HYPERPARAMETER-FREE SPARSE REGRESSION OF GROUPED VARIABLES

Ted Kronvall, Stefan Ingi Adalbjörnsson, Santhosh Nadig, and Andreas Jakobsson

Dept. of Mathematical Statistics, Lund University, Sweden

ABSTRACT

In this paper, we introduce a novel framework for semi-parametric estimation of an unknown number of signals, each parametrized by a group of components. Via a reformulation of the covariance fitting criteria, we formulate a convex optimization problem over a grid of candidate representations, promoting solutions with only a few active groups. Utilizing the covariance fitting allows for a hyperparameter-free estimation procedure, highly robust against coherency between candidates, while still allowing for a computationally efficient implementation. Numerical simulations illustrate how the proposed method offers a performance similar to the group-LASSO for incoherent dictionaries, and superior performance for coherent dictionaries.

Index Terms— covariance fitting, group sparsity, convex optimization, multi-pitch estimation.

1. INTRODUCTION

The notable recent theoretical and algorithmic development in sparse reconstruction modeling, compressive sensing, and the related areas, have resulted in a convenient toolbox of methods, allowing practitioners to relatively easily tackle a wide range of problems (see, e.g., [1–4]). Commonly, the problems considered can be either transformed into, or well approximated by, a sum of an unknown (small) number of vectors selected from a very large set of possible vectors, or, as we will consider in this paper, by a small number of subsets, or groups, of such possible vectors [5–13]. However, for most such algorithms, there is one or more tuning parameters that need to be set, typically used to control the level of sparsity desired in the solution. Although such tuning parameters may in many cases be chosen using cross-validation or by some other heuristics, such solutions may in other cases be time-consuming, computationally complex, and/or difficult. In a recent effort to formulate solutions free from such such hyperparameters, the semi-parametric sparse iterative covariance estimate (SPICE) method was introduced in [14] for the estimation of line spectra. The method has been shown to offer accurate and robust estimates, as well as being closely related to the least absolute deviation LASSO and the square-root LASSO methods [15–19]. In this work, we extend on these formulations, allowing also for group sparse cases, such as relevant, for instance, in the formulation of multi-pitch estimation problems (see, e.g., [8]). We show that the resulting estimator may be efficiently implemented using a sequence of simple optimization problems, computable in closed form.

2. GROUP SPARSITY VIA COVARIANCE FITTING

Consider a length N complex-valued measurement, constituting a mix of C sources, each parametrized by a group of L_c components, sampled in noise, such that (see also, e.g., [8,9])

$$\mathbf{y} = \sum_{c=1}^{C} \mathbf{s}_c + \mathbf{e}' \tag{1}$$

where $\mathbf{e}' \in \mathbb{C}^N$ denotes the noise components, and where

$$\mathbf{s}_c = \sum_{\ell=1}^{L_k} \mathbf{a}(\theta_c, \ell) x_{\theta_c,\ell}, \quad \mathbf{s}_c \in \mathbb{C}^N \tag{2}$$

is the parametrization of the c:th source, with $\mathbf{a}(\theta_c, n) \in \mathbb{C}^N$ denoting the signal response vector, or steering vector, fully described by θ_c and n, and $x_{\theta_c,n}$ its corresponding complex-valued amplitude. Thus, the signal is fully parametrized by the (unknown) parameters

$$\{\theta_c, x_{\theta_c,1}, \ldots, x_{\theta_c,L_c}\}_{c=1,\ldots C} \tag{3}$$

Typically, also the number of sources, C, and the model order of each source, L_c, are unknown, complicating the problem. Using a sparse reconstructing framework, we proceed to introduce a dictionary of possible candidates over θ, i.e., consisting of θ_k, for $k = 1 \ldots K$, with $K \gg C$ selected large enough to ensure that some of the K candidates well coincides with the true parameters. To simplify the notation, we hereafter simply use $(\cdot)_k$ in place of $(\cdot)_{\theta_k}$, allowing (1) to expressed compactly as

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{A}_k \mathbf{x}_k + \mathbf{e} \tag{4}$$

where $\mathbf{e}$ is analogue to $\mathbf{e}'$, and

This work was supported in part by the Swedish Research Council and the Royal Physiographic Society in Lund.

Algorithm 1 The proposed group-SPICE algorithm

```
initialize $j \leftarrow 0$, and set
for all $(k, \ell)$ do
    $p_{k,\ell}{}^{(0)} = \frac{|\mathbf{a}_{k,\ell}^H \mathbf{y}|^2}{||\mathbf{a}_{k,\ell}||^4}$
end for
repeat
    covariance update:
    $\mathbf{R} = \sum_{k=1}^{K+N} p_{k,\ell}{}^{(j)} ||\mathbf{a}_{k,\ell}||^2, \; \mathbf{z} = \mathbf{R} \backslash \mathbf{y}$
    power update:
    for all $(k, \ell)$ do
        $r_{k,\ell} = |\mathbf{a}_{k,\ell}^H \mathbf{z}|$, and
        $p_{k,\ell}{}^{(j+1)} = \frac{\left(p_{k,\ell}{}^{(j)} r_{k,\ell}\right)^{2/3} \left(\sum_{\ell=1}^{L_k} \left(p_{k,\ell}{}^{(j)} r_{k,\ell}\right)^{4/3}\right)^{1/4}}{\sqrt{||\mathbf{w}_k||_2}}$
    end for
    $j \leftarrow j + 1$
until convergence
for all $(k, \ell)$ do
    $\hat{x}_{k,\ell} = p_{k,\ell}{}^{(\text{end})} \, \mathbf{a}_{k,\ell}^H \mathbf{z}$
end for
```

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{a}_{k,1} & \dots & \mathbf{a}_{k,L_k} \end{bmatrix} \tag{5}$$

implying that

$$\mathbf{\Sigma} = \mathbb{E}(\mathbf{e}\mathbf{e}^H) = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_N \end{bmatrix} \tag{6}$$

where $\mathbb{E}(\cdot)$ denotes the expectation. Assuming that the phases of $x_{k,l}$ are independent and uniformly distributed on $[0,\ 2\pi]$, the covariance matrix of the measurement, $\mathbf{R} = \mathbb{E}(\mathbf{y}\mathbf{y}^H)$, may thus be expressed as

$$\mathbf{R} = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} |x_{k,\ell}|^2 \mathbf{a}_{k,\ell} \mathbf{a}_{k,\ell}^H + \mathbf{\Sigma} \triangleq \mathbf{A}\mathbf{P}\mathbf{A}^H \tag{7}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \cdots & \mathbf{A}_K & \mathbf{I} \end{bmatrix} \tag{8}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \mathbf{P}_2 & 0 & \cdots & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \sigma_1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \sigma_N \end{bmatrix} \tag{9}$$

$$\mathbf{P}_k = \text{diag}(\mathbf{p}_k) \tag{10}$$

$$\mathbf{p}_k = \begin{bmatrix} |x_{k,1}|^2 & \cdots & |x_{k,L_k}|^2 \end{bmatrix}^\top \tag{11}$$

$$\triangleq \begin{bmatrix} p_{k,1} & \cdots & p_{k,L_k} \end{bmatrix}^\top \tag{12}$$

with diag$(\cdot)$ denoting the diagonal matrix with diagonal elements corresponding to its input argument, and $(\cdot)^\top$ the matrix transpose. For notational simplicity, let

$$p_{k,l} = p_k = \sigma_{k-K}, \quad k = K+1, \ldots, K+N, \; \ell = 1 \tag{13}$$

implying that $L_{K+1} = \cdots = L_{K+N} = 1$. In order to propose a group sparse solution using the covariance fitting criterion (see, e.g., [10], and the references therein), we here proceed to derive an estimator minimizing

$$f = ||\mathbf{R}^{-1/2}(\hat{\mathbf{R}} - \mathbf{R})||_{\mathcal{F}}^2 \tag{14}$$

$$\propto \mathbf{y}^H \mathbf{R}^{-1} \mathbf{y} + \text{tr}(\mathbf{R})/||\mathbf{y}||^2 \tag{15}$$

with $||\cdot||_{\mathcal{F}}$ denoting the column-wise ℓ_2-norm, tr$(\cdot)$ the trace, $\hat{\mathbf{R}} = \mathbf{y}\mathbf{y}^H$, and where (15) is formed by completing the square. It may be noted that (15) may be bound from above as

$$\frac{\text{tr}(\mathbf{R})}{||\mathbf{y}||^2} = \sum_{k=1}^{K+N} \sum_{\ell=1}^{L_k} p_{k,\ell} \overbrace{||\mathbf{a}_{k,\ell}||_2^2/||\mathbf{y}||^2}^{w_{k,\ell}} \tag{16}$$

$$= \sum_{k=1}^{K+N} \mathbf{p}_k^\top \Big[\overbrace{\; w_{k,1} \quad \cdots \quad w_{k,L_k} \;}^{\mathbf{w}_k} \Big]^\top \tag{17}$$

$$= \sum_{k=1}^{K+N} \langle \mathbf{p}_k, \mathbf{w}_k \rangle \leq \sum_{k=1}^{K+N} ||\mathbf{p}_k||_2 ||\mathbf{w}_k||_2 \tag{18}$$

using the Cauchy-Schwartz inequality in forming (18), and where $\langle (\cdot), (\cdot) \rangle$ denotes the inner product. Thus, instead of minimizing (15), the solution may be found by instead minimizing an upper bound of the problem by solving

$$\begin{aligned} \underset{\mathbf{P}}{\text{minimize}} \quad & \mathbf{y}^H \mathbf{R}^{-1} \mathbf{y} + \sum_{k=1}^{K+N} ||\mathbf{p}_k||_2 ||\mathbf{w}_k||_2 \\ \text{s.t.} \quad & \mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H \quad p_{k,\ell} \geq 0, \; \forall p, \ell \end{aligned} \tag{19}$$

This optimization problem contains two expressions, where the first describes the cost of the error, i.e., the misfit between the data and the model, and where the second adds a cost to every non-zero parameter, $p_{k,\ell}$. In comparison with the SPICE method, the minimization in (19), increases the cost of activating components in a new candidate group, k', instead of activating a component within an already active group, k, given that these components model the same data behavior, which thus should promote a group sparse solution.

3. EFFICIENT IMPLEMENTATION

The optimization problem in (19) is convex, and thus enjoys favorable properties, such that any local optima is also the global optimum, that there exists a well defined theory stating necessary and sufficient considitions for optimality, as well as efficient computational methods. Here, we present one such solver. As the first part of the cost function in (19) contains is non-separable in the estimation parameters, $p_{k,\ell}$, one must resort to numerical approximations to solve it directly. Instead, reminiscent to [10], we propose to split the optimization problem into two simpler convex subproblems, each of which a has closed-form solution, and then to solve these iteratively. To make the optimization problem separable in P, we need to introduce an auxiliary variable $\mathbf{Q}$, which fulfills $\mathbf{AQ} = \mathbf{I}$. Given an initial value of $\mathbf{P}$, we may solve for $\mathbf{Q}$ by reformulating (19) as

$$\begin{aligned} \underset{\mathbf{Q}}{\text{minimize}} \quad & \mathbf{y}^H \mathbf{Q}^H \mathbf{P}^{-1} \mathbf{Q} \mathbf{y} \\ \text{s.t.} \quad & \mathbf{AQ} = \mathbf{I} \end{aligned} \tag{20}$$

Utilizing the Shur complement condition for positive semidefiniteness [20], one may verify that (20) is solved for

$$\hat{\mathbf{Q}} = \mathbf{P}\mathbf{A}^H \mathbf{R}^{-1} \tag{21}$$

Also, inputing this back into (20) gives the original expression in (19), why we may equivalently reformulate (19) given $\mathbf{Q}$ as

$$\begin{aligned} \underset{\mathbf{P}}{\text{minimize}} \quad & h = \boldsymbol{\beta}^H \mathbf{P}^{-1} \boldsymbol{\beta} + \sum_{k=1}^{K+N} ||\mathbf{p}_k||_2 ||\mathbf{w}_k||_2 \\ \text{s.t.} \quad & \boldsymbol{\beta} = \mathbf{Q}\mathbf{y} \\ & p_{k,\ell} \geq 0, \forall (k, \ell) \end{aligned} \tag{22}$$

Using some linear algebra, h may be expressed as

$$h = \sum_{k=1}^{K+N} \left(\sum_{\ell=1}^{L_k} \frac{|\beta_{k,\ell}|^2}{p_{k,\ell}} + ||\mathbf{p}_k||_2 ||\mathbf{w}_k||_2 \right) \triangleq \sum_{k=1}^{K+N} h_k \tag{23}$$

and is thus a fully separable problem in the $K + N$ groups. Before deriving a minimizer for h, we ask the reader to notice two aspects of the optimization problem. First, the problem is constrained to non-negative solutions, but the second expression in h is non-differentiable for $p_{k,\ell} = 0$. Second, whenever any parameter do become zero, $h \to \infty$ if $\beta_{k,l} \neq 0$. To that end, the Karush-Kuhn-Tucker (KKT) conditions state that we obtain the global optimum (solving each separable problem h_k) by setting the gradient of it's Lagrangian to zero and solving for each parameter. However, if we may guarantee that the solution will lie in the interior of the feasible set of the problem, i.e., $p_{k,\ell} > 0, \forall \ell$, we may equivalently solve it as an unconstrained problem. To show that this in fact is the case, assume that $p_{k,\ell} > 0$. Subproblem h_k is thus differentiable w.r.t. $p_{k,\ell}$ and we may solve

$$\frac{\partial h_k}{\partial p_{k,\ell}} = -\frac{|\beta_{k,\ell}|^2}{p_{k,\ell}^2} + \frac{p_{k,\ell} ||\mathbf{w}_k||_2}{||\mathbf{p}_k||_2} = 0 \tag{24}$$

yielding

$$p_{k,\ell}^3 = \frac{|\beta_{k,\ell}|^2 ||\mathbf{p}_k||_2}{||\mathbf{w}_k||_2} \tag{25}$$

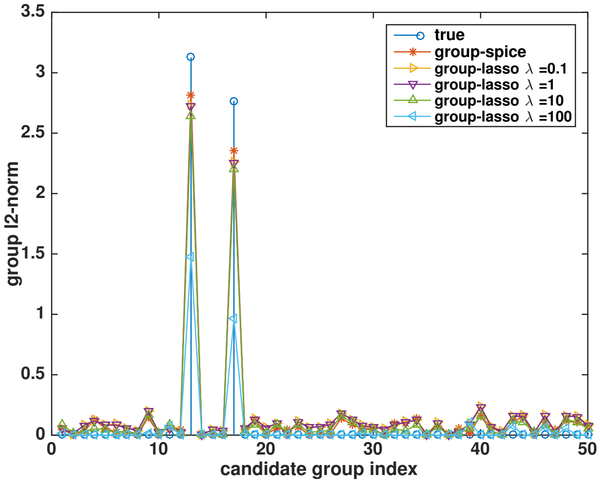

Fig. 1. The group-SPICE and group-LASSO estimates using a dictionary with coherency multiplier $\kappa = 1$, i.e., an incoherent dictionary. The y-axis displays the ℓ_2-norm of the solution for the k:th group.

from which we get that

$$||\mathbf{p}_k||_2 = \left(\sum_{\ell=1}^{L_k} \left(\frac{|\beta_{k,\ell}|^2}{||\mathbf{w}_k||_2} \right)^{2/3} \right)^{3/4} \tag{26}$$

Plugged back into (25), we obtain the estimate

$$\hat{p}_{k,\ell} = \frac{|\beta_{k,\ell}|^{2/3} \left(\sum_{\ell=1}^{L_k} |\beta_{k,\ell}|^{4/3} \right)^{1/4}}{||\mathbf{w}_k||_2^{1/2}} \tag{27}$$

which is valid $\forall (k, \ell)$ in the parameter set. We may thus conclude that whenever $\beta_{k,\ell} \neq 0$, the solution is guaranteed to lie in the interior of the feasible set, and so (27) is also the solution to (22). From (21), we get

$$\beta_{k,\ell} = p_{k,l}\, \mathbf{a}_{k,\ell}^H \mathbf{R}^{-1} \mathbf{y} \tag{28}$$

from which we may conclude that $\beta_{k,\ell} = 0$ if either $p_{k,l} = 0$, or if the steering vector is orthogonal to $\mathbf{R}^{-1}\mathbf{y}$, at which point we set $p_{k,l} = 0$ and exclude it from further estimation. To finally estimate the response vector $\hat{\mathbf{x}}$ from $\hat{\mathbf{P}}$, we apply the LMMSE estimator formula [18], i.e.,

$$\hat{x}_{k,\ell} = \hat{p}_{k,\ell}\, \mathbf{a}_{k,\ell}^H \hat{\mathbf{R}}^{-1} \mathbf{y} \tag{29}$$

and we may thus use the final estimation iterate of $\beta_{k,l}$ as our estimate of the steering response. Algorithm 1 summarizes the proposed method, called the group-SPICE.

4. NUMERICAL RESULTS

In this section, we proceed to evaluate the performance of the proposed group-SPICE estimator, comparing its performance of that of the closely related group-LASSO estimator,

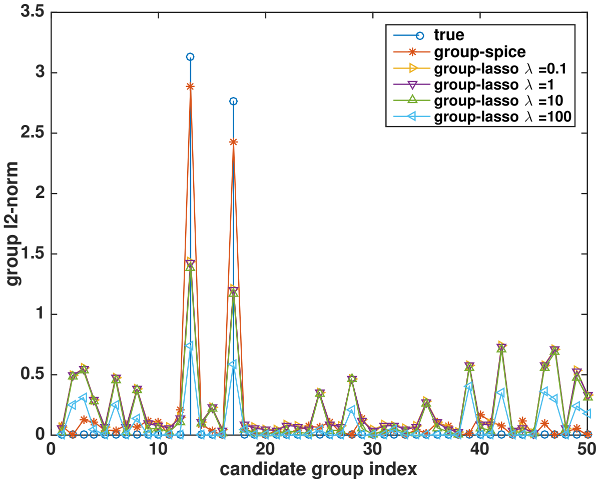

Fig. 2. The group-SPICE and group-LASSO estimates using a dictionary with coherency multiplier $\kappa = 2$. The y-axis displays the ℓ_2-norm of the solution for the k:th group.

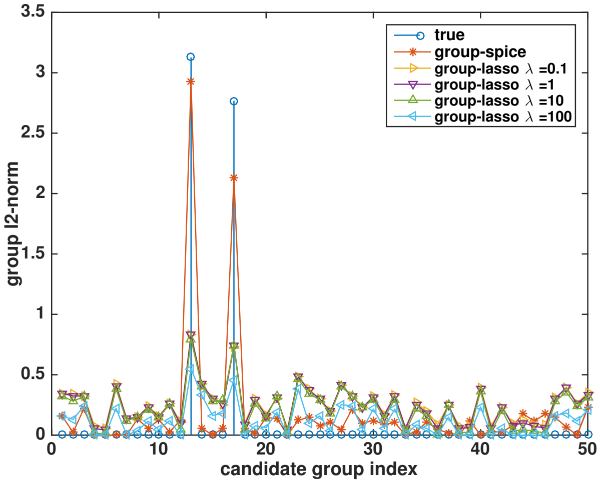

Fig. 3. Group-SPICE and Group-LASSO estimates using a dictionary with coherency multiplier $\kappa = 4$. The y-axis displays the ℓ_2-norm of the solution for the k:th group.

for varying degrees of dictionary coherence. We consider a signal consisting of $N = 100$ samples of a grouped signal, containing $C = 2$ groups, with (a maximum of) $L = 10$ components in each. The groups are chosen at random, here $k = [14,\ 18]$, from a total of $K = 50$ candidates in the dictionary, and are corrupted by a white Gaussian noise. The dictionary is based on a circular-symmetric independent Gaussian matrix, denoted $\mathbf{A}_\kappa \in \mathbb{C}^{N\times(PL/\kappa)}$, which is shown to be incoherent with high probability [21]. Here, $\kappa \in \mathbb{N}_+$ denotes a coherency multiplier, which we use to illustrate the performance of coherent dictionaries. Thus, we choose the dictionary as a random permutation of the columns in a horizontal stacking of the same matrix $\mathbf{A}_\kappa$ κ times, i.e.,

$$\mathbf{A} \sim \mathcal{U}\left([\ \overbrace{\mathbf{A}_\kappa \quad \mathbf{A}_\kappa \quad \cdots \quad \mathbf{A}_\kappa}^{\#\kappa}\ ] \right) \tag{30}$$

where $\mathcal{U}(\cdot)$ denotes random permutation. This has the effect that any component in an active group will also be present $\kappa - 1$ times more in an other or in the same candidate group. Hence, the dictionary will exhibit a high degree of coherence, except when $\kappa = 1$. Figures 1 - 3 illustrate the estimates of group-SPICE and the group-LASSO, the latter with the regularization parameter λ arbitrarily set at some different orders of magnitude. As can be seen from the figures, group-SPICE offers a preferable reconstruction as compared to the group-LASSO for the examined regularization levels, in particular for the coherent dictionaries. Figure 3 in particular show that this effect becomes more pronounced for high levels of coherency, for which the performance of the group-LASSO deteriorates significantly, whereas the group-SPICE estimates remain unaffected.

5. REFERENCES

[1] M. Elad, *Sparse and Redundant Representations*. Springer, 2010.

[2] D. Donoho, “Compressed Sensing,” *IEEE Trans. Inf. Theory*, vol. 52, pp. 1289–1306, 2006.

[3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *The Annals of Statistics*, vol. 32, pp. 407–499, April 2004.

[4] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

[5] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “A Sparse-Group Lasso,” *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[6] D. Malioutov, M. Cetin, and A. S. Willsky, “A Sparse Signal Reconstruction Perspective for Source Localization With Sensor Arrays,” *IEEE Trans. Signal Process.*, vol. 53, pp. 3010–3022, August 2005.

[7] X. Tan, W. Roberts, J. Li, and P. Stoica, “Sparse Learning via Iterative Minimization With Application to MIMO Radar Imaging,” *IEEE Trans. Signal Process.*, vol. 59, pp. 1088–1101, March 2011.

[8] S. I. Adalbjörnsson, T. Kronvall, S. Burgess, K. Åström, and A. Jakobsson, “Sparse Localization of Harmonic Audio Sources,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 117–129, Jan. 2016.

[9] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, “Multi-Pitch Estimation Exploiting Block Sparsity,” *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[10] P. Stoica, P. Babu, and J. Li, “SPICE : a novel covariance-based sparse estimation method for array processing,” *IEEE Trans. Signal Process.*, vol. 59, pp. 629 –638, Feb. 2011.

[11] R. Gribonval and E. Bacry, “Harmonic decomposition of audio signals with matching pursuit,” *IEEE Trans. Signal Process.*, vol. 51, pp. 101–111, jan. 2003.

[12] M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies, “Sparse representations of polyphonic music,” *Signal Processing*, vol. 86, pp. 417–431, March 2006.

[13] J. J. Fuchs, “On the Use of Sparse Representations in the Identification of Line Spectra,” in *17th World Congress IFAC*, (Seoul), pp. 10225–10229, jul 2008.

[14] 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 Trans. Signal Process.*, vol. 59, pp. 35–47, Jan 2011.

[15] P. Babu, *Spectral Analysis of Nonuniformly Sampled Data and Applications*. PhD thesis, Uppsala University, 2012.

[16] P. Stoica and P. Babu, “SPICE and LIKES: Two hyperparameter-free methods for sparse-parameter estimation,” *Signal Processing*, vol. 92, pp. 1580–1590, July 2012.

[17] C. R. Rojas, D. Katselis, and H. Hjalmarsson, “A Note on the SPICE Method,” *IEEE Trans. Signal Process.*, vol. 61, pp. 4545–4551, Sept. 2013.

[18] P. Stoica, D. Zachariah, and L. Li, “Weighted SPICE: A Unified Approach for Hyperparameter-Free Sparse Estimation,” *Digit. Signal Process.*, vol. 33, pp. 1–12, October 2014.

[19] D. Zachariah and P. Stoica, “Online Hyperparameter-Free Sparse Estimation Method,” *IEEE Trans. Signal Process.*, vol. 63, pp. 3348–3359, July 2015.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[21] E. Elhamifar and R. Vidal, “Block-Sparse Recovery via Convex Optimization,” *IEEE Transactions on Signal Processing*, vol. 60, pp. 4094–4107, Aug 2012.