Spatial location priors for Gaussian model based reverberant audio source separation

Ngoc Duong, Emmanuel Vincent, Rémi Gribonval

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Spatial location priors for Gaussian model based reverberant audio source separation

Ngoc QK Duong[1], Emmanuel Vincent[2*] and Rémi Gribonval[3]

Abstract

We consider the Gaussian framework for reverberant audio source separation, where the sources are modeled in the time-frequency domain by their short-term power spectra and their spatial covariance matrices. We propose two alternative probabilistic priors over the spatial covariance matrices which are consistent with the theory of statistical room acoustics and we derive expectation-maximization algorithms for maximum a posteriori (MAP) estimation. We argue that these algorithms provide a statistically principled solution to the permutation problem and to the risk of overfitting resulting from conventional maximum likelihood (ML) estimation. We show experimentally that in a semi-informed scenario where the source positions and certain room characteristics are known, the MAP algorithms outperform their ML counterparts. This opens the way to rigorous statistical treatment of this family of models in other scenarios in the future.

Keywords: Audio source separation; Spatial covariance; EM algorithm; Probabilistic priors; Inverse-Wishart; Gaussian

1 Introduction

We consider the task of reverberant audio source separation, that is, to extract individual sound sources from a multichannel microphone array recording. Many approaches have been proposed in the literature, which typically operate in the time-frequency domain via the short-time Fourier transform (STFT) [1-3]. One category of approaches models the mixture STFT coefficients as the product of the source STFT coefficients and complex-valued *mixing vectors*, which are estimated by frequency-domain independent component analysis (FDICA) [4,5] or by clustering [6,7]. In under-determined conditions when the number of sources is greater than the number of channels, the source STFT coefficients are then obtained via binary masking [6], soft masking [7], or ℓ_1-norm minimization [8]. Lately, a Gaussian framework has emerged where the mixture STFT coefficients are modeled as a function of the power spectra and the *spatial covariance matrices* of the sources, and separation is achieved by multichannel Wiener filtering [9-11]. These covariance matrices may equivalently be expressed as the outer product of *subsource mixing matrices*, which reduce to mixing vectors when the spatial covariance matrices have rank 1 [12]. Full-rank matrices have been shown to improve separation performance in reverberant conditions by modeling not only the spatial position of the sources but also their spatial width [11].

While a number of deterministic [12-14] and probabilistic [15-17] priors have been proposed over the source spectra, the mixing vectors and the source spatial covariance matrices are usually estimated in an unconstrained manner. The lack of a constraint relating these quantities across frequency causes a *permutation problem*, which has been coped with by reordering the estimates in each frequency bin while keeping their value [7,18]. More crucially, the estimated values of the mixing vectors and the source spatial covariance matrices in a given frequency bin are likely to suffer from *overfitting* when the corresponding sources are little active in that bin.

Building upon the studies for instantaneous mixtures in [19,20] and the deterministic subspace constraints in [21,22], a few algorithms have been designed that exploit soft penalties or probabilistic priors over the mixing vectors for increased estimation accuracy. These algorithms typically target semi-informed scenarios such as formal meetings or in-car speech where the spatial locations of the sources are known and they rely on the assumption

*Correspondence: emmanuel.vincent@inria.fr
[2]Inria, 54600 Villers-lès-Nancy, France
Full list of author information is available at the end of the article

that the mixing vectors are close to the steering vectors representing the direct path from the sources to the microphones. Squared Euclidean penalties over the blocking vectors are a common choice for FDICA [21,23]. An inverse-Wishart prior over the outer product of the mixing vectors was also employed in [24]. These penalties and priors were not designed according to the actual statistics of reverberation. Moreover, to the best of our knowledge, no such priors have been designed for full-rank matrices.

In this article, we propose two probabilistic priors over the source spatial covariance matrices or the subsource mixing matrices which are consistent with the theory of statistical room acoustics. One of them was briefly introduced in our preliminary paper [25]. We extend the two Gaussian expectation-maximization (EM) algorithms in [12,26] so as to perform maximum a posteriori (MAP) estimation. We then compare the resulting separation performance with conventional maximum likelihood (ML) estimation and with two baseline approaches in an underdetermined full-rank semi-informed scenario where the source positions and certain room characteristics are known. For clarity, we do not assume any other constraint on the model parameters, which allows us to assess the improvement resulting from these priors alone.

The structure of the article is as follows. In Section 2, we recall the Gaussian framework for audio source separation and we present a result of the theory of statistical room acoustics. We introduce an EM algorithm using an inverse-Wishart prior in Section 3 and an EM algorithm using a Gaussian prior in Section 4. We evaluate their separation performance in Section 5 and we conclude in Section 6.

2 Gaussian modeling and statistical room acoustics

2.1 Gaussian modeling for source separation

Let us consider a mixture signal $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^{\mathrm{T}}$ recorded by an array of I microphones. Denoting by J the number of sources, the mixing process is expressed as [27]

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{1}$$

where $\mathbf{c}_j(t) = [c_{1j}(t), \ldots, c_{Ij}(t)]^{\mathrm{T}}$ is the *spatial image* of the jth source, which is its contribution to the signals recorded at the microphones. The STFT coefficients $\mathbf{c}_j(n,f)$ of the source spatial images in each time frame n and each frequency bin f are modeled as zero-mean Gaussian random vectors

$$\mathbf{c}_j(n,f) \sim \mathcal{N}(\mathbf{0}, v_j(n,f)\,\mathbf{R}_j(f)) \tag{2}$$

where $v_j(n,f)$ are scalar nonnegative *variances* encoding the short-term power spectra of the sources and $\mathbf{R}_j(f)$ are $I \times I$ spatial covariance matrices encoding their spatial position and their spatial width [9,11].

Under the assumption that the sources are uncorrelated, the mixture covariance matrix $\Sigma_{\mathbf{x}}(n,f)$ is equal to

$$\Sigma_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(f). \tag{3}$$

The log-likelihood is then given by [26]

$$\log \mathcal{L} = \sum_{nf} -\mathrm{tr}(\Sigma_{\mathbf{x}}^{-1}(n,f)\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)) - \log|\pi\,\Sigma_{\mathbf{x}}(n,f)| \tag{4}$$

where tr(.) and |.| denote the trace and the determinant of a square matrix, and $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$ is the *empirical mixture covariance matrix* obtained by local averaging of $\mathbf{x}(n,f)\mathbf{x}^{\mathrm{H}}(n,f)$ over the neighborhood of each time-frequency bin

$$\widehat{\mathbf{R}}_{\mathbf{x}}(n,f) = \sum_{n',f'} w_{nf}^2(n',f')\mathbf{x}(n',f')\mathbf{x}^{\mathrm{H}}(n',f') \tag{5}$$

where w_{nf} is a bi-dimensional window specifying the shape of the neighborhood [26].

Source separation can then be achieved by estimating the model parameters $\theta = \{v_j(n,f), \mathbf{R}_j(f)\}$ in the ML sense and by deriving the spatial images of all sources in the minimum mean square error sense via multichannel Wiener filtering of the mixture STFT coefficients $\mathbf{x}(n,f)$

$$\widehat{\mathbf{c}}_j(n,f) = v_j(n,f)\mathbf{R}_j(f)\Sigma_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \tag{6}$$

2.2 A result from the theory of statistical room acoustics

In a scenario such as in [21-23], the distance and the orientation of the sources and the microphones with respect to each other (aka, the scene geometry) are assumed to be known but their absolute position in the room is unknown. According to the theory of statistical room acoustics [28,29], the mean spatial covariance matrix of a source over all possible source and microphone positions and orientations, under the constraint that the scene geometry remains fixed, can be expressed as

$$\boldsymbol{\mu}_{\mathbf{R}_j}(f) = \mathbf{d}_j(f)\mathbf{d}_j^{\mathrm{H}}(f) + \sigma_{\mathrm{rev}}^2\Omega(f) \tag{7}$$

where $.^{\mathrm{H}}$ denotes conjugate transposition. The first term of this expression models the contribution of direct sound, where

$$\mathbf{d}_j(f) = \begin{pmatrix} \frac{1}{\sqrt{4\pi}\,r_{1j}} e^{-2i\pi f \frac{r_{1j}}{c}} \\ \vdots \\ \frac{1}{\sqrt{4\pi}\,r_{Ij}} e^{-2i\pi f \frac{r_{Ij}}{c}} \end{pmatrix} \tag{8}$$

is the steering vector representing the direct paths from the source to the microphones, with c the sound velocity and r_{ij} the distance from the jth source to the ith microphone. The second term of this expression models the contribution of echoes and reverberation, which are assumed to come from all possible directions on average over all absolute positions: σ^2_{rev} is the power of echoes and reverberation and $\Omega(f)$ is the covariance matrix of a diffuse sound field.

The entries $\Omega_{ii'}(f)$ of $\Omega(f)$ depend on the microphone directivity patterns and on the distance $d_{ii'}$ between the ith and the i'th microphone. For omnidirectional microphones, this quantity can be shown to be real-valued and equal to [28],

$$\Omega_{ii'}(f) = \frac{\sin(2\pi f d_{ii'}/c)}{2\pi f d_{ii'}/c}. \tag{9}$$

Moreover, the power of the reverberant part within a parallelepipedic room with dimensions L_x, L_y, L_z is given by

$$\sigma^2_{\text{rev}} = \frac{4\beta^2}{\mathcal{A}(1-\beta^2)} \tag{10}$$

where $\mathcal{A}$ is the total wall area and β the wall reflection coefficient computed from the room reverberation time T_{60} via Eyring's formula [29],

$$\beta = \exp\left\{-\frac{13.82}{(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z})cT_{60}}\right\}. \tag{11}$$

In order to match the physics of reverberation, a prior over the source spatial covariance matrices or over the subsource mixing matrices should lead to a mean spatial covariance matrix $\boldsymbol{\mu}_{\mathbf{R}_j}(f)$ satisfying the constraint (7). This is not the case of the prior in [24], whose mean is equal to $\mathbf{d}_j(f)\mathbf{d}_j^{\text{H}}(f) + \epsilon\mathbf{I}_I$ with $\mathbf{I}_I$ the identity matrix of size I and ϵ a small constant. Isotropic Gaussian priors over the subsource mixing matrices would not satisfy this constraint either due to the interchannel correlation introduced by $\Omega(f)$. Fixed spatial covariance matrices set to the value in (7) were employed for single source localization in [29] and for source separation in [30]. Later work confirmed that the model (7) is valid on average over all absolute positions in the room but that $\mathbf{R}_j(f)$ varies with the absolute position so that it must be estimated from the observed mixture signal [11].

3 Source image-based EM algorithms

3.1 General EM algorithm

Assuming that the spatial covariance matrices $\mathbf{R}_j(f)$ are full-rank, ML estimation can be achieved using the source image-based EM (SIEM) algorithm in [26] where the spatial images $\{\mathbf{c}_j(n,f)\}_{n,f}$ of all sources in all time-frequency bins are considered as *hidden data*. Strictly speaking, this algorithm is a generalized form of EM [31] because the M step increases but does not maximize the expectation of the log-likelihood of the hidden data. Since the priors proposed hereafter pertain to the spatial covariance matrices only, MAP estimation can be achieved via the same algorithm except for the corresponding update in the M step.

The resulting EM updates are listed in Algorithm 1. In the E step, the Wiener filter $\mathbf{W}_j(n,f)$ and the second-order raw moment $\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$ of the spatial images of all sources are computed[a]. In the M step $v_j(n,f)$ and $\mathbf{R}_j(f)$ are updated. In the ML case, the update for $\mathbf{R}_j(f)$ in (17) is given by [26]

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)}{v_j(n,f)} \tag{12}$$

where N is the total number of time frames.

Algorithm 1 SIEM algorithm [26]

E step:

$$\Sigma_{\mathbf{c}_j}(n,f) = v_j(n,f)\mathbf{R}_j(f) \tag{13}$$

$$\mathbf{W}_j(n,f) = \Sigma_{\mathbf{c}_j}(n,f)\Sigma_{\mathbf{x}}^{-1}(n,f) \tag{14}$$

$$\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f) = \mathbf{W}_j(n,f)\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)\mathbf{W}_j^{\text{H}}(n,f) + (\mathbf{I}_I - \mathbf{W}_j(n,f))\Sigma_{\mathbf{c}_j}(n,f) \tag{15}$$

M step:

$$v_j(n,f) = \frac{1}{I}\text{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)) \tag{16}$$

$$\text{Update } \mathbf{R}_j(f). \tag{17}$$

Given this algorithm, we now consider the design of suitable priors over $\mathbf{R}_j(f)$. In addition to the physical constraint (7), the priors must satisfy practical engineering constraints: they must be defined over the space of Hermitian positive definite matrices, have a small number of parameters, have a closed-form mean and result in closed-form EM updates. The inverse-Wishart and the Wishart distributions satisfy these constraints. In this paper we present only the inverse-Wishart prior since we observed experimentally that the Wishart prior results in poorer separation performance compared to both the ML algorithm and the MAP algorithm using the inverse-Wishart prior.

3.2 Inverse-Wishart prior

The inverse-Wishart distribution is the *conjugate prior* for the likelihood (4) of our model. This prior is defined as

$$\mathbf{R}_j(f) \sim \mathcal{IW}(\Psi_j(f), m) \tag{18}$$

where

$$\mathcal{IW}(\mathbf{R}|\Psi,m)=\frac{|\Psi|^m|\mathbf{R}|^{-(m+I)}e^{-\mathrm{tr}(\Psi\mathbf{R}^{-1})}}{\pi^{I(I-1)/2}\prod_{i=1}^{I}\Gamma(m-i+1)} \tag{19}$$

is the inverse-Wishart density over Hermitian positive definite matrices $\mathbf{R}$ with positive definite inverse scale matrix Ψ, m degrees of freedom, and mean $\Psi/(m-I)$ [32], with Γ the gamma function. This density, its mean, and its variance are finite for $m > I-1$, $m > I$, and $m > I+1$, respectively. We fix the inverse scale matrix $\Psi_j(f)$ as

$$\Psi_j(f)=(m-I)\boldsymbol{\mu}_{\mathbf{R}_j}(f) \tag{20}$$

so that the mean of $\mathbf{R}_j(f)$ is consistent with (7). The deviation allowed from the mean is controlled by the so-called number of degrees of freedom m, which is not necessarily an integer.

3.3 Learning the hyper-parameter

In order to obtain the best fit between this prior and the actual prior distribution of spatial covariance matrices, we learn the number of degrees of freedom m from training data. We assume that m depends on the distance and the orientation of the microphones with respect to each other (aka, the array geometry) and on the distance from the source to the center of the array, but not on the source direction of arrival. Given the microphone array geometry and the source distance, we generate training signals $\mathbf{c}_p(t)$ indexed by p for a number of microphone array positions and orientations and for a number of source directions of arrival by convolving the corresponding room impulse responses with a single-channel signal. We derive the spatial covariance matrix $\mathbf{R}_p(f)$ associated with each training signal in an *oracle* fashion [30] by alternately applying (16) and (12) to the empirical covariance matrices $\widehat{\mathbf{R}}_{\mathbf{c}_p}(n,f)$ computed as in (5). Such training data can be generated in any practical scenario where the source separation system is to be deployed in fixed known environments, where the impulse responses can be pre-recorded or simulated via the *image method* [33].

Since $\mathbf{R}_p(f)$ is measured only up to an arbitrary nonnegative scaling factor $\alpha_p(f)$, we jointly estimate the number of degrees of freedom m and the scaling factors in the ML sense by maximizing

$$\begin{aligned}\mathcal{L}_{\mathcal{IW}}&=\prod_{p,f}p(\mathbf{R}_p(f)|\alpha_p(f),\Psi_p(f),m)\\&=\prod_{p,f}J_{\alpha_p(f)}\mathcal{IW}(\alpha_p(f)\mathbf{R}_p(f)|\Psi_p(f),m)\end{aligned} \tag{21}$$

where $J_{\alpha_p(f)}=\alpha_p^{I^2}(f)$ is the Jacobian of the scaling transform and $\Psi_p(f)$ is the inverse scale matrix in (20) which depends on p. Maximization with respect to m can be achieved using a nonlinear optimization technique [34], where the optimal scaling factors for a given m are given by

$$\alpha_p(f)=\frac{\mathrm{tr}(\Psi_p(f)\mathbf{R}_p^{-1}(f))}{Im}. \tag{22}$$

The values of m learned for the geometrical setting and the reverberation times tested in Section 5 are shown in Table 1.

3.4 MAP EM update

Given the hyper-parameters $\Psi_j(f)$ and m, the spatial covariance matrices $\mathbf{R}_j(f)$ can be estimated in the MAP sense in step (17) of Algorithm 1 by maximizing the expectation of the log-posterior of the hidden data

$$\begin{aligned}Q_{\mathcal{IW}}=\gamma\sum_{j,f}\log\mathcal{IW}(\mathbf{R}_j(f)|\Psi_j(f),m)+\sum_{j,n,f}\\-\mathrm{tr}(\Sigma_{\mathbf{c}_j}^{-1}(n,f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f))-\log|\pi\,\Sigma_{\mathbf{c}_j}(n,f)|\end{aligned} \tag{23}$$

where γ is a trade-off hyper-parameter determining the strength of the prior. Strictly speaking, MAP estimation corresponds to $\gamma=1$. However, as in other fields of signal processing [35], a larger strength parameter is needed in practice in order to balance the absolute values of the prior and the likelihood, and this generalized rule is loosely referred to as MAP. By computing the partial derivatives of $Q_{\mathcal{IW}}$ with respect to each entry of $\mathbf{R}_j(f)$ and equating them to zero, we obtain the MAP update

$$\mathbf{R}_j(f)=\frac{1}{\gamma(m+I)+N}\left(\gamma\Psi_j(f)+\sum_{n=1}^{N}\frac{\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)}{v_j(n,f)}\right). \tag{24}$$

When $\gamma=0$, the contribution of the prior is excluded and (24) becomes equal to the ML update in (12). The setting of γ will be discussed in Section 5.3.

4 Subsource-based EM algorithm

4.1 General EM algorithm

Besides the SIEM algorithm, an alternative subsource-based EM (SSEM) algorithm was proposed for ML estimation in [12] that applies to spatial covariance matrices of any rank R_j. This algorithm relies on the non-unique representation of the source spatial images as $\mathbf{c}_j(n,f)=\mathbf{H}_j(f)\mathbf{s}_j(n,f)$, where the entries $s_{jr}(n,f)$, $r=1,\ldots,R_j$, of $\mathbf{s}_j(n,f)$ are uncorrelated complex-valued subsource coefficients distributed as $s_{jr}(n,f)\sim\mathcal{N}(0,v_j(n,f))$ and $\mathbf{H}_j(f)$

Table 1 Learned values of the prior hyper-parameters

T_{60}	50 ms	130 ms	250 ms	500 ms
m	2.1	2.1	3.4	5.3
σ_1^2	0.009	0.033	0.068	0.148
σ_2^2	0.002	0.024	0.063	0.139

is an $I \times R_j$ complex-valued subsource mixing matrix satisfying the constraint [12]

$$\mathbf{R}_j(f) = \mathbf{H}_j(f)\mathbf{H}_j^{\mathrm{H}}(f). \tag{25}$$

This subsource mixing matrix reduces to a mixing vector in the particular case when $\mathbf{R}_j(f)$ has rank 1. Overall, the mixture STFT coefficients are written as

$$\mathbf{x}(n,f) = \mathbf{H}(f)\mathbf{s}(n,f) + \mathbf{b}(n,f) \tag{26}$$

where $\mathbf{s}(n,f) = [\,s_{11}(n,f),\ldots,s_{1R_1}(n,f),\ldots,s_{JR_j}(n,f)]^{\mathrm{T}}$ is an $R\times 1$ vector of subsource coefficients with $R = \sum_{j=1}^{J} R_j$, $\mathbf{H}(f) = [\,\mathbf{H}_1(f),\ldots,\mathbf{H}_J(f)]$ is an $I \times R$ mixing matrix and $\mathbf{b}(n,f)$ is a small Gaussian noise with covariance matrix $\Sigma_{\mathbf{b}}(n,f) = \sigma_b^2(f)\mathbf{I}_I$ required by the EM algorithm. The log-likelihood (4) can then be maximized by considering the set $\{\mathbf{x}(n,f), s_j(n,f)\}_{j,n}$ of observed mixture STFT coefficients and hidden subsource STFT coefficients in all time-frequency bins as *complete data*. Once again, it turns out that MAP estimation can be achieved via the same algorithm except for the mixing matrix update in the M step.

The details of one iteration are summarized in Algorithm 2, where $\mathcal{R}_j$ denotes the set of subsource indices associated with the jth source and $\tilde{v}_r(n,f) = v_j(n,f)$ if and only if $r \in \mathcal{R}_j$. In the E step, the Wiener filter $\mathbf{W}_j(n,f)$ and the second-order cross-moments $\widehat{\mathbf{R}}_{\mathbf{s}}(n,f)$ and $\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f)$ are computed. In the M step $v_j(n,f)$ and $\mathbf{H}(f)$ are updated. In the ML case, the update for $\mathbf{H}(f)$ in (34) is given by [12]

$$\mathbf{H}(f) = \left(\sum_{n=1}^{N} \widehat{\mathbf{R}}_{\mathbf{xs}}(n,f)\right)\left(\sum_{n=1}^{N} \widehat{\mathbf{R}}_{\mathbf{s}}(n,f)\right)^{-1}. \tag{27}$$

Algorithm 2 SSEM algorithm [12]

E step:

$$\Sigma_{\mathbf{s}}(n,f) = \mathrm{diag}\left([\,\tilde{v}_r(n,f)]_{r=1}^{R}\right) \tag{28}$$

$$\Sigma_{\mathbf{x}}(n,f) = \mathbf{H}(f)\Sigma_{\mathbf{s}}(n,f)\mathbf{H}^{\mathrm{H}}(f) + \Sigma_{\mathbf{b}}(n,f) \tag{29}$$

$$\mathbf{W}(n,f) = \Sigma_{\mathbf{s}}(n,f)\mathbf{H}^{\mathrm{H}}(f)\Sigma_{\mathbf{x}}^{-1}(n,f) \tag{30}$$

$$\widehat{\mathbf{R}}_{\mathbf{s}}(n,f) = \mathbf{W}(n,f)\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)\mathbf{W}^{\mathrm{H}}(n,f) + (\mathbf{I}_R - \mathbf{W}(n,f)\mathbf{H}(f))\,\Sigma_{\mathbf{s}}(n,f) \tag{31}$$

$$\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f) = \widehat{\mathbf{R}}_{\mathbf{x}}(n,f)\mathbf{W}^{\mathrm{H}}(n,f) \tag{32}$$

M step:

$$v_j(n,f) = \frac{1}{R_j}\sum_{r\in\mathcal{R}_j} [\,\widehat{\mathbf{R}}_{\mathbf{s}}(n,f)]_{rr} \tag{33}$$

Update $\mathbf{H}(f)$. (34)

4.2 Gaussian prior

The design of a suitable prior over $\mathbf{H}(f)$ is subject to the same practical engineering constraints as above, which leads us to propose a Gaussian prior. We model each column $\mathbf{h}_{jr}(f)$, $r = 1,\ldots,R_j$, of $\mathbf{H}_j(f)$ as a complex-valued Gaussian random vector

$$\mathbf{h}_{jr}(f) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f), \Sigma_{\mathbf{h}_{jr}}(f)) \tag{35}$$

with mean $\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f)$ and covariance $\Sigma_{\mathbf{h}_{jr}}(f)$. Following the assumption in Section 2.2, echoes and reverberation cancel out on average over all orientations in the room so that they appear only in the covariance, while only the part corresponding to direct sound appears in the mean. Without loss of generality, let us select $\mathbf{H}_j(f)$ such that direct sound is concentrated in the first subsource of each source, i.e., the first subsource includes direct sound, echoes, and reverberation, while the other subsources include echoes and reverberation only[b]. The mean and the covariance of the prior can then be expressed as

$$\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f) = \begin{cases} \mathbf{d}_j(f) & \text{if } r = 1 \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{36}$$

$$\Sigma_{\mathbf{h}_{jr}}(f) = \sigma_r^2\Omega(f) \tag{37}$$

where the echo and reverberation power of all subsources sums up to the total power in (10):

$$\sum_{r=1}^{R_j} \sigma_r^2 = \sigma_{\mathrm{rev}}^2. \tag{38}$$

Contrary to the inverse-Wishart prior whose variance is governed by a single hyper-parameter m, this prior involves $R_j - 1$ free hyper-parameters σ_r^2, $r = 2,\ldots,R_j$, which makes it potentially more flexible as soon as $I \geq R_j > 2$. The priors are distinct, however, in the sense that the Gaussian prior does not generalize the inverse-Wishart prior whatever the choice of the hyper-parameters.

4.3 Learning the hyper-parameters

In order to fit the actual distribution of subsource mixing matrices, we learn these free hyper-parameters from training data. The training data consist of the spatial covariance matrices $\mathbf{R}_p(f)$ computed in Section 3.3 for different positions p, from which we derive the corresponding subsource mixing matrices $\mathbf{H}_p(f)$ by singular value decomposition $\mathbf{R}_p(f) = \mathbf{H}_p(f)\mathbf{H}_p^{\mathrm{H}}(f)$ such that the columns of $\mathbf{H}_p(f)$ are orthogonal and sorted by decreasing norm.

These columns $\mathbf{h}_{pr}(f)$ are observed only up to an arbitrary scale common to all r and an arbitrary phase rotation specific to each r. Phase rotations do not affect the learned variances σ_r^2 for $r > 1$, since the corresponding means

$\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f)$ are zero. Multiplying $\mathbf{H}_p(f)$ by a global complex-valued factor $\alpha_p(f)$ is hence sufficient to address this indeterminacy. Denoting by

$$\underline{\mathbf{h}}_p(f) = \begin{pmatrix} \mathbf{h}_{p1}(f) \\ \vdots \\ \mathbf{h}_{pR_j}(f) \end{pmatrix} \tag{39}$$

the $IR_j \times 1$ vectorization of $\mathbf{H}_p(f)$ with mean

$$\boldsymbol{\mu}_{\underline{\mathbf{h}}_p}(f) = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{h}_{p1}}(f) \\ \vdots \\ \boldsymbol{\mu}_{\mathbf{h}_{pR_j}}(f) \end{pmatrix} \tag{40}$$

and covariance

$$\Sigma_{\underline{\mathbf{h}}_p}(f) = \begin{pmatrix} \Sigma_{\mathbf{h}_{p1}}(f) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \Sigma_{\mathbf{h}_{pR_j}}(f) \end{pmatrix}, \tag{41}$$

the hyper-parameters and the multiplication factors are jointly estimated in the ML sense by maximizing

$$\begin{aligned} \mathcal{L}_G &= \prod_{pf} p(\underline{\mathbf{h}}_p(f)|\alpha_p(f), \boldsymbol{\mu}_{\underline{\mathbf{h}}_p}(f), \Sigma_{\underline{\mathbf{h}}_p}(f)) \\ &= \prod_{pf} J_{\alpha_p(f)} \mathcal{N}(\alpha_p(f)\underline{\mathbf{h}}_p(f)|\boldsymbol{\mu}_{\underline{\mathbf{h}}_p}(f), \Sigma_{\underline{\mathbf{h}}_p}(f)) \end{aligned} \tag{42}$$

where $J_{\alpha_p(f)} = |\alpha_p(f)|^{2I^2}$ is the Jacobian of the multiplication. Maximization is achieved using a nonlinear optimization technique, where the optimal multiplication factors as a function of the hyper-parameters are found as

$$\alpha_p(f) = \frac{-|b| - (|b|^2 - 4ac)^{1/2}}{2a} \frac{b}{|b|} \tag{43}$$

where

$$\begin{aligned} a &= -\underline{\mathbf{h}}_p^{\mathrm{H}}(f)\Sigma_{\underline{\mathbf{h}}_p}^{-1}(f)\underline{\mathbf{h}}_p(f) \\ b &= \underline{\mathbf{h}}_p^{\mathrm{H}}(f)\Sigma_{\underline{\mathbf{h}}_p}^{-1}(f)\boldsymbol{\mu}_{\underline{\mathbf{h}}_p}(f) \\ c &= I^2. \end{aligned} \tag{44}$$

The values of σ_1^2 and σ_2^2 learned in the setting of Section 5 ($R_j = I = 2$) are displayed in Table 1.

4.4 MAP EM update

Similarly to (39), let us denote by $\underline{\mathbf{h}}(f)$ the vectorization of $\mathbf{H}(f)$ as an $IR \times 1$ column vector. The prior distribution (35) translates into

$$\underline{\mathbf{h}}(f) \sim \mathcal{N}(\boldsymbol{\mu}_{\underline{\mathbf{h}}}(f), \Sigma_{\underline{\mathbf{h}}}(f)), \tag{45}$$

where $\boldsymbol{\mu}_{\underline{\mathbf{h}}}(f)$ is the $IR \times 1$ vector obtained by concatenating $\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f)$ for all j, r; and $\Sigma_{\underline{\mathbf{h}}}(f)$ is the $IR \times IR$ block-diagonal matrix whose entries are equal to $\Sigma_{\mathbf{h}_{jr}}(f)$ for all j, r.

The MAP update for $\mathbf{H}(f)$ is derived by maximizing the expectation of the log-posterior of the complete data that is equal up to a constant to (see Equation 18 in [12]) for the expression of the expectation of the log-likelihood)

$$\begin{aligned} Q_{\mathcal{G}} = \gamma \log \mathcal{N}(\underline{\mathbf{h}}(f)|\boldsymbol{\mu}_{\underline{\mathbf{h}}}(f), \Sigma_{\underline{\mathbf{h}}}(f)) \\ + \sum_{n,f} -\frac{1}{\sigma_b^2(f)} \mathrm{tr}[\,\widehat{\mathbf{R}}_{\mathbf{x}}(n,f) - \mathbf{H}(f)\widehat{\mathbf{R}}_{\mathbf{xs}}^{\mathrm{H}}(n,f) \\ - \widehat{\mathbf{R}}_{\mathbf{xs}}(n,f)\mathbf{H}^{\mathrm{H}}(f) + \mathbf{H}(f)\widehat{\mathbf{R}}_{\mathbf{s}}(n,f)\mathbf{H}^{\mathrm{H}}(f)] \end{aligned} \tag{46}$$

where γ is a trade-off hyper-parameter determining the strength of the prior. By rewriting the matrix quadratic form in the log-likelihood term of (46) as a vector quadratic form in terms of $\underline{\mathbf{h}}(f)$ and by computing the gradient of $Q_{\mathcal{G}}$ and equating it to zero, we obtain

$$\begin{aligned} \underline{\mathbf{h}}(f) = \left(\gamma \Sigma_{\underline{\mathbf{h}}}^{-1}(f) + \frac{1}{\sigma_b^2(f)} \sum_{n=1}^{N} (\widehat{\mathbf{R}}_{\mathbf{s}}(n,f) \otimes \mathbf{I}_I)^{\mathrm{T}}\right)^{-1} \\ \times \left(\gamma \Sigma_{\underline{\mathbf{h}}}^{-1}(f)\boldsymbol{\mu}_{\underline{\mathbf{h}}}(f) + \frac{1}{\sigma_b^2(f)} \sum_{n=1}^{N} \mathrm{vec}(\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f))\right) \end{aligned} \tag{47}$$

where $.^{\mathrm{T}}$ denotes transposition, $\otimes$ is the Kronecker product and vec(.) concatenates the columns of a matrix into a single column vector. The mixing matrix $\mathbf{H}(f)$ is then obtained by devectorizing $\underline{\mathbf{h}}(f)$. This update boils down to the ML update (27) when $\gamma = 0$.

5 Experimental evaluation

We evaluate the performance of the proposed MAP estimation algorithms compared to the conventional ML estimation algorithms and to two baseline approaches for the separation of two-channel convolutive mixtures of three sources. We target a semi-informed scenario where the relative positions of the sources and the microphones are known, but nothing is known about their absolute position in the room nor about the source signals. The reverberant character of the data calls for the use of full-rank spatial covariance matrices and subsource mixing matrices, i.e., $R_j = 2$ for all j. We do not constrain the source variances $v_j(n,f)$, so as to measure the improvement due to the priors alone. The full Matlab code for our experiments can be downloaded from [36].

5.1 Data

The proposed priors can be applied in any scenario where the source separation system is to be deployed in fixed, known environments, where the impulse responses can be pre-recorded or simulated. In the following, we use simulated mixtures so as to test a wide range of room reverberation times. The use of simulated data is widespread in audio source separation and it has been shown to yield comparable separation performance to real-world data in

general [37]. As a matter of fact, the results of the ML algorithms reported below are comparable to those previously reported on real-world recordings in Figure six in [11].

The positions of the sources and the microphones in the test data are illustrated in Figure 1. The room dimensions are 4.45 × 3.55 × 2.5 m as in [37], and the microphone spacing and the source-to-microphone distances are fixed to $d = 5$ and $r = 50$ cm, respectively. We generated room impulse responses via the image method [33] using the Roomsimove toolbox[c] for four reverberation times: $T_{60} =$ 50, 130, 250, or 500 ms, which we convolved with 10 s speech signals sampled at 16 kHz. For each T_{60}, 6 mixture signals were generated using speech signals from the Signal Separation and Evaluation Campaign (SiSEC) [37]: 2 mixtures of English and Japanese male speech, 2 mixtures of English and Japanese female speech, and 2 mixtures of male and female speech, resulting in 24 mixture signals in total.

Training data were generated in a similar fashion by simulating room impulse responses for 20 random source directions of arrival for each of 20 random microphone pair positions and orientations for the same d and r as above. This resulted in a total of 400 source image signals indexed by p for each T_{60}.

5.2 Learned hyper-parameter values

Regarding training, preliminary experiments showed that the functions (21) and (42) are concave in practice. Hence, we maximized them using Matlab's fmincon optimizer (Mathworks Inc., Natick, MA, USA). The resulting hyper-parameter values are shown in Table 1.

As expected, the total power of echoes and reverberation $\sigma^2_{\text{rev}} = \sigma^2_1 + \sigma^2_2$ strongly increases with T_{60}, such that the direct-to-reverberant ratio is 14 dB lower when $T_{60} = 500$ ms than when $T_{60} = 50$ ms. The variance of the inverse-Wishart prior, which is inversely related to m [32], decreases with T_{60}. The ratio $\sigma^2_1/\sigma^2_{\text{rev}}$ decreases with T_{60}, which indicates that the echoic and reverberant part of the impulse responses becomes more and more diffuse.

5.3 Tested algorithms and evaluation criteria

In addition to the proposed MAP versions of SIEM and SSEM (*MAP inverse-Wishart* and *MAP Gaussian*), we consider the conventional ML versions of these algorithms where the initial values of $\mathbf{R}_j(f)$ and $\mathbf{H}(f)$ are either set to $\boldsymbol{\mu}_{\mathbf{R}_j}(f)$ and $\boldsymbol{\mu}_{\mathbf{H}}(f)$ given the scene geometry (*ML geom. init*) or blindly estimated via hierarchical clustering followed by permutation alignment as detailed in [11] (*ML blind init*). Subsequent permutation alignment of the sources after convergence of the ML algorithms was found not to improve performance and therefore it is not used in the following. For comparison, we evaluate two baseline approaches, namely, *binary masking* and *ℓ_0-norm minimization*, using the reference software in [38] where the mixing matrix in each frequency bin is estimated by hierarchical clustering followed by permutation alignment [11].

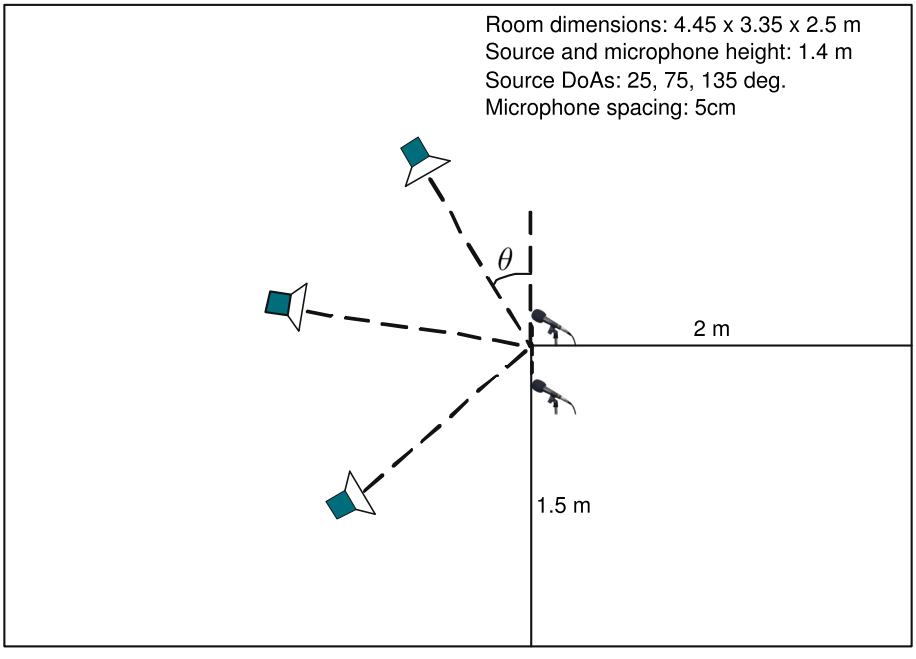

Figure 1 Room geometric setting for testing data.

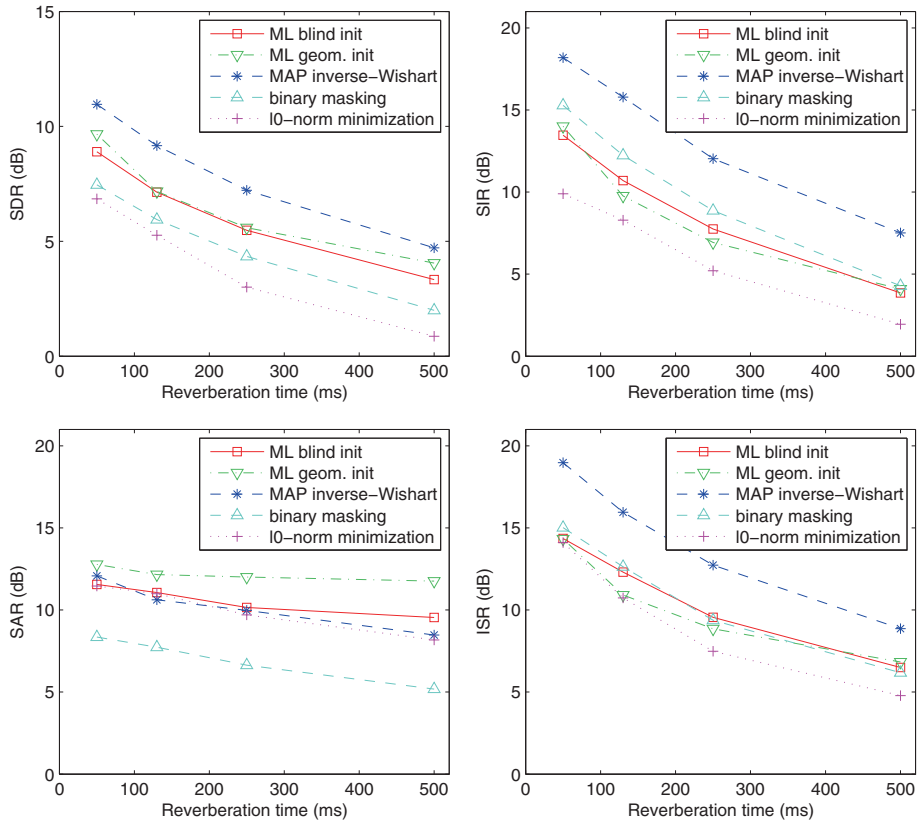

Figure 2 Separation performance of the SIEM algorithms as a function of reverberation time.

In order to assess the respective impact of the priors on solving the permutation problem and on reducing overfitting, we also report an upper bound on the performance of the MAP and the *ML geom. init* algorithms with *oracle* permutation alignment. In each frequency bin, the best possible permutation is found by considering all possible permutations of the estimated sources and by selecting the one that leads to the smallest mean square error compared to the true source signals in that bin.

We computed the STFT with half-overlapping sine windows of length 1,024 and the empirical mixture covariance using a window w_{nf} of size 3×3 as in [26]. The trade-off parameter γ does not significantly affect the results but we observed that $\gamma = 100$ and $\gamma = 10$ are good choices for SIEM and SSEM respectively on average. The number of iterations was fixed to 10 for SIEM and 30 for SSEM, since the convergence of SSEM is typically slower.

The priors did not significantly increase running time. Indeed, the MAP inverse-Wishart update has the same computational complexity as the ML SIEM update. The MAP Gaussian update has greater complexity than the ML SSEM update, but it occurs only once per iteration in each frequency bin, in contrast with the updates in the E step which occur in each time frame. For a typical number of time frames N, the computational complexity is therefore dominated by the E step, regardless of the priors.

We evaluated the separation quality via the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR), and source image-to-spatial distortion ratio (ISR) criteria in decibels (dB) [37], which account respectively for overall distortion, residual crosstalk, musical noise, and target distortion. These criteria were computed using version 3.0 of the BSS Eval toolbox[d] and averaged over all sources and all mixtures for each T_{60}.

Table 2 SDR (dB) of the SIEM algorithms with estimated vs. oracle permutation

	T_{60}	50 ms	130 ms	250 ms	500 ms
Estimated	*ML geom. init*	9.7	7.2	5.6	4.1
permutation	*MAP inverse-Wishart*	11.0	9.2	7.2	4.7
Oracle	*ML geom. init*	9.7	7.2	5.8	4.4
permutation	*MAP inverse-Wishart*	11.0	9.2	7.6	5.1

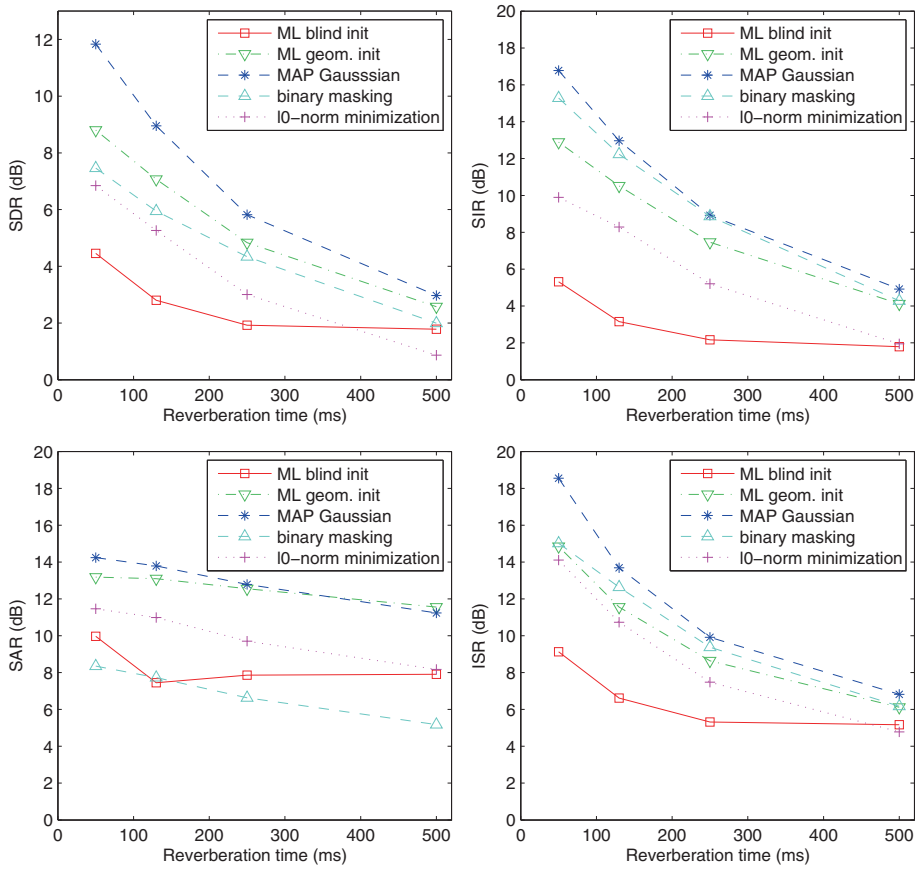

Figure 3 Separation performance of the SSEM algorithms as a function of reverberation time.

5.4 Results for source image-based EM algorithms

The results of the SIEM algorithms and the baselines are compared in Figure 2. *Binary masking* and *ℓ_0-norm minimization* provide lower SDR than all other algorithms for all reverberation conditions. *ML geom. init* results in better performance than *ML blind init* in terms of SDR and SAR for all T_{60}. Overall, *MAP inverse-Wishart* outperforms all other algorithms for all considered T_{60} in terms of SDR, SIR, and ISR. For instance, at $T_{60} = 250$ ms, it improves the SDR by 1.7, 1.6, 2.8, and 4.2 dB compared to *ML blind init*, *ML geom init*, *binary masking*, and *ℓ_0-norm minimization*, respectively. This confirms the benefit of the proposed inverse-Wishart spatial location prior and the associated MAP algorithm.

These results are shown against the corresponding upper bounds obtained with oracle permutation alignment in Table 2. By comparing the first two lines with the last two lines, it appears that *ML geom. init* and *MAP inverse-Wishart* both solve the permutation problem at low reverberation times up to $T_{60} = 130$ ms and that little SDR improvement from 0.2 to 0.4 dB is to be expected from better permutation at higher reverberation times. By contrast, comparison of the third and the fourth lines of the table indicates that even if the permutation problem were solved, *MAP inverse-Wishart* would still outperform *ML geom. init* by 1.8 dB at $T_{60} = 250$ ms, which can be attributed to better robustness to overfitting.

5.5 Results for subsource-based EM algorithms

The results of the SSEM algorithms are depicted in Figure 3. Again, *ML geom init* results in significantly better performance than *ML blind init* in terms of all criteria for all T_{60}, and it also offers higher SDR than *binary masking* and *ℓ_0-norm minimization* for all T_{60}. But the best performance is achieved by *MAP Gaussian* in terms of

Table 3 SDR (dB) of the SSEM algorithms with estimated vs. oracle permutation

T_{60}		50 ms	130 ms	250 ms	500 ms
Estimated	*ML geom. init*	8.8	7.1	4.8	2.6
permutation	*MAP Gaussian*	11.8	9.0	5.8	3.0
Oracle	*ML geom. init*	9.1	7.3	5.3	3.0
permutation	*MAP Gaussian*	11.9	9.0	6.0	3.3

all criteria and for all T_{60}, except in terms of SAR at $T_{60} = 500$ ms. For instance, at $T_{60} = 250$ ms, *MAP Gaussian* improves the SDR by 3.9, 1.0, 1.4, and 2.8 dB compared to *ML blind init*, *ML geom. init*, *binary masking*, and ℓ_0*-norm minimization*, respectively. This confirms the benefit of the proposed Gaussian spatial location prior and the associated MAP algorithm.

These results are shown against the corresponding upper bounds obtained with oracle permutation alignment in Table 3. Again, *MAP Gaussian* significantly outperforms *ML geom. init* in the oracle case, meaning that the overfitting issue in ML estimates is better addressed in MAP estimates with a proper prior. On the other hand, it can be seen that *MAP Gaussian* does not fully solve the permutation problem at medium and high reverberation conditions, but that the gap with the oracle permutation is small and slightly smaller than for *ML geom. init*.

6 Conclusions

We considered two classes of source separation algorithms grounded on the emerging Gaussian EM framework. In contrast with classical ML estimation of the spatial parameters, we proposed two priors exploiting a result from the theory of statistical room acoustics and we derived closed-form MAP updates. The SIEM algorithm with an inverse-Wishart prior and the SSEM algorithm with a Gaussian prior were shown to outperform their ML counterparts for all room reverberation times in a semi-informed scenario. We showed that this performance improvement can be mostly attributed to the greater robustness to overfitting of MAP compared to ML. The proposed MAP algorithms also provide a solution to the problem of permutation of the source estimates that is consistent with the statistics of sound fields. The resulting permutations and those obtained by ML estimation initialized with the known geometric setting are, however, comparably good.

The results in this paper can readily be used in certain real-world scenarios where the source positions are known from, e.g., physical constraints or visual input, and the reverberation characteristics can be learned from the environment [21-23]. Perhaps more importantly, they constitute a first step towards full Bayesian treatment of this family of models in other blind or semi-blind scenarios in the future. In addition to blind estimation of the source positions and possibly of the microphone distance and directivity [39], robustness to erroneous estimation of these hyper-parameters, and blind estimation of the hyper-parameters σ^2_{rev}, m and σ^2_r both pose significant challenges, which go beyond the scope of this paper. Future work will concentrate on these challenges by extending blind techniques for room reverberation time estimation [40]. Usage of the proposed Gaussian prior, which is also valid for rank-1 mixing vectors, may also be explored in the context of FDICA, with the difficulty of translating this prior into a prior over the blocking vectors which are usually considered as parameters in this context instead.

Endnotes

[a]Note that in order to yield nonzero likelihood, $v_j(n,f)$ must be nonzero for at least one source j. $\mathbf{\Sigma_x}(n,f)$ in (14) is therefore the sum of Hermitian positive semi-definite matrices, at least one of which is definite, so it is Hermitian positive definite and invertible.

[b]If several $\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f)$ are nonzero multiples of $\mathbf{d}_j(f)$, a unitary transform can be applied to $\mathbf{H}_j(f)$ in (25) such that only the first one remains nonzero.

[c]http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip
This toolbox provides a command-line interface which, in contrast with the original GUI by D. R. Campbell, allows generation of a large amount of data.

[d]http://bass-db.gforge.inria.fr/bss_eval/

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by the EUREKA Eurostars i3Dmusic project funded by Oseo. Most of it was done while the first two authors were with Inria Rennes.

Author details
[1]Technicolor Rennes Research & Innovation Center, 35510 Cesson-Sévigné, France. [2]Inria, 54600 Villers-lès-Nancy, France. [3]Inria, 35042 Rennes Cedex, France.

Received: 3 April 2013 Accepted: 30 August 2013
Published: 23 September 2013

References

1. P O'Grady, B Pearlmutter, ST Rickard, Survey of sparse and non-sparse methods in source separation. Int. J. Imaging Syst. Technol. **15**, 18–33 (2005)
2. S Makino, TW Lee, H Sawada, *Blind Speech Separation*. (Springer, Berlin, 2007)
3. E Vincent, MG Jafari, SA Abdallah, MD Plumbley, ME Davies, in *Machine Audition: Principles, Algorithms and Systems*. Probabilistic modeling paradigms for audio source separation (IGI Global, Hershey, 2010), pp. 162–185
4. P Smaragdis, Blind separation of convolved mixtures in the frequency domain. Neurocomputing. **22**, 21–34 (1998)
5. H Sawada, S Araki, S Makino, in *Blind Speech Separation*. Frequency-domain blind source separation (Springer, Berlin, 2007), pp. 47–78
6. O Yilmaz, ST Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)
7. H Sawada, S Araki, S Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. Audio Speech Lang. Process. **19**(3), 516–527 (2011)
8. S Winter, W Kellermann, H Sawada, S Makino, MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization. EURASIP J. Adv. Signal Process. **2007**, 024717. doi:10.1155/2007/24717
9. C Févotte, JF Cardoso, in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models (Mohonk, NY, 16–19 October 2005), pp. 78–81

10. A Ozerov, C Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. Audio Speech Lang. Process. **18**(3), 550–563 (2010)
11. NQK Duong, E Vincent, R Gribonval, Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Trans. Audio Speech Lang. Process. **18**(7), 1830–1840 (2010)
12. A Ozerov, E Vincent, F Bimbot, A general flexible framework for the handling of prior information in audio source separation. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1118–1133 (2012)
13. L Benaroya, F Bimbot, R Gribonval, Audio source separation with a single sensor. IEEE Trans. Audio Speech Lang. Process. **14**, 191–199 (2006)
14. C Févotte, N Bertin, JL Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
15. T Virtanen, AT Cemgil, SJ Godsill, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Bayesian extensions to non-negative matrix factorisation for audio signal modelling (Las Vegas, 30 March to 4 April 2008), pp. 1825–1828
16. O Dikmen, AT Cemgil, Gamma Markov random fields for audio source modeling. IEEE Trans. Audio Speech Lang. Process. **18**(3), 589–601 (2010)
17. K Itoyama, M Goto, K Komatani, T Ogata, HG Okuno, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Simultaneous processing of sound source separation and musical instrument identification using Bayesian spectral modeling (Prague, 22–27 May 2011), pp. 3816–3819
18. H Sawada, R Mukai, S Araki, S Makino, A robust, A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech Audio Process. **12**(5), 530–538 (2004)
19. KH Knuth, in *Proceedings of the International Workshop on Independent Component Analysis and Source Separation (ICA)*. A Bayesian approach to source separation (Aussois, January 1999), pp. 283–288
20. AT Cemgil, C Févotte, SJ Godsill, Variational and stochastic inference for Bayesian source separation. Digit. Signal Process. **17**, 891–913 (2007)
21. L Parra, C Alvino, Geometric source separation: merging convolutive source separation with geometric beamforming. IEEE Trans. Audio Speech Lang. Process. **10**(6), 352–362 (2002)
22. M Knaak, S Araki, S Makino, Geometrically constrained independent component analysis. IEEE Trans. Audio Speech Lang. Process. **15**(2), 715–726 (2007)
23. K Reindl, Y Zheng, A Schwarz, S Meier, R Maas, A Sehr, W Kellermann, A stereophonic acoustic signal extraction scheme for noisy and reverberant environments. Comput. Speech Lang. **27**(3), 726–745 (2013)
24. T Otsuka, K Ishiguro, H Sawada, HG Okuno, in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. Bayesian unification of sound source localization and separation with permutation resolution (Toronto, 22–26 July 2012), pp. 2038–2045
25. NQK Duong, E Vincent, R Gribonval, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. An acoustically-motivated spatial prior for under-determined reverberant source separation (Prague, 22–27 May 2011), pp. 9–12
26. NQK Duong, E Vincent, R Gribonval, in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation (St. Malo, 27–30 September 2010), pp. 73–80
27. JF Cardoso, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Multidimensional independent component analysis (Seattle, May 1998), pp. 1941–1944
28. H Kuttruff, *Room Acoustics*, 4th edn. (Spon Press, New York, 2000)
29. T Gustafsson, BD Rao, M Trivedi, Source localization in reverberant environments: modeling and statistical analysis. IEEE Trans. Speech Audio Process. **11**, 791–803 (2003)
30. NQK Duong, E Vincent, R Gribonval, in *Proceedings on the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Spatial covariance models for under-determined reverberant audio source separation (Mohonk, 18–21 October 2009), pp. 129–132
31. G McLachlan, T Krishnan, *The EM Algorithm and Extensions*. (Wiley, New York, 1997)
32. D Maiwald, D Kraus, Calculation of moments of complex Wishart and complex inverse-Wishart distributed matrices. IEEE Proc. Radar Sonar Navigation. **147**, 162–168 (2000)
33. JB Allen, DA Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
34. J Nocedal, SJ Wright, *Numerical Optimization*. (Springer, New York, NY, 1999)
35. A Ogawa, K Takeda, F Itakura, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. Balancing acoustic and linguistic probabilities (Seattle, 1998), pp. I–181–184
36. NQK Duong, E Vincent, R Gribonval, Matlab code for Gaussian model based audio source separation using spatial location priors. http://www.loria.fr/~evincent/spatial_priors.zip
37. E Vincent, S Araki, F Theis, G Nolte, P Bofill, H Sawada, A Ozerov, V Gowreesunker, D Lutter, NQK Duong, The Signal Separation Campaign (2007-2010): achievements and remaining challenges. Signal Process. **92**, 1928–1936 (2012)
38. E Vincent, S Araki, P Bofill, in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*. Signal Separation Evaluation Campaign: a community-based approach to large-scale evaluation (Paraty, 15–18 March 2009), pp. 734–741
39. K Hasegawa, N Ono, S Miyabe, S Sagayama, *Blind estimation of locations and time offsets for distributed recording devices*, (St. Malo, 27–30 September 2010), pp. 57–64
40. ND Gaubitch, H Löllmann, M Jeub, T Falk, PA Naylor, P Vary, M Brookes, in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*. Performance comparison of algorithms for blind reverberation time estimation from speech (Aachen, 4–6 September 2012), pp. 1–4

doi:10.1186/1687-6180-2013-149
Cite this article as: Duong *et al.*: **Spatial location priors for Gaussian model based reverberant audio source separation.** *EURASIP Journal on Advances in Signal Processing* 2013 **2013**:149.