Variational Bayesian Inference of Line Spectra

Badiu, Mihai Alin; Hansen, Thomas Lundgaard; Fleury, Bernard Henri

Published in:
I E E E Transactions on Signal Processing

DOI (link to publication from Publisher):
10.1109/TSP.2017.2655489

Publication date:
2017

Document Version
Accepted author manuscript, peer reviewed version

Variational Bayesian Inference of Line Spectra

Mihai-Alin Badiu, Thomas Lundgaard Hansen, and Bernard Henri Fleury

***Abstract*—In this paper, we address the fundamental problem of line spectral estimation in a Bayesian framework. We target model order and parameter estimation via variational inference in a probabilistic model in which the frequencies are continuous-valued, i.e., not restricted to a grid; and the coefficients are governed by a Bernoulli-Gaussian prior model turning model order selection into binary sequence detection. Unlike earlier works which retain only point estimates of the frequencies, we undertake a more complete Bayesian treatment by estimating the posterior probability density functions (pdfs) of the frequencies and computing expectations over them. Thus, we additionally capture and operate with the uncertainty of the frequency estimates. Aiming to maximize the model evidence, variational optimization provides analytic approximations of the posterior pdfs and also gives estimates of the additional parameters. We propose an accurate representation of the pdfs of the frequencies by mixtures of von Mises pdfs, which yields closed-form expectations. We define the algorithm VALSE in which the estimates of the pdfs and parameters are iteratively updated. VALSE is a gridless, convergent method, does not require parameter tuning, can easily include prior knowledge about the frequencies and provides approximate posterior pdfs based on which the uncertainty in line spectral estimation can be quantified. Simulation results show that accounting for the uncertainty of frequency estimates, rather than computing just point estimates, significantly improves the performance. The performance of VALSE is superior to that of state-of-the-art methods and closely approaches the Cramér-Rao bound computed for the true model order.**

***Index Terms*—Line spectral estimation, complex sinusoids, model order selection, Bayesian inference, von Mises distribution, super-resolution, Bernoulli-Gaussian model, sparse estimation**

I. Introduction

The problem of line spectral estimation (LSE) [1], i.e. extracting the parameters of a superposition of complex exponential functions from noisy measurements is fundamental in numerous disciplines in engineering, physics, and natural sciences. To quote a few examples, solutions to this problem have applications to range and direction estimation in sonar and radar, channel estimation in wireless communications, speech analysis, spectroscopy, molecular dynamics, power electronics, geophysical exploration.

In LSE, the original signal $\mathbf{x} = (x_0, \dots, x_{N-1})^{\mathrm{T}} \in \mathbb{C}^N$ is a superposition of K complex sinusoids, i.e.,

$$x_n = \sum_{k=1}^{K} \alpha_k e^{j\omega_k n}, \tag{1}$$

where $\alpha_k \in \mathbb{C}$ and $\omega_k \in [-\pi, \pi)$ are the complex amplitude and (angular) frequency, respectively, of the kth component. We are given the vector $\mathbf{y}$ containing $M \leq N$ noisy measurements of those components of $\mathbf{x}$ with indices in $\mathcal{M} \subseteq \{0, \dots, N-1\}$, $|\mathcal{M}| = M$. Defining the function $\mathbf{a} : [-\pi, \pi) \to \mathbb{C}^M$, $\omega \to \mathbf{a}(\omega) = (e^{j\omega m} \mid m \in \mathcal{M})^{\mathrm{T}}$ and the vector $\boldsymbol{\epsilon}$ representing additive noise, we write

$$\mathbf{y} = \sum_{k=1}^{K} \alpha_k \mathbf{a}(\omega_k) + \boldsymbol{\epsilon}. \tag{2}$$

The problem of LSE involves estimating the number K of sinusoidal components, also referred to as model order selection, and their associated parameters α_k and ω_k. Even if the model order K is given, LSE is still nontrivial because of the nonlinear dependency of (2) on the frequencies.[1]

A. Prior Work

Under the assumption of known K, the ω_k's are traditionally estimated using the maximum-likelihood (ML) technique or subspace methods, such as [2], [3]. The ML method involves the hard task of maximizing a nonconvex function that has a multimodal shape with a sharp global maximum. The maximizer is typically searched using iterative algorithms (e.g., [4]–[6]) which, however, require accurate initialization and, at best, are guaranteed to converge to a local optimum. Nonetheless, the performance of the ML technique is superior to that of subspace methods, the difference being evident especially when the sample size M or alternatively the signal-to-noise ratio (SNR) are small. Since K is typically unknown in practice, the model order is conventionally selected based on an information criterion, which comprises a data term representing the fitting error and a penalty term that increases with the model order (see [7] and references therein). Assuming a range of potential model orders, the parameters corresponding to each possible order are estimated using, e.g., one of the aforementioned methods. Finally, the tradeoff between fitting error and model complexity is made by selecting the configuration that minimizes the criterion. Scanning a range of model orders can be computationally expensive. Also, in non-asymptotic regimes (particularly limited M or SNR), information criteria tend to provide a wrong model order. A comprehensive review of classical approaches can be found in [1].

A more recent approach to LSE is dictionary-based model estimation, see [8] and the references therein. In this approach, nonlinear estimation of the frequencies is avoided by discretizing the range $[-\pi, \pi)$ into a finite set (grid) of samples that represent the candidate frequency estimates. The signal model (2) is then approximated with a linear system comprising a so-called dictionary matrix (whose columns are

This work was supported by the research project VIRTUOSO (funded by Intel Mobile Communications, Anite, Telenor, Aalborg University, and the Danish National Advanced Technology Foundation) and the Danish Council for Independent Research under grant IDs DFF–5054-00212 and DFF–4005-00549.

The authors are with the Department of Electronic Systems, Aalborg University, Denmark (e-mail: {mib,tlh,bfl}@es.aau.dk).

[1]When K and the frequencies are given, the complex amplitudes can be easily estimated with the linear least-squares method.

given by $\mathbf{a}(\cdot)$ evaluated at the grid samples) and a vector of weights. Thus, the original nonlinear problem is replaced by a linear inverse problem to which a sparse solution is sought. The nonzero entries of the sparse estimate of the weight vector encode the model order and parameter estimates. There is a plethora of techniques that can be used for sparse signal representation, see the detailed survey [9]. However, restricting the candidate frequency estimates to a discrete grid induces spectral leakage due to the model mismatch. Consequently, $\mathbf{x}$ can admit only an approximately sparse representation (or may be even incompressible) in a finite dictionary [10], [11]. On the one hand, a denser grid provides a better sparse approximation and higher accuracy of frequency estimation. On the other hand, increasing the grid density makes the dictionary columns highly coherent, which might affect the sparse reconstruction capability, and boosts the computational complexity. To alleviate the mismatch issues, several approaches are conceived, e.g.: in [11], the concept of structured sparsity is utilized to inhibit closely-spaced frequency estimates; the method in [12] starts with a coarse grid and heuristically iterates between estimating the weights and placing a finer grid around the location of the non-zero weight estimates; in [13]–[16], a less fine grid is used as a baseline and the dictionary matrix is modified to include auxiliary interpolation functions.

In the quest for gridless methods which work directly with continuously parameterized dictionaries, i.e., dictionaries whose parameter ranges in $[-\pi, \pi)$, several works depart from using a static dictionary given by a fixed grid. By including the parameters that dictate the dictionary in the estimation problem, they obtain dynamic dictionary algorithms in which the candidate frequencies and hence the dictionary columns are gradually refined. In [8], two such algorithms are designed based on the ℓ_p regularized least squares objective by adding a penalty term to prohibit closely spaced frequencies and respectively imposing a hard constraint on the minimum distance between frequencies. The algorithms approximately solve the involved nonlinear estimation and still require an initial grid [8]. A different line of works adopts the Bayesian framework and augments the probabilistic model of sparse Bayesian learning (SBL) [17], [18] to incorporate the candidate frequencies. In SBL, a sparse weight vector is promoted by selecting a parameterized/hierarchical prior model for its entries [17], [18]. Estimation in the augmented model is performed using variational inference methods [19]–[21] or maximization of the marginalized posterior pdf [22]. Common to all existing SBL-based approaches is that they restrict to compute point estimates of the frequencies (i.e., MAP/ML estimates), which implies nontrivial maximization of highly multimodal functions (similar to classical ML frequency estimation) in each iteration. The maximization is accomplished approximately by using a grid followed by refinement with Newton's method or interpolation. Another limitation is that, while providing good reconstruction performance, the SBL-based methods reportedly overestimate the model order, i.e., they consistently output additional spurious components (artifacts) of small power [19], [21].

A different gridless approach that avoids the frequency discretization issues is based on the atomic norm (equivalently, the total variation norm), which is the continuous analog of the ℓ_1 norm and allows for working with an infinite, continuous dictionary. In this way, it is shown that for the noiseless case the frequencies can be perfectly recovered from complete data ($M = N$) [23] or incomplete data ($M < N$) [24], as long as they are well separated. In [25], the atomic norm soft thresholding (AST) method, which solves a convex program, is proposed for LSE from noisy, complete data. AST is generalized to the incomplete data case in [26]. Given that AST requires knowledge of the noise variance, the grid-based SPICE method [27] (which minimizes a covariance matrix fitting criterion) is extended in [26] to gridless SPICE (GLS). GLS is applicable to both complete and incomplete data cases without knowledge of the noise power and is equivalent to atomic-norm-based methods; however, it has the limitations of frequency splitting and inaccurate model order estimation [26]. To overcome the two drawbacks, a GLS-based framework is proposed in [26], in which: GLS is used as a method to estimate the covariance matrix of $\mathbf{y}$, based on which the model order is selected using the SORTE algorithm [28] and the frequencies are estimated with MUSIC [2].

An important limitation of atomic-norm-based techniques is that they require the frequencies be sufficiently separated in order to be recovered. Enhanced matrix completion (EMaC) [29] and reweighted atomic-norm minimization (RAM) [30] are two recent algorithms that are reported to improve the resolution capability of atomic-norm methods.

B. Contribution

In this paper, we propose a method for LSE from the measurement model (2) by following the approach of sparse Bayesian inference including estimation of continuous-valued frequencies. The key development that sets our work apart from the related methods [19]–[22] is that, instead of retaining only point estimates of the frequencies, we seek a more complete Bayesian treatment by estimating pdfs of the frequencies and computing expectations over them. Our basic motivation is that, in general, a fully Bayesian approach is expected to show benefits, especially in the situations where sample sizes or SNRs are limited. The fully Bayesian approach naturally allows for representing and operating with the uncertainty of the frequency estimates, in addition to only that of the weights as considered so far. In particular, our approach involves computing expectations of $e^{jn\Theta}$, rather than just evaluating the phasor at a certain point estimate. The uncertainty impacts all other estimates and also the criterion for accepting a component in the estimated model (through the estimates involved) and therefore the model order estimate. Our results show that accounting for the uncertainty of frequency estimates with the fully Bayesian approach proves to be essential for improving model estimation performance. A second distinction from related works is that we employ a Bernoulli-Gaussian hierarchical model for the weights [31] instead of the typical SBL prior model [17], [18]. By analyzing the component-acceptance criteria induced by the two models, we observe that the Bernoulli-Gaussian model is more resilient to insertion of small spurious components.

We provide our probabilistic formulation of LSE in Section II. Since exact inference in the proposed model requires computations that do not admit closed-form analytical expressions, we take the variational approach[2] to: compute approximate posterior pdfs of the frequencies and weights; attempt MAP detection of the binary vector of the hierarchical model; and target ML estimation of the noise variance and parameters of the Bernoulli-Gaussian model. The variational optimization problem consists in maximizing a lower bound on the model evidence over the pdfs and parameters of interest. In Section III, we derive implicit expressions for local maximizers, which are to be updated iteratively. To enable closed-form expectations over the approximate pdfs of the frequencies, we show in Section IV that these pdfs can be very well represented by mixtures of von Mises pdfs (see also Appendices B and C). In Section V, we propose a specific initialization and schedule of iterations that define the variational LSE (VALSE) algorithm. VALSE has several attractive features: it is fully automated (i.e., does not include parameters to be tuned, as all necessary parameters are learned from the data); it converges because each step increases the lower bound on the model evidence; it has the ability to easily incorporate prior knowledge about the frequencies (through a von Mises pdf or a mixture of such pdfs); it provides posterior distributions based on which uncertainty in LSE can be quantified. In Section VI the performance of VALSE is evaluated and compared against state-of-art methods through computer simulations. Finally, Section VII concludes the paper.

II. Bayesian Formulation and Variational Approach

Given the difficulty of not knowing the model order K in (2), for the design of our Bayesian estimator we propose a probabilistic model consisting of a superposition of N (i.e. the dimension of the original signal $\mathbf{x}$ in (1)) complex sinusoids that have random frequencies and weights. Since we want that eventually only K of those components have nonzero weights, we use a sparsity-promoting prior model for the weights. Inference in the following model ideally would recover the K true frequencies and corresponding nonzero weights and yield zero weights for the excessive $N-K$ components. Concretely, we assume that the measurement vector $\mathbf{y}$ is a realization of a random process described by

$$\boldsymbol{Y} = \sum_{i=1}^{N} W_i \mathbf{a}(\Theta_i) + \boldsymbol{U}. \tag{3}$$

The complex weights $\boldsymbol{W} = [W_1, \ldots, W_N]^{\mathrm{T}}$ are governed by independent Bernoulli variables $\boldsymbol{S} = [S_1, \ldots, S_N]^{\mathrm{T}}$ such that the elements of $\boldsymbol{W} \mid \boldsymbol{S}$ are independent and (S_i, W_i) has a Bernoulli-Gaussian distribution. That is,

$$p_{W_i|S_i}(w_i \mid s_i; \tau) = (1-s_i)\delta(w_i) + s_i f_{\mathrm{CN}}(w_i; 0, \tau) \tag{4}$$

and $p_{S_i}(s_i) = \rho^{s_i}(1-\rho)^{(1-s_i)}$. Since $S_i = 0$ implies that $W_i = 0$, the probability ρ controls how likely it is for the ith component to be "active" (i.e. its weight to be nonzero). In (4), $W_i \mid S_i = 1$ has a zero-mean Gaussian pdf with variance τ.[3] In this paper, $f_{\mathrm{CN}}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the complex univariate/multivariate Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The frequencies $\boldsymbol{\Theta} = [\Theta_1, \ldots, \Theta_N]^{\mathrm{T}}$ have the prior pdf $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \prod_i p_{\Theta_i}(\theta_i)$. As justified in Section IV, p_{Θ_i} is a von Mises pdf, or a mixture of such pdfs if one wants to model a more sophisticated, possibly multimodal distribution; the lack of prior knowledge can be represented by setting the concentration parameter of the von Mises pdf to zero. We assume that the components of the noise $\boldsymbol{U}$ are iid complex Gaussian with mean zero and variance ν, which gives the likelihood

$$p_{\boldsymbol{Y}|\boldsymbol{\Theta},\boldsymbol{W}}(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{w}; \nu) = f_{\mathrm{CN}}(\mathbf{y}; \sum_i w_i \mathbf{a}(\theta_i), \nu \mathbf{I}). \tag{5}$$

The model parameters are collectively denoted by $\boldsymbol{\beta} = \{\nu, \rho, \tau\}$.

We can relate model (3) to a sparse approximation problem in which, given the frequencies $\boldsymbol{\Theta} = \boldsymbol{\theta}$, $\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1) \ldots \mathbf{a}(\theta_N)]$ is the dictionary matrix and we need to infer the weights $\boldsymbol{W}$ from $M \leq N$ data samples. Using sparsity-promoting hierarchical models for $\boldsymbol{W}$ is a common Bayesian approach to find sparse solutions to ill-posed problems in compressed sensing. While the Bayesian treatment of LSE [19]–[22] typically uses the SBL prior model [17], [18], the Bernoulli-Gaussian model [31], [33] has not been used in the LSE context before. In the Bernoulli-Gaussian model, the binary vector $\boldsymbol{S} = [S_1, \ldots, S_N]^{\mathrm{T}}$ represents the support of the weights $\boldsymbol{W}$. Contrary to the standard sparse estimation problem, in our context the dictionary is parameterized by the frequencies that are to be inferred as well.

We would like to compute mean and circular mean estimates of $\boldsymbol{W}$ and $\boldsymbol{\Theta}$, respectively, based on the posterior pdf

$$p_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}(\boldsymbol{\theta}, \mathbf{w}, \mathbf{s} \mid \mathbf{y}; \boldsymbol{\beta}) = \frac{p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}(\mathbf{y}, \boldsymbol{\theta}, \mathbf{w}, \mathbf{s}; \boldsymbol{\beta})}{p_{\boldsymbol{Y}}(\mathbf{y}; \boldsymbol{\beta})}. \tag{6}$$

In (6), the joint pdf in the numerator is the likelihood (5) times the prior pdfs defined above, i.e.

$$\begin{aligned} &p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}(\mathbf{y}, \boldsymbol{\theta}, \mathbf{w}, \mathbf{s}; \boldsymbol{\beta}) \\ &= p_{\boldsymbol{Y}|\boldsymbol{\Theta},\boldsymbol{W}}(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{w}; \nu) \prod_{i=1}^{N} p_{\Theta_i}(\theta_i) p_{W_i|S_i}(w_i \mid s_i) p_{S_i}(s_i), \end{aligned} \tag{7}$$

while the denominator $p_{\boldsymbol{Y}}(\mathbf{y}; \boldsymbol{\beta})$, called the model evidence (or marginal likelihood of $\boldsymbol{\beta}$), is the marginal of the joint pdf and acts as a normalizing constant. Fig.1 illustrates the factor graph representation of (7). The sought estimates unfortunately require operations (high-dimensional integrals, summation over 2^N possible values of $\mathbf{s}$) that cannot be performed analytically. Therefore we use variational inference to compute

[2] Variational methods are deterministic inference techniques which provide analytical approximations of posterior pdfs, unlike the stochastic method of Markov chain Monte Carlo (MCMC) sampling. The convergence of MCMC methods can be prohibitively slow and difficult to diagnose. MCMC sampling has been previously used for LSE, see [32] and the references therein.

[3] While $p_{W_i|S_i}(w_i \mid s_i = 1)$ should model some prior knowledge about the amplitudes, for the design of our estimator we select a zero-mean Gaussian pdf mainly for computational convenience (see Sec. III-B). In fact, in the simulation experiments we generate the complex amplitudes in (1) from a distribution different than Gaussian.

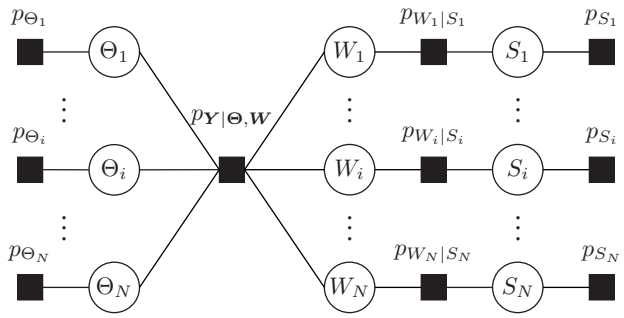

Fig. 1. Factor graph representation of the joint pdf (7).

a surrogate pdf $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$ that should approximate (6) well and at the same time enable tractable estimation.

The variational approach builds on the fact that, for any postulated pdf $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$, the log model evidence can be expressed as [34, Ch. 10]

$$\ln p_{\boldsymbol{Y}}(\mathbf{y};\boldsymbol{\beta}) = \mathcal{D}_{\mathrm{KL}}(q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}} || p_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}) + \mathcal{L}(q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}). \tag{8}$$

The first term in (8) is the Kullback-Leibler divergence of $p_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$ from $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$,[4] while the functional $\mathcal{L}$ reads

$$\mathcal{L}(q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}) = \mathrm{E}_{q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}}\left[\ln \frac{p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}(\mathbf{y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S};\boldsymbol{\beta})}{q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}(\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S} \mid \mathbf{y})}\right]. \tag{9}$$

Given that $p_{\boldsymbol{Y}}(\mathbf{y};\boldsymbol{\beta})$ is constant w.r.t. $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$ and $\mathcal{D}_{\mathrm{KL}} \geq 0$, minimizing the divergence is equivalent to maximizing $\mathcal{L}$ and tightening it as a lower bound to the log model evidence. The KL divergence vanishes only when $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}} = p_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$, in which case $\mathcal{L}$ attains its maximum value, $\ln p_{\boldsymbol{Y}}(\mathbf{y};\boldsymbol{\beta})$. Nonetheless, as we already mentioned, working with the posterior pdf (6) is intractable so we have to restrict the family of candidate pdfs.

We postulate that $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$ factors as

$$\begin{aligned} &q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}(\boldsymbol{\theta},\mathbf{w},\mathbf{s} \mid \mathbf{y}) \\ &= \prod_{i=1}^{N} q_{\Theta_i|\boldsymbol{Y}}(\theta_i \mid \mathbf{y})\, q_{\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{S}}(\mathbf{w} \mid \mathbf{y},\mathbf{s})\, q_{\boldsymbol{S}|\boldsymbol{Y}}(\mathbf{s} \mid \mathbf{y}). \end{aligned} \tag{10}$$

That is, we assume that the frequencies are *a posteriori* independent (mutually and of the other variables).[5] Furthermore, we consider that $q_{\boldsymbol{S}|\boldsymbol{Y}}$ has all its mass at $\hat{\mathbf{s}}$, i.e., $q_{\boldsymbol{S}|\boldsymbol{Y}}(\mathbf{s} \mid \mathbf{y}) = \delta(\mathbf{s} - \hat{\mathbf{s}})$, where the function δ equals 1 when $\mathbf{s} = \hat{\mathbf{s}}$ and 0 otherwise. The simplifying restrictions define a family of pdfs and our goal is to search for the member which maximizes the lower bound $\mathcal{L}$.

The estimates of interest are computed from $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$ as follows. Since Θ_i is an angle, its estimate $\hat{\theta}_i$ is defined so as to give the mean direction of $e^{j\Theta_i}$ [35]:

$$\hat{\theta}_i = \arg\left(\mathrm{E}_{q_{\Theta_i|\boldsymbol{Y}}}[e^{j\Theta_i}]\right), \quad i \in \{1,\ldots,N\}. \tag{11}$$

The estimates $\mathrm{E}_{q_{\Theta_i|\boldsymbol{Y}}}[e^{jn\Theta_i}]$, $n \in \{0,\ldots,N-1\}$ are central in this work. Their magnitudes are ≤ 1 with equality if, and only if $q_{\Theta_i|\boldsymbol{Y}}$ is the Dirac delta distribution. A broad $q_{\Theta_i|\boldsymbol{Y}}$ signifying high uncertainty gives a small magnitude, and vice versa. Those estimates with indices in $\mathcal{M}$ give the elements of $\hat{\mathbf{a}}_i = \mathrm{E}_{q_{\Theta_i|\boldsymbol{Y}}}[\mathbf{a}(\Theta_i)]$; similarly, $\|\hat{\mathbf{a}}_i\|_2^2 \leq M$. The mean and covariance estimates of the weights are defined as

$$\hat{\mathbf{w}} = \mathrm{E}_{q_{\boldsymbol{W}|\boldsymbol{Y}}}[\boldsymbol{W}] \text{ and } \hat{\mathbf{C}} = \mathrm{E}_{q_{\boldsymbol{W}|\boldsymbol{Y}}}\left[\boldsymbol{W}\boldsymbol{W}^{\mathrm{H}}\right] - \hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathrm{H}}. \tag{12}$$

Given that $q_{\boldsymbol{S}|\boldsymbol{Y}} = \delta(\mathbf{s} - \hat{\mathbf{s}})$, the posterior pdf of $\boldsymbol{W}$ is

$$q_{\boldsymbol{W}|\boldsymbol{Y}}(\mathbf{w} \mid \mathbf{y}) = q_{\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{S}}(\mathbf{w} \mid \mathbf{y}, \hat{\mathbf{s}}). \tag{13}$$

Intuitively, the closer $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$ is to $p_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$, the better the estimates (11) and (12) approximate the estimates which we would have computed from (6), if we could. The forms of the pdfs and the support estimate $\hat{\mathbf{s}}$ in the r.h.s. of (10) result from maximizing the lower bound $\mathcal{L}$. When the parameters in $\boldsymbol{\beta}$ are unknown, we target their ML estimates also by maximizing the lower bound to the log marginal likelihood $\ln p_{\boldsymbol{Y}}(\mathbf{y};\boldsymbol{\beta})$.

Finally, based on $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{w}}$, we define the estimates of the quantities in the original superposition (1). Let $\mathcal{S}$ be the set of indices of the non-zero components of $\mathbf{s}$, i.e.

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

Analogously, we define $\hat{\mathcal{S}}$ based on $\hat{\mathbf{s}}$. The estimate of the model order is the cardinality of $\hat{\mathcal{S}}$:

$$\hat{K} = |\hat{\mathcal{S}}|. \tag{14}$$

We define the reconstructed signal $\hat{\mathbf{x}} \triangleq (\hat{x}_1,\ldots,\hat{x}_N)^{\mathrm{T}}$ as the expectation of the signal part in the r.h.s. of (3) over $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$, which gives

$$\hat{x}_n = \sum_{i \in \hat{\mathcal{S}}} \hat{w}_i\, \mathrm{E}_{q_{\Theta_i|\boldsymbol{Y}}}[e^{jn\Theta_i}], \quad n \in \{1,\ldots,N\}. \tag{15}$$

The components of $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{w}}$ with indices in $\hat{\mathcal{S}}$ give the estimates of the frequencies and amplitudes in (1).

III. Solution to the Variational Optimization Problem

We now turn to maximizing the lower bound $\mathcal{L}(q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}})$ in (9) with $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$ of the form (10). Except for restricting $q_{\boldsymbol{S}|\boldsymbol{Y}}$ to give probability one to some sequence $\hat{\mathbf{s}}$, we do not impose any constraints on the forms of the factors in (10). That is, the forms of the approximate posterior pdfs result from variational optimization and are dictated by the likelihood (5) and prior pdfs. As maximizing $\mathcal{L}$ over all factors simultaneously is not viable, we perform alternating optimization: $\mathcal{L}$ is maximized over each of the factors $q_{\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$, $q_{\Theta_i|\boldsymbol{Y}}$, $i = 1,\ldots,N$, in turn while keeping the others fixed. Consequently, the form of each factor is implicit because it depends on the other factors.

Upon their initialization, we iteratively cycle through the factors and replace them one by one with a revised expression. Such a scheme is guaranteed to converge to some local maximum of $\mathcal{L}$ [34, Ch. 10]. In the following we derive the factor expressions that correspond to the fixed-point of the scheme. A specific initialization and scheduling of updates are proposed in Sec. V.

[4] The KL divergence of a pdf p from a pdf q (both defined on some set $\mathcal{X}$) is $\mathcal{D}_{\mathrm{KL}}(q||p) = \int_{\mathcal{X}} q(x) \ln \frac{q(x)}{p(x)}\, \mathrm{d}x$.

[5] The assumed factorization of $q_{\boldsymbol{\Theta}|\boldsymbol{Y}}$ is also referred to as a naïve mean field approximation.

A. Inferring the frequencies $\boldsymbol{\Theta}$

For each $i = 1, \ldots, N$, maximizing $\mathcal{L}$ in (9) w.r.t. the factor $q_{\Theta_i|\boldsymbol{Y}}$ gives [34, Ch. 10, p. 466]

$$\ln q_{\Theta_i|\boldsymbol{Y}}(\theta_i \mid \mathbf{y}) = \mathrm{E}_{\sim\theta_i}\left[\ln p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}(\mathbf{y}, \theta_i, \boldsymbol{\Theta}_{\sim i}, \boldsymbol{W}, \boldsymbol{S}; \boldsymbol{\beta})\right] + \text{const.}$$

where the expectation is taken over $q_{\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}} \prod_{j\neq i} q_{\Theta_j|\boldsymbol{Y}}$, the joint pdf $p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}$ is given by (7) and the constant ensures normalization of the pdf. We further write only the terms that depend on θ_i, i.e.

$$\ln q_{\Theta_i|\boldsymbol{Y}}(\theta_i \mid \mathbf{y}) = \mathrm{E}_{\sim\theta_i}\left[\ln p_{\boldsymbol{Y}|\boldsymbol{\Theta},\boldsymbol{W}}(\mathbf{y} \mid \theta_i, \boldsymbol{\Theta}_{\sim i}, \boldsymbol{W}; \nu)\right] + \ln p_{\Theta_i}(\theta_i) + \text{const.}$$

Plugging the Gaussian form of the likelihood (5) in the above expression and carrying out the required expectations, we finally obtain

$$q_{\Theta_i|\boldsymbol{Y}}(\theta_i \mid \mathbf{y}) \propto p_{\Theta_i}(\theta_i) \exp\left\{\Re\left(\boldsymbol{\eta}_i^{\mathrm{H}} \mathbf{a}(\theta_i)\right)\right\} \tag{16}$$

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

$$\boldsymbol{\eta}_i = \frac{2}{\nu}\left(\mathbf{y} - \sum_{l\in\hat{\mathcal{S}}\setminus\{i\}} \hat{w}_l \hat{\mathbf{a}}_l\right)\hat{w}_i^* - \frac{2}{\nu}\sum_{l\in\hat{\mathcal{S}}\setminus\{i\}} \hat{C}_{l,i}\hat{\mathbf{a}}_l \tag{17}$$

when $i \in \hat{\mathcal{S}}$, and $\boldsymbol{\eta}_i = \mathbf{0}$ otherwise. The second factor in the r.h.s. of (16) is an approximation of the marginal likelihood of θ_i; it is an extremely multimodal function, see Sec. IV. According to (17), the likelihood favors values of θ_i for which the angle between $\hat{w}_i\mathbf{a}(\theta_i)$ and the residual signal (after canceling the interference from the other components) is small.[6] Interestingly, the likelihood corresponds to coherent estimation of Θ_i from the residual signal when the weight is fixed to $\hat{w}_i$. At the same time, it penalizes (to an extent given by the cross-variance of the weights) values that result in small angle between $\mathbf{a}(\theta_i)$ and $\hat{\mathbf{a}}_l$ of the other components in the model. Naturally, when $i \notin \hat{\mathcal{S}}$ (i.e. $\hat{s}_i = 0$), only the prior information comes into play in (16).

The pdf (16) does not yield analytic expressions for $\mathrm{E}_{q_{\Theta_i|\boldsymbol{Y}}}[\mathbf{a}(\Theta_i)]$. In Section IV, we show that $q_{\Theta_i|\boldsymbol{Y}}$ in (16) is well approximated by a mixture of von Mises pdfs, which gives a closed-form approximation of $\mathrm{E}_{q_{\Theta_i|\boldsymbol{Y}}}[\mathbf{a}(\Theta_i)]$.

B. Inferring the weights $\boldsymbol{W}$ and support $\boldsymbol{S}$

We next maximize $\mathcal{L}$ w.r.t. $q_{\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}(\mathbf{w}, \mathbf{s} \mid \mathbf{y})$ when $q_{\Theta_i|\boldsymbol{Y}}$, $i = 1, \ldots, N$, are kept fixed. Since $q_{\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}(\mathbf{w}, \mathbf{s} \mid \mathbf{y})$ is restricted in (10) to give the marginal pmf $q_{\boldsymbol{S}|\boldsymbol{Y}}(\mathbf{s} \mid \mathbf{y}) = \delta(\mathbf{s} - \hat{\mathbf{s}})$, we cannot anymore use the factor-update expression corresponding to free-form optimization [34, Ch. 10, p. 466]. So we will explicitly carry out the maximization of $\mathcal{L}$.

Plugging the postulated pdf (10) in (9) we obtain

$$\mathcal{L}(q_{\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{S}}, \hat{\mathbf{s}}) = \text{const.} - \mathrm{E}_{q_{\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{S}}}\Big\{\ln q_{\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{S}}(\boldsymbol{W} \mid \mathbf{y}, \hat{\mathbf{s}}) - \mathrm{E}_{q_{\boldsymbol{\Theta}|\boldsymbol{Y}}}\left[\ln p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}(\mathbf{y}, \boldsymbol{\Theta}, \boldsymbol{W}, \hat{\mathbf{s}}; \boldsymbol{\beta})\right]\Big\}.$$

[6] The angle ϕ between two complex vectors $\mathbf{u}$ and $\mathbf{v}$ satisfies $\cos(\phi) = \frac{\Re(\mathbf{u}^{\mathrm{H}}\mathbf{v})}{\|\mathbf{u}\|\|\mathbf{v}\|}$.

Let us introduce the pdf

$$t(\mathbf{w}; \hat{\mathbf{s}}) = \frac{1}{Z(\hat{\mathbf{s}})} \exp\left\{\mathrm{E}_{q_{\boldsymbol{\Theta}|\boldsymbol{Y}}}\left[\ln p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}(\mathbf{y}, \boldsymbol{\Theta}, \mathbf{w}, \hat{\mathbf{s}}; \boldsymbol{\beta})\right]\right\}$$

where $p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}$ is given by (7) and $Z(\hat{\mathbf{s}})$ is the normalizing constant obtained by integrating the exponential over $\mathbf{w}$. We can now write

$$\mathcal{L}(q_{\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{S}}, \hat{\mathbf{s}}) = -\mathcal{D}_{\mathrm{KL}}(q_{\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{S}} || t) + \ln Z(\hat{\mathbf{s}}) + \text{const.} \tag{18}$$

Inspecting (18), for any $\hat{\mathbf{s}}$ the maximum of $\mathcal{L}$ over $q_{\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{S}}$ is attained when the KL divergence vanishes. Thus, $\mathcal{L}$ has its maximum at

$$q_{\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{S}}(\mathbf{w} \mid \mathbf{y}, \hat{\mathbf{s}}) = t(\mathbf{w}; \hat{\mathbf{s}}) \quad \text{and} \quad \hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \ln Z(\mathbf{s}). \tag{19}$$

To compute $\mathrm{E}_{q_{\boldsymbol{\Theta}|\boldsymbol{Y}}}\left[\ln p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}(\mathbf{y}, \boldsymbol{\Theta}, \mathbf{w}, \mathbf{s}; \boldsymbol{\beta})\right]$ required for $t(\mathbf{w}; \hat{\mathbf{s}})$ and $Z(\mathbf{s})$ in (19), we use (7), together with (5) and (4), and obtain an expression that is quadratic in $\mathbf{w}$, given that all $p_{W_i|S_i}(w_i \mid s_i = 1)$ are Gaussian. We define the matrix $\mathbf{J}$ with elements $J_{ii} = M$ and $J_{ij} = \hat{\mathbf{a}}_i^{\mathrm{H}}\hat{\mathbf{a}}_j$, $i, j = 1, \ldots, N$, $j \neq i$, and the vector $\mathbf{h} = \left[\hat{\mathbf{a}}_1^{\mathrm{H}}\mathbf{y}, \ldots, \hat{\mathbf{a}}_N^{\mathrm{H}}\mathbf{y}\right]^{\mathrm{T}}$. From (13) and (19), we obtain

$$q_{\boldsymbol{W}|\boldsymbol{Y}}(\mathbf{w} \mid \mathbf{y}) = f_{\mathrm{CN}}\left(\mathbf{w}_{\hat{\mathcal{S}}}; \hat{\mathbf{w}}_{\hat{\mathcal{S}}}, \hat{\mathbf{C}}_{\hat{\mathcal{S}}}\right) \prod_{i\notin\hat{\mathcal{S}}} \delta(w_i),$$

where the mean and covariance matrix of the Gaussian posterior pdf of $\boldsymbol{W}_{\hat{\mathcal{S}}}$ are

$$\hat{\mathbf{w}}_{\hat{\mathcal{S}}} = \nu^{-1}\hat{\mathbf{C}}_{\hat{\mathcal{S}}}\mathbf{h}_{\hat{\mathcal{S}}} \quad \text{and} \quad \hat{\mathbf{C}}_{\hat{\mathcal{S}}} = \nu\left(\mathbf{J}_{\hat{\mathcal{S}}} + \frac{\nu}{\tau}\mathbf{I}\right)^{-1}. \tag{20}$$

The mean is the LMMSE estimate of $\boldsymbol{W}_{\hat{\mathcal{S}}}$ assuming $\mathbf{s} = \hat{\mathbf{s}}$. For $i \notin \hat{\mathcal{S}}$, the measurements are noninformative and, conveniently, $q_{W_i|\boldsymbol{Y}}(w_i \mid \mathbf{y}) = p_{W_i|S_i}(w_i \mid s_i = 0) = \delta(w_i)$, i.e., $\hat{w}_i = 0$.

From (19), the sequence $\hat{\mathbf{s}}$ (which determines $\hat{\mathcal{S}}$) is the maximizer of

$$\ln Z(\mathbf{s}) = \ln\det\left(\mathbf{J}_{\mathcal{S}} + \frac{\nu}{\tau}\mathbf{I}\right)^{-1} + \frac{1}{\nu}\mathbf{h}_{\mathcal{S}}^{\mathrm{H}}\left(\mathbf{J}_{\mathcal{S}} + \frac{\nu}{\tau}\mathbf{I}\right)^{-1}\mathbf{h}_{\mathcal{S}} + \|\mathbf{s}\|_0 \ln\frac{\rho\nu}{(1-\rho)\tau} + \text{const.} \tag{21}$$

Since maximizing the nonlinear function (21) is NP-hard, in Appendix A we propose a suboptimal procedure which is guaranteed to converge to a local maximum of $\ln Z(\mathbf{s})$.

According to Appendix A, a sinusoidal component (we drop the index i for the moment) is admitted only if the posterior mean $\hat{w}$ and variance $\hat{C}$ of its weight (for $\hat{s} = 1$) satisfy

$$\frac{|\hat{w}|^2}{\hat{C}} > \ln\left(\tau/\hat{C}\right) + \ln\frac{1-\rho}{\rho}. \tag{22}$$

It is interesting to relate (22) to the test $|\tilde{w}|^2/\tilde{C} > 1$ obtained in [36] for the SBL prior model of the weights [17], [18], where $\tilde{w}$ and $\tilde{C}$ are the mean and variance of the posterior divided by the prior. The SBL prior model is often used for estimating superimposed signals [19]–[22] and, reportedly, the resulting estimators output additional spurious components (artifacts) of small power. Since $|\tilde{w}|^2/\tilde{C}$ can be viewed as an SNR of the component, the threshold can be heuristically

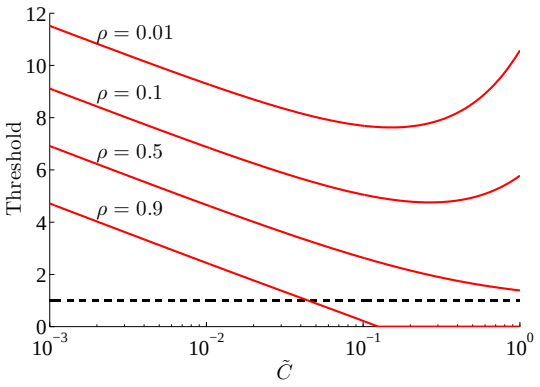

Fig. 2. Activation thresholds vs. weight variance for: the Bernoulli-Gaussian model (solid curves) with $\tau = 1$ and different values of ρ, and for the SBL prior model (dashed line). The activation test is satisfied by the points lying above the given curve.

increased such that a higher SNR is required [19], [36]. For the Bernoulli-Gaussian prior model we express (22) as

$$\frac{|\tilde{w}|^2}{\tilde{C}} > \left(1 + \tilde{C}/\tau\right) \ln\left[\left(1 + \tau/\tilde{C}\right) \frac{1-\rho}{\rho}\right] \qquad (23)$$

where we used $p_{W|S}(w \mid s = 1) = f_{\mathrm{CN}}(w; 0, \tau)$, which gives $\hat{C}^{-1} = \tilde{C}^{-1} + \tau^{-1}$ and $\hat{C}^{-1}\hat{w} = \tilde{C}^{-1}\tilde{w}$. Thus, the threshold (23) is not constant but depends on ρ, τ and also $\tilde{C}$. The latter dependence makes the method more resilient to insertion of artifacts, because, as shown in Fig. 2, the threshold increases with smaller variance, unlike for the SBL model where it stays the same.

C. Estimating the model parameters

The noise variance ν is often unknown in practice. Also, it might be unclear how to set the parameters ρ and τ of the Bernoulli-Gaussian prior model. We show that learning the parameters can be easily included in the variational approach.

The lower bound (9) now additionally depends on $\boldsymbol{\beta} = \{\nu, \rho, \tau\}$. We alternate between maximizing $\mathcal{L}(q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}, \boldsymbol{\beta})$ over $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$ for $\boldsymbol{\beta}$ fixed to $\hat{\boldsymbol{\beta}}$ (according to the previous subsections) and over $\boldsymbol{\beta}$ for fixed $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$. In the latter step,

$$\mathcal{L}(\boldsymbol{\beta}) = \mathrm{E}_{q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}}\left[\ln p_{\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}}(\mathbf{y}, \boldsymbol{\Theta}, \boldsymbol{W}, \boldsymbol{S}; \boldsymbol{\beta})\right] + \text{const.}$$

where we write only the term depending on $\boldsymbol{\beta}$. The joint pdf and the approximate posterior pdf are given by (7) and (10), respectively. Based on the forms of the likelihood (5) and prior pdfs defined in Sec. II, we obtain

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{\nu}\left[2\Re\left(\hat{\mathbf{w}}_{\hat{\mathcal{S}}}^{\mathrm{H}}\mathbf{h}_{\hat{\mathcal{S}}}\right) - \hat{\mathbf{w}}_{\hat{\mathcal{S}}}^{\mathrm{H}}\mathbf{J}_{\hat{\mathcal{S}}}\hat{\mathbf{w}}_{\hat{\mathcal{S}}} - \mathbf{y}^{\mathrm{H}}\mathbf{y} - \mathrm{tr}\left(\mathbf{J}_{\hat{\mathcal{S}}}\hat{\mathbf{C}}_{\hat{\mathcal{S}}}\right)\right] \\ - M\ln\nu - \frac{1}{\tau}\left[\hat{\mathbf{w}}_{\hat{\mathcal{S}}}^{\mathrm{H}}\hat{\mathbf{w}}_{\hat{\mathcal{S}}} + \mathrm{tr}\left(\hat{\mathbf{C}}_{\hat{\mathcal{S}}}\right)\right] - \|\hat{\mathbf{s}}\|_0 \ln\tau \\ + \|\hat{\mathbf{s}}\|_0 \ln\rho + (N - \|\hat{\mathbf{s}}\|_0)\ln(1-\rho) + \text{const.}\end{aligned}$$

We can carry out $\arg\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$ independently over each parameter. Equating the partial derivatives to zero gives unique solutions that correspond to the global maximum (the second-order derivatives are strictly negative). Specifically, we obtain

$$\begin{aligned}\hat{\nu} = \frac{1}{M}\|\mathbf{y} - \sum_{i\in\hat{\mathcal{S}}} \hat{w}_i\hat{\mathbf{a}}_i\|_2^2 + \frac{1}{M}\,\mathrm{tr}\left(\mathbf{J}_{\hat{\mathcal{S}}}\hat{\mathbf{C}}_{\hat{\mathcal{S}}}\right) \\ + \sum_{i\in\hat{\mathcal{S}}} |\hat{w}_i|^2\left(1 - \|\hat{\mathbf{a}}_i\|_2^2/M\right).\end{aligned} \qquad (24)$$

Thus, $\hat{\nu}$ takes into account not only the fitting error, but also the uncertainty of weight estimation (through $\hat{\mathbf{C}}_{\hat{\mathcal{S}}}$) and of frequency estimation (via $\hat{\mathbf{a}}_i$). Regarding the latter, the sharper $q_{\Theta_i|\boldsymbol{Y}}$, the closer $\|\hat{\mathbf{a}}_i\|_2^2$ is to M and therefore the smaller the contribution to $\hat{\nu}$. For ρ and τ we obtain the estimates

$$\hat{\rho} = \frac{\|\hat{\mathbf{s}}\|_0}{N} \qquad \text{and} \qquad \hat{\tau} = \frac{\hat{\mathbf{w}}_{\hat{\mathcal{S}}}^{\mathrm{H}}\hat{\mathbf{w}}_{\hat{\mathcal{S}}} + \mathrm{tr}\left(\hat{\mathbf{C}}_{\hat{\mathcal{S}}}\right)}{\|\hat{\mathbf{s}}\|_0}. \qquad (25)$$

Naturally, $\hat{\rho}$ is given by the number of nonzero components of $\hat{\mathbf{s}}$ and $\hat{\tau}$ is the averaged second-moment of the weights corresponding to those components.

IV. Approximating $q_{\Theta_i|\boldsymbol{Y}}$ by a Mixture of von Mises pdfs

In this section, after providing some background on the von Mises distribution, we show that any pdf of the form $\exp\left(\Re\left(\boldsymbol{\eta}^{\mathrm{H}}\mathbf{a}(\theta)\right)\right)$, such as $q_{\Theta_i|\boldsymbol{Y}}$ in (16), can be well represented by a mixture of von Mises pdfs (MVM). The proposed approximation enables easy computation of expectations over such pdfs. We exploit the MVM approximation in the initialization of our algorithm as well, since the exponential of the periodogram also has the said form.

A. The von Mises distribution

Among the distributions on the unit circle, the von Mises (VM) distribution is of significant importance, its role being similar to that of the Gaussian distribution on the line [35]. The pdf of the VM distribution of a random angle Θ is

$$f_{\mathrm{VM}}(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\theta-\mu)}.$$

The parameters μ and κ are the mean direction and concentration parameter, respectively, and $I_p(\cdot)$ is the modified Bessel function of the first kind and order p. The pdf is symmetrical about its single mode, which is at $\Theta = \mu$. The VM pdf can also be parameterized in terms of $\eta = \kappa e^{j\mu}$:

$$f_{\mathrm{VM}}(\theta; \eta) = \frac{1}{2\pi I_0(|\eta|)} \exp\left(\Re\{\eta^* e^{j\theta}\}\right).$$

The properties of circular distributions are completely determined by the characteristic function, $\varphi_p \triangleq \mathbb{E}[e^{jp\Theta}]$, $p \in \mathbb{Z}$ [35]. The characteristic function of the VM distribution is

$$\varphi_p = e^{jp\mu}\frac{I_p(\kappa)}{I_0(\kappa)}, \quad p \in \mathbb{Z}. \qquad (26)$$

The moments of circular distributions are the moments of $e^{j\Theta}$, i.e., values of the characteristic function. The first moment of the VM distribution, $\varphi_1 = e^{j\mu}A(\kappa)$, gives the mean direction μ and the mean resultant length $A(\kappa) = I_1(\kappa)/I_0(\kappa)$.

The multiplication of two VM pdfs gives

$$f_{\text{VM}}(\theta;\eta_1)f_{\text{VM}}(\theta;\eta_2) \propto f_{\text{VM}}(\theta;\eta) \tag{27}$$

with $\eta = \eta_1+\eta_2$. That is, the result is proportional to a VM pdf with mean direction $\arg(\eta_1+\eta_2)$ and concentration $|\eta_1+\eta_2|$. Thus, the family of VM pdfs is closed under multiplication.

B. The proposed MVM approximation

In the following, we drop the frequency index i for convenience. We write (16) as

$$q_{\Theta|\boldsymbol{Y}}(\theta \mid \mathbf{y}) \propto p_\Theta(\theta) \prod_{m\in\mathcal{M}} \exp\left(\Re\{\eta_m^* e^{jm\theta}\}\right), \tag{28}$$

where the entries of $\boldsymbol{\eta}$ have the polar form $\eta_m = \kappa_m\, e^{j\mu_m}$. When $0 \in \mathcal{M}$ the factor in (28) corresponding to $m = 0$ is a constant, so we can just remove this index from $\mathcal{M}$. Also, when $1 \in \mathcal{M}$, the factor indexed by $m = 1$ has the form of a von Mises (VM) pdf $f_{\text{VM}}(\theta;\eta_1)$ with mean direction μ_1 and concentration κ_1. Furthermore, the factors indexed by $m > 1$ have the form of m-fold wrapped VM pdfs. Thus, we can write (28) as

$$q_{\Theta|\boldsymbol{Y}}(\theta \mid \mathbf{y}) \propto p_\Theta(\theta) \prod_{m\in\mathcal{M}} f_{\text{VM}}(m\theta;\eta_m). \tag{29}$$

In Appendix B we show that a wrapped VM pdf can be very well approximated by an appropriate MVM obtained by matching their characteristic functions. Employing the result (52), we approximate each of the m-fold wrapped VM pdfs in (29) by a mixture of m VM pdfs, i.e.,

$$f_{\text{VM}}(m\theta;\eta_m) \simeq \sum_{r=0}^{m-1} \frac{1}{m} f_{\text{VM}}(\theta;\tilde{\eta}_{m,r}), \tag{30}$$

where $\tilde{\eta}_{m,r} = \tilde{\kappa}_m e^{j\tilde{\mu}_{m,r}}$. The m components of the MVM have equal amplitudes and concentrations. The value $\tilde{\kappa}_m$ of the latter is the solution to

$$\frac{I_m(\tilde{\kappa}_m)}{I_0(\tilde{\kappa}_m)} = \frac{I_1(\kappa_m)}{I_0(\kappa_m)} \tag{31}$$

where $I_p(\cdot)$ is the modified Bessel function of the first kind and order p. We show in Appendix B that an approximate solution to the transcendental equation (31) can be easily found. The means $\tilde{\mu}_{m,r}$, $r = 0,\ldots,m-1$, are given by

$$\tilde{\mu}_{m,r} = \frac{\mu_m + 2\pi r}{m}, \tag{32}$$

i.e., they are evenly distributed around the circle, $2\pi/m$ apart. The higher the concentration parameter of the wrapped VM pdf, the better its approximation (30). As illustrated in Fig. 3, the approximation is very tight even for moderate values of the concentration and still good for small concentrations.

The proposed approximation enables us to exploit the fact that the family of VM pdfs is closed under multiplication. To that end, we conveniently choose the prior pdf of Θ to be $p_\Theta(\theta) = f_{\text{VM}}(\theta;\eta_{\text{a}})$, with $\eta_{\text{a}} = \kappa_{\text{a}} e^{j\mu_{\text{a}}}$.[7,8] Replacing (30) in (29) we obtain that $q_{\Theta|\boldsymbol{Y}}(\theta \mid \mathbf{y})$ is an MVM. Specifically, let us write $\mathcal{M} = \{m_1, m_2, \ldots, m_M\} \subseteq \{1,\ldots,N-1\}$ and define $\mathcal{R} = \{1,\ldots,m_1\}\times\ldots\times\{1,\ldots,m_M\}$. Using the multi-index $\mathbf{r} = (r_1,\ldots,r_M) \in \mathcal{R}$, we have

$$q_{\Theta|\boldsymbol{Y}}(\theta \mid \mathbf{y}) = \frac{1}{Z_\theta} \sum_{\mathbf{r}\in\mathcal{R}} \exp\left\{\Re\left(\xi_{\mathbf{r}}^* e^{j\theta}\right)\right\} \tag{33}$$

with

$$\xi_{\mathbf{r}} = \eta_{\text{a}} + \tilde{\eta}_{m_1,r_1} + \ldots + \tilde{\eta}_{m_M,r_M} \tag{34}$$

and the normalizing constant $Z_\theta = 2\pi \sum_{\mathbf{r}\in\mathcal{R}} I_0(|\xi_{\mathbf{r}}|)$. We explicitly express (33) as an MVM where the amplitude, mean and concentration of each of the mixture's components are given by the corresponding parameter $\xi_{\mathbf{r}}$:

$$q_{\Theta|\boldsymbol{Y}}(\theta \mid \mathbf{y}) = \sum_{\mathbf{r}\in\mathcal{R}} \frac{2\pi I_0(|\xi_{\mathbf{r}}|)}{Z_\theta} f_{\text{VM}}(\theta;\xi_{\mathbf{r}}). \tag{35}$$

The number $m_1 \times \ldots \times m_M$ of components in (35) can be intractable. For the component with index $\mathbf{r}$ to have an important contribution to (35), its amplitude and concentration must be high, i.e., $|\xi_{\mathbf{r}}|$ be large. Based on the observation that only a small fraction of them contribute significantly to the mass of $q_{\Theta|\boldsymbol{Y}}$, in the following we propose two heuristic methods for representing (35) by a limited number of components.

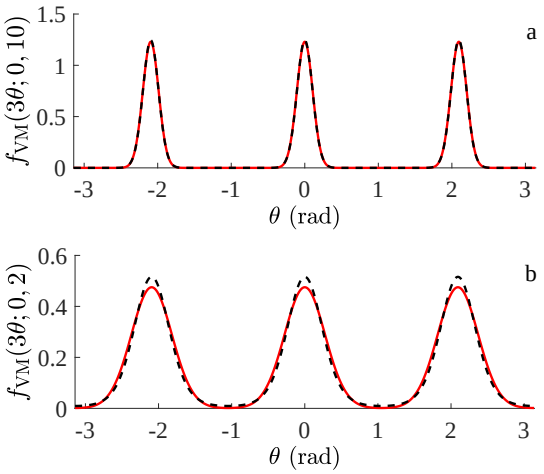

Fig. 3. Illustration of the approximation (30). The pdf $f_{\text{VM}}(3\theta;\kappa\, e^{j\cdot 0})$ of a 3-fold wrapped VM distribution (dashed curve) is approximated by a mixture of 3 von Mises pdfs (solid curve); in (a) $\kappa = 10$ for which (31) gives $\tilde{\kappa} \approx 85.78$; in (b) $\kappa = 2$ for which (31) gives $\tilde{\kappa} \approx 13.02$.

C. Heuristic 1

The first heuristic is a greedy procedure aiming to find and represent $q_{\Theta|\boldsymbol{Y}}$ by only the D most dominant components in (35). The idea is to progressively construct an approximation of (28) by sweeping through the index set $\mathcal{M}$ and including in the approximation one additional index in each step. In step p, $1 \le p \le M$, we have a "partial" posterior pdf given by the factors in (28) with indices $\{m_1,\ldots,m_p\}$, i.e., only p measurements are taken into account. The partial pdf is an MVM with $m_1\times\ldots\times m_p$ components parameterized by $\eta_{\text{a}}+\tilde{\eta}_{m_1,r_1}+\ldots+\tilde{\eta}_{m_p,r_p}$. As outlined in Algorithm 1, in each step the heuristic procedure retains from the "partial" posterior

[7] When we do not have any prior information about Θ, we can set the concentration $\kappa_{\text{a}} = 0$, in which case the prior pdf becomes the uniform circular pdf $p_\Theta(\theta) = 1/(2\pi)$.

[8] Alternatively, we can select an MVM prior, if we wish to use a multimodal distribution.

Algorithm 1 Heuristic 1

Input: $\mathcal{M}$, $\boldsymbol{\eta}$, η_0 and D
Output: $\boldsymbol{\xi}$

1: Compute all $\tilde{\eta}_{m,r} = \tilde{\kappa}_m e^{j\tilde{\mu}_{m,r}}$ in (31) and (32)
2: $\boldsymbol{\xi}^{(1)} \leftarrow (\eta_{\mathrm{a}} + \tilde{\eta}_{m_1,r} \mid 0 \le r \le m_1 - 1)$
3: **for** $p = 2$ to M **do**
4: $\quad \boldsymbol{\xi}^{(p)} \leftarrow D$ elements of $\left\{\xi_d^{(p-1)} + \tilde{\eta}_{m_p,r}\right\}_{d,r}$ with largest magnitudes[9]
5: **end for**
6: **return** $\boldsymbol{\xi} = \boldsymbol{\xi}^{(M)}$

(at most) D components having the highest concentration parameters. The complexity of the greedy search is $\mathcal{O}(DMN)$. The algorithm outputs the D parameters in $\boldsymbol{\xi}$ which give

$$q_{\Theta|\boldsymbol{Y}}(\theta \mid \mathbf{y}) \approx \sum_{d=1}^{D} \frac{2\pi I_0(|\xi_d|)}{\tilde{Z}_\theta} f_{\mathrm{VM}}(\theta; \xi_d) \tag{36}$$

where $\tilde{Z}_\theta = 2\pi \sum_{d=1}^{D} I_0(|\xi_d|)$. Now we can compute expectations in closed-form. Using (36) and (26), we obtain

$$\hat{\mathbf{a}} = \frac{2\pi}{\tilde{Z}_\theta} \sum_{d=1}^{D} \mathrm{diag}\left(I_{m_1}(|\xi_d|), \ldots, I_{m_M}(|\xi_d|)\right) \mathbf{a}(\arg(\xi_d)).$$

Similarly, the frequency estimate $\hat{\theta}$ defined in (11) is given by

$$\hat{\theta} = \arg\left(\frac{2\pi}{\tilde{Z}_\theta} \sum_{d=1}^{D} I_1(|\xi_d|) e^{j \arg(\xi_d)}\right).$$

D. Heuristic 2

The second approach is to search for the most dominant component of the mixture (35), i.e., that with index $\arg\max_{\mathbf{r} \in \mathcal{R}} |\xi_{\mathbf{r}}|$. Then, we represent (35) by a single von Mises pdf based on a second-order Taylor approximation around the mean $\bar{\theta}$ of that component. The intuition is that, with sufficient SNR and number M of measurements, the pdf (35) would peak somewhere in the neighborhood of $\bar{\theta}$.

Given that for each m, $|\tilde{\eta}_{m,r}|$ does not depend on r, see (30), to maximize $|\xi_{\mathbf{r}}|$ we have to look for that $\mathbf{r}$ for which the phases of the terms of (34) are best aligned. Such an alignment is searched in a greedy way by Algorithm 2 whose complexity is $\mathcal{O}(MN)$. Without loss of generality we assume $m_1 > \ldots > m_M$. The algorithm maintains a number m_1 of candidates and proceeds in a progressive manner. In step p, the lth candidate $\xi_l^{(p)}$, $1 \le l \le m_1$, is obtained by adding the term whose phase is closest to that of $\xi_l^{(p-1)}$, i.e., having the index

$$r_l^{(p)} = \underset{0 \le r \le m_p - 1}{\arg\max} \left| \xi_l^{(p-1)} + \tilde{\kappa}_{m_p} \exp\left(j \tfrac{\mu_{m_p} + 2\pi r}{m_p}\right) \right|.$$

The closed-form update is given by lines 5 and 6 of Algorithm 2 where $[\cdot]$ is the nearest-integer function. We set $\bar{\theta} = \arg \xi_{l^\star}^{(M)}$ with $l^\star = \arg\max_{1 \le l \le m_1} |\xi_l^{(M)}|$. Denoting the exponent of (28) by $f(\theta) = \Re\left(\eta_{\mathrm{a}}^* e^{j\theta} + \sum_{m \in \mathcal{M}} \eta_m^* e^{jm\theta}\right)$, we make a second-order Taylor approximation of $f(\theta)$ around

[9] $\boldsymbol{\xi}^{(p)}$ has less than D components when $m_1 \times \ldots \times m_p < D$.

Algorithm 2 Heuristic 2

Input: $\mathcal{M}$, $\boldsymbol{\eta}$ and η_0
Output: $\hat{\theta}$ and $\hat{\kappa}$

1: Compute all $\tilde{\eta}_{m,r} = \tilde{\kappa}_m e^{j\tilde{\mu}_{m,r}}$ in (31) and (32)
2: $\boldsymbol{\xi}^{(1)} \leftarrow (\eta_{\mathrm{a}} + \tilde{\eta}_{m_1,r} \mid 0 \le r \le m_1 - 1)$
3: **for** $p = 2$ to M **do**
4: $\quad$ **for** $l = 1$ to m_1 **do**
5: $\quad\quad r_l^{(p)} = \left[\frac{m_p \arg\left(\xi_l^{(p-1)}\right) - \mu_{m_p}}{2\pi}\right]$
6: $\quad\quad \xi_l^{(p)} \leftarrow \xi_l^{(p-1)} + \tilde{\kappa}_{m_p} \exp\left(j \frac{\mu_{m_p} + 2\pi r_l^{(p)}}{m_p}\right)$
7: $\quad$ **end for**
8: **end for**
9: Determine $l^\star = \arg\max_l |\xi_l^{(M)}|$ and set $\bar{\theta} = \arg \xi_{l^\star}^{(M)}$
10: **return** $\hat{\theta} = \bar{\theta} - \frac{f'(\bar{\theta})}{f''(\bar{\theta})}$ and $\hat{\kappa} = A^{-1}\left(\exp\left(0.5/f''(\bar{\theta})\right)\right)$

$\bar{\theta}$. Then we use the properties of the wrapped normal distribution [35, p. 50] and its similarity to the von Mises distribution [35, p. 38] to arrive at

$$q_{\Theta|\boldsymbol{Y}}(\theta \mid \mathbf{y}) \approx f_{\mathrm{VM}}(\theta; \hat{\eta})$$

with $\hat{\eta} = \hat{\kappa} e^{j\hat{\theta}}$, $\hat{\theta} = \bar{\theta} - \frac{f'(\bar{\theta})}{f''(\bar{\theta})}$ and $\hat{\kappa} = A^{-1}\left(e^{0.5/f''(\bar{\theta})}\right)$. A useful approximation of the inverse of the function $A(\cdot) = I_1(\cdot)/I_0(\cdot)$ is given in [35, pp. 85–86]. Finally, we can easily obtain the expected value of $\mathbf{a}(\Theta)$,

$$\hat{\mathbf{a}} = \mathrm{diag}\left(\tfrac{I_{m_1}(\hat{\kappa})}{I_0(\hat{\kappa})}, \ldots, \tfrac{I_{m_M}(\hat{\kappa})}{I_0(\hat{\kappa})}\right) \mathbf{a}(\hat{\theta}).$$

E. Illustrative examples

To illustrate the effectiveness of the proposed approximation, we give a few simple examples where exact pdfs of Θ of the form $\exp\left\{\Re\left(\boldsymbol{\eta}^{\mathrm{H}} \mathbf{a}(\theta)\right)\right\}$, i.e., as in (16), occur.

1) Coherent estimation: Let us consider the estimation of the frequency of a single sinusoid when we know its weight w. In this case, the posterior pdf is

$$p_{\Theta|\boldsymbol{Y},W}(\theta \mid \mathbf{y}, w) \propto \exp\left\{\Re\left(\frac{2}{\nu} w \mathbf{y}^{\mathrm{H}} \mathbf{a}(\theta)\right)\right\}. \tag{37}$$

Fig. 4a and 4b display snapshots of the pdf (37) and its approximations for different settings of $\mathcal{M}$ and SNR. When the number of measurements and SNR are low (Fig. 4a top), the pdf is spread across its domain. The MVM approximation (35) has $1 \times 2 = 2$ components; Heuristic 1 follows closely the exact pdf by keeping both components of the mixture, while the single VM pdf given by Heuristic 2 captures the dominant mode. Increasing the SNR (Fig. 4a bottom) makes the pdf more concentrated and both approximations are tight (in Heuristic 1, one component of the MVM has amplitude almost one and therefore the other is irrelevant). The pdf becomes more concentrated also by increasing the number of measurements, even though the SNR is low (Fig. 4b top). Even though the approximation (35) has 10! components, among the $D = 1000$ components output by Heuristic 1 only one is relevant. In the case of incomplete data (Fig. 4b bottom), the pdf (37) can have several significant modes. Among the

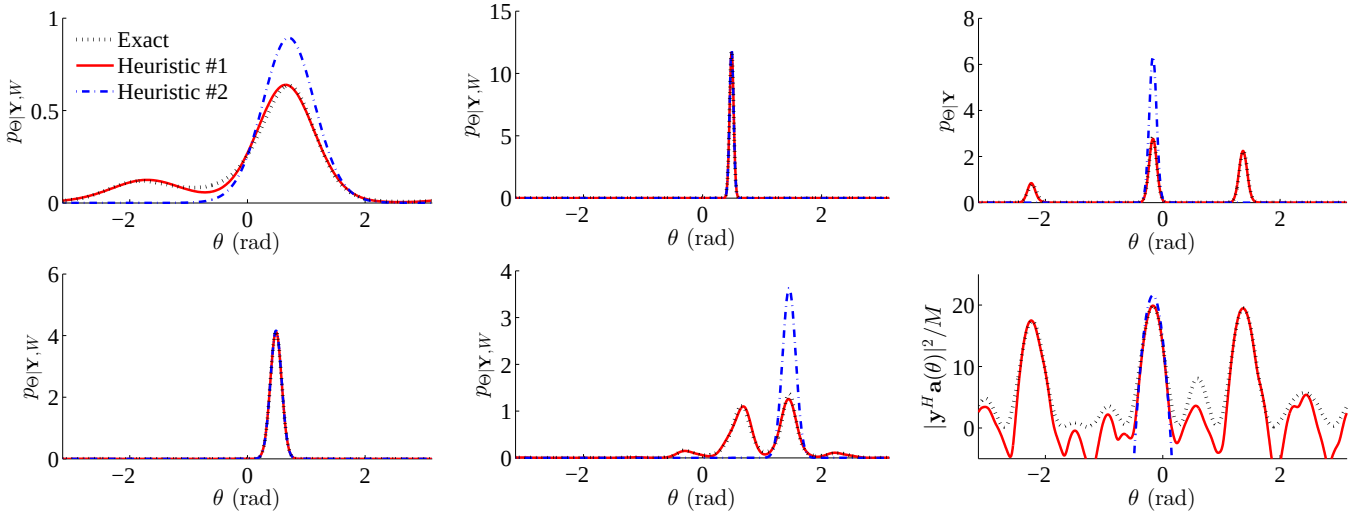

Fig. 4. (a) Snapshot of the pdf (37) and its approximations for $\theta = 0.5$, $\mathcal{M} = \{1, 2\}$ and SNR $= 0$ dB (top), SNR $= 10$ dB (bottom). (b) Snapshot of the pdf (37) and its approximations for $\theta = 0.5$, SNR $= 0$ dB and $\mathcal{M} = \{1, \ldots, 10\}$ (top), $\mathcal{M} = \{1, 7, 10\}$ (bottom). (c) Snapshot of $p_{\Theta|\boldsymbol{Y}}$ in (38) and its approximations for $K = 3$, $\theta_1 = -2.28$, $\theta_2 = -0.04$, $\theta_3 = 1.39$, $\mathcal{M} = \{0, \ldots, 9\}$, SNR $= 3$ dB (top); using a log scale (bottom).

$D = 1 \times 7 \times 10 = 70$ components provided by Heuristic 1 (with this setting of D, all components in (35) are kept), only 12 have amplitudes larger than 10^{-3}. Heuristic 2 captures the largest mode and misses the mass containing the true θ.

2) Noncoherent estimation: Without knowing the weight of the sinusoid, we can marginalize $p_{\Theta,W|\boldsymbol{Y}}$ and, assuming an improper "flat" prior of W, obtain

$$p_{\Theta|\boldsymbol{Y}}(\theta \mid \boldsymbol{y}) = \int f_{\mathrm{CN}}(\mathbf{y}; w\mathbf{a}(\theta), \nu)\, \mathrm{d}w \propto \exp\left(\frac{|\mathbf{y}^{\mathrm{H}}\mathbf{a}(\theta)|^2}{\nu M}\right). \tag{38}$$

The exponent of (38) is in fact the periodogram scaled by $1/\nu$. We write (38) in a form favorable for the MVM approximation. First, let us define $\mathcal{M}' = \{m - n \mid m, n \in \mathcal{M}, m > n\}$ with cardinality M' and the vector-valued function $\mathbf{a}' : [-\pi, \pi) \to \mathbb{C}^{M'}$, $\omega \to \mathbf{a}'(\omega) \triangleq (e^{j\omega m} \mid m \in \mathcal{M}')^{\mathrm{T}}$. By simply developing $|\mathbf{y}^{\mathrm{H}}\mathbf{a}(\theta)|^2$ we arrive at

$$p_{\Theta|\boldsymbol{Y}}(\theta \mid \boldsymbol{y}) \propto \exp\left(\Re\left(\tfrac{2}{\nu}\boldsymbol{\gamma}^{\mathrm{H}}\mathbf{a}'(\theta)\right)\right) \tag{39}$$

where, for each $t = 1, \ldots, M'$, $\gamma_t = \frac{1}{M}\sum_{(k,l)\in\mathcal{T}_t} y_k y_l^*$ with $\mathcal{T}_t = \{(k, l) \mid 1 \le k, l \le M, m_k - m_l = t\}$.[10] Given (39), we can approximate $p_{\Theta|\boldsymbol{Y}}$ as an MVM (35). In the log domain the approximation provides a representation of the periodogram.

As an illustration, we take $K = 3$ and plot a snapshot of $p_{\Theta|\boldsymbol{Y}}$ (Fig. 4c top) and the log of $p_{\Theta|\boldsymbol{Y}}$ scaled so as to give the periodogram (Fig. 4c bottom). We can see again the good agreement between the approximations and the exact curves. The three lobes corresponding to each of the sinusoids are very well represented by Heuristic 1 while Heuristic 2 picks up the highest lobe. Due to the exponentiation, $p_{\Theta|\boldsymbol{Y}}$ is significant only at the values of θ for which $\mathbf{a}(\theta)$ is well aligned with $\mathbf{y}$.

[10] Actually, $\boldsymbol{\gamma}$ is the sample autocovariance of $\mathbf{y}$.

V. The VALSE Algorithm

We define a schedule for iteratively updating the factors of $q_{\boldsymbol{\Theta},\boldsymbol{W},\boldsymbol{S}|\boldsymbol{Y}}$ and estimates $\hat{\nu}$, $\hat{\rho}$, $\hat{\tau}$ derived in Sec. III, and propose an initialization of the scheme.[11] The resulting algorithm, which we dub variational line spectral estimation (VALSE), is outlined in Algorithm 3. Since each step increases the lower bound (9), the algorithm converges to some local maximum of $\mathcal{L}$. The stopping criterion can be defined in terms of the relative change of some quantity (e.g., $\hat{\mathbf{x}}$) from one iteration to the next or a maximum number of iterations.

Algorithm 3 Outline of the VALSE algorithm

Input: Signal vector $\mathbf{y}$, set $\mathcal{M}$ of measurement indices
Output: Model order estimate $\hat{K}$, frequency and amplitude estimates $\{(\hat{\omega}_k, \hat{\alpha}_k)\}_{k=1}^{\hat{K}}$, reconstructed signal $\hat{\mathbf{x}}$

1: Initialize $\hat{\nu}$, $\hat{\rho}$, $\hat{\tau}$ and $q_{\Theta_i|\boldsymbol{Y}}$, $i \in \{1, \ldots, N\}$; compute $\hat{\mathbf{a}}_i$
2: **repeat**
3: Update $\hat{\mathbf{s}}$, $\hat{\mathbf{w}}_{\hat{\mathcal{S}}}$ and $\hat{\mathbf{C}}_{\hat{\mathcal{S}}}$ (Algorithm 4)
4: Update $\hat{\nu}$ (24), $\hat{\rho}$ and $\hat{\tau}$ (25)
5: For all $i \in \hat{\mathcal{S}}$, update $\boldsymbol{\eta}_i$ (17) and $\hat{\mathbf{a}}_i$ (Sec. IV)
6: **until** stopping criterion
7: **return** $\|\hat{\mathbf{s}}\|_0$, $\hat{\boldsymbol{\theta}}_{\hat{\mathcal{S}}}$, $\hat{\mathbf{w}}_{\hat{\mathcal{S}}}$ and $\hat{\mathbf{x}}$ (15)

While several initialization schemes can be imagined, we choose to initialize $\{q_{\Theta_i|\boldsymbol{Y}}\}_{i=1}^N$ in a sequential manner. In the first step, we assign $q_{\Theta_1|\boldsymbol{Y}}$ the noncoherent pdf form (38) and initialize the parameter estimates. For the latter, we use $\boldsymbol{\gamma}$ in (39) (whose entries are estimates of the autocovariane function) to build a Toeplitz estimate of $\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{H}}]$. Then, we initialize $\hat{\nu}$ with the average of the lower quarter of the eigenvalues of

[11] The alternating minimization scheme generates sequences of factors and estimates which, for notational convenience, we will not index with iteration numbers. It is to be understood that the update of one quantity depends on the most recent updates of the rest of the quantities.

that matrix. Given that $\mathrm{E}[\mathbf{y}^{\mathrm{H}}\mathbf{y}]/M = \rho N\tau+\nu$, we set $\hat{\rho} = 0.5$ and let $\hat{\tau} = (\mathbf{y}^{\mathrm{H}}\mathbf{y}/M - \hat{\nu})/(\hat{\rho}N)$. Then, in step i, when we have initialized the first $i-1$ pdfs, we compute the estimates $\{\hat{w}_k\}_{k=1}^{i-1}$ and the residual $\mathbf{z}_{i-1} = \mathbf{y} - \sum_{k=1}^{i-1} \hat{w}_k \hat{\mathbf{a}}_k$. Initializing $q_{\Theta_i|\boldsymbol{Y}} \propto \exp\left\{|\mathbf{z}_{i-1}^{\mathrm{H}}\mathbf{a}(\theta)|^2/(\nu M)\right\}$, we can represent $q_{\Theta_i|\boldsymbol{Y}}$ as an MVM (see (38) and (39)) and compute $\hat{\mathbf{a}}_i$.[12]

The complexity per iteration is dominated by the maximization of $\ln Z(\mathbf{s})$ needed in line 3 (realized by Algorithm 4) and the approximation of $\{q_{\Theta_i|\boldsymbol{Y}}\}_{i\in\hat{\mathcal{S}}}$ by mixtures of von Mises pdfs required in line 5 (using Algorithm 1 or Algorithm 2). According to the analysis in Appendix A, the maximization has complexity $\mathcal{O}(N\hat{K}^3)$ (actually, $\mathcal{O}(N\hat{K}^2)$ during most of the iterations of VALSE). As indicated in Sec. IV, the complexity of the MVM approximation is $\mathcal{O}(DMN)$ with Heuristic 1 and $\mathcal{O}(MN)$ with Heuristic 2; thus, the update of the pdfs of all frequencies with indices in $\hat{\mathcal{S}}$ has complexity $\mathcal{O}(\hat{K}DMN)$, respectively $\mathcal{O}(\hat{K}MN)$.

VI. Simulation Experiments

In this section, we use computer simulations to assess the performance of the VALSE algorithm and state-of-the-art methods under different scenarios.

A. Setup, metrics and algorithms

Referring to (1), the K values $\{\omega_k\}_{k=1}^K$ of the angular frequencies are generated one-by-one: ω_k is drawn from $\mathcal{U}(-\pi,\pi)$ until a minimum (wrap-around) distance $\Delta\omega$ is ensured between ω_k and each of the $k-1$ previously generated values. The complex amplitudes $\{\alpha_k\}_{k=1}^K$ are generated randomly by drawing their magnitudes from $\mathcal{N}(1, 0.1)$ and phases from $\mathcal{U}(-\pi,\pi)$. The noise samples contaminating the observations (2) are independent and zero-mean complex Gaussian distributed.

The following metrics are evaluated by averaging from 500 independent trials: the normalized mean square error of signal reconstruction, $\mathrm{E}[\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2/\|\mathbf{x}\|_2^2]$; the success rate, which we compute as the empirical probability of $\hat{K} = K$; and frequency estimation error. For a given simulation point, the frequency estimation error is evaluated only for the algorithms that provide a success rate ≥ 0.1 by averaging only the trials in which all those algorithms output $\hat{K} = K$. The assignment of estimated components to the true ones is performed according to the Munkres' (or Hungarian) algorithm [37] with the cost being the squared error of frequency estimates. We also report the runtime per trial for different problem sizes as an indicator of the complexity of the methods. The Cramér-Rao lower bounds (CRLB) on the reconstruction and frequency estimation errors are computed by assuming K is known.

We present the results for VALSE using Heuristic 2 to compute the estimates $\{\hat{\mathbf{a}}_i\}$ in line 5 of Algorithm 3. In general, we obtain very similar performances with Heuristic 1 and Heuristic 2, the latter being significantly faster. Even though in the tough conditions of low SNR and/or few measurements Heuristic 1 provides better representation of the pdfs (see Fig. 4), we observed that in those conditions Heuristic 1 has the tendency to underestimate K and provide slightly lower success rate than Heuristic 2. We assume no prior information about the frequencies is available, so we set $p_{\Theta_i}(\theta_i) = 1/(2\pi)$, $i = 1,\ldots,N$. Algorithm 3 stops at iteration t if $\|\hat{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}^{(t-1)}\|/\|\hat{\mathbf{x}}^{(t-1)}\| < 10^{-6}$ or the number of iterations reaches 5000.

We also introduce a variant of our algorithm, called "VALSE-pt", which operates with point estimates of the frequencies (as in the traditional approach). VALSE-pt additionally assumes that $q_{\Theta_i|\boldsymbol{Y}}(\theta_i \mid \mathbf{y}) = \delta(\theta_i - \hat{\theta}_i)$ for all i, which gives that $\hat{\theta}_i$ is the maximizer of (16) and $\hat{\mathbf{a}}_i = \mathbf{a}(\hat{\theta}_i)$. We obtain $\hat{\theta}_i = \arg\max \Re\left(\boldsymbol{\eta}_i^{\mathrm{H}}\mathbf{a}(\theta_i)\right)$ numerically. Except for the computation of $\hat{\mathbf{a}}_i$ in line 5 of Algorithm 3, all the other steps and settings of VALSE-pt and VALSE are identical.

For comparison, we evaluate the following state-of-the-art methods described in the Introduction: atomic-norm soft-thresholding[13] (AST) [25]—only applicable in the complete data case, the gridless-SPICE-based framework[14] (GLS) [26], enhanced matrix completion[15] (EMaC) [29] and reweighted atomic-norm minimization[16] (RAM) [30]. To configure algorithm-specific parameters, AST, EMaC and RAM require knowledge of the noise power. For each of these three methods we use the noise-variance estimation in [25], which computes $\hat{\nu}$ by averaging a lower part of the eigenvalues of an estimate of $\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{H}}]$. EMaC and RAM require an upper bound on the ℓ_2 norm of the noise vector in order to search only among candidate solutions whose distances to the measurement $\mathbf{y}$ are less than the bound; we set this bound to $\sqrt{(M + 2\sqrt{M})\hat{\nu}}$, as suggested in [30].

B. Estimation from complete data

Fig. 5 displays the results of estimating $K = 5$ sinusoidal components from $M = N = 21$ measurements at different SNR values. The distance between any two frequencies is at least $\Delta\omega = \frac{2\pi}{N}$ radians. VALSE outperforms the reference methods at all SNR values and shows excellent performance at SNR $\geq$ 10 dB, where the reconstruction and frequency estimation errors are very close to the CRLB and the success rate is almost one. AST and GLS estimate the model order accurately as well in high SNR, but their success rates decrease earlier. The success rate of AST seems to saturate at a value slightly below one and degrades faster than that of GLS when the SNR decreases. In Fig. 5a, the gap between AST and VALSE increases for SNR $\geq$ 10 dB, while GLS maintains a constant gap of about 0.5 dB. A similar behavior can be also observed for the frequency estimation error in Fig. 5c.

[12] In the initialization we use Heuristic 2 to compute the $\hat{\mathbf{a}}_i$'s because, when the sinusoidal components have similar powers, Heuristic 1 will capture contributions from the signal components that are not yet initialized, while Heuristic 2 picks up the strongest one (see Fig. 4c).

[13] The software is available at https://github.com/badrinarayan/astlinespec. We used the implementation via ADMM.

[14] The software was provided by the authors of [26]. We used the implementation via ADMM [26].

[15] We used the software available at http://www2.ece.ohio-state.edu/~chi/research.html. The implementation uses the SDPT3 solver. Based on the "cleaned" signal output by EMaC, we perform model order and parameter estimation using Root-MUSIC and Akaike information criterion.

[16] The software was provided by the authors of [30]. The implementation uses the SDPT3 solver.

We have also evaluated the EMaC and RAM algorithms. Since they did not provide significant improvements over AST and for the clarity of the figures, we do not show those results. In fact, we observed that EMaC and RAM perform well only at high SNR (above 20-25 dB) where their success rates approach one; still, in this SNR regime, EMaC shows worse signal reconstruction and frequency estimation than AST, while RAM provides slight improvement over AST. For SNR $<$ 20 dB, both EMaC and RAM are outperformed by AST in all metrics. Our explanation for their not so good performance in the low-to-moderate SNR region is that, according to our observations, their performance is quite sensitive to the setting of the upper bound on the ℓ_2 norm of the noise vector and therefore to the accuracy of the noise variance estimate.

The gap between the success rates of VALSE-pt and VALSE is due to the former's tendency to overestimate K more heavily. For example, at SNR $=$ 15 dB, VALSE outputs $\hat{K} = 6$ in 5 out of the 500 simulation trials, while VALSE-pt outputs $\hat{K} = 6$, 7 and 10 components in 53, 7 and respectively 1 trials. The discrepancy between their performance comes from the way in which $\hat{\mathbf{a}}_i$ is computed in line 5 of Algorithm 3, since this is the only difference between the two algorithms. VALSE computes $\hat{\mathbf{a}}_i = \mathrm{E}_{q_{\Theta_i|\boldsymbol{Y}}}[\mathbf{a}(\Theta_i)]$, which involves the expectations of the phasors $e^{jn\Theta_i}$. The more concentrated $q_{\Theta_i|\boldsymbol{Y}}$, the closer $|\,\mathrm{E}_{q_{\Theta_i|\boldsymbol{Y}}}[e^{jn\Theta_i}]|$ is to one and $\|\hat{\mathbf{a}}_i\|_2$ to $\sqrt{M}$. Therefore, the uncertainty in frequency estimation captured by $q_{\Theta_i|\boldsymbol{Y}}$ is reflected in $\hat{\mathbf{a}}_i$. Consequently, the uncertainty impacts all the other estimates, which in turn determine the component-acceptance criterion, and therefore influences the model order estimate. On contrary, VALSE-pt assigns $\hat{\mathbf{a}}_i = \mathbf{a}(\hat{\theta}_i)$ and thus puts full certainty on the phasors' estimates. Loosely speaking, VALSE-pt might include excessive components because it "overtrusts" them—this is what we also observe experimentally.

C. Estimation from incomplete data

We now study the performance when the measurement data is incomplete, i.e., $M < N$. We consider the estimation of $K = 3$ sinusoids when $N = 20$ and SNR $=$ 10 dB. The frequencies are separated by at least $\Delta\omega = \frac{2\pi}{N}$. Based on the previous analysis, in the comparison we include only the GLS method. The results in Fig. 6a and 6b show that, for $M \geq 14$, VALSE estimates $\mathbf{x}$ very accurately (close to the CRLB) and selects the correct model order with a rate close to one. On contrary, the reconstruction errors of GLS and VALSE-pt are 1–2 dB larger in that range of M. GLS provides a good estimation of K, although the success rate is always lower than that of VALSE and decreases earlier when reducing M. VALSE-pt shows a significantly lower success rate (again due to overestimation). When the algorithms estimate K correctly, both VALSE and VALSE-pt provide very accurate frequency estimation, while GLS has larger errors.

D. Resolution capability

Next, we evaluate the performance of resolving $K = 2$ sinusoids that are closely-spaced in frequency. We draw ω_1 from $\mathcal{U}(-\pi, \pi)$ and set $\omega_2 = \omega_1 + \Delta\omega$ (i.e., we impose an exact separation of $\Delta\omega$, and not a minimum one as in the previous experiments). Fig. 7 shows results for $M = N = 51$, SNR $=$ 10 dB and $0.1 \times \frac{2\pi}{N} \leq \Delta\omega \leq 2 \times \frac{2\pi}{N}$. We observe that, for $\Delta\omega > 0.5 \times \frac{2\pi}{N}$, VALSE and GLS reconstruct the signal similarly well and estimate the correct model order with high probability (the success rate of VALSE seems to cap at about 0.95 while that of GLS comes very close to 1). When the two frequencies are separated by less than $0.5 \times \frac{2\pi}{N}$, VALSE shows a significantly higher success rate compared to GLS; the reconstruction performance of the latter also degrades considerably. Fig. 7c shows that VALSE estimates the frequencies accurately in the whole range of $\Delta\omega$. AST, EMaC and RAM provide significantly lower performance in reconstructing the signal and selecting the model order, which is inline with our observations in the first experiment.

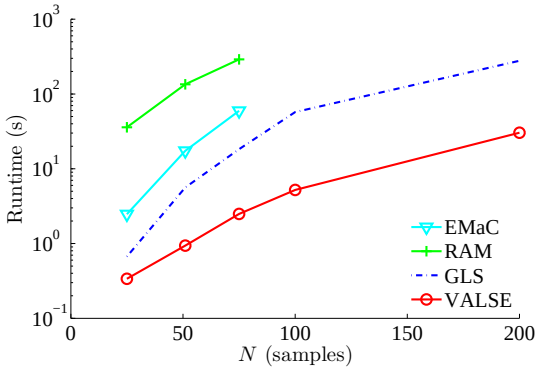

Fig. 8. Scaling of the runtime with the problem size. The simulation points correspond to the following (N, M, K) triples: $(25, 15, 2)$, $(51, 30, 4)$, $(75, 45, 6)$, $(100, 60, 8)$ and $(200, 120, 16)$.

E. Scaling with the problem size

To obtain an indication of how the complexity of VALSE scales with the dimension of the problem, we evaluate the runtime for different sizes N. We consider an incomplete-data scenario in which the number M of measurements and model order K scale with N and SNR $=$ 20 dB. The following (N, M, K) triples are investigated: $(25, 15, 2)$, $(51, 30, 4)$, $(75, 45, 6)$, $(100, 60, 8)$ and $(200, 120, 16)$. The results in Fig. 8 clearly show that VALSE is computationally advantageous compared to the benchmark methods. While RAM's runtime becomes quickly prohibitive, followed by EMaC and GLS, VALSE is about 10 times faster than GLS when N increases.

VII. Conclusions

In this paper, we treated line spectral estimation (LSE) as Bayesian inference in a probabilistic model of the frequencies and coefficients. The latter were modeled by a Bernoulli-Gaussian distribution, which turned model order selection into detection of a binary sequence. To circumvent the deadlock of exact inference we resorted to the variational approach in which an approximate (surrogate) posterior pdf was computed analytically by maximizing a lower bound on the model

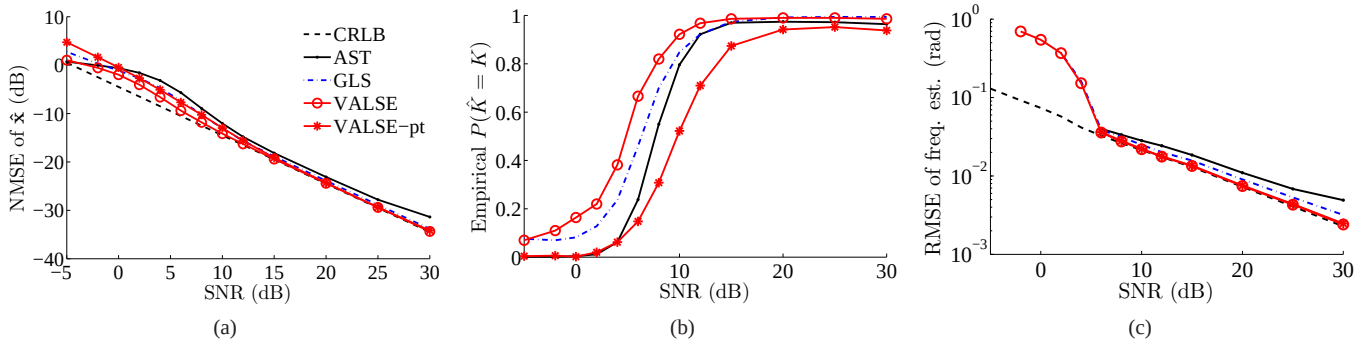

Fig. 5. Performance vs. SNR for $M = N = 21$ samples, $K = 5$ and minimum separation $\Delta\omega = \frac{2\pi}{N}$: (a) normalized MSE of the reconstructed signal; (b) success rate of model order estimation; (c) root MSE for frequency estimation.

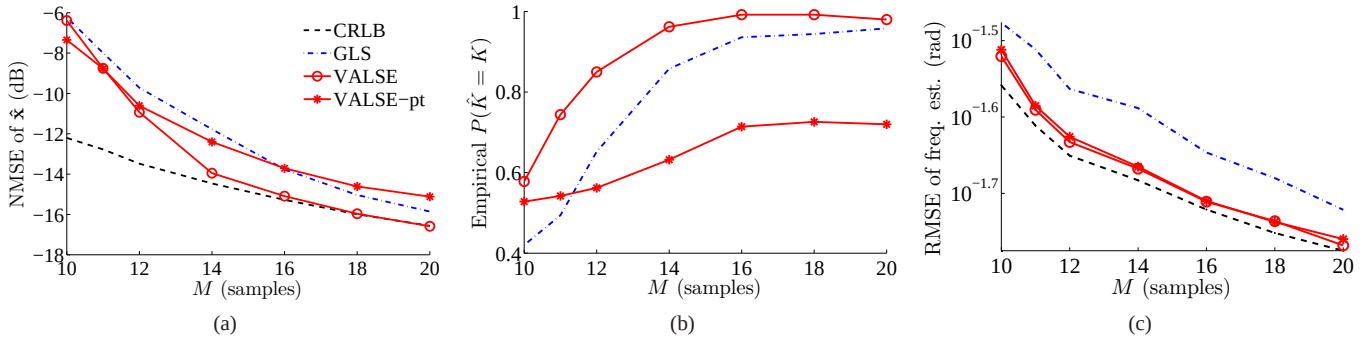

Fig. 6. Performance vs. M for $K = 3$, $N = 20$ samples, SNR $= 10$ dB and minimum separation $\Delta\omega = \frac{2\pi}{N}$: (a) normalized MSE of the reconstructed signal; (b) success rate of model order estimation; (c) root MSE for frequency estimation.

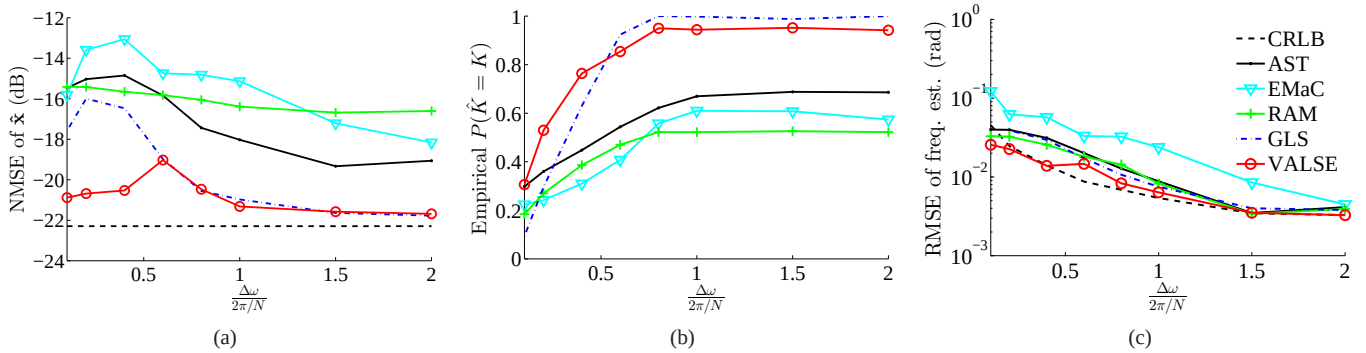

Fig. 7. Performance of resolving $K = 2$ sinusoids separated by small $\Delta\omega$; $M = N = 51$ samples, SNR $= 10$ dB and $0.1 \times \frac{2\pi}{N} \leq \Delta\omega \leq 2 \times \frac{2\pi}{N}$: (a) normalized MSE of the reconstructed signal; (b) success rate of model order estimation; (c) root MSE for frequency estimation.

evidence. Contrary to related works which compute point estimates of the frequencies, we considered estimating and working with their posterior probability density functions (pdfs). We showed that these pdfs can be very well approximated by mixtures of von Mises pdfs, which enables computation of closed-form expectations. In fact, our simulations show that the representation by one von Mises pdf seems appropriate. The resulting VALSE algorithm increases the lower bound on the model evidence in each step and hence is convergent. Since all the parameters are estimated, VALSE does not require any fine tuning by the user. Simulation results advocate our fully Bayesian approach of representing and operating with the uncertainty in frequency estimation, as we obtain significantly improved performance compared to just using point estimates. VALSE shows an excellent performance (often close to the Cramér-Rao bound), consistently better than the benchmark.

Our method can straightforwardly include prior knowledge about the frequencies in the form of von Mises pdfs or mixtures of such pdfs if multimodal distribution are desired. The fact that VALSE conveniently represents posterior distributions allows for estimating the uncertainty in the estimation. Also, the pdfs can be subsequently used as prior pdfs in applications that rely on line spectral estimation. As an outlook, we expect that finding a better variational approximation, in which the

surrogate pdf does not fully factorize over the frequencies yet still facilitates tractable inference, would further improve performance, especially in situations where the frequencies are very closely spaced.

APPENDIX A
FINDING A LOCAL MAXIMUM OF $\ln Z(\mathbf{s})$

To find the globally optimal binary sequence $\mathbf{s}$ would require 2^N evaluations of $\ln Z(\mathbf{s})$ given by (21). Inspired by the iterative search strategy proposed in [31], we seek a locally optimal solution in a progressive manner. In step p, the utility of the reference sequence $\mathbf{s}^{(p)}$ is compared to the N utilities corresponding to N test sequences. Specifically, the kth test sequence $\mathbf{t}_k$ is obtained by flipping the kth location of $\mathbf{s}^{(p)}$. The change $\Delta_k^{(p)} = \ln Z(\mathbf{t}_k) - \ln Z(\mathbf{s}^{(p)})$ is evaluated for each $k = 1, \ldots, N$, and the test sequence giving the highest positive change is used as the reference sequence $\mathbf{s}^{(p+1)}$ in the next step. If $\Delta_k^{(p)} < 0$ for all $k = 1, \ldots, N$, then the search stops and we set $\hat{\mathbf{s}} = \mathbf{s}^{(p)}$. The search starts with a certain initial reference sequence $\mathbf{s}^{(0)}$ and converges in a finite number of steps to some locally optimal sequence. Although (21) involves a matrix inversion, the changes $\Delta_k^{(p)}$, $k = 1, \ldots, N$, can be efficiently computed in each step p as follows.

Assume we change a sequence $\mathbf{s}$ into a sequence $\mathbf{s}'$ by flipping the bit at the kth location. When $k \notin \mathcal{S}$, i.e. $s_k = 0$, $s'_k = 1$ and $\mathcal{S}' = \mathcal{S} \cup \{k\}$, we say the kth location is activated. Using the formulas for block-matrix determinant and block-wise matrix inversion, we write $\ln Z(\mathbf{s}') - \ln Z(\mathbf{s})$ as

$$\Delta_k = \ln \frac{v_k}{\tau} + \frac{|u_k|^2}{v_k} + \ln \frac{\rho}{1-\rho}. \tag{40}$$

where

$$\begin{aligned} v_k &= \nu \left(M + \frac{\nu}{\tau} - \nu^{-1} \mathbf{j}_k^{\mathrm{H}} \hat{\mathbf{C}}_{\mathcal{S}} \mathbf{j}_k \right)^{-1} \\ u_k &= \nu^{-1} v_k (h_k - \mathbf{j}_k^{\mathrm{H}} \hat{\mathbf{w}}_{\mathcal{S}}) \end{aligned} \tag{41}$$

with $\mathbf{j}_k = (J_{ik} \mid i \in \mathcal{S})^{\mathrm{T}}$. Upon changing $\mathbf{s}$ into $\mathbf{s}'$, we use rank-one updates for the mean and covariance of the weights

$$\hat{w}'_i = \begin{cases} u_k, & i = k, \\ \hat{w}_i - \hat{\mathbf{c}}_i^{\mathrm{H}} \mathbf{j}_k u_k, & i \in \mathcal{S}. \end{cases} \tag{42}$$

and

$$\begin{pmatrix} \hat{\mathbf{C}}'_{\mathcal{S}} & \hat{\mathbf{c}}'_k \\ \hat{\mathbf{c}}'^{\mathrm{H}}_k & \hat{C}'_{kk} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{C}}_{\mathcal{S}} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} + v_k \begin{pmatrix} \hat{\mathbf{C}}_{\mathcal{S}} \mathbf{j}_k \\ -1 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{C}}_{\mathcal{S}} \mathbf{j}_k \\ -1 \end{pmatrix}^{\mathrm{H}} \tag{43}$$

Thus, by activating the kth component, the posterior mean and variance of W_k are u_k and v_k, respectively.

In the case of deactivation, i.e., $s_k = 1$, $s'_k = 0$ and $\mathcal{S}' = \mathcal{S} \setminus \{k\}$, the change $\ln Z(\mathbf{s}') - \ln Z(\mathbf{s})$ is given by

$$\Delta_k = -\ln \frac{\hat{C}_{kk}}{\tau} - \frac{|\hat{w}_k|^2}{\hat{C}_{kk}} - \ln \frac{\rho}{1-\rho}. \tag{44}$$

We can again develop efficient updates: for all $i, j \in \mathcal{S}'$,

$$\hat{w}'_i = \hat{w}_i - \frac{\hat{C}_{ik}}{\hat{C}_{kk}} \hat{w}_k \quad \text{and} \quad \hat{C}'_{ij} = \hat{C}_{ij} - \frac{\hat{C}_{ik} \hat{C}_{kj}}{\hat{C}_{kk}}. \tag{45}$$

The iterative maximization is given by Algorithm 4. The most expensive computation is to obtain u_k for all $k \notin \mathcal{S}$ (in line 3). It requires $\mathcal{O}((N-l)l^2)$ operations, where $l = \|\mathbf{s}\|_0$ is the current number of active locations. If in line 1 we initialize $\mathbf{s} = \mathbf{0}$ (i.e., $l = 0$), the algorithm will execute the while loop $\mathcal{O}(\hat{K})$ times to output $\hat{\mathbf{s}}$, where $\hat{K} = \|\hat{\mathbf{s}}\|_0$. This gives the overall complexity $\mathcal{O}(N\hat{K}^3)$. However, in line 1 we can initialize $\mathbf{s}$ with $\hat{\mathbf{s}}$ from the previous iteration of VALSE (Algorithm 3). In this case, we observed that in each iteration of VALSE (except for the first one), the number of locations of $\hat{\mathbf{s}}$ that are changed by Algorithm 4 is very small (in fact, often zero!). Thus, empirically, the complexity of Algorithm 4 during most of the iterations of VALSE is $\mathcal{O}(N\hat{K}^2)$.

Algorithm 4 Algorithm for maximizing $\ln Z(\mathbf{s})$

Input: $\mathbf{J}$, $\mathbf{h}$, ν and ρ
Output: $\hat{\mathbf{s}}$, $\hat{\mathbf{w}}_{\hat{\mathcal{S}}}$ and $\hat{\mathbf{C}}_{\hat{\mathcal{S}}}$

1: Initialize $\mathbf{s}$ and compute $\hat{\mathbf{w}}_{\mathcal{S}}$ and $\hat{\mathbf{C}}_{\mathcal{S}}$ (20)
2: **while** true **do**
3: For each $k \notin \mathcal{S}$, compute u_k and v_k (41), and Δ_k (40)
4: For each $k \in \mathcal{S}$, compute Δ_k (44)
5: **if** $\{k \mid \Delta_k > 0\} \neq \emptyset$ **then**
6: $k_* = \arg\max_k \Delta_k$
7: If $s_{k_*} = 0$ compute (42) (43), else compute (45)
8: $s_{k_*} \leftarrow s_{k_*} \oplus 1$
9: **else**
10: **break**
11: **end if**
12: **end while**
13: **return** $\hat{\mathbf{s}} = \mathbf{s}$, $\hat{\mathbf{w}}_{\hat{\mathcal{S}}} = \hat{\mathbf{w}}_{\mathcal{S}}$ and $\hat{\mathbf{C}}_{\hat{\mathcal{S}}} = \hat{\mathbf{C}}_{\mathcal{S}}$

APPENDIX B
APPROXIMATION OF WRAPPED VON MISES DISTRIBUTIONS

The N-fold wrapped VM distribution is invariant under the transformation $\Theta \mapsto \Theta + \frac{2\pi}{N}$ [35, p. 52]. Its pdf is $f_{\mathrm{VM}}(N\theta; \eta)$, for some $\eta = \kappa e^{jN\mu}$. The N modes of the pdf have equal amplitudes and are evenly distributed around the circle, i.e., they are at $\mu + 2\pi n/N$, $n = 0, \ldots, N-1$. We show that such a distribution is well approximated by an appropriate mixture of von Mises distributions (MVM) obtained by matching their characteristic functions. Our result extends the one in [35, p. 54] which proposes the approximation for $N = 2$.

The characteristic function φ'_p, $p \in \mathbb{Z}$, of a random variable having an N-fold wrapped VM distribution is

$$\begin{aligned} \varphi'_p &= \int_0^{2\pi} e^{jp\theta} \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos N(\theta - \mu)} \, \mathrm{d}\theta \\ &= \frac{e^{jp\mu}}{2\pi I_0(\kappa)} \sum_{n=0}^{N-1} \int_{2\pi n/N}^{2\pi(n+1)/N} e^{jp\theta} \, e^{\kappa \cos N\theta} \, \mathrm{d}\theta \\ &= \frac{e^{jp\mu}}{2\pi I_0(\kappa)} \sum_{n=0}^{N-1} e^{j2\pi \frac{p}{N} n} \int_0^{2\pi/N} e^{jp\theta} \, e^{\kappa \cos N\theta} \, \mathrm{d}\theta \\ &= \frac{e^{jp\mu}}{2\pi I_0(\kappa)} \frac{1}{N} \sum_{n=0}^{N-1} e^{j2\pi \frac{p}{N} n} \int_0^{2\pi} e^{j\frac{p}{N}\theta} \, e^{\kappa \cos \theta} \, \mathrm{d}\theta \end{aligned} \tag{46}$$

The sum of a geometrical progression in (46) amounts to

$$\frac{1}{N} \sum_{n=0}^{N-1} e^{j2\pi \frac{p}{N} n} = \begin{cases} 1, & \text{if } p \bmod N = 0, \\ 0, & \text{else.} \end{cases} \tag{47}$$

Finally, we obtain the characteristic function

$$\varphi'_p = \begin{cases} e^{jp\mu}\frac{I_{p/N}(\kappa)}{I_0(\kappa)}, & \text{if } p \bmod N = 0, \\ 0, & \text{else.} \end{cases} \quad (48)$$

Given the properties of the N-fold wrapped VM pdf, we choose the approximating MVM pdf

$$\frac{1}{2\pi I_0(\tilde{\kappa})N}\sum_{n=0}^{N-1} e^{\tilde{\kappa}\cos(\theta-\tilde{\mu}-2\pi n/N)},$$

i.e., the pdf has N components with equal amplitudes, evenly spaced means $\tilde{\mu}+2\pi n/N$, $n=0,\ldots,N-1$, and concentration parameters equal to $\tilde{\kappa}$. The characteristic function φ''_p, $p\in\mathbb{Z}$, is obtained as

$$\begin{aligned}\varphi''_p &= \int_0^{2\pi} e^{jp\theta}\frac{1}{2\pi I_0(\tilde{\kappa})N}\sum_{n=0}^{N-1} e^{\tilde{\kappa}\cos(\theta-\tilde{\mu}-2\pi n/N)}\,\mathrm{d}\theta \\ &= \frac{e^{jp\tilde{\mu}}}{2\pi I_0(\tilde{\kappa})}\frac{1}{N}\sum_{n=0}^{N-1} e^{j2\pi\frac{p}{N}n}\int_0^{2\pi} e^{jp\theta}\,e^{\tilde{\kappa}\cos\theta}\,\mathrm{d}\theta\end{aligned}$$

Using (47) again, we obtain

$$\varphi''_p = \begin{cases} e^{jp\tilde{\mu}}\frac{I_p(\tilde{\kappa})}{I_0(\tilde{\kappa})}, & \text{if } p \bmod N = 0, \\ 0, & \text{else.} \end{cases} \quad (49)$$

We want to find $\tilde{\mu}$ and $\tilde{\kappa}$ that provide a good match between φ'_p and φ''_p, for all $p\in\mathbb{Z}$. Setting $\tilde{\mu}=\mu$, we obtain $\arg\{\varphi'_p\} = \arg\{\varphi''_p\}$, for all p. As to the magnitudes, we equate the first nonzero values of the characteristic functions, i.e., $\varphi'_N = \varphi''_N$, and obtain the transcendental equation in $\tilde{\kappa}$

$$\frac{I_N(\tilde{\kappa})}{I_0(\tilde{\kappa})} = \frac{I_1(\kappa)}{I_0(\kappa)} \quad (50)$$

To show that (50) yields a good approximation $|\varphi''_p| \simeq |\varphi'_p|$ for any p, we make use of the fact that the VM distribution characterized by μ and κ can be well approximated by a wrapped normal distribution with mean direction μ and mean resultant length $\rho = A(\kappa) \triangleq I_1(\kappa)/I_0(\kappa)$ [35]. While the approximation is tight for large κ, it is still satisfactory for intermediate values of κ. Therefore, the characteristic function $e^{jp\mu}\rho^{p^2}$ of the wrapped normal distribution approximates that one of a VM distribution. Based on (26), we can thus write:

$$\frac{I_p(\kappa)}{I_0(\kappa)} \simeq \rho^{p^2}, \quad (51)$$

for all $p\in\mathbb{Z}$ and $\rho = A(\kappa)$. Next, we define $\tilde{\rho}$ by $\tilde{\rho}^{N^2} = I_N(\tilde{\kappa})/I_0(\tilde{\kappa})$. According to (50), $\tilde{\rho}^{N^2}=\rho$. Thus, using (51), we obtain that, for all $p\in\mathbb{Z}$, $p \bmod N = 0$,

$$\frac{I_{p/N}(\kappa)}{I_0(\kappa)} \simeq \rho^{p^2/N^2} \simeq (\tilde{\rho})^{p^2} \simeq \frac{I_p(\tilde{\kappa})}{I_0(\tilde{\kappa})}.$$

In conclusion, setting $\tilde{\mu}=\mu$ and solving (50) for $\tilde{\kappa}$, we find a good approximation $\varphi''_p \simeq \varphi'_p$ for all p and thus can write

$$f_{\mathrm{VM}}(N\theta; N\mu, \kappa) \simeq \frac{1}{N}\sum_{n=0}^{N-1} f_{\mathrm{VM}}(\theta; \mu+2\pi n/N, \tilde{\kappa}). \quad (52)$$

An approximate solution to (50) can be found by using (51) to arrive at $[A(\tilde{\kappa})]^{N^2} = A(\kappa)$, where $A(\cdot) = I_1(\cdot)/I_0(\cdot)$. Approximations of the function $A(\cdot)$ and its inverse are well studied, see [35, p. 40] and [35, pp. 85–86].

REFERENCES

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

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, Mar 1986.

[3] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, Jul 1989.

[4] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 10, pp. 1553–1560, Oct 1988.

[5] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 4, pp. 477–489, Apr 1988.

[6] B. H. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, and K. I. Pedersen, "Channel parameter estimation in mobile radio environments using the SAGE algorithm," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 3, pp. 434–450, Mar 1999.

[7] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, July 2004.

[8] C. D. Austin, J. N. Ash, and R. L. Moses, "Dynamic dictionary algorithms for model order and parameter estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 5117–5130, Oct 2013.

[9] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 948–958, June 2010.

[10] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2182–2195, May 2011.

[11] M. F. Duarte and R. G. Baraniuk, "Spectral compressive sensing," *Applied and Computational Harmonic Analysis*, vol. 35, no. 1, pp. 111–129, Jul. 2013.

[12] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, Aug 2005.

[13] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, "Recovery of sparse translation-invariant signals with continuous basis pursuit," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4735–4744, Oct 2011.

[14] Z. Yang, L. Xie, and C. Zhang, "Off-grid direction of arrival estimation using sparse Bayesian inference," *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 38–43, Jan 2013.

[15] L. Hu, J. Zhou, Z. Shi, and Q. Fu, "A fast and accurate reconstruction algorithm for compressed sensing of complex sinusoids," *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5744–5754, Nov 2013.

[16] K. Fyhn, M. F. Duarte, and S. H. Jensen, "Compressive parameter estimation for sparse translation-invariant signals using polar interpolation," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 870–881, Feb 2015.

[17] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[18] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, Aug 2004.

[19] D. Shutin and B. H. Fleury, "Sparse variational Bayesian SAGE algorithm with application to the estimation of multipath wireless channels," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3609–3623, Aug 2011.

[20] L. Hu, Z. Shi, J. Zhou, and Q. Fu, "Compressed sensing of complex sinusoids: An approach based on dictionary refinement," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3809–3822, July 2012.

[21] D. Shutin, W. Wang, and T. Jost, "Incremental sparse Bayesian learning for parameter estimation of superimposed signals," in *Proc. of the 10th International Conference on Sampling Theory and Applications (SampTA)*, Bremen, Germany, July 2013.

[22] T. L. Hansen, M. A. Badiu, B. H. Fleury, and B. D. Rao, "A sparse Bayesian learning algorithm with dictionary parameter estimation," in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014 IEEE 8th*, June 2014, pp. 385–388.

[23] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.
[24] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7465–7490, Nov 2013.
[25] B. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *Signal Processing, IEEE Transactions on*, vol. 61, no. 23, pp. 5987–5999, Dec 2013.
[26] Z. Yang and L. Xie, "On gridless sparse methods for line spectral estimation from complete and incomplete data," *Signal Processing, IEEE Transactions on*, vol. 63, no. 12, pp. 3139–3153, June 2015.
[27] P. Stoica, P. Babu, and J. Li, "New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 35–47, Jan 2011.
[28] Z. He, A. Cichocki, S. Xie, and K. Choi, "Detecting the number of clusters in n-way probabilistic clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2006–2021, 2010.
[29] Y. Chen and Y. Chi, "Robust spectral compressed sensing via structured matrix completion," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6576–6601, Oct 2014.
[30] Z. Yang and L. Xie, "Enhancing sparsity and resolution via reweighted atomic norm minimization," *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 995–1006, Feb 2016.
[31] J. J. Kormylo and J. Mendel, "Maximum likelihood detection and estimation of Bernoulli-Gaussian processes," *IEEE Transactions on Information Theory*, vol. 28, no. 3, pp. 482–488, May 1982.
[32] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2667–2676, Oct 1999.
[33] C. Soussen, J. Idier, D. Brie, and J. Duan, "From bernoulli–gaussian deconvolution to sparse signal restoration," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4572–4584, Oct 2011.
[34] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
[35] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons, 2000.
[36] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, "Fast variational sparse Bayesian learning with automatic relevance determination for superimposed signals," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6257–6261, Dec 2011.
[37] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society of Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, March 1957.

Mihai-Alin Badiu received the Dipl.-Ing., M.S. and Ph.D. degrees in electrical engineering from the Technical University of Cluj-Napoca, Romania, in 2008, 2010 and 2012, respectively. Since 2012, he has been with the Department of Electronic Systems, Aalborg University, Denmark, where he is currently holding a post-doc fellowship from the Danish Council for Independent Research. From 2008 to 2010 he was a research assistant at the Technical University of Cluj-Napoca. In 2011 he was a visiting PhD researcher at Aalborg University, Denmark. In 2016–2017 he was a visiting postdoctoral researcher at Aston University, Birmingham, United Kingdom. His research interests are in the fields of machine learning, wireless networks and signal processing.

Thomas Lundgaard Hansen received the B.Sc. and M.Sc. (cum laude) in electrical engineering from Aalborg University, Denmark, in 2011 and 2014, respectively. Since 2014 he has been a Ph.D. fellow at Aalborg University. During 2013 and 2015 he was a visiting scholar at University of California, San Diego. He is the recipient of the best student paper award at the 2014 IEEE Sensor Array and Multichannel Signal Processing workshop and also received an award from IDA Efondet for his masters thesis. His research interests include signal processing, machine learning and wireless communications.

Bernard Henri Fleury (M'97–SM'99) received the Diplomas in electrical engineering and in mathematics in 1978 and 1990 respectively and the Ph.D. degree in electrical engineering in 1990 from the Swiss Federal Institute of Technology Zurich (ETHZ), Zurich, Switzerland.

Since 1997, he has been with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark, as a Professor of Communication Theory. From 2000 till 2014 he was Head of Section, first of the Digital Signal Processing Section and later of the Navigation and Communications Section. From 2006 to 2009, he was partly affiliated as a Key Researcher with the Telecommunications Research Center Vienna (ftw.), Vienna, Austria. During 1978–1985 and 1992–1996, he was a Teaching Assistant and a Senior Research Associate, respectively, with the Communication Technology Laboratory, ETHZ. Between 1988 and 1992, he was a Research Assistant with the Statistical Seminar at ETHZ.

Prof. Fleury's research interests cover numerous aspects within communication theory, signal processing, and machine learning, mainly for wireless communication systems and networks. His current scientific activities include stochastic modeling and estimation of the radio channel, especially for large systems (operating in large bandwidths, equipped with large antenna arrays, etc.), especially when these systems operate in harsh conditions, e.g. in highly time-varying environments; iterative message-passing processing, with focus on the design of efficient feasible architectures for wireless receivers; localization techniques in wireless terrestrial systems; and radar signal processing. Prof. Fleury has authored and coauthored nearly 150 publications and is co-inventor of 6 filed or published patents in these areas. He has developed, with his staff, a high-resolution method for the estimation of radio channel parameters that has found a wide application and has inspired similar estimation techniques both in academia and in industry.