hat{\boldsymbol{\phi}}(t)$ that is a neighborhood of $\boldsymbol{\phi}(t)$. Then the problem is to learn an algebraic formalization that leads to the upper bound of the following equation

$$\left\|\mathrm{NMSE} - \widehat{\mathrm{NMSE}}\right\|^2 = \left\|\mathrm{tr}\left\{(\boldsymbol{\Sigma}_{\delta_\phi} - \bar{\boldsymbol{\Sigma}}_{\delta_\phi})\mathrm{Cov}(\mathbf{W}^\top)\right\}\right\|^2, \tag{10}$$

where $\boldsymbol{\Sigma}_{\delta_\phi}$ is covariance matrix of δ_ϕ and $\bar{\boldsymbol{\Sigma}}_{\delta_\phi}$ is the corresponding counterpart. □

Problem 1 implies that if we present a closed-from expression on NMSE, the performance loss of recovered sources can be minimized approximately by reducing the discrepancy between the $\boldsymbol{\Sigma}_{\delta_\phi}$ and its empirical counterpart. In other words, $\hat{\boldsymbol{\phi}}(t)$ is expected to extract the nonlinearity of the mixing function so as to the NMSE can be minimized as the sample size tends to be infinity.

The performance analysis of (10) can be concluded from the derivation of two parts. One is to derive an iterative expression of the coefficient matrix $\mathbf{W}$ that to be estimated are modeled as deterministic but depend on the data. The details derivation are described in Section 3. Another part showed in Section 4 aims to establish a closed-form expression of discrepancy between the truth model and its counterpart under the finite sample size.

III. Estimate Coefficient Matrix $\mathbf{W}$

We now turn to estimate coefficient matrix $\mathbf{W}$ of (10), which lead to an iterative expression of coefficient matrix $\mathbf{W}$. The analysis first focus is on the result of the behavior of the iterative solution of ML estimation.

A. Maximum-Likelihood (ML) Estimation

Given the source signals $\{\mathbf{s}(t)\}_{t=1}^T$ that are assumed to be drawn independently from a multivariate Gaussian distribution. We thus can estimate the nuisance parameters by ML estimation. The log-likelihood function is given by

$$\begin{aligned}\log p\left(\{\mathbf{s}(t)\}_{t=1}^T\right) &= \log\prod_{t=1}^{T}p(\mathbf{s}(t)) \\ &= -\frac{T}{2}\log\det(\boldsymbol{\Sigma}_s) - \frac{1}{2}\sum_{t=1}^{T}\mathbf{s}(t)^\top\boldsymbol{\Sigma}_s^-\mathbf{s}(t) - \frac{nT}{2}\log 2\pi. \end{aligned} \tag{11}$$

where $\det(\boldsymbol{\Sigma}_s)$ indicates the determinant of $\boldsymbol{\Sigma}_s$. The first equality comes from the assumption of independence of sources $\mathbf{s}(t)$ and $\mathbf{s}(t')$ for $t \neq t'$. The second equality follows the Gaussian distribution with the zero vector mean and the

covariance matrix is $\mathbf{\Sigma}_s$. The above equation can be rewritten by using trace trick

$$\begin{aligned}
&\log p\left(\{\mathbf{s}(t)\}_{t=1}^T\right) \\
&= -\frac{T}{2}\log\det(\mathbf{\Sigma}_s) - \frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}\left\{\mathbf{\Sigma}_s^{-}\mathbf{s}(t)\mathbf{s}(t)^\top\right\} - \frac{nT}{2}\log 2\pi \\
&= -\frac{T}{2}\log\det(\mathbf{\Sigma}_s) - \frac{T}{2}\mathrm{tr}\left\{\bar{\mathbf{\Sigma}}_s\mathbf{\Sigma}_s^{-}\right\} - \frac{nT}{2}\log 2\pi \\
&= -\frac{T}{2}\log\frac{\det(\mathbf{\Sigma}_s)}{\det(\bar{\mathbf{\Sigma}}_s)} - \frac{T}{2}\mathrm{tr}\left\{\bar{\mathbf{\Sigma}}_s\mathbf{\Sigma}_s^{-}\right\} - \kappa_1, \qquad (12)
\end{aligned}$$

where $\kappa_1 = -\frac{T}{2}\log\det(\bar{\mathbf{\Sigma}}_s) - \frac{nT}{2}\log 2\pi$ denotes the term, which is irrelevant to the maximization of the likelihood with respect to its parameter. The results of the first equality are from the property $\mathbf{a}^\top\mathbf{\Sigma}\mathbf{a} = \mathrm{tr}\{\mathbf{\Sigma}\mathbf{a}\mathbf{a}^\top\}$ for any vector $\mathbf{a}$ and matrix $\mathbf{\Sigma}$ with appropriate dimensions. Then, using the definition of an empirical counterpart in (8) and the property of $\mathrm{tr}\{\mathbf{AB}\} = \mathrm{tr}\{\mathbf{BA}\}$, we have the second equality. To obtain a similar form with Kullback-Leibler divergence in Definition 1, the third equality is a derivation of a simple operation.

Definition 1. Let $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ be two $n \times n$ positive definite matrices. The Kullback-Leibler divergence $\mathcal{KL}(\mathbf{\Sigma}_1\|\mathbf{\Sigma}_2)$ measures the difference between two multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma}_1)$ and $\mathcal{N}(0, \mathbf{\Sigma}_2)$, which is given by

$$\mathcal{KL}(\mathbf{\Sigma}_1\|\mathbf{\Sigma}_2) = \frac{1}{2}\left(\mathrm{tr}\{\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-}\} - \log\frac{\det(\mathbf{\Sigma}_1)}{\det(\mathbf{\Sigma}_2)} - n\right). \quad (13)$$

We set $\mathbf{\Sigma}_1 = \bar{\mathbf{\Sigma}}_s$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_s$. Using the definition of Kullback-Leibler divergence as a measure to two matrices, then the log-likelihood of (12) can be rewritten as

$$\begin{aligned}
\log p\left(\{\boldsymbol{s}(t)\}_{t=1}^T\right) &= -T\mathcal{KL}(\bar{\mathbf{\Sigma}}_s\|\boldsymbol{R}_s) - \frac{nT}{2} - \kappa_1 \\
&= -T\mathcal{KL}(\mathbf{W}\bar{\mathbf{\Sigma}}_\phi\mathbf{W}^\top\|\mathbf{\Sigma}_s) + \kappa_2, \qquad (14)
\end{aligned}$$

where $\kappa_2 = -\frac{T}{2}\log\det(\bar{\mathbf{\Sigma}}_s) - \frac{nT}{2}\log 2\pi - \frac{nT}{2}$ denotes the term which is irrelevant to the maximization of the likelihood with respect to its parameters. The stationary point of Kullback-Leibler divergence in (14) leads to the way for estimating the covariance matrix $\mathbf{\Sigma}_s$.

B. Estimate $\mathbf{\Sigma}_s$ for Fixed $\mathbf{W}$

To analyze $\bar{\mathbf{\Sigma}}_\phi$, we fix the coefficient matrix $\mathbf{W}$ first. Thus, maximizing log-likelihood (12) is equivalent to minimizing the Kullback-Leibler, which is given by

$$\max_{\mathbf{\Sigma}_s}\log p\left(\{\mathbf{s}(t)\}_{t=1}^T|\mathbf{\Sigma}_s, \mathbf{W}\right) = \min_{\mathbf{\Sigma}_s}\mathcal{KL}(\mathbf{W}\bar{\mathbf{\Sigma}}_\phi\mathbf{W}^\top\|\mathbf{\Sigma}_s). \quad (15)$$

To derive an estimator $\bar{\mathbf{\Sigma}}_\phi$ for fixed $\mathbf{W}$, the definition of Kullback-Leibler in (13) can be rewritten in the form as

$$\begin{aligned}
\mathcal{KL}(\mathbf{\Sigma}_1\|\mathbf{\Sigma}_2) &= \frac{1}{2}\mathrm{tr}\{\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-} - \mathbf{I}\} - \frac{1}{2}\log\det(\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-}) \\
&= \frac{1}{2}\mathrm{tr}\{\mathbf{\Sigma}_2^{-\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-\frac{1}{2}} - \mathbf{I}\} - \frac{1}{2}\log\det(\mathbf{\Sigma}_2^{-\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-\frac{1}{2}}) \\
&= \frac{1}{2}\mathrm{tr}\{\boldsymbol{\mathcal{R}} - \mathbf{I}\} - \frac{1}{2}\log\det(\boldsymbol{\mathcal{R}}). \qquad (16)
\end{aligned}$$

We note that the fact, for a positive-definite matrix $\boldsymbol{\mathcal{R}}$, $\mathrm{tr}(\boldsymbol{\mathcal{R}} - \mathbf{I})$ is an upper bound for $\log\det\boldsymbol{\mathcal{R}}$, which is attained if and only if $\boldsymbol{\mathcal{R}}$ is the identity. This follows immediately from the inequality $\log x \le x - 1$ which is valid for all $x > 0$. Thus, the equality has the minimization value if and only if $x = 1$.

Therefore, minimizing the right-hand side of (15) can be understood as diagonalization of covariance matrix $\bar{\mathbf{\Sigma}}_\phi$ by matrix $\mathbf{W}$. Since we assume the source signals $\mathbf{s}(t)$ are independent of $\mathbf{s}(t')$ for any $t \neq t'$, then we have the covariance matrix $\mathbf{\Sigma}_{s_i,s_j} = \mathbf{0}_{n\times n}$ for $i \neq j$. That is, $\mathcal{KL}(\mathbf{W}\bar{\mathbf{\Sigma}}_\phi\mathbf{W}^\top\|\mathbf{\Sigma}_s) \ge 0$ is satisfied with equality if and only if $\mathbf{\Sigma}_s^{ML} = \mathrm{diag}\{\mathbf{W}\bar{\mathbf{\Sigma}}_\phi\mathbf{W}^\top\}$, where $\mathrm{diag}\{\cdot\}$ is the operator of the diagonalization. Then (15) takes the form

$$\begin{aligned}
&\max_{\mathbf{\Sigma}_s}\log p(\{\mathbf{s}(t)\}_{t=1}^T|\mathbf{\Sigma}_s, \mathbf{W}) \\
&\qquad = -T\mathcal{KL}(\mathrm{diag}\{\mathbf{W}\bar{\mathbf{\Sigma}}_\phi\mathbf{W}^\top\}\|\mathbf{\Sigma}_s) + \kappa_2, \qquad (17)
\end{aligned}$$

where $\kappa_2 = -\frac{T}{2}\log\det(\bar{\mathbf{\Sigma}}_s) - \frac{nT}{2}\log 2\pi - \frac{nT}{2}$ is an irrelevant term with respect to the parameter. $\bar{\mathbf{\Sigma}}_s$ is empirical counterpart that can be determined by the data.

C. Estimate $\mathbf{W}$ for Fixed $\mathbf{\Sigma}_s$

Now let us consider to estimate the coefficient matrix $\mathbf{W}$ by characterizing the optimal solution of the log-likelihood (17). For this purpose, we calculate the derivative of the likelihood function $\mathcal{J}(\mathbf{W}, \mathbf{\Sigma}_s) = \log p(\{\mathbf{s}(t)\}_{t=1}^T|\mathbf{\Sigma}_s, \mathbf{W})$ with respect to $\mathbf{W}$ for fixed $\mathbf{\Sigma}_s$. Assume that we are at the point $\mathbf{W}$ and wish to find a direction for a small matrix increment $\delta\mathbf{W}$ such that the value in

$$\mathcal{J}(\mathbf{W} + \delta\mathbf{W}, \mathbf{\Sigma}_s) = \mathcal{KL}(\mathbf{W}\bar{\mathbf{\Sigma}}_\phi\mathbf{W}^\top\|\mathbf{\Sigma}_s)\big|_{\mathbf{W}=\mathbf{W}+\delta\mathbf{W}}, \quad (18)$$

is minimized under the constraint that the squared norm $\|\delta\mathbf{W}\|^2$ is constant. This is a natural requirement, as any step in a gradient algorithm for minimization must consist of the direction of the step and the length. Keeping the length constant, we search for the optimal direction.

Let us require that the displacement $\delta\mathbf{W}$ is always proportional to $\mathbf{W}$ itself, $\delta\mathbf{W} = \mathbf{D}\mathbf{W}$. We thus have the Taylor series of $\mathcal{J}(\mathbf{W} + \mathbf{D}\mathbf{W})$ that be expressed by

$$\begin{aligned}
\mathcal{J}(\mathbf{W}(\mathbf{I} + \mathbf{D}), \mathbf{\Sigma}_s) &= \mathcal{J}(\mathbf{W}, \mathbf{\Sigma}_s) \\
&\quad + \mathrm{tr}\left\{(\nabla\mathcal{J}(\mathbf{W}, \mathbf{\Sigma}_s))^\top\mathbf{D}\right\} + \mathcal{O}(\mathbf{D}\mathbf{W}) \qquad (19)
\end{aligned}$$

The multiplier for $\mathbf{D}$ or matrix $\nabla\mathcal{J}(\mathbf{W}, \mathbf{\Sigma}_s)$ is called the natural gradient. It is the usual matrix gradient multiplied by $\mathbf{W}^\top$.

The largest decrement in the value of $\mathcal{J}(\mathbf{W}(\mathbf{I} + \mathbf{D}) - \mathcal{J}(\mathbf{W})$ is now obviously obtained when the term $\mathrm{tr}\{\nabla\mathcal{J}(\mathbf{W}, \mathbf{\Sigma}_s)\mathbf{W}^\top)^\top\mathbf{D}\}$ is minimized, which happens when $\mathbf{D}$ is proportional to $-\nabla\mathcal{J}(\mathbf{W}, \mathbf{\Sigma}_s)\mathbf{W}^\top$. The natural gradient algorithm has the form

$$\mathbf{W} \leftarrow \mathbf{W} - \nabla\mathcal{J}(\mathbf{W}, \mathbf{\Sigma}_s)\mathbf{W}^\top\mathbf{W}, \qquad (20)$$

where the symbol $\leftarrow$ means substitution, i.e., the value of the right-hand side is computed and substituted on the left-hand

Algorithm 1 Estimate coefficient matrix $\mathbf{W}$ using iterative procedure.

Initialization: Choose initial estimations $\mathbf{W}^{(0)}$ and $\mathbf{\Sigma}_s^{(0)} = \text{diag}\{\mathbf{W}^{(0)}\bar{\mathbf{\Sigma}}_\phi(\mathbf{W}^{(0)})^\top\}$.

for $t = 1$ **do**
  Calculate $\mathbf{\Sigma}_s$ from $\mathbf{\Sigma}_s^{(t)} = \text{diag}\{\mathbf{W}^{(t-1)}\bar{\mathbf{\Sigma}}_\phi(\mathbf{W}^{(t-1)})^\top\}$
  Update $\mathbf{W}$ using $\mathbf{W}^{*(t)} = \left[2\mathbf{I} - \frac{1}{2}(\mathbf{\Sigma}_s^{(t-1)})^-(\bar{\mathbf{\Sigma}}_s^\top + \bar{\mathbf{\Sigma}}_s)\right]\mathbf{W}^{(t-1)}$
  $\mathbf{W}^{(t)} = \frac{\mathbf{W}^{*(t)}}{\|\mathbf{W}^{*(t)}\|}$
  Check for convergence
  if $\|\mathbf{W}^{(t)} - \mathbf{W}^{(t-1)}\| \le \epsilon$ **then**
    Break
  else
    $t = t + 1$
  end if
end for

side. The natural gradient update of the coefficient matrix is given by

$$-\nabla\mathcal{J}(\mathbf{W}, \mathbf{\Sigma}_s)\mathbf{W}^\top\mathbf{W} = \left[\mathbf{I} - \frac{1}{2}\mathbf{\Sigma}_s^-(\bar{\mathbf{\Sigma}}_s^\top + \bar{\mathbf{\Sigma}}_s)\right]\mathbf{W}. \quad (21)$$

Additionally, using natural gradient updates have a faster convergence. The iterative procedure to obtain the coefficient matrix $\mathbf{W}$ can be described in algorithm 1.

IV. The Estimation for Covariance Matrix

In this section, we establish a closed-form expression of discrepancy between the covariance matrix $\mathbf{\Sigma}_{\delta_\phi}$ and its counterpart under the finite sample size.

Theorem 1. Let $\bar{\mathbf{\Sigma}}_{\delta_\phi}$ be an estimator of the $k \times k$ covariance matrix $\mathbf{\Sigma}_{\delta_\phi}$ on the finite sample size. For the constants $C_1, C_2 > 0$, we have

$$\mathbb{E}\left\|\bar{\mathbf{\Sigma}}_{\delta_\phi} - \mathbf{\Sigma}_{\delta_\phi}\right\|^2 \le C_2\frac{m + \log k}{T} + C_1^2\left(\frac{p}{2}\right)^{-2\alpha}, \quad (22)$$

where m is the range of non-zero elements in the blocks and parameter α specifies the rate of decay for the elements of covariance as they move away from the diagonal. p is an arbitrary integer with $1 \le p \le k$. □

The above problem can be analyzed by separating as bias part and variance part in terms of

$$\mathbb{E}\|\bar{\mathbf{\Sigma}}_{\delta_\phi} - \mathbf{\Sigma}_{\delta_\phi}\|^2 \le \underbrace{\|\mathbb{E}[\bar{\mathbf{\Sigma}}_{\delta_\phi}] - \mathbf{\Sigma}_{\delta_\phi}\|^2}_{bias} + \underbrace{\mathbb{E}\|\bar{\mathbf{\Sigma}}_{\delta_\phi} - \mathbb{E}[\bar{\mathbf{\Sigma}}_{\delta_\phi}]\|^2}_{variance}. \quad (23)$$

In the remainder of this section, we shall derive approximate expressions for the bias and variance respectively, and use these to investigate how the NMSE will behave.

A. The Analysis of Bias Part

We first prove the risk upper bound for the bias part. The derivation of the procedure is inspired by the idea of convergence bound under the spectral norm.

The bias part of the analysis is almost identical to spectral norm and matrix of bounded that is almost identical to [12].

Definition 2. Let $\rho(\mathbf{A})$ be a spectral radius of $\mathbf{A}$. If $\mathbf{A} \in \mathbb{R}^{k\times k}$ is a symmetric matrix with eigenvalues $\lambda_1, \cdots, \lambda_k$, then $\|\mathbf{A}\|_2 = \rho(\mathbf{A})$ has a definition as

$$\|\mathbf{A}\|_2 = \rho(\mathbf{A}) := \max_{1\le i\le k}\{|\lambda_i|\}. \quad (24)$$

Definition 3. For any matrix $\mathbf{A} \in \mathbb{R}^{k\times k}$ be a square matrix of size $k \times k$. The matrix norm $\|\mathbf{A}\|$ is defined by $\|\mathbf{A}\| := \max_{1\le i\le k}\sum_{j=1}^k |a_{ij}|$.

Theorem 2. *For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{k\times k}$. Its $\|\mathbf{A}\|_2$ is bounded by matrix norm in the terms of*

$$\|\mathbf{A}\|_2 \le \|\mathbf{A}\| = \max_{1\le i\le k}\sum_{j=1}^k |a_{ij}|. \quad (25)$$

Proof: Assume λ_i is an arbitrary eigenvalue of matrix $\mathbf{A}$ and $\mathbf{v}_i$ is its corresponding eigenvector. Then we have

$$|\lambda_i|\|\mathbf{v}_i\| = \|\lambda_i\mathbf{v}_i\| = \|\mathbf{A}\mathbf{v}_i\| \le \|\mathbf{A}\|\|\mathbf{v}_i\|. \quad (26)$$

Since $\mathbf{v}_i$ is a non-zero vector $\|\mathbf{v}_i\| \ne 0$, then $|\lambda_i| \le \|\mathbf{A}\|$. Since λ_i is an arbitrary eigenvalue, then we have $\rho(\mathbf{A}) \le \|\mathbf{A}\|_1$. ■

This result is considered to be used for a convergence of the bias part in (23), $\mathbb{E}[\bar{\mathbf{\Sigma}}_{\delta_\phi}] - \mathbf{\Sigma}_{\delta_\phi}$ by its l_1 norm.

Considering infinity sample size, the convariance matrix $\mathbf{\Sigma}_{\delta_\phi}$ is defined by

$$\mathbf{\Sigma}_{\delta_\phi} = (\sigma_{ij})_{1\le i,j\le k} = \mathbb{E}\left[\boldsymbol{\delta}_\phi(t)\boldsymbol{\delta}_\phi(t)^\top\right], \quad (27)$$

where σ_{ij} is the element of the convariance matrix. Since the data points are finite in practice, the definition of empirical counterpart using the estimator is given by

$$\bar{\mathbf{\Sigma}}_{\delta_\phi} = (\omega_{ij}\sigma_{ij})_{k\times k} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\delta}_\phi(t)\boldsymbol{\delta_\phi}(t)^\top, \quad (28)$$

where the ω_{ij} is weight. Without loss of generality, the weight ω_{ij} can be defined as

$$\omega_{ij} = \begin{cases} 1, & \text{when } |i-j| < \frac{p}{2}, \\ 2 - \frac{2|i-j|}{p}, & \text{when } \frac{p}{2} \le |i-j| < p, \\ 0, & \text{otherwise.} \end{cases} \quad (29)$$

where p is an arbitrary integer with $1 \le p \le k$. As noted in [13], the estimated covariance matrices $\mathbf{\Sigma}_{k\times k} = [\sigma_{ij}]_{1\le i,j\le k}$ can be considered over the following parameter space.

$$\left\{\mathbf{\Sigma} : \max_i \sum_j \left\{|\sigma_{ij}| : |i-j| \ge \frac{p}{2}\right\} \le C_1\left(\frac{p}{2}\right)^{-\alpha}\right\}, \quad (30)$$

where $C_1 > 0$ is a constant and p is an even integer with $1 \le p \le k$. The parameter α essentially specifies the rate

of decay for the covariances σ_{ij} as they move away from the diagonal, can be viewed as an analog of the smoothness parameter in non-parametric function estimation problems.

Thus, the bias part $\mathbb{E}[\bar{\mathbf{\Sigma}}_{\delta_\phi}] - \mathbf{\Sigma}_{\delta_\phi}$ can be expressed in the form of

$$\mathbb{E}[\bar{\mathbf{\Sigma}}_{\delta_\phi}] - \mathbf{\Sigma}_{\delta_\phi} = ((\omega_{ij} - 1)\sigma_{ij})_{k\times k}\,, \tag{31}$$

where $\omega_{ij} \in [0, 1]$. Since the operator norm of a symmetric matrix is bounded by its l_1 norm, in which $\omega_{ij} = 1$ when $|i - j| < \frac{p}{2}$, then

$$\begin{aligned}\left\|\mathbb{E}[\bar{\mathbf{\Sigma}}_{\delta_\phi}] - \mathbf{\Sigma}_{\delta_\phi}\right\|^2 &\le \left[\max_{1\le i\le k}\sum_j |(\omega_{ij} - 1)\sigma_{ij}|\right]^2 \\ &\le \left[\max_{1\le i\le k}\sum_{j:|i-j|\ge\frac{p}{2}} |\sigma_{ij}|\right]^2 \\ &\le C_1^2\left(\frac{p}{2}\right)^{-2\alpha}. \end{aligned}\tag{32}$$

B. The analysis of variance part

Next, we consider the upper bound of the variance part. This paper uses the tapering estimator of the covariance matrix as in [13]. The derivation of the idea is inspired by chernoff bounds. The approach write a matrix as a sum of many small block matrices along the diagonal, where the block matrices are given by

$$\mathbf{M}_l^{(m)} = (\sigma_{ij}\Omega\{l \le i < l + m, l \le j < l + m\})_{k\times k}\,, \tag{33}$$

where $\Omega\{l \le i < l + m, l \le j < l + m\}$ is an indicator that assigns the value one to the elements in this range of matrix. Without loss of generality, we assume that k is divisible by m. By setting $\mathbf{S}^{(m)}$ as $\mathbf{S}^{(m)} = \sum_{l=1-m}^{k}\mathbf{M}_l^{(m)}$, the tapering estimator can be written as

$$\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)} = \frac{1}{m_h}(\mathbf{S}^{(m)} - \mathbf{S}^{(m_h)}), \tag{34}$$

where $m_h = \frac{m}{2}$. The performance of the estimator $\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)}$ depends on the choice of parameter m. From the above equation, we can see that the estimation $\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)}$ can be written as the sum of a large number of small disjoint block matrices.

Lemma 1. Let $\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)}$ be an estimator, which is defined in (34). Then we have

$$\left\|\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)} - \mathbb{E}[\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)}]\right\| \le m\mathcal{N}_l^{(m)}, \tag{35}$$

where $\mathcal{N}_l^{(m)} = \max_{1\le l\le k}\|\mathbf{M}_l^{(m)} - \mathbb{E}[\mathbf{M}_l^{(m)}]\|$. □

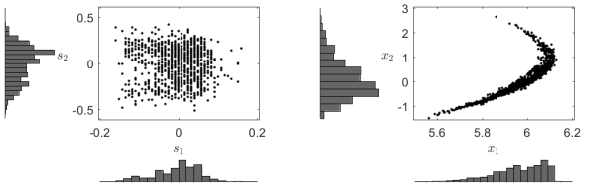
Lemma 2. Assume the distribution of $\phi(x_i)$ is sub-Gaussian in the sense that there is $\rho > 0$ such that

$$\mathbb{P}\{|\mathbf{v}^\top(\boldsymbol{\phi}(\mathbf{x}_i) - \mathbb{E}[\boldsymbol{\phi}(\mathbf{x}_i)])| > t\} \le \exp(-t^2\rho/2), \tag{36}$$

where $t > 0$ and $\|\mathbf{v}\|_2 = 1$. Then there is a constant $\rho_1 > 0$ such that

$$\mathbb{P}\{\mathcal{N}_l^{(m)} > 0\} \le 2k5^m\exp(-nx^2\rho_1), \tag{37}$$

for all $0 < x < \rho_1$ and $1 - m \le l \le p$.



(a) Source signals (b) Mixture signals

Fig. 2: The scatter plots of the source signals and mixture signals. (a) The source signals use the "CHiME3" dataset in TABLE I. (b) The mixture signals are nonlinearly mixed by PNL mixture function.

Thus, set $x = \sqrt{\frac{\log k+m}{T}}$. Since x is bounded by $0 < x < \rho_1$, then we have $\|\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)} - \mathbb{E}[\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)}]\|$ is bounded by a constant as

$$\left\|\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)} - \mathbb{E}[\hat{\mathbf{\Sigma}}_{\delta_\phi}^{(m)}]\right\| \le C_2\frac{\log k + m}{T}. \tag{38}$$

V. Simulation Results

In this section, we present some illustrative examples to demonstrate the validity of the computed bounds.

A. Data and Evaluation Equation

The source signals used for the following simulations include four real-world audio data. They are publicly available [14]. Each one has its own advantages, depending on whether one is interested in a variety of environments, in a number of data points, or in the overlap. For instance, the data "AMI" has two kinds of sound from the cable news and network news. Another data "Multitrack" was mixed with two anonymous singers. The length of the samples was varied to assess how the amount of data affects the performance of the algorithm. The general properties of the datasets are summarized in TABLE I.

These source signals are nonlinearly mixed by the post-nonlinear mixture (PNL), which was used in [9], [10].

The post-nonlinear mixtures constitute a particularly interesting example with the theoretical separability characterized by weak indeterminacy. The sources were the first subject to a linear mixture $\mathbf{z}(t) = \mathbf{A}\mathbf{s}(t)$, where $\mathbf{A}$ is a 2×2 mixing matrix

$$\mathbf{A} = \begin{pmatrix} -0.2261 & -0.1189 \\ -0.1706 & -0.2836 \end{pmatrix}. \tag{39}$$

Then each mixture component is generated from a nonlinear, invertible transformation, as the form of

$$\begin{aligned} x_1(t) &= (z_2(t) + 3z_1(t) + 6)\cos(1.5\pi)z_1(t), \\ x_2(t) &= (z_2(t) + 3z_1(t) + 6)\sin(1.5\pi)z_1(t). \end{aligned}\tag{40}$$

The sources are plotted in Fig. 2: (a). The mixture components are shown in Fig. 2: (b), where we can see the distortions caused by the nonlinearities.

TABLE I: Descriptions of real-world data [14].

Name	Scenario	Duration(s)	Sample Size	Overlap
AMI[1]	News	6	50,000	yes
CHiME3[2]	Talker	6	50,000	yes
Nonspeech[3]	TV order	10	160,000	no
Multitrack[4]	Theater	147	6,482,701	yes

To measure the performance of recovered sources, the normalized mean squared error (NMSE) was used [15], which has the definition

$$\mathrm{NMSE}(\mathbf{s}_i, \hat{\mathbf{s}}_i) = 10\log_{10}\left(\frac{1}{n}\sum_{i=1}^{n}\min_{\delta}\frac{\|\mathbf{s}_i - \delta\hat{\mathbf{s}}_i\|_2^2}{\|\mathbf{s}_i\|_2^2}\right), \quad (41)$$

where $\hat{\mathbf{s}}_i$ denotes the estimation of the source signal $\mathbf{s}_i$, and δ is a scalar reflecting the scalar ambiguity.

In addition, parameter determination is still an open problem [16]. The closed-form expressions of NMSE in (22) depends on the choice of parameter. We determine the parameter $\alpha = 0.1$ and other parameters are empirically determines as in traditional approaches [13].

B. Experiment Description and Results

Example 1: In this example, the numerical results show the convergence behavior of the iterative procedure to derive the coefficient matrix $\mathbf{W}$. The details are described in Algorithm 1. The choice of the threshold ϵ determines the convergence rate, in terms of an effective iterative time. As a general rule, a large threshold ϵ is preferred during the learning to promote the fast convergence. In contrast, a small ϵ is suggested as convergence to minimize variance. The iterative procedure is performed under the threshold varies from 10^{-6} to 10^{-2} by a step of 10^{-1}. The evaluation metric has been shown in (41). To reduce the randomness effect, 100 times Monte Carlo simulations are performed to evaluate the bound on NMSE for recovered sources shown in Fig. 3.

As seen from the appearance in Fig. 3, the objective is to show the theoretical NMSE changed during the iteration. The number of iteration is determined when the stopping threshold is achieved. The curves are labeled with the signal-to-noise power ratio (SNR) in decibels. As illustrated in this figure, the asymptotic closed-form expressions of NMSE are pertinent from threshold values. Thus, the divergence between two consecutive coefficient matrices $\mathbf{W}$ converges to threshold of approximately zero, which means that the final results will not be changed drastically. Besides, the asymptotic conditions are reached fast even for a small threshold.

Example 2: This example contains the comparison of experimental and theoretical performance on three kinds of real-world data with different noise intensities. We fixed threshold as $\epsilon = 10^{-4}$ that implies a much higher stopping criterion between two consecutive coefficient matrices. Both algorithm performance and theoretical bound are performed under different SNRs, which varied from 5 dB to 45 dB by a step of 10 dB. One of the expected outcomes is the curve of NMSE that tends to track the curve of bound even the SNR changed. We keep other parameters the same as in Example 1. We repeated 20 trials and averaged results.

The curves are labeled with different dataset, such as "AMI", "CHiME3" and "Nonspeech". The experimental and theoretical performance are marked as dash-dotted line and dashed line, respectively. As we can see from Fig. 5, the gap between the experimental result and theoretical result becomes smaller when the values of SNR is much higher. That is, the

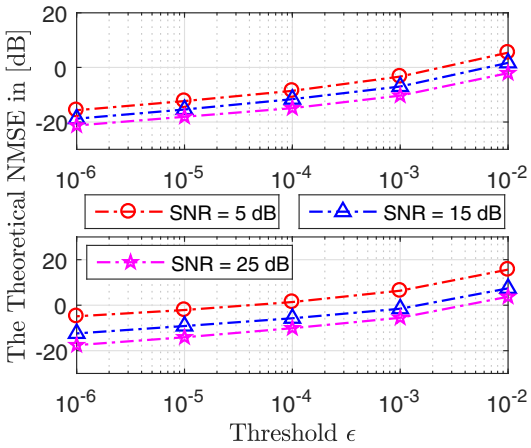


Fig. 3: The bound on NMSE of estimated signals versus of different values of threshold ϵ. The top figure uses "AMI" dataset. The bottom figure uses "CHiME3" dataset. Both of them are nonlinearly mixed by PNL mixture function.

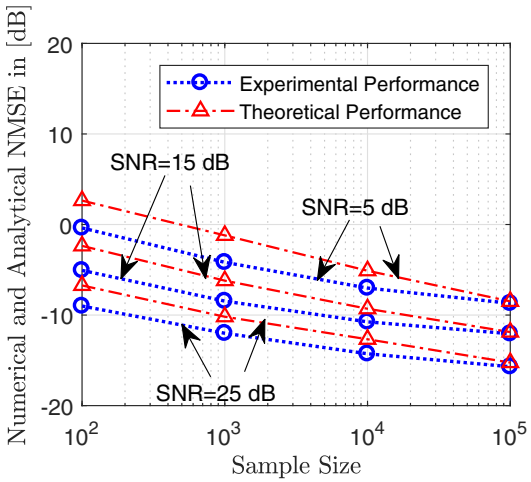


Fig. 4: The performance of experimental NMSE and theoretical bound versus of different SNR intensities. The dash-dotted line represents the theoretical curve. The dashed line represents the experimental curve.

[1]https://research.ics.aalto.fi/ica/newindex.shtml
[2]http://laslab.org/SpeechSeparationChallenge/
[3]http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html
[4]http://www.cambridge-mt.com/ms-mtk.htm

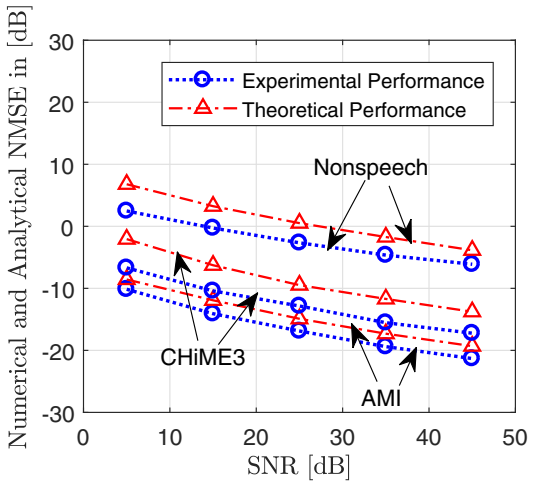


Fig. 5: The convergence behavior of NMSE versus the sample size.

experimental result follows the trends of the theoretical result more faithfully.

Example 3: In order to show the convergence behavior, this example compares the experimental and theoretical performance versus of the sample size when the threshold is set as $\epsilon = 10^{-4}$. The curves are labeled with the noise level that is kept constant at 15 dB and 25 dB, respectively. The evaluation metric has been shown in (41). To reduce the randomness effect, 20 times of Monte Carlo simulations are performed to evaluate the experimental and theoretical performance versus of sample size.

In Fig. 5, with each fixed noise intensities, the experimental curve are pertinent to the size of the samples. They all converge to a fixed value asymptotically when the sample size increased. We can see the experimental performance converges close to the theoretical performance as the number of samples increases, even if there is a model mismatch, which demonstrates the flexibility of the model. Especially, the convergence behavior in terms of noise intensity is unaffected.

VI. Conclusions

In this paper, we provide a upper bound on the normalized mean squared error (NMSE) of the estimated signals, which includes the closed-form expressions of NMSE, as well as proposing a new algebraic formalization that leads to the upper bound on the performance measure. The analysis stems from the performance of a mismatched estimator that accesses the finite sample size. The idea is inspired by the derivation of two parts. One is to derive an iterative expression of the coefficient matrix $\mathbf{W}$, which is to be estimated as deterministic but depends on the data. Another part aims to establish a closed-form expression of discrepancy between the truth model and its counterpart under the finite sample size. A major focus is on the derivation of the estimation of covariance matrices, which can be treated approximately under an asymptotic conditions.

Appendix A
Asymptotic Expression for NMSE

Without loss of generality, (6) can be equally rewritten as

$$\begin{aligned}\widehat{\mathrm{NMSE}} &= \mathrm{tr}\left\{\bar{\mathbf{\Sigma}}_{\delta_\phi}\mathbf{W}^\top\mathbf{W}\right\} \\ &= \mathrm{tr}\left\{\mathbf{\Sigma}_{\delta_\phi}\mathbf{W}^\top\mathbf{W}\right\} + \mathrm{tr}\left\{\delta\mathbf{\Sigma}_{\delta_\phi}\mathbf{W}^\top\mathbf{W}\right\}, \qquad \text{(A.1)}\end{aligned}$$

where the first equality derived from the property of $\mathrm{tr}\{\mathbf{ABC}\} = \mathrm{tr}\{\mathbf{BCA}\}$. The last equality is due to the definition of error

$$\bar{\mathbf{\Sigma}}_{\delta_\phi} = \mathbf{\Sigma}_{\delta_\phi} + \delta\mathbf{\Sigma}_{\delta_\phi}. \qquad \text{(A.2)}$$

Taking expectation of (A.1), we obtain

$$\begin{aligned}\mathbb{E}\{\widehat{\mathrm{NMSE}}\} &= \mathrm{tr}\left\{\mathbf{\Sigma}_{\delta_\phi}\mathbb{E}[\mathbf{W}^\top\mathbf{W}]\right\} + \mathbb{E}\left[\mathrm{tr}\{\delta\mathbf{\Sigma}_{\delta_\phi}\mathbf{W}^\top\mathbf{W}\}\right], \\ &= \mathrm{tr}\left\{\mathbf{\Sigma}_{\delta_\phi}\mathrm{Cov}(\mathbf{W}^\top)\right\} + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \qquad \text{(A.3)}\end{aligned}$$

Under an asymptotic condtions, i.e. $T \to \infty$, the covariance $\bar{\mathbf{\Sigma}}_{\delta_\phi}$ converges. As for the convergence rate, $\delta\mathbf{\Sigma}_{\delta_\phi}$ is proportional to $1/\sqrt{T}$. The detail derivation can be found in [17].

References

[1] J.-F. Cardoso, "Blind Signal Separation: Statistical Principles," *Proceeding of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.
[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* . JOHN WILEY & SONS, INC, 2001.
[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, LTD, 2002.
[4] A. Hyvärinen and H. Morioka, "Nonlinear ICA of Temporally Dependent Stationary Sources," in *Proc. of Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, vol. 54, 20–22 Apr. 2017, pp. 460–469.
[5] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Nonlinear Independent Components Estimation," in *arXiv:1410.8516 [cs.LG]*, Apr. 2015.
[6] A. Hyvärinen and P. Pajunen, "Nonlinear Independent Component Analysis: Existence and Uniqueness Results," *Neural Networks*, vol. 12, no. 3, pp. 429–439, Sep. 1999.
[7] H. H. Yang, S.-I. Amari, and A. Cichocki, "Information Theoretic Approach to Blind Separation of Sources in Nonlinear Mixture," *Signal Processing*, vol. 64, no. 3, pp. 291–300, Feb. 1998.
[8] G. Marques and L. Almeida, "Separation of Nonlinear Mixtures Using Pattern Repulsion," in *In Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA)*, Feb. 1999, pp. 277–282.
[9] E. Bahram, B.-Z. Massoud, R. Bertrand, and J. Christian, "Blind Source Separation in Nonlinear Mixtures: Separability and a Basic Algorithm," *IEEE Trans. on Signal Processing*, vol. 65, no. 16, pp. 4339–4352, Aug. 2017.
[10] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. Academies Press, Feb. 2010.
[11] L. Wang and T. Ohtsuki, "Nonlinear Blind Source Separation Unifying Vanishing Component Analysis and Temporal Structure," *IEEE Access*, vol. 6, no. 1, pp. 2169–3536, Dec. 2018.
[12] P. J. Bickel and E. Levina, "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, Feb. 2008.
[13] T. T. Cai and H. H. Zhou, "Optimal Rates of Convergence for Sparse Covariance Matrix Estimation," *The Annals of Statistics*, vol. 40, no. 5, pp. 2389–2420, Mar. 2012.
[14] J. Roux and E. Vincent, "A Categorization of Robust Speech Processing Datasets," Cambridge, MA, USA, Tech.Rep.TR2014–116, Aug. 2014.
[15] L. Zhen, D. Peng, Z. Yi, Y. Xiang, and P. Chen, "Underdetermined Blind Source Separation Using Sparse Coding," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 28, no. 12, pp. 3102–3108, Dec. 2017.

[16] C. Hou, F. Nie, X. Li, D. Yi, and Y. W, "Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[17] D. Lahat and C. Jutten, "Joint Independent Subspace Analysis Using Second-Order Statistics," *IEEE Trans. on Signal Processing*, vol. 64, no. 18, pp. 4891–4904, Sep. 2016.