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

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 LaTeX.

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

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 aknowledge my other friends who has left S2 and I enjoyed woking 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*

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) Wideband 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/FCS.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. *4 th 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., Haghighi, 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., Haghighi, 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

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-hard [1] 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

[1] A 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 is 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 dicretization 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 $t = 1, 2, \ldots, T$. The expression

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

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 $\mathbf{a}(\theta) : \Theta \to \mathbb{C}^m$. The function $\mathbf{a}(\theta)$ is called a representation for the dictionary $\mathcal{A}$ if

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

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 sound[1]. 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 $t = 1, 2, \ldots$ by $\mathbf{x}(t) = [x_1(t)\ x_2(t) \ldots x_m(t)]^T$, 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.

[1] The 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

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

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

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

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

$$\mathbf{y} = \mathbf{\Psi}\mathbf{s}, \tag{2.5}$$

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 $s_1, s_2, \ldots, s_n$, corresponding to the columns $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_n$ 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

$$\mathbf{y} = \sum_k s_k \boldsymbol{\psi}_k. \tag{2.6}$$

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

$$\mathbf{x} = \sum_k s_k \mathbf{a}_k, \tag{2.7}$$

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 $X_1, X_2, \ldots, X_R$ and a number of output ones $Y_1, \ldots, Y_L$ 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

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

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

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

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

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

called the spectrum. Then, the expression in (2.1) can be equivalently written as[2]

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

Note that while in (2.1) the amplitudes $\{s_1,\ldots,s_n\}$, together with the set of bases $\{\mathbf{a}_1,\ldots,\mathbf{a}_n\}$ 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 $\tilde{\mathcal{A}}=\{\tilde{\mathbf{a}}_1,\tilde{\mathbf{a}}_2,\ldots,\tilde{\mathbf{a}}_N\}$ of $\mathcal{A}$, known as a grid, and only store the on-grid spectrum, $\tilde{s}_k(t)=\tilde{s}(\mathbf{a}_k,t)$. In a case, where the dictionary is represented by an index set Θ, this can be performed by discretizing Θ, to obtain $\tilde{\Theta}=\{\tilde{\theta}_1,\tilde{\theta}_2,\ldots,\tilde{\theta}_N\}$.

[2]For 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 $\boldsymbol{\theta}^{(n)}, \{\mathbf{s}^{(n)}(t)\}$:

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

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

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

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

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

where $\mathbf{A}^{\dagger}(\boldsymbol{\theta}^{(n)})$ denotes the Moore-Penrose pseudoinverse of the matrix $\mathbf{A}(\theta^{(n)}) = [\mathbf{a}(\theta_1)\ \mathbf{a}(\theta_1) \dots \mathbf{a}(\theta_n)]$, 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

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

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

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

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

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

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

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

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 $n = 1, 2, \ldots, \text{Spark}(\mathcal{A}) - 1$ 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 $\tilde{s}(\mathbf{a}, t)$ by $\tilde{s}_t$, we can write the relation in (2.11), in an abstract form, as

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

where $\mathscr{A}$ denotes the linear operator transforming the spectrum into the observed vector. Notice that the transformation by $\mathscr{A}$ 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, $\mathscr{A}$ 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 $\mathscr{W}$ 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 $\mathscr{W}$ 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 $\tilde{s}(\mathbf{a}) = \mathbf{w}^H(\mathbf{a})\mathbf{x}$. Now, it is intended to devise a beamformer $\mathscr{W}$, such that its application to (3.8) leads to

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

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)

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

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.

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

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

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

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

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

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

$$\hat{\mathbf{R}} = \mathbf{U}^H \mathbf{\Lambda} \mathbf{U} \tag{3.14}$$

where $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots \mathbf{u}_m]$ is a unitary matrix and $\mathbf{\Lambda}$ is the diagonal elements of the singular values $\lambda_1, \lambda_2, \ldots, \lambda_m$, written in a descending order. Then, the solution to (3.13) is given by $\mathcal{R} = (\mathbf{u}_1, \mathbf{u}_2 \ldots, \mathbf{u}_n)$, 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

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

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 $\theta_1, \theta_2, \ldots, \theta_n$ and $s_1, s_2, \ldots, s_n$, where $n < (\mathrm{Spark}(\mathcal{A}) - 1)/2$, guaranteeing the uniqueness of the ideal decomposition. We denote the ML estimates by $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ and $\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_n$, 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 $k = 1, 2, \ldots, n$,

$$|\Delta\theta_k| \to_p 0, \quad |\Delta s_k| \to_p 0 \tag{3.17}$$

as $\sigma \to 0$, where $\Delta\theta_k = \hat{\theta}_k - \theta_k$, $\Delta s_k = \hat{s}_k - s_k$ and $\to_p$ 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 noiselss 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

[1] Although 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

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

where $\mathbf{d}(\theta) = \mathrm{d}\mathbf{a}(\theta)/\mathrm{d}\theta$. 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 $(\Delta\boldsymbol{\theta}_{\mathrm{ML}}, \Delta\mathbf{s}_{\mathrm{ML}}) = \mathbf{P}_{\Omega}\mathbf{n}$, where $\mathbf{P}_{\Omega}$ 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 $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$, obtained by MUSIC. The true parameters are $\theta_1, \ldots, \theta_n$ and the true amplitudes are sampled from white, centered and uncorrelated sources. The order $n <$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

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

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 $\mathbf{A}(\theta)$. 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 vector [2] 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 $\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_m$ are the eigenvectors of $\hat{\mathbf{R}}$, sorted in the descending order of their corresponding singular values[3].

[2] This 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.

[3] The 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 $k = 1, 2 \ldots, n$ and defining $\Delta\theta_k = \hat{\theta}_k - \theta_k$, we obtain that

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

which using Taylor expansion and after straightforward calculations leads to

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

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

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

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.

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

where λ_l for $l = 1, 2 \ldots, m$ 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 $\boldsymbol{\theta}_r$ 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

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

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

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

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

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

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

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

It turns out that the optimization in (4.3) is convex on the convex set $\Psi(\mathcal{A}) \times \ldots \times \Psi(\mathcal{A})$. 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 $\tilde{\mathcal{A}} = \{\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \ldots, \tilde{\mathbf{a}}_N\}$ and restrict the spectrum to $\tilde{\mathcal{A}}$. We denote $\tilde{\mathbf{s}}(t) = [\tilde{s}_1(t)\ \tilde{s}_2(t) \ldots \tilde{s}_N(t)]^T$, where $\tilde{s}_k(t) = \tilde{s}(\tilde{\mathbf{a}}_k, t)$. Then, the BP optimization is written as

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

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

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

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

following approximate optimization

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

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 $\tilde{S}^{(k-1)}$, but also $\tilde{S}^{(k-2)}$. 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

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

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

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

The SPICE solves (4.8) for $\{p_k\}$ and $\{\tilde{\mathbf{s}}(t) \in \mathbb{C}^N\}$, alternatingly, 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 resent 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

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

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 $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ in the null space of $\mathbf{A}$, the following condition holds*

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

where $Supp(\mathbf{s}_0)$ 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 $I \subset \{1, 2, \ldots, N\}$ of n indexes and any nonzero vector $\boldsymbol{\nu}$ in the null space of $\mathbf{A}$ the following relation holds.

$$\sum_{k \in I} |\nu_k| < \sum_{k \notin I} |\nu_k| \tag{4.11}$$

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 therefor 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 $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathcal{A}$, we have that

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

where $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and the largest singular value of their arguments, respectively. Note that if $\delta = 0$, then the basis $\mathbf{a}_1, \ldots, \mathbf{a}_n$ 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 follows[1]: 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 $\mathbf{\Phi} = [\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_n]$ and a unit vector $\mathbf{x} \in \mathbb{C}^n$, we may conclude that

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

for a proper value of c and δ. Then, it is possible to show by the union bound that[2]

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

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

[1] See [125] for more details.

[2] For 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

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

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

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

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]:

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

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 alternatingly 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

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

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.

[3] The 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

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

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 $t, t+1, \ldots, t+T$:

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

where $p_{S_{t+1}|S_t}(s_1 \mid s_0)$ is called the transitional probability density. If the transitional probability density is constant, i.e. $p_{S_{t+1}|S_t}(s_1 \mid s_0) = Q(s_1 \mid s_0)$, 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 $p_{\mathbf{x}(t)|S(t)}(\mathbf{x}(t) \mid s(t))$, in short denoted by $p(\mathbf{x}(t) \mid s(t))$. At a certain time instant t, the question of interest is to estimate a group of parameters based on the observations $\mathbf{x}(1), \ldots, \mathbf{x}(t)$. For example, the entire state trajectory can be estimated by the Maximum a' Posteriori[1] (MAP) estimator as

$$(\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_t) = \arg \max_{s_1,s_2,\ldots,s_t}$$
$$p(s_1, s_2, \ldots, s_t \mid \mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T)) =$$
$$\arg \max_{(s_1,s_2,\ldots,s_t)}$$
$$p(\mathbf{x}(1) \mid s_1) p(\mathbf{x}(2) \mid s_2) \ldots p(\mathbf{x}(t) \mid s_t) \times p(s_1) \times$$
$$Q(s_2 \mid s_1) Q(s_3 \mid s_2) \ldots Q(s_t \mid s_{t-1}) \tag{5.3}$$

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

$$V_t(s_t) = \max_{s_1,s_2,\ldots,s_{t-1}}$$
$$p(\mathbf{x}(1) \mid s_1) p(\mathbf{x}(2) \mid s_2) \ldots p(\mathbf{x}(t-1) \mid s_{t-1}) \times p_{S_1}(s_1) \times$$
$$Q(s_2 \mid s_1) Q(s_3 \mid s_2) \ldots Q(s_t \mid s_{t-1}) \tag{5.4}$$

[1] See [142].

which may be obtained by

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

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

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

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 know 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

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

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

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

where

$$p_{S_t|\mathbf{X}^{(t-1)}}(s_t \mid \mathbf{X}^{(t-1)}) = \int_{s_{t-1} \in \mathcal{S}} Q(s_t \mid s_{t-1}) \mathrm{d}P_{S_{t-1}|\mathbf{X}^{(t-1)}}(s_{t-1} \mid \mathbf{X}^{(t-1)}) \tag{5.9}$$

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 $\mathbf{a}_1, \ldots, \mathbf{a}_n$ and amplitudes $s_1, s_2, \ldots, s_n$, can be represented by a state

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

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 $\boldsymbol{\theta}(t) = (\theta_1(t), \theta_2(t), \ldots, \theta_n(t))$ and $\mathbf{s}(t) = (s_1(t), s_2(t), \ldots, s_n(t))$. 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 ($Q(S_t \mid S_{t-1})$) are considered. For example, the linear case

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

where $\mathbf{H}_\theta$ and $\mathbf{H}_s$ are known (and often identity). Moreover, $\mathbf{w}_\theta$ 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 $p(\mathbf{x}(t) \mid S_t)$ 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 $p(s_t \mid \mathbf{X}^{(t-1)})$ and $p(s_t \mid \mathbf{X}^t)$ 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,

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

where we used the notation $\hat{s}_{t|t-1} = (\{\hat{\theta}_k(t \mid t-1)\}, \{\hat{s}_k(t \mid t-1)\})$ and[2] $\mathbf{d}(\theta) = \mathrm{d}\mathbf{a}(\theta)/\mathrm{d}\theta$. 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

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

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 $\tau = 0, 1, \ldots, t$, 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

[2] If 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

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

where

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

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

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.

[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. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[16] E. J. Candes 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.

[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 *IEEEInt. 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.

[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.

[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. Syrnp. IriJbrrn. 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.

[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/k2)," 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 l1 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 Mathematique*, 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.

[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.

[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.