Bai *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021: 14
https://doi.org/10.1186/s13636-021-00200-z

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Acoustic DOA estimation using space alternating sparse Bayesian learning

Zonglong Bai[1,2*], Liming Shi[2], Jesper Rindom Jensen[2], Jinwei Sun[1] and Mads Græsbøll Christensen[2]

Abstract

Estimating the direction-of-arrival (DOA) of multiple acoustic sources is one of the key technologies for humanoid robots and drones. However, it is a most challenging problem due to a number of factors, including the platform size which puts a constraint on the array aperture. To overcome this problem, a high-resolution DOA estimation algorithm based on sparse Bayesian learning is proposed in this paper. A group sparse prior based hierarchical Bayesian model is introduced to encourage spatial sparsity of acoustic sources. To obtain approximate posteriors of the hidden variables, a variational Bayesian approach is proposed. Moreover, to reduce the computational complexity, the space alternating approach is applied to push the variational Bayesian inference to the scalar level. Furthermore, an acoustic DOA estimator is proposed to jointly utilize the estimated source signals from all frequency bins. Compared to state-of-the-art approaches, the high-resolution performance of the proposed approach is demonstrated in experiments with both synthetic and real data. The experiments show that the proposed approach achieves lower root mean square error (RMSE), false alert (FA), and miss-detection (MD) than other methods. Therefore, the proposed approach can be applied to some applications such as humanoid robots and drones to improve the resolution performance for acoustic DOA estimation especially when the size of the array aperture is constrained by the platform, preventing the use of traditional methods to resolve multiple sources.

Keywords: Sparse Bayesian learning, Acoustic DOA estimation, Sound source localization

1 Introduction

Acoustic direction-of-arrival (DOA) estimation is a key technology in audio signal processing where it enables source localization for humanoid robots [1, 2], drones [3, 4], teleconferencing [5, 6], and hearing aids [7]. The goal of acoustic DOA estimation is to obtain the bearing angle of acoustic waves generated by sound sources using a microphone array. According to the Rayleigh criterion [8], the resolution of traditional DOA estimation approaches (e.g., the classical beamforming (CBF)[1] based approach and the steered-response power phase transform (SRP-PHAT) method [9]) is limited by the array aperture. Therefore, for some applications like humanoid robots and drones with a small platform size, the traditional approaches suffer in scenarios with multiple sources simultaneously present. Although methods such as the minimum variance distortionless response (MVDR) [8, 10], multiple signal classification (MUSIC) [11], and estimation of signal parameters via the rotational invariance technique (ESPRIT) [12] can offer a high-resolution performance, they are sensitive to calibration errors and errors in the assumed or estimated signal statistics [13, 14]. The robustness of the MVDR and MUSIC methods have been studied in the presence of array errors [15–17]. However, these studies rely on asymptotic properties, i.e., high signal-to-noise ratio (SNR) scenarios and large numbers of snapshots. Thus, these studies do not apply when only a small number of snapshots is available.

[1]In this paper, the CBF is referred to as delay and sum beamforming.

*Correspondence: baizonglong@gmail.com
[1]School of Instrument Science and Engineering, Harbin Institute of Technology, Xidazhi Street 92, 150006 Harbin, China
[2]CREATE, Aalborg University, Rendsburggade 14 9000 Aalborg, Denmark

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Sparse signal recovery-based DOA estimation methods have enjoyed much success in recent decades by exploiting the sparsity of sources in the spatial domain [18, 19]. These approaches are attractive because (1) they offer robustness against noise and limitations in data quality [18], (2) they have a good performance with a small number of snapshots [20], (3) they offer a higher resolution performance than MVDR and MUSIC methods [21, 22], and (4) they have the capability to resolve coherent sources [23]. In [18], the source localization problem was first formulated as an over-complete basis representation problem. To estimate the source amplitudes, an l_1-norm based singular value decomposition (SVD) method was proposed. In [24], a complex least absolute shrinkage and selection operator (cLASSO) method was proposed for DOA estimation. In [25], a re-weighted regularized sparse recovery method was proposed for DOA estimation with unknown mutual coupling. All these methods are based on convex optimization theory, that is, the signals are recovered by solving a regularized optimization problem. They have a good performance with a properly chosen regularization factor, but the regularization factor needs to be determined empirically [26].

Because of its self-regularization nature and its ability to quantify uncertainty, the sparse Bayesian learning (SBL)-based methods have attracted a lot of attention in sparse signal recovery and compressed sensing. The SBL principle was originally proposed in [27] for obtaining sparse solutions to regression and classification tasks. The SBL algorithm was applied to the compressed sensing in [28], and an SBL-based Bayesian compressed sensing method using Laplace priors was proposed in [29]. More recently, a scalable mean-field SBL was proposed in [30]. In [31], an SBL-based DOA estimation method with predefined grids was proposed. In that paper, the DOA estimation is formulated as a sparse signal recovery and compressed sensing problem. To obtain refined estimates of the DOA, an off-grid DOA estimation method was proposed in [32]. In [21], a multi-snapshot SBL (MSBL) method was proposed for the multi-snapshot DOA estimation problem. The method was further applied to sound source localization and speech enhancement in [22]. To reduce the computational complexity of the wide-band approach, a computationally efficient DOA estimation method was proposed in [33] based on a sparse Bayesian framework. Additionally, some of our previous works are related to this paper. In [34], we proposed an SBL method with compressed data for sound source localization. The results show that the SBL method offers an excellent estimation accuracy for sound source localization even with low data quality. In [35], we proposed an SBL-based acoustic reflector localization method, which models the acoustic reflector localization problem as a sparse signal recovery problem. It shows that the SBL-based method offers a more robust performance for basis mismatch compared to the state-of-the-art methods. However, a common drawback of these approaches is that the traditional SBL-based approaches are computationally complex due to the matrix inversion operation required for updating the covariance matrix of the source signals.

Computationally efficient SBL algorithms have also been proposed in various applications. For example, in [36], a basis adding/deleting scheme based on the marginal distribution was proposed. In [37], an inverse free SBL method was proposed by relaxing the evidence lower bound. In [38], a space alternating variational estimation (SAVE) algorithm was proposed to push the variational Bayesian inference (VBI) based SBL to a scalar level. The experimental results show that the SAVE approach has a faster convergence and a lower minimum mean square error (MMSE) performance than other fast SBL algorithms.

Based on this, we propose a space alternating SBL-based acoustic DOA estimation method for high-resolution estimation in this paper. A hierarchical Bayesian framework with group sparse priors is built to model multiple measurement vector (multi-snapshot) signals. As direct calculation of the posterior distribution is not possible, variational Bayesian inference is applied to infer all hidden variables in the proposed model. Furthermore, we extend the SAVE method [38] to the multiple measurement vector (MMV) case to reduce the computational complexity of the algorithm. The proposed algorithm can be applied to each frequency bin independently. To jointly utilize the recovered signals from all frequency bins, a complex Gaussian mixture model (CGMM) based expectation–maximization (EM) algorithm is proposed. We refer to the proposed method as the SAVE-MSBL-EM method.

The rest of this paper is organized as follows: In Section 2, we pose the narrow-band acoustic DOA estimation problem as a sparse signal recovery problem with an over-complete dictionary. Moreover, under the assumption that the DOAs of all sources do not change in a frame, a hierarchical Bayesian framework is built by exploiting the group sparsity of the MMV source signals. In Section 3, the SAVE-MSBL algorithm is proposed to infer all the hidden variables in the hierarchical Bayesian model for one frequency bin. Then, the CGMM-based EM algorithm is formulated to deal with the wide-band acoustic DOA estimation. In Section 4, we evaluate the performance of the proposed algorithm using both synthetic data and real data. Finally, we provide our conclusions in Section 5.

Note that vectors and matrices are represented using bold lowercase and uppercase letters, respectively. The superscripts $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ denote the transpose and conjugate transpose operator, respectively. Moreover, $L \times L$ identity matrix is denoted as $\mathbf{I}_L$. The l_p norm and

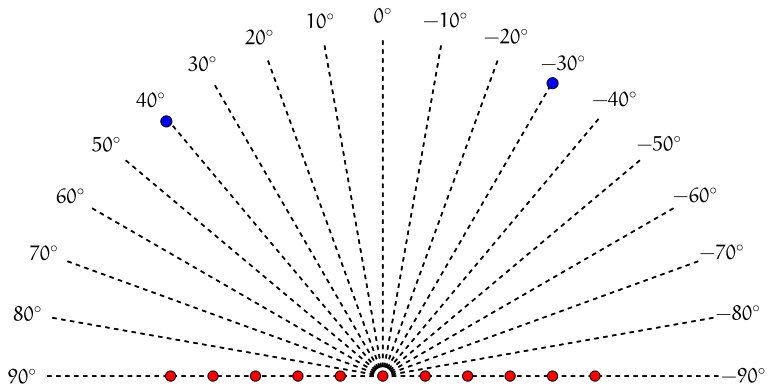

Fig. 1 The candidate DOAs in the target space. The red circles denote the microphones. The blue circles denote the positions of the sound sources

Frobenius norm are represented using $\|\cdot\|_p$ and $\|\cdot\|_F$, respectively.

2 Signal model

2.1 Problem formulation

The problem considered in this paper can be stated as follows. We consider the scenario that P sound sources exist in the far-field of an arbitrary microphone array with M microphones which are used to record the signals. The center point of the microphone array is denoted as O. All the microphones are assumed to be omnidirectional and synchronized. As it is shown in [18, 22, 33], the DOA estimation problem can be formulated as a sparse signal recovery problem using an over-complete dictionary with basis vectors containing the DOA information. Let $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_K]^{\mathrm{T}}$ denote a set of candidate DOAs, where K denotes the total number of candidate DOAs. The signal model for the fth ($1 \le f \le F$) frequency bin of one frame can be expressed as

$$\boldsymbol{X}_f = \boldsymbol{A}_f \boldsymbol{S}_f + \boldsymbol{N}_f, \tag{1}$$

where

$$\begin{aligned}
\boldsymbol{X}_f &= \left[\boldsymbol{x}_{f,1}, \boldsymbol{x}_{f,2}, \cdots, \boldsymbol{x}_{f,L}\right], \\
\boldsymbol{x}_{f,l} &= \left[x_{f,l,1}, x_{f,l,2}, \cdots, x_{f,l,M}\right]^{\mathrm{T}}, \\
\boldsymbol{A}_f &= \left[\boldsymbol{a}_{f,1}, \boldsymbol{a}_{f,2}, \cdots, \boldsymbol{a}_{f,K}\right], \\
\boldsymbol{a}_{f,k} &= \left[1, e^{-j\omega_f \tau_{k2}}, \cdots, e^{-j\omega_f \tau_{kM}}\right]^{\mathrm{T}}, \\
\boldsymbol{S}_f &= \left[\boldsymbol{s}_{f,1}, \boldsymbol{s}_{f,2}, \cdots, \boldsymbol{s}_{f,K}\right]^{\mathrm{T}}, \\
\boldsymbol{s}_{f,k} &= \left[s_{f,k,1}, s_{f,k,2}, \cdots, s_{f,k,L}\right]^{\mathrm{T}},
\end{aligned}$$

F is the total number of frequency bins, $\boldsymbol{X}_f \in \mathbb{C}^{M\times L}$ is a collection of signal snapshots in the frequency-domain with $x_{f,l,m}$ being the signal at the fth frequency bin, lth snapshot, and mth microphone. We refer to the matrix $\boldsymbol{X}_f$ as one frame and $\boldsymbol{x}_{f,l} \in \mathbb{C}^M$ as one snapshot, $l \in [1, 2, \cdots, L]$ is the index of the snapshots[2]. The matrix $\boldsymbol{A}_f \in \mathbb{C}^{M\times N}$ is the dictionary for the fth frequency bin with the basis vector $\boldsymbol{a}_{f,k} \in \mathbb{C}^M$ representing the array response for the direction θ_k, ω_f is the fth angular frequency, and τ_{km} is the relative time delay of source k between microphone m and the array center point O. Moreover, $\boldsymbol{S}_f \in \mathbb{C}^{K\times L}$ is a collection of the source signals with $\boldsymbol{s}_{f,k}$ being the kth row. The noise matrix $\boldsymbol{N}_f \in \mathbb{C}^{M\times L}$ is defined similarly to $\boldsymbol{S}_f$. Assuming that several sound sources are active in one frame, let $\boldsymbol{\theta}_s$ ($\boldsymbol{\theta}_s \subset \boldsymbol{\theta}$) denote the true DOA set and $\boldsymbol{k}_s$ ($\boldsymbol{k}_s \subset [1, 2, \cdots, K]$) denote the true index set. Based on the above definition and the signal model in (1), $\boldsymbol{S}_f$ is an all-zero matrix except for the elements of the rows within the ground truth index set $\boldsymbol{k}_s$. An example is given in Fig. 1, which uses a uniform linear array (ULA). In this example, the target space is sampled uniformly with an interval of 10°. Two sources are located at −30° and 40°, respectively. Thus, when the two sources are active simultaneously, only the elements in the two rows of $\boldsymbol{S}_f$ corresponding to the bearing angles −30° and 40° are non-zero.

Based on (1), to obtain the DOA estimator, we can first recover the source signal, $\boldsymbol{S}_f$, given the MMV, $\boldsymbol{X}_f$, and the predefined dictionary, $\boldsymbol{A}_f$, using MMV sparse signal recovery methods, and then find the row index set of the non-zero elements, which indicates the acoustic DOAs. We assume that the sound sources are static or move slowly such that the direction of the sound sources do not change within the snapshots in a frame. We further assume that the number of active sound sources P is very small compared to the number of candidate DOAs K, i.e., $P \ll K$. As a result, the sound source signal, $\boldsymbol{S}_f$, is a signal matrix with group sparsity and the algorithms for sparse signal recovery can be applied [18, 19]. In this paper, we propose a space alternating MSBL method to improve the

[2]Here, a snapshot refers to the array data in one observation window.

estimation performance by exploiting the group sparsity of $\boldsymbol{S}_f$.

2.2 Probabilistic models

The SBL method is a widely used sparse signal reconstruction method. It is a probabilistic parameter estimation approach based on a hierarchical Bayesian framework. It learns the sparse signal from the over-complete observation model, resulting in a robust maximum likelihood estimation method [27, 39]. Like other Bayesian algorithms, SBL estimates model parameters by maximizing the posterior with a sparse prior. However, instead of adding a specialized model prior, SBL encourages sparsity by using a hierarchical framework that controls the scaling of Gaussian priors through updating individual parameters of each model [27, 40].

2.2.1 Sparse signal model

Following the SBL method proposed in [27], a hierarchical Bayesian framework is used to model the signal matrix, $\boldsymbol{S}_f$. For the sake of brevity, we omit the dependency of random variables on the subscript, f, where appropriate. First, we assume that the candidate sources are independent to each other. Then, a multivariate complex Gaussian distribution is used to describe the kth candidate source signal $\boldsymbol{s}_k$ with zero mean and a covariance matrix $\lambda_k^{-1}\mathbf{I}_L$, i.e.,

$$p(\boldsymbol{S}|\boldsymbol{\lambda}) = \prod_{k=1}^{K} \mathcal{CN}\left(\boldsymbol{s}_k|\mathbf{0}, \lambda_k^{-1}\mathbf{I}_L\right), \tag{2}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_K]^T$ is the hyper-parameter vector, λ_k is the hyper-parameter related to the amplitude of the kth candidate source signal $\boldsymbol{s}_k$, e.g., the amplitude of $\boldsymbol{s}_k$ is 0 when $\lambda_k \to \infty$. Moreover, $\mathbf{I}_L$ is the $L \times L$ identity matrix, $\mathcal{CN}(\cdot)$ denotes the complex Gaussian distribution and λ_k is the precision of $\boldsymbol{s}_k$. Note that, for each candidate DOA (e.g., the kth DOA), an individual precision λ_k is used, but the precision λ_k is set to the same for the signal in different snapshots, thereby encouraging group sparsity [41].

The motivation is that the DOAs of the sound sources, as well as the set of active sources, are assumed to not change within a frame. For different candidate DOAs, different precisions are used to encourage the sparsity (see [18, 19] for further details).

In the second layer of the hierarchy, we assume that the precision variables are independent and follow gamma distributions, i.e.,

$$p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} \mathcal{G}\left(\lambda_k|1, \gamma_k\right), \tag{3}$$

where $\mathcal{G}(a, b)$ denotes the gamma distribution with the shape parameter a and the rate parameter b. There are two reasons for this particular choice of prior distribution: (1) the gamma distribution is a conjugate prior for the variable λ_k in the complex Gaussian distribution, leading to a tractable posterior, and (2) the marginal distribution $\int p(\boldsymbol{S}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\gamma})d\boldsymbol{\lambda}$ is a Student's t distribution encouraging sparsity [27].

To facilitate the inference of $\boldsymbol{\gamma}$, we further assume that the variables in $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_k, \cdots, \gamma_K]^{\mathrm{T}}$ follow i.i.d. gamma distributions, i.e.,

$$p(\boldsymbol{\gamma}) = \prod_{k=1}^{K} \mathcal{G}(\gamma_k|a, b), \tag{4}$$

where a and b are model parameters that will be treated as hyper-parameters.

2.2.2 Likelihood function and noise model

Under the assumption of circular symmetric complex Gaussian noises, the likelihood function can be written as

$$p(\boldsymbol{X}|\boldsymbol{S}, \rho) = \prod_{l=1}^{L} \mathcal{CN}\left(\boldsymbol{x}_l|\boldsymbol{A}\boldsymbol{s}_l, \rho^{-1}\mathbf{I}_M\right), \tag{5}$$

where ρ denotes the noise precision.

For tractability, we assume that ρ follows a gamma distribution as follows

$$p(\rho) = \mathcal{G}(\rho|c, d), \tag{6}$$

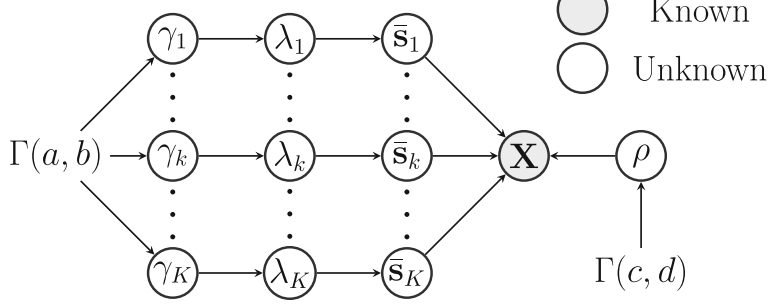

Fig. 2 Probabilistic graphical model

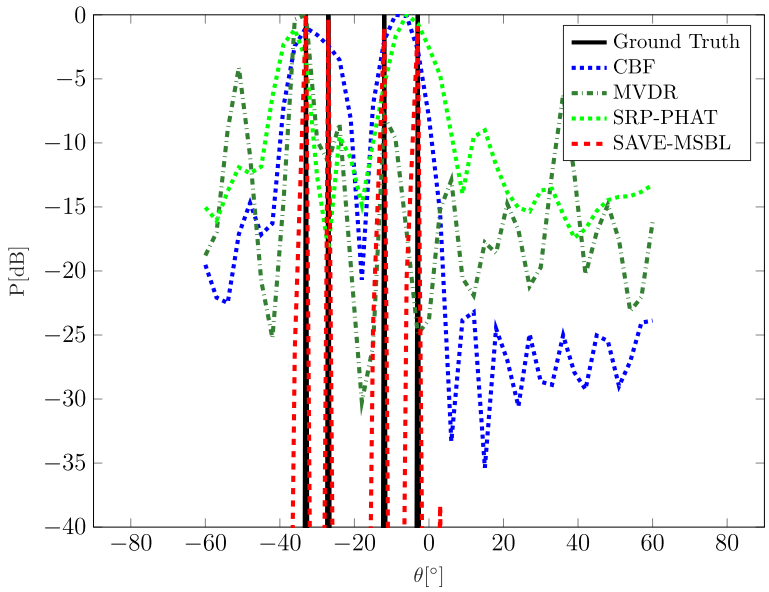

Fig. 3 The resolution performance for different methods

where c and d are modeling parameters.

The hierarchical Bayesian model is built using (2), (3), (4), (5) and (6), and the graphical model is shown in Fig. 2.

3 Bayesian inference using space alternating variational estimation

3.1 Variational Bayesian inference

Let $\boldsymbol{\Theta} = \{\boldsymbol{S}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \rho\}$ denote the set of hidden variables. Based on the graphical model shown in Fig. 2, the joint pdf can be written as

$$p(\boldsymbol{X}, \boldsymbol{\Theta}) = p(\boldsymbol{X}|\boldsymbol{S}, \rho)p(\boldsymbol{S}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\rho). \tag{7}$$

A closed-form expression of the full posterior $p(\boldsymbol{\Theta}|\boldsymbol{X})$ requires computation of the marginal pdf $(\boldsymbol{X})$, which is intractable. In this paper, VBI is therefore applied to obtain an approximation of true posterior using a factorized distribution [42, 43]

$$q(\boldsymbol{\Theta}) = q(\rho)\left(\prod_{k=1}^{K} q(\boldsymbol{s}_k)q(\lambda_k)q(\gamma_k)\right), \tag{8}$$

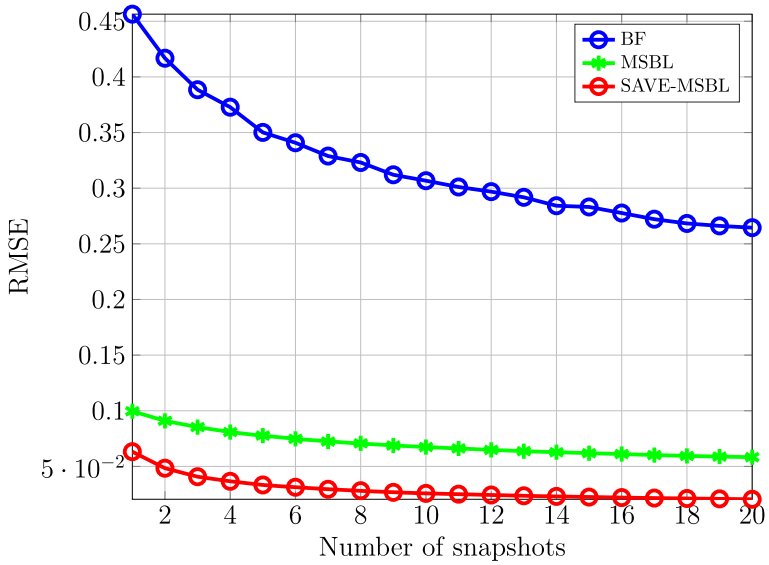

Fig. 4 Recovery accuracy with different numbers of snapshots

Table 1 Parameter setup

	Parameter
Room dimensions in meter	10 × 9 × 8
Reverberation time in seconds	0.25
Reflection order	2
RIR length in samples	1024
Refresh rate of the AIR	128
Sound velocity in meters/second	340
Sampling frequency in kHz	16

where $q(\boldsymbol{\Theta})$ is an approximation of the full posterior $p(\boldsymbol{\Theta}|\boldsymbol{X})$. For notational simplicity, the dependency of the approximated posterior on the observed signal $\boldsymbol{X}$ is omitted. Note that, instead of pursuing the full posterior $q(\boldsymbol{S})$ of the source signals, a factorial form of the posterior $\prod_{k=1}^{K} q(\boldsymbol{s}_k)$ is used to reduce the computational complexity. This is an extension to the SAVE proposed in the single measurement vector (SMV) scenario [38]. When $L = 1$, the proposed approximation model (8) reduces to the model in SAVE. We also assume that the approximate posteriors have the same functional forms as the priors for all the hidden variables. For example, both the prior $p(\boldsymbol{s}_k|\lambda_k)$ and posterior $q(\boldsymbol{s}_k)$ are complex Gaussian. The VBI approach minimizes the Kullback–Leibler (KL) divergence between $p(\boldsymbol{\Theta}|\boldsymbol{X})$ and $q(\boldsymbol{\Theta})$ by maximizing the following variational objective:

$$\mathcal{L} = \mathrm{E}_{q(\boldsymbol{\Theta})}\left[\ln p(\boldsymbol{X}, \boldsymbol{\Theta})\right] - \mathrm{E}_{q(\boldsymbol{\Theta})}\left[\ln q(\boldsymbol{\Theta})\right],$$

where $\mathrm{E}_q[\cdot]$ denotes the expectation operator over the distribution q, i.e., $\mathrm{E}_{q(x)}[p(x)] = \int q(x)p(x)\mathrm{d}x$.

Since the prior and likelihood of all nodes of the model shown in Fig. 2 fall within the conjugate exponential family, the VBI can be written as [42, 43]

$$\ln q(\boldsymbol{\Theta}_i) = \mathrm{E}_{q(\boldsymbol{\Theta}_{\bar{i}})}\left[\ln p(\boldsymbol{S}, \boldsymbol{\Theta})\right] + \mathrm{C}, \tag{9}$$

where C is a constant and $\boldsymbol{\Theta}_i$ denotes one of the variables in the factorized distribution (8), such as $\boldsymbol{s}_k$. The notation $\boldsymbol{\Theta}_{\bar{i}}$ denotes the hidden variable set $\boldsymbol{\Theta}$ excluding $\boldsymbol{\Theta}_i$.

3.2 The logarithm of the joint distribution

As shown in (9), the logarithmic form of the joint distribution is required for VBI. Substituting (2), (3), (4), (5), and (6) into (7), we have

$$\begin{aligned}\ln p(\boldsymbol{X}, \boldsymbol{\Theta}) = {} & ML\ln\rho - \rho\|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|_F^2 + \\ & L\sum_{k=1}^{K}\ln\lambda_k - \sum_{k=1}^{K}\lambda_k\boldsymbol{s}_k\boldsymbol{s}_k^{\mathrm{H}} + \sum_{k=1}^{K}\ln\gamma_k - \\ & \sum_{k=1}^{K}\gamma_k\lambda_k + (a-1)\sum_{k=1}^{K}\ln\gamma_k - \\ & b\sum_{k=1}^{K}\gamma_k + (c-1)\ln\rho - d\rho + \mathrm{C},\end{aligned} \tag{10}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Next, we present the approximate posterior by substituting (10) into (9).

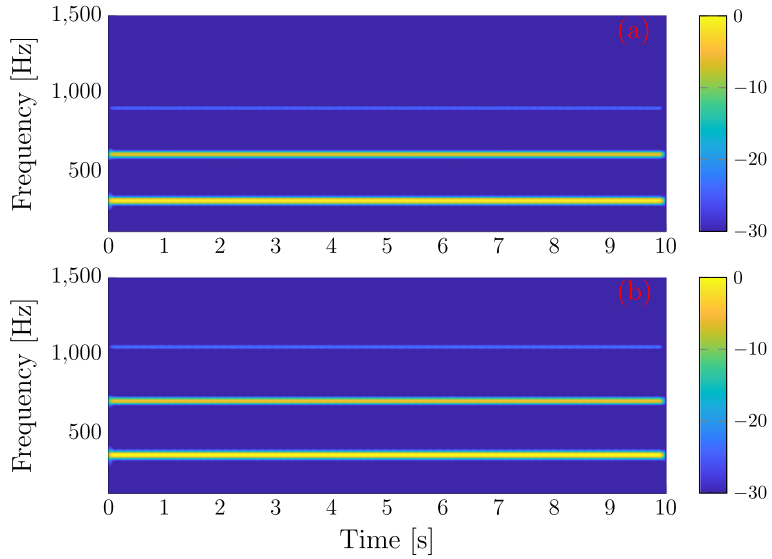

Fig. 5 The spectrograms of the two sources. **a** The spectrogram of source 1. **b** The spectrogram of source 2

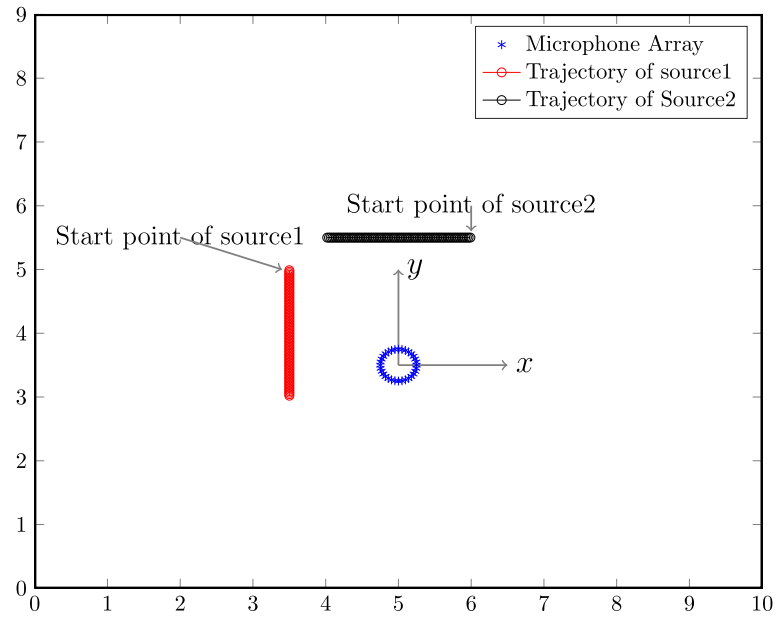

Fig. 6 Illustration of the first virtual room setup with two moving sources for the simulation 1

3.2.1 Update of s_k

The approximate posterior of $\boldsymbol{s}_k$ can be written as [3]

$$\begin{aligned}\ln q(\boldsymbol{s}_k) = &-\operatorname{tr}\left[\boldsymbol{s}_k^{\mathrm{H}}\left(\langle\rho\rangle\,\boldsymbol{a}_k^{\mathrm{H}}\boldsymbol{a}_k+\langle\lambda_k\rangle\right)\boldsymbol{s}_k-\right.\\ &\langle\rho\rangle\,\boldsymbol{s}_k^{*}\boldsymbol{a}_k^{\mathrm{H}}\left(\boldsymbol{X}-\boldsymbol{A}_{\bar{k}}\left\langle\boldsymbol{S}_{\bar{k}}\right\rangle\right)-\\ &\left.\langle\rho\rangle\left(\boldsymbol{X}-\boldsymbol{A}_{\bar{k}}\left\langle\boldsymbol{S}_{\bar{k}}\right\rangle\right)^{\mathrm{H}}\boldsymbol{a}_k\boldsymbol{s}_k^{\mathrm{T}}\right]+\mathrm{C},\end{aligned}\tag{11}$$

where

$$\begin{aligned}\left\langle\boldsymbol{S}_{\bar{k}}\right\rangle =&\mathrm{E}_{q(\boldsymbol{S}_{\bar{k}})}\left[\boldsymbol{S}_{\bar{k}}\right]\\ =&\left[\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_{k-1},\boldsymbol{\mu}_{k+1}\cdots,\boldsymbol{\mu}_K\right]^T,\\ \langle\rho\rangle =&\mathrm{E}_{q(\rho)}[\,\rho],\ \langle\lambda_k\rangle=\mathrm{E}_{q(\lambda_k)}[\,\lambda_k],\end{aligned}$$

and $\langle\cdot\rangle$ is the shorthand of the expectation operator $\mathrm{E}_q[\,\cdot]$. Moreover, $\operatorname{tr}[\,\cdot]$ denotes the trace operator, $\boldsymbol{a}_k$ denotes the kth column of $\boldsymbol{A}$, $\boldsymbol{A}_{\bar{k}}$ is the matrix $\boldsymbol{A}$ with the kth column $\boldsymbol{a}_k$ being removed, and $\boldsymbol{S}_{\bar{k}}$ is the matrix $\boldsymbol{S}$ with the kth row $\boldsymbol{s}_k^{\mathrm{T}}$ being removed. From (11), it can be shown that $q(\boldsymbol{s}_k)=\mathcal{CN}\left(\boldsymbol{s}_k|\boldsymbol{\mu}_k,\sigma_k^2\mathbf{I}\right)$, where

$$\sigma_k^2=\left(M\,\langle\rho\rangle+\langle\lambda_k\rangle\right)^{-1},\tag{12}$$

$$\boldsymbol{\mu}_k=\sigma_k^2\,\langle\rho\rangle\left(\boldsymbol{X}-\boldsymbol{A}_{\bar{k}}\left\langle\boldsymbol{S}_{\bar{k}}\right\rangle\right)^{\mathrm{T}}\boldsymbol{a}_k^{*},\tag{13}$$

where the property $\boldsymbol{a}_k^{\mathrm{H}}\boldsymbol{a}_k=M$ is used. Note that the mean $\{\boldsymbol{\mu}_k\}$ is updated based on the space alternating approach [38, 44], where the newest estimates are always used.

3.2.2 Update of λ, γ and ρ

The approximate posteriors for $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$ and ρ can be derived in a similar way as $\boldsymbol{s}_k$, and we only give the results here.

[3] See Appendix A: Derivation of (11) for more derivation details.

Update $q(\lambda_k)$: $q(\lambda_k)=\mathcal{G}(\alpha_{\lambda_k},\beta_{\lambda_k})$, where

$$\begin{aligned}&\alpha_{\lambda_k}=1+L,\ \beta_{\lambda_k}=\boldsymbol{\mu}_k^H\boldsymbol{\mu}_k+L\sigma_k^2+\langle\gamma_k\rangle,\\ &\langle\lambda_k\rangle=\frac{\alpha_{\lambda_k}}{\beta_{\lambda_k}}.\end{aligned}\tag{14}$$

Update $q(\gamma_k)$: $q(\lambda_k)=\mathcal{G}\left(\alpha_{\gamma_k},\beta_{\gamma_k}\right)$, where

$$\alpha_{\gamma_k}=1+a,\ \beta_{\gamma_k}=\langle\lambda_k\rangle+b,\ \langle\gamma_k\rangle=\frac{\alpha_{\gamma_k}}{\beta_{\gamma_k}}.\tag{15}$$

Update $q(\rho)$: $q(\rho)=\mathcal{G}(\alpha_\rho,\beta_\rho)$, where

$$\begin{aligned}&\alpha_\rho=ML+c,\\ &\beta_\rho=\|\boldsymbol{X}-\mathbf{A}\,\langle\boldsymbol{S}\rangle\,\|_F^2+L\operatorname{tr}\left[\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A}\right]+d,\\ &\quad=\|\boldsymbol{X}-\mathbf{A}\,\langle\boldsymbol{S}\rangle\,\|_F^2+ML\sum_{k=1}^{K}\sigma_k^2+d,\\ &\langle\rho\rangle=\frac{\alpha_\rho}{\beta_\rho},\end{aligned}\tag{16}$$

where $\boldsymbol{\Sigma}=\operatorname{diag}[\,\sigma_1^2,\cdots,\sigma_2^2,\cdots,\sigma_K^2]$ and $\operatorname{diag}[\,\cdot]$ denotes a diagonal matrix.

We refer to the proposed algorithm as SAVE-MSBL. By using the space alternating approach, the computationally complex matrix inversion operation of the traditional MSBL [19] can be avoided. Moreover, instead of using the above formulas directly, we can further reduce the computational complexity by introducing a temporary matrix $\widehat{\boldsymbol{X}}$, which can be seen as an approximation of $\boldsymbol{X}$. By removing or adding the terms $\boldsymbol{a}_k\boldsymbol{\mu}_k^T$, the two terms $\boldsymbol{A}_{\bar{k}}\left\langle\boldsymbol{S}_{\bar{k}}\right\rangle$ and $\mathbf{A}\,\langle\boldsymbol{S}\rangle$ in (13) and (16) can be updated using $\boldsymbol{a}_k\boldsymbol{\mu}_k^T$, resulting in a computationally efficient implementation. The pseudocode for the proposed method

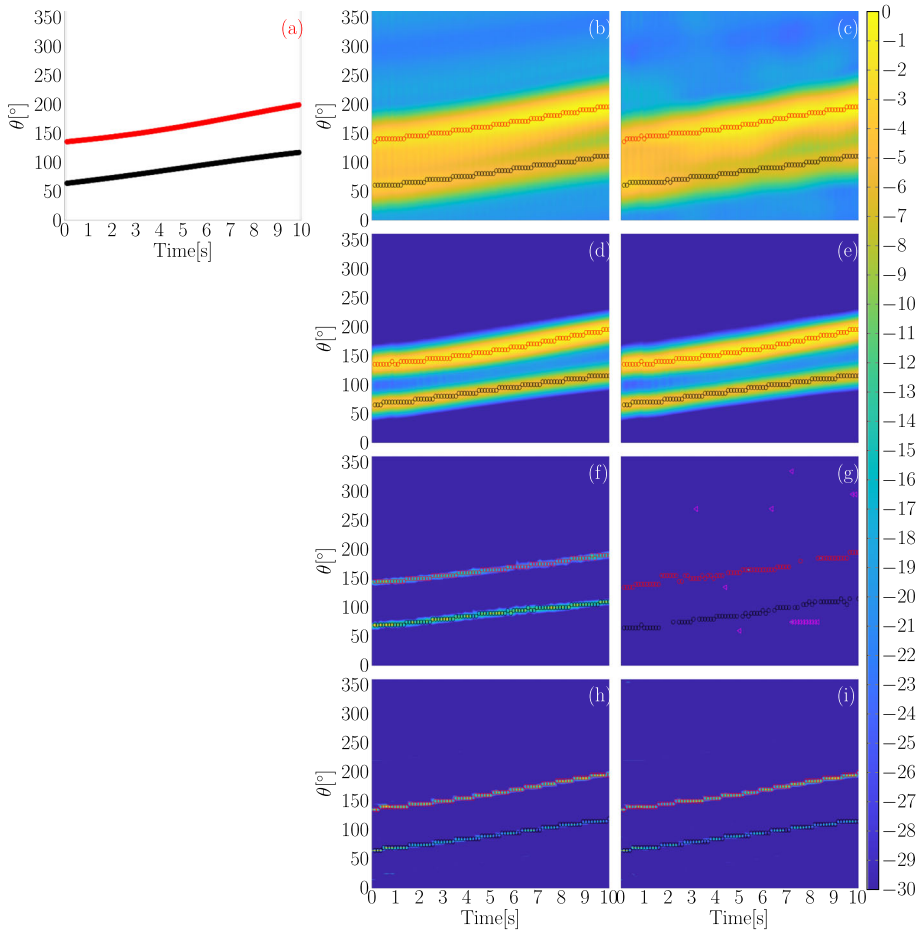

Fig. 7 Estimation results for the first virtual room setup. For the CBF, MVDR ,and SRP-PHAT methods, the estimation results are shown using the spatial spectrum of all frames. For the proposed SAVE-MSBL-EM method, the estimation results are shown using the weights $\boldsymbol{w}$ of all frames. **a** The trajectories of two source. **b** Estimation result of CBF method in free field. **c** Estimation result of CBF method in low reverberation. **d** Estimation result of SRP-PHAT method in free field. **e** Estimation result of SRP-PHAT method in low reverberation. **f** Estimation result of MVDR method in free field. **g** Estimation result of MVDR method in low reverberation. **h** Estimation result of the proposed SAVE-MSBL-EM method in free field. **i** Estimation result of the proposed SAVE-MSBL-EM method in low reverberation

is shown Algorithm 1. Note that the proposed SAVE-MSBL algorithm can be applied to each frequency bin independently.

3.3 CGMM-based acoustic DOA estimator

Up to this point, the posteriors of the source signals (i.e., $\{q(\boldsymbol{s}_{f,k})\}$) from all the frequency bins are obtained independently. The source signals $\boldsymbol{s}_{f,k}$ can be estimated using the MMSE estimator, i.e.,

$$\widehat{\boldsymbol{s}}_{f,k} = \boldsymbol{\mu}_{f,k}, \tag{17}$$

where $\widehat{\boldsymbol{s}}_{f,k}$ denotes the estimate of the source signal. In this section, we propose an acoustic DOA estimator, jointly utilizing the estimated source signals from all the frequency bins, based on the CGMM model. By fitting the observations and estimates of the source signals to the CGMM model, the weighting parameters can be obtained using the EM algorithm. The weighting parameter of each mixture component in the CGMM can be seen as the probability that there is an active acoustic source at the corresponding candidate location. With a known number of sources, the DOA estimates for all the sources can be obtained using peak-picking on the weighting parameters.

Table 2 Performance of all the methods for simulation 1

		CBF	SRP-PHAT	MVDR	SAVE-MSBL-EM
FF	RMSE[°]	4.8	2.9	2.1	1.7
	FA[%]	0	0	0	0
	MD[%]	0	0	0	0
RB	RMSE[°]	5.4	3.0	3.9	1.8
	FA[%]	0	0	10.2	0
	MD[%]	0	0	14	0

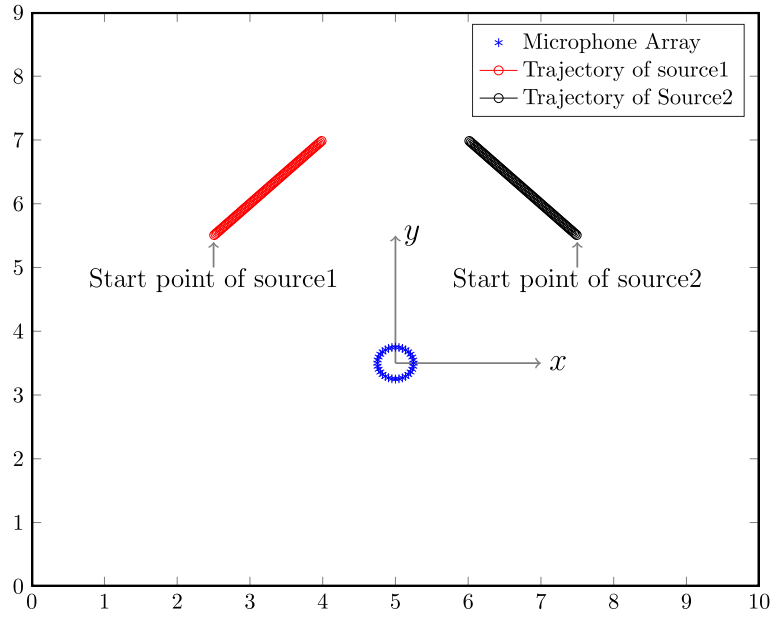

Fig. 8 Illustration of the second virtual room setup with two moving sources for the simulation 2

Algorithm 1 The proposed SAVE-MSBL algorithm

Initialize the modeling parameters a, b, c and d.
Initialize the threshold of stopping the iteration err_{max} and the maximum iteration number I_{max}.
Initialize $\{\langle \boldsymbol{\mu}_k \rangle\}$, $\{\langle \lambda_k \rangle\}$, $\{\langle \gamma_k \rangle\}$, $\langle \rho \rangle$ and $\{\sigma_k^2\}$.
Set the error $err = 0$ and the iteration counter $I = 0$.
Initialize the temporary matrix $\widehat{\boldsymbol{X}} = \boldsymbol{A} \langle \boldsymbol{S} \rangle$.
while $err > err_{max}$ and $I < I_{max}$ **do**
  for $k = 1, 2, \cdots, K$ **do**
    $\boldsymbol{\lambda}^{\text{old}} = \boldsymbol{\lambda}$.
    Update $\widehat{\boldsymbol{X}}$ with $\widehat{\boldsymbol{X}} \leftarrow \widehat{\boldsymbol{X}} - \boldsymbol{a}_k \boldsymbol{\mu}_k^T$.
    Update σ_k^2 using (12).
    Update $\boldsymbol{\mu}_k$ using $\boldsymbol{\mu}_k = \sigma_k^2 \langle \rho \rangle \left(\boldsymbol{X} - \widehat{\boldsymbol{X}}\right)^{\mathrm{T}} \boldsymbol{a}_k^*$
    Update $\widehat{\boldsymbol{X}}$ with $\widehat{\boldsymbol{X}} \leftarrow \widehat{\boldsymbol{X}} + \boldsymbol{a}_k \boldsymbol{\mu}_k^T$.
    Update $\langle \lambda_k \rangle$ using (14).
    Update $\langle \gamma_k \rangle$ using (15).
  end for
  Update $\langle \rho \rangle$ using $\langle \rho \rangle = \frac{ML+c}{\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|_F^2 + ML \sum_{k=1}^K \sigma_k^2 + d}$.
  Update the iteration number $I = I + 1$.
  Update the error $err = \frac{\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{\text{old}}\|_2}{\|\boldsymbol{\lambda}^{\text{old}}\|_2}$.
end while

Inspired by the Gaussian mixture model [45, 46] and the probabilistic steered-response power (SRP) model [47, 48], we assume that $\boldsymbol{x}_{f,l}$ follows a CGMM distribution with estimated source signals $\boldsymbol{s}_{f,k}$, i.e.,

$$p(\boldsymbol{x}_{f,l}; \boldsymbol{w}) = \sum_{k=1}^{K} w_k \mathcal{CN}\left(\boldsymbol{x}_{f,l} | \boldsymbol{a}_{f,k} \mu_{f,k,l}, \eta\right),$$

where η is an empirically chosen small value, and $w_k \geq 0$ is the weighting parameter for the kth complex Gaussian component with the constraint $\sum_{k=1}^K w_k = 1$. Then, the distribution of the observation set for all frequency bins can be expressed as

$$p(\boldsymbol{Y}; \boldsymbol{w}) = \prod_{f=1}^{F} \sum_{k=1}^{K} w_k \left[\prod_{l=1}^{L} \mathcal{CN}\left(\boldsymbol{x}_{f,l} | \boldsymbol{a}_{k,f} \mu_{f,k,l}, \eta\right) \right], \quad (18)$$

where $\boldsymbol{Y} = \{\boldsymbol{X}_f\}_{f=1}^F$ is the observation set for all frequency bins. Once (18) is maximized, each weight w_k represents the probability of an acoustic source being active in the direction θ_k. However, it is intractable to maximize the function in (18) due to its high dimensionality. Therefore, an EM procedure is applied to deal with this maximization problem. Following [42], we introduce a set of hidden variables $\boldsymbol{z} = \{\boldsymbol{r}_f\}_{f=1}^F$. The $\boldsymbol{r}_f$ contains binary random variables with only one particular element $r_{f,k}$ being 1 while the others are all zeros. The variable $r_{f,k}$ can be seen as an indicator associated with the acoustic source from the direction θ_k at the fth frequency bin. Assuming $p(r_{f,k} = 1) = w_k$, we can write the joint distribution as follows:

$$p(\boldsymbol{z}; \boldsymbol{w}) = \prod_{f=1}^{F} \prod_{k=1}^{K} w_k^{r_{f,k}}. \quad (19)$$

The conditional distribution of the observation set $\boldsymbol{Y}$ given $\boldsymbol{z}$ is

$$p(\boldsymbol{Y}|\boldsymbol{z}) = \prod_{f=1}^{F}\prod_{k=1}^{K}\left[\prod_{l=1}^{L}\mathcal{CN}(\boldsymbol{x}_{f,l}|\boldsymbol{a}_{k,f}\mu_{f,k,l},\eta)\right]^{r_{f,k}}. \quad (20)$$

Then, the joint distribution can be derived from (19) and (20) using Bayes' rule, i.e.,

$$\begin{aligned} p(\boldsymbol{Y},\boldsymbol{z};\boldsymbol{w}) &= p(\boldsymbol{Y}|\boldsymbol{z})p(\boldsymbol{z};\boldsymbol{w}) \\ &= \prod_{f=1}^{F}\prod_{k=1}^{K}\left[w_k\prod_{l=1}^{L}\mathcal{CN}(\boldsymbol{x}_{f,l}|\boldsymbol{a}_{k,f}\mu_{f,k,l},\eta)\right]^{r_{f,k}}. \end{aligned} \quad (21)$$

3.3.1 E-step

In the E-step, we use the current parameter $\hat{\boldsymbol{w}}^{\text{old}}$ to update the posterior mean of the hidden variable denoted as $\mathrm{E}[\,r_{f,k}|\boldsymbol{Y};\hat{\boldsymbol{w}}^{\text{old}}]$. From (21), the E-step can be written as

$$\begin{aligned} Q(\boldsymbol{w};\hat{\boldsymbol{w}}^{\text{old}}) &= \mathrm{E}[\,r_{f,k}|\boldsymbol{Y};\hat{\boldsymbol{w}}^{\text{old}}] \\ &= \sum_{f=1}^{F}\sum_{k=1}^{K}\mathrm{E}\left[r_{f,k}|\boldsymbol{Y};\hat{\boldsymbol{w}}^{\text{old}}\right]\left[\ln\hat{w}_k^{\text{old}}+\phi_{f,k,l}\right], \end{aligned} \quad (22)$$

where

$$\begin{aligned} \phi_{f,k,l} &= \sum_{l=1}^{L}\left[\ln\mathcal{CN}(\boldsymbol{x}_{f,l}|\boldsymbol{a}_{k,f}\mu_{f,k,l},\eta)\right] \\ &= \sum_{l=1}^{L}\left\{-M\ln\eta-\frac{1}{\eta}\left[\|\boldsymbol{x}_{f,l}-\boldsymbol{a}_{k,f}\mu_{f,k,l}\|^2\right]\right\}, \end{aligned}$$

where $\mu_{f,k,l}$ is obtained using Algorithm 1.

Therefore, the expected value $\mathrm{E}[\,r_{f,k}|\boldsymbol{Y};\hat{\boldsymbol{w}}^{\text{old}}]$ is given by [42, 49]

$$\mathrm{E}\left[r_{f,k}|\boldsymbol{Y};\hat{\boldsymbol{w}}^{\text{old}}\right] = \frac{\hat{w}_k^{\text{old}}\exp\left(\phi_{f,k,l}\right)}{\sum_{\tilde{k}=1}^{K}\hat{w}_{\tilde{k}=1}^{\text{old}}\exp\left(\phi_{f,\tilde{k},l}\right)} = \langle r_{f,k}\rangle. \quad (23)$$

3.3.2 M-step

In the M-step, the required parameter $\boldsymbol{w}$ is updated through a constrained maximization of (22), i.e.,

$$\begin{aligned} \hat{\boldsymbol{w}}^{\text{new}} &= \arg\max_{\boldsymbol{w}} Q\left(\boldsymbol{w};\boldsymbol{w}^{\text{old}}\right) \\ &s.t.\sum_{k=1}^{K}w_k = 1; 0 < w_k < 1. \end{aligned} \quad (24)$$

Therefore, the M-step can be stated as

$$\hat{w}_k^{\text{new}} = \frac{\sum_{f=1}^{F}\langle r_{f,k}\rangle}{\sum_{f=1}^{F}\sum_{\tilde{k}=1}^{K}\left\langle r_{f,\tilde{k}}\right\rangle} = \frac{1}{F}\sum_{f=1}^{F}\langle r_{f,k}\rangle. \quad (25)$$

Given an initial value for the parameter $\boldsymbol{w}$, the EM algorithm iterates between the E-step in (23) and the M-step in (25) until convergence. The EM algorithm is summarized in Algorithm 2.

Algorithm 2 The EM algorithm

Initialize the threshold of error err_0 and the parameter η.
while Convergence criterion not meet **do**
  for $k = 1, \cdots, K$ **do**
    Update $\mathrm{E}[\,r_{f,k}|\boldsymbol{Y};\hat{\boldsymbol{w}}^{\text{old}}]$ using (23).
    Update the weight $\hat{w}_k$ using (25).
  end for
end while

4 Results and discussion

In this section, we first investigate the computational complexity of the proposed SAVE-MSBL-EM method. Then, we test the performance of our proposed SAVE-MSBL-EM algorithm using both synthetic data and real data from the LOCATA dataset[4]. The performance of the different methods are tested in three different scenarios. In the first scenario, we test the recovery accuracy and the resolution performance using narrow-band sources and a ULA. In the second part, we consider a complicated scenario with closely spaced sources in a virtual room. Last, the proposed method is tested using real data.

4.1 Computational complexity analysis

We first analyze the computational complexity of the proposed SAVE-MSBL algorithm by counting the number of mathematical multiplication/division operations in each iteration. As can be seen from Algorithm 1, in each "*for*" loop, the complexity of the proposed algorithm mainly depends on the update of the temporary matrix $\bar{\boldsymbol{X}}$ and $\boldsymbol{\mu}_k$, which is $\mathcal{O}(ML)$. The computational complexity of updating $\langle\rho\rangle$ is $\mathcal{O}(ML)$. Therefore, the computational complexity of the proposed algorithm for each iteration is $\mathcal{O}(KML)$. If we consider the variational Bayesian inference without the space alternating approach, the computational complexity is $\mathcal{O}(M^3L^3)$. Thus, the space alternating approach leads to a significant reduction on the computational complexity. Moreover, the computational complexity of MSBL proposed in [19] is $\mathcal{O}(KM^2)$. Therefore, the proposed method is faster than the MSBL method when $L < M$. Since the SVD approach can be utilized for data reduction [18], the condition $L < M$ is met in most cases. For the EM algorithm, the computational complexity is $\mathcal{O}(KML)$ for one frequency bin. Thus, the computational complexity of the proposed SAVE-MSBL-EM method is $\mathcal{O}(KML)$ for each frequency bin.

[4] The LOCATA dataset is publicly available at https://www.locata.lms.tf.fau.de/

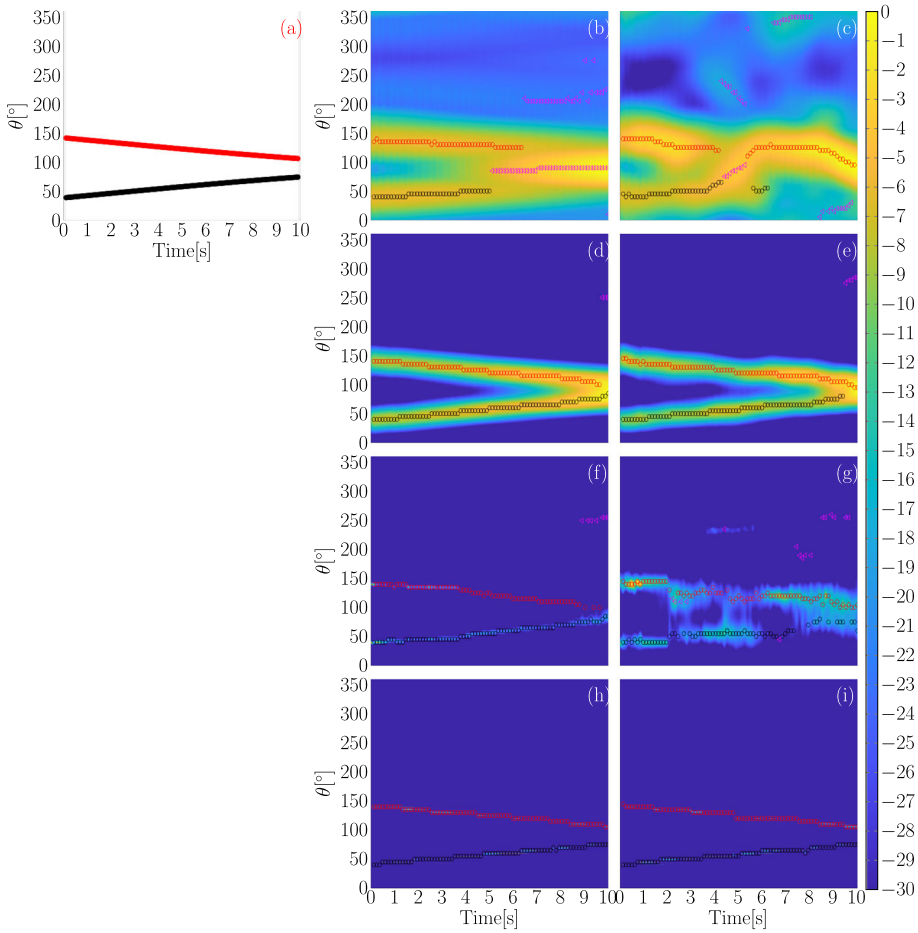

Fig. 9 Estimation results for the second virtual room setup. For the CBF, MVDR and SRP-PHAT methods, the estimation results are shown using the spatial spectrum of all frames. For the proposed SAVE-MSBL-EM method, the estimation results are shown using the weights $\boldsymbol{w}$ of all frames. **a** The trajectories of two source. **b** Estimation result of CBF method in free field. **c** Estimation result of CBF method in low reverberation. **d** Estimation result of SRP-PHAT method in free field. **e** Estimation result of SRP-PHAT method in low reverberation. **f** Estimation result of MVDR method in free field. **g** Estimation result of MVDR method in low reverberation. **h** Estimation result of the proposed method in free field. **i** Estimation result of the proposed method in low reverberation

We further measure the computational complexity using the "cputime" function provided by MATLAB. The computer is equipped with an i7-8700 processor. The clock rate is 3.19 GHz. The operation system is Windows 10. The software is MATLAB 2019a. We test the computational complexity for one frequency bin. The number of iterations is fixed to 100, the number of candidate DOAs is set to 41, the number of microphones is set to 15, the number of snapshots is set to 10, and the number of Monte-Carlo experiments is set to 1000. For a single frequency bin, the time consumption of the proposed SAVE-MSBL-EM method and the MSBL proposed in [19] are 0.08 and 0.25 s, respectively, i.e., the proposed method is faster than the MSBL method by a factor of $\sim$ 3. Note that, in practice, the time consumption for the acoustic DOA estimation algorithm is proportional to the number of frequency bins.

Table 3 Performance of all the methods for simulation 2

		CBF	SRP-PHAT	MVDR	SAVE-MSBL-EM
FF	RMSE[°]	4.1	2.7	3.6	2.3
	FA[%]	42.7	1.8	4.2	0
	MD[%]	43.9	3.0	5.4	1.2
RB	RMSE[°]	6.3	2.9	5.9	2.5
	FA[%]	25.9	3.0	10.2	0
	MD[%]	31.9	4.2	11.4	1.2

4.2 Experimental results

The methods used for comparison in this section are summarized as following: CBF refers to classical beamforming

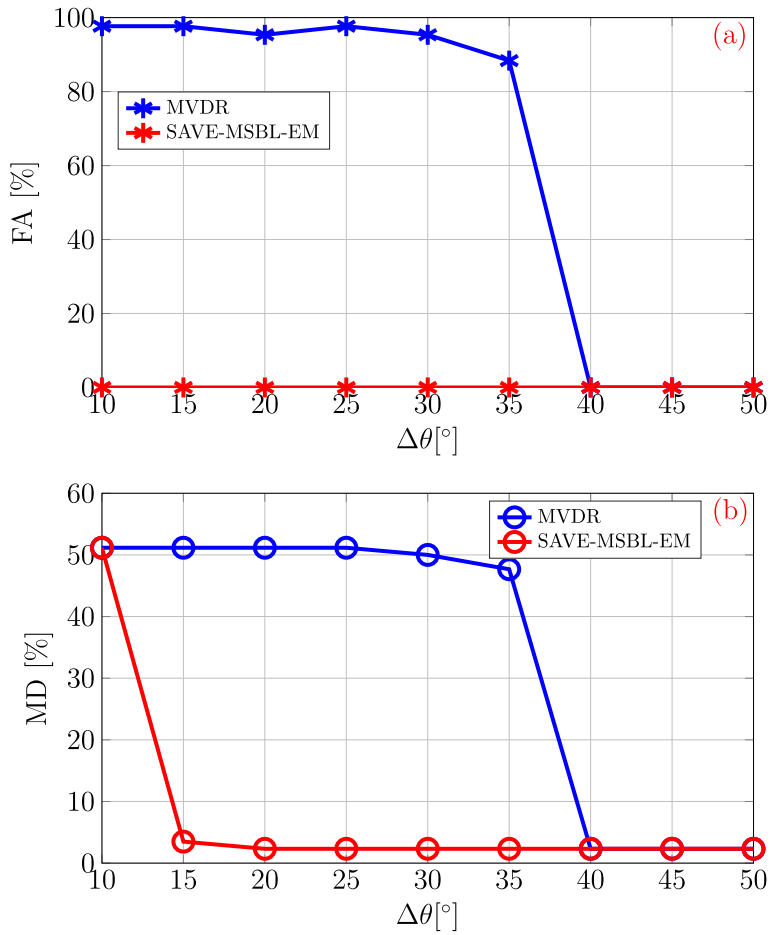

Fig. 10 Simulation results. **a** False alarm rate in low-reverberation environment. **b** Miss-detected rate in low-reverberation environment

based method which is widely used in practice; SRP-PHAT is another widely used method for sound source localization especially in reverberant environments [9]; and MVDR is a method offering high-resolution performance [10]. Note that the implementation of the MVDR method is based on the observed signal statistics. Moreover, MSBL refers to the multiple snapshots SBL method for narrow-band signals proposed in [19]. MSBL-EM is an acoustic DOA estimator which combines the MSBL algorithm and proposed EM algorithm. Furthermore, SAVE-MSBL is the proposed method for narrow-band signals and SAVE-MSBL-EM is the proposed method for acoustic DOA estimation. For the MSBL method, the threshold for stopping the iteration err_{max} is set to $1e-10$. For the proposed SAVE-MSBL-EM method,

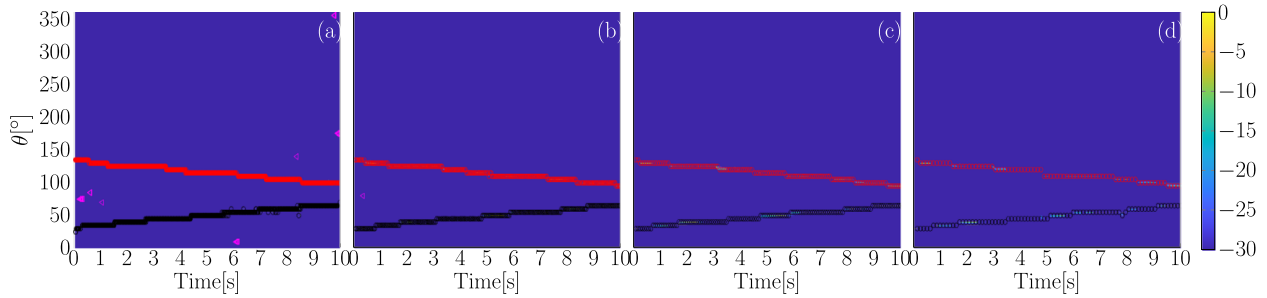

Fig. 11 The performance of the proposed method versus different number of snapshots L in a frame. **a** $L = 1$. **b** $L = 5$. **c** $L = 10$. **d** $L = 15$

Table 4 Performance of the proposed method versus L

	1	5	10	15
RMSE[°]	3.3	2.5	2.5	2.5
FA[%]	1.8	0.2	0	0
MD[%]	2.4	1.3	1.2	1.2

the modeling parameters a, b, c, and d are all set to $1e-3$, the parameter η is set to 0.1, the threshold for the SAVE-MSBL algorithm err_{max} is set to $1e-10$, and the threshold for the EM algorithm err_0 is set to $1e-3$.

4.2.1 Recovery performance analysis using a ULA

In this section, we test the recovery performance of the proposed SAVE-MSBL algorithm using four acoustic sources comprising pure sinusoidal signals. Two assumptions are made in this simulation: (1) all the acoustic sources are located in the far-field of the microphone array and (2) the power of all the acoustic sources are equal. The frequencies of all the sources are set to 1 kHz. For each source, the initial phase is generated randomly. Assume that a ULA with 15 omni-directional microphones is used to receive the signals. The distance between adjacent microphones is set to 0.05 m in this simulation. The microphone array data are generated by assigning different time delays according to the true bearing angles of the sources. White Gaussian noise is added to the clean array data and the SNR is set to 10 dB. The sampling frequency is set to 16 kHz. The time-domain data are converted to the frequency-domain using the short-time Fourier transform (STFT). The temporal length of the snapshot is set to 1024. The length of the increment for the snapshots is set to 256, i.e., the overlap is 75%. The length of the FFT is set to 2048. The number of snapshots is set to 10. As the frequencies of all sources are 1 kHz, only the frequency bin whose center frequency is 1kHz is used for the estimation. We define the fan-shaped horizontal plane in the range from $-60°$ to $60°$ as the target space (see Fig. 1). The target space is uniformly separated with a grid interval of $3°$, i.e., the number of grid points is 41 and the array response matrix (dictionary) is pre-computed according to these grid points. Moreover, the bearing angles of four pure sinusoidal sources are $-33°$, $-27°$, $-12°$, and $-3°$, respectively. Figure 3 shows the estimation results of the CBF, MVDR, SRP-PHAT, and SAVE-MSBL methods.

It can be seen that the CBF and SRP-PHAT methods fail to separate the two sources located at $-33°$ and $-27°$, but the MVDR and proposed SAVE-MSBL methods still work in this case.

We now proceed to test the performance of the proposed method with respect to the number of snapshots. The number of Monte-Carlo runs is 1000. The recovery accuracy is measured by the root-mean-square-error (RMSE), defined as

$$e_{rec} = \left(\sqrt{\frac{1}{N_{MC}L} \sum_{i=1}^{N_{MC}} \frac{\|\hat{\boldsymbol{S}} - \boldsymbol{S}\|_F^2}{\|\boldsymbol{S}\|_F^2}} \right), \tag{26}$$

where $\hat{\boldsymbol{S}}$ is the recovered signal, $\boldsymbol{S}$ is the true signal, $\|\cdot\|_F$ denotes the Frobenius norm, L is the number of snapshots, and N_{MC} is the number of Monte-Carlo experiments. We compare the proposed method with the CBF method in [6] and one of the widely used MSBL algorithms proposed in [19]. The results of the RMSEs of the recovered signals are illustrated in Fig. 4. It can be seen that the recovery performance of all the methods improve dramatically as the number of snapshots increases in the range from 1 to 3. Moreover, the simulation result shows that the proposed SAVE-MSBL method achieves better recovery accuracy compared with the CBF and MSBL methods.

4.2.2 Simulation with virtual room

In this part, we test the resolution performance of the proposed method with respect to different intervals of bearing angles between two sources. The synthetic array data are generated using the "signal-generator"[5] with a virtual room. Note that the "signal-generator" is designed for the moving source scenario. The room setup is summarized in Table 1.

In this virtual room, a uniform circular array (UCA) with 32 omni-directional microphones is used to record the signals. The center position of the UCA is $(5, 3.5, 3)$ m. The radius of the UCA is set to 0.25 m. Two acoustic sources are used. Both of them play uninterrupted harmonic signals. The fundamental frequencies of the two sources are 300 Hz and 350 Hz, respectively. The spectrograms of the two sound sources are shown in Fig. 5.

We assume the sound sources are moving on a horizontal plane where the microphone array is located in. The horizontal plane is separated into 73 grid points from $0°$ to $360°$ with an angle interval $5°$, where $0°$ is in the positive direction of the x-axis and $90°$ is in the positive direction of the y-axis. For simulation 1, the trajectories of the two sources are illustrated in Fig. 6. The first source moves along the negative direction of y-axis while the second source moves along the negative direction of x-axis. The original positions of the first and second sound sources are $(3.5, 5, 3)$ m and $(6, 5.5, 3)$ m. The end positions are $(3.5, 3, 3)$ m and $(4, 5.5, 3)$ m, respectively. The true DOA trajectories of the two sources with respect to the microphone array are shown in Fig. 7(a).

According to the simulation setup, the time-domain array signals can be generated using the "signal-generator." Then, the received array signals are first segmented

[5] The "signal-generator" for synthetic array data generation is online available: https://www.audiolabs-erlangen.de/fau/professor/habets/software.

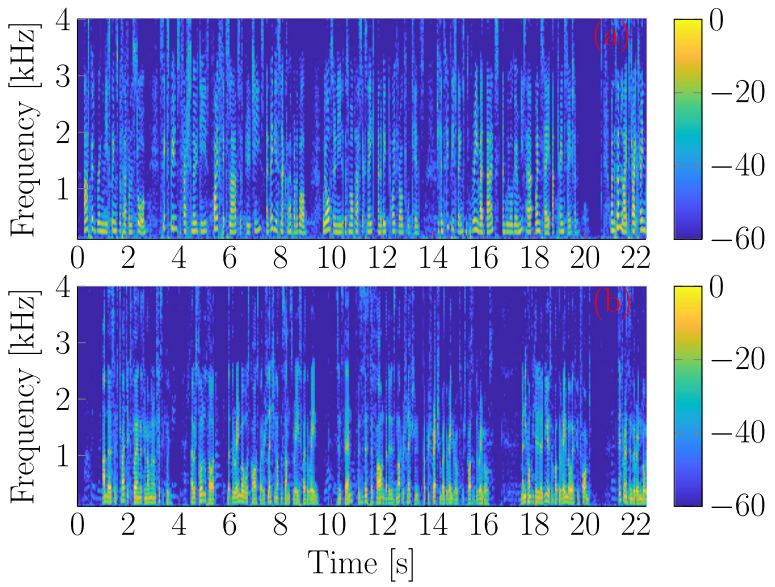

Fig. 12 The spectrograms of the two sources. **a** The spectrogram of source 1. **b** The spectrogram of source 2

into a batch of snapshots with 87.5% overlap. By applying the fast Fourier transform (FFT) on each snapshot, the time-domain array signals are converted to the frequency-domain array data. Then, the frequency-domain array data is segmented into several frames with L consecutive snapshots grouped as one frame. In the first and second simulations, L is set to 15. The effect of L is discussed in the last part of this subsection. Note that the SVD approach is used for data reduction in this paper. After applying acoustic DOA estimation methods for each frame, we find the peaks for each frame and label these peaks according to the ground truth DOAs of the two sources. The error range is set to 15°, i.e., if the minimum error between the estimated angle and all ground truth angles is larger than 15°, we label the peak as a false estimate. In this paper, we use the black and red circles to denote estimates of the first source and the second source, respectively. Moreover, we use magenta triangles to denote false estimates.

To quantitatively show the difference of the resolution performance between the proposed SAVE-MSBL-EM method and other methods, the RMSE, the false alarm (FA) rate, and the miss-detected (MD) rate are used to measure the recovery performance. The RMSE is defined as

$$e = \sqrt{\frac{1}{N_c}\sum_{i=1}^{N_c}\left|\tilde{\theta}_i - \theta_i\right|^2}, \tag{27}$$

where N_c is the total number of correct estimates, $\tilde{\theta}_i$ is the ith correct estimate, and θ_i is the ith true bearing angle. Following [50], the FA rate is defined as the percent of sources that are falsely estimated out of the total number of sources and the MD rate is defined as the percent of sources that are miss-detected out of the total number of sources, i.e.,

$$FA = \frac{N_F}{N_T} \times 100\%, \quad MD = \frac{N_{M1} + N_{M2}}{2N_T} \times 100\%,$$

where N_F is the number of sources with false estimation, N_T is the total number of sources for all frames, and N_{M1} and N_{M2} are the miss-detected number of the first source and the second source, respectively. Note that two continuous harmonic sound signals are used in this simulation. Thus, two active sources exist in each frame.

We consider two reverberation conditions for all the methods: the free-field (no reverberation) and low-reverberation conditions (RT60 = 0.25 s). For the CBF, MVDR, and SRP-PHAT methods, the estimation results are shown using the spatial spectrums of all frames. For the proposed SAVE-MSBL-EM method, the estimation results are shown using the weight, $\boldsymbol{w}$, of all frames. For comparison, all the data are normalized frame by frame and displayed using color maps.

In simulation 1, the estimation results of the CBF method in free-field and low-reverberation environments are shown in Figs. 7b and c, respectively. The estimation results of the different methods in both the free field and low reverberation conditions are shown in Fig. 7b–i. The RMSE, FA, and MD are shown in Table 2. Note that "FF" refers to the free-field condition and "RB" refers to the reverberation environment. It can be seen that all the methods perform well under the free-field condition. In the presence of reverberation, the good accuracy

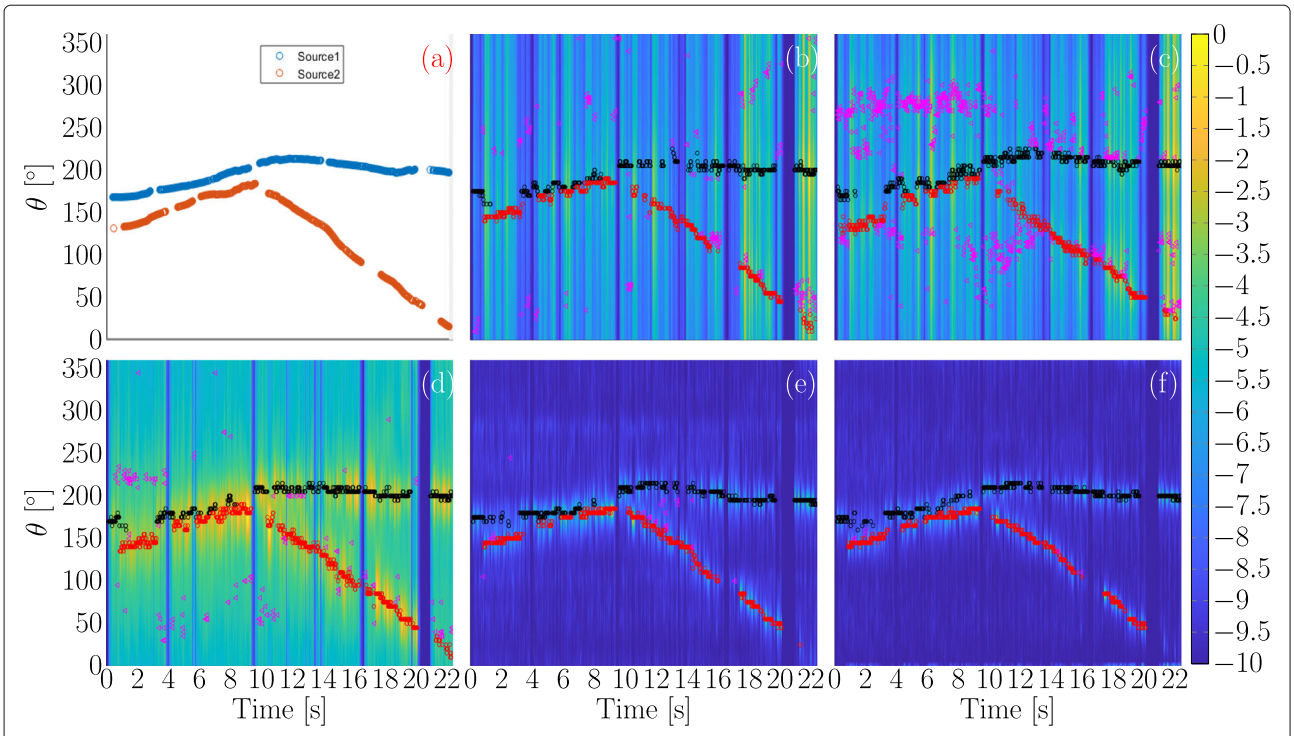

Fig. 13 Estimation results for the real data. For the CBF, MVDR and SRP-PHAT methods, the estimation results are shown using the spatial spectrum of all frames. For the MSBL-EM and proposed SAVE-MSBL-EM method, the estimation results are shown using the weights $\boldsymbol{w}$ of all frames. All the data are normalized frame by frame. **a** The truth trajectory of two source. **b** The results of CBF based method. **c** The results of MVDR based method. **d** The results of SRP-PHAT method. **e** The results of MSBL-based method. **f** The results of the proposed method

performance of the CBF, SRP-PHAT, and proposed SAVE-MSBL-EM method are retained but the MVDR method degrades considerably.

To further verify the performance of the proposed SAVE-MSBL-EM method in terms of resolution, another scenario is considered. In this case, all of the setup remains the same except the trajectories of the two sources. We refer to this simulation as simulation 2. The original position of the first source is (2.5, 5.5, 3) m while the second is (7.5, 5.5, 3) m. The end positions are (4, 7, 3) m and (6, 7, 3) m, respectively. Figure 8 shows the trajectories of the two sources in the virtual room.

The true bearing angles of the two sources with respect to the microphone array are illustrated in Fig. 9a. The estimation results of the CBF, SRP-PHAT, MVDR, and SAVE-MSBL-EM methods in the free-field environment are shown in Figs. 9b, d, f, and h, respectively, while the results for the low reverberation condition are shown in Figs. 9c, e, g, and i, respectively. The RMSE, FA, and MD are summarized in Table 3.

From Figs. 9b, c, d, e, f, and g, it can be seen that the performance of the CBF, SRP-PHAT and MVDR methods degrade dramatically as two sound sources move closer. However, the proposed SAVE-MSBL-EM method retains an accurate estimation performance for the acoustic DOA estimation. In this case, the proposed SAVE-MSBL-EM method offers higher resolution performance than other methods.

We then test the performance of the proposed method and MVDR method using static sources and the results are shown in Fig. 10.

The microphone array signals are generated using the "rir-generator"[6]. The distance between the sound sources and the microphone array center is set to 3 m. We tested the FA rate with different bearing intervals between the two sound sources in the low reverberation condition (RT60 = 0.25 s). Figure 10(a) depicts the FA rates of the MVDR method and the proposed algorithm.

It can be seen that the proposed SAVE-MSBL-EM algorithm has a lower FA rate in the interval range from 15° to 40°. Figure 10(b) shows he MD rates of two algorithms. Compared with the MVDR method, the proposed method has a lower MD rate in the range from 15° to 40°. From Figs. 7, 9 and 10, we can thus conclude that the proposed SAVE-MSBL-EM method provides a better resolution performance than the CBF, SRP-PHAT, and MVDR methods in both free-field and low-reverberation conditions.

[6]The RIR generator is publicly available at: https://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator.

To test the effect of the frame (window) length L on the localization performance, we conduct a simulation for different number of snapshots L. The simulation setup is the same as that of simulation 2, that is, the trajectories of the two sources and the true bearing angles of the two sources with respect to the microphone array are shown in Figs. 8 and 9a, respectively. The simulation is conducted in the reverberation environment (RT60 = 0.25 s). The results are illustrated in Fig. 11. The RMSE, FA, and MD are shown in Table 4.

It can be seen that the proposed method works for all snapshot numbers. However, the localization performance degrades if the number of snapshots is small, e.g., the FA and MD in Figs. 11a and b are higher than the FA and MD in Figs. 11c and d.

4.2.3 Real data experiments

The LOCATA dataset provides a series of microphone array data recorded in the Computing Laboratory of the Department of Computer Science of Humboldt University Berlin [51]. The room size is 7.1 × 9.8 × 3 m, with the reverberation time RT60 = 0.55 s. In this paper, we use the "benchmark2" microphone array data in task #6 to test the high-resolution performance of the proposed method. The number of microphones of the "benchmark2" array is 12. Two speakers are moving and continuously speaking with short pauses. The spectrograms of the two sources recorded with one microphone are illustrated in Fig. 12.

In this experiment, we just consider the azimuth angle estimation with the elevation angle fixed at 90°. The target plane is uniformly separated into 73 grid points from −180° to 180° with a uniform interval of 5°. The true positions and sound source signals of two sources are provided by the LOCATA dataset. We applied a voice activity detector [52] to these source signals to obtain ground-truth voice activity information of the two sound sources. Figure 13a shows the true trajectories of the two sources. We also applied the voice activity detector to the microphone array signals to obtain the voice activity information of each frame. Similar to the simulation part, we find two peaks for each voice active frame and label these peaks according to the true source position. Note that a threshold δ is set to judge the existence of peaks, i.e, if the amplitude of peaks is less than δ, this estimated peak is considered as an invalid estimate. The black circles and red circles denote the true DOAs of the first and second sources, respectively. The magenta triangles denote the false estimates.

The estimation results of the CBF, MVDR, SRP-PHAT, and MSBL-EM methods are shown in Figs. 13b, c, d, and e, respectively. Moreover, the estimation results of the proposed SAVE-MSBL-EM method is shown in Fig. 13f. From Figs. 13b–d, it can be seen that the two sources can hardly be separated in the time range from 6 to 10 s using the CBF, SRP-PHAT, and MVDR methods. However, the proposed SAVE-MSBL-EM method can separate two sources successfully, indicating a higher resolution than the CBF, SRP-PHAT, and MVDR methods (see Fig. 13f). Comparing Fig. 13e and f, it can be seen that the proposed SAVE-MSBL-EM method achieves better recovery performance than MSBL-EM method in the time range

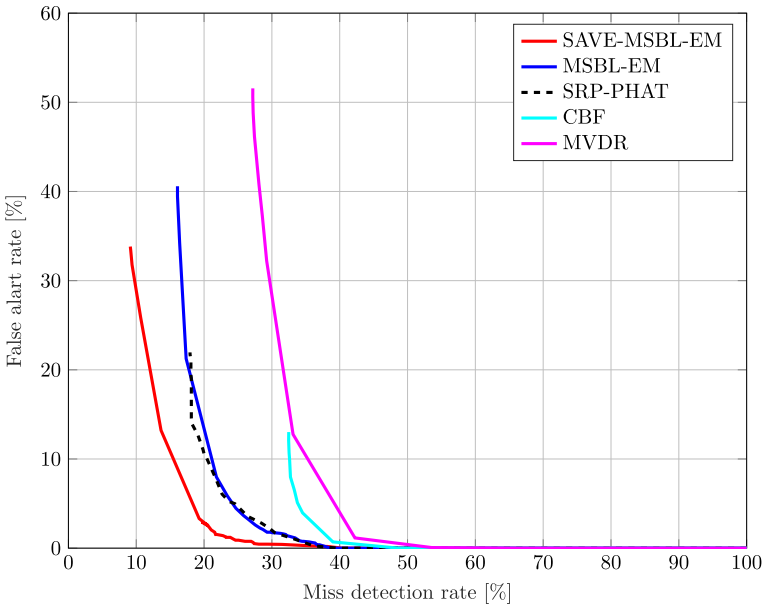

Fig. 14 The MD rate versus FA rate by varying the peak selection threshold

Table 5 Results for the real data

	FA [%]	MD [%]	RMSE [°]
CBF	12.7	32.5	3.4
SRP-PHAT	10.1	20.3	3.2
MVDR	51.4.	27.2	3.7
MSBL	1.7	30.1	3.1
SAVE-MSBL-EM	0.4	29.6	2.9

from 8 to 10 s. To evaluate the performance of all the methods, the MD rate versus FA rate is computed by varying the peak selection threshold (see Fig. 14). For all the curves in Fig. 14, the closer to the left-bottom the better. It can be seen that the proposed SAVE-MSBL-EM method achieves better performance than state-of-the-art methods.

We further report the estimation result for a fixed peak selection threshold $\delta = -40$ dB (see Table 5). It can be seen that the proposed SAVE-MSBL-EM method outperforms other methods especially for the FA rate and RMSE. The reason is that the proposed method successfully resolves the two sources while the others are failing in the range from 6 to 10 s. The results indicate that the proposed SAVE-MSBL-EM method provides a higher resolution performance than state-of-the-art methods also in real conditions where all assumptions of the proposed method might not hold.

5 Conclusion

In this paper, we propose a space alternating MSBL method for acoustic DOA estimation that offers a high-resolution performance. First, we build a group sparse prior based hierarchical Bayesian framework for the MMV signal model by exploiting the group sparsity of candidate source amplitude matrix. Then, the computational efficient SAVE-MSBL algorithm is proposed to infer all hidden variables in the Bayesian model. Moreover, an EM algorithm is proposed to deal with the acoustic DOA estimation problem. In the experimental parts, the performance of the proposed method is investigated using both synthetic and real data. The results show that the proposed method has lower RMSE and FA rate than state-of-the-art methods in both free-field and low-reverberation conditions. As a result, the proposed method can be applied to some applications (e.g., humanoid robots and drones) to improve the resolution performance for acoustic DOA estimation.

Appendix A: Derivation of (11)

According to Eq. 9 and Eq. 10, the signal $\boldsymbol{s}_k$ can be updated using the space alternating approach as follows:

$$
\begin{aligned}
\ln q(\boldsymbol{s}_k) &= \mathrm{E}_{q(\boldsymbol{\Theta}/\boldsymbol{s}_k)}\left[p(\boldsymbol{X},\boldsymbol{\Theta})\right] \\
&= \mathrm{E}_{q(\boldsymbol{\Theta}/\boldsymbol{s}_k)}\left[-\rho\|\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S}\|_F^2 - \sum_{k=1}^{K}\lambda_k \boldsymbol{s}_k^{\mathrm{H}}\boldsymbol{s}_k\right] \\
&= -\mathrm{E}_{q(\boldsymbol{\Theta}/\boldsymbol{s}_k)}\left[\mathrm{tr}\left[\rho\,(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S})^{\mathrm{H}}\right.\right. \\
&\quad \left.\left.\times(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S}) + \lambda_k \boldsymbol{s}_k^{\mathrm{H}}\boldsymbol{s}_k\right]\right] + \mathrm{C} \\
&= -\mathrm{E}_{q(\boldsymbol{\Theta}/\boldsymbol{s}_k)}\left[\mathrm{tr}\left[\rho\left(\boldsymbol{X}-\boldsymbol{A}_{\bar{k}}\boldsymbol{S}_{\bar{k}}-\boldsymbol{a}_k\boldsymbol{s}_k^{\mathrm{T}}\right)^{\mathrm{H}}\right.\right. \\
&\quad \left.\left.\times\left(\boldsymbol{X}-\boldsymbol{A}_{\bar{k}}\boldsymbol{S}_{\bar{k}}-\boldsymbol{a}_k\boldsymbol{s}_k^{\mathrm{T}}\right)+\lambda_k\boldsymbol{s}_k^{\mathrm{H}}\boldsymbol{s}_k\right]\right] + \mathrm{C} \\
&= -\mathrm{E}_{q(\boldsymbol{\Theta}/\boldsymbol{s}_k)}\left[\mathrm{tr}\left[\boldsymbol{s}_k^{\mathrm{H}}\left(\rho\boldsymbol{a}_k^{\mathrm{H}}\boldsymbol{a}_k+\lambda_k\right)\boldsymbol{s}_k - \rho\boldsymbol{s}_k^{*}\boldsymbol{a}_k^{\mathrm{H}}\right.\right. \\
&\quad \left.\left.\times\left(\boldsymbol{X}-\boldsymbol{A}_{\bar{k}}\boldsymbol{S}_{\bar{k}}\right) - \rho\left(\boldsymbol{X}-\boldsymbol{A}_{\bar{k}}\boldsymbol{S}_{\bar{k}}\right)^{\mathrm{H}}\boldsymbol{a}_k\boldsymbol{s}_k^{\mathrm{T}}\right]\right] + \mathrm{C} \\
&= -\mathrm{tr}\left[\boldsymbol{s}_k^{\mathrm{H}}\left(\langle\rho\rangle\,\boldsymbol{a}_k^{\mathrm{H}}\boldsymbol{a}_k+\langle\lambda_k\rangle\right)\boldsymbol{s}_k + \langle\rho\rangle\,\boldsymbol{s}_k^{*}\boldsymbol{a}_k^{\mathrm{H}}\left(\boldsymbol{X}-\boldsymbol{A}_{\bar{k}}\langle\boldsymbol{S}_{\bar{k}}\rangle\right)\right. \\
&\quad \left.-\langle\rho\rangle\left(\boldsymbol{X}-\boldsymbol{A}_{\bar{k}}\times\langle\boldsymbol{S}_{\bar{k}}\rangle\right)^{\mathrm{H}}\boldsymbol{a}_k\boldsymbol{s}_k^{\mathrm{T}}\right] + \mathrm{C},
\end{aligned}
$$

where $\boldsymbol{\Theta}/\boldsymbol{s}_k$ denotes the set of variables with $\boldsymbol{s}_k$ removed, C denotes a constant. Note that $\boldsymbol{A}\boldsymbol{S}$ can be rewritten as $\boldsymbol{A}_{\bar{k}}\boldsymbol{S}_{\bar{k}} + \boldsymbol{a}_k\boldsymbol{s}_k^{\mathrm{T}}$.

Abbreviations

DOA: Direction-of-arrival; CBF: Classical beamforming; SRP-PHAT: Steered-response power phase transform; MVDR: Minimum variance distortionless response; MUSIC: Multiple signal classification; ESPRIT: Estimation of signal parameters via rotational invariance technique; SNR: Signal-to-noise ratio; SVD: Singular value decomposition; cLASSO: Complex least absolute shrinkage and selection operator; SBL: Sparse Bayesian learning ; MSBL: Multi-snapshot sparse Bayesian learning; SAVE: Space alternating variational estimation; VBI: Variational Bayesian inference; MMSE: Minimum mean square error; RMSE: Root-mean-square-error; MMV: Multiple measurement vector; CGMM: Complex Gaussian mixture model; EM: Expectation–maximization; ULA: Uniform linear array; UCA: Uniform circular array; KL: Kullback-Leibler; FFT: Fourier transform; FA: False alarm; MD: Miss-detected

Acknowledgements

The authors would like to thank Zhilin Zhang for providing the source code of the MSBL approach.

Authors' contributions

Z. Bai and L. Shi conceptualized the study and run the experiments. M. G. Christensen, J. R. Jensen, and J. Sun edited the manuscript. All the authors read and approved the final manuscript.

Funding

This work was supported by the China Scholarship Council, grant ID.201806120176.

Availability of data and materials

The software for microphone array data generation is from "International Audio Laboratories Erlangen" and is online available: https://www.audiolabs-erlangen.de/home. The LOCATA data originates from the"IEEE-AASP Challenge on Acoustic Source Localization and Trackin" and can be found under the following link: https://www.locata.lms.tf.fau.de/.

Competing interests

The authors declare that they have no competing interests.

Received: 14 September 2020 Accepted: 27 January 2021
Published online: 06 April 2021

References

1. J. Hornstein, M. Lopes, J. Santos-Victor, F. Lacerda, in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Sound localization for humanoid robots - building audio-motor maps based on the HRTF (IEEE, Beijing, 2006), pp. 1170–1176
2. C. Rascon, I. Meza, Localization of sound sources in robotics: a review. Robot. Auton. Syst. **96**, 184–210 (2017)
3. M. Strauss, P. Mordel, V. Miguet, A. Deleforge, in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. DREGON: dataset and methods for UAV-embedded sound source localization (IEEE, Madrid, 2018)
4. A. Deleforge, D. D. Carlo, M. Strauss, R. Serizel, L. Marcenaro, Audio-based search and rescue with a drone: highlights from the IEEE signal processing cup 2019 student competition. IEEE Signal Proc. Mag. **36**(5), 138–144 (2019)
5. J. M. Valin, F. Michaud, J. Rouat, in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*. Robust 3D localization and tracking of sound sources using beamforming and particle filtering (IEEE, Toulouse, 2006), pp. 841–844
6. C. Zhang, D. Florencio, D. E. Ba, Z. Zhang, Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. IEEE Trans. Multimed. **10**(3), 538–548 (2008)
7. M. Farmani, M. S. Pedersen, Z.-H. Tan, J. Jensen, Informed sound source localization using relative transfer functions for hearing aid applications. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(3), 611–623 (2017)
8. H. L. Van Trees. One, in *Part IV of Detection, Estimation, and Modulation Theory*. Optimum array processing (John Wiley and Sons, New York, 2004), pp. 21–53
9. J. H. DiBiase, H. F. Silverman, M. S. Brandstein, in *Microphone arrays*. Robust localization in reverberant rooms (Springer, Berlin, Heidelberg, 2001), pp. 164–180
10. V. Krishnaveni, T. Kesavamurthy, A. B, Beamforming for direction-of-arrival (DOA) estimation-a survey. Int. J. Comput. Appl. **61**(11), 4–11 (2013)
11. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
12. R. Roy, T. Kailath, ESPRIT-estimation of signal parameters via rotational invariance techniques. IEEE Trans. Acoustics Speech Sig. Process. **37**(7), 984–995 (1989)
13. H. Cox, R. Zeskind, M. Owen, Robust adaptive beamforming. IEEE Trans Acoustics Speech Sig. Process. **35**(10), 1365–1376 (1987)
14. D. D. Feldman, L. J. Griffiths, A projection approach for robust adaptive beamforming. IEEE Trans Sig. Process. **42**(4), 867–876 (1994)
15. M. Pardini, F. Lombardini, F. Gini, The hybrid Cramér–Rao bound on broadside DOA estimation of extended sources in presence of array errors. IEEE Trans Sig. Process. **56**(4), 1726–1730 (2008)
16. A. Khabbazibasmenj, S. A. Vorobyov, A. Hassanien, Robust adaptive beamforming based on steering vector estimation with as little as possible prior information. IEEE Trans Sig. Process. **60**(6), 2974–2987 (2012)
17. A. L. Kintz, I. J. Gupta, A modified MUSIC algorithm for direction of arrival estimation in the presence of antenna array manifold mismatch. IEEE Trans. Antennas Propag. **64**(11), 4836–4847 (2016)
18. D. Malioutov, M. Cetin, A. S. Willsky, A sparse signal reconstruction perspective for source localization with sensor arrays. IEEE Trans. Sig. Process. **53**(8), 3010–3022 (2005)
19. D. P. Wipf, B. D. Rao, An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. IEEE Trans. Sig. Process. **55**(7), 3704–3716 (2007)
20. S. Fortunati, R. Grasso, F. Gini, M. S. Greco, K. LePage, Single-snapshot DOA estimation by using compressed sensing. EURASIP J. Adv. Sig. Process. **2014**(1), 1-17 (2014)
21. P. Gerstoft, C. F. Mecklenbrauker, A. Xenaki, S. Nannuru, Multisnapshot sparse Bayesian learning for DOA. IEEE Sig. Process. Lett. **23**(10), 1469–1473 (2016)
22. A. Xenakia, J. B. Boldt, M. G. Christensen, Sound source localization and speech enhancement with sparse Bayesian learning beamforming. J. Acoust. Soc. Am. **143**(6), 3912–3921 (2018)
23. A. Xenaki, P. Gerstoft, K. Mosegaard, Compressive beamforming. J. Acoust. Soc. Am. **136**(1), 260–271 (2014)
24. C. F. Mecklenbräuker, P. Gerstoft, E. Zöchmann, c–LASSO and its dual for sparse signal estimation from array data. Sig. Process. **130**, 204–216 (2017)
25. X. Wang, D. Meng, M. Huang, L. Wan, Reweighted regularized sparse recovery for DOA estimation with unknown mutual coupling. IEEE Commun. Lett. **23**(2), 290–293 (2019)
26. Z. Yang, J. Li, P. Stoica, L. Xie, C. Rama, T. Sergios. One, in *Academic Press Library in Signal Processing*. Sparse methods for direction-of-arrival estimation, vol. 7, (New York, 2018), pp. 509–581
27. M. E. Tipping, A. Smola, Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. **59**(1), 211–244 (2001)
28. S. Ji, Y. Xue, L. Carin, Bayesian compressive sensing. IEEE Trans. Sig. Process. **56**(6), 2346–2356 (2008)
29. S. D. Babacan, R. Molina, A. K. Katsaggelos, Bayesian compressive sensing using laplace priors. IEEE Trans. Image Process. **19**(1), 53–63 (2010)
30. B. Worley, Scalable mean-field sparse bayesian learning. IEEE Trans. Sig. Process. **67**(24), 6314–6326 (2019)
31. D. Wipf, S. Nagarajan, in *Proceedings of the 24th International Conference on Machine Learning - ICML 07*. Beamforming using the relevance vector machine (ACM Press, New York, USA, 2007), pp. 1–8
32. Z. Yang, L. Xie, C. Zhang, Off-grid direction of arrival estimation using sparse Bayesian inference. IEEE Trans. Sig. Process. **61**(1), 38–43 (2013)
33. L. Zhao, X. Li, L. Wang, G. Bi, Computationally efficient wide-band DOA estimation methods based on sparse Bayesian framework. IEEE Trans. Veh. Technol. **66**(12), 11108–11121 (2017)
34. Z. Bai, J. Sun, J. R. Jensen, M. G. Christensen, in *2019 27th European Signal Processing Conference (EUSIPCO)*. Indoor sound source localization based on sparse Bayesian learning and compressed data (IEEE, A Coruna, Spain, 2019), pp. 1–5
35. Z. Bai, J. R. Jensen, J. Sun, M. G. Christensen, in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. A sparse Bayesian learning based RIR reconstruction method for acoustic TOA and DOA estimation (IEEE, New York, USA, 2019), pp. 1–5
36. M. E. Tipping, A. Faul, J. J. T. Avenue, J. J. T. Avenue, in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Fast marginal likelihood maximisation for sparse Bayesian models (JMLR, Key West, 2003), pp. 3–6
37. H. Duan, L. Yang, J. Fang, H. Li, Fast inverse-free sparse Bayesian learning via relaxed evidence lower bound maximization. IEEE Sig. Process. Lett. **24**(6), 774–778 (2017)
38. C. K. Thomas, D. Slock, in *2018 26th European Signal Processing Conference (EUSIPCO)*. Space alternating variational Bayesian learning for LMMSE filtering (IEEE, Rome, Italy, 2018), pp. 1–5
39. D. P. Wipf, B. D. Rao, Sparse Bayesian learning for basis selection. IEEE Trans. Sig. Process. **52**(8), 2153–2164 (2004)
40. Z. Zhang, B. D. Rao, Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning. IEEE J. Sel. Top. Sig. Process. **5**(5), 912–926 (2011)
41. J. Huang, T. Zhang, The benefit of group sparsity. Ann. Stat. **38**(4), 1978–2004 (2010). https://doi.org/10.1214/09-aos778
42. C. M. Bishop, in *Pattern recognition and machine learning*. Approximate inference (Springer, New York, 2006), pp. 472–485
43. D. G. Tzikas, A. C. Likas, N. P. Galatsanos, The variational approximation for Bayesian inference. IEEE Sig. Process. Mag. **25**(6), 131–146 (2008)
44. J. A. Fessler, A. O. Hero, Space-alternating generalized expectation-maximization algorithm. IEEE Trans. Sig. Process. **42**(10), 2664–2677 (1994)
45. Y. Dorfan, S. Gannot, Tree-based recursive expectation-maximization algorithm for localization of acoustic sources. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(10), 1692–1703 (2015)
46. X. Li, Y. Ban, L. Girin, A. P. Xavier, R. Horaud, Online localization and tracking of multiple moving speakers in reverberant environments. IEEE J. Sel. Top. Sig. Process. **13**(1), 88–103 (2019)
47. S. T. Birchfield, D. K. Gillmor, in *IEEE International Conference on Acoustics Speech and Signal Processing*. Fast Bayesian acoustic localization (IEEE, Palo Alto, California, 2002), pp. 1–4
48. J. Traa, D. Wingate, N. D. Stein, P. Smaragdis, Robust source localization and enhancement with a probabilistic steered response power model. IEEE/ACM Trans. Audio Speech. Lang. Process. **24**(3), 493–503 (2016)
49. R. D. Nowak, Distributed EM algorithms for density estimation and clustering in sensor networks. IEEE Trans. Sig. Process. **51**(8), 2245–2253 (2003)
50. Y. Dorfan, G. Hazan, S. Gannot, in *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. Multiple acoustic sources localization using distributed expectation-maximization algorithm (IEEE, Villers-les-Nancy, France, 2014), pp. 1–5

51. H. W. Lollmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, W. Kellermann, in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. The LOCATA challenge data corpus for acoustic source localization and tracking (IEEE, Sheffield, 2018), pp. 410–414
52. J. Sohn, N. S. Kim, W. Sung, A statistical model-based voice activity detection. IEEE Sig. Process. Lett. **6**(1), 1–3 (1999)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com