Source Separation with Side Information Based on Gaussian Mixture Models With Application in Art Investigation

Zahra Sabetsarvestani[1], Francesco Renna[2], Franz Kiraly[3], Miguel Rodrigues[1]
[1]Department of Electronic and Electrical Engineering, University College London,
[2]Instituto de Telecomunicações and Faculdade de Ciências da Universidade do Porto
[3]Department of Statistical Science, University College London

Abstract—In this paper, we propose an algorithm for source separation with side information where one observes the linear superposition of two source signals plus two additional signals that are correlated with the mixed ones. Our algorithm is based on two ingredients: first, we learn a Gaussian mixture model (GMM) for the joint distribution of a source signal and the corresponding correlated side information signal; second, we separate the signals using standard computationally efficient conditional mean estimators.

The paper also puts forth new recovery guarantees for this source separation algorithm. In particular, under the assumption that the signals can be perfectly described by a GMM model, we characterize necessary and sufficient conditions for reliable source separation in the asymptotic regime of low-noise as a function of the geometry of the underlying signals and their interaction. It is shown that if the subspaces spanned by the innovation components of the source signals with respect to the side information signals have zero intersection, provided that we observe a certain number of linear measurements from the mixture, then we can reliably separate the sources; otherwise we cannot.

Our proposed framework – which provides a new way to incorporate side information to aid the solution of source separation problems where the decoder has access to linear projections of superimposed sources and side information – is also employed in a real-world art investigation application involving the separation of mixtures of X-ray images. The simulation results showcase the superiority of our algorithm against other state-of-the-art algorithms.

I. INTRODUCTION

Blind source separation (BSS) – where the aim is to recover individual unobserved source signals from some mixture of these signals – is a very relevant problem arising in various applications such as:

- *Audio source separation* [1]–[4] – also usually referred to as the "cocktail party" problem – involving the extraction of one or more source signals of interest from the observation of several audio recordings which contain different linear combinations of the source signals.
- *Image separation* [5]–[7] involving the decomposition of a given image into different components, usually associated with different semantic meanings or different

phenomena or sources. For example, in astronomical imaging one often wishes to separate stars from filaments for further analysis [7], [8]. In this case, image acquisition can be done via the use of compressive sensing, thus leading to compressive source separation frameworks.

- *Hyperspectral Unmixing* involving the decomposition of mixed pixels into endmembers and corresponding abundance maps of endmembers. In hyperspectral imaging, one pixel consists of the reflectance spectra of several materials where each coefficient in the mixture corresponds to abundance of each material [9]. Hyperspectral imaging is often used to determine what materials are present in a scene [10], [11].
- *Multi-user digital communications* [12]–[14] where one wishes to extract a message associated with a single user given the superposition of signals associated with various users conveyed over a communications channel [12].

These intrinsically ill-defined inverse problems, which often require additional prior knowledge/assumptions about the mixed signals, such as sparsity [15]–[17], have been the subject of intensive research for the past two decades [15], [16], [18], [19].

The availability of side information in general inverse problems – beyond source separation – is indeed known to bring about various benefits. For example, in compressive sensing, where the aim is to reconstruct high-dimensional signals from low-dimensional measurements, it is now well-established that the availability of side information can lead to improved reconstruction accuracy [20]–[24]. Likewise, in the separation of mixed signals, it is also known that the availability of side information – again, in the form of other signals correlated with the signals of interest– can also lead to better separation accuracy [6]. However, current approaches to source separation with side information also have various limitations [2]–[4], [6] including:

- *Performance*: As it will be shown by the numerical results reported in Section VI, informed source separation (ISS) approaches such as [6], [16] exhibit poor separation performance, leading to significant cross-interference between the separated sources. This is possibly due to the inability of separating morphologically similar contents or due to challenges in learning discriminative dictionar-

ies operating at different scales.

- *Complexity*: Informed source separation approaches such as [6] also exhibit high computational complexity, both during the training phase and the testing phase.
- *Guarantees*: Current informed source separation approaches also lack separation guarantees, i.e., conditions for identifiability of the individual sources given the mixed one in terms of the underlying properties of the individual and mixed signals.

Motivated by these limitations, we build upon our recent work [21], [24] to propose a new approach to separate linearly mixed signals in the presence of side information, leveraging a joint Gaussian mixture model (GMM) for the joint distribution of the source and side information signals. Our proposed approach can address various limitations associated with previous approaches:

- First, our method outperforms existing ones such as [6], [16]. This is due, in part, to the fact that GMM priors have been shown to deliver outstanding results in various applications such as image processing [25], [26], video compression [27] and dictionary learning [28], [29]. The GMMs can be seen as the Bayesian counterpart of the union-of-subspaces model, where each subspace corresponds to the image of the (possibly low rank) covariance matrix of each Gaussian component within the GMM [21], [24], [29], [30].
- Second, our method is also less complex than existing ones such as [6]. This is, in turn, due to the fact that *i*) we learned a GMM model using very efficient algorithms such as an expectation maximization (EM) algorithm [31] or the non-parametric approach described in [28] and *ii*) one can also find a closed-form solution for the signal separation problem when assuming that the signals are described by a GMM model.
- Third, our approach is very flexible in modeling a wide range of signals such as patches extracted from natural images and videos, portions of hyperspectral data cubes, speech features and so on [21], [25], [29]. Note also that GMM distributions can approximate any distribution with arbitrary precision (by adding further components to the mixture) [32].
- Finally, GMM models are amenable to mathematical analysis, enabling one to determine identifability conditions for a wide range of problems [21], [29].

A. Contributions

Our main contributions are:

1) We propose a new approach to source separation with side information based on GMM models of the source and side information signals.
2) We characterize the identifiability properties associated with source separation with side information – in the form of necessary and sufficient conditions for reliable separation in the asymptotic regime of low noise – as a function of the geometry of the source signals, the side information signals, and their interaction.

3) We provide a number of synthetic results showcasing that our theoretical identifiability conditions align with practice.
4) Finally, we provide a number of results associated with a real-world problem – involving the separation of mixed X-ray signals – that demonstrate that our approach leads to better results than competing ones, in terms of performance and complexity [6], [16].

Part of this work was previously presented in [33]. This work extends significantly [33] by adding the following contributions.

- New theoretical results: the theorems proved in this work are now valid for the more general case in which GMMs can contain non-zero mean Gaussian components. On the other hand, the results in [33] were valid only for the case of zero-mean distributions.
- A real-world application of the proposed source separation framework is studied in this work. In particular, an image separation problem with side information for art investigation is considered.
- Detailed proofs of all theoretical results are reported in this work.

B. Organization

The remainder of the paper is organized as follows. In Section II, we review the relevant source separation literature, including approaches to source separation with and without side information. Section III proposes our approach to source separation with side information. Section IV provides identifiability conditions associated with source separation with side information. Sections V and VI offer a number of results with synthetic and real data showcasing the merits of our approach. Finally, conclusions are drawn in Section VII. The proofs of the main results appear in the Appendices.

C. Notation

We now set up the notation used in the remainder of the paper. Matrices are denoted by upper-case bold characters ($\mathbf{A}$) and vectors are denoted by lower-case bold letters ($\mathbf{a}$). The identity matrix of dimension $n \times n$ is denoted by $\mathbf{I}_n$ and $\mathbf{0}_{m \times n}$ represents the all zero matrix of dimension $m \times n$. The operators transpose, trace, rank, and Moore-Penrose pseudo inverse are denoted by $(\cdot)^T$, $\mathrm{tr}(\cdot)$, $\mathrm{rank}(\cdot)$, and $(\cdot)^{\dagger}$, respectively. $\mathrm{Im}(\cdot)$ and $\mathrm{Null}(\cdot)$ denote the image and null space of a matrix, respectively, and $\dim(\cdot)$ denotes the dimension of a linear subspace. The Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbb{E}(\cdot)$ denotes the expectation operator.

II. RELATED WORK

The source separation problem has attracted considerable attention in the past two decades. We review main approaches to blind source separation and informed source separation.

A. Blind Source Separation

The BSS problem involves the estimation of individual sources merely from observing their mixture. This highly ill-posed problem has been approached by imposing different constraints on the source signals, e.g. non-Gaussianity [18], low-rank [34], sparsity [15], [17], or prior distributions [35], [36].

A widely used technique for BSS is independent component analysis (ICA), which assumes the individual source signals are non-Gaussian and independent. The separation is done by minimizing the mutual information between source signals, where separation guarantees – up to a permutation and scaling – can also be provided if the number of linear mixtures exceeds the number of sources, and at most one of the sources is Gaussian [37]. Various algorithms have been developed to implement ICA, e.g., FastICA [38], RADICAL [39], JADE [40].

Sparsity is also often used to enable BSS approaches [15], [17]. In particular, sparsity-based BSS algorithms exploit the fact that different source signals can often be described as linear combinations of a few atoms from a dictionary. For example, morphological component analysis (MCA) [16] is a well-known sparsity-based source separation algorithm that takes advantage of both sparsity and morphological diversity to separate a linear mixture onto its different components. More specifically, it assumes that each source signal has a sparse representation over a specific dictionary that is not sparse over the dictionaries of the other components.

Finally, it is also common to impose priors on the source signals to carry out BSS within a Bayesian framework [15], [35], [41]. In [42], a unified Bayesian inference framework based on Markov chain Monte Carlo algorithm is proposed in order to recover the source signals from the mixture, based on a hidden Markov model of the wavelet coefficients associated with the source signals. Reference [35] studies the image separation problem where the prior density of pixels are constructed using Markov random fields (MRF) based on a statistical model of the gradient image. The MRF is then solved numerically using a modified Gibbs sampling.

B. Informed Source Separation

In turn, ISS aims to recover the source signals from a mixture of them in the presence of some additional information which can be incorporated to aid the task in hand. Broadly speaking, the ISS algorithms can be divided into two main categories: signal processing based algorithms [2]–[4] and learning based algorithms [43]–[45].

A relatively recent signal processing based algorithm relies on the assumption that the source signals are known in the encoding (mixing) stage. This framework is motivated by some key audio applications where mixing and demixing can be processed separately by cooperative users [3]. In the encoder some descriptors – which consist in the characteristic parameters of the source signals – are extracted from each source and embedded into the mixture. In the decoder the descriptors are extracted to be used to separate each source signal from the mix signal.

On the other hand, learning based source separation algorithms rely on the availability of training data to carry out the source separation task. Different levels of supervision have been considered: semi-supervised setting and fully-supervised setting.

1) Semi-supervised setting: In the semi-supervised setting, it has been assumed that the available training data only contains examples of source signals [44], [46] (or occasionally the source signals plus additional side information [6]); for example, in separating instruments from music recordings, one can have access to examples of one or more solo recordings. The training data is used to learn a generative model for source signals. This generative model is then used to carry out the separation task.

A number of semi-supervised source separation methods have been proposed in the last decade. Non-negative Matrix Factorization (NMF) [44], [46]–[48], a popular method for source separation, tackles the problem in a semi-supervised setting where it has been assumed that one has access to examples of signals at a training stage. This prior information is exploited to pre-compute the dictionaries that accurately represent the signals [44]. The learned dictionaries are then used to solve the separation task.

More recently, [6] proposes a semi-supervised algorithm based on a coupled dictionary approach for X-ray image separation exploiting side information. In this scenario, the X-ray signals to be separated have similar morphological characteristics, thus, one of the fundamental assumptions for the use of MCA is not verified. The proposed approach [6] instead couples the two available imaging modalities, X-ray and RGB, by using a coupled dictionary learning algorithm to capture both the inherent similarities and the discrepancies among heterogeneous correlated data. The coupled dictionaries are learned from a set of training data which contains examples of pairs of X-ray and RGB images. Then, in the separation stage, sparsity constraints are imposed to solve the source separation problem.

2) Fully-supervised setting: In the fully-supervised setting, it is assumed that we have access to a set of training data containing pairs of mixtures and corresponding individual signals where the source separation problem boils down to learning a mapping from mixture to unmixed signals [43], [45].

In this setting, the algorithms exploit full supervision to learn a mapping from the mixture to the individual source components (e.g. [43], [45]). Unfortunately, in various applications, one rarely has access to full supervision in order to take advantage of fully-supervised source separation approaches.

In this paper, we consider a semi-supervised (compressive) source separation framework where we only have access to examples of the source signals and, in addition, other signals correlated with the source signals.

III. PROPOSED SOURCE SEPARATION APPROACH

Our approach aims to decompose a linear mixture of two signals into its constituents from a set of linear (possibly compressive) observations:

$$\mathbf{v} = \mathbf{\Phi}(\mathbf{x}_1 + \mathbf{x}_2) + \mathbf{n}, \tag{1}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n_x}$ represents a random linear mixing matrix (operator) drawn from a rotationally invariant distribution. Note that random projections, as those computed via the observation matrix $\boldsymbol{\Phi}$, have been shown to capture salient features of natural signals in an efficient way and without requiring adaptation to the specific application [49], [50]. Moreover, rotationally invariant distributions guarantee incoherence between the projections and the domain where the signals of interest have a sparse representation with high probability. This fact leads to the possibility of guaranteeing stable recovery of the signal of interest from a reduced number of projections [51]. The vector $\mathbf{n}$ is a zero-mean white Gaussian noise, i.e., $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$, where $\sigma^2 > 0$ represents the noise variance, $\mathbf{v} \in \mathbb{R}^m$ is a vector representing the noisy mixed signal, and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{n_x}$ are vectors representing the signals of interest. It is assumed throughout that the number of measurements m extracted from the linear mixture is such that $m < n_x$.

We also consider that the decoder has access to additional side information signals

$$\mathbf{u}_1 = \mathbf{y}_1 + \mathbf{n}_1 \tag{2}$$
$$\mathbf{u}_2 = \mathbf{y}_2 + \mathbf{n}_2, \tag{3}$$

where $\mathbf{y}_1 \in \mathbb{R}^{n_{y_1}}$ is correlated with $\mathbf{x}_1$, $\mathbf{y}_2 \in \mathbb{R}^{n_{y_2}}$ is correlated with $\mathbf{x}_2$, and $\mathbf{n}_1 \in \mathbb{R}^{n_{y_1}}$ and $\mathbf{n}_2 \in \mathbb{R}^{n_{y_2}}$ represent zero-mean Gaussian noise, i.e., $\mathbf{n}_1 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_{y_1}})$ and $\mathbf{n}_2 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_{y_2}})$. We further assume that $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \mathbf{n}_1$ and $\mathbf{n}_2$ are statistically independent. Note that, although the proposed analysis could be generalized to the case when side information signals and the mixture are affected by noise with different variances, in order to simplify the exposition, all noise variances are equal to the same value σ^2.

Given the model in (1), (2), and (3), we now propose a source separation algorithm that is based on two main ingredients. The first ingredient in our separation algorithm involves learning a joint GMM model connecting each individual signal to the corresponding side information signal, given a number of pairs of samples of these signals. In particular, we consider the sets of labels $\mathcal{I} = \{1, \dots, \alpha\}$ and $\mathcal{J} = \{1, \dots, \beta\}$, where $\mathcal{I}$ and $\mathcal{J}$ are associated with the pairs $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$, which are formed from the source signals and side information signals, respectively. Then, we model $\mathbf{x}_1$ and $\mathbf{y}_1$ via a joint Gaussian mixture distribution, i.e.,

$$p(\mathbf{x}_1, \mathbf{y}_1) = \sum_{i \in \mathcal{I}} p_I(i) p(\mathbf{x}_1, \mathbf{y}_1 | I = i), \tag{4}$$

where $p_I(i), i \in \mathcal{I}$ is the probability distribution of the label I over the set of values $\mathcal{I}$ and $p(\mathbf{x}_1, \mathbf{y}_1 | I = i)$ is a multivariate Gaussian distribution

$$p(\mathbf{x}_1, \mathbf{y}_1 | I = i) = \mathcal{N}(\boldsymbol{\mu}_{x_1 y_1}^{(i)}, \bar{\boldsymbol{\Sigma}}_{x_1 y_1}^{(i)}), \tag{5}$$

with mean $\boldsymbol{\mu}_{x_1 y_1}^{(i)}$ and covariance matrix:

$$\bar{\boldsymbol{\Sigma}}_{x_1 y_1}^{(i)} = \begin{bmatrix} \boldsymbol{\Sigma}_{x_1}^{(i)} & \boldsymbol{\Sigma}_{x_1 y_1}^{(i)} \\ \boldsymbol{\Sigma}_{y_1 x_1}^{(i)} & \boldsymbol{\Sigma}_{y_1}^{(i)} \end{bmatrix}. \tag{6}$$

Also, conditioned on a particular label value $I = i$, the vectors $\mathbf{x}_1$ and $\mathbf{y}_1$ are assumed to be Gaussian distributed.

In the same way, we also model $\mathbf{x}_2$ and $\mathbf{y}_2$ via a Gaussian mixture distribution, i.e.,

$$p(\mathbf{x}_2, \mathbf{y}_2) = \sum_{j \in \mathcal{J}} p_J(j) p(\mathbf{x}_2, \mathbf{y}_2 | J = j), \tag{7}$$

where $p_J(j), j \in \mathcal{J}$ is the probability distribution of the label J over the set of values $\mathcal{J}$ and

$$p(\mathbf{x}_2, \mathbf{y}_2 | J = j) = \mathcal{N}(\boldsymbol{\mu}_{x_2 y_2}^{(j)}, \bar{\boldsymbol{\Sigma}}_{x_2 y_2}^{(j)}), \tag{8}$$

with mean $\boldsymbol{\mu}_{x_2 y_2}^{(j)}$ and covariance matrix:

$$\bar{\boldsymbol{\Sigma}}_{x_2 y_2}^{(j)} = \begin{bmatrix} \boldsymbol{\Sigma}_{x_2}^{(j)} & \boldsymbol{\Sigma}_{x_2 y_2}^{(j)} \\ \boldsymbol{\Sigma}_{y_2 x_2}^{(j)} & \boldsymbol{\Sigma}_{y_2}^{(j)} \end{bmatrix}. \tag{9}$$

Also, conditioned on a particular label value $J = j$, the vectors $\mathbf{x}_2$ and $\mathbf{y}_2$ are assumed to be Gaussian distributed.

Conditioned on the component labels $(I, J) = (i, j) \in \mathcal{L} = \mathcal{I} \times \mathcal{J}$, the joint distribution of the two source signals $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1^T \mathbf{x}_2^T \end{bmatrix}^T$ and of the two side information signals $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1^T \mathbf{y}_2^T \end{bmatrix}^T$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{xy}^{(ij)}, \boldsymbol{\Sigma}_{xy}^{(ij)})$, with mean

$$\boldsymbol{\mu}_{xy}^{(ij)} = \begin{bmatrix} \boldsymbol{\mu}_{x_1 y_1}^{(i)}{}^T & \boldsymbol{\mu}_{x_2 y_2}^{(j)}{}^T \end{bmatrix}^T \tag{10}$$

and covariance:

$$\boldsymbol{\Sigma}_{xy}^{(ij)} = \begin{bmatrix} \boldsymbol{\Sigma}_{x_1}^{(i)} & \mathbf{0} & \boldsymbol{\Sigma}_{x_1 y_1}^{(i)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{x_2}^{(j)} & \mathbf{0} & \boldsymbol{\Sigma}_{x_2 y_2}^{(j)} \\ \boldsymbol{\Sigma}_{y_1 x_1}^{(i)} & \mathbf{0} & \boldsymbol{\Sigma}_{y_1}^{(i)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{y_2 x_2}^{(j)} & \mathbf{0} & \boldsymbol{\Sigma}_{y_2}^{(j)} \end{bmatrix}. \tag{11}$$

These joint GMM models can be easily learned using an EM algorithm.

The second ingredient in our source separation algorithm involves leveraging the GMM to separate the two source signals given the mixed signal and the associated side information signals. In particular, we can write the mean-squared error associated with the recovery of the two individual source signals as follows:

$$\text{MSE}_{\mathbf{x}|\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2}(\sigma^2) = \mathbb{E}\left[\|\mathbf{x} - f(\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2)\|^2\right], \tag{12}$$

where $f(\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2)$ represents an estimator delivering an estimate of $\mathbf{x}_1$ and $\mathbf{x}_2$ given $\mathbf{v}, \mathbf{u}_1$ and $\mathbf{u}_2$, and, likewise, we can also write the minimum mean-squared error (MMSE) associated with the recovery of the two signals – where the minimum is with respect to all possible estimators of $\mathbf{x}_1$ and $\mathbf{x}_2$ given $\mathbf{v}, \mathbf{u}_1$ and $\mathbf{u}_2$ – as follows:

$$\text{MMSE}_{\mathbf{x}|\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2}(\sigma^2) = \mathbb{E}\left[\|\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2]\|^2\right], \tag{13}$$

where $\mathbb{E}[\mathbf{x}|\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2]$ corresponds to the conditional mean estimator of $\mathbf{x}_1$ and $\mathbf{x}_2$ given $\mathbf{v}, \mathbf{u}_1$ and $\mathbf{u}_2$ that can also be written in closed-form as in (14)-(15), at the top of next page [21], [28].

This conditional mean estimator – which we also adopt to reconstruct the original source signals given the mixed signal and the associated side information signals – has various advantages. It exhibits a computational complexity increasing linearly with the number of components associated with the joint GMM, involving only computation of simple algebraic operations and matrix inversion operations (with the matrices

$$\mathbb{E}[\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2] = \sum_{i\in\mathcal{I},j\in\mathcal{J}} \pi_{I,J}(i,j)\left(\begin{bmatrix}\boldsymbol{\mu}_{x_1}^{(i)}\\\boldsymbol{\mu}_{x_2}^{(j)}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\Sigma}_{x_1}^{(i)} & \mathbf{0} & \boldsymbol{\Sigma}_{x_1y_1}^{(i)} & \mathbf{0}\\\mathbf{0} & \boldsymbol{\Sigma}_{x_2}^{(j)} & \mathbf{0} & \boldsymbol{\Sigma}_{x_2y_2}^{(j)}\end{bmatrix}\boldsymbol{\Phi}_0^T(\boldsymbol{\Phi}_0\boldsymbol{\Sigma}_{xy}^{(ij)}\boldsymbol{\Phi}_0^T + \sigma^2\mathbf{I})^{-1}(\mathbf{w} - \boldsymbol{\Phi}_0\boldsymbol{\mu}_{xy}^{(ij)})\right),$$
(14)

where $\boldsymbol{\mu}_{x_1y_1}^{(i)} = \begin{bmatrix}\boldsymbol{\mu}_{x_1}^{(i)}\\\boldsymbol{\mu}_{y_1}^{(i)}\end{bmatrix}$, $\boldsymbol{\mu}_{x_2y_2}^{(j)} = \begin{bmatrix}\boldsymbol{\mu}_{x_2}^{(j)}\\\boldsymbol{\mu}_{y_2}^{(j)}\end{bmatrix}$, $\boldsymbol{\Phi}_0 = \begin{bmatrix}\boldsymbol{\Phi} & \boldsymbol{\Phi} & \mathbf{0} & \mathbf{0}\\\mathbf{0} & \mathbf{0} & \mathbf{I}_{n_{y_1}} & \mathbf{0}\\\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{n_{y_2}}\end{bmatrix}$, $\mathbf{w} = \begin{bmatrix}\mathbf{v}\\\mathbf{u}_1\\\mathbf{u}_2\end{bmatrix}$, and

$$\pi_{I,J}(i,j) = \frac{p_I(i)p_J(j)\mathcal{N}(\mathbf{w};\boldsymbol{\Phi}_0\boldsymbol{\mu}_{xy}^{(ij)},\boldsymbol{\Phi}_0\boldsymbol{\Sigma}_{xy}^{(ij)}\boldsymbol{\Phi}_0^T + \sigma^2\mathbf{I})}{\sum_{(i,j)\in\mathcal{L}} p_I(i)p_J(j)\mathcal{N}(\mathbf{w};\boldsymbol{\Phi}_0\boldsymbol{\mu}_{xy}^{(ij)},\boldsymbol{\Phi}_0\boldsymbol{\Sigma}_{xy}^{(ij)}\boldsymbol{\Phi}_0^T + \sigma^2\mathbf{I})}.$$
(15)

dimensions scaling with m, n_x, n_{y_1}, and n_{y_2}, so of moderate size for standard patch sizes). A few tens of components were shown to be sufficient to model patches extracted from images or frames of a video [21], [29]. In contrast, competing approaches such as [6] can exhibit much higher computational complexity.

In summary, our procedure to separate two mixed signals given side information involves two main steps:

1) Learning a GMM model connecting each individual signal and the side information given a number of samples associated with the individual signals and the side information. Once again, this can be done efficiently using a well-known EM algorithm [31] or the non-parametric approach described in [28].
2) Employing a simple conditional mean estimator to recover the two individual signals given the linear mixture signal and the side information signals. This can be done via the closed-form expression reported in (14)-(15).

A. Learning a GMM model

In order to learn the GMM priors that will be leveraged to separate the signals of interest from the observed mixture, we assume that we have access to N_θ samples of the source signals $\mathbf{x}_1,\mathbf{x}_2$ and the associated side information signals $\mathbf{y}_1,\mathbf{y}_2$. From these samples, $2N_\theta$ vectors are generated as:

$$\boldsymbol{\theta}_1^q = \left[(\mathbf{x}_1^q)^T,(\mathbf{y}_1^q)^T\right]^T, \quad \boldsymbol{\theta}_2^q = \left[(\mathbf{x}_2^q)^T,(\mathbf{y}_2^q)^T\right]^T,$$

for $q = 1,\ldots,N_\theta$. Then, the vectors $\{\boldsymbol{\theta}_1^q\}_{q=1}^{N_\theta}$ are used to determine the parameters of the GMM prior modeling $\mathbf{x}_1$ and $\mathbf{y}_1$, whereas the vectors $\{\boldsymbol{\theta}_2^q\}_{q=1}^{N_\theta}$ are used to determine the parameters of the GMM prior modeling $\mathbf{x}_2$ and $\mathbf{y}_2$.

First, the number of Gaussian components in each GMM, i.e., the values α and β, are fixed. Then, the GMM parameters are learned using an EM algorithm [31] tailored to this problem. Namely, the EM algorithm is applied independently to the sets of vectors $\{\boldsymbol{\theta}_1^q\}_{q=1}^{N_\theta}$, and $\{\boldsymbol{\theta}_2^q\}_{q=1}^{N_\theta}$. In the following, we will provide details of the application of the EM algorithm to the training set $\{\boldsymbol{\theta}_1^q\}_{q=1}^{N_\theta}$. Similar steps can be applied to the training set $\{\boldsymbol{\theta}_2^q\}_{q=1}^{N_\theta}$.

The EM algorithm is an iterative algorithm that starts from some initial estimate of the parameters and then proceeds to iteratively update the parameters until convergence. Each iteration consists of two main steps: the expectation step and maximization step.

- Expectation step: In this step, we use the GMM parameters obtained as output of the previous iteration of the EM to compute the membership weights c_i^q for all data points $\boldsymbol{\theta}_1^q$, $\forall q \in \{1,\cdots,N_\theta\}$ and for all mixture components $\forall i \in \{1,\cdots,\alpha\}$ (see [31] for details).
- Maximization step: In this step, the calculated membership weights are used to update the parameter estimates. Let N_i denote the sum of membership weights for component i, defined as $N_i = \sum_{q=1}^{N_\theta} c_i^q$. This is the effective number of data samples associated with component i. Then, the parameters of the GMM are updated with the following expressions:

$$p_I(i) \leftarrow N_i/N_\theta \tag{16}$$

$$\boldsymbol{\mu}_{x_1y_1}^{(i)} \leftarrow \frac{\sum_{q=1}^{N_\theta} c_i^q\boldsymbol{\theta}_1^q}{N_i} \tag{17}$$

$$\bar{\boldsymbol{\Sigma}}_{x_1y_1}^{(i)} \leftarrow \frac{\sum_{q=1}^{N_\theta} c_i^q(\boldsymbol{\theta}_1^q - \boldsymbol{\mu}_{x_1y_1}^{(i)})(\boldsymbol{\theta}_1^q - \boldsymbol{\mu}_{x_1y_1}^{(i)})^T}{N_i}. \tag{18}$$

Then, expectation and maximization steps are applied alternatively until convergence and the GMM parameters obtained from the last iteration are retained. The details of the use of the EM algorithm in the estimation of the GMM models describing the vectors $\mathbf{x}_1,\mathbf{y}_1$ and $\mathbf{x}_2,\mathbf{y}_2$ are reported in Algorithm 1.

Algorithm 1 Learning Algorithm

Input: $\{\boldsymbol{\theta}_1^q\}_{q=1}^{N_\theta}$, $\{\boldsymbol{\theta}_2^q\}_{q=1}^{N_\theta}$, α, β, N_{iter}
Output: $\bar{\boldsymbol{\Sigma}}_{x_1y_1}^{(i)}$, $\bar{\boldsymbol{\Sigma}}_{x_2y_2}^{(j)}$, $\boldsymbol{\mu}_{x_1y_1}^{(i)}$, $\boldsymbol{\mu}_{x_2y_2}^{(j)}$, $p_I(i),p_J(j)$, $\forall(i,j)$
Initialization: Initialize $p_I(i)$, $\boldsymbol{\mu}_{x_1y_1}^{(i)}$, $\bar{\boldsymbol{\Sigma}}_{x_1y_1}^{(i)}$, $p_J(j)$, $\boldsymbol{\mu}_{x_2y_2}^{(j)}$, $\bar{\boldsymbol{\Sigma}}_{x_2y_2}^{(j)}$, $\forall(i,j)$
for $n = 1,\ldots,N_{\text{iter}}$ **do**
$\quad$ Expectation: compute the membership weights c_i^q of $\boldsymbol{\theta}_1^q$ and c_j^q of $\boldsymbol{\theta}_2^q$ with the current parameters, $\forall(i,j)$
$\quad$ Maximization: Update $p_I(i)$, $\boldsymbol{\mu}_{x_1y_1}^{(i)}$, $\bar{\boldsymbol{\Sigma}}_{x_1y_1}^{(i)}$ using (16)-(18) and update $p_J(j)$, $\boldsymbol{\mu}_{x_2y_2}^{(j)}$, $\bar{\boldsymbol{\Sigma}}_{x_2y_2}^{(j)}$ with equations akin to (16)-(18), $\forall(i,j)$
end for

B. Separating GMM Signals

The separation algorithm takes as input the observation vectors $\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2$, the projection matrix $\mathbf{\Phi}$, and the GMM parameters, which are assumed to be known by the decoder. The algorithm gives as output the estimates of the signals of interest $\mathbf{x}_1$ and $\mathbf{x}_2$. The pseudo code of the separation algorithm is detailed in Algorithm 2.

Algorithm 2 Separation Algorithm

Input: $\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{\Phi}, \mathbf{\Sigma}_{xy}^{(ij)}, \mu_{xy}^{(ij)}, p_I(i), p_J(j), \forall(i,j) \in \mathcal{L}$
Output: $\mathbb{E}[\mathbf{x}|\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2]$
for $\forall(i,j) \in \mathcal{L}$ **do**
 Find the posterior probability of each component
 label given the observation vector $\pi_{I,J}(i,j)$ from (15)
end for
Compute the estimate $\mathbb{E}[\mathbf{x}|\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2]$ using (14)

IV. Proposed Separation Approach: Identifiability Conditions

Our proposed separation approach also comes with identifiability guarantees. In particular, we now consider both necessary and sufficient conditions for

$$\lim_{\sigma^2 \to 0} \mathrm{MMSE}_{\mathbf{x}|\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2}(\sigma^2) = 0, \qquad (19)$$

entailing reliable separation of both source signals in the sense that the individual MMSEs associated with the reconstruction of $\mathbf{x}_1$ given $\mathbf{v}, \mathbf{u}_1$, and $\mathbf{u}_2$ and the reconstruction of $\mathbf{x}_2$ given $\mathbf{v}, \mathbf{u}_1$, and $\mathbf{u}_2$ approach zero in the asymptotic regime where $\sigma^2 \to 0$. This asymptotic regime is relevant in many practical signal and image processing scenarios where the variance of noise can be considered to be -60 dB or less [29]. We also consider that the relevant signals obey (exactly) the GMM models in (4)-(6) and (7)-(9). Note that the analysis of the separation error induced by the process of learning signal and side information distributions is not considered in this work. In fact, when a sufficient number of training samples is available, the adopted GMM distributions are shown to approximate any distribution with arbitrary precision, by adding the appropriate number of components to each mixture [32].

The identifiability conditions are expressed in terms of the number of measurements m and quantities that are related to the geometry of the signals and the measurement matrix. In particular, we will be using the following quantities:

- $r_{x_1}^{(i)} = \mathrm{rank}(\mathbf{\Sigma}_{x_1}^{(i)})$ represents the dimension of the subspace spanned by the source signal $\mathbf{x}_1$ and $r_{y_1}^{(i)} = \mathrm{rank}(\mathbf{\Sigma}_{y_1}^{(i)})$ represents the dimension of the subspace spanned by the side information signal $\mathbf{y}_1$, identified by component labels $I = i$.
- $r_{x_2}^{(j)} = \mathrm{rank}(\mathbf{\Sigma}_{x_2}^{(j)})$ represents the dimension of the subspace spanned by the source signal $\mathbf{x}_2$ and $r_{y_2}^{(j)} = \mathrm{rank}(\mathbf{\Sigma}_{y_2}^{(j)})$ represents the dimension of the subspace spanned by the side information signal $\mathbf{y}_1$, identified by component labels $J = j$.
- $r_{x_1 y_1}^{(i)} = \mathrm{rank}(\bar{\mathbf{\Sigma}}_{x_1 y_1}^{(i)})$ represents the dimension of the subspace spanned collectively by the source signal $\mathbf{x}_1$

and the side information signal $\mathbf{y}_1$, identified by the labels $I = i$.
- $r_{x_2 y_2}^{(j)} = \mathrm{rank}(\bar{\mathbf{\Sigma}}_{x_2 y_2}^{(j)})$ represents the dimension of the subspace spanned collectively by the source signal $\mathbf{x}_2$ and the side information signal $\mathbf{y}_2$, identified by the labels $J = j$.
- $r_{xy}^{(ij)} = \mathrm{rank}(\mathbf{\Sigma}_{xy}^{(ij)})$ represents the dimension of the subspace spanned collectively by the source signals and the side information signals identified by the component labels $(I, J) = (i, j)$. The superscripts are dropped when the results hold for all possible choice of labels or when the meaning is clear from the context.

Such conditions are also provided for the scenario where the covariance matrices $\mathbf{\Sigma}_{x_1}^{(i)}$, $\mathbf{\Sigma}_{y_1}^{(i)}$, $\mathbf{\Sigma}_{x_1 y_1}^{(i)}$, $\mathbf{\Sigma}_{x_2}^{(j)}$, $\mathbf{\Sigma}_{y_2}^{(j)}$ and $\mathbf{\Sigma}_{x_2 y_2}^{(j)}$ have low rank, smaller than or equal to the ambient dimensions. Moreover, the images of the covariance matrices corresponding to different Gaussian components are assumed to be drawn from random ensembles. Namely, on recalling the definition of a Grassmann manifold as the space of all k-dimensional linear subspaces of a given ambient linear space [52], we assume that the ranges of the covariance matrices are subspaces drawn from a continuous distribution defined over the corresponding Grassmann manifold. Note that the assumption on the subspaces associated with covariance matrices is plausible as it reflects well the behaviour of many real data ensembles for various applications such as face recognition, digits classification, etc. [53]. Moreover, it simplifies the statement of some of our theoretical results.

The identifiability conditions for GMM signals build upon identifiability conditions for Gaussian signals. We therefore treat each case separately, denoting the MMSE associated with the separation of Gaussian signals by $\mathrm{MMSE}_{\mathbf{x}|\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2}^{\mathrm{G}}$ and the MMSE associated with GMM signals by $\mathrm{MMSE}_{\mathbf{x}|\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2}^{\mathrm{GMM}}$.

The theoretical results in the next sections are based on the following Lemma, which states the geometrical interplay between the measurement matrix $\mathbf{\Phi}_0$ and the covariance matrices of source and side information signals

Lemma 1. Consider the following compact notation for the measurement model in (1)-(3):

$$\mathbf{w} = \mathbf{\Phi}_0 \mathbf{s} + \mathbf{n}_0, \qquad (20)$$

where $\mathbf{n}_0 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{m + n_{y_1} + n_{y_2}})$ and,

$$\mathbf{\Phi}_0 = \begin{bmatrix} \mathbf{\Phi} & \mathbf{\Phi} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{n_{y_1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{n_{y_2}} \end{bmatrix}, \ \mathbf{s} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \ \mathbf{w} = \begin{bmatrix} \mathbf{v} \\ \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}. \ (21)$$

Let $\mathbf{\Sigma}_{xy}$ be the matrix defined in (11), and $\mathbf{\Phi}_0$ be the matrix defined in (21) where $\mathbf{\Phi} \in \mathbb{R}^{m \times n_x}$ is a random matrix drawn from a rotationally invariant distribution. Then, with probability one, the rank of the matrix $\mathbf{\Phi}_0 \mathbf{\Sigma}_{xy}$ is given by:

$$r = \min\{r_{xy} - \mathcal{D}_{x|y}, m + r_{y_1} + r_{y_2}\}, \qquad (22)$$

where

$$\mathcal{D}_{x|y} = \dim(\mathrm{Im}(\mathbf{\Sigma}_{x_1|y_1}) \cap \mathrm{Im}(\mathbf{\Sigma}_{x_2|y_2})). \qquad (23)$$

Proof. See Appendix A.   □

A. Gaussian Signals

We first consider conditions for reliable separation of Gaussian distributed signals in the presence of side information.

Theorem 1. Consider the measurement model in (1)-(3), where $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ are drawn from the joint Gaussian distributions described in (5) and (8), respectively, and $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ are statistically independent. Then, with probability one, it holds:

$$\lim_{\sigma^2 \to 0} \mathrm{MMSE}^{\mathrm{G}}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2) = 0 \iff$$
$$m \geq r_{xy} - r_{y_1} - r_{y_2} \text{ and } \mathcal{D}_{x|y} = 0, \qquad (24)$$

where

$$\mathcal{D}_{x|y} = \dim(\mathrm{Im}(\mathbf{\Sigma}_{x_1|y_1}) \cap \mathrm{Im}(\mathbf{\Sigma}_{x_2|y_2}))$$

and

$$\mathbf{\Sigma}_{x_1|y_1} = \mathbf{\Sigma}_{x_1} - \mathbf{\Sigma}_{x_1 y_1} \mathbf{\Sigma}_{y_1}^{\dagger} \mathbf{\Sigma}_{y_1 x_1},$$
$$\mathbf{\Sigma}_{x_2|y_2} = \mathbf{\Sigma}_{x_2} - \mathbf{\Sigma}_{x_2 y_2} \mathbf{\Sigma}_{y_2}^{\dagger} \mathbf{\Sigma}_{y_2 x_2}.$$

Proof. See Appendix B. □

Remark 1. In the case where side information is not available, i.e., $\mathbf{y}_1 = \mathbf{y}_2 = \mathbf{0}$, irrespective of the number of measurements, reliable separation of the source signals is only feasible provided that the range spaces associated to the source signals have no overlap, i.e.,

$$\mathcal{D}_x = \dim(\mathrm{Im}(\mathbf{\Sigma}_{x_1}) \cap \mathrm{Im}(\mathbf{\Sigma}_{x_2})) = 0.$$

Under this condition, $r_{x_1} + r_{x_2}$ measurements are necessary and sufficient to drive the MMSE to zero in the low-noise regime.

Interestingly, Theorem 1 shows that the presence of side information not only reduces the number of measurements, necessary and sufficient for reliable separation, but equally importantly also relaxes the condition on the interaction between the subspaces associated with the two source signals. More intuitively, given that the dimension of the subspaces associated with the source signals conditioned on the side information can shrink, we can still reliably separate the source signals given the side information provided the "conditioned" subspaces do not overlap even if the original ones do (as suggested by the conditions in (24)).

Moreover under condition $\mathcal{D}_{x|y} = 0$, Theorem 1 shows that we can reliably separate the source signals in the presence of the side information, provided that we observe at least $r_{xy} - r_{y_1} - r_{y_2}$ measurements extracted from the mixture, i.e., the dimension of the projected mixture is equal to or greater than the sum of the dimensions of two spaces spanned by components of the source signals which are most uncorrelated with respect to side information signals.

In other terms, when side information is available to the decoder, this can be effectively used to extract all information contained in such signals that is correlated with the signals of interest. Therefore, reliable separation in the presence of side information is achieved when the projections of the components of the source signals which are most uncorrelated with the corresponding side information do not overlap.

B. GMM Signals

We now consider conditions for reliable separation of GMM signals in the presence of side information. In this case, the analysis challenge relates to the absence of closed-form expressions for the MMSE in (13) associated with GMM sources. Therefore, to derive necessary conditions, we will be working with the following MMSE lower bound, which we denote by $\mathrm{MMSE}^{\mathrm{LB}}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2)$, which is given by:

$$\mathrm{MMSE}^{\mathrm{GMM}}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2) = \mathbb{E}\left[\|\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2]\|^2\right]$$
$$= \sum_{(i,j)\in\mathcal{L}} p_I(i) p_J(j) \, \mathbb{E}\left[\|\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2]\|^2 | I=i, J=j\right]$$
$$\geq \sum_{(i,j)\in\mathcal{L}} p_I(i) p_J(j) \, \mathrm{MMSE}^{\mathrm{G}(i,j)}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2)$$
$$= \mathrm{MMSE}^{\mathrm{LB}}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2),$$

where $\mathrm{MMSE}^{\mathrm{G}(i,j)}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2)$ represents the MMSE associated with the recovery of the Gaussian component corresponding to the label (i,j) of the GMM signal given the measurements of the linear mixture and given the side information. The inequality is the consequence of the optimality of the MMSE estimator for the Gaussian sources and side information. The analysis of the MMSE lower bound leads immediately to the following theorem.

Theorem 2. Consider the measurement model in (1)-(3), where the source signals $\mathbf{x}_1$, $\mathbf{x}_2$ and the side information signals $\mathbf{y}_1$, $\mathbf{y}_2$ are drawn from the joint GMM distribution described in Section III. Then, with probability one, it holds:

$$\lim_{\sigma^2 \to 0} \mathrm{MMSE}^{\mathrm{GMM}}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2) = 0 \Rightarrow$$
$$m \geq r_{xy}^{(ij)} - r_{y_1}^{(i)} - r_{y_2}^{(j)} \qquad (25)$$
$$\text{and } \mathcal{D}_{x|y}^{(ij)} = 0 \quad \forall (i,j) \in \mathcal{L}, \qquad (26)$$

where

$$\mathcal{D}_{x|y}^{(ij)} = \dim(\mathrm{Im}(\mathbf{\Sigma}_{x_1|y_1}^{(i)}) \cap \mathrm{Im}(\mathbf{\Sigma}_{x_2|y_2}^{(j)}))$$

and

$$\mathbf{\Sigma}_{x_1|y_1}^{(i)} = \mathbf{\Sigma}_{x_1}^{(i)} - \mathbf{\Sigma}_{x_1 y_1}^{(i)} \mathbf{\Sigma}_{y_1}^{(i)\dagger} \mathbf{\Sigma}_{y_1 x_1}^{(i)},$$
$$\mathbf{\Sigma}_{x_2|y_2}^{(j)} = \mathbf{\Sigma}_{x_2}^{(j)} - \mathbf{\Sigma}_{x_2 y_2}^{(j)} \mathbf{\Sigma}_{y_2}^{(j)\dagger} \mathbf{\Sigma}_{y_2 x_2}^{(j)}.$$

Proof. See appendix C. □

In turn, to derive sufficient conditions, we will be working with an MMSE upper bound, $\mathrm{MSE}^{\mathrm{CS}}(\sigma^2)$, associated with a specific two-step classify and separate (CS) decoder. The upper bound is due to the sub-optimality of the classify and separate decoder, which means that, due to the definition of the MMSE, it holds $\mathrm{MMSE}^{\mathrm{GMM}}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2) \leq \mathrm{MSE}^{\mathrm{CS}}(\sigma^2)$.

The CS decoder operates in two main steps as follows:

1) Classification step: In the first step, an estimate $(\hat{I}, \hat{J})$ of the component labels associated to the source signals and side information signals is obtained via the maximum *a posteriori* classifier:

$$(\hat{I}, \hat{J}) = \arg \max_{(i,j)\in\mathcal{L}} p(I=i, J=j|\mathbf{w})$$

$$= \arg \max_{(i,j)\in\mathcal{L}} p_{I,J}(i,j)\, p(\mathbf{w}|I=i,J=j).$$

2) Separation step: In the second step, an estimate $\hat{\mathbf{x}}$ of the vector $\mathbf{x}$ is obtained using the Gaussian conditional mean estimator associated to the Gaussian component with labels $(\hat{I},\hat{J})$:

$$\hat{\mathbf{x}}^{(\hat{I}\hat{J})}(\mathbf{w}) = \boldsymbol{\mu}_x^{(\hat{I}\hat{J})} + \mathbf{W}_x^{(\hat{I}\hat{J})}(\mathbf{w} - \boldsymbol{\Phi}_0 \boldsymbol{\mu}_{xy}^{(\hat{I}\hat{J})})$$

where

$$\boldsymbol{\mu}_x^{(\hat{I}\hat{J})} = \left[{\boldsymbol{\mu}_{x_1}^{(\hat{I})}}^T \quad {\boldsymbol{\mu}_{x_2}^{(\hat{J})}}^T \right]^T$$

and

$$\mathbf{W}_x^{(\hat{I}\hat{J})} = \begin{bmatrix} \boldsymbol{\Sigma}_{x_1}^{(\hat{I})} & \mathbf{0} & \boldsymbol{\Sigma}_{x_1 y_1}^{(\hat{I})} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{x_2}^{(\hat{J})} & \mathbf{0} & \boldsymbol{\Sigma}_{x_2 y_2}^{(\hat{J})} \end{bmatrix} \boldsymbol{\Phi}_0^T$$
$$\cdot (\sigma^2 \mathbf{I} + \boldsymbol{\Phi}_0 \boldsymbol{\Sigma}_{xy}^{(\hat{I}\hat{J})} \boldsymbol{\Phi}_0^T)^{-1}.$$

The analysis of the MMSE upper bound leads to the following theorem.

Theorem 3. Consider the measurement model in (1)-(3), where the source signals $\mathbf{x}_1$, $\mathbf{x}_2$ and the side information signals $\mathbf{y}_1$, $\mathbf{y}_2$ are drawn from the the joint GMM distribution described in Section III. Then, with probability one, it holds:

$$m > r_{xy}^{(ij)} - r_{y_1}^{(i)} - r_{y_2}^{(j)} \tag{27}$$

$$\text{and } \mathcal{D}_{x|y}^{(ij)} = 0 \quad \forall (i,j) \in \mathcal{L} \Rightarrow \tag{28}$$

$$\lim_{\sigma^2 \to 0} \mathrm{MMSE}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}^{\mathrm{GMM}}(\sigma^2) = 0$$

Proof. See appendix D. $\qquad\square$

The results from Theorems 2 and 3 state that, in order to achieve reliable separation, the spaces spanned by conditional covariances, i.e., the space spanned by signal components which are not correlated with the side informations, ought to have no intersection for all possible label pairs (I, J). Moreover, the measurements extracted from the mixture should be enough to capture the components of source signals which are not correlated with the side informations for all Gaussian components.

Notably, the provided conditions for reliable separations are tight, as the necessary conditions are only one measurement away from the sufficient conditions. In other terms, when using the proposed sufficient conditions for reliable separation to gauge the number of measurements m used for a given separation problem, the provided analysis guarantees that such number differs at most one measurement from the optimal solution.

V. NUMERICAL RESULTS

We now provide results with synthetic data showcasing the interplay between the number of linear measurements from the mixture and the properties of the individual components of the mixture impact on the quality of separation. In particular, these results also demonstrate that our theory is able to predict the number of linear observations required for reliable separation.

In our simulations, we use random measurement matrices whose entries are i.i.d., Gaussian random variables with zero
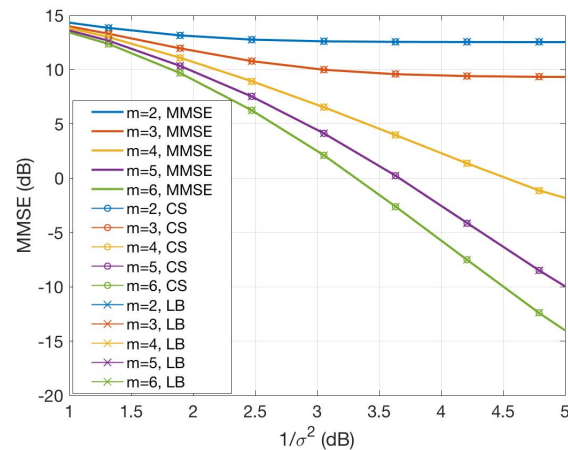

Fig. 1: MMSE associated with the separation of two GMM sources vs. $1/\sigma^2$ for different number of random measurements $m = 2$ to $m = 6$. The actual MMSE is represented by solid lines, the CS upper bound is represented by circled solid lines, and the lower bound by dashed lines. The lines corresponding to the actual MMSE, the CS upper bound and the lower bound are almost completely overlapping over the considered noise range.

mean and unit variance, which have been normalized so that it holds $\boldsymbol{\Phi}\boldsymbol{\Phi}^T = \mathbf{I}$.

We assume that the joint distribution of $(\mathbf{x}_1, \mathbf{y}_1)$ and the joint distribution of $(\mathbf{x}_2, \mathbf{y}_2)$ are modeled via a zero-mean Gaussian mixture model. Conditioned on a component labels i or j, the covariance matrices for the pairs source signal and side information signal are generated according to the model adopted in [21], where the correlation between any two Gaussian vectors is expressed in terms of a common component and innovation components. In the following description, in order to simplify notation, superscripts are dropped, as the covariance matrix construction is repeated similarly for all labels i. Then, on denoting by $\mathbf{P}_{cx_1}, \mathbf{P}_{cy_1}$ the matrices modulating the common components of the signal of interest and side information, respectively, and by $\mathbf{P}_{ix_1}, \mathbf{P}_{iy_1}$ the matrices modulating the corresponding innovation components, the covariance matrices of the signal of interest and the side information can be written as $\boldsymbol{\Sigma}_{x_1} = \mathbf{P}_{cx_1}\mathbf{P}_{cx_1}^T + \mathbf{P}_{ix_1}\mathbf{P}_{ix_1}^T$ and $\boldsymbol{\Sigma}_{y_1} = \mathbf{P}_{cy_1}\mathbf{P}_{cy_1}^T + \mathbf{P}_{iy_1}\mathbf{P}_{iy_1}^T$, respectively. Moreover, the joint covariance matrix of the pair $(\mathbf{x}_1, \mathbf{y}_1)$ is given by:

$$\bar{\boldsymbol{\Sigma}}_{x_1 y_1} = \begin{bmatrix} \mathbf{P}_{cx_1}\mathbf{P}_{cx_1}^T + \mathbf{P}_{ix_1}\mathbf{P}_{ix_1}^T & \mathbf{P}_{cx_1}\mathbf{P}_{cy_1}^T \\ \mathbf{P}_{cy_1}\mathbf{P}_{cx_1}^T & \mathbf{P}_{cy_1}\mathbf{P}_{cy_1}^T + \mathbf{P}_{iy_1}\mathbf{P}_{iy_1}^T \end{bmatrix}. \tag{29}$$

The matrices $\mathbf{P}_{cx_1} \in \mathbb{R}^{n_x \times r_{c_1}}, \mathbf{P}_{ix_1} \in \mathbb{R}^{n_x \times r_{ix_1}}, \mathbf{P}_{cy_1} \in \mathbb{R}^{n_{y_1} \times r_{c_1}}$, and $\mathbf{P}_{iy_1} \in \mathbb{R}^{n_{y_1} \times r_{iy_1}}$ have i.i.d., zero-mean, Gaussian entries. Therefore, the corresponding ranks are, with probability one, $r_{x_1} = \min\{n_x, r_{c_1} + r_{ix_1}\}, r_{y_1} = \min\{n_{y_1}, r_{c_1} + r_{iy_1}\}$, and $r_{x_1 y_1} = \min\{n_x + n_{y_1}, r_{c_1} + r_{ix_1} + r_{iy_1}\}$. A similar construction is adopted to generate the joint covariance matrices for $(\mathbf{x}_2, \mathbf{y}_2)$.

Fig. 1 shows the MMSE associated with the separation of two-components GMM signals, i.e., $\alpha = \beta = 2$, with

Fig. 2: Panels from the Ghent Altarpiece: (left) open panel, (centre) closed panel, (right) corresponding X-ray images containing a mixture of components.

TABLE I: Running time comparison.

Algorithm	Train time	Test time
CDL	1.5×10^5s	1.1×10^4s
GMM	1.65×10^3s	5.03s

dimensions $n_{x_1} = n_{x_2} = 10$, in the presence of side information where $r_{x_1 y_1} = r_{x_2 y_2} = 6$, $r_{x_1} = r_{x_2} = 4$ and $r_{y_1} = r_{y_2} = 4$ for all component labels. We report the actual values of the separation error $\text{MMSE}^{\text{GMM}}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2)$, the lower bound $\text{MMSE}^{\text{LB}}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}(\sigma^2)$ and the upper bound $\text{MSE}^{\text{CS}}(\sigma^2)$ associated to the classify and separate decoder. We observe that reliable separation is achieved with $m = 4$ for the lower bound, the upper bound and the actual MMSE, although the curves offsets are slightly different. The fact that both upper and lower bounds approach zero with the same number of measurements relates to the performance of the MAP classifier in the presence of side information. More specifically, when we fully observe the side information and the associated subspaces are distinguishable, the classification can be reliably performed based on the side information, regardless of m. In this case, the MMSE associated to the classify and reconstruct decoder perfectly captures the features of the actual MMSE.

Then, it is possible to observe that the results reported in Fig. 1 indicate that the theory presented in Section IV allows to gauge the number of observation of the linear mixture required to obtain reliable separation.

VI. APPLICATION: SEPARATION OF SUPER-IMPOSED X-RAY IMAGES GIVEN RGB SIDE INFORMATION

We finally use our algorithm to address an image separation problem arising in the context of the well-known Ghent Altarpiece. This altarpiece, one of the most admired masterpieces in the history of art, is a polyptych on wood panels attributed to the brothers Hubert and Jan van Eyck. It consists of a series of panels organized in two vertical tiers, each with double sets of foldable wings comprising inner and outer panel painting. The closed and open views of the master piece are shown in Fig 2.

The masterpiece was documented with various imaging modalities including visual microphotography, X-radiography

Fig. 3: Image set from a double sided panel of the Ghent Altarpiece which we used as the test data to evaluate the performance of our proposed algorithm. The first and second images are the photographs of the closed and open panels respectively and the third column is the corresponding X-ray scan.
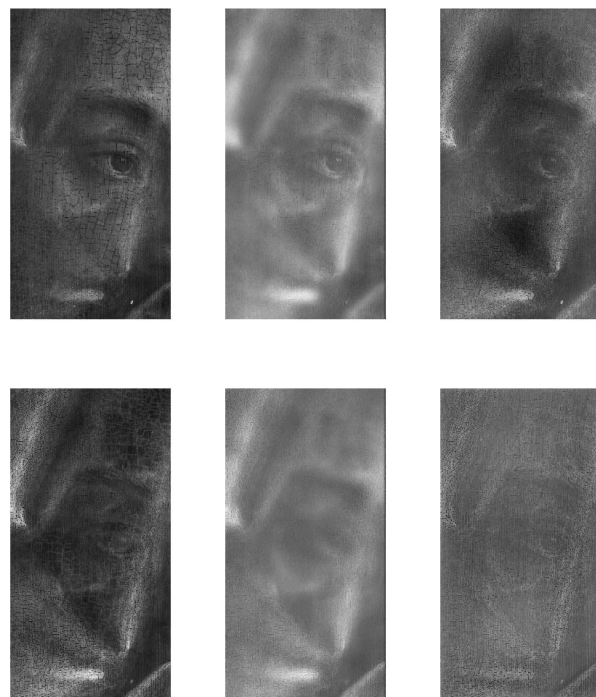

Fig. 4: Visual evaluation of the proposed algorithm in comparison with other algorithms [6], [16] in the separation of X-ray images in Fig 3; first row separated side 1; second row separated side 2; first column, the results of the proposed algorithm; second column, the results of the multi-modal dictionary learning algorithm [6]; third column, the results of the MCA algorithm with fixed dictionaries [16].

and infrared reflectography[1]. The X-ray is a common tool in art investigation, which – by penetrating through the painting – reveals information about its inner structures, including underpaintings, composition of the materials and cracks in the different layers. Due to the X-ray penetration property, X-ray

[1] http://closertovaneyck.kikirpa.be

TABLE II: Table of performances for the source separation algorithms on synthetically added X-ray images with PSNR and SSIM error measures. The rows present the algorithms and the columns present the error measures for each mixture pair.

	Image	Mixture 1		Mixture 2		Mixture 3		Mixture 4		Mixture 5		Mixture 6	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CDL	X-ray 1	23.1932	0.7297	19.1268	0.6001	16.4117	0.7538	24.2636	0.7336	**25.7920**	0.6337	**21.7206**	0.7781
	X-ray 2	21.7015	0.4448	19.5387	0.3394	21.7062	0.5783	24.6145	0.6823	**26.3819**	0.6452	**24.1172**	0.5858
GMM	X-ray 1	**23.5703**	**0.7412**	**25.0642**	**0.7548**	**22.4417**	**0.8805**	**25.5121**	**0.8626**	24.8206	**0.8421**	21.2675	**0.8569**
	X-ray 2	**24.6499**	**0.8387**	**23.3593**	**0.8039**	**23.7488**	**0.9075**	**24.7298**	**0.8537**	24.4782	**0.8210**	22.0572	**0.8970**

scans of double sided panels are, in fact, mixtures of the X-rays of both sides. The image separation challenge then relates to the separation of the mixture of the X-ray images taken from double-sided panels, leveraging the visual scan of each side as the side information.

We approach this problem using our proposed method. In particular, our goal is to decompose a vector $\mathbf{v}$ representing a patch of size 8×8 corresponding to the X-ray mixture image onto two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ representing patches of size 8×8 corresponding to the X-ray images of each side, given vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ relating to the corresponding patches from the RGB scans of each side. Note that, in this set of experiments, the mixture is fully observed, i.e., $\boldsymbol{\Phi} = \mathbf{I}$.

Concretely, we first learn a GMM model to describe the pairs of vectors $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ with $\alpha = \beta = 15$. This joint GMM model is learned from 8 image pairs – each consisting of one X-ray scan and the corresponding visual scan – taken from single-sided panels of the Ghent Altarpiece using an EM algorithm [54]. Stability of the chosen EM procedure is guaranteed by forcing the covariance matrices in the GMM to be approximately low rank with the introduction of a small diagonal perturbation.[2]

We then use such GMM model to separate the mixed $\mathbf{x}_1$ and $\mathbf{x}_2$ from their superposition given $\mathbf{u}_1$ and $\mathbf{u}_2$ using the conditional mean estimator appearing in (14)-(15). This process is also repeated for every patch – where each two adjacent patches are overlapping by 4 pixels – enabling us to separate the mixed X-ray image shown in Fig. 3 onto its constituents. It is important to mention that the images used as training and test sets have been registered (aligned) using multi-modal image registration algorithms [56].

The results obtained with the proposed algorithm are depicted in Fig. 4. We also compare the separation results of the proposed algorithm against those achieved with a state-of-the-art algorithm in separating X-ray images which uses coupled dictionary learning [6] and MCA [16]. For both the proposed algorithm and the method in [6], training data are obtained by extracting 8×8 patches from the same images.

We follow the setting in the original work on the same dataset [6], where we set the number of sparse common component to 10, the number of sparse innovative components to 8, and the number of each dictionary atoms to 256. We report the best performing multi-scale approach which recursively decomposes the X-ray and visual images into low and high pass bands. The high frequency component of the mixed X-

ray image at each scale is then separated patch-by-patch. It is worth mentioning that, as opposed to the results reported in [6], we do not remove the crack patterns from the painting images for any of the algorithms. For MCA algorithms – which do not take advantage of the availability of the RGB images associated with each side of the panel – we use fixed curvelet and wavelet dictionaries.

Although a third-party evaluation has not been considered in the comparison, a visual analysis of the results reported in Fig. 4 suggests that the proposed algorithm outperforms the other two algorithms in capturing some fine details of the images (e.g., eye lashes). Moreover, the proposed algorithm is not affected by the smoothing effect which appears for the algorithms based on the coupled dictionary learning and MCA.

It is also interesting to observe that the proposed method offers significant advantages in terms of computational complexity, both during the training and testing phases. Table I displays the train and test time of the proposed algorithm (referred as GMM) in comparison to coupled dictionary learning algorithm (referred as CDL), where one obtains over 100-fold improvements during the training phase and over 1000-fold gain during the testing phase. This is due to the efficiency of the EM algorithm in comparison to coupled dictionary learning in the training phase. Also, in test phase the GMM algorithm exploits an efficient closed form expression as opposed to CDL which uses an iterative algorithm to solve a regularized optimization problem. Both algorithms have been implemented in Matlab and have been run on the same machine with 12 CPU cores. The reported test time is measured on separating a mixture of X-ray images of size 256×256.

Finally, due to the lack of ground truth data, it is also of interest to evaluate the performance of our algorithm quantitatively. We have thus generated simulated mixtures using the available training data taken from single sided panels. In particular, we extract 6 image patches of size 256×256, from the X-ray images of single sided panels, depicting content similar to the images in Fig. 3, i.e, 6 image patches visualizing part of a face and 6 images patches visualizing a piece of fabric. We then assessed the separation performance of the proposed algorithm against the coupled dictionary algorithm [6] by measuring the peak-signal-to-noise-ratio (PSNR) and structural similarity index metric (SSIM). The results are reported in Table II. Compared to the state-of-the-art coupled dictionary learning algorithm, the proposed algorithm provides a considerably better separation quality both in terms of PSNR and SSIM except for simulated Mixtures 5 and 6. This could be due to the fact that the images in mixture 5 and mixture 6

---

[2]Note that the analysis carried out for exactly low-rank GMM models can be shown to provide valuable information in determining the behavior of approximately low-rank data [30], [55].

have less high frequency components.

VII. CONCLUSION

In this paper, we have proposed a new approach to source separation with side information. In particular, our main contributions are:

1) A new framework to source separation with side information leveraging a GMM of both the individual source signals of interest and the side information signals.
2) A characeterization of necessary and sufficient conditions for reliable separation of source signals from a linear mixture in the presence of side information.
3) An application of the approach to image separation problems arising in art investigation.

When compared with other source separation algorithms leveraging the availability of side information, as [6], the proposed approach exhibits a clear advantage in terms of training and testing time, and it guarantees visually superior separation results.

APPENDIX A
PROOF OF LEMMA 1

We start by considering the following rank:

$$\text{rank}(\mathbf{\Phi}_0\mathbf{\Sigma}_{xy}) = \text{rank}\left(\begin{bmatrix}\mathbf{\Phi} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_{y_1}+n_{y_2}}\end{bmatrix}\right.$$
$$\left.\begin{bmatrix}\mathbf{\Sigma}_{x_1} & \mathbf{\Sigma}_{x_2} & \mathbf{\Sigma}_{x_1y_1} & \mathbf{\Sigma}_{x_2y_2} \\ \mathbf{\Sigma}_{y_1x_1} & \mathbf{0} & \mathbf{\Sigma}_{y_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{y_2x_2} & \mathbf{0} & \mathbf{\Sigma}_{y_2}\end{bmatrix}\right). \quad (30)$$

We define:

$$\mathbf{M} = \begin{bmatrix}\mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H}\end{bmatrix}, \quad (31)$$

with

$$\mathbf{E} = \begin{bmatrix}\mathbf{\Sigma}_{x_1} & \mathbf{\Sigma}_{x_2}\end{bmatrix} \quad \mathbf{F} = \begin{bmatrix}\mathbf{\Sigma}_{x_1y_1} & \mathbf{\Sigma}_{x_2y_2}\end{bmatrix}$$
$$\mathbf{G} = \begin{bmatrix}\mathbf{\Sigma}_{y_1x_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{y_2x_2}\end{bmatrix} \quad \mathbf{H} = \begin{bmatrix}\mathbf{\Sigma}_{y_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{y_2}\end{bmatrix}.$$

Now, by applying the rule for the computation of the rank of the Schur complement [57] we have:

$$\text{rank}(\mathbf{M}) = \text{rank}(\mathbf{H}) + \text{rank}(\mathbf{M}/\mathbf{H}) \quad (32)$$
$$\text{rank}(\mathbf{M}/\mathbf{H}) = \text{rank}(\mathbf{E} - \mathbf{F}\mathbf{H}^\dagger\mathbf{G}) \quad (33)$$

where $\mathbf{M}/\mathbf{H}$ is the Schur complement of block $\mathbf{H}$ in matrix $\mathbf{M}$ and

$$\mathbf{E} - \mathbf{F}\mathbf{H}^\dagger\mathbf{G} =$$
$$\begin{bmatrix}\mathbf{\Sigma}_{x_1} - \mathbf{\Sigma}_{x_1y_1}\mathbf{\Sigma}_{y_1}^\dagger\mathbf{\Sigma}_{y_1x_1} & \mathbf{\Sigma}_{x_2} - \mathbf{\Sigma}_{x_2y_2}\mathbf{\Sigma}_{y_2}^\dagger\mathbf{\Sigma}_{y_2x_2}\end{bmatrix}. \quad (34)$$

On leveraging the rank equality for block matrices [58], one can write:

$$\text{rank}(\mathbf{E} - \mathbf{F}\mathbf{H}^\dagger\mathbf{G}) = r_{x_1y_1} - r_{y_1} + r_{x_2y_2} - r_{y_2} - \mathcal{D}_{x|y}. \quad (35)$$

Then, by substituting $\text{rank}(\mathbf{H}) = r_{y_1} + r_{y_2}$ and the rank expression (35) in (32), we have:

$$\text{rank}(\mathbf{M}) = r_{x_1y_1} + r_{x_2y_2} - \mathcal{D}_{x|y} = r_{xy} - \mathcal{D}_{x|y}. \quad (36)$$

Now we can apply similar techniques as those used in the proof of Lemma 5 in [21] to calculate the rank in (30). Moreover, on defining the matrix $\mathbf{\Phi}_1$ as:

$$\mathbf{\Phi}_1 = \begin{bmatrix}\mathbf{\Phi} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_{y_1}+n_{y_2}}\end{bmatrix}, \quad (37)$$

and by applying the Sylvester's rank equality [52], we can write:

$$r = \text{rank}(\mathbf{\Phi}_1\mathbf{M}) = \text{rank}(\mathbf{M}) - \dim(\text{Im}(\mathbf{M}) \cap \text{Null}(\mathbf{\Phi}_1)). \quad (38)$$

Now, we define matrix $\mathbf{\Psi} \in \mathbb{R}^{n_x-m \times n_x}$ as the matrix whose columns form the basis for $\text{Null}(\mathbf{\Phi})$. Since $\mathbf{\Phi}$ is randomly drawn from a rotationally invariant distribution, column range space of $\mathbf{\Psi}$ is drawn from a uniform distribution on the Grassmann manifold of $(n_x - m)$ dimensional subspaces in $\mathbb{R}^{n_x}$. One can also show that $\begin{bmatrix}\mathbf{\Psi}^T\mathbf{0}_{n_{y_1}+n_{y_2} \times n_x-m}^T\end{bmatrix}^T$ form the basis for the $\text{Null}(\mathbf{\Phi}_1)$ and:

$$r = \text{rank}(\mathbf{M}) - \dim\left(\text{Im}(\mathbf{M}) \cap \text{Im}\left(\begin{bmatrix}\mathbf{\Psi} \\ \mathbf{0}\end{bmatrix}\right)\right) \quad (39)$$

$$= \text{rank}\left(\begin{bmatrix}\mathbf{A} & \mathbf{\Psi} \\ \mathbf{B} & \mathbf{0}\end{bmatrix}\right) - (n_x - m), \quad (40)$$

where we have *i*) partitioned the matrix $\mathbf{M}$ to $\mathbf{M} = \begin{bmatrix}\mathbf{A}^T\mathbf{B}^T\end{bmatrix}^T$ where

$$\mathbf{A} = \begin{bmatrix}\mathbf{\Sigma}_{x_1} & \mathbf{\Sigma}_{x_2} & \mathbf{\Sigma}_{x_1y_1} & \mathbf{\Sigma}_{x_2y_2}\end{bmatrix} \quad (41)$$
$$\mathbf{B} = \begin{bmatrix}\mathbf{\Sigma}_{y_1x_1} & \mathbf{0} & \mathbf{\Sigma}_{y_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{y_2x_2} & \mathbf{0} & \mathbf{\Sigma}_{y_2}\end{bmatrix}, \quad (42)$$

ii) used the fact that

$$\text{rank}\left(\begin{bmatrix}\mathbf{A} & \mathbf{\Psi} \\ \mathbf{B} & \mathbf{0}\end{bmatrix}\right) = \text{rank}\left(\begin{bmatrix}\mathbf{A} \\ \mathbf{B}\end{bmatrix}\right) + \text{rank}\left(\begin{bmatrix}\mathbf{\Psi} \\ \mathbf{0}\end{bmatrix}\right)$$
$$- \dim\left(\text{Im}\left(\begin{bmatrix}\mathbf{A} \\ \mathbf{B}\end{bmatrix}\right) \cap \text{Im}\left(\begin{bmatrix}\mathbf{\Psi} \\ \mathbf{0}\end{bmatrix}\right)\right), \quad (43)$$

and *iii*) used the fact that $\text{rank}(\mathbf{\Psi}) = n_x - m$. In order to compute the rank in (40), we apply the generalized singular value decomposition (GSVD) [59] on the matrices $\mathbf{A}$ and $\mathbf{B}$ as defined above. The GSVD implies that there exist two orthogonal matrices $\mathbf{U} \in \mathbb{R}^{n_x \times (2n_x+n_{y_1}+n_{y_2})}$ and $\mathbf{V} \in \mathbb{R}^{(n_{y_1}+n_{y_2}) \times (2n_x+n_{y_1}+n_{y_2})}$ and a non-singular matrix $\mathbf{X} \in \mathbb{R}^{(2n_x+n_{y_1}+n_{y_2}) \times (2n_x+n_{y_1}+n_{y_2})}$ such that:

$$\mathbf{U}^T\mathbf{A}\mathbf{X} = \begin{bmatrix}\mathbf{\Lambda}_A & \mathbf{0}_{n_x \times (2n_x+n_{y_1}+n_{y_2}-r_{AB})}\end{bmatrix}, \quad (44)$$
$$\mathbf{V}^T\mathbf{B}\mathbf{X} = \begin{bmatrix}\mathbf{\Lambda}_B & \mathbf{0}_{(n_{y_1}+n_{y_2}) \times (2n_x+n_{y_1}+n_{y_2}-r_{AB})}\end{bmatrix}, \quad (45)$$

where

$$\mathbf{\Lambda}_A = \begin{matrix} & \overset{r_{AB}-r_B}{} & \overset{s_{AB}}{} & \overset{r_{AB}-r_A}{} & \\ \begin{pmatrix}\mathbf{I} & & \\ & \mathbf{D}_A & \\ & & \mathbf{0}\end{pmatrix} & & & & \begin{matrix}r_{AB}-r_B \\ s_{AB} \\ n-r_A\end{matrix}\end{matrix} \quad (46)$$

$$\mathbf{\Lambda}_B = \begin{matrix} & \overset{r_{AB}-r_B}{} & \overset{s_{AB}}{} & \overset{r_{AB}-r_A}{} & \\ \begin{pmatrix}\mathbf{0} & & \\ & \mathbf{D}_A & \\ & & \mathbf{I}\end{pmatrix} & & & & \begin{matrix}r_{AB}-r_B \\ s_{AB} \\ n-r_A\end{matrix}\end{matrix} \quad (47)$$

$$\quad (48)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2020.2964195, IEEE Transactions on Signal Processing

12

and $r_A = \text{rank}(\mathbf{A})$, $r_B = \text{rank}(\mathbf{B})$, $r_{AB} = \text{rank}(\mathbf{M}) = \text{rank}(\begin{bmatrix} \mathbf{A}^T \mathbf{B}^T \end{bmatrix}^T)$, $s_{AB} = r_A + r_B - r_{AB}$. Furthermore $\mathbf{D}_A$ and $\mathbf{D}_B$ are diagonal matrices such that $\mathbf{D}_A = \text{diag}(\alpha_1, ..., \alpha_{s_{AB}})$, $\mathbf{D}_B = \text{diag}(\beta_1, ..., \beta_{s_{AB}})$ and $\alpha_i^2 + \beta_i^2 = 1$, for $i = 1, ..., s_{AB}$. Hence, we can write:

$$r_{\boldsymbol{\Psi}} = \text{rank}\left(\begin{bmatrix} \mathbf{U}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^T \end{bmatrix} \begin{bmatrix} \mathbf{A} & \boldsymbol{\Psi} \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \right) \quad (49)$$

$$= \text{rank}\left(\begin{bmatrix} \boldsymbol{\Lambda}_A & \mathbf{0} & \boldsymbol{\Psi}' \\ \boldsymbol{\Lambda}_B & \mathbf{0} & \mathbf{0}, \end{bmatrix} \right), \quad (50)$$

where the range of $\boldsymbol{\Psi}' = \mathbf{U}^T \boldsymbol{\Psi}$ is uniformly distributed over the Grassmann manifold of $n_x - m$ dimensional subspaces of $\mathbb{R}^{n_x}$. Now, by considering the first $r_{AB} - r_B$ columns of the matrix in (50) together with its last $n_x - m$ columns and by using the fact that the columns of $\boldsymbol{\Psi}'$ form a random subspace in $\mathbb{R}^{n_x}$, we can conclude that, with probability one, we can select from such columns, $\min\{r_{AB} - r_B + n_x - m, n_x\}$ independent columns which are also independent from the other r_B non-zero columns of the same matrix. Therefore, we obtain:

$$r_{\boldsymbol{\Psi}} = \min\{r_{AB} - r_B + n_x - m, n_x\} + r_B. \quad (51)$$

By substituting $r_{AB} = \text{rank}(\mathbf{M}) = r_{xy} - \mathcal{D}_{x|y}$, $r_B = r_{y_1} + r_{y_2}$ in (51) and then substituting (51) in (40), we have:

$$r = \min\{r_{xy} - \mathcal{D}_{x|y}, m + r_{y_1} + r_{y_2}\}. \quad (52)$$

Note that the expression for r_B follows from a column permutation in $\mathbf{B}$ and using the fact that $\text{rank}(\begin{bmatrix} \boldsymbol{\Sigma}_{y_1 x_1} & \boldsymbol{\Sigma}_{y_1} \end{bmatrix}) = r_{y_1}$ and $\text{rank}(\begin{bmatrix} \boldsymbol{\Sigma}_{y_2 x_2} & \boldsymbol{\Sigma}_{y_2} \end{bmatrix}) = r_{y_2}$. $\square$

APPENDIX B
PROOF OF THEOREM 1

The proof is based on reformulating the MMSE associated to the reconstruction of source signals in terms of the MMSE associated to the reconstruction of the signals conditioned on observing the side-information signals:

$$\text{MMSE}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}^{\text{G}}(\sigma^2) = \mathbb{E}\left[\|\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{w}]\|^2\right] = \quad (53)$$

$$\text{MMSE}(\mathbf{z}_1, \mathbf{z}_2 | \mathbf{w}_z), \quad (54)$$

where $\mathbf{z}_1 \sim p(\mathbf{x}_1 | \mathbf{y}_1)$, $\mathbf{z}_2 \sim p(\mathbf{x}_2 | \mathbf{y}_2)$ and $\mathbf{w}_z = \boldsymbol{\Phi}(\mathbf{z}_1 + \mathbf{z}_2) + \mathbf{n}$. The equality in (53) follows from taking the expectation of (53) with respect to the variables $(\mathbf{x}_1 | \mathbf{y}_1)$, $(\mathbf{x}_2 | \mathbf{y}_2)$ and $\mathbf{y}_1, \mathbf{y}_2$, separately. Since $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ are jointly Gaussian, then

$$\mathbf{z}_1 \sim \mathbf{x}_1 | \mathbf{y_1} \sim \mathcal{N}(\boldsymbol{\mu}_{x_1|y_1}, \boldsymbol{\Sigma}_{x_1|y_1}), \ \mathbf{z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_{z_1}, \boldsymbol{\Sigma}_{z_1}) \quad (55)$$

$$\mathbf{z}_1 \sim \mathbf{x}_2 | \mathbf{y_2} \sim \mathcal{N}(\boldsymbol{\mu}_{x_2|y_2}, \boldsymbol{\Sigma}_{x_2|y_2}), \ \mathbf{z}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{z_2}, \boldsymbol{\Sigma}_{z_2}) \quad (56)$$

where,

$$\boldsymbol{\mu}_{z_1} = \boldsymbol{\mu}_{x_1|y_1} = \boldsymbol{\mu}_{x_1} + \boldsymbol{\Sigma}_{x_1 y_1} \boldsymbol{\Sigma}_{y_1}^\dagger (\mathbf{y}_1 - \boldsymbol{\mu}_{y_1}) \quad (57)$$

$$\boldsymbol{\mu}_{z_2} = \boldsymbol{\mu}_{x_2|y_2} = \boldsymbol{\mu}_{x_2} + \boldsymbol{\Sigma}_{x_2 y_2} \boldsymbol{\Sigma}_{y_2}^\dagger (\mathbf{y}_2 - \boldsymbol{\mu}_{y_2}) \quad (58)$$

$$\boldsymbol{\Sigma}_{z_1} = \boldsymbol{\Sigma}_{x_1|y_1} = \boldsymbol{\Sigma}_{z_1} = \boldsymbol{\Sigma}_{x_1} - \boldsymbol{\Sigma}_{x_1 y_1} \boldsymbol{\Sigma}_{y_1}^\dagger \boldsymbol{\Sigma}_{y_1 x_1} \quad (59)$$

$$\boldsymbol{\Sigma}_{z_2} = \boldsymbol{\Sigma}_{x_2|y_2} = \boldsymbol{\Sigma}_{z_2} = \boldsymbol{\Sigma}_{x_2} - \boldsymbol{\Sigma}_{x_2 y_2} \boldsymbol{\Sigma}_{y_2}^\dagger \boldsymbol{\Sigma}_{y_2 x_2}. \quad (60)$$

More specifically, $\boldsymbol{\Sigma}_{x_1|y_1}$ and $\boldsymbol{\Sigma}_{x_2|y_2}$ are the conditional covariances of the conditional random variables $\mathbf{x}_1 | \mathbf{y}_1$ and

$\mathbf{x}_2 | \mathbf{y}_2$, respectively. By applying the results in Appendix D [21], one can show that , $r_{z_1} = \text{rank}(\boldsymbol{\Sigma}_{z_1}) = r_{x_1 y_1} - r_{y_1}$ and $r_{z_2} = \text{rank}(\boldsymbol{\Sigma}_{z_2}) = r_{x_2 y_2} - r_{y_2}$. We define the signal $\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix}$ where $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ and

$$\boldsymbol{\mu}_z = \begin{bmatrix} \boldsymbol{\mu}_{z_1} \\ \boldsymbol{\mu}_{z_2} \end{bmatrix} \qquad \boldsymbol{\Sigma}_z = \begin{bmatrix} \boldsymbol{\Sigma}_{z_1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{z_2} \end{bmatrix}. \quad (61)$$

Now, the problem can be formulated as recovering $\mathbf{z}$ from the observation in $\mathbf{w}_z$. Our results on the separation of two Gaussian sources in Theorem 1 imply that we can reliably reconstruct $\mathbf{z}$ from $\mathbf{w}_z$ if and only if $\dim(\text{Im}(\boldsymbol{\Sigma}_{z_2}) \cap \text{Im}(\boldsymbol{\Sigma}_{z_1})) = 0$ and $m \geq r_{z_1} + r_{z_2}$. This directly leads to the necessary and sufficient conditions for the reliable separation of $\mathbf{x}_1$ and $\mathbf{x}_2$ from $\mathbf{w}$ which are

- $m > r_{x_1 y_1} + r_{x_2 y_2} - r_{y_1} - r_{y_2}$
- $\dim(\text{Im}(\boldsymbol{\Sigma}_{x_1|y_1}) \cap \text{Im}(\boldsymbol{\Sigma}_{x_2|y_2})) = 0$. $\square$

APPENDIX C
PROOF OF THEOREM 2

This immediately follows from applying Theorem 2 on the lower bound $\text{MMSE}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}^{\text{LB}}$. In particular, Theorem 2 implies that if $\text{MMSE}_{\mathbf{x}|\mathbf{v},\mathbf{u}_1,\mathbf{u}_2}^{\text{G}} \to 0$ for $\forall(i, j) \in \mathcal{L}$ then the conditions (24) are verified for $\forall(i, j) \in \mathcal{L}$. $\square$

APPENDIX D
PROOF OF THEOREM 3

This proof is based on the proof in [29, Appendix C] where the upper bound to the MMSE associated to the use of a classify and separate decoder described in Section IV-B is used. Conditioned on the component labels $(I, J) = (i, j)$, the source signals and side information signals jointly follow a Gaussian distribution. The decoder separates the source signals by using the Wiener filter associated to the estimated component -which is the output of the MAP classifier- $(\hat{I} = i, \hat{J} = j)$ as follows:

$$\hat{\mathbf{x}}^{(ij)}(\mathbf{w}) = \boldsymbol{\mu}_x^{(ij)} + \mathbf{W}_x^{(ij)}(\mathbf{w} - \boldsymbol{\Phi}_0 \boldsymbol{\mu}_{xy}^{(ij)})$$

where

$$\boldsymbol{\mu}_x^{(ij)} = \begin{bmatrix} \boldsymbol{\mu}_{x_1}^{(i)T} & \boldsymbol{\mu}_{x_2}^{(j)T} \end{bmatrix}^T$$

and

$$\mathbf{W}_x^{(ij)} = \begin{bmatrix} \boldsymbol{\Sigma}_{x_1}^{(i)} & \mathbf{0} & \boldsymbol{\Sigma}_{x_1 y_1}^{(i)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{x_2}^{(j)} & \mathbf{0} & \boldsymbol{\Sigma}_{x_2 y_2}^{(j)} \end{bmatrix} \boldsymbol{\Phi}_0^T$$
$$\cdot (\sigma^2 \mathbf{I} + \boldsymbol{\Phi}_0 \boldsymbol{\Sigma}_{xy}^{(ij)} \boldsymbol{\Phi}_0^T)^{-1}.$$

On using the law of total probability, we can write;

$$\text{MSE}^{\text{CS}} \leq \sum_{ij} p_{I,J}(i,j) \mathbb{E}[\|\mathbf{x} - \hat{\mathbf{x}}^{(ij)}(\mathbf{w})\|^2 | I = i, J = j]$$

$$+ \sum_{ij} p_{I,J}(i,j) \sum_{(i'j') \neq (ij)} p(\hat{I} = i', \hat{J} = j' | I = i, J = j)$$

$$\mathbb{E}[\|\mathbf{x} - \hat{\mathbf{x}}^{(i'j')}(\mathbf{w})\|^2 | I = i, J = j].$$

Under the conditions in (24), Theorem 1 implies that:

$$\lim_{\sigma^2 \to 0} \mathbb{E}[\|\mathbf{x} - \hat{\mathbf{x}}^{(ij)}(\mathbf{w})\|^2 | I = i, J = j] = 0.$$

In the following, by leveraging an analysis of the MAP classifier akin to that provided in [60] (see supplementary material), we will prove that,

$$\lim_{\sigma^2 \to 0} p(\hat{I} = i', \hat{J} = j' | I = i, J = j)$$
$$\mathbb{E}[\|\mathbf{x} - \hat{\mathbf{x}}^{(i'j')}(\mathbf{w})\|^2 | \hat{I}_1 = i', \hat{J} = j', I = i, J = j] = 0. \tag{62}$$

We consider two different cases:

1)

$$r_{y_1}^{(i,i')} + r_{y_2}^{(j,j')} > r_{y_1}^{(i)} + r_{y_2}^{(j)}, \; r_{y_1}^{(i')} + r_{y_2}^{(j')},$$

or

$$r_{z_1}^{(i,i')} + r_{z_2}^{(j,j')} - \mathcal{D}_{x|y}^{(ij,i'j')} > $$
$$r_{z_1}^{(i)} + r_{z_2}^{(j)} - \mathcal{D}_{x|y}^{(ij)}, r_{z_1}^{(i')} + r_{z_2}^{(j')} - \mathcal{D}_{x|y}^{(i'j)}. \tag{63}$$

In this case, under the assumption that

$$m > \min\{r_{z_1}^{(i)} + r_{z_2}^{(j)} - \mathcal{D}_{x|y}^{(ij)}, r_{z_1}^{(i')} + r_{z_2}^{(j')} - \mathcal{D}_{x|y}^{(i'j')}\},$$

Theorem **??** implies that

$$\lim_{\sigma^2 \to 0} p(\hat{I} = i', \hat{J} = j' | I = i, J = j) = 0.$$

Therefore, (62) is verified, as the term that does not depend on the error probability can be upper bounded by a constant when, $\sigma^2 \to 0$.

2)

$$r_{z_1}^{(i,i')} + r_{z_2}^{(j,j')} - \mathcal{D}_{x|y}^{(ij,i'j')} = r_{z_1}^{(i)} + r_{z_2}^{(j)} - \mathcal{D}_{x|y}^{(ij)} = $$
$$r_{z_1}^{(i')} + r_{z_2}^{(j')} - \mathcal{D}_{x|y}^{(i'j')} \tag{64}$$

and

$$r_{y_1}^{(i,i')} + r_{y_2}^{(j,j')} = r_{y_1}^{(i)} + r_{y_2}^{(j)} = r_{y_1}^{(i')} + r_{y_2}^{(j')}.$$

Although in this case the classification probability is not guaranteed to converge to zero, it is possible to observe that such conditions are not verified with probability one. The last statement follows from the fact that, under the assumption that subspaces associated to the image of the covariance matrices are drawn from continuous distributions defined over the Grassman manifold, the subspaces associated to the $\text{Im}(\mathbf{\Sigma}_{z_1}^i + \mathbf{\Sigma}_{z_2}^j)$ and $\text{Im}(\mathbf{\Sigma}_{z_1}^{i'} + \mathbf{\Sigma}_{z_2}^{j'})$ are random subspaces whose dimensions are defined and strictly smaller than the ambient dimension n_x. Since

$$m > \min\{r_{z_1}^{(i)} + r_{z_2}^{(j)} - \mathcal{D}_{x|y}^{(ij)}, r_{z_1}^{(i')} + r_{z_2}^{(j')} - \mathcal{D}_{x|y}^{(i'j')}\}$$

and $m < n_x$, these two subspaces will not be completely overlapping, i.e., with probability one, it holds,

$$r_{z_1}^{(i,i')} + r_{z_2}^{(j,j')} - \mathcal{D}_{x|y}^{(ij,i'j')} > $$
$$r_{z_1}^{(i)} + r_{z_2}^{(j)} - \mathcal{D}_{x|y}^{(ij)}, r_{z_1}^{(i')} + r_{z_2}^{(j')} - \mathcal{D}_{x|y}^{(i'j')}. \tag{65}$$

This concludes the proof. $\square$

REFERENCES

[1] A. Aïssa-El-Bey, K. Abed-Meraim, and Y. Grenier, "Underdetermined blind audio source separation using modal decomposition," *EURASIP J. Audio Speech Music Process.*, vol. 2007, no. 1.

[2] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937 – 1949, 2012.

[3] M. Parvaix, L. Girin, and J. M. Brossier, "A watermarking-based method for informed source separation of audio signals with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1464–1475, Aug 2010.

[4] S. Gorlow and S. Marchand, "Informed separation of spatial images of stereo music recordings using second-order statistics," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2013, pp. 1–6.

[5] U. R. Abeyratne, A. P. Petropulu, and J. M. Reid, "Higher order spectra based deconvolution of ultrasound images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 42, no. 6, pp. 1064–1075, Nov 1995.

[6] N. Deligiannis, J. F. C. Mota, B. Cornelis, M. R. D. Rodrigues, and I. Daubechies, "Multi-modal dictionary learning for image separation with application in art investigation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 751–764, Feb 2017.

[7] G. Kutyniok and W.-Q. Lim, "Image separation using wavelets and shearlets," in *Curves and Surfaces*, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 416–430.

[8] J. Cardoso, M. Le Jeune, J. Delabrouille, M. Betoule, and G. Patanchon, "Component separation with flexible models—application to multichannel astrophysical observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 5, pp. 735–746, Oct 2008.

[9] L. Parra, C. Spence, P. Sajda, A. Ziehe, and K.-R. Müller, "Unmixing hyperspectral data," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, ser. NIPS'99. Cambridge, MA, USA: MIT Press, 1999, pp. 942–948. [Online]. Available: http://dl.acm.org/citation.cfm?id=3009657.3009790

[10] C. Lin and J. M. Bioucas-Dias, "New theory for unmixing ill-conditioned hyperspectral mixtures," in *10th IEEE Sensor Array and Multichannel Signal Processing Workshop, SAM 2018, Sheffield, South Yorkshire, United Kingdom, July 8-11, 2018*, 2018, pp. 430–434.

[11] C.-H. Lin and J. M. Bioucas-Dias, "Linear spectral unmixing via matrix factorization: Identifiability criteria for sparse abundances," in *2018 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018, Valencia, Spain, July 22-27, 2018*, 2018, pp. 6155–6158.

[12] M. Brandt-Pearce, "Signal separation using fractional sampling in multiuser communications," *IEEE Transactions on Communications*, vol. 48, no. 2, pp. 242–251, Feb 2000.

[13] M. Feng and K. Kammeyer, "Blind source separation for communication signals using antenna arrays," in *ICUPC '98. IEEE 1998 International Conference on Universal Personal Communications. Conference Proceedings (Cat. No.98TH8384)*, vol. 1, Oct 1998, pp. 665–669 vol.1.

[14] J. R. Treichler, M. G. Larimore, and J. C. Harp, "Practical blind demodulators for high-order QAM signals," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 1907–1926, Oct 1998.

[15] C. Fevotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2174–2188, Nov 2006.

[16] J. Bobin, J. L. Starck, J. Fadili, and Y. Moudden, "Sparsity and morphological diversity in blind source separation," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2662–2674, Nov 2007.

[17] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13, no. 4, pp. 863–882, 2001.

[18] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

[19] E. M. Grais and H. Erdogan, "Regularized nonnegative matrix factorization using gaussian mixture priors for supervised single channel source separation," *Computer Speech & Language*, vol. 27, no. 3, pp. 746 – 762, 2013, special Issue on Speech Separation and Recognition in Multisource Environments.

[20] J. F. C. Mota, N. Deligiannis, and M. R. D. Rodrigues, "Compressed sensing with prior information: Strategies, geometry, and bounds," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4472–4496, July 2017.

[21] F. Renna, L. Wang, X. Yuan, J. Yang, G. Reeves, R. Calderbank, L. Carin, and M. R. D. Rodrigues, "Classification and reconstruction of high-dimensional signals from low-dimensional features in the presence of side information," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6459–6492, Nov 2016.

[22] P. Song, J. F. C. Mota, N. Deligiannis, and M. R. D. Rodrigues, "Measurement matrix design for compressive sensing with side information at the encoder," in *2016 IEEE Statistical Signal Processing Workshop (SSP)*, June 2016, pp. 1–5.

[23] P. Song, J. Mota, N. Deligiannis, and M. R. D. Rodrigues, "Coupled dictionary learning for multimodal image super-resolution," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2016, pp. 162–166.

[24] M. Chen, F. Renna, and M. R. D. Rodrigues, "Compressive sensing with side information: How to optimally capture this extra information for GMM signals?" *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2314–2329, May 2018.

[25] G. Yu and G. Sapiro, "Statistical compressed sensing of Gaussian mixture models," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5842–5858, Dec 2011.

[26] G. Yu, G. Sapiro, and S. Mallat, "Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity," *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2481–2499, May 2012.

[27] J. Yang, X. Yuan, X. Liao, P. Llull, D. J. Brady, G. Sapiro, and L. Carin, "Video compressive sensing using Gaussian mixture models," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4863–4878, Nov 2014.

[28] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin, "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6140–6155, Dec 2010.

[29] F. Renna, R. Calderbank, L. Carin, and M. Rodrigues, "Reconstruction of signals drawn from a Gaussian mixture via noisy compressive measurements," *IEEE Trans. on Signal Processing*, vol. 62, no. 9, pp. 2265–2277, May 2014.

[30] J. Sokolić, F. Renna, R. Calderbank, and M. R. D. Rodrigues, "Mismatch in the classification of linear subspaces: Sufficient conditions for reliable classification," *IEEE Transactions on Signal Processing*, vol. 64, no. 12, pp. 3035–3050, June 2016.

[31] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[32] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.

[33] Z. Sabetsarvestani, F. Renna, F. Kiraly, and M. R. D. Rodrigues, "Source separation in the presence of side information: Necessary and sufficient conditions for reliable de-mixing," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov 2018, pp. 351–355.

[34] P. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 57–60.

[35] K. Kayabol, E. E. Kuruoglu, and B. Sankur, "Bayesian separation of images modeled with MRFs using MCMC," *IEEE Transactions on Image Processing*, vol. 18, no. 5, pp. 982–994, May 2009.

[36] J. W. Miskin, "Ensemble learning for independent component analysis," In Advances In Independent Component Analysis, Tech. Rep., 2000.

[37] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. Orlando, FL, USA: Academic Press, Inc., 2010.

[38] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

[39] H. Xu, C. H. Chen, F. Cong, L. Yang, and X. Shi, "Independent component analysis based on nonparametric density estimation on time-frequency domain," in *2005 IEEE Workshop on Machine Learning for Signal Processing*, Sept 2005, pp. 171–176.

[40] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proceedings F - Radar and Signal Processing*, vol. 140, no. 6, pp. 362–370, Dec 1993.

[41] D. B. Rowe, "A Bayesian approach to blind source separation," *Journal of Interdisciplinary Mathematics*, vol. 5, no. 1, pp. 49–76, 2002.

[42] Y. Wu, Y. Chi, and R. Calderbank, "Compressive blind source separation," in *2010 IEEE International Conference on Image Processing*, Sept 2010, pp. 89–92.

[43] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[44] P. Sprechmann, A. M. Bronstein, and G. Sapiro, *Supervised non-negative matrix factorization for audio source separation*. Cham: Springer International Publishing, 2015, pp. 407–420. [Online]. Available: https://doi.org/10.1007/978-3-319-20188-7_16

[45] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Dec 2014.

[46] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[47] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 1, pp. I–65–I–68, 2007.

[48] M. Miron, J. J. Carabias-Orti, J. J. Bosch, E. Gómez, and J. Janer, "Score-informed source separation for multichannel orchestral recordings," *JECE*, vol. 2016, Dec. 2016.

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

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

[51] E. J. Candés and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, March 2008.

[52] C. D. Meyer, *Matrix analysis and applied linear algebra*. Siam, 2000, vol. 71.

[53] M. Chen, W. Carson, M. Rodrigues, R. Calderbank, and L. Carin, "Communications inspired linear discriminant analysis," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ser. ICML'12. USA: Omnipress, 2012, pp. 1507–1514.

[54] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[55] H. Reboredo, F. Renna, R. Calderbank, and M. R. D. Rodrigues, "Bounds on the number of measurements for reliable compressive classification," *IEEE Transactions on Signal Processing*, vol. 64, no. 22, pp. 5778–5793, Nov 2016.

[56] I. Arganda-Carreras, C. O. S. Sorzano, R. Marabini, J. M. Carazo, C. Ortiz-de Solorzano, and J. Kybic, "Consistent and elastic registration of histological sections using vector-spline regularization," in *Computer Vision Approaches to Medical Image Analysis*, R. R. Beichel and M. Sonka, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 85–95.

[57] F. Zhang, *The Schur Complement and Its Applications*, ser. Numerical Methods and Algorithms. New York: Springer, 2005, vol. 4.

[58] Y. Tian and G. P. Styan, "Rank equalities for idempotent matrices with applications," *Journal of Computational and Applied Mathematics*, vol. 191, no. 1, pp. 77 – 97, 2006.

[59] C. C. Paige and M. A. Saunders, "Towards a generalized singular value decomposition," *SIAM Journal on Numerical Analysis*, vol. 18, no. 3, pp. 398–405, 1981.

[60] H. Reboredo, F. Renna, A. R. Calderbank, and M. R. D. Rodrigues, "Compressive classification of a mixture of gaussians: Analysis, designs and geometrical interpretation," *CoRR*, vol. abs/1401.6962, 2014. [Online]. Available: http://arxiv.org/abs/1401.6962