Recovery of Independent Sparse Sources From Linear Mixtures Using Sparse Bayesian Learning

Seyyed Hamed Fouladi[1], Sung-En Chiu[2] , Bhaskar D. Rao[2] and Ilangko Balasingham[1]

***Abstract*—Classical algorithms for the multiple measurement vector (MMV) problem assume either independent columns for the solution matrix or certain models of correlation among the columns. The correlation structure in the previous MMV formulation does not capture the signals well for some applications like photoplethysmography (PPG) signal extraction where the signals are independent and linearly mixed in a certain manner. In practice, the mixtures of these signals are observed through different channels. In order to capture this structure, we decompose the solution matrix into multiplication of a sparse matrix composed of independent components, and a linear mixing matrix. We derive a new condition that guarantees a unique solution for this linear mixing MMV problem. The condition can be much less restrictive than the conditions for the typical MMV problem in previous works. We also propose two novel sparse Bayesian learning (SBL) algorithms, independent component analysis sparse Bayesian learning (ICASBL) and fast independent component sparse Bayesian learning (FASTICASBL), which capture the linear mixture structure. Analysis of the global and local minima of the ICASBL cost function is also provided, and similar to the typical SBL cost function it is shown that the local minima are sparse and that the global minima have maximum sparsity. Experimental results show that the proposed algorithms outperform traditional approaches and can recover the signal with fewer number of measurements in the linear mixing MMV setting.**

***Index Terms*—Sparse Bayesian learning (SBL), multiple measurement vectors (MMV), compressed sensing (CS), independent component analysis (ICA).**

I. Introduction

Sparse signal recovery and compressed sensing are being increasingly used in the area of signal reconstruction [1]–[4]. Both topics are applicable in a wide range of applications such as imaging [5], [6], biomedical signal processing [7]–[9], radar signal processing [10]–[12], and remote sensing [13]. A typical noiseless single measurement vector (SMV) problem employs the following model,

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}, \tag{1}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$ ($M \ll N$) is a known dictionary matrix, $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ is an unknown sparse vector with r nonzero elements, and $\boldsymbol{y} \in \mathbb{R}^{M\times 1}$ is the measurement vector. In the SMV problem, the task is to estimate the vector $\boldsymbol{x}$. To ensure a unique global solution, the number of nonzero entries in $\boldsymbol{x}$ has to be less than the following threshold [14],

$$\|\boldsymbol{x}\|_0 < \frac{\text{Spark}(\boldsymbol{\Phi})}{2}, \tag{2}$$

where $\|\boldsymbol{x}\|_0$ denotes the number of nonzero elements of the vector $\boldsymbol{x}$ and $\text{Spark}(\boldsymbol{\Phi})$ is defined as the smallest number of columns of $\boldsymbol{\Phi}$ that are linearly dependent [14].

There are many applications such as electroencephalogram (EEG) [15] and photoplethysmogram (PPG) [7] signal estimation where a sequence of the measurement vectors are available (See Section VIII-D for more details about rPPG and PPG applications.). In such cases, the model in (1) can be generalized into the multiple measurement vector (MMV) problem

$$\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{X}, \tag{3}$$

where $\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{y}_1\boldsymbol{y}_2\cdots\boldsymbol{y}_L\end{bmatrix} \in \mathbb{R}^{M\times L}$ is the observation matrix consisting of L measurement vectors, and $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_L\end{bmatrix} \in \mathbb{R}^{N\times L}$ is an unknown matrix (in this paper, we consider $L \ll N$). One approach for solving the MMV problem is to consider it as multiple SMV problems and solve for each column of $\boldsymbol{X}$ separately. However, a key assumption in the MMV model is that the support (i.e. indexes of nonzero entries) of every column in $\boldsymbol{X}$ is identical (referred as the common sparsity assumption in literature [16]), which enables us to solve the MMV problem by the following optimization

$$\min_{\boldsymbol{X}} \mathcal{R}(\boldsymbol{X}), \text{ subject to } \boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{X}, \tag{4}$$

where $\mathcal{R}(\boldsymbol{X})$ denotes a count of the number of nonzero rows of $\boldsymbol{X}$. This MMV optimization can recover uniquely $\boldsymbol{X}$ with less restrictive condition [17],

$$\mathcal{R}(\boldsymbol{X}) < \frac{\text{Spark}(\boldsymbol{\Phi}) + \text{Rank}(\boldsymbol{Y}) - 1}{2}. \tag{5}$$

The condition in (5) is less restrictive than the condition (2) in SMV whenever $\text{Rank}(\boldsymbol{Y}) > 1$.

In signal reconstruction for applications like PPG or remote PPG (rPPG), the desired signal is often contaminated with artifacts and noise which are independent [18]–[20]. The independent components are mixed and observed through some channels or sensors. Mathematically speaking, each column of $\boldsymbol{Y}$ represents a linear mixture of these signals; i.e.,

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{A}, \tag{6}$$

where $\boldsymbol{Z} \in \mathbb{R}^{M\times L}$ whose columns, $\boldsymbol{z}_i$, include the non-sparse sources, for example PPG and rPPG signals in time domain, and $\boldsymbol{A} \in \mathbb{R}^{L\times L}$ is an unknown full-rank mixing matrix. The

[1] Seyyed Hamed Fouladi and Ilangko Balasingham are with the Department of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim, Norway. (hamed.fouladi@ntnu.no; ilangkob@medisin.uio.no)

[2] Sung-En Chiu and Bhaskar D. Rao are with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407 USA (e-mail: suchiu@eng.ucsd.edu; brao@ucsd.edu).

problem in (6) is called instantaneous blind source separation (BSS) problem. The problem of estimating $\boldsymbol{Z}$ or $\boldsymbol{A}$ from $\boldsymbol{Y}$ has been well discussed in BSS and independent component analysis (ICA) literature [21]. Many algorithms were proposed to solve (6) by finding the sources as independent as possible using an information-theoretic cost function such as minima of KullbackLeibler divergence or maximization of cumulants [22]–[25], or high-order cross-cumulants [26]–[28]. ICA is applied first for extracting rPPG [18], [29] and PPG [19], [30] signals, and then discrete cosine domain (DCT) or fast Fourier transformation (FFT) is employed to obtain the heart rate. However, the resolution of FFT or DCT and the performance of ICA algorithms decrease when a small number of samples are available. This results in a poor estimate of the heart rate [31], [32].

The vectors $\boldsymbol{z}_i$ can be sparse in a domain $\mathcal{D}$ with the basis in dictionary matrix $\boldsymbol{\Phi}$. For instance PPG and rPPG signals are sparse in DCT or FFT domain. One can write

$$\boldsymbol{Z} = \boldsymbol{\Phi}\boldsymbol{S}, \tag{7}$$

where $\boldsymbol{S} = \begin{bmatrix}\boldsymbol{s}_1 \boldsymbol{s}_2 \cdots \boldsymbol{s}_L\end{bmatrix} \in \mathbb{R}^{N\times L}$ is an unknown source matrix including L source vectors. Using (6) and (7), $\boldsymbol{X}$ including linear mixtures of the sources in domain $\mathcal{D}$ can be written as

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}. \tag{8}$$

Considering the model in (8) for the solution matrix leads to a new class of MMV problems. It will be explained in Section III that the uniqueness condition for recovery of $\boldsymbol{X}$ can be further relaxed compared to the general MMV model.

In [33]–[36], algorithms based on generalized morphological component analysis (GMCA) were proposed which takes advantage of the sparse representation of $\boldsymbol{Z}$ using the basis in $\boldsymbol{\Phi}$ to estimate $\boldsymbol{Z}$ and $\boldsymbol{A}$. In classical instantaneous BSS problem, estimating either $\boldsymbol{Z}$ or equivalently $\boldsymbol{A}$ solves the problem in (6), but there are an infinite number of $\boldsymbol{S}$ which satisfy (7) or an infinite number of $\boldsymbol{X}$ which satisfy (3). In PPG or rPPG application, using classical BSS algorithms, which estimate $\boldsymbol{Z}$ and $\boldsymbol{A}$, may not be useful. This is because the recovery of the locations of nonzero elements in $\boldsymbol{S}$ or nonzero rows of $\boldsymbol{X}$ is the main parameter of interest because these locations include the heart rate frequency information. The union of the supports of $\boldsymbol{s}_i$ or the support of $\boldsymbol{X}$ involves the nonzero elements' locations. In Section III, we show mathematically that recovery of $\boldsymbol{S}$ requires more restrictive conditions compared to recovery of $\boldsymbol{X}$. Therefore, it is vital to have the exact recovery of the support for $\boldsymbol{X}$.

Many different algorithms have been proposed to recover the sparsest vector $\boldsymbol{x}$ for the SMV problem. One approach is to obtain a sparse representation through a greedy search algorithm like orthogonal matching pursuit (OMP). It was shown in [37], [38] that under specific conditions, OMP could find the sparsest representation of the signal. In addition, a popular approach is a regularization framework where a regularization penalty is introduced to promote sparsity. An example of this is using the ℓ_p norm to approximate ℓ_0 norm which is a popular analytical approach. In particular, when p is equal to 1, this leads to a convex optimization problem [39], [40]. Other regularization penalty functions, with fewer theoretical guarantees, lead to the reweighted ℓ_1 and ℓ_2 norm minimization algorithms [3], [41]–[43] which practically provide superior sparse recovery performance over ℓ_1 norm minimization. In addition to the regularization framework, many algorithms have been developed based on a Bayesian framework [44]–[47]. One of the important group of Bayesian approaches was proposed in [48], [49] and extended in [50]–[54]. In [51], the sparse Bayesian learning (SBL) algorithm was introduced for the SMV problem. One advantage of SBL is that all the local minima are sparse ($\|\boldsymbol{x}_*\|_0 \leq M$). Moreover, the SBL algorithm enjoyed having fewer number of local minima than the classical algorithms like the focal underdetermined system solver (FOCUSS) family [3], [16].

It has been shown that sparse recovery performance can be considerably improved compared to SMV algorithms while using multiple measurement vectors [16], [55], [56]. Consequently, many SMV algorithms were extended to address the MMV problem. For example, variants of OMP were proposed for recovering MMV that shared a joint sparsity pattern [57], [58]. In addition to the extensions of the OMP approaches, the regularization frameworks were developed for joint sparse recovery [16], [55]. Among the MMV approaches, Bayesian algorithms have attracted much attention due to their recovery performance. In the family of Bayesian algorithms, the extensions of SBL were developed for MMV [50], [59]–[61]. For instance, a block sparse Bayesian learning framework was presented in [60], which transformed an MMV problem to an SMV problem in order to capture the structure of nonzero rows of $\boldsymbol{X}$. The algorithm was extended in [59]. In [60], it was shown that the performance of sparse recovery was improved by modeling the structure of the nonzero rows of $\boldsymbol{X}$.

Among the algorithms for MMV, Bayesian approach offers superior performance [62] as well as flexibility by capturing different models for $\boldsymbol{X}$ motivated by the needs for the various applications [53], [60], [63]. Most of existing MMV algorithms assume that each row of $\boldsymbol{X}$ is independent and identically distributed (i.i.d). This is not suitable for many real-world scenarios, since practically rows of $\boldsymbol{X}$ will have certain structures, like temporal structure. In [53], it was shown that the recovery performance of exiting algorithms was affected by the temporal structure. To overcome this problem, the AR-SBL algorithm was proposed in [53] where each row of $\boldsymbol{X}$ was modeled as a first order autoregressive (AR) process with the AR coefficients learned from the observations. In [60], the authors generalized the correlation model to be arbitrary and developed the temporal multiple sparse Bayesian learning (T-MSBL) algorithm which performed well for some real world signals [64]. In this algorithm, the covariance of the i'th row of $\boldsymbol{X}$, was estimated in the learning procedure in order to capture the temporal structure. However, these approaches might not be accurate in some other applications, where $\boldsymbol{X}$ is a results of mixtures of independent sparse components, such as in image separation [33], [65], [66], and PPG signal extraction [7], [8], [67].

Motivated by applications such as PPG and rPPG, we first introduce a new model for $\boldsymbol{X}$ with a new correlation structure

in (8) for the MMV problem. Based on the model, we obtain a new uniqueness condition for recovery of $\boldsymbol{X}$ from the observation matrix $\boldsymbol{Y}$. This condition can be much less restrictive than the conditions obtained in the previous works for the general MMV model. In addition, a novel algorithm is proposed called independent component analysis sparse Bayesian learning (ICASBL) based on the independence assumption on the sources in $\boldsymbol{S}$. Analysis of the global and local minima of the ICASBL cost function is also provided, similar to the typical SBL formulation that ensures sparsity of local minima and the sparsest property of global minima. For practical use, a fast algorithm is needed. For this purpose, we propose fast independent component analysis sparse Bayesian learning (FASTICASBL) by using some approximations. The simulation results show the superiority of sparse recovery performance of our proposed algorithms ICASBL and FASTICASBL.

The main contributions of this paper are as follows:

- A new class of MMV model for capturing linear mixtures of independent sparse signals.
- A new condition for recovering $\boldsymbol{X}$ uniquely in the linear mixing MMV model. The new uniqueness condition is less restrictive than that of the general MMV model in certain regime.
- A corresponding Bayesian algorithm, ICASBL, for solving the linear mixing MMV problem.
- A fast version of ICASBL algorithm.
- Analytical proof that ICASBL cost function enjoys sparsity of local minima and sparset global minima, similar to the original SBL.

The rest of the paper is organized as follow. Section II introduces the new model for $\boldsymbol{X}$ and the formulation of the sparse recovery problem based on the new structure. In Section III, the uniqueness condition for $\boldsymbol{X}$ based on the new model is presented. In Section IV, the ICASBL algorithm is presented. FASTICASBL algorithm is derived in Section VI. Section VII provides analytical results on the global and local minima of the ICASBL cost function. Section VIII provides experimental results to support the efficacy of the developed algorithms. Finally, conclusions are drawn in the last section.

Notations:

- $\|\boldsymbol{s}\|_0, \|\boldsymbol{s}\|_1, \|\boldsymbol{s}\|_2, \|\boldsymbol{A}\|_{\mathcal{F}}$ denote the ℓ_0 norm of the vector $\boldsymbol{s}$, the ℓ_1 norm of $\boldsymbol{s}$, the ℓ_2 norm of $\boldsymbol{s}$ and the Frobenius norm of matrix $\boldsymbol{A}$, respectively.
- $\mathcal{R}(\boldsymbol{X})$ denotes the number of nonzero rows in the matrix $\boldsymbol{X}$.
- $\text{diag}\{a_1, \cdots, a_L\}$ denotes a diagonal matrix with diagonal elements being $a_1, \cdots, a_L$.
- For a matrix $\boldsymbol{A}$ and a vector $\boldsymbol{s}$, $\boldsymbol{A}_{(i,j)}$, and $\boldsymbol{s}_{(i)}$ denote the element that lies in the i-th row and the j-th column of $\boldsymbol{A}$ and i-th element of $\boldsymbol{s}$ respectively.
- $\boldsymbol{A} \otimes \boldsymbol{B}$ represents the Kronecker product of the two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. $\text{Tr}(\boldsymbol{A})$ denotes the trace of $\boldsymbol{A}$. $\boldsymbol{A}^T$ denotes the transpose of $\boldsymbol{A}$.
- $\text{vec}(\boldsymbol{A})$ denotes the vectorization of $\boldsymbol{A}$ formed by stacking its columns into a single column vector.
- $\boldsymbol{A}^{(k)}$, $\boldsymbol{a}^{(k)}$ and $\Theta^{(k)}$ show matrix $\boldsymbol{A}$, the vector $\boldsymbol{a}$ and the set $\Theta^{(k)}$ updated in the k'th step of the proposed algorithm, respectively.

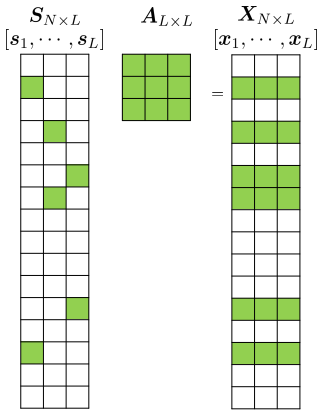

Fig. 1. Visual representation of the structure in (8).

II. Problem Statement

The targeted MMV problem to be solved contains the solution matrix $\boldsymbol{X}$ which includes linear mixtures of the sources in domain $\mathcal{D}$, can be written as

$$\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{X}, \ \boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}. \tag{9}$$

The multiplication of the source matrix $\boldsymbol{S}$ and the mixing matrix $\boldsymbol{A}$ results in $\boldsymbol{X}$, which consists of L linear mixtures of the sources. The i'th column of $\boldsymbol{S}$ consists of a sparse vector with r_i number of nonzero elements, i.e. $\mathcal{R}(\boldsymbol{s}_i) = \|\boldsymbol{s}_i\|_0 = r_i$. There is no other structure imposed on $\boldsymbol{S}$ other than the fact that each column of $\boldsymbol{S}$ is sparse. Since $\boldsymbol{A}$ is a full-rank matrix and independent from $\boldsymbol{S}$, this implies that $\text{Rank}(\boldsymbol{X}) = \text{Rank}(\boldsymbol{S})$.

Fig. 1 shows visually the structure described by (8) where each row of $\boldsymbol{S}$ can have fewer number of nonzero elements compared to $\boldsymbol{X}$ while both $\boldsymbol{S}$ and $\boldsymbol{X}$ have the same number of nonzero rows. Later, it will be shown that recovering $\boldsymbol{S}$ and $\boldsymbol{A}$ with the structure in (8) can help to recover $\boldsymbol{X}$ with fewer number of measurements M.

The sparsest representation of $\boldsymbol{S}$ is the solution to the following optimization problem,

$$(\boldsymbol{P}_0): \ \min_{\boldsymbol{S},\boldsymbol{A}} \sum_{i=1}^{L} \|\boldsymbol{s}_i\|_0, \ \text{subject to } \boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A}, \boldsymbol{X} = \boldsymbol{S}\boldsymbol{A} \text{ and } \sqrt{\sum_{j=1}^{L} \left(\boldsymbol{A}_{(i,j)}\right)^2} = 1. \tag{10}$$

In general, solving $(\boldsymbol{P}_0)$ requires enumerating all the subsets of set $\{1, \cdots, N\}$ for all $\boldsymbol{s}_i$. The complexity of such a subset-search algorithm grows exponentially with the dictionary size N. The global minimum of (10) leads to $\boldsymbol{S}$ which is as sparse as possible.

Remark 1. *In (10), the solution $\boldsymbol{S}$ has the least number of nonzero elements. Since the aim of the MMV problem is to find $\boldsymbol{X}$, in this paper $\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}$ is also called as the solution.*

Remark 2. *In this work it is assumed that the following condition holds*

$$\text{Rank}(\boldsymbol{Y}) = \text{Rank}(\boldsymbol{X}).$$

III. Uniqueness in The ℓ_0 Norm Minimization

We now discuss two aspects of the MMV problem in (10). They are conditions for uniqueness for $\boldsymbol{X}$, and recovery of $\boldsymbol{A}$ and $\boldsymbol{S}$ from $\boldsymbol{X}$.

Initially, a condition is found that leads to a unique solution for $\boldsymbol{X}$ by minimizing the ℓ_0 norm in (10) for a known matrix $\boldsymbol{A}$. Then, we derive a condition for the case with an unknown matrix $\boldsymbol{A}$. It is shown that these conditions can be less restrictive than the condition derived for the typical MMV problem in (5) assuming row sparsity.

A. Uniqueness condition for recovering matrix $\boldsymbol{X}$

This subsection deals with the uniqueness condition for $\boldsymbol{X}$ while minimizing (10). In previous MMV works, $\mathcal{R}(\boldsymbol{X})$ is minimized, however, in this paper we minimize $\sum_{i=1}^{L}\|\boldsymbol{s}_i\|_0$. The value of $\mathcal{R}(\boldsymbol{X})$ depends on the locations of the nonzero elements of $\boldsymbol{S}$. $\mathcal{R}(\boldsymbol{X})$ is equal to or larger than the sparsity of each column of $\boldsymbol{S}$ and is equal to or less than the summation of nonzero elements of $\boldsymbol{S}$; i.e., $\max_i\{r_i\} \leq \mathcal{R}(\boldsymbol{X}) \leq \sum_{j=1}^{L} r_j$. The lower bound is obtained when the support of each column of $\boldsymbol{S}$ is a subset of the support of the column of $\boldsymbol{S}$ with the highest number of nonzero elements. The upper bound is attained when there is no common support between the columns of $\boldsymbol{S}$. This shows that $\boldsymbol{X}$ is less sparse than $\boldsymbol{S}$; i.e., $\mathcal{R}(\boldsymbol{X}) \geq r_i$ for all i.

In the following two theorems, the sufficient conditions are found for a unique $\boldsymbol{X}$ to the problem $(\boldsymbol{P}_0)$ if $\boldsymbol{A}$ is known, and later for an unknown $\boldsymbol{A}$. An example is provided which shows solving (10) can be less restrictive than minimizing $\mathcal{R}(\boldsymbol{X})$.

Theorem 1. *Let source matrix $\tilde{\boldsymbol{S}} \in \mathbb{R}^{N\times L}$ with columns $\tilde{\boldsymbol{s}}_i \in \mathbb{R}^{N\times 1}$ and each column is a sparse vector with r_i nonzero elements. If matrix $\tilde{\boldsymbol{A}}$ is known and full-rank, matrix $\boldsymbol{X}$ will be a unique solution to the problem $(\boldsymbol{P}_0)$, if $\boldsymbol{Y} = \boldsymbol{\Phi}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}$, and*

$$\max_i\{r_i\} < \frac{\text{Spark}(\boldsymbol{\Phi})}{2}. \tag{11}$$

Furthermore $\boldsymbol{X} = \tilde{\boldsymbol{X}}$.

Proof. See appendix A. □

If $\boldsymbol{A}$ is known, the MMV problem in (3) and (8) can be changed into L SMV problems. Therefore, the source with the highest number of nonzero elements imposes the most restrictive condition as given in (11).

Now a sufficient condition is found such that the solution, $\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}$, to $(\boldsymbol{P}_0)$ is unique when $\boldsymbol{A}$ is unknown and full-rank.

Theorem 2. *Let source matrix $\tilde{\boldsymbol{S}} \in \mathbb{R}^{N\times L}$ with columns $\tilde{\boldsymbol{s}}_i \in \mathbb{R}^{N\times 1}$ and each column is a sparse vector with r_i nonzero elements. With an unknown full-rank matrix $\tilde{\boldsymbol{A}}$, matrix $\tilde{\boldsymbol{X}}$ will be a unique solution to the problem $(\boldsymbol{P}_0)$, if $\boldsymbol{Y} = \boldsymbol{\Phi}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}$, and*

$$\max_i\{r_i\} + \sum_{j=1}^{L} r_j < \text{Spark}(\boldsymbol{\Phi}). \tag{12}$$

Proof. See appendix B. □

By the following example, it is shown that the condition obtained in (12) can be much less restrictive than the condition given in (5).

B. Example

Let vectors $\boldsymbol{s}_i$ for $i \in \{1, \cdots, L\}$ be r-sparse and matrix $\boldsymbol{A}$ be full-rank. Assume that the columns of $\boldsymbol{S}$ do not have common support. Then the number of nonzero rows of $\boldsymbol{X}$ is Lr. For solving (4), the condition (5) has to be satisfied. For this example it is given by

$$L(2r-1)+1 < \text{Spark}(\boldsymbol{\Phi}), \tag{13}$$

while if the problem (10) is solved, then, using (12), the unique condition for this example is given by

$$r + Lr < \text{Spark}(\boldsymbol{\Phi}). \tag{14}$$

The left hand side of (13) is larger than (14) if $L > 1$ and $r > 1$. Fig. 1 shows the problem (10) when the columns of the matrix $\boldsymbol{S}$ have no common support. It can be seen that $\boldsymbol{S}$ is sparser than $\boldsymbol{X}$.

C. Condition for estimating the matrices $\boldsymbol{S}$ and $\boldsymbol{A}$ up to a permutation and a scale

The uniqueness condition for recovery of $\boldsymbol{X}$ from $\boldsymbol{Y}$ was discussed. Although $\boldsymbol{X}$ can be uniquely recovered under the condition (12), infinite number of solutions for $\boldsymbol{S}$ and $\boldsymbol{A}$ exist. For example if the solution is written as $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}$, the multiplication of the matrices $\tilde{\boldsymbol{S}}' = \tilde{\boldsymbol{S}}\boldsymbol{D}^{-1}\boldsymbol{P}^{-1}$ and $\tilde{\boldsymbol{A}}' = \boldsymbol{P}\boldsymbol{D}\tilde{\boldsymbol{A}}$ also gives $\tilde{\boldsymbol{X}}$, where $\boldsymbol{P}$ is a permutation matrix and $\boldsymbol{D}$ is a diagonal matrix. In a practical scenario, any solution like $\{\tilde{\boldsymbol{S}}', \tilde{\boldsymbol{A}}'\}$ is acceptable because the target is to extract the desired sources up to a scale. The scale ambiguity can be easily avoided by assuming a constant for the norm of each row of $\boldsymbol{A}$. In this paper, the ℓ_2 norm of each row of matrix $\boldsymbol{A}$ is set to 1 to avoid the scale ambiguity.

Now lets discuss the condition under which the matrices $\boldsymbol{S}$ and $\boldsymbol{A}$ can be estimated up to a permutation and a scale.

Theorem 3. *Let $\{\tilde{\boldsymbol{S}}, \tilde{\boldsymbol{A}}\}$ be a solution to the problem (10). Let $\tilde{\boldsymbol{s}}_i$ and $\tilde{\boldsymbol{s}}_j$ be r_i-sparse and r_j-sparse, respectively, with no common support for $i \neq j$. Any other global minima will be obtained by $\boldsymbol{S}_* = \tilde{\boldsymbol{S}}\boldsymbol{P}$ for a permutation matrix $\boldsymbol{P}$ if the condition (12) is satisfied.*

Proof. See Appendix C. □

If there is no further assumption on source and mixing matrices, the permutation and scale ambiguities are unavoidable. However, since the sources are separated from each other, all the possible solutions are acceptable.

Remark 3. *In order to obtain a unique $\boldsymbol{X}$ from optimizing (10), there is no need for $\tilde{\boldsymbol{s}}_i$ with no common support.*

In the following sections, we propose two algorithms that recover the sources $\tilde{\boldsymbol{S}}$, $\tilde{\boldsymbol{A}}$, and $\tilde{\boldsymbol{X}}$.

IV. Independent Component Analysis Based Sparse Bayesian Learning

A Bayesian framework is used to estimate $\boldsymbol{S}$ and $\boldsymbol{A}$. This is motivated by the flexibility of the Bayesian framework for incorporating structure and the success of the SBL family of algorithms. This work develops an SBL type algorithm for the problem of interest. Assuming the observations are contaminated with noise, the observation model is as follows

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A} + \boldsymbol{N}, \\ \boldsymbol{X} &= \boldsymbol{S}\boldsymbol{A}, \end{aligned} \tag{15}$$

where matrix $\boldsymbol{N} \in \mathbb{R}^{M\times L}$ is the observation noise matrix.

It is convenient to convert the matrix formulation to an equivalent vector problem for tractability and for algorithm development. For this purpose the following definitions are useful.

Definition 1. *A linear transformation* $\text{vec}_{m,n}(.) : \mathbb{R}^{m\times n} \rightarrow \mathbb{R}^{mn\times 1}$ *of a matrix* $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ *converts the matrix into a column vector as*

$$\begin{aligned} &\text{vec}_{m,n}(\boldsymbol{B}) = \left[\boldsymbol{B}_{(1,1)}, \cdots, \boldsymbol{B}_{(m,1)}, \boldsymbol{B}_{(m,2)}, \cdots, \boldsymbol{B}_{(m,n)}\right]^T, \\ &\text{vec}_{m,n}^{-1}\left(\left[\boldsymbol{B}_{(1,1)}, \cdots, \boldsymbol{B}_{(m,1)}, \boldsymbol{B}_{(m,2)}, \cdots, \boldsymbol{B}_{(m,n)}\right]^T\right) = \boldsymbol{B}. \end{aligned}$$

If the operator vec is applied to both sides of (15), it results

$$\begin{aligned} \text{vec}_{N,M}(\boldsymbol{Y}^T) =& \text{vec}_{N,M}((\boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A})^T) + \text{vec}_{N,M}(\boldsymbol{N}^T) \\ =& (\boldsymbol{\Phi}\otimes\boldsymbol{A}^T)\text{vec}_{N,L}(\boldsymbol{S}^T) + \text{vec}_{N,M}(\boldsymbol{N}^T), \end{aligned} \tag{16}$$

where the operator $\otimes$ denotes Kronecker product.

For the sake of simplicity, let $\boldsymbol{y}^v = \text{vec}_{L,M}(\boldsymbol{Y}^T)$, $\boldsymbol{s}^v = \text{vec}_{L,N}(\boldsymbol{S}^T)$, $\boldsymbol{n}^v = \text{vec}_{L,M}(\boldsymbol{N}^T)$ and $\boldsymbol{\phi}_{\boldsymbol{A}} = \boldsymbol{\Phi}\otimes\boldsymbol{A}^T$, then, we can write (16) as

$$\boldsymbol{y}^v = \boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{s}^v + \boldsymbol{n}^v. \tag{17}$$

This leads to an SMV problem where the dictionary matrix $\boldsymbol{\phi}_{\boldsymbol{A}}$ is not known because of its dependence on the unknown $\boldsymbol{A}$. To model the noise, let elements in the noise vector be independent and identically Gaussian distributed, i.e., $\boldsymbol{n}^v_{(i)} \sim \mathcal{N}(0,\sigma^2)$, where $\boldsymbol{n}^v_{(i)}$ is the i'th element in $\boldsymbol{n}$ and σ^2 is the noise variance. The likelihood of model (17) is obtained by

$$p(\boldsymbol{y}^v|\boldsymbol{s}^v;\boldsymbol{A},\sigma^2) = (2\pi\sigma^2)^{-\frac{LM}{2}}\exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{y}^v - \boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{s}^v\right\|_2^2\right), \tag{18}$$

In order to find a sparse $\boldsymbol{s}^v$ to (18), a prior distribution for $\boldsymbol{s}^v$ is necessary. Following the SBL approach [51], it is assumed a Gaussian prior for $\boldsymbol{s}^v$ as

$$p(\boldsymbol{S};\boldsymbol{\Gamma}) = p(\boldsymbol{s}^v;\boldsymbol{\Gamma}) = \prod_{i=1}^{L}\prod_{j=1}^{N}(2\pi\gamma_{ij})^{-\frac{1}{2}}\exp\left(-\frac{\boldsymbol{S}^2_{(i,j)}}{2\gamma_{ij}}\right), \tag{19}$$

where γ_{ij} are hyperparameters controlling the prior variance of each element of matrix $\boldsymbol{S}$ and matrix $\boldsymbol{\Gamma}$ is

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & & & \\ & \gamma_{12} & & \\ & & \ddots & \\ & & & \gamma_{LN} \end{bmatrix}. \tag{20}$$

By proper choice of a prior on the hyperparameters $\boldsymbol{\Gamma}$, one can impose a super-Gaussian prior on $\boldsymbol{S}$ which are known to be sparsity inducing [45]. However, based on the past experience with SBL a non-informative prior has been found to be sufficient and useful [51], [60]. This is equivalent to, from an algorithmic point of view, considering $\boldsymbol{\Gamma}$ as a deterministic unknown. This is the approach adopted in this work. For fixed values of the hyperparameters governing the prior on $\boldsymbol{s}^v$, the posterior density of the sources given the measurements is Gaussian and is given by

$$\begin{aligned} p(\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta) =& (2\pi)^{-\frac{LN}{2}}\left|\boldsymbol{\Sigma}_{\boldsymbol{s}^v}\right|^{-\frac{1}{2}} \\ & \exp\left(-\frac{1}{2}(\boldsymbol{s}^v - \boldsymbol{\mu})^T\boldsymbol{\Sigma}_{\boldsymbol{s}^v}^{-1}(\boldsymbol{s}^v - \boldsymbol{\mu})\right), \end{aligned} \tag{21}$$

where the hyperparameters $\Theta = \{\boldsymbol{\Gamma}, \sigma^2, \boldsymbol{A}\}$ and the mean, $\boldsymbol{\mu}$, is

$$\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}_{\boldsymbol{s}^v}{\boldsymbol{\phi}_{\boldsymbol{A}}}^T\boldsymbol{y}^v, \tag{22}$$

and the covariance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{s}^v}$, is

$$\boldsymbol{\Sigma}_{\boldsymbol{s}^v} = (\sigma^{-2}\boldsymbol{\phi}_{\boldsymbol{A}}^T\boldsymbol{\phi}_{\boldsymbol{A}} + \boldsymbol{\Gamma}^{-1})^{-1}. \tag{23}$$

Consequently, given the hyperparameters $\boldsymbol{\Gamma}, \sigma^2$, matrix $\boldsymbol{A}$, and the observations, the maximum aposteriori probability (MAP)/minimum mean squared error (MMSE) estimate of $\boldsymbol{s}^v$ is given by

$$\hat{\boldsymbol{s}^v} = \boldsymbol{\mu} = \left(\boldsymbol{\phi}_{\boldsymbol{A}}^T\boldsymbol{\phi}_{\boldsymbol{A}} + \sigma^2\boldsymbol{\Gamma}^{-1}\right)^{-1}\boldsymbol{\phi}_{\boldsymbol{A}}^T\boldsymbol{y}^v. \tag{24}$$

The estimation of the source matrix is readily obtained as

$$\hat{\boldsymbol{S}} = \text{vec}_{N,L}^{-1}(\hat{\boldsymbol{s}^v}). \tag{25}$$

The sparsity of $\hat{\boldsymbol{s}^v}$ is controlled by γ_{ij} in $\boldsymbol{\Gamma}$. During the estimation procedure, when γ_{ij} tends to zero, the associated $(L(i-1)+j)$'th element of $\hat{\boldsymbol{s}^v}$ tends to zero or the associated element in i'th row and j'th column of $\hat{\boldsymbol{S}}$ tends to zero.

Evidence maximization or Type-II maximum likelihood is exploited for estimating the hyperparameters [48]. The hyperparameters can be learned from marginalizing the observations over $\boldsymbol{s}^v$ and then performing ML optimization. In the following section, we discuss the learning of hyperparameters in detail.

V. Hyperparameter Estimation

In order to find the hyperparameters Θ, the expectation maximization (EM) algorithm is utilized for maximizing $p(\boldsymbol{y}^v;\boldsymbol{\Gamma},\sigma^2,\boldsymbol{A})$.

$$\begin{aligned} p(\boldsymbol{y}^v;\boldsymbol{\Gamma},\boldsymbol{A},\sigma^2) =& \int p(\boldsymbol{y}^v|\boldsymbol{s}^v;\boldsymbol{A},\sigma^2)p(\boldsymbol{s}^v;\boldsymbol{\Gamma})d\boldsymbol{s}^v \\ =& (2\pi)^{-\frac{LM}{2}}\left|\boldsymbol{\Sigma}_{\boldsymbol{y}^v}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{y}^v)^T\boldsymbol{\Sigma}_{\boldsymbol{y}^v}^{-1}\boldsymbol{y}^v\right), \end{aligned} \tag{26}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{y}^v} = \sigma^2\boldsymbol{I} + \boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{\Gamma}\boldsymbol{\phi}_{\boldsymbol{A}}^T$. Minimizing $\mathcal{L}(\Theta)$ in the following equation is the same as minimizing $-\log p(\boldsymbol{y}^v;\boldsymbol{\Theta})$, because they are equal up to an additive constant,

$$\mathcal{L}(\Theta) = \log|\boldsymbol{\Sigma}_{\boldsymbol{y}^v}| + (\boldsymbol{y}^v)^T\boldsymbol{\Sigma}_{\boldsymbol{y}^v}^{-1}\boldsymbol{y}^v, \tag{27}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{y}^v} = \sigma^2\boldsymbol{I} + \boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{\Gamma}\boldsymbol{\phi}_{\boldsymbol{A}}^T$. The actual EM formulation proceeds by treating $\boldsymbol{s}^v$ as hidden variables and then maximizing the following objective function with respect to the parameters,

$$\max_{\boldsymbol{\Gamma},\sigma^2,\boldsymbol{A}} \mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta}\left\{p\left(\boldsymbol{y}^v,\boldsymbol{s}^v;\boldsymbol{\Gamma},\sigma^2\right)\right\}, \tag{28}$$

where $p\left(\boldsymbol{y}^v,\boldsymbol{s}^v;\boldsymbol{\Gamma},\sigma^2\right) = p\left(\boldsymbol{y}^v|\boldsymbol{s}^v;\boldsymbol{A},\sigma^2\right)p\left(\boldsymbol{s}^v;\boldsymbol{\Gamma}\right)$. The parameters are updated as follows. Detailed derivation of the parameter update procedure is presented in Appendix D. Hyperparameters γ_{ij} are updated as

$$\gamma_{ij}^{(k+1)} = \left(\boldsymbol{\Sigma}_{\boldsymbol{s}^v}^{(k)}\right)_{(i+j)} + \left(\boldsymbol{\mu}_{(i+j)}^{(k)}\right)^2. \tag{29}$$

The variance σ^2 is updated as

$$\begin{aligned}\left(\sigma^2\right)^{(k+1)} &= \frac{\|\boldsymbol{y}-\boldsymbol{\phi}_{\boldsymbol{A}^{(k)}}\boldsymbol{\mu}^{(k)}\|_2^2}{ML} \\ &+ \frac{\left(\sigma^2\right)^{(k)}\sum_{i=1}^{N}\sum_{j=1}^{L}\left(1-\left(\gamma_{ij}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{s}^v}^{(k)}\right)_{(i+j)}\right)}{ML}.\end{aligned} \tag{30}$$

In order to update $\boldsymbol{A}$, the term $\mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta^{(k)}}\left\{\|\boldsymbol{y}^v-\boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{s}^v\|_2^2\right\}$ is minimized with respect to $\boldsymbol{A}$ leading to the following update rule

$$\boldsymbol{A}^{(k+1)} = \left(\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}\right\}\right)^{-1}\times\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\right\}\boldsymbol{Y}, \tag{31}$$

where $\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\right\}$ is given by

$$\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\right\} = \text{vec}_{L,N}^{-1}\left(\boldsymbol{\mu}^{(k)}\right)\boldsymbol{\Phi}^T, \tag{32}$$

and the (i,j)th element of the matrix $\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}\right\}$ is given by

$$\left(\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}\right\}\right)_{(i,j)} = \text{Tr}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{s}_j\boldsymbol{s}_i^T\right\}\right). \tag{33}$$

The elements of $\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{s}_j\boldsymbol{s}_i^T\right\}$ can be obtained by

$$\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}_{(i,j)}\boldsymbol{S}_{(v,p)}\right\} = \left(\boldsymbol{\Sigma}_{\boldsymbol{s}^v}^{(k)}\right)_{(i',j')} + \boldsymbol{\mu}_{(i')}^{(k)}\boldsymbol{\mu}_{(j')}^{(k)}, \tag{34}$$

where $i' = j+L(i-1)$ and $j' = p+L(v-1)$. After obtaining $\boldsymbol{A}^{(k+1)}$, the ℓ_2 norm of each row of $\boldsymbol{A}^{(k+1)}$ is set to 1.

We refer to the algorithm including learning rules (23), (24), (29), (30) and (31) as independent component analysis based sparse Bayesian learning (ICASBL).

VI. Fast Independent Component Analysis Sparse Bayesian Learning

We will show in Section VIII that the proposed ICASBL algorithm performs well in terms of recovery performance. However, the algorithm is not fast because the dimension of the parameters to be learned jointly is high as a result of using the vectorization operator. The dictionary matrix $\boldsymbol{\Phi}$ in MMV problem is $M\times N$ while the defined dictionary matrix for ICASBL, $\boldsymbol{\phi}_{\boldsymbol{A}}$, is $ML\times NL$. To develop a fast variant, we change the MMV problem in (15) to L SMV problems by a right multiplication of $\boldsymbol{A}^{-1}$ to both sides of (15),

$$\bar{\boldsymbol{Y}} = \boldsymbol{\Phi}\boldsymbol{S} + \bar{\boldsymbol{N}}, \tag{35}$$

where $\bar{\boldsymbol{N}} = \boldsymbol{N}\boldsymbol{A}^{-1}$ and $\bar{\boldsymbol{Y}}$ has independent columns. For each column of $\bar{\boldsymbol{Y}}$, one can write

$$\bar{\boldsymbol{y}}_i = \boldsymbol{\Phi}\boldsymbol{s}_i + \bar{\boldsymbol{n}}_i, \tag{36}$$

where $\bar{\boldsymbol{y}}_i$ and $\bar{\boldsymbol{n}}_i$ are the i'th column of $\bar{\boldsymbol{Y}}$ and the i'th column of $\bar{\boldsymbol{N}}$, respectively. By ignoring the correlation among the columns of $\bar{\boldsymbol{N}}$, we consider (36) as L SMV problems for $1\le i\le L$, i.e., we assume

$$\begin{aligned}\mathbf{E}\{\bar{\boldsymbol{n}}_{i_{(j)}}\bar{\boldsymbol{n}}_{i_{(k)}}\} &\approx 0, \quad j\neq k, \\ \text{Cov}\left(\bar{\boldsymbol{n}}_i\right) &\approx \bar{\sigma}^2\boldsymbol{I},\end{aligned} \tag{37}$$

where $\bar{\sigma}^2$ is estimated in the learning procedure. By this assumption, the columns of $\bar{\boldsymbol{Y}}$ are independent. The learning rules are now similar to the SBL approach in [51] and are listed below. The posterior density of $\boldsymbol{s}_i$ given $\bar{\boldsymbol{y}}_i$ is given by

$$\begin{aligned}p(\boldsymbol{s}_i|\bar{\boldsymbol{y}}_i;\boldsymbol{\Gamma}_i,\bar{\sigma}^2) &= (2\pi)^{-\frac{LN}{2}}\left|\boldsymbol{\Sigma}_{\boldsymbol{s}_i}\right|^{-\frac{1}{2}} \\ &\exp\left(-\frac{1}{2}(\boldsymbol{s}_i-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_{\boldsymbol{s}_i}^{-1}(\boldsymbol{s}_i-\boldsymbol{\mu}_i)\right),\end{aligned} \tag{38}$$

where

$$\boldsymbol{\mu}_i = \hat{\boldsymbol{s}}_i = \left(\bar{\sigma}^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{\Gamma}_i^{-1}\right)^{-1}\boldsymbol{\Phi}^T\bar{\boldsymbol{y}}_i, \tag{39}$$

and $\boldsymbol{\Gamma}_i$ (using the MATLAB notations),

$$\boldsymbol{\Gamma}_i = \boldsymbol{\Gamma}(i:L:i+L(N-1)) = \text{diag}([\gamma_{1i}\cdots\gamma_{Ni}]). \tag{40}$$

The covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{s}_i} = \left(\bar{\sigma}^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{\Gamma}_i^{-1}\right)^{-1}, \tag{41}$$

where

$$\gamma_{ji}^{(k+1)} = \left(\boldsymbol{\Sigma}_{\boldsymbol{s}_i}^{(k)}\right)_{(j,j)} + \left(\hat{\boldsymbol{s}}_{i(j)}^{(k)}\right)^2 = \left(\boldsymbol{\Sigma}_{\boldsymbol{s}_i}^{(k)}\right)_{(j,j)} + \left(\boldsymbol{\mu}_{i(j)}^{(k)}\right)^2. \tag{42}$$

The parameter $\bar{\sigma}^2$ is updated by

$$\begin{aligned}\left(\bar{\sigma}^2\right)^{(k+1)} &= \frac{\|\bar{\boldsymbol{Y}}^{(k)}-\boldsymbol{\Phi}\boldsymbol{S}^{(k)}\|_{\mathcal{F}}^2}{ML} \\ &+ \frac{\left(\bar{\sigma}^2\right)^{(k)}\sum_{i=1}^{N}\sum_{j=1}^{L}\left(1-\left(\gamma_{ji}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{s}_i}^{(k)}\right)_{(j,j)}\right)}{ML},\end{aligned} \tag{43}$$

where $\|.\|_F$ denotes Frobenius norm.

The matrix $\boldsymbol{A}$ can be updated by (31), (32) and (33). By considering $\bar{\boldsymbol{Y}}^{(k)} = \boldsymbol{Y}\left(\boldsymbol{A}^{(k)}\right)^{-1}$ and the approximation for

the covariance of $\bar{\boldsymbol{N}}$, the following approximation can be made using (36) to develop updates for $\boldsymbol{A}$,

$$\begin{aligned}\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}_{(i,j)}\boldsymbol{S}_{(v,p)}\right\} =&\mathbf{E}_{\boldsymbol{S}|\tilde{\boldsymbol{Y}}^{(k)};\Theta^{(k)}}\left\{\boldsymbol{S}_{(i,j)}\boldsymbol{S}_{(v,p)}\right\}\\ \approx&\boldsymbol{S}^{(k)}_{(i,j)}\boldsymbol{S}^{(k)}_{(v,p)},\\ =&\boldsymbol{\mu}^{(k)}_{j_{(i)}}\boldsymbol{\mu}^{(k)}_{p_{(v)}} \text{ for } j\neq p, \qquad (44)\\ \mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}_{(i,j)}\boldsymbol{S}_{(v,p)}\right\} =&\left(\boldsymbol{\Sigma}^{(k)}_{\boldsymbol{s}_j}\right)_{(i,v)}+\boldsymbol{S}^{(k)}_{(i,j)}\boldsymbol{S}^{(k)}_{(v,p)},\\ =&\left(\boldsymbol{\Sigma}^{(k)}_{\boldsymbol{s}_j}\right)_{(i,v)}+\boldsymbol{\mu}^{(k)}_{j_{(i)}}\boldsymbol{\mu}^{(k)}_{p_{(v)}},\\ &\text{for } j=p. \qquad (45)\end{aligned}$$

Using (44) and (45), the approximation of $\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}\right\}$ is obtained. After obtaining $\boldsymbol{A}^{(k+1)}$, we set the ℓ_2 norm of each row of $\boldsymbol{A}^{(k+1)}$ to 1.

We refer to the sparse recovery algorithm based on (39), (41), (42), (43) and (31) as fast ICASBL (FASTICASBL). According to the approximation in (37), we expect the FASTICASBL algorithm to perform well when σ^2 is small or goes to zero.

VII. Analysis of Global Minima and Local Minima

In the following subsections, the properties of the cost function $\mathcal{L}$ is analyzed.

A. Analysis of the Global Minima

In our analysis, $\tilde{\boldsymbol{S}}$ is considered as the sparsest solution with matrix $\tilde{\boldsymbol{A}}$ providing $\tilde{\boldsymbol{X}}=\tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}$. The following theorem shows that the other solution of (10) other than $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{A}}$ can be obtained by a permutation matrix and a diagonal matrix, i.e. $\boldsymbol{S}_*=\tilde{\boldsymbol{S}}\boldsymbol{P}$ and $\boldsymbol{A}_*=\boldsymbol{P}^{-1}\tilde{\boldsymbol{A}}$, which leads to the unique matrix $\tilde{\boldsymbol{X}}$. All the solutions in the form of $\boldsymbol{S}_*$ and $\boldsymbol{A}_*$ are acceptable. Using the following theorem, it is shown that $\boldsymbol{S}_*$ are the global minima of the cost function $\mathcal{L}$ in (27).

Theorem 4. *Let the noise variance $\sigma^2\to 0$, matrix $\boldsymbol{A}$ be a full-rank matrix, and the sparsest solution to (10) be $\tilde{\boldsymbol{S}}$ with r_i-sparse columns with no common support, and the variance vector $\tilde{\boldsymbol{\gamma}}=\text{diag}(\tilde{\boldsymbol{\Gamma}})$. Assume that the condition (12) is satisfied giving a unique solution for (10) which is denoted by $\tilde{\boldsymbol{X}}$. The global minimum of (26) is achieved at $\hat{\boldsymbol{\gamma}}=\text{diag}(\hat{\boldsymbol{\Gamma}})=\boldsymbol{\gamma}_*$ which is used to obtain $\hat{\boldsymbol{s}}=\text{vec}(\hat{\boldsymbol{S}}_*)$ by (24) where the matrix $\boldsymbol{S}_*$ is obtained only by permuting the matrix $\tilde{\boldsymbol{S}}$.*

Proof. See Appendix E. □

Theorem 4 ensures that the objective function being minimized has the desired global minima.

B. Analysis of Local Minima

In this subsection we discuss about the local minima property of the cost function $\mathcal{L}$ in (27) with respect to $\boldsymbol{\gamma}$, in which $\boldsymbol{\phi}_{\boldsymbol{A}}=\boldsymbol{\Phi}\otimes\boldsymbol{A}^T$ for a fixed $\boldsymbol{A}$. The lines of proofs for the lemmas and the theorem in this subsection are similar to the proofs in [51] and are extended with some considerations for the cost function $\mathcal{L}$ in our work.

Lemma 1. $\log(\boldsymbol{\Sigma}_{\boldsymbol{y}^v})=\log(|\sigma^2\boldsymbol{I}+\boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{\Gamma}\boldsymbol{\phi}_{\boldsymbol{A}}^T|)$ *is concave with respect to* $\boldsymbol{\gamma}$, *where* $\boldsymbol{\gamma}=\text{diag}(\boldsymbol{\Gamma})$.

This can be proved by the composition property of concave functions [68].

Lemma 2. *The term $(\boldsymbol{y}^v)^T\boldsymbol{\Sigma}_{\boldsymbol{y}^v}^{-1}\boldsymbol{y}^v$ equals a constant C for all $\boldsymbol{\gamma}$ satisfying the ML linear constraints $\boldsymbol{b}=\boldsymbol{G}\boldsymbol{\gamma}$.*

$$\boldsymbol{b}\triangleq\boldsymbol{y}^v-\sigma^2\boldsymbol{u}, \qquad (46)$$

$$\boldsymbol{G}\triangleq\boldsymbol{\phi}_{\boldsymbol{A}}\text{diag}(\boldsymbol{\phi}_{\boldsymbol{A}}^T\boldsymbol{u}), \qquad (47)$$

where $\boldsymbol{G}$ is a full-rank matrix and $\boldsymbol{u}$ is any fixed vector such that $(\boldsymbol{y}^v)^T\boldsymbol{u}=C$.

Proof. See Appendix F. □

Lemma 1 and Lemma 2 are used to find a bound for local minima in the following theorem.

Theorem 5. *Every local minimum of $\mathcal{L}$ is achieved at a sparse solution, i.e., $\|\boldsymbol{\gamma}\|_0<ML$, irrespective of the presence of noise.*

Proof. See Appendix G. □

This shows that the local minima of $\mathcal{L}$ have less than ML nonzero elements. Since the global minima of $\mathcal{L}$ have the number of nonzero elements less than M, we are more interested in the local minima with the number of nonzero elements less than M.

Lemma 3. *If σ^2 goes to* 0, *for every local minimum which satisfies $\|\hat{\boldsymbol{\gamma}}\|_0\leq M$, the i'th nonzero element of $\hat{\boldsymbol{\gamma}}$ is obtained by*

$$\hat{\gamma}_{(i)}=\left(\hat{s^v}_{(i)}\right)^2, \quad \text{for } i \text{ s.t. } \hat{\gamma}_{(i)}\neq 0, \qquad (48)$$

where $\hat{\boldsymbol{s}^v}$ is the basic feasible solution to $\boldsymbol{y}^v=\boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{s}^v$.

Proof. See Appendix H. □

By Lemma 3, a closed form for $\hat{\gamma}_{(i)}$ is achieved at the global minimum.

Here we mention the differences between the local minima analysis of $\boldsymbol{\gamma}$ for T-MSBL and ICASBL. Let $\boldsymbol{\gamma}_T\in\mathbb{R}^N$ be a vector including hyperparameters for the T-MSBL algorithm. The nonzero elements of the estimated vector $\boldsymbol{\gamma}_T$ corresponds to the nonzero rows of $\boldsymbol{X}$. Based on Theorem 2 in [60], $\|\boldsymbol{\gamma}_T\|_0$ is less or equal than ML. This bound is the same as what we obtained in Theorem 5 for $\boldsymbol{\gamma}$ at local minima. However, the size of $\boldsymbol{\gamma}_T$ is N by 1 while the size of $\boldsymbol{\gamma}$ in Theorem 5 is NL by 1. Unlike $\boldsymbol{\gamma}$ in this work, the bound for $\boldsymbol{\gamma}_T$ is only meaningful when $ML<N$.

VIII. Experimental Results

A simulation study is conducted in order to evaluate the performance of our algorithms compared to other algorithms. All the experiments include 500 independent trials. In each trial, every entry of the dictionary matrix $\boldsymbol{\Phi}\in\mathbb{R}^{M\times N}$ is generated by a Gaussian distribution with unit variance as in [51]. The matrix $\tilde{\boldsymbol{S}}\in\mathbb{R}^{N\times L}$ is the source matrix whose columns are sources. The matrix $\tilde{\boldsymbol{A}}$ is generated by a Gaussian distribution where the norm of each row of $\tilde{\boldsymbol{A}}\in\mathbb{R}^{L\times L}$ is

set to be one. The matrix $\tilde{\boldsymbol{X}}$ is obtained by $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}$. The observation matrix $\boldsymbol{Y}$ is constructed by $\boldsymbol{Y} = \boldsymbol{\Phi}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}+\boldsymbol{N}$ where $\boldsymbol{N}$ is a zero-mean Gaussian noise with variance σ^2.

Two criteria are used to evaluate the algorithms. One criterion is the percentage of the support truly recovered, i.e.,

$$\alpha = \frac{e_{\hat{\boldsymbol{X}}}}{\mathcal{R}(\tilde{\boldsymbol{X}})}, \tag{49}$$

where $e_{\hat{\boldsymbol{X}}}$ is the number of row indexes in the support of $\tilde{\boldsymbol{X}}$ that is not in the support of $\hat{\boldsymbol{X}}$. The solution matrix $\tilde{\boldsymbol{X}}$ has K nonzero rows with K being calculated in each experiment after generating $\tilde{\boldsymbol{X}}$. The support of $\hat{\boldsymbol{X}}$ includes the indexes of the K rows with the largest ℓ_2 norms (Note that K is not an input to the proposed algorithms). Then, α is calculated. In order to evaluate the performance of support recovery, the mean and median of α are calculated in (49) over all trials. The median is also used because some values of α may be unusually large for a small number of runs which leads to the calculated mean to be non-informative. In noisy scenarios, the normalized mean square error (MSE) is also used as a measure of performance given by

$$\text{Normalized MSE} = \frac{\|\hat{\boldsymbol{X}} - \tilde{\boldsymbol{X}}\|_{\mathcal{F}}^2}{\|\tilde{\boldsymbol{X}}\|_{\mathcal{F}}^2}. \tag{50}$$

Normalized MSE and α defined in (49) and (50) describe the estimation performance of $\tilde{\boldsymbol{X}}$. In order to evaluate performance of the estimate of $\tilde{\boldsymbol{A}}$, Amari error, e_A, defined in [69] is used:

$$e_A = \sum_{i=1}^{L}\left(\sum_{j=1}^{L}\frac{|\boldsymbol{H}_{(i,j)}|}{\max_k |\boldsymbol{H}_{(i,k)}|} - 1\right) + \sum_{j=1}^{L}\left(\sum_{i=1}^{L}\frac{|\boldsymbol{H}_{(i,j)}|}{\max_k |\boldsymbol{H}_{(k,j)}|} - 1\right), \tag{51}$$

where the matrix $\boldsymbol{H}$ equals $\tilde{\boldsymbol{A}}\hat{\boldsymbol{A}}^{-1}$. Intuitively, Amari error can be employed to measure how much matrix $\boldsymbol{H}$ is close to $\boldsymbol{DP}$ for a diagonal matrix $\boldsymbol{D}$ and a permutation matrix $\boldsymbol{P}$. Since the traditional MMV algorithms do not estimate $\boldsymbol{A}$, one of the successful BSS algorithms called EBM in [70] is used to obtain $\hat{\boldsymbol{A}}$ from $\hat{\boldsymbol{X}}$.

In the experiments, our ICASBL and FASTICASBL algorithms are compared with the following algorithms:

- T-MSBL, proposed in [60]. [1]
- Compressive multiple signal classification (CSMUSIC), proposed in [71]. [2]
- The regularized Multiple FOCUSS (MFOCUSS), the regularized MFOCUSS algorithm proposed in [16].[3] The re-weighted p-norm of $\boldsymbol{X}$ appears in the MFOCUSS cost function. We set $p = 0.8$, as suggested by the authors.
- We also set $p = 1$ for MFOCUSS to evaluate the performance of the reweighing ℓ_1 minimization.

For trials, the noise variances are known for all algorithms except ICASBL, FASTICASBL and T-MSBL.

[1]The MATLAB code was downloaded at http://sccn.ucsd.edu/~zhang/TMSBL_code.zip

[2]The MATLAB code was downloaded at http://bispl.weebly.com/compressive-music.html

[3]The MATLAB code was downloaded at http://dsp.ucsd.edu/~zhilin/MFOCUSS.m

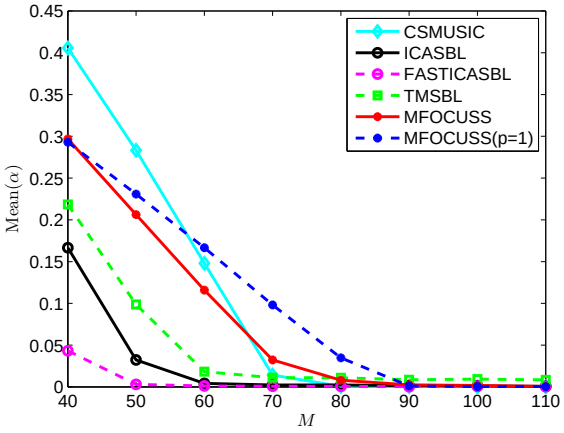

Fig. 2. Mean of α in terms of number of measurements M.

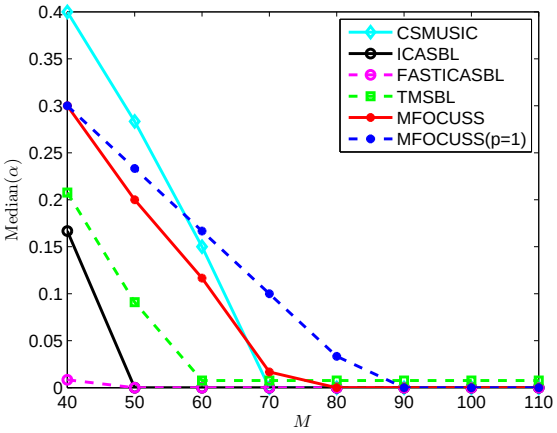

Fig. 3. Median of α in terms of number of measurements M.

The experiments have been carried out using the MATLAB software on an Intel Core i7 CPU 2.7 GHz processor and 16 GB RAM.

A. Recovery of sources with different numbers of measurements

In this experiment, the performance of the algorithms is studied in terms of number of measurements. The dictionary matrix $\boldsymbol{\Phi}$ is of size $M \times N$, where $N = 125$ and M is increased from 40 to 110. The matrix $\tilde{\boldsymbol{S}}$ is of the size 125×6 where each column has 10 nonzero elements. The location of the nonzero elements of $\tilde{\boldsymbol{s}}_i$ for $1 \leq i \leq 6$, are independent uniform random integer variables from 1 to 125.

In Fig. 2, the mean of α over all runs is presented. As expected, all algorithms performed better as the number of measurements M increases. The ICASBL approach outperforms the other approaches especially when the number of samples is low. This means that if the solution matrix $\tilde{\boldsymbol{X}}$ can be decomposed in to two matrices $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{A}}$, where matrix $\tilde{\boldsymbol{S}}$ has independent components, then the performance of support recovery is increased dramatically. Fig. 3 shows the median of α in terms of number of measurements. The median for

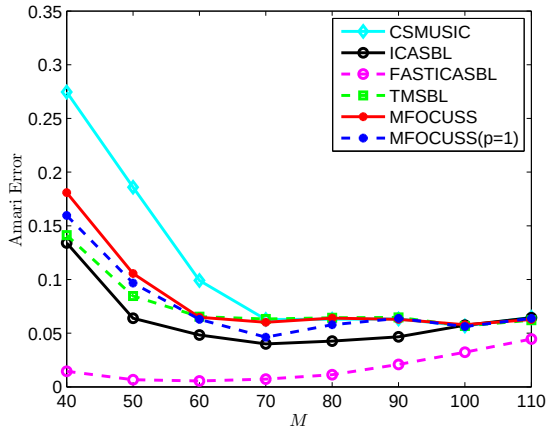

Fig. 4. Amari error in terms of number of measurements M.

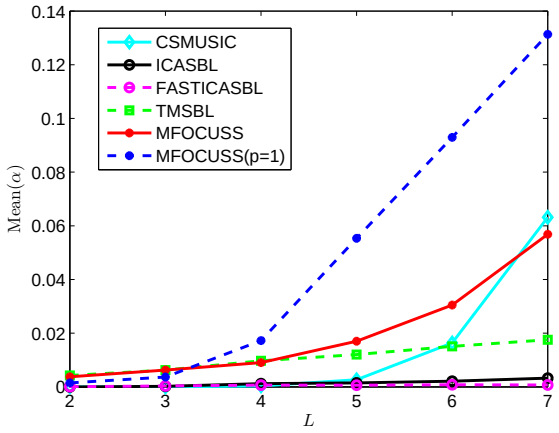

Fig. 5. Mean of α in terms of number of sources L.

the ICASBL is zero for the scenarios with M more than 60. Fig. 4 shows Amari error in terms of number of measurements. Our algorithms estimate the mixing matrix better than other algorithms. There are two facts that need to be considered. First, the mixing matrix estimation accuracy depends on how well $\tilde{\boldsymbol{X}}$ is estimated. Second, the estimation performance of mixing matrix depends on the number of independent samples available [72]; i.e., $\mathcal{R}(\boldsymbol{S})$. The correlation structure among source samples deteriorates the ICA methods' performance or mixing matrix estimation accuracy [73]. Therefore, increasing M does not necessarily improve the performance of mixing matrix estimation.

B. Multiple measurement vectors with different number of independent sources

This experiment evaluates the performance of algorithms when different number of sources are available. The dictionary matrix $\boldsymbol{\Phi}$ is 70×125. The matrix $\tilde{\boldsymbol{S}}$ is of the size $125 \times L$ where each column has 10 nonzero elements. L is increased from 2 to 6. The location of the nonzero elements of $\tilde{\boldsymbol{s}}_i$ for $2 \leq i \leq 7$, are uniform random integer variables from 1 to 125. The value of each nonzero element is generated from a Laplacian distribution with unit variance.

Fig. 5 and Fig. 6 show the mean and the median of α in terms of the number of independent components increasing from 2 to 7. Since each column of $\tilde{\boldsymbol{S}}$ is independent from the other columns, higher number of nonzero rows of $\tilde{\boldsymbol{X}}$ is obtained by increasing the number of independent sources. Therefore, the higher number of independent components, or higher $\sum_{i=1}^{L} r_i$, results in worse recovery performance for different algorithms while the number of measurements M is the same (See Theorem 2). Both proposed algorithms outperform the other algorithms as the number of measurements is fixed. The ICASBL and FASTICASBL could improve the quality of recovery dramatically compared to other approaches as the number of independent components becomes higher.

Fig. 7 shows the average time consumption for each algorithm, i.e. if t_i is the processing time for an algorithm to solve the problem in the i'th trial, then, the average

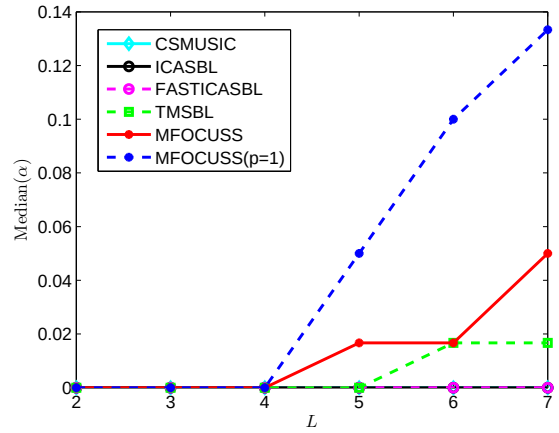

Fig. 6. Median of α in terms of number of sources L.

time consumption is $\frac{1}{N_t}\sum_{i=1}^{N_t} t_i$ where N_t is number of trials. It can be seen that the performance improvement by ICASBL and FASTICASBL is at the expense of increased time consumption. The time consumption grows as the number of independent sources increases. In comparison with TMSBL algorithm, the FASTICASBL and ICASBL algorithms have higher number of hyperparameters to be estimated.

C. Recovery ability at different noise levels

In this experiment, the algorithms are evaluated under noisy observations condition. The signal to noise ratio (SNR), i.e. $\text{SNR} = \|\boldsymbol{\Phi}\tilde{\boldsymbol{X}}\|_{\mathcal{F}}^2/\|\boldsymbol{N}\|_{\mathcal{F}}^2$, is in the range from 40dB to -5dB. The dictionary matrix $\boldsymbol{\Phi}$ is 40×125. The matrix $\tilde{\boldsymbol{S}}$ is of the size 125×3, where each column has 10 nonzero elements. The location of the nonzero elements of $\tilde{\boldsymbol{s}}_i$ for $1 \leq i \leq 3$, are uniform random integer variables from 1 to 125.

Fig. 8, Fig. 9 and Fig. 10 show the average of α, median of α and normalized MSE over 100 runs in terms of SNR respectively. As the SNR increases, the performance for all approaches is improved. In the noisy scenario, ICASBL outperforms all the algorithms and the performance of FASTICASBL decreased compared to the noiseless scenario. The

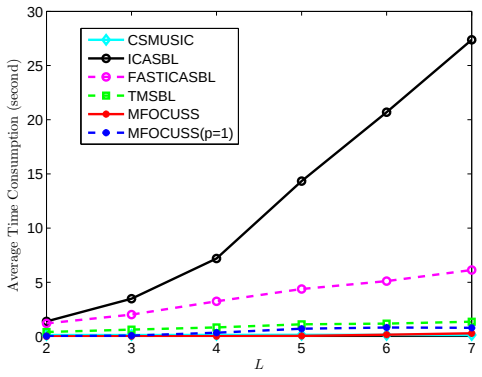

Fig. 7. Average Time Consumption in terms of number of sources L.

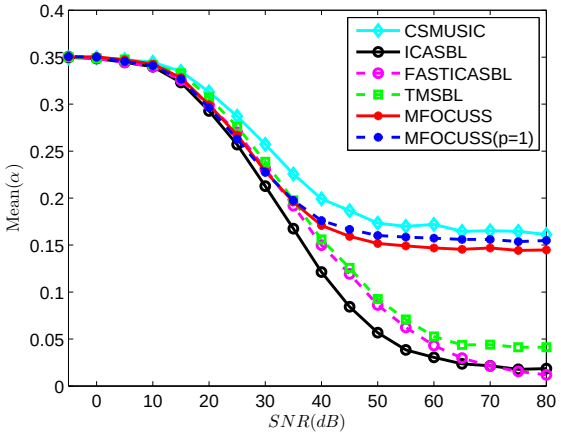

Fig. 8. Mean of α in terms of SNR.

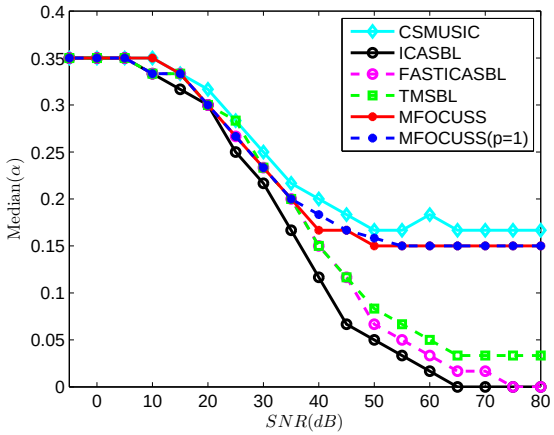

Fig. 9. Median of α in terms of SNR.

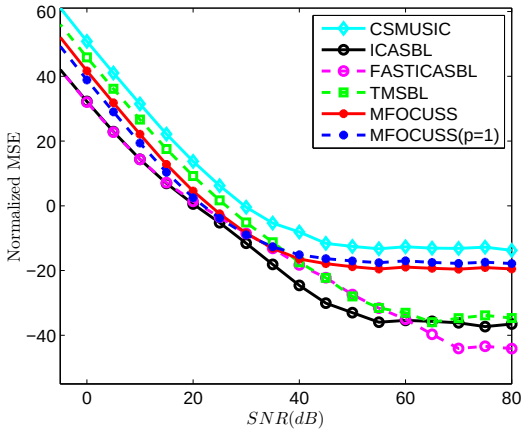

Fig. 10. Normalized MSE in terms of SNR.

deterioration can be attributed to the approximation applied on the covariance of noise in (37). In Fig 8, it can be seen that FASTICASBL provides a better performance when SNRs are larger than 70 dB.

Fig. 11 shows the Amari error in terms of SNR. All algorithms have better performances as SNR increases. Both proposed algorithms outperform competing algorithms for the SNR values larger than 10 dB.

D. Real Data

In this experiment we evaluate the algorithms for a real application. First we describe briefly the problem of heart rate estimation from a face recording.

PPG is an electro-optic scheme for measuring the tissue blood volume pulses [74]–[76]. The vital signs such as heart rate and respiratory rate can be estimated using extracted PPG. PPG can be monitored by pulse oximeter or remotely using a video camera recording showing a part of the skin, for example, the face.

The reason that rPPG can be measured by a recording is that the blood's hemoglobin absorbs light differently than other tissues over the time. When arterial blood volume changes during the cardiac cycles, light absorption of the human skin fluctuates [77]. Non-contact PPG or rPPG captures the color variations in time during the recording. Heart rate can be estimated by recording tiny color variations along with minor light intensity variation of the skin.

One of the main issues is the estimation of the PPG signal spectrum to find the frequency related to heart rate. FFT and DCT have low resolution when a low number of samples is available [31], for example, in the case of short videos [32]. In [7], [31], it has been shown that the sparsity assumption on PPG and rPPG signals results in a higher accuracy.

Generally a framework for heart rate estimation employs the following steps.

- Region of interest (ROI) registration: This step deals with finding a part of the recording suitable for rPPG estimation.
- ROI tracking: The ROI is found in the video frames.
- Preprocessing, noise and motion artifacts removal.
- Heart rate estimation.

The proposed algorithms can be employed in heart rate estimation. Since detection and tracking are not the purpose of this paper, we used face recording of a person sitting still in front of the video camera. The frame rate was 30 frames per second (fps). The forehead was selected as ROI. After

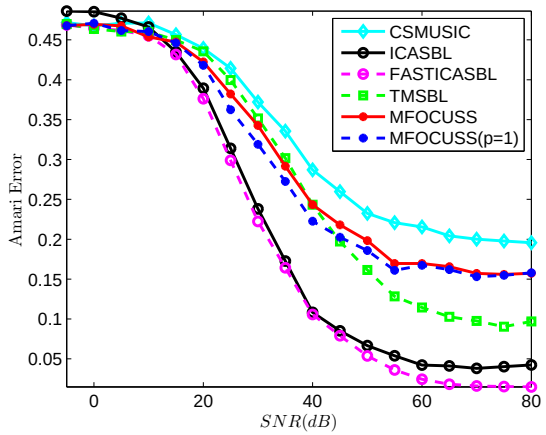

Fig. 11. Amari error in terms of SNR.

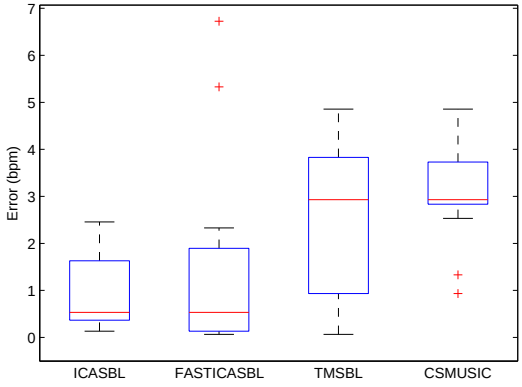

Fig. 12. The error of heart rate estimation.

tracking ROI, in the preprocessing step, three one-dimensional signals from RGB channels are extracted. Human heart rate frequency in resting condition is typically in the range of 40 beat per minute (bpm) to 180 bpm, where the signals were filtered in order to remove undesired coefficients. In this experiment we used partially known support assumption by removing the coefficients in the range of [0 - 0.6] Hz and [3 - 15] Hz. $\boldsymbol{\Phi}$ is the DCT transformation matrix including the columns corresponds to [0.6 - 3] Hz. We estimate the matrix $\hat{\boldsymbol{S}}$ whose columns are the estimated PPG signal, noise and motion artifacts in DCT domain. The matrix $\hat{\boldsymbol{A}}$ is the estimated mixture coefficients. A pulse oximeter measuring heart rate frequency was used as a reference.

The error is simply obtained by the absolute value of the difference of the estimated heart rate by the algorithms and that measured by the pulse oximeter. Fig. 12 shows the error of heart rate estimation by the algorithms. The red line in the middle of each box shows the median of the error. The edges of the box are the 25th percentile, q_1, and the 75th percentile, q_3. The sign "+" shows outliers which represent the values larger than $q_3 + 1.5(q_3 - q_1)$. As it can be seen in Fig. 12, the proposed algorithms outperformed CSMUSIC and TMSBL. Since MFOCUSS algorithm failed in the experiment, we removed the related results for a clearer presentation.

IX. Conclusion

We addressed a multiple measurement vector (MMV) model in practical scenarios, where the columns of the solution $\tilde{\boldsymbol{X}}$ were dependent and the number of measurements vectors was small. It was shown that existing algorithms performed poorly when the columns of the solution $\tilde{\boldsymbol{X}}$ could be decomposed into a sparse matrix $\tilde{\boldsymbol{S}}$ with independent columns and a full-rank square matrix $\tilde{\boldsymbol{A}}$. To address the problem, we considered the blind source separation (BSS) model for matrix $\boldsymbol{X}$ which could be decomposed into product of a source matrix with independent columns and a mixing matrix. Based on the model, we derived a new condition for uniqueness in the ℓ_0 norm minimization problem in (10). Based on this framework, we developed two algorithms; i.e., independent component analysis sparse Bayesian learning (ICASBL) and fast independent component analysis sparse Bayesian learning (FASTICASBL). The latter was faster but involves some approximations. Experiments showed that the proposed algorithms had superior performance compared to state-of-the-art algorithms. Based on the experimental results, FASTICASBL performed better than ICASBL and all other algorithms in noiseless scenarios, while ICASBL was the most successful algorithm in the noisy case. Theoretical analysis was also provided showing that the proposed sparse Bayesian learning (SBL) objective function had desirable global and local minima properties.

Appendix A
Proof of Theorem 1

Proof. If the two sides of (3) are multiplied with $\tilde{\boldsymbol{A}}^{-1}$, we obtain

$$\boldsymbol{Y}' = \boldsymbol{\Phi}\tilde{\boldsymbol{S}}, \tag{52}$$

where $\boldsymbol{Y}'$ is $\boldsymbol{Y}\tilde{\boldsymbol{A}}^{-1}$. Columns of $\boldsymbol{Y}'$ are independent from each other and the columns of $\tilde{\boldsymbol{S}}$ are also independent. Therefore, (11) is L SMV problems. All columns of $\tilde{\boldsymbol{S}}$ will be uniquely recovered, if

$$\|\tilde{\boldsymbol{s}}_i\|_0 < \frac{\text{Spark}(\boldsymbol{\Phi})}{2} \quad \text{for all } i. \tag{53}$$

The most restrictive condition would correspond to the column of $\boldsymbol{S}$ with the highest number of nonzero elements; i.e., $\max_i\{r_i\} < \text{Spark}(\boldsymbol{\Phi})/2$. The inequality in (53) proves the theorem. □

Appendix B
Proof of Theorem 2

Assume matrices $\{\tilde{\boldsymbol{S}}, \tilde{\boldsymbol{A}}\}$ and $\{\check{\boldsymbol{S}}, \check{\boldsymbol{A}}\}$ denote two solutions to the problem $(\boldsymbol{P}_0)$ such that $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}$ and $\check{\boldsymbol{X}} = \check{\boldsymbol{S}}\check{\boldsymbol{A}}$, where $\tilde{\boldsymbol{S}} \in \mathbb{R}^{N\times L}$, $\check{\boldsymbol{S}} \in \mathbb{R}^{N\times L}$, $\tilde{\boldsymbol{A}} \in \mathbb{R}^{L\times L}$ and $\check{\boldsymbol{A}} \in \mathbb{R}^{L\times L}$. The matrices $\tilde{\boldsymbol{S}}$ and $\check{\boldsymbol{S}}$ have the same number of nonzero elements. Note that unlike previous MMV works, since $\mathcal{R}(\boldsymbol{X})$ is not minimized, $\mathcal{R}(\tilde{\boldsymbol{X}})$ and $\mathcal{R}(\check{\boldsymbol{X}})$, are not necessarily the same.

One can write using (3),

$$\boldsymbol{0} = \boldsymbol{\Phi}\left(\tilde{\boldsymbol{X}} - \check{\boldsymbol{X}}\right). \tag{54}$$

By multiplying $\tilde{\boldsymbol{A}}^{-1}$ to both sides of (54), the following equation is obtained

$$\mathbf{0} = \boldsymbol{\Phi}\left(\tilde{\boldsymbol{S}} - \check{\boldsymbol{X}}\tilde{\boldsymbol{A}}^{-1}\right) = \boldsymbol{\Phi}\boldsymbol{V}, \tag{55}$$

where $\tilde{\boldsymbol{S}} - \check{\boldsymbol{X}}\tilde{\boldsymbol{A}}^{-1} = \boldsymbol{V}$. Using (55), one can write that

$$\begin{aligned}\|\boldsymbol{v}_i\|_0 \leq & \max_i \{\|\tilde{\boldsymbol{s}}_i\|_0\} + \mathcal{R}\left(\check{\boldsymbol{X}}\tilde{\boldsymbol{A}}^{-1}\right) \\ \leq & \max_i\{r_i\} + \sum_{j=1}^{L} r_j,\end{aligned} \tag{56}$$

where $\boldsymbol{v}_i$ is the i'th column of $\boldsymbol{V}$. The condition on Spark($\boldsymbol{\Phi}$) in (12); i.e., $\max_i\{r_i\} + \sum_{j=1}^{L} r_j < \text{Spark}(\boldsymbol{\Phi})$, implies that for any vector $\boldsymbol{v}$ such that $\boldsymbol{v} \in \text{Ker}(\boldsymbol{\Phi})$, the number of nonzero elements of $\boldsymbol{v}$ must be larger than $\max_i\{r_i\} + \sum_{j=1}^{L} r_j$. Using (56), the number of nonzero elements of $\boldsymbol{v}_i$ is less than Spark($\boldsymbol{\Phi}$). Therefore, $\boldsymbol{V} = \mathbf{0}$ or $\check{\boldsymbol{X}} = \check{\boldsymbol{X}}$. This completes the theorem.

Appendix C
Proof of Theorem 3

Let $\{\tilde{\boldsymbol{S}}, \tilde{\boldsymbol{A}}\}$ and $\{\check{\boldsymbol{S}}, \check{\boldsymbol{A}}\}$ be the solutions to (10). If the condition in (12) is satisfied, Theorem 2 proves that the matrix $\tilde{\boldsymbol{X}}$, obtained by minimizing (10), is unique. Therefore,

$$\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}} = \check{\boldsymbol{S}}\check{\boldsymbol{A}}. \tag{57}$$

By multiplying $\check{\boldsymbol{A}}^{-1}$ to the both sides of (57), it is obtained that $\tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}\check{\boldsymbol{A}}^{-1} = \check{\boldsymbol{S}}$. Since the columns of $\tilde{\boldsymbol{S}}$ have disjoint supports and the number of nonzero elements of $\tilde{\boldsymbol{S}}$ and $\check{\boldsymbol{S}}$ are equal, $\tilde{\boldsymbol{A}}\check{\boldsymbol{A}}^{-1}$ has one nonzero element in each row. Since $\tilde{\boldsymbol{A}}\check{\boldsymbol{A}}^{-1}$ is full-rank and square, one can write $\tilde{\boldsymbol{A}}\check{\boldsymbol{A}}^{-1} = \boldsymbol{P}\boldsymbol{D}$ for a permutation matrix $\boldsymbol{P}$ and a diagonal matrix $\boldsymbol{D}$. Due to the fact that the ℓ_2 norm of each row of $\tilde{\boldsymbol{A}}$ and $\check{\boldsymbol{A}}$ equals 1, then, $\boldsymbol{D}$ is a unit matrix. This completes the theorem.

Appendix D
Details of Parameter Update

Here the details of the derivation of hyperparameters are described. We maximize $\mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta}\left\{p\left(\boldsymbol{y}^v, \boldsymbol{s}^v; \boldsymbol{\Gamma}, \sigma^2\right)\right\} = \mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta}\left\{p\left(\boldsymbol{y}^v|\boldsymbol{s}^v; \boldsymbol{A}, \sigma^2\right) p\left(\boldsymbol{s}^v; \boldsymbol{\Gamma}\right)\right\}$ using fix point iteration in terms of hyperparameters by treating $\boldsymbol{s}^v$ as hidden variable

$$\begin{aligned}\gamma_{ij}^{(k+1)} = & \arg\max_{\gamma_{ij}>0} \mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta^{(k)}}\left\{p\left(\boldsymbol{y}^v, \boldsymbol{s}^v; \Theta^{(k)}\right)\right\} \\ = & \arg\max_{\gamma_{ij}>0} \mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v,\Theta^{(k)}}\left\{p(\boldsymbol{s}^v; \boldsymbol{\Gamma}^{(k)})\right\} \\ = & \mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}_{(i,j)}^2\right\} \\ = & \mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta^{(k)}}\left\{\left(\boldsymbol{s}^v_{((i-1)N+j)}\right)^2\right\}.\end{aligned} \tag{58}$$

Likewise, an update rule for σ^2 can be incorporated as [51]

$$\begin{aligned}\left(\sigma^2\right)^{(k+1)} = & \frac{\|\boldsymbol{y} - \boldsymbol{\phi}_{\boldsymbol{A}^{(k)}}\boldsymbol{\mu}^{(k)}\|_2^2}{ML} \\ + & \frac{\left(\sigma^2\right)^{(k)}\sum_{i=1}^{L}\sum_{j=1}^{L}\left(1 - \left(\gamma_{ij}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{s}^v}^{(k)}\right)_{(i+j)}\right)}{ML}.\end{aligned} \tag{59}$$

Only the term $p\left(\boldsymbol{y}^v|\boldsymbol{s}^v; \boldsymbol{A}, \sigma^2\right)$ depends on $\boldsymbol{A}$, we can estimate $\boldsymbol{A}$ by minimizing $\mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta^{(k)}}\left\{\|\boldsymbol{y}^v - \boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{s}^v\|_2^2\right\}$,

$$\begin{aligned}&\mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta^{(k)}}\left\{\|\boldsymbol{y}^v - \boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{s}^v\|_2^2\right\} = \mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta^{(k)}}\left\{\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A}\|_{\mathcal{F}}^2\right\} \\ &= \mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\text{Tr}\left(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A}\right)\left(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A}\right)^T\right\} \\ &= \mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\text{Tr}(-2\boldsymbol{Y}^T\boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A} + \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S})\right\}.\end{aligned} \tag{60}$$

The matrix $\boldsymbol{A}$ is obtained by solving the following equation

$$\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\nabla_{\boldsymbol{A}}\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A}\|_{\mathcal{F}}^2\right\} = 0. \tag{61}$$

Using (60), one can write (61) as

$$\begin{aligned}&\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\nabla_{\boldsymbol{A}}\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A}\|_{\mathcal{F}}^2\right\} \\ &= \mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{-2\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{Y} + 2\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{A}\right\} \\ &= -2\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\right\}\boldsymbol{Y} + 2\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}\right\}\boldsymbol{A}.\end{aligned} \tag{62}$$

Then, the matrix $\boldsymbol{A}$ is updated by the following equation

$$\boldsymbol{A}^{(k+1)} = \left(\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}\right\}\right)^{-1}\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\right\}\boldsymbol{Y}.$$

We find $\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\right\}$ as

$$\begin{aligned}\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\right\} = & \mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\right\}\boldsymbol{\Phi}^T \\ = & \text{vec}_{L,N}^{-1}\left(\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\text{vec}_{L,N}\left(\boldsymbol{S}^T\right)\right\}\right)\boldsymbol{\Phi}^T \\ = & \text{vec}_{L,N}^{-1}\left(\boldsymbol{\mu}\right)\boldsymbol{\Phi}^T.\end{aligned} \tag{63}$$

With some algebraic effort, the i'th and j'th element of the matrix $\boldsymbol{\kappa} = \mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}\right\}$ is found as

$$\begin{aligned}\boldsymbol{\kappa}_{(i,j)} = & \mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{s}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{s}_j\right\} \\ = & \mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\text{Tr}\left(\boldsymbol{s}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{s}_j\right)\right\} \\ = & \text{Tr}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{s}_j\boldsymbol{s}_i^T\right\}\right).\end{aligned} \tag{64}$$

In order to find $\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{s}_j\boldsymbol{s}_i^T\right\}$, the term $\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}_{(i,j)}\boldsymbol{S}_{(v,p)}\right\}$ should be obtained

$$\begin{aligned}&\mathbf{E}_{\boldsymbol{S}|\boldsymbol{Y};\Theta^{(k)}}\left\{\boldsymbol{S}_{(i,j)}\boldsymbol{S}_{(v,p)}\right\} = \mathbf{E}_{\boldsymbol{s}^v|\boldsymbol{y}^v;\Theta^{(k)}}\left\{\boldsymbol{s}^v_{(j+L(i-1))}\boldsymbol{s}^v_{(p+L(v-1))}\right\} \\ &= \left(\boldsymbol{\Sigma}_{\boldsymbol{s}^v}^{(k)}\right)_{((j+L(i-1)),(p+L(v-1)))} + \boldsymbol{\mu}^{(k)}_{(j+L(i-1))}\boldsymbol{\mu}^{(k)}_{((p+L(v-1)))}.\end{aligned} \tag{65}$$

Appendix E
Proof of Theorem 4

For the sake of convenience we consider the equivalent model (17) instead of (15). The cost function from (27) includes two terms: the logarithm of determinant of $\boldsymbol{\Sigma}_{\boldsymbol{y}^v}$ and $\left(\boldsymbol{y}^v\right)^T\boldsymbol{\Sigma}_{\boldsymbol{y}^v}^{-1}\boldsymbol{y}^v$. While the later term is strictly greater than zero and bounded by some finite bound C for all $\boldsymbol{\Gamma}$, $\boldsymbol{A}$ and σ^2, the first term goes to minus infinity by reducing the determinant of $\boldsymbol{\Sigma}_{\boldsymbol{y}^v}$ to zero. As such, the minimum of $\mathcal{L}(\Theta)$ occurs whenever

$$|\boldsymbol{\Sigma}_{\boldsymbol{y}^v}| = |\sigma^2\boldsymbol{I} + \boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{\Gamma}\boldsymbol{\phi}_{\boldsymbol{A}}^T| = 0, \tag{66}$$

while maintaining some finite bound C such that

$$0 \leq \left(\boldsymbol{y}^v\right)^T\left(\sigma^2\boldsymbol{I} + \boldsymbol{\phi}_{\boldsymbol{A}}\boldsymbol{\Gamma}\boldsymbol{\phi}_{\boldsymbol{A}}^T\right)^{-1}\boldsymbol{y}^v \leq C. \tag{67}$$

For demonstrating that $\tilde{\gamma}$ or equivalently $\tilde{\mathbf{\Gamma}}$ satisfies the condition in (66), we find a bound for $\|\tilde{\gamma}\|_0$. Using (12) one can write for $\sum_{i=1}^{L}\|\tilde{\boldsymbol{s}}_i\|_0$ that

$$\begin{aligned}\max_i\{r_i\}+\sum_{j=1}^{L}r_j <&\text{Spark}(\mathbf{\Phi})\leq M+1\\ \sum_{j=1}^{L}r_j <&M+1-\max_i\{r_i\}\\ \|\tilde{\gamma}\|_0 <&M.\end{aligned}\tag{68}$$

It shows that Rank$(\tilde{\mathbf{\Gamma}})$ is less than M. The condition (66) is satisfied with $\mathbf{\Gamma}=\tilde{\mathbf{\Gamma}}$, $\sigma^2=0$ and Rank$(\tilde{\mathbf{\Gamma}})<$ Rank$(\boldsymbol{\phi}_{\boldsymbol{A}})=$ Rank$(\mathbf{\Phi}\otimes\boldsymbol{A})=ML$. This happens because

$$\text{Rank}(\mathbf{\Sigma}_{\boldsymbol{y}^v})=\text{Rank}(\boldsymbol{\phi}_{\boldsymbol{A}}\tilde{\mathbf{\Gamma}}\boldsymbol{\phi}_{\boldsymbol{A}}^T)\leq\|\tilde{\gamma}\|_0<M.\tag{69}$$

Therefore, $\mathbf{\Sigma}_{\boldsymbol{y}^v}\in\mathbb{R}^{ML\times ML}$ is not full-rank, and its determinant is zero as in (66) which completes the proof. It is shown in Theorem 3 that if the condition (12) is satisfied the other global minimum of (10) are different up to permutation of the solution $\tilde{\boldsymbol{S}}$. These solutions are also acceptable due to unavoidable scale and permutation ambiguities which emanate from BSS problem if no further assumption is made on sources and the mixing matrix. Therefore, the other acceptable solutions are $\boldsymbol{S}_*=\tilde{\boldsymbol{S}}\boldsymbol{P}$ with the variance $\gamma_*=\text{diag}(\mathbf{\Gamma}_*)$ and $\boldsymbol{A}_*=\boldsymbol{P}^{-1}\tilde{\boldsymbol{A}}$ for some permutation matrix $\boldsymbol{P}$. These solutions are also global minima of $\mathcal{L}$ in (27). This is simply true because $\boldsymbol{Y}=\mathbf{\Phi}\boldsymbol{S}_*\boldsymbol{A}_*$ and the number of nonzero elements of $\boldsymbol{S}_*$ equals to $\tilde{\boldsymbol{S}}$, or $\|\gamma_*\|_0=\|\tilde{\gamma}\|_0$.

Appendix F
Proof of Lemma 2

The proof follows along the lines of Lemma 2 in [51]. If $\boldsymbol{A}$ is fixed, the constraint $(\boldsymbol{y}^v)^T(\sigma^2\boldsymbol{I}+\boldsymbol{\phi}_{\boldsymbol{A}}\mathbf{\Gamma}\boldsymbol{\phi}_{\boldsymbol{A}}^T)^{-1}\boldsymbol{y}^v=C$ is subsumed by the constraint $(\sigma^2\boldsymbol{I}+\boldsymbol{\phi}_{\boldsymbol{A}}\mathbf{\Gamma}\boldsymbol{\phi}_{\boldsymbol{A}}^T)^{-1}\boldsymbol{y}^v=\boldsymbol{u}$. By some manipulations, we will obtain

$$\begin{aligned}\boldsymbol{y}^v-\sigma^2\boldsymbol{u}=&(\sigma^2\boldsymbol{I}+\boldsymbol{\phi}_{\boldsymbol{A}}\mathbf{\Gamma}\boldsymbol{\phi}_{\boldsymbol{A}}^T)\boldsymbol{u}-\sigma^2\boldsymbol{u}\\ =&(\boldsymbol{\phi}_{\boldsymbol{A}}\mathbf{\Gamma}\boldsymbol{\phi}_{\boldsymbol{A}}^T)\boldsymbol{u}\\ =&\boldsymbol{\phi}_{\boldsymbol{A}}\text{diag}\left(\boldsymbol{\phi}_{\boldsymbol{A}}^T\boldsymbol{u}\right)\gamma.\end{aligned}\tag{70}$$

This completes the proof.

Appendix G
Proof of Theorem 5

The proof follows along the lines of Theorem 2 in [51] by considering the fact that the

$$\text{Rank}(\boldsymbol{\phi}_{\boldsymbol{A}})=\text{Rank}(\mathbf{\Phi}\otimes\boldsymbol{A}^T)\leq ML.\tag{71}$$

Consider the optimization problem

$$\begin{aligned}\min:\quad & f(\gamma)\\ \text{subject to}\quad & \boldsymbol{G}\gamma=\boldsymbol{b},\gamma\geq\boldsymbol{0},\end{aligned}\tag{72}$$

where $\boldsymbol{b}$ and $\boldsymbol{G}$ are defined in (46) and (47), and $f(\gamma)=\log(|\mathbf{\Sigma}_{\boldsymbol{y}^v}|)$. It can be seen that the optimization problem (72) is optimizing a concave function over a closed, bounded convex polytope. Clearly, for a fixed $\boldsymbol{A}$, any local minimum of $\mathcal{L}$, e.g. γ^*, must also be a local minimum (72) optimization problem with

$$C=(\boldsymbol{y}^v)^T\boldsymbol{u}=(\boldsymbol{y}^v)^T((\sigma^*)^2\boldsymbol{I}+\boldsymbol{\phi}_{\boldsymbol{A}}\mathbf{\Gamma}^*\boldsymbol{\phi}_{\boldsymbol{A}}^T)^{-1}\boldsymbol{y}^v.\tag{73}$$

According to [78, Theorem 6.5.3], the minimum of (40) is achieved at an extreme point. Furthermore, based on the Theorem in [78, Chapter 2.5] the extreme point is a basic feasible solution, i.e., a solution with at most ML nonzero values. Therefore, the local minima are at sparse solutions.

Appendix H
Proof of Lemma 3

The proof follows along the lines of Lemma 3 in [51]. For the sake of simplicity we assume that $\|\hat{\gamma}\|_0=M$, then, there are M basis of $\boldsymbol{\phi}_{\hat{\boldsymbol{A}}}$ corresponding to nonzero elements of $\hat{\gamma}$, where $\boldsymbol{\phi}_{\hat{\boldsymbol{A}}}=\mathbf{\Phi}\otimes\hat{\boldsymbol{A}}^T$ and $\hat{\boldsymbol{A}}$ is the estimated mixing matrix. Lets denote $\bar{\mathbf{\Gamma}}$, $\bar{\gamma}$ and $\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}$ as the nonzero diagonal elements of $\hat{\mathbf{\Gamma}}$, the nonzero elements of $\hat{\gamma}$ and the corresponding columns of $\boldsymbol{\phi}_{\hat{\boldsymbol{A}}}$ respectively. The likelihood can be written as

$$\begin{aligned}L=&\log\left(\left|\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}\bar{\mathbf{\Gamma}}\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}^T\right|\right)+\boldsymbol{y}^T\left(\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}\bar{\mathbf{\Gamma}}\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}^T\right)^{-1}\boldsymbol{y}\\ =&\log\left(\left|\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}\right|\left|\bar{\mathbf{\Gamma}}\right|\left|\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}^T\right|\right)+\boldsymbol{y}^T\left(\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}\bar{\mathbf{\Gamma}}\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}^T\right)^{-1}\boldsymbol{y}\\ =&2\log\left(\left|\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}\right|\right)+\sum_{i=1}^{M}\log\bar{\gamma}_{(i)}+\sum_{i=1}^{M}\frac{\bar{\boldsymbol{s}}_{(i)}^2}{\bar{\gamma}_{(i)}},\end{aligned}\tag{74}$$

where $\bar{\boldsymbol{s}}_{(i)}^2=\bar{\boldsymbol{\phi}}_{\hat{\boldsymbol{A}}}^{-1}\boldsymbol{y}$. Then, the gradient of (74) is formed as

$$\frac{\partial L}{\partial\bar{\gamma}_{(i)}}=\frac{1}{\bar{\gamma}_{(i)}}-\frac{\bar{\boldsymbol{s}}_{(i)}^2}{\bar{\gamma}_{(i)}^2}.\tag{75}$$

By setting (75) to zero, we obtain $\bar{\gamma}_{(i)}=\bar{\boldsymbol{s}}_{(i)}^2$ and clearly, $\|\hat{\boldsymbol{s}}\|_0=\|\hat{\gamma}\|_0$ which completes the lemma.

References

[1] Y. C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, 1997.

[4] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[5] J. Romberg, "Imaging via compressive sampling," *IEEE Signal Process. Magazine*, vol. 25, no. 2, pp. 14–20, 2008.

[6] M. F. Duarte, M. A. Davenport, D. Takbar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE signal Process. magazine*, vol. 25, no. 2, pp. 83–91, 2008.

[7] Z. Zhang, "Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 8, pp. 1902–1910, 2015.

[8] Z. Zhang, Z. Pi, and B. Liu, "TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 522–531, 2015.

[9] S. Aviyente, "Compressed sensing framework for EEG compression," in *Statistical Signal Process., 2007. SSP'07. IEEE/SP 14th Workshop on*. IEEE, 2007, pp. 181–184.

[10] T. Strohmer and B. Friedlander, "Compressed sensing for MIMO radar-algorithms and performance," in *Signals, Syst. and Computers, 2009 Conf. Record of the Forty-Third Asilomar Conf. on*. IEEE, 2009, pp. 464–468.

[11] Y. Yu, A. P. Petropulu, and H. V. Poor, "MIMO radar using compressive sampling," *IEEE J. of Sel. Topics in Signal Process.*, vol. 4, no. 1, pp. 146–163, 2010.

[12] S. Gogineni and A. Nehorai, "Target estimation using sparse modeling for distributed MIMO radar," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5315–5325, 2011.

[13] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," in *Proc. of the 5th Int. Conf. on Inf. Process. in sensor networks*. ACM, 2006, pp. 134–142.

[14] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[15] A. Shukla and A. Majumdar, "Row-sparse blind compressed sensing for reconstructing multi-channel EEG signals," *Biomed. Signal Process. and Control*, vol. 18, pp. 174–178, 2015.

[16] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, 2005.

[17] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, 2006.

[18] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, 2011.

[19] B. S. Kim and S. K. Yoo, "Motion artifact reduction in photoplethysmography using independent component analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 566–568, 2006.

[20] J. Yao and S. Warren, "A short study to assess the potential of independent component analysis for motion artifact separation in wearable pulse oximeter signals," in *Engineering in Medicine and Biology Soc., 2005. IEEE-EMBS 2005. 27th Annual Int. Conf. of the*. IEEE, 2005, pp. 3585–3588.

[21] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley Interscience, New York, 2001.

[22] A. Cichocki, "Blind signal processing methods for analyzing multichannel brain signals," *International Journal of Bioelectromagnetism*, vol. 6, no. 1, pp. 22–27, 2004.

[23] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[24] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal processing letters*, vol. 4, no. 4, pp. 112–114, 1997.

[25] S. Cruces, A. Cichocki, and L. De Lathauwer, "Thin QR and SVD factorizations for simultaneous blind signal extraction," in *Signal Processing Conference, 2004 12th European*. IEEE, 2004, pp. 217–220.

[26] A. Cichocki and R. Unbehauen, "Robust neural networks with online learning for blind identification and blind separation of sources," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 43, no. 11, pp. 894–906, 1996.

[27] A. Cichocki, R. Unbehauen, and E. Rummert, "Robust learning algorithm for blind separation of signals," *Electronics Letters*, vol. 30, no. 17, pp. 1386–1387, 1994.

[28] A. Cichocki, R. E. Bogner, L. Moszczyński, and K. Pope, "Modified herault–jutten algorithms for blind separation of sources," *Digital signal processing*, vol. 7, no. 2, pp. 80–93, 1997.

[29] G. De Haan and A. Van Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological measurement*, vol. 35, no. 9, p. 1913, 2014.

[30] F. Peng, Z. Zhang, X. Gou, H. Liu, and W. Wang, "Motion artifact removal from photoplethysmographic signals by combining temporally constrained independent component analysis and adaptive filter," *Biomedical engineering online*, vol. 13, no. 1, p. 50, 2014.

[31] S. H. Fouladi, I. Balasingham, T. A. Ramstad, and K. Kansanen, "Accurate heart rate estimation from camera recording via music algorithm," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 7454–7457.

[32] S. H. Fouladi, I. Balasingham, K. Kansanen, and T. A. Ramstad, "Extracting remote photoplethysmogram signal from endoscopy videos for vessel and capillary density recognition," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 227–230.

[33] J. Bobin, J.-L. Starck, J. M. Fadili, and Y. Moudden, "Sparsity and morphological diversity in blind source separation," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2662–2674, 2007.

[34] F. Feng and M. Kowalski, "An unified approach for blind source separation using sparsity and decorrelation," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 1736–1740.

[35] J. Bobin, J.-L. Starck, Y. Moudden, and M. J. Fadili, "Blind source separation: The sparsity revolution," *Advances in Imaging and Electron Physics*, vol. 152, no. 1, pp. 221–302, 2008.

[36] C. Chenot, J. Bobin, and J. Rapin, "Robust sparse blind source separation," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2172–2176, 2015.

[37] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[38] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[39] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[40] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. of the Royal Statistical Soc. Series B (Methodological)*, pp. 267–288, 1996.

[41] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[42] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Acoust., speech and signal Process., 2008. ICASSP 2008. IEEE Int. Conf. on*. IEEE, 2008, pp. 3869–3872.

[43] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, 1999.

[44] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, 2008.

[45] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 53–63, 2010.

[46] J. P. Vila and P. Schniter, "An empirical-Bayes approach to recovering linearly constrained non-negative sparse signals," *IEEE Trans. Signal Process.*, vol. 62, no. 18, pp. 4689–4703, 2014.

[47] R. Giri and B. D. Rao, "Bootstrapped sparse Bayesian learning for sparse signal recovery," in *Signals, Syst. and Computers, 2014 48th Asilomar Conf. on*. IEEE, 2014, pp. 1657–1661.

[48] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001.

[49] A. C. Faul and M. E. Tipping, "Analysis of sparse Bayesian learning," *Adv. in neural Inf. Process. Syst.*, vol. 1, pp. 383–390, 2002.

[50] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3704–3716, 2007.

[51] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.

[52] D. P. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6236–6255, 2011.

[53] Z. Zhang and B. D. Rao, "Sparse signal recovery in the presence of correlated multiple measurement vectors," in *Acoust. Speech and Signal Process. (ICASSP), 2010 IEEE Int. Conf. on*. IEEE, 2010, pp. 3986–3989.

[54] K. Qiu and A. Dogandzic, "Variance-component based sparse signal reconstruction and model selection," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 2935–2952, 2010.

[55] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[56] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, 2010.

[57] J. A. Tropp, A. C. Gilbert and M. J. Strauss, "Simultaneous sparse approximation via greedy pursuit," in *Acoust., Speech, and Signal Process., 2005. Proc.(ICASSP'05). IEEE Int. Conf. on*, vol. 5. IEEE, 2005, pp. v–721.

[58] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.

[59] Z. Zhang and B. D. Rao, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 2009–2015, 2013.

[60] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE J. of Sel. Topics in Signal Process.*, vol. 5, no. 5, pp. 912–926, 2011.

[61] Z. Zhang and B. D. Rao, "Recovery of block sparse signals using the framework of block sparse Bayesian learning," in *Acoust., Speech and Signal Process. (ICASSP), 2012 IEEE Int. Conf. on*. IEEE, 2012, pp. 3345–3348.

[62] Z. Zhang, "Comparison of sparse signal recovery algorithms with highly coherent dictionary matrices: the advantage of T-MSBL," *Research note*, 2012.

[63] S.-E. Chiu and B. Rao, "Correlation learning on joint support recovery for more sources than measurements," in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2016 IEEE*. IEEE, 2016, pp. 1–5.

[64] Z. Zhang, T.-P. Jung, S. Makeig, and B. D. Rao, "Compressed sensing of EEG for wireless telemonitoring with low energy consumption and inexpensive hardware," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 221–224, 2013.

[65] J. Bobin, Y. Moudden, J.-L. Starck, and M. Elad, "Morphological diversity and source separation," *IEEE Signal Process. Lett.*, vol. 13, no. 7, pp. 409–412, 2006.

[66] M. J. Fadili, J.-L. Starck, J. Bobin, and Y. Moudden, "Image decomposition and separation using sparse representations: An overview," *Proc. of the IEEE*, vol. 98, no. 6, pp. 983–994, 2010.

[67] Z. Zhang, "Heart rate monitoring from wrist-type photoplethysmographic (PPG) signals during intensive physical exercise," in *Signal and Inf. Process. (GlobalSIP), 2014 IEEE Global Conf. on*. IEEE, 2014, pp. 698–702.

[68] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[69] A. Cichocki and H. Yang, "A new learning algorithm for blind signal separation," *Adv. in neural Inf. Process. Syst.*, vol. 8, pp. 757–763, 1996.

[70] X.-L. Li and T. Adali, "Independent component analysis by entropy bound minimization," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5151–5164, 2010.

[71] J. M. Kim, O. K. Lee, and J. C. Ye, "Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing," *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 278–301, 2012.

[72] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[73] Y.-m. Cheung and L. Xu, "Dual multivariate auto-regressive modeling in state space for temporal signal separation," *IEEE Transactions on Syst., Man, and Cybernetics, Part B (Cybernetics)*, vol. 33, no. 3, pp. 386–398, 2003.

[74] A. B. Hertzman, "Observations on the finger volume pulse recorded photoelectrically," *Am. J. Physiol.*, vol. 119, pp. 334–335, 1937.

[75] A. Kamal, J. Harness, G. Irving, and A. Mearns, "Skin photoplethysmographya review," *Computer methods and programs in biomedicine*, vol. 28, no. 4, pp. 257–269, 1989.

[76] A. Reisner, P. A. Shaltis, D. McCombie, and H. H. Asada, "Utility of the photoplethysmogram in circulatory monitoring," *Anesthesiology: The Journal of the American Society of Anesthesiologists*, vol. 108, no. 5, pp. 950–958, 2008.

[77] T. Aoyagi, N. Kobayashi, and T. Sasaki, "Apparatus for determining the concentration of a light-absorbing material in blood," May 23 1989, uS Patent 4,832,484.

[78] D. G. Luenberger and Y. Ye, *Linear and nonlinear programming*. Springer, 2015, vol. 228.