Aalborg Universitet

A Sparse Bayesian Learning Algorithm With Dictionary Parameter Estimation

Hansen, Thomas Lundgaard; Badiu, Mihai Alin; Fleury, Bernard Henri; Rao, Bhaskar D.

A Sparse Bayesian Learning Algorithm
With Dictionary Parameter Estimation

Thomas L. Hansen*, Mihai A. Badiu*, Bernard H. Fleury* and Bhaskar D. Rao[†]

* Dept. of Electronic Systems, Aalborg University, Fr. Bajers Vej 7,
DK-9220, Aalborg, Denmark – {tlh,in_mba, bfl}@es.aau.dk

[†] Dept. of Electrical and Computer Engineering, University of California at San Diego,
9500 Gilman Drive, La Jolla, CA 92093, USA – brao@ucsd.edu

Abstract—This paper concerns sparse decomposition of a noisy signal into atoms which are specified by unknown continuous-valued parameters. An example could be estimation of the model order, frequencies and amplitudes of a superposition of complex sinusoids. The common approach is to reduce the continuous parameter space to a fixed grid of points, thus restricting the solution space. In this work, we avoid discretization by working directly with the signal model containing parameterized atoms. Inspired by the "fast inference scheme" by Tipping and Faul we develop a novel sparse Bayesian learning (SBL) algorithm, which estimates the atom parameters along with the model order and weighting coefficients. Numerical experiments for spectral estimation with closely-spaced frequency components, show that the proposed SBL algorithm outperforms state-of-the-art subspace and compressed sensing methods.

I. INTRODUCTION

Suppose we have a length-N signal $\boldsymbol{x}$ given by a weighted sum of $K \ll N$ elementary functions, so-called signal atoms:

$$\boldsymbol{x} = \sum_{i=1}^{K} \boldsymbol{\psi}(\theta_i)\alpha_i, \qquad (1)$$

where $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_K]^{\text{T}}$ is a vector of atom (dictionary) parameters and $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K]^{\text{T}}$ is a vector of weighting coefficients. We specify the atoms by the vector-valued function $\boldsymbol{\psi} : [0, 1) \rightarrow \mathbb{C}^N$.[1] We take noisy measurements $\boldsymbol{y} \in \mathbb{C}^N$ as $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w}$, where $\boldsymbol{w} \in \mathbb{C}^N$ is a zero-mean complex Gaussian noise vector with independent and identical distributed (i.i.d.) entries of variance λ^{-1}.

The problem of estimating K, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ given the atom specification $\boldsymbol{\psi}(\cdot)$ and observation $\boldsymbol{y}$ is ubiquitous in signal processing. When the Fourier atom is considered, i.e. $\boldsymbol{\psi}(\theta_i) = [\text{e}^{j2\pi\theta_i 0}, \cdots, \text{e}^{j2\pi\theta_i(N-1)}]^{\text{T}}$, the problem reduces to the line spectral estimation problem, which has a wide range of applications, e.g. direction of arrival estimation. The problem also arises in the context of compressed sensing (CS) reconstruction: When a sensing matrix $\boldsymbol{\Phi}$ is employed, the atoms in our model become $\boldsymbol{\psi}(\cdot) = \boldsymbol{\Phi}\tilde{\boldsymbol{\psi}}(\cdot)$, where $\tilde{\boldsymbol{\psi}}(\cdot)$ is the atom specified without a sensing matrix.

A common approach, particularly in CS [1], is to discretize the parameter space into a (uniform) grid of $M \geq N$ values on $[0, 1)$. Then, a sparse representation of $\boldsymbol{x}$ is sought in the

[1]Any function $\psi(\cdot)$ having a connected and bounded domain in $\mathbb{R}$ can be expressed in this form.

dictionary obtained by evaluating $\boldsymbol{\psi}(\cdot)$ at the M grid points. However, the mismatch between the atoms in the dictionary and the true atoms limits the accuracy of the fixed dictionary approach. To combat the severity of the atom mismatch a finer grid can be employed, leading to two undesired effects: *a)* the dictionary becomes increasingly coherent, rendering the estimation problem ill-posed [1] and *b)* the larger size of the dictionary results in higher computational cost of estimation.

Model-based CS can be applied to mitigate the coherence issue occurring with fixed dictionaries, see for example [2], [3]. If the objective is to reconstruct the signal vector $\boldsymbol{x}$, and not to find a decomposition into atoms, "analysis sparsity" [4], [5] alleviates the need for dictionary incoherence. All these coherence-controlling methods suffer from high computational complexity when a fine grid is used.

Recent works [6] formulate our estimation problem as a total variation norm minimization problem. The strongly related parallel development [7] uses a similar atomic norm minimization. Minimizing the ℓ_1-norm promotes sparse estimates. The idea in [6], [7] is to generalize the ℓ_1-norm from vectors to the real line, such that minimizing the norm promotes a sparse signal on the real line, i.e. a sum of spikes. For Fourier atoms, [6], [7] rewrite the norm minimization as a semi-definite program. The theoretical analysis of these works requires the frequencies in $\boldsymbol{\theta}$ to be well-separated, e.g. by at least $2/N$ in [6].

Another recently-proposed approach is to introduce a complementary dictionary which characterizes the basis mismatch, e.g. [8], [9].

For Fourier atoms (1) reduces to the line spectral estimation problem and many methods have been proposed; see [10] for a list of references. The most prominent of these methods are the so-called subspace methods [11], e.g. ESPRIT [12].

In this paper, we devise a sparse Bayesian learning (SBL) algorithm for estimating K, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. Since most SBL methods (e.g. [13]–[16]) are developed for discrete dictionaries, they suffer from the mentioned drawbacks when applied to our problem. Therefore, we instead use the parameterized model (1) to devise our algorithm. Specifically, we extend the sparse prior model proposed in [13] and devise an inference scheme, inspired by [14], which estimates K, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. A parallel development is found in [17], which uses a variational Bayesian method for estimation.
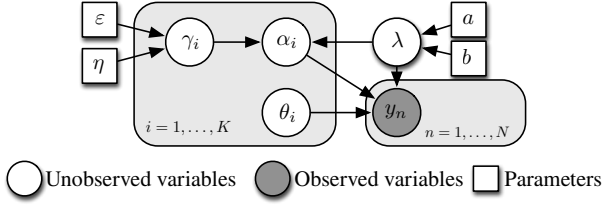
Fig. 1. Bayesian network of the probabilistic model.

II. PROBABILISTIC MODELLING

Our probabilistic model is an extension of that in [13] to include modelling of the atom parameter vector $\boldsymbol{\theta}$. The Bayesian network representation of our model is shown in Fig. 1. We note that the model in [13] is itself obtained by modifying the model in [15] such that $\boldsymbol{\alpha}$ depends on λ. This modification is employed since it gives certain benefits in computational complexity as discussed in [13].[2]

The following treatment is valid for both the real and complex cases.[3] The observation $\boldsymbol{y}$ is taken in white Gaussian noise with variance λ^{-1}:

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \lambda) = \mathrm{N}(\boldsymbol{y}|\boldsymbol{\Psi}(\boldsymbol{\theta})\boldsymbol{\alpha}, \lambda^{-1}\mathbf{I}), \tag{2}$$

where the dictionary matrix $\boldsymbol{\Psi}(\boldsymbol{\theta}) = [\boldsymbol{\psi}(\theta_1), \cdots, \boldsymbol{\psi}(\theta_K)]$ contains the atoms as columns. The noise precision λ is modelled as gamma distributed with shape a and rate b: $p(\lambda) = \mathrm{Ga}(\lambda|a, b)$. The atom parameter vector $\boldsymbol{\theta} \in [0, 1)^K$ has i.i.d. uniformly distributed entries: $p(\boldsymbol{\theta}) = \prod_{i=1}^{K} \mathrm{unif}(\theta_i|0, 1)$. The coefficients $\boldsymbol{\alpha}$ are modelled through a two-layer hierarchical specification. The first layer is a zero-mean Gaussian distribution, i.e. $p(\boldsymbol{\alpha}|\boldsymbol{\gamma}, \lambda) = \mathrm{N}(\boldsymbol{\alpha}|\boldsymbol{0}, \lambda^{-1}\boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. The vector $\boldsymbol{\gamma}$ constitute the second layer. Its entries are modelled as i.i.d. gamma distributed: $p(\boldsymbol{\gamma}) = \prod_{i=1}^{K} \mathrm{Ga}(\gamma_i|\varepsilon, \eta)$.

III. BAYESIAN INFERENCE

To infer on $(K, \boldsymbol{\theta}, \boldsymbol{\alpha})$, we proceed by applying Type-II estimation [18], i.e. we use Bayesian inference to find point estimates $(\hat{K}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda})$ of the parameters $(K, \boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda)$ and find the estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ as the mode of $p(\boldsymbol{\alpha}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda})$. From Bayes rule, $p(\boldsymbol{\alpha}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda)$ turns out to be a normal density:

$$p(\boldsymbol{\alpha}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda) = \mathrm{N}(\boldsymbol{\alpha}|\boldsymbol{\mu}, \lambda^{-1}\boldsymbol{\Sigma}), \tag{3}$$

where

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{H}}(\boldsymbol{\theta})\boldsymbol{y}, \tag{4}$$

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Psi}^{\mathrm{H}}(\boldsymbol{\theta})\boldsymbol{\Psi}(\boldsymbol{\theta}) + \boldsymbol{\Gamma}^{-1}\right)^{-1}. \tag{5}$$

We find $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda})$ as an approximation of the maximum a posteriori (MAP) estimate

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda})_{\mathrm{MAP}} = \underset{(\boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda)}{\arg\max} \ln p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda|\boldsymbol{y}). \tag{6}$$

[2]Note that with this modification, the entries of $\boldsymbol{\gamma}$ represent the ratio of the weight variances to the noise variance, i.e. signal-to-noise ratios.

[3] The multivariate normal density is parameterized to encompass both the real ($\rho = \frac{1}{2}$) and complex ($\rho = 1$) cases:

$$\mathrm{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\frac{\rho}{\pi}\right)^{\rho\dim(\boldsymbol{x})} |\boldsymbol{\Sigma}|^{-\rho} \exp\left(-\rho(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Following the steps of [13], we proceed by writing the log-posterior by its factors:

$$\ln p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda|\boldsymbol{y}) \propto^{\mathrm{e}} \ln p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda)p(\boldsymbol{\theta})p(\boldsymbol{\gamma})p(\lambda), \tag{7}$$

where $x \propto^{\mathrm{e}} y$ denotes $x = y + \mathrm{const}$. The marginal likelihood can be found by marginalizing the coefficient vector $\boldsymbol{\alpha}$ out:

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda) = \mathrm{N}(\boldsymbol{y}|\boldsymbol{0}, \lambda^{-1}\mathbf{B}), \tag{8}$$

where $\mathbf{B} = \left(\mathbf{I} - \boldsymbol{\Psi}(\boldsymbol{\theta})\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{H}}(\boldsymbol{\theta})\right)^{-1} = \mathbf{I} + \boldsymbol{\Psi}(\boldsymbol{\theta})\boldsymbol{\Gamma}\boldsymbol{\Psi}^{\mathrm{H}}(\boldsymbol{\theta})$. Using this result, we can rewrite the log-posterior (7) such that one set of parameters $(\theta_i, \gamma_i, \lambda)$ appear explicitly:

$$\begin{aligned} \ln p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda|\boldsymbol{y}) \propto^{\mathrm{e}} &\ \rho N \ln \lambda - \rho \ln |\mathbf{B}_{-i}| \\ &- \rho \ln (1 + \gamma_i s_i) - \rho \lambda \boldsymbol{y}^{\mathrm{H}} \mathbf{B}_{-i}^{-1} \boldsymbol{y} + \frac{\rho \lambda |q_i|^2}{\gamma_i^{-1} + s_i} \\ &+ \sum_{k=1}^{K} \left\{ (\varepsilon - 1) \ln \gamma_k - \eta \gamma_k \right\} + (a - 1) \ln \lambda - b\lambda, \end{aligned} \tag{9}$$

where $\mathbf{B}_{-i} = \mathbf{I} + \boldsymbol{\Psi}(\boldsymbol{\theta}_{-i})\boldsymbol{\Gamma}_{-i}\boldsymbol{\Psi}^{\mathrm{H}}(\boldsymbol{\theta}_{-i})$ and $\boldsymbol{\Gamma}_{-i} = \mathrm{diag}(\boldsymbol{\gamma}_{-i})$. The notation $\boldsymbol{a}_{-i}$ denotes a vector $\boldsymbol{a}$ with the ith component removed. We have further defined the quantities

$$s_i \triangleq \boldsymbol{\psi}^{\mathrm{H}}(\theta_i)\mathbf{B}_{-i}^{-1}\boldsymbol{\psi}(\theta_i) \quad \text{and} \quad q_i \triangleq \boldsymbol{\psi}^{\mathrm{H}}(\theta_i)\mathbf{B}_{-i}^{-1}\boldsymbol{y}. \tag{10}$$

To estimate the parameters $\{\theta_i, \gamma_i\}_{i=1,\dots,\hat{K}}$ and λ we iteratively maximize (9) with respect to one parameter, while keeping all other parameters fixed at their current estimate. To do so, we take partial derivatives of (9) w.r.t λ and γ_i and solve for the roots. Following a procedure similar to that in [15], to analyse the stationary points, we get the updates

$$\hat{\lambda} = \frac{\rho N + a - 1}{\rho \boldsymbol{y}^{\mathrm{H}} \hat{\mathbf{B}}^{-1} \boldsymbol{y} - b} \tag{11}$$

$$\hat{\gamma}_i = \begin{cases} \frac{-(2\varepsilon - 2 - \rho)\hat{s}_i - \rho\hat{\lambda}|\hat{q}_i|^2 - \sqrt{\Delta}}{2(\varepsilon - 1 - \rho)\hat{s}_i^2} & \rho\hat{\lambda}|\hat{q}_i|^2 > \delta, \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

where $\Delta = \left((2\varepsilon - 2 - \rho)\hat{s}_i + \rho\hat{\lambda}|\hat{q}_i|^2\right)^2 - 4(\varepsilon - 1)(\varepsilon - 1 - \rho)\hat{s}_i^2$ and $\delta = \left(2 + \rho - 2\varepsilon + 2\sqrt{(1 - \varepsilon)(1 + \rho - \varepsilon)}\right)\hat{s}_i$.

It is not tractable to obtain similar closed-from expression for updating $\hat{\theta}_i$. We instead use Newton's method for unconstrained optimization, which iteratively updates $\hat{\theta}_i$ as

$$\hat{\theta}_i^{\mathrm{new}} = \hat{\theta}_i^{\mathrm{old}} - l'\left(\hat{\theta}_i^{\mathrm{old}}\right) \Big/ l''\left(\hat{\theta}_i^{\mathrm{old}}\right), \tag{13}$$

where $l'(\theta_i)$ and $l''(\theta_i)$ are the first and second partial derivatives of (9) with respect to θ_i as given in the Appendix.

Algorithm 1 combines update equations (11), (12) and (13) into a constructive sequential algorithm for estimating $(K, \boldsymbol{\theta}, \boldsymbol{\alpha})$. The algorithm follows the idea of the fast inference scheme [14] and starts with an empty model and iteratively adds atoms into the model. Due to the highly-nonlinear form of (9) as a function of θ_i, the search for candidate atoms to include in the model is aided by a grid search. We have found that using the Jeffreys prior for γ_i and λ, obtained as the special case $\varepsilon = \eta = a = b = 0$, yields good performance.

Algorithm 1: SBL with Dictionary Parameter Estimation

Input: Signal measurement $\boldsymbol{y}$.

Output: Estimates of the model order $\hat{K}$, atom parameters in $\hat{\boldsymbol{\theta}} \in [0,1)^{\hat{K}}$ and coefficients in $\hat{\boldsymbol{\alpha}} \in \mathbb{C}^{\hat{K}}$.

Parameters: Prior parameters ε, a, b ($\eta = 0$ assumed.)

1   $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) \leftarrow$ Empty vectors; $\hat{\lambda} \leftarrow 100$

2   **while** *Stopping criterion not met* **do**

3   **if** *Iteration number* $\in \{0, 5, 10, 20, 30, \ldots\}$ **then**

4   **for** $i = 0, \ldots, 3N - 1$ **do**

5         $\hat{\theta}_{\text{candidate}} \leftarrow$ Calc. from (13), with $\hat{\theta}^{\text{old}} = \frac{i}{3N}$.

6         $\hat{\gamma}_{\text{candidate}} \leftarrow$ Calculate from (12).

7   **if** $\gamma_{\text{candidate}} > 0$ **then**

8   Append $(\theta_{\text{candidate}}, \gamma_{\text{candidate}})$ to $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$.

9   **end**

10   **end**

11   **end**

12   **for** $i = 1, \ldots, ||\hat{\boldsymbol{\theta}}||_0$ **do**

13       $\hat{\theta}_i \leftarrow$ Update from (13).

14       $\hat{\gamma}_i \leftarrow$ Update from (12).

15   **if** $\hat{\gamma}_i == 0$ **then**

16   Remove component i from $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$.

17   **end**

18   **end**

19     $\hat{\lambda} \leftarrow$ Update from (11).

20   **end**

21   $\hat{\boldsymbol{\alpha}} \leftarrow \hat{\boldsymbol{\mu}}$ $\hat{\boldsymbol{\mu}}$ calculated from (4) based on $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$.

We have observed that applying the update (13) more than once in lines 5 and 13, does not improve the overall convergence rate significantly for Fourier atoms. As a stopping criterion in the kth iteration we use $||\hat{\boldsymbol{\theta}}^k - \hat{\boldsymbol{\theta}}^{k-1}||_\infty < 10^{-6}/N$. To reduce computational complexity, the matrix $\hat{\boldsymbol{\Sigma}}$ can be iteratively updated as atoms are added and removed from the model, by using the expressions in [16]. With these updates, the algorithm has computational complexity per iteration of $\mathcal{O}(N^2 \hat{K})$ when searching for new atoms to add to the model and $\mathcal{O}(N \hat{K}^2)$ otherwise.

IV. NUMERICAL EXPERIMENTS

To investigate the performance of our algorithm, we conduct a numerical experiment with Fourier atoms. The simulation code is available online at http://doi.org/srp. We use $N = 100$ measurements and $K = 10$ signal components. We generate $\boldsymbol{\theta}$ consisting of 5 pairs of frequencies. The distance $d(\theta_i, \theta_{i+1})$ between two paired frequencies is i.i.d. uniform random on the interval $[0.7/N, 1/N]$. The distance metric $d(\cdot, \cdot)$ is the wrap-around distance on the interval $[0, 1)$. The pairs are located randomly such that any set of frequencies (θ_i, θ_j) which are not paired are separated by at least $d(\theta_i, \theta_j) \geq 1.5/N$. The complex amplitudes of the signal components in $\boldsymbol{\alpha}$ are generated i.i.d. with uniform random phase on $[0, 2\pi)$ and amplitudes drawn from a normal density of mean 1 and variance 0.1.

We measure performance in terms of normalized mean-squared error (MSE) of the reconstructed signal $\hat{\boldsymbol{x}} = \boldsymbol{\Psi}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\alpha}}$ and the following performance metric for $\hat{\boldsymbol{\theta}}$:

$$\beta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \triangleq \frac{1}{K} \sum_{i=1}^{K} \left(\min_{\hat{\theta} \in \hat{\boldsymbol{\theta}}} d(\hat{\theta}, \theta_i) \right)^2, \tag{14}$$

where the notation $\hat{\theta} \in \hat{\boldsymbol{\theta}}$ signifies that $\hat{\theta}$ is given as one of the entries in $\hat{\boldsymbol{\theta}}$. Further, we show histograms of the model order estimation error at $15\,$dB SNR for some of the algorithms. All reported values are averaged over 100 trials.

We compare our algorithm with the fixed dictionary SBL algorithm [13], which our algorithm is based on, also using Jeffreys priors for γ_i and λ ($\varepsilon = \eta = a = b = 0$). Further, we compare with a selection of state-of-the-art estimators: total variation norm minimization via semi-definite programming (SDP) [6]; a complex reformulation of continuous basis pursuit (CCBP) from [9]; the integer program employed as part of spectral iterative hard thresholding (SIHT IP) [2]; "analysis" basis-pursuit denoising (BPDN) [5]; and ESPRIT [12].

Fixed dictionary SBL, SIHT IP and analysis BPDN are given a fixed dictionary of size $M = 10N$. Due to computational feasibility, CCBP is given a smaller dictionary of size $M = 3N$. ESPRIT is given the true model order, i.e. $\hat{K} = K$, and SIHT IP is given $\hat{K} = 3K$ as the model order. The regularization parameters of SDP and CCBP are set based on the true noise variance λ. The above choices of $\hat{K}$ and regularization parameters are based on preliminary numerical investigations to obtain the best performance of the algorithms.

For the reconstruction MSE, we compare with an oracle estimator which computes a least-squares solution to find $\hat{\boldsymbol{\alpha}}$ from the true set of frequencies. The Cramér-Rao bound for $\hat{\boldsymbol{\theta}}$ (CRB) is calculated as in [19] and averaged over all simulated signal realizations. The CRB is a lower bound on $\beta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, assuming that the matching of an estimate $\hat{\theta}$ to each of the true frequencies θ_i in the minimization of (14) is one-to-one.

The simulation results are shown in Fig. 2. We see that our proposed SBL algorithm is superior to all other algorithms at SNR $\geq 5\,$dB. At SNR $\geq 15\,$dB ESPRIT performs similarly to our algorithm in terms of reconstruction MSE. The proposed SBL algorithm attains the CRB of $\beta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ at SNR $\geq 5\,$dB, while ESPRIT only almost does so at SNR $\geq 20\,$dB.

All algorithms which rely on a fixed dictionary (fixed dict. SBL, CCBP and SIHT IP) show a saturation effect in $\beta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$; beyond a certain SNR threshold the performance does not improve since the resolution limit of the dictionary is reached. Notice how CCBP performs similar to fixed dict. SBL in terms of $\beta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, even though it uses a much smaller dictionary.

For fixed dict. SBL, SDP and SIHT IP the saturation effect in terms of $\beta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ does not give rise to a corresponding saturation in reconstruction MSE. These three estimators are, however, significantly overestimating K in high SNR (see Fig. 2c and remember that $\hat{K} = 3K$ for SIHT IP). Our algorithm has small model order estimation error for all SNR values greater than $0\,$dB and still out-performs ESPRIT which is given the true model order.

(a) Normalized reconstruction MSE.   (b) Performance metric for $\hat{\boldsymbol{\theta}}$.   (c) Mean model order estimation error with shaded regions giving the $20-80\,\%$ quantiles.
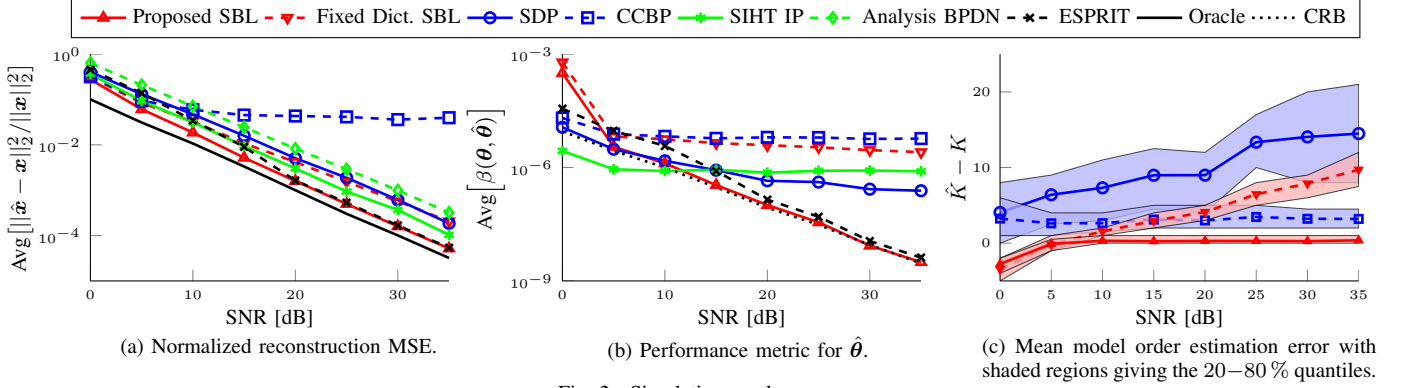
Fig. 2. Simulation results.

V. CONCLUSION

In this paper we addressed the so-called off-grid sparse decomposition problem which consists in decomposing a noisy signal into atoms specified by unknown continuous-valued parameters. We found that a convenient way to avoid the undesired effects of a discretized atom parameter space is to consider the atom parameters as unknown variables to be estimated along with the model order and atom coefficients.

Thus, we proposed a novel SBL algorithm based on an extension of the probabilistic model in [13]. Inspired by the constructive scheme of [14], we devised update expressions which provide a simple criterion for inclusion of an atom into the estimated model; candidate atoms for inclusion are identified via a grid search. Unlike for the rest of the variables, the updates of the atom parameters cannot be computed in closed-form, so we resorted to Newton's method to update these estimates.

The numerical results show that our algorithm is superior to the reference algorithms for spectral estimation with closely-spaced frequency components. This is remarkable since our algorithm estimates the model order, while many of the reference algorithms are given the true value. An interesting aspect for further research is to reduce the computational demands of the algorithm, in particular that connected with the search for new atoms to include in the model estimate.

APPENDIX: PARTIAL DERIVATIVES OF (9) W.R.T. θ_i

$$l'(\theta_i) = \frac{\rho\hat{\lambda}\hat{\gamma}_i}{1+\hat{\gamma}_i\hat{s}_i}\frac{\partial|\hat{q}_i|^2}{\partial\theta_i} - \left(\frac{\rho\hat{\gamma}_i}{1+\hat{\gamma}_i\hat{s}_i} + \frac{\rho\hat{\lambda}\hat{\gamma}_i^2|\hat{q}_i|^2}{(1+\hat{\gamma}_i\hat{s}_i)^2}\right)\frac{\partial\hat{s}_i}{\partial\theta_i}$$

$$l''(\theta_i) = \left(\hat{\lambda}\frac{\partial^2|\hat{q}_i|^2}{\partial\theta_i^2} - \frac{\partial^2\hat{s}_i}{\partial\theta_i^2}\right)\frac{\rho\hat{\gamma}_i}{1+\hat{\gamma}_i\hat{s}_i} + \left(\frac{\partial\hat{s}_i}{\partial\theta_i}\right)^2\frac{2\rho\hat{\lambda}\hat{\gamma}_i^3|\hat{q}_i|^2}{(1+\hat{\gamma}_i\hat{s}_i)^3}$$
$$+ \left(\left(\frac{\partial\hat{s}_i}{\partial\theta_i}\right)^2 - 2\hat{\lambda}\frac{\partial\hat{s}_i}{\partial\theta_i}\frac{\partial|\hat{q}_i|^2}{\partial\theta_i} - \hat{\lambda}|\hat{q}_i|^2\frac{\partial^2\hat{s}_i}{\partial\theta_i^2}\right)\frac{\rho\hat{\gamma}_i^2}{(1+\hat{\gamma}_i\hat{s}_i)^2}$$

$$\frac{\partial\hat{s}_i}{\partial\theta_i} = 2\,\mathrm{Re}\left\{\boldsymbol{\psi}^{\mathrm{H}}(\theta_i)\hat{\mathbf{B}}_{-i}^{-1}\frac{\partial\boldsymbol{\psi}(\theta_i)}{\partial\theta_i}\right\}$$

$$\frac{\partial|\hat{q}_i|^2}{\partial\theta_i} = 2\,\mathrm{Re}\left\{\hat{q}_i\boldsymbol{y}^{\mathrm{H}}\hat{\mathbf{B}}_{-i}^{-1}\frac{\partial\boldsymbol{\psi}(\theta_i)}{\partial\theta_i}\right\}$$

$$\frac{\partial^2\hat{s}_i}{\partial\theta_i^2} = 2\,\mathrm{Re}\left\{\boldsymbol{\psi}^{\mathrm{H}}(\theta_i)\hat{\mathbf{B}}_{-i}^{-1}\frac{\partial^2\boldsymbol{\psi}(\theta_i)}{\partial\theta_i^2} + \frac{\partial\boldsymbol{\psi}^{\mathrm{H}}(\theta_i)}{\partial\theta_i}\hat{\mathbf{B}}_{-i}^{-1}\frac{\partial\boldsymbol{\psi}(\theta_i)}{\partial\theta_i}\right\}$$

$$\frac{\partial^2|\hat{q}_i|^2}{\partial\theta_i^2} = 2\,\mathrm{Re}\left\{\boldsymbol{y}^{\mathrm{H}}\hat{\mathbf{B}}_{-i}^{-1}\left[\frac{\partial\boldsymbol{\psi}(\theta_i)}{\partial\theta_i}\frac{\partial\boldsymbol{\psi}^{\mathrm{H}}(\theta_i)}{\partial\theta_i} + \frac{\partial^2\boldsymbol{\psi}(\theta_i)}{\partial\theta_i^2}\boldsymbol{\psi}^{\mathrm{H}}(\theta_i)\right]\hat{\mathbf{B}}_{-i}^{-1}\boldsymbol{y}\right\}.$$

For Fourier atoms we have $\frac{\partial\boldsymbol{\psi}(\theta_i)}{\partial\theta_i} = \mathbf{D}\boldsymbol{\psi}(\theta_i)$ and $\frac{\partial^2\boldsymbol{\psi}(\theta_i)}{\partial\theta_i^2} = \mathbf{D}^2\boldsymbol{\psi}(\theta_i)$, where $\mathbf{D} = \mathrm{diag}([0,\,j2\pi,\,\cdots,\,j2\pi(N-1)])$.

REFERENCES

[1] T. Strohmer, "Measure what should be measured: Progress and challenges in compressive sensing," *IEEE Signal Process. Lett.*, vol. 19, pp. 887–893, Dec. 2012.

[2] M. Duarte and R. Baraniuk, "Spectral compressive sensing," *Appl. and Computational Harmonic Anal.*, vol. 35, pp. 111–129, Jul. 2013.

[3] A. Fannjiang and W. Liao, "Coherence pattern-guided compressive sensing with unresolved grids," *SIAM J. Imaging Sci.*, vol. 5, pp. 179–202, 2012.

[4] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse problems*, vol. 23, no. 3, p. 947, Jun. 2007.

[5] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. and Computational Harmonic Anal.*, vol. 31, pp. 59–73, Jul. 2011.

[6] E. J. Candès and C. Fernandez-Granda, "Super-resolution from noisy data," *J. of Fourier Anal. and Applicat.*, vol. 19, no. 6, pp. 1229–1254, Dec. 2013.

[7] B. N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5987–5999, Dec. 2013.

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

[9] K. Fyhn, M. F. Duarte, and S. H. Jensen, "Compressive parameter estimation for sparse translation-invariant signals using polar interpolation," 2013, submitted to *IEEE Trans. Signal Process.*, arXiv:1305.3483.

[10] P. Stoica, "List of references on spectral line analysis," *Signal Process.*, vol. 31, no. 3, pp. 329–340, Apr. 1993.

[11] S.-Y. Kung, K. S. Arun, and B. D. V. Rao, "State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem," *J. of the Optical Soc. of America*, vol. 73, pp. 1799–1811, Dec. 1983.

[12] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 984–995, Jul. 1989.

[13] T. L. Hansen, P. B. Jørgensen, N. L. Pedersen, C. N. Manchón, and B. H. Fleury, "Bayesian compressed sensing with unknown measurement noise level," in *Asilomar Conf. Signals, Syst., and Computers*, Nov. 2013.

[14] M. E. Tipping and A. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *Proc. 9th Int. Workshop Artificial Intell. and Stat.*, Jan. 2003, pp. 3–6.

[15] N. L. Pedersen, C. N. Manchón, M.-A. Badiu, D. Shutin, and B. H. Fleury, "Sparse estimation using Bayesian hierarchical prior modeling for real and complex models," 2013, submitted to *IEEE Trans. Signal Process.*, arXiv:1108.4324.

[16] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 92–106, Jan. 2009.

[17] D. Shutin, W. Wang, and T. Jost, "Incremental sparse Bayesian learning for parameter estimation of superimposed signals," in *Proc. 10th Int. Conf. Sampling Theory and Applicat. (SampTA)*, Jul. 2013.

[18] D. P. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6236–6255, Sep. 2011.

[19] P. Stoica and N. Arye, "MUSIC, maximum likelihood, and Cramèr-Rao bound," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 5, pp. 720–741, May 1989.