A NUMERICAL IMPLEMENTATION OF GRIDLESS COMPRESSED SENSING

*Ashkan Panahi**, *Mats Viberg**, *Babak Hassibi*†

* Signal processing Group, Chalmers University of Technology, Gothenburg, Sweden
† Department of Electrical Engineering, Caltech, Pasadena, USA

ABSTRACT

Atomic norm denoising has been recently introduced as a generalization of the Least Absolute Shrinkage and Selection Operator (LASSO) to overcome the problem of off-grid parameters. The method has been found to possess many interesting theoretical properties. However, its implementation has been only discussed in a special case of spectral line estimation by uniform sampling. In this paper, we propose a general numerical method to solve the atomic norm denoising problem. The complexity of the proposed algorithm is proportional to the complexity of a single-parameter search in the parameter space and thus in many interesting cases, including frequency estimation it enjoys fast realization.

Index Terms— Atomic norm, LASSO, multiple parameter estimation, implementation

1. INTRODUCTION

Many problems concerning multiple parameter estimation can be formulated as linear sparse estimation problems and solved by the so-called compressive sensing techniques [1, 2, 3]. These problems are commonly refered to as Atomic Decomposition (AD) [4]. The Least Absolute Shrinkage and Selection Operator is a good example of the AD solving compressive sensing techniques. Many interesting properties, such as guaranteed performance and robustness against model uncertainty have been discovered for the LASSO [5, 6]. The LASSO is convex. Hence, it can be implemented by efficient convex optimization techniques [7, 5]. However, its implementation usually involves discretizing the parameter space and selecting from the resulting finite set of discretized parameters, also called the grid [8].

The effect of discretization has been previously studied [9]. It has been observed that when the true parameters are not on the grid, multiple on-grid neighbor parameters are selected by LASSO [10, 2]. In general, it is difficult to distinguish and combine these estimates to achieve a consistent set of parameter estimates. This is known in the literature as the off-grid problem. Different approaches to the off-grid problem have been considered and discussed [11, 12, 13]. One promising approach, considered here is to generalize LASSO to admit a continuum of parameters.

Recently, Atomic Norm DeNoising (ANDN) has been introduced as a generalization of the LASSO, without a need to discretize [14]. It has been shown that this approach inherits the LASSO properties [15, 14]. In fact, it can be shown that the two are identical under any desired precision for a sufficiently dense grid. In particular, this implies that ANDN has a guaranteed performance but suffers resolution limit. In [16, 14] an implementation of ANDN is considered for the special case of frequency estimation by uniform samples, which involves Semidefinie Programing, imposing a high computational burden. An interesting generalization of this approach to the same problem with some prior information can be found in [17]. However, the general implementation of ANDN has not been addressed yet.

In this paper, we propose a technique to implement ANDN for a large group of AD problems. In particular, we show that this algorithm converges to the exact solution of ANDN and the computational complexity is proportional to that of searching over the entire parameter space. Hence, the complexity may be low if either the parameter space is low dimensional, e.g. in the single-dimensional frequency estimation problem, or the parameter space is well-structured.

Our approach is to show that the ANDN has a non-convex parametric representation. However, a global optimality condition for ANDN and thus for this specific non-convex representation has already been introduced [18, 16, 14]. We show that a simple correction of a cyclic coordinate descent based implementation converges to a point satisfying the global optimality condition of ANDN. We demonstrate the numerical results for the case of spectral line estimation with random and fixed sampling schemes.

2. PROBLEM FORMULATION

Consider a set of $m-$dimensional candidate basis vectors $\mathcal{A} \subset \mathbb{C}^m$, known as dictionary and a given data vector $\mathbf{x} \in \mathbb{C}^m$. Suppose that the data is obtained by the following model:

$$\mathbf{x} = \mathbf{a}_1 r_1 + \mathbf{a}_2 r_2 + \ldots + \mathbf{a}_n r_n + \mathbf{n}, \qquad (1)$$

This work was supported in part by the Swedish research council (VR) and in part by the Ericsson Research Foundation.

where $\mathbf{a}_k \in \mathcal{A}$, $r_k > 0$ are real numbers and $\mathbf{n}$ is a random centered Gaussian measurement noise vector with covariance $\sigma^2\mathbf{I}$. The problem of interest herein is to retrieve the basis vectors $\mathbf{a}_k$ as well as the coefficients r_k when the order n is unknown. This is called the Atomic Decomposition (AD) problem [14, 4]. A more familiar form of AD is when the data model is given by

$$\mathbf{x} = \mathbf{b}_1 s_1 + \mathbf{b}_2 s_2 + \ldots + \mathbf{b}_n s_n + \mathbf{n}, \tag{2}$$

where $s_k \in \mathbb{C}$ and $\mathbf{b}_k$ belongs to a dictionary $\mathcal{B}$. However, (2) can be simply written as (1) by defining $r_k = |s_k|$, $s_k = r_k e^{j\phi_k}$ and $\mathbf{a}_k = \mathbf{b}_k e^{j\phi_k}$. Then, the corresponding dictionary in (1) is given by

$$\mathcal{A} = \{\mathbf{b}e^{j\phi} \mid \mathbf{b} \in \mathcal{B},\ 0 \leqslant \phi < 2\pi\}. \tag{3}$$

The AD problem, especially in its latter form in (2) has been long addressed and discussed in different contexts. An interested reader is referred to [19, 20, 21] for more details. We introduce the well-known example of spectral line estimation, where the data $\mathbf{x}$ is obtained by sampling a continuous signal at instants $t_1, t_2, \ldots, t_m$ and the dictionary in the form of (2) is given by

$$\mathcal{B} = \{\mathbf{b}(\omega) = [e^{j\omega t_1}\ e^{j\omega t_2}\ \ldots e^{j\omega t_m}]^T \mid\ -W \leqslant \omega \leqslant W\}. \tag{4}$$

The parameter W is the system bandwidth. We remind that the same problem can be written as (1) using the dictionary given by (3).

The exact solution of AD is generally intractable for large m or n, and the previous techniques do not guarantee such an exact result. A more recent approach in the context of AD is the ANDN, which is closely connected to the sparsity-based linear estimation methods, especially LASSO. The exact recovery by ANDN is well studied and many interesting properties has been discovered. The ANDN is defined as the optimization

$$\min_{\mathbf{y}\in\mathbb{C}^m} \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{y}\|_{\mathcal{A}} \tag{5}$$

where λ is a suitable positive number and the atomic norm $\|\mathbf{y}\|_{\mathcal{A}}$ is defined as

$$\begin{aligned}\|\mathbf{y}\|_{\mathcal{A}} = \min_{n} \min_{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_n} \min_{r_1,r_2,\ldots,r_n\geqslant 0} & r_1 + r_2 + \ldots + r_n \\ \text{s.t.} & \\ \mathbf{y} = \sum_{k=1}^{n} \mathbf{a}_k r_k & \end{aligned} \tag{6}$$

We refer to the solution of (5) by $\bar{\mathbf{y}}$. Replacing $\mathbf{y}$ by $\bar{\mathbf{y}}$ in (6), we denote the parameters at the resulting optimal point by $\bar{n}$, $\bar{\mathbf{a}}_k$ and $\bar{r}_k$, respectively. These are the resulting estimates of the ANDN approach.

Note that the order n is also variable. However, by the so-called Caratheodory Theorem for convex hulls, it is possible to show that (6) attains its minimum with $n \leqslant 2m$. In fact, it suffices to solve (6) only for a fixed order $n \geqslant 2m$, since a smaller order can be simply obtained by setting some parameters r_k (e.g. the first ones) to zero. For the reasons later discussed, we will fix $n = 2m+1$, but for this section let us discuss a variable n.

The optimization in (5) is convex. Thus, its Karush Kuhn Tucker (KKT) condition implies global optimality. The cost in (5) is nondifferentiable. Thus, its KKT condition can be written as

$$\mathbf{x} - \bar{\mathbf{y}} \in \lambda\partial\|\bar{\mathbf{y}}\|_{\mathcal{A}} \tag{7}$$

where we remind that the parameters with bar denote the optimal point in (5) and (6). The subdifferential is given by [14]

$$\partial\|\bar{\mathbf{y}}\|_{\mathcal{A}} = \{\mathbf{z} \mid \sup_{\mathbf{a}\in\mathcal{A}} \Re(\mathbf{z}^H\mathbf{a}) \leqslant 1,\ \forall k\ \Re(\bar{\mathbf{a}}_k^H\mathbf{z}) = 1\} \tag{8}$$

Convexity is not well-defined for (6) due to variable order, but when n is fixed the remaining optimization is non-convex. Let us also rewrite (7) more explicitly as

$$\sup_{\mathbf{a}\in\mathcal{A}} \Re(\bar{\mathbf{z}}^H\mathbf{a}) \leqslant 1, \quad \forall k\ \Re(\bar{\mathbf{a}}_k^H\bar{\mathbf{z}}) = 1 \tag{9}$$

where we define

$$\bar{\mathbf{z}} = \mathbf{x} - \bar{\mathbf{y}} = \mathbf{x} - \sum_{k=1}^{\bar{n}} \bar{\mathbf{a}}_k\bar{r}_k \tag{10}$$

The surprising fact about ANDN is that any set of parameters $\{\bar{\mathbf{a}}_k, \bar{r}_k\}_{k=1}^{\bar{n}}$ satisfying (9) provides a global optimum of (6) [14, 18]. Note that neither (9) is the KKT condition of (6), nor (6) is convex. Still, (9) provides the global optimality of (6).

Due to (9), the bases $\bar{\mathbf{a}}$ can be found as the global maxima of the function $f(\mathbf{a}) = \Re(\bar{\mathbf{z}}^H\mathbf{a})$ given $\bar{\mathbf{y}}$. Together with a technique to find $\bar{\mathbf{y}}$, this constitutes the frequency estimation method in [15]. However, the estimator of $\bar{\mathbf{y}}$ in [15] heavily relies on the special structure of frequency estimation with uniform samples ($t_k = kT$) and cannot be generalized.

3. IMPLEMENTING ANDN

In this section, we develop a general algorithm which exactly solves the ANDN, without relying on the structure of the dictionary. We first substitute the definition of the atomic norm (6) in (5) to obtain.

$$\begin{aligned}\min_{\mathbf{y}\in\mathbb{C}^m} \min_{n} \min_{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_n} \min_{r_1,r_2,\ldots,r_n\geqslant 0} & \tfrac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 + \lambda\sum_{k=1}^{n} r_k \\ \text{s.t.} & \\ \mathbf{y} = \sum_{k=1}^{n} \mathbf{a}_k r_k & \end{aligned} \tag{11}$$

which can be simplified to

$$\min_{n} \min_{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_n} \min_{r_1,r_2,\ldots,r_n\geqslant 0} \frac{1}{2}\|\mathbf{x} - \sum_{k=1}^{n} \mathbf{a}_k r_k\|_2^2 + \lambda\sum_{k=1}^{n} r_k \tag{12}$$

Next, note that the order can be fixed to $2m+1$ as the solution of ANDN satisfies $n \leqslant 2m$. The smaller dimensions are obtained by setting some elements of r_k to zero. We obtain the following non convex optimization problem, whose global optimum, by the previous discussion, coincides with ANDN. Hence, it is characterized by the optimality condition in (9).

$$\min_{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_{2m+1}} \min_{r_1,r_2,\ldots,r_{2m+1}\geqslant 0} \frac{1}{2}\|\mathbf{x}-\sum_{k=1}^{2m+1}\mathbf{a}_k r_k\|_2^2+\lambda\sum_{k=1}^{2m+1} r_k \tag{13}$$

3.1. Cyclic Coordinate Descent Algorithm

The optimization in (13) can be locally solved by a Cyclic Coordinate Descent (CCD) technique [22]. CCD consists of iteration cycles, where each pair of $(\mathbf{a}_l, r_l)$ is updated at an individual iteration by solving (13), keeping the others constant. In this case, the solution has the following closed form

$$\hat{\mathbf{a}}_l = \arg\max_{\mathbf{a}\in\mathcal{A}} \Re(\mathbf{a}^H\mathbf{z}^{(l)}), \quad \hat{r}_l = \frac{(\Re(\hat{\mathbf{a}}_l^H\mathbf{z}^{(l)})-\lambda)_+}{\|\hat{\mathbf{a}}_l\|_2^2} \tag{14}$$

where $(.)_+$ denotes the positive part function and

$$\mathbf{z}^{(l)} = \mathbf{x} - \sum_{k\neq l}\mathbf{a}_k r_k \tag{15}$$

The CCD algorithm is summarized in Algorithm 1.

Algorithm 1 The CCD algorithm

Require: A starting point $\{\mathbf{a}_k\}$ and $\{r_k\}$.
  repeat
    for $l = 1 : n$ **do**
      Replace $(\mathbf{a}_l, r_l)$ with $(\hat{\mathbf{a}}_l, \hat{r}_l)$ in (14).
    end for
  until Convergence

Each iteration of CCD decreases the cost in (13). Thus, the algorithm converges to a stopping point. It is not clear that this stopping point is a global optimum, but remember that this can be verified by the conditions in (9). In this manner, the following theorem can be obtained.

Theorem 1. *The stopping point of CCD consisting of the parameters $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_{2m+1}$ and $\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \ldots, \hat{\mathbf{a}}_{2m+1}$ satisfies* (9)*, and therefore is a global optimum of ANDN if at least one of the elements $\hat{r}_k$ is zero.*

Proof. The stopping point is a local optimal point and satisfies the KKT condition for the minimization in (13), which implies that

$$\begin{cases} \Re(\hat{\mathbf{a}}_k^H\hat{\mathbf{z}}) = 1 & \hat{r}_k > 0 \\ \Re(\hat{\mathbf{a}}_k^H\hat{\mathbf{z}}) \leqslant 1 & \hat{r}_k = 0 \end{cases} \tag{16}$$

where

$$\hat{\mathbf{z}} = \mathbf{x} - \sum_{k=1}^{n}\hat{\mathbf{a}}_k\hat{r}_k \tag{17}$$

Now, assume that $\hat{r}_l = 0$. Note that in this case $\hat{\mathbf{z}}^{(l)} = \hat{\mathbf{z}}$. On the other hand, updating $(\hat{\mathbf{a}}_l, \hat{r}_l)$ by (14) does not change the pair, which implies that

$$\max_{\mathbf{a}\in\mathcal{A}} \Re(\mathbf{a}^H\hat{\mathbf{z}}) = \max_{\mathbf{a}\in\mathcal{A}} \Re(\mathbf{a}^H\hat{\mathbf{z}}^{(l)}) \leqslant \lambda \tag{18}$$

and together with (16) shows that (7) holds. □

3.2. Correcting CCD

We cannot generally establish global optimality of the CCD algorithm, except when the stopping point contains an inactive element. Now, we show that a different step can be added to CCD, which ensures that the stopping point contains an inactive element. Clearly, the additional step may not increase the cost in (13), to provide convergence. In practice, we apply this step after a fixed number L of CCD cycles.

Consider a step in the algorithm where the parameters are given by $\{\mathbf{a}_k, r_k\}$. Fixing $\{\mathbf{a}_k\}$, we update $\{r_k\}$ such that at the same time $\sum_k r_k$ does not increase, $\sum_k \mathbf{a}_k r_k$ does not change and one of the parameters r_k is set to zero. Then, the cost in (13) does not increase. Hence, the combination of this step with CCD converges to a stopping point of CCD, containing a zero parameter. To do so, consider real parameters u_k such that at least one of them is nonzero and

$$\sum_{k=1}^{n} u_k\mathbf{a}_k = 0 \quad \sum_{k=1}^{n} u_k \geqslant 0 \tag{19}$$

The choice $n = 2m+1$ ensures existence of u_k, which can be obtained for example by the Gaussian elimination technique. Then $\{r_l\}$ is updated to

$$\hat{r}_l = r_l - \alpha u_l \tag{20}$$

where

$$\alpha = \min_{l|u_l>0} \frac{u_l}{r_l} \tag{21}$$

If the minimum is achieved at index l_0 then

$$\alpha = \frac{u_{l_0}}{r_{l_0}} \rightarrow \hat{r}_{l_0} = u_{l_0} - \alpha r_{l_0} = 0 \tag{22}$$

This ensures that at least one element will be zero after the updating procedure. Thus, the entire algorithm in Algorithm 2 stops at a point, where by Theorem 1 (9) holds. Thus, the stopping point is a global optimal point of ANDN.

4. NUMERICAL RESULTS

Now, we examine our proposed Algorithm 2 in some experimental scenarios. We consider the spectral line estimation

Algorithm 2 The overall algorithm.

Require: A starting point, and the number of CCD cycles L.

repeat

Run L cycles of the CCD algorithm in Algorithm 1 to obtain $\{\mathbf{a}_k, r_k\}$.

Fix $\{\mathbf{a}_k\}$ and replace $\{r_k\}$ by $\{\hat{r}_k\}$ given in (20).

until Convergence

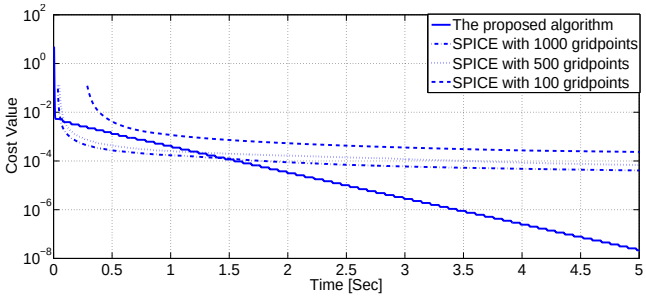

Fig. 1. Cost value for the proposed algorithm compared to SPICE at different iterations. The difference to the true optimal value is plotted.

problem introduced in Section 2. The m sampling instants are selected uniformly randomly in the time interval of $[0\ m]$, which provides the average sampling rate of 1 sample/sec. The correcting step in Algorithm 2 is applied after each $L = 10$ cycles of CCD. All techniques are implemented by MATLAB.

We compare our technique to a fast grid based implementation of LASSO, called Sparse Parameter estimation by Iterative Covariance Estimation (SPICE), which has interesting numerical properties [23]. Figure 1 shows the resulting cost during the iterative procedure versus the CPU runtime. The number of samples was set to $m = 10$. The true signal consisted of two frequencies at $0,\ 4\pi/m$. This choice guarantees noiseless perfect recovery. The other parameters were calculated such that the exact solution of ANDN could be predicted. In this manner, Figure 1 shows the difference to the exact optimal cost in a logarithmic scale. As seen, SPICE is initially faster than the proposed one, but then substantially reduces speed. This is because the nearby highly coherent atoms are difficult to be driven to zero. Note that due to discretization an error floor proportional to the number of grids exists, but cannot be achieved in the depicted time interval. It is also interesting to see that the proposed algorithm has a stepwise behavior. This shows that the first iteration in each cycle has a more prominent effect on the cost reduction.

In a different experiment we considered uniform samples, where the SemiDefinite Programming (SDP) technique of [15, 14] can be employed to exactly solve ANDN. Figure 2 shows the result for different values of m. Two frequencies are again considered at $0,\ 5\pi/m$ and the other parameters are

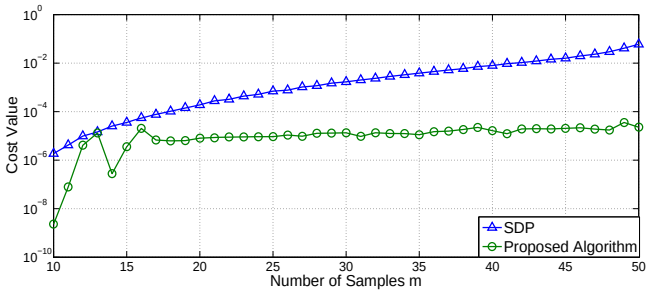

Fig. 2. Cost value for the proposed algorithm compared to SDP with different number of observations. The difference to the true optimal value is plotted.

properly selected. For the SDP approach, 1000 iterations of the ADMM implementation in [14] was considered and the total time was calculated. Then, the proposed algorithm was applied for the equal time and the resulting final cost values were compared. As seen, the SDP error diverges for larger values of m. This implies that more iterations are needed for larger m. The error of the current technique remains constant. However, the runtime grows linearly with m, which is not depicted, due to lack of space.

5. CONCLUSION

In this paper, we proposed a numerical algorithm to solve ANDN. We introduced a parametric form of ANDN and introduced a CCD algorithm to solve it. Then, we showed that introducing an additional step guarantees convergence to the global optimal point.

We compared our technique to SPICE and the SDP approach in [15, 14], in the context of spectral line estimation. The latter only admits uniform samples. The former is more general, but needs discretization. As shown here, the discretization not only introduces a quantization error, but also reduces the convergence speed and precision. More investigation on similar implementation methods reveals that this generally holds for the so-called semi-parametric formulation, where a large number of parameters are simultaneously optimized. In this case, the optimization over parameters with a highly similar effect needs a very high numerical precision and a considerable amount of computation. On the contrary, our parametric form does not need such considerations. The SDP approach is precise, but still has a higher computational complexity compared to our technique.

The simple CCD step considered here is not generally regarded among the fastest possible optimization techniques. Hence, it is suspected that more sophisticated correction steps may increase the speed of the current technique. For example, a derivative based approach such as the Newton method can be examined. However, this is postponed to a future study.

6. REFERENCES

[1] Moshe Mishali and Yonina C Eldar, "From theory to practice: Sub-nyquist sampling of sparse wideband analog signals," *IEEE J. Select. Topics Signal Processing*, vol. 4, pp. 375–391, Apr. 2010.

[2] DM Malioutov, "A sparse signal reconstruction perspective for source localization with sensor arrays," M.S. thesis, MIT, July 2003.

[3] Huajian Yao, Peter Gerstoft, Peter M Shearer, and Christoph Mecklenbräuker, "Compressive sensing of the tohoku-oki mw 9.0 earthquake: Frequency-dependent rupture modes," *Geophys. Res. Lett.*, vol. 38, Oct. 2011.

[4] Scott S. Chen, David L. Donoho, and Michael A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, Dec. 1998.

[5] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., Series B (Methodological)*, vol. 58, pp. 267–288, Jan. 1996.

[6] D.L. Donoho and Y. Tsaig, "Fast solution of 1-norm minimization problems when the solution may be sparse," *IEEE Trans. Intell. Transport. Syst.*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[7] Michael R Osborne, Brett Presnell, and Berwin A Turlach, "A new approach to variable selection in least squares problems," *IMA journal of numerical analysis*, vol. 20, no. 3, pp. 389–403, 2000.

[8] P. Stoica and P. Babu, "Sparse estimation of spectral lines: Grid selection problems and their solutions," *IEEE Trans. Signal Processing*, vol. 60, pp. 962–967, Feb. 2012.

[9] Yuejie Chi, Louis L Scharf, Ali Pezeshki, and A Robert Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Trans. Signal Processing*, 2011.

[10] A. Panahi and M. Viberg, "Maximum aposteriory based regularization parameter selection," in *IEEE Int. Conf. Acoust. Speech, Signal Processing*, 2011.

[11] Hao Zhu, Geert Leus, and Georgios B Giannakis, "Sparsity-cognizant total least-squares for perturbed compressive sampling," *Signal Processing, IEEE Transactions on*, vol. 59, no. 5, pp. 2002–2016, 2011.

[12] Zhao Tan, Peng Yang, and Arye Nehorai, "Joint sparse recovery method for compressed sensing with structured dictionary mismatch," *arXiv preprint arXiv:1309.0858*, 2013.

[13] Chaitanya Ekanadham, Daniel Tranchina, and Eero P Simoncelli, "Recovery of sparse translation-invariant signals with continuous basis pursuit," *IEEE Trans. Signal Processing*, 2011.

[14] Badri Narayan Bhaskar and Benjamin Recht, "Atomic norm denoising with applications to line spectral estimation," in *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*. IEEE, 2011, pp. 261–268.

[15] Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht, "Compressive sensing off the grid," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 778–785.

[16] Ashkan Panahi, "Parameter estimation using sparse modeling: Algorithms and performance analysis," *Licentiate thesis, Chalmers University of Technology, Department of Signals and Systems*, Sept. 2012.

[17] Kumar Vijay Mishra, Myung Cho, Anton Kruger, and Weiyu Xu, "Off-the-grid spectral compressed sensing with prior information," *arXiv preprint arXiv:1311.0950*, 2013.

[18] Emmanuel J Candès and Carlos Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[19] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Mag.*, vol. 13, pp. 67 –94, July 1996.

[20] Petre Stoica and Nehorai Arye, "Music, maximum likelihood, and cramer-rao bound," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 5, pp. 720–741, 1989.

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

[22] Mokhtar S Bazaraa, Hanif D Sherali, and C M Shetty, *Nonlinear Programming: Theory and Algorithms*, Wiley, 3 edition, 2006.

[23] Petre Stoica, Prabhu Babu, and Jian Li, "Spice: A sparse covariance-based estimation method for array processing," *Signal Processing, IEEE Transactions on*, vol. 59, no. 2, pp. 629–638, 2011.