

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Parameter Estimation and Filtering Using Sparse Modeling

ASHKAN PANAHI



Department of Signals and Systems
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2015

Parameter Estimation and Filtering Using Sparse Modeling

ASHKAN PANAHI

ISBN 978-91-7597-213-8

© ASHKAN PANAHI, 2015.

Doktorsavhandlingar vid Chalmers tekniska högskola

Serie nr 3894

ISSN 0346-718X

Signal Processing Group

Department of Signals and Systems

CHALMERS UNIVERSITY OF TECHNOLOGY

SE-412 96 Göteborg

Sweden

Telephone: +46 (0)31 – 772 1000

Typeset by the author using L^AT_EX.

Chalmers Reproservice
Göteborg, Sweden 2015

Abstract

Sparsity-based estimation techniques deal with the problem of retrieving a data vector from an undercomplete set of linear observations, when the data vector is known to have few nonzero elements with unknown positions. It is also known as the atomic decomposition problem, and has been carefully studied in the field of compressed sensing. Recent findings have led to a method called basis pursuit, also known as Least Absolute Shrinkage and Selection Operator (LASSO), as a numerically reliable sparsity-based approach. Although the atomic decomposition problem is generally NP-hard, it has been shown that basis pursuit may provide exact solutions under certain assumptions. This has led to an extensive study of signals with sparse representation in different domains, providing a new general insight into signal processing. This thesis further investigates the role of sparsity-based techniques, especially basis pursuit, for solving parameter estimation problems.

The relation between atomic decomposition and parameter estimation problems under a so-called separable model has also led to the application of basis pursuit to these problems. Although simulation results suggest a desirable trend in the behavior of parameter estimation by basis pursuit, a satisfactory analysis is still missing. The analysis of basis pursuit has been found difficult for several reasons, also related to its implementation. The role of the regularization parameter and discretization are common issues. Moreover, the analysis of estimates with a variable order, in this case, is not reducible to multiple fixed-order analysis. In addition to implementation and analysis, the Bayesian aspects of basis pursuit and combining prior information have not been thoroughly discussed in the context of parameter estimation.

In the research presented in this thesis, we provide methods to overcome the above difficulties in implementing basis pursuit for parameter estimation. In particular, the regularization parameter selection problem and the so-called off-grid effect is addressed. We develop numerically stable algorithms to avoid discretization and study homotopy-based solutions for complex-valued problems. We use our continuous estimation algorithm, as a framework to analyze the basis pursuit. Moreover, we introduce finite set based mathematical tools to perform the analysis. Finally, we study the Bayesian aspects of basis pursuit. In particular, we introduce and study a recursive Bayesian filter for tracking the sparsity pattern in a variable

ABSTRACT

parameter estimation setup.

Keywords: Sparsity based techniques, parameter estimation, compressed sensing, off-grid effect, continuous basis pursuit, sparsity based tracking

Acknowledgment

I would like to take the opportunity to thank people who directly, or indirectly had influence on the preparation of this thesis. I should first thank my supervisor Prof. Mats Viberg for giving me the opportunity to study as a PhD student with the signal processing group. I should also thank him for believing in me, his guidance and encouragement throughout my research. Many thanks to my co-supervisor Lennart Svensson for all interesting discussions and useful suggestions. Special thanks to the head of SP group, Prof. Thomas McKelvey for all joyful talks and interesting technical discussions. Also, I would like to thank my adviser, Prof. Babka Hassibi during my visit at Caltech. I appreciate introducing challenging, but very interesting, research topics, and the time he spent for all the detailed technical discussions. I should also thank Prof. Christoph Mecklenbrä uker for our collaboration and joyful talks. Thanks to Prof. Yonina Eldar for her comments at my Licentiate examination, which I later found very useful in my research, and finally many thanks to Prof. Peter Stoica for your hospitality during my very short visit at the university of Uppsala and all the careful later comments.

Many thanks to my other co-authors, Prof. Tomas Ericsson, Dr. Giuseppe Durisi, Dr. Mark C. Reed, Dr. Peter Gerstoft, Dr. Marie Strö m, M.Reza Khanzadi, Amin Movahed, Dr. Kasra Haghighi, Dr. Moslem Rashidi and Christos Thrampoulidis. I really enjoyed working with you and hope that we can continue our collaboration for future.

I would like to thank my colleagues at S2, especially the people at the signal processing group. Also, I would like to acknowledge other members of the group, Prof. Irene Gu and Dr. Thomas Rylander as well as students at S2, especially my friends Reza, Ayca, Marie, Yinan, Abu, Maryam, Tomas, Malin, Lars, Johan(s), Erik, Nikolaos, Xinling, Yixia and many others. I should also acknowledge my other friends who has left S2 and I enjoyed working with. Thanks to Sima, Mazyar, Kasra, Lotfollah, Hamidreza, Mohsen, Mohammad, Livia, Panagiota, Johnny, Ali, Mitra, Sahar, Roozbeh and others.

Last but not least, I would like to thank my partner and my dearest friend, Negar. I very appreciate your kindness and your encouragement during all the busy and tough days and thank you for all the beautiful moments we have had together.

List of Publications

This thesis is based on the following three appended papers:

Paper 1

Panahi A., Viberg M. and Hassibi B. A Numerical Approach to Gridless Compressed Sensing, to be submitted to IEEE Transactions on Signal Processing.

Paper 2

Panahi A. and Viberg M., Performance Analysis of Parameter Estimation Using LASSO, to be submitted to IEEE Transactions on Signal Processing.

Paper 3

Panahi A. and Viberg M., A Novel Sparsity-Based Approach to Recursive Estimation of Dynamic Parameter Sets, to be submitted to IEEE Transactions on Signal Processing.

Other Publications

Panahi, A., Viberg, M. and Hassibi, B. (2015) A numerical Implementation of Gridless Compressed Sensing. *2015 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015; Brisbane; Australia; 19 April 2015 through 24 April 2015*

Panahi, A., Ström M. and Viberg, M. (2015) Wideband Waveform Design for Robust Target Detection. *2015 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015; Brisbane; Australia; 19 April 2015 through 24 April 2015*

LIST OF PUBLICATIONS

- Thrampoulidis, C, Panahi, A. and Hassibi B. (2015) Precise Error Analysis for the LASSO. *2015 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015; Brisbane; Australia; 19 April 2015 through 24 April 2015*
- Movahed, A., Panahi, A. and Reed, M. (2014) Recovering signals with variable sparsity levels from the noisy 1-bit compressive measurements. *2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014; Florence; Italy; 4 May 2014 through 9 May 2014*
DOI: 10.1109/ICASSP.2014.6854847.
- Panahi, A., Ström, M. and Viberg, M. (2014) Basis pursuit over continuum applied to range-Doppler estimation problem. *IEEE 8th Sensor Array and Multichannel Signal Processing Workshop, SAM 2014; A Coruna; Spain; 22 June 2014 through 25 June 2014* DOI: 10.1109/SAM.2014.6882421.
- Panahi, A. and Viberg, M. (2014) Gridless compressive sensing. *2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014; Florence; Italy; 4 May 2014 through 9 May 2014* DOI: 10.1109/ICASSP.2014.6854228.
- Ström, M., Panahi, A., Viberg, M. and Falk, K. (2014) Wide-band waveform design for clutter suppression. *2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop, SAM 2014; A Coruna; Spain; 22 June 2014 through 25 June 2014* DOI: 10.1109/SAM.2014.6882400.
- Khanzadi, M., Kuylenstierna, D., Panahi, A., Eriksson, T. and Zirath, H. (2013) Calculation of the Performance of Communication Systems from Measured Oscillator Phase Noise. *IEEE Transactions on Circuits and Systems Part 1: Regular Papers* 61, nr. 5, s. 1553-1565. DOI: 10.1109/TCSI.2013.2285698.
- Mecklenbrauker, C., Gerstoft, P., Panahi, A. and Viberg, M. (2013) Sequential Bayesian Sparse Signal Reconstruction Using Array Data. *IEEE Transactions on Signal Processing* 61, nr. 24, s. 6344-6354. DOI: 10.1109/tsp.2013.2282919.
- Panahi, A. and Viberg, M. (2013) A novel method of DOA tracking by penalized least squares. *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, CAMSAP 2013* DOI: 10.1109/CAMSAP.2013.6714007.
2012

- Khanzadi, M., Panahi, A., Kuylenstierna, D. and Eriksson, T. (2012) A model-based analysis of phase jitter in RF oscillators. *Proceedings - 66th IEEE International Frequency Control Symposium, IFCS 2012, Baltimore, 21-24 May 2012* DOI: 10.1109/IFCS.2012.6243677.
- Movahed, A., Panahi, A. and Durisi, G. (2012) A robust RFPI-based 1-bit compressive sensing reconstruction algorithm. *IEEE Information Theory Workshop (ITW), Lausanne, 3-7 September 2012* DOI: 10.1109/ITW.2012.6404739.
- Panahi, A. and Viberg, M. (2012) A robust l1 penalized DOA estimator. *46th Asilomar Conference on Signals, Systems and Computers*. DOI: 10.1109/ACSSC.2012.6489394.
- Panahi, A. and Viberg, M. (2012) Fast Candidate Points Selection in the LASSO Path. *IEEE Signal Processing Letters* 19, nr. 2, s. 79-82. DOI: 10.1109/LSP.2011.2179534.
- Panahi, A. (2012) Parameter Estimation Using Sparse Modeling: Algorithms and Performance Analysis. Lic. Thesis. *Chalmers University of Technology*. 2012
- Panahi, A. and Viberg, M. (2011) Fast LASSO based DOA tracking. *4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011*
- Panahi, A. and Viberg, M. (2011) Maximum A Posteriori Based Regularization Parameter Selection. *2011 IEEE International Conference on Acoustics, Speech, and Signal Processing*
- Panahi, A. and Viberg, M. (2011) On the resolution of the LASSO-based DOA estimation method. *Proceedings - 2011 International ITG Workshop on Smart Antennas, WSA 2011* DOI: 10.1109/WSA.2011.5741938.
- Rashidi Avendi, M., Haghghi, K., Panahi, A. and Viberg, M. (2011) A NLLS based sub-Nyquist rate Spectrum Sensing for Wideband Cognitive Radio. *Fifth IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks 2011(DySPAN 2011)*
- Khanzadi, M., Haghghi, K., Panahi, A. and Eriksson, T. (2010) A Novel Cognitive Modulation Method Considering the Performance of Primary User. *Wireless Advanced (WiAD), 2010 6th Conference on* DOI: 10.1109/WIAD.2010.5544942.

Contents

Abstract	i
List of Publications	v
Contents	ix

I Introductory Chapters

1 Introduction	1
1.1 Thesis Outline	3
1.1.1 Introductory Part	3
2 Atomic Decomposition Problem	5
2.1 Mathematical Modeling	5
2.2 Atomic Decomposition in Practice	6
2.2.1 Sensor-Based Estimation Problems	6
2.2.2 Compressive Image Acquisition	7
2.2.3 Learning Gaussian Models	8
2.3 Spectral Representation	9
3 Solutions to the Atomic Decomposition Problem	11
3.1 General Approaches	11
3.1.1 Parametric Approaches	11
3.1.2 Spectral-Based Approaches	13
3.1.3 Subspace-Based Approaches	14
3.2 Model Order Selection Problem	15
3.3 Analysis	16
3.3.1 Analysis of Maximum Likelihood in a High SNR Case	17
3.3.2 Analysis of MUSIC in a Large Sample Size Case . . .	18

4	Sparsity-Based Atomic Decomposition	21
4.1	Basis Pursuit	21
4.1.1	Implementing Basis Pursuit	22
4.1.2	Regularization Parameter Selection	25
4.2	Analysis of Basis Pursuit for Large Dimensions	25
4.2.1	Null Space Property	26
4.2.2	Restricted Isometry Property	27
4.2.3	Error Analysis	28
4.3	The Off-Grid Problem	28
4.4	Other Approaches	30
5	Dynamic Atomic Decomposition	31
5.1	Recursive Bayesian Estimator	32
5.2	Filtering Theory for Atomic Decomposition	33
5.2.1	Extended Kalman Filter	34
5.3	Sparsity-Based Filtering	35
6	Thesis Contributions	37
6.1	Summary of Appended Papers	37
6.2	Suggestions for Future Work	38
	References	41

II Included Papers

Paper 1	A Numerical Approach to Gridless Compressed Sensing	59
1	Introduction	59
1.1	Problem Formulation	61
2	Examples of Atomic Decomposition	62
2.1	Spectral Line Estimation	62
2.2	Far-field Narrow-Band Direction-of-Arrival Estimation	62
2.3	Narrow-band Radar Delay-Doppler Estimation	63
3	Prior Art: Atomic Norm Denoising	63
3.1	Optimality Condition For ANDN	64
3.2	Implementing ANDN for Frequency Estimation	65
4	Contribution	66
4.1	Cyclic Coordinate Descent Algorithm	66
4.2	Correcting CCD	68
4.3	How to Select the Cycle	70
5	Numerical Results and Comparison to Related Works	70

5.1	Related Works: Basis Pursuit	70
5.2	Numerical Result	73
6	Conclusion	80
7	Appendix: Properties of the Atomic Norm	81
	References	81
Paper 2 Performance Analysis of Parameter Estimation Using LASSO		87
1	Introduction	87
2	Mathematical Modeling	90
2.1	The Principle of Sparsity	92
2.2	The Sensor Array Example	92
3	LASSO, Parametric LASSO and CLASS	93
3.1	Preliminaries on Asymptotic Analysis	95
3.2	CLASS Solution	97
3.3	Dual Convergence Properties	99
3.4	First Order Linearization	99
4	Statistical Results	101
4.1	Ideal Consistency	103
4.2	Statistical Properties of Perturbations	104
5	Numerical Results	106
5.1	Evaluation of Theoretical Performance	106
5.2	Comparison with Other Methods	107
6	Concluding Remarks	108
7	Appendix: LASSO Topology on ADP space	109
8	Appendix: Proof of Theorem 5	111
9	Appendix: Proof of Theorem 6	115
10	Appendix: Proof of Theorem 7	116
11	Appendix: Proof of Theorem 8	118
12	Appendix: Proof of Theorem 9	119
	References	119
Paper 3 A Novel Sparsity-Based Approach to Recursive Estimation of Dynamic Parameter Sets		127
1	Introduction	127
1.1	Literature Survey	129
1.2	Motivation	130
1.3	Mathematical Notation	131
2	Problem Formulation	131
2.1	Observation Model	131
2.2	Time Evolution Model	133
3	Recursive Bayesian Filtering	134

CONTENTS

3.1	Calculating the MAP Hyper-State Estimate	135
3.2	Update Step	136
3.3	Prediction Step Approximation	137
4	Numerical Results and Comparison to Related Works	138
4.1	Related Studies	139
4.2	Numerical Results	142
5	Concluding Remarks	144
6	Appendix: Calculus of Random Finite Sets	146
6.1	Functional Representation	146
6.2	Integration	147
7	Appendix: RFS Local Approximation	147
7.1	Poisson Approximation	148
7.2	Extended Laplace's Method	148
8	Appendix: Perturbative KL-based Projection	149
	References	150

Part I

Introductory Chapters

Chapter 1

Introduction

The last two decades witnessed the advent of so-called sparsity-based techniques, which concern a variety of different signal processing applications. They have been originally introduced and studied for the specific purpose of data acquisition, where they are often referred to as Compressed Sensing (CS). The sparsity-based techniques were soon discovered to be useful in many different applications with similar mathematical representations [1–7]. Here, we refer to this model as Atomic Decomposition (AD), which does not imply any specific application. The atomic decomposition model leads to NP-hard¹ computational problems. Accordingly, approximate techniques are since long considered in the literature. These techniques are developed and presented in different contexts and under different model representations. The AD formalism provides an occasion to present them in a unified framework.

Sparsity-based techniques appeared first in the context of image processing, where they were applied to the so-called Compressed Sensing (CS) problem [9,10]. The invention of ℓ_1 regularization and the convexifying technique had a great impact on the later developments in this field [11–13]. The ℓ_1 -regularization, known as Basis Pursuit (BP) or Least Absolute Shrinkage and Selection Operator (LASSO) rapidly received attention in the machine learning and data acquisition societies, where pioneering studies showed interesting characteristics of LASSO [10,14–18]. It was, for example, shown that BP can be solved in a polynomial time to provide ideal signal recovery

¹A NP-hard problem is informally defined as the one, being as difficult as the most complex problem in the family of Non-deterministic Polynomial (NP) problems. The NP family consists of the problems, which can be solved in a so-called non-deterministic (or oracle) computing machine in a polynomial time. However, the simulation of a non-deterministic machine in a deterministic one (such as ordinary computers) generally needs an exponentially growing amount of computation, which indicates a higher amount of complexity for the NP problems. Nevertheless, it is not still clear, whether NP problems can be polynomially solved in a deterministic machine or not. See [8] for more details.

for certain large matrices.

The sparsity-based techniques are also getting more popular in parameter estimation problems with an underlying atomic decomposition model, where the model is often referred to as separable [19, 20]. This approach was first introduced and studied in the pioneering studies by Fuchs [21], Stoica [22] and others [23, 24]. Although it is generally believed that BP has unique estimation properties, its theoretical analysis has found to be difficult. Particularly, its super-resolution properties as well as technical issues, such as the choice of regularization parameter and the effect of the grid is still under question. The computational aspects of sparsity based techniques should also be discussed. It is often observed that sparsity based methods need a higher computational demand than other parameter estimation techniques.

A great potential is observed in applying BP to problems with Bayesian prior information [25, 26]. In the case of large dimensions, where the Bayesian interpretation is replaced by the deterministic concept of typicality, this is currently being widely studied under the title of structured sparsity [27–31]. However, this potential has not been exploited in the parameter estimation case. A few papers have addressed the weighted BP approach, but the general principles of weighted BP design is not well-understood.

Accordingly, this thesis is devoted to investigating the particular application of BP to separable parametric estimation problems with an AD nature. The following issues are highlighted throughout this work:

Implementation Issues

The implementation of BP is well discussed, and usually involves discretization [32–37]. The current grid-based implementation of BP limits its potential to provide accurate parameter estimates. For example, the discrete nature of BP leads to the so-called off-grid problem, restricting its resolution [38–41]. Different studies suggest techniques to mitigate the off-grid effect [42–46]. In this thesis, we discuss a framework, under which the discretization step can be avoided and a continuous sparsity-based estimator is obtained. In this regard, the outcome of this thesis is a numerical method which guarantees global convergence. This method implements a continuous extension of LASSO, referred to in the literature as Atomic Norm DeNoising (ANDN) [44]. Throughout this study, we have also developed other implementation techniques to treat the order selection and noiseless estimation, for which the reader is also referred to [47].

Analysis of Parameter Estimation

We also provide a parametric analysis of LASSO, which is suitable for the application of interest herein. The analysis is difficult for multiple reasons. For example, the effect of the grid complicates the analysis of BP. Due to the unpredictable relation of the regularization parameter to the order, it is also impossible to analyze the estimates for a fixed order. To respond to the above, we consider the continuous framework, developed for implementation, and present the analysis of the estimates, obtained by BP or (ANDN) in a high SNR scenario. This also includes the miss detection properties.

Application to Dynamic Parameter Estimation

Finally, we address the dynamic parameter estimation problem [48–51]. In the problems of interest herein, a dynamic model for the parameters of interest leads to another NP-hard problem, called data association. This is mainly due to the variable order of the parameter set. We present methods to utilize the sparsity-based estimation framework to simplify calculations. In particular, we investigate re-weighting schemes for BP to incorporate the information from past to the current estimation problem in a recursive Bayesian framework. In this context, we have examined a number of different approaches, for which the reader is also referred to [52–54].

1.1 Thesis Outline

This thesis includes two main parts. In the first part, an introduction to the topics of interest in this research is presented. The second part consists of three papers, summarizing our main contributions. More details about the first part is presented in the sequel.

1.1.1 Introductory Part

In Chapter 2, the problem of Atomic Decomposition (AD) is presented and mathematically formulated. A number of popular examples of AD are introduced. AD can be derived using two different mathematical representations, namely parametric and spectral, the latter of which leads to sparsity based techniques. This is clarified in Section 2.3.

Chapter 3 discusses the previous atomic decomposition techniques, mainly developed in the field of parameter estimation, but widely used in a larger range of applications. We refer to some of the more popular approaches. A typical analysis of popular AD solutions is included in Chapter 3. Main

issues and related research, such as model order selection and statistical analysis of these techniques are also considered in this chapter.

In Chapter 4, different sparsity-based techniques are discussed. The focus is mainly on Basis Pursuit (BP). The main difficulties in applying BP to parametric estimation are introduced. Moreover, the previous analysis of these techniques is considered, which mainly revolves around large matrix-based atomic decomposition. The lack of relation between these studies and the parametric approaches, introduced in Chapter 3 is addressed in this chapter.

The extension of these methods to dynamic models is also considered and briefly discussed in Chapter 5, where also the possibility of sparse estimation under time evolution is presented. Finally, Chapter 6 introduces the papers, included in the second part of the thesis and clarifies the main contributions in each of them.

Chapter 2

Atomic Decomposition Problem

2.1 Mathematical Modeling

Consider a set of m -dimensional complex-valued bases $\mathcal{A} \subset \mathbb{C}^m$, referred to as the dictionary, and a sequence of complex-valued observation vectors $\{\mathbf{x}(t) \in \mathbb{C}^m\}$ for $t = 1, 2, \dots, T$. The expression

$$\mathbf{x}(t) = \sum_{k=1}^n s_k(t) \mathbf{a}_k + \mathbf{n}(t) \quad (2.1)$$

is called an atomic decomposition, where the vectors $\{\mathbf{a}_k \in \mathcal{A}\}$ are the bases incorporated in the decomposition, and the coefficients $\{s_k(t) \in \mathbb{C}\}$ are called amplitudes. The term $\mathbf{n}(t) \in \mathbb{C}^m$ denotes either the observation noise or the modeling error at time t . It is assumed to be a centered, temporally white and circularly symmetric Gaussian vector with covariance matrix $\sigma^2 \mathbf{I}$, where σ^2 is the noise variance. The number of incorporated bases n is known as the order of the decomposition.

Often in practice, the set \mathcal{A} is indexed by real numbers. Take a d -dimensional real-valued index set $\Theta \subseteq \mathbb{R}^d$ and consider an injective function $\mathbf{a}(\theta) : \Theta \rightarrow \mathbb{C}^m$. The function $\mathbf{a}(\theta)$ is called a representation for the dictionary \mathcal{A} if

$$\mathcal{A} = \{\mathbf{a}(\theta) \mid \theta \in \Theta\} \quad (2.2)$$

We mainly consider a case, where the index set Θ is closed, connected and bounded; and the function $\mathbf{a}(\theta)$ is smooth. In this case, \mathcal{A} is a d -dimensional manifold embedded in \mathbb{C}^m .

When the observation noise $\mathbf{n}(t)$ is zero, or equivalently $\sigma = 0$, the atomic decomposition in (2.1) is called noiseless. Given a sequence of observations $\{\mathbf{x}(t)\}$, a noiseless atomic decomposition with the smallest order is referred to as an ideal atomic decomposition. Clearly, an ideal decomposition of order n is the unique ideal decomposition if any set of $2n$ bases in

\mathcal{A} is linearly independent. The smallest number of linearly dependent basis vectors in \mathcal{A} is denoted by $\text{Spark}(\mathcal{A})$. Thus, any ideal decomposition of an order smaller than or equal to $(\text{Spark}(\mathcal{A}) - 1)/2$ is unique. Throughout this thesis, we always assume that this condition holds.

Given the sequence of observations, the atomic decomposition problem is to provide an AD with a suitable order and noise level. For the reasons, discussed in Chapter 3, this cannot be easily formulated in mathematical terms. We postpone a more detailed discussion to Section 3.2.

2.2 Atomic Decomposition in Practice

The AD problem concerns a large and increasing range of applications. Here, we consider few more popular examples, with different dictionary characteristics. In the first example, the dictionary is a low-dimensional manifold, while in the second one, the dictionary is finite. The third example shows a dictionary with a weak (high-dimensional) representation. In the latter chapters, we focus on cases similar to the first example, though, to some extent, the arguments are applicable to the other two examples.

2.2.1 Sensor-Based Estimation Problems

In this setup, the state θ of a finite number of unknown objects are to be estimated by sensing a scalar field at the position of a finite number of sensors. The field can be, for example, electromagnetic or sound¹. The state may also include, the object's position, velocity, etc; depending on the application of interest. Although this setup includes many different problems, depending on the choice of parameters, it can always be written in the atomic decomposition form, as long as the field superposition law holds [24, 55–59]. Denoting the local field observations at discrete time $t = 1, 2, \dots$ by $\mathbf{x}(t) = [x_1(t) \ x_2(t) \ \dots \ x_m(t)]^T$, where $x_k(t)$ represents the observation from the k^{th} sensor, the relation in (2.1) holds, where s_k characterizes the local field at the objects position and $\mathbf{a}_k = \mathbf{a}(\theta_k)$ represents the relation between s_k and the observation field, and is obtained by the field equation.

We take a more specific example, where θ includes only the direction of an object with respect to the origin of a fixed coordinate system. For simplicity, only a planar case is considered. We further assume that the sensors are located in the vicinity of the origin, constituting a sensor array. In contrast, the sources are relatively far. The scalar field is electromagnetic.

¹The electromagnetic field is vector-valued. However, the sensing apparatus of interest herein usually observe a scalar projection of the vector-field, which can be interpreted as an individual scalar field with similar dynamics to the electromagnetic wave.

It is originated from narrow-band sources, such that the field fluctuation at any point is represented by a narrow-band signal centered around the frequency f_0 , corresponding to the wavelength $d_0 = c/(2\pi f_0)$, where c is the speed of light. Taking $\{x_k(t)\}$ and $\{s_l(t)\}$ as the baseband complex envelope of their corresponding fields, we obtain that

$$\mathbf{a}(\theta) = \begin{bmatrix} e^{j\frac{2\pi\rho_1}{d_0}\cos(\theta-\theta_1)} \\ e^{j\frac{2\pi\rho_2}{d_0}\cos(\theta-\theta_2)} \\ \vdots \\ e^{j\frac{2\pi\rho_m}{d_0}\cos(\theta-\theta_m)} \end{bmatrix} \quad -\pi \leq \theta < \pi, \quad (2.3)$$

where (ρ_k, θ_k) is the polar coordinate of the k^{th} sensor [59]. In this case, the dictionary is represented by $\mathbf{a}(\theta)$. Hence, it is a one-dimensional manifold, called the array manifold.

In a case, where $\theta_k = 0$ and $\rho_k = (k-1)d_0/2$, the array is called half-wavelength Uniform Linear Array (ULA). Then, defining the electrical angle $\phi = \pi \cos(\theta)$, the basis representation in (2.3) is simplified to

$$\mathbf{a}(\phi) = \begin{bmatrix} 1 \\ e^{j\phi} \\ e^{j2\phi} \\ \vdots \\ e^{j(m-1)\phi} \end{bmatrix} \quad -\pi \leq \phi < \pi, \quad (2.4)$$

The dictionary in (2.4) is also known as the Fourier manifold, which is related to the problem of estimating spectral lines (finite number of frequency components) of a signal by observing m uniform samples of it [5, 60, 61].

2.2.2 Compressive Image Acquisition

In this setup, the goal is to compress and store a high-resolution image. It is well known that images have sparse representations in certain domains. This means that denoting by \mathbf{y} the vectorized 2D image intensity values, the following relation holds

$$\mathbf{y} = \mathbf{\Psi}\mathbf{s}, \quad (2.5)$$

where the vector \mathbf{s} is assumed to contain few non-zero elements [7, 62–64]. The number of non-zero elements in \mathbf{s} is denoted by $\|\mathbf{s}\|_0$. Suppose that \mathbf{s} contains exactly n non-zero elements, denoted by s_1, s_2, \dots, s_n , corresponding to the columns $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_n$ of $\mathbf{\Psi}$, respectively. Note that the indexes

of s and $\boldsymbol{\psi}$, do not represent their place in the vector \mathbf{s} and the matrix $\boldsymbol{\Psi}$, respectively. Then, (2.5) can be compactly represented by

$$\mathbf{y} = \sum_k s_k \boldsymbol{\psi}_k. \quad (2.6)$$

It is generally difficult to obtain a generic transform $\boldsymbol{\Psi}$ based on the physical process of imaging. Thus, different heuristic transforms are considered. The FFT, wavelet and curvelet transforms are popular examples, for the details of which the reader is referred to [65, 66]. It is also possible to append two domains $\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2$ to obtain an overcomplete domain $[\boldsymbol{\Psi}_1 \ \boldsymbol{\Psi}_2]$ [12]. To reduce the complexity of image processing, it is further suggested to apply a linear compression $\boldsymbol{\Phi}$ to the data vector \mathbf{y} to obtain $\mathbf{x} = \boldsymbol{\Phi}\mathbf{y}$, with a substantially smaller dimension than \mathbf{y} . This is generally known as compressed sensing [7, 10, 67]. In this case, the model in (2.6) yields to

$$\mathbf{x} = \sum_k s_k \mathbf{a}_k, \quad (2.7)$$

where $\mathbf{a}_k = \boldsymbol{\Phi}\boldsymbol{\psi}_k$ is the corresponding column in $\mathbf{A} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ to s_k . In practice, the observation noise should also be included in (2.7), leading to the AD model with the dictionary \mathcal{A} , comprising of the columns of \mathbf{A} .

2.2.3 Learning Gaussian Models

In this case, the relation between a number of input random variables X_1, X_2, \dots, X_R and a number of output ones Y_1, \dots, Y_L is to be discovered. For simplicity, the variables are assumed to be centered Gaussian. Then, the relation is simply expressed by the cross-correlation matrix $\mathbf{M} = (M_{r,l})$, where $M_{r,l} = \mathcal{E}(X_r Y_l)$. Using the SVD, we obtain that

$$\mathbf{M} = \sum_{k=1}^n s_k \mathbf{u}_k \mathbf{v}_k^H, \quad (2.8)$$

where \mathbf{u}_k and \mathbf{v}_k are the left and right singular vectors, respectively, corresponding to the positive singular value s_k of \mathbf{M} . The parameter n denotes the rank of \mathbf{M} . Note that although the bases \mathbf{u}_k and \mathbf{v}_k should satisfy a set of orthogonality conditions, this can be neglected as long as only the rank of \mathbf{M} is considered. Then, the model in (2.8) is an AD with positive amplitudes s_k , where the dictionary is the set of all rank-1 matrices, given by

$$\mathcal{A} = \{\mathbf{u}\mathbf{v}^H \mid \mathbf{u} \in \mathbb{C}^R, \mathbf{v} \in \mathbb{C}^L, \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1\} \quad (2.9)$$

Note that $m = RL$ and $d = m + n - 2$. This problem is useful, for example in social network learning. It can be applied after the compressed sensing

process (obtaining few linear combinations), where it is sometimes referred to as matrix completion [68, 69] or the Netflix prize problem [70, 71].

2.3 Spectral Representation

The AD model in (2.1) can be generally represented in a different way. According to (2.1), define the function

$$\tilde{s}(\mathbf{a}, t) = \begin{cases} s_k(t) & \mathbf{a} = \mathbf{a}_k \\ 0 & \text{Otherwise} \end{cases}, \quad (2.10)$$

called the spectrum. Then, the expression in (2.1) can be equivalently written as²

$$\mathbf{x}(t) = \mathbf{n}(t) + \sum_{\mathbf{a} \in \mathcal{A}} \tilde{s}(\mathbf{a}, t) \mathbf{a} \quad (2.11)$$

Note that while in (2.1) the amplitudes $\{s_1, \dots, s_n\}$, together with the set of bases $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$ provide the representation, the expression in (2.11) only relies on $\tilde{s}(\mathbf{a}, t)$. The former can still be obtained from the latter by taking the set of bases corresponding to the nonzero values, also known as the support, of the spectrum.

The methods utilizing the formalism in (2.11) are known as spectral techniques. Mathematically speaking, the expression in (2.11) is only interesting when the spectrum is sparse, i.e. it has a finite support. However, many spectral techniques deal with non-sparse, and often continuous spectra. Nevertheless, those techniques should include a sparsifying step, sometimes referred to as focusing. If the underlying dictionary \mathcal{A} is equipped by a topology, the focusing step may simply consist of identifying the set of local maxima in the spectrum \mathcal{A} as the support.

Another issue with spectral techniques is that the spectrum needs to be stored. One solution is to consider a large finite subset $\tilde{\mathcal{A}} = \{\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \dots, \tilde{\mathbf{a}}_N\}$ of \mathcal{A} , known as a grid, and only store the on-grid spectrum, $\tilde{s}_k(t) = \tilde{s}(\tilde{\mathbf{a}}_k, t)$. In a case, where the dictionary is represented by an index set Θ , this can be performed by discretizing Θ , to obtain $\tilde{\Theta} = \{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_N\}$.

²For the rigorous definition of summation over infinite sets, see [72]. In short, summation of positive values is defined as the supremum (maximum) of all the summations over finite subsets of the original set. The summation of real numbers is performed by dividing the summation over the positive and the negative part. The summation of complex values is performed by decomposing the values to the real and imaginary part and so on.

Chapter 3

Solutions to the Atomic Decomposition Problem

3.1 General Approaches

There are several approaches to solve the ADP problem. Some of them use special structure of certain AD problems. Hence, they may not be generally applied. Examples of the latter can be found, for example, in [73–75]. Here, we focus on techniques that are applicable to any AD problem. However, the quality of their result clearly depends on the structure of the dictionary they are applied to.

3.1.1 Parametric Approaches

Methods that directly provide estimates for the parameters in (2.1) are called parametric. Usually, the dictionary is represented by a label parameter θ . Then, the parametric approaches provide estimates for $\{\theta_k\}$ and $\{s_k(t)\}$. In this case, the atomic decomposition problem can be studied from a statistical perspective. If the order $n < \text{Spark}(\mathcal{A})$ is known, the ADP is equivalent to estimating a vector of parameters $\boldsymbol{\theta}^{(n)} = [\theta_1 \ \theta_2 \ \dots \ \theta_n]^T$ as well as $\mathbf{s}^{(n)}(t) = [s_1(t) \ s_2 \ \dots, s_n(t)]^T$. Due to the statistical assumptions on the noise, we obtain the following likelihood function for the parameters $\boldsymbol{\theta}^{(n)}, \{\mathbf{s}^{(n)}(t)\}$:

$$L(\boldsymbol{\theta}^{(n)}, \{\mathbf{s}^{(n)}(t)\}; \{\mathbf{x}(t)\}) = p(\{\mathbf{x}(t)\} | \boldsymbol{\theta}^{(n)}, \{\mathbf{s}^{(n)}(t)\}) = \frac{1}{(\pi\sigma^2)^{mT}} e^{-\frac{\sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) \right\|_2^2}{\sigma^2}} \quad (3.1)$$

Thus, the Maximum Likelihood estimates are given by the global minimum of the following optimization problem

$$(\hat{\boldsymbol{\theta}}^{(n)}, \{\hat{\mathbf{s}}^{(n)}(t)\}) = \arg \min_{\boldsymbol{\theta}^{(n)}, \{\mathbf{s}^{(n)}(t)\}} \sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) \right\|_2^2 \quad (3.2)$$

The ML estimates in (3.2) are optimal in a statistical sense, but it is difficult to obtain them by solving (3.2). The optimization is often highly nonlinear and contains a large number of local minima. Nevertheless, many optimization techniques are considered to solve (3.2) locally [76–78].

Note that the optimization in (3.2) can be solved for $\{\mathbf{s}(t)\}$ to obtain

$$\hat{\mathbf{s}}^{(n)}(t) = \mathbf{A}^\dagger(\hat{\boldsymbol{\theta}}^{(n)}) \mathbf{x}(t), \quad (3.3)$$

where $\mathbf{A}^\dagger(\boldsymbol{\theta}^{(n)})$ denotes the Moore-Penrose pseudoinverse of the matrix $\mathbf{A}(\boldsymbol{\theta}^{(n)}) = [\mathbf{a}(\theta_1) \ \mathbf{a}(\theta_2) \ \dots \ \mathbf{a}(\theta_n)]$, and we used the fact that \mathbf{A} has a singleton null-space $\{\mathbf{0}\}$ due to $n < \text{Spark}(\mathcal{A})$. Substituting (3.3) into (3.2) and simplifying the result leads to

$$\hat{\boldsymbol{\theta}}^{(n)} = \arg \min_{\boldsymbol{\theta}^{(n)}} \text{Tr} \left(\mathbf{P}_{\mathbf{A}(\boldsymbol{\theta}^{(n)})}^\perp \hat{\mathbf{R}} \right) \quad (3.4)$$

where $\hat{\mathbf{R}} = \sum_{t=1}^T \mathbf{x}(t) \mathbf{x}^H(t) / T$ is the data sample covariance matrix and $\mathbf{P}_{\mathbf{A}(\boldsymbol{\theta})}^\perp = \mathbf{I} - \mathbf{A}(\boldsymbol{\theta}) \mathbf{A}^\dagger(\boldsymbol{\theta})$ is the projection matrix into the orthogonal complement of the range space of $\mathbf{A}(\boldsymbol{\theta})$.

Standard optimization techniques such as cyclic coordinate descent, gradient descent or Newton's method have been applied to both (3.4) and (3.2). In every case, achieving the global optimum has been observed to depend highly on the choice of the initial point [77, 79]. However, a specific application of cyclic coordinate descent to (3.2), called RELAX, has gained attention for its simplicity and good performance [80]. As a cyclic coordinate descent realization, RELAX iteratively performs cycles, consisting of n iterations, at the k^{th} of which, only parameters $\theta_k, \{s_k(t)\}$ are updated by minimizing (3.2). This yields to the following updating rule

$$\begin{aligned} \theta_k &\leftarrow \hat{\theta}_k = \arg \max_{\theta} \frac{\mathbf{a}^H(\theta) \hat{\mathbf{R}}_k \mathbf{a}(\theta)}{\|\mathbf{a}(\theta)\|_2^2} \\ s_k(t) &\leftarrow \frac{\mathbf{a}^H(\hat{\theta}_k) \mathbf{x}(t)}{\|\mathbf{a}(\hat{\theta}_k)\|_2^2} \end{aligned} \quad (3.5)$$

where defining $\mathbf{z}_k(t) = \mathbf{x}(t) - \sum_{l \neq k} \mathbf{a}(\theta_l) s_l(t)$, we denote

$$\hat{\mathbf{R}}_k = \frac{\sum_{t=1}^T \mathbf{z}_k(t) \mathbf{z}_k^H(t)}{T} \quad (3.6)$$

The RELAX method may also be interpreted as a Space-Altering Generalized Expectation (SAGE) maximization [81], where at the k^{th} iteration of each cycle the parameter space $(\boldsymbol{\theta}, \{\mathbf{s}(t)\})$ is augmented by

$$\{\mathbf{y}(t)\} = \mathbf{a}(\theta_k)s_k(t) + \mathbf{n}(t) \quad (3.7)$$

In the same manner as SAGE, one can utilize an Expectation Maximization (EM) algorithm to solve (3.2) through augmenting the parameter set by the set $\{\mathbf{y}_k(t) = \mathbf{a}(\theta_k)s_k(t) + \mathbf{n}_k(t)\}$, where $\mathbf{n}_k(t)$ is a noise term with variance σ_k^2 , such that $\sigma^2 = \sum_k \sigma_k^2$ [82, 83].

More generally when the order n is unknown, one of the solutions from the estimates $(\hat{\boldsymbol{\theta}}^{(n)}, \{\hat{\mathbf{s}}^{(n)}(t)\})$ for $n = 1, 2, \dots, \text{Spark}(\mathcal{A}) - 1$ is selected, by for example an Information Criterion (IC) or a statistical test [84, 85]. These techniques are discussed in detail, in Section 3.2.

3.1.2 Spectral-Based Approaches

The spectral formulation of atomic decomposition in (2.11) can be exploited to obtain the desired AD. Note that denoting the spectrum $\tilde{\mathbf{s}}(\mathbf{a}, t)$ by \tilde{s}_t , we can write the relation in (2.11), in an abstract form, as

$$\mathbf{x}(t) = \mathcal{A}\tilde{s}_t + \mathbf{n}(t) \quad (3.8)$$

where \mathcal{A} denotes the linear operator transforming the spectrum into the observed vector. Notice that the transformation by \mathcal{A} is well-defined if the spectrum \tilde{s}_t is sparse, and generally does not have an interesting analytical extension on the entire space of spectra (including non-sparse ones). Hence, \mathcal{A} does not generally possess interesting properties over the space of spectra. For example, it does not have a pseudo-inverse. Nevertheless, the possibility of inverting the relation in (3.8) by multiplying by a dual linear operator \mathcal{W} has been considered. In the field of sensor array processing, where the spectrum \tilde{s}_t has a spatial interpretation, this is generally known as beamforming [19, 86–88]. The operator \mathcal{W} is known as a beamformer. Mathematically speaking, a beamformer is represented by a collection of vectors $\{\mathbf{w}(\mathbf{a}) \in \mathbb{C}^m\}$. It acts on an observation vector \mathbf{x} to produce a spectrum $\tilde{\mathbf{s}}(\mathbf{a}) = \mathbf{w}^H(\mathbf{a})\mathbf{x}$. Now, it is intended to devise a beamformer \mathcal{W} , such that its application to (3.8) leads to

$$\mathcal{W}\mathbf{x}(t) = \mathcal{W}\mathcal{A}\tilde{s}_t + \mathcal{W}\mathbf{n}(t) \approx \tilde{s}_t \quad (3.9)$$

Apart from the noise effect, the precision of the approximation above is generally limited. For example, the result of beamforming is not sparse, and often leads to a blurred spectrum. This is sometimes referred to as the

spectral leakage effect [89]. The lack of rigorous statistical foundation, has also motivated for different heuristic design frameworks, discussed below.

From one perspective, the beamforming design is closely related to the filter design problem. In this case, the element $\mathbf{w}(\mathbf{a})$ is interpreted as a linear filter, removing the effect of every basis \mathbf{a}' in \mathcal{A} from \mathbf{x} , except $\mathbf{a}' = \mathbf{a}$. The matched filtering criterion suggests to consider a filter $\mathbf{w}_{\text{mf}}(\mathbf{a})$, maximizing the output Signal to Noise Ratio (SNR)

$$\begin{aligned} \mathbf{w}_{\text{mf}}(\mathbf{a}) &= \arg \min \sigma^2 \|\mathbf{w}\|_2^2 \\ &\text{subject to } \mathbf{w}^H \mathbf{a} = 1 \\ &= \frac{\mathbf{a}}{\|\mathbf{a}\|_2^2} \end{aligned} \quad (3.10)$$

where the last equality follows from the Cauchy-Schwarz inequality. This is also known as the conventional beamforming technique. Since the matched filter does not consider the filtering aspects, it is expected that it provides poor results in terms of resolution. In fact, the result of the matched filter is inconsistent, when local-maximum-based focusing is considered. However, it turns out that the uncertainty principle prevents the improvement of the matched filter by a generic design. This is well-known in the linear filter design literature as the windowing effect, and motivated to incorporate the observed data in the beamformer design. This is generally known as adaptive beamforming [87,90,91]. Perhaps, the most popular adaptive beamformer is the Minimum Variance Distortionless Response (MVDR), also known as the Capon beamformer [92,93]. The idea in MVDR is to learn the minimum-variance projection $\mathbf{w}_{\text{MVDR}}^H(\mathbf{a})\mathbf{x}(t)$, maintaining a constant correlation with \mathbf{a} . Since variance is not observable, the sample variance is instead used.

$$\begin{aligned} \mathbf{w}_{\text{MVDR}}(\mathbf{a}) &= \arg \min \sum_{t=1}^T |\mathbf{w}^H \mathbf{x}(t)|^2 \\ &\text{subject to } \mathbf{w}^H \mathbf{a} = 1 \\ &= \frac{\hat{\mathbf{R}}^{-1} \mathbf{a}}{\mathbf{a}^H \hat{\mathbf{R}}^{-1} \mathbf{a}}, \end{aligned} \quad (3.11)$$

where $\hat{\mathbf{R}} = \sum_{t=1}^T \mathbf{x}(t)\mathbf{x}^H(t)/T$. The Capon beamformer is consistent in a high SNR or when T is large, but requires a full-rank sample covariance $\hat{\mathbf{R}}$. Thus, it is not applicable to a case with few data snapshots.

3.1.3 Subspace-Based Approaches

The subspace techniques are motivated by the observation that the basis estimation process in the AD problem is equivalent to finding the linear subspace \mathcal{R} , spanned by these bases. Once this subspace is found, the

condition $n < \text{Sparke}(\mathcal{A}) - 1$ guarantees that no other base $\mathbf{a} \in \mathcal{A}$ resides in \mathcal{R} , since otherwise, \mathcal{A} will include $n + 1 < \text{Spark}(\mathcal{A})$ linearly dependent bases.

The relation between the AD and the subspace estimation problems is clearly seen in (3.4), where the projection matrix into \mathcal{R} is considered. Now, we may rewrite (3.4) as

$$\begin{aligned} & \max_{\mathcal{R}} \text{Tr} \left(\mathbf{P}_{\mathcal{R}} \hat{\mathbf{R}} \right) \\ & \text{subject to } \mathcal{R} \in \{(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \mid \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \in \mathcal{A}\}, \end{aligned} \quad (3.12)$$

where $\mathbf{P}_{\mathcal{R}}$ is the projection matrix into \mathcal{R} , and $(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ denotes the linear span of the bases $\mathbf{a}_1, \dots, \mathbf{a}_n$. The subspace may be simply estimated by relaxing (3.12) to obtain

$$\begin{aligned} & \max_{\mathcal{R}} \text{Tr} \left(\mathbf{P}_{\mathcal{R}} \hat{\mathbf{R}} \right) \\ & \text{subject to } \mathcal{R} \in \{(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \mid \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \in \mathbb{C}^m\} \end{aligned} \quad (3.13)$$

The solution to (3.13) is found by obtaining the Singular Value Decomposition of $\hat{\mathbf{R}}$ as

$$\hat{\mathbf{R}} = \mathbf{U}^H \mathbf{\Lambda} \mathbf{U} \quad (3.14)$$

where $\mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_m]$ is a unitary matrix and $\mathbf{\Lambda}$ is the diagonal elements of the singular values $\lambda_1, \lambda_2, \dots, \lambda_m$, written in a descending order. Then, the solution to (3.13) is given by $\mathcal{R} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)$, the subspace spanned by the singular vectors, corresponding to the n largest singular values. This solution is known as the signal subspace, while its orthogonal complement is often referred to as the noise subspace. Finally, the closest bases to the subspace \mathcal{R} is selected. For this, the Multiple Signal Classification technique suggests to calculate the spectrum as

$$P(\mathbf{a}) = \frac{1}{\|\mathbf{a}\|_2^2 - \sum_{k=1}^n |\mathbf{a}^H \mathbf{u}_k|^2} = \frac{1}{\sum_{k=n+1}^m |\mathbf{a}^H \mathbf{u}_k|^2} \quad (3.15)$$

and take the largest local maxima as the estimates [78]. The MUSIC technique is consistent and offers high-resolution at high SNR or large T . However, it is sensitive to the noise model and the precision of the sample covariance matrix.

3.2 Model Order Selection Problem

The techniques discussed in section 3.1 are based on the assumption that the order n is known. In a case, where the order is unknown, those techniques

can still be used over a range of orders, but eventually one of their solutions should be selected. This is called Model Order Selection (MOS). The model order selection problem can be put in a statistical framework. However, this needs careful considerations. To elaborate on this, suppose that the ML principle is to be applied. It is simple to see that the result is obtained by extending the minimization in (3.2) over the space of orders $n \in \mathbb{N}$. For a fixed n , denote the minimum in (3.2) by V_n . Then, the ML principle selects the minimum value of V_n . On the other hand, it is simple to see that V_n is monotonically decreasing. Hence, the ML solution is the largest possible order.

Different approaches are proposed in the literature to tackle the tendency to over estimate the model order. For example, the Minimum Description Length proposes a different framework, inspired by the research on data compression [94]. Another simple approach is to use statistical inference techniques and obtain tests to decide on the model order. The Generalized likelihood ratio test (GLRT) is a popular example [85, 95]. The popular statistical techniques focus on the following Bayesian framework [96, 97]:

$$\hat{n} = \min_n V_n + kn \tag{3.16}$$

where the constant k varies among different techniques, according to their underlying problem formulation. For example, the Akaike Information Criterion (AIC) suggests to apply $k = \sigma^2(3T + 1)$ [97, 98]. Other information criteria such as the Bayesian Information Criterion (BIC) [99] are also introduced. Although the AIC criterion requires a large number of observations, it is commonly used in practice. However, the parameter k needs to be tuned.

3.3 Analysis

In this section, we review a statistical analysis for the ADP problem from a parameter estimation point of view. In the problems of interest herein, the dictionary is labeled by a parameter θ and the error in terms of θ is considered. In general, the analysis is complicated. This is not only because of the nonlinear nature of estimation, but also due to the fact that it is difficult to quantify the estimation precision in a variable order scenario. Hence, the analysis is usually restricted to nearly ideal scenarios, where it is remarkably simplified by Taylor expansion. There are three main near ideal scenarios: asymptotically low noise σ^2 , large sample size T and large dictionary dimensions. The latter concerns a case, where a well-related family of ADP problems of different size are considered. For example, the

sensor array example may be analyzed for a large number sensors [100]. Another example of this case is considered in the next Chapter, where the conventional analysis of basis pursuit is reviewed. Here, we focus on the two first cases. For the low-noise case, we consider a single-snapshot AD problem, where only the ML approach works. For the case with a large sample size, we consider the analysis of MUSIC.

3.3.1 Analysis of Maximum Likelihood in a High SNR Case

We consider a case, where a single-snapshot data $\mathbf{x} = \mathbf{x}(1)$ is analyzed by the ML rule. We assume that the dictionary is indexed by θ and the true AD is given by $\theta_1, \theta_2, \dots, \theta_n$ and s_1, s_2, \dots, s_n , where $n < (\text{Spark}(\mathcal{A}) - 1)/2$, guaranteeing the uniqueness of the ideal decomposition. We denote the ML estimates by $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n$ and $\hat{s}_1, \hat{s}_2, \dots, \hat{s}_n$, where we assume that the order is known.

The first step of the analysis is to show that the low-noise assumption indeed leads to a near-ideal case. That is to show that for every $k = 1, 2, \dots, n$,

$$|\Delta\theta_k| \rightarrow_p 0, \quad |\Delta s_k| \rightarrow_p 0 \quad (3.17)$$

as $\sigma \rightarrow 0$, where $\Delta\theta_k = \hat{\theta}_k - \theta_k$, $\Delta s_k = \hat{s}_k - s_k$ and \rightarrow_p denotes convergence in probability [101]. Note that since θ_k is deterministic, the convergence can be expressed in the distribution sense as well. Obtaining consistency is not straightforward and it might not be generally true, even if uniqueness is guaranteed in a noiseless case. However, the assumptions in Section 2.1 guarantee convergence in the case of interest herein.¹ Now, we assume that consistency holds. The second step is to consider a sufficiently small noise variance such that in the Taylor expansion of (3.2), the terms of an order higher than 2 can be neglected. Using (2.1) and after straightforward

¹Although the exact proof is omitted in favor of simplicity, a sketch is given in the sequel. Denote the relation between the vector $\boldsymbol{\theta}$ and the linear subspace spanned by its corresponding bases by \mathcal{L} . The range of this correspondence is a closed subspace of the Grassman manifold, known as the Union of Subspaces (UoS). It is not difficult to show that the ML rule induces a neighborhood relation on the UoS, under which \mathcal{L} is continuous. Note that every continuous bijection on a compact set is also bi-continuous, i.e. it is inversely continuous. Thus, \mathcal{L} is inversely continuous. The estimates converge to their true values as the estimated subspace converges to the true subspace under the ML-induced topology. The compactness of the label set is crucial in this proof. For example, the case in (2.4) does not satisfy the compactness of the index set, thus violating the proof assumptions. As a result, a jump from π to $-\pi$ may occur in the estimation problem. The solution is either to restrict the analysis to the true parameters with a local isomorphism, or consider a modified metric, respecting the topology on the label set Θ , induced by the process of indexing.

manipulations, this leads to the following approximate ML optimization

$$(\Delta\boldsymbol{\theta}_{\text{ML}}, \Delta\mathbf{s}_{\text{ML}}) = \arg \min_{\Delta\boldsymbol{\theta}, \Delta\mathbf{s}} \left\| \mathbf{n} - \sum_{k=1}^n \mathbf{a}(\theta_k) \Delta s_k - \sum_{k=1}^n \mathbf{d}(\theta_k) s_k \Delta \theta_k \right\|_2^2, \quad (3.18)$$

where $\mathbf{d}(\theta) = \mathbf{d}\mathbf{a}(\theta)/\mathbf{d}\theta$. Define the linear operator Ω as

$$\Omega(\Delta\boldsymbol{\theta}, \Delta\mathbf{s}) = \sum_{k=1}^n \mathbf{a}(\theta_k) \Delta s_k + \sum_{k=1}^n \mathbf{d}(\theta_k) s_k \Delta \theta_k \quad (3.19)$$

Then, the optimization in (3.18) is an ordinary LS problem and can be solved to obtain $(\Delta\boldsymbol{\theta}_{\text{ML}}, \Delta\mathbf{s}_{\text{ML}}) = \mathbf{P}_\Omega \mathbf{n}$, where \mathbf{P}_Ω is the orthogonal projection operator into the range space of Ω . Explicit terms for the error can be found in [102].

3.3.2 Analysis of MUSIC in a Large Sample Size Case

Now, we consider the estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n$, obtained by MUSIC. The true parameters are $\theta_1, \dots, \theta_n$ and the true amplitudes are sampled from white, centered and uncorrelated sources. The order $n < \text{Spark}(\mathcal{A}) - 1$ is known. Again, we need to establish the two analysis steps. For the first step, we note that by the strong law of large numbers

$$\lim_{T \rightarrow \infty} \hat{\mathbf{R}} = \mathbf{R} = \mathcal{E}(\mathbf{x}(t)\mathbf{x}^H(t)) = \mathbf{A}(\boldsymbol{\theta})\boldsymbol{\Sigma}\mathbf{A}^H(\boldsymbol{\theta}) + \sigma^2\mathbf{I} = \mathbf{R}_s + \sigma^2\mathbf{I} \quad (3.20)$$

where $\boldsymbol{\Sigma}$ is the amplitude correlation matrix. Note that the SVD of \mathbf{R} is obtained by only adding the term σ^2 to the singular values of \mathbf{R}_s , and letting the subspaces remain unchanged. Now, it is clear that if the MUSIC method is applied to \mathbf{R} , the subspace obtained by the n largest singular values of \mathbf{R} , coincides with that of \mathbf{R}_s , corresponding to the range space of $\mathbf{A}(\boldsymbol{\theta})$. Thus, the MUSIC method, in this case, calculates the parameters exactly. It is also simple to see that convergence for \mathbf{R} leads to convergence of the subspace, guaranteeing a vanishing error vector² denoted by $\Delta\boldsymbol{\theta}$.

For the second step, we consider a small error in $\hat{\mathbf{R}}$, denoted by $\Delta\mathbf{R} = \hat{\mathbf{R}} - \mathbf{R}$. Since the error converges to zero, we can use Taylor expansion similar to Section 3.3.1. Note that the MUSIC estimates are the local maxima of the spectrum $p(\theta, \hat{\mathbf{R}}) = \sum_{k=1}^n |\mathbf{a}^H(\theta)\hat{\mathbf{u}}_k|^2$ where $\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \dots, \hat{\mathbf{u}}_m$ are the eigenvectors of $\hat{\mathbf{R}}$, sorted in the descending order of their corresponding singular values³.

²This is again obtained by noting that the covariance \mathbf{R} is a continuous map from the compact space of a bounded number of bases. Thus, it is bicontinuous and the two spaces are isomorphic.

³The matrix $\hat{\mathbf{R}}$ is symmetric positive semidefinite. Thus its singular vectors coincide with its eigenvectors. Furthermore, the singular values are the squared eigenvalues.

Denoting the estimates by $\hat{\theta}_k$ for $k = 1, 2, \dots, n$ and defining $\Delta\theta_k = \hat{\theta}_k - \theta_k$, we obtain that

$$\Delta\theta_k = \arg \max_{\Delta\theta} p(\theta_k + \Delta\theta_k, \mathbf{R} + \Delta\mathbf{R}) \quad (3.21)$$

which using Taylor expansion and after straightforward calculations leads to

$$\Delta\theta_r = \frac{\frac{\partial p}{\partial \theta}(\theta_r, \mathbf{R} + \Delta\mathbf{R})}{\frac{\partial^2 p}{\partial \theta^2}(\theta_r, \mathbf{R})} \quad (3.22)$$

The denominator is simple to calculate to obtain $\frac{\partial^2 p}{\partial \theta^2}(\theta_r, \mathbf{R}) = -2\|\mathbf{P}_{\mathbf{A}(\theta)}^\perp \mathbf{d}(\theta_r)\|_2^2$. We can further simplify the result in (3.22) by introducing $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$ as the eigenvectors of \mathbf{R} and $\Delta\mathbf{u}_k = \hat{\mathbf{u}}_k - \mathbf{u}_k$. Then, linearization leads to

$$\frac{\partial p}{\partial \theta}(\theta_r, \mathbf{R} + \Delta\mathbf{R}) = 2\Re \left[\sum_{k=1}^n \mathbf{d}^H(\theta_r) (\mathbf{u}_k \Delta\mathbf{u}_k^H + \Delta\mathbf{u}_k \mathbf{u}_k^H) \mathbf{a}(\theta_r) \right] \quad (3.23)$$

Note that the variation $\Delta\mathbf{u}_k$ is a result of the variation $\Delta\mathbf{R}$. Up to first order, this can be analytically calculated to obtain.

$$\Delta\mathbf{u}_k = \sum_{l \in \{1, 2, \dots, m\} \setminus \{k\}} \frac{\mathbf{u}_l \mathbf{u}_l^H}{\lambda_k - \lambda_l} \Delta\mathbf{R} \mathbf{u}_k \quad k = 1, 2, \dots, n \quad (3.24)$$

where λ_l for $l = 1, 2, \dots, m$ is the singular value corresponding to \mathbf{u}_l and it is assumed that λ_k is simple (has algebraic multiplicity 1). Plugging (3.24) into (3.23) and combining the result to (3.22), the relation between θ_r and $\Delta\mathbf{R}$ is obtained.

It is often desirable to identify the statistics of the error $\Delta\theta_r$. Note that by the central limit theorem, it is simple to see that $\Delta\mathbf{R}$ is asymptotically centered Gaussian and the error terms $\Delta\theta_r$ are linear functions of $\Delta\mathbf{R}$. Hence, they are also centered and jointly Gaussian and can be totally identified by the correlation elements $\mathcal{E}(\Delta\theta_k \Delta\theta_l)$. This can be performed using (3.24), (3.23) and (3.22), and noting that the correlation elements of $\Delta\mathbf{R}$ are given by a 2×2 tensor \mathcal{T} defined as

$$\begin{aligned} \mathcal{T}(\mathbf{e}_1^H, \mathbf{e}_2^H, \mathbf{e}_3, \mathbf{e}_4) &= \mathcal{E}(\mathbf{e}_1^H \Delta\mathbf{R} \mathbf{e}_3 \mathbf{e}_2^H \Delta\mathbf{R} \mathbf{e}_4) \\ &= \frac{1}{T} \mathcal{E}(\mathbf{e}_1^H \mathbf{x}(t) \mathbf{e}_2^H \mathbf{x}(t) \mathbf{x}^H(t) \mathbf{e}_3 \mathbf{x}^H(t) \mathbf{e}_4) \end{aligned} \quad (3.25)$$

This shows that the error covariance decreases with rate $1/T$. More detailed results can be found in [102].

Chapter 4

Sparsity-Based Atomic Decomposition

This chapter presents a relatively recent approach to solving the atomic decomposition problem, known as sparsity-based estimation. Similar to the spectral techniques, the idea is to use the spectral representation in (2.11). In Section 3.1.2, we discussed linear spectral estimation and argued that the linear operators, the beamformers, may not directly provide a sparse spectrum. In contrast to beamforming, the sparsity-based techniques exploit nonlinear estimators to obtain sparse spectra. Let us take a sequence of spectra $\{\tilde{s}_t \in \Psi(\mathcal{A})\}$, where $\Psi(\mathcal{A})$ denotes the set of all sparse spectra on (\mathcal{A}) . Denote $\text{Supp}(\{\tilde{s}_t\}) = \{\mathbf{a} \mid \exists t, \tilde{s}_t(\mathbf{a}) \neq 0\}$ and define $\|\{\tilde{s}_t\}\|_0$ as the cardinality of $\text{Supp}(\{\tilde{s}_t\})$. Sparsity means that $\|\{\tilde{s}_t\}\|_0 < \infty$. Now, it is clear through the relation between (2.11) and (2.1) that $\|\{\tilde{s}_t\}\|_0$ also denotes the order n of the atomic decomposition corresponding to $\{\tilde{s}_t\}$. We can also rewrite the overall procedure of atomic decomposition by ML in (3.2) and the MOS procedure in (3.16) as

$$\min_{\{\tilde{s}_t \in \Psi(\mathcal{A})\}} \sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{\mathbf{a}} \tilde{s}_t(\mathbf{a}, t) \mathbf{a} \right\|_2^2 + k \|\{\tilde{s}_t\}\|_0, \quad (4.1)$$

where $k > 0$ is a suitable constant. In this chapter, we focus on approximate techniques to solve (4.1).

4.1 Basis Pursuit

One method to solve (4.1) is to approximate its cost by a convex function. For example, it is proposed to substitute the term $\|\{\tilde{s}_t\}\|_0$ by $\|\{\tilde{s}_t\}\|_1$,

defined by

$$\|\{\tilde{s}_t\}\|_1 = \sum_{\mathbf{a} \in \mathcal{A}} \sqrt{\sum_{t=1}^T |s(\mathbf{a}, t)|^2}. \quad (4.2)$$

The result is called Basis Pursuit [12] (BP) or Least Absolute Shrinkage and Selection Operator (LASSO) [13], given by

$$\min_{\{\tilde{s}_t \in \Psi(\mathcal{A})\}} \sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{\mathbf{a}} \tilde{s}(\mathbf{a}, t) \mathbf{a} \right\|_2^2 + \lambda \|\{\tilde{s}_t\}\|_1 \quad (4.3)$$

It turns out that the optimization in (4.3) is convex on the convex set $\Psi(\mathcal{A}) \times \dots \times \Psi(\mathcal{A})$. Thus, any local minimum point is the global optimal point. Note that the parameter k is replaced by $\lambda > 0$, which essentially plays a similar role as k , i.e. it controls the order of the solution. However, the relation between λ and the order is complicated. Nevertheless, similar ideas to that of the MOS problem can be applied to the problem of selecting λ [21].

4.1.1 Implementing Basis Pursuit

In essence, the optimization in (4.3) is nonparametric, which complicates its numerical evaluation. There are different methods to tackle this problem, many of which are not compatible with the sparsity assumption on the spectrum. A promising approach is to take a discretization $\tilde{\mathcal{A}} = \{\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \dots, \tilde{\mathbf{a}}_N\}$ and restrict the spectrum to $\tilde{\mathcal{A}}$. We denote $\tilde{\mathbf{s}}(t) = [\tilde{s}_1(t) \ \tilde{s}_2(t) \ \dots \ \tilde{s}_N(t)]^T$, where $\tilde{s}_k(t) = \tilde{s}(\tilde{\mathbf{a}}_k, t)$. Then, the BP optimization is written as

$$\min_{\{\tilde{\mathbf{s}}(t) \in \mathbb{C}^N\}} \sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{k=1}^N \tilde{s}_k(t) \tilde{\mathbf{a}}_k \right\|_2^2 + \lambda \sum_{k=1}^N \sqrt{\sum_{t=1}^T |\tilde{s}_k(t)|^2} \quad (4.4)$$

It is easy to show that the optimization (4.4) has a solution with few nonzero elements, corresponding to a linearly independent set of bases. Once this solution is obtained, the atomic decomposition bases are selected as the ones, corresponding to nonzero elements in $\tilde{\mathbf{s}}(t)$. Since the solution for the amplitudes $s_k(t)$ is biased, it is instead suggested to recalculate $s_k(t)$ by using the LS solution in (3.3). This is called debiasing.

Convex Optimization

The optimization in (4.4) is convex and can be solved by general convex optimization techniques. The difficulty with (4.4) is in the non-smooth

behavior of the cost function. In fact, the desired solution of BP is at a singularity point, promoting sparsity. On the other hand, the numerical solution of such optimization problems has been under extensive study for decades, resulting in strong convex optimization solvers such as SeDuMi [103] and SDPT3 [104], used in the CVX toolbox [37, 105]. Also, note that the BP problem can be represented in different dual forms, including constraints. Thus, cone and linear programming techniques are commonly used for solving BP. In this work, we focus on the form introduced in (4.4).

Specific Approaches

The special structure of LASSO allows for special type of optimization techniques. We explain some of these techniques in the sequel.

The so-called homotopy-based techniques rely on the observation that the solution path of BP (4.4), resulting from modifying the value of λ (and keeping other parameters unchanged) is continuous. If the problem is real valued and based on single snapshot ($T = 1$), it is further shown that the path is piecewise linear [11, 32, 106]. In the solution path, the transition points are related to adding and removing new non-zero positions with small amplitudes. The position of each transition point can be predicted from the previous transition point, leading to a recursive optimization technique by following the homotopy path. In [33], it is shown that the complexity of this method equals that of solving an ordinary LS of size n . However, in the case of complex-valued parameters, multi-snapshot data or a continuous dictionary, the path is not piecewise linear anymore, but it is still piecewise smooth. We have considered a generalization of the homotopy method to these cases in [36]. The main advantage of the homotopy techniques is that they provide flexibility in selecting the regularization parameter, since they essentially provide the solutions for every possible value of λ , in a tractable way.

The Iterative Soft Thresholding Algorithm (ISTA) provides an iterative optimization technique, where the optimal point is updated at each iteration, based on locally approximating the cost function [35, 107, 108]. Rewrite (4.4) as

$$\min_{\tilde{\mathbf{S}}} \Phi_{\text{LS}}(\tilde{\mathbf{S}}) + \lambda \|\tilde{\mathbf{S}}\|_1 \quad (4.5)$$

where $\tilde{\mathbf{S}} = [\tilde{\mathbf{s}}(1) \dots \tilde{\mathbf{s}}(NT)]$ is a matrix representation of $\{\tilde{\mathbf{s}}(t)\}$ and $\Phi_{\text{LS}}(\tilde{\mathbf{S}})$ denotes the first LS part in (4.4). In the k^{th} iteration, the ISTA solves the

following approximate optimization

$$\begin{aligned} \tilde{\mathbf{S}}^{(k)} = \arg \min_{\tilde{\mathbf{S}}} & \Phi_{\text{LS}}(\tilde{\mathbf{S}}^{(k-1)}) + \nabla^T \Phi_{\text{LS}}(\tilde{\mathbf{S}}^{(k-1)}) \left(\tilde{\mathbf{S}} - \tilde{\mathbf{S}}^{(k-1)} \right) \\ & + \frac{1}{\alpha_k} \left\| \tilde{\mathbf{S}} - \tilde{\mathbf{S}}^{(k-1)} \right\|_F^2 + \lambda \|\tilde{\mathbf{S}}\|_1 \end{aligned} \quad (4.6)$$

where $\tilde{\mathbf{S}}^{(k)}$ denotes the estimate at the k^{th} iteration and α_k is the stepsize, insuring stability of the algorithm. The optimization in (4.6) has simple closed-form solution, which can be found in [108].

As a first-order programming technique, the ISTA typically has a slow convergence rate. It is proposed in [34] to apply the so-called Nesterov's gradient acceleration technique ([109]) to improve ISTA, resulting in the Fast Iterative Soft Thresholding Algorithm (FISTA). The Nesterov's acceleration technique suggests to incorporate, not only the previous $\tilde{\mathbf{S}}^{(k-1)}$, but also $\tilde{\mathbf{S}}^{(k-2)}$. The associated Nesterov's theorem states that this method achieves the convergence bound for the generic, first-order, convex optimization techniques [110, 111]. The Approximate Message Passing (AMP) algorithm is a similar algorithm to FISTA, derived under more statistical assumptions on the dictionary \mathcal{A} . The AMP algorithm is developed for the cases, where the dictionary set consists of the columns of a dictionary matrix, whose entries are generated independently by a Gaussian distribution [111–114]. However, some universality considerations suggest that it is also useful for other types of "sample" dictionaries. Note that this setup is less relevant to our consideration than that of the other techniques. Due to their simple calculations at each iteration, both FISTA and AMP are suitable in problems with a large dimension.

The SParse Iterative Covariance based Estimator (SPICE) is a different approach to solving BP [22, 115]. It exploits the interesting observation that

$$\sqrt{\sum_{t=1}^T |\tilde{s}_k(t)|^2} = \frac{1}{2} \min_{p_k > 0} \frac{\sum_{t=1}^T |\tilde{s}_k(t)|^2}{p_k} + p_k \quad (4.7)$$

Hence, the optimization in (4.4) can be written as

$$\min_{\{\tilde{\mathbf{s}}(t) \in \mathbb{C}^N\}, \{p_k\}} \sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{k=1}^N \tilde{s}_k(t) \tilde{\mathbf{a}}_k \right\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^N \frac{\sum_{t=1}^T |\tilde{s}_k(t)|^2}{p_k} + p_k \quad (4.8)$$

The SPICE solves (4.8) for $\{p_k\}$ and $\{\tilde{\mathbf{s}}(t) \in \mathbb{C}^N\}$, alternately, where both steps have closed form solutions, found in [22]. This method has a good speed of convergence, but needs a higher amount of calculations at each iteration. Thus, it is not suitable for problems with a large dictionary dimension m .

4.1.2 Regularization Parameter Selection

The regularization parameter λ in (4.3) and (4.4) plays an important role in the process of estimation by BP. Changing λ , typically leads to a remarkable effect in the AD estimate. However, λ is often identified by its role in the order selection. In general, it is not clear how to select λ . Notice that even if the order n is known, there is no simple way to decide on a value of λ , leading to the desirable order n . In this case, the homotopy techniques provide an opportunity to sweep a large range of λ values to select the desired estimate. In a more general case, the situation is more or less similar to MOS, where it is not clear how the selection should be performed.

Similar attempts to MOS can be considered for selecting λ . For example, a statistical perspective can be employed. This, for example, has led to the cross-validation approaches [12, 116]. More elaborate studies considered the regularization parameter selection as a hyper parameter estimation, where the BP estimator is treated as a Bayesian estimator with a Laplacian prior [25, 26, 117]. In [25], the Laplacian prior is also expanded in a hierarchical way and the estimation of λ is performed by considering non-informative priors for the hierarchical model. We have considered the Bayesian aspects of regularization parameter selection in [118].

More recent suggestions on the choice of λ is provided by the analysis of BP in the asymptotic cases. For example, the recent error analysis for the large random matrix based AD problem, provided an asymptotically optimal value of λ , for which the ℓ_2 error is minimized [119]. We have also considered the role of regularization parameter in a parametric AD scenario, where the SNR is high. Our semi-parametric results also lead to an approximate optimal value for the regularization parameter in Paper 2.

4.2 Analysis of Basis Pursuit for Large Dimensions

The application of BP originated from the field of image processing, where AD problems, related to large matrices were involved. Later, the technique was found useful in other application fields, concerning large matrices. For this reason, the analysis of BP traditionally revolves around dictionaries obtained by large matrices and the compressive characteristics of the AD problem. Here, we refer to the main outcomes of this type of analysis. For simplicity, a single-snapshot case is considered and the dictionary is obtained as the columns of an $m \times N$ dictionary \mathbf{A} .

As mentioned in Section 3.3, the analysis is pursued in two stages. In the first one, convergence to the ideal estimates is considered in an asymptotic

case. In the second one, a near optimal analysis is provided. The analysis, presented here is carried out in a high SNR regime, where the first stage is referred to as the ideal atomic decomposition. In Section 2.1, we show that the uniqueness of the ideal decomposition is guaranteed by the condition $n < (\text{Spark}(\mathcal{A}) - 1)/2$. It is not clear that BP is generally able to recover an ideal decomposition under the above assumption. It turns out that BP does not guarantee the recovery of the ideal decomposition under the Spark condition only. Hence, stronger conditions are necessary. However, selecting the regularization parameter is not an issue, since the high SNR case is naturally related to a vanishingly small choice of λ . In the limit, when λ shrinks to zero, the BP optimization in (4.4) approaches

$$\begin{aligned} & \min_{\tilde{\mathbf{s}}} \sum_k |\tilde{s}_k| \\ & \text{subject to } \mathbf{x} = \mathbf{A}\tilde{\mathbf{s}}, \end{aligned} \quad (4.9)$$

known as the noiseless BP optimization. The ideal decomposition question is that under which assumptions the optimization in (4.9), where \mathbf{x} is generated by $\mathbf{x} = \mathbf{A}\tilde{\mathbf{s}}_0$ and $\|\tilde{\mathbf{s}}\|_0 < (\text{Spark}(\mathcal{A}) - 1)/2$, leads to the true $\tilde{\mathbf{s}}_0$ as the solution.

4.2.1 Null Space Property

The null-space property identifies a necessary and sufficient condition for the ideal decomposition question, which can be expressed as follows [67, 120, 121]:

Theorem 1. *For any observation $\mathbf{x} = \mathbf{A}\tilde{\mathbf{s}}_0$, the solution to (4.9) is given by $\tilde{\mathbf{s}} = \tilde{\mathbf{s}}_0$ if, for any non-zero vector $\boldsymbol{\nu} = (\nu_1, \dots, \nu_N)$ in the null space of \mathbf{A} , the following condition holds*

$$\sum_{k \in \text{Supp}(\mathbf{s}_0)} |\nu_k| < \sum_{k \notin \text{Supp}(\mathbf{s}_0)} |\nu_k| \quad (4.10)$$

where $\text{Supp}(\mathbf{s}_0)$ denotes the set of indexes, corresponding to the n nonzero elements of \mathbf{s}_0 . In particular, the optimization (4.9) can recover any ideal decomposition of order n , if and only if for any subset $I \subset \{1, 2, \dots, N\}$ of n indexes and any nonzero vector $\boldsymbol{\nu}$ in the null space of \mathbf{A} the following relation holds.

$$\sum_{k \in I} |\nu_k| < \sum_{k \notin I} |\nu_k| \quad (4.11)$$

This is known as the n -null space property.

4.2.2 Restricted Isometry Property

The null-space property is not practically useful, since it is difficult to verify, or intuitively understand. Other stronger conditions are therefore developed, implying the null space property. These conditions are easier to verify, at least for a certain type of matrices. One condition, frequently considered in practice, is based on the mutual coherence, given by the maximum cosine of the angle between two distinct bases [61, 122]. However, a condition on the mutual coherence provides too conservative results. For this reason, the restricted isometry property is introduced [15, 123]. A dictionary \mathcal{A} is said to satisfy the n -restricted isometry property with restricted isometry constant δ if, for any choice of n distinct bases $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathcal{A}$, we have that

$$(1 - \delta) \leq \sigma_{\min}([\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_n]) \leq \sigma_{\max}([\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_n]) \leq (1 + \delta), \quad (4.12)$$

where σ_{\min} and σ_{\max} denote the smallest and the largest singular value of their arguments, respectively. Note that if $\delta = 0$, then the basis $\mathbf{a}_1, \dots, \mathbf{a}_n$ is orthonormal (isometric). For an infinite dictionary, the n -RIP constant δ_n is larger than 1, since in that case, one can always find a subset of n bases with an arbitrarily high mutual coherence.

If the n -RIP constant is small enough, the dictionary also satisfies the n -null space property, thus guaranteeing perfect recovery. For example, in [15] the bound $\delta < \sqrt{2} - 1$ is obtained. This is improved in [124]. It is also generally NP-hard to verify the RIP condition. However, a large body of results are provided, identifying cases, where randomly generated large matrices satisfy a suitable RIP condition. The underlying argument in these works is as follows¹: Assume that the desired order n , the size of dictionary N and the dictionary dimension m grow to infinity; and the dictionary is generated randomly with independent entries, such that for a random matrix $\Phi = [\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_n]$ and a unit vector $\mathbf{x} \in \mathbb{C}^n$, we may conclude that

$$\Pr(\left| \|\Phi \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| > \delta) < e^{-cm} \quad (4.13)$$

for a proper value of c and δ . Then, it is possible to show by the union bound that²

$$\Pr\left(\max_{\|\mathbf{x}\|=1} \left| \|\Phi \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| > \delta\right) < e^{-c'm} \quad (4.14)$$

where c' is another positive constant. Since there exist $\binom{N}{n} \leq (eN/n)^n$ combinations of bases, the union bound gives that the probability of the

¹See [125] for more details.

²For this, one needs to take an exponentially growing number of maximally separated points on a unit sphere and use the triangle inequality for an arbitrary point \mathbf{x} on the sphere.

n -RIP constant being larger than δ is bounded by

$$e^{-c'm} \times \left(\frac{eN}{n}\right)^n = e^{n \log(\frac{eN}{n}) - c'm} \quad (4.15)$$

Thus the probability goes to zero if the number of measurements (the size of the observation vector) m grows faster than $n \log(N/n)$, which is a popular result. The threshold rate $n \log(N/n)$ has also been shown to be maximal in this setup.

4.2.3 Error Analysis

Recently, the second step of error analysis has been taken by two independent research groups [126, 127] and [18, 119, 128]. These papers establish results for randomly generated Gaussian dictionaries, though it is empirically observed that the result is universal for a large family of random dictionaries [129]. It is shown that the error term $\|\tilde{\mathbf{s}} - \tilde{\mathbf{s}}_0\|_2$, where $\tilde{\mathbf{s}}$ is obtained by (4.4) has a deterministic limit when dimensions grow. The study in [126] is based on AMP and demonstrates more characteristics of the error. The work in [18] utilizes the so-called comparison inequalities and considers a more general framework than the AMP-based approach and LASSO.

4.3 The Off-Grid Problem

Remember that for the infinite dictionaries, the parametric form of BP in (4.4) is obtained by considering a discretization. When the data vector is obtained by basis vectors, excluded from the discretized basis frame $\tilde{\mathcal{A}}$, the so-called off-grid problem occurs. If the discretization is fine enough, such that an excluded base can be approximated by nearby elements in $\tilde{\mathcal{A}}$, and the true order is small enough, the off-grid effect is not severe, but still degrades the high SNR properties of estimates. Usually, the off-grid base is approximated by multiple nearby on-grid elements, which we refer to as its cloud. In a high-SNR case, the cloud for each base is easily distinguished in the exact solution of BP. Once a cloud is calculated, its elements should be combined to obtain a focused solution.

To tackle the off-grid effect, some techniques have recently been considered for a case, where the bases are represented by a real number θ . To explain the main idea, we focus on the single-snapshot case. Using the Taylor expansion, we obtain

$$\mathbf{a}(\theta) \approx \mathbf{a}(\tilde{\theta}_l) + \mathbf{d}(\tilde{\theta}_l)\Delta\theta \quad (4.16)$$

where $\tilde{\theta}_l$ is the nearest element to θ in $\tilde{\mathcal{A}}$ and $\Delta\theta = \theta - \theta_l$. Then, (2.1) can be written as

$$\mathbf{x} \approx \sum_k \left(\mathbf{a}(\tilde{\theta}_{l_k}) + \mathbf{d}(\tilde{\theta}_{l_k})\Delta\theta_k \right) s_k + \mathbf{n} = \sum_k \mathbf{a}(\tilde{\theta}_{l_k})s_k + \mathbf{d}(\tilde{\theta}_{l_k})\beta_k + \mathbf{n} \quad (4.17)$$

where θ_{l_k} is the nearest grid point to θ_k and $\Delta\theta_k = \theta_{l_k} - \theta_k$. Moreover, $\beta_k = s_k\Delta\theta_k$.

Accordingly, the Sparse Total Least Square (S-TLS) approach suggests to solve the following optimization ³ [45]:

$$\min_{\{\Delta\tilde{\theta}_l, \tilde{s}_l\}} \frac{1}{2} \left\| \mathbf{x}_k - \sum_{l=1}^N \left(\mathbf{a}(\tilde{\theta}_l) + \mathbf{d}(\tilde{\theta}_l)\Delta\tilde{\theta}_l \right) \tilde{s}_l \right\|_2^2 + \lambda \sum_{l=1}^N |\tilde{s}_l| + \frac{\mu}{2} \sum_l |\Delta\tilde{\theta}_l|^2 \quad (4.18)$$

where μ is practically a tuning parameter. The S-TLS method can be solved exactly with the method, explained in [45]. It can also be solved by alternately minimizing over $\{\Delta\tilde{\theta}_k\}$ and $\{\tilde{s}_k\}$.

Another approach is to use the last expression in (4.17), where the relation between β_k and s_k is generally non-convex. An exception is when $s_k > 0$ is real and $\Delta\theta_k$ is bounded in a convex set. In a general case, the nonconvex relation can be convexified to obtain the following optimization

$$\min_{\{\tilde{\beta}_l, \tilde{s}_l\}} \frac{1}{2} \left\| \mathbf{x}_k - \sum_{l=1}^N \mathbf{a}(\tilde{\theta}_l)\tilde{s}_l + \mathbf{d}(\tilde{\theta}_l)\tilde{\beta}_l \right\|_2^2 + \lambda \sum_{l=1}^N \sqrt{|\tilde{s}_l|^2 + |\tilde{\beta}_l|^2} \quad (4.19)$$

which is referred to in [38] as the Joint LASSO (J-LASSO) optimization. The J-LASSO optimization is convex and can be solved by off-the-shelf optimization techniques, or simplified methods [38, 46].

In all of the above techniques, the final result still suffers from a defocused cloud of estimates. In [46], it is suggested to use the following merging technique. Denoting by $\{\hat{s}_l, \hat{\theta}_l\}$, the cloud related a true set of parameters (s, θ) , it is proposed to combine the cloud by

$$\hat{s} = \sum_l \hat{s}_l \quad \hat{\theta} = \frac{\sum_l |\hat{s}_l| \hat{\theta}_l}{\sum_l |\hat{s}_l|} \quad (4.20)$$

to obtained weighted average estimates, which has an interesting physical interpretation as center of gravity.

³The original definition in [45] is slightly different. It is based on an unstructured basis perturbation \mathbf{e} instead of $\mathbf{d}(\theta)\Delta\theta$.

4.4 Other Approaches

The problem of atomic decomposition has a long history, and has been discussed in a variety of different applications. The sparsity-based approaches are relatively recent. However, different approaches are also discussed in this context. One of the first approaches is Matching Pursuit (MP), which is a forward stagewise algorithm, i.e. it selects a new base at each stage [130]. Having an ADP estimate and the remainder term at a given stage, the next stage adds a new pair (\mathbf{a}, s) to the AD by taking the largest projection of the remainder vector onto the basis vectors \mathbf{a} . The previous parameter estimates do not change. Orthogonal Matching Pursuit (OMP) modifies MP by replacing the remainder vector by the projection vector into the orthogonal complement of the linear span of the previous estimates [131]. Inspired by basis pursuit, the Dantzig Selector (DS) was introduced in [132], which promotes stronger sparsity than BP. Inspired by different numerical implementations of the BP, other modified approaches have also been introduced. For example, the homotopy implementation and its modifications is usually referred to as Least Angle Regression (LARS), first termed by Efron [33]. The approximate message passing technique has also introduced the belief propagation ideas to the field of sparse regression [133]. The SPICE approach has also been extended by Stoica to obtain the LIKelihood based Estimation of Sparse parameters (LIKES) [115], the Iterative Adaptive Approach (IAA) [134, 135] and Sparse Learning via Iterative Minimization (SLIM) [136]. The idea of weighted ℓ_1 regularization is further frequently discussed [137]. Finally, regularization by the so-called $p < 1$ semi-norm is also studied. A good example of the latter is the FOCal Underdetermined System Solver (FOCUSS) [138]. It should be remembered, though, that for $p < 1$ the norm is not convex.

Chapter 5

Dynamic Atomic Decomposition

In this chapter, we consider a generalization of the atomic decomposition model, introduced in (2.1). Here, we assume that the bases \mathbf{a}_k may vary by time, such that the data model is given by

$$\mathbf{x}(t) = \sum_{k=1}^{n(t)} \mathbf{a}_k(t) s_k(t) + \mathbf{n}(t) \quad (5.1)$$

This is the case in applications such as sensor array processing, seismology and medical tomography. It is further assumed that the bases are temporally correlated, such that the observations at different time instants can be combined to improve the estimation performance at a certain time instant. In this manner, different types of questions can be considered. For example, the filtering problem concerns estimating the AD at a time instant t , based on the vectors $\mathbf{x}(t')$, observed up to time $t \geq t'$. Although the focus here is on filtering, it should be noted that other types of problems also exist, depending on the amount of observation data presented for a specific estimate. The problem of estimating the parameter trajectories is also widely considered.

To obtain a desired AD, the process of filtering is vague, unless clear statistical assumptions on the temporal relation of the parameters are made. On the other hand, the main characteristics of the dynamic AD model in (5.1) is its dynamic parameter size (order). Hence, the temporal models of AD are complicated. The sparsity-based techniques have recently been applied to simplify these types of problems. However, it seems problematic to rely on the spectral model to express the temporal relation. Toward this goal, simple steps are taken in [48, 51, 139–141]. In the sequel, we first present a general framework for statistical filtering and then relate it to the dynamic AD problem.

5.1 Recursive Bayesian Estimator

In this section, we present the general theory of filtering by a Bayesian recursion. Later, we relate this to the AD problem. Although, a variety of different statistical models may be considered, a fairly general and popular one is the state space based model. Consider a system, described by the state S , belonging to a state space \mathcal{S} . Suppose a sequence of observations $\{\mathbf{x}(t)\}$ is obtained by the system at the corresponding states $\{S_t = S(t)\}$. The state space model assumes that the statistics of the state at a time instant $t + 1$ is completely identified by the previous state at the time instant t . Mathematically speaking, this is described by a Markov Chain (MC) process given by the following joint distribution over an arbitrary time window $t, t + 1, \dots, t + T$:

$$p_{S_t, S_{t+1}, \dots, S_{t+T-1}}(s_t, s_{t+1}, \dots, s_{t+T-1}) = p_{S_t}(s_t) \times p_{S_{t+1}|S_t}(s_{t+1} | s_t) p_{S_{t+2}|S_{t+1}}(s_{t+2} | s_{t+1}) \dots p_{S_{t+T-1}|S_{t+T-2}}(s_{t+T-1} | s_{t+T-2}) \quad (5.2)$$

where $p_{S_{t+1}|S_t}(s_1 | s_0)$ is called the transitional probability density. If the transitional probability density is constant, i.e. $p_{S_{t+1}|S_t}(s_1 | s_0) = Q(s_1 | s_0)$, for a fixed function Q , then the MC is called time homogeneous. We only consider time homogeneous systems. According to the state space model, the observation vector $\mathbf{x}(t)$ is inclusively determined by the state S_t through a conditional distribution $p_{\mathbf{x}(t)|S(t)}(\mathbf{x}(t) | s(t))$, in short denoted by $p(\mathbf{x}(t) | s(t))$. At a certain time instant t , the question of interest is to estimate a group of parameters based on the observations $\mathbf{x}(1), \dots, \mathbf{x}(t)$. For example, the entire state trajectory can be estimated by the Maximum a' Posteriori¹ (MAP) estimator as

$$\begin{aligned} (\hat{s}_1, \hat{s}_2, \dots, \hat{s}_t) &= \arg \max_{s_1, s_2, \dots, s_t} \\ p(s_1, s_2, \dots, s_t | \mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(T)) &= \\ \arg \max_{(s_1, s_2, \dots, s_t)} & \\ p(\mathbf{x}(1) | s_1) p(\mathbf{x}(2) | s_2) \dots p(\mathbf{x}(t) | s_t) \times p(s_1) \times & \\ Q(s_2 | s_1) Q(s_3 | s_2) \dots Q(s_t | s_{t-1}) & \end{aligned} \quad (5.3)$$

It is seen that the final optimization in (5.3) can be efficiently solved in a recursive way. Define

$$\begin{aligned} V_t(s_t) &= \max_{s_1, s_2, \dots, s_{t-1}} \\ p(\mathbf{x}(1) | s_1) p(\mathbf{x}(2) | s_2) \dots p(\mathbf{x}(t-1) | s_{t-1}) \times p_{S_1}(s_1) \times & \\ Q(s_2 | s_1) Q(s_3 | s_2) \dots Q(s_t | s_{t-1}) & \end{aligned} \quad (5.4)$$

¹See [142].

which may be obtained by

$$V_t(s_t) = \max_{s_{t-1}} V_{t-1}(s_{t-1})Q(s_t | s_{t-1})p(\mathbf{x}(t-1) | s_{t-1}) \quad (5.5)$$

where the maximum point is denoted by $\hat{s}_{t-1}(s_t)$. For the final time t , we can write

$$\hat{s}_t = \arg \max_{s_t} p(\mathbf{x}(t) | s_t)V_t(s_t) \quad (5.6)$$

The estimates at the previous times $t' < t$ can also be found backward recursively as $\hat{s}_{t'} = \hat{s}_{t'}(\hat{s}_{t'+1})$. This is known as the Viterbi algorithm [143], which is closely related to the Bellman recursive decision algorithm [144].

Another case of interest is when only S_t is under question at time t . Then, the (MAP) estimator is given by

$$\hat{s}_t = \max_{s_t} p_{S_t | \mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t)}(s_t | \mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t)) \quad (5.7)$$

Interestingly, the MAP estimator is again solved in a recursive way. For simplicity, define $\mathbf{X}^{(t)} = [\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t)]$. Then, by the Bayes rule, we have that

$$p_{S_t | \mathbf{X}^{(t)}}(s_t | \mathbf{X}^{(t)}) = \frac{p_{S_t | \mathbf{X}^{(t-1)}}(s_t | \mathbf{X}^{(t-1)})p(\mathbf{x}(t) | s_t)}{p(\mathbf{x}(t) | \mathbf{X}^{(t-1)})} \quad (5.8)$$

where

$$p_{S_t | \mathbf{X}^{(t-1)}}(s_t | \mathbf{X}^{(t-1)}) = \int_{s_{t-1} \in \mathcal{S}} Q(s_t | s_{t-1})dP_{S_{t-1} | \mathbf{X}^{(t-1)}}(s_{t-1} | \mathbf{X}^{(t-1)}) \quad (5.9)$$

The steps in (5.8) and (5.9) are known as update and prediction, respectively. The overall algorithm by recursively applying them is referred to as Recursive Bayesian Filtering (RBF) [145, 146]. Similarity is observed between the Viterbi algorithm and RBF. In a sense, both approaches follow the evolution of a spectrum over the state space. The difference is that the Viterbi algorithm uses recursive optimization, while RBF employs integration.

5.2 Filtering Theory for Atomic Decomposition

Now, we discuss the application of filtering to the AD model. The first issue is to define the state space. Clearly, the state space is given by the

set of all desirable decompositions for a single snapshot $T = 1$. Mathematically speaking, such a decomposition, consisting of the bases $\mathbf{a}_1, \dots, \mathbf{a}_n$ and amplitudes s_1, s_2, \dots, s_n , can be represented by a state

$$S = \{(\mathbf{a}_1, s_1), (\mathbf{a}_1, s_2), \dots, (\mathbf{a}_n, s_n)\}. \quad (5.10)$$

We are particularly interested in decompositions of an order $n < (\text{Spark}(\mathcal{A}) - 1)/2$. Thus, \mathcal{S} is identified as the collection of all finite sets of $\mathcal{A} \times \mathbb{C}$ with a cardinality smaller than $n < (\text{Spark}(\mathcal{A}) - 1)/2$. The second issue is to define probability densities on \mathcal{S} , which essentially gives a random finite set characteristic to the state. To the best of our knowledge, this approach has not been directly applied to the problem of our interest. However, the random finite set theory is well-studied in mathematics and also applied in the signal processing literature, for example in target tracking [147–149].

Simpler models for ADP is obtained, when the order n is assumed to be fixed and the dictionary is labeled by θ . In this case, similar to the parametric approaches, the state S_t is given by two vectors $\boldsymbol{\theta}(t) = (\theta_1(t), \theta_2(t), \dots, \theta_n(t))$ and $\mathbf{s}(t) = (s_1(t), s_2(t), \dots, s_n(t))$. Then, $\mathcal{S} = (\mathbb{R}^d)^n \times \mathbb{C}^n$, where we remind that $\theta \in \mathbb{R}^d$. Simple models for the parameter evolution ($Q(S_t | S_{t-1})$) are considered. For example, the linear case

$$\begin{aligned} \boldsymbol{\theta}(t+1) &= \mathbf{H}_\theta \boldsymbol{\theta}(t) + \mathbf{w}_\theta(t) \\ \mathbf{s}(t+1) &= \mathbf{H}_s \mathbf{s}(t) + \mathbf{w}_s(t) \end{aligned} \quad (5.11)$$

where \mathbf{H}_θ and \mathbf{H}_s are known (and often identity). Moreover, \mathbf{w}_θ and \mathbf{w}_s are two independent, uncorrelated, white, centered Gaussian processes, known as the process noise and the observation noise, respectively. The observation model $p(\mathbf{x}(t) | S_t)$ is given by (5.1), where $\mathbf{a}_k(t) = \mathbf{a}(\theta_k(t))$. Given the evolution and the observation model, it is possible to obtain a RBF. However, it is generally difficult to solve the integrals and store the posteriors in a computing machine. Thus, different approximate solutions are considered.

5.2.1 Extended Kalman Filter

If the parameter variation is small at each time and the SNR is high, it is possible to approximate the filter. We assume that the distributions $p(s_t | \mathbf{X}^{(t-1)})$ and $p(s_t | \mathbf{X}^t)$ are Gaussian with mean $\hat{s}_{t|t-1}, \hat{s}_{t|t}$ and covariance matrices $\mathbf{P}_{t|t-1}, \mathbf{P}_{t|t}$, respectively. Notice that by Taylor expansion,

$$\mathbf{x}(t) \approx \sum_{k=1}^n \mathbf{a}(\hat{\theta}_k(t | t-1)) s_k + \mathbf{d}(\hat{\theta}_k(t | t-1)) (\theta_k - \hat{\theta}_k(t | t-1)) \hat{s}_k(t | t-1) \quad (5.12)$$

where we used the notation $\hat{s}_{t|t-1} = (\{\hat{\theta}_k(t | t - 1)\}, \{\hat{s}_k(t | t - 1)\})$ and² $\mathbf{d}(\theta) = \mathbf{d}\mathbf{a}(\theta)/\mathbf{d}\theta$. Using the approximate relation in (5.12) and the Gaussian assumption, (5.9) and (5.8) are solved to obtain the so called Kalman Filter (KF) [150], recursively updating the parameters $\hat{s}_{t|t-1}$, $\hat{s}_{t|t}$ and variance $\mathbf{P}_{t|t-1}$, $\mathbf{P}_{t|t}$. The approach of obtaining a KF by linear approximations is called Extended Kalman Filter (EKF) [151]. The EKF can also be improved by other techniques such as Unscented Kalman (UKF) Filter [152], but due to their locality, they generally have poor results in highly nonlinear cases.

5.3 Sparsity-Based Filtering

The problem of tracking a dynamic set of parameters is also addressed in a sparsity-based framework. The focus in these works has been on the finite dictionary case, where a sparse vector $\tilde{\mathbf{s}}(t)$ represents the spectrum $\tilde{s}(\theta, t)$ [48, 51, 140, 141]. It is difficult to connect the model in (5.11) to this setup and often no clear statistical assumptions on the dynamics of $\tilde{\mathbf{s}}(t)$ are made. Thus, lacking a rigorous Bayesian framework motivated to replace the RBF approach by heuristic methods. We present two popular examples below. We assume that a sparse vector $\tilde{\mathbf{s}}(t)$ slowly evolves in time. This means that both, the sparsity pattern (support) and the amplitudes of this vector may slowly vary by time. For the methods presented herein, no more rigorous statistical assumptions are made.

The study in [48] considers a more general setup than the AD model, where the observation at time instant t is obtained by $\mathbf{x}(t) = \mathbf{A}_t \tilde{\mathbf{s}}(t) + \mathbf{n}(t)$, where the dictionary \mathbf{A}_t may also vary by time. Assuming that the sparse vector $\tilde{\mathbf{s}}(t)$ evolves slowly and motivated by the Recursive Least Squares method, [48] suggests to obtain the estimate at time t by solving the following optimization

$$\arg \min_{\tilde{\mathbf{s}}} \sum_{\tau=0}^t \gamma_{t,\tau} \|\mathbf{x}(\tau) - \mathbf{A}_\tau \tilde{\mathbf{s}}\|_2^2 + \lambda \|\tilde{\mathbf{s}}\|_1 \quad (5.13)$$

where $\gamma_{t,\tau} > 0$ is a predefined sequence of weights that usually decreases with increasing time difference $t - \tau$. The idea with (5.13) is that at each time instant, the vector $\tilde{\mathbf{s}}(t)$ is assumed to be constant for the time interval $\tau = 0, 1, \dots, t$, and the modeling error induced by such an assumption is reflected by the weight $\gamma_{t,\tau}$. The optimization in (5.13) can also be written

²If θ is multidimensional, the derivative should be replaced by gradient, and its corresponding manipulations should be replaced by proper tensorial ones. This is neglected in favor of simplicity.

as

$$\arg \min_{\tilde{\mathbf{s}}} \tilde{\mathbf{s}}(t)^H \tilde{\mathbf{R}}_t \tilde{\mathbf{s}}(t) + 2\Re(\mathbf{z}(t)^H \tilde{\mathbf{s}}(t)) + \lambda \|\tilde{\mathbf{s}}\|_1 \quad (5.14)$$

where

$$\begin{aligned} \tilde{\mathbf{R}}_t &= \sum_{\tau=0}^t \gamma_{t,\tau} \mathbf{A}_\tau^H \mathbf{A}_\tau \\ \mathbf{z}(t) &= \sum_{\tau=0}^t \gamma_{t,\tau} \mathbf{A}_\tau^H \mathbf{x}(\tau) \end{aligned} \quad (5.15)$$

The interesting fact about this method is that if one selects $\gamma_{t,\tau} = \beta^{t-\tau}$ for a given value of β , then $\tilde{\mathbf{R}}_t$ and $\mathbf{z}(t)$ can be recursively calculated.

Another approach is introduced in [51] for models, where the sparsity pattern varies slowly by time. Clearly, this is not generally compatible with the model introduced in (5.11). Still, it is useful in particular applications such as MRI imaging. The idea is that obtaining a new observation, a Kalman filter iteration is applied over the previously estimated support. Then, a statistical test is performed to detect support change. If a support change is detected, a sparsity-based estimation technique, such as LASSO or Dantzig selector [132] is applied over the off-support elements. The new support is added to the previous one and the Kalman step is corrected by taking the new support. Finally the indexes corresponding to small elements is removed from the support.

As seen, the above techniques are not based on clear statistical assumptions and do not follow the general RBF methodology. Thus, it is difficult to discuss their performance. We have considered this problem and provided approximate techniques to apply RBF to the sparsity-based tracking problem [52, 54]. In this thesis, Paper 3 is related to this topic.

Chapter 6

Thesis Contributions

In this thesis, we study the application of the basis pursuit approach in parameter estimation problems, which can be represented by atomic decomposition. First we study the regularization parameter selection and its Bayesian aspects. Later, we consider deterministic selection of the parameters, which motivates investigating the homotopy methods in complex valued problems. Next, we provide methods to overcome the off-grid problem and formulate a continuous extension of BP, closely related to atomic norm de-noising. We develop a numerical approach to solve the extended BP. The method is guaranteed to converge to the global optimum with a moderate computational effort. Using the framework of continuous extension, we present the analysis of LASSO in a high-SNR scenario. We also utilize the continuous BP framework to develop a random finite set based Bayesian interpretation for sparsity-based estimation. Considering dynamic set of parameters, we used this approach to design improved recursive Bayesian filters, avoiding the NP-hard problem of association.

6.1 Summary of Appended Papers

Paper 1 proposes a numerical implementation of the continuously extended BP, in the recently developed framework of atomic norm de-noising. The paper includes comparisons with other techniques, proposed to alleviate the off-grid effect. The design of the proposed algorithm is presented, such that global convergence is evident. Numerical results on the speed of convergence are also included.

Paper 2 presents the analysis of BP, by linking BP, in a case with a highly dense grid, to the continuous framework, developed in Paper 1. New mathematical tools are developed to perform analysis in a high-SNR scenario. According to the variable order of estimates, these tools essentially

formulate the perturbation theory of finite sets and connect it to the existing terminology in the field of parameter estimation. In this paper, interesting properties of BP, such as its resolution limit and the biasing effect of the absolute shrinkage operator, as well as the choice of regularization parameter is discussed.

Motivated by the findings in Papers 1 and 2, we later considered the continuous extension of BP as a finite set estimator and attempted to interpret it in a Bayesian sense. This finally led to a framework presented in Paper 3, where the developed Bayesian method was incorporated in a RFS-based recursive Bayesian filter to enhance estimation of dynamic parameter sets. We present results suggesting an improvement in estimation performance.

6.2 Suggestions for Future Work

Today, the sparsity-based estimation area is highly active. The need for applying the sparsity-based estimation methods to emerging applications with a potentially unaccustomed data model, naturally calls for further research on adapting the existing techniques to these applications. Furthermore, our analysis shows deficiency in parameter estimation by the existing sparse estimation techniques. Accordingly, we propose the following possibilities for a future study.

From our current understanding of sparsity-based parameter estimation techniques, it is clear that the convex methods, such as LASSO lead to statistically inefficient estimates, due to a structured model mismatch. There are opportunities, such as re-weighting to improve the result in the literature. Their relation with parameter estimation and our continuous interpretation of LASSO can be clarified in a future study. Note that this study focused on the parametric aspects of LASSO, while many proposed improvements essentially deal with the spectral interpretation of LASSO.

Another important issue is to consider different observation models, such as the ones representing practical observation impairments. The phase retrieval and the 1-bit compressed sensing are popular examples. While little is known about the general behavior of this type of problems, the parameter estimation perspective not only frames them into a more practical framework, but also provides a new opportunity to analyze them.

Last but not least, we propose to study the role of dictionary learning techniques in parameter estimation. Dictionary learning is the process of simultaneously learning the dictionary and atomic decomposition from a sequence of observed data. A great potential is observed in parametric dictionary learning as it is simply seen to be related to the well-know family of blind estimation problems, such as blind deconvolution, blind source

6.2. SUGGESTIONS FOR FUTURE WORK

separation and blind channel estimation. Again, the mixture of parametric and sparsity-based estimation perspectives is seen to be highly useful in developing related techniques.

References

- [1] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, “Genome-wide association analysis by lasso penalized logistic regression,” *Bioinformatics*, vol. 25, pp. 714–721, Mar. 2009.
- [2] H. Konno and H. Yamazaki, “Mean-absolute deviation portfolio optimization model and its applications to tokyo stock market,” *Manage. Sci.*, vol. 37, pp. 519–531, May 1991.
- [3] W. Tu and S. Sun, “Spatial filter selection with lasso for EEG classification,” in *Advanced Data Mining and Applications*, Chongqing, China, 2010, pp. 142–149.
- [4] M. Mishali and Y. 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.
- [5] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, “Beyond nyquist: Efficient sampling of sparse bandlimited signals,” *IEEE Trans. Inform. Theory*, vol. 56, pp. 520–544, Jan. 2010.
- [6] H. Yao, P. Gerstoft, P. M. Shearer, and C. Mecklenbräuker, “Compressive sensing of the tohoku-oki mw 9.0 earthquake: Frequency-dependent rupture modes,” *Geophys. Res. Lett.*, vol. 38, Oct. 2011.
- [7] M. Lustig, D. Donoho, and J. M. Pauly, “Sparse MRI: The application of compressed sensing for rapid MR imaging,” *Resonance Med. Mag.*, vol. 58, pp. 1182–1195, Dec. 2007.
- [8] J. Van Leeuwen, *Handbook of theoretical computer science, Vol B: Formal models and semantics*. Elsevier, 1990, vol. 137.
- [9] R. G. Baraniuk, “Compressive sensing [lecture notes],” *IEEE Signal Processing Mag.*, vol. 24, pp. 118–121, July 2007.
- [10] D. Donoho, “Compressed sensing,” *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

REFERENCES

- [11] D. Donoho and Y. Tsaig, “Fast solution of 1-norm minimization problems when the solution may be sparse,” *IEEE Trans. Inform. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.
- [12] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic Decomposition by Basis Pursuit,” *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, Dec. 1998.
- [13] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Stat. Soc., Series B (Methodological)*, vol. 58, pp. 267–288, Jan. 1996.
- [14] E. J. Candès and Y. Plan, “Near-ideal model selection by ℓ_1 minimization,” *Ann. Stat.*, vol. 37, pp. 2145–2177, Oct. 2009.
- [15] E. J. Candès and T. Tao, “Decoding by linear programming,” *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
- [16] E. J. Candès and Y. Plan, “A Probabilistic and RIPless Theory of Compressed Sensing,” *IEEE Trans. Inform. Theory*, vol. 57, no. 11, pp. 7235–7254, Nov. 2011.
- [17] A. Maleki, L. Anitori, Z. Yang, and R. G. Baraniuk, “Asymptotic analysis of complex lasso via complex approximate message passing (camp),” *CoRR*, vol. abs/1108.0477, 2011.
- [18] S. Oymak, C. Thrampoulidis, and B. Hassibi, “The squared-error of generalized lasso: A precise analysis,” *arXiv preprint arXiv:1311.0830*, 2013.
- [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] S. Theodoridis and R. Chellappa, *Academic Press Library in Signal Processing: Array and Statistical Signal Processing*. Academic Press, 2013, vol. 3.
- [21] J. J. Fuchs, “Detection and estimation of superimposed signals,” in *IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 3, May 1998, pp. 1649–1652 vol.3.
- [22] P. Stoica, P. Babu, and J. Li, “Spice: A sparse covariance-based estimation method for array processing,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 2, pp. 629–638, 2011.

- [23] D. Malioutov, M. Çetin, and A. S. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 3010–3022, 2005.
- [24] M. A. Herman and T. Strohmer, “High-resolution radar via compressed sensing,” *Signal Processing, IEEE Transactions on*, vol. 57, no. 6, pp. 2275–2284, 2009.
- [25] M. Figueiredo, “Adaptive sparseness for supervised learning,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1150 – 1159, Sept. 2003.
- [26] M. Yuan and Y. Lin, “Efficient empirical bayes variable selection and estimation in linear models,” *Journal of the American Statistical Association*, vol. 100, no. 472, 2005.
- [27] J. Huang, T. Zhang, and D. Metaxas, “Learning with structured sparsity,” *The Journal of Machine Learning Research*, vol. 12, pp. 3371–3412, 2011.
- [28] H. Waki, S. Kim, M. Kojima, and M. Muramatsu, “Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity,” *SIAM Journal on Optimization*, vol. 17, no. 1, pp. 218–242, 2006.
- [29] G. Yu, G. Sapiro, and S. Mallat, “Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity,” *Image Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 2481–2499, 2012.
- [30] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, “Recovering sparse signals using sparse measurement matrices in compressed dna microarrays,” *IEEE, J. Select. Topics Signal Processing*, vol. 2, pp. 275–285, June 2008.
- [31] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, “Simultaneously structured models with application to sparse and low-rank matrices,” *arXiv preprint arXiv:1212.3753*, 2012.
- [32] M. R. Osborne, B. Presnell, and B. Turlach, “A new approach to variable selection in least squares problems,” 1999.
- [33] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, “Least angle regression,” *Ann. Stat.*, vol. 32, pp. 407–499, Apr. 2004.

REFERENCES

- [34] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [35] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communications on pure and applied mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [36] A. Panahi and M. Viberg, “Fast candidate points selection in the lasso path,” *IEEE Signal Processing Lett.*, vol. 19, no. 2, pp. 79–82, Feb. 2012.
- [37] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 1.21,” <http://cvxr.com/cvx>, Apr. 2011.
- [38] Z. Tan, P. Yang, and A. Nehorai, “Joint sparse recovery method for compressed sensing with structured dictionary mismatch,” *arXiv preprint arXiv:1309.0858*, 2013.
- [39] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, “Sensitivity to basis mismatch in compressed sensing,” *IEEE Trans. Signal Processing*, 2011.
- [40] P. Zhao and B. Yu, “On model selection consistency of lasso,” *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.
- [41] E. Candes and C. Fernandez-Granda, “Towards a mathematical theory of super-resolution,” *arXiv preprint arXiv:1203.5871*, 2012.
- [42] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, “Recovery of sparse translation-invariant signals with continuous basis pursuit,” *IEEE Trans. Signal Processing*, 2011.
- [43] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, “Compressive sensing off the grid,” in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 778–785.
- [44] B. N. Bhaskar and B. 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.
- [45] H. Zhu, G. Leus, and G. B. Giannakis, “Sparsity-cognizant total least-squares for perturbed compressive sampling,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 5, pp. 2002–2016, 2011.

- [46] Z. Tan and A. Nehorai, “Sparse direction of arrival estimation using co-prime arrays with off-grid targets,” *Signal Processing Letters, IEEE*, vol. 21, no. 1, pp. 26–29, 2014.
- [47] M. V. Ashkan Panahi, “Gridless compressive sensing,” in *IEEE Int. Conf. Acoust. Speech, Signal Processing*, 2014.
- [48] D. Angelosante and G. Giannakis, “RLS-weighted lasso for adaptive estimation of sparse signals,” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, Apr. 2009, pp. 3245–3248.
- [49] N. Vaswani and W. Lu, “Modified-CS: Modifying compressive sensing for problems with partially known support,” *Signal Processing, IEEE Transactions on*, vol. 58, no. 9, pp. 4595–4607, 2010.
- [50] Y. Kopsinis, K. Slavakis, and S. Theodoridis, “Online sparse system identification and signal reconstruction using projections onto weighted balls,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 3, pp. 936–952, 2011.
- [51] N. Vaswani, “Kalman filtered compressed sensing,” in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 893–896.
- [52] A. Panahi and M. Viberg, “Fast lasso based doa tracking,” in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on*. IEEE, 2011, pp. 397–400.
- [53] —, “A novel method of doa tracking by penalized least squares,” in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2013 IEEE 5th International Workshop on*. IEEE, 2013, pp. 61–64.
- [54] C. F. Mecklenbrauker, P. Gerstoft, A. Panahi, and M. Viberg, “Sequential bayesian sparse signal reconstruction using array data,” *Signal Processing, IEEE Transactions on*, vol. 61, no. 24, pp. 6344–6354, 2013.
- [55] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin, “Sparsity and compressed sensing in radar imaging,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1006–1020, 2010.

REFERENCES

- [56] D. Malioutov, M. Çetin, and A. S. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 3010–3022, 2005.
- [57] D. Model and M. Zibulevsky, “Signal reconstruction in sensor arrays using sparse representations,” *Signal Processing*, vol. 86, no. 3, pp. 624–638, 2006.
- [58] V. Cevher, M. Duarte, and R. G. Baraniuk, “Distributed target localization via spatial sparsity,” in *European Signal Processing Conference (EUSIPCO)*, 2008.
- [59] J.-J. Fuchs, “On the application of the global matched filter to doa estimation with uniform circular arrays,” *Signal Processing, IEEE Transactions on*, vol. 49, no. 4, pp. 702–709, 2001.
- [60] M. Unser, “Sampling-50 years after shannon,” *Proc. IEEE*, vol. 88, no. 4, pp. 569–587, 2000.
- [61] Y. C. Eldar and M. Mishali, “Robust recovery of signals from a structured union of subspaces,” *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.
- [62] J. S. Lim, “Two-dimensional signal and image processing,” *Englewood Cliffs, NJ, Prentice Hall, 1990, 710 p.*, vol. 1, 1990.
- [63] J.-L. Starck, M. Elad, and D. L. Donoho, “Image decomposition via the combination of sparse representations and a variational approach,” *Image Processing, IEEE Transactions on*, vol. 14, no. 10, pp. 1570–1582, 2005.
- [64] E. Van Den Berg and M. P. Friedlander, “Probing the pareto frontier for basis pursuit solutions,” *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.
- [65] S. G. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, no. 7, pp. 674–693, 1989.
- [66] E. J. Candes, D. L. Donoho *et al.*, *Curvelets: A surprisingly effective nonadaptive representation for objects with edges*. DTIC Document, 1999.

- [67] D. Donoho and X. Huo, “Uncertainty principles and ideal atomic decomposition,” *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 2845–2862, nov 2001.
- [68] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.
- [69] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.
- [70] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, “Large-scale parallel collaborative filtering for the netflix prize,” in *Algorithmic Aspects in Information and Management*. Springer, 2008, pp. 337–348.
- [71] G. Takács, I. Pilászy, B. Németh, and D. Tikk, “Matrix factorization and neighbor based algorithms for the netflix prize problem,” in *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 2008, pp. 267–274.
- [72] G. B. Folland, *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013.
- [73] R. Roy and T. Kailath, “Esprit-estimation of signal parameters via rotational invariance techniques,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 7, pp. 984–995, 1989.
- [74] B. D. Rao and K. Hari, “Performance analysis of root-music,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 12, pp. 1939–1949, 1989.
- [75] B. Ottersten, P. Stoica, and R. Roy, “Covariance matching estimation techniques for array signal processing applications,” *Digital Signal Processing*, vol. 8, no. 3, pp. 185–210, 1998.
- [76] P. Stoica and K. Sharman, “Maximum likelihood methods for direction-of-arrival estimation,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 7, pp. 1132–1143, 1990.
- [77] B. Ottersten, M. Viberg, P. Stoica, and A. Nehorai, “Exact and large sample ML techniques for parameter estimation and detection in array processing,” in *Radar Array Processing*, Haykin, Litva, and Shepherd, Eds. Berlin: Springer-Verlag, 1993, pp. 99–151.

REFERENCES

- [78] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, 1986.
- [79] T. Abatzoglou, "A fast maximum likelihood algorithm for frequency estimation of a sinusoid based on newton's method," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 1, pp. 77–89, 1985.
- [80] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," *Signal Processing, IEEE Transactions on*, vol. 44, no. 2, pp. 281–295, 1996.
- [81] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *Signal Processing, IEEE Transactions on*, vol. 42, no. 10, pp. 2664–2677, 1994.
- [82] P. J. Chung and J. F. Böhme, "Comparative convergence analysis of em and sage algorithms in doa estimation," *Signal Processing, IEEE Transactions on*, vol. 49, no. 12, pp. 2940–2949, 2001.
- [83] F. Dellaert, "The expectation maximization algorithm," *Georgia Institute of Technology, Technical Report Number GIT-GVU-02-20*, 2002.
- [84] S. Konishi and G. Kitagawa, *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.
- [85] P. Stoica, Y. Selén, and J. Li, "On information criteria and the generalized likelihood ratio test of model order selection," *Signal Processing Letters, IEEE*, vol. 11, no. 10, pp. 794–797, 2004.
- [86] J. C. Chen, K. Yao, and R. E. Hudson, "Source localization and beamforming," *Signal Processing Magazine, IEEE*, vol. 19, no. 2, pp. 30–39, 2002.
- [87] J. Li and P. Stoica, *Robust adaptive beamforming*. Wiley Online Library, 2006.
- [88] N. Wagner, Y. C. Eldar, and Z. Friedman, "Compressed beamforming in ultrasound imaging," *Signal Processing, IEEE Transactions on*, vol. 60, no. 9, pp. 4643–4657, 2012.
- [89] M. Hawkes and A. Nehorai, "Acoustic vector-sensor beamforming and capon direction estimation," *Signal Processing, IEEE Transactions on*, vol. 46, no. 9, pp. 2291–2304, 1998.

- [90] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *Signal Processing, IEEE Transactions on*, vol. 51, no. 2, pp. 313–324, 2003.
- [91] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *Antennas and Propagation, IEEE Transactions on*, vol. 30, no. 1, pp. 27–34, 1982.
- [92] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.
- [93] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *Signal Processing, IEEE Transactions on*, vol. 51, no. 7, pp. 1702–1715, 2003.
- [94] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of statistics*, pp. 416–431, 1983.
- [95] Z. Lu and A. Zoubir, "Source enumeration in array processing using a two-step test," *IEEE Transactions on Signal Processing*, 2015.
- [96] T. Soderstrom, "On model structure testing in system identification," *International Journal of Control*, vol. 26, no. 1, pp. 1–18, 1977.
- [97] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *Signal Processing Magazine, IEEE*, vol. 21, no. 4, pp. 36–47, 2004.
- [98] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *Proc. 2nd Int. Symp. IriJbrn. Theory*, pp. 267–281, 1973.
- [99] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [100] M. Viberg, B. Ottersten, and A. Nehorai, "Performance analysis of direction finding with large arrays and finite data," *Signal Processing, IEEE Transactions on*, vol. 43, no. 2, pp. 469–477, 1995.
- [101] P. Billingsley, *Probability and measure*. John Wiley & Sons, 2008.
- [102] P. Stoica and N. Arye, "Music, maximum likelihood, and cramer-rao bound," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 5, pp. 720–741, 1989.

REFERENCES

- [103] J. F. Sturm, “Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones,” *Optimization methods and software*, vol. 11, no. 1-4, pp. 625–653, 1999.
- [104] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, “Sdpt3: a matlab software package for semidefinite programming, version 1.3,” *Optimization methods and software*, vol. 11, no. 1-4, pp. 545–581, 1999.
- [105] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [106] M. R. Osborne, B. Presnell, and B. A. Turlach, “On the lasso and its dual,” *Journal of Computational and Graphical statistics*, vol. 9, no. 2, pp. 319–337, 2000.
- [107] S. Wright, R. Nowak, and M. Figueiredo, “Sparse reconstruction by separable approximation,” *IEEE Trans. Signal Processing*, vol. 57, no. 7, pp. 2479–2493, July 2009.
- [108] M. A. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 586–597, 2007.
- [109] Y. Nesterov, “A method of solving a convex programming problem with convergence rate $o(1/k^2)$,” in *Soviet Mathematics Doklady*, vol. 27, no. 2, 1983, pp. 372–376.
- [110] Y. Nesterov, A. Nemirovskii, and Y. Ye, *Interior-point polynomial algorithms in convex programming*. SIAM, 1994, vol. 13.
- [111] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.
- [112] M. A. Maleki, *Approximate message passing algorithms for compressed sensing*. Stanford University, 2010.
- [113] M. Bayati and A. Montanari, “The dynamics of message passing on dense graphs, with applications to compressed sensing,” *Information Theory, IEEE Transactions on*, vol. 57, no. 2, pp. 764–785, 2011.
- [114] S. Som and P. Schniter, “Compressive imaging using approximate message passing and a markov-tree prior,” *Signal Processing, IEEE Transactions on*, vol. 60, no. 7, pp. 3439–3448, 2012.

- [115] P. Stoica and P. Babu, “Spice and likes: Two hyperparameter-free methods for sparse-parameter estimation,” *Signal Processing*, vol. 92, no. 7, pp. 1580–1590, 2012.
- [116] P. Boufounos, M. F. Duarte, and R. G. Baraniuk, “Sparse signal reconstruction from noisy compressive measurements using cross validation,” in *Statistical Signal Processing, 2007. SSP’07. IEEE/SP 14th Workshop on*. IEEE, 2007, pp. 299–303.
- [117] T. Park and G. Casella, “The bayesian lasso,” *J. Amer. Stat. Assoc.*, vol. 103, pp. 681–686, 2008.
- [118] A. Panahi and M. Viberg, “Maximum aposteriory based regularization parameter selection,” in *IEEE Int. Conf. Acoust. Speech, Signal Processing*, 2011.
- [119] C. Thrampoulidis, A. Panahi, D. Guo, and B. Hassibi, “Precise error analysis of the ℓ_2 -lasso,” *arXiv preprint arXiv:1502.04977*, 2015.
- [120] D. L. Donoho and M. Elad, “Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.
- [121] B. Recht, W. Xu, and B. Hassibi, “Null space conditions and thresholds for rank minimization,” *Mathematical programming*, vol. 127, no. 1, pp. 175–202, 2011.
- [122] Z. Ben-Haim, Y. C. Eldar, and M. Elad, “Coherence-based performance guarantees for estimating a sparse vector under random noise,” *Signal Processing, IEEE Transactions on*, vol. 58, no. 10, pp. 5030–5043, 2010.
- [123] E. J. Candès, “The restricted isometry property and its implications for compressed sensing,” *Comptes Rendus Mathématique*, vol. 346, no. 9, pp. 589–592, 2008.
- [124] J. D. Blanchard, C. Cartis, and J. Tanner, “Compressed sensing: How sharp is the restricted isometry property?” *SIAM review*, vol. 53, no. 1, pp. 105–125, 2011.
- [125] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, “A simple proof of the restricted isometry property for random matrices,” *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

REFERENCES

- [126] M. Bayati and A. Montanari, “The lasso risk for gaussian matrices,” *Information Theory, IEEE Transactions on*, vol. 58, no. 4, pp. 1997–2017, 2012.
- [127] D. L. Donoho, A. Maleki, and A. Montanari, “The noise-sensitivity phase transition in compressed sensing,” *Information Theory, IEEE Transactions on*, vol. 57, no. 10, pp. 6920–6941, 2011.
- [128] C. Thrampoulidis, A. Panahi, and B. Hassibi, “Asymptotically exact error analysis for the generalized ℓ_2^2 -lasso,” *arXiv preprint arXiv:1502.06287*, 2015.
- [129] S. B. Korada and A. Montanari, “Applications of the lindeberg principle in communications and statistical learning,” *Information Theory, IEEE Transactions on*, vol. 57, no. 4, pp. 2440–2450, 2011.
- [130] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.
- [131] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Trans. Inform. Theory*, vol. 53, pp. 4655–4666, Dec. 2007.
- [132] E. Candes and T. Tao, “The dantzig selector: statistical estimation when p is much larger than n ,” *The Annals of Statistics*, pp. 2313–2351, 2007.
- [133] D. Needell and J. A. Tropp, “Cosamp: Iterative signal recovery from incomplete and inaccurate samples,” *Elsevier, Appl. Comput. Harmon. Anal.*, vol. 26, pp. 301–321, May 2009.
- [134] P. Stoica, J. Li, and J. Ling, “Missing data recovery via a nonparametric iterative adaptive approach,” *Signal Processing Letters, IEEE*, vol. 16, no. 4, pp. 241–244, 2009.
- [135] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer, “Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares,” *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 46, no. 1, pp. 425–443, 2010.
- [136] 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,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 1, pp. 35–47, 2011.

- [137] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted l_1 minimization,” *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.
- [138] I. F. Gorodnitsky and B. D. Rao, “Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm,” *Signal Processing, IEEE Transactions on*, vol. 45, no. 3, pp. 600–616, 1997.
- [139] M. Salman Asif and J. Romberg, “Dynamic updating for minimization,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 2, pp. 421–434, 2010.
- [140] J. Jin, Y. Gu, and S. Mei, “A stochastic gradient approach on compressive sensing signal reconstruction based on adaptive filtering framework,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 2, pp. 409–420, 2010.
- [141] M. Lustig, J. M. Santos, D. L. Donoho, and J. M. Pauly, “kt sparse: High frame rate dynamic mri exploiting spatio-temporal sparsity,” in *Proceedings of the 13th Annual Meeting of ISMRM, Seattle*, vol. 2420, 2006.
- [142] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 1998, vol. 31.
- [143] A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967.
- [144] R. Bellman, “Dynamic programming and lagrange multipliers,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 42, no. 10, p. 767, 1956.
- [145] Y.-C. Ho and R. Lee, “A bayesian approach to problems in stochastic estimation and control,” *Automatic Control, IEEE Transactions on*, vol. 9, no. 4, pp. 333–339, 1964.
- [146] N. Bergman, “Recursive bayesian estimation,” *Department of Electrical Engineering, Linköping University, Linköping Studies in Science and Technology. Doctoral dissertation*, vol. 579, 1999.
- [147] R. P. Mahler, “Multitarget bayes filtering via first-order multitarget moments,” *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 39, no. 4, pp. 1152–1178, 2003.

REFERENCES

- [148] S. S. Blackman, “Multiple-target tracking with radar applications,” *Dedham, MA, Artech House, Inc., 1986, 463 p.*, vol. 1, 1986.
- [149] Y. Bar-Shalom, P. K. Willett, and X. Tian, “Tracking and data fusion,” *A Handbook of Algorithms. Yaakov Bar-Shalom*, 2011.
- [150] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Fluids Engineering*, vol. 82, no. 1, pp. 35–45, 1960.
- [151] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.
- [152] S. J. Julier and J. K. Uhlmann, “New extension of the kalman filter to nonlinear systems,” in *AeroSense’97*. International Society for Optics and Photonics, 1997, pp. 182–193.

Part II

Included Papers

Paper 1

A Numerical Approach to Gridless Compressed Sensing

A. Panahi, M. Viberg and B. Hassibi

To be Submitted to IEEE transactions on signal processing

A Numerical Approach to Gridless Compressed Sensing

A. Panahi, M. Viberg and B. Hassibi

Abstract

The advent of sparsity based techniques has provided a new potential in parameter estimation scenarios, where the problem can be viewed as atomic decomposition. This includes a large range of applications, but the performance of these techniques is limited by the structured model mismatch, induced by discretization, which is also known as the off-grid effect. The atomic norm denoising approach provides an opportunity to apply the sparsity-based techniques without the off-grid effect. However, there is no general numerical solution to ANDN. In this work, we propose a numerical method to atomic norm denoising. We show by numerical studies that the proposed technique is fast and the resulting estimates reflect the desired properties of the sparsity-based techniques. We also introduce a heuristic combination technique for the traditional grid-based technique and identify cases where the combining technique may outperform the exact ANDN approach.

1 Introduction

The advent of sparsity-based techniques has led to new opportunities for numerically stable calculations in certain type of estimation problems. It is observed that many problems of interest, commonly referred to as Atomic Decomposition, can be represented by sparse linear models and solved by non-linear estimators, with guaranteed performance [1–5]. This has not only influenced the field of parametric inference, but has also revolutionized the insight into data acquisition. Accordingly, the field of Compressed Sensing (CS) has been devoted to study reliable acquisition techniques, associated with later desired sparsity-based data processes [6, 7].

Although the pioneering studies on CS have provided nearly optimal schemes with numerically reliable methods [8–10], difficulty arises when high precision in parametric models is desired. This is due to the fundamental inconsistency in the application of sparsity techniques to continuous models, leading to the so-called off grid problem [11, 12]. According to the

description of the off-grid problem, the sparsity based model requires discretization of the parameter space. While the average estimation precision is proportional to the level of discretization, the numerical complexity grows polynomially with it. From this perspective, the current CS techniques are remarkably slower than the other existing methods. This paper proposes a general method to elevate the computational speed and overcome the off-grid problem for high precision CS problems, maintaining the guaranteed performance of the previous techniques.

Our method is inspired by the Least Absolute Shrinkage and Selection Operator (LASSO), a central object in the analysis and the application of CS [1, 13]. The LASSO, estimates a sparse vector through optimizing a statistically derived convex objective function. This guarantees good performance and interesting numerical properties through convex optimization [14, 15]. Moreover, the special structure of LASSO allows to apply specific numerical methods to solve it. Familiar examplers are the iterative thresholding technique [16, 17], the shooting algorithm [18], SParse Iterative Covariance-based Estimation (SPICE) [19]. Nevertheless, the LASSO needs discretization, when applied to continuous models, thus leading to the off-grid effect. This usually means that an off-grid parameter is estimated by multiple nearby grid points. In general, it is difficult to distinguish and combine theses estimates, and achieve a consistent one.

The off grid problem has been considered and discussed by different previous studies. For example, [20, 21] alleviate the off grid effect by including the sample derivatives of the manifold in the dictionary. However, multiple estimates are still observed due to the higher order effect of the manifold and the discrete nature of the method. The work in [22] considers a similar approach, formulated in a total-Least-square setup. A different type of solution is considered by interpreting the LASSO estimates as a result of quantizing a set of continuous ones [23]. The iterative re-sampling scheme discussed in [24] can be regarded as such. The exact continuous estimation procedure is also formulated in [25], where the authors introduce the idea of Atomic Norm DeNoising (ANDN) and discuss its relationship with LASSO. They proved similar properties to that of the LASSO estimates for the continuous estimates of ANDN. In fact, it is shown in [26] that the two are identical under any desired precision for sufficiently fine discretization. In particular, this implies that ANDN has a guaranteed performance under a resolution limit.

The work in [27] also points out to the numerical properties of ANDN, but only provides an implementation for a special case of frequency estimation, where the specific algebraic structure of the manifold allows particular manipulations. However, this method cannot be generalized to the com-

mon CS cases, where very few assumptions on the manifold can be made. Moreover, the resulting algorithm is often slow. Some other limited implementations are also considered in [23]. The general implementation of ANDN is not addressed yet. Accordingly, we hereby propose a general technique to numerically solve ANDN. The method is fast in a large variety of CS scenarios and does not depend on the structure of the manifold. However, special structure or low dimension of the manifold may further enhance its speed. Our approach is to first show that ANDN has a non-convex parametric representation and study the Cyclic Coordinate Descent (CCD) technique to numerically solve the latter. Then, appealing to the well-known global optimality condition for ANDN, we seek for a correcting step in CCD, such that the stopping point of the overall algorithm fulfills the condition and thus, guaranteeing global convergence. We compare different modifying steps from the perspective of convergence speed. We also apply our technique to different atomic decomposition setups and CS scenarios, which demonstrates the superiority of our technique both by speed and by precision.

1.1 Problem Formulation

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

$$\mathbf{x} = \mathbf{b}_1 s_1 + \mathbf{b}_2 s_2 + \dots + \mathbf{b}_n s_n + \mathbf{n}, \quad (1)$$

where $s_k \in \mathbb{C}$ and \mathbf{b}_k belong to the dictionary \mathcal{B} 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 s_k when the order n is unknown. This is called the Atomic Decomposition (AD) problem [1, 27]. It is convenient to write (1) as

$$\mathbf{x} = \mathbf{b}_1 e^{j\phi_1} r_1 + \mathbf{b}_2 e^{j\phi_2} r_2 + \dots + \mathbf{b}_n e^{j\phi_n} r_n + \mathbf{n}, \quad (2)$$

where (ϕ_k, r_k) is the polar representation of the complex number s_k , and define $\mathbf{a}_k = \mathbf{b}_k e^{j\phi_k}$ to obtain

$$\mathbf{x} = \mathbf{a}_1 r_1 + \mathbf{a}_2 r_2 + \dots + \mathbf{a}_n r_n + \mathbf{n}, \quad (3)$$

Then, we define

$$\mathcal{A} = \{\mathbf{b} e^{j\phi} \mid \mathbf{b} \in \mathcal{B}, 0 \leq \phi < 2\pi\}. \quad (4)$$

Clearly finding a desirable representation in (3) with $\mathbf{a}_k \in \mathcal{A}$, and $r_k > 0$ leads to a unique representation in (1) by recombining the phase part of \mathbf{a}_k to r_k to obtain s_k .

2 Examples of Atomic Decomposition

The AD problem in (1) has been long addressed and discussed in different contexts. An interested reader is referred to [1, 28–30] for more details. Here, we introduce some well-known examples. We introduce the corresponding models in the sense of (1), but remind that the same problem can be written as (3) using the dictionary transform given by (4).

2.1 Spectral Line Estimation

In spectral line estimation, the data \mathbf{x} is obtained by sampling a continuous signal at instants t_1, t_2, \dots, t_m and the dictionary in the form of (1) is given by

$$\mathcal{B} = \{\mathbf{b}(\omega) = [e^{j\omega t_1} \ e^{j\omega t_2} \ \dots \ e^{j\omega t_m}]^T \mid -W \leq \omega \leq W\}. \quad (5)$$

The parameter W is the system bandwidth. Each basis vector $\mathbf{b}(\omega)$ in the dictionary consists of the samples of a sinusoidal signal at frequency ω . Solving the AD problem for the spectral line estimation model leads to the best signal fit by a finite number of sinusoids. The direct advantage is that if $m \gg n$, the signal can be stored with less effort, leading to a large amount of compression. With some assumptions on the underlying signal and a proper choice of the sampling times t_k , it is shown that the continuous signal may be completely acquired by the above scheme of sampling and atomic decomposition, storing only the resulting frequencies and corresponding amplitudes s_k . This is called compressed sensing.

2.2 Far-field Narrow-Band Direction-of-Arrival Estimation

The Direction-Of-Arrival (DOA) estimation problem is related to a scenario, where a physical field (e.g. electromagnetic, or sonar waves) is generated by n sources, each corresponding to a narrowband signal. For a short time interval, a narrowband signal can be expressed by its frequency and its complex amplitude, the later of which representing the magnitude and the phase shift in the signal. The field is locally sensed by a set of m sensors, resulting in a vector of observations $\mathbf{x} = [x_1, x_2, \dots, x_m]^T$. The sources are assumed to be far from the vicinity of the sensors, so that only their relative angles, i.e. DOAs, can be estimated from data. Taking a local coordinate system at the position of the sensors and assuming sources with equal and known frequency, the relationship between the data vector \mathbf{x} and the sources

is given by (1), where s_k is the complex amplitude of the k^{th} source and

$$\mathbf{b}_k = \mathbf{b}(\theta_k) = [e^{j2\pi\rho_1 \cos(\theta_k - \eta_1)} \ e^{j2\pi\rho_2 \cos(\theta_k - \eta_2)} \ \dots \ e^{j2\pi\rho_m \cos(\theta_k - \eta_m)}]^T \quad (6)$$

Moreover, θ_k is the angle coordinate (DOA) of the k^{th} source and (ρ_l, η_l) are the total polar coordinate of the l^{th} sensor in units of wavelength (for the first coordinate) in the selected coordinate system. Solving the AD problem for this case gives the DOA estimates. Note that when the sensors are linearly placed and uniformly separated with a half-wavelength, the coordinate system can be chosen, such that $\eta_k = 0$ and $\rho_k = (k - 1)/2$. In this case, the DOA estimation model is identical to the frequency estimation model by defining $\omega = \pi \cos(\theta)$ and $W = \pi$.

2.3 Narrow-band Radar Delay-Doppler Estimation

Although a large variety of radar detection scenarios can be modeled as AD, we consider one of the simplest models, where a single continuous narrow-band pulse $\xi(t) = \psi(t) \exp j\omega_0 t$ is transmitted, where $\psi(t)$ is a baseband signal and ω_0 is a large carrier frequency. The signal is reflected by multiple targets and is received by a sensor. The sample vector \mathbf{x} of the received signal at times t_1, t_2, \dots, t_m can be written as the model in (1), where s_k represents the reflection coefficient of the k^{th} sensor and

$$\mathbf{b}_k = \mathbf{b}_k(\mu_k, \tau_k) = [\psi(t_1 - \tau_k)e^{-j\mu_k\omega_0 t_1} \ \psi(t_2 - \tau_k)e^{-j\mu_k\omega_0 t_2} \ \dots \ \psi(t_m - \tau_k)e^{-j\mu_k\omega_0 t_m}]^T \quad (7)$$

where t_k is the time delay associated with the distance (range) of the target and μ_k is the Doppler shift related to its velocity. The AD in this case gives the delay and Doppler (distance and velocity) estimates. If the baseband waveform is constant, i.e. $\psi(t) = 1$, then this AD problem simplifies to the spectral line estimation problem by minor change of variables.

3 Prior Art: Atomic Norm Denoising

The ANDN is closely connected to BP, but does not need to define grid. The exact recovery by ANDN is well studied and many interesting properties has been discovered. The ANDN is based on the transformed model in (3) and

the definition of the atomic norm $\|\cdot\|_{\mathcal{A}}$ associated with the dictionary \mathcal{A} as

$$\begin{aligned} \|\mathbf{y}\|_{\mathcal{A}} = \min_n \min_{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n} \min_{r_1, r_2, \dots, r_n} r_1 + r_2 + \dots + r_n \\ \text{s.t.} \\ \mathbf{y} = \sum_{k=1}^n \mathbf{a}_k r_k \end{aligned} \quad (8)$$

The properties of the atomic norm have also been studied and it is shown by the so-called Caratheodory theorem for convex hulls that (8) attains its minimum at $n \leq 2m$. We also give a simple proof for this in Appendix 7.

The ANDN method is to solve the optimization

$$\min_{\mathbf{y} \in \mathbb{C}^m} \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{y}\|_{\mathcal{A}} \quad (9)$$

where λ is a suitable positive number. We refer to the solution of (9) by $\bar{\mathbf{y}}$. Replacing \mathbf{y} by $\bar{\mathbf{y}}$ in (8), 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. There is no general numerical method to solve (9) and the objective of this work is to provide one. However, if (9) is numerically solved, it is shown (and we also obtain below) that for any \mathbf{a} , the relation $\Re(\mathbf{a}^H(\mathbf{x} - \bar{\mathbf{y}})) \leq \lambda$ holds and $\{\bar{\mathbf{a}}_k\}$ is the set of solutions for \mathbf{a} to $\Re(\mathbf{a}^H(\mathbf{x} - \bar{\mathbf{y}})) = \lambda$.

3.1 Optimality Condition For ANDN

The optimization in (9) is convex. Thus, its Karush Kuhn Tucker (KKT) condition implies global optimality. The cost in (9) is also nondifferentiable. Thus, its KKT condition can be written as

$$\mathbf{x} - \bar{\mathbf{y}} \in \lambda \partial \|\bar{\mathbf{y}}\|_{\mathcal{A}} \quad (10)$$

where we remind that the parameters with bar denote the optimal point in (9) and (8). The subdifferential is given by [27]

$$\partial \|\bar{\mathbf{y}}\|_{\mathcal{A}} = \{\mathbf{z} \mid \sup_{\mathbf{a} \in \mathcal{A}} \Re(\mathbf{z}^H \mathbf{a}) \leq 1, \forall k \Re(\bar{\mathbf{a}}_k^H \mathbf{z}) = 1\} \quad (11)$$

Convexity is not well-defined for (8) due to variable order, but when n is fixed the remaining optimization is non-convex. Let us also rewrite (10) more explicitly as

$$\sup_{\mathbf{a} \in \mathcal{A}} \Re(\bar{\mathbf{z}}^H \mathbf{a}) \leq \lambda, \quad \forall k \Re(\bar{\mathbf{a}}_k^H \bar{\mathbf{z}}) = \lambda \quad (12)$$

where we define

$$\bar{\mathbf{z}} = \mathbf{x} - \bar{\mathbf{y}} = \mathbf{x} - \sum_{k=1}^{\bar{n}} \bar{\mathbf{a}}_k \bar{r}_k \quad (13)$$

The surprising fact about ANDN, phrased as the theorem below, is that (12) fully characterizes (8).

Theorem 2. *any set of parameters $\{\bar{\mathbf{a}}_k, \bar{r}_k\}_{k=1}^{\bar{n}}$ satisfying (12) provides a global optimum of (8) when \mathbf{y} is replaced by $\bar{\mathbf{y}}$.*

Proof. See [27, 31]. □

Note that neither (12) is the KKT condition of (8), nor (8) is convex. Still, (12) provides the global optimality of (8). It is now clear from (12) that, as we previously stated, if $\bar{\mathbf{y}}$ is known (thus so being $\bar{\mathbf{z}}$), the basis estimates are the global maxima of the spectrum $\mathbf{a}^H \bar{\mathbf{z}}$.

3.2 Implementing ANDN for Frequency Estimation

Although there is no general method to implement ANDN, for a special case of frequency estimation, it turns out that the atomic norm has a useful representation as follows, which enables implementation through semidefinite cone programming. Let us first define the so called Toeplitz transform $T : \mathbb{C}^m \rightarrow \mathbb{C}^{m \times m}$

$$T(\mathbf{w}) = \begin{pmatrix} w_1 & w_2 & \dots & w_m \\ w_2^* & w_1 & \dots & w_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ w_m^* & w_{m-1}^* & \dots & w_1 \end{pmatrix} \quad (14)$$

Take the dictionary as in (5) with $t_k = k - 1$ for $k = 1, 2, \dots, m$ and $W = \pi$. This corresponds to the well-know frequency estimation problem with uniform samples. Then, according to [25], we have that

$$\begin{aligned} \|\mathbf{x}\|_{\mathcal{A}} &= \max_{t, \mathbf{w}} \frac{1}{2}(t + w_1) \\ &\text{s.t} \\ &\begin{pmatrix} T(\mathbf{w}) & \mathbf{w} \\ \mathbf{w}^H & t \end{pmatrix} \succeq 0 \end{aligned} \quad (15)$$

where w_1 is the first element in \mathbf{w} and $\succeq 0$ means that the matrix is positive semidefinite. This can be solved by the off-the-shelf algorithms or the ADMM approach in [25].

4 Contribution

In this section we develop an algorithm which exactly solves the ANDN, without relying on the structure of the dictionary. We first substitute the definition of the atomic norm (8) in (9) to obtain.

$$\begin{aligned} \min_{\mathbf{y} \in \mathbb{C}^m} \min_n \min_{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n} \min_{r_1, r_2, \dots, r_n} \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \sum_{k=1}^n r_k \\ \text{s.t.} \\ \mathbf{y} = \sum_{k=1}^n \mathbf{a}_k r_k \end{aligned} \quad (16)$$

which can be simplified to

$$\min_n \min_{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n} \min_{r_1, r_2, \dots, r_n} \frac{1}{2} \|\mathbf{x} - \sum_{k=1}^n \mathbf{a}_k r_k\|_2^2 + \lambda \sum_{k=1}^n r_k \quad (17)$$

Next, note that the order can be fixed to $2m + 1$. To elaborate on this, note that the optimization in (8) always attains its minimum at $n \leq 2m + 1$. If the minimum in (8) occurs at a lower order n than $2m + 1$, say at r_1, r_2, \dots, r_n with some corresponding bases, the optimization with exactly $2m + 1$ elements can also attain the same minimum by defining $r_{n+1} = r_{n+2} = \dots = r_{2m+1} = 0$.

We obtain the following non convex optimization problem, whose global optimum, by the previous discussion, coincides with ANDN.

$$\min_{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{2m+1}} \min_{r_1, r_2, \dots, r_{2m+1}} \frac{1}{2} \|\mathbf{x} - \sum_{k=1}^{2m+1} \mathbf{a}_k r_k\|_2^2 + \lambda \sum_{k=1}^{2m+1} r_k \quad (18)$$

Notice that the condition in (12) is a necessary and sufficient for optimality in ANDN. Hence, the fact that the global optimum of (18) coincides with the ANDN implies that (12) is exactly equivalent to global optimality in (18). In short, we obtain a non-convex, for which we derived a non-KKT equivalent global optimality condition. Although this does not provide a numerical method, guaranteed to globally converge, it instead provides means to assess global optimality by verifying the condition at the possible stopping points of a numerical algorithm under question.

4.1 Cyclic Coordinate Descent Algorithm

We now examine the local search methods on (18). We focus on the Cyclic Coordinate Descent (CCD) [32]. The direct application of CCD to BP in (29) is called the shooting method [18]. The CCD method consists of iteration cycles, where each pair of (\mathbf{a}_l, r_l) is updated at an individual iteration

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 (19).

end for
until Convergence

by solving (18), keeping the others constant. In this case, direct calculation yields to the following closed form solution

$$\hat{\mathbf{a}}_l = \arg \max_{\mathbf{a} \in \mathcal{A}} \Re(\mathbf{a}^H \mathbf{z}^{(l)}), \quad \hat{r}_l = \frac{(\Re(\hat{\mathbf{a}}_l^H \mathbf{z}^{(l)}) - \lambda)_+}{\|\hat{\mathbf{a}}_l\|_2^2} \quad (19)$$

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

$$\mathbf{z}^{(l)} = \mathbf{x} - \sum_{k \neq l} \mathbf{a}_k r_k \quad (20)$$

The CCD algorithm is summarized in Algorithm 1.

Each iteration of CCD decreases the cost in (18). 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 (12). In this manner, the following theorem can be obtained.

Theorem 3. *The stopping point of CCD consisting of the parameters $\hat{r}_1, \hat{r}_2, \dots, \hat{r}_{2m+1}$ and $\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \dots, \hat{\mathbf{a}}_{2m+1}$ satisfies (12), 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 (18), which implies that

$$\begin{cases} \Re(\hat{\mathbf{a}}_k^H \hat{\mathbf{z}}) = 1 & \hat{r}_k > 0 \\ \Re(\hat{\mathbf{a}}_k^H \hat{\mathbf{z}}) \leq 1 & \hat{r}_k = 0 \end{cases} \quad (21)$$

where

$$\hat{\mathbf{z}} = \mathbf{x} - \sum_{k=1}^n \hat{\mathbf{a}}_k \hat{r}_k \quad (22)$$

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 (19) at the convergence point does not change the pair, which implies that

$$\max_{\mathbf{a} \in \mathcal{A}} \Re(\mathbf{a}^H \hat{\mathbf{z}}) = \max_{\mathbf{a} \in \mathcal{A}} \Re(\mathbf{a}^H \hat{\mathbf{z}}^{(l)}) \leq \lambda \quad (23)$$

and together with (21) shows that (10) holds. \square

4.2 Correcting CCD

We cannot generally establish global optimality of the CCD algorithm, except when the stopping point contains an inactive (zero) 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 (18), to guarantee convergence. In practice, we apply this step after a fixed number L of CCD cycles.

Linear Extremizer

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 (18) 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

$$\sum_{k=1}^n u_k \mathbf{a}_k = 0 \quad \sum_{k=1}^n u_k \geq 0 \quad (24)$$

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

$$\hat{r}_l = r_l - \alpha u_l \quad (25)$$

where

$$\alpha = \min_{l|u_l \geq 0} \frac{u_l}{r_l} \quad (26)$$

First, note that the value of α in (26) ensures that the resulting elements \hat{r}_l in (25) are non-negative. Second, assuming that the minimum is achieved at index l_0 , we have that

$$\alpha = \frac{u_{l_0}}{r_{l_0}} \rightarrow \hat{r}_{l_0} = u_{l_0} - \alpha r_{l_0} = 0 \quad (27)$$

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 (12) holds. Thus, the stopping point is a global optimal point of ANDN.

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 (25).

until Convergence

Algorithm 3 The boosted 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 $\{\mathbf{a}_k, r_k\}$.

Writing $\{\mathbf{a}_k = e^{j\phi_k} \mathbf{b}_k\}$, fix \mathbf{b}_k and solve (28). Denoting the minimum by \hat{s}_k , replace $\{r_k\}$ and $\{e^{j\phi_k}\}$ by $\{|\hat{s}_k|\}$ and $\{s_k/|s_k|\}$, respectively, the latter of which only applies to nonzero \hat{s}_k .

Apply linear extremizer by fixing $\{\mathbf{a}_k\}$ and replacing $\{r_k\}$ by $\{\hat{r}_k\}$ given in (25).

until Convergence

A Boosting Step

The linear extremizer, introduced in Section 4.2 guarantees global convergence, but does not reduce the cost significantly. In a case, where the model is originally transformed from (1) to (3) and to enhance the convergence rate, we can also include the following step before applying the linear extremizer. Note that due to (4), the parameters $\{\mathbf{a}_k, r_k\}$ at a certain step, are equivalent to the expanded set $\{\mathbf{b}_k, e^{j\phi_k}, r_k\}$. We may fix \mathbf{b}_k and minimize over $\{e^{j\phi_k}, r_k\}$, which can also be recombined into $s_k = r_k e^{j\phi_k}$. Noticing that $r_k = |s_k|$, the optimization in (18) with fixed $\{\mathbf{b}_k\}$ may be written as

$$\min_{s_1, s_2, \dots, s_{2m+1}} \frac{1}{2} \left\| \mathbf{x} - \sum_{k=1}^{2m+1} \mathbf{b}_k s_k \right\|_2^2 + \lambda \sum_{k=1}^{2m+1} |s_k| \quad (28)$$

It is immediately observed that the optimization in (28) is identical to (29) when $\tilde{\mathbf{B}}$ is replaced by $\{\mathbf{b}_k\}$. Thus, it can be solved by ISTA, SPICE or any other technique. The boosted algorithm is summarized in Algorithm 3.

Note that since solving (28) is complex, it should not be frequently used in order to achieve the best convergence speed. Thus, selecting the correct value for L is crucial.

4.3 How to Select the Cycle

Another issue with the above algorithm is that the desired order of the ANDN solution is often small, while the parameter space of the algorithm is of dimension $2m + 1$, which is usually large. In practice, this leads to a substantial speed reduction. Note that the order $2m + 1$ is necessary to apply the CCD correction step, without which the algorithm may potentially stop at a local minima. A solution is to alter the order in the CCD cycle to emphasis on the sparse nature of the set of optimization parameters. In the begining, the parameters r_k are sorted in the descending order. Then, the the CCD cycle in Algorithm 1 is applied only over the first two elements (r_1, \mathbf{a}_1 and r_2, \mathbf{a}_2). Once this stage stops by convergence, the third element is added and Algorithm 1 runs again with three elements. This procedure continues until all $2m + 1$ are involved. As seen, this methods gives priority to the largest entries, thus promoting sparsity.

5 Numerical Results and Comparisson to Related Works

5.1 Related Works: Basis Pursuit

The exact solution of AD is generally intractable for large m or n , and the previous techniques do not guarantee such an exact result. There is a variety of different techniques to solve AD, which an interested reader may find in [28]. We focus here on two recent sparsity approaches called Basis Pursuit (BP), also known as Least Absolute Shrinkage and Selection Operator (LASSO), and Atomic Norm DeNoising (ANDN). The BP method is to select a large, but finite subset of \mathcal{B} , say $\tilde{\mathcal{B}} = \{\tilde{\mathbf{b}}_1, \dots, \tilde{\mathbf{b}}_N\}$, and apply the following optimization

$$\min_{\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_N} \frac{1}{2} \left\| \mathbf{x} - \sum_k \tilde{s}_k \tilde{\mathbf{b}}_k \right\|_2^2 + \lambda \sum_k |\tilde{s}_k| \quad (29)$$

where $\lambda > 0$ is a design parameter. The exact solution of (29) is sparse, i.e. it includes few nonzero elements \tilde{s}_k . The best representing bases in (1) are then given by the bases $\tilde{\mathbf{b}}_k$ corresponding to nonzero elements \tilde{s}_k .

Implementing LASSO

Although BP is convex and can be solved by the off-the-shelf convex optimization techniques, there are faster implementations relying on its special structure. In the sequel, we discuss two methods, which can be also applied

to complex-valued models of interest herein. The first method is called Iterative Shrinkage-Thresholding Algorithm (ISTA). At the r^{th} iteration, the ISTA locally approximates the BP cost by a so-called proximal function around the vector $\tilde{\mathbf{s}}^{(r-1)} = [\tilde{s}_1^{(r-1)}, \tilde{s}_2^{(r-1)}, \dots, \tilde{s}_N^{(r-1)}]^T$, calculated from the previous iterations. By minimizing the proximal function, it updates $\tilde{\mathbf{s}}^{(r-1)}$ to the minimum point of the proximal cost. Without stating more details, the resulting update step is given by

$$\tilde{\mathbf{s}}^{(r)} = \mathcal{T}_{\lambda\eta}(\tilde{\mathbf{s}}^{(r-1)} + \eta\mathbf{B}^H(\mathbf{x} - \mathbf{B}\tilde{\mathbf{s}}^{(r-1)})) \quad (30)$$

where η is a suitable step length, $\mathbf{B} = [\tilde{\mathbf{b}}_1 \tilde{\mathbf{b}}_2 \dots \tilde{\mathbf{b}}_N]$ and $\mathcal{T}_\alpha(\cdot)$ denotes the elementwise application of the following so called shrinkage operator

$$\mathcal{T}_\alpha(re^{j\phi}) = (r - \alpha)_+ e^{j\phi} \quad (31)$$

Although it is theoretically shown that the method can be arbitrarily slow in special cases, it has been so-far successful in practice. Some modifications to ISTA has also been proposed, but the improvement is normally limited [33].

The second technique is called SParse Iterative Covariance-based Estimation (SPICE) method. It is based on the following identity

$$|\tilde{s}_k| = \frac{1}{2} \min_{\tilde{p}_k > 0} \frac{|\tilde{s}_k|^2}{\tilde{p}_k} + \tilde{p}_k \quad (32)$$

Substituting (32) in (29), the SPICE method alternatively minimizes the resulting cost function with respect to $\{\tilde{s}_k\}$ and $\{\tilde{p}_k\}$ parameters. This leads to the following iteration

$$\begin{aligned} \tilde{s}_k^{(r)} &= \tilde{p}_k^{(r-1)} \tilde{\mathbf{b}}_k^H \left(\lambda \mathbf{I} + \sum_l \tilde{p}_l^{(r-1)} \tilde{\mathbf{b}}_l \tilde{\mathbf{b}}_l^H \right)^{-1} \mathbf{x} \\ \tilde{p}_k^{(r)} &= |\tilde{s}_k^{(r)}| \end{aligned} \quad (33)$$

The SPICE technique is generally fast as the off-support elements exponentially decrease.

The Off-grid Effect

As seen, the BP method selects only from the grid $\tilde{\mathcal{B}}$. If a true atomic decomposition exists, e.g. in the DOA estimation case, and if the true parameters are related to off-grid bases (i.e. the ones out of $\tilde{\mathcal{B}}$), then BP is unable to retrieve the exact model. However, BP most often selects few neighbor on-grid bases of each off-grid one. We call these estimates the associated cloud to the former off-grid parameter. In practice, it is not difficult to cluster the LASSO estimates and distinguish the clouds, since the

resulting spectrum is highly sparse. It is common to combine the elements of a resulting cloud to mitigate the off-grid effect. For illustration, take the DOA examples mentioned in Section 2.2. As estimating bases correspond to finding DOAs, a cloud is represented by a number of 'on-grid DOAs', say $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{n_1}$ and their corresponding estimated amplitudes $\hat{s}_1, \hat{s}_2, \dots, \hat{s}_{n_1}$. This cloud is usually replaced by the following simple weighted average DOA

$$\hat{\theta}' = \frac{\hat{\theta}_1|\hat{s}_1| + \hat{\theta}_2|\hat{s}_2| + \dots + \hat{\theta}_{n_1}|\hat{s}_{n_1}|}{|\hat{s}_1| + |\hat{s}_2| + \dots + |\hat{s}_{n_1}|} \quad (34)$$

with corresponding amplitude $\hat{s}_1 + \hat{s}_2 + \dots + \hat{s}_n$. This simple method has grounds in statistical analysis of the BP for asymptotically fine grid and high SNR. However, it usually needs a high grid size, leading to high complexity. Accordingly, other corrections to BP are considered. As an example, we explain the so-called Sparse Total Least Square (S-TLS) in the sequel. Again, we take the DOA example for simplicity. The keystone in developing S-TLS is to approximate the effect of an off-grid base, say $\mathbf{a}(\theta)$, by Taylor expansion to obtain

$$\mathbf{b}(\theta) \approx \mathbf{b}(\tilde{\theta}_i) + \mathbf{d}(\tilde{\theta}_i)(\theta - \theta_i) \quad (35)$$

where $\mathbf{d}(\theta) = d\mathbf{b}(\theta)/d\theta$ and $\tilde{\theta}_i$ denotes the nearest grid point to θ . Then, the model in (1) can be approximated by

$$\mathbf{x} \approx \sum_k \tilde{s}_{i_k} (\tilde{\mathbf{b}}_{i_k} + \epsilon_k \tilde{\mathbf{d}}_{i_k}) + \mathbf{n} \quad (36)$$

where $\tilde{\mathbf{d}}_{i_k}$ is the corresponding derivative to $\tilde{\mathbf{b}}_{i_k}$ and i_k is the index of the nearest grid point to \mathbf{b}_k . Accordingly, the BP is corrected to

$$\frac{1}{2} \left\| \mathbf{x} - \sum_k \tilde{s}_k (\tilde{\mathbf{b}}_k + \tilde{\epsilon}_k \tilde{\mathbf{d}}_k) \right\|_2^2 + \lambda \sum_k |\tilde{s}_k| + \frac{\mu}{2} \sum_k (\tilde{\epsilon}_k)^2 \quad (37)$$

where the penalty term $\frac{\mu}{2} \sum_k (\tilde{\epsilon}_k)^2$ promotes small deviations in accordance with the approximation in (35). In [22] the value of μ is suggested to be 1, based on a statistical discussion, but since STLS is an approximate technique for DOA estimation μ may vary. The S-TLS is solved by alternatively minimizing for $\{\tilde{s}_k\}$ and $\{\tilde{\epsilon}_k\}$ parameters. Fixing $\{\tilde{\epsilon}_k\}$, the optimization over $\{\tilde{s}_k\}$ becomes a regular BP, which can be solved by the previously mentioned techniques (ISTA, SPICE, etc). The optimization over $\{\tilde{\epsilon}_k\}$ for fixed $\{\tilde{s}_k\}$ is quadratic. Hence, it can be exactly solved.

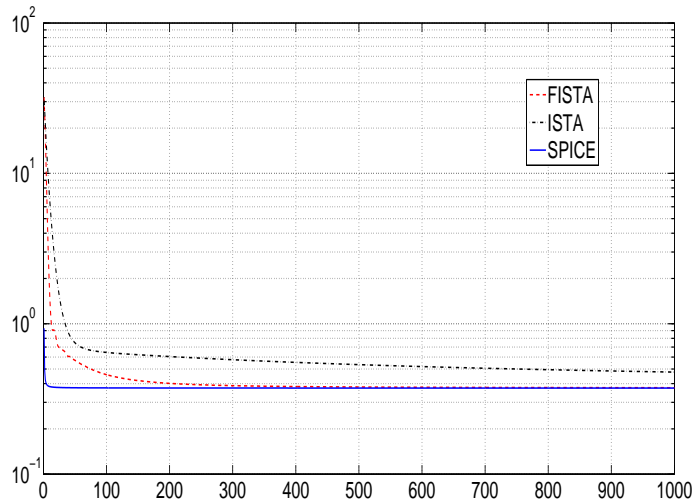


Figure 1: The comparison of the BP cost function at different iterations of ISTA, FISTA and SPICE.

5.2 Numerical Result

We now compare the proposed algorithm to the traditional application of BP with clustering and combining the elements in each cluster. We also consider STLS introduced in Section 5.1. For implementation of BP, we consider SPICE. The reason is shown in Figure 1, where the optimized cost of different techniques, in a typical scenario, are compared over different iterations. Clearly, SPICE has better convergence properties. However, as Figure 2 depicts, the resulting SPICE spectrum after 1000 iterations is still non-sparse. The reason is that nearby grid points to a specific true parameter have highly similar contributions. Thus, it is difficult to distinguish them. We use the process in Section 5.1 to combine the neighbor points in the spectrum. We take the peak points (local maxima) of the resulting spectrum as the centroids of the clouds. Then, each grid point belongs to the cloud with nearest centroid. Once the clouds are formed in this way, the estimates are found by (34). It is surprising to observe that this combining procedure yields to estimates with remarkable statistical properties. In fact, the next results show that the combining procedure can even improve the properties of LASSO (or ANDN) in certain cases. Later, we consider 400 iterations of SPICE.

Frequency Estimation

In this part, we consider the problem of frequency estimation, explained in Section 2.1. We first consider uniform sampling, where ANDN can be also

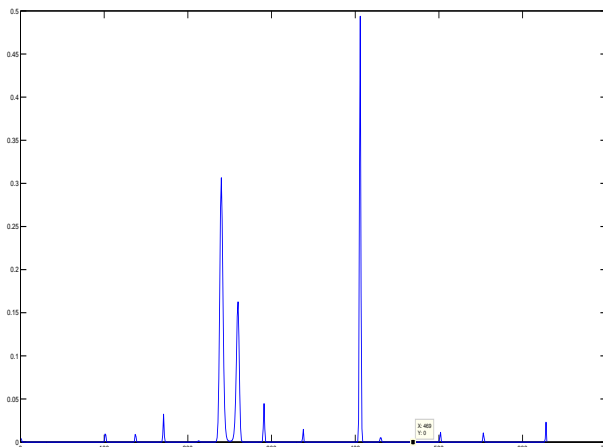


Figure 2: A typical spectrum $|\tilde{s}_k|$ of SPICE after 1000 iterations.

solved by the method explained in Section 3.2. Moreover, we consider the numerical implementation of the latter, explained in [25]. Figure 3 concerns a case, where a noiseless signal, comprising of two frequency components, is sampled at times $t_k = k$ for $k = 1, 2, \dots, m$. For different values of m , the frequencies were fixed to 0 and $4\pi/m$, with amplitudes 1 and $1j$, respectively. This choice maintains a fixed amount of coherence between the basis vectors, corresponding to the components, and Roughly speaking, presents a constant amount of difficulty to the techniques. In this manner, we ensure that the result in Figure 3 merely reflects the numerical difficulty over an increasing dimension. The proposed technique and the SDP approach ran over an equal amount of time defined by 1000 iterations of the SDP approach. Both techniques were set to terminate when the cost value was not improved more than 10^{-4} . As seen in Figure 3, the proposed approach always achieves the termination precision, but the optimality gap increases for the SDP technique. This implies that the SDP technique requires an increasing amount of time and iterations to converge, while the proposed technique converges within the required time period.

We next focused on the effect of noise on the estimates. We compare the proposed technique to SPICE with the combining step and STLS. For STLS in (37), we use $\mu = 100$. Again m uniform samples were considered and m was fixed to 20. The frequencies were fixed to 0 and $4 * \pi/m$ and the amplitudes were fixed to 1 and $1i$, respectively. The variance of noise varied and the performance of the methods was measured by selecting the two largest components as the estimates and assuming the others as false alarm. For SPICE and STLS, we used a uniform grid with separation 0.01

5. NUMERICAL RESULTS AND COMPARISON TO RELATED WORKS

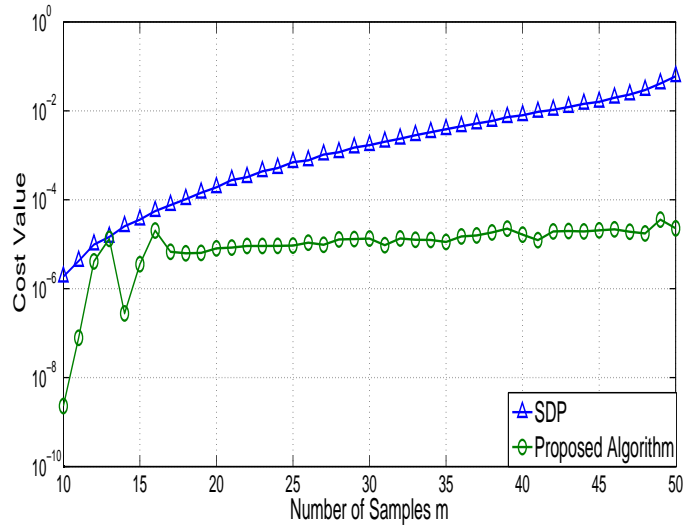


Figure 3: The optimality gap at termination for the proposed method compared to the SDP implementation.

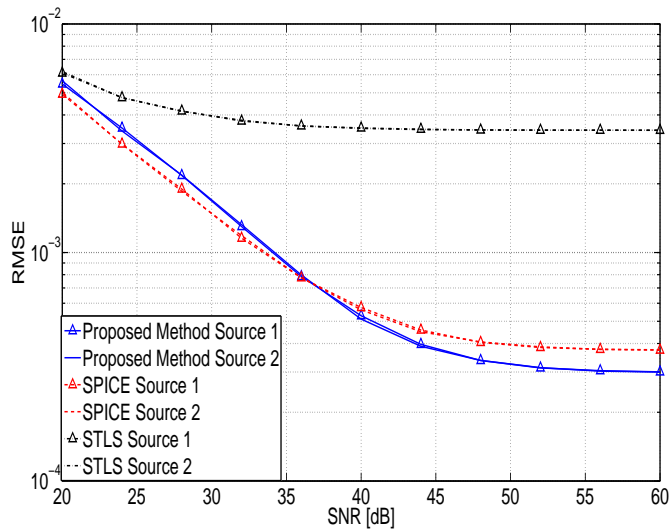


Figure 4: Root mean squared error for different techniques at different SNRs.

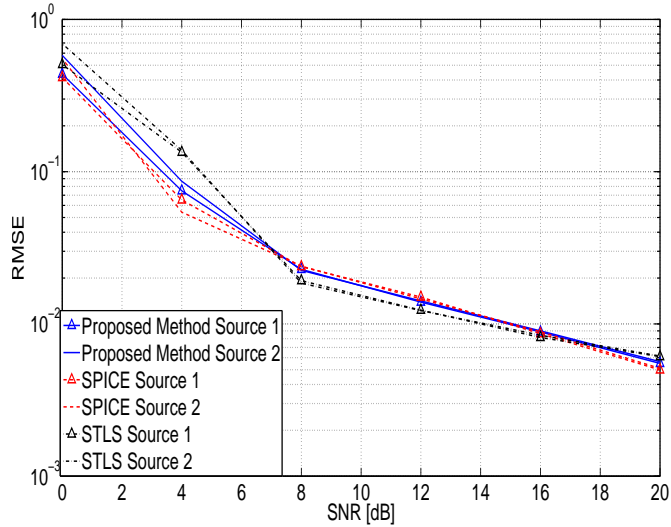


Figure 5: Root mean squared error for different techniques at different SNRs.

in the range of $[-\pi \pi]$. For all methods, $\lambda = 0.1$ is utilized. Figures 4 and 5 show the Root Mean Squared Error (RMSE) in two different ranges of SNR. As expected, the off-grid effect dominates at very high SNR and the proposed method outperforms the other techniques. However, all methods reach an error floor due to the biasing effect of λ . In a medium SNR range (approximately 16 to 35 dB), where noise dominates, SPICE have a better performance. We believe that this is due to the averaging effect of the combination procedure in (34), which suppresses the noise effect. It is interesting to see that at a low SNR range the STLS technique outperforms the other two.

Figures 6 and 7 depict the false alarm rate in the same SNR ranges as in the previous results. As seen, the number of false alarms for ANDN (or BP) grows with SNR, since the regularization parameter, controlling the false alarm rate, is fixed to 0.1. On the contrary, the number of STLS estimates decreases by increasing SNR. It is also interesting to observe that SPICE has a slightly better false alarm properties. This is due to the fact that BP and ANDN do not have a built-in mechanism to limit the source dynamic range, thus their exact optimal point contains very small amplitudes. However, limited numerical resource prevents distinguishing very small amplitudes. In this sense, limiting the number of iterations for SPICE serves as an intrinsic thresholding process which decreases false alarm.

The plots in Figures 8 and 9 show results for a different setup, where a signal with two frequency components 0 and 0.1π are observed by samples uniformly randomly selected in the time interval $[0 10]$. The figures show

5. NUMERICAL RESULTS AND COMPARISON TO RELATED WORKS

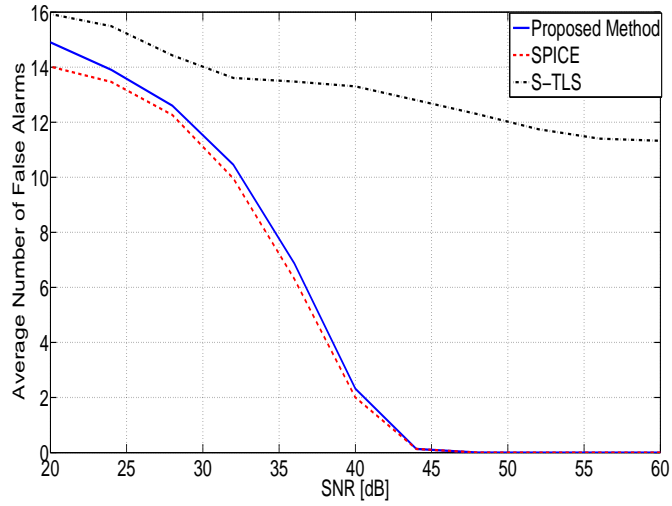


Figure 6: Average false alarm rate for different techniques at different SNRs.

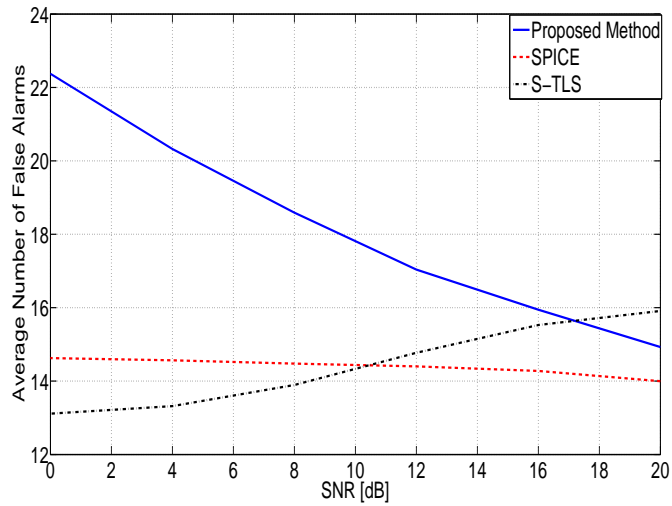


Figure 7: Average false alarm rate for different techniques at different SNRs.

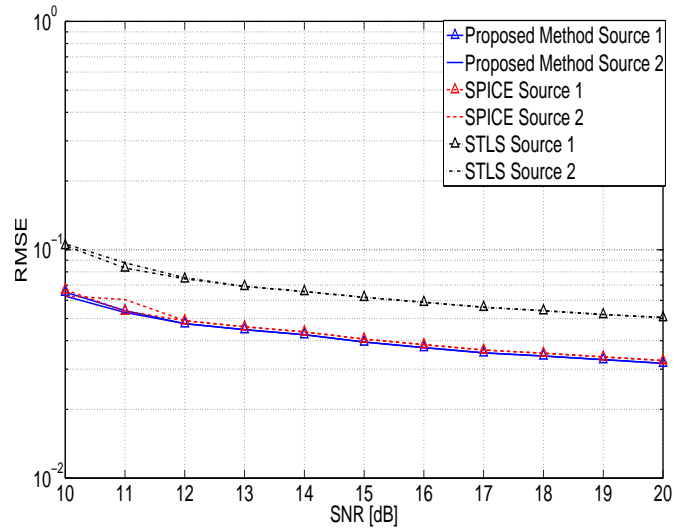


Figure 8: Root mean squared error for different techniques by different number of random samples.

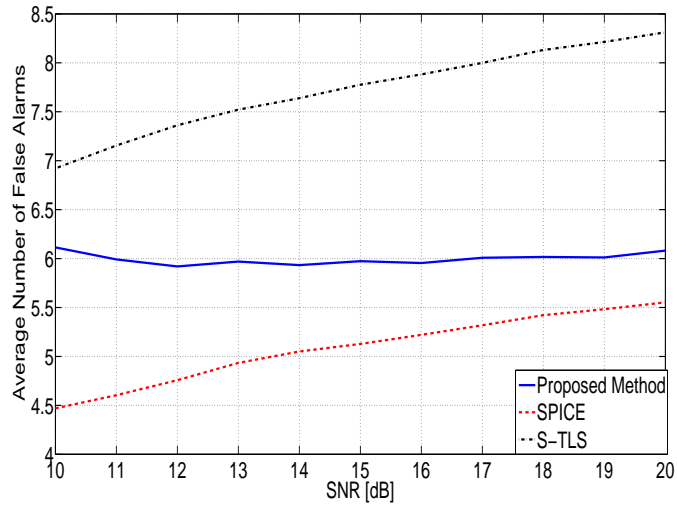


Figure 9: Average false alarm rate for different techniques by different number of random samples.

5. NUMERICAL RESULTS AND COMPARISON TO RELATED WORKS

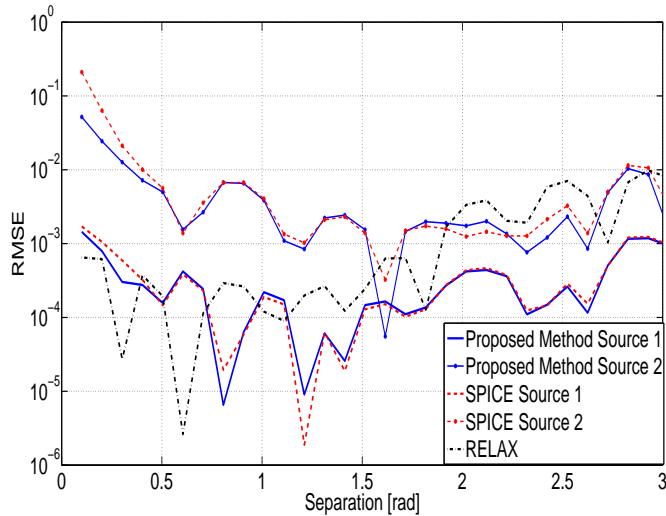


Figure 10: Root mean square error for different techniques versus DOA separation angle.

RMSE and false alarm rate versus variable number of samples. As seen, the proposed method is slightly better than the SPICE method and remarkably superior to STLS in the RMSE sense. However, the SPICE technique again improves the false alarm properties of ANDN, but the gap decreases as the number of observations increase.

DOA Estimation with a Circular Array

Linear arrays are popular in the context of DOA estimation and their corresponding model is equivalent to spectral line estimation, studied in the prequel. On the other hand, the utilization of circular arrays for DOA estimation is emerging in literature. Thus, we consider a uniform circular array, represented by $\eta_k = \frac{2\pi(k-1)}{m}$ for $k = 1, 2, \dots, m$ and $\rho_k = \rho$ for fixed values ρ and m . In this study, we consider $m = 15$ and $\rho = 1/(4 \sin(\pi/m))$, corresponding to half wavelength separation between nearby elements. This scenario is difficult, due to high coherence (side lobe level) of the resulting dictionary. We consider two sources with highly different level of amplitudes, namely $s_1 = 1$ and $s_2 = 0.1j$. The true DOA for the first source is fixed at $\theta_1 = 0$, while the position of the second source is variable. SNR was set to 20dB for the weaker source. As a comparison, we also consider the estimation result by RELAX and AIC enumeration, which is a popular technique. Figure 10 shows the result. As seen and expected, the estimation performance rapidly changes by the DOA separation. The interesting observation is that the RELAX technique can detect only the stronger source,

because the coherence between the bases corresponding to the two DOAs masks the effect of the weaker source. However, the ANDN approach is well known to lead to a mechanism, retrieving the latter. However, a bias is induced from the larger source, which decreases the estimation performance on the smaller source.

6 Conclusion

In this paper, we considered the problem of atomic decomposition, concerning a variety of different applications. We introduced spectral line and sensor/radar-related estimation problems as such applications. We focused on the sparsity based estimation techniques, where discretization influences the estimation process. More specifically, we discussed the atomic norm denoising technique, which does not include discretization. Our contribution here was to develop a general method to numerically solve atomic norm denoising. Previously, atomic norm denoising was only considered for polynomial-based models such as frequency estimation with uniform samples.

We rigorously showed that the developed technique may only stop at the global optimum of ANDN, which is well known to provide a remarkable performance. By numerical experiments we also verified that the technique converges faster than the other state-of-the-art ADMM implementation of ANDN for the frequency estimation problem.

Once the exact solution of ANDN was provided by the proposed algorithm, it was compared to other sparsity based techniques. We considered SPICE and S-TLS as well as the popular parametric approach RELAX. It is clearly observed that the ANDN approach often overestimates the number of parameters, when an infinite dynamic range is considered. However, the extra estimates usually possess remarkably (10 to 100 times) smaller amplitudes than the ones corresponding to true parameters. The second observation was that although the exact ANDN solution has a smaller RMSE in very high SNR ranges, the combining technique over the grid-based BP can improve its precision in certain cases. This reveals a new general potential in the spectral and grid based techniques, which should be a subject of future study.

Finally, the difficult scenario of DOA estimation, considered here, once again confirms the strength of sparsity based techniques such as ANDN in difficult scenarios and high dynamic range of variables.

7 Appendix: Properties of the Atomic Norm

We show that the atomic norm defined in (8) exists and is obtained by $n \leq 2m$. For the first part, consider the following optimization

$$\sup t \quad \text{s.t.} \quad t\mathbf{y} \in \text{conv}(\mathcal{A}) \quad (38)$$

where $\text{conv}(\mathcal{A})$ denotes the convex hull of the dictionary \mathcal{A} . Denoting the supremum by t_0 , we observe that t_0 is finite, since the set $\text{conv}(\mathcal{A})$ is bounded. Moreover $t_0\mathbf{y} \in \text{conv}(\mathcal{A})$, since $\text{conv}\mathcal{A}$ is closed. Now, note that for any set of parameters $\{r_k\}$ and bases $\{\mathbf{a}_k\}$ in (8), we have that

$$\frac{\mathbf{y}}{\sum_l r_l} = \sum_k \mathbf{a}_k \frac{r_k}{\sum_l r_l} \in \text{conv}(\mathcal{A}) \quad (39)$$

Thus,

$$\frac{1}{\sum_l r_l} \leq t_0 \rightarrow \sum_l r_l \geq \frac{1}{t_0} \quad (40)$$

On the other hand, $t_0\mathbf{y} \in \text{conv}(\mathcal{A})$ implies that there exists constants $\lambda_1, \dots, \lambda_n$ and bases $\mathbf{a}_{01}, \dots, \mathbf{a}_{0n}$ such that $\sum_k \lambda_k = 1$ and

$$t_0\mathbf{y} = \sum_k \mathbf{a}_{0k} \lambda_k \quad (41)$$

Define $r_{0k} = \lambda_k/t_0$. It is now easy to see that due to (40), $\|\mathbf{y}\|_{\mathcal{A}} = 1/t_0$ and $\{r_{0k}, \mathbf{a}_{0k}\}$ is a global minimum point in (8).

For the second part, assume without loss of generality that $\{r_{0k}, \mathbf{a}_{0k}\}$ is a global minimum point in (8) with shortest order n . If $n \geq 2m + 1$, we can fix \mathbf{a}_{0k} and apply the extremizer step in Section 4.2 to update r_{0k} . Note that the update sets one of the elements to zero without increasing the cost $\sum_k r_{0k}$. Removing this element and its corresponding basis, we observe that the resulting set of parameters also minimizes (8), but has a smaller order, contradicting the order minimality of the former parameters. This shows that $n \leq 2m$.

References

- [1] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic Decomposition by Basis Pursuit,” *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, Dec. 1998.

- [2] M. A. Herman and T. Strohmer, “High-resolution radar via compressed sensing,” *Signal Processing, IEEE Transactions on*, vol. 57, no. 6, pp. 2275–2284, 2009.
- [3] Y. C. Eldar and M. Mishali, “Robust recovery of signals from a structured union of subspaces,” vol. 55, no. 11, pp. 5302–5316, 2009.
- [4] D. Malioutov, M. Çetin, and A. S. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 3010–3022, 2005.
- [5] H. Yao, P. Gerstoft, P. M. Shearer, and C. Mecklenbräuker, “Compressive sensing of the tohoku-oki mw 9.0 earthquake: Frequency-dependent rupture modes,” *Geophys. Res. Lett.*, vol. 38, Oct. 2011.
- [6] R. G. Baraniuk, “Compressive sensing [lecture notes],” vol. 24, pp. 118–121, July 2007.
- [7] D. Donoho, “Compressed sensing,” vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [8] D. Donoho and X. Huo, “Uncertainty principles and ideal atomic decomposition,” vol. 47, no. 7, pp. 2845–2862, nov 2001.
- [9] S. Oymak, C. Thrampoulidis, and B. Hassibi, “The squared-error of generalized lasso: A precise analysis,” *arXiv preprint arXiv:1311.0830*, 2013.
- [10] E. J. Candès and Y. Plan, “Near-ideal model selection by ℓ_1 minimization,” *Ann. Stat.*, vol. 37, pp. 2145–2177, Oct. 2009.
- [11] P. Zhao and B. Yu, “On model selection consistency of lasso,” *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.
- [12] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, “Sensitivity to basis mismatch in compressed sensing,” 2011.
- [13] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Stat. Soc., Series B (Methodological)*, vol. 58, pp. 267–288, Jan. 1996.
- [14] D. Donoho and Y. Tsaig, “Fast solution of 1-norm minimization problems when the solution may be sparse,” vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

- [15] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 1.21,” <http://cvxr.com/cvx>, Apr. 2011.
- [16] M. A. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 586–597, 2007.
- [17] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communications on pure and applied mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [18] W. J. Fu, “Penalized regressions: the bridge versus the lasso,” *Journal of computational and graphical statistics*, vol. 7, no. 3, pp. 397–416, 1998.
- [19] P. Stoica, P. Babu, and J. Li, “Spice: A sparse covariance-based estimation method for array processing,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 2, pp. 629–638, 2011.
- [20] Z. Tan and A. Nehorai, “Sparse direction of arrival estimation using co-prime arrays with off-grid targets,” *Signal Processing Letters, IEEE*, vol. 21, no. 1, pp. 26–29, 2014.
- [21] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, “Recovery of sparse translation-invariant signals with continuous basis pursuit,” 2011.
- [22] H. Zhu, G. Leus, and G. B. Giannakis, “Sparsity-cognizant total least-squares for perturbed compressive sampling,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 5, pp. 2002–2016, 2011.
- [23] M. V. Ashkan Panahi, “Gridless compressive sensing,” in *IEEE Int. Conf. Acoust. Speech, Signal Processing*, 2014.
- [24] D. Malioutov, “A sparse signal reconstruction perspective for source localization with sensor arrays,” Master’s thesis, MIT, July 2003.
- [25] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, “Compressive sensing off the grid,” in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 778–785.
- [26] A. Panahi, “Parameter estimation using sparse modeling: Algorithms and performance analysis,” *Licentiate thesis, Chalmers University of Technology, Department of Signals and Systems*, Sept. 2012.

- [27] B. N. Bhaskar and B. 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.
- [28] H. Krim and M. Viberg, “Two decades of array signal processing research: the parametric approach,” vol. 13, pp. 67–94, July 1996.
- [29] P. Stoica and N. Arye, “Music, maximum likelihood, and cramer-rao bound,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 5, pp. 720–741, 1989.
- [30] P. Stoica and Y. Selen, “Model-order selection: a review of information criterion rules,” *Signal Processing Magazine, IEEE*, vol. 21, no. 4, pp. 36–47, 2004.
- [31] 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.
- [32] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed. Wiley, 2006.
- [33] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

Paper 2

Performance Analysis of Parameter Estimation Using LASSO

A. Panahi, M. Viberg

To be Submitted to IEEE transactions on signal processing

Performance Analysis of Parameter Estimation Using LASSO

A. Panahi, M. Viberg

Abstract

The Least Absolute Shrinkage and Selection Operator (LASSO) has gained attention in a wide class of continuous parametric estimation problems with promising results. It has been a subject of research for more than a decade. Due to the nature of LASSO, the previous analyses have been non-parametric. This ignores useful information and makes it difficult to compare LASSO to traditional estimators. In particular, the role of the regularization parameter and super-resolution properties of LASSO have not been well-understood yet. The objective of this work is to provide a new insight into this context by introducing LASSO as a parametric technique of a varying order. This provides us theoretical expressions for the LASSO-based estimation error and false alarm rate in the asymptotic case of high SNR and dense grids. For this case, LASSO is compared to maximum likelihood and conventional beamforming. It is found that LASSO loses performance due to the regularization term, but the amount of loss, in terms of mean squared error, is practically negligible with a proper choice of the regularization parameter. Thus, we provide suggestions on the selection of the regularization parameter. Without loss of generality, we present the comparative numerical results in the context of Direction of Arrival (DOA) estimation using a sensor array.

1 Introduction

The last two decades witnessed rapid emergence of sparse data models and their corresponding techniques in many traditional signal processing areas [1–5]. Although the basic principles of sparsity are easily recognized in many conventional methods, more exotic approaches such as ℓ_1 penalized least squares, well known as LASSO [6] (Least Absolute Shrinkage and Selection Operator), basis pursuit [7] or global matched filter [8], and its variants [9–12] have been unknown to the majority of the researchers until recently. Regarding these new techniques, it is natural to question how

these sparsity-based methods improve the conventional techniques. This is especially important since the new methods demand substantially higher computational effort in their basic form. In addition, many related questions such as the role of the regularization parameter and the effect of off-grid parameters in LASSO remain unclear. Hence, the current study is devoted to an analysis of LASSO, which provides both a framework to compare LASSO to traditional techniques and a deeper insight into the above questions.

LASSO is a smart solution to the Atomic Decomposition Problem (ADP), first formulated by Chen and Donoho [7]. Many other techniques such as matching pursuit [13] and orthogonal matching pursuit [14], Least Angle Regression (LARS) [15], and Compressive Sampling Matching Pursuit (CoSaMP) [16] are also developed to solve the ADP. The ADP naturally appears in various problems, e.g. the ones dealing with physical fields [5,17,18]. It invokes sparsity, since it may be viewed as a specific type of the so called Union-of-Subspaces (UoS) model, where subspaces are constructed from a set of dictionary bases [19]. However, LASSO is only well-defined for a finite dictionary case, while the problems of interest herein are normally related to infinite dictionaries. Examples of such are frequency and spectrum estimation [20,21], sensor array analysis [22], image processing [23,24], tomography [17,25] and seismology [5]. In practice, this is rectified by adopting a dictionary sampling (discretization) scheme, which provides a set of *quantized* estimates [26]. This is similar in spirit to the so-called spectral based techniques such as matched filter banks [27], but different in that LASSO provides a strongly sparse spectrum [22]. Another difficulty arises in selecting the LASSO Regularization Parameter (RP). In essence, this reflects the freedom in selecting the model order. However, this is particularly difficult since the relation between RP and the model order is generally complex [28].

Many other pioneering works have considered analyzing LASSO, commonly focusing on an information theoretic aspect, widely referred to as compressive sensing (CS) [29]. In other words, LASSO has been studied as a *decoder*, which together with a random linear encoding scheme provides a capacity achieving (optimal) compression rate in asymptotically large setups. However, as we show here, the asymptotic analysis, such as the ones in [23,29–32] rely on techniques which neglect useful information of LASSO, making them unsuccessful in explaining various effects such as discretization and regularization. Consequently, the final results of a CS study is incompatible with a classical analysis of an estimation problem. The same concern is also observed in some other works, e.g. [33,34]. In the above studies, it is observed that the difficulty appears since the sparsity pattern (support) is expressed implicitly (non-parametric). Thus, we suggest to fill the above gap by providing an analysis, where the support is explicitly ex-

pressed by parameters (parametric). In the previous literature, one may find similar attempts such as [35]. However, the considered metrics therein are not sufficient for the above mentioned practical interest.

From a parametric point of view, an ADP is a variable-order problem, where LASSO simultaneously provides an estimator and an order selection scheme. Taking this perspective and similar to many classical studies, we consider an individual case analysis, enabling comparison to the Cramer Rao Bound (CRB) [36]. This also brings a new insight into the problem of RP selection as an order selection technique. We also consider an asymptotically high SNR analysis to enjoy linearization techniques. We further address the discretization problem. Similar to the spectral-based techniques, our approach is to find an intermediate continuous estimator, of which the LASSO estimates can be regarded as a quantization. In simpler words, we show that the LASSO estimates converge to the intermediate estimates, called Continuous LASSO (CLASS) estimates, when we employ an increasingly dense discretization. The idea of CLASS is rapidly emerging in the ongoing research literature [37,38]. Thus, the current work can also be considered as an analysis of the more recent techniques of solving the ADP. Clearly, the implementation aspect of CLASS is irrelevant to the current study as it only serves as a bridge to analyze LASSO. The implementation of CLASS is addressed in [38,39]. The LASSO error is then identified as the combination of the CLASS error and the trivial quantization noise imposed by discretization.

Employing the above, we obtain the following results. First, we find the explicit relation between error, noise and the RP. This confirms that the RP introduces an undesired bias. However, unlike the Fourier-based techniques, the bias is proportional to the noise and vanishes in the noiseless case. Another important observation is that the behavior of LASSO in the noiseless case is completely independent of the signal power. Note that in presence of sources with high dynamic range, other state of the art techniques such as RELAX [40] and SAGE [41] are well known to behave poorly. Then, we discuss a certain strategy of RP selection and formulate the overall mean squared error (MSE) corresponding to the selected strategy. These results generally show that although LASSO does not achieve the CRB due to the regularization induced bias, in many occasions the degradation is negligible.

In summary, the novel ideas and results of this paper are the following:

- We introduce a framework, enabling to compare LASSO with other parameter estimation techniques.
- We provide asymptotic expressions for the LASSO estimation error in our developed framework and calculate its statistical properties, such

as bias and MSE.

- Based on the expressions, we provide some suggestions for the selection of the RP.
- We compare the resulting expressions to the error of the previously analyzed techniques, namely RELAX and conventional beamforming, as well as the CRB. We conclude that while the LASSO technique is substantially more robust assuming high dynamic range of amplitudes, in many practical situations, it loses a negligible amount of performance due the biasing effect of regularization.

2 Mathematical Modeling

Consider a closed index set $\Theta \subset \mathbb{R}$ and a collection of complex *basis* vectors $\mathbf{a}(\theta) \in \mathbb{C}^m$ indexed by the elements $\theta \in \Theta$. For our purpose, it suffices to assume that $\mathbf{a}(\theta)$ is a smooth function of θ , where it is referred to as a *manifold*. In most applications of interest, the dependence of $\mathbf{a}(\theta)$ on θ is non-linear. Consider a set of n indexes $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_n]$ and its corresponding discrete-time *complex amplitudes* $\mathbf{s}(t) = [s_1(t), s_2(t), \dots, s_n(t)] \in \mathbb{C}^n$ for $t = 1, 2, \dots, T$. We refer to the expression

$$\mathbf{x}(t) = \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) \quad (1)$$

as an atomic synthesis. For simplicity we denote a synthesis such as (1) by $\{(\theta_k, \{s_k(t)\})\}$. In many cases, the synthesized vectors $\mathbf{x}(t)$ correspond to a sequence of observed data and the pair of indexes and amplitudes are to be estimated. This is called an atomic decomposition problem (ADP) [7]. We call a basis manifold $\mathbf{a}(\theta)$ *regular* if any arbitrary sequence $\{\mathbf{x}(t)\}$ can be decomposed as in (1). We assume that $\mathbf{a}(\theta)$ is regular throughout this paper. Generally speaking, the order n may or may not be known. In either case, the model in (1) may be insufficient to uniquely infer the decomposition $\{(\theta_k, \{s_k(t)\})\}$ from observations $\{\mathbf{x}(t)\}$. If n is unknown, the ADP model is commonly accompanied by the principle of parsimony, stating that the smallest order n , usually referred to as data rank, is always preferable. In this case, the corresponding synthesis is often referred to as an ideal ADP. Of course, this is not generally appealing in presence of noise, which is shortly discussed.

It is also useful to consider the smallest number n_0 of linearly dependent atoms $\{\mathbf{a}(\theta_k)\}$, which is sometimes denoted by $\text{Spark}(\mathbf{a})$ [42]. In this case, the manifold $\mathbf{a}(\theta)$ is also called n_0 -ambiguous. Clearly, this is only possible

if $n_0 \leq m + 1$. Moreover, one can easily construct an $m + 1$ -ambiguous manifold in an m -dimensional space, which is often referred to as an unambiguous. A practical example of such is the Uniform Linear Array (ULA) manifold, discussed in Section 2.2. The following simple but fundamental result according to [42] formulates the uniqueness of the ideal ADP.

Theorem 4. *If a manifold is n_0 -ambiguous, each set of observations $\{\mathbf{x}(t)\}$ has at most one ideal decomposition $\{(\theta_k, \{s_k(t)\})\}$ of an order $n < n_0/2$.*

Another issue with ADP is that the observed data is normally noisy. Thus, it is more practical to assume a model of the following form

$$\mathbf{x}(t) = \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) + \mathbf{n}(t) \quad (2)$$

where the noise vector $\mathbf{n}(t)$ is assumed to be a white and circularly symmetric, complex-valued Gaussian process throughout this study. Given T data snapshots $\{\mathbf{x}(k)\}$, the problem of interest here is to estimate the signal parameters θ and their corresponding amplitudes s . If the model order is unknown, it also needs to be estimated. The focus in this paper is to assess the quality of the parameter estimates $\hat{\theta}$.

For the noisy model in (2), the Least Squares (LS) solution of ADP for a given order n is given by

$$\min_{\mathcal{M}_n} \sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) \right\|_2^2 \quad (3)$$

where \mathcal{M}_n denotes the set of all irreducible decompositions $\{(\theta_k, \{s_k(t)\})\}$ of order n . The LS solution in (3) coincides with the Maximum Likelihood (ML) estimator, providing interesting statistical properties [36, 43].

In ignorance of the order n , the previous statement of the principle of parsimony should be utilized with extra care. Note that unlike (1), the model in (2) admits any order n . However, lower order expressions associate with higher magnitude residual. Thus, a more sophisticated modification of (3) should be considered. This is usually referred to as the Model Order Selection (MOS) problem, which essentially establishes a balance between the residual level and the order [44, 45].

2.1 The Principle of Sparsity

Solving the LS problem in (3) and MOS have been previously considered. For reviews on different aspects of the problem, see [46, 47], where the accuracy of the estimates for $\{\theta_k\}$ and $\{s_k(t)\}$ are also discussed in different

asymptotic cases. Such analysis and any numerical method to solve (3) by assigning and iteratively updating values to parameters $\{\theta_k\}$ is commonly called *parametric*. Despite their theoretical accuracy, the parametric approaches suffer from numerical deficiency, which has motivated for alternative approaches. LASSO is regarded as such a *non-parametric* method, which has its roots in what we refer to as the principle of sparsity, explained below.

The principle of sparsity simply refers to the fact that in (1), the parameters corresponding to zero amplitude are ignorable. Note that taking any decomposition $A = \{(\theta_k, \{s_k(t)\})\}$ of an order n , one may define $\boldsymbol{\theta}_s = \{\theta_{k_1}, \theta_{k_2}, \dots, \theta_{k_r}\}$, the subset of $\{\theta_k\}$ comprising the elements θ_k for which $s_k(t) \neq 0$ for at least a single snapshot t . Equivalently, $\theta_k \notin \boldsymbol{\theta}_s$ implies that $s_k(t) = 0$ for every time index t . The set $\boldsymbol{\theta}_s$ and its number of elements are called the *support* and *cardinality* of the decomposition, and are denoted by $\text{Supp}(A)$ and $\|A\|_0$, respectively. Then, define the *reduced* decomposition $B = \{(\theta_{k_l}, \{s_{k_l}(t)\})\}$. Note that

$$\mathbf{x}(t) = \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) = \sum_{\theta_k \in \boldsymbol{\theta}_s} \mathbf{a}(\theta_k) s_k(t) + \underbrace{\sum_{\theta_k \notin \boldsymbol{\theta}_s} \mathbf{a}(\theta_k) s_k(t)}_0 = \sum_{\theta_k \in \boldsymbol{\theta}_s} \mathbf{a}(\theta_k) s_k(t) \quad (4)$$

Thus, B synthesizes the same vectors $\mathbf{x}(t)$ as A and the difference between B and A is practically unimportant. We call the reduced decomposition B the *root* of the original decomposition A . If a decomposition $\{(\theta_k, \{s_k(t)\})\}$ is such that for any k the amplitude $s_k(t)$ is nonzero for at least one time index t , i.e. it is only reduced to itself, it is called an irreducible decomposition.

A high-order decomposition with a low cardinality is called a sparse decomposition. It can be naturally reduced to a low-order decomposition. Accordingly, LASSO is based on finding a sparse decomposition, which reduces to the ideal decomposition.

2.2 The Sensor Array Example

Finally in this section, we introduce a practical illustrative example which we also consider later. We consider the planar Direction Of Arrival (DOA) estimation problem, in which a set of m sensors listen to n far and narrow band sources and decide on their directions. The received data is modeled by (1), where the basis manifold is given by ([8])

$$\mathbf{a}(\theta) = \begin{bmatrix} e^{j\frac{2\pi}{d}r_1 \cos(\theta-\rho_1)} & e^{j\frac{2\pi}{d}r_2 \cos(\theta-\rho_2)} & \dots & \\ & & & e^{j\frac{2\pi}{d}r_m \cos(\theta-\rho_m)} \end{bmatrix}^T, \quad (5)$$

in which (r_i, ρ_i) is the polar coordinate pair of the i^{th} sensor ($i = 1, 2, \dots, m$) and d is the wavelength at the central frequency. Then, the goal is to estimate $\{\theta_k\}$ which represents the directions given $\{\mathbf{x}(t)\}$. Obviously, the problem is defined in a complex-valued space of variables. The manifold in (5) is not necessarily unambiguous. An important unambiguous case, which we focus on later is the half-wavelength ($r_i = \frac{(i-1)d}{2}$) Uniform Linear Array (ULA). Note that a linear array means that $\rho_i = 0$. In this case, it is more convenient to write (5) in terms of the *electrical angle* $\phi = \pi \cos \theta$. The ULA manifold resembles the classical Fourier basis, when represented in terms of the electrical angle. Thus, the sensor array example essentially includes other applications such as frequency estimation and sampling.

The ULA manifold is unambiguous. This is easily seen by taking any combination of m distinct bases indexed by electrical angles $\phi_1, \phi_2, \dots, \phi_m$ and noting that the matrix

$$[\mathbf{a}(\phi_1) \ \mathbf{a}(\phi_2) \ \dots \ \mathbf{a}(\phi_m)] = \begin{bmatrix} 1 & 1 & \dots & 1 \\ e^{j\phi_1} & e^{j\phi_2} & \dots & e^{j\phi_m} \\ e^{j2\phi_1} & e^{j2\phi_2} & \dots & e^{j2\phi_m} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j(m-1)\phi_1} & e^{j(m-1)\phi_2} & \dots & e^{j(m-1)\phi_m} \end{bmatrix} \quad (6)$$

is a Vandermonde matrix and thus its columns are linearly independent as long as they are distinct.

3 LASSO, Parametric LASSO and CLASS

In the previous section we formulated atomic decomposition by LS and discussed the principle of sparsity. Sparsity does not directly simplify the computational procedure of obtaining the decomposition. Instead, it provides a framework to obtain better approximate results. For example, greedy algorithms such as Matching Pursuit (MP) [13] and Orthogonal Matching Pursuit (OMP) [14] can be applied due to the principle of sparsity. They basically select bases from Θ iteratively. Despite their wide application, they are shown to lack robustness in certain examples. This motivated a different approach by introducing an approximate optimization, whose solution is related to ADP and simple to obtain. A fairly general framework in this matter is to consider convex optimization, where LASSO is a good example. To solve the ADP, the LASSO method suggests to consider a finite, but large discretization (grid) $\tilde{\Theta} = \{\tilde{\theta}^1, \tilde{\theta}^2, \dots, \tilde{\theta}^N\}$ of Θ and assign parameters $\{\tilde{s}^k(t)\}$ to $\tilde{\theta}^k$. The parameters $\tilde{\theta}^k$ are known. However, if the

decomposition $\{\tilde{\theta}^k, \{\tilde{s}^k(t)\}\}$ is sparse it can be reduced to a good approximation of any desired decomposition. To ensure sparsity, LASSO considers the following optimization

$$\min_{\{\tilde{s}_k(t)\}} \frac{1}{2} \sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{k=1}^N \mathbf{a}(\tilde{\theta}_k) \tilde{s}_k(t) \right\|_2^2 + \lambda \|\{\tilde{s}_k(t)\}\|_{2,1} \quad (7)$$

where

$$\|\{\tilde{s}_k(t)\}\|_{2,1} = \sum_{k=1}^N \sqrt{\sum_{t=1}^T |\tilde{s}_k(t)|^2} \quad (8)$$

This is sometimes called group-LASSO to emphasise on its multi-snapshot nature. The regularization parameter $\lambda > 0$ controls the cardinality of the estimate while the order is fixed to N (see Section 2.1). However, the analytical relation between λ and the cardinality is difficult to obtain. It is also easy to show that group-LASSO always has a solution of a bounded cardinality independently of the order N and different grid choices. For a single snapshot case, $T = 1$, group-LASSO is simplified to the more familiar LASSO optimization. We usually apply the term LASSO to also refer to group-LASSO when there is no risk of confusion. Using Lagrangian duality and noting that group-LASSO is a convex optimization, (7) can be written in different equivalent forms. In this paper, we always refer to (7) as the canonical form.

LASSO and group-LASSO can be numerically solved in polynomial time by off-the-shelf optimization techniques such as interior point. There is also a variety of different heuristic numerical methods to decrease its computational complexity [15, 48]. It has been observed that all of these numerical techniques run into numerical problems when LASSO is applied to a dense grid with small λ . Thus, a relatively coarse grid is applied in practice, which leads to the so-called off-grid problem as the true parameters can be relatively distant from the grid points. There has been different attempts to overcome the off-grid effect [34, 49]. To the best of our knowledge no related general study is available. A natural attempt to isolate the effect of discretization in an analysis is to extend LASSO to admit a continuum instead. This has been central in many recent studies such as [33, 34, 37], where the idea of atomic norm regularization is proposed and limited numerical approaches are discussed. In [38] an extension with a general implementation is also proposed. The numerical implementation is not a concern in the current study and as we shortly discuss, all the above extensions are theoretically equivalent. Here, we develop this unique extension again as a parametric method instead, and rename it as Continuous LASSO (CLASS) to emphasis on its parametric aspects. Then, we first relate the original

LASSO with discretization to CLASS by an asymptotic analysis and next provide a statistical analysis of CLASS.

3.1 Preliminaries on Asymptotic Analysis

The analysis herein is carried out based on some asymptotic assumptions. We have already pointed out the issue of grid density. Here we clarify the assumptions under which the analysis holds. In short, our analysis admits a case with sufficiently dense grid and highly small noise variance σ . To formulate the density of a grid, one may find the following definition useful.

Definition 1. A finite grid $G \subset \Theta$ is called δ -dense if for any $\theta \in \Theta$ there exists a close sample $\tilde{\theta}^k \in G$ such that $|\theta - \tilde{\theta}^k| < \delta$. In other words,

$$\max_{\theta \in \Theta} \min_{\tilde{\theta}^k \in \tilde{\Theta}} |\theta - \tilde{\theta}^k| < \delta \quad (9)$$

The asymptotic analysis of LASSO over dense grids, connects it to CLASS. We shortly show that the group-LASSO estimate for a δ -dense grid approaches a fixed solution when δ tends to zero, i.e. when the grid is densified. The limit, CLASS, is independent of other properties of the grid, and coincides with the result of the approaches in [33, 34, 37]. Although one may find this result intuitively trivial, the real difficulty here is in the mathematical development, which once accomplished, provides us with a strong tool to take further steps.

The main issue with the mathematical development is that LASSO provides a solution of varying, but not always desirable order. In practice, the off-grid effect leads to a remarkably overestimated order, where each true parameter is replaced by a group of nearby estimates, later referred to as a cloud. Thus, an extra care should be taken on evaluating the accuracy of the estimates. We define a proper distance (or more formally a topology) on the space of atomic decompositions, satisfying our practical concerns. This is given in Appendix 7.

We also consider a high Signal-to-Noise Ratio (SNR) scenario, where the ideal decomposition is close to the noiseless ADP in (1). For the noiseless case, the LS term in LASSO can be replaced by an equality constraint, which simplifies (7) to

$$\begin{aligned} \min_{\{\tilde{s}_k(t)\}} \|\{\tilde{s}_k(t)\}\|_{2,1} \\ \text{s.t} \\ \mathbf{x}(t) = \sum_{k=1}^N \mathbf{a}(\tilde{\theta}^k) \tilde{s}_k(t) \end{aligned} \quad (10)$$

This is known as the noiseless group-LASSO. It is well-known that the solution of the noiseless LASSO is also the limit solution of (7) when λ approaches zero. This, so-called *homotopy* rule suggests that in a high-SNR scenario a small value of λ should be utilized. We later discuss selection of λ in more detail.

Let us summarize the above. Considering the asymptotic high SNR analysis, our strategy is to first verify if a noiseless synthesis in (1) can be recovered perfectly by the noiseless LASSO, and then analyze the estimate by Taylor expansion for a small amount of noise. Note that to overcome the discretization effect, we eventually need to instead characterize the limit estimates for infinitely dense grids, called CLASS estimates. We also find the following well-known result, characterizing the solution of (7) and (10) useful in our analysis. Note that group-LASSO is convex and thus its local optimality condition by the Karush-Kuhn-Tucker (KKT) theorem guarantees global optimality. The following theorem provides the KKT condition for (7) and (10), which characterizes the LASSO solution.

Lemma 1. Consider a sequence $\{\tilde{s}_k(t)\}$ and define

$$\tilde{p}_k = \sqrt{\sum_t |\tilde{s}_k(t)|^2} \quad (11)$$

Then, $\{\tilde{s}_k(t)\}$ is an optimal point in (10) if and only if there exists a *dual verifier sequence* $\mathbf{z}(1), \mathbf{z}(2), \dots, \mathbf{z}(T) \in \mathbb{C}^m$ such that

$$\tilde{p}_k \neq 0 \rightarrow \mathbf{a}^H(\tilde{\theta}^k)\mathbf{z}(t) = \frac{\tilde{s}_k(t)}{\tilde{p}_k} \quad (12)$$

and

$$\sqrt{\sum_{t=1}^T \left| \mathbf{a}^H(\tilde{\theta}^k)\mathbf{z}(t) \right|^2} \leq 1, \quad (13)$$

for $k = 1, 2, \dots, N$. Moreover, $\{\tilde{s}_k(t)\}$ is a solution to (7) if the dual verifiers also satisfy

$$\lambda \mathbf{z}(t) = \mathbf{x}(t) - \sum_{k=1}^N \mathbf{a}(\tilde{\theta}^k) \tilde{s}_k(t) \quad (14)$$

Proof. First consider the noiseless case in (10). Take $\{\tilde{s}_k(t)\}$ as the optimal solution. Applying the KKT theorem, we obtain that there exists a sequence of vectors $\mathbf{z}(1), \mathbf{z}(2), \dots, \mathbf{z}(T) \in \mathbb{C}^m$ such that for each k ,

$$\{\mathbf{a}^H(\tilde{\theta}^k)\mathbf{z}(t)\}_{t=1}^T \in \partial \|\{\tilde{s}_k(t)\}\|_2 \quad (15)$$

where $\partial\|\{\tilde{s}_k(t)\}\|_2$ denotes the subdifferential of the multivariable 2-norm function $\|\{\tilde{s}_k(t)\}\|_2$. Note that

$$\partial\|\{\tilde{s}_k(t)\}\|_2 = \begin{cases} \left\{ \left\{ \eta_k(t) \right\}_{t=1}^T \mid \sum_{t=1}^T |\eta(t)|^2 \leq 1 \right\} & \tilde{p}_k = 0 \\ \left\{ \left\{ \eta_k(t) = \frac{\tilde{s}_k(t)}{\tilde{p}_k} \right\}_{t=1}^T \right\} & \tilde{p}_k \neq 0 \end{cases} \quad (16)$$

Plugging (16) into (15) proves the theorem. The noisy case in (7) is similarly proved by the KKT Theorem and using (16). \square

3.2 CLASS Solution

Now, we show that the solutions to LASSO (7) and (10) have unique limits when the density of the grid G increases. We first introduce the limit as a solution to a parametric optimization, called CLASS, and then show convergence. The main idea is that CLASS somehow generalizes LASSO to the case of a continuum. Of course, implementation aspects of CLASS falls beyond our concern as CLASS serves only as an analytical asymptotic tool. However, we remind once again that CLASS has also been considered by other researchers as a numerical approach and the implementation aspects of CLASS is an ongoing research [34, 39].

The CLASS formalism relies on the fact that any sparse decomposition can be expressed by its parameters over the support only, since other parameters are zero and uninteresting according to the principle of sparsity. Take an arbitrary collection $\{\tilde{s}_k(t)\}$ in the search space of LASSO over a fixed grid $\tilde{\Theta} = \{\tilde{\theta}^k\}$. Remember that this more precisely corresponds to the atomic decomposition $\{\tilde{\theta}^k, \{\tilde{s}_k(t)\}\}$, where $\{\tilde{\theta}^k\}$ is always neglected as it is known. Define its support $\{\theta_1 = \tilde{\theta}^{k_1}, \theta_2 = \tilde{\theta}^{k_2}, \dots, \theta_n = \tilde{\theta}^{k_n}\}$ and note that the reduced atomic decomposition $\{\theta_l = \tilde{\theta}^{k_l}, \{\tilde{s}_{k_l}(t)\}\}$ entirely represents the original decomposition. Thus, the search space of LASSO can be equivalently represented by the space of all possible reduced decompositions, i.e. the space of all irreducible representations $\{\theta_l, \{s_l(t)\}\}$ with $\theta_l \in \tilde{\Theta}$. Let us denote this space by $\tilde{\mathcal{M}}$. Then LASSO can be written as

$$\min_{\tilde{\mathcal{M}}} \frac{1}{2} \sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) \right\|_2^2 + \lambda \|\{s_k(t)\}\|_{2,1} \quad (17)$$

where

$$\|\{s_k(t)\}\|_{2,1} = \sum_{k=1}^n \sqrt{\sum_{t=1}^T |s_k(t)|^2} \quad (18)$$

Now, the generalization comes with relaxing the requirement that θ_k lies on the grid $\tilde{\Theta}$. Then, $\tilde{\mathcal{M}}$ is replaced by the set \mathcal{M} of all decompositions over

Θ :

$$\min_{\mathcal{M}} \frac{1}{2} \sum_{t=1}^T \left\| \mathbf{x}(t) - \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) \right\|_2^2 + \lambda \|\{s_k(t)\}\|_{2,1} \quad (19)$$

We call (19) parametric LASSO or Continuous LASSO when Θ is a continuum. While the reader may verify by simple calculations that the above is a different representation of the atomic norm denoising technique introduced in [37], it is also simple to show that the total variation formalism in [33] always leads to the same result as CLASS. Still, it is not clear that CLASS has a solution, since there is no restriction on the order n and the cost may decrease unboundedly. An independent argument on the existence of the CLASS solution is included in Appendix 7 to point out to some other useful technical facts. However, the reader may refer to [37] as well. Thus, we may consider that the solution of CLASS exists. One can similarly consider the following parametric extension of the noiseless LASSO in (7), which we call noiseless CLASS:

$$\begin{aligned} & \min_{\mathcal{M}} \|\{s_k(t)\}\|_{2,1} \\ & \text{s.t.} \\ & \mathbf{x}(t) = \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) \end{aligned} \quad (20)$$

Now, we state the convergence theorem, which ties CLASS to the conventional LASSO with a finite grid. We actually provide a stronger convergence property which includes all later asymptotic concerns, including the noise effect and the regularization parameter.

Theorem 5. *Consider a regular manifold $\mathbf{a}(\theta)$, arbitrary observations $\mathbf{x}(t)$, perturbations $\{\mathbf{n}(t)\}$, a grid $\tilde{\Theta}$ and $\lambda > 0$. For any desired precision $\epsilon > 0$, there exists a positive real δ such that if $\|\mathbf{n}(t)\|_2 \leq \delta$, $\tilde{\Theta}$ is δ -dense and $\lambda < \delta$, then any group-LASSO estimate for $\{\mathbf{x}(t) + \mathbf{n}(t)\}$ by $\tilde{\Theta}$ and λ are in an ϵ -neighborhood of a noiseless CLASS estimate of $\{\mathbf{x}(t)\}$.*

Proof. See Appendix 8. □

In simple words the solution of LASSO with a dense grid, small noise and regularization parameter is in the sense discussed in Appendix 7 close to the noiseless solution.

3.3 Dual Convergence Properties

Theorem 5 shows that the solution to the noisy group LASSO is arbitrarily close to the ideal noiseless CLASS in an asymptotic case. However, to analyze LASSO in the asymptotic case, we need to characterize these solutions.

We have already done this for LASSO in Lemma 1. Here, we extend this to the CLASS solution and provide convergence properties for the dual verifier vectors $\{\mathbf{z}(t)\}$. Once we provide these results we can characterize small perturbations by Taylor expansion, which is discussed in the next section.

Theorem 6. *A decomposition such as $\{(\theta_k, \{s_k(t)\})\}_{k=1}^n$ is a solution to noiseless CLASS with $\{\mathbf{x}(t)\}$ if and only if defining $p_k = \sqrt{\sum_t |s_k(t)|^2}$, there exists a sequence of dual verifier vectors $\{\mathbf{z}(t)\}$ such that*

$$\mathbf{a}^H(\theta_k)\mathbf{z}(t) = \frac{s_k(t)}{p_k} \quad (21)$$

and

$$\forall \theta \in \Theta \quad \sum_{t=1}^T |\mathbf{a}^H(\theta)\mathbf{z}(t)|^2 \leq 1 \quad (22)$$

Furthermore, for each arbitrary precision ϵ there exists $\delta > 0$ such that if $\tilde{\Theta}$ is δ -dense, $\lambda < \delta$ and $\|\mathbf{n}(t)\|_2 < \delta$ then any set of dual verifiers $\{\mathbf{z}_0(t)\}$ for their corresponding group LASSO over $\{\mathbf{x}(t) + \mathbf{n}(t)\}$ satisfies $\|\mathbf{z}(t) - \mathbf{z}_0(t)\|_2 < \epsilon$ for a set of dual verifiers $\{\mathbf{z}(t)\}$ corresponding to a solution of noiseless CLASS.

Proof. The proof is given in Appendix 9. □

3.4 First Order Linearization

We finally arrive at the crucial step of calculating the approximate LASSO error in a high SNR and dense grid case. We later develop conditions under which, the true parameters are exactly identical to the solution of noiseless CLASS. For the time being, we treat the noiseless CLASS solution as the desired estimate. Thus, the error is only associated with noise, grid and regularization parameter λ . Then, Theorem 5 shows that the error is infinitesimal in the vicinity of the ideal setup, i.e. when noise and λ are small and the grid is dense. This allows for the application of a Taylor expansion. However, due to the unfamiliar role of the grid and the unspecified order of the estimates, a careful study is necessary. Let us start from the result of Theorem 5. Take a solution $A = \{(\theta_k, \{s_k(t)\})\}$ corresponding to a δ -dense grid G , δ -small noise terms $\mathbf{n}(t)$ and $\lambda < \delta$. Suppose that δ is small such that Theorem 5 guarantees that A is in an ϵ -neighborhood of a noiseless CLASS solution $A_0 = \{(\theta_{l,0}, \{s_{l,0}(t)\})\}$. The definition of neighborhood allows that some indexes of A , associated to an infinitesimal amplitude, lie outside the ϵ -neighborhood of the elements of A_0 . We call them *false alarm*. More formally, an index θ_k is a false alarm if $|\theta_k - \theta_{l,0}| > \epsilon$ holds for all l . Clearly, for such an index $|s_k(t)| < \epsilon$ also holds. Note that the definition of

false alarm depends on the neighborhood size ϵ . Should there be a risk of confusion, we may refer to the term ϵ -false alarms for clarity. The other estimates, also called "detections" (or ϵ -detections) can be assigned uniquely to a close noiseless estimate in a sufficiently small neighborhood. However, the definition of neighborhood also allows for multiple detections assigned to the same index. We call this the dispersion effect, which might be related, for example, to discretization. Finally, the detections related to the same index are somehow subject to an overall estimation error (shift). Our analysis will characterize the above three asymptotic elements of estimation; false alarm, dispersion and the overall estimation error.

To formulate the asymptotic behavior of LASSO in the above sense, we first need to review some basic definitions. Consider again the above solution A in a sufficiently small ϵ -neighborhood of the noiseless solution A_0 such that each index θ_k in A is either a false alarm or is uniquely located in an ϵ -neighborhood of an index $\theta_{l,0}$. We refer to all elements θ_k in the neighborhood of a specific element $\theta_{l,0}$ as its corresponding *cloud*. We basically show that each cloud may consist of at most 2 elements of zero or first order. To elaborate on this, consider the third largest element in each cloud and denote the maximum amplitude of these elements by δ_3 . Then, we show that δ_3 vanishes up to first order with respect to δ . Finally, we define the "overall" effect of each cloud by the following parameters:

$$\sigma_l(t) = \sum_{k||\theta_k - \theta_{l,0}| < \epsilon} s_k(t) - s_{l,0}(t) \quad (23a)$$

$$\pi_l = \frac{1}{p_{l,0}} \sum_{k||\theta_k - \theta_{l,0}| < \epsilon} p_k(\theta_k - \theta_{l,0}) \quad (23b)$$

where $p_k = \sqrt{\sum_t |s_k(t)|^2}$ and $p_{l,0} = \sqrt{\sum_t |s_{l,0}(t)|^2}$. In fact, it is simple by Taylor expansion to see that the first order properties of any estimator is well expressed by the above parameters, where $\sigma_k(t)$ is complex-valued and $\pi_k(t)$ is real-valued. Note that, in general the characteristics of σ and π do not completely reveal the properties of individual indexes and amplitudes in each cloud, which after all, depend on the circumstances (e.g. discretization) under which the cloud is produced. Now, define

$$g = \frac{1}{2} \sum_t \left\| \mathbf{n}(t) - \sum_l (\mathbf{a}_l \sigma_l(t) + \mathbf{d}_l s_{l,0}(t) \pi_l) - \sum_p \mathbf{a}(\bar{\theta}_p) \bar{s}_p(t) \right\|_2^2 + \lambda \sum_p \sqrt{\sum_t |\bar{s}_p(t)|^2} + \lambda \sum_{l,t} \Re(\gamma_l^*(t) \sigma_l(t)) \quad (24)$$

where

$$\begin{aligned} \mathbf{a}_l &= \mathbf{a}(\theta_{l,0}) \\ \mathbf{d}_l &= \frac{d\mathbf{a}}{d\theta}(\theta_{l,0}) \\ \gamma_l(t) &= \frac{s_{l,0}(t)}{\sqrt{\sum_t |s_{l,0}(t)|^2}} \end{aligned} \quad (25)$$

This is a function of $\{\pi_l, \{\sigma_l(t)\}\}$ and an arbitrary decomposition $\bar{A} = \{\bar{\theta}_p, \{\bar{s}_p(t)\}\}$. The following theorem identifies the first order perturbation of the solution in terms of the above definitions.

Theorem 7. *Consider LASSO with a δ -dense grid, and $\lambda < \delta$ over observations with small perturbation $\|\mathbf{n}(t)\| < \delta$ such that any solution A lies in a small ϵ -neighborhood of a noiseless solution A_0 .*

a) *Minimizing g in (24) gives the first-order perturbation of the noisy solution: Consider the optimization*

$$\min_{\{\pi_l \in \mathbb{R}, \{\sigma_l(t) \in \mathbb{C}\}, \{\bar{\theta}_p, \{\bar{s}_p(t)\}\}} g \quad (26)$$

There exists a minimum point $\bar{\pi}_l, \bar{\sigma}_l(t)$ and \bar{A} such that up to first order, $\pi_l, \sigma_l(t)$ and false alarms are identical to $\bar{\pi}_l, \bar{\sigma}_l(t)$ and \bar{A} .

b) *There exists a solution of LASSO for which the maximum false amplitude δ_3 vanishes up to first order, i.e. $\delta_3 = o(\delta)$.*

Proof. See Appendix 10 for proof and more details. \square

Theorem 7 may be regarded as the central contribution of this work. Once this is established, characterizing the high SNR properties of LASSO boils down to analyzing the minimizers of the linearized criterion g . The next section provides such an analysis, where we use the linearization result to give a statistical analysis of LASSO estimates in presence of a white Gaussian noise.

4 Statistical Results

In the previous section, we developed results characterizing the LASSO estimates in an asymptotic case. In this section, we connect those results to practice. We shortly address the statistical effect of noise and grid on the estimation procedure. We also deal with a more fundamental question of consistency. Recall that the linearization results characterize the deviation from the noiseless solution, but we have not yet discussed the own properties of the noiseless solution. Many previous studies have considered this and

what we correspondingly state in the sequel is more or less a restatement of the results in [33,34] for special cases, which is derived more systematically as a part of a general framework resulting from Theorem 7. In fact, Theorem 7 is central in the entire discussion of the current section, which readily characterizes the first order deviation from a noiseless solution. It only remains to investigate the statistical properties of the deviation in a given scenario. Hence, it seems rational to spend a bit of effort first to learn more about the consequences of Theorem 7.

Let us start by some simplifying definitions. Consider a noiseless solution $A = \{\theta_k, \{s_k(t)\}\}$ and its corresponding parameters $\mathbf{a}_k = \mathbf{a}(\theta_k)$, $\mathbf{d}_k = d\mathbf{a}/d\theta(\theta_k)$ and $\gamma_k(t) = s_k(t)/p_k$, where $p_k = \sqrt{\sum_t |s_k(t)|^2}$. Define $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_n]$ and $\mathbf{D} = [\mathbf{d}_1 \ \mathbf{d}_2 \ \dots \ \mathbf{d}_n]$ as well as $\mathbf{A}^\dagger = (\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H$ and $\mathbf{P} = \mathbf{I} - \mathbf{A} \mathbf{A}^\dagger$, where \mathbf{I} denotes the unit matrix. Finally, define

$$\begin{aligned} \xi_{l,k} &= \sum_{t=1}^T s_{l,0}^*(t) s_{k,0}(t) \\ \mathbf{R} &= \Re [(\mathbf{D}^H \mathbf{P} \mathbf{D}) \odot \Xi] \end{aligned} \quad (27)$$

where \odot denotes elementwise product and Ξ is the matrix of the elements $\xi_{l,k}$. Denote by $\boldsymbol{\xi}_k$ the k^{th} column of Ξ .

Now let us try to solve (26). Note that fixing the false alarm \bar{A} , the optimization over the π and σ parameters is quadratic of the following form

$$\min_{\{\pi_l(t), \sigma_l(t)\}} \frac{1}{2} \sum_t \left\| \boldsymbol{\nu}(t) - \sum_l (\mathbf{a}_l \sigma_l(t) + \mathbf{d}_l s_{l,0}(t) \pi_l) \right\|_2^2 + \lambda \sum_{l,t} \Re(\gamma_l^*(t) \sigma_l(t)) \quad (28)$$

where

$$\boldsymbol{\nu}(t) = \mathbf{n}(t) - \sum_p \mathbf{a}(\bar{\theta}_p) \bar{s}_p(t) \quad (29)$$

and the constant terms are neglected. The solution to this can easily be found by differentiation as

$$\boldsymbol{\sigma}(t) = \mathbf{A}^\dagger (\boldsymbol{\nu}(t) - \sum_l \mathbf{d}_l s_{l,0}(t) \pi_l) - \lambda (\mathbf{A}^H \mathbf{A})^{-1} \boldsymbol{\gamma}(t) \quad (30)$$

$$\boldsymbol{\pi} = \mathbf{R}^{-1} (\boldsymbol{\omega} + \lambda \boldsymbol{\delta}) \quad (31)$$

where $\boldsymbol{\sigma}(t)$, $\boldsymbol{\pi}$, $\boldsymbol{\omega}$ and $\boldsymbol{\delta}$ denote the vectors with $\sigma_k(t)$, π_k , ω_k and δ_k as elements respectively such that

$$\omega_k = \Re(\boldsymbol{\zeta}_k^H \mathbf{P} \mathbf{d}_k) \quad \delta_k = \Re(\boldsymbol{\xi}_k^T \mathbf{A}^\dagger \mathbf{d}_k) \quad (32)$$

where

$$\boldsymbol{\zeta}_k = \sum_{t=1}^T s_{k,0}^*(t) \boldsymbol{\nu}(t) \quad (33)$$

On the other hand, fixing $\{\pi_k, \{\sigma_k(t)\}\}$, the optimization over false alarm is a LASSO problem. We have found it both difficult and practically uninteresting to fully analyze the properties of the false alarm solution as a random atomic decomposition, or more restrictively, a random finite set. Instead, we only study the occurrence of false alarm, which is to identify when \bar{A} is nonempty in (26). Note that when \bar{A} is empty $\boldsymbol{\nu}(t) = \mathbf{n}(t)$ and according to Theorem 6 the following relation equivalently holds for any θ .

$$\sum_t \left| \mathbf{a}^H(\theta) \left(\mathbf{n}(t) - \sum_l (\mathbf{a}_l \sigma_l(t) + \mathbf{d}_l s_{l,0}(t) \pi_l) \right) \right|^2 \leq \lambda^2 \quad (34)$$

where $\sigma_k(t)$ and π_k are given by (30). We define the probability of false alarm (PFA) as $\text{PFA} = \Pr(\bar{A} \neq \emptyset)$.

4.1 Ideal Consistency

Based on the above, we now provide a sufficient condition for a true decomposition to be exactly retrieved by noiseless CLASS. Note that Theorem 6 gives a necessary and sufficient condition for our purpose. However, it is not straightforward to verify it by introducing dual verifiers. What we do in the sequel is to propose a certain choice of dual verifiers based on (34) which is easier to verify and still identifies a large class of consistent cases.

Considering (34), it is easy to recognize that some individual terms depend on noise and vanish in the noiseless case. This concerns a case where the noise data are processed by noisy LASSO, a dense grid and small λ . Then,

$$\boldsymbol{\sigma}(t) = -\lambda \mathbf{A}^\dagger \left(\sum_l \mathbf{d}_l s_{l,0}(t) \pi_{0,l} \right) - \lambda (\mathbf{A}^H \mathbf{A})^{-1} \boldsymbol{\gamma}(t) = \lambda \boldsymbol{\sigma}_0(t) \quad (35)$$

$$\boldsymbol{\pi}_0 = \mathbf{R}^{-1} \boldsymbol{\delta} \quad (36)$$

where $\boldsymbol{\pi}_0$ is a vector consisting of the elements $\pi_{0,l}$ and (34) can also be characterized by

$$\lambda^2 \sum_t \left| \mathbf{a}^H(\theta) \left(\sum_l (\mathbf{P} \mathbf{d}_l \gamma_l(t) \pi_{0,l} - \mathbf{A} (\mathbf{A}^H \mathbf{A})^{-1} \boldsymbol{\gamma}(t)) \right) \right|^2 \leq \lambda^2 \quad (37)$$

As (34) identifies existence of false alarm, (37) naturally identifies a case where application of noisy LASSO results in no false alarm. We call this case purely consistent. However, (37) implies pure consistency only when consistency is priorly established. Fortunately, it can also be seen that (37) automatically implies consistency as the vector

$$\mathbf{z}(t) = \sum_k \sigma_{k,0}(t) \mathbf{a}_k + \sum_k \pi_{k,0} s_{k,0}(t) \mathbf{d}_k \quad (38)$$

would then satisfy Theorem 6 by direct calculation. The following theorem summarizes and completes the above discussion.

Theorem 8. a) *A decomposition $A = \{\theta_k, \{s_k(t)\}\}$ is consistent, i.e. LASSO estimates for its corresponding observation by a sufficiently small noise is arbitrarily close to A if (34) holds, in which case it is also purely consistent.*

b) *Any consistent decomposition is a subset of a purely consistent decomposition.*

Proof. See Appendix 11. □

4.2 Statistical Properties of Perturbations

Let us assume that a consistent true decomposition is observed by the model in (2) and the noise perturbation is so small and the grid is so dense that Theorem 7 characterizes the estimation error. Thus, we may analyze the statistical properties of the solution of (26) to understand the statistical behavior of the LASSO solution. It may be readily seen that the overall error properties π, σ as well as PFA are linked to false alarm, which subsequently depends on the choice of λ . On the other hand, the method of selecting λ is not inherent in the machinery of LASSO. Thus, we examine the previous results in terms of an arbitrary λ in some example cases. As already stated, the results are given in terms of the π, σ parameters and PFA.

We will discuss in the following two different cases of interest. In the first case, the true order is known and λ is adapted to provide an estimate of correct order. In the second one, the order is unknown and λ is fixed to meet a certain PFA criterion. In either case, (34) is useful as it characterizes when no false alarm occurs.

Known Order and Adaptive Regularization

When the number of parameters is known, λ may be selected based on the given data set to provide a correct number of estimates. In this case no false alarm is observed and thus λ satisfies (34). To investigate the best performance, we select smallest such value of λ and denote it by λ_b . In this case, λ becomes a function of the noise realization. Hence, it is a random variable. Remember that now $\boldsymbol{\nu}(t) = \mathbf{n}(t)$. Thus, the expressions for π and σ and their corresponding statistics can be easily calculated. The following theorem summarizes the final expressions.

Theorem 9. a) *The λ_b may be calculated by*

$$\lambda_b = \max_{\theta} \Lambda(\theta) \tag{39}$$

where $\Lambda(\theta)$ is the unique positive solution of the following equation for λ .

$$\sum_t \left| \mathbf{a}^H(\theta) \left(\mathbf{n}(t) - \sum_l (\mathbf{a}_l \sigma_l(t) + \mathbf{d}_l s_{l,0}(t) \pi_l) \right) \right|^2 = \lambda^2 \quad (40)$$

and π_l and σ_l are given in (30) applying $\boldsymbol{\nu} = \mathbf{n}$.

b) When the regularization parameter is selected as λ_b , the estimates have the following first-order statistical properties:

$$\mathcal{E}(\boldsymbol{\pi}) = \mathcal{E}(\lambda_b) \mathbf{R}^{-1} \boldsymbol{\delta} \quad (41)$$

$$\mathcal{E}(\boldsymbol{\sigma}) = \mathcal{E}(\lambda_b) (\mathbf{A}^H \mathbf{A})^{-1} \boldsymbol{\gamma}(t) \quad (42)$$

$$\text{Cov}(\boldsymbol{\pi}) = \mathbf{R}^{-1} \Re [(\mathbf{D}^H \mathbf{P} \mathbf{C} \mathbf{P} \mathbf{D}) \odot \Xi] \mathbf{R}^{-1} + \text{Var}(\lambda_b) \mathbf{R}^{-1} \boldsymbol{\delta} \boldsymbol{\delta}^T \mathbf{R}^{-1} \quad (43)$$

where \mathbf{C} denotes the covariance of the noise $\mathbf{n}(t)$.

Proof. The proof is given in Appendix 12. \square

The expression in (43) has an interesting interpretation. The first term is recognized as the error covariance of the ML θ estimates, which is also the Cramer-Rao Bound (CRB) in the high SNR case [46]. The second term, proportional to the regularization parameter, is the additional contribution due to the regularization. Note that in absence of dispersion, i.e. when each parameter corresponds to single estimate, the π parameters are equivalent to θ and thus the current analysis shows that ML is a special case of LASSO, where no dispersion and no regularization exists. We remind that in presence of dispersion, only the π parameters can be calculated by a first order approximation.

Unknown Order and fixed λ

When the order is unknown, λ may be fixed to set a balance between PFA and error parameters in the absence of false alarm. Although a data driven λ is still a valid choice, it remains out of scope of the current analysis. When false alarm occurs, there is no agreed definition of the performance. Thus, we consider the average error in $\boldsymbol{\pi}, \boldsymbol{\sigma}$ only in absence of false alarm, which can be mathematically written as

$$MSE_f = \mathcal{E}(\boldsymbol{\pi} \boldsymbol{\pi}^T \mid NFA) \quad (44)$$

where NFA denotes the event that no false alarm occurs. Together with PFA , the above constitutes the performance measure. Unfortunately, we have not been able to provide analytical expressions for this case, since assuming NFA changes the posterior distribution of $\mathbf{n}(t)$ in a non-tractable way. In the next chapter we show numerical calculations based on a Monte Carlo method for this case.

5 Numerical Results

We have previously formulated a parametric approach to analyze LASSO and provided the details for a high-SNR scenario. In this section, we examine our previous derivations in the case of ADP applied to DOA estimation. The numerical results can be categorized into two groups. In the first, the theoretical results are calculated by Monte Carlo techniques, reminding that some expectations could not be analytically calculated in the previous derivation. The second group compares the theoretical performance to that of some alternative methods. We consider the CLASS (atomic-norm denoising) implementation in [34, 37], only considering the frequency estimation problem (ULA in our case) with uniform samples.

5.1 Evaluation of Theoretical Performance

Equations (41) and the definition of MSE_f and false alarm in (44) and (34), respectively constitutes the analysis. However, evaluating them in practice needs a complicated numerical procedure. In particular, we are interested in calculating the first and second order statistics of λ_b as well as MSE_f and PFA by a Monte Carlo method, which provides the results in Figures 1 and 2.

Taking a closer look at the definition of λ_b in (39), one may suspect that under certain practical assumptions, many terms in (39) can be neglected such that λ_b can be approximated by λ_f given by

$$\lambda_f = \sqrt{\max_{\theta \in \Theta} \sum_t |\mathbf{a}^H(\theta) \mathbf{n}(t)|^2} \quad (45)$$

The statistics of λ_f is widely considered in the design of Constant-False-Alarm-Rate (CFAR) estimators. Note that unlike λ_b , λ_f is independent of the true decomposition, while still depending on the noise realization. The statistics of λ_f can also be analytically expressed in some asymptotic cases. Figure 1 shows the evaluated expected value for different dimensions of observation m , where LASSO is applied to data from a ULA (Fourier) manifold explained in Section 2. The true DOAs are fixed at electrical angles $[0 \ 2.5\pi/m]$ with corresponding amplitudes $[1 \ 1]$. The results are taken over 10000 trials. Figure 2 shows the variance with a similar setup. As seen, λ_f may be considered in practice as a good approximate value especially for a high number of sensors, where the relative error decreases.

For the case of fixed λ we calculated MSE_f and PFA by another MC experiment. We compared two different choices of DOA separation, namely $2.5\pi/m$ and $2.7\pi/m$, both with unit amplitudes. A single snapshot was

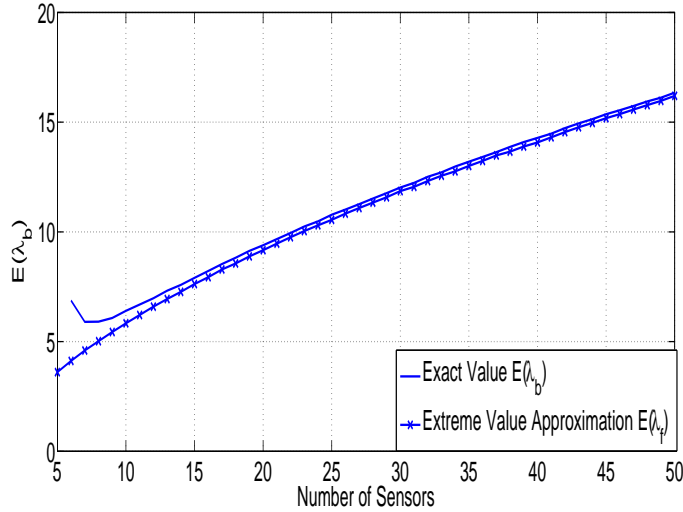


Figure 1: Mean value of λ_b compared to λ_f for different number of sensors.

considered and the SNR and m were set to 10 dB and 10 respectively. Figure 3 shows the resulting average PFA for different values of λ over 10^5 realizations. As seen, better separated sources need smaller value of λ to achieve a required PFA.

Then, Figure 4 shows the trade-off between MSE_f and PFA in the above scenarios. As seen, the error dramatically decreases by increasing the separation. Reaching to the separation of $3\pi/m$, the error practically approaches the Cramer-Rao bound in the desirable range of PFA . The same trend is observed when the number of sensors increases from 10 to 20.

5.2 Comparison with Other Methods

We finally compared the LASSO performance to that of ML (see (2)) with exhaustive search [47] and Conventional BeamForming (CBF) [50]. Figures 5 and 6 compare the estimate Mean Squared Errors and variances of three different estimators; CLASS, ML and CBF, respectively. The setup is similar to the one in Figures 2 and 3, while the number of sensors m is fixed to 15. The results are the average of the outcomes of 100 trials at each noise level. We see that while the asymptotic variances of CLASS and ML methods coincide, the CLASS estimator has a higher asymptotic MSE. We conclude that CLASS modifies the solution of ML mostly by adding a bias term in the very high SNR regime. However, as SNR decreases, the MSE of CLASS reaches the one for the ML estimator in the SNR regime between -2 and 5 dBs. There is a significant (almost 3 dB) difference between threshold edge of LASSO and ML. Note that ML with exhaustive search is not prac-

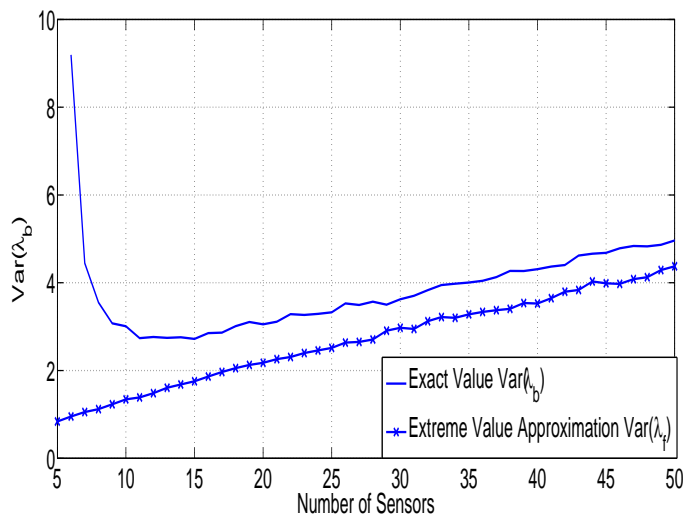


Figure 2: The variance of λ_b compared to λ_f for different number of sensors.

tical and the difference might be less with a more realistic implementation.

6 Concluding Remarks

This work was devoted to analysis of LASSO as a non-linear estimator of a parametric signal. The main idea here was to parameterize the support, which brought a parametric interpretation of LASSO. To meet the continuous estimation requirements, the parametric LASSO was modified to CLASS. This is similar in nature to the approach in [37]. The parametric CLASS estimates were then analyzed by linearization in a high-SNR case and related to the original estimates by LASSO. The numerical implementation of CLASS was out of the scope of the current work. However, [37] also provides a CLASS implementation for a specific case, which we employed for numerical validation.

The above approach enabled to analyze LASSO more deeply from a deterministic point of view, which is of a great interest in applications, where LASSO is utilized to estimate parameters, such as radar localization. Although, important properties of LASSO, especially the role of the RP, presented in a limited number of scenarios, the current work provides a framework for further investigations. The MSE calculations also provide a new insight to the role of the RP. With our approach, we were able to calculate MSE and the false alarm rate, which commonly characterize an estimator of varying order in the high-SNR case. The process of false alarms

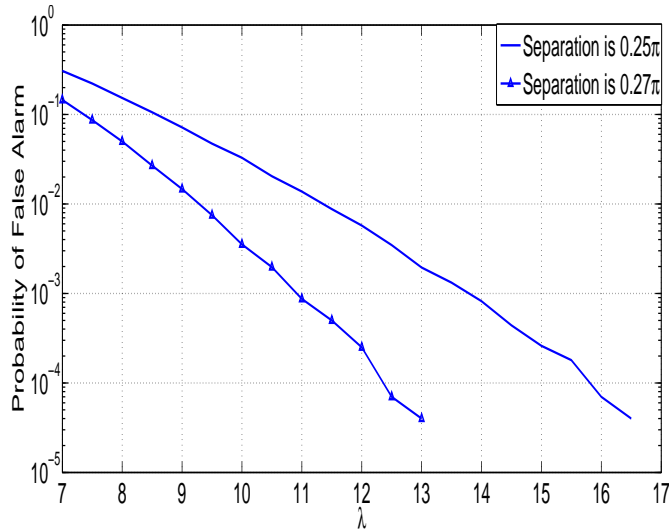


Figure 3: The PFA for different values of λ compared for different DOA separations.

were more generally characterized, but we left a more accurate investigation for a future work.

The current theoretical and numerical results suggest that LASSO provides a good trade-off between error and PFA, under some considerations about resolution. This is verified for a fixed-RP scenario. However, we suspect that employing a thresholding scheme would reduce PFA more, thus further improving the properties of LASSO. However, the numerical implementation may be crucial for the performance, and should therefore be the subject of a future study.

7 Appendix: LASSO Topology on ADP space

This part includes the definition of distance between atomic decompositions. Despite its complex technical definition it implies a natural concept, which easily follows from the analysis of LASSO.

Definition 2. (LASSO-topology)

a) Consider an irreducible decomposition $A = \{\{s_k(t)\}, \theta_k\}_{k=1}^n$ and another arbitrary decomposition $\bar{A} = \{\{\bar{s}_k(t)\}, \bar{\theta}_k\}_{k=1}^{\bar{n}}$. Let $I_k = (\theta_k - \epsilon, \theta_k + \epsilon)$ be the ϵ -ball at θ_k . Then, \bar{A} is said to be in an ϵ -neighborhood of A if

1. The ϵ -balls I_k cover all indexes of \bar{A} , i.e. $\{\bar{\theta}_k\} \subset \bigcup_{l=1}^n I_l$

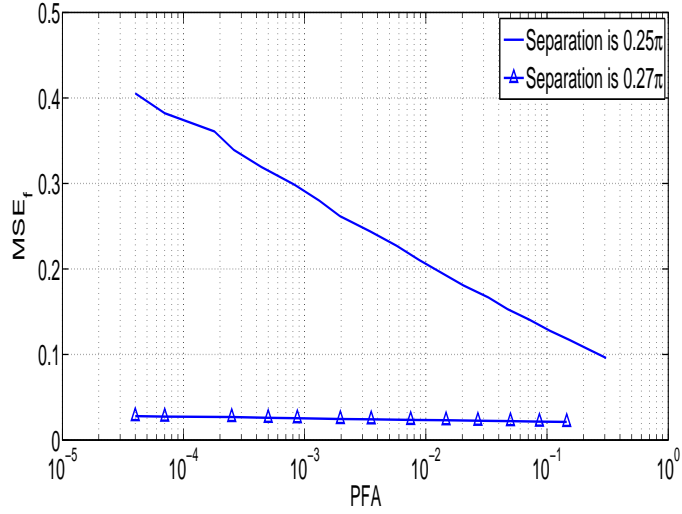


Figure 4: The PFA versus MSE for different values of λ compared for different DOA separations.

- For each interval I_k at each time index t , we have that

$$\left| s_l(t) - \sum_{k|\bar{\theta}_k \in I_l} \bar{s}_k(t) \right| < \epsilon \quad (46)$$

- For each \bar{k} and k the relation $\bar{\theta}_k \in I_k$ implies that there exists $\alpha_{k,\bar{k}} > 0$ such that

$$\forall t \quad |\alpha_{k,\bar{k}} s_k(t) - \bar{s}_{\bar{k}}(t)| < \epsilon \quad (47)$$

b) Two arbitrary decompositions B and \bar{B} are called ϵ -similar and shown by $A \sim_{\epsilon} \bar{A}$ if there exists an irreducible decomposition A such that both B and \bar{B} are in ϵ -neighborhood of A .

Figure 7 illustrates the concept of ϵ -neighborhood, where a decomposition is represented by a set of arrows, whose amplitudes show s , while their position denote θ . As seen, the definition does not restrict the orders. Condition 1 guarantees that the elements of $\{\bar{\theta}_k\}$ are concentrated around the elements of $\{\theta\}$. Then, Condition 2 provides that \bar{A} leads to a close synthesis to A through the model in (1). Finally, Condition 3 guarantees that the LASSO cost values in (7) for A and \bar{A} are close.

8 Appendix: Proof of Theorem 5

The proof is based on the following elements:

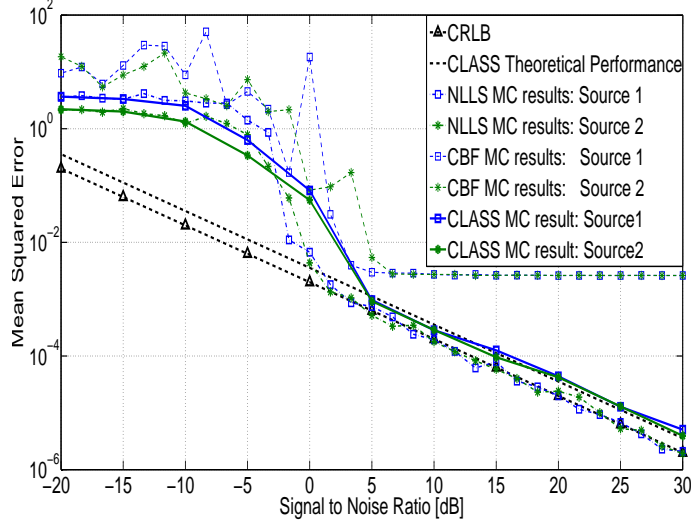


Figure 5: The statistical MSE for different methods versus input SNRs. The estimation is based on one snapshot measurement of two sources separated by $\Delta\theta = \frac{4\pi}{m}$, and waveform values $s_1 = s_2 = 1$.

1. For a regular manifold, there exists a finite subset $\{\theta_{b,1}, \theta_{b,2}, \dots, \theta_{b,p}\}$, such that the matrix $\mathbf{B} = [\mathbf{a}(\theta_{b,1}), \mathbf{a}(\theta_{b,2}), \dots, \mathbf{a}(\theta_{b,p})]$ is full rank.
2. The order of the LASSO and CLASS estimates are always bounded by $2mT$, i.e. the estimates are in \mathcal{M}_{2mT} .
3. For any R and n , the set \mathcal{M}_n^R of all decompositions with an order smaller than n and amplitudes smaller than R , i.e. $|s_k(t)| < R$, is compact in the regular topology of fixed dimension.
4. Define the synthesis function $f : \mathcal{M} \rightarrow \mathbf{C}^{m \times T}$ such that for any $X = \{\mathbf{x}(t)\}$ and $A = \{\theta_k, \{s_k(t)\}\}$ the relation $X = f(A)$ implies (1). Also define $\ell(A) = \|\{s_k(t)\}\|_{2,1}$. Then f and ℓ are continuous.
5. For arbitrary observations $X = \{\mathbf{x}(t)\}$, define also $\phi_{\tilde{\Theta}}(X)$ and $\phi(X)$ as the optimal values of the noiseless LASSO optimization in (10) and the noiseless CLASS in (20). Then, from the sparsity principle, we obtain that

$$\begin{aligned}
 \phi_{\tilde{\Theta}}(X) &= \min_{\tilde{\mathcal{M}}} \|\{s_k(t)\}\|_{2,1} \\
 &\quad \text{s.t.} \\
 \mathbf{x}(t) &= \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t)
 \end{aligned} \tag{48}$$

where the minimal point corresponds to the solution of LASSO (10) .

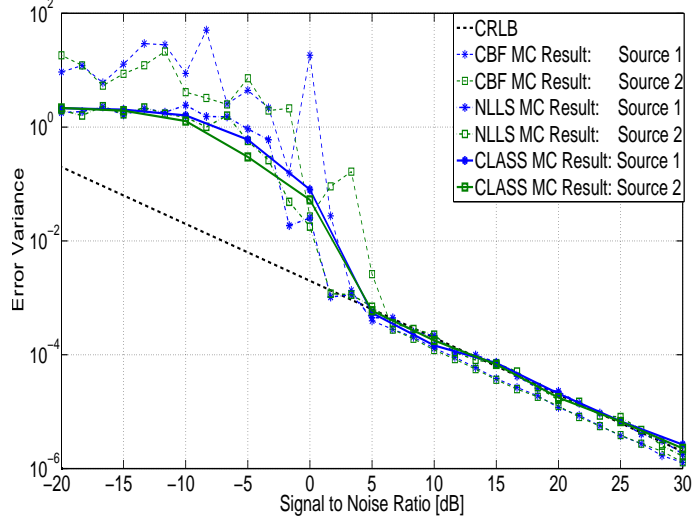


Figure 6: The statistical variance for different methods in different input SNRs. The estimation is based on one snapshot measurement of two sources separated by $\Delta\theta = \frac{4\pi}{m}$, and waveform values $s_1 = s_2 = 1$.

6. The function $\phi_{\tilde{\Theta}}$ is convex and conic, i.e. for any observation sets X, Y and $\alpha > 0$, we have that $\phi_{\tilde{\Theta}}(X + Y) \leq \phi_{\tilde{\Theta}}(X) + \phi_{\tilde{\Theta}}(Y)$ and $\phi_{\tilde{\Theta}}(\alpha X) = \alpha\phi_{\tilde{\Theta}}(X)$.

Using the above, the proof is straightforward. Note that from Observation 1, for any X the solution A to the group-LASSO optimization as well as the noiseless optimization is in \mathcal{M}_{2mT}^R , where R is a sufficiently large upper-bound on the amplitudes which only depends on X . If X is further bounded, then R is a constant.

Now, assume conversely that the theorem does not hold. This means that there exists an $\epsilon > 0$ such that for any of the values $\delta_r = 1/r$ there exists $X^{(r)} = \{\mathbf{x}(t) + \mathbf{n}^{(r)}(t)\}$, a δ_r -dense grid G_r and $\lambda_r < \delta_r$ such that $\|\mathbf{n}^{(r)}(t)\| < \delta_r$, but their corresponding group-LASSO solution A_r is out of the ϵ -neighborhood of any solution A to the noiseless CLASS for X . Since $X^{(r)}$ is bounded, there exists a fixed R , such that $A_r \in \mathcal{M}_{2mT}^R$. Now, from the second observation, we may assume without loss of generality that the sequence A_r has a limit $\bar{A} \in \mathcal{M}_{2mT}^R$, since otherwise one may take a converging subsequence. But \bar{A} is also out of the ϵ -neighborhood of any solution A of noiseless CLASS. We finally show in the sequel that in fact \bar{A} is contrarily equal to a solution A , which completes the proof.

To show that \bar{A} is a minimizer of noiseless CLASS, first note that $\mathbf{a}(\theta)$ is a continuous function over a compact set Θ . Thus, it is uniformly continuous. This means that for each value $\mu > 0$ there exists a $\delta > 0$ such that $|\theta_1 - \theta_2| < \delta$ implies that $\|\mathbf{a}(\theta_1) - \mathbf{a}(\theta_2)\| \leq \mu$. Fix a μ and corresponding

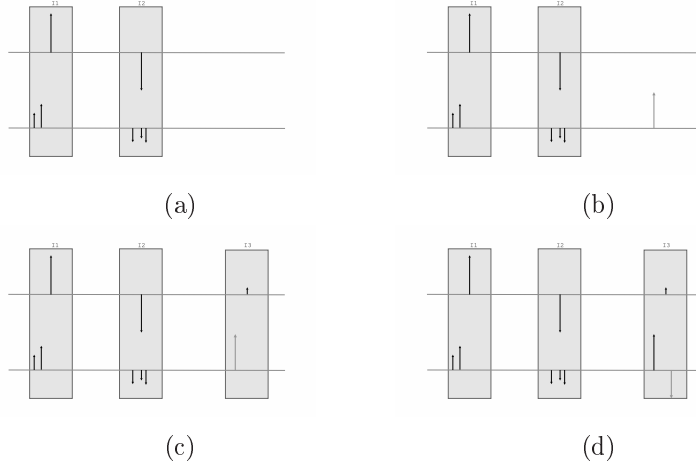


Figure 7: An illustration of the neighborhood concept: Two decompositions are shown, each by vertical arrows on a horizontal axes. The position of arrows shows θ , while their amplitude shows s . The two decompositions are neighbor in a), while in b),c) and d) Conditions 1,2 and 3 are violated, respectively.

δ . Consider the noiseless CLASS solution $A = \{(\theta_k, \{s_k(t)\})\}$ of X . As G_r is δ_r -dense, for each θ_k , there exists a $\hat{\theta}_k \in G_r$ such that $|\hat{\theta}_k - \theta_k| < \delta_r$. For sufficiently large r , this implies that $|\hat{\theta}_k - \theta_k| < \delta_r < \delta$, which further implies that $\|\mathbf{a}(\hat{\theta}_k) - \mathbf{a}(\theta_k)\| < \mu$. Take the approximate on-grid estimate $\hat{A}_r = \{(\hat{\theta}_k, \{s_k(t)\})\} \in \mathcal{M}_{G_r}$ and define $\hat{X}_r = \{\hat{\mathbf{x}}^{(r)}(t)\} = f(\hat{A}_r)$. Note that

$$\phi_r(\hat{X}_r) \leq \ell(\hat{A}_r) = \ell(A) = \phi(X) \quad (49)$$

where $\phi_r = \phi_{G_r}$ and the right-hand side of the inequality is the cost calculated at \hat{A}_r . On the other hand, for large r

$$\|\hat{\mathbf{x}}^{(r)}(t) - \mathbf{x}(t)\| = \left\| \sum_{k=1}^n (\mathbf{a}(\theta_k) - \mathbf{a}(\hat{\theta}_k)) s_k(t) \right\| \leq \mu \sum_{k=1}^n |s_k(t)| \leq \mu L \quad (50)$$

where L is a proper upper bound for $\sum_{k=1}^n |s_k(t)|$ over time, which always exists as the cost is bounded. This shows that

$$\lim_{r \rightarrow \infty} \hat{\mathbf{x}}^{(r)}(t) = \mathbf{x}(t) \quad (51)$$

Note also that the group-LASSO in the parametric form (17) can also be written as

$$\begin{aligned} \min_{\mathcal{M}_{G, \{\mathbf{y}(t)\}}} \frac{1}{2} \sum_{t=1}^T \|\mathbf{x}(t) - \mathbf{y}(t)\|_2^2 + \lambda \|\{s_k(t)\}\|_{2,1} \\ \text{s.t.} \\ \mathbf{y}(t) = \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t) \end{aligned} \quad (52)$$

which can be simplified to

$$\min_{Y = \{\mathbf{y}(t)\}} \frac{1}{2} \sum_{t=1}^T \|\mathbf{x}(t) - \mathbf{y}(t)\|_2^2 + \lambda \phi_G(Y) \quad (53)$$

Consider, $X = X^{(r)}$, $\lambda = \lambda_r$ and $G = G_r$. Then, the optimal point in (52) and (53) is given by A_r and $Y^{(r)} = \{\mathbf{y}^{(r)}(t)\} = f(A_r) = f(A_r)$ respectively. Then,

$$\begin{aligned} \frac{1}{2} \sum_{t=1}^T \|\mathbf{x}(t) + \mathbf{n}^{(r)}(t) - \mathbf{y}^{(r)}(t)\|_2^2 + \lambda_r \phi_r(Y^{(r)}) \\ \leq \frac{1}{2} \sum_{t=1}^T \|\mathbf{x}(t) + \mathbf{n}^{(r)}(t) - \hat{\mathbf{x}}^{(r)}(t)\|_2^2 + \lambda_r \phi_r(\hat{X}^{(r)}) \end{aligned} \quad (54)$$

The right hand side is the cost in (53) evaluated at $\hat{X}^{(r)}$. Then, using (49), we obtain that

$$\frac{1}{2} \sum_{t=1}^T \|\mathbf{x}(t) - \mathbf{y}^{(r)}(t)\|_2^2 \leq \frac{1}{2} \sum_{t=1}^T \|\mathbf{x}(t) - \hat{\mathbf{x}}^{(r)}(t)\|_2^2 + T\delta_r^2 + \lambda_r \phi(X) \quad (55)$$

Letting r tend to infinity, we get that

$$\lim_{r \rightarrow \infty} \mathbf{y}^{(r)}(t) = \mathbf{x}(t) \quad (56)$$

Until now, we have found observations Y^r converging to X such that A_r is the noiseless-LASSO solutions of Y^r over G_r , i.e $\ell(A_r) = \phi_r(Y_r)$. Note that from the continuity of f and ℓ we have that $f(\bar{A}) = X$ and

$$\ell(\bar{A}) = \lim_{r \rightarrow \infty} \ell(A_r) \quad (57)$$

Define $E_r = \{\mathbf{e}^{(r)}(t) = \mathbf{y}^{(r)}(t) - \hat{\mathbf{x}}^{(r)}(t)\}$. Then, E_r tends to zero as r tends to infinity. Then from observation 6,

$$\ell(A_r) = \phi_r(Y_r) \leq \phi_r(\hat{X}_r) + \phi_r(E_r) \leq \ell(\hat{A}_r) + \phi_r(E_r) = \ell(A) + \phi_r(E_r) \quad (58)$$

The final observation is that $\phi_r(E_r)$ vanishes as r tends to infinity. To see this consider the set in observation 1 and note that for an arbitrary μ and a large r there exists indexes $\hat{\theta}_{b,k}^r$ such that $\|\mathbf{a}(\hat{\theta}_{b,k}^r) - \mathbf{a}(\theta_{b,k})\| < \mu$. Define $\hat{\mathbf{B}}^r = [\mathbf{a}(\hat{\theta}_{b,1}^r) \dots \mathbf{a}(\hat{\theta}_{b,p}^r)]$. As the set of full-rank matrices is open, μ can be selected such that $\hat{\mathbf{B}}^r$ is full rank. Then,

$$\phi_r(E_r) \leq \sum_t \|(\hat{\mathbf{B}}^r)^\dagger \mathbf{e}^{(r)}(t)\|_2 \quad (59)$$

which tends to zero as $\mathbf{e}^{(r)}$ vanishes and the pseudo inverse $(\hat{\mathbf{B}}^r)^\dagger$ stays bounded in the μ -neighborhood of \mathbf{B} . Finally as $\phi_r(E_r)$ tends to zero, taking the limit of (58) and combining with (57) we conclude that.

$$\ell(\bar{A}) \leq \ell(A) \quad (60)$$

which shows that \bar{A} is a minimizer of the noiseless CLASS.

9 Appendix: Proof of Theorem 6

First, note that for a regular manifold $\mathbf{a}(\theta)$, the set

$$L = \left\{ \{\mathbf{z}(t)\} \mid \forall \theta \quad \sum_{t=1}^T |\mathbf{a}^H(\theta) \mathbf{z}(t)|^2 \leq 1 \right\} \quad (61)$$

is compact. For any grid G , define

$$L_G = \left\{ \{\mathbf{z}(t)\} \mid \forall \theta \in G \quad \sum_{t=1}^T |\mathbf{a}^H(\theta) \mathbf{z}(t)|^2 \leq 1 \right\} \quad (62)$$

Then there exists a value δ such that for every δ -dense grid G , the set L_G is compact. Furthermore, for any value $\mu > 0$, there exists a δ value such that for every δ -dense grid G ,

$$L \subseteq L_G \subseteq L^\mu \quad (63)$$

where L^μ denotes the union of all closed μ -neighborhoods of elements in L . Note also that L^μ is compact.

Now, consider an arbitrary solution $A = \{(\theta_k, \{s_k(t)\})\}$ of the noiseless CLASS. Take a sequence of $\delta_r = 1/r$ -dense grids G_r such that $\theta_k \in G_r$ for all k and r . Then, clearly $A \in \mathcal{M}_{G_r}$ and thus it minimizes noiseless group-LASSO over G_r . From Theorem 1, this means that there exists a sequence of dual vectors $Z_r = \{\mathbf{z}_r(t)\} \in L_{G_r}$ such that

$$\mathbf{a}^H(\theta_k) \mathbf{z}_r(t) = \frac{s_k(t)}{p_k} \quad (64)$$

Note that for any fixed μ and sufficiently large r we have that $Z_r \in L^\mu$. Thus, Z_r has a subsequence converging to a point $Z \in L^\mu$. Then

$$Z \in \bigcap_{\mu} L^\mu = L \quad (65)$$

since the choice of μ is arbitrary. clearly Z also satisfies the other condition in (64).

Conversely, suppose that there exists $Z_r = \{\mathbf{z}(t)\} \in L$ satisfying (64) for A . Then, we show that A is the global minimum of noiseless-CLASS. Take any other decomposition $B = \{(\theta'_l, \{s'_l(t)\})\}$ with $f(B) = f(A) = X$. Take the grid $G = \{\theta'_l\} \cup \{\theta_k\}$. Note that taking the dual verifiers in Z , the conditions of Theorem 1 for the noiseless case is satisfied. Thus, A a minimizer of noiseless LASSO for grid G and input X , which implies that $\ell(A) \leq \ell(B)$.

Finally, let us prove convergence. Suppose conversely that taking $\delta_r = 1/r$, there exists a sequence of primal solutions A_r with corresponding dual parameters Z_r to the group lasso with a perturbed input $\{\mathbf{x}(t) + \mathbf{n}_r(t)\}$ where $\|\mathbf{n}_r(t)\|_2 \leq \delta_r$, $\lambda_r < \delta_r$ and over the δ_r -dense grid, such that Z_r is not in a ϵ -neighborhood of any dual vector of the noiseless CLASS solution. But since Z_r can be bounded in a compact set for large enough r and due to Theorem 5, the sequences has a subsequence converging to A and Z respectively. Since Z_r is ϵ -distant from any dual solution of noiseless CLASS, the limit is so. But, it is simple to check that the conditions of the current theorem holds for Z , which implies that Z is a dual for A . This is a contradiction and completes the proof.

10 Appendix: Proof of Theorem 7

First, let us explain part (a) with more details. Convergence means that:

For any $\omega > 0$, there exists a $\delta > 0$ such that if the grid is δ -dense, $\lambda < \delta$, and perturbations satisfy $\|\mathbf{n}(t)\| < \delta$ and the solution A is in ϵ -neighborhood of a noiseless solution A_0 such that $\epsilon < \delta$, then the ϵ -false alarms are in a $\omega\delta$ -neighborhood of \bar{A} and $|\pi_l - \bar{\pi}_l| < \omega\delta$ and $|\sigma_l(t) - \bar{\sigma}_l(t)| < \omega\delta$ hold.

Now, to prove this, we follow the following steps:

1. Suppose that $G = (\boldsymbol{\pi}, \boldsymbol{\sigma}, \bar{A})$ minimizes g for a certain choice of $\mathbf{n}(t)$, and true parameters and $H = (\boldsymbol{\pi}', \boldsymbol{\sigma}', \bar{A}')$ is another non-optimal point. Then, there exists a constant K depending only on true parameters such that

$$g(\boldsymbol{\pi}', \boldsymbol{\sigma}', \bar{A}') - g(\boldsymbol{\pi}, \boldsymbol{\sigma}, \bar{A}) \geq (G, H)^2 \quad (66)$$

2. Consider ϵ and δ such that the solution of LASSO optimization with a δ -dense grid $\tilde{\Theta}$, $\|\mathbf{n}(t)\| < \delta$ and $\lambda < \delta$ is in an ϵ -neighborhood of the noiseless (true) solution. Denote by $\boldsymbol{\pi}_m, \boldsymbol{\sigma}_m, \bar{A}_m$ the corresponding parameters to the optimal point of LASSO with the optimal cost f_{\min} . Then,

$$|f_{\min} - g(\boldsymbol{\pi}_m, \boldsymbol{\sigma}_m, \bar{A}_m)| < K_1 \epsilon \|\mathbf{n}\| \delta \quad (67)$$

3. Consider the same setup as above and remember that $\boldsymbol{\pi}, \boldsymbol{\sigma}$ and \bar{A} minimize g . Take the optimization

$$\begin{aligned} & \min \|\{\tilde{s}_k(t)\}\|_{1,2} \\ & \text{s.t.} \\ & \sigma_l(t) = \sum_{k|\|\tilde{\theta}^k - \theta_{l,0}\| < \epsilon} \tilde{s}_k(t) - s_{l,0}(t) \\ & \pi_l \gamma_l(t) = \sum_{k|\|\tilde{\theta}^k - \theta_{l,0}\| < \epsilon} \tilde{s}_k(t) (\tilde{\theta}^k - \theta_{l,0}) \end{aligned} \quad (68)$$

and note that it has a solution $\{\tilde{s}_k(t)\}$ with only two active elements in each cloud. Take this solution and calculate the original LASSO cost f at this point. Then

$$|f - g(\boldsymbol{\pi}, \boldsymbol{\sigma}, \bar{A})| < K_2 \epsilon \|\mathbf{n}\| \delta \quad (69)$$

4. Putting (69) and (67) together, it is simple to conclude that

$$g(\boldsymbol{\pi}_m, \boldsymbol{\sigma}_m, \bar{A}_m) - g(\boldsymbol{\pi}, \boldsymbol{\sigma}, \bar{A}) < K_3 \epsilon \delta^2 \quad (70)$$

5. Now, if (a) is not correct then there exists a ω such that for any arbitrary δ there exists a δ -exact case such that $d(G, Gm) > \omega \delta$. Consider now that $\epsilon < K\omega^2/K_3$ we get from (69) that

$$K\omega^2 \delta^2 < g(\boldsymbol{\pi}_m, \boldsymbol{\sigma}_m, \bar{A}_m) - g(\boldsymbol{\pi}, \boldsymbol{\sigma}, \bar{A}) < K_3 \epsilon \delta^2 \quad (71)$$

which leads to $K\omega^2/K_3\omega^2 < \epsilon$ and contradicts to the choice of ϵ . Thus, (a) holds.

6. Relations (69) and (67) imply

$$f - f_{\min} < K\epsilon \delta^2 \quad (72)$$

7. Similar to step 1 if $\|\{\tilde{s}(t) - \tilde{s}_m(t)\}\|_{\infty} = d$, then one can conclude that

$$f - f_{\min} > K_4 d^2 \quad (73)$$

8. Finally for any ω and sufficiently small δ , the relation $\|\{\tilde{s}(t) - \tilde{s}_m(t)\}\|_{\infty} < \omega \delta$ must hold otherwise (72) and (73) will contradict again for small choice of ϵ and δ . Then, $\delta_3 < \|\{\tilde{s}(t) - \tilde{s}_m(t)\}\|_{\infty}$ proves the result.

11 Appendix: Proof of Theorem 8

For part (a), it is easy to plug (38) in Theorem 6 and check by direct calculation that (34) ensures optimality of the true parameters,

For (b), since $\{\theta_k, \{s_k(t)\}\}$ is consistent the optimization

$$\begin{aligned} & \min_{\{\mathbf{z}(t)\}} \sum_t \|\mathbf{z}(t)\|_2^2 \\ & \text{s.t.} \\ & \sum_t |\mathbf{z}^H(t)\mathbf{a}(\theta)|_2^2 \leq 1 \quad \mathbf{a}^H(\theta_k)\mathbf{z}(t) = \gamma_k(t) = \frac{s_k(t)}{\sqrt{\sum_t |s_k(t)|^2}} \end{aligned} \quad (74)$$

is feasible and has solution \mathbf{z}' . It is simple to see that from the KKT theorem \mathbf{z}' can be written as

$$\mathbf{z}' = \sum_l \mathbf{a}(\theta'_l) r_l \gamma'_l(t) + \sum_k \mathbf{a}(\theta_k) u_k \quad (75)$$

where $\{\theta'_l\}$ is the set of all peaks of the spectrum $|\mathbf{a}^H(\theta)\mathbf{z}'|$, thus including θ_k , and r_l, u_k are suitable dual parameters. This shows that $\mathbf{z}'(t)$ is in the range space of \mathbf{A}' consisting of $\mathbf{a}(\theta'_l)$ as columns, i.e

$$\mathbf{z}(t) = \mathbf{A}' \boldsymbol{\sigma}'(t) \quad (76)$$

Furthermore,

$$\mathbf{a}^H(\theta'_l)\mathbf{z}'(t) = \gamma'_l(t) \rightarrow \mathbf{A}'^H \mathbf{z}(t) = \boldsymbol{\gamma}'(t) \quad (77)$$

and

$$\frac{\partial \sum_t |\mathbf{z}^H(t)\mathbf{a}(\theta)|_2^2}{\partial \theta} \Big|_{\theta=\theta'_l} = 0 \rightarrow \sum_t \Re(\gamma'_l(t) \mathbf{d}^H(\theta'_l)\mathbf{z}(t)) = 0 \quad (78)$$

It is easy by direct calculation to show that (76),(77) and (78) may only hold if $\boldsymbol{\sigma}'$ is equal to $\boldsymbol{\sigma}_0$ in (35) if \mathbf{A} and $\boldsymbol{\gamma}$ are replaced by their primed counterparts and the resulting $\boldsymbol{\delta}$ is zero. Then, similar to part (a), the optimality condition directly leads to (37) which establishes pure consistency for $\{\theta'_l, \{s_l(t) = \gamma'_l(t)\}\}$.

12 Appendix: Proof of Theorem 9

a) By definition, λ_b can be written as

$$\begin{aligned}\lambda_b &= \min\{\lambda \mid \forall \theta \sum_t \left| \mathbf{a}^H(\theta) \left(\mathbf{n}(t) - \sum_l (\mathbf{a}_l \sigma_l(t) + \mathbf{d}_l \gamma_l(t) \pi_l) \right) \right|^2 \leq \lambda^2\} \\ &= \min \underbrace{\bigcap_{\theta} \left\{ \lambda \mid \sum_t \left| \mathbf{a}^H(\theta) \left(\mathbf{n}(t) - \sum_l (\mathbf{a}_l \sigma_l(t) + \mathbf{d}_l \gamma_l(t) \pi_l) \right) \right|^2 \leq \lambda^2 \right\}}_{S_{\theta}}\end{aligned}\quad (79)$$

Note that the term $\mathbf{a}_l \sigma_l(t) + \mathbf{d}_l \gamma_l(t) \pi_l$ is linear in λ . Thus, $S_{\theta} = \{\lambda \mid P_{\theta}(\lambda) \leq 0\}$ where $P_{\theta}(\lambda)$ is a quadratic function of λ . Note that if the case is purely consistent the leading term in P_{θ} can be shown by calculation to be negative. Furthermore $P_{\theta}(0) > 0$. Thus, P_{θ} has exactly one positive root $\Lambda(\theta)$, given by (40), and $S_{\theta} = [\Lambda(\theta) \infty)$, leading to

$$\lambda_b = \min_{\theta} \bigcap [\Lambda(\theta) \infty) = \min_{\theta} [\max_{\theta} \Lambda(\theta) \infty) = \max_{\theta} \Lambda(\theta) \quad (80)$$

b) The result follows from direct calculation and noting that the expression $\mathcal{E}(n(t)\lambda_b(n(t)))$ is zero. To see this follow the following steps

1. Note that $\lambda_b = \lambda_b(\{\mathbf{n}(t)\})$ is a conic function of the noise, i.e. the expression $\lambda_b(\{\alpha \mathbf{n}(t)\})$ equals $|\alpha| \lambda_b(\{\mathbf{n}(t)\})$.

2. Then,

$$\mathcal{E}(\mathbf{n}(t) \mid \lambda_b) = \lambda_b \mathcal{E}(\mathbf{n}(t) \mid \lambda_b = 1) \quad (81)$$

3. Note that $\mathcal{E}(\mathbf{n}(t) \mid \lambda_b = 1) = 0$, since

$$0 = \mathcal{E}(\mathbf{n}(t)) = \mathcal{E}_{\lambda_b}(\mathcal{E}(\mathbf{n}(t) \mid \lambda_b)) = \mathcal{E}(\mathbf{n}(t) \mid \lambda_b = 1) \mathcal{E}(\lambda_b) \quad (82)$$

4. Finally,

$$\mathcal{E}(n(t)\lambda_b) = \mathcal{E}_{\lambda_b}(\lambda_b \mathcal{E}(\mathbf{n}(t) \mid \lambda_b)) = \mathcal{E}(\mathbf{n}(t) \mid \lambda_b = 1) \mathcal{E}(\lambda_b^2) = 0 \quad (83)$$

References

- [1] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, pp. 714–721, Mar. 2009.

- [2] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, "Recovering sparse signals using sparse measurement matrices in compressed dna microarrays," *IEEE, J. Select. Topics Signal Processing*, vol. 2, pp. 275–285, June 2008.
- [3] W. Tu and S. Sun, "Spatial filter selection with lasso for EEG classification," in *Advanced Data Mining and Applications*, Chongqing, China, 2010, pp. 142–149.
- [4] H. Konno and H. Yamazaki, "Mean-absolute deviation portfolio optimization model and its applications to tokyo stock market," *Manage. Sci.*, vol. 37, pp. 519–531, May 1991.
- [5] H. Yao, P. Gerstoft, P. M. Shearer, and C. Mecklenbräuker, "Compressive sensing of the tohoku-oki mw 9.0 earthquake: Frequency-dependent rupture modes," *Geophys. Res. Lett.*, vol. 38, Oct. 2011.
- [6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., Series B (Methodological)*, vol. 58, pp. 267–288, Jan. 1996.
- [7] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, Dec. 1998.
- [8] J. J. Fuchs, "On the application of the global matched filter to doa estimation with uniform circular arrays," vol. 49, no. 4, pp. 702–709, Apr. 2001.
- [9] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Select. Topic. Signal Processing*, vol. 1, pp. 586–597, Dec. 2007.
- [10] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.
- [11] T. Park and G. Casella, "The bayesian lasso," *J. Amer. Stat. Assoc.*, vol. 103, pp. 681–686, 2008.
- [12] H. Zayyani, M. Babaie-Zadeh, and C. Jutten, "Bayesian pursuit algorithm for sparse representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 1549–1552.

- [13] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” vol. 41, pp. 3397–3415, Dec. 1993.
- [14] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” vol. 53, pp. 4655–4666, Dec. 2007.
- [15] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, “Least angle regression,” *Ann. Stat.*, vol. 32, pp. 407–499, Apr. 2004.
- [16] D. Needell and J. A. Tropp, “Cosamp: Iterative signal recovery from incomplete and inaccurate samples,” *Elsevier, Appl. Comput. Harmon. Anal.*, vol. 26, pp. 301–321, May 2009.
- [17] M. Lustig, D. Donoho, and J. M. Pauly, “Sparse MRI: The application of compressed sensing for rapid MR imaging,” *Resonance Med. Mag.*, vol. 58, pp. 1182–1195, Dec. 2007.
- [18] J. Provost and F. Lesage, “The application of compressed sensing for photo-acoustic tomography,” *Medical Imaging, IEEE Transactions on*, vol. 28, no. 4, pp. 585–594, 2009.
- [19] Y. C. Eldar and M. Mishali, “Robust recovery of signals from a structured union of subspaces,” vol. 55, no. 11, pp. 5302–5316, 2009.
- [20] M. Mishali and Y. 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.
- [21] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, “Beyond nyquist: Efficient sampling of sparse bandlimited signals,” vol. 56, pp. 520–544, Jan. 2010.
- [22] D. Malioutov, M. Çetin, and A. S. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 3010–3022, 2005.
- [23] R. G. Baraniuk, “Compressive sensing [lecture notes],” vol. 24, pp. 118–121, July 2007.
- [24] M. Arigovindan, M. Suhling, P. Hunziker, and M. Unser, “Variational image reconstruction from arbitrarily spaced samples: A fast multiresolution spline solution,” vol. 14, pp. 450–460, Apr. 2005.

- [25] P. Milanfar, W. C. Karl, and A. S. Willsky, "A moment-based variational approach to tomographic reconstruction," vol. 5, pp. 459–470, Mar. 1996.
- [26] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *Signal Processing, IEEE Transactions on*, vol. 57, no. 6, pp. 2275–2284, 2009.
- [27] P. Stoica, A. Jakobsson, and J. Li, "Matched-filter bank interpretation of some spectral estimators," *Signal Processing*, vol. 66, no. 1, pp. 45–59, 1998.
- [28] P. Zhao and B. Yu, "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.
- [29] D. Donoho, "Compressed sensing," vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [30] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Ann. Stat.*, vol. 37, pp. 2145–2177, Oct. 2009.
- [31] E. J. Candès and Y. Plan, "A Probabilistic and RIPless Theory of Compressed Sensing," vol. 57, no. 11, pp. 7235–7254, Nov. 2011.
- [32] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized lasso: A precise analysis," *arXiv preprint arXiv:1311.0830*, 2013.
- [33] E. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *arXiv preprint arXiv:1203.5871*, 2012.
- [34] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressive sensing off the grid," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 778–785.
- [35] Z. Ben-Haim and Y. C. Eldar, "The cramer-rao bound for sparse estimation," *arXiv preprint arXiv:0905.4378*, 2009.
- [36] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," vol. 38, no. 10, pp. 1783–1795, 1990.
- [37] B. N. Bhaskar and B. 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.

- [38] M. V. Ashkan Panahi, “Gridless compressive sensing,” in *IEEE Int. Conf. Acoust. Speech, Signal Processing*, 2014.
- [39] V. M. H. B. Panahi, Ashkan, “A numerical approach to gridless compressed sensing,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.
- [40] J. Li and P. Stoica, “Efficient mixed-spectrum estimation with applications to target feature extraction,” *Signal Processing, IEEE Transactions on*, vol. 44, no. 2, pp. 281–295, 1996.
- [41] J. A. Fessler and A. O. Hero, “Space-alternating generalized expectation-maximization algorithm,” *Signal Processing, IEEE Transactions on*, vol. 42, no. 10, pp. 2664–2677, 1994.
- [42] D. L. Donoho and M. Elad, “Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.
- [43] L. L. Scharf, *Statistical signal processing*. Addison-Wesley Reading, MA, 1991, vol. 98.
- [44] P. Stoica and Y. Selen, “Model-order selection: a review of information criterion rules,” *Signal Processing Magazine, IEEE*, vol. 21, no. 4, pp. 36–47, 2004.
- [45] J. Rissanen, “A universal prior for integers and estimation by minimum description length,” *The Annals of statistics*, pp. 416–431, 1983.
- [46] P. Stoica and N. Arye, “Music, maximum likelihood, and cramer-rao bound,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 5, pp. 720–741, 1989.
- [47] B. Ottersten, M. Viberg, P. Stoica, and A. Nehorai, “Exact and large sample ML techniques for parameter estimation and detection in array processing,” in *Radar Array Processing*, Haykin, Litva, and Shepherd, Eds. Berlin: Springer-Verlag, 1993, pp. 99–151.
- [48] M. R. Osborne, B. Presnell, and B. Turlach, “A new approach to variable selection in least squares problems,” 1999.
- [49] H. Zhu, G. Leus, and G. B. Giannakis, “Sparsity-cognizant total least-squares for perturbed compressive sampling,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 5, pp. 2002–2016, 2011.

- [50] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," vol. 13, pp. 67–94, July 1996.

Paper 3

A Novel Sparsity-Based Approach to Recursive Estimation of Dynamic Parameter Sets

A. Panahi, M. Viberg

To be Submitted to IEEE transactions on signal processing

A Novel Sparsity-Based Approach to Recursive Estimation of Dynamic Parameter Sets

A. Panahi, M. Viberg

Abstract

We consider the problem of estimating a variable number of parameters with a dynamic nature. A familiar example is finding the position of moving targets using sensor array observations. The problem is challenging in cases where either the observations are not reliable or the parameters evolve rapidly. Inspired by the sparsity based techniques, we introduce a novel Bayesian model for the problems of interest and study its associated recursive Bayesian filter. We propose an algorithm approximating the Bayesian filter, maintaining a reasonable amount of calculations. We compare by numerical evaluation the resulting technique to state-of-the-art algorithms in different scenarios. In a scenario with a low SNR, the proposed method outperforms other complex techniques.

1 Introduction

Estimating a dynamic set of parameters is a highly useful and wide area of research, with a long and fruitful history [1]. Indeed, noticing the ever increasing application of the Kalman filter and its variants to many newly developed technologies is enough to understand the importance of this topic. In this context, the quest for modified techniques usually concerns cases where either the currently existing methods fail to meet the computational limitations, or result in an insufficient precision. The latter may also be either due to an inconsistent model, on which the technique is based, or simply because of improper approximations. From this perspective, one finds certain estimation problems, for example the ones concerning data generated by a sensor array, more challenging. The reason is that the associated models, being capable of capturing the desired properties of the parameters, are so complicated that standard design methods by them lead to computationally intractable techniques. Thus, appealing to proper approximations is inevitable in those cases. This paper addresses these problems and aims to provide a modified approximate estimation technique. The emphasis here

is on maintaining a low computational complexity, while maintaining the statistical properties of the estimates.

The central idea in estimating a time varying parameter is that a parameter following a well-structured temporal model has locally correlated samples. Thus, they can be fused to improve the quality of estimation for a specific sample. This is particularly known as parameter filtering [2]. The basic ideas of filtering can be easily observed in the pioneering studies of Wiener, initiating the field of adaptive filtering [3]. Later, the seminal work of Kalman framed adaptive filtering into a rigorous statistical context, and showed a case, where statistically efficient estimates could be exactly calculated by a recursive method [4]. Soon after Kalman, Ho and Lee generalized this idea to the so called Markov Chain (MC) models, comprising of parameter evolution and measurement models [5]. Their solution is generally called Recursive Bayesian Filtering (RBF). The main advantage of the RBF is that it is highly adaptive to different application specifications, including a non-stationary behavior [6]. However, it requires storing and integrating posterior densities. Approximate techniques such as the Extended Kalman Filter (EKF) [7,8] and Unscented Kalman Filter (UKF) [9] are commonly used to perform this. Due to their local nature, they perform poorly, when multi-modal distributions are considered. The advent of statistical sampling and Monte Carlo methods provided an alternative method of implementing recursive Bayesian filters, by the so called Markov Chain Monte Carlo (MCMC) method. The resulting filter is generally known as the particle filter [10–12].

The difficulty arises in applying the above to problems such as radar detection, where the data is generated by a sensor array. This is due to multiple reasons, discussed in the sequel. The first reason is that a MC model is not directly applicable. To elaborate on this, note that the corresponding measurement model for data generated by a sensor array consists of two distinct set of parameters, known as amplitude and position parameters. In many applications, the amplitudes evolve rapidly in time, resulting in highly uncorrelated samples. Thus, only in the sense of position parameters one may perceive a Bayesian filter. To remain in the realm of RBF, it is still necessary to handle the amplitudes in a Bayesian manner. The second reason is that the observation model of the applications of interest is nonlinear, and estimation through them usually leads to the local minima problem. In the same manner, nonlinearity results in posterior multimodality, which not only complicates estimation, but also makes the posterior calculation difficult. The third reason is related to the fact that the time evolution model of the position parameters concerns varying order. Take the radar example. In the course of observation, it is perceivable that some

targets may be introduced or removed from the observation scene. In a more elaborate model, a single target may spawn multiple future targets. A MC model capturing the dynamics of such a system is complex and its corresponding sequential Bayesian filter can only be derived in an abstract form. To reduce the computational cost without introducing too much error, this filter needs to be approximately parametrized. This is generally a challenging task.

1.1 Literature Survey

Due to the above, one may find different approaches in the literature to recursive filtering of the sensor array data. According to different representations of the problem of interest, these methods are developed under different names. More specifically, the parametric (Kalman filter-based), spectral-based and subspace-based representations give rise to filtering techniques under similar titles. Some spectral based techniques can be found, e.g in [13–15]. The subspace tracking approaches have also been recently studied and applied in the literature [16–18]. The semi-parametric sparsity-based techniques are also rapidly emerging in literature under the title of sparsity tracking [19–21]. The filtering techniques can be also categorized from a different perspective. Many studies consider a case where preliminary parameter estimates are provided, relying only on their corresponding data. This is called target tracking and is favorable in occasions, such as some radar detection problems, where only the preliminary estimates are accessible for process [22–24]. In contrast to target tracking, the recent attempts to directly use sensor data to perform parameter filtering is often referred to as Track-Before-Detect (TBD), but this is not a generic term [25, 26].

The above techniques deal with the aforementioned difficulties in different ways. The target tracking and the subspace tracking techniques do not suffer from lack of amplitude models, while other TBD approaches either assume a specific amplitude model, depending on the application or eliminate them by assuming a Bayesian model and integration [27]. The amplitude models usually involve hyper-parameters, for which simple time evolution models are considered. The parametric formulation is the most precise likelihood based approach [27], but is numerically sensitive to nonlinearity. The Joint Probabilistic Data Association (JPDA) and Probabilistic MultiHypothesis Tracker PMHT [28] methods are popular examples of parametric target tracking [29, 30]. Instead, the methods leading to spectral estimation such as subspace-based and sparsity-based techniques trade off precision in favor of numerical stability. Moreover, particle filtering is nowadays a common approach to overcome multi-modality [31]. Concerning the issue

with variable order, many related studies consider a fairly general model, where the parameters have a fixed probability to survive, disappear or appear at the next time instant. In the recent literature, this is formulated as a Random Finite Set (RFS) model, also considered here and referred to as the standard model [32]. The RFS based representation not only provides a formal definition of the time evolution model, but also suggests certain approximation techniques. For example the Probability Hypothesis Density (PHD) filter provides a method to overcome the so-called data association problem in target tracking through approximating the RFS-based posteriors by a Poisson process [33]. The data association problem is due to the fact that the preliminary estimates are not generally labeled by their corresponding true parameter. More elaborate examples of such can be found in [34,35].

1.2 Motivation

In the problems of interest herein, the RBF approach needs to be approximated and the performance of all the techniques in the prequel is limited by the precision of their underlying approximation. From this perspective, these techniques can be divided into three groups: The locality based approaches such as EKF and UKF, the ones based on stochastic sampling, i.e. particle filters, and other model-based approximations such as the ones in the PHD filtering. The latter is normally based on minimizing the Kullback-Leibler (KL) distance between the resulting posteriors and a parametrized model set, which is applicable only if the minimization has a tractable solution. Clearly, the choice of approximation depends on the type of filter. For example, a locality based approximation is not appropriate for parametric filtering, where multiple local minima are present. In general, particle filters are always applicable, but need a higher computational effort (number of particles) than the other techniques to provide the desired precision. The precision of the methods such as the PHD filter depends on how well the approximate model fits to the exact one. Practically speaking, this restricts such methods to a high SNR or a slowly varying case. Moreover, the target tracking performance is also dependent on the quality of the preliminary estimates, which considerably limits the SNR range of application for these techniques.

In this paper, we study a different opportunity provided by the findings in the field of sparsity-based estimation, especially the Least Absolute Shrinkage and Selection Operator (LASSO) [36–38]. Recently, the inspiring work of Stoica et al in [38,39] has provided an important Bayesian insight into this approach, which we slightly modify here to fit the RFS framework.

Using this model for observation and considering the standard RFS based time evolution model, we investigate on the resulting RBF. The RBF is again intractable and needs approximation. On the other hand, it is observed that the convexity of LASSO yields to unimodality of the posterior distributions. Thus, it is favorable to use local approximations, similar to EKF. We develop a local expansion technique performed on the abstract space of finite sets and apply it to the proposed RFS, leading to a tractable filter.

1.3 Mathematical Notation

In this paper, \mathbb{R} , \mathbb{R}_+ and \mathbb{C} refer to the set of real, non-negative real and complex numbers, respectively. The notation $\text{Tr}(\cdot)$ denotes the trace operator and $|\cdot|$ shows either the absolute (in the case of a numerical argument), or the cardinality (in the case of a set argument) of the argument. Moreover, $(\cdot)_+$ denotes the positive part of its real argument. We also define an assignment R between finite sets A and B as a subset of $A \times B$ satisfying the following conditions

- $\forall (a_1, b_1) \in R, (a_2, b_2) \in R; \quad a_1 = a_2 \rightarrow b_1 = b_2$
- $\forall (a_1, b_1) \in R, (a_2, b_2) \in R; \quad b_1 = b_2 \rightarrow a_1 = a_2$

Moreover, we define the domain sets of R as the elements in A and B , included in R , i.e.

- $d_1(R) = \{a \in A \mid \exists b \in B, (a, b) \in R\}$
- $d_2(R) = \{b \in B \mid \exists a \in A, (a, b) \in R\}$

Throughout the paper, $+$ and $-$ subscripts or superscripts denote parameter values right after and before an observation, respectively. The primed parameters are usually related to the ones at a previous time instant. The notation $p(\cdot)$ denotes the probability density function (pdf) of its argument and $Q(\cdot, \cdot)$ represents the transitional probability between consecutive samples.

2 Problem Formulation

2.1 Observation Model

Consider a compact subset $\Theta \subset \mathbb{R}$ and a smooth basis manifold $\mathbf{a} : \Theta \rightarrow \mathbb{C}^m$. Further, consider a vector data set $\{\mathbf{x}(t) \in \mathbb{C}^m\}_{t=1}^{\infty}$, observed through the

following model:

$$\mathbf{x}(t) = \sum_{k=1}^{n_t} \mathbf{a}(\theta_k(t)) s_k(t) + \mathbf{n}(t) \quad (1)$$

where t is the time index, the sets $\{\theta_k(t) \in \Theta\}$ and $\{s_k(t) \in \mathbb{C}\}$ are called position and amplitude parameters, respectively and $\{\mathbf{n}(t)\}$ denotes the additive noise, assumed to be a centered Gaussian, white and stationary process, with covariance matrix $\sigma^2 \mathbf{I}$. Notice that n_t , the number of parameters involved in modeling $\mathbf{x}(t)$, also known as order, can be variable in time and is seldom a priori known. The aim is mainly to estimate n_t and the position parameters ($\{\theta_k(t)\}$), since they often carry the desired information. However, since the model in (1) is linear in the amplitude parameters, once the position parameters are replaced by their estimates, a standard linear estimator such as Ordinary Least Squares (OLS) may be used to estimate the amplitudes. The problem of estimating n_t is often called model order selection.

More formally, the observation model in (1) can be written as

$$p(\mathbf{x}(t) | S_t) = \frac{1}{(\pi\sigma^2)^m} e^{-\frac{\|\mathbf{x}(t) - \sum_{k=1}^{n_t} \mathbf{a}(\theta_k(t)) s_k(t)\|_2^2}{\sigma^2}} \quad (2)$$

where the finite set S_t , given by

$$S_t = \{(\theta_1(t), s_1(t)), (\theta_2(t), s_2(t)), \dots, (\theta_{n_t}(t), s_{n_t}(t))\}. \quad (3)$$

represents the state. We further assume that the amplitudes $s_k(t)$ are distributed by a centered Gaussian pdf with variance $\mathcal{I}_k(t)$, which we refer to as intensity. This can be formally written as

$$p(s_k | \mathcal{I}_k(t)) = \frac{1}{\pi \mathcal{I}_k(t)} e^{-\frac{|s_k|^2}{\mathcal{I}_k(t)}} \quad (4)$$

After straightforward manipulations, combining (2) and (4), and integrating over s_k leads to the following likelihood function in terms of the position and intensity parameters.

$$p(\mathbf{x}(t) | \bar{S}_t) = \frac{1}{\pi^m \det(\mathbf{R}(t))} e^{-\mathbf{x}^H(t) \mathbf{R}^{-1}(t) \mathbf{x}(t)} \quad (5)$$

where

$$\bar{S}_t = \{(\theta_1(t), \mathcal{I}_1(t)), (\theta_2(t), \mathcal{I}_2(t)), \dots, (\theta_{n_t}(t), \mathcal{I}_{n_t}(t))\} \quad (6)$$

is a new state representation, here called the hyper-state, and

$$\mathbf{R}(t) = \mathbf{R}(\bar{S}_t) = \sigma^2 \mathbf{I} + \sum_{k=1}^{n_t} \mathcal{I}_k(t) \mathbf{a}(\theta_k(t)) \mathbf{a}^H(\theta_k(t)) \quad (7)$$

The recent findings [39] in the field of sparsity-based estimation suggest to substitute the determinant term with an exponential function to obtain

$$p(\mathbf{x}(t) | \bar{S}_t) \propto e^{-\mathbf{x}^H(t)\mathbf{R}^{-1}(t)\mathbf{x}(t) - \lambda_0 \sum_k \mathcal{I}_k} \quad (8)$$

where λ_0 is related to the average number of parameters, and practically treated as a design parameter. Considered in this work, the model in (8) leads to a convex ML estimator, known as SParse Iterative Covariance-based Estimation (SPICE). Moreover, the convexity of the negative log-likelihood leads to unimodality of the posterior distributions.

2.2 Time Evolution Model

For the applications of interest herein, it is not suitable to consider an evolution model for the state S_t . Instead, a motion model for the hyperstate \bar{S}_t is considered. The motion model is a Markov chain, represented by the transitional probability density $Q(\bar{S}, \bar{S}') = p(\bar{S}_{t+1} = \bar{S} | \bar{S}_t = \bar{S}')$. It assigns to any pair of finite sets (\bar{S}, \bar{S}') a value, quantifying the likelihood of \bar{S}' being followed by \bar{S} . Note that we consider a temporally constant transition function Q , corresponding to a stationary Markov Chain (MC). Then, the joint p.d.f. of the sequence of state sets over an arbitrary window $\{t_1, t_1 + 1, \dots, t_2\}$ of time is given by

$$p(\bar{S}_{t_1}, \bar{S}_{t_1+1}, \bar{S}_{t_1+2}, \dots, \bar{S}_{t_2}) = p_{t_1}(\bar{S}_{t_1})Q(\bar{S}_{t_1+1}, \bar{S}_{t_1})Q(\bar{S}_{t_1+2}, \bar{S}_{t_1+1}) \dots Q(\bar{S}_{t_2}, \bar{S}_{t_2-1}) \quad (9)$$

where $p_{t_1}(\bar{S}_{t_1})$ denotes the marginal state distribution at the initial time t_1 . We focus on a specific transition probability, associated with a case, where the elements of S_t may first independently disappear with a small probability α . Then, the surviving elements may be modified by scalar models $p_0(\theta_{t+1} = \theta | \theta_t = \theta')$, $p_1(\mathcal{I}_{t+1} = \mathcal{I} | \mathcal{I}_t = \mathcal{I}')$, and finally some new independent elements may be added according to a Poisson process with the hypothesis density function $\delta(\theta, \mathcal{I})$. This means that a new parameter may independently appear in a small neighborhood \mathcal{N} of a point (θ, \mathcal{I}) with probability $\delta(\theta, \mathcal{I})d(\mathcal{N})$, where $d(\mathcal{N})$ is the volume (Lebesgue measure) of \mathcal{N} . Note that

$$\delta = \int_{\Theta \times \mathbb{R}_+} \delta(\theta, \mathcal{I})d\theta d\mathcal{I} < \infty. \quad (10)$$

is the average rate of parameter birth, here assumed to be small. Then, the transition probability $Q(\bar{S}, \bar{S}')$ is given by

$$Q(\bar{S}_{t+1} = \bar{S}, \bar{S}_t = \bar{S}') = e^{-\delta} \sum_{R \in \mathcal{T}(\bar{S}, \bar{S}')} \alpha^{|\bar{S}'| - |R|} (1 - \alpha)^{|R|} \prod_{(\theta, \mathcal{I}, \theta', \mathcal{I}') \in R} p_0(\theta | \theta') p_1(\mathcal{I} | \mathcal{I}') \prod_{\theta \notin d_1(R)} \delta(\theta) \quad (11)$$

where each summand is defined by an assignment R between the elements of \bar{S} and the elements of \bar{S}' . Note that $|\bar{S}'| - |R|$ is the number of removed parameters from \bar{S}' , and the set $\theta \notin d_1(R)$ contains the newly introduced parameters in \bar{S} . Hence, the three product terms in the summand evaluate the probabilities of survival, alteration and birth, respectively and according to the assignment R . The question of interest herein is to provide a filter, estimating the set \bar{S}_t at each time t based on the observations $\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t)$, the observation model in (8) and the MC motion model given by the transition probability in (11).

3 Recursive Bayesian Filtering

The model in (9) enables us to solve exactly the desired estimation problem in a recursive way. Denoting $X^{(t)} = [\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t)]$, we observe that the best estimate, in the Maximum A Posterior (MAP) density, for \bar{S}_t based on the observations up to time t , i.e. $X^{(t)}$ is given by maximizing the conditional likelihood $p(\bar{S}_t | X^{(t)})$. The special form of the MC model in (9) allows to recursively calculate $p(\bar{S}_t | X^{(t)})$ by applying the Bayes rule:

$$p(\bar{S}_t | X^{(t)}) = \frac{p(\bar{S}_t, \mathbf{x}(t) | X^{(t-1)})}{p(\mathbf{x}(t) | X^{(t-1)})} = \frac{p(\mathbf{x}(t) | \bar{S}_t) p(\bar{S}_t | X^{(t-1)})}{\int_{\mathcal{S}} p(\mathbf{x}(t) | \bar{S}_t = \bar{S}) p(\bar{S}_t = \bar{S} | X^{(t-1)}) d\bar{S}} \quad (12)$$

where \mathcal{S} denotes the entire space of the hyper-states, discussed in Appendix 6, and

$$p(\bar{S}_t | X^{(t-1)}) = \int_{\mathcal{S}} Q(\bar{S}_t, \bar{S}_{t-1} = \bar{S}) p(\bar{S}_{t-1} = \bar{S} | X^{(t-1)}) d\bar{S} \quad (13)$$

The resulting recursion is simple: Given the conditional distribution $p(\bar{S}_{t-1} | X^{(t-1)})$ at time instant $t - 1$, calculate the prediction distribution $p(\bar{S}_t | X^{(t-1)})$ by (13). Then, use (12) to update the conditional distribution to $p(\bar{S}_t | X^{(t)})$. As seen, the denominator in (12) is independent of \bar{S}_t . Thus, it

can be replaced by any other scaling factor, without affecting the final result of MAP estimation, simplifying the calculations. This is called recursive Bayesian filtering.

The difficulty in the above method is to store the conditional distribution and calculate the integral in (13). Our method here is to consider the following family of approximate distributions, parametrized by an arbitrary symmetric positive semidefinite matrix $\hat{\mathbf{R}}$ and an arbitrary positive weight function $\lambda : \Theta \rightarrow \mathbb{R}_+$ as follows

$$p(\bar{S}; \hat{\mathbf{R}}, \lambda) = \exp -\text{Tr}(\hat{\mathbf{R}}\mathbf{R}^{-1}) - \sum_{(\theta, \mathcal{I}) \in \bar{S}} \lambda(\theta)\mathcal{I} \quad (14)$$

where

$$\mathbf{R} = \mathbf{R}(\bar{S}) = \sigma^2\mathbf{I} + \sum_{(\theta, \mathcal{I}) \in \bar{S}} \mathcal{I}\mathbf{a}(\theta)\mathbf{a}^H(\theta) \quad (15)$$

We approximate the posteriors by selecting the closest distribution in the KL sense in this family. We denote the parameters of the closest distribution to $p(\bar{S}_t | X^{(t-1)})$ and $p(\bar{S}_t | X^{(t)})$ by $(\hat{\mathbf{R}}_t^-, \lambda_t^-)$ and $(\hat{\mathbf{R}}_t^+, \lambda_t^+)$, respectively.

The distribution in (14) is necessarily unimodal. Moreover, when $\hat{\mathbf{R}}$ and λ are large, it is highly concentrated around its global maximal point, called the Maximum A Posterior (MAP) hyper-state estimate. When the updated distribution $p(\bar{S}_t | X^{(t)})$ is considered, the resulting MAP estimate is the filter output (the desired estimate). When, $p(\bar{S}_t | X^{(t-1)})$ is instead considered, the MAP estimate is called the predicted hyper state.

3.1 Calculating the MAP Hyper-State Estimate

One of the advantages with the above choice of approximate distribution is that it simplifies calculating the maximum a posterior estimate. When the posterior $p(\bar{S}_t | X^{(t)})$ is calculated and approximated by parameters $(\hat{\mathbf{R}}_t^+, \lambda_t^+)$, the hyper-state MAP estimate is calculated by

$$\hat{S}_t = \arg \max_{\bar{S} \in \mathcal{S}} p(\bar{S}_t | \hat{\mathbf{R}}_t^+, \lambda_t^+) \quad (16)$$

Similarly, the MAP predicted hyper-state is defined by

$$\hat{S}_t^- = \arg \max_{\bar{S} \in \mathcal{S}} p(\bar{S}_t | \hat{\mathbf{R}}_t^-, \lambda_t^-) \quad (17)$$

Both optimizations in (17) and (16) yield to

$$\hat{S}_t = \arg \min_{\bar{S} \in \mathcal{S}} \text{Tr} \left[\left(\sigma^2 \mathbf{I} + \sum_{(\theta, \mathcal{I}) \in \bar{S}} \mathcal{I} \mathbf{a}(\theta) \mathbf{a}^H(\theta) \right)^{-1} \hat{\mathbf{R}}_t^\pm \right] + \sum_{(\theta, \mathcal{I}) \in \bar{S}} \mathcal{I} \lambda_t^\pm(\theta) \quad (18)$$

where the plus and negative sign is for (16) and (17), respectively. The optimization in (18) is a type of sparsity-based estimator and can be solved fast and precisely, with the so called weighted SPICE technique. First, a fine grid $\{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_N\}$ over Θ is considered. Then, the following convex optimization is solved and the non-zero elements are selected as the estimates.

$$\min_{(\tilde{\mathcal{I}}_1, \tilde{\mathcal{I}}_2, \dots, \tilde{\mathcal{I}}_N) \geq 0} \text{Tr} \left[\left(\sigma^2 \mathbf{I} + \sum_{k=1}^N \tilde{\mathcal{I}}_k \mathbf{a}(\tilde{\theta}_k) \mathbf{a}^H(\tilde{\theta}_k) \right)^{-1} \hat{\mathbf{R}}_t^\pm \right] + \sum_{k=1}^N \tilde{\mathcal{I}}_k \lambda_t^\pm(\tilde{\theta}_k) \quad (19)$$

The optimization in (19) can be solved either by the off-the-shelf techniques, such as the CVX toolbox, or by the specific technique explained in [38].

3.2 Update Step

Assume that at a certain time instant t , the posterior $p(\bar{S}_t | X^{(t-1)})$ is approximated by parameters $(\hat{\mathbf{R}}_t^-, \lambda_t^-)$. Once the vector $\mathbf{x}(t)$ is observed, the posterior is changed according to (12), which using (8), results in

$$\begin{aligned} p(\bar{S}_t | X^{(t)}) &\propto \\ e^{-\mathbf{x}^H(t) \mathbf{R}^{-1}(t) \mathbf{x}(t) - \text{Tr}(\hat{\mathbf{R}}_t^- \mathbf{R}^{-1}(t)) - \sum_{(\theta, \mathcal{I}) \in \bar{S}_t} (\lambda_t^-(\theta) \mathcal{I} + \lambda_0 \mathcal{I})} &= \\ \exp \left\{ -\text{Tr} \left[\mathbf{R}^{-1}(t) \left(\hat{\mathbf{R}}_t^- + \mathbf{x}(t) \mathbf{x}^H(t) \right) \right] \right. & \\ \left. - \sum_{(\theta, \mathcal{I}) \in \bar{S}_t} (\lambda_t^-(\theta) + \lambda_0) \mathcal{I} \right\} & \end{aligned} \quad (20)$$

We obtain that

$$\begin{aligned} \hat{\mathbf{R}}_t^+ &= \hat{\mathbf{R}}_t^- + \mathbf{x}(t) \mathbf{x}^H(t) \\ \lambda_t^+(\theta) &= \lambda_t^-(\theta) + \lambda_0 \end{aligned} \quad (21)$$

3.3 Prediction Step Approximation

Now, consider occasions where the posterior $p(\bar{S}_{t+1} | X^{(t)})$ is to be calculated by (13). Assume that the posterior $p(\bar{S}_t | X^{(t)})$ is approximated by the parameters $\hat{\mathbf{R}}_t^+$ and λ_t^+ , and that these parameters are large enough, such that the corresponding posterior is highly concentrated around the filter output \hat{S}_t . In this case and according to Appendix 7, \bar{S}_t is a result of perturbing the parameters of the MAP hyper-state estimate with a Gaussian perturbation, followed by adding extra elements $(\theta_k^f, \mathcal{I}_k^f)$, distributed by a Poisson distribution. For simplicity, let us denote $\hat{S}_t = \{(\theta_1, \mathcal{I}_1), \dots, (\theta_n, \mathcal{I}_n)\}$ and denote by $\Delta\theta_k$ and $\Delta\mathcal{I}_k$ the perturbations in θ_k and \mathcal{I}_k , respectively. Then, according to the extended Laplace's method, derived in Appendix 7.2, we may approximate $p(\bar{S}_t | X^{(t)})$:

$$\begin{aligned} \Delta\theta_k &\sim \mathcal{N}(\mathbf{0}, G_k^{-1}), \quad \Delta\mathcal{I}_k \sim \mathcal{N}(\mathbf{0}, H_k^{-1}) \\ \{(\theta_k^f, \mathcal{I}_k^f)\} &\sim \text{Poisson}(\omega(\theta, \mathcal{I})) \end{aligned} \quad (22)$$

where

$$\begin{aligned} G_k &= \frac{\partial^2 \text{Tr}(\hat{\mathbf{R}}_t^+ \mathbf{R}^{-1})}{\partial \theta_k^2}, \quad H_k = \frac{\partial^2 \text{Tr}(\hat{\mathbf{R}}_t^+ \mathbf{R}^{-1})}{\partial \mathcal{I}_k^2} \\ \omega(\theta, \mathcal{I}) &= \exp\left(\frac{\mathbf{a}(\theta)^H \mathbf{R}_+^{-1}(t) \hat{\mathbf{R}}_t^+ \mathbf{R}_+^{-1}(t) \mathbf{a}(\theta) \mathcal{I}}{1 + \mathbf{a}(\theta)^H \mathbf{R}_+^{-1}(t) \mathbf{a}(\theta) \mathcal{I}} - \lambda_t^+(\theta) \mathcal{I}\right) \end{aligned} \quad (23)$$

Simple calculations show that after applying time evolution by (13), the approximation in (22) still holds, but the parameters G_k, H_k and $\omega(\theta, \mathcal{I})$ are updated (See [33] for the Poisson Process under time evolution) to

$$\begin{aligned} G'_k &= \frac{G_k}{1 + \sigma_\theta^2 G_k}, \quad H'_k = \frac{H_k}{1 + \sigma_{\mathcal{I}}^2 H_k} \\ \omega'(\theta, \mathcal{I}) &= (1 - \alpha) \int_{\Theta \times \mathbb{R}_+} \omega(\theta', \mathcal{I}') p_0(\theta | \theta') p_1(\mathcal{I} | \mathcal{I}') d\theta d\mathcal{I} + \\ &\quad \delta(\theta, \mathcal{I}) \end{aligned} \quad (24)$$

respectively, where σ_θ^2 and $\sigma_{\mathcal{I}}^2$ are the perturbation variance, associated with the time evolution models p_0 and p_1 , given by $\text{Var}(\theta_t | \theta_t = \theta_k)$ and $\text{Var}(\mathcal{I}_t | \mathcal{I}_t = \mathcal{I}_k)$ respectively. This represents the posterior distribution after time evolution. Now, we project this distribution on the desired space of parametrized distributions by $\hat{\mathbf{R}}$ and λ . We perform this by taking the minimum KL distance. Although the process is generally intractable, assuming that time evolution is small, i.e. the hyper-state does not change fast, the process can be easily performed by perturbation theory. Appendix 8, establishes this relation. Here, we consider the final result, where limited computational complexity is also considered. The simplified prediction

steps can be represented by

$$\begin{aligned}\hat{\mathbf{R}}_{t+1}^- &= \frac{\sum_k \frac{1}{1+\sigma_\theta^2 G_k} + \sum_k \frac{1}{1+\sigma_\theta^2 H_k}}{2n} \hat{\mathbf{R}}_t^+ \\ \lambda_{t+1}^-(\theta) &= \left[\lambda_t^+(\theta) - \right. \\ &\quad \left. \frac{\delta_1(\theta)}{2} (\lambda(\theta) - \mathbf{a}^H(\theta) \mathbf{R}_+^{-1}(t) \hat{\mathbf{R}}_t^+ \mathbf{R}_+^{-1}(t) \mathbf{a}(\theta)) \right]_+ \end{aligned} \quad (25)$$

where $\delta_1(\theta) = \int_{\mathbb{R}_+} \mathcal{I} \delta(\theta, \mathcal{I}) d\mathcal{I}$. The overall proposed algorithm is summarized in Algorithm 4.

Algorithm 4 The proposed algorithm.

Require: A positive definite matrix \mathbf{C} and a positive function $\delta_1(\theta)$.

Initialize by a symmetric positive definite matrix $\hat{\mathbf{R}}_1^-$ and a positive function $\lambda_1^-(\theta)$.

Set $t = 1$.

repeat

Observe $\mathbf{x}(t)$ and calculate $\hat{\mathbf{R}}_t^+$ and λ_t^+ from (21).

Calculate $\hat{\mathcal{S}}_t$ by solving its corresponding SPICE optimization in (19) and selecting nonzero elements. Calculate $\mathbf{R}_+(t) = \mathbf{R}(\hat{\mathcal{S}}_t)$.

Calculate $\hat{\mathbf{R}}_{t+1}^-$ and λ_{t+1}^- from (25).

Set $t \leftarrow t + 1$.

until Required.

4 Numerical Results and Comparison to Related Works

In this section, we examine the method developed in Section 3 in a number of selected scenarios and compare the results on the synthetic data to other filtering technique. We consider the problem of Direction of Arrival (DOA) estimation with a Uniform Linear Array (ULA), where the position parameter is the direction (angle) of an electromagnetic source and the amplitude is the complex envelope of the electromagnetic wave transmitted by it and the observation vector is the signal measured at a ULA of $m = 20$ sensors. The DOA is often reparametrized for simplicity, introducing the electrical angle, which we utilize here. Then, (1) holds with

$$\mathbf{a}(\theta) = [1 \ e^{j\theta} \ e^{2j\theta} \ \dots \ e^{(m-1)j\theta}] \quad (26)$$

where $\theta \in \Theta = [-\pi \ \pi]$ is the electrical angle.

In all simulations, we use a Gaussian MC model for parameter evolution, i.e.

$$p_0(\theta | \theta') = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} e^{-\frac{(\theta-\theta')^2}{2\sigma_\theta^2}} \quad (27)$$

and

$$p_1(\mathcal{I} | \mathcal{I}') = \frac{1}{\sqrt{2\pi\sigma_{\mathcal{I}}^2}} e^{-\frac{(\mathcal{I}-\mathcal{I}')^2}{2\sigma_{\mathcal{I}}^2}} \quad (28)$$

We also perform the calculations over the spectra (e.g. $\lambda(\theta)$) in the recursive algorithms of interest, by taking a uniform grid $\tilde{\Theta}$ over Θ with minimum separation 0.01. This results in 629 grid points. The average false alarm power $\delta_1(\theta)$ is also selected uniformly over Θ , i.e. $\delta(\theta) = \delta$.

4.1 Related Studies

In the literature, there is a number of different studied approaches, applicable to the problem of interest herein. We briefly review some of the more popular ones, considered her for comparison.

Sliding Window Techniques

In the simplest case, a temporal window is considered, which is generally defined by a window function w_τ for $\tau = 0, 1, \dots$. At a given time t , the following optimization is solved

$$\begin{aligned} (\hat{\theta}_1(t), \hat{\theta}_2(t), \dots, \hat{\theta}_n(t)) = \arg \min_{\theta_1, \theta_2, \dots, \theta_n} \min_{s_1(t), s_2(t), \dots, s_n(t)} \\ \sum_{\tau=1}^T w_\tau \left\| \mathbf{x}(t-\tau) - \sum_{k=1}^n \mathbf{a}(\theta_k) s_k(t-\tau) \right\|_2^2 \end{aligned} \quad (29)$$

Then, the position parameters $\hat{\theta}_k(t)$ of the global minimum point is the filter output. Notice that the summation in the cost of (29) is over time, but the parameters θ_k are not time dependent. The motivation for (29) is that the error in assuming constant position parameters can be approximately modeled by the increase in the noise variance with the factors $\{w_{\Delta t}\}$. Optimizing (29) is equivalent to solving the ML estimator for such an approximate model. Also, note that the order n is fixed. In practice, where the order is typically unknown and variable, (29) is solved for a variety of orders. This can be efficiently done, e.g. by the RELAX technique [40]. Denoting by V_n the optimal value of (29), the order and its corresponding solution is selected by a rule over the collection $\{V_n\}$, generally called information criterion. We consider a popular choice of information criterion,

given by minimizing

$$\min_n V_n + kn \quad (30)$$

where k is a design parameter. The choice of k for asymptotic cases and other information criteria are discussed in [41]. When $w_{\Delta t} = \delta_{0,\Delta t}$, i.e it is non zero, only when $\Delta = 0$, the optimization in (29) simplifies to the exact ML estimator based on the observation model. We refer to this as the "instantaneous" estimator.

The inner optimization in (29) can be solved analytically to obtain

$$(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n) = \arg \max_{\theta_1, \theta_2, \dots, \theta_n} \text{Tr} \left(\hat{\mathbf{R}}_t \mathbf{P}_{\mathbf{A}(\theta_1, \dots, \theta_n)} \right) \quad (31)$$

where $\hat{\mathbf{R}}_t = \sum_{\Delta t=1}^t w_{\Delta t} \mathbf{x}(t - \Delta t) \mathbf{x}^H(t - \Delta t)$ is the windowed sample correlation matrix, $\mathbf{A}(\theta_1, \dots, \theta_n) = [\mathbf{a}(\theta_1), \dots, \mathbf{a}(\theta_n)] = \mathbf{A}$ and $\mathbf{P}_{\mathbf{A}(\theta_1, \dots, \theta_n)} = \mathbf{A}(\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H$ is the projection matrix into the range space of \mathbf{A} , also known as the signal space. Solving (31) is still difficult, but the following approximate technique can be used: First, the closest projection matrix $\hat{\mathbf{P}}_T$ to $\hat{\mathbf{R}}_T$ in the Frobenius distance is found as

$$\hat{\mathbf{P}}_T = \mathbf{U}_{n,T} \mathbf{U}_{n,T}^H \quad (32)$$

where $\mathbf{U}_{n,T}$ is the collection of the eigenvectors related to the n largest eigenvalues of $\hat{\mathbf{R}}_T$. Then, the closest bases $\mathbf{a}(\theta)$ to the range space of $\hat{\mathbf{P}}_T$ is selected by taking the local minima of the spectrum $u_T(\theta) = \|\mathbf{a}(\theta) - \hat{\mathbf{P}}_T \mathbf{a}(\theta)\|_2^2$. This technique is called MUltiple Signal Classification (MUSIC).

Target Tracking Techniques

From one perspective, the target tracking techniques are to enhance the quality of estimates provided by other methods, such as the instantaneous estimates. Suppose that an instantaneous estimator is utilized to obtain a preliminary set of estimates $Z_t = \{\hat{\theta}_1(t), \hat{\theta}_2(t), \dots, \hat{\theta}_{\hat{n}_t}(t)\}$. Then, the estimates can be related to X_t through the analysis of the instantaneous estimator, leading to a conditional pdf $p(Z_t | X_t)$. As seen, the resulting model is again RFS based. Most often, the following approximate relation, very similar to the evolution model in (11) is considered.

$$p(Z_t | X_t) = e^{-\mu} \sum_{R \in \mathcal{T}(Z_t, X_t)} \beta^{|R|} (1 - \beta)^{|X_t| - |R|} \prod_{(\hat{\theta}, \theta) \in R} p_1(\hat{\theta} | \theta) \prod_{\theta \notin d_1(R)} \mu(\theta) \quad (33)$$

where β is the probability of detection of a parameter, $p_1(\hat{\theta} | \theta)$ is the distribution of an estimates $\hat{\theta}$, corresponding to the true parameter θ , and

$\mu(\theta)$ is the hypothesis density for the false alarm (false detection) process, assumed to be a Poisson process. Note that

$$\mu = \int_{\Theta} \mu(\theta) d\theta < \infty \quad (34)$$

is the average false alarm rate. Given (11) and (33), we may use (12) and (13) to obtain a recursive filter, called target tracking filter. The exact result is generally numerically intractable. To maintain a limited amount of calculations in the course of target tracking, the method of Probability Hypothesis Density (PHD) [33] approximates the resulting posterior distributions by the Poisson process, leading to the following steps: Denoting by D_t^+ and D_t^- , the PHDs for the updated and predicted posteriors, respectively, the prediction in (13) is exactly resolved to give

$$D_t^+(\theta) = \alpha \int_{\Theta} p_0(\theta | \theta') D_{t-1}^-(\theta') d\theta' + \delta(\theta) \quad (35)$$

and the closest approximation in the Kullback-Leibler sense to the result of the calculations in (12) is found to be

$$D_t^-(\theta) = (1 - \beta) D_t^+(\theta) + \sum_{\hat{\theta} \in Z_t} \frac{\beta p_1(\hat{\theta} | \theta) D_t^+(\theta)}{\beta \int_{\Theta} p_1(\hat{\theta} | \theta) D_t^+(\theta) d\theta + \mu(\hat{\theta})} \quad (36)$$

The final estimates are given by local maxima of $D_t^-(\theta)$.

Subspace-Based Techniques

Another type of recursive filters is introduced, based on the subspace techniques such as the previously introduced MUSIC method. The idea is to replace $X_t = \{\theta_k(t)\}$ by the subspace \mathcal{X} , spanned by the bases $\{\mathbf{a}(\theta_k(t))\}$. The subspace is represented by a projection matrix $\mathbf{P}(t)$. An effective way to estimate $\mathbf{P}(t)$, also considered here is to solve

$$\mathbf{P}(t) = \arg \min_{\mathbf{P}} \|\mathbf{x}(t) - \mathbf{P}\mathbf{x}(t)\|_2^2 + \alpha \|\mathbf{P} - \mathbf{P}(t-1)\|_F^2 \quad (37)$$

where $\mathbf{P}(t-1)$ is the estimate at the previous time instant and α is a design parameter. Once $\mathbf{P}(t)$ is calculated, the parameter estimates are obtained by the MUSIC technique. Note that this technique is loosely tied to the statistical model, stated in Section 2, though it enjoys a remarkably low computational complexity.

4.2 Numerical Results

Now, we consider the introduced techniques and the proposed one in some scenarios. For the PHD observation model, we also choose

$$p(\hat{\theta} | \theta) = \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{-\frac{(\hat{\theta}-\theta)^2}{2\sigma_e^2}} \quad (38)$$

where we treat σ_e as a tuning parameter. The instantaneous estimator for the target tracking technique is RELAX with the information criterion in (30) and $k = 3$.

Two Crossing Targets

In this setup, two moving sources $(\theta_1(t), \theta_2(t))$ were considered. They moved according to the equations $\theta_1 = -\pi/2 + 0.01\pi t$ and $\theta_2 = \pi/2 - 0.01\pi t$ for $t = 1, 2, \dots, T = 100$. Their corresponding amplitudes were randomly generated by the standard Gaussian distribution. The noisy observations were obtained by adding centered, uncorrelated Gaussian noise to the observations, with variance 0.25, providing $\text{SNR} \approx 6\text{dB}$.

The proposed technique was applied by $\lambda = 2$ and $\sigma = 0.5$, together with the time evolution parameters $\delta_1(\theta) = 0.1$ and $\sigma_\theta = \sigma_{\mathcal{I}} = 0.03$. We also considered instantaneous estimation by RELAX and enhanced the results by PHD filtering. For the latter case, the parameters $\beta = 0.99$, $\alpha = 0.01$, $\delta(\theta) = 10^{-4}$, $\mu = 0.04$ and $\sigma_e = 0.01$ are selected. Moreover, the subspace technique in (37) is used with $\alpha = 2$, adjusted for the best result.

In terms of missed detection, false alarm and error, figures 1, 2 and 3 depict the average quality of the resulting estimates over time, respectively. At a specific time, the number of false alarms, and missed detections are simply calculated as the number of exceeding or lacking parameters, namely $(\hat{n}_t - n_t)_+$ and $(n_t - \hat{n}_t)_+$, respectively. The error is calculated by adding the square error over the best assignment between estimates and the true parameters.

As seen, the instantaneous RELAX estimator typically has a high false alarm rate. The PHD filter substantially improve both the false alarm, and the error properties of the RELAX method, but increases the missed detection rate. Changing the parameters of the PHD filter modifies the trade off between false alarm and missed detection, but may not improve both. On the other hand, the proposed technique has improved miss-detection properties, but slightly increases the error level. This is due to the mismatch between the exact model in (1) and the applied one in (8), which is well known to result in biased estimates. It is clearly seen that the proposed technique initially needs about 40 samples to achieve its steady behavior,

4. NUMERICAL RESULTS AND COMPARISON TO RELATED WORKS

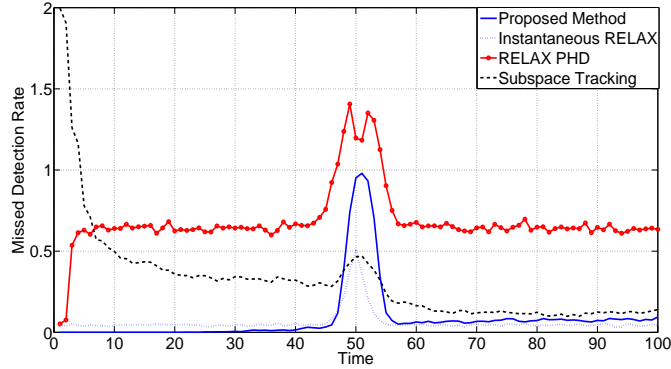


Figure 1: Missed detection rate in the deterministic crossing setup, averaged over 16000 trials.

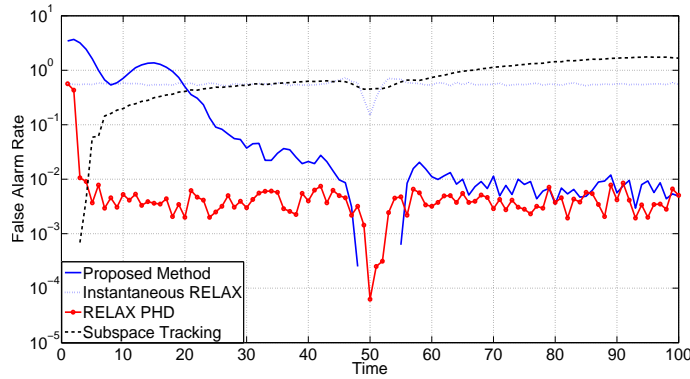


Figure 2: False alarm rate in the deterministic crossing setup, averaged over 16000 trials.

but later rapidly adapts itself to a varying environment. This may imply an improper choice of initial parameters. Finally, notice that the proposed technique provides better results at the crossing point, suggesting that the proposed method relies more on the time correlation of parameters. This can also be seen from the fact that in Figure 3, the proposed technique corresponds to a smoother curve than the other techniques, showing higher temporal correlation between the estimates.

Single Target with a Sudden Change

In a different setup, we considered a single target θ . The target is assumed to be at rest for the first 100 samples, i.e. $\theta(t) = -\pi/2$ for $t = 1, \dots, 100$. Next, it started a linear movement with an impulsive initial position change, given by $\theta(t) = 3\pi/2 - 0.01\pi t$ for $t = 101, \dots, 2000$.

The proposed technique was compared to sliding window, with the win-

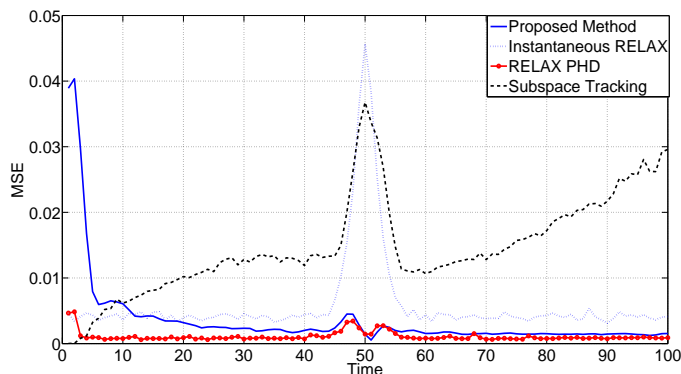


Figure 3: Mean square error in the deterministic crossing setup, averaged over 16000 trials.

dow function $w_\tau = \eta^\tau$. This choice generally simplifies the calculations, since it leads to a recursive evaluation of the matrix $\hat{\mathbf{R}}_t$ as

$$\hat{\mathbf{R}}_{t+1} = \eta \hat{\mathbf{R}}_t + \mathbf{x}(t)\mathbf{x}(t)^T \quad (39)$$

where $\hat{\mathbf{R}}_t$ is defined in (31). It is interesting to see that the overall recursive calculation of $\hat{\mathbf{R}}_t^+$ in the proposed algorithm is similar to (39), when the forgetting factor is replaced by a time-varying parameter. We also used the SPICE technique to solve (31) or equivalently (29), leading to the same optimization in (19), when $\hat{\mathbf{R}}_t^\pm$ and $\lambda(\theta)$ are replaced by $\hat{\mathbf{R}}_t$ and $\lambda_0/(1 - \eta)$, respectively. From this perspective, the proposed method is a sliding window technique with a SPICE estimator, where adaptive forgetting factor and weights are utilized.

Figures 4, 5, 6 depict the average missed detection, false alarm and error results, respectively, where the same parameters as the previous setup and $\eta = 0.8$ were used. In the initial stationary phase, the sliding window technique outperforms the proposed one, since the setup fits the assumptions of the sliding window. In terms of error, both techniques rapidly adapt to the sudden change, but the sliding window has a longer transient in terms of false alarm rate.

5 Concluding Remarks

In this paper, the problem of filtering a variable number of parameters in difficult scenarios was discussed. We used a recent modified Bayesian model in [39] and related it to a RFS-based evolution model to obtain a consistent representation for our problem of interest. Next, we approximated the corresponding recursive Bayesian filter to our model, and obtained a tractable

5. CONCLUDING REMARKS

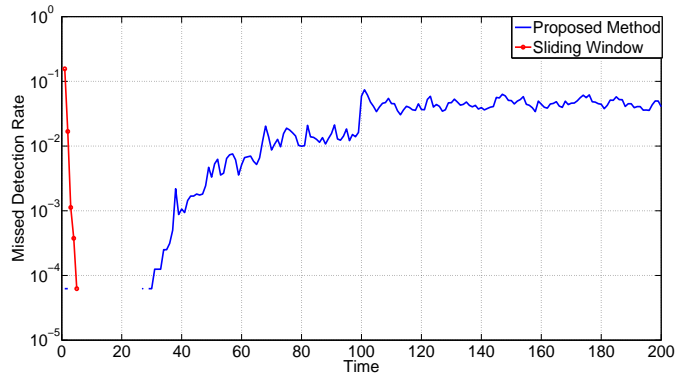


Figure 4: Missed Detection rate in the sudden movement setup, averaged over 16000 trials.

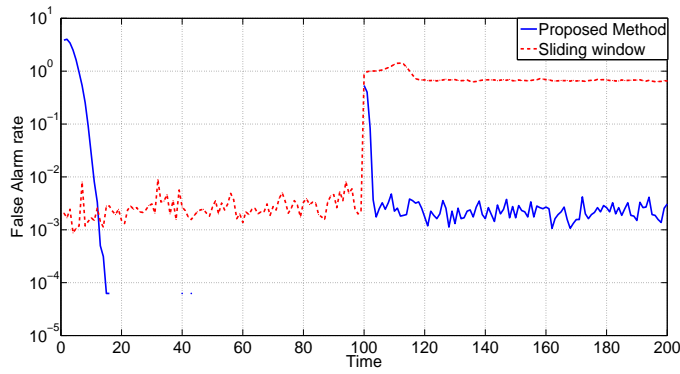


Figure 5: False alarm rate in the sudden movement setup, averaged over 16000 trials.

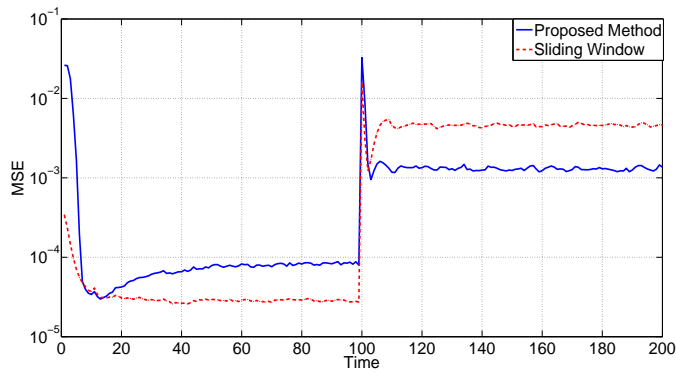


Figure 6: Mean square error in the in the sudden movement setup, averaged over 16000 trials.

filter. We simplified the design to avoid heavy computations. This led to a filter based on two components; An approximate data covariance matrix,

and a weight function, controlling miss-detection over the space of parameters.

As the numerical experiments suggest, the technique is more robust to observation impairments and is more flexible against rapid movements. Our approach exploits, and is highly connected to the SPICE technique. Hence, it exhibits similar behavior. For example, it has a relatively short convergence rate and provides consistent estimates, but the effect of noise is not symmetric on the estimates. Mathematically speaking, the estimates have a small statistical bias, proportional to the noise power. The method also exhibits a robust behavior in a low SNR regime.

Herein, the emphasis was on simplifying calculations at each recursion by avoiding difficulties with the grid-based spectral manipulations and instead combining approximate information of different time instants to maintain performance. As observed by simulations, this is favorable in a low SNR case, where fusing multiple observations is necessary to obtain a reliable estimate. However, the method might be improved if complexity is not an immediate concern and a more complex approximation is desirable. Moreover, the possibility of bootstrapping and the application of particle filters should not be ruled out.

6 Appendix: Calculus of Random Finite Sets

6.1 Functional Representation

To perform RBF, we need to calculate posteriors over finite sets, involving integration over RFS densities. Here, we review how this can be accomplished. In general, the probability distributions over the set of all finite sets can be represented by a sequence of real functions. For example, the marginal state distribution $p_t(\bar{S}_t)$ may be represented by the function sequence $\{q_t^{(n)} : \mathbb{R}^n \times \mathbb{R}_+^n \rightarrow \mathbb{R}_+\}$ defined by

$$q_t^{(n)}(\theta_1, \dots, \theta_n, \mathcal{I}_1, \dots, \mathcal{I}_n) = p_{t_1}(\bar{S}_t = \{(\theta_1, \mathcal{I}_1), \dots, (\theta_n, \mathcal{I}_n)\}) \quad (40)$$

Note that the functions $q_t^{(n)}$ are symmetric under the permutation of the pairs $(\theta_k, \mathcal{I}_k)$, since the state set is invariant under such a transform. Moreover, for a fixed n ,

$$\int_{\mathbb{R}^n \times \mathbb{R}_+^n} q_t^{(n)}(\theta_1, \dots, \theta_n, \mathcal{I}_1, \dots, \mathcal{I}_n) d^n \theta d^n \mathcal{I} = n! \times p(n_t = n) \quad (41)$$

The reason is that the left hand side integration hits each set S_t of order n exactly $n!$ times by different permutations of parameters, but does not hit

a set S_t of a different order. In the same manner, the transition probability Q can be expressed by the following sequence of functions

$$q^{(n,n')}(\theta_1, \dots, \theta_n, \mathcal{I}_1, \dots, \mathcal{I}_n, \theta'_1, \dots, \theta'_{n'}, \mathcal{I}'_1, \dots, \mathcal{I}'_{n'}) = Q(\bar{S} = \{(\theta_k, \mathcal{I}_k)\}, \bar{S}' = \{(\theta'_k, \mathcal{I}'_k)\}) \quad (42)$$

6.2 Integration

In general integration over the set of random finite sets can be explained in terms of the above functional representation. Consider the marginal distribution over the step of finite sets \bar{S}_t , represented by sequence of functions $q_t^{(n)}$ and take a function $f(\bar{S}) : \mathcal{S} \rightarrow \mathbb{R}$. Then, we have that

$$\int_{\mathcal{S}} f(\bar{S}) d\bar{S} = \sum_{n=0}^{\infty} \frac{1}{n!} \int_{\Theta^n \times \mathbb{R}_+^n} q^n(\theta_1, \dots, \theta_n, \mathcal{I}_1, \dots, \mathcal{I}_n) d^n \theta d^n \mathcal{I} \quad (43)$$

Notice how division by $n!$ cancels the aforementioned effect of multiple recalculation. Other integrations such as marginalization in (13) can be carried out in a similar manner. For example, suppose that the posterior $p(\bar{S}_t | X^{(t)})$ is represented by functions $q_0^{(n)}$ at a certain time t . Then, the integration in (13) yields to

$$p(\bar{S}_{t+1} = \{(\theta_k, \mathcal{I}_k)\} | X^{(t)}) = \sum_{n'=0}^{\infty} \frac{1}{n'!} \int_{\Theta^{n'} \times \mathbb{R}_+^{n'}} q_0^{(n,n')} q_0^{n'}(\theta'_1, \dots, \theta'_{n'}, \mathcal{I}'_1, \dots, \mathcal{I}'_{n'}) d^{n'} \theta' d^{n'} \mathcal{I}' \quad (44)$$

where the similar argument of $q^{(n,n')}$ to (42) is neglected.

7 Appendix: RFS Local Approximation

Consider a distribution in the family, given by (14), and suppose that the parameters $\hat{\mathbf{R}}$ and λ are large. Take $\hat{S} = \{(\theta_1, \mathcal{I}_1), \dots, (\theta_n, \mathcal{I}_n)\}$ as the maximum probability point. Then, a large deviation from \hat{S} leads to a considerable probability reduction. Thus, we may assume that the deviation is small. Hence, local Taylor expansion can be applied. Note that a small deviation from the set \bar{S} includes small perturbations leading to a typical hyper-state set

$$\bar{S} = \{(\theta_k + \Delta\theta_k, \mathcal{I}_k + \Delta\mathcal{I}_k)\} \cup \{(\theta_1^f, \mathcal{I}_1^f), \dots, (\theta_{n_f}^f, \mathcal{I}_{n_f}^f)\} \quad (45)$$

where the parameters, indexed with f are additional. Furthermore, the parameters $\Delta\theta_k$, $\Delta\mathcal{I}_k$ and \mathcal{I}_k^f are assumed to be small. The negative log-density function is written as

$$\begin{aligned}
 & -\log p(\bar{S}; \hat{\mathbf{R}}, \lambda) = \\
 & \text{Tr} \left[\hat{\mathbf{R}} \left(\sigma^2 \mathbf{I} + \sum_k (\mathcal{I}_k + \Delta\mathcal{I}_k) \mathbf{a}(\theta_k + \Delta\theta_k) \mathbf{a}^H(\theta_k + \Delta\theta_k) + \right. \right. \\
 & \quad \left. \left. \sum_k \mathcal{I}_k^f \mathbf{a}(\theta_k^f) \mathbf{a}^H(\theta_k^f) \right)^{-1} \right] + \\
 & \quad \sum_k \lambda(\theta_k + \Delta\theta_k) (\mathcal{I}_k + \Delta\mathcal{I}_k) + \sum_k \lambda(\theta_k^f) \mathcal{I}_k^f
 \end{aligned} \tag{46}$$

We may now apply the Taylor expansion.

7.1 Poisson Approximation

Due to the local minimality of \hat{S} , it turns out that the effect of $\Delta\theta_k$ and $\Delta\mathcal{I}_k$ vanish up to the first order. This means the negative log-distribution can be written as

$$\begin{aligned}
 & -\log p(\bar{S}; \hat{\mathbf{R}}, \lambda) = \\
 & \text{Tr} \left[\hat{\mathbf{R}} \left(\mathbf{R}_0 + \sum_k \mathcal{I}_k^f \mathbf{a}(\theta_k^f) \mathbf{a}^H(\theta_k^f) \right)^{-1} \right] + \\
 & \quad \sum_k \lambda(\theta_k) (\mathcal{I}_k) + \sum_k \lambda(\theta_k^f) \mathcal{I}_k^f
 \end{aligned} \tag{47}$$

where $\mathbf{R}_0 = \mathbf{R}(\hat{S})$. Using the matrix inversion lemma and neglecting the cross-product terms $\mathcal{I}_k^f \mathcal{I}_l^f$, we obtain

$$\begin{aligned}
 & -\log p(\bar{S}; \hat{\mathbf{R}}, \lambda) = \\
 & -\log p(\hat{S}; \hat{\mathbf{R}}, \lambda) + \sum_k \left(\lambda(\theta_k^f) \mathcal{I}_k^f - \frac{\mathbf{a}^H(\theta_k^f) \mathbf{R}_0^{-1} \hat{\mathbf{R}} \mathbf{R}_0^{-1} \mathbf{a}(\theta_k^f) \mathcal{I}_k^f}{1 + \mathbf{a}^H(\theta_k^f) \hat{\mathbf{R}} \mathbf{R}_0^{-1} \mathbf{a}(\theta_k^f) \mathcal{I}_k^f} \right)
 \end{aligned} \tag{48}$$

This shows that up to the first order, the behavior of the RFS can be identified by the Poisson process of additional elements $(\theta_k^f, \mathcal{I}_k^f)$ with density

$$w(\theta, \mathcal{I}) = e^{-\left(\lambda(\theta) \mathcal{I} - \frac{\mathbf{a}^H(\theta) \mathbf{R}_0^{-1} \hat{\mathbf{R}} \mathbf{R}_0^{-1} \mathbf{a}(\theta) \mathcal{I}}{1 + \mathbf{a}^H(\theta) \hat{\mathbf{R}} \mathbf{R}_0^{-1} \mathbf{a}(\theta) \mathcal{I}} \right)} \tag{49}$$

7.2 Extended Laplace's Method

To capture the behavior of $\Delta\theta_k$ and $\Delta\mathcal{I}_k$, we need to consider the higher order terms. However, we neglect the cross-product terms in favor of numerical simplicity, and according to the fact they are often smaller due to low

coherency in the basis manifold. Then after straightforward calculations, we obtain that

$$\begin{aligned}
 & -\log p(\bar{S}; \hat{\mathbf{R}}, \lambda) = \\
 & -\log p(\hat{S}; \hat{\mathbf{R}}, \lambda) - \sum_k \log w(\theta_k^f, \mathcal{I}_k^f) - \\
 & \frac{1}{2} \sum_k (\Delta\theta_k)^2 G_k + (\Delta\mathcal{I}_k)^2 H_k
 \end{aligned} \tag{50}$$

where

$$\begin{aligned}
 G_k &= -\frac{\partial^2}{\partial\theta_k^2} \text{Tr} \left[\hat{\mathbf{R}} \left(\sigma^2 \mathbf{I} + \sum_k \mathcal{I}_k \mathbf{a}(\theta_k) \mathbf{a}^H(\theta_k) \right)^{-1} \right] \\
 H_k &= -\frac{\partial^2}{\partial\mathcal{I}_k^2} \text{Tr} \left[\hat{\mathbf{R}} \left(\sigma^2 \mathbf{I} + \sum_k \mathcal{I}_k \mathbf{a}(\theta_k) \mathbf{a}^H(\theta_k) \right)^{-1} \right]
 \end{aligned} \tag{51}$$

This implies that $\Delta\theta_k \sim \mathcal{N}(0, G_k^{-1})$ and $\Delta\mathcal{I}_k \sim \mathcal{N}(0, H_k^{-1})$.

8 Appendix: Perturbative KL-based Projection

Suppose that the distribution $p(\bar{S}_{t+1} | X^{(t)})$ is calculated as

$$p(\bar{S}_{t+1} = \bar{S} | X^{(t)}) \approx p(\bar{S}_t = \bar{S} | X^{(t)}) + \Delta p(\bar{S}) = p(\bar{S}; \hat{\mathbf{R}}_t^+, \lambda_t^+) + \Delta p(\bar{S}) \tag{52}$$

The question of interest is to find the perturbation in parameters minimizing the KL distance between $p(\bar{S}_{t+1} = \bar{S} | X^{(t)})$ and the parametric model, i.e to solve

$$\begin{aligned}
 & \arg \min_{\Delta\hat{\mathbf{R}}, \Delta\lambda} - \int_{\bar{S}} \left(p(\bar{S}; \hat{\mathbf{R}}_t^+, \lambda_t^+) + \Delta p(\bar{S}) \right) \\
 & \log \left(p(\bar{S}; \hat{\mathbf{R}}_t^+ + \Delta\hat{\mathbf{R}}, \lambda_t^+ + \Delta\lambda) \right) d\bar{S}
 \end{aligned} \tag{53}$$

Although this can be generally solved up to the first order, by the technique explained below, we restrict $\Delta\hat{\mathbf{R}}$ to be $\gamma\hat{\mathbf{R}}_t^+$ for $\gamma > 0$ to simplify calculations, and also to ensure positive semi-definiteness. After Taylor expansion, and performing the minimization, we obtain that

$$\gamma = - \frac{\int_{\bar{S}} \frac{\partial \log p(\bar{S}; (1+\gamma)\hat{\mathbf{R}}_t^+, \lambda_t^+)}{\partial\gamma} \Big|_{\gamma=0} \Delta p(\bar{S}) d\bar{S}}{\int_{\bar{S}} p(\bar{S}; \hat{\mathbf{R}}_t^+, \lambda_t^+) \frac{\partial^2 \log p(\bar{S}; (1+\gamma)\hat{\mathbf{R}}_t^+, \lambda_t^+)}{\partial\gamma^2} \Big|_{\gamma=0} d\bar{S}} \tag{54}$$

and

$$\Delta\lambda(\theta) = -\frac{\int_{\mathcal{S}} \frac{\partial \log p(\bar{S}; \hat{\mathbf{R}}_t^+, \lambda_t^+)}{\partial \lambda(\theta)} \Delta p(\bar{S}) d\bar{S}}{\int_{\mathcal{S}} p(\bar{S}; \hat{\mathbf{R}}_t^+, \lambda_t^+) \frac{\partial^2 \log p(\bar{S}; \hat{\mathbf{R}}_t^+, \lambda_t^+)}{\partial \lambda(\theta)^2} d\bar{S}} \quad (55)$$

We obtain the desired update by the above relations. We further simplify this relation in favor of low complexity. Using the approximation in (50) and after straightforward manipulations, we get that

$$\gamma = \frac{\sum_k \frac{\frac{\partial G_k}{\partial \gamma}}{G_k^2} \Delta G_k + \sum_k \frac{\frac{\partial H_k}{\partial \gamma}}{H_k^2} \Delta H_k + \int_{\Theta \times \mathbb{R}_+} \frac{\partial \log \omega}{\partial \gamma} \Delta \omega d\theta d\mathcal{I}}{\sum_k \frac{\left(\frac{\partial G_k}{\partial \gamma}\right)^2}{G_k^2} + \sum_k \frac{\left(\frac{\partial H_k}{\partial \gamma}\right)^2}{H_k^2} + \int_{\Theta \times \mathbb{R}_+} \left(\frac{\partial \log \omega}{\partial \gamma}\right)^2 \omega d\theta d\mathcal{I}} \quad (56)$$

and

$$\Delta\lambda(\theta) = \frac{\int_{\mathbb{R}_+} \frac{\partial \log \omega}{\partial \lambda(\theta)} \Delta \omega d\mathcal{I}}{\int_{\mathbb{R}_+} \left(\frac{\partial \log \omega}{\partial \lambda(\theta)}\right)^2 \omega d\mathcal{I}} \quad (57)$$

This can be further simplified noting that the terms $\log \omega$, G_k and H_k are linear in $1 + \gamma$ thus their partial derivative with respect to γ equals their value at $\gamma = 0$, leading to

$$\gamma = \frac{\sum_k \frac{\Delta G_k}{G_k} + \sum_k \frac{\Delta H_k}{H_k} + \int_{\Theta \times \mathbb{R}_+} \log \omega \times \Delta \omega d\theta d\mathcal{I}}{2n + \int_{\Theta \times \mathbb{R}_+} (\log \omega)^2 \omega d\theta d\mathcal{I}} \quad (58)$$

According to the empirical observation that the terms involving ω are substantially smaller than the other terms, we simplify the calculations more by neglecting them to obtain

$$\gamma = \frac{\sum_k \frac{\Delta G_k}{G_k} + \sum_k \frac{\Delta H_k}{H_k}}{2n} \quad (59)$$

The expression in (57) can also be simplified by considering that $\Delta \omega \approx \delta$, and approximating ω as

$$\omega(\theta, \mathcal{I}) \approx e^{-\mathcal{I}(\lambda(\theta) - \mathbf{a}^H(\theta) \mathbf{R}_+^{-1}(t) \hat{\mathbf{R}}_t^+ \mathbf{R}_+^{-1}(t) \mathbf{a}^H(\theta))} \quad (60)$$

Simple calculations lead to

$$\Delta\lambda = -\frac{1}{2}(\lambda(\theta) - \mathbf{a}^H(\theta) \mathbf{R}_+^{-1}(t) \hat{\mathbf{R}}_t^+ \mathbf{R}_+^{-1}(t) \mathbf{a}^H(\theta)) \int_{\mathbb{R}_+} \mathcal{I} \delta(\theta, \mathcal{I}) d\mathcal{I} \quad (61)$$

References

- [1] Z. Chen, “Bayesian filtering: From kalman filters to particle filters, and beyond,” *Statistics*, vol. 182, no. 1, pp. 1–69, 2003.
- [2] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.
- [3] T. Kailath, *Lectures on Wiener and Kalman filtering*. Springer, 1981.
- [4] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Fluids Engineering*, vol. 82, no. 1, pp. 35–45, 1960.
- [5] Y.-C. Ho and R. Lee, “A bayesian approach to problems in stochastic estimation and control,” *Automatic Control, IEEE Transactions on*, vol. 9, no. 4, pp. 333–339, 1964.
- [6] B. D. Anderson and J. B. Moore, *Optimal filtering*. Courier Corporation, 2012.
- [7] S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Kalman filtering and neural networks*. Wiley Online Library, 2001.
- [8] L. Ljung, “Asymptotic behavior of the extended kalman filter as a parameter estimator for linear systems,” *Automatic Control, IEEE Transactions on*, vol. 24, no. 1, pp. 36–50, 1979.
- [9] S. J. Julier and J. K. Uhlmann, “New extension of the kalman filter to nonlinear systems,” in *AeroSense’97*. International Society for Optics and Photonics, 1997, pp. 182–193.
- [10] G. Kitagawa, “Monte carlo filter and smoother for non-gaussian nonlinear state space models,” *Journal of computational and graphical statistics*, vol. 5, no. 1, pp. 1–25, 1996.
- [11] J. Carpenter, P. Clifford, and P. Fearnhead, “Improved particle filter for nonlinear problems,” *IEE Proceedings-Radar, Sonar and Navigation*, vol. 146, no. 1, pp. 2–7, 1999.
- [12] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. Wan, “The unscented particle filter,” in *NIPS*, 2000, pp. 584–590.
- [13] S. Lototsky, R. Mikulevicius, and B. L. Rozovskii, “Nonlinear filtering revisited: a spectral approach,” *SIAM Journal on Control and Optimization*, vol. 35, no. 2, pp. 435–461, 1997.

- [14] P. Heidenreich, L. A. Cirillo, and A. M. Zoubir, "Morphological image processing for fm source detection and localization," *Signal Processing*, vol. 89, no. 6, pp. 1070–1080, 2009.
- [15] L. Rankine, M. Mesbah, and B. Boashash, "If estimation for multicomponent signals using image processing techniques in the time–frequency domain," *Signal Processing*, vol. 87, no. 6, pp. 1234–1250, 2007.
- [16] M. Moonen, P. Van Dooren, and J. Vandewalle, "A singular value decomposition updating algorithm for subspace tracking," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 4, pp. 1015–1038, 1992.
- [17] G. W. Stewart, "An updating algorithm for subspace tracking," *Signal Processing, IEEE Transactions on*, vol. 40, no. 6, pp. 1535–1541, 1992.
- [18] B. Yang, "Projection approximation subspace tracking," *Signal Processing, IEEE Transactions on*, vol. 43, no. 1, pp. 95–107, 1995.
- [19] C. F. Mecklenbrauker, P. Gerstoft, A. Panahi, and M. Viberg, "Sequential bayesian sparse signal reconstruction using array data," *Signal Processing, IEEE Transactions on*, vol. 61, no. 24, pp. 6344–6354, 2013.
- [20] N. Vaswani, "Kalman filtered compressed sensing," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 893–896.
- [21] M. Lustig, J. M. Santos, D. L. Donoho, and J. M. Pauly, "kt sparse: High frame rate dynamic mri exploiting spatio-temporal sparsity," in *Proceedings of the 13th Annual Meeting of ISMRM, Seattle*, vol. 2420, 2006.
- [22] Y. Bar-Shalom, P. K. Willett, and X. Tian, "Tracking and data fusion," *A Handbook of Algorithms*. Yaakov Bar-Shalom, 2011.
- [23] S. S. Blackman, "Multiple-target tracking with radar applications," *Dedham, MA, Artech House, Inc., 1986, 463 p.*, vol. 1, 1986.
- [24] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classification and tracking from real-time video," in *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*. IEEE, 1998, pp. 8–14.

- [25] S. J. Davey, M. G. Rutten, and B. Cheung, "A comparison of detection performance for several track-before-detect algorithms," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, p. 41, 2008.
- [26] S. M. Tonissen and R. J. Evans, "Performance of dynamic programming techniques for track-before-detect," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 32, no. 4, pp. 1440–1451, 1996.
- [27] S. M. Tonissen and Y. Bar-Shalom, "Maximum likelihood track-before-detect with fluctuating target amplitude," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 34, no. 3, pp. 796–809, 1998.
- [28] P. Willett, Y. Ruan, and R. Streit, "Pmht: problems and some solutions," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 38, no. 3, pp. 738–754, 2002.
- [29] Y. Bar-Shalom and E. Tse, "Tracking in a cluttered environment with probabilistic data association," *Automatica*, vol. 11, no. 5, pp. 451–460, 1975.
- [30] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *Oceanic Engineering, IEEE Journal of*, vol. 8, no. 3, pp. 173–184, 1983.
- [31] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 174–188, 2002.
- [32] B.-N. Vo, S. Singh, and A. Doucet, "Sequential monte carlo methods for multitarget filtering with random finite sets," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 41, no. 4, pp. 1224–1245, 2005.
- [33] R. P. Mahler, "A theoretical foundation for the stein-winter" probability hypothesis density (phd)" multitarget tracking approach," DTIC Document, Tech. Rep., 2000.
- [34] L. Svensson, D. Svensson, M. Guerriero, and P. Willett, "Set jpda filter for multitarget tracking," *Signal Processing, IEEE Transactions on*, vol. 59, no. 10, pp. 4677–4691, 2011.
- [35] D. F. Crouse, P. Willett, L. Svensson, D. Svensson, and M. Guerriero, "The set mht," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*. IEEE, 2011, pp. 1–8.

REFERENCES

- [36] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Stat. Soc., Series B (Methodological)*, vol. 58, pp. 267–288, Jan. 1996.
- [37] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic Decomposition by Basis Pursuit,” *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, Dec. 1998.
- [38] P. Stoica, P. Babu, and J. Li, “Spice: A sparse covariance-based estimation method for array processing,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 2, pp. 629–638, 2011.
- [39] P. Stoica and P. Babu, “Spice and likes: Two hyperparameter-free methods for sparse-parameter estimation,” *Signal Processing*, vol. 92, no. 7, pp. 1580–1590, 2012.
- [40] J. Li and P. Stoica, “Efficient mixed-spectrum estimation with applications to target feature extraction,” *Signal Processing, IEEE Transactions on*, vol. 44, no. 2, pp. 281–295, 1996.
- [41] P. Stoica, Y. Selen, and J. Li, “On information criteria and the generalized likelihood ratio test of model order selection,” vol. 11, no. 10, pp. 794 – 797, oct. 2004.