Signal Processing 161 (2019) 155–164

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Grid-less variational Bayesian line spectral estimation with multiple measurement vectors

Jiang Zhu[a,*], Qi Zhang[a], Peter Gerstoft[b], Mihai-Alin Badiu[c,d], Zhiwei Xu[a]

[a] Ocean College, Zhejiang University, Zhoushan, China
[b] Electrical and Computer Engineering, University of California, San Diego, USA
[c] Department of Engineering Science, University of Oxford, UK
[d] Department of Electronic Systems, Aalborg University, Denmark

ARTICLE INFO

Article history:
Received 29 November 2018
Revised 28 March 2019
Accepted 29 March 2019
Available online 30 March 2019

Keywords:
Variational Bayesian inference
Multiple snapshot
Line spectral estimation
Von mises distribution
Grid-less
Sequential estimation

ABSTRACT

Line spectral estimation (LSE) with multiple measurement vector (MMV) is studied utilizing the Bayesian variational inference. Motivated by the recent grid-less variational line spectral estimation (VALSE) method, we develop the MMV VALSE (MVALSE). The MVALSE shares the advantages of the VALSE method, such as automatically estimating the model order, noise variance, weight variance, and providing the uncertainty of the frequency estimates. The MVALSE can be viewed as applying the VALSE with single measurement vector to each snapshot, and combining the intermediate data appropriately. Furthermore, the MVALSE is developed to perform sequential estimation. Numerical results demonstrate the effectiveness of the MVALSE method, compared to the state-of-the-art MMV methods.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Line spectral estimation (LSE), i.e., recovering the parameters of a superposition of complex exponential functions is one of the classical problems in signal processing fields [1], which has many applications such as channel estimation in wireless communications [2], direction of arrival estimation in radar systems [3], speech analysis and so on. Traditional methods for solving the LSE problem include periodogram, MUSIC, ESPRIT and maximum likelihood (ML) method [1].

In the past decades, sparse methods for LSE have been popular due to the development of sparse signal representation and compressed sensing theory, which can be clarified into on-grid, off-grid and grid-less methods [4]:

- On-grid: The on-grid method assumes that the frequencies locate on the grids. By discretizing the continuous frequency into a finite set of grid points, the nonlinear problem can be formulated as a linear problem. ℓ_1 optimization [5], sparse iterative covariance-based estimation (SPICE) [6–8], sparse Bayesian learning (SBL) [9] are main sparse methods. Compared to classical methods, the on-grid methods perform better by utilizing the sparsity in the frequency domain. However, the frequencies might not lie on the grids. Thus on-grid methods suffer from grid mismatch and spectral leakage.
- Off-grid: To overcome the grid mismatch problem, off-grid methods gradually refine the dynamic dictionary. In [10], a Newtonalized orthogonal matching pursuit (NOMP) method is proposed, where a Newton step and feedback are utilized to refine the frequency estimation. In addition, the NOMP algorithm is also extended to deal with multiple measurement vector (MMV) [11]. In [12], the iterative reweighted approach (IRA) via majorization-minimization (MM) is proposed, which effectively overcomes the grid mismatch problem and achieves a super-resolution accuracy. Later in [13], a prior knowledge of the frequency distribution aided iterative reweighted algorithm is proposed. A different off-grid approach is based on the Bayesian framework and SBL [16,17] is adopted. In [14], an SBL based method is proposed which jointly estimates the grid and grid bias, while in [15], a Newton method is proposed to refine the frequency estimates. In [18], variational inference method is proposed. In [19], maximization of the marginalized posterior probability density function (PDF) is performed. For all these approaches, only point estimates of the frequency are computed in each iteration, which is similar to the classical

[*] Corresponding author.
E-mail addresses: jiangzhu16@zju.edu.cn (J. Zhu), zhangqi13@zju.edu.cn (Q. Zhang), pgerstoft@ucsd.edu (P. Gerstoft), mib@es.aau.dk (M.-A. Badiu), xuzw@zju.edu.cn (Z. Xu).

https://doi.org/10.1016/j.sigpro.2019.03.024
0165-1684/© 2019 Elsevier B.V. All rights reserved.

ML methods. Another limitation is that these methods usually overestimates the model order [18]. In [20], a low complexity superfast LSE methods are proposed based on fast Toeplitz matrix inversion algorithm.

- Grid-less: To completely overcome the grid mismatch problem, grid-less methods which work directly with continuously parameterized dictionaries was proposed [21–28]. For the SMV case, the atomic norm based method has been proposed in the noiseless case [21]. In [22,23], the atomic soft thresholding (AST) method is proposed in the noisy case. Since AST method requires knowledge of the noise variance, the gridless SPICE (GLS) method is proposed without knowledge of noise power [23]. In [24], an exact discretization-free method called sparse and parametric approach (SPA) is proposed for uniform and sparse linear arrays, which is based on the well-established covariance fitting criterion. In [26], two approaches based on atomic norm minimization and structured covariance estimation are developed in the MMV case, and the benefit of including MMV is demonstrated. To further improve the resolution of the atomic norm based methods, enhanced matrix completion (EMac) [29] and reweighted atomic-norm minimization (RAM) [30] are proposed and the resolution capability is improved numerically. These grid-less based methods involve solving a semidefinite programming (SDP) problem [31], whose computation complexity is prohibitively high for large-scale problems. In [32], by treating the frequencies as random variables, a grid-less variational line spectral estimation (VALSE) algorithm is proposed. This work is closely related with [32], and details are introduced in the ensuing subsection.

1.1. Main contributions and comparisons to related work

In [32], a grid-less variational line spectral estimation algorithm is proposed, where PDFs of the frequencies are estimated, instead of retaining only the point estimates of the frequencies. This more complete Bayesian approach allows to represent and operate with the frequency uncertainty, in addition to only that of the weights. We rigorously develop the variational Bayesian inference method for LSE in the MMV setting, which is especially important in array signal processing. Meanwhile, the derived MVALSE reveals close relationship to the VALSE algorithm, which is suitable for parallel processing. The prior information may be given from past experience, and is particularly useful for low SNR or few samples are available [33]. For sequential estimation, the output of the PDF of the frequencies from the previous observations can be employed as the prior of the frequency, and sequential MVALSE (Seq-MVALSE) is proposed. Substantial experiments are conducted to illustrate the competitive performance of the MVALSE method and its application to DOA problems, compared to other sparse based approaches.

1.2. Notation

Let $\mathcal{S}\subset\{1,\cdots,N\}$ be a subset of indices and $|\mathcal{S}|$ denote its cardinality. For a matrix $\mathbf{A}\in\mathbb{C}^{M\times N}$, let $\mathbf{A}_{:,\mathcal{S}}$ denote the submatrix with the columns indexed by $\mathcal{S}$, let $\mathbf{a}_i^{\rm T}$ denote the ith row of $\mathbf{A}$ and $\mathbf{A}_{\mathcal{S},:}$ denote the submatrix with the rows of $\mathbf{A}$ indexed by $\mathcal{S}$. For a matrix $\mathbf{J}\in\mathbb{C}^{N\times N}$, let $\mathbf{J}_{\mathcal{S},\mathcal{S}}$ denote the submatrix by choosing both the rows and columns of $\mathbf{J}$ indexed by $\mathcal{S}$. Let $(\cdot)_{\mathcal{S}}^*$, $(\cdot)_{\mathcal{S}}^{\rm T}$ and $(\cdot)_{\mathcal{S}}^{\rm H}$ be the conjugate, transpose and Hermitian transpose operator of $(\cdot)_{\mathcal{S}}$, respectively. Let $\mathbf{I}_L$ denote the identity matrix of dimension L. Let $||\cdot||_{\rm F}$ denote the Frobenius norm. For a vector $\mathbf{x}$, let $\|\mathbf{x}\|_0$ denote the number of nonzero elements, and sometimes we let $[\mathbf{x}]_i$ or x_i denote its ith element. Similarly, let $[\mathbf{B}]_{i,j}$ or B_{ij} denote the (i, j)th element of $\mathbf{B}$, and let $\mathbf{B}_{i,:}$ and $\mathbf{B}_{:,j}$ denote the ith row and jth column of $\mathbf{B}$, respectively. Let Re{ · } return the real part. Let $\mathcal{CN}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ denote the complex normal distribution of $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and let $\mathcal{VM}(\theta,\mu,\kappa)$ denote the von Mises distribution of θ with mean direction μ and concentration parameter κ.

2. Problem setup

For the line spectral model with L snapshots, the measurements $\mathbf{Y}\in\mathbb{C}^{M\times L}$ consist of a superposition of K complex sinusoids corrupted by the additive white Gaussian noise (AWGN) $\mathbf{U}$:

$$\mathbf{Y}=\sum_{k=1}^{K}\mathbf{a}(\widetilde{\theta}_k)\widetilde{\mathbf{w}}_k^{\rm T}+\mathbf{U},\tag{1}$$

where M is the number of measurements. The complex weights over the L snapshots and the frequency of the kth component are $\widetilde{\mathbf{w}}_k\in\mathbb{C}^{L\times 1}$ and respectively $\widetilde{\theta}_k\in[-\pi,\pi)$. The elements of the noise $\mathbf{U}\in\mathbb{C}^{M\times L}$ are i.i.d. and $U_{ij}\sim\mathcal{CN}(U_{ij};0,\nu)$, and $\mathbf{a}(\widetilde{\theta}_k)=[1,{\rm e}^{{\rm j}\widetilde{\theta}_k},\cdots,{\rm e}^{{\rm j}(M-1)\widetilde{\theta}_k}]^{\rm T}$.

Since the number of complex sinusoids K is generally unknown, the measurements $\mathbf{Y}$ is assumed to consist of a superposition of known N components with $N>K$ [32], i.e.,

$$\mathbf{Y}=\sum_{i=1}^{N}\mathbf{a}(\theta_i)\mathbf{w}_i^{\rm T}+\mathbf{U}=\mathbf{A}\mathbf{W}+\mathbf{U},\tag{2}$$

where $\mathbf{A}=[\mathbf{a}(\theta_1),\cdots,\mathbf{a}(\theta_N)]\in\mathbb{C}^{M\times N}$, $\mathbf{a}(\theta_i)$ denotes the ith column of $\mathbf{A}$, $\mathbf{w}_i^{\rm T}$ denote the ith row of $\mathbf{W}\in\mathbb{C}^{N\times L}$. Since $N>K$, binary hidden variables $\mathbf{s}=[s_1,\ldots,s_N]^{\rm T}$ are introduced with probability mass function $p(\mathbf{s};\lambda)=\prod_{i=1}^{N}p(s_i;\lambda)$, where $s_i\in\{0,1\}$ and

$$p(s_i;\lambda)=\lambda^{s_i}(1-\lambda)^{(1-s_i)}.\tag{3}$$

We assume $p(\mathbf{W}|\mathbf{s};\tau)=\prod_{i=1}^{N}p(\mathbf{w}_i|s_i;\tau)$, where $p(\mathbf{w}_i|s_i;\ \tau)$ is Bernoulli-Gaussian distributed

$$p(\mathbf{w}_i|s_i;\tau)=(1-s_i)\delta(\mathbf{w}_i)+s_i\mathcal{CN}(\mathbf{w}_i;\mathbf{0},\tau\mathbf{I}_L),\tag{4}$$

where $\delta(\cdot)$ is the Dirac delta distribution and τ is the variance of the elements of $\mathbf{w}_i$ corresponding to the active component. Eqs. (3) and (4) show that λ controls the probability of the ith component to be active, i.e., $p(s_i=1)=\lambda$. The prior distribution $p(\boldsymbol{\theta})$ of the frequency $\boldsymbol{\theta}=[\theta_1,\ldots,\theta_N]^{\rm T}$ is $p(\boldsymbol{\theta})=\prod_{i=1}^{N}p(\theta_i)$, where $p(\theta_i)$ is encoded through the von Mises distribution [34], p. 36]

$$p(\theta_i)=\mathcal{VM}(\theta_i;\mu_{0,i},\kappa_{0,i})=\frac{1}{2\pi I_0(\kappa_{0,i})}{\rm e}^{\kappa_{0,i}\cos(\theta-\mu_{0,i})},\tag{5}$$

where $\mu_{0,i}$ and $\kappa_{0,i}$ are the mean direction and concentration parameters of the prior of the ith frequency θ_i, $I_p(\cdot)$ is the modified Bessel function of the first kind and the order p [34], p. 348]. Note that $\kappa_{0,i}=0$ corresponds to the uninformative prior distribution $p(\theta_i)=1/(2\pi)$ [32].

For measurement model (2), the likelihood $p(\mathbf{Y}|\mathbf{A}\mathbf{W};\nu)$ is

$$p(\mathbf{Y}|\mathbf{A}\mathbf{W};\nu)=\prod_{i,j}\mathcal{CN}(Y_{ij};[\mathbf{A}\mathbf{W}]_{i,j},\nu).\tag{6}$$

Let $\boldsymbol{\beta}=\{\nu,\ \lambda,\ \tau\}$ and $\boldsymbol{\Phi}=\{\boldsymbol{\theta},\mathbf{W},\mathbf{s}\}$ be the model and estimated parameters. Given the above statistical model, the type II maximum likelihood (ML) estimation of the model parameters $\hat{\boldsymbol{\beta}}_{\rm ML}$ is

$$\hat{\boldsymbol{\beta}}_{\rm ML}=\underset{\boldsymbol{\beta}}{\rm argmax}\int p(\mathbf{Y},\boldsymbol{\Phi};\boldsymbol{\beta}){\rm d}\mathbf{s}{\rm d}\mathbf{W}{\rm d}\boldsymbol{\theta},\tag{7}$$

where $p(\mathbf{Y},\boldsymbol{\Phi};\boldsymbol{\beta})\propto p(\mathbf{Y}|\mathbf{A}\mathbf{W};\nu)\prod_{i=1}^{N}p(\theta_i)p(\mathbf{w}_i|s_i;\tau)p(s_i;\lambda)$. Then the minimum mean square error (MMSE) estimate $\boldsymbol{\Phi}_{\rm MMSE}$ of the parameters $\mathbf{\Phi}$ is

$$\widehat{\mathbf{\Phi}}_{\rm MMSE} = {\rm E}[\mathbf{\Phi}|\mathbf{Y};\widehat{\boldsymbol{\beta}}_{\rm ML}], \tag{8}$$

where the expectation is taken with respect to the PDF

$$p(\mathbf{\Phi}|\mathbf{Y};\widehat{\boldsymbol{\beta}}_{\rm ML}) \propto p(\mathbf{Y}|\mathbf{AW};\widehat{\nu}_{\rm ML})\prod_{i=1}^{N}p(\theta_i)p(\mathbf{w}_i|s_i;\widehat{\tau}_{\rm ML})p(s_i;\widehat{\lambda}_{\rm ML}). \tag{9}$$

However, computing both the ML estimate of $\boldsymbol{\beta}$ (7) and the MMSE estimate of $\mathbf{\Phi}$ (8) are intractable. Thus an iterative algorithm is designed in the following.

3. MVALSE Algorithm

In this section, a mean field variational Bayes method is proposed to find an approximate PDF $q(\mathbf{\Phi}|\mathbf{Y})$ by minimizing the Kullback-Leibler (KL) divergence ${\rm KL}(q(\mathbf{\Phi}|\mathbf{Y})||p(\mathbf{\Phi}|\mathbf{Y}))$ [35], p. 732]

$${\rm KL}(q(\mathbf{\Phi}|\mathbf{Y})||p(\mathbf{\Phi}|\mathbf{Y})) = \int q(\mathbf{\Phi}|\mathbf{Y})\ln\frac{q(\mathbf{\Phi}|\mathbf{Y})}{p(\mathbf{\Phi}|\mathbf{Y})}{\rm d}\boldsymbol{\theta}{\rm d}\mathbf{W}{\rm d}\mathbf{s}. \tag{10}$$

For any assumed PDF $q(\mathbf{\Phi}|\mathbf{Y})$, the log marginal likelihood (model evidence) $\ln p(\mathbf{Y};\boldsymbol{\beta})$ is [35], pp. 732-733]

$$\ln p(\mathbf{Y};\boldsymbol{\beta}) = {\rm KL}(q(\mathbf{\Phi}|\mathbf{Y})||p(\mathbf{\Phi}|\mathbf{Y})) + \mathcal{L}(q(\mathbf{\Phi}|\mathbf{Y});\boldsymbol{\beta}), \tag{11}$$

where

$$\mathcal{L}(q(\mathbf{\Phi}|\mathbf{Y});\boldsymbol{\beta}) = {\rm E}_{q(\mathbf{\Phi}|\mathbf{Y})}\left[\ln\frac{p(\mathbf{Y},\mathbf{\Phi};\boldsymbol{\beta})}{q(\mathbf{\Phi}|\mathbf{Y})}\right]. \tag{12}$$

For a given data $\mathbf{Y}$, $\ln p(\mathbf{Y};\boldsymbol{\beta})$ is a constant, thus minimizing the KL divergence is equivalent to maximizing $\mathcal{L}(q(\mathbf{\Phi}|\mathbf{Y});\boldsymbol{\beta})$ in (11). Therefore we maximize $\mathcal{L}(q(\mathbf{\Phi}|\mathbf{Y});\boldsymbol{\beta})$ in the sequel.

For the factored PDF $q(\mathbf{\Phi}|\mathbf{Y})$, the following assumptions are made:

- Given $\mathbf{Y}$, the frequencies $\{\theta_i\}_{i=1}^N$ are mutually independent.
- The posterior of the binary hidden variables $q(\mathbf{s}|\mathbf{Y})$ has all its mass at $\widehat{\mathbf{s}}$, i.e., $q(\mathbf{s}|\mathbf{Y}) = \delta(\mathbf{s}-\widehat{\mathbf{s}})$.
- Given $\mathbf{Y}$ and $\mathbf{s}$, the frequencies and weights are independent.

As a result, $q(\mathbf{\Phi}|\mathbf{Y})$ can be factored as

$$q(\mathbf{\Phi}|\mathbf{Y}) = \prod_{i=1}^{N}q(\theta_i|\mathbf{Y})q(\mathbf{W}|\mathbf{Y},\mathbf{s})\delta(\mathbf{s}-\widehat{\mathbf{s}}). \tag{13}$$

Due to the factorization property of (13), the frequencies $\boldsymbol{\theta}$ can be estimated from the marginal distribution $q(\mathbf{\Phi}|\mathbf{Y})$ as [34], pp. 26]

$$\widehat{\theta}_i = \arg({\rm E}_{q(\theta_i|\mathbf{Y})}[{\rm e}^{{\rm j}\theta_i}]), \tag{14a}$$

$$\widehat{\mathbf{a}}_i = {\rm E}_{q(\theta_i|\mathbf{Y})}[\mathbf{a}(\theta_i)],\ i\in\{1,\ldots,N\}, \tag{14b}$$

where $\arg(\cdot)$ returns the angle. In Section 3.1, $q(\theta_i|\mathbf{Y})$ is approximated as a von Mises distribution. For von Mises distribution $\mathcal{VM}(\theta;\mu,\kappa)$ (5), $\arg({\rm E}_{\mathcal{VM}(\theta;\mu,\kappa)}[{\rm e}^{{\rm j}\theta}]) = \arg({\rm e}^{{\rm j}\mu}\frac{I_1(\kappa)}{I_0(\kappa)}) = \mu = {\rm E}_{\mathcal{VM}(\theta;\mu,\kappa)}[\theta]$. Therefore, $\widehat{\theta}_i$ is also the mean direction of θ for von Mises distribution. Besides, ${\rm E}[{\rm e}^{{\rm j}m\theta}] = {\rm e}^{{\rm j}m\mu}I_m(\kappa)/I_0(\kappa)$[1]

Given that $q(\mathbf{s}|\mathbf{Y}) = \delta(\mathbf{s}-\widehat{\mathbf{s}})$, the posterior PDF of $\mathbf{W}$ is

$$q(\mathbf{W}|\mathbf{Y}) = \int q(\mathbf{W}|\mathbf{Y},\mathbf{s})\delta(\mathbf{s}-\widehat{\mathbf{s}}){\rm d}\mathbf{s} = q(\mathbf{W}|\mathbf{Y};\widehat{\mathbf{s}}). \tag{15}$$

For the given posterior PDF $q(\mathbf{W}|\mathbf{Y})$, the mean and covariance of the weights are estimated as

$$\widehat{\mathbf{w}}_i = {\rm E}_{q(\mathbf{W}|\mathbf{Y})}[\mathbf{w}_i], \tag{16a}$$

$$\widehat{\mathbf{C}}_{i,j} = {\rm E}_{q(\mathbf{W}|\mathbf{Y})}[\mathbf{w}_i\mathbf{w}_j^{\rm H}] - \widehat{\mathbf{w}}_i\widehat{\mathbf{w}}_j^{\rm H},\ i,j\in\{1,\ldots,N\}. \tag{16b}$$

Let $\mathcal{S}$ be the set of indices of the non-zero components of s, i.e.,

$$\mathcal{S} = \{i|1\leq i\leq N, s_i=1\}.$$

Analogously, $\widehat{\mathcal{S}}$ is defined based on $\widehat{\mathbf{s}}$. The model order is estimated as the cardinality of $\widehat{\mathcal{S}}$, i.e.,

$$\widehat{K} = |\widehat{\mathcal{S}}|.$$

According to (2), the noise-free signal is reconstructed as

$$\widehat{\mathbf{X}} = \sum_{i\in\widehat{\mathcal{S}}}\widehat{\mathbf{a}}_i\widehat{\mathbf{w}}_i^{\rm T}.$$

Maximizing $\mathcal{L}(q(\mathbf{\Phi}|\mathbf{Y}))$ with respect to all the factors is also intractable. Similar to the Gauss-Seidel method [36], $\mathcal{L}$ is optimized over each factor $q(\theta_i|\mathbf{Y})$, $i=1,\ldots,N$ and $q(\mathbf{W},\mathbf{s}|\mathbf{Y})$ separately with the others being fixed. Let $\mathbf{z} = (\theta_1,\ldots,\theta_N,(\mathbf{W},\mathbf{s}))$ be the set of all latent variables. Maximizing $\mathcal{L}(q(\mathbf{\Phi}|\mathbf{Y});\boldsymbol{\beta})$ (12) with respect to the posterior approximation $q(\mathbf{z}_d|\mathbf{Y})$ of each latent variable $\mathbf{z}_d$, $d=1,\ldots,N+1$ yields [35], pp. 735, Eq. (21.25)]

$$\ln q(\mathbf{z}_d|\mathbf{Y}) = {\rm E}_{q(\mathbf{z}\setminus\mathbf{z}_d|\mathbf{Y})}[\ln p(\mathbf{Y},\mathbf{z})] + {\rm const}, \tag{17}$$

where the expectation is with respect to all the variables $\mathbf{z}$ except $\mathbf{z}_d$ and the constant ensures normalization of the PDF. In the following, we detail the procedures.

3.1. Inferring the frequencies

For each $i=1,\ldots,N$, we maximize $\mathcal{L}$ with respect to the factor $q(\theta_i|\mathbf{Y})$. For $i\notin\mathcal{S}$, we have $q(\theta_i|\mathbf{Y}) = p(\theta_i)$. According to (17), for $i\in\mathcal{S}$, the optimal factor $q(\theta_i|\mathbf{Y})$ can be calculated as

$$\ln q(\theta_i|\mathbf{Y}) = {\rm E}_{q(\mathbf{z}\setminus\theta_i|\mathbf{Y})}\left[\ln p(\mathbf{Y},\mathbf{\Phi};\boldsymbol{\beta})\right] + {\rm const}. \tag{18}$$

In Appendix A.1, it is shown that

$$q(\theta_i|\mathbf{Y}) \propto \underbrace{p(\theta_i)}_{(a)}\underbrace{\exp({\rm Re}\{\boldsymbol{\eta}_i^{\rm H}\mathbf{a}(\theta_i)\})}_{(b)}, \tag{19}$$

where the complex vector $\boldsymbol{\eta}_i$ is given by

$$\boldsymbol{\eta}_i = \frac{2}{\nu}\left(\mathbf{Y} - \sum_{j\in\widehat{\mathcal{S}}\setminus\{i\}}\widehat{\mathbf{a}}_j\widehat{\mathbf{w}}_j^{\rm T}\right)\widehat{\mathbf{w}}_i^* - \frac{2}{\nu}\sum_{j\in\widehat{\mathcal{S}}\setminus\{i\}}{\rm tr}(\widehat{\mathbf{C}}_{j,i})\widehat{\mathbf{a}}_j \tag{20}$$

for $i\in\widehat{\mathcal{S}}$, and $\boldsymbol{\eta}_i = \mathbf{0}$ otherwise, which is consistent with the results in [32] for the SMV case. In order to obtain the approximate posterior distribution of $\mathbf{W}$, as shown in the next subsection, (14b) needs to be computed. While it is hard to obtain the analytical results for the PDF (19), heuristic 2 from [32] is used to obtain a von Mises approximation. For the second frequency, the prior can be similarly chosen from the set $\{p(\theta_i)\}_{i=1}^N$ with the first selected prior being removed. For the other frequencies, the steps follow similarly.

It is worth noting that for the prior distribution (5), when $\kappa_{0,i}$ tends to infinity, $p(\theta_i) = \delta(\theta_i-\mu_{0,i})$. Consequently, the signal model (2) is a sum over deterministic frequencies $\mu_{0,i}$, i.e., $\mathbf{Y} = \sum_{i=1}^N\mathbf{a}(\mu_{0,i})\mathbf{w}_i^{\rm T} + \mathbf{U}$. Thus, in this case, the MVALSE algorithm is a complete grid based method. When $\kappa_{0,i}=0$, $p(\theta_i) = \frac{1}{2\pi}$ corresponding to the uninformative prior, the MVALSE is a complete grid-less based method. Thus, by varying $\kappa_{0,i}$, the prior of the MVALSE algorithm provides a trade-off between on-grid and grid-less methods.

[1] As $I_m(\kappa)/I_0(\kappa)<1$ for $m\in 1,\cdots,M-1$, the magnitudes of the elements of ${\rm E}_{q(\theta_i|\mathbf{Y})}[\mathbf{a}(\theta_i)]$ are less than 1. An alternative approach is to assume the following posterior PDF $\delta(\theta_i-\widehat{\theta}_i)$ which corresponds to the point estimates of the frequencies, and let $\widehat{\mathbf{a}}_i$ be $\mathbf{a}(\widehat{\theta}_i)$, which yields the VALSE-pt algorithm [32]. Numerical results show that the performance of VALSE-pt is slightly worse than that of VALSE algorithm [32]. Here we use (14b) to estimate $\mathbf{a}(\theta_i)$.

3.2. Inferring the weights and support

Next $q(\theta_i|\mathbf{Y}), i=1,\ldots,N$ are fixed and $\mathcal{L}$ is maximized w.r.t. $q(\mathbf{W},\,\mathbf{s}|\mathbf{Y})$. Define the matrices $\mathbf{J}$ and $\mathbf{H}$ as

$$J_{ij}=\begin{cases}M, & i=j\\ \widehat{\mathbf{a}}_i^{\rm H}\widehat{\mathbf{a}}_j, & i\neq j\end{cases}, \quad i,j\in\{1,2,\cdots,N\}, \tag{21a}$$

$$\mathbf{H}=\widehat{\mathbf{A}}^{\rm H}\mathbf{Y}. \tag{21b}$$

According to (17), $q(\mathbf{W},\,\mathbf{s}|\mathbf{Y})$ can be calculated as

$$\begin{aligned}\ln q(\mathbf{W},\mathbf{s}|\mathbf{Y}) &= {\rm E}_{q(\mathbf{z}\backslash(\mathbf{W},s)|\mathbf{Y})}\left[\ln p(\mathbf{Y},\boldsymbol{\Phi};\boldsymbol{\beta})\right]+{\rm const}\\ &={\rm E}_{q(\boldsymbol{\theta}|\mathbf{Y})}\left[\sum_{i=1}^{N}\ln p(s_i)+\ln p(\mathbf{W}|\mathbf{s})+\ln p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{W})\right]+{\rm const}\\ &=||\mathbf{s}||_0\ln\frac{\lambda}{1-\lambda}+||\mathbf{s}||_0L\ln\frac{1}{\pi\tau}-\frac{1}{\tau}{\rm tr}(\mathbf{W}_{\mathcal{S},:}\mathbf{W}_{\mathcal{S},:}^{\rm H})\\ &\quad+\frac{2}{\nu}{\rm Re}\{{\rm tr}(\mathbf{W}_{\mathcal{S},:}^{\rm H}\mathbf{H}_{\mathcal{S},:})\}-\frac{1}{\nu}{\rm tr}(\mathbf{W}_{\mathcal{S},:}^{\rm H}\mathbf{J}_{\mathcal{S},\mathcal{S}}\mathbf{W}_{\mathcal{S},:})+{\rm const}\\ &={\rm tr}\Big((\mathbf{W}_{\mathcal{S},:}-\widehat{\mathbf{W}}_{\mathcal{S},:})^{\rm H}\widehat{\mathbf{C}}_{\mathcal{S},0}^{-1}(\mathbf{W}_{\mathcal{S},:}-\widehat{\mathbf{W}}_{\mathcal{S},:})\Big)+{\rm const},\end{aligned} \tag{22}$$

where

$$\widehat{\mathbf{W}}_{\mathcal{S},:}=\nu^{-1}\widehat{\mathbf{C}}_{\mathcal{S},0}\mathbf{H}_{\mathcal{S},:}, \tag{23a}$$

$$\widehat{\mathbf{C}}_{\mathcal{S},0}=\left(\frac{\mathbf{J}_{\mathcal{S},\mathcal{S}}}{\nu}+\frac{\mathbf{I}_{|\mathcal{S}|}}{\tau}\right)^{-1}. \tag{23b}$$

From (13), the posterior approximation $q(\mathbf{W},\,\mathbf{s}|\mathbf{Y})$ can be factored as the product of $q(\mathbf{W}|\mathbf{Y},\,\mathbf{s})$ and $\delta(\mathbf{s}-\widehat{\mathbf{s}})$. According to the formulation of (22), for a given $\widehat{\mathbf{s}}$, $q(\mathbf{W}|\mathbf{Y})$ is a complex Gaussian distribution, i.e.,

$$\begin{aligned}q(\mathbf{W}|\mathbf{Y};\widehat{\mathbf{s}}) &= \frac{1}{(\pi^{||\widehat{s}||_0}\det(\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0}))^L}\\ &\quad\times\exp\left[-{\rm tr}\Big((\mathbf{W}_{\widehat{\mathcal{S}},:}-\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:})^{\rm H}\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0}^{-1}(\mathbf{W}_{\widehat{\mathcal{S}},:}-\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:})\Big)\right]\prod_{i\notin\widehat{\mathcal{S}}}\delta(\mathbf{w}_i)\end{aligned} \tag{24}$$

$$=\prod_{l=1}^{L}\mathcal{CN}(\mathbf{w}_{\widehat{\mathcal{S}},l};\widehat{\mathbf{w}}_{\widehat{\mathcal{S}},l},\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0})\prod_{i\notin\widehat{\mathcal{S}}}\delta(w_{i,l}). \tag{25}$$

From (25), it can be seen that each column of $\mathbf{W}_{\widehat{\mathcal{S}},:}$ is independent and is a complex Gaussian distribution. This is convenient for parallel execution, as described in Section 4.

To calculate $q(\mathbf{W}|\mathbf{Y})$, $\widehat{\mathbf{s}}$ has to be given. Plugging the postulated PDF $q(\boldsymbol{\Phi}|\mathbf{Y})$ (13) in (12), one has

$$\begin{aligned}\ln Z(\widehat{\mathbf{s}}) &\triangleq \mathcal{L}(q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y});\widehat{\mathbf{s}})={\rm E}_{q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y})}\left[\ln\tfrac{p(\mathbf{Y},\boldsymbol{\theta},\mathbf{W},\mathbf{s};\widehat{\mathbf{s}})}{q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y};\widehat{\mathbf{s}})}\right]\\ &={\rm E}_{q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y})}\left[\sum_{i=1}^{N}\ln p(s_i)+\ln p(\mathbf{W}|\mathbf{s})+\ln p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{W})\right.\\ &\quad\left.-\ln q(\mathbf{W}|\mathbf{Y})\right]+{\rm const}\\ &=-L\ln\det\left(\mathbf{J}_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}+\frac{\nu}{\tau}\mathbf{I}_{|\widehat{\mathcal{S}}|}\right)+||\widehat{\mathbf{s}}||_0\ln\frac{\lambda}{1-\lambda}\\ &\quad+\nu^{-1}{\rm tr}\Big(\mathbf{H}_{\widehat{\mathcal{S}},:}^{\rm H}(\mathbf{J}_{\widehat{\mathcal{S}}}+\frac{\nu}{\tau}\mathbf{I}_{|\widehat{\mathcal{S}}|})^{-1}\mathbf{H}_{\widehat{\mathcal{S}},:}\Big)\\ &\quad+||\widehat{\mathbf{s}}||_0L\ln\frac{\nu}{\tau}+{\rm const}.\end{aligned} \tag{26}$$

Thus $\widehat{\mathbf{s}}$ should be chosen to maximize $\ln Z(\widehat{\mathbf{s}})$ (26).

The computation cost of enumerative method to find the globally optimal binary sequence $\mathbf{s}$ of (26) is $O(2^N)$, which is impractical for typical values of N. Here a greedy iterative search strategy similar to [32] is proposed. For a given $\widehat{\mathbf{s}}$, we update it as follows: For each $k=1,\cdots,N$, calculate $\Delta_k=\ln Z(\widehat{\mathbf{s}}^k)-\ln Z(\widehat{\mathbf{s}})$, where $\widehat{\mathbf{s}}^k$ is the same as $\widehat{\mathbf{s}}$ except that the kth element of $\widehat{\mathbf{s}}$ is flipped. Let $k_{\max}=\underset{k}{\operatorname{argmax}}\,\Delta_k$. If $\Delta_{k_{\max}}>0$, we update $\widehat{\mathbf{s}}$ with the $k_{\max}$th element flipped, and $\widehat{\mathbf{s}}$ is updated, otherwise $\widehat{\mathbf{s}}$ is kept, and the algorithm is terminated. In fact, Δ_k can be easily calculated and the details are provided in Appendix A.2.

Since each step increases the objective function (which is bounded) and $\mathbf{s}$ can take a finite number of values (at most 2^N), the method converges in a finite number of steps to some local optimum. If deactive is not allowed and $\hat{\mathbf{s}}^0$ is initialized as $\mathbf{0}_N$, then it can be proved that finding a local maximum of $\ln Z(\widehat{\mathbf{s}})$ costs only $O(\widehat{K})$ steps. In general, numerical experiments show that $O(\widehat{K})$ steps is often enough to find the local optimum.

3.3. Estimating the model parameters

After updating the frequencies and weights, the model parameters $\boldsymbol{\beta}=\{\nu,\ \lambda,\ \tau\}$ are estimated via maximizing the lower bound $\mathcal{L}(q(\boldsymbol{\Phi}|\mathbf{Y});\boldsymbol{\beta})$ for fixed $q(\boldsymbol{\Phi}|\mathbf{Y})$. In Appendix A.3, it is shown that

$$\begin{aligned}\mathcal{L}(q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y});\boldsymbol{\beta}) &= {\rm E}_{q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y})}\left[\ln\tfrac{p(\mathbf{Y},\boldsymbol{\theta},\mathbf{W},\mathbf{s};\boldsymbol{\beta})}{q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y})}\right]\\ &=-\frac{1}{\nu}\Big[||\mathbf{Y}-\widehat{\mathbf{A}}_{:,\widehat{\mathcal{S}}}\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}||_{\rm F}^2+L{\rm tr}(\mathbf{J}_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0})\Big]\\ &\quad-\frac{1}{\tau}[{\rm tr}(\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}^{\rm H})+L{\rm tr}(\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0})]\\ &\quad+||\widehat{\mathbf{s}}||_0(\ln\frac{\lambda}{1-\lambda}-L\ln\tau)+N\ln(1-\lambda)-ML\ln\nu+{\rm const}.\end{aligned} \tag{27}$$

Setting $\frac{\partial\mathcal{L}}{\partial\nu}=0,\ \frac{\partial\mathcal{L}}{\partial\lambda}=0,\ \frac{\partial\mathcal{L}}{\partial\tau}=0,$ we have

$$\begin{aligned}\widehat{\nu} &= ||\mathbf{Y}-\widehat{\mathbf{A}}_{:,\widehat{\mathcal{S}}}\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}||_{\rm F}^2/(ML)+{\rm tr}(\mathbf{J}_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0})/M\\ &\quad+\sum_{i\in\widehat{\mathcal{S}}}\sum_{l=1}^{L}|\widehat{W}_{il}|^2(1-||\widehat{\mathbf{a}}_i||_2^2/M)/L,\\ \widehat{\lambda} &= \frac{||\widehat{\mathbf{s}}||_0}{N},\qquad \widehat{\tau}=\frac{{\rm tr}(\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}^{\rm H})+L{\rm tr}(\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0})}{L||\widehat{\mathbf{s}}||_0}.\end{aligned} \tag{28}$$

3.4. The MVALSE algorithm

Now the details of updating the assumed posterior $q(\boldsymbol{\theta},\,\mathbf{W},\,\mathbf{s}|\mathbf{Y})$ have been given and summarized in Algorithm 1. For the proposed algorithm, the initialization is important for the performance of the algorithm. The schemes that we initialize $\widehat{\nu}$, $\widehat{\lambda}$, $\widehat{\tau}$ and $q(\theta_i|\mathbf{Y})$, $i\in\{1,\cdots,N\}$ are below.

Algorithm 1 Outline of MVALSE algorithm with MMV setting.

Input: Signal matrix $\mathbf{Y}$

Output: The model order estimate $\widehat{K}$, frequencies estimate $\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{S}}}$, complex weights estimate $\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}$ and reconstructed signal $\widehat{\mathbf{X}}$

```
Initialize $\widehat{\nu}$, $\widehat{\lambda}$, $\widehat{\tau}$ and $q_{\theta_i|\mathbf{Y}}, i\in\{1,\cdots,N\}$; compute $\widehat{\mathbf{a}}_i$
repeat
    Update $\widehat{\mathbf{s}}$, $\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}$ and $\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0}$ (Sec.3.2)
    Update $\widehat{\nu}$, $\widehat{\lambda}$, $\widehat{\tau}$ (28)
    Update $\boldsymbol{\eta}_i$ and $\widehat{\mathbf{a}}_i$ for all $i\in\widehat{\mathcal{S}}$ (Sec.3.1)
until stopping criterion is satisfied
return $\widehat{K}$, $\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{S}}}$, $\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}$ and $\widehat{\mathbf{X}}$
```

First, initialize $q(\theta_1|\mathbf{Y})$ as $q(\theta_1|\mathbf{Y}) \propto \exp(\frac{||\mathbf{Y}^{\rm H}\mathbf{a}(\theta_1)||_2^2}{\nu M})$, which can be simplified as the form similar to (19): By defining $\mathcal{M}' = \{m - n \mid m, n \in \{0, 1, \cdots, M-1\}, m > n\}$ with cardinality $M' = M - 1$ and $\mathbf{a}' : [-\pi, \pi) \to \mathbb{C}^{M'}, \theta \to \mathbf{a}'(\theta) \triangleq ({\rm e}^{{\rm j}\theta m} \mid m \in \mathcal{M}')^{\rm T}$. Obviously $\mathbf{a}(\theta) = [1; \mathbf{a}'(\theta)]$. For each $t = 1, \cdots, M'$, by constructing γ_t as $\gamma_t = \frac{1}{M}\sum_{(k,l)\in\mathcal{T}_t} \mathbf{Y}_{k,:}\mathbf{Y}_{l,:}^{\rm H}$ with $\mathcal{T}_t = \{(k, l) \mid 1 \le k, l \le M, m_k - m_l = t\}$, $q(\theta_1|\mathbf{Y})$ can be re-expressed as

$$q(\theta_1|\mathbf{Y}) \propto \exp\left({\rm Re}\left\{\frac{2}{\nu}\gamma^{\rm H}\mathbf{a}'(\theta_1)\right\}\right). \tag{29}$$

Then $\hat{\mathbf{a}}_1 = {\rm E}[\mathbf{a}(\theta_1)]$ can be calculated. Since $J_1 = M$ and $\mathbf{H}_1$ (21) can be calculated. According to (23b) and (23a), $\hat{\mathbf{w}}_1$ is calculated. Then we update $q(\theta_2|\mathbf{Y}) \propto \exp(\frac{||\mathbf{Y}_1^{\rm H}\mathbf{a}(\theta_2)||_2^2}{\nu M})$ with $\mathbf{Y}_1 = \mathbf{Y} - \hat{\mathbf{a}}_1\hat{\mathbf{w}}_1^{\rm T}$. Following the previous steps, $q(\theta_i|\mathbf{Y})$, $\hat{\mathbf{a}}_i$ and $\hat{\mathbf{w}}_i$ are all initialized. As for the model parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_{M-1}]^{\rm T} \in \mathbb{C}^{M-1}$ is used to build a Toeplitz estimate of ${\rm E}[\mathbf{Y}\mathbf{Y}^{\rm H}]$. Let $L\hat{\nu}$ be the average of the lower quarter of the eigenvalues of that matrix, and $\hat{\tau}$ is initialized as $\hat{\tau} = ({\rm tr}[\mathbf{Y}^{\rm H}\mathbf{Y}]/M - L\hat{\nu})/(\hat{\lambda}N)$, where the active probability λ is initialized as $\hat{\lambda} = 0.5$.

The complexity of MVALSE algorithm is dominated by the two steps [32]: the maximization of $\ln Z(\mathbf{s})$ and the approximations of the posterior PDF $q(\theta|\mathbf{Y})$ by mixtures of von Mises PDFs. For the maximization of $\ln Z(\mathbf{s})$, if $\mathcal{S}$ is initialized such that $|\hat{\mathcal{S}}| = 0$ and deactivation is not allowed, it can be proved that the greedy iterative search strategy needs at most N steps to converge. For the general case where deactive is allowed, numerical experiments show that $O(N)$ steps is enough to converge. For each step, the computational complexity is $O(N^2 + NL)$ due to the matrix multiplication. Therefore, the computational complexity is $O(N^4 + N^3L)$. For the approximations of the posterior PDF $q(\theta|\mathbf{Y})$ by mixtures of von Mises PDFs, the Heuristic 2 method [32], Section IV.D] is adopted and the computational complexity is $O(N^2M + M^2N + N^2L + MNL)$. In conclusion, the dominant computational complexity of the MVALSE is $O[(N^4 + N^3L) \times T]$ with T being the number of iterations as M is close to N.

4. MVALSE With parallel processing

The MVALSE Algorithm 1 is compared with the VALSE algorithm [32]. The MMV is decoupled as L SMVs. For each SMV, we perform the VALSE algorithm and obtain $\boldsymbol{\eta}_{i,l}$ according to [32], Eq. (17)] for the lth snapshot, i.e.,

$$\boldsymbol{\eta}_{i,l} = \frac{2}{\nu}\left(\mathbf{y}_l - \sum_{j\in\hat{\mathcal{S}}\backslash\{i\}} \hat{\mathbf{a}}_j[\hat{\mathbf{w}}_j^{\rm T}]_l\right)[\hat{\mathbf{w}}_i^*]_l - \frac{2}{\nu}\sum_{j\in\hat{\mathcal{S}}\backslash\{i\}} [\hat{\mathbf{C}}_{j,i}]_{l,l}\hat{\mathbf{a}}_j, \tag{30}$$

where $[\hat{\mathbf{C}}_{j,i}]_{l,l}$ denotes the (l, l)th element of $\hat{\mathbf{C}}_{j,i}$, $[\hat{\mathbf{w}}_j^{\rm T}]_l$ denotes the lth element of $\hat{\mathbf{w}}_j^{\rm T}$. From (20), $\boldsymbol{\eta}_i$ is the sum of $\boldsymbol{\eta}_{i,l}$ for all the snapshots, i.e., $\boldsymbol{\eta}_i = \sum_{l=1}^{L}\boldsymbol{\eta}_{i,l}$, and now each $\boldsymbol{\eta}_{i,l}$ is updated as $\boldsymbol{\eta}_i$. We use $\boldsymbol{\eta}_i$ to obtain estimates $\hat{\theta}_i$ and $\hat{\mathbf{a}}_i$ [32]. In addition, we update the weights and their covariance (23) by applying the SMV VALSE. Let $\hat{\mathbf{w}}_{\hat{\mathcal{S}},l}^{\rm T}$ be the estimated weights of the lth snapshot, the whole weight matrix $\hat{\mathbf{W}}_{\hat{\mathcal{S}},:}$ (23) can be constructed as $[\hat{\mathbf{w}}_{\hat{\mathcal{S}},1}^{\rm T}; \cdots; \hat{\mathbf{w}}_{\hat{\mathcal{S}},L}^{\rm T}]$. It is worth noting that Eq. (25) reveals that for different snapshots, the weight vectors are uncorrelated. Besides, the covariance of the weights for each snapshot is the same, which means that the common covariance of the weight can be fed to the SMV VALSE. For updating $\mathcal{S}$ under the active case, according to [32], Eq. (40)], the changes $\Delta_{k,l}$ for the lth snapshot is

$$\Delta_{k,l} = \ln\frac{\nu_k}{\tau} + \frac{|[\mathbf{u}_k]_l|^2}{\nu_k} + \ln\frac{\lambda}{1-\lambda}. \tag{31}$$

Thus (39) can also be expressed as

$$\Delta_k = \sum_{l=1}^{L}\Delta_{k,l} - (L-1)\ln\frac{\lambda}{1-\lambda}, \tag{32}$$

which can be viewed as a sum of the results $\Delta_{k,l}$ from the VALSE in SMVs, minus an additional constant term $(L-1)\ln\frac{\lambda}{1-\lambda}$. Similarly, for the deactive case, (42) can be viewed as a sum of the results (Eq. (44) in [32]) from the VALSE in SMVs, plus an additional constant term $(L-1)\ln\frac{\lambda}{1-\lambda}$. The additional constant terms can not be neglected because we need to determine the sign of (39) and (42) to update $\mathcal{S}$. For the lth snapshot, running the VALSE algorithm yields the model parameters estimates

$$\begin{aligned}\hat{\nu}_l &= ||\mathbf{y}_l - \hat{\mathbf{A}}_{\hat{\mathcal{S}}}[\hat{\mathbf{W}}_{\hat{\mathcal{S}},:}]_{:,l}||^2/M + {\rm tr}(\mathbf{J}_{\hat{\mathcal{S}},\hat{\mathcal{S}}}\hat{\mathbf{C}}_{\hat{\mathcal{S}},0})/M \\ &\quad + \sum_{i\in\hat{\mathcal{S}}}|\hat{w}_{il}|^2(1 - ||\hat{\mathbf{a}}_i||_2^2/M), \\ \hat{\tau}_l &= \frac{\|[\hat{\mathbf{W}}_{\hat{\mathcal{S}},:}]_{:,l}\|^2 + {\rm tr}(\hat{\mathbf{C}}_{\hat{\mathcal{S}},0})}{||\hat{\mathbf{s}}||_0}.\end{aligned} \tag{33}$$

According to (28), model parameters estimates $\hat{\nu}$ and $\hat{\tau}$ are updated as the average of their respective estimates, i.e., $\hat{\nu} = \sum_{l=1}^{L}\hat{\nu}_l/L$ and $\hat{\tau} = \sum_{l=1}^{L}\hat{\tau}_l/L$, where $\hat{\nu}_l$ and $\hat{\tau}_l$ denote the estimate of the lth SMV VALSE, and $\hat{\lambda}$ can be naturally estimated.

5. MVALSE For sequential estimation (Seq-MVALSE)

The previous MVALSE algorithm is designed to process a batch of data. In fact, MVALSE is very suitable for sequential estimation [38]. We develop the Seq-MVALSE algorithm for sequential estimation, which is very natural as MVALSE outputs conjugate priors of the frequency. Suppose that the whole data $\mathbf{Y} = [\mathbf{Y}_{g_1}, \mathbf{Y}_{g_2}, \cdots, \mathbf{Y}_{g_G}]$ is partitioned into G groups, where $g_1 + g_2 + \cdots + g_G = L$. For the first group with data $\mathbf{Y}_{g_1}$, we perform the MVALSE and obtain the posterior PDF of the frequencies. Then the posterior PDF of the frequencies can be viewed as the prior of the frequencies, and the MVALSE is performed with data $\mathbf{Y}_{g_2}$. Following the previous steps, Seq-MVALSE can be obtained for sequential estimation. The Seq-MVALSE is summarized as Algorithm 2.

Algorithm 2 Outline of Seq-MVALSE.

Input: Signal matrix $\mathbf{Y} = [\mathbf{Y}_{g_1}, \mathbf{Y}_{g_2}, \cdots, \mathbf{Y}_{g_G}]$

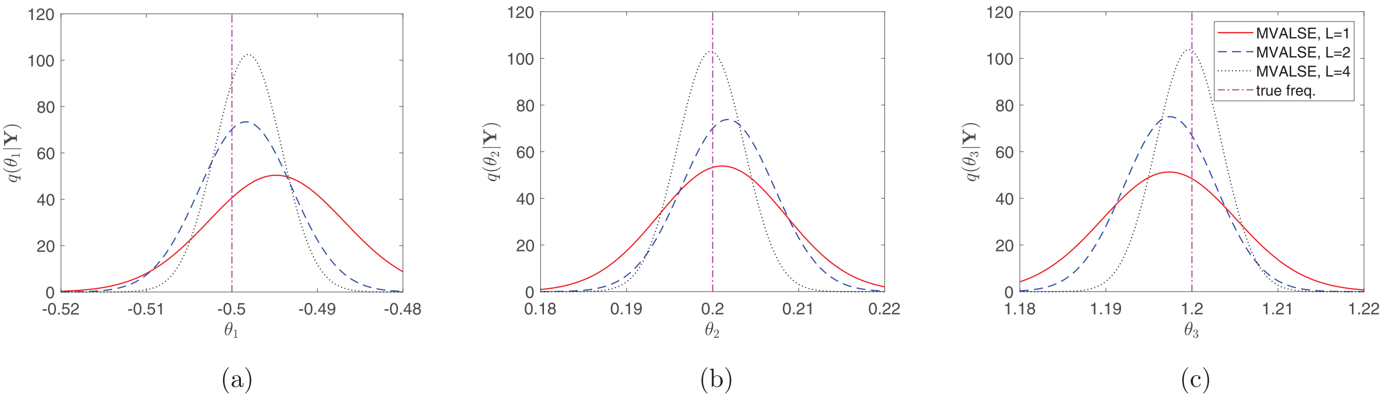
Output: The model order estimate $\hat{K}$, frequencies estimate $\hat{\boldsymbol{\theta}}_{\hat{\mathcal{S}}}$, complex weights estimate $\hat{\mathbf{W}}_{\hat{\mathcal{S}},:}$ and reconstructed signal $\hat{\mathbf{X}}$

```
Initialize $\hat{\nu}, \hat{\lambda}, \hat{\tau}$ and $q_{\theta_i|\mathbf{Y}_{g_1}}, i \in \{1, \cdots, N\}$; compute $\hat{\mathbf{a}}_i$
for $j = 1, \cdots, G$ do
    Run the MVALSE algorithm with data $\mathbf{Y}_{g_j}$, and output the posterior PDF $p(\boldsymbol{\theta}|\mathbf{Y}_{g_j})$.
    Set $p(\boldsymbol{\theta}|\mathbf{Y}_{g_j})$ as the prior distribution of the next data group.
end for
Return $\hat{K}$, $\hat{\boldsymbol{\theta}}_{\hat{\mathcal{S}}}$, $\hat{\mathbf{W}}_{\hat{\mathcal{S}},:}$ and $\hat{\mathbf{X}}$
```

6. Numerical simulation

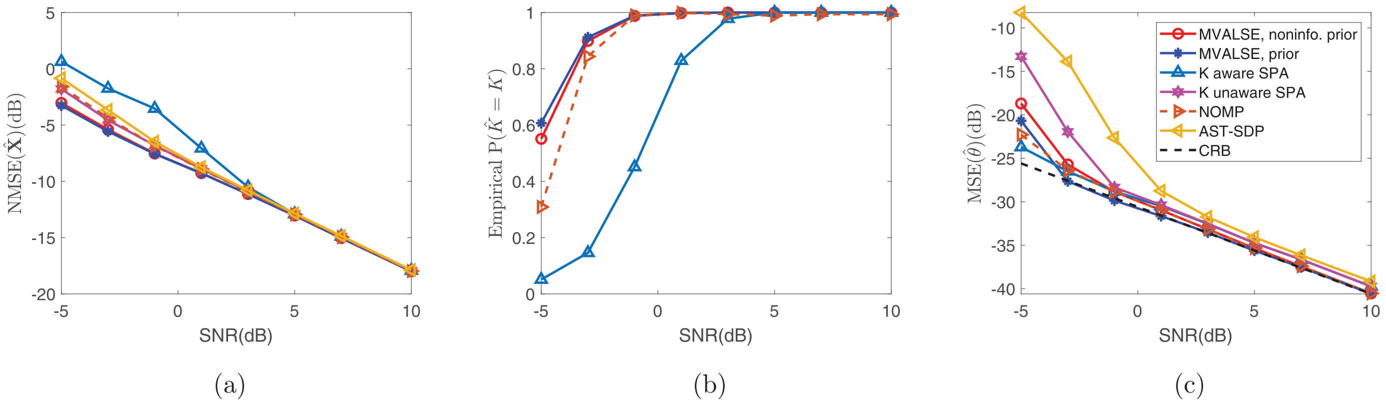
In this section, substantial numerical simulations are performed to substantiate the MVALSE algorithm.

For the signal generation, the frequencies are generated as follows unless stated otherwise: First, K distributions are uniformly picked from N von Mises distributions (5) with $\mu_{0,i} = (2i - 1 - N)/(N+1)\pi$ and $\kappa_{0,i} = 10^3$, $i = 1, \cdots, N$ without replacement. The frequencies $\{\theta_i\}_{i=1}^{K}$ are generated from the selected von Mises distribution and the minimum wrap-around distance is greater than



(a) (b) (c)

Fig. 1. The posterior PDF of the frequencies generated by MVALSE for $M = N = 20$ and SNR = 10 dB.



(a) (b) (c)

Fig. 2. Performance of algorithms versus SNR for $M = 20$ and $L = 8$.

$\Delta\theta = \frac{2\pi}{N}$. The elements of $\mathbf{W}$ are drawn i.i.d. from $\mathcal{CN}(1, 0.1)$. Other parameters are: $K = 3,\ N = 20$. The standard deviation of the von Mises distribution $\approx 1/\sqrt{\kappa_{0,i}} \approx 0.0316$ and the distance between the adjacent frequencies is $\mu_{0,i+1} - \mu_{0,i} = 0.3 \leq 2\pi/N$. Thus the MVALSE with prior is almost a grid based method.

We define signal-to-noise ratio (SNR) as $\text{SNR} \triangleq 20\log(||\mathbf{A}(\widetilde{\boldsymbol{\theta}})\widetilde{\mathbf{W}}^{\text{T}}||_{\text{F}}/||\widetilde{\mathbf{U}}||_{\text{F}})$ and the normalized mean square error (NMSE) of $\widehat{\mathbf{X}}$ and MSE of $\widehat{\boldsymbol{\theta}}$ are $\text{NMSE}(\widehat{\mathbf{X}}) \triangleq 20\log(||\widehat{\mathbf{X}} - \mathbf{A}(\widetilde{\boldsymbol{\theta}})\widetilde{\mathbf{W}}^{\text{T}}||_{\text{F}}/||\mathbf{A}(\widetilde{\boldsymbol{\theta}})\widetilde{\mathbf{W}}^{\text{T}}||_{\text{F}})$ and $\text{MSE}(\widehat{\boldsymbol{\theta}}) \triangleq 20\log(||\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}||_2)$, the correct model order estimated probability $\text{P}(\widehat{K} = K)$ are adopted as the performance metrics. The MSE of the frequency is calculated only when both the model order is correctly estimated and $\text{MSE}(\widehat{\boldsymbol{\theta}}) \leq 0$ (dB). The Algorithm 1 stops when $||\widehat{\mathbf{X}}^{(t-1)} - \widehat{\mathbf{X}}^{(t)}||_{\text{F}}/||\widehat{\mathbf{X}}^{(t-1)}||_{\text{F}} < 10^{-5}$ or $t > 5000$, where t is the number of iteration.

As for performance comparison, the SPA method [24], the atomic norm minimization algorithm proposed in [26] called AST-SDP, the Newtonized orthogonal matching pursuit (NOMP) method [10,11] and the Cramér-Rao bound (CRB) derived in [11] are chosen. For the SPA approach, the denoised covariance matrix is obtained firstly and the MUSIC method is used to avoid frequency splitting phenomenon, where the MUSIC method is provided by MATLAB *rootmusic* and the optimal sliding window W is empirically found, $W = 12$. For the NOMP method, the termination condition is set such that the probability of model order overestimate is 1% [11]. All results are averaged over 10^3 Monte Carlo (MC) trials unless stated otherwise.

At first, we apply MVALSE to obtain the posterior PDF of the frequencies, see Fig. 1. It can be seen that as the number of snapshots increases, the posterior PDFs become more concentrated, i.e., the uncertainties becomes smaller.

6.1. *Performance of MVALSE*

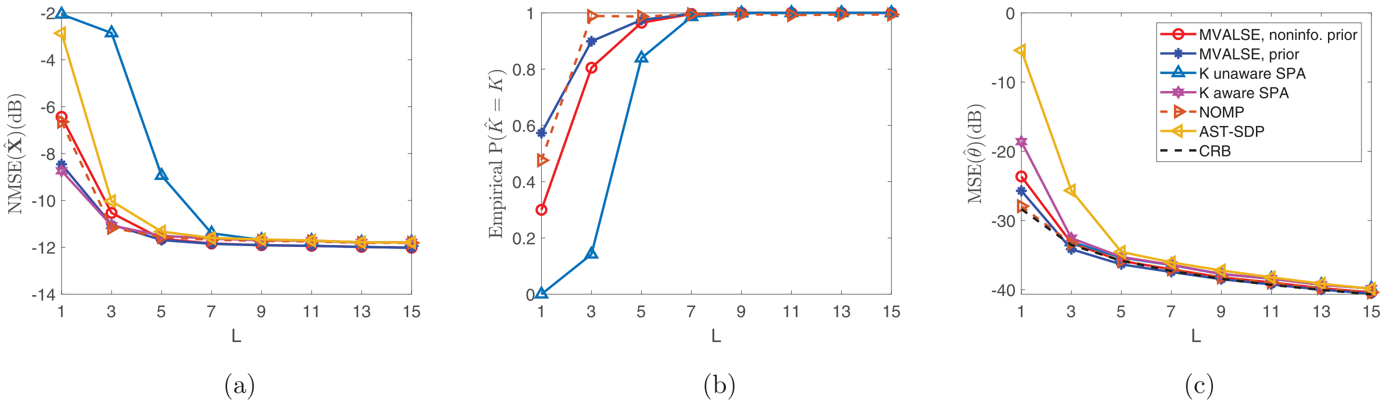
In this section, the performance of MVALSE algorithm is evaluated by varying SNR, the number of snapshots L or the number of observations M.

6.1.1. *Estimation with varied SNR*

The performance in terms of model order estimation accuracy and frequency estimation error by varying SNR is presented in Fig. 2. In Fig. 2(a), as the SNR increase, $\text{NMSE}(\widehat{\mathbf{X}})$ decreases. When SNR ≥ 5 dB, $\text{NMSE}(\widehat{\mathbf{X}})$ are almost identical for all the algorithms. It can be seen that utilizing the prior information improves the performance of the VALSE algorithm. In Fig. 2(b), the MVALSE algorithm achieves the highest probability of correct model order estimation, compared with NOMP and SPA algorithms. For the frequency estimation error, it is seen that MVALSE with prior approaches the CRB firstly. The MSE of frequency of the K aware SPA is larger than that of K unaware SPA. The reason is that the MSE of frequency is averaged only when both the model order is correct and $\text{MSE}(\widehat{\boldsymbol{\theta}}) \leq 0$ (dB), which makes the MSE of frequency of the K unaware SPA algorithm lower.

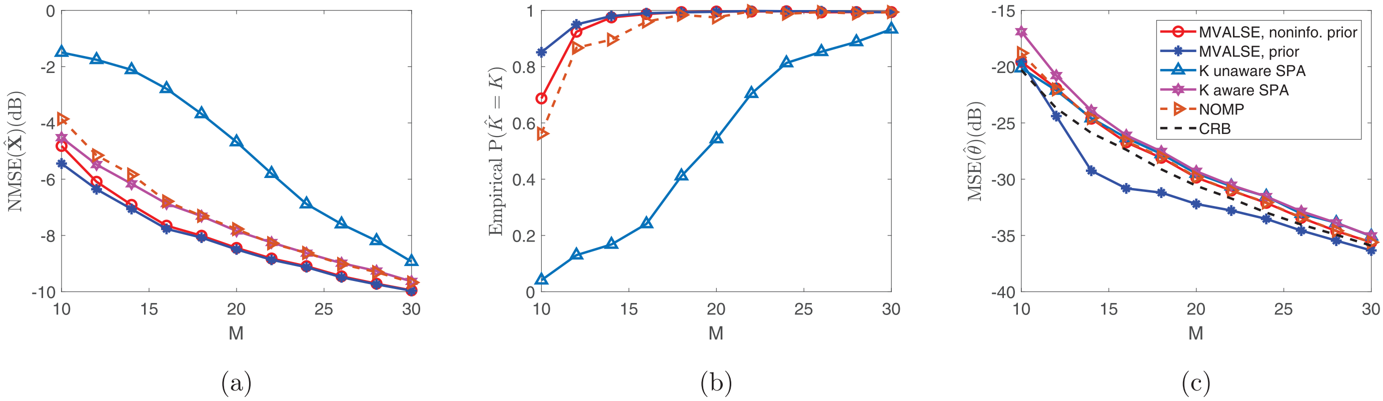
6.1.2. *Estimation with varied L*

In this subsection, we examine the estimation performance by varying the number of snapshots L, see Fig. 3. In Fig. 3(a), as the number of snapshots L increases, $\text{NMSE}(\widehat{\mathbf{X}})$ decreases and finally becomes stable. pFrom Fig. 3(b) and (c), we can see that when $L \geq 3$, the NOMP algorithm achieves the highest probability of correct model order estimation, while its NMSE is higher than that of both MVALSE with prior and K aware SPA methods. For the frequency estimation error in Fig. 3(c), all the algorithms approach to the CRLB as L increases.



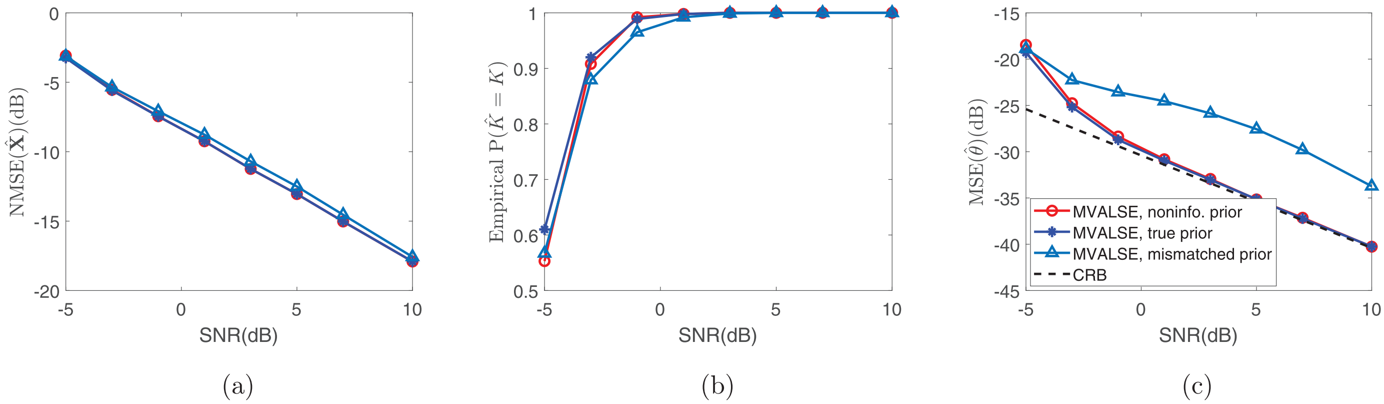
(a) (b) (c)

Fig. 3. Performance of algorithms versus snapshots L for SNR = 2dB and $M = 30$.



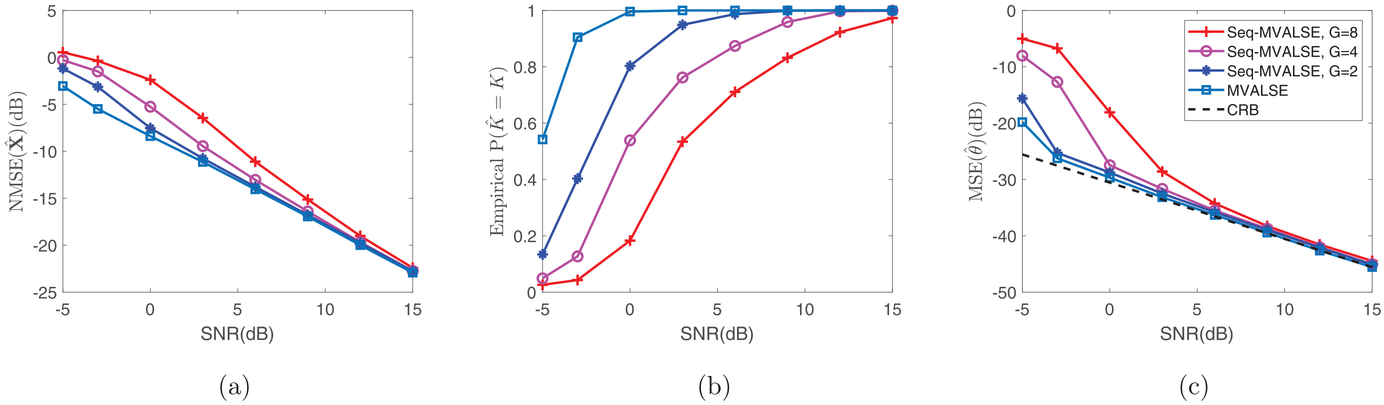
(a) (b) (c)

Fig. 4. Performance of algorithms versus M for SNR = 0 dB and $L = 8$.



(a) (b) (c)

Fig. 5. Performances of MVALSE, MAVLSE with true prior, MVALSE with mismatched prior for $L = 8$, $K = 3$ and $M = 20$.



(a) (b) (c)

Fig. 6. Performance of MVALSE for sequential estimation versus SNR for $L = 8$ and $M = 20$.

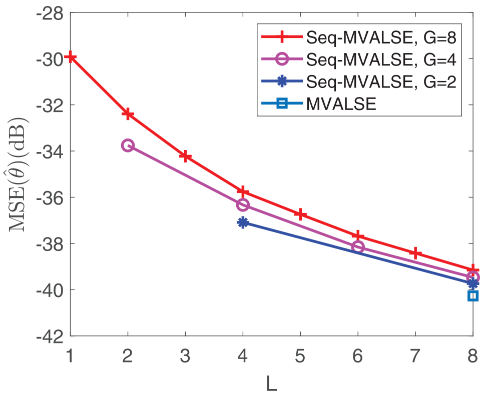


Fig. 7. NMSE of frequency estimation of MVALSE for sequential estimation by varying L. We set SNR = 10 dB and the number of measurements $M = 20$.

6.1.3. Estimation with varied m

The performance is examined by varying the number of measurements per snapshots, see Fig. 4. Note that AST-SDP are not presented for the poor performance. In Fig. 4(a) and (b), MVALSE achieves the best performance. As for the performance in terms of frequency estimation error, MVALSE with prior is lower than CRB in Fig. 4(c). In 4(b), the model order probability of all the algorithms are close to 1 except the K aware SPA algorithm for $M \geq 20$. Meanwhile, MVALSE and NOMP also works well and asymptotically approach the CRB.

6.1.4. MVALSE With mismatched prior

This subsection investigates the mismatched case, i.e., MVALSE utilizing the prior which is not matched with the signal generated prior. The true $\kappa_{0,i}$ is $\kappa_{0,i} = 10^3$, while the MVALSE with mismatched prior uses $\kappa_{\mathrm{mis},i} = 10^4$. Fig. 5 shows that MAVSLE with mismatched prior has some performance degradation, while MVALSE with true prior has some performance improvement, compared to the MVALSE with uninformative prior.

6.2. Sequential estimation

The performance of Seq-MVALSE is evaluated with number of snapshots $L = 8$. The snapshots are uniformly partitioned into G groups (G is 1, 4, or 8). Note that for $G = 1$, we perform MVALSE. We set $K = 3$, $M = 20$ and $N = 20$.

Two numerical experiments are conducted to investigate the performance. For the first numerical experiment in Fig. 6, the SNR is varied. As the SNR increases, the performances of all algorithms improve. In addition, comparing with MVALSE, Seq-MVALSE has some performance degradation. As G decreases, the performances of Seq-MVALSE improve. For the second numerical experiment in Fig. 7, the performance is investigated with the whole number of snapshots fixed as 8. It can be seen that the algorithm improves as the data arrives. For the fixed number of snapshots, the performance of Seq-MVALSE algorithm improves as G decreases.

6.3. Application: DOA estimation

The performance of MVALSE for DOA estimation is evaluated in this experiment. Let $\boldsymbol{\phi} \in \mathbb{R}^K$ denote the DOAs. For the DOA estimation problem where K narrow band far-field signals impinging onto an M-element uniform linear array (ULA) with half wavelength spacing, i.e., $d = \lambda/2$, the DOA estimation problem can be formulated as the LSE with $\boldsymbol{\theta} = \frac{2\pi d}{\lambda}\sin(\boldsymbol{\phi}) = \pi\sin(\boldsymbol{\phi})$. We generate the frequencies $\boldsymbol{\theta}$ such that the DOAs are $[-2, 5, 12]^\circ$. We set $M = 40$, $L = 20$ and $K = 3$. Since EPUMA outperforms many subspace based DOA estimators, especially for small sample scenarios [37], we compare the MVALSE with EPUMA. Similar to [37], the root MSE (RMSE) $\mathrm{RMSE} \triangleq \sqrt{\sum_{i=1}^{K}(\hat{\phi}_i - \phi_i)^2}$ is used to characterize the performance of the algorithms, where $\hat{\boldsymbol{\phi}}$ denotes the output of the algorithm.

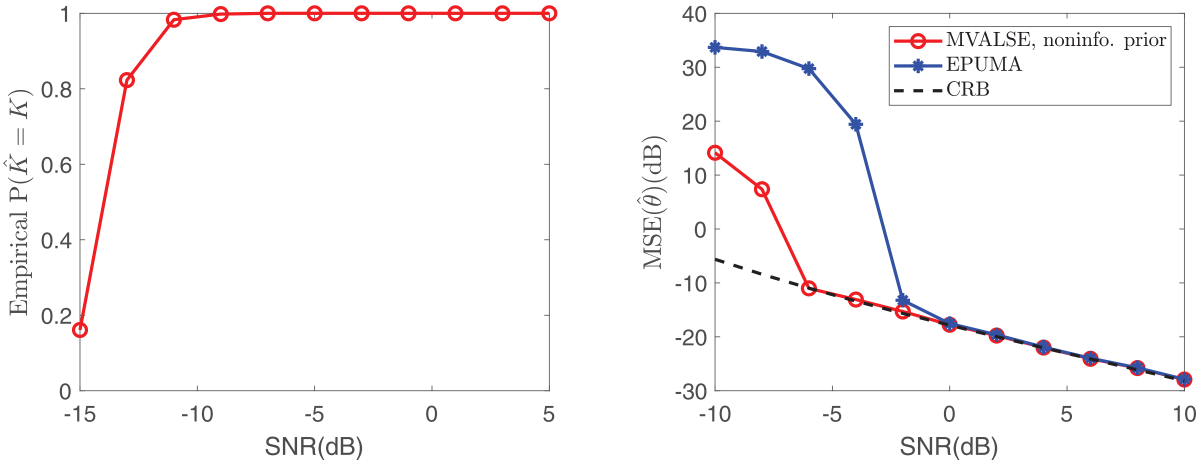
Fig. 8 shows that MVALSE performs better than EPUMA. Both algorithms approach the CRB as SNR increases.

7. Conclusion

The MVALSE algorithm is developed to jointly estimate the frequencies and weight coefficients for MMV. The MVALSE estimates the posterior PDF of the frequencies. The performance of the MVALSE method with von Mises prior PDFs for the frequencies is studied. It is also shown that the derived MVALSE is closely related to the VALSE algorithm, which is suitable for parallel processing. Furthermore, the MVALSE is extended to perform sequential estimation. Substantial experiments are conducted to illustrate the competitive performance of the MVALSE method and its application to DOA problems, compared to other approaches.

Declaration of interests

None.



(a)

(b)

Fig. 8. Performance of MVALSE algorithm for DOA estimation. We set SNR = 10 dB, the number of measurements is $M = 40$ and the number of snapshots is $L = 20$.

Acknowledgement

The authors thank Qian Cheng for sharing the EPUMA code to help us make the performance comparison in the DOA experiments. This work is supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ18F010001 and National Natural Science Foundation of China (NSFC 61674128, 61731019).

Appendix A

A.1. Derivation of $q(\theta_i|\mathbf{Y})$

Substituting (14) and (16) in (18), $\ln q(\theta_i|\mathbf{Y})$ is obtained as

$$\begin{aligned}\ln q(\theta_i|\mathbf{Y}) &= \mathrm{E}_{q(\mathbf{z}\backslash\theta_i|\mathbf{Y})}[\ln p(\mathbf{Y},\boldsymbol{\Phi})] + \text{const}\\ &= \mathrm{E}_{q(\mathbf{z}\backslash\theta_i|\mathbf{Y})}[\ln(p(\boldsymbol{\theta})p(\mathbf{s})p(\mathbf{W}|\mathbf{s})p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{W}))] + \text{const}\\ &= \mathrm{E}_{q(\mathbf{z}\backslash\theta_i|\mathbf{Y})}\left[\sum_{j=1}^{N}\ln p(\theta_j)+\sum_{j=1}^{N}\ln p(s_j)+\ln p(\mathbf{W}|\mathbf{s})+\ln p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{W})\right]\\ &\quad + \text{const}\\ &= \ln p(\theta_i) + \mathrm{E}_{q(\mathbf{z}\backslash\theta_i|\mathbf{Y})}[\nu^{-1}||\mathbf{Y}-\mathbf{A}_{:,\widehat{\mathcal{S}}}\mathbf{W}_{\widehat{\mathcal{S}},:}||_{\mathrm{F}}^{2}] + \text{const}\\ &= \ln p(\theta_i) + 2\nu^{-1}\mathrm{Re}\left\{\widehat{\mathbf{w}}_i^{\mathrm{T}}\mathbf{Y}^{\mathrm{H}}\mathbf{a}(\theta_i)\right\}\\ &\quad - 2\nu^{-1}\mathrm{Re}\left\{\mathrm{E}_{q(\mathbf{z}\backslash\theta_i|\mathbf{Y})}\left[(\mathbf{w}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{H}}_{\widehat{\mathcal{S}}\backslash\{i\},:}\mathbf{A}^{\mathrm{H}}_{:,\widehat{\mathcal{S}}\backslash\{i\}})\mathbf{a}(\theta_i)\right]\right\} + \text{const}\\ &\overset{a}{=} \ln p(\theta_i) + \mathrm{Re}\left\{\boldsymbol{\eta}_i^{\mathrm{H}}\mathbf{a}(\theta_i)\right\},\end{aligned} \tag{34}$$

where $\overset{a}{=}$ utilizes (16), and the complex vector $\boldsymbol{\eta}_i$ is given in (20). Thus $q(\theta_i|\mathbf{Y})$ is obtained in (19).

A.2. Finding a local maximum of $\ln Z(\mathbf{s})$

Finding the globally optimal binary sequence $\mathbf{s}$ of (26) is hard in general. As a result, a greedy iterative search strategy is adopted [32]. We proceed as follows: In the pth iteration, we obtain the kth test sequence $\mathbf{t}_k$ by flipping the kth element of $\mathbf{s}^{(p)}$. Then we calculate $\Delta_k^{(p)} = \ln Z(\mathbf{t}_k) - \ln Z(\mathbf{s}^{(p)})$ for each $k = 1,\cdots,N$. If $\Delta_k^{(p)} < 0$ holds for all k we terminate the algorithm and set $\widehat{\mathbf{s}} = \mathbf{s}^{(p)}$, else we choose the t_k corresponding to the maximum $\Delta_k^{(p)}$ as $\mathbf{s}^{(p+1)}$ in the next iteration.

When $k \notin \mathcal{S}$, that is, $s_k = 0$, we activate the kth component of $\mathbf{s}$ by setting $s_k' = 1$. Now, $\mathcal{S}' = \mathcal{S}\cup\{k\}$.

$$\begin{aligned}\Delta_k &= \ln Z(\mathbf{s}') - \ln Z(\mathbf{s})\\ &= L\left(\ln\det\left(\mathbf{J}_{\mathcal{S},\mathcal{S}}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}|}\right)-\ln\det\left(\mathbf{J}_{\mathcal{S}',\mathcal{S}'}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}'|}\right)\right)+\ln\frac{\lambda}{1-\lambda}+L\ln\frac{\nu}{\tau}\\ &\quad+\nu^{-1}\mathrm{tr}\left(\mathbf{H}^{\mathrm{H}}_{\mathcal{S}',:}\left(\mathbf{J}_{\mathcal{S}',\mathcal{S}'}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}'|}\right)^{-1}\mathbf{H}_{\mathcal{S}',:}-\mathbf{H}^{\mathrm{H}}_{\mathcal{S},:}\left(\mathbf{J}_{\mathcal{S},\mathcal{S}}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}|}\right)^{-1}\mathbf{H}_{\mathcal{S},:}\right).\end{aligned} \tag{35}$$

Let $\mathbf{j}_k = \mathbf{J}_{\mathcal{S},k}$ denote the kth column of $\mathbf{J}_{\mathcal{S},\mathcal{S}}$ and $\mathbf{h}_k^{\mathrm{T}} = \mathbf{H}_{k,:}$ denote the kth row of $\mathbf{H}$. Generally, $\mathbf{j}_k$ and $\mathbf{j}_k^{\mathrm{T}}$ should be inserted into the kth column and kth row of $\mathbf{J}_{\mathcal{S}}$, respectively, and M is inserted into (k, k)th of $\mathbf{J}_{\mathcal{S},\mathcal{S}}$ to obtain $\mathbf{J}_{\mathcal{S}',\mathcal{S}'}$. By using the block-matrix determinant formula, one has

$$\begin{aligned}\ln\det(\mathbf{J}_{\mathcal{S}',\mathcal{S}'}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}'|}) &= \ln\det\left(\mathbf{J}_{\mathcal{S},\mathcal{S}}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}|}\right)\\ &\quad+\ln\left(M+\frac{\nu}{\tau}-\mathbf{j}_k^{\mathrm{H}}\left(\mathbf{J}_{\mathcal{S},\mathcal{S}}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}|}\right)^{-1}\mathbf{j}_k\right).\end{aligned} \tag{36}$$

Similarly, $\mathbf{h}_k^{\mathrm{T}}$ is inserted into the kth row of $\mathbf{H}_{\mathcal{S},:}$. By the block-wise matrix inversion formula, one has

$$\begin{aligned}\mathrm{tr}\left[\mathbf{H}^{\mathrm{H}}_{\mathcal{S}',:}\left(\mathbf{J}_{\mathcal{S}',\mathcal{S}'}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}'|}\right)^{-1}\mathbf{H}_{\mathcal{S}',:}\right] &= \mathrm{tr}\left[\mathbf{H}^{\mathrm{H}}_{\mathcal{S},:}\left(\mathbf{J}_{\mathcal{S},\mathcal{S}}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}|}\right)^{-1}\mathbf{H}_{\mathcal{S},:}\right]\\ &\quad+\nu\frac{\mathbf{u}_k^{\mathrm{H}}\mathbf{u}_k}{\nu_k},\end{aligned} \tag{37}$$

where

$$\begin{aligned}\nu_k &= \nu\left(M+\frac{\nu}{\tau}-\mathbf{j}_k^{\mathrm{H}}\left(\mathbf{J}_{\mathcal{S},\mathcal{S}}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}|}\right)^{-1}\mathbf{j}_k\right)^{-1},\\ \mathbf{u}_k &= \nu^{-1}\nu_k\left(\mathbf{h}_k^{*}-\mathbf{H}^{\mathrm{H}}_{\mathcal{S},:}\left(\mathbf{J}_{\mathcal{S},\mathcal{S}}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}|}\right)^{-1}\mathbf{j}_k\right).\end{aligned} \tag{38}$$

Inserting (36) and (37) into (35), Δ_k can be simplified as

$$\Delta_k = L\ln\frac{\nu_k}{\tau}+\frac{\mathbf{u}_k^{\mathrm{H}}\mathbf{u}_k}{\nu_k}+\ln\frac{\lambda}{1-\lambda}. \tag{39}$$

Given that $\mathbf{s}$ is changed into $\mathbf{s}'$, the mean $\widehat{\mathbf{W}}'_{\mathcal{S}',:}$ and covariance $\widehat{\mathbf{C}}_{\mathcal{S}',0}$ of the weights can be updated from (23), i.e.,

$$\widehat{\mathbf{C}}_{\mathcal{S}',0} = \nu(\mathbf{J}_{\mathcal{S}',\mathcal{S}'}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}'|})^{-1}, \tag{40a}$$

$$\widehat{\mathbf{W}}'_{\mathcal{S}',:} = \nu^{-1}\widehat{\mathbf{C}}'_{\mathcal{S}',0}\mathbf{H}_{\mathcal{S}',:}. \tag{40b}$$

In fact, the matrix inversion can be avoided when updating $\widehat{\mathbf{W}}'_{\mathcal{S}',:}$ and $\widehat{\mathbf{C}}_{\mathcal{S}',0}$. It can be shown that

$$\begin{aligned}\widehat{\mathbf{C}}_{\mathcal{S}',0} &= \begin{pmatrix}\widehat{\mathbf{C}}'_{\mathcal{S}'\backslash k,0} & \widehat{\mathbf{c}}'_{k,0}\\ \widehat{\mathbf{c}}'^{\mathrm{H}}_{k,0} & \widehat{C}'_{kk,0}\end{pmatrix} = \nu\begin{pmatrix}\mathbf{J}_{\mathcal{S},\mathcal{S}}+\frac{\nu}{\tau}\mathbf{I}_{|\mathcal{S}|} & \mathbf{j}_k\\ \mathbf{j}_k^{\mathrm{H}} & M+\frac{\nu}{\tau}\end{pmatrix}^{-1}\\ &= \nu\begin{pmatrix}\nu\widehat{\mathbf{C}}^{-1}_{\mathcal{S},0} & \mathbf{j}_k\\ \mathbf{j}_k^{\mathrm{H}} & M+\frac{\nu}{\tau}\end{pmatrix}^{-1}\\ &= \begin{pmatrix}\widehat{\mathbf{C}}_{\mathcal{S},0}+\frac{\nu_k}{\nu^2}\widehat{\mathbf{C}}_{\mathcal{S},0}\mathbf{j}_k\mathbf{j}_k^{\mathrm{H}}\widehat{\mathbf{C}}_{\mathcal{S},0}- & \frac{\nu_k}{\nu}\widehat{\mathbf{C}}_{\mathcal{S},0}\mathbf{j}_k\\ -\frac{\nu_k}{\nu}\mathbf{j}_k^{\mathrm{H}}\widehat{\mathbf{C}}_{\mathcal{S},0} & \nu_k\end{pmatrix}.\end{aligned} \tag{41}$$

Furthermore, the weight $\widehat{\mathbf{W}}'_{\mathcal{S}',:}$ is updated as

$$\begin{aligned}\widehat{\mathbf{W}}'_{\mathcal{S}',:} &= \begin{pmatrix}\widehat{\mathbf{W}}'_{\mathcal{S}'\backslash k,:}\\ \widehat{\mathbf{w}}'^{\mathrm{T}}_k\end{pmatrix} = \nu^{-1}\begin{pmatrix}\widehat{\mathbf{C}}'_{\mathcal{S}'\backslash k,0} & \widehat{\mathbf{c}}'_{k,0}\\ \widehat{\mathbf{c}}'^{\mathrm{H}}_{k,0} & \widehat{C}'_{kk,0}\end{pmatrix}\begin{pmatrix}\mathbf{H}_{\mathcal{S}'\backslash k,:}\\ \mathbf{h}_k^{\mathrm{T}}\end{pmatrix}\\ &= \begin{pmatrix}\widehat{\mathbf{W}}_{\mathcal{S},:}-\nu^{-1}\widehat{\mathbf{C}}_{\mathcal{S},0}\mathbf{j}_k\mathbf{u}_k^{\mathrm{H}}\\ \mathbf{u}_k^{\mathrm{H}}\end{pmatrix}.\end{aligned}$$

It can be seen that after activating the kth component, the posterior mean and variance of $\mathbf{w}_k$ are $\mathbf{u}_k$ and $\nu_k\mathbf{I}_L$, respectively.

For the deactive case with $\mathbf{s}_k = 1$, $\mathbf{s}_k' = 0$ and $\mathcal{S}' = \mathcal{S}\backslash\{k\}$, $\Delta_k = \ln Z(\mathbf{s}') - \ln Z(\mathbf{s})$ is the negative of (39), i.e.,

$$\Delta_k = -L\ln\frac{\nu_k}{\tau}-\frac{\mathbf{u}_k^{\mathrm{H}}\mathbf{u}_k}{\nu_k}-\ln\frac{\lambda}{1-\lambda}. \tag{42}$$

Similar to (41), the posterior mean and covariance update equation from $\mathcal{S}'$ to $\mathcal{S}$ case can be rewritten as

$$\begin{pmatrix}\widehat{\mathbf{C}}'_{\mathcal{S}',0}+\frac{\nu_k}{\nu^2}\widehat{\mathbf{C}}'_{\mathcal{S}',0}\mathbf{j}_k\mathbf{j}_k^{\mathrm{H}}\widehat{\mathbf{C}}'_{\mathcal{S}',0} & -\frac{\nu_k}{\nu}\widehat{\mathbf{C}}'_{\mathcal{S}',0}\mathbf{j}_k\\ -\frac{\nu_k}{\nu}\mathbf{j}_k^{\mathrm{H}}\widehat{\mathbf{C}}'_{\mathcal{S}',0} & \nu_k\end{pmatrix} = \begin{pmatrix}\widehat{\mathbf{C}}_{\mathcal{S}\backslash k,0} & \widehat{\mathbf{c}}_{k,0}\\ \widehat{\mathbf{c}}^{\mathrm{H}}_{k,0} & \widehat{C}_{kk,0}\end{pmatrix} \tag{43}$$

$$\begin{pmatrix}\widehat{\mathbf{W}}'_{\mathcal{S}',:}-\nu^{-1}\widehat{\mathbf{C}}'_{\mathcal{S}',0}\mathbf{j}_k\mathbf{u}_k^{\mathrm{H}}\\ \mathbf{u}_k^{\mathrm{H}}\end{pmatrix} = \begin{pmatrix}\widehat{\mathbf{W}}_{\mathcal{S}\backslash k,:}\\ \widehat{\mathbf{w}}_k^{\mathrm{T}}\end{pmatrix}, \tag{44}$$

where $\widehat{\mathbf{c}}_{k,0}$ denotes the column of $\widehat{\mathbf{C}}_{\mathcal{S},0}$ corresponding to the kth component. According to (43) and (44), one has

$$\widehat{\mathbf{C}}'_{\mathcal{S}',0}+\frac{\nu_k}{\nu^2}\widehat{\mathbf{C}}'_{\mathcal{S}',0}\mathbf{j}_k\mathbf{j}_k^{\mathrm{H}}\widehat{\mathbf{C}}'_{\mathcal{S}',0} = \widehat{\mathbf{C}}_{\mathcal{S}\backslash k,0}, \tag{45a}$$

$$-\frac{\nu_k}{\nu}\widehat{\mathbf{C}}'_{\mathcal{S}',0}\mathbf{j}_k = \widehat{\mathbf{c}}_{k,0} \tag{45b}$$

$$\nu_k = \widehat{C}_{kk,0}, \tag{45c}$$

$$\widehat{\mathbf{W}}'_{\mathcal{S}',:} - \nu^{-1}\widehat{\mathbf{C}}'_{\mathcal{S}',0}\mathbf{j}_k\mathbf{u}_k^{\mathrm{H}} = \widehat{\mathbf{W}}_{\mathcal{S}\backslash k,:}, \tag{45d}$$

$$\mathbf{u}_k^{\mathrm{H}} = \widehat{\mathbf{w}}_k^{\mathrm{T}}. \tag{45e}$$

Thus, $\widehat{\mathbf{C}}'_{\mathcal{S}',0}$ can be updated by substituting (45b) and (45c) in (45a), i.e.,

$$\widehat{\mathbf{C}}'_{\mathcal{S}',0} = \widehat{\mathbf{C}}_{\mathcal{S}\backslash k,0} - \frac{\nu_k}{\nu^2}\widehat{\mathbf{C}}'_{\mathcal{S}',0}\mathbf{j}_k\mathbf{j}_k^{\mathrm{H}}\widehat{\mathbf{C}}'_{\mathcal{S}',0} = \widehat{\mathbf{C}}_{\mathcal{S}\backslash k,0} - \frac{\widehat{\mathbf{c}}_{k,0}\widehat{\mathbf{c}}_{k,0}^{\mathrm{H}}}{\widehat{C}_{kk,0}}. \tag{46}$$

Similarly, $\widehat{\mathbf{W}}'_{\mathcal{S}',:}$ can be updated by substituting (45b) and (45e) in (45d), i.e.,

$$\widehat{\mathbf{W}}'_{\mathcal{S}',:} = \nu^{-1}\widehat{\mathbf{C}}'_{\mathcal{S}',0}\mathbf{j}_k\mathbf{u}_k^{\mathrm{H}} + \widehat{\mathbf{W}}_{\mathcal{S}\backslash k,:} = \widehat{\mathbf{W}}_{\mathcal{S}\backslash k,:} - \frac{\widehat{\mathbf{c}}_{k,0}}{\widehat{C}_{kk,0}}\widehat{\mathbf{w}}_k^{\mathrm{T}}. \tag{47}$$

According to $\nu_k = \widehat{C}_{kk,0}$ (45c) and $\mathbf{u}_k^{\mathrm{H}} = \widehat{\mathbf{w}}_k^{\mathrm{T}}$ (45e), Δ_k (42) can be simplified as

$$\Delta_k = -L\ln\frac{\widehat{C}_{kk,0}}{\tau} - \frac{\mathbf{w}_k^{\mathrm{H}}\mathbf{w}_k}{\widehat{C}_{kk,0}} - \ln\frac{\lambda}{1-\lambda}. \tag{48}$$

A.3. Estimation of model parameters

Plugging the postulated PDF (13) in (12), one has

$$\begin{aligned}
\mathcal{L}(q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y});\boldsymbol{\beta}) &= \mathrm{E}_{q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y})}\left[\ln\frac{p(\mathbf{Y},\boldsymbol{\theta},\mathbf{W},\mathbf{s};\boldsymbol{\beta})}{q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y})}\right] \\
&= \mathrm{E}_{q(\boldsymbol{\theta},\mathbf{W},\mathbf{S}|\mathbf{Y})}\left[\sum_{i=1}^{N}\ln p(s_i) + \ln p(\mathbf{W}|\mathbf{s}) + \ln p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{W})\right] + \mathrm{const} \\
&= ||\widehat{\mathbf{s}}||_0\ln\lambda - ||\widehat{\mathbf{s}}||_0\ln(1-\lambda) + ||\widehat{\mathbf{s}}||_0 L\ln\frac{1}{\pi\tau} \\
&\quad - \mathrm{E}_{q(\mathbf{W}|\mathbf{Y})}\left[\frac{1}{\tau}\mathrm{tr}(\mathbf{W}_{\widehat{\mathcal{S}},:}\mathbf{W}^{\mathrm{H}}_{\widehat{\mathcal{S}},:})\right] + \mathrm{const} \\
&\quad + ML\ln\frac{1}{\pi\nu} - \frac{1}{\nu}\mathrm{tr}(\mathbf{Y}^{\mathrm{H}}\mathbf{Y}) + \frac{2}{\nu}\mathrm{Re}\{\mathrm{tr}(\widehat{\mathbf{W}}^{\mathrm{H}}_{\widehat{\mathcal{S}},:}\mathbf{H}_{\widehat{\mathcal{S}},:})\} \\
&\quad - \frac{1}{\nu}\mathrm{E}_{q(\mathbf{W}|\mathbf{Y})}[\mathrm{tr}(\mathbf{W}^{\mathrm{H}}_{\widehat{\mathcal{S}},:}\mathbf{J}_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}\mathbf{W}_{\widehat{\mathcal{S}},:})].
\end{aligned}$$

Substituting $\mathrm{E}_{q(\mathbf{W}|\mathbf{Y})}[\mathrm{tr}(\mathbf{W}_{\widehat{\mathcal{S}},:}\mathbf{W}^{\mathrm{H}}_{\widehat{\mathcal{S}},:})] = \mathrm{tr}(\widehat{\mathbf{W}}^{\mathrm{H}}_{\widehat{\mathcal{S}},:}\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}) + L\mathrm{tr}(\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0})$ and $\mathrm{E}_{q(\mathbf{W}|\mathbf{Y})}[\mathrm{tr}(\mathbf{W}^{\mathrm{H}}_{\widehat{\mathcal{S}},:}\mathbf{J}_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}\mathbf{W}_{\widehat{\mathcal{S}},:})] = \mathrm{tr}(\mathbf{J}_{\widehat{\mathcal{S}},\widehat{\mathcal{S}}}(\widehat{\mathbf{W}}_{\widehat{\mathcal{S}},:}\widehat{\mathbf{W}}^{\mathrm{H}}_{\widehat{\mathcal{S}},:} + L\widehat{\mathbf{C}}_{\widehat{\mathcal{S}},0}))$ in the above equation, $\mathcal{L}(q(\boldsymbol{\theta},\mathbf{W},\mathbf{s}|\mathbf{Y});\boldsymbol{\beta})$ is obtained as (27).

References

[1] P. Stoica, R.L. Moses, Spectral Analysis of Signals, Upper Saddle River, NJ, USA: Prentice-Hall, 2005.

[2] T.L. Hansen, P.B. Jørgensen, M.A. Badiu, B.H. Fleury, An iterative receiver for OFDM with sparsity-based parametric channel estimation, IEEE Trans. Signal Process. 66 (20) (2018) 5454–5469.

[3] B. Ottersten, M. Viberg, T. Kailath, Analysis of subspace fitting and ML techniques for parameter estimation from sensor array data, IEEE Trans. Signal Process. 40 (1992) 590–600.

[4] Z. Z. Yang, J. Li, P. P. Stoica, L. Xie, Sparse methods for direction-of-arrival estimation, Acad. Press Lib. Signal Process. 7 (2018) 509–581.

[5] D. Malioutov, M. Cetin, A. Willsky, A sparse signal reconstruction perspective for source localization with sensor arrays, IEEE Trans. Signal Process. 53 (8) (2005). 3010–2022.

[6] P. Stoica, P. Babu, J. Li, New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data, IEEE Trans. Signal Process. 59 (1) (2011a) 35–47.

[7] P. Stoica, P. Babu, J. Li, SPICE: A sparse covariance-based estimation method for array processing, IEEE Trans. Signal Process. 59 (2) (2011b) 629–638.

[8] P. Stoica, P. Babu, SPICE and LIKES: two hyperparameter-free methods for sparse-parameter estimation, Signal Process. 92 (7) (2012) 1580–1590.

[9] P. Gerstoft, C.F. Mecklenbräuker, A. Xenaki, S. Nannuru, Multisnapshot sparse Bayesian learning for DOA, IEEE Signal Process. Lett. 23 (10) (2016) 1469–1473.

[10] B. Mamandipoor, D. Ramasamy, U. Madhow, Newtonized orthogonal matching pursuit: frequency estimation over the continuum, IEEE Trans. Signal Process. 64 (19) (2016) 5066–5081.

[11] J. Zhu, L. Han, R.S. Blum, Z. Xu, Newtonized orthogonal matching pursuit for line spectrum estimation with multiple measurement vectors, 2018,. arXiv: 1802.01266.

[12] J. Fang, F. Wang, Y. Shen, H. Li, R.S. Blum, Superresolution compressed sensing for line spectral estimation: an iterative reweighted approach, IEEE Trans. Signal Process. 64 (18) (2016) 4649–4662.

[13] F. Wang, J. Fang, H. Li, Prior knowledge aided super-resolution line spectral estimation: an iterative reweighted algorithm, ICASSP (2017).

[14] L. Hu, J. Zhou, Z. Shi, Q. Fu, A fast and accurate reconstruction algorithm for compressed sensing of complex sinusoids, IEEE Trans. Signal Process. 61 (22) (2013) 5744–5754.

[15] L. Hu, Z. Shi, J. Zhou, Q. Fu, Compressed sensing of complex sinusoids: an approach based on dictionary refinement, IEEE Trans. Signal Process. 60 (7) (2012) 3809–3822.

[16] M.E. Tipping, Sparse bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (2001) 211–244.

[17] D.P. Wipf, B.D. Rao, Sparse bayesian learning for basis selection, IEEE Trans. Signal Process. 52 (8) (2004) 2153–2164.

[18] D. Shutin, B.H. Fleury, Sparse variational Bayesian SAGE algorithm with application to the estimation of multipath wireless channels, IEEE Trans. Signal Process. 59 (8) (2011) 3609–3623.

[19] T.L. Hansen, M.A. Badiu, B.H. Fleury, B.D. Rao, A sparse bayesian learning algorithm with dictionary parameter estimation, in Proc. IEEE 8th Sensor Array Multichannel Signal Process. Workshop (2014) 385–388.

[20] T.L. Hansen, B.H. Fleury, B.D. Rao, Superfast line spectral estimation, IEEE Trans. Signal Process. 66 (10) (2018) 2511–2526.

[21] G. Tang, B. Bhaskar, P. Shah, B. Recht, Compressed sensing off the grid, IEEE Trans. Inf. Theory 59 (11) (2013) 7465–7490.

[22] B. Bhaskar, G. Tang, B. Recht, Atomic norm denoising with applications to line spectral estimation, IEEE Trans. Signal Process. 61 (23) (2013) 5987–5999.

[23] Z. Yang, L. Xie, On gridless sparse methods for line spectral estimation from complete and incomplete data, IEEE Trans. Signal Process. 63 (12) (2015) 3139–3153.

[24] Z. Yang, L. Xie, C. Zhang, A discretization-free sparse and parametric approach for linear array signal processing, IEEE Trans. Signal Process. 62 (19) (2014) 4959–4973.

[25] Z. Yang, L. Xie, Continuous compressed sensing with a single or multiple measurement vectors, IEEE Workshop on Statistical Signal Processing (2014) 288–291.

[26] Y. Li, Y. Chi, Off-the-grid line spectrum denoising and estimation with multiple measurement vectors, IEEE Trans. Signal Process. 64 (5) (2016) 1257–1269.

[27] X. Angeliki, P. Gerstoft, Grid-free compressive beamforming, J Acoust. Soc. Am. 137 (2015) 1923–1935.

[28] Y. Park, P. Gerstoft, W. Seong, Grid-free compressive mode extraction, J Acoust. Soc. Am. 145 (2019) 1427–1442.

[29] Y. Chen, Y. Chi, Robust spectral compressed sensing via structured matrix completion, IEEE Trans. Inf. Theory 60 (10) (2014) 6576–6601.

[30] Z. Yang, L. Xie, Enhancing sparsity and resolution via reweighted atomic norm minimization, IEEE Trans. Signal Process. 64 (4) (2016) 995–1006.

[31] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, New York, 2004.

[32] M.A. Badiu, T.L. Hansen, B.H. Fleury, Variational bayesian inference of line spectral, IEEE Trans. Signal Process. 65 (9) (2017) 2247–2261.

[33] D. Zachariah, P. Wirfält, M. Jansson, S. Chatterjee, Line spectrum estimation with probabilistic priors, Signal Processing 93 (11) (2013) 2969–2974.

[34] K.V. Mardia, P.E. Jupp, Directional Statistics, Wiley, New York, NY, USA, 2000.

[35] K.P. Murphy, Machine Learning a Probabilistic Perspective, MIT Press, Cambridge, USA, 2012.

[36] D.P. Bertsekas, J.N. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods, Athenan Scientific, Massachusetts, 1997.

[37] C. Qian, L. Huang, N.D. Sidiropoilos, H.C. So, Enhanced PUMA for direction-of-arrival estimation and its performance analysis, IEEE Trans. Signal Process. 64 (16) (2016) 4127–4137.

[38] C.F. Mecklenbrauker, P. Gerstoft, A. Panahi, M. Viberg, Sequential Bayesian sparse signal reconstruction using array data, IEEE Trans. Signal Process. 61 (24) (2013) 6344–6354.