Underdetermined Instantaneous Audio Source Separation via Local Gaussian Modeling

Emmanuel Vincent, Simon Arberet, and Rémi Gribonval

METISS Group, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France
{emmanuel.vincent,simon.arberet,remi.gribonval}@irisa.fr

Abstract. Underdetermined source separation is often carried out by modeling time-frequency source coefficients via a fixed sparse prior. This approach fails when the number of active sources in one time-frequency bin is larger than the number of channels or when active sources lie on both sides of an inactive source. In this article, we partially address these issues by modeling time-frequency source coefficients via Gaussian priors with free variances. We study the resulting maximum likelihood criterion and derive a fast non-iterative optimization algorithm that finds the global minimum. We show that this algorithm outperforms state-of-the-art approaches over stereo instantaneous speech mixtures.

1 Introduction

Underdetermined source separation is the problem of recovering the single-channel source signals $s_j(t)$, $1 \leq j \leq J$, underlying a multichannel mixture signal $x_i(t)$, $1 \leq i \leq I$, with $I < J$. The mixing process can be modeled in the time-frequency domain via the Short-Term Fourier Transform (STFT) as

$$\mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{s}(n,f) \tag{1}$$

where $\mathbf{s}(n,f)$ is the vector of source STFT coefficients in time-frequency bin (n,f), $\mathbf{x}(n,f)$ is the vector of mixture STFT coefficients in that bin, and $\mathbf{A}(f)$ is a complex mixing matrix. This problem can be addressed by first estimating the mixing matrices then computing the Maximum *A Posteriori* (MAP) source coefficients given some prior distribution and inverting the STFT. For audio data, a common sparse prior such as the Laplacian [1], a mixture of Gaussians [2] or a generalized Gaussian [3], is usually assumed for all source coefficients. This model suffers from two issues. Firstly, a maximum number of I nonzero coefficients can often be recovered in each time-frequency bin, with the $J - I$ remaining coefficients being estimated as zero [1,3]. Secondly, the corresponding columns of the mixing matrix must point towards the closest directions to the observed mixture direction. A proof is given in [4] for a Laplacian prior.

In this paper, we aim to overcome both issues in the popular setting of stereo ($I = 2$) instantaneous mixtures, where the mixing matrices $\mathbf{A}(f)$ are equal to the same real-valued matrix $\mathbf{A}$ for all f. We assume that $\mathbf{A}$ is known and that its

columns are pairwise linearly independent, *i.e.* the sources have different directions. We build upon the Statistically Sparse Decomposition Principle (SSDP) presented in [4], which addresses the second issue using the correlation between the mixture channels but provides poor separation performance due to time-domain modeling and constraining of the number of nonzero sources per bin.

The structure of the paper is as follows. We apply the SSDP to the time-frequency domain in Section 2 and prove that it implicitly assumes a local Gaussian source model in Section 3. In Section 4, we extend this model to a larger number of nonzero sources and derive a new source separation algorithm. We evaluate its performance on speech data in Section 5 and conclude in Section 6.

2 Time-frequency statistically sparse decomposition

The SSDP is based on the empirical multichannel covariance matrix of the mixture over short time frames. In the time-frequency domain, we define this quantity over the neighborhood of each time-frequency bin (n, f) instead as

$$\widehat{\mathbf{R}}_{\mathbf{xx}}(n,f) = \frac{1}{\sum_{n',f'} w(n'-n, f'-f)} \sum_{n',f'} w(n'-n, f'-f)\mathbf{x}(n',f')\mathbf{x}(n',f')^H \tag{2}$$

where w is a bi-dimensional window specifying the shape of the neighborhood and ^{H} denotes the conjugate transpose of a matrix. In the rest of the paper, bin indexes (n, f) are dropped for the sake of legibility. The quantity (2) has long been exploited by mixing matrix estimation methods, *e.g.* [5,6], to obtain accurate direction estimates by selecting the bins where a single source is active. These bins are characterized by the fact that the cross-correlation between the mixture channels, also termed interchannel coherence, is high.

More generally, the cross-correlation is higher when the active sources have close directions. This fact can be exploited for source separation as follows. Let us assume that the number of active sources in each time-frequency bin is equal to two. For each pair of source indexes (j_1, j_2), the empirical covariance matrix of these sources may be defined as [4]

$$\widehat{\mathbf{R}}_{\mathbf{s}_{j_1 j_2}\mathbf{s}_{j_1 j_2}} = \mathbf{A}_{j_1 j_2}^{-1}\, \widehat{\mathbf{R}}_{\mathbf{xx}}\, (\mathbf{A}_{j_1 j_2}^{-1})^T \tag{3}$$

where $\mathbf{A}_{j_1 j_2}$ is the 2×2 matrix composed of the columns $\mathbf{A}_j$ of $\mathbf{A}$ indexed by $j \in \{j_1, j_2\}$ and ^{T} denotes transposition. The best pair of active sources may be selected via the SSDP [4]

$$(\widehat{j}_1, \widehat{j}_2) = \arg\min_{j_1, j_2} \frac{|\widehat{R}_{s_{j_1} s_{j_2}}|}{\sqrt{\widehat{R}_{s_{j_1} s_{j_1}}\, \widehat{R}_{s_{j_2} s_{j_2}}}} \tag{4}$$

with $\widehat{R}_{s_{j_k} s_{j_l}}$ denoting the (k, l)-th entry of $\widehat{\mathbf{R}}_{\mathbf{s}_{j_1 j_2}\mathbf{s}_{j_1 j_2}}$. The source STFT coefficients are then estimated by local mixing inversion as

$$\begin{cases} \widehat{\mathbf{s}}_{\widehat{j}_1 \widehat{j}_2}(n,f) = \mathbf{A}_{\widehat{j}_1 \widehat{j}_2}^{-1}\, \mathbf{x}(n,f) \\ \widehat{s}_j(n,f) = 0 \qquad\qquad\qquad\qquad \text{for all } j \notin \{\widehat{j}_1, \widehat{j}_2\}. \end{cases} \tag{5}$$

3 Interpretation as constrained local Gaussian modeling

This algorithm admits the following probabilistic interpretation. Let us assume that the source coefficients follow independent zero-mean Gaussian priors over the neighborhood of each time-frequency bin (n, f) whose variances v_j depend on that bin. This assumption appears well suited to audio signals, which are typically non-sparse over small time-frequency regions but non-stationary hence sparse over larger regions. Given this model, the mixture coefficients in a given neighborhood follow a zero-mean Gaussian prior with covariance matrix

$$\mathbf{R_{xx}} = \mathbf{A}\,\mathbf{Diag}(\mathbf{v})\,\mathbf{A}^T. \tag{6}$$

where the operator **Diag** applied to a vector denotes the diagonal matrix whose entries are those of the vector. The log-likelihood of the source variances is equal up to a constant to minus the Kullback-Leibler (KL) divergence $KL(\widehat{\mathbf{R}}_{\mathbf{xx}}|\mathbf{R_{xx}})$ between the empirical and expected mixture covariances [7][1] with

$$KL(\widehat{\mathbf{R}}|\mathbf{R}) = \frac{1}{2}[\mathrm{tr}(\mathbf{R}^{-1}\widehat{\mathbf{R}}) - \log\det(\mathbf{R}^{-1}\widehat{\mathbf{R}})] - 1. \tag{7}$$

Assuming that at most two sources have nonzero variance, their indexes and variances may be estimated in the Maximum Likelihood (ML) sense by

$$(\widehat{j}_1, \widehat{j}_2, \widehat{\mathbf{v}}_{\widehat{j}_1\widehat{j}_2}) = \arg\min_{j_1,j_2,\mathbf{v}_{j_1j_2}\geq\mathbf{0}} KL(\widehat{\mathbf{R}}_{\mathbf{xx}}|\mathbf{R_{xx}}). \tag{8}$$

The KL divergence is invariant under invertible linear transforms. When applied to $\mathbf{A}_{j_1j_2}^{-1}$, this property yields

$$KL(\widehat{\mathbf{R}}_{\mathbf{xx}}|\mathbf{R_{xx}}) = KL(\widehat{\mathbf{R}}_{\mathbf{s}_{j_1j_2}\mathbf{s}_{j_1j_2}}|\mathbf{Diag}(\mathbf{v}_{j_1j_2})) \tag{9}$$

$$= \frac{1}{2}\left[\frac{\widehat{R}_{s_{j_1}s_{j_1}}}{v_{j_1}} + \frac{\widehat{R}_{s_{j_2}s_{j_2}}}{v_{j_2}} - \log\frac{\widehat{R}_{s_{j_1}s_{j_1}}\widehat{R}_{s_{j_2}s_{j_2}} - |\widehat{R}_{s_{j_1}s_{j_2}}|^2}{v_{j_1}v_{j_2}}\right] - 1. \tag{10}$$

By finding the zeroes of the partial derivatives of this expression with respect to v_{j_1} and v_{j_2}, we get

$$\begin{cases} \widehat{v}_{j_1} = \widehat{R}_{s_{j_1}s_{j_1}} \quad \text{and} \quad \widehat{v}_{j_2} = \widehat{R}_{s_{j_2}s_{j_2}} \\ (\widehat{j}_1, \widehat{j}_2) = \arg\min_{j_1,j_2} -\frac{1}{2}\log\left(1 - \dfrac{|\widehat{R}_{s_{j_1}s_{j_2}}|^2}{\widehat{R}_{s_{j_1}s_{j_1}}\widehat{R}_{s_{j_2}s_{j_2}}}\right). \end{cases} \tag{11}$$

This criterion is equivalent to (4), hence the SSDP does estimate the two sources with nonzero variance in the ML sense. In addition, the ML variances of these sources are equal to the diagonal entries of the empirical source covariance matrix. It can also be shown that the MAP source coefficients given the ML source variances are obtained via (5).

[1] This relation holds provided that $\widehat{\mathbf{R}}_{\mathbf{xx}}$ has full rank. We consider the KL divergence because of its well-known invariance and nonnegativity properties. However it can be shown from the expression of the log-likelihood that the following derivations remain true otherwise.

4 Minimally constrained local Gaussian modeling

While the SSDP allows the separation of sources not pointing to close directions, the number of nonzero source coefficients that can be estimated in each time-frequency bin remains constrained to two. The above probabilistic interpretation provides a natural way of relaxing this constraint by assuming that the source coefficients follow independent zero-mean Gaussian priors over the neighborhood of each time-frequency bin, whose variances v_j are free. This model has been exploited in the context of determined mixtures, albeit with the different goal of estimating the mixing matrix given estimates of the source variances [7]. In the under-determined context, ML variance estimates are obtained by

$$\widehat{\mathbf{v}} = \arg\min_{\mathbf{v}\geq\mathbf{0}} KL(\widehat{\mathbf{R}}_{\mathbf{xx}}|\mathbf{R}_{\mathbf{xx}}) \tag{12}$$

and MAP source coefficients are classically derived via the Wiener filter

$$\widehat{\mathbf{s}}(n,f) = \mathbf{Diag}(\widehat{\mathbf{v}})\,\mathbf{A}^T\,(\mathbf{A}\,\mathbf{Diag}(\widehat{\mathbf{v}})\,\mathbf{A}^T)^{-1}\,\mathbf{x}(n,f). \tag{13}$$

The above vector minimization problem may be solved via standard iterative optimization techniques based on the gradient. However these methods are computationally intensive and the result may be a local minimum or one of several possible global minima. We avoid these issues by characterizing the minima. We show below that global minima with three or more nonzero entries satisfy the equality $\mathbf{R}_{\mathbf{xx}} = \Re(\widehat{\mathbf{R}}_{\mathbf{xx}})$, where $\Re$ denotes the real part of a complex matrix. If no vector satisfies this equality, the global minima consequently have two nonzero entries and can be obtained via the SSDP as shown in Section 3. This suggests an efficient way of computing the global minima: find the vectors $\mathbf{v} \geq \mathbf{0}$ such that $\mathbf{R}_{\mathbf{xx}} = \Re(\widehat{\mathbf{R}}_{\mathbf{xx}})$ and, if none exists, apply the SSDP instead. We also study below the cases where several solutions arise and propose minimal constraints to select a single solution. The reader is advised to skip the proofs of the following lemmas on first reading and to proceed directly with the details of Algorithm 1 at the end of this section.

Lemma 1. *The KL divergence criterion is always larger than* $KL(\widehat{\mathbf{R}}_{\mathbf{xx}}|\Re(\widehat{\mathbf{R}}_{\mathbf{xx}}))$ *and equal to that value if and only if* $\mathbf{R}_{\mathbf{xx}} = \Re(\widehat{\mathbf{R}}_{\mathbf{xx}})$.

Proof. Since the mixing matrix $\mathbf{A}$ is real-valued, $\mathbf{R}_{\mathbf{xx}}$ is real-valued and admits a real-valued square root $\mathbf{R}_{\mathbf{xx}}^{1/2}$. The matrix $\mathbf{R}_{\mathbf{xx}}^{-1/2}\widehat{\mathbf{R}}_{\mathbf{xx}}\mathbf{R}_{\mathbf{xx}}^{-1/2}$ is Hermitian, hence using the commutativity of the trace $\mathrm{tr}(\mathbf{R}_{\mathbf{xx}}^{-1}\widehat{\mathbf{R}}_{\mathbf{xx}}) = \mathrm{tr}(\mathbf{R}_{\mathbf{xx}}^{-1/2}\widehat{\mathbf{R}}_{\mathbf{xx}}\mathbf{R}_{\mathbf{xx}}^{-1/2}) = \mathrm{tr}(\mathbf{R}_{\mathbf{xx}}^{-1/2}\Re(\widehat{\mathbf{R}}_{\mathbf{xx}})\mathbf{R}_{\mathbf{xx}}^{-1/2}) = \mathrm{tr}(\mathbf{R}_{\mathbf{xx}}^{-1}\Re(\widehat{\mathbf{R}}_{\mathbf{xx}}))$. By combining this equality with (7), we get $KL(\widehat{\mathbf{R}}_{\mathbf{xx}}|\mathbf{R}_{\mathbf{xx}}) = KL(\Re(\widehat{\mathbf{R}}_{\mathbf{xx}})|\mathbf{R}_{\mathbf{xx}}) + \log\det(\widehat{\mathbf{R}}_{\mathbf{xx}}^{-1}\Re(\widehat{\mathbf{R}}_{\mathbf{xx}}))$. The second term of this equation does not depend on $\mathbf{v}$, while the first term is nonnegative and equal to zero if and only if $\mathbf{R}_{\mathbf{xx}} = \Re(\widehat{\mathbf{R}}_{\mathbf{xx}})$ by property of the KL divergence. □

Lemma 2. *If* $\mathbf{v}$ *is a local minimum of the criterion with* $K \geq 3$ *nonzero entries* $v_{j_1}, \ldots, v_{j_K}$*, then* $\mathbf{v}$ *is a global minimum and satisfies* $\mathbf{R}_{\mathbf{xx}} = \Re(\widehat{\mathbf{R}}_{\mathbf{xx}})$.

Proof. The gradient of the criterion is given by

$$\frac{\partial KL(\widehat{\mathbf{R}}_{\mathbf{xx}}|\mathbf{R}_{\mathbf{xx}})}{\partial v_j} = \langle \mathbf{E}, \mathbf{A}_j \mathbf{A}_j^T \rangle \quad \text{where } \mathbf{E} = \frac{1}{2}[\mathbf{R}_{\mathbf{xx}}^{-1}(\mathbf{R}_{\mathbf{xx}} - \Re(\widehat{\mathbf{R}}_{\mathbf{xx}}))\mathbf{R}_{\mathbf{xx}}^{-1}] \quad (14)$$

and $\langle ., . \rangle$ is the Euclidean dot product over the space $S_2(\mathbb{R})$ of real-valued symmetric 2×2 matrices. If $\mathbf{v}$ is a local extremum, the gradient is zero for all entries v_j that are not boundaries of the optimization domain. Hence $\mathbf{E}$ is orthogonal to the matrices $\mathbf{A}_j \mathbf{A}_j^T$, $j \in \{j_1, j_2, j_3\}$.

Let us consider the 3×3 matrix $\mathbf{B}_{j_1 j_2 j_3}$ whose columns consist of the upper triangular entries of the latter matrices:

$$\mathbf{B}_{j_1 j_2 j_3} = \begin{pmatrix} A^2_{1_{j_1}} & A^2_{1_{j_2}} & A^2_{1_{j_3}} \\ A^2_{2_{j_1}} & A^2_{2_{j_2}} & A^2_{2_{j_3}} \\ A_{1_{j_1}} A_{2_{j_1}} & A_{1_{j_2}} A_{2_{j_2}} & A_{1_{j_3}} A_{2_{j_3}} \end{pmatrix}. \quad (15)$$

By computing and factoring the determinant of $\mathbf{B}_{j_1 j_2 j_3}$, we get

$$\det \mathbf{B}_{j_1 j_2 j_3} = \det \mathbf{A}_{j_1 j_2} \det \mathbf{A}_{j_2 j_3} \det \mathbf{A}_{j_3 j_1}. \quad (16)$$

Since the columns $\mathbf{A}_j$ of $\mathbf{A}$ are pairwise linearly independent, all the terms of this equation are nonzero and the columns of $\mathbf{B}_{j_1 j_2 j_3}$ form a basis of $\mathbb{R}^3$. This is equivalent to $\mathbf{A}_j \mathbf{A}_j^T$, $j \in \{j_1, j_2, j_3\}$, being a basis of $S_2(\mathbb{R})$.

We deduce from the above results that $\mathbf{E} = \mathbf{0}$ hence $\mathbf{R}_{\mathbf{xx}} = \Re(\widehat{\mathbf{R}}_{\mathbf{xx}})$. Therefore $\mathbf{v}$ is a global minimum of the criterion according to lemma 1. □

Lemma 3. *The matrix equality* $\mathbf{R}_{\mathbf{xx}} = \Re(\widehat{\mathbf{R}}_{\mathbf{xx}})$ *can be equivalently rewritten as*

$$\mathbf{B}_{j_1 \dots j_K} \mathbf{v}_{j_1 \dots j_K} = \widehat{\mathbf{w}} \quad (17)$$

where $\mathbf{v}_{j_1 \dots j_K}$ *is the vector of nonzero entries of* $\mathbf{v}$*,*

$$\mathbf{B}_{j_1 \dots j_K} = \begin{pmatrix} A^2_{1_{j_1}} & \dots & A^2_{1_{j_K}} \\ A^2_{2_{j_1}} & \dots & A^2_{2_{j_K}} \\ A_{1_{j_1}} A_{2_{j_1}} & \dots & A_{1_{j_K}} A_{2_{j_K}} \end{pmatrix} \quad \textit{and} \quad \widehat{\mathbf{w}} = \begin{pmatrix} \widehat{R}_{x_1 x_1} \\ \widehat{R}_{x_2 x_2} \\ \Re(\widehat{R}_{x_1 x_2}) \end{pmatrix}. \quad (18)$$

Proof. From (6), $\mathbf{R}_{\mathbf{xx}} = \Re(\widehat{\mathbf{R}}_{\mathbf{xx}})$ is equivalent to $\sum_{k=1}^{K} v_{j_k} \mathbf{A}_{j_k} \mathbf{A}_{j_k}^T = \Re(\widehat{\mathbf{R}}_{\mathbf{xx}})$. By rearranging the upper triangular terms of this matrix equality into vectors, this is in turn equivalent to (17). □

Lemma 4. *With* $J \geq 4$ *sources, if the criterion admits a global minimum with* $K \geq 3$ *nonzero entries, then there exists a global minimum with* $K \leq 3$ *nonzero entries. Moreover, if* $\mathbf{A}$ *is nonnegative and there is a global minimum with* $K = 3$ *nonzero entries, then there are several global minima with* $K \leq 3$ *nonzero entries.*

Proof. Let $\mathbf{v}$ be a global minimum of the criterion with $K \geq 4$ nonzero entries. According to lemma 2 and its proof, $\mathbf{v}$ satisfies (17) and $\mathbf{B}_{j_1 \dots j_K}$ has rank 3. The null space of $\mathbf{B}_{j_1 \dots j_K}$ is therefore of dimension $K - 3 > 0$. Let $\mathbf{z}$ be a vector such

that $\mathbf{z}_{j_1 \ldots j_K} \neq \mathbf{0}$ belongs to that null space and $z_j = 0$ for all $j \notin \{j_1, \ldots, j_K\}$. We define the vector $\mathbf{v}'$ as

$$\mathbf{v}' = \mathbf{v} - \frac{v_{j_l}}{z_{j_l}} \mathbf{z} \quad \text{with } l = \arg \min_{k, z_{j_k} \neq 0} \frac{v_{j_k}}{|z_{j_k}|}. \tag{19}$$

The entries of this vector are given by $v'_{j_k} = v_{j_k} - v_{j_l} z_{j_k} / z_{j_l}$. Clearly, $v'_j = 0$ for all $j \notin \{j_1, \ldots, j_K\}$ and $v'_{j_l} = 0$. If z_{j_k} and z_{j_l} have different signs, then $z_{j_k}/z_{j_l} \leq 0$ and $v'_{j_k} \geq v_{j_k} \geq 0$. If z_{j_k} and z_{j_l} have the same sign, then $v'_{j_k} = v_{j_k} - v_{j_l} |z_{j_k}| / |z_{j_l}| \geq 0$ given (19). Hence $\mathbf{v}'$ has nonnegative entries and at most $K-1$ positive entries. In addition, $\mathbf{B}_{j_1 \ldots j_{l-1} j_{l+1} \ldots j_K} \mathbf{v}'_{j_1 \ldots j_{l-1} j_{l+1} \ldots j_K} = \mathbf{B}_{j_1 \ldots j_K} \mathbf{v}'_{j_1 \ldots j_K} = \mathbf{B}_{j_1 \ldots j_K} \mathbf{v}_{j_1 \ldots j_K} - v_{j_l}/z_{j_l} \mathbf{B}_{j_1 \ldots j_K} \mathbf{z}_{j_1 \ldots j_K} = \widehat{\mathbf{w}} - v_{j_l}/z_{j_l} \mathbf{0} = \widehat{\mathbf{w}}$. This shows that $\mathbf{v}'$ is a global minimum of the criterion with $K' \leq K-1$ nonzero entries. By recurrently applying the above construction, we find a global minimum $\mathbf{v}''$ with $K'' \leq 3$ nonzero entries.

Let us now assume that $\mathbf{A}$ is nonnegative and $K'' = 3$. We denote by j_1, j_2, j_3 the nonzero entries of $\mathbf{v}''$ and by j_4 any other index. Since the matrix $\mathbf{B}_{j_1 j_2 j_3 j_4}$ is nonnegative and all its 3×3 submatrices have rank 3, the non-null vectors of its null space have no zero entry and both positive and negative entries. Let $\mathbf{z}'$ be a vector such that $\mathbf{z}'_{j_1 j_2 j_3 j_4} \neq \mathbf{0}$ belongs to that null space, $z'_{j_4} < 0$ and $z'_j = 0$ for all $j \notin \{j_1, j_2, j_3, j_4\}$. We define the vector $\mathbf{v}'''$ as

$$\mathbf{v}''' = \mathbf{v}'' - \frac{v''_{j_l}}{z'_{j_l}} \mathbf{z}' \quad \text{with } l = \arg \min_{k, z'_{j_k} > 0} \frac{v''_{j_k}}{z'_{j_k}}. \tag{20}$$

Similarly to above, it can be proved that $\mathbf{v}'''$ is a global minimum of the criterion with $K''' \leq 3$ nonzero entries indexed by some $j \in \{j_1, j_2, j_3, j_4\}$ and $j \neq j_l$. □

Lemma 4 shows that ML estimation of the source variances is an ill-posed problem with $J \geq 4$ sources. Appropriate constraints must be set over the source variances in order to obtain a unique solution. While probabilistic hyperpriors may model flexible constraints, the resulting MAP solution may not match any of the ML solutions, so that the benefit of characterizing ML solutions is lost. Instead, we select the sparsest ML solution: we restrict the optimization domain to vectors with $K \leq 3$ nonzero entries and select the ML solution with minimum l_p norm $\|\widehat{\mathbf{v}}\|_p$ [3] in case several ML solutions can be found in this domain.

Given these constraints and the characterization of ML source variances in Section 3 and lemma 3, we perform source separation in each time-frequency bin (n, f) via the following fast global optimization algorithm.

Algorithm 1

1. *Compute the empirical mixture covariance* $\widehat{\mathbf{R}}_{\mathbf{xx}}$ *in* (2) *and derive the vector* $\widehat{\mathbf{w}}$ *in* (18).
2. *Compute the candidate source variances* $\mathbf{v}_{j_1 j_2 j_3} = \mathbf{B}^{-1}_{j_1 j_2 j_3} \widehat{\mathbf{w}}$ *for all triplets of source indexes* $\{j_1, j_2, j_3\}$*, with* $\mathbf{B}_{j_1 j_2 j_3}$ *defined in* (15).
3. *If some candidates have positive entries only, then they are solutions of the ML estimation problem. Select the one with minimum* l_p *norm among these and derive the MAP source coefficients via* (13).

4. *Otherwise, compute the empirical source covariance matrices* $\widehat{\mathbf{R}}_{\mathbf{s}_{j_1 j_2}\mathbf{s}_{j_1 j_2}} = \mathbf{A}_{j_1 j_2}^{-1}\,\widehat{\mathbf{R}}_{\mathbf{xx}}\,(\mathbf{A}_{j_1 j_2}^{-1})^T$ *for all pairs of source indexes* $\{j_1, j_2\}$. *Select the ML pair via* (4) *and estimate the MAP source coefficients via* (5).

5 Experimental results

We evaluated this algorithm over the speech data in [8]. The number of sources J was varied from 3 to 6. For each J, a nonnegative mixing matrix was computed from [9], given an angle of $50 - 5J$ degrees between successive sources, and ten instantaneous mixtures were generated from different source signals resampled at 8 kHz. The STFT was computed with a sine window of length 512 (64 ms). The bi-dimensional window w defining time-frequency neighborhoods was chosen as the outer product of two rectangular or Hanning windows with variable length. The l_p norm exponent was set to $p \to 0$ [3]. The results were evaluated via the Signal-to-Distortion Ratio (SDR) defined in [10]. The best results were achieved for w chosen as the outer product of two Hanning windows of length 3. The computation time was then between 1.8 and 3.7 times the mixture duration depending on J, using Matlab on a 1.2 GHz dual core CPU.

Figure 1 compares the average SDR achieved by the proposed algorithm, the time-frequency domain SSDP in Section 2 and two state-of-the-art algorithms: l_p norm minimization [3] and DUET [11]. The proposed algorithm outperforms all other algorithms whatever the number of sources. Nevertheless, it should be noted that its performance remains about 10 dB below the theoretical upper bound obtained by local mixing inversion (5) given the best pair of active sources [8] and 11 dB below the theoretical upper bound obtained by Wiener filtering (13) given the true sources variances.

This algorithm was submitted to the 2008 Signal Separation Evaluation Campaign with the same parameters, except a STFT window length of 1024 and step size of 256. Mixing matrices were estimated via the software in [6].

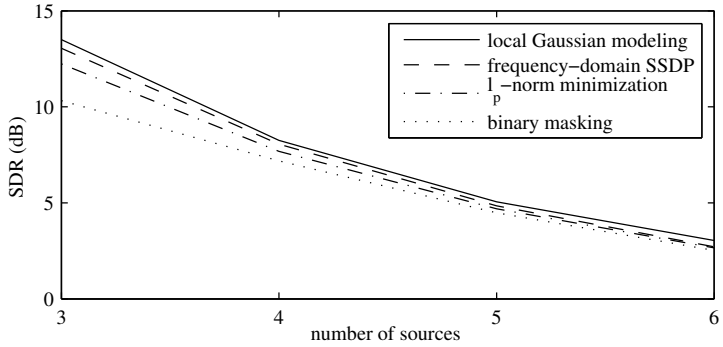

Fig. 1. Source separation performance over stereo instantaneous speech mixtures.

6 Conclusion

In this paper, we proposed a new source separation algorithm for stereo instantaneous mixtures based on the modeling of source STFT coefficients via local Gaussian priors with minimally constrained variances. This algorithm can estimate up to three nonzero source coefficients in each bin, as opposed to two for state-of-the-art methods, and provides improved separation performance. This suggests that local mixture covariance can be successfully exploited for underdetermined source separation in addition to mixing matrix estimation. Further work includes the generalization of this algorithm to convolutive mixtures with $I \geq 2$ channels. A larger improvement is expected, since up to $I(I+1)/2$ nonzero source coefficients could be estimated in each time-frequency bin. Local nongaussian source priors could also be investigated.

References

1. Zibulevsky, M., Pearlmutter, B.A., Bofill, P., Kisilev, P.: Blind source separation by sparse decomposition in a signal dictionary. In: Independent Component Analysis : Principles and Practice. Cambridge Press (2001) 181–208
2. Davies, M.E., Mitianoudis, N.: Simple mixture model for sparse overcomplete ICA. IEE Proceedings on Vision, Image and Signal Processing **151**(1) (2004) 35–43
3. Vincent, E.: Complex nonconvex l_p norm minimization for underdetermined source separation. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA). (2007) 430–437
4. Xiao, M., Xie, S., Fu, Y.: A statistically sparse decomposition principle for underdetermined blind source separation. In: Proc. Int. Symp. on Intelligent Signal Processing and Communication Systems (ISPACS). (2005) 165–168
5. Belouchrani, A., Amin, M.G., Abed-Meraïm, K.: Blind source separation based on time-frequency signal representations. IEEE Trans. on Signal Processing **46**(11) (1998) 2888–2897
6. Arberet, S., Gribonval, R., Bimbot, F.: A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA). (2006) 536–543
7. Pham, D.T., Cardoso, J.F.: Blind separation of instantaneous mixtures of non stationary sources. IEEE Trans. on Signal Processing **49**(9) (2001) 1837–1848
8. Vincent, E., Gribonval, R., Plumbley, M.D.: Oracle estimators for the benchmarking of source separation algorithms. Signal Processing **87**(8) (2007) 1933–1950
9. Pulkki, V., Karjalainen, M.: Localization of amplitude-panned virtual sources I: stereophonic panning. Journal of the Audio Engineering Society **49**(9) (2001) 739–752
10. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Trans. on Audio, Speech and Language Processing **14**(4) (2006) 1462–1469
11. Yılmaz, O., Rickard, S.T.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. on Signal Processing **52**(7) (2004) 1830–1847