Thesis for the Degree of Licentiate of Engineering

Parameter Estimation Using Sparse Modeling: Algorithms and Performance Analysis

Ashkan Panahi

Department of Signals and Systems
Chalmers University of Technology
Göteborg, Sweden 2012

Parameter Estimation Using Sparse Modeling: Algorithms and Performance Analysis

Ashkan Panahi

Technical report number: R010/2012
ISSN 1403-266X

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 2012

Abstract

The idea of representing a signal in a classical computing machine has played a central role in the field of signal processing. The last two decades have witnessed an important breakthrough in this by taking all possible linear transforms and domains into account. The current observations show the possibility of reconstructing a sparse signal by few measurements through linear transforms without the knowledge of the subspace where the signal resides.

This work is devoted to the application of such compressive sensing techniques to estimate a set of parameters. We try to address the main conventional ideas of estimation, especially as a regression problem, and connect these ideas to the recently developed technique by domain sparsity. We also review the conventional method of applying the so called Least Absolute Shrinkage and Selection Operator (LASSO) technique to solve estimation by domain sparsity, which looks inappropriate as a continuous estimation solution. In return, we try to develop a framework for the continuous estimation and address its unsolved problems to a concerned reader.

We also introduce a practical method of implementing the continuous LASSO as a successful attempt to solve convex variational problems. We introduce this method in the context of Direction of Arrival (DOA) estimation using an array of sensors by spatial sparsity, which gives us the possibility of analyzing the aforementioned Compressive Sensing (CS) techniques from a different perspective of statistics. The introductory parts contain the essential issues in DOA estimation, which are more or less common in all regression problems. We also review the Bayesian aspects of the LASSO based estimation briefly.

Keywords: Complex LASSO, continuous LASSO, convex variational optimization, DOA estimation, spatial sparsity, compressed sensing, linear regression

List of Publications

This thesis is based on the following three appended papers:

Paper 1

A. Panahi and M. Viberg, Maximum A posteriori Based Regularization Parameter Selection, *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.2452-2455, 22-27 May 2011, Prague, Czech.

Paper 2

A. Panahi and M. Viberg, Fast Candidate Points Selection in the LASSO Path, *Signal Processing Letters, IEEE* , vol.19, no.2, pp.79-82, Feb. 2012.

Paper 3

A. Panahi and M. Viberg, Performance Analysis of LASSO-Based Signal Parameter Estimation, Submitted to *Transactions on Signal Processing, IEEE*.

Other Publications

A. Movahed, A. Panahi, and G. Durisi, A robust RFPI-based 1-bit compressive sensing reconstruction algorithm, in *Proc. IEEE Inf. Theory Workshop (ITW)*, Lausanne, Switzerland, Sep. 2012, to appear.

Panahi, A. ; Viberg, M. A Robust ℓ_1 penalized DOA estimator *accepted to Asilomar Conference on Signals, Systems, and Computers*, 2012.

Khanzadi, M. R. ; Panahi, A. ; Kuylenstierna, D. et al. (2012). A model-based analysis of phase jitter in RF oscillators *IEEE International Frequency Control Symposium*, 2012.

Panahi, A. ; Viberg, M. Fast LASSO based DOA tracking *4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2011. p. 397-400.

Panahi, A. ; Viberg, M. On the resolution of the LASSO-based DOA estimation method *Proceedings of International ITG Workshop on Smart Antennas, WSA* 2011.

Rashidi Avendi, M. ; Haghighi, K. ; Panahi, A. et al. A NLLS based sub-Nyquist rate Spectrum Sensing for Wideband Cognitive Radio *Fifth IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2011, p. 545-551.

Khanzadi, M. R. ; Haghighi, K. ; Panahi, A. et al. A Novel Cognitive Modulation Method Considering the Performance of Primary User *Wireless Advanced (WiAD), 6th Conference on.*, 2010, p. 1-6.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

The signal processing field consists of an expanding set of problems about analyzing and operating different types of signals by mathematical techniques. Such problems have deep roots in many branches of mathematics such as analysis, probability theory, statistics, optimization, numerical methods, etc. Obviously, the possibility of such operations is limited by the computational effort and the research field is strongly influenced by the technological and theoretical improvements in the process speed. Nowadays, the signal processing techniques are mostly dependent upon the Digital Signal Processing (DSP) device family as a fast, simple, and universal solution to the computation demand.

The signal operation more or less consists of two phases. In the primary phase of modeling, there is usually a little information about the signal structure. In this case, the aim is to find a feasible representation to meet the computational constraints. In connection to the application, there always exists a primal representation of the signal as a function in a physical measurement domain (e.g temporal with a single time coordinate, spatial with three real coordinates, etc). It is also assumed that the value of this function at any single point of the domain can be measured and stored. However, the representation is inapplicable since it is impossible to sense and store the values for every such point in a digital computer. From this point of view, the modeling problem is rather a mathematical one to find a dense finite dimensional functional subspace in the domain of typical signals. The popular such subspaces are usually provided by Fourier, wavelet, and time frequency transforms [1]. However, transformations are only applicable if the signal is priorly sensed. The sensing procedure can be thought of as a group of applicable maps from the signal to the real line (e.g filtering and sampling). The modeling procedure can then be broken into multiple iterations of sieving the model subspace by studying the parameters of the measured signals of the current maps and updating them, which

leads to a *signal domain*, in which the desired signals can be represented by finite, but sometimes relatively high, numbers of parameters. An example is when a sound signal is first sampled and observed to have a bounded level, quantized and observed to contain a bounded frequency range, and finally cosine transformed in a DSP device. The modeling process might continue in a statistical manner connecting the big data representation to a smaller *parameter set* of reasonable size with a reasonable loss of precision. We only note that the "reasonable size" is often determined by the DSP speed. Once an applicable domain with a statistical model is obtained, as the second step, the signal can be analyzed and manipulated by applying mathematical algorithms to the small parameter set. The typical such processes are filtering, estimation, compresson, etc.

Accordingly, the modern signal processing problems deal with the cases in which either it is not easy to find a small parameter set, or the statistical model is not simple to handle. A particular but fairly common case of the former problem is when the signal representation in a signal domain contains a large vector, which contains only few non zero elements. However, the positions of such *active* elements are unknown. This is known as a *sparse* representation. If the signal domain is indexed, which is often the case, one may suggest to keep only the active parameters with their indexes forming a much smaller parameter set. This is a common method of data compression. However, working in such a low dimensional space normally leads to hard nonlinear problems. A more recent approach is then to focus on the high dimensional sparse representation and use the so called Compressive Sensing (CS) technique [2,3] to decrease the transformation cost. An example is the sampling transformation from the continuous time domain to the discrete one which could be extremely costly in practice. The CS idea proposes methods with much simpler linear samples followed by quantization and a digital processing on the samples [4]. A more recent technique also combines the sampling and quantization steps in a so called one-bit CS framework [5].

Finally, to address the second difficulty, we introduce the problem of Direction of Arrival (DOA) estimation based on the narrow-band, far-field models for which the solution could be extremely hard due to the highly nonlinear nature of the model. A promissing method to deal with such a problem is to perform the inverse procedure of the above and change the signal domain into the one, for which the estimation is extremely simple. However, we then translate the difficulties to the modeling phase. As it is shown in [6] and restated in Chapter 2, the measured data in a DOA estimation framework can be represented by the sparse pseudo-inverse of a truncated Fourier series, which is not a linear transform anymore. This can be practically done by the well known CS technique of Least Absolute

Shrinkage and Selection Operator (LASSO) [7]. This has proved to be a viable approach to DOA estimation, although several open issues remain, some of which are addressed in this thesis. We remark that although the presentation is done in the context of DOA estimation for clarity, most of the results and methods are applicable in a wider class of signal parameter estimation contexts.

1.1 Aim of thesis

This work explains, clarifies, and implements the idea of estimating DOAs by representing data in a sparse domain and applying the LASSO technique which was popularized in [6]. It further shows the result of applying such method comparing to the other techniques. We present solutions for different difficulties in defining and implementing the LASSO-based estimation method as follows.

First, as a Least Square (LS) optimization regularized by the ℓ_1 shrinkage and selection operator, any noisy LASSO implementation includes finding the best value of the regularization parameter. Although, there is a relation between the estimated number of sources and this parameter, the relation could not be expressed or solved analytically or by a relatively simple numerical method.

Second, although the convex LASSO technique guarantees a convergent solution of the estimator through convex optimization techniques, such an implementation is not suitable for the simultaneous estimation of the parameters and the regularization parameter. The Least Angle Regression (LARS) implementation of LASSO is not applicable due to the complex-valued nature of the problem and the complexity still grows fast by the size of grid.

Third, the LASSO estimator can be interpreted as a Bayesian estimator with a Laplacian prior over the sparse signal domain. The Bayesian properties of such a method are not known in connection to the compressed index representation.

Finally, defining a continuous estimator based on LASSO is hard. Note that the new representation is expected to be sparse in a continuous domain which is not well-defined in a functional variation theory. A solution might be to introduce a measure-based theory of estimation instead of a functional-based one.

Paper 1 concerns the first and the third problems. It also tries to examine LASSO as a Bayesian estimator rather than just a computationally efficient one. Paper 1 also tries to utilize such a Bayesian interpretation to get a better DOA modeling. Yet, the Bayesian properties of LASSO are still vague and inapplicable. In Paper 2 we address the first and the second problems by introducing a generalization of LARS for complex valued problems. In paper 3 a completed implementation and a precise definition of LASSO-based estimation is given under the name of Continuous LASSO (CLASS).

1.2 Thesis outline

In the mathematical derivation of the technique, we try to be as precise as possible. Thus, some proofs need a background of advanced mathematics such as complex analysis, measure and integration theories. The reader is assumed to be familiar to such concepts. However, such discussions may be omitted by an unfamiliar reader without losing the consistency.

A good review of estimation theory introducing general concepts of unbiasedness, admissibility, Bayesian estimator, Maximum A Posterior (MAP) rule, etc can be found in [8]. The reader is also encouraged to study [9] for an introduction to the nonlinear optimization, especially for convex non-smooth problems. The basic rules of electromagnetic models can be found in [10]. We start the introductory part by introducing the concept of DOA estimation and the conventional models as well as the new sparse formalism in Chapter 2. We then move to the solutions and implementation issues in Chapter 3. Finally, we show an attempt to generalize the continuous sparse regression in a practically suitable framework in Chapter 4.

Chapter 2

DOA Estimation

This work concerns the problem of localizing a set of sources by receiving their transmitted data through a set of localized sensors. It has found various direct and indirect applications in many engineering fields [11–13]. An interesting subclass of such an estimation problem is when the maximal sensor separation is much smaller than the minimal distance between the sources and the sensors, in which case the set of sensors is referred to as a *sensor array*, or simply an *array*. One may imagine a simple example of an earth observer watching stars, in which case the array is defined by the retinal cells, which are obviously much closer to each other compared to the distance of a star from earth. Our biological observatory example reveals some further aspects of the estimation problem. The observer sky may be assumed to be projected to a sphere centered at the observer's location. In fact, the sky objects were assumed to lie on the surface of such a celestial sphere and move by an Aristotelean model for centuries! We conclude that the information about the source distances could not be retrieved even with the extremely powerful processor of the human brain. Thus, the question reduces to estimating the direction of sources in the case of observing by an array. We refer to this problem as the Direction of Arrival (DOA) estimation.

2.1 The Narrow-band, far-field DOA Model

Now, we proceed to express the problem by mathematics and numbers. First, note that we concern the case that the information is carried by an electromagnetic wave. However, the final model is acceptable for a wider class of problems. Assume a set of n analog data signals $s_i(t)$ for $i = 1, \ldots, n$ transmitted from point sources at locations represented by the position vectors $\mathbf{y}_i$ in a 3-dimensional Euclidean space respectively. Now, take a set

of m sensors at the locations $\mathbf{x}_j$ in the same coordinates for $j = 1, \ldots, m$. Without loss of generality, we may assume that the coordinate origin is close to the array.

The exact relation between the measured and the transmitted information is governed by Maxwell's equations. However, the fact that the sensors and sources are far apart simplifies the analysis. Note that from the electromagnetic superposition rule, a generated observable field $\phi_j(t) = \phi(\mathbf{y}_j, t)$ (e.g. electrical field) at the j^{th} sensor point can be written as $\phi_j(t) = \sum_{i=1}^{n} \phi_j^i(t)$, where $\phi_j^i(t) = \phi^i(\mathbf{y}_j, t)$ denotes the field introduced by i^{th} source. Also, note that the relation between $s_i(t)$ and $\phi_j(t)$ is given by a Linear, Time-Invariant (LTI) system. By the Maxwell laws, the fields $\phi^i(\mathbf{x}, t)$ satisfy the wave equation

$$\nabla^2 \phi^i(\mathbf{x}, t) - \frac{1}{c^2} \frac{\partial^2 \phi^i(\mathbf{x}, t)}{\partial t^2} = s_i(t)\delta(\mathbf{x} - \mathbf{x}_i), \tag{2.1}$$

where ∇^2, δ and c denote the Laplace operator, the Dirac impulse function and the wave speed respectively [10]. Note that fixing $\mathbf{x}$, this equation and the additional causality condition give a causal LTI system, with $s_i(t)$ as the input and $\phi^i(\mathbf{x}, t)$ as the output. The impulse response of the system $h_i(\mathbf{x}, t)$ can be found by setting $s_i(t) = \delta(t)$ and assuming $h_i(\mathbf{x}, t) = 0$ for $t < 0$ as the initial condition. Taking the multidimensional Fourier transform with respect to $\mathbf{x}$ in (2.1) and solving the time differential equation, we get $h_i(\mathbf{x}, t) = h_i'(\mathbf{x}, t)u(t)$, where

$$h_i'(\mathbf{x}, t) = \int_{\mathbb{R}^3} \exp(j\mathbf{k}^T(\mathbf{x} - \mathbf{x}_i)) \frac{\sin(c\|\mathbf{k}\|t)}{2c\|\mathbf{k}\|} \mathrm{d}^3\mathbf{k}, \tag{2.2}$$

and where $u(t)$ is the Heavyside step function [1]. Taking the Fourier transform, the transfer function can also be written as

$$H_i'(\mathbf{x}, \omega) = \int_{\mathbb{R}^3} \exp(j\mathbf{k}^T(\mathbf{x} - \mathbf{x}_i)) \frac{\delta(\omega - c\|\mathbf{k}\|) - \delta(\omega + c\|\mathbf{k}\|)}{4jc\|\mathbf{k}\|} \mathrm{d}^3\mathbf{k}, \tag{2.3}$$

Take a coordinate system for $\mathbf{k}$ whose z axis is along $\mathbf{x} - \mathbf{x}_i$. Note that the radial component of the corresponding polar coordinate r is equal to $\|\mathbf{k}\|$. Denoting $R = \|\mathbf{x} - \mathbf{x}_i\|$, (2.3) can be simplified by expanding in polar coordinates as

$$H_i'(\mathbf{x}, \omega) = \int_0^\infty \int_\Omega \exp\left(jrR\cos\theta\right) \frac{\delta(\omega - cr) - \delta(\omega + cr)}{4jc} r \mathrm{d}r\mathrm{d}\Omega, \tag{2.4}$$

where $\mathrm{d}\Omega = \sin\theta \mathrm{d}\theta\mathrm{d}\phi$ is the infinitesimal element of the solid angle. After some manipulations, (2.4) can be simplified to

$$H_i'(\mathbf{x}, \omega) = \frac{\sin\frac{\omega R}{c}}{2jRc}, \tag{2.5}$$

which in return can be inversely transformed to $h'_i(\mathbf{x},t) = \frac{\delta(t-\frac{R}{c})-\delta(t+\frac{R}{c})}{4Rc}$. Thus,

$$h_i(\mathbf{x},t) = h'_i(\mathbf{x},t)u(t) = \frac{\delta(t-\frac{R}{c})}{4Rc},$$
$$H_i(\mathbf{x},\omega) = \frac{e^{-j\omega\frac{R}{c}}}{4Rc}. \tag{2.6}$$

Together, we conclude that the received signal at the j^{th} sensor can be written as

$$\phi_j(t) = \sum_{i=1}^{n} \frac{s_i(t-\frac{R_{ij}}{c})}{4R_{ij}c},$$
$$\Phi_j(\omega) = \sum_{i=1}^{n} \frac{e^{-j\omega\frac{R_{ij}}{c}}}{4R_{ij}c} S_i(\omega), \tag{2.7}$$

where $R_{ij} = \|\mathbf{x}_i - \mathbf{y}_j\|$.

2.1.1 Narrow-band Model

A narrow-band signal is one whose Fourier transform has a narrow support interval $[\omega_0 - \delta\ \omega_0 + \delta]$. A real-valued narrow-band signal $s(t)$ can be expressed as $s(t) = \Re(e^{j\omega_0 t}\tilde{s}(t))$, where $\tilde{s}(t)$ is a low-pass signal known as the *complex envelope* of $s(t)$. If a narrow-band signal is transformed by an LTI system with a transfer function $H(\omega)$, the result is another narrow-band signal $x(t)$ with the same frequency support band. The transfer function can be expanded by tailor series around ω_0 as $H(\omega) = H(\omega_0) + (\omega-\omega_0)\frac{\mathrm{d}H}{\mathrm{d}\omega}(\omega_0) + \frac{1}{2}(\omega-\omega_0)^2\frac{\mathrm{d}^2H}{\mathrm{d}\omega^2}(\omega_0) + \ldots$ for $|\omega - \omega_0| < \delta$, which implies that

$$\tilde{x}(t) = H(\omega_0)\tilde{s}(t) + \frac{\mathrm{d}H}{\mathrm{d}\omega}(\omega_0)\frac{\mathrm{d}\tilde{s}}{\mathrm{d}t}(t) + \ldots. \tag{2.8}$$

In this work, we are interested in the zeroth order of such an expansion with respect to the bandwidth 2δ. In this case, we simply get $\tilde{x}(t) = H(\omega_0)\tilde{s}(t)$ and (2.7) is written as

$$\tilde{\phi}_j(t) = \sum_{i=1}^{n} \frac{e^{j\omega_0\frac{R_{ij}}{c}}}{4R_{ij}c}\tilde{s}_i(t). \tag{2.9}$$

2.1.2 Far-field Approximation

At this point, we refer to the definition of an array as a small-size group of sensors far departed from the set of sources. In this case, neither the location parameters R_{ij}, nor the original signals $s(t)$ are given or retrievable from the received data. Instead, one may introduce a time shifted and scaled

observable version of these signals as references and express the model in terms of such vectors. Assume that there exists an imaginary sensor at the origin. For narrow-band transmitted signals, the received one at this point is given by

$$\phi(0,t) = \sum_{i=1}^{n} \frac{e^{j\omega_0 \frac{\|\mathbf{x}_i\|}{c}}}{4\|\mathbf{x}_i\|c} \tilde{s}_i(t) = \sum_{i=1}^{n} \tilde{s}_{\text{ref},i}(t), \tag{2.10}$$

where we introduce $\tilde{s}_{\text{ref},i}(t) = \frac{e^{j\omega_0 \frac{\|\mathbf{x}_i\|}{c}}}{4\|\mathbf{x}_i\|c} \tilde{s}_i(t)$. Noting that $\|\mathbf{y}_j\|$ is much smaller than $\|\mathbf{x}_i\|$, we get

$$R_{ij} = \|\mathbf{x}_i - \mathbf{y}_j\| = \|\mathbf{x}_i\| - \frac{\mathbf{x}_i^T \mathbf{y}_j}{\|\mathbf{x}_i\|} + O(\frac{\|\mathbf{y}_j\|}{\|\mathbf{x}_i\|}), \tag{2.11}$$

which implies that (2.9) can be approximated by

$$\tilde{\phi}_j(t) = \sum_{i=1}^{n} e^{j\boldsymbol{\nu}_i^T \mathbf{y}_j \frac{\omega_0}{c}} \tilde{s}_{\text{ref},i}(t) \tag{2.12}$$

up to the zeroth order of $\frac{\|\mathbf{y}_j\|}{\|\mathbf{x}_i\|}$, where $\boldsymbol{\nu}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$ is the unit vector along the DOA. The factor $k = \frac{\omega_0}{c} = \frac{2\pi}{\lambda}$ if often called the wavenumber and $\mathbf{k} = -k\boldsymbol{\nu}$ is known as the wave vector.

2.1.3 Further Simplifications

In this work, we are interested in the far field model introduced in (2.12). The receive signal $x_j(t)$ from the j^{th} sensor is contaminated with noise, i.e. $x_j(t) = \tilde{\phi}_j(t) + n_j(t)$, where $n_j(t)$ is the additive noise process at the j^{th} sensor. From now on, we usually drop the tilde notation for the complex envelope of the narrow band signals. Furthermore, the given signals are sampled at $t = nT$, in which case we may denote $s(n) = s(nT)$ for any signal s. However, the variable t often denotes the discrete time as well. Accordingly, the received sample vector $\mathbf{x}(t) = [x_1(t)\ x_2(t) \dots x_m(t)]^T$ for $t = 1, 2, \dots, T$ can be written as

$$\mathbf{x}(t) = \sum_{i=1}^{n} \mathbf{a}(\boldsymbol{\nu}_i) s(t) + \mathbf{n}(t), \tag{2.13}$$

where

$$\mathbf{a}(\boldsymbol{\nu}) = \begin{bmatrix} e^{j\boldsymbol{\nu}^T \mathbf{y}_1 \frac{\omega_0}{c}} \\ e^{j\boldsymbol{\nu}^T \mathbf{y}_2 \frac{\omega_0}{c}} \\ \vdots \\ e^{j\boldsymbol{\nu}^T \mathbf{y}_m \frac{\omega_0}{c}} \end{bmatrix} \tag{2.14}$$

is known as the *steering vector* of the DOA corresponding to $\boldsymbol{\nu}$. Note that $\frac{2\pi c}{\omega_0} = \lambda$ is the wavelength at the central frequency. In summary, the goal of DOA estimation is to find estimates of the parameters $\boldsymbol{\nu}_i$ given a set of observations $\mathbf{x}(t)$ and the model in (2.13). However, we mostly interested in a planar case, in which the source and sensor positions can be represented by a planar polar coordinate, and the DOA $\boldsymbol{\nu}$ can be represented by the angle coordinate θ. Assuming the i^{th} sensor at the position (θ_i, ρ_i) in the polar system, the product $\boldsymbol{\nu}^T \mathbf{y}_i$ is equal to $\rho_i \cos(\theta - \theta_i)$ and the steering vector of the DOA θ can be written as

$$\mathbf{a}(\theta) = \begin{bmatrix} e^{j2\pi \frac{\rho_1}{\lambda} \cos(\theta - \theta_1)} \\ e^{j2\pi \frac{\rho_2}{\lambda} \cos(\theta - \theta_2)} \\ \vdots \\ e^{j2\pi \frac{\rho_m}{\lambda} \cos(\theta - \theta_m)}. \end{bmatrix} \tag{2.15}$$

A further simplified case is when the planar array is linear, i.e. $\theta_i = 0$ and the sensors are separated uniformly with separation d, in which the coordinate system can be chosen such that $\rho_i = (i-1)d$ for $i = 1, \ldots, n$. Accordingly, the steering vector can be written as

$$\mathbf{a}_{\text{ULA}}(\theta) = \begin{bmatrix} 1 \\ e^{j2\pi \frac{d}{\lambda} \cos(\theta)} \\ e^{j2\pi \frac{2d}{\lambda} \cos(\theta)} \\ \vdots \\ e^{j2\pi \frac{(m-1)d}{\lambda} \cos(\theta)}. \end{bmatrix} \tag{2.16}$$

Finally, note that for a ULA the term $\phi = 2\pi \frac{d}{\lambda} \cos(\theta)$, known as the *electrical angle*, determines the steering vector. The ULA estimation is ambiguous in this case. To see this, note that $\mathbf{a}_{\text{ULA}}(\theta) = \mathbf{a}_{\text{ULA}}(-\theta)$. Thus, the estimation is limited to a subregion of the DOAs. However, it is simple to check that the maximal region of the upper half plain is attained when $d = \frac{\lambda}{2}$. In this work, we mostly focus on the half-wavelength ULA for analysis, while most of the methodology is applicable to a more general case. The steering vector may be indexed by electrical angle as $\mathbf{a}_{\text{ULA}}(\phi)$ or simply $\mathbf{a}(\phi)$ whenever there is no risk of confusion. The model in (2.13) can also be written as

$$\mathbf{x}(t) = \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t) + \mathbf{n}(t), \tag{2.17}$$

where $\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1)\ \mathbf{a}(\theta_2) \ldots \mathbf{a}(\theta_n)]$ for a collection of DOAs $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.

2.2 DOA Estimation

There is often a statistical model for the noise process $\mathbf{n}(t)$, which describes the statistical relation between the received data and the DOAs. In the simplest case, the noise process is a centered, circularly symmetric, white and uncorrelated vector Gaussian process, which means that the joint probability density function $p(\mathbf{n}(t_1), \mathbf{n}(t_2), \ldots, \mathbf{n}(t_k))$ for any time instance sequence $t_1, t_2, \ldots, t_k$ can be written as

$$p(\mathbf{n}(t_1), \mathbf{n}(t_2), \ldots, \mathbf{n}(t_k)) = \frac{1}{(\pi\sigma^2)^{mk}} e^{-\frac{\sum_{i=1}^{k} \|\mathbf{n}(t_i)\|^2}{\sigma^2}}. \tag{2.18}$$

The relation (2.18) implies that

$$\begin{aligned} \mathcal{E}(\mathbf{n}(t_1)\mathbf{n}(t_2)^H) &= \delta(t_1 - t_2)\sigma^2\mathbf{I} \\ \mathcal{E}(\mathbf{n}(t_1)\mathbf{n}(t_2)^T) &= \mathbf{0}. \end{aligned} \tag{2.19}$$

By estimating DOAs, we actually attempt to estimate a finite set $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$ of DOAs. Taking Θ_0 as the collection of all such sets, it is simple to check that the problem is ill-posed. This can be seen by noting that (2.17) has many mostly high dimensional solutions even with $\mathbf{n}(t) = 0$. Thus, in a Bayesian manner, some prior models are necessary. Note that the parameters $\mathbf{s}(t)$ are treated as *latent* or *hidden* variables. One way to deal with such a problem is to estimate the hidden parameters jointly with the DOAs. In this case, we may introduce Θ as the collection of all sets containing a finite number of pairs $(\theta_i, \{\mathbf{s}_i(t)\}_{t=1}^T)$. We will return to such representations later. However, the much simpler conventional method is to treat $\boldsymbol{\theta}$ as a vector in $\mathbb{R}^n$ with an unknown dimension n, which is referred to as the *model order*. The estimation is then performed for each model order separately and the best order is selected through a procedure known as the *model order selection*. In the following we review this method briefly.

2.3 Maximum Likelihood Estimation

The first step in performing the DOA estimation is to express the likelihood function. The relations (2.17) and (2.18) enable us to derive the likelihood function $p(\mathbf{N}, \mathbf{X}|\boldsymbol{\theta}, \mathbf{S})$, where $\mathbf{N} = [\mathbf{n}(1), \mathbf{n}(2), \ldots, \mathbf{n}(t)]$, $\mathbf{S} = [\mathbf{s}(1), \mathbf{s}(2), \ldots, \mathbf{s}(t)]$ and $\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(t)]$ as

$$p(\mathbf{N}, \mathbf{X}|\boldsymbol{\theta}, \mathbf{S}) = p(\mathbf{N}) \left[\prod_{t=1}^{T} \delta(\mathbf{x}(t) - \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t) + \mathbf{n}(t)) \right], \tag{2.20}$$

where $\delta(x)$ denotes the Dirac delta function so that $\int_{x\in B} \delta(x)\mathrm{d}x = 1$ if $0 \in B$, otherwise it is zero. Furthermore, $p(\mathbf{N})$ is the noise probability density function given by (2.18) with $t_i = i$ for $i = 1, \ldots, n$. It is now simple to check that the pdf $p(\mathbf{N}, \mathbf{X}|\boldsymbol{\theta}, \mathbf{S})$ can be marginalized with repect to $\mathbf{N}$, to get $p(\mathbf{X}|\boldsymbol{\theta}, \mathbf{S}) = \int_{\mathbb{C}^{m\times T}} p(\mathbf{N}, \mathbf{X}|\boldsymbol{\theta}, \mathbf{S})\mathrm{d}\mathbf{N}$ as,

$$p(\mathbf{X}|\boldsymbol{\theta}, \mathbf{S}) = \frac{1}{(\pi\sigma^2)^{mT}} e^{-\frac{\sum_{t=1}^{T} \|\mathbf{x}(t)-\mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t)\|^2}{\sigma^2}}, \tag{2.21}$$

also known as the likelihood function. One straightforward method is the Maximum Likelihood (ML) estimator [14] given by

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{S}}) = \arg\max_{\boldsymbol{\theta},\mathbf{S}} p(\mathbf{X}|\boldsymbol{\theta}, \mathbf{S}) = \arg\min_{\boldsymbol{\theta},\mathbf{S}} \sum_{t=1}^{T} \|\mathbf{x}(t) - \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t)\|^2, \tag{2.22}$$

which solving for $\mathbf{S}$, can be written as [15]

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathrm{Tr}[\mathbf{P}^{\perp}_{\mathbf{A}(\boldsymbol{\theta})}\mathbf{R}] \tag{2.23}$$

where $\mathbf{R} = \frac{1}{T}\sum_{i=1}^{T} \mathbf{x}(t)\mathbf{x}^H(t)$ is the sample correlation matrix and $\mathbf{P}^{\perp}_{\mathbf{A}(\boldsymbol{\theta})}$ denotes the orthogonal projection matrix.

2.4 Model Order Selection

It is well known [8] that the ML estimator of (2.23) is consistent in an asymptotic case where the number of snapshots T tends to infinity if the model order n is known. As we previously explained, when n is unknown, the estimation is performed for each model order and an order is chosen based on these estimates. Let us analyze this case in more details. In such an asymptotic case we have

$$\mathbf{R} \to \mathbf{A}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}\mathbf{A}(\boldsymbol{\theta}_0)^H + \sigma^2\mathbf{I}, \tag{2.24}$$

where $\boldsymbol{\Sigma} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{s}_0(t)\mathbf{s}_0^H(t)$. Note that the zero index denotes the true values. Substituting (2.24) in (2.23), we get

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathrm{Tr}[\mathbf{P}^{\perp}_{\mathbf{A}(\boldsymbol{\theta})}\mathbf{A}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}\mathbf{A}(\boldsymbol{\theta}_0)^H] + \sigma^2(m-n). \tag{2.25}$$

Note that since the estimation is made on a compact set of variables the cost function for finite T in (2.23) tends uniformly to the one in (2.25) almost surely. This guarantees that the ML estimate also tends in probability to

the optimal point in (2.25) for each n. It is simple to check that since Σ is positive semidefinite the first term in (2.25) is always positive so that it achieves its minimum at a nonzero value. For $n \geq n_0$ this is only possible if $\mathbf{P}^{\perp}_{\mathbf{A}(\boldsymbol{\theta})}\mathbf{A}(\boldsymbol{\theta}_0) = 0$. If $n = n_0$, this implies that $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. However, for $n > n_0$ the solution is not unique but always contains the true DOAs if we limit $n < \frac{m}{2}$.

One simple way to choose n is to interpret the problem as a hypothesis testing. In this case, the order is decided by minimizing (2.25) with respect to n [16]. One may note that this gives an inconsistent estimated model due to the term $-n\sigma^2$ in (2.25). In other words, the hypothesis testing result tends to overestimate the model order. One simple way to overcome this problem is to add a linear term Kn with respect to the model order. It is simple to see that there exists a value of K giving a consistent estimation. However, this value depends on the level of the noise variance σ^2 and the source correlation matrix Σ. If these are known values, the consistent values of K are bounded by

$$\sigma^2 + \min_{\boldsymbol{\theta}_0, \boldsymbol{\theta}, n(\boldsymbol{\theta}) < n(\boldsymbol{\theta}_0)} \frac{\mathrm{Tr}[\mathbf{P}^{\perp}_{\mathbf{A}(\boldsymbol{\theta})}\mathbf{A}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}\mathbf{A}(\boldsymbol{\theta}_0)^H]}{n(\boldsymbol{\theta}) - n(\boldsymbol{\theta}_0)} > K > \sigma^2. \tag{2.26}$$

Note that the minimum at the left hand side of (2.26) equals zero and the above method is inconsistent without any further assumptions. To see this, consider a case in which $\boldsymbol{\theta}_0$ includes two very close DOAs. It is now possible to take θ as a DOA vector with one less element but very close to θ_0. It is now obvious that the two bounds get close as the two elements of θ_0 merge. This shows a fundamental ambiguity of the vectorized formulation of the DOA estimation. A DOA vector with equal entries is identical to the shorter one resulting by removing repeated entries. Our example also shows a way to deal with the problem. To avoid such ambiguous cases, we need to limit the resolution so that it gets impossible for two DOAs to get very close. The smaller the resolution limit, the wider the consistent region in (2.26). Note also that choosing a K value in the middle of the range is preferred, since it gives a deeper minimum point in the asymptotic cost with respect to n, which can be obtained practically with less number of snapshots T. In summary the linear additional term, which is equivalent to a Bayesian hypothesis testing with an exponential prior $p(n) = e^{-Kn}$ establishes a trade off between the number of snapshots and the resolution limit. Further analysis around the asymptotic point $T = \infty$ reveals that the level of this trade off itself depends on the level of SNR $\frac{\Sigma}{\sigma^2}$. To illustrate this, take the extreme case of $T = 1$ and assume that $\sigma = 0$. By a method similar to the above argument, it is simple to check that the estimator (2.23) is consistent and assuming a finite resolution, there exists a consistent range

for K. It is now simple to see that the uniform minimality of the correct order is preserved by increasing the noise level until a certain level.

If Σ and σ^2 are unknown, which is a common case, one may attempt to use an estimates of these parameters. One other approach might be to integrate out the unknown parameters by some Bayesian priors and perform a similar procedure as above, which is known as the Bayesian Likelihood Ratio Test (BLRT) [17]. We are not going to consider BLRT tests in this work. Then, one simple way to receive such estimates is to maximize the likelihood function with respect to σ to get

$$\hat{\boldsymbol{\sigma}}^2_{\mathrm{ML}} = \frac{\mathrm{Tr}[\mathbf{P}^{\perp}_{\mathbf{A}(\hat{\boldsymbol{\theta}})} R]}{m}, \tag{2.27}$$

which in fact gives a Generalized LRT (GLRT) for the model order [17]. The estimate of Σ can also be found by

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{s}}(t)\hat{\mathbf{s}}^H(t) = \mathbf{A}^{\dagger}(\hat{\boldsymbol{\theta}})\mathbf{R}\mathbf{A}^{\dagger}(\hat{\boldsymbol{\theta}})^H, \tag{2.28}$$

where $\mathbf{A}^{\dagger}$ shows the *pseudo-inverse* of $\mathbf{A}$. Although $\hat{\Sigma}$ is a consistent estimator of Σ, $\hat{\sigma}^2_{\mathrm{ML}}$ is asymptotically biased. It is simple to check that by replacing the denominator of (2.27) by $m - n$, we will get an unbiased estimator. Thus we may use

$$\hat{\sigma}^2_{\mathrm{unbiased}} = \frac{\mathrm{Tr}[\mathbf{P}^{\perp}_{\mathbf{A}(\hat{\boldsymbol{\theta}})} R]}{m - n} \tag{2.29}$$

to find a consistent range in (2.26).

2.4.1 Information Criteria

We previously explained the main rule of selecting a model order, that is to use the principle of consistency in asymptotic cases. Still, finding a consistent range of the parameters describing the Bayesian model order priors is mainly a subjective task. A remarkable development of these ideas have been provided by introducing the information criteria, pioneered by the prominent work of Akaike [18]. Note that the term "information" or "information theory" merely represents those aspects of the stochastic data revealed by observing an asymptotically long sequence of measurements. It is also worth noting that the deterministic DOA estimation problem does not fit completely to the following framework of information theory, since the number of free parameters grows with T as the source vector is an independent unknown vector at each snapshot. Still, information criteria might be tried to the problem.

Kullback Leibler Distance

Let us assume that in a general estimation setup the observed data follow a true distribution $f(\mathbf{x})$, while we are to choose from an indexed set of candidate ones $f(\mathbf{x};\theta)$. If a long sequence of outcomes $\mathbf{x}_t$ is observed the normalized logarithm of the joint likelihood function

$$\frac{1}{N}\sum_{t=1}^{T}\log f(\mathbf{x}_t;\theta), \tag{2.30}$$

tends to the expected value

$$H(f(\mathbf{x}), f(\mathbf{x};\theta)) = \mathcal{E}_x(\log f(X;\theta)) = \int_X f(\mathbf{x})\log f(\mathbf{x};\theta)\mathrm{d}\mathbf{x}, \tag{2.31}$$

under some mild conditions. It is simple to check that the asymptotic likelihood is maximized if for some θ, we have $f(\mathbf{x};\theta) = f(\mathbf{x})$. For any arbitrary θ, and the function $g(\mathbf{x}) = f(\mathbf{x};\theta)$ the difference between the maximal value $H(f,f)$ and $H(f,g)$ is called the *Kullback Leibler* distance between f and g and is denoted by $D(f||g)$. In other words,

$$D(f||g) = \int_X f(\mathbf{x})\log\frac{f(\mathbf{x})}{g(\mathbf{x})}\mathrm{d}\mathbf{x} \tag{2.32}$$

is a positive quantity which is zero if an only if $f = g$. Now, it is simple to see that the maximum likelihood estimation asymptotically tends to minimizing $D(f(\mathbf{x})||f(\mathbf{x};\theta))$ although we are not aware of the true distribution $f(\mathbf{x})$.

Akaike Information Criterion

When the candidate models are of different orders, the above argument still holds true. The best model order is the one containing the closest model in the KL sense. However, the strategy of maximizing the likelihood function over all model orders by a finite number of snapshots needs a high number of snapshots to meet the asymptotic consistency conditions. Denote the ML estimate and the minimum KL distance point of the order n by $\hat{\theta}_n$ and $\bar{\theta}_n$ respectively. From the above argument, $\hat{\theta}_n \to \bar{\theta}_n$ as $T \to \infty$. A more careful analysis reveals that $\hat{\theta}_n = \bar{\theta}_n + \nu_n$ up to the first order of T, where

$$\nu_n \sim \mathcal{N}\left(0, \frac{1}{T}\mathbf{J}_n^{-1}\right), \tag{2.33}$$

where $\mathbf{J}_n$ is the *Fisher information* matrix given by

$$\mathbf{J}_n = -\mathcal{E}_x\left(\frac{\partial^2\log f(\mathbf{x};\bar{\theta}_n)}{\partial\theta\partial\theta^T}\right) = -\int_X f(\mathbf{x})\frac{\partial^2\log f(\mathbf{x};\bar{\theta}_n)}{\partial\theta\partial\theta^T}\mathrm{d}\mathbf{x}, \tag{2.34}$$

Then, the minimum Negative Log-Likelihood (NLL) can be approximated by

$$-\sum_{t=1}^{T}\log f(\mathbf{x}_t;\hat{\theta}_n) = -\sum_{t=1}^{T}\log f(\mathbf{x}_t;\bar{\theta}_n) - \sum_{t=1}^{T}\nu_n^T\frac{\partial \log f(\mathbf{x}_t;\bar{\theta}_n)}{\partial\theta} - \frac{1}{2}\sum_{t=1}^{T}\nu_n^T\frac{\partial^2 \log f(\mathbf{x}_t;\bar{\theta}_n)}{\partial\theta\partial\theta^T}\nu_n \tag{2.35}$$

up to the first order of T. For large enough T one may assume that values are concentrated about their expectation so that the first term is close to T times the minimum KL distance of order n, the second term is zero and the third term will be simple to compute to be close to n. The third bias term slows down the convergence process. To speed up the convergence rate Akaike et. al. [18] proposed to add the term n to NLL to reach

$$AIC(n) = -\sum_{t=1}^{T}\log f(\mathbf{x}_t;\hat{\theta}_n) + n \tag{2.36}$$

and find the minimum.

Minimum Description Length

The KL divergence in (2.32) has another inspiring view. Let us consider the problem of coding a set of data, which is to express a long sequence of data $\mathbf{x}_t$ $(t = 1, 2, \ldots)$ by a finite alphabet. Note that if the data is continuous this is impossible, in which case we are instead interested in the rate of increase in the code length by growing the quantization level. This is known as the *differential information* [19]. We assume an i.i.d. model with a density $f(\mathbf{x})$ for the data. Then, by Shannon's theory of information it can be shown that for each f there exists a lossless Shannon code reaching the minimal average length of $T\mathcal{E}_x\log f(\mathbf{x})$. Furthermore, if the source encoder corresponding to the model $g(\mathbf{x})$ is instead used, the average code length is increased by $TD(f||g)$. This means that the principle of minimum KL distance is equivalent to the Minimum Description Length (MDL) in this scenario. The MDL principle first appeared as a non-Bayesian alternative to the Minimum Message Length (MML) rule [20]. It roughly states that the best model is the one which assigns the shortest Shannon code to the observed data.

Let us denote the set of all models sharing an identical order by a class. First, assume a scenario in which there only exists one model order, i.e. there is one class. It is now simple to see that the MDL intra-class principle simply means to choose the smallest Negative Log Likelihood (NLL), which gives the ML estimate. Now, if there exists multi-class models, the MDL idea needs justification. Note that to incorporate MDL rule in the model order selection strategy, each class should be assigned to an encoder with the

shortest average length, known as the *universal encoder*. It is simple to see that encoding each message with the Shannon encoder of its own ML model may not give a lossless code. Moreover, averaging is not well defined in this context, since there is no Bayesian information about unknown parameters. To deal with the above difficulties, Rissanen proposed the Shannon code of a distribution uniformly optimal over the class in a minimax manner as a universal code [21]. It turns out that this universal model is given by the Normalized Maximum Likelihood (NML) distribution.

$$f_{\mathrm{NML}}(\mathbf{x}) = \frac{1}{A} f(\mathbf{x}; \hat{\theta}_{\mathrm{ML}}(\mathbf{x})), \tag{2.37}$$

where A is a suitable normalizing factor (if it exists). We avoid mathematical complication and only state the final result that under mild conditions the universal code length of each class for T snapshots of data is approximately given by [21]

$$-\sum_{t=1}^{T} \log f(\mathbf{x}_t; \hat{\theta}_n) + \frac{n}{2} \log T \tag{2.38}$$

In [22] the possibility of utilizing information criteria in model order decision is discussed and it is shown that unlike the AIC rule, the MDL principle is consistent in the asymptotically many snapshot case.

2.5 DOA modeling by Spatial Sparsity

In this section, we introduce a different idea in estimating DOAs. This is a result of reinterpreting the model in (2.17). The simple fact behind this new interpretation is that there is no distinction between a case where there is no source along some direction θ and a case where the source along θ transmits the null signal $s(t) = 0$. Thus, one may assume in (2.17) that $\boldsymbol{\theta}$ includes every possible direction, but most corresponding sources transmit the null vector. Thus, the vector $\mathbf{s}$ becomes a sparse one. There are some difficulties in defining such an infinite dimensional vector, which traditionally led to defining a discretized space of DOAs [6]. However, we later introduce an alternative solution is Chapter 4, which is proper for the theoretical analysis.

Assume a grid of DOAs $\boldsymbol{\theta}^g$, from which we are to choose the estimates. We hope that for a dense grid these discretized estimates tend to the continuous ones. Then, the model (2.17) could approximately be written as

$$\mathbf{x}(t) = \mathbf{A}(\boldsymbol{\theta}^g)\mathbf{s}^g(t) + \mathbf{n}(t), \tag{2.39}$$

where $\mathbf{s}^g(t)$ is the long grid source vector whose elements are zero except when they correspond to a true DOA θ_i, in which case they are equal to the

true sources $s_i(t)$. The model in (2.39) is ill-posed, since the matrix $\mathbf{A}^g = \mathbf{A}(\boldsymbol{\theta}^g)$ has a huge null space. In return, the number of nonzero elements in $\mathbf{s}^g$, also known as ℓ_0 *semi-norm* may be bounded. This strongly resembles the compressive sensing setup [2]. However, the theoretical conditions of the latter do not generally fit to the current problem. Nevertheless, the model in (2.39) with ℓ_0 constraint is obviously equivalent to the model in (2.17) and the compressed sensing techniques of resolving the inverse problem are well suited to it. The overall ML estimator in this sparse regression setup can be written as

$$\begin{aligned} \{\hat{\mathbf{s}}^g(t)\}_{t=1}^T = \arg\min_{\mathbf{s}^g(t)} \sum \|\mathbf{x} - \mathbf{A}(\theta^g)\mathbf{s}^g(t)\|_2^2 \\ \text{subject to } n < n_0, \end{aligned} \tag{2.40}$$

where n is the number of indexes, the corresponding source of which is nonzero in at least one snapshot, and n_0 is an appropriate upper bound. Introducing the spatially sparse model of DOA, we finish this chapter and postpone the implementation solutions to the next one.

Chapter 3

Implementation Issues

In Chapter 2 we introduced the ML estimator of the DOAs in the Narrowband Far-field model of transmission. We showed that the estimation can be implemented by minimizing a group of Negative Log-Likelihood functions. We reviewed two such methods. In the first attempt, the DOAs were modeled by a vector of a given dimensionality n and the procedure is followed by a model order selection. In the second one, the likelihood is expressed based on spatial sparsity and a regression expression pairs, which includes every dimensionality. The process of maximizing the likelihood is identical to a joint deterministic estimation of the model order and the DOAs, which is not desirable as we have already discussed. The alternative solution is to modify the region of optimization in order to confine the dimension of the optimal point. However, this should be done in a way to sustain the convexity properties. In this Chapter we will discuss the ℓ_1 constraint which is the hart of many modern sparse estimation procedures.

3.1 The Vector Based Solutions

The ML criterion in (2.22) can be solved by different optimization techniques. Here, we introduce two such algorithms. We focus on the one-snapshot case for simplicity, thus omitting the important class of subspace algorithms [15].

3.1.1 Cyclic Coordinate Descent Algorithm

The optimization techniques are generally designed based on the properties of the cost function. A much less restrictive property is continuity, under which the coordinate descent algorithm has been introduced. The algorithm breaks the problem into multiple sequential steps of optimizing the cost over lines (line search), which are chosen by alternating the directions of a

coordinate system through the candidate point at each step. For simplicity, we take the simple coordinate system of $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ for DOAs, where $\mathbf{e}_i = (0, 0, \ldots, 0, 1, 0, 0, \ldots, 0)$ with one at the i^{th} element. Take an arbitrary DOA vector $\boldsymbol{\theta}$ and maximize the likelihood over the points $\boldsymbol{\theta} + \alpha\mathbf{e}_i$ and $\mathbf{s}+\beta\mathbf{e}_j$, for $\alpha, \beta \in \mathbb{R}$. Note that the element θ_i and s_j may change arbitrarily due to this subproblem. Minimizing with respect to s_j provides

$$(s_j)_{\min} = \begin{cases} \mathbf{a}^H(\theta_j)[\mathbf{n} - \mathbf{a}(\theta_i)s_i] & i \neq j \\ \mathbf{a}^H(\theta_i)\mathbf{n} & i = j \end{cases}, \tag{3.1}$$

where $\mathbf{n} = \mathbf{x} - \mathbf{A}(\boldsymbol{\theta})\mathbf{s}$. Substituting to (2.22), we get

$$(\theta_i)_{\min} = \begin{cases} \arg\min\limits_{\theta} \|\mathbf{P}_i^{\perp}(\mathbf{n} - \mathbf{a}(\theta)s_i)\|_2 & i \neq j \\ \arg\max\limits_{\theta} |\mathbf{a}^H(\theta)\mathbf{n}| & i = j \end{cases}, \tag{3.2}$$

Substituting the minimum points to $\boldsymbol{\theta}$ and $\mathbf{s}$ and alternating with respect to i, j, we get the complete algorithm.

A very important simplification in the above is to restrict the alternation over the indexes $i = j$. This algorithm has a different view as a Space Alternating Generalized Expectation (SAGE) maximization algorithm [23] and is better known as the RELAX algorithm in a more technical context [24]. A further simplification is to take the local maximum points of the spectrum $|\mathbf{a}^H(\theta)\mathbf{x}|$ as the DOA estimates, which is known as the conventional beamforming algorithm. There also exists many variants of these algorithms.

3.1.2 Newton's Method

A stronger assumption is that the cost function is twice differentiable, under which many interesting results may be deduced. Such a function has a quadratic approximation around each point $\mathbf{u} = [\boldsymbol{\theta}^T, \mathbf{s}^T]^T$ as

$$F(\mathbf{u} + \mathrm{d}\mathbf{u}) = F(\mathbf{u}) + \nabla F(\mathbf{u})^T \mathrm{d}\mathbf{u} + \frac{1}{2}\mathrm{d}\mathbf{u}^T \mathbf{H} \mathrm{d}\mathbf{u} \tag{3.3}$$

where $\mathbf{H}$ and ∇F are the Hessian matrix and the gradient vector respectively. Minimizing (3.3) at each step over a small neighborhood $\|\mathrm{d}\mathbf{u}\| \leq \epsilon$, we get the next step. This is the *quadratic trust region* method. There are different ways to perform the local optimization. However, if $\mathbf{H}$ is positive definite and $\|\mathbf{H}^{-1}\nabla F\| \leq \epsilon$ then the next point is given by $\mathbf{u} - \mathbf{H}^{-1}\nabla F$. The conditions are usually met when the initial point is close enough to an optimal point with a positive definite Hessian. In this case, the algorithm

is known as the method of Newton. Practically, this method follows an exhaustive search on a coarse grid to assure starting from a close point, which might not be tractable.

In a more general setup, the quadratic trust region method might be solved by a Quasi-Newton or conjugate gradient method [9]. The Hessian could also be modified to avoid instability of inversion, resulting in the Levenberg-Marquardt method. However, for the current study of DOA problems there exists many non-optimal fixed points of all such methods starting from an arbitrary initial point. For example a big class of such points for the Newton's method is identified by $\nabla F = 0$, which can be simplified to

$$\begin{aligned} \mathbf{a}^H(\theta_i)\mathbf{n} &= 0 \\ \Re(\mathbf{d}^H(\theta_i)\mathbf{n}) &= 0, \end{aligned} \tag{3.4}$$

where $\mathbf{d}(\theta) = \frac{\mathrm{d}\mathbf{a}}{\mathrm{d}\theta}$.

3.2 Why Convex Programming?

As we observed earlier, solving the ML optimization for the problem of DOA estimation is hard, and the vector based optimization by the current techniques, based on the line and local searches do not guarantee convergence to the global optimum point. In this section, we try to answer the converse question, namely, when is the convergence to the optimum point guaranteed?

A wide answer to the above question is provided by the Zangwill's theorem. All deterministic algorithms of optimizing a function f on a set S are defined by a set-valued algorithmic map $A : S \to 2^S$, where 2^S denotes the power set of S. The algorithm generates a sequence x_n starting from an arbitrary point x_0 such that $x_n \in A(x_{n-1})$ for $n = 1, 2, \ldots$. The Zangwill's theorem states that any accumulation point of x_n is a global optimum if *a)* $f(x_{n-1}) \geq f(x_n)$ and for non optimal x_{n-1}, $f(x_{n-1}) > f(x_n)$ and *b)* The mapping A is closed everywhere on S.

It is now possible to see that the previous examples of maximizing the likelihood function do not meet the above requirements. For example the coordinate descent algorithm (RELAX) violates the condition *a*. This implies that there exists a non-optimal point in the optimization region for which the coordinate descent does not decrease the cost anymore. The situation for the gradient descent and the Newton method is even worse, since they also suffer lack of continuity. See [9] for more details.

A well known class of optimization methods for which the coordinate descent algorithm meets Zangwill's conditions is the linear programming,

in which the cost and constraints are expressed by linear functions. The constrained region is also expressed by a finite intersection of half spaces. In fact, the finiteness of the number of half spaces is unnecessary and one may assume a case in which the region S is an intersection of a family of half spaces. It is simple to show that this is the case if and only if S is convex. There also exists a standard method to convert any optimization to one with a linear cost. Note that minimizing $f(x)$ for $x \in S$ is equivalent to minimizing a variable y subject to the inequality $y \geq f(x)$ for some $x \in S$. It is now simple to see that the new region, known as the epigraph of f, is convex if and only if f is a convex function. Thus, the convex programming inherits many of the nice properties of the linear programming.

3.3 ℓ_1 Constraint

The DOA model by the spatial sparsity in (2.39) is a straightforward conversion of (2.17), while it had a great impact on the implementation of the ML estimator. A fairly simple approximation to (2.39) is to modify the constraint to one with better implementation properties. According to the above, one may propose a convex constraint with an "edged" level set such that the optimization tends to choose the corners, i.e. the sparse points. These are all served by the ℓ_1 norm, which is defined as the sum of the absolute values of the elements of a vector $\mathbf{s}$, and is denoted by $\|\mathbf{s}\|_1$. Accordingly, the approximated ML estimator of (2.40) for a single snapshot can be written as

$$\begin{aligned} \hat{\mathbf{s}}^g = \arg\min_{\mathbf{s}^g} \sum \|\mathbf{x} - \mathbf{A}(\theta^g)\mathbf{s}^g(t)\|_2^2 \\ \text{subject to} \|\mathbf{s}^g\|_1 < C, \end{aligned} \tag{3.5}$$

where C is a suitable constant. The optimization in (3.5) is known as the Least Absolute Shrinkage and Selection Operation (LASSO). The LASSO optimization can be written in different equivalent forms by the Lagrange's duality. A most common form is the following

$$\hat{\mathbf{s}}^g = \arg\min_{\mathbf{s}^g} \frac{1}{2} \sum \|\mathbf{x} - \mathbf{A}(\theta^g)\mathbf{s}^g(t)\|_2^2 + \lambda\|\mathbf{s}\|_1, \tag{3.6}$$

which we will refer to as the standard form. The parameter λ is known as the Regularization Parameter (RP), which controls the trade off between the level of sparsity and the error level.

The standard LASSO optimization is an unconstrained convex one. Thus, it can be solved by any standard convex optimization technique [25]. Once the solution is formed, the DOAs can be estimated as the ones corresponding to the active indexes in $\hat{\mathbf{s}}$. However, it should be noted that

since the true DOA value may lie out of the grid point set, a number of active sources appear around the true value (or an estimate), which should be combined to one. One simple way is to take the peak only or to perform an averaging.

3.4 Bayesian Interpretation of LASSO

A very familiar concept since the invention of LASSO has been the Bayesian interpretation of the ℓ_1 penalty term. It is straightforward to see that the LASSO technique can be regarded as a Maximum A-posteriori Probability (MAP) estimator with a Laplacian prior density of the sources

$$p(\mathbf{s}^g;\mu) = \left(\frac{\mu^2}{2\pi}\right)^N e^{-\mu\|\mathbf{s}^g\|_1}, \tag{3.7}$$

where N is the number of grid points, the dimensionality of $\mathbf{s}^g$. This view, has a great impact. Recalling the model order selection problem in Section 2.4, one may identify the similarities to the current problem. The RP plays the role of a relaxed model order, which can be estimated in a hierarchical Bayesian manner. However, the characteristics of the estimate will be affected by the number of grid points N in (3.7) which seems irrelevant. Furthermore, the Laplacian prior in (3.7) is not a *conjugate prior*, meaning that the resulting posterior is not a Laplacian anymore. In the case of real-valued regression, a wise solution is made in [26] by introducing a Gibbs sampler of the Laplacian prior which has a convincing empirical justification. The important observation behind is that the Laplacian prior can be interpreted as a marginalized distribution of a centered Gaussian with an exponential hyper-prior on the variance, i.e.

$$\frac{\mu}{2}e^{-\mu|s|} = \int_0^\infty \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{|s|^2}{2\tau}} \frac{\mu^2}{2} e^{-\frac{\mu^2\tau}{2}} \mathrm{d}\tau, \tag{3.8}$$

from which it has already been proposed to substitute the exponential hyper-prior by the improper non-informative Jeffreys' prior $p(\tau) = \frac{1}{\tau}$ to remove the RP [27]. Instead, in [26] a Gamma hyper-prior on the RP is proposed, which in return introduces a new problem of matching the Gamma distribution parameters. Nevertheless, the above ideas have not been, and are not straightforward to be, extended to the complex case of DOA estimation, for which there has not been any genuine proposed Bayesian solution.

Chapter 4

Grid-less LASSO

In Chapter 2 we constructed a sparse framework to estimate the DOAs by a Narrow-band and far-field model. The approach was to discretize the space of DOAs and to select the active grid points by the LASSO method. However, the estimation performance depends on the choice of the grid. There has also been attempts to involve the off-grid parameters (the deviation of the true DOAs from the grid) in the estimation procedure by considering the Taylor expansion. Assuming the true DOAs as $\boldsymbol{\theta}_0$ and the closest grid points as $\boldsymbol{\theta}_0^g$, for a fine grid we have

$$\mathbf{A}(\boldsymbol{\theta}_0) \approx \mathbf{A}(\boldsymbol{\theta}_0^g) + \mathbf{D}(\boldsymbol{\theta}_0^g)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^g), \tag{4.1}$$

where $\mathbf{D}$ contains the first order derivatives of the steering vectors in $\mathbf{A}$. The direct application of such an approach leads to non-convex optimizations, which are hard and worthless to solve. Instead, we are going to answer a different question: "What is the result of adding every DOA to the dictionary $\mathbf{A}^g$?". It is natural to think that under some conditions the summation terms tend to integrals and LASSO becomes a variational optimization. Although this *Integral LASSO (I-LASSO)* might be interesting from theoretical point of view, it is hopeless from the computational perspective, for the variational optimization can only be solved directly by some finite element technique, which includes discretization again! During our research, we came up with a different generalization which develops the essence of Paper 3. Here, we give a deeper illustration.

The LASSO optimization can conceptually be performed in two stages. Define the active indexes $I(\mathbf{s}^g)$ as the set of all indexes with nonzero corresponding elements in a given vector $\mathbf{s}^g$. Then, one may optimize LASSO first by minimizing over $\mathbf{s}^g$ keeping I constant and then optimizing over all possible index sets I. Obviously, this can not be a feasible method, since the second stage needs a combinatorial search. In other words, if one could perform LASSO with the above procedure, there would be no need for LASSO,

since one could perform the same procedure to solve the exact ML. However, as a generalization, one may assume that the vector $\mathbf{s}^g$ grows as large as possible so that it contains every possible direction. Denote the set of all possible DOAs by S. Then, the collection of the active indexes becomes an arbitrary subset of S. It is not hard to see that the number of active indexes for any choice of the grid is always bounded by a unique bound. A closer look reveals that this bound is in fact the number of sensors m. Thus, it is natural to assume that our continuous extension of LASSO also achieves its optimal solution at a point with a *finite* active subset. Accordingly, the generalized LASSO, which we refer to as the CLASS optimization could be expressed as a two stage minimization, in the first of which the LASSO is performed over an arbitrary but fixed finite active DOA set $\boldsymbol{\theta}$. Then, the second stage is to minimize over all possible finite active sets. Still, the method looks infeasible. However, we try to show the superior properties of this definition in this chapter. The most important observation is that this generalization and the variational approach are equivalent. Thus, the generalized KKT conditions of one could be used for the other.

In the following, we clarify the above idea. The next sections contain mathematical material of higher level and might be ignored by an unfamiliar reader. However, this new setup is the key to the later development of the sparse estimation techniques to the author's belief.

4.1 Construction of the Regression Space

Consider the regression model in (2.17) with a general given manifold $\mathbf{a}(\theta)$ defining the regressors, where θ belongs to a compact set S. As we explained in the above, this model is ambiguous under permutation and introduction of repeated elements and the regressors are defined uniquely by the *finite set* of DOAs $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_n\}$. The cardinality of $\boldsymbol{\theta}$ is shown by $n(\boldsymbol{\theta}) \in \mathbb{N}$. The parameter vector is also defined uniquely by a function $\mathbf{s} : \boldsymbol{\theta} \to \mathbb{C}$ at each time. We refer to the pair of $(\boldsymbol{\theta}, \mathbf{s})$ as a *primal regression pair*. Let us denote the set of all primal regression pairs by Ψ_p. Note that (2.17) is also ambiguous under adding a null source. Let us define the *support* of a pair $\psi = (\boldsymbol{\theta}, \mathbf{s})$ as $I(\psi) = \{\theta \in \boldsymbol{\theta} | \mathbf{s}(\theta) \neq 0\}$. If $I = \boldsymbol{\theta}$, the pair is called irreducible. Two primal pairs ψ_1 and ψ_2 are equivalent, denoted by $\psi_1 \sim \psi_2$ if they have the same support I and $\mathbf{s}_1(\theta) = \mathbf{s}_2(\theta)$ for all $\theta \in I$. The quotient set, the set of all equivalence classes $\Psi = \Psi_p / \sim$ is called the regression space. Obviously, each element of Ψ is an equivalent class, which can be expressed by its unique irreducible element. We refer to each element of ψ as an *expression*.

We are going to show that similar to the vector representation, (2.17) is

an affine function of expressions. Accordingly, we need to give a definition of the addition operator on Ψ. take two pairs $\psi_1, \psi_2 \in \Psi_p$. Define $(\boldsymbol{\theta}, \mathbf{s}) = \psi_1 + \psi_2$, where $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \cup \boldsymbol{\theta}_2$ and

$$\mathbf{s}(\theta) = \begin{cases} \mathbf{s}_1(\theta) & \theta \in \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \\ \mathbf{s}_2(\theta) & \theta \in \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \\ \mathbf{s}_1(\theta) + \mathbf{s}_2(\theta) & \theta \in \boldsymbol{\theta}_1 \cap \boldsymbol{\theta}_2 \end{cases} \tag{4.2}$$

It is simple to see that the addition of primal pairs is invariant under equivalence, which means that + is well defined on the regression space. It is also simple to define the scalar product as $\alpha(\boldsymbol{\theta}, \mathbf{s}) = (\boldsymbol{\theta}, \alpha\mathbf{s})$, which is also invariant under $\sim$. We are not going to mention invariance later and it will always be assumed implicitly. However, the reader may check this property whenever necessary. The space Ψ with the addition and scalar product operators form a vector space. The set $B = \{(\{\theta\}, \{(\theta, 1)\}) | \theta \in S\}$ forms a basis set for such a space. Note that each element in Ψ can be written as a linear combination of a *finite* number of elements in B. Furthermore, a linear function $F : \Psi \to \mathbb{C}$ can be written as

$$F(\boldsymbol{\theta}, \mathbf{s}) = \sum_{\theta \in \boldsymbol{\theta}} f(\theta)\mathbf{s}(\theta) \tag{4.3}$$

where $f : \mathbb{R} \to \mathbb{C}$ is an arbitrary function. This forms the dual space of Ψ.

It only remains to define a topology on Ψ, which is rather a technical task. Up to now, we have defined a vector space, in which the CLASS optimization will be defined as a convex optimization, for which we are to devise an algorithm "converging" to the optimal point. The convergence definition is obviously dependent upon the definition of being "relatively close or far", which is given by the *topology*. On the other hand, we have already introduced a group of "standard" procedures (e.g. gradient descent, cyclic coordinate descent, Newton's method, etc) for any such vector space with a topology. However, the convergence of these methods are conditioned by some further properties of the topology.

While it is straightforward to define a metrizable topology on Ψ_p, i.e. a one based on the definition of a distance between any two point regarding the Hausdorff distance on the hyperspace of finite subsets of S [28], it is difficult to define one that is invariant under $\sim$ with the desired properties of convergence. As an example, define the $p-$norm on Ψ as

$$\|(\boldsymbol{\theta}, \mathbf{s})\|_p = \left(\sum_{\theta \in \boldsymbol{\theta}} |\mathbf{s}(\theta)|^p \right)^{\frac{1}{p}} \tag{4.4}$$

It is now simple to see that for an arbitrary linear function $F : \Psi \to \mathbb{C}$ we have

$$|F(\psi_1) - F(\psi_2)| \leq \left(\sum_{\theta \in \boldsymbol{\theta}} |f(\theta)|^q \right)^{\frac{1}{q}} \|\psi_1 - \psi_2\|_p \tag{4.5}$$

by Hölder's inequality, which can not bound the increment for very close points even with very regular functions f, since the dimension of $\boldsymbol{\theta}$ is unbounded. Thus, almost all linear functions are not continuous with respect to the $p-$norm. However, if we restrict Ψ to any bounded cardinality, the usual behavior of the $p-$metric is retrieved. The only exception is when $p = 1$, in which case the linear function is continuous if an only if f is bounded on any compact set. As can be seen, the $p-$metric does not take the DOA separation into account. As a result, the practical methods based on this topology end up with an explosion of the number of regressors with no real convergence, since these methods do not have any mechanism of updating DOAs other than introducing a new one.

Accordingly, we address a so called hit and miss topology which combines the topological properties of the DOA and parameter vectors. We should note that in general a topology is not necessarily expressed by a distance function. In fact a topology is well defined by its *open* sets. Then, a neighborhood of a point is an open set containing the point. Roughly speaking, two points are close if each one is included in most open neighborhoods of the other. A sequence of points also converge to a desired point if each neighborhood of the desired point contains all the sequence except an initial part.

Take open sets $C \subseteq S$ and $T \subseteq \mathbb{C}$ and a closed set $D \subseteq S$. Define

$$\begin{aligned} H(C,T) &= \{\psi = (\boldsymbol{\theta}, \mathbf{s}) | \sum_{\theta \in \boldsymbol{\theta} \cap C} \mathbf{s}(\theta) \in T\} \\ M(D,T) &= \{\psi = (\boldsymbol{\theta}, \mathbf{s}) | \sum_{\theta \in \boldsymbol{\theta} \cap D} |\mathbf{s}(\theta)| \in T\}. \end{aligned} \tag{4.6}$$

Then, we define the P topology as the topology generated by every $H(C,T)$ and $M(D,T)$. Such hit and miss topologies have already been used in the tracking context by Random Finite Sets (RFS), e.g. the PHD filter [29]. It is simple to see which type of distance the P topology implies. Take a point $\psi = (\boldsymbol{\theta}, \mathbf{s})$ and a very "small" neighborhood constructed by $\bigcup_i H(C_i, T_i) \cap M(D,T)$, where C_i is a very small open subset of S containing $\theta_i \in \boldsymbol{\theta}$, $D = \left(\bigcup_i C_i \right)^c$, and T is a very small neighborhood of 0. The sets T_i are also small neighborhoods of $\mathbf{s}(\theta_i)$ respectively. Now, it is simple to see that a close regression pair $\psi' = (\boldsymbol{\theta}', \mathbf{s}')$ is in this neighborhood if the sum of the sources in the *cluster* $C_i \cap \boldsymbol{\theta}'$ is close to $\mathbf{s}(\theta_i)$, and the elements outside

all clusters are very weak. Roughly speaking, ψ' is very close to ψ if it could approximately be broken into multiple clusters, each of which could be *merged* to the regression pair $(\theta_i, \mathbf{s}(\theta_i))$. It is a reasonable definition. Suppose a radar receiver gives a set of estimated DOAs and corresponding messages. It is logical to assume that this estimate is a result of neglecting weak sources, due to the noise and merging the close sources, due to the finite resolution. Thus, the $P-$topology is a suitable choice for our purpose.

It is also simple to check that a linear function is continuous with respect to P if and only if $f(\theta)$ is continuous. As the final and the most important word, if S is compact, the P space admits the Heine-Borel property, that is every closed and bounded subset of Ψ under P topology is compact, i.e. every covering of it has a finite sub-cover. This establishes many important properties such as Bolzano-Weierstrass's theorem, which enables us to use the algorithms, such as gradient descent in the CLASS optimization.

Once a topology is established on Ψ, it is also straightforward to define the Borel $\sigma-$algebra. The further definitions such as integrable functions and probability spaces follow from the definition of the σ-algebra. Although these tools are necessary in order to design new estimation methods similar to the RFS theory and PhD filtering, we stop constructing at this point.

4.2 ℓ_1 Constraint on the Regression Space

As we stated in Chapter 2, the likelihood is convex on the space of generalized sparse source vectors. However, we need to confine the search to a suitable region to avoid undesired solutions. To this end, we introduced the ℓ_1 constraint in Chapter 3, which confines the search to a convex set while maximizing the chance of getting a low dimensional optimal solution. Let us define $L_1(c) = \left\{(\psi_1, \psi_2, \dots, \psi_T) \in \Psi^T | \sum_{\theta \in \boldsymbol{\theta}} \sqrt{\sum_{t=1}^{T} |\mathbf{s}(\theta, t)|^2} \leq c \right\}$, where c is a positive real and $\boldsymbol{\theta} = \bigcup_t \boldsymbol{\theta}_t$. Note that for $T = 1$, the set $L_1(c) = \{(\psi \in \Psi | \|\psi\|_1 \leq c\}$ is the level set of the 1-norm. Now, we can define the constrained optimization

$$\begin{aligned} \min & \sum_{t=1}^{T} \|\mathbf{x}(t) - \sum_{\theta \in \boldsymbol{\theta}} \mathbf{a}(\theta)\mathbf{s}(\theta, t)\|_2^2 \\ & (\psi_1, \dots, \psi_T) \in L_1(c) \end{aligned} \tag{4.7}$$

as the generalization of LASSO for multiple snapshots. It is possible to show that $L_1(c)$ is closed on Ψ^T with respect to the $p-$norm topology but not compact, while it is closed and compact under P. However, the CLASS optimization is not continuous under P. The aim of Paper 3 is to investigate the

performance of the well defined optimization of (4.7) as a DOA estimator. The optimization (4.7) is traditionally referred to as the Complex Group Least Absolute Shrinkage and Selection Operation (CG-LASSO). It is usual to write down the LASSO optimization as an unconstrained one with a Lagrange dual variable λ. Define $\|\psi_1, \psi_2, \ldots, \psi_T\|_{1,2} = \sum_{\theta \in \boldsymbol{\theta}} \sqrt{\sum_{t=1}^{T} |\mathbf{s}(\theta, t)|^2}$ Then, the Lagrange dual problem can be written as

$$\min \frac{1}{2} \sum_{t=1}^{T} \|\mathbf{x}(t) - \sum_{\theta \in \boldsymbol{\theta}} \mathbf{a}(\theta)\mathbf{s}(\theta, t)\|_2^2 + \lambda \|\psi_1, \psi_2, \ldots, \psi_T\|_{1,2}, \tag{4.8}$$

where $\lambda > 0$. By the saddle point optimality conditions a global solution $\psi_1, \psi_2, \ldots, \psi_T$ of (4.8) is also an optimum point for (4.7) with $c = \|\psi_1, \psi_2, \ldots, \psi_T\|_{1,2}$. We continue to generalize the LASSO idea to the regression space in Section 4.1 in the two proceeding sections. However, these parts could also be neglected due to the higher level of mathematics involved.

4.3 LASSO Implementation Difficulties on Regression Space

We conclude this section by addressing the implementation issues of the LASSO algorithm on Ψ. The main point here is to illustrate the importance of compactness versus Zangwill's theorem properties. It is well known that the essentially important issues for the optimization techniques are the compactness of the feasible region and the continuity of the function. However, in finite-dimensional spaces the compactness is strongly related to the closedness property. Note that the Zangewill's theorem can be concluded on Ψ only if Bolzano-Weierstrass's theorem holds true and the region is closed. Thus, the superior properties of the convex programing is established only in such a case. Let us try the steepest descent algorithm on the $1-$norm topology for $T = 1$, in which case L_1 is not compact anymore. Take a point $\psi \in \Psi$ and take the cost in (4.8), F as a continuous convex function on Ψ. Note that

$$\begin{aligned} F'(\psi, \psi_0) &= \lim_{\epsilon \to 0} \frac{F(\psi + \epsilon\psi_0) - F(\psi)}{\epsilon} \\ &= \sum_{\theta \in \boldsymbol{\theta}_0 \cup \boldsymbol{\theta}} \Re(\mathbf{s}_0^*(\theta)\mathbf{a}^H(\theta)\mathbf{n}) + \lambda \sum_{\theta \in \boldsymbol{\theta}} \Re\left(\frac{\mathbf{s}(\theta)^* \mathbf{s}_0(\theta)}{|\mathbf{s}(\theta)|}\right) \\ &+ \lambda \sum_{\theta \in \boldsymbol{\theta}_0 - \boldsymbol{\theta}} |\mathbf{s}_0(\theta)|, \end{aligned} \tag{4.9}$$

where $\mathbf{n} = \mathbf{x} - \sum_{\theta \in \boldsymbol{\theta}} \mathbf{a}(\theta)\mathbf{s}(\theta)$. The derivative in the direction ψ_0 is defined at any point ψ. Now, the steepest direction in the $1-$norm topology is given by the minimum point of $F'(\psi, \psi_0)$, where $\|\psi\|_1 \leq 0$. It is not hard to see that due to the polyhedral nature of this problem, the minimum occurs at an extreme point $\psi_0 = (\{\theta_c\}, \{(\theta_c, e^{j\phi})\})$, where θ_c and ϕ are arbitrary numbers. For such extreme point, the minimum directional derivative over all values of ϕ and fixing θ_c is given by

$$g(\theta_c) = \begin{cases} \lambda - |\mathbf{a}^H(\theta_c)\mathbf{n}| & \theta_c \notin \boldsymbol{\theta} \\ -|\lambda \frac{\mathbf{s}(\theta_c)}{|\mathbf{s}(\theta_c)|} - \mathbf{a}^H(\theta_c)\mathbf{n}| & \theta_c \in \boldsymbol{\theta} \end{cases} \tag{4.10}$$

Thus, the steepest descent direction is recognized by the minimum of $g(\theta)$. Once this value is defined at a certain iteration, the point ψ can be updated to $\psi + \epsilon\psi_0$, where ϵ is the step size and ψ_0 corresponds to the minimum point of $g(\theta)$. It is interesting to see that the well known *matching pursuit* is derived from the above algorithm letting $\lambda = 0$. Note also that the fixed point of this algorithm is given by a point for which $g \geq 0$ everywhere. This results in the following optimality conditions

$$\begin{aligned} &|\mathbf{a}^H(\theta_c)\mathbf{n}| \leq \lambda && \theta_c \notin \boldsymbol{\theta} \\ &\lambda \tfrac{\mathbf{s}(\theta_c)}{|\mathbf{s}(\theta_c)|} = \mathbf{a}^H(\theta_c)\mathbf{n} && \theta_c \in \boldsymbol{\theta}, \end{aligned} \tag{4.11}$$

which could be shown to hold true only for the global minimum point (Paper 3). However, the above algorithm will not converge to the fixed point by the steepest descent, since the generated sequence of the algorithmic map need not necessarily contain an accumulation point due to non-compactness of L_1. In fact, the algorithm almost never converges to any point, and the dimensionality of the result always increases. One may simply think about other ways to combine close DOAs to control the number of regressors.

The alternative approach is to use the P topology instead. Although we did not introduce any explicit metric on P, it is still possible to define a group of $\epsilon-$neighborhoods $\mathcal{N}_\epsilon(\psi, \theta_1, \theta_2, \ldots, \theta_k)$ of any point $\psi = (\boldsymbol{\theta}_0, \mathbf{s}_0) \in \Psi$ as follows

$$\mathcal{N}_\epsilon(\psi, \theta_1, \theta_2, \ldots, \theta_k) = \left\{ (\boldsymbol{\theta}, \mathbf{s}) \middle| \forall i, \left| \sum_{\theta \in \boldsymbol{\theta}_0 \cap \mathcal{N}_\epsilon(\theta_i)} \mathbf{s}_0(\theta) - \sum_{\theta \in \boldsymbol{\theta} \cap \mathcal{N}_\epsilon(\theta_i)} \mathbf{s}(\theta) \right| < \epsilon, \sum_{\theta \in \boldsymbol{\theta} - \bigcup \mathcal{N}_\epsilon(\theta_i)} |\mathbf{s}(\theta)| < \epsilon \right\} \tag{4.12}$$

where $\mathcal{N}_\epsilon(\theta_i)$ denotes the neighborhood of an arbitrary DOA $\theta_i \in \Theta$, such that $\sum_{\theta \in \boldsymbol{\theta}_0 - \bigcup \mathcal{N}_\epsilon(\theta_i)} |\mathbf{s}(\theta)| < \epsilon$. Let us analyze the behavior of F in a close

neighborhood of ψ_0. We observe for $\psi \in \mathcal{N}_\epsilon(\psi_0)$,

$$F(\psi) = F(\psi_0) + \lambda(\|\psi\|_1 - \|\psi_0\|_1) + \sum_i \Re(\delta_i^* \mathbf{a}^H(\theta_i)\mathbf{n}) + \sum_i \Re(\eta_i^* \mathbf{d}^H(\theta_i)\mathbf{n}) + \sum_{\theta \in \boldsymbol{\theta} \setminus \bigcup \mathcal{N}_\epsilon(\theta_i)} \Re(\mathbf{s}(\theta)^* \mathbf{a}^H(\theta)\mathbf{n}) \tag{4.13}$$

where $\delta_i = \sum_{\theta \in \boldsymbol{\theta} \cap \mathcal{N}_\epsilon(\theta_i)} \mathbf{s}(\theta) - \sum_{\theta \in \boldsymbol{\theta}_0 \cap \mathcal{N}_\epsilon(\theta_i)} \mathbf{s}_0(\theta)$, $\eta_i = \sum_{\theta \in \boldsymbol{\theta} \cap \mathcal{N}_\epsilon(\theta_i)} \mathbf{s}(\theta)(\theta - \theta_i)$, and $\mathbf{n}$ is as given above. First, note that F is not a continuous function anymore, since the $1-$norm is non-continuous. However, the approximation in (4.13) could be minimized to provide a trust region optimization method. This is a polyhedral optimization and it could be shown that it does not include any descent direction such that the optimization always gives a solution if $\epsilon < \frac{\lambda}{\max_\theta \|\mathbf{d}^H(\theta)\mathbf{n}\|}$. It is also simple to show that the fixed point for such an algorithm is given by the optimal points in (4.11), which characterizes a globally optimal point. It is interesting to note that the polyhedral optimization solution occurs at an extreme point, which means that at each step either a new DOA is introduced by updating $\theta \in \boldsymbol{\theta} \setminus \bigcup \mathcal{N}_\epsilon(\theta_i)$ or a group $\boldsymbol{\theta} \cap \mathcal{N}_\epsilon(\theta_i)$ is shrinked to a DOA at $\theta_i \pm \epsilon$ with the corresponding waveform value $s = \sum_{\theta \in \boldsymbol{\theta}_0 \cap \mathcal{N}_\epsilon(\theta_i)} \mathbf{s}_0(\theta) \pm \epsilon$. As can be seen, the second proposed method in this section might be successful in general ignoring the complexity of clustering. However, due to the complexity of this method we propose a different homotopic technique based on the interesting properties of the introduced problem in Paper 3.

Chapter 5

Conclusions and Future Work

This work was devoted to investigating the sparse regression by ℓ_1 penalties over a continuous manifold. We chose the DOA model as a target of our analysis, while it could be extended to any proper manifold. We discussed first the traditional techniques of solving such a problem by the ML estimator and the model order selection methods. We later discussed the computational shortcoming of the traditional methods in a general setup. Then, we introduced the spatial sparsity framework, which we attempted to well define as an estimation on a relaxed regression space, leading to a convex optimization. We showed that the over-complete nature of the latter necessitates some regulatory conditions. We introduced the ℓ_1 penalties and regions as such a regularization, resulting in the LASSO technique when combined with the original regression problem. We reviewed the solution of LASSO by discretizing the parameter space and addressed the problem of RP selection. We further provided some hints to perform the ℓ_1 penalized ML over the generalized regression space.

Based on the presented introduction the current research may be continued in several ways. A well know method of implementing LASSO is to use homotopy, i.e. to follow the solution by changing RP. The homotopic solution of LASSO on the continuous regression space have already been implemented with some minor deficiencies at the saddle-points. There might also exist better optimization techniques on the regression space as a further research topic. However, to the author's opinion, the Bayesian interpretation of LASSO leaves plenty of room to investigate. The theory of Bayesian priors on the continuous LASSO space is not discussed and there has been almost no attempts to characterize the LASSO posteriors. Furthermore, the Bayesian properties of the so called *adaptive LASSO* [30], as a result of assigning individual RPs to different sparse elements are not well known. This will hopefully be unified in a "tracking theory of regression", for which we have taken some initial steps in [31]. As the final note,

the theoretical importance of this work should be addressed. Note that the probabilistic arguments leading to the RIP condition are inadequate for the analysis. Thus, a wider set of questions about perfect recovery and compression, such as the ones about short length codes might be answered following the current line of reasoning.

Chapter 6

Summary of Papers

Part 2 contains three selected papers reflecting the main contributions of this research.

The first paper is our first attempt to understand the Bayesian behavior of LASSO. It is well known that the ML estimator of the RP with the model in (3.7) has naive properties. We tried to correct this model by restricting the Laplacian model only over the true sources and estimating the RP. The result show a remarkable improvement in the range of consistency of the model order. However, the investigation to find a genuine and general model is still in progress.

The second work proposes a fast numerical solution to the complex LASSO. A very similar idea in the real regression case is the LARS algorithm [32], based on the former observation that the LASSO path consists of linear pieces and one may jump from one edge to the other by predicting the next edge point. Although the linear pieces do not exist in the complex case, the edges do. Thus, we generalized the stagewise idea in LARS to the complex case and introduced the so called SPS-LASSO technique.

The third contribution is devoted to the theoretical analysis of the LASSO based estimation by giving the estimation error variance and consistency conditions. This paper builds the first steps to understand the grid-less theory of LASSO, which we referred to as the CLASS algorithm. Based on these findings, we show that the CLASS algorithm, as the best instance of the ℓ_1 penalized estimation has a fundamental resolution to be consistent and compared to the ML estimator its error is increased by introducing a bias term.

References

[1] R. W. S. Alan V. Oppenheim, *Digital Signal Processing*. Prentice-Hall, 1975.

[2] D. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289 –1306, april 2006.

[3] E. J. Candes and Y. Plan, "A Probabilistic and RIPless Theory of Compressed Sensing," *Information Theory, IEEE Transactions on*, vol. 57, no. 11, pp. 7235–7254, Nov. 2011.

[4] M. Mishali, Y. Eldar, and A. Elron, "Xampling: Signal acquisition and processing in union of subspaces," *Signal Processing, IEEE Transactions on*, vol. 59, no. 10, pp. 4719 –4734, oct. 2011.

[5] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *Submitted*, Submitted.

[6] D. Malioutov, M. Cetin, and A. Willsky, "Source localization by enforcing sparsity through a laplacian prior: an svd-based approach," *IEEE Workshop on Statistical Signal Processing*, pp. 573 – 576, sept.-1 oct. 2003.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B, (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[8] E. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. Springer, 1998.

[9] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms; 3rd ed.* Newark, NJ: Wiley, 2006.

[10] J. D. Jackson, *Classical Electrodynamics*, 3rd ed. Wiley, 1999.

[11] H. L. V. Trees, *Detection, Estimation, and Modulation Theory, Part III: Radar-Sonar Signal Processing and Gaussian Signals in Noise*. John Wiley and Sons, 1968.

[12] H. V. Poor and G. W. Wornell, *Wireless Communications: Signal Processing Perspectives*. Prentice Hall, 1998.

[13] E. A. Robinson and S. Treitel, *Geophysical Signal Analysis*. Prentice Hall, 1980.

[14] 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, 7 1990.

[15] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *Signal Processing Magazine, IEEE*, vol. 13, no. 4, pp. 67–94, jul 1996.

[16] P. Stoica, Y. Selen, 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, oct. 2004.

[17] 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, 7 2004.

[18] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *Proc. 2nd Int. Symp. of Inf. Theory*, pp. 267–281, 1973.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley and Sons, 2006.

[20] C. S. Wallace, *Statistical And Inductive Inference By Minimum Message Length*. Springer, 2005.

[21] J. Rissanen, "Information and complexity in statistical modeling," *Information Theory Workshop, ITW Punta del Este. IEEE*, pp. 351 –351, march 2006, 10.1109/ITW.2006.1633845.

[22] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 387 – 392, apr 1985.

[23] J. Fessler and A. Hero, "Space-alternating generalized expectation-maximization algorithm," *Signal Processing, IEEE Transactions on*, vol. 42, no. 10, pp. 2664 –2677, oct 1994.

[24] 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, feb 1996.

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

[26] T. Park and G. Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, pp. 681–686, 2008.

[27] M. Figueiredo, "Adaptive sparseness for supervised learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1150 – 1159, sept. 2003.

[28] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Springer, 2011.

[29] R.Mahler, "An introduction to multisource-multitarget statistics and applications," Lockheed Martin, Monograph, March 2000.

[30] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[31] 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*, dec. 2011, pp. 397 –400.

[32] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.