Analysis, Algorithms and Applications of Compressed Sensing

YANG ZAI

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirements for the degree of
Doctor of Philosophy

2013

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

. . . .

Date   YANG ZAI

Acknowledgements

There were many people who helped me during my Ph.D, and I would like to take this opportunity to thank them.

First and foremost, I would like to express my appreciation and gratitude to my supervisors, Professor XIE Lihua and Professor ZHANG Cishen (currently at Swinburne University of Technology, Australia), for years of invaluable guidance, encouragement, and financial support. Professor Zhang was the first person who introduced me to the scientific world. Thanks to Professor Xie for his endless patience, thoughtful discussions, and the many opportunities and the freedom he has offered that allowed me to find the research topics I was most interested in. This thesis would not have been possible without their enthusiastic help.

Thanks also to Dr. LU Wenmiao, Dr. Arian Maleki and Professor Richard G. Baraniuk for their collaborations and fruitful discussions.

I would also like to thank several fellow graduate students, postdocs and visiting scholars for their help and support in both research and life. In particular, thanks to YOU Keyou, LIU Shuai, XIAO Nan, GAO Tingting, HU Jinwen, MENG Wei, YU Chengpu, DENG Jun, GUO Yuqian, Fang Jianyin, ZHOU Siwang, ZHAO Dongya, QU Xiaomei and LI Wuquan.

In the end, I would like to thank my parents and my wife, ZHANG Xia, for their endless love, support and self-sacrifices. They have always believed in me, even more than I myself do. Special thanks to Xia for her encouragement whenever I was frustrated. This thesis would not have been possible without her support.

Abstract

Compressed sensing (CS) is an emerging research area which studies the problem of recovering a high dimensional sparse signal from its low dimensional linear samples. While CS can acquire a signal with a sub-Nyquist sampling rate, it requires a nonlinear signal reconstruction strategy which is computationally expensive compared to the simple linear reconstruction procedure in the traditional Nyquist sampling case. Given the exact information of the sampling/sensing system, existing results have shown that the signal of interest can be accurately recovered via convex relaxation (ℓ_1 minimization) and/or other approaches provided the signal is sufficiently sparse.

To understand the sparsity-undersampling tradeoff is essential in CS, which is the first task of this thesis. Phase transition is currently the most precise tool to describe the tradeoff in the sense that it provides a condition that is both necessary and sufficient. While existing results are focused on the real valued signal and system case, we extend it to the complex case, discovering a new phase transition curve of ℓ_1 minimization which is positioned well above the real one. On the other hand, the sampling system may not be exactly known *a priori*, the signal recovery performance in the presence of system perturbations is also studied in this thesis. In particular, a structured sensing matrix perturbation is studied which has practical relevance. Under mild sufficient conditions, it is shown that a sparse signal can be recovered by ℓ_1 minimization and the recovery error is at most proportional to the measurement noise level, which is similar to the standard CS result. In the special noise free case, the recovery is exact provided that the signal is sufficiently sparse with respect to the perturbation level.

Secondly, computationally efficient algorithms for CS are developed in the thesis. ℓ_1 minimization has favorable guarantees of signal recovery accuracy but most solvers are slower than expected for practical problems of high dimension. In the thesis, two first-order algorithms are proposed for the ℓ_1 minimization problem in CS: one exactly solves the problem and the other is a relaxed version of the first one. The latter can be considered as a modified iterative soft thresholding algorithm and is easy to implement. Sparse Bayesian learning (SBL) is another popular approach to the sparse signal recovery. In SBL, the signal sparsity information is exploited by assuming a sparsity-inducing prior for the signal that is then estimated using Bayesian inference. In the thesis, a new sparsity-inducing prior is introduced and efficient algorithms are developed for the signal recovery. The main algorithm is shown to produce a sparser solution than existing SBL methods while preserving their desirable properties. Moreover, a Bayesian framework is introduced in the thesis for the CS problem with quantized measurements. All proposed algorithms are compared with state-of-the-art algorithms through numerical simulations.

At last, applications of CS to direction of arrival (DOA) estimation and motion correction in magnetic resonance imaging (MRI) are presented by exploiting the spatial sparsity of the source signals and the image sparsity in an appropriate domain, respectively. The proposed approach to the DOA estimation resolves the problem that the true DOAs may not lie on the discretized sampling grid which is typically required in existing methods. The new method can maintain high estimation accuracy even under a very coarse sampling grid. While existing applications of CS to MRI are focused on accelerating the scanning process by undersampling k-space data, a sparsity-driven approach to the motion correction in MRI is presented for the first time in the thesis.

Symbols and Acronyms

Symbols

$\mathbb{C}$	set of complex numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}^+$	set of positive real numbers
$\mathbb{R}^N$	set of real N dimensional vectors
$\mathbb{R}^{M\times N}$	set of real $M \times N$ dimensional matrices
$\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{\Phi}$, $\cdots$	matrices
$\boldsymbol{x}$, $\boldsymbol{y}$, $\cdots$	vectors
x, y, $\cdots$	scalars
$\boldsymbol{A}_i$	ith column of $\boldsymbol{A}$
A_{ij}	ijth entry of $\boldsymbol{A}$
$\boldsymbol{A}^T$	transpose of $\boldsymbol{A}$
$\boldsymbol{A}^H$	complex transpose of $\boldsymbol{A}$
$\boldsymbol{A}'$	transpose for real-valued $\boldsymbol{A}$ and complex transpose for complex-valued $\boldsymbol{A}$
x_i	ith entry of $\boldsymbol{x}$
$\left\|\cdot\right\|_2$	spectral norm for a matrix, or Euclidean norm for a vector
$\left\|\cdot\right\|$	determinant for a matrix, or absolute value for a scalar, or cardinality for a set
$\operatorname{diag}(\cdot)$	a diagonal matrix if applied to a vector with its diagonal entries formed by the vector, or if applied to a matrix, a vector composed of diagonal entries of the matrix

$\mathrm{supp}\{\boldsymbol{x}\}$	support of $\boldsymbol{x}$ defined as $\{i : x_i \neq 0\}$
$\|\boldsymbol{x}\|_p$	ℓ_p norm ($p \geq 1$) or pseudo-norm ($0 < p < 1$) of $\boldsymbol{x}$ defined as $\left(\sum_i \|x_i\|^p\right)^{\frac{1}{p}}$
$\|\boldsymbol{x}\|_0$	ℓ_0 pseudo-norm of $\boldsymbol{x}$ defined as the number of nonzero entries
$\|\boldsymbol{x}\|_\infty$	ℓ_∞ norm of $\boldsymbol{x}$ defined as $\max\{\|x_i\|\}$
$\langle \boldsymbol{x},\ \boldsymbol{y} \rangle$	inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$
$\boldsymbol{x} \succeq \boldsymbol{y}$	$x_i \geq y_i$ for all i
$E\{\cdot\}$	expectation
$Var\{\cdot\}$	variance

Acronyms

CS	compressed sensing
SSR	sparse signal representation
DOA	direction of arrival
MRI	magnetic resonance imaging
BP	basis pursuit
BPDN	basis pursuit denoising
OMP	orthogonal matching pursuit
IHT	iterative hard thresholding
IST	iterative soft thresholding
SBL	sparse Bayesian learning
OGSBL	off-grid SBL
RAAR	relaxed averaged alternating reflections
SRAAR	sparse RAAR
MAP	maximum *a posteriori*
AWGN	additive white Gaussian noise
i.i.d.	independently and identically distributed

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Motivation and Objective

Compressed sensing (CS) [1, 2], also known as compressive sensing or compressive sampling, is an emerging technique which has attracted a lot of researchers due to its potentially wide applications. In conventional wisdom, a signal of interest needs to be sampled at the Nyquist rate for its complete recovery. Since most natural signals of interest are sparse/compressible under an appropriate basis, the sampled signal is then represented in the associated basis and only a few significant coefficients are retained for signal reconstruction with little energy loss, which is known as the compression technique (consider the JPEG 2000 standard for image compression). While the separate "sampling" and "compression" steps can prove to be wasteful of sensing and sampling resources, CS attempts to do them at a single step as its name suggests and accesses the success of signal recovery with the sparsity/compressibility. In particular, CS acquires a much smaller number of compressive linear measurements for reconstructing a high dimensional signal. The CS theory guarantees exact or accurate signal reconstruction given that the signal of interest is sufficiently sparse.

While progresses have been made in the past few years, many problems in CS are still open or need further studies. For example, current results in CS are mainly focused on the ideal situation where the sensing system is exactly known. This may not be the case in practice. The theoretical performance of CS in the non-ideal case needs to be studied. Since CS can acquire a signal with a sub-Nyquist sampling rate, it requires a nonlinear signal reconstruction strategy which is typically computationally expensive. Hence, computationally efficient and accurate algorithms are needed for sparse signal recovery in CS. The inherent reason of the success of CS is the sparsity while sparse signals are ubiquitous in practice. That means, CS has potential applications in many research areas. In summary, objectives of the thesis include:

1. To analyze the performance of CS with/without system uncertainties;
2. To propose accurate and efficient algorithms for sparse signal recovery in CS;
3. To apply CS to solve practical problems.

1.2 Main Contributions

The main contributions of the thesis are listed as follows:

1. We discover a new phase transition curve, which provides precise sparsity-undersampling tradeoff of ℓ_1 minimization, for the complex valued CS problem where both the signal of interest and the sensing matrix are complex valued.
2. We analyze the CS problem subject to a structured perturbation in the sensing matrix which has practical relevance. Under mild conditions, we show that a sparse signal can be recovered by solving an ℓ_1 minimization problem and the recovery error is at most proportional to the measurement noise level, which is similar to the standard CS result. In the special noise free case, the recovery is exact provided that the signal is sufficiently sparse with respect to the perturbation level.

3. We study two popular approaches to sparse signal recovery in CS including convex relaxation and sparse Bayesian learning (SBL). For the convex relaxation method, a first-order algorithm is proposed which can be considered as a modified iterative soft thresholding algorithm and is easy to implement. For SBL, we introduce a new sparsity-inducing prior and develop efficient algorithms for signal recovery. We show that the main algorithm produces a sparser solution than existing SBL methods while preserving their desirable properties. Their performances are demonstrated for 1D synthetic data and 2D practical images.

4. We study sparse signal recovery from quantized compressive measurements. A novel variational Bayesian inference based algorithm is presented, which unifies the multi- and 1-bit CS processing and is applicable to various cases of noiseless/noisy environment and unsaturated/saturated quantizer. The quantization error is decoupled from the measurement noise, modeled as a random variable and then estimated jointly with the signal being recovered. The novel characterization of the quantization error results in improved performance of the algorithm.

5. We present an accurate and efficient algorithm for applying CS to direction of arrival (DOA) estimation. The new method resolves the problem that the true DOAs are not exactly on the discretized sampling grid and can maintain high estimation accuracy under a very coarse sampling grid.

6. We propose a novel application of CS to motion correction in magnetic resonance imaging (MRI). An iterative algorithm is proposed to jointly reconstruct the image content and estimate the motion which solves the motion correction problem without acquiring additional measurement data or modifying the sampling sequence.

Moreover, for the purpose of reproducible research, the Matlab codes of the algorithms developed in this thesis have been made downloadable online at the author's personal website, see https://sites.google.com/site/zaiyang0248.

1.3 Outline of the Thesis

In Chapter 2 we review some of the existing results on theories, algorithms and applications of CS. The theoretical results include theoretical guarantees of existing sparse signal recovery algorithms, which are deterministic in the sense that they can be applied as long as the sensing matrix satisfies some condition, and the phase transition theory that is specialized for random matrices. The algorithms mentioned in this chapter include those with theoretical guarantees, e.g., convex relaxation and greedy pursuits, and SBL. Applications are focused on the DOA estimation and MRI that are closely related to the thesis.

The following chapters that present our main contributions are divided into two parts: Part I (Chapters 3–7) is on analysis and algorithms of CS and Part II (Chapters 8 and 9) is on applications.

In Chapter 3 we propose first-order algorithms for convex relaxation (ℓ_1 minimization) approach to CS. We introduce a method called orthonormal expansion to reformulate the basis pursuit problem and present two algorithms (named as ONE-L1) based on convex optimization: one exactly solves the problem and the other is a relaxed version of the first one. The relationship between the relaxed version and iterative soft thresholding algorithm is studied. This chapter is mainly based on the paper that has been published in *IEEE Transactions on Signal Processing* [3].

In Chapter 4 we study the phase transition of ℓ_1 minimization in the complex setting where both the signal and sensing system are complex valued while existing phase transition theory is focused on the popular real setting. By extending the algorithms presented in Chapter 3 to the complex case and extensive numerical experiments we discover a new phase transition curve for the complex valued CS problem which complements the existing phase transition results. This chapter is mainly based on the paper that has been published in *IEEE Signal Processing Letters* [4]. Some of the results mentioned in this chapter were published in [5,6].

In Chapter 5 we study the CS problem subject to a structured matrix perturbation. We propose a nonconvex ℓ_1 minimization problem to recover sparse signals in which the parameters specifying the perturbation are jointly recovered. We provide theoretical guarantees of the proposed problem for recovery of sparse and compressible signals in both noisy and noiseless environments. An alternating direction algorithm is presented to solve the ℓ_1 minimization problem. This chapter is mainly based on the paper that has been published in *IEEE Transactions on Signal Processing* [7]. Some of the results mentioned in this chapter were published in [8].

In Chapter 6 we present a new SBL approach to sparse signal recovery in CS, where we introduce a new sparsity-inducing prior and develop efficient algorithms for signal recovery. We analyze the main algorithm and reveal that it produces a sparser solution than its existing SBL peers while preserving their desirable properties.

In Chapter 7 SBL is used to study the quantized CS problem where the new challenge is the existence of the quantization error. We present a Bayesian framework which unifies the multi- and 1-bit CS problems and is able to deal with various cases of noiseless/noisy environment and unsaturated/saturated quantizer. After that we propose a novel algorithm for signal recovery within the framework based on variational Bayesian inference. This chapter is mainly based on the paper that has been published in *IEEE Transactions on Signal Processing* [9]. Some of the results mentioned in this chapter were published in [10].

In Chapter 8 we study the DOA estimation problem from a sparse representation perspective and are concerned about the practical situation where the true DOAs are not on the discretized sampling grid. An off-grid model is obtained using linear approximation and then two approaches to off-grid DOA estimation are proposed based on the model. In the first approach, we show that the off-grid DOA estimation problem can be formulated as the perturbed CS problem studied in Chapter 5 and thus the ℓ_1 optimization approach presented in Chapter 5 can be applied straightforward. In the other, SBL is adopted to do the estimation in an iterative manner to reduce the computational workload. This chapter is mainly based on the papers that have been published in *IEEE Transactions on Signal Processing* [7, 11].

In Chapter 9 we solve the motion correction problem in MRI by exploiting the MR image sparsity/compressibility. We formulate the translational motions as unknown parameters in the sensing system and seek for a motion estimate such that the motion-compensated image has the maximal sparsity/compressibility. This chapter is inspired by Chapter 5. Some of the results mentioned in this chapter have been published in [12].

In Chapter 10 we conclude the thesis and highlight some potential future research directions.

Chapter 2

Literature Review

The recently developed compressed sensing (CS) theory and methods [1, 2] can achieve acquisition of information contained within a huge volume of data using only a small number of measurement samples. Different from the classical Shannon-Nyquist sampling theorem which requires that the sampling frequency be twice as high as the bandwidth of a signal in order to reconstruct its complete information, the CS theory accesses the success of signal recovery with the sparsity. There are different definitions of sparse signals. We consider a simple one which corresponds to so-called exactly sparse signals. Approximately sparse (or compressible) signals will be defined elsewhere when needed. A signal $\boldsymbol{x}^o \in \mathbb{R}^N$ of length N is called K-sparse under a basis $\boldsymbol{\Psi} \in \mathbb{R}^{N\times N}$ if all but at most a number of $K \ll N$ entries of its coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^N$ are zeros with $\boldsymbol{x}^o = \boldsymbol{\Psi\theta}$. Sparse/compressible signals are ubiquitous in reality. As an example shown in Fig. 2.1, a natural image can be compressively represented in a wavelet domain with most of the energy concentrating at a few number of wavelet coefficients. Without loss of generality we assume that $\boldsymbol{\Psi}$ is an identity matrix, i.e., $\boldsymbol{x}$ is sparse in the canonical basis, since for a general basis $\boldsymbol{\Psi}$ it can be absorbed into the following introduced sensing matrix $\boldsymbol{A}$. Rather than observing directly the original sparse signal $\boldsymbol{x}^o$, a number of M, $K < M \ll N$, linear measurements are acquired in CS as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^o, \tag{2.1}$$

where $\boldsymbol{y} \in \mathbb{R}^M$ is the measurement vector and $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ denotes the sensing/measurement matrix. Without considering the signal structure, the recovery of $\boldsymbol{x}^o$ from $\boldsymbol{y}$ is generally ill-posed since there are less linear equations than the unknown variables. By accounting for the signal sparsity, naturally, we want to seek for the sparsest solution by solving the combinatorial optimization problem

$$\min \|\boldsymbol{x}\|_0, \text{ subject to } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \tag{2.2}$$

where $\|\boldsymbol{x}\|_0$ denotes the pseudo ℓ_0 norm which counts the number of nonzero entries of $\boldsymbol{x}$. Unfortunately, the ℓ_0 optimization cannot be used in practice since it is non-deterministic polynomial-time (NP) hard. Alternative, efficient approaches are needed. Chen *et al.* [13] propose to solve the basis pursuit (BP) problem

$$\min \|\boldsymbol{x}\|_1, \text{ subject to } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \tag{2.3}$$

which is a tight convex relaxation of the ℓ_0 optimization problem and can be solved in a polynomial time. In fact, the use of ℓ_1 optimization for obtaining a sparse solution can date back to [14] for seismic data recovery. BP was empirically observed to give good performance, however, rigorous analysis had not been provided until the last decade.

2.1 Deterministic Results

We review some of the existing deterministic results in this section which mainly fall into the following two categories: one based on mutual coherence [15] and the other based on restricted isometry property (RIP) [16]. By deterministic results we mean that they can be applied once the sensing matrix satisfies some condition.

Definition 2.1. *The mutual coherence of a matrix $\boldsymbol{A}$, $\mu(\boldsymbol{A})$, is the largest absolute*

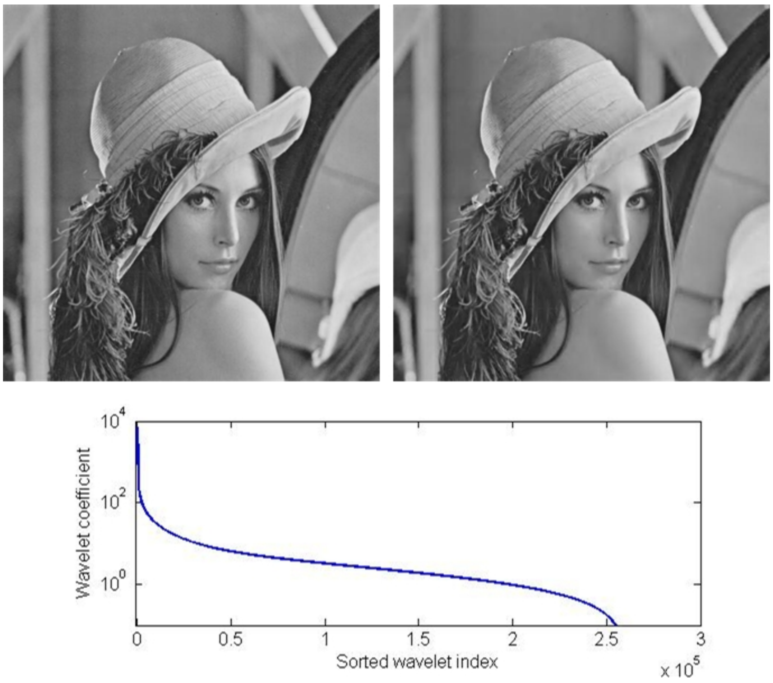


Figure 2.1: Upper left: original Lena image. Upper right: Lena image reconstructed from the largest (in amplitude) 10% of the wavelet coefficients. Lower: sorted absolute values of the wavelet coefficients of the original Lena image.

correlation between any two columns of $\boldsymbol{A}$*, i.e.,*

$$\mu(\boldsymbol{A}) = \max_{i \neq j} \frac{|\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|}{\|\boldsymbol{a}_i\|_2 \|\boldsymbol{a}_j\|_2}. \tag{2.4}$$

Definition 2.2. *The* k*-restricted isometry constant (RIC) of a matrix* $\boldsymbol{A}$*,* $\delta_k(\boldsymbol{A})$*, is the smallest number such that the inequality*

$$(1 - \delta_k(\boldsymbol{A})) \|\boldsymbol{v}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{v}\|_2^2 \leq (1 + \delta_k(\boldsymbol{A})) \|\boldsymbol{v}\|_2^2$$

holds for all k*-sparse vectors* $\boldsymbol{v}$*.* $\boldsymbol{A}$ *is said to satisfy the* k*-RIP with constant* $\delta_k(\boldsymbol{A})$ *if* $\delta_k(\boldsymbol{A}) < 1$*.*

Note that it is usually easy to compute the coherence of a matrix. But it may become infeasible to compute the RIC in the case of a large k (which is related to and typically larger than the signal sparsity) since the computational complexity increases exponentially with respect to k. Since the coherence considers only the relationship between any two columns of the sensing matrix while the RIP is on much more (a number adaptive to the signal sparsity) columns, the RIP may provide better results for the sparse signal recovery, which will be demonstrated later. Without ambiguity, we write μ and δ_k instead of $\mu\left(\boldsymbol{A}\right)$ and $\delta_k\left(\boldsymbol{A}\right)$ for short in the remaining of the chapter.

2.1.1 Convex Relaxation

Convex relaxation (or ℓ_1 optimization) has been a most studied method for the sparse signal recovery. The following theorem is the first result which demonstrates the equivalence between the BP and the ℓ_0 optimization.

Theorem 2.1 ([15]). *Assume that $\left\|\boldsymbol{x}^o\right\|_0 \leq K$ and $\mu < \frac{1}{2K-1}$. Then $\boldsymbol{x}^o$ is the unique solution of the ℓ_0 optimization and the BP.*

Theorem 2.1 states that under some conditions the solution of the NP-hard ℓ_0 optimization can be efficiently obtained by solving the convex BP problem. An RIP-based result is as follows.

Theorem 2.2 ([17]). *Assume that $\left\|\boldsymbol{x}^o\right\|_0 \leq K$ and $\delta_{2K} < \sqrt{2}-1$. Then $\boldsymbol{x}^o$ is the unique solution of the ℓ_0 optimization and the BP.*

After the work by Candès [17], the sufficient RIP condition has been improved, e.g., to $\delta_{2K} < \frac{3}{4+\sqrt{6}}$ by Foucart [18]. Other types of RIP conditions are also provided, e.g., $\delta_K < 0.307$ in [19]. Moreover, in the presence of measurement noises and/or when the signal of interest is not exactly sparse the convex relaxation can produce faithful signal estimates by solving the basis pursuit denoising (BPDN) problem

$$\min\left\|\boldsymbol{x}\right\|_1, \text{ subject to } \left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\right\|_2 \leq \epsilon, \tag{2.5}$$

where ϵ is an upper bound of the noise energy. Readers are referred to [17] or Chapter 5 for more details.

2.1.2 Other Approaches

At the beginning of the development of CS, convex relaxation methods (BP and BPDN) were solved using linear programming or second order cone programming while these solvers are usually slower than expected when dealing with practical problems of high dimension, e.g., the recovery of a 512 × 512 image. Other efficient approaches are developed as well for the sparse signal recovery and those with guaranteed performance include orthogonal matching pursuit (OMP) [20], compressive sampling matching pursuit (CoSaMP) [21], subspace pursuit (SP) [22] and iterative hard thresholding (IHT) [23]. We focus on OMP and IHT.

OMP is a greedy pursuit algorithm and has been studied by many researchers due to its concise implementation which is presented in Algorithm 2.1 below. It is proven in [24] that OMP provides the same theoretical guarantee as BP in terms of the mutual coherence, i.e., a K-sparse signal can be exactly recovered by OMP if $\mu < \frac{1}{2K-1}$. The first RIP-based analysis of OMP appears in [25], showing that if $\delta_{K+1} < \frac{1}{3\sqrt{K}}$, then OMP can exactly recover a K-sparse signal within K iterations. The RIP condition has been relaxed in later studies. For example, [26] shows that $\delta_{K+1} < \frac{1}{\sqrt{K}+1}$ is sufficient for the exact recovery within K iterations. However, it is clear that such RIP conditions where the upper bounds decay with respect to the sparsity level K are inferior to that for BP in Theorem 2.2 where an RIC bounded by a constant is sufficient. By allowing OMP to perform more than K iterations, [27] proves that the exact recovery of a K-sparse signal can be obtained within $30K$ iterations if $\delta_{31K} < \frac{1}{3}$, where the coefficients before K are not optimal and expected to be decreased in the future.

On the other hand, IHT starts with $\boldsymbol{x}^0 = \boldsymbol{0}$ and carries out the following iteration

$$\boldsymbol{x}^{l+1} = H_K\left(\boldsymbol{x}^l + \boldsymbol{A}^T\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^l\right)\right), \tag{2.6}$$

where $H_K(\boldsymbol{v})$ denotes a hard thresholding operator which sets all but the largest (in amplitude) K entries of a vector $\boldsymbol{v}$ to zero. The following result is proven in [23]:

Algorithm 2.1: Orthogonal matching pursuit (OMP)

```
Input: $\boldsymbol{A}$, $\boldsymbol{y}$
1. initialize: $\boldsymbol{r}^0 = \boldsymbol{y}$, $\Omega = \emptyset$, $l = 0$
2. while not converged do
3. match: $\boldsymbol{h}^l = \boldsymbol{A}^T \boldsymbol{r}^l$
4. identify: $\Omega^{l+1} = \Omega^l \cup \arg\max_j \left\{ \left| h_j^l \right| \right\}$
5. update: $\boldsymbol{x}^{l+1} = \arg\min_{\boldsymbol{z}, \text{supp}(\boldsymbol{z}) \subset \Omega^{l+1}} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{z} \right\|_2$
           $\boldsymbol{r}^{l+1} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{l+1}$
           $l = l + 1$
6. end while
Output: $\widehat{\boldsymbol{x}} = \boldsymbol{x}^l = \arg\min_{\boldsymbol{z}, \text{supp}(\boldsymbol{z}) \subset \Omega^l} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{z} \right\|_2$
```

if $\delta_{2K} < \frac{1}{\sqrt{8}}$, then IHT can exactly recover a K-sparse signal. Moreover, IHT has stable performance in the presence of measurement noises and when the signal of interest is not exactly sparse. Though IHT has comparable theoretical guarantees as BP with a more concise implementation, one disadvantage of IHT is that it requires the signal sparsity level in advance while it is not needed by BP. In fact, the same disadvantage is also shared among the above mentioned OMP, CoSaMP and SP.

2.1.3 Number of Measurements for Guaranteed Performance

In the last two subsections, various sufficient conditions are provided for BP and other approaches to guarantee the recovery of the sparse signal of interest, however, it is unclear how many measurements are sufficient for the sparse signal recovery. Random sensing matrices, especially sub-gaussian ones, are typically considered to answer the question.

Definition 2.3. *A random variable X is called sub-gaussian iff there exist constants c, $C > 0$ such that $P(|X| > t) < C \exp(-ct^2)$.*[1]

Examples of sub-gaussian random variables include Gaussian and all bounded random variables. A sub-gaussian matrix means that all entries of the matrix are drawn

[1] The notations c and C will be repetitively used in the thesis whose values may change from instance to instance.

independently and identically from a sub-gaussian distribution. The following theorems reveal the relationship between the mutual coherence as well as the RIP and the number of measurements.

Theorem 2.3 ([28]). *Suppose that the elements of matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ are drawn i.i.d. from a sub-gaussian distribution with mean zero and variance M^{-1}. Then there exists a constant $C > 0$ such that the following holds for every number $\bar{\mu} \in (0, 1)$ and $\epsilon \in (0, 1)$:*

$$\textit{if } M \geq C\bar{\mu}^{-2} \log \frac{N}{\epsilon}, \quad \textit{then } P\left(\mu < \bar{\mu}\right) \geq 1 - \epsilon^2.$$

Theorem 2.4 ([29]). *Suppose that the elements of matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ are drawn i.i.d. from a sub-gaussian distribution with mean zero and variance M^{-1}. Then there exists constants c, $C > 0$ such that the following holds for every sparsity level $1 \leq K \leq N$ and every number $\bar{\delta} \in (0, 1)$:*

$$\textit{if } M \geq C\bar{\delta}^{-2} K \log \frac{eN}{K}, \quad \textit{then } P\left(\delta_K < \bar{\delta}\right) \geq 1 - 2\exp\left(-c\bar{\delta}^2 M\right).$$

For both convex relaxation and OMP, the results presented previously show that the mutual coherence of the sensing matrix is required to be on the order of K^{-1}, denoted by $\sim K^{-1}$, for recovering a sparse signal with the sparsity level of K. According to Theorem 2.3, a number of $M \sim K^2 \log N$ random sub-gaussian measurements are required to guarantee its recovery. Meanwhile, a (moderately small) constant RIC can guarantee the signal recovery according to the RIP-based results for all of the BP, OMP and IHT. As a result, a number of $M \sim K \log \frac{N}{K}$ random measurements are sufficient following from Theorem 2.4. So, the RIP-based results are superior to the ones based on the coherence in terms of the number of measurements required for guaranteed performance. Before closing this section, we note that the RIP-based results provide only sufficient conditions and it is shown in [30] that such conditions can be quite conservative. In the next section, a sufficient and necessary condition will be introduced in the case of Gaussian sensing matrices.

2.2 Phase Transition

The theory of CS is mainly focused on studying how aggressively a sparse signal can be undersampled while still preserving all information for its recovery. The existing results based on the mutual incoherence and the RIP are introduced in the last section which provide sufficient conditions for the sparse signal recovery and are conservative in practice [30]. So far the theory of phase transition [4, 6, 31–33] is the most precise one owing to its necessary and sufficient condition for measuring the sparsity-undersampling tradeoff performance. Before our work [4–6], the study of phase transition is focused on the real setting where the sensing matrix and the sparse signal are real valued. Our main interest is the sparsity-undersampling tradeoff of the BP problem in (2.3). Some variants will be discussed later. Consider sensing matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ with i.i.d. random Gaussian entries. Denote by $\delta = M/N$ the sampling ratio and by $\rho = K/M$ the sparsity ratio, where K denotes the sparsity level of the signal $\boldsymbol{x}$ of interest. Then the sparsity-undersampling tradeoff of the BP can be precisely described by δ and ρ when the problem dimension approaches infinity. In particular, the plane of (δ, ρ) is divided by a phase transition curve into two phases, a 'success' phase where BP successfully recovers the sparse signal and a 'failure' phase where the original signal cannot be recovered by solving BP, both with an overwhelming probability (see Fig. 2.2).

So far three different approaches have been developed to calculate the phase transition curve of BP, including combinatorial geometry [31], null space method [32] and state evolution [33]. Different formulae are derived based on different approaches, however, they coincide with each other numerically. Since the formulae are complicated, their expressions are omitted. An interesting scenario is when

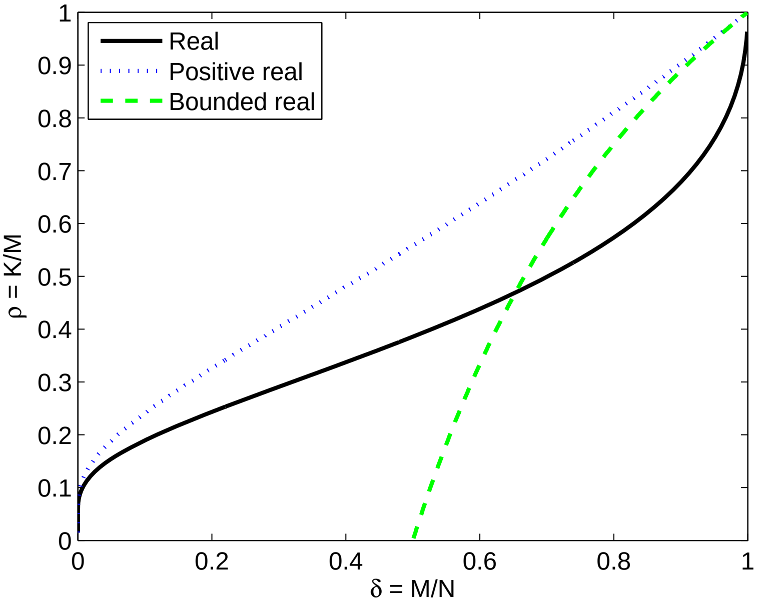


Figure 2.2: Phase transition curves of CS with three different settings derived in [31–33].

$\delta \to 0$, i.e., in the extreme undersampling case. It is shown in [31] and [33] that $\rho^R(\delta) \sim [2\ln(1/\delta)]^{-1}$ as $\delta \to 0$, where $\rho^R(\delta)$ denotes the real phase transition of BP and by the expression $a \sim b$ we mean $a = b + o(b)$. This means that, the sparsity ratio stays positive whenever $\delta > 0$, i.e., for any given degree of undersampling, there exists a signal of sufficient sparsity leading to successful reconstruction using BP. Informally, we can conclude that a K-sparse signal can be reconstructed provided that the number of random Gaussian measurements $M \geq 2K \ln \frac{N}{M}$. Notice that the number of measurements required is about as the same order as that derived based on the RIP in Subsection 2.1.3 but with a significantly reduced coefficient. Numerical simulations have shown that the observed phase transition matches the theoretical curve even for modestly large N, e.g., $N = 1000$ [3, 34], where the phase transition for finite-N is defined as the value of ρ at which the original signal is successfully recovered with the probability of 50%. Moreover, it is practically observed that the Gaussian condition on $\boldsymbol{A}$ can be considerably relaxed, resulting in the well known observed universality of phase transitions [34].

Another two phase transition curves have also been derived respectively in other two different settings. One imposes that the sparse signal of interest is positive valued and thus the following BP problem is solved:

$$\min_{\boldsymbol{x}\succeq\mathbf{0}} \sum_i x_i, \text{ subject to } \boldsymbol{A}\boldsymbol{x}=\boldsymbol{y}, \tag{2.7}$$

where $\succeq$ denotes $\geq$ with an elementwise operation. The other assumes that all of the nonzero entries of the sparse signal fall into the interval $(0,1]$ and the signal is recovered by solving the following feasibility problem:

$$\text{find } \boldsymbol{x}\in[0,1]^N, \text{ subject to } \boldsymbol{A}\boldsymbol{x}=\boldsymbol{y}. \tag{2.8}$$

The phase transition curves are displayed in Fig. 2.2. Since additional knowledge is taken into the problem solving, the phase transition of (2.7) is positioned well above that of the standard BP. But they are equal asymptotically when $\delta\to 0$ [35], i.e., the value of positivity constraints evaporates at extreme undersampling. The phase transition of (2.8) has a simple form: $\rho(\delta)=\max\left(0,2-\delta^{-1}\right)$, i.e., the reconstruction is possible only when $\delta>0.5$.

Complex data are involved in numerous CS applications, e.g., MRI [36], radar imaging [37] and source localization [11], where both the signal and sensing matrix are complex valued. Hence, sparsity-undersampling tradeoff in the complex setting deserves further studies. Chapter 4 presents our first result on the phase transition in the complex domain, where a new phase transition is discovered which is positioned well above the one in the real setting. Its theoretical derivation is later carried out in [6].

2.3 Sparse Bayesian Learning

Bayesian approaches to CS, known as Bayesian CS [38], was originated from the area of machine leaning and introduced by Tipping [39] for obtaining sparse solutions to regression and classification tasks that use models which are linear in the parameters, coined as relevance vector machine (RVM) or sparse Bayesian learning (SBL). SBL is built upon a statistical perspective where the sparsity information is exploited by assuming a sparsity-inducing prior for the signal of interest that is then estimated via Bayesian inference. Its theoretical performance is analyzed by Wipf and Rao [40]. After being introduced into CS by Ji *et al.* [38], this technique has become a popular approach to CS and other sparsity-related problems, see a list of references [11, 41–51]. It is noted that Chapters 6, 7 and 8 of the thesis are built upon this approach. Main research topics in SBL for CS include (a) developing efficient sparsity-inducing priors, (b) incorporating additional signal structures in the prior besides sparsity and (c) designing fast and accurate inference algorithms.

Many sparsity-inducing priors have been studied in the literature. In [44, 50], a spike-and-slab prior [52] is applied which is a mixture of a point mass at zero and a continuous distribution elsewhere and fits naturally for sparse signals. A typical inference scheme for such a prior is a Markov chain Monte Carlo (MCMC) method [53] due to the lack of closed-form expressions of Bayesian estimators. As a result, the inference process may suffer from computational difficulties because a large number of samples are required to approximate the posterior distribution and the convergence is typically slow. A popular class of sparsity-inducing priors is introduced in a hierarchical framework where a complex prior is composed of two or more simple distributions. For example, a Student's t-prior (or Gaussian-inverse gamma prior) is used in the basic SBL [39] that is composed of a Gaussian prior in the first layer and a gamma prior in the second. A Laplace (Gaussian-exponential) prior is used in [46]. A Gaussian-gamma prior is recently studied in [54] that generalizes the Laplace prior. Two popular inference methods for the hierarchical priors are evidence procedure [55], e.g., in [39, 46], and variational Bayesian inference [56], e.g., in [48]. Both the methods are approximations of Bayesian inference since the exact inference is intractable. In an evidence procedure, the signal estimator has a simple expression in which some unknown hyperparameters are involved and estimated iteratively by maximizing their evidence. In a variational Bayesian inference method, the posterior distribution is approximated using some family of tractable distributions followed by computation of an optimal distribution within the family. To circumvent high-dimensional matrix inversions, a fast algorithm framework is developed in [57] for evidence procedure and also adopted in [46]. When additional signal structures besides sparsity are known *a priori*, they can be incorporated into the signal prior for possibly improved performance. For example, [44] exploits the wavelet structure of images and [50] studies cluster structured sparse signals. SBL is also applied to the multiple-measurement-vectors (MMV) case where joint sparsity which exists among the two or more signals of interest is exploited [41, 43, 49].

Though formulated from a different perspective, SBL is related to other approaches to CS. Consider the observation model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ where $\boldsymbol{e}$ represents an additive white Gaussian noise (AWGN). Let $p(\boldsymbol{x})$ be the prior for $\boldsymbol{x}$. Then, a maximum *a posteriori* (MAP) estimator of $\boldsymbol{x}$ coincides with a solution to a regularized least-squares problem with $-\log p(\boldsymbol{x})$ (up to a scale) being the regularization term, which bridges SBL and optimization methods. For example, a Laplace prior corresponds to the widely studied ℓ_1 minimization. A prior corresponding to the nonconvex ℓ_p $(0 < p < 1)$ norm is studied in [45]. The fast algorithm in [57] is related to the greedy pursuit method. In fact, it is a greedy algorithm using a different support modification criterion. Unlike OMP and StOMP, it allows deletion of irrelevant basis vectors that may have been added to the solution support in earlier steps.

2.4 CS with Practical Considerations

The observation model is $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ in standard CS. The main problem to explore is under what conditions and how the signal $\boldsymbol{x}$ can be recovered given the measured samples $\boldsymbol{y}$, the sensing matrix $\boldsymbol{A}$ and possibly the noise information as well. It is noted that $\boldsymbol{y}$ and $\boldsymbol{A}$ are assumed to be exactly known, which may not be satisfied in practice. We focus on the gap between standard CS and practical situations in this section and review some existing results on these issues.

2.4.1 Measurement Quantization

The real-valued measurement $\boldsymbol{y}$ is assumed in the standard CS framework which has infinite bit precision. But it is known that measurement quantization is inevitable for practical considerations, e.g., data storage and transmission. During the quantization process, each continuous-valued measurement is quantized into some value in a finite set. A new challenge for the sparse signal recovery is thus the existence of quantization errors. The noise free case with a uniform unsaturated quantizer is studied in [1, 58, 59]. A solver with quantization consistency is recommended in [1] that corresponds to replacing the ℓ_2 norm in the BPDN problem by the ℓ_∞ norm. A BPDN solver is used in [58] which treats the quantization errors as additive noises with bounded energy. A family of solvers, named as basis pursuit dequantizer of moment p (BPDQ$_p$), which includes BPDN and that in [1] as special cases is studied in [59] where the ℓ_2 norm in BPDN is replaced by an ℓ_p norm with $2 \leq p \leq +\infty$. By characterizing the quantization errors as independent random variables uniformly distributed in a common interval, it is shown in [59] that the optimal signal recovery accuracy is obtained at some finite $p \geq 2$. But unfortunately, the optimal p cannot be explicitly given in practice. Note that the common uniform distribution assumption is crucial to obtain the results in [59]. As a result, it is unclear whether the results in [59] can be extended to a general quantizer case where such an assumption fails. It is obvious that both BPDN and BPDQ$_p$ are inappropriate in the case of a saturated quantizer since data saturation may lead to large or even unbounded quantization errors which deteriorate their performance. To deal with the data saturation, Laska *et al.* [60] propose two modified versions of BPDN which either reject saturated measurements or incorporate them into signal recovery. While quantization errors and measurement noises are coupled in most existing methods (some methods, e.g., BPDQ$_p$, consider only the noise free case to avoid such a problem), e.g., in [60], they are separately studied by Zymnis *et al.* [61] where the authors seek to find a signal estimate that maximizes the likelihood of the quantized measurements while the ℓ_1 norm is used to promote the signal sparsity. The resulting algorithm is quoted as ℓ_1-regularized maximum likelihood (L1RML).

An extreme case of quantized CS is so-called 1-bit CS where each quantized measurement keeps only the sign information of the real measurement and thus uses just one bit. The 1-bit CS framework is proposed in [62] and has attracted many research interests because it possesses many merits. For example, a 1-bit quantizer is a simple comparator that tests whether the measurement is above or below zero, leading to an easy implementation and a fast quantization process. A measurement noise can be neglected in 1-bit CS as long as it does not change the sign of the measurement. It is shown in [63] that to acquire just one bit for each measurement is optimal in the presence of heavy noises. The 1-bit case is quite different from the multi-bit case since all measurements are saturated in 1-bit CS and the signal scaling information is lost. A common approach to the signal scaling problem is to impose that the signal to be recovered has a fixed unit norm and then search for the signal on the unit hyper-sphere rather than in the whole space. Such a constraint is nonconvex and brings new challenges to algorithm design. Existing algorithms based on this constraint include renormalized fixed point iteration (RFPI) [62], matching sign pursuit (MSP) [64], restricted-step shrinkage (RSS) [65] and binary iterative hard thresholding (BIHT) [66]. Convex formulations of the 1-bit CS problem have been recently proposed by Plan and Vershynin [67, 68]. They show in [67] that a linear program can decode the noiseless case with guaranteed signal recovery accuracy under similar mild conditions as in conventional CS. In [68] they introduce a seemingly unrelated convex program for the noisy case and show similar results. It is noted that both the BIHT and the convex program in [68] which deal with the noisy 1-bit CS problem require the signal sparsity information (BIHT needs the sparsity level K and CVXP requires a proper upper bound for the signal's ℓ_1 norm). While the multi- and 1-bit CS problems are studied separately in the literature due to their big difference, we present a Bayesian framework that unifies both the cases in Chapter 7.

2.4.2 Sensing Matrix Perturbation

The sensing matrix $\boldsymbol{A}$ is assumed known *a priori* in standard CS, which is, however, not always the case in practical situations. For example, a matrix perturbation can be caused by quantization during implementation. In source separation [69, 70] the sensing matrix (or mixing system) is usually unknown and needs to be estimated, and thus estimation errors exist. In source localization such as direction of arrival (DOA) estimation [11, 71] and radar imaging [37, 72], the sensing matrix (or overcomplete dictionary) is constructed via discretizing one or more continuous parameters, and errors exist typically in the sensing matrix since the true source locations may not be on a discretized sampling grid.

There have been recent active studies on the CS problem where the sensing matrix is unknown or subject to an unknown perturbation. Gleichman and Eldar [73] introduce a concept named as blind CS where the sensing matrix is assumed unknown. Herman and Strohmer [74] analyze the effect of a general matrix perturbation and show that the signal recovery is robust to the perturbation in the sense that the recovery error grows linearly with the perturbation level. Similar robust recovery results are reported in [75, 76]. It is demonstrated in [76, 77] that the signal recovery may suffer from a large error under a large perturbation. In addition, the existence of recovery error caused by the perturbed sensing matrix is independent of the sparsity of the original signal. Algorithms have also been proposed to deal with sensing matrix perturbations. For example, Zhu *et al.* [78] propose a sparse total least-squares approach to alleviating the effect of perturbation where they explore the structure of the perturbation to improve recovery performance. In Chapter 8, we formulate the off-grid DOA estimation problem from a sparse Bayesian inference perspective and iteratively recover the source signal and the matrix perturbation. It is noted that existing algorithmic results provide no guarantees on the signal recovery accuracy when there exist perturbations in the sensing matrix. In Chapter 5, we study the compressed sensing problem subject to a structured perturbation in the sensing matrix. Theoretical guarantees are provided for the signal recovery performance via an ℓ_1 optimization approach. Its practical relevance to the DOA estimation is studied in Chapter 8.

2.5 Applications of CS

It is always of great interest to acquire a signal using reduced measurements which can still guarantee its recovery since each measurement may have a potential time and/or power cost. The inherent reason of the success of CS is the signal sparsity. Fortunately, most natural signals can be sparsely represented in some basis, rendering that CS has been successfully applied to many problems in the past few years. Current applications include single-pixel camera [79], medical imaging [36,80], communications [81, 82] and computational biology [83], to name just a few. In this section we focus on the applications of CS to array processing, specifically, direction of arrival (DOA) estimation [84], and magnetic resonance imaging (MRI) [85] which are closely related to the thesis.

2.5.1 DOA Estimation

Source localization using sensor arrays [84] has been an active research area for decades. We focus on the narrowband far-field source case where the wave front is assumed to be planar and the angle/direction information is to be estimated, known as the DOA estimation problem. MUSIC is the most successful method among conventional DOA estimation techniques in the case of a large number of snapshots and uncorrelated source signals [86, 87]. The research on the DOA estimation has been advanced in recent years owing to the development of methods based on sparse signal representation (SSR) or CS [88, 89]. Note that the DOAs of interest lie in a continuous range while CS is applicable to recovery of discrete signals. Hence, discretization of the DOA range is used in [88, 89], where a fixed sampling grid is selected which serves as the set of all candidates of DOA estimates. Then by assuming that all true (unknown) DOAs are exactly on the selected grid, an SSR problem can be formulated where the DOAs of interest constitute the support of the sparse signal to be recovered.

Though existing CS-based approaches have shown their improvements in the DOA estimation, e.g., their success in the case of limited snapshots, there are still difficulties in practical situations where the true DOAs are not on the sampling grid. On one hand, a dense sampling grid is necessary for accurate DOA estimation to reduce the gap between the true DOA and its nearest grid point since the estimated DOAs are constrained on the grid. On the other hand, too dense a sampling grid leads to a highly coherent matrix that violates the condition for the sparse signal recovery. Chapter 8 concerns this problem where we propose novel approaches to off-grid DOA estimation.

2.5.2 MRI

MRI is a noninvasive medical imaging technique to display detailed internal structure of the body. High-quality MR images enable physicians to assess the health condition of the body and determine the presence of certain diseases. MRI uses no ionizing radiation and does not have the associated potential harmful effects. Instead, a strong magnetic field is utilized to align the nuclear magnetization of hydrogen atoms in water of the body. Then MRI adopts radio frequency fields to change the alignment of the magnetization, which causes the hydrogen nuclei to generate a magnetic field that can be detected by the scanner. Finally, additional magnetic fields are used to manipulate this signal to gather enough information, known as k-space, to construct an image of the body. The k-space data can be formulated as a 2D or 3D Fourier transform of the image, and then inverse Fourier transform will reconstruct the image.

The patient is required to stay still during the k-space data collection process. Imaging speed is thus important in many MRI applications since patient motion happens very likely in a long scanning process. However, there are many limitations, including physical and physiological constraints, to speed up data collection in MRI.

Hence, one research area is to seek for methods to reconstruct high-quality images from reduced (sub-Nyquist sampled) k-space data while each of which has a potential time cost. Current applications of CS to MRI are focused on this by exploiting the image sparsity under, for example, a wavelet domain. A short list of publications include [36, 90–94].

Motion correction is another active research topic in MRI since a small amount of motions in practice may cause severe imaging artifacts. Existing methods [95–99] resolve the problem by acquiring more k-space data which prolongs the scanning process and may possibly introduce more motions. While CS has been vastly applied to MRI for imaging acceleration, few results have appeared on its application to the motion correction problem. Chapter 9 presents our first sparsity-based approach to MRI motion correction.

2.6 Conclusion

In this chapter we have reviewed many of the existing results of CS, including main approaches to the sparse signal recovery in CS (e.g., convex relaxation, greedy pursuit and SBL), their theoretical guarantees (those based on RIP, coherence and the phase transition theory) and practical applications of CS (e.g., DOA estimation and MRI). We have also discussed some open problems or those that need to be further studied. The following chapters are focused on solving these problems.

Part I

Analysis and Algorithms

Chapter 3

Orthonormal Expansion ℓ_1-Minimization Algorithms

Many existing results of CS have been reviewed in Chapter 2. Among both the theoretical and algorithmic results, convex relaxation (or ℓ_1 minimization) plays an important role for the sparse signal recovery. Since the ℓ_1 norm is non-smooth at the origin, algorithm design remains a major problem with the convex relaxation methods for dealing with practical problems of high dimension. In this chapter we study practically useful algorithms for the sparse signal recovery based on ℓ_1 minimization. Notice that besides the basis pursuit (BP) formulation in the noise-free case which takes the form

$$\text{(BP)} \qquad \min_{\boldsymbol{x}} \left\|\boldsymbol{x}\right\|_1, \text{ subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}, \tag{3.1}$$

two other formulations in the noisy case are basis pursuit denoising (BPDN) and a regularized form (QP) which are, respectively,

$$\text{(BPDN)} \qquad \min_{\boldsymbol{x}} \left\|\boldsymbol{x}\right\|_1, \text{ subject to } \left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right\|_2 \leq \epsilon, \tag{3.2}$$

$$\text{(QP)} \qquad \min_{\boldsymbol{x}} \left\{\lambda \left\|\boldsymbol{x}\right\|_1 + \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right\|_2^2\right\}. \tag{3.3}$$

In the above, ϵ is an upper bound of the noise level and $\lambda > 0$ is a regularization parameter. It is well known that BPDN and QP are equivalent with appropriate choices of ϵ and λ. In general, λ decreases as ϵ decreases. In the limiting case of $\lambda, \epsilon \to 0$, both BPDN and QP converges to the BP for the noiseless case. Since QP seems to be easier to solve than BPDN and BP, many algorithms have been proposed to solve QP or based on that. Convex optimization methods include ℓ_1-magic [100], interior-point method [101], conjugate gradient method [36], fixed-point continuation [102] and NESTA [103] based on Nesterov's smoothing technique [104] with continuation.

This chapter is based on [3] and mainly focused on solving the BP in noiseless CS. We consider the case where $\boldsymbol{AA'}$ is an identity matrix, i.e., the rows of $\boldsymbol{A}$ are orthonormal. This consists of most fast transforms in CS, e.g., discrete cosine transform (DCT), discrete Fourier transform (DFT) and some wavelet transforms, e.g., Haar wavelet transform. This case has also been studied in other algorithms, e.g., NESTA. We introduce a novel method called orthonormal expansion to reformulate BP. The exact OrthoNormal Expansion ℓ_1 minimization (eONE-L1) algorithm is then proposed to exactly solve BP based on the augmented Lagrange multiplier (ALM) method [105, 106]. The relaxed ONE-L1 (rONE-L1) algorithm is further developed to simplify eONE-L1. We show that rONE-L1 converges at least exponentially and is in the form of modified iterative soft thresholding (see details in Subsection 3.1.2). In the case of strictly sparse signals and noise-free measurements, numerical simulations show that rONE-L1 has the same sparsity-undersampling tradeoff as BP. We further compare the proposed algorithms with existing methods including FPC-AS [107], AMP [33] and NESTA.

3.1 Preliminaries

3.1.1 Soft Thresholding Operator

For $w \in \mathbb{R}$, the soft thresholding of w with a threshold $\lambda \in \mathbb{R}^+$ is defined as:

$$S_\lambda(w) = \operatorname{sgn}(w) \cdot (|w| - \lambda)^+, \tag{3.4}$$

where $(\cdot)^+ = \max(\cdot, 0)$ and

$$\operatorname{sgn}(w) = \begin{cases} w/|w|, & w \neq 0; \\ 0, & w = 0. \end{cases} \tag{3.5}$$

The operator $S_\lambda(\cdot)$ can be extended to vector variables by its element-wise operation.

The soft thresholding operator can be applied to solve the following ℓ_1-norm regularized least square problem [108], i.e.,

$$S_\lambda(\boldsymbol{w}) = \arg\min_{\boldsymbol{v}} \left\{ \lambda \|\boldsymbol{v}\|_1 + \frac{1}{2} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2 \right\}, \tag{3.6}$$

where $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^M$, and $S_\lambda(\boldsymbol{w})$ is the unique minimizer.

3.1.2 Iterative Soft Thresholding

One popular approach to QP is iterative soft thresholding (IST) of the form (stating from $\boldsymbol{x}_0 = 0$) [108, 109]

$$\boldsymbol{x}_{t+1} = S_\lambda(\boldsymbol{x}_t + \boldsymbol{A}'\boldsymbol{z}_t), \quad \boldsymbol{z}_t = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_t, \tag{3.7}$$

where $'$ denotes the transpose operator. IST has a concise form and is easy to implement, but its convergence can be very slow [109], especially for small λ. To improve its speed, a fixed-point continuation (FPC) strategy is presented in [102], where λ is decreased in a continuation scheme and a q-linear convergence rate is achieved. Further, FPC-AS [107] improves the performance of FPC by introducing an active set inspired by greedy pursuit methods. An alternative approach to improving the speed of IST is to use an aggressive continuation where λ is decreased in each iteration:

$$\boldsymbol{x}_{t+1} = S_{\lambda_t}\left(\boldsymbol{x}_t + \boldsymbol{A}'\boldsymbol{z}_t\right), \quad \boldsymbol{z}_t = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_t. \tag{3.8}$$

However, the algorithm of this form typically has a worse sparsity-undersampling tradeoff than BP [110]. In the case of recovering exactly sparse signals, such a disadvantage is overcome by approximate message passing (AMP) [33] which has the same sparsity-undersampling tradeoff as BP and is a modified IST:

$$\boldsymbol{x}_{t+1} = S_{\lambda_t}\left(\boldsymbol{x}_t + \boldsymbol{A}'\boldsymbol{z}_t\right), \quad \boldsymbol{z}_t = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_t + \frac{N\left\|\boldsymbol{x}_t\right\|_0}{M}\boldsymbol{z}_{t-1}. \tag{3.9}$$

3.2 ONE-L1 Algorithms

3.2.1 Problem Description

Consider the ℓ_1-minimization problem BP with the sampling matrix $\boldsymbol{A}$ satisfying that $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix. We say that $\boldsymbol{A}$ is a partially orthonormal matrix hereafter as its rows are usually randomly selected from an orthonormal matrix in practice, e.g. partial-DCT matrix. Hence, there exists another partially orthonormal matrix $\boldsymbol{B} \in \mathbb{R}^{(N-M)\times N}$, whose rows are orthogonal to those of $\boldsymbol{A}$, such that $\boldsymbol{\Phi} = \begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{bmatrix}$ is orthonormal. Let $\boldsymbol{p} = \boldsymbol{\Phi}\boldsymbol{x}$. The BP is then equivalent to

$$(\mathrm{BP}^o) \qquad \min_{\boldsymbol{x},\boldsymbol{p},\Gamma(\boldsymbol{p})=\boldsymbol{y}} \left\|\boldsymbol{x}\right\|_1, \text{ subject to } \boldsymbol{\Phi}\boldsymbol{x} = \boldsymbol{p},$$

where $\Gamma(\boldsymbol{p})$ is an operator projecting the vector $\boldsymbol{p}$ onto its first M entries.

In (BP^o), the sampling matrix $\boldsymbol{A}$ is expanded into an orthonormal matrix $\boldsymbol{\Phi}$. It corresponds to the scenario where the full sampling is carried out with the sampling matrix $\boldsymbol{\Phi}$ and $\boldsymbol{p}$ is the vector containing all measurements. Note that only $\boldsymbol{y}$, as a part of $\boldsymbol{p}$, is actually observed. To expand the sampling matrix $\boldsymbol{A}$ into an orthonormal matrix $\boldsymbol{\Phi}$ is a key step to show that the ALM method exactly solves (BP^o) and, hence, BP. The next subsection describes the proposed algorithm, referred to as orthonormal expansion ℓ_1-minimization (ONE-L1).

3.2.2 ALM Based ONE-L1 Algorithms

In this subsection we solve the ℓ_1-minimization problem (BP^o) using the ALM method [105, 106]. The ALM method is similar to the quadratic penalty method except an additional Lagrange multiplier term. Compared with the quadratic penalty method, the ALM method has some salient properties, e.g., the ease of parameter tuning and the convergence speed. The augmented Lagrangian function is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}, \mu) = \|\boldsymbol{x}\|_1 + \langle \boldsymbol{p} - \boldsymbol{\Phi}\boldsymbol{x}, \boldsymbol{\nu} \rangle + \frac{\mu}{2} \|\boldsymbol{p} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2, \tag{3.10}$$

where Lagrange multiplier $\boldsymbol{\nu} \in \mathbb{R}^N$, $\mu \in \mathbb{R}^+$ and $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}'\boldsymbol{v} \in \mathbb{R}$ is the inner product of $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{R}^N$. (3.10) can be expressed as follows:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}, \mu) = \|\boldsymbol{x}\|_1 + \frac{\mu}{2} \left\|\boldsymbol{p} - \boldsymbol{\Phi}\boldsymbol{x} + \mu^{-1}\boldsymbol{\nu}\right\|_2^2 - \frac{1}{2\mu} \|\boldsymbol{\nu}\|_2^2 . \tag{3.11}$$

Subsequently, we have the following optimization problem (SP):

$$(SP) \qquad \min_{\boldsymbol{x}, \boldsymbol{p}, \Gamma(\boldsymbol{p}) = \boldsymbol{y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}, \mu).$$

Instead of solving (SP), let us consider the two related problems

$$(SP_1) \qquad \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}, \mu),$$

and

$$(SP_2) \qquad \min_{\boldsymbol{p}, \Gamma(\boldsymbol{p}) = \boldsymbol{y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}, \mu).$$

Note that problem (SP_1) is similar to the ℓ_1-regularized problem in (3.6). In general, (SP_1) cannot be directly solved using the soft thresholding operator as that in (3.6) since there is a matrix product of Φ and $\boldsymbol{x}$ in the term of ℓ_2-norm. However, the soft thresholding operator does apply to the special case where $\boldsymbol{\Phi}$ is orthonormal. Given $\|\boldsymbol{\Phi u}\|_2 = \|\boldsymbol{u}\|_2$ for any $\boldsymbol{u} \in \mathbb{R}^N$, we can apply the soft thresholding to obtain

$$S_{\mu^{-1}}\left(\boldsymbol{\Phi}'\left(\boldsymbol{p} + \mu^{-1}\boldsymbol{\nu}\right)\right) = \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}, \mu). \tag{3.12}$$

To solve (SP_2), we let $\partial_{\overline{\Gamma}(\boldsymbol{p})}\mathcal{L}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}, \mu) = 0$ to obtain $\overline{\Gamma}(\boldsymbol{p}) = \overline{\Gamma}\left(\boldsymbol{\Phi x} - \mu^{-1}\boldsymbol{\nu}\right)$, i.e.,

$$\begin{bmatrix} \boldsymbol{y} \\ \overline{\Gamma}\left(\boldsymbol{\Phi x} - \mu^{-1}\boldsymbol{\nu}\right) \end{bmatrix} = \arg\min_{\boldsymbol{p}, \Gamma(\boldsymbol{p}) = b} \mathcal{L}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}, \mu), \tag{3.13}$$

where $\overline{\Gamma}(\cdot)$ is the operator projecting the variable to its last $N - M$ entries. As a result, an iterative solution of (SP) is stated in the following lemma.

Lemma 3.1. *For fixed $\boldsymbol{\nu}$ and μ, the iterative algorithm given by*

$$\boldsymbol{x}^{j+1} = S_{\mu^{-1}}\left(\boldsymbol{\Phi}'\left(\boldsymbol{p}^j + \mu^{-1}\boldsymbol{\nu}\right)\right), \tag{3.14}$$

$$\boldsymbol{p}^{j+1} = \begin{bmatrix} \boldsymbol{y} \\ \overline{\Gamma}\left(\boldsymbol{\Phi x}^{j+1} - \mu^{-1}\boldsymbol{\nu}\right) \end{bmatrix} \tag{3.15}$$

converges to an optimal solution of (SP).

Proof. Denote $\mathcal{L}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}, \mu)$ as $\mathcal{L}(\boldsymbol{x}, \boldsymbol{p})$, for simplicity. By the optimality and uniqueness of $\boldsymbol{x}^{j+1}$ and $\boldsymbol{p}^{j+1}$, we have $\mathcal{L}(\boldsymbol{x}^{j+1}, \boldsymbol{p}^{j+1}) \le \mathcal{L}(\boldsymbol{x}^j, \boldsymbol{p}^j)$ and the equality holds if and only if $(\boldsymbol{x}^{j+1}, \boldsymbol{p}^{j+1}) = (\boldsymbol{x}^j, \boldsymbol{p}^j)$. Hence, the sequence $\{\mathcal{L}(\boldsymbol{x}^j, \boldsymbol{p}^j)\}$ is bounded and converges to a constant L^*, i.e., $\mathcal{L}(\boldsymbol{x}^j, \boldsymbol{p}^j) \to L^*$ as $j \to +\infty$. Since the sequence $\{\boldsymbol{x}^j\}$ is also bounded by $\|\boldsymbol{x}^j\|_1 \le \mathcal{L}(\boldsymbol{x}^j, \boldsymbol{p}^j) + \frac{1}{2\mu}\|\boldsymbol{\nu}\|_2^2$, there exists a sub-sequence $\{\boldsymbol{x}^{j_i}\}_{i=1}^{+\infty}$ such that $\boldsymbol{x}^{j_i} \to \boldsymbol{x}_s^*$ as $i \to +\infty$, where $\boldsymbol{x}_s^*$ is an accumulation point of $\{\boldsymbol{x}^j\}$. Correspondingly, $\boldsymbol{p}^{j_i} \to \boldsymbol{p}_s^* = \begin{bmatrix} \boldsymbol{y} \\ \overline{\Gamma}\left(\boldsymbol{\Phi x}^* - \mu^{-1}\boldsymbol{\nu}\right) \end{bmatrix}$ and $\mathcal{L}(\boldsymbol{x}_s^*, \boldsymbol{p}_s^*) = L^*$.

We then show that $(\boldsymbol{x}_s^*, \boldsymbol{p}_s^*)$ is a fixed point of the algorithm. Since it holds that $\boldsymbol{p}_s^* = \arg\min_{\boldsymbol{p},\Gamma(\boldsymbol{p})=\boldsymbol{y}} \mathcal{L}(\boldsymbol{x}_s^*, \boldsymbol{p})$, we need only to show $\boldsymbol{x}_s^* = \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{p}_s^*)$. Suppose there exists $\overline{\boldsymbol{x}}_s \neq \boldsymbol{x}_s^*$ such that $\overline{\boldsymbol{x}}_s = \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{p}_s^*)$. By (3.14), (3.15) and [108, Lemma 2.2], we have $\|\boldsymbol{x}^{j_i+1} - \overline{\boldsymbol{x}}_s\|_2 \leq \|\boldsymbol{x}^{j_i} - \boldsymbol{x}_s^*\|_2 \to 0$, i.e., $\boldsymbol{x}^{j_i+1} \to \overline{\boldsymbol{x}}_s$, as $i \to +\infty$. Meanwhile, $\boldsymbol{p}^{j_i+1} \to \overline{\boldsymbol{p}}_s$. Hence, $\mathcal{L}(\boldsymbol{x}^{j_i+1}, \boldsymbol{p}^{j_i+1}) \to \mathcal{L}(\overline{\boldsymbol{x}}_s, \overline{\boldsymbol{p}}_s) < L^*$, which is because $(\overline{\boldsymbol{x}}_s, \overline{\boldsymbol{p}}_s) \neq (\boldsymbol{x}_s^*, \boldsymbol{p}_s^*)$ and contradicts $\mathcal{L}(\boldsymbol{x}^j, \boldsymbol{p}^j) \to L^*$, resulting in that $(\boldsymbol{x}_s^*, \boldsymbol{p}_s^*)$ is a fixed point. Moreover, it follows from $\|\boldsymbol{x}^{j_i+q} - \boldsymbol{x}_s^*\|_2 \leq \|\boldsymbol{x}^{j_i} - \boldsymbol{x}_s^*\|_2 \to 0$ for any positive integer q, that $\boldsymbol{x}^j \to \boldsymbol{x}_s^*$, as $j \to +\infty$.

Note that orthonormal matrix $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix}$ and $\boldsymbol{\Phi}'\boldsymbol{\Phi} = \boldsymbol{A}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}$. We can obtain

$$\boldsymbol{x}_s^* = S_{\mu^{-1}}\left(\boldsymbol{x}_s^* + \boldsymbol{A}'\left(\boldsymbol{y} + \mu^{-1}\Gamma(\boldsymbol{\nu}) - \boldsymbol{A}\boldsymbol{x}_s^*\right)\right). \tag{3.16}$$

Meanwhile, (SP) is equivalent to

$$\begin{aligned} &\min_{\boldsymbol{x}} \mathcal{L}\left(\boldsymbol{x}, \begin{bmatrix} \boldsymbol{y} \\ \overline{\Gamma}\left(\boldsymbol{\Phi}\boldsymbol{x} - \mu^{-1}\boldsymbol{\nu}\right) \end{bmatrix}\right) \\ &= \|\boldsymbol{x}\|_1 + \frac{\mu}{2}\left\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} - \mu^{-1}\Gamma(\boldsymbol{\nu})\right\|_2^2 - \frac{1}{2\mu}\|\boldsymbol{\nu}\|_2^2. \end{aligned} \tag{3.17}$$

By [108, Proposition 3.10], $\boldsymbol{x}_s^*$ is an optimal solution of the problem in (3.17) and equivalently, $(\boldsymbol{x}_s^*, \boldsymbol{p}_s^*)$ is an optimal solution of (SP). ■

Remark 3.1.

1. *Lemma 1 shows that to solve problem (SP) is equivalent to solve, iteratively, problems (SP_1) and (SP_2).*
2. *Reference [108, Proposition 3.10] only deals with the special case $\|\boldsymbol{A}\|_2 < 1$ and it is, in fact, straightforward to extend the result to arbitrary $\boldsymbol{A}$.*

Following from the framework of the ALM method [106] and Lemma 3.1, the ALM based ONE-L1 algorithm is outlined in Algorithm 3.1, where $(\boldsymbol{x}_t^*, \boldsymbol{p}_t^*)$ is the optimal solution to (SP) in the tth iteration and $\boldsymbol{\nu}_t^*$ is the corresponding Lagrange multiplier. The convergence of Algorithm 3.1 is stated in the following theorem.

Algorithm 3.1: Exact ONE-L1 Algorithm via ALM Method

```
Input: Expanded orthonormal matrix $\boldsymbol{\Phi}$ and observed sample data $\boldsymbol{y}$.
1.  $\boldsymbol{x}_0^* = \boldsymbol{0}$; $\boldsymbol{p}_0^* = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}$; $\boldsymbol{\nu}_0^* = \boldsymbol{0}$; $\mu_0 > 0$; $t = 0$.
2.  while not converged do
3.      Lines 4-9 solve $\left(\boldsymbol{x}_{t+1}^*, \boldsymbol{p}_{t+1}^*\right) = \arg\min_{(\boldsymbol{x},\boldsymbol{p},\Gamma(\boldsymbol{p})=\boldsymbol{y})} \mathcal{L}\left(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\nu}_t^*, \mu_t\right)$;
4.      $\boldsymbol{x}_{t+1}^0 = \boldsymbol{x}_t^*$, $\boldsymbol{p}_{t+1}^0 = \boldsymbol{p}_t^*$, $j = 0$;
5.      while not converged do
6.          $\boldsymbol{x}_{t+1}^{j+1} = S_{\mu_t^{-1}}\left(\boldsymbol{\Phi}'\left(\boldsymbol{p}_{t+1}^{j} + \mu_t^{-1}\boldsymbol{\nu}_t^*\right)\right)$;
7.          $\boldsymbol{p}_{t+1}^{j+1} = \begin{bmatrix} \boldsymbol{y} \\ \overline{\Gamma}\left(\boldsymbol{\Phi}\boldsymbol{x}_{t+1}^{j+1} - \mu_t^{-1}\boldsymbol{\nu}_t^*\right) \end{bmatrix}$;
8.          set $j = j + 1$;
9.      end while
10. $\boldsymbol{\nu}_{t+1}^* = \boldsymbol{\nu}_t^* + \mu_t\left(\boldsymbol{p}_{t+1}^* - \boldsymbol{\Phi}\boldsymbol{x}_{t+1}^*\right)$;
11. choose $\mu_{t+1} > \mu_t$;
12. set $t = t + 1$;
13. end while
Output: $(\boldsymbol{x}_t^*, \boldsymbol{p}_t^*)$.
```

Theorem 3.1. *Any accumulation point $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ of sequence $\{(\boldsymbol{x}_t^*, \boldsymbol{p}_t^*)\}_{t=1}^{+\infty}$ of Algorithm 3.1 is an optimal solution of (BP^o) and the convergence rate with respect to the outer iteration loop index t is at least $O\left(\mu_{t-1}^{-1}\right)$ in the sense that*

$$\left|\|\boldsymbol{x}_t^*\|_1 - \boldsymbol{x}^\dagger\right| = O\left(\mu_{t-1}^{-1}\right),$$

where $\boldsymbol{x}^\dagger = \|\boldsymbol{x}^*\|_1$.

Proof. We first show that the sequence $\{\boldsymbol{\nu}_t^*\}$ is bounded. By the optimality of $\left(\boldsymbol{x}_{t+1}^*, \boldsymbol{p}_{t+1}^*\right)$ we have

$$0 \in \partial_{\boldsymbol{x}}\mathcal{L}\left(\boldsymbol{x}_{t+1}^*, \boldsymbol{p}_{t+1}^*, \boldsymbol{\nu}_t^*, \mu_t\right) = \partial\left\|\boldsymbol{x}_{t+1}^*\right\|_1 - \boldsymbol{\Phi}'\boldsymbol{\nu}_{t+1}^*, \tag{3.18}$$

$$0 = \partial_{\overline{\Gamma}(\boldsymbol{p})}\mathcal{L}\left(\boldsymbol{x}_{t+1}^*, \boldsymbol{p}_{t+1}^*, \boldsymbol{\nu}_t^*, \mu_t\right) = \overline{\Gamma}\left(\boldsymbol{\nu}_{t+1}^*\right), \tag{3.19}$$

where $\partial_{\boldsymbol{x}}$ denotes the partial differential operator with respect to $\boldsymbol{x}$ resulting in a set of subgradients. Hence, $\boldsymbol{\Phi}'\boldsymbol{\nu}_{t+1}^* \in \partial\left\|\boldsymbol{x}_{t+1}^*\right\|_1$. It follows that $\left\|\boldsymbol{\Phi}'\boldsymbol{\nu}_{t+1}^*\right\|_\infty \leq 1$ and

$\{\boldsymbol{\nu}_t^*\}$ is bounded. By $\boldsymbol{x}^\dagger \geq \mathcal{L}\left(\boldsymbol{x}_{t+1}^*, \boldsymbol{p}_{t+1}^*, \boldsymbol{\nu}_t^*, \mu_t\right)$,

$$\begin{aligned} \left\|\boldsymbol{x}_{t+1}^*\right\|_1 &= \mathcal{L}\left(\boldsymbol{x}_{t+1}^*, \boldsymbol{p}_{t+1}^*, \boldsymbol{\nu}_t^*, \mu_t\right) - \frac{1}{2\mu_t}\left(\left\|\boldsymbol{\nu}_{t+1}^*\right\|_2^2 - \left\|\boldsymbol{\nu}_t^*\right\|_2^2\right) \\ &\leq \boldsymbol{x}^\dagger - \frac{1}{2\mu_t}\left(\left\|\boldsymbol{\nu}_{t+1}^*\right\|_2^2 - \left\|\boldsymbol{\nu}_t^*\right\|_2^2\right). \end{aligned} \tag{3.20}$$

Since $\{\boldsymbol{\nu}_t^*\}$ is bounded,

$$\left\|\boldsymbol{x}_{t+1}^*\right\|_1 \leq \boldsymbol{x}^\dagger + O\left(\mu_t^{-1}\right). \tag{3.21}$$

For any accumulation point $\boldsymbol{x}^*$ of $\boldsymbol{x}_t^*$, without loss of generality, we have $\boldsymbol{x}_t^* \to \boldsymbol{x}^*$ as $t \to +\infty$. Hence, $\|\boldsymbol{x}^*\|_1 \leq \boldsymbol{x}^\dagger$. In the mean time, $\boldsymbol{p}_{t+1}^* = \boldsymbol{\Phi}\boldsymbol{x}_{t+1}^* + \mu_t^{-1}\left(\boldsymbol{\nu}_{t+1}^* - \boldsymbol{\nu}_t^*\right) \to \boldsymbol{p}^*$ and $\boldsymbol{\Phi}\boldsymbol{x}^* = \boldsymbol{p}^*$ result in that $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ is an optimal solution to (BP^o).

Moreover, by $\boldsymbol{x}_{t+1}^* = \boldsymbol{\Phi}'\left[\boldsymbol{p}_{t+1}^* - \mu_t^{-1}\left(\boldsymbol{\nu}_{t+1}^* - \boldsymbol{\nu}_t^*\right)\right]$ and

$$\boldsymbol{x}^\dagger = \min_{\boldsymbol{\Phi}\boldsymbol{x}=\boldsymbol{p}, \Gamma(\boldsymbol{p})=\boldsymbol{y}} \|\boldsymbol{x}\|_1 = \min_{\boldsymbol{p}, \Gamma(\boldsymbol{p})=\boldsymbol{y}} \|\boldsymbol{\Phi}'\boldsymbol{p}\|_1 \leq \left\|\boldsymbol{\Phi}'\boldsymbol{p}_{t+1}^*\right\|_1, \tag{3.22}$$

we have $\left\|\boldsymbol{x}_{t+1}^*\right\|_1 \geq \boldsymbol{x}^\dagger - O\left(\mu_t^{-1}\right)$, which establishes the theorem with (3.21). ■

Algorithm 3.1 contains, respectively, an inner and an outer iteration loops. Theorem 3.1 presents only the convergence rate of the outer loop. A natural way to speed up Algorithm 3.1 is to terminate the inner loop without convergence and use the obtained inner-loop solution as the initialization for the next iteration. This is similar to a continuation strategy and can be realized with reasonably set precision and step size μ_t [106]. When the continuation parameter μ_t increases very slowly, in a few iterations, the inner loop can produce a solution with high accuracy. In particular, for the purpose of fast and simple computing, we may update the variables in the inner loop only once before stepping into the outer loop operation. This results in a relaxed version of exact ONE-L1 algorithm (eONE-L1), namely relaxed ONE-L1 algorithm (rONE-L1) outlined in Algorithm 3.2.

Theorem 3.2. *The iterative solution* $(\boldsymbol{x}_t, \boldsymbol{p}_t)$ *of Algorithm 3.2 converges to a feasible solution* $(\boldsymbol{x}^f, \boldsymbol{p}^f)$ *of (BP^o) if* $\sum_{t=1}^{+\infty} \mu_t^{-1} < +\infty$. *It converges at least exponentially to* $(\boldsymbol{x}^f, \boldsymbol{p}^f)$ *if* $\{\mu_t\}$ *is an exponentially increasing sequence.*

Algorithm 3.2: Relaxed ONE-L1 Algorithm

```
Input: Expanded orthonormal matrix $\boldsymbol{\Phi}$ and observed sample data $\boldsymbol{y}$.
$\boldsymbol{x}_0 = \boldsymbol{0}$; $\boldsymbol{p}_0 = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}$; $\boldsymbol{\nu}_0 = \boldsymbol{0}$; $\mu_0 > 0$; $t = 0$.
while not converged do
    $\boldsymbol{x}_{t+1} = S_{\mu_t^{-1}}\left(\boldsymbol{\Phi}'\left(\boldsymbol{p}_t + \mu_t^{-1}\boldsymbol{\nu}_t\right)\right)$;
    $\boldsymbol{p}_{t+1} = \begin{bmatrix} \boldsymbol{y} \\ \overline{\Gamma}\left(\boldsymbol{\Phi}\boldsymbol{x}_{t+1} - \mu_t^{-1}\boldsymbol{\nu}_t\right) \end{bmatrix}$;
    $\boldsymbol{\nu}_{t+1} = \boldsymbol{\nu}_t + \mu_t\left(\boldsymbol{p}_{t+1} - \boldsymbol{\Phi}\boldsymbol{x}_{t+1}\right)$;
choose $\mu_{t+1} > \mu_t$;
set $t = t + 1$;
end while
Output: $(\boldsymbol{x}_t, \boldsymbol{p}_t)$.
```

Proof. We show first that sequences $\{\hat{\boldsymbol{\nu}}_t\}$ and $\{\boldsymbol{\nu}_t\}$ are bounded, where $\hat{\boldsymbol{\nu}}_t = \boldsymbol{\nu}_{t-1} + \mu_{t-1}\left(\boldsymbol{p}_{t-1} - \boldsymbol{\Phi}\boldsymbol{x}_t\right)$. By the optimality of $\boldsymbol{x}_{t+1}$ and $\boldsymbol{p}_{t+1}$ we have

$$0 \in \partial_{\boldsymbol{x}}\mathcal{L}\left(\boldsymbol{x}_{t+1}, \boldsymbol{p}_t, \boldsymbol{\nu}_t, \mu_t\right) = \partial\left\|\boldsymbol{x}_{t+1}\right\|_1 - \boldsymbol{\Phi}'\hat{\boldsymbol{\nu}}_{t+1}, \tag{3.23}$$

$$0 = \partial_{\overline{\Gamma}(\boldsymbol{p})}\mathcal{L}\left(\boldsymbol{x}_{t+1}, \boldsymbol{p}_{t+1}, \boldsymbol{\nu}_t, \mu_t\right) = \overline{\Gamma}\left(\boldsymbol{\nu}_{t+1}\right). \tag{3.24}$$

Hence, $\left\|\boldsymbol{\Phi}'\hat{\boldsymbol{\nu}}_{t+1}\right\|_\infty \leq 1$ and it follows that $\{\hat{\boldsymbol{\nu}}_t\}$ is bounded. Since $\boldsymbol{\nu}_{t+1} = \hat{\boldsymbol{\nu}}_{t+1} + \mu_t\left(\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\right)$, we obtain $\Gamma\left(\boldsymbol{\nu}_{t+1}\right) = \Gamma\left(\hat{\boldsymbol{\nu}}_{t+1}\right)$. This together with $\overline{\Gamma}\left(\boldsymbol{\nu}_{t+1}\right) = 0$ results in $\left\|\boldsymbol{\nu}_{t+1}\right\|_2 \leq \left\|\hat{\boldsymbol{\nu}}_{t+1}\right\|_2$ and the boundedness of $\{\boldsymbol{\nu}_t\}$. By $\boldsymbol{p}_{t+1} - \boldsymbol{p}_t = \mu_t^{-1}\left(\boldsymbol{\nu}_{t+1} - \hat{\boldsymbol{\nu}}_{t+1}\right)$, we have $\left\|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\right\|_2 \leq C\mu_t^{-1}$ with C being a constant. Then $\{\boldsymbol{p}_t\}$ is a Cauchy sequence if $\sum_{t=1}^{+\infty}\mu_t^{-1} < +\infty$, resulting in $\boldsymbol{p}_t \to \boldsymbol{p}^f$ as $t \to +\infty$. In the mean time, $\boldsymbol{x}_t \to \boldsymbol{x}^f$, $\boldsymbol{\Phi}\boldsymbol{x}^f = \boldsymbol{p}^f$. Hence, $\left(\boldsymbol{x}^f, \boldsymbol{p}^f\right)$ is a feasible solution of (BP^o). Suppose that $\{\mu_t\}$ is an exponentially increasing sequence, i.e., $\mu_{t+1} = r\mu_t$ with $r > 1$. By the boundedness of $\{\boldsymbol{\nu}_t\}$ and $\{\hat{\boldsymbol{\nu}}_t\}$ we have

$$\begin{aligned} \left\|\boldsymbol{p}_t - \boldsymbol{p}^f\right\|_2 &= \left\|\sum_{i=t}^{+\infty}\left(\boldsymbol{p}_i - \boldsymbol{p}_{i+1}\right)\right\|_2 \leq \sum_{i=t}^{+\infty}\left\|\boldsymbol{p}_i - \boldsymbol{p}_{i+1}\right\|_2 \\ &\leq C\mu_t^{-1}\sum_{i=0}^{+\infty} r^{-i} = O\left(\mu_t^{-1}\right). \end{aligned} \tag{3.25}$$

Hence, $\{\boldsymbol{p}_t\}$ converges at least exponentially to $\boldsymbol{p}^f$ since $\left\{\mu_t^{-1}\right\}$ exponentially converges to 0, and the same result holds for $\{\boldsymbol{x}_t\}$. ■

Remark 3.2. *It is shown in Theorem 3.2 that faster growth of* $\{\mu_t\}$ *can result in faster convergence of* $\{\boldsymbol{x}_t\}$. *Intuitively, the reduced number of iterations for the inner loop problem* (SP) *may result in some error from the optimal solution* x_t^* *of the inner loop. This will likely affect the accuracy of the final solution* x^f *for BP. Therefore, the growth speed of* $\{\mu_t\}$ *provides a tradeoff between the convergence speed of the algorithm and the precision of the final solution, which will be illustrated in Section 3.3 through numerical simulations.*

3.2.3 Relationship Between rONE-L1 and IST

The studies and applications of IST type algorithms have been very active in recent years because of their concise presentations. This subsection considers the relationship between rONE-L1 and IST. Note that $\overline{\Gamma}(\boldsymbol{\nu}_t) = 0$ in Algorithm 3.2 and $\boldsymbol{\Phi}'\boldsymbol{\Phi} = \boldsymbol{A}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}$. After some derivations, it can be shown that the rONE-L1 algorithm is equivalent to the following iteration (starting from $\boldsymbol{x}_t = 0$, as $t \leq 0$, and $\boldsymbol{z}_t = 0$, as $t < 0$):

$$\begin{aligned} \boldsymbol{x}_{t+1} &= S_{\lambda_t}\left(\boldsymbol{x}_t + \boldsymbol{A}'\boldsymbol{z}_t\right), \\ \boldsymbol{z}_t &= \boldsymbol{y} - \boldsymbol{A}\left[(1+\kappa_t)\,\boldsymbol{x}_t - \kappa_t\boldsymbol{x}_{t-1}\right] + \kappa_t\boldsymbol{z}_{t-1}, \end{aligned} \tag{3.26}$$

where $\lambda_t = \mu_t^{-1}$ and $\kappa_t = \frac{\mu_{t-1}}{\mu_t}$. Compared with the general form of IST in (3.8), one more term $\kappa_t\boldsymbol{z}_{t-1}$ is added when computing the current residual $\boldsymbol{z}_t$ in rONE-L1. Moreover, a weighted sum $(1+\kappa_t)\boldsymbol{x}_t - \kappa_t\boldsymbol{x}_{t-1}$ is used instead of the current solution $\boldsymbol{x}_t$. It will be shown later that these two changes essentially improve the sparsity-undersampling tradeoff.

Remark 3.3. *Equations in (3.26) show that the expansion from the partially orthonormal matrix* $\boldsymbol{A}$ *to orthonormal* $\boldsymbol{\Phi}$ *is not at all involved in the actual implementation and computation of rONE-L1. The same claim also holds for eONE-L1 algorithm. Nevertheless, the orthonormal expansion is a key instrumentation in the derivation and analysis of Algorithms 3.1 and 3.2.*

3.2.4 Implementation Details

As noted in Remark 3.3, the expansion from $\boldsymbol{A}$ to $\boldsymbol{\Phi}$ is not involved in the computing of ONE-L1 algorithms. In our implementations, we consider using exponentially increasing μ_t, i.e., fixing $r > 1$ and $\mu_t = r^t \mu_0$. Let $Q_\alpha(\cdot)$ be an α-quantile operator and $\mu_0 = 1/Q_\alpha\left(|\boldsymbol{A}'\boldsymbol{y}|\right)$, with $|\cdot|$ applying to the vector variable elementwise, μ_0^{-1} being the threshold in the first iteration and $\alpha = 0.99$. In eONE-L1, a large r can speed up the convergence of the outer loop iteration according to Theorem 3.1. However, simulations show that a larger r can result in more iterations in the inner loop. We use $r = 1 + M/N$ as default. In rONE-L1, the value of r provides a tradeoff between the convergence speed of the algorithm and the precision of the final solution. Our recommendation of r to achieve the optimal sparsity-undersampling tradeoff is $r = \min\left(1 + 0.04M/N, 1.02\right)$, which will be illustrated in Section 3.3.1.

An iterative algorithm needs a termination criterion. The eONE-L1 algorithm is considered converged if $\frac{\|\boldsymbol{A}\boldsymbol{x}_t^* - \boldsymbol{y}\|_2}{\|b\|_2} < \tau_1$ with τ_1 being a user-defined tolerance. The inner iteration is considered converged if $\frac{\left\|\boldsymbol{x}_t^{j+1} - \boldsymbol{x}_t^j\right\|_2}{\left\|\boldsymbol{x}_t^j\right\|_2} < \tau_2$. In our implementation, the default values are $(\tau_1, \tau_2) = (10^{-5}, 10^{-6})$. The rONE-L1 algorithm is considered converged if $\frac{\|\boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2} < \tau$, with $\tau = 10^{-5}$ as default.

3.3 Numerical Simulations

3.3.1 Sparsity-Undersampling Tradeoff

This subsection considers the sparsity-undersampling tradeoff of rONE-L1 in the case of strictly sparse signals and noise-free measurements. Phase transition is a measure of the sparsity-undersampling tradeoff in this case. Let the sampling ratio be $\delta = M/N$ and the sparsity ratio be $\rho = K/M$, where K is a measure of sparsity of $\boldsymbol{x}$, and we call that $\boldsymbol{x}$ is K-sparse if at most K entries of $\boldsymbol{x}$ are nonzero. As $K, M, N \to \infty$ with fixed δ and ρ, the behavior of the phase transition of BP is controlled by (δ, ρ) (see more details in Section 2.2). We denote this theoretical curve by $\rho = \rho^R(\delta)$, which is plotted in Fig 3.1.

We estimate the phase transition of rONE-L1 using a Monte Carlo method as in [33, 34]. Two matrix ensembles are considered, including Gaussian with $N = 1000$ and partial-DCT with $N = 1024$. Here the finite-N phase transition is defined as the value of ρ at which the success probability to recover the original signal is 50%. We consider 33 equispaced values of δ in $\{0.02, 0.05, \cdots, 0.98\}$. For each δ, 21 equispaced values of ρ are generated in the interval $\left[\rho^R(\delta) - 0.1, \rho^R(\delta) + 0.1\right]$. Then $M = 20$ random problem instances are generated and solved with respect to each combination of (δ, ρ), where $M = \lceil \delta N \rceil$, $K = \lceil \rho M \rceil$, and nonzero entries of sparse signals are generated from the standard Gaussian distribution. Success is declared if the relative root mean squared error (relative RMSE) $\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}^o\|_2}{\|\boldsymbol{x}^o\|_2} < 10^{-4}$, where $\hat{\boldsymbol{x}}$ is the recovered signal. The number of success among M experiments is recorded. Finally, a generalized linear model is used to estimate the phase transition as in [34].

The experiment result is presented in Fig. 3.1. The observed phase transitions using the recommended value of r strongly agree with the theoretical result of BP. It shows that the rONE-L1 algorithm has the optimal sparsity-undersampling tradeoff in the sense of ℓ_1 minimization.

3.3.2 Comparison with IST

The rONE-L1 algorithm can be considered as a modified version of IST in (3.8). In this subsection we compare the sparsity-undersampling tradeoff and speed of these two algorithms. A similar method is adopted to estimate the phase transition of IST, which is implemented using the same parameter values as rONE-L1. Only nine values of δ in $\{0.1, 0.2, \cdots, 0.9\}$ are considered with the partial-DCT matrix ensemble for time consideration. Another choice of $r = 1 + 0.2\delta$ is considered besides the recommended one. Correspondingly, the phase transition of rONE-L1 with $r = 1 + 0.2\delta$ is also estimated.

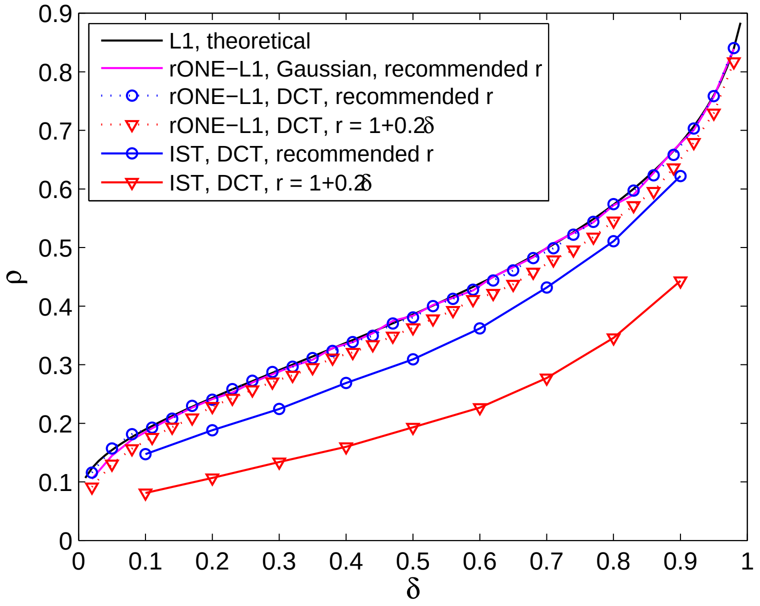


Figure 3.1: Observed phase transitions of rONE-L1, and comparison with those of IST. Notice that, 1) the observed phase transitions of rONE-L1 with the recommended r strongly agree with the theoretical calculation based on BP; 2) rONE-L1 has significantly enlarged success phases compared with IST.

The observed phase transitions are shown in Fig. 3.1. As a modified version of IST, obviously, rONE-L1 makes a great improvement over IST in the sparsity-undersampling tradeoff. Meanwhile, comparison of the averaged number of iterations of the two algorithms shows that rONE-L1 is also faster than IST. For example, as $\delta = 0.2$ and the recommended r is used, rONE-L1 is about 6 times faster than IST.

3.3.3 Comparison with AMP, FPC-AS and NESTA in Noise-free Case

In this subsection, we report numerical simulation results comparing rONE-L1 with state-of-the-art algorithms, including AMP, FPC-AS and NESTA, in the case of sparse signals and noise-free measurements. Our experiments used FPC-AS v.1.21, NESTA v.1.1, and AMP codes provided by the author. We choose parameter values for FPC-AS and NESTA such that each method produces a solution with approximately the same precision as that produced by rONE-L1. In our experiments we consider the recovery of exactly sparse signals from partial-DCT measurements. We set $N = 2^{14}$ and $\delta = 0.2$, and an 'easy' case where $\rho = 0.1$ and a 'hard' case where $\rho = 0.22$ are considered, respectively.[1] Twenty random problems are created and solved for each algorithm with each combination of (δ, ρ), and the minimum, maximum and averaged relative RMSE, number of calls of $\boldsymbol{A}$ and $\boldsymbol{A}'$, and CPU time usages are recorded. All experiments are carried on Matlab v.7.7.0 on a PC with a Windows XP system and a 3GHz CPU. Default parameter values are used in eONE-L1 and rONE-L1.

AMP: terminating if $\frac{\|\boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2} < 10^{-5}$.

FPC-AS: $\lambda = 2 \times 10^{-6}$ and $gtol = 1 \times 10^{-6}$, where $gtol$ is the termination criterion on the maximum norm of sub-gradient. FPC-AS solves the problem QP.

NESTA: $\lambda = 2 \times 10^{-6}$, $\epsilon = 0$ and the termination criterion $tolvar = 1 \times 10^{-8}$. NESTA solves BPDN using the Nesterov algorithm [104], with continuation.

Our experiment results are presented in Table 3.1. In both 'easy' and 'hard' cases, rONE-L1 is much faster than eONE-L1. In the 'easy' case, the proposed rONE-L1 algorithm takes the most number of calls of $\boldsymbol{A}$ and $\boldsymbol{A}'$, except that of eONE-L1, due to a conservative setting of r. But this number of calls (515.4) is very close to that of NESTA (468.9), and furthermore, the CPU time usage of rONE-L1 (2.14 s) is less than that of NESTA (2.70 s) because of its concise implementation. In the 'hard' case, rONE-L1 has the second best performance with significantly less CPU time than that of AMP and NESTA. AMP has the second worst CPU time and the worst accuracy as the dynamic threshold in each iteration depends on the mean squared error of the current iterative solution, which cannot be calculated accurately in the implementation.

[1] Here 'easy' and 'hard' refer to the difficulty degree in recovering a sparse signal from a specific number of measurements. The setting $(\delta, \rho) = (0.2, 0.22)$ is close to the phase transition of BP.

Table 3.1: Comparison Results of ONE-L1 Algorithms with State-of-the-art Methods. The column of "# calls $\boldsymbol{A}$ & $\boldsymbol{A}'$" presents the values of mean(minimum, maximum).

ρ	Method	# calls $\boldsymbol{A}$ & $\boldsymbol{A}'$	CPU time (s)	Error (10^{-5})
0.1	eONE-L1	1819 (1522,2054)	5.62 (4.67,6.52)	0.42 (0.11,0.94)
	rONE-L1	515.4 (286,954)	2.14 (1.19,3.92)	1.08 (0.53,1.30)
	AMP	222.7 (216,234)	0.80 (0.76,0.86)	1.02 (0.85,1.15)
	FPC-AS	150.2 (135,170)	0.50 (0.44,0.56)	1.13 (1.07,1.23)
	NESTA	468.9 (458,484)	2.70 (2.55,2.98)	1.05 (0.99,1.13)
0.22	eONE-L1	9038 (7270,11194)	28.5 (22.0,35.8)	1.87 (0.46,2.66)
	rONE-L1	722.3 (440,972)	2.61 (1.63,3.93)	1.80 (1.37,3.05)
	AMP	1708 (1150,2252)	6.21 (4.19,9.11)	10.5 (6.96,15.8)
	FPC-AS	589.4 (476,803)	2.10 (1.65,2.80)	1.96 (1.46,3.60)
	NESTA	1084 (890,1244)	6.47 (5.22,7.50)	2.90 (1.62,3.98)

3.3.4 Recovery of 2D Image

This subsection demonstrates the efficiency of rONE-L1 in the general CS setting where the signal of interest is approximately sparse and measurements are contaminated with noise. We seek to reconstruct the Mondrian image of size 256 × 256, shown in Fig. 3.2, from its noise-contaminated partial-DCT coefficients. This image presents a challenge as its wavelet expansion is compressible but not exactly sparse. The sampling pattern, which is inspired by magnetic resonance imaging (MRI) and is shown in Fig. 3.2, is adopted since most energy of the image concentrates at low-frequency components after the DCT transform. The measurement vector $\boldsymbol{y}$ contains $M = 7419$ DCT measurements ($\delta = 0.113$). White Gaussian noise with standard deviation $\sigma = 1$ is then added. We set $\epsilon = \sqrt{M + 2\sqrt{2M}}\sigma$. Haar wavelet with a decomposition level 4 is chosen as the sparsifying transform $\mathcal{W}$. Hence, the problem to be solved is BPDN with $\boldsymbol{A} = \mathcal{F}_p\mathcal{W}'$, where $\mathcal{F}_p$ is the partial-DCT transform. The reconstructed image is $\hat{\boldsymbol{H}} = \mathcal{W}'\hat{\boldsymbol{x}}$ with $\hat{\boldsymbol{x}}$ being the reconstructed wavelet coefficients and reconstruction error is calculated as $\frac{\|\hat{\boldsymbol{H}}-\boldsymbol{H}^o\|_{\mathrm{F}}}{\|\boldsymbol{H}^o\|_{\mathrm{F}}}$, where $\boldsymbol{H}^o$ is the original image and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. We compare the performance of rONE-L1 with NESTA and FPC-AS.

Remark 3.4. *AMP is omitted for its poor performance in this approximately-sparse-signal case. For AMP, the value of the dynamic threshold λ_t and the term $\|x_t\|_0$ in*

(3.9) depend on the condition that the signal to reconstruct is strictly sparse.

In such a noisy measurement case, an exact solution for BPDN is not sought after in the rONE-L1 simulation. The computation of the rONE-L1 algorithm is set to terminate if $\frac{\|\boldsymbol{A}\boldsymbol{x}_t-\boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2} \leq \tau = \frac{\epsilon}{\|b\|_2}$, i.e., rONE-L1 outputs the first $\boldsymbol{x}_t$ when it becomes a feasible solution of BPDN.

FPC-AS: $\lambda = 1 \times 10^{-3}$, $gtol = 1 \times 10^{-3}$, $gtol\_scale\_x = 1 \times 10^{-6}$ and the maximum number of iterations for subspace optimization $sub\_mxitr = 10$. The parameters are set according to [107, Section 4.4].

NESTA: $\lambda = 1 \times 10^{-4}$, and $tolvar = 1 \times 10^{-6}$. The parameters are tuned to achieve the minimum reconstruction error.

Fig. 3.2 shows the experiment results where rONE-L1, FPC-AS and NESTA produce faithful reconstructions of the original image. The rONE-L1 algorithm produces a reconstruction error (0.0741) lower than that of FPC-AS (0.0809) with comparable computation times (11.1 s and 11.4 s, respectively). While NESTA results in a slightly lower reconstruction error (0.0649), it incurs about twice more computation time (29.4 s).

3.4 Conclusion

In this chapter, we presented novel algorithms for the sparse signal recovery in CS. The proposed rONE-L1 algorithm which is based on the augmented Lagrange multiplier method and heuristic simplification can be considered as a modified IST with an aggressive continuation strategy. Their performance was demonstrated for 1D sparse signals and 2D images and compared with existing methods. In the next chapter, the ONE-L1 algorithms will be extended to deal with CS problems involving complex valued data and applied to discover a new phase transition in the complex domain.

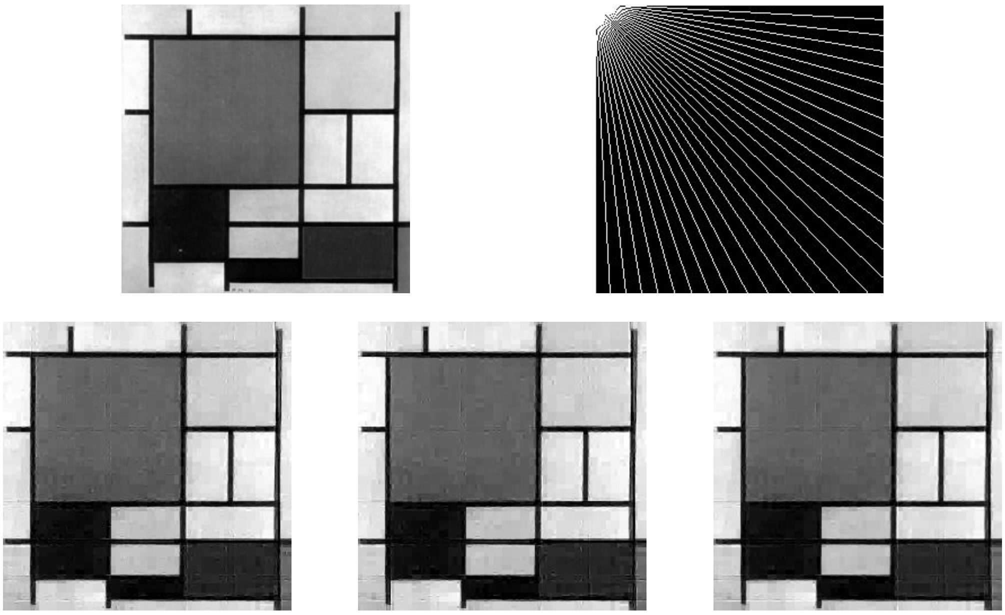


Figure 3.2: An example of 2D image reconstruction from noise-contaminated partial-DCT measurements. Upper left: original Mondrian image; upper right: sampling pattern. The lower three are reconstructed images respectively by rONE-L1 (lower left, error: 0.0741, time: 11.1 s), FPC-AS (lower middle, error: 0.0809, time: 11.4 s) and NESTA (lower right, error: 0.0649, time: 29.4 s).

Chapter 4

Phase Transition in the Complex Domain

As mentioned in Chapter 2, phase transition plays an important role in the CS theory since it describes the exact sparsity-undersampling tradeoff of the ℓ_1 minimization approach under some technical assumptions. The existing phase transition results have been reviewed in Chapter 2 which are focused on the real setting where the sparse signal of interest and the sensing matrix are both real-valued. Since complex data are involved in numerous CS applications, e.g., MRI [36], radar imaging [37] and source localization [11], where both the signal and sensing matrix are complex valued, sparsity-undersampling tradeoff in such a complex setting deserves further studies. For convenience the CS problem in the real and complex settings are referred to as real valued CS (RVCS) and complex valued CS (CVCS) hereafter, respectively. This chapter presents the first result of the phase transition in the complex setting. The significance of our result is twofold: 1) discovering a new phase transition curve for the complex setting so to complement the existing phase transition theory (see Fig. 4.1); 2) providing insights into performance of CS applications involving complex data. Our result is obtained by extending the ONE-L1 algorithms presented in Chapter 3 to the CVCS and applying them to the empirical evaluation of the phase transition of CVCS. Notice that the discovered complex phase transition is

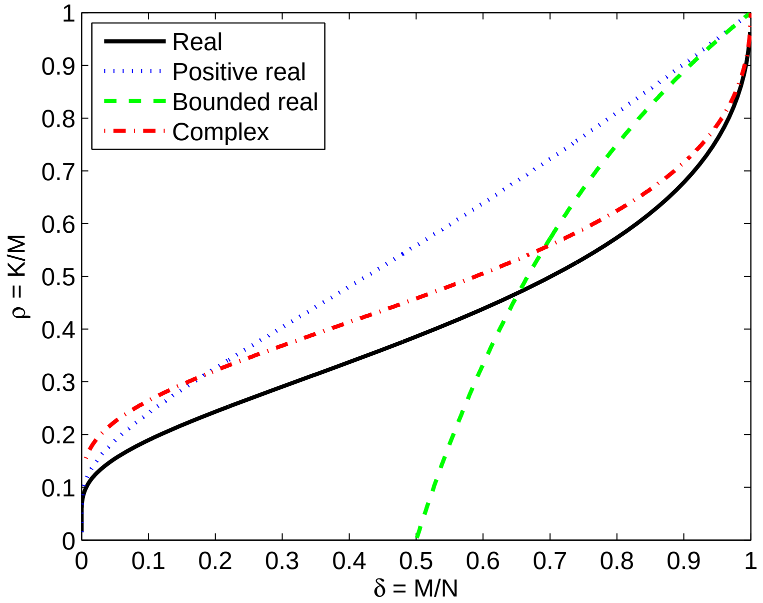


Figure 4.1: Phase transition curves of CS with four different settings. The ones in the real settings are derived in [31–33] (see details in Section 2.2) and the one in the complex setting is discovered due to our work presented in this chapter.

independent of the specific algorithm applied, e.g., ONE-L1 algorithms that we use. However, the computationally efficient ONE-L1 algorithms provide an effective means for empirically exploring the phase transition of CVCS. This chapter is mainly based on [4–6].

4.1 ONE-L1 Algorithms for CVCS

The empirical complex phase transition is evaluated based on the ONE-L1 algorithms presented in Chapter 3. The ONE-L1 algorithms solve the BP problem in RVCS where $\boldsymbol{x}$, $\boldsymbol{A}$ and $\boldsymbol{y}$ are all real valued. Assume that $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$. The BP is firstly transformed into an equivalent form and then the augmented Lagrange multiplier (ALM) method is applied to solve it. The obtained result leads to the exact ONE-L1 (eONE-L1) algorithm for iterative computation of the optimal solution. While eONE-L1 has an inner iteration loop embedded in the outer loop iteration, its relaxed version, rONE-L1, further speeds up the computation by simplifying the inner loop iteration into a single update. The rONE-L1 algorithm is numerically optimal in terms of the sparsity-undersampling tradeoff under reasonable parameter settings, i.e., its phase transition result matches that of the BP. It has been shown that rONE-L1 is of iterative thresholding type and is very fast, with appropriate settings of the regulation variable, in comparison with other state-of-the-art algorithms.

An important operation in the ONE-L1 algorithms is the soft thresholding operator $S_\lambda(\boldsymbol{w})$ defined in (3.4) which solves an ℓ_1 regularized least squares problem (see (3.6)). To extend the ONE-L1 algorithms to CVCS, the complex sensing matrix $\boldsymbol{A} \in \mathbb{C}^{M \times N}$ is also assumed to satisfy $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$ with $'$ denoting the conjugate transpose. The soft thresholding operator $S_\lambda(\boldsymbol{w})$, for a complex vector $\boldsymbol{w}$, is defined the same as in the real case and is the optimal solution to the same ℓ_1 regularized least squares problem. Let $\Re$ and $\Im$ be operators taking, respectively, the real and imaginary parts of a variable. The augmented Lagrangian function in Chapter 3 is modified, for the CVCS, as

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{y}, \mu) = \|\boldsymbol{x}\|_1 + \Re\langle \boldsymbol{d} - \boldsymbol{\Phi}\boldsymbol{x}, \boldsymbol{y}\rangle + \frac{\mu}{2}\|\boldsymbol{d} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2,$$

where $\boldsymbol{y} \in \mathbb{C}^N$ is the Lagrange multiplier vector, $\mu > 0$ and $\langle \boldsymbol{u}_1, \boldsymbol{u}_2 \rangle = \boldsymbol{u}_1'\boldsymbol{u}_2 \in \mathbb{C}$ is the inner product of $\boldsymbol{u}_1$, $\boldsymbol{u}_2 \in \mathbb{C}^N$. The modified augmented Lagrangian function allows a straightforward extension of the derivation and optimization steps in Chapter 3 to the complex BP problem, yielding the optimal solution in the same form as that of the real valued BP. As a result, the ONE-L1 algorithms can be directly applicable to the CVCS problem. Readers are referred to Chapter 3 for detailed algorithm steps and the optimality and convergence analysis.

4.2 Phase Transition of CVCS

4.2.1 Evaluation of Complex Phase Transition via ONE-L1

Section 4.1 has shown that the ONE-L1 algorithms can be extended and applied to the CVCS problem. The eONE-L1 achieves the optimal solution of BP and rONE-L1 is numerically optimal and exponentially converges, which are applied in this subsection to empirically explore the sparsity-undersampling tradeoff of BP. The implementations of ONE-L1 algorithms follow the same procedure as their real versions in Chapter 3. We fix $r > 1$ and let $\mu_{t+1} = r \cdot \mu_t$. The regulation parameter r is set to $r = 1 + \delta$ in eONE-L1 and $r = \min\left(1 + 0.04\delta, 1.02\right)$ is chosen in rONE-L1. The success of recovering the original signal is stated if the relative root mean squared error (RRMSE) $\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}^o\right\|_2 / \left\|\boldsymbol{x}^o\right\|_2 < 10^{-4}$, where $\boldsymbol{x}^o$ and $\hat{\boldsymbol{x}}$ are the original and recovered signals, respectively. Meanwhile, the failure in solving BP using ONE-L1 is stated if $\left\|\hat{\boldsymbol{x}}\right\|_1 \geq \left(1 + 10^{-5}\right)\left\|\boldsymbol{x}^o\right\|_1$ and $\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}^o\right\|_2 / \left\|\boldsymbol{x}^o\right\|_2 \geq 10^{-4}$.

Inspired by the estimation of the real phase transition in [33, 34], we first set a complex matrix ensemble, e.g., Gaussian, and dimension N. A grid of (δ, ρ) is generated in the plane $[0,1] \times [0,1]$ with equispaced $\delta \in \{0.02, 0.05, \cdots, 0.98\}$ and $\rho \in \left\{\rho^R\left(\delta\right) + 0.01(i - 21) : i = 1, 2, \cdots, 41\right\}$ with respect to δ, where $\rho^R\left(\delta\right)$ denotes the theoretical real phase transition as shown in Fig. 4.3. For each combination of (δ, ρ), $M = 20$ random problem instances are generated and solved with $M = \lceil \delta N \rceil$ and $K = \lceil \rho M \rceil$. The number of success S among M instances is recorded. After data acquisition, a generalized linear modal (GLM) with a logistic link is used to estimate the phase transition.

We now explore the sparsity-undersampling tradeoff of BP with partial-Fourier sampling. Four values of signal length N are considered, including 1024, 2048, 4096 and 8192. When $N = 1024$, both eONE-L1 and rONE-L1 are used to estimate the phase transition of BP. The rONE-L1 algorithm is used for other N. Few failures in solving the BP occur when using rONE-L1 (see Table 4.1).

Table 4.1: Number of Failures and CPU Time Consumptions. The "Number of failures" column presents #failures(#problem instances solved).

Ensemble	N	Method	Number of failures			CPU time
			$\delta = 0.02$	$\delta = 0.05$	Others	(hours)
Fourier	1024	exact	0(480)	0(340)	0(8180)	6.26
Fourier	1024	relaxed	14(440)	4(340)	0(8280)	0.327
Fourier	2048	relaxed	1(320)	1(320)	0(7460)	0.594
Fourier	4096	relaxed	4(240)	0(260)	0(6980)	1.01
Fourier	8192	relaxed	1(260)	0(240)	0(6660)	2.15
Gaussian	1000	relaxed	16(420)	0(340)	0(8240)	5.56
Bernoulli	1000	relaxed	9(420)	2(360)	0(8340)	5.60
Ternary	1000	relaxed	9(360)	2(280)	0(8240)	5.65

The observed success rates of our experiments are shown in Fig. 4.2, where the phase-transition performance can be observed. The phase transitions occur at about the same location and larger signal length N leads to sharper phase transition, which is consistent with the behavior of the real phase transition. Fig. 4.3 presents the estimated phase transitions of partial-Fourier sampling. The five observed phase transitions of BP, estimated respectively by eONE-L1 with $N = 1024$ and rONE-L1 with $N = 1024$, 2048, 4096 and 8192, coincide with each other and are higher than the real phase transition of BP. To sum up the observations of Figs. 4.2 and 4.3 we can state:

Finding 4.1. *For complex signals and the partial-Fourier matrix ensemble with large dimension N, we observe that*

I. *BP exhibits phase transition in the plane of (δ, ρ), and a larger N can result in a sharper phase transition.*

II. *the complex phase transition of BP is higher than the real phase transition with a considerably enlarged success phase.*

4.2.2 Universality of Phase Transitions of CVCS

The observed universality of phase transitions of BP across different matrix ensembles in the real case has been studied in [34] and the same result is stated in [33] and

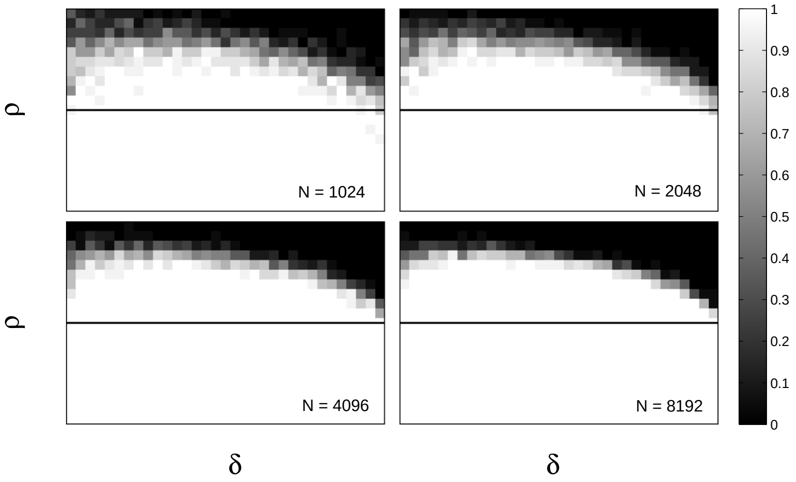


Figure 4.2: Observed success rates of partial-Fourier sampling. Four images show results of rONE-L1 with $N = 1024$, 2048, 4096 and 8192, respectively. Each pixel refers to the success rate S/M at the corresponding coordinate (δ, ρ). Sampling ratio δ ranges from 0.02 (left) to 0.98 (right) with an interval 0.03; Sparsity ratio ρ ranges from $\rho^R(\delta) - 0.1$ (bottom) to $\rho^R(\delta) + 0.1$ (top) with an interval 0.01 with respect to each δ (only part of the experiment results are shown). The middle line in each image refers to the real phase transition.

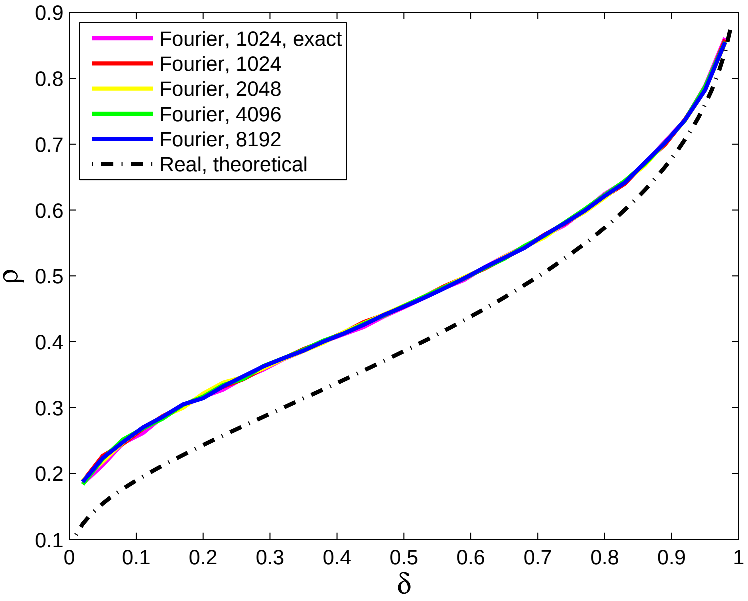


Figure 4.3: Observed phase transitions of partial-Fourier sampling. The upper five curves are observed phase transitions estimated by eONE-L1 with $N = 1024$ and rONE-L1 with $N = 1024$, 2048, 4096 and 8192, respectively. The lower is the theoretical real phase transition of BP.

Chapter 3. In this subsection, we examine whether the same property holds in the complex case. Apart from the partial-Fourier matrix ensemble, three other complex matrix ensembles are considered with signal length $N = 1000$, including Gaussian, Bernoulli and Ternary. All random matrices have i.i.d. real and imaginary parts following the corresponding distributions. Bernoulli refers to equally likely being 0 or 1, and Ternary is equally likely to be -1, 0 or 1. Notice that a matrix generated from these matrix ensembles may not be partially orthonormal as required by the ONE-L1 algorithms. This problem can be resolved by left multiplying both sides of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ with an invertible matrix, e.g., using QR decomposition, so the transformed equation meets the partially orthonormal condition and preserves the same solution space.

Few failures are observed in our experiment when using rONE-L1 to solve BP (see Table 4.1). Fig. 4.4 presents the observed phase transitions of BP with different matrix ensembles. Like the universality of phase transitions in the real case, we have the following finding.

Finding 4.2. *For complex signals and a number of complex matrix ensembles with large dimension N, BP exhibits the same phase transition.*

4.2.3 Connection to Block-Sparse CS

The complex system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ can be rewritten into a real valued formulation as $\boldsymbol{A}_r\boldsymbol{x}_r = \boldsymbol{y}_r$ with $\boldsymbol{A}_r = \begin{bmatrix} \Re\boldsymbol{A} & -\Im\boldsymbol{A} \\ \Im\boldsymbol{A} & \Re\boldsymbol{A} \end{bmatrix}$, $\boldsymbol{x}_r = \begin{bmatrix} \Re\boldsymbol{x} \\ \Im\boldsymbol{x} \end{bmatrix}$ and $\boldsymbol{y}_r = \begin{bmatrix} \Re\boldsymbol{y} \\ \Im\boldsymbol{y} \end{bmatrix}$. By the K-sparsity of $\boldsymbol{x}$, $\Re\boldsymbol{x}$ and $\Im\boldsymbol{x}$ are jointly K-sparse in the sense that they are both K-sparse and share the same support. On the other hand, after proper permutations of entries of $\boldsymbol{x}_r$ as well as the corresponding columns of $\boldsymbol{A}_r$, $\boldsymbol{A}_r\boldsymbol{x}_r = \boldsymbol{y}_r$ can be recast into a block-sparse CS (BSCS) problem [111] with block size 2. The blocked signal of $\boldsymbol{x}_r$ is K-block-sparse with at most K nonzero block entries and its $\ell_{2,1}$-norm as defined in [111] is equivalent to the ℓ_1 norm of $\boldsymbol{x}$ in the CVCS. Thus

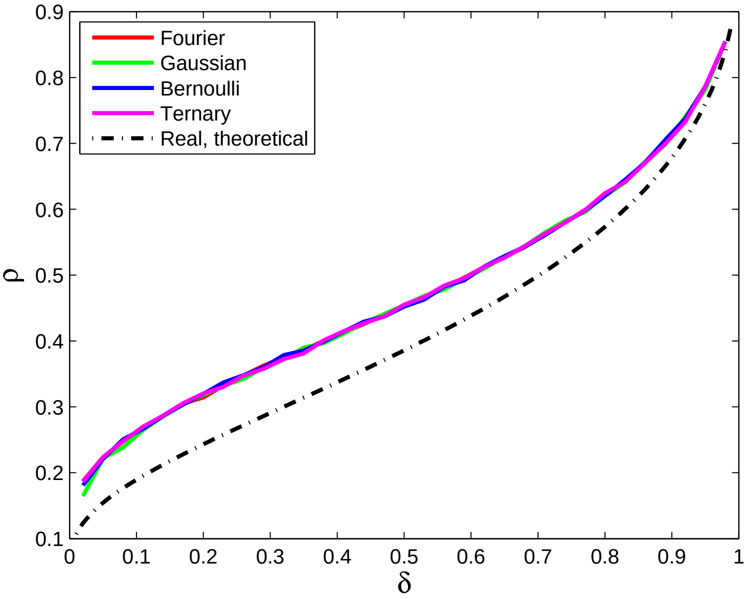


Figure 4.4: Observed universality of phase transitions of BP in the complex domain. The signal length $N = 8192$ for partial-Fourier matrix ensemble and $N = 1000$ for the other three matrix ensembles.

the CVCS problem is strongly connected to the BSCS problem via the real valued reformation $\boldsymbol{A}_r\boldsymbol{x}_r = \boldsymbol{y}_r$. Their only difference is that $\boldsymbol{A}_r$ is subject to a structured constraint whereas entries of the sensing matrix of the BSCS problem are assumed to be independent in general.

A comparison of the phase transition of CVCS with the theoretical phase transition of BSCS with block size 2 in [111] is presented in Fig. 4.5. It is shown that the observed phase transition of CVCS coincides with that of BSCS.

In [111], the theoretical phase transition of BSCS is derived under the condition that the null space of the real valued random sensing matrix is distributed uniformly in the Grassmanian with respect to the Haar measure. It is known that the real Gaussian matrix ensemble satisfies such a condition [112]. We now provide an analysis in the following proposition to show that such a condition is not satisfied in the CVCS problem. It therefore clarifies that the phase transition of the CVCS studied in this chapter is not a special case of the phase transition result of BSCS in [111].

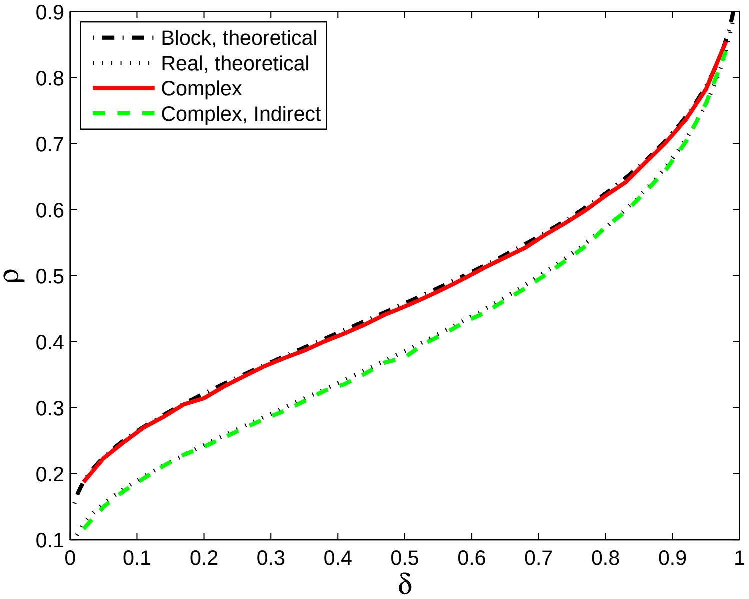


Figure 4.5: Coincidence between our observed complex phase transition and the block-sparse phase transition with block size 2. The indirect phase transition curve refers to minimizing $\|\boldsymbol{x}_r\|_1$ subject to $\boldsymbol{Ax} = \boldsymbol{y}$.

Proposition 4.1. *For any complex random matrix* $\boldsymbol{A}$ *associated with the CVCS problem, the null space of* $\boldsymbol{A}_r$ *is not distributed uniformly in the Grassmanian with respect to the Haar measure.*

Proof. Let $\mathcal{N}(\boldsymbol{A}_r)$ denote the null space of $\boldsymbol{A}_r$. Suppose $\begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{v} \end{bmatrix} \in \mathcal{N}(\boldsymbol{A}_r)$ with $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^N$. It is obvious that $\begin{bmatrix} -\boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} \in \mathcal{N}(\boldsymbol{A}_r)$ and thus it is always possible to choose a basis for $\mathcal{N}(\boldsymbol{A}_r)$ in the form $\begin{bmatrix} \boldsymbol{W} & -\boldsymbol{V} \\ \boldsymbol{V} & \boldsymbol{W} \end{bmatrix}$ with $\boldsymbol{W}, \boldsymbol{V} \in \mathbb{R}^{N\times(N-M)}$ (columns of $\boldsymbol{W} + j\boldsymbol{V} \in \mathbb{C}^{N\times(N-M)}$ in fact compose a basis for the null space of $\boldsymbol{A}$). As a result, $\mathcal{N}(\boldsymbol{A}_r)$ is not distributed uniformly in the Grassmanian. ■

Proposition 4.1 shows that the existing phase transition theory of BSCS cannot explain our obtained empirical phase transition of the CVCS. Intuitively, the coincidence of the phase transition of CVCS with that of BSCS may be interpreted as the universality of phase transitions of BSCS across different matrix ensembles. Fig. 4.5 also shows that the observed phase transition by minimizing $\|\boldsymbol{x}_r\|_1$, subject to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ and without taking into account the block-sparsity, matches the real phase transition of BP. It further explains that the higher successful rate of CVCS is obtained by incorporating the block-sparsity factor into its problem solving.

In the above, it has been shown that the CVCS problem is connected to BSCS with a special structure of the sensing matrix $\boldsymbol{A}_r$. It is interesting to consider that the sensing matrix in CVCS is further specialized into a block-diagonal matrix, i.e., $\Im\boldsymbol{A} = \boldsymbol{0}$ in $\boldsymbol{A}_r$ or, equivalently, $\boldsymbol{A} \in \mathbb{R}^{M\times N}$. Such a case refers to the jointly sparse CS or the case of multiple measurement vectors (MMV). In comparison with the standard real CS or the case of single measurement vector (SMV), the recovery performance can be improved in the MMV case under some assumptions on the distribution of entries of $\boldsymbol{x}$ [113]. On the other hand, the analysis of the null space property in [114] shows that there is little performance improvement in the worst case of signal recovery, such as an $\boldsymbol{x}$ with identical real and imaginary parts. So, in this special case of jointly sparse CS, the observed universality of phase transitions of BSCS does not hold.

4.3 Rigorous Derivation of the Complex Phase Transition

After we obtained the empirical results of the phase transition presented above, a rigorous analysis has been carried out in [6] with collaborators. In our method, we extend the approximate message passing (AMP) algorithm originally proposed in [33] for RVCS to solve complex valued BP and obtain the complex AMP algorithm (CAMP). We then generalize the state evolution framework which was introduced in [33] for the analysis of AMP, to the complex setting. Using the state evolution, we derive accurate formulae for the phase transition of both BP and CAMP. Our results are theoretically proven for the case of i.i.d. Gaussian sensing matrices and coincide with our empirical observation in this chapter. In the extreme undersampling regime it holds that $\rho^C(\delta) \sim [\ln(1/2\delta)]^{-1}$, $\delta \to 0$, where $\rho^C(\delta)$ denotes the complex phase transition curve as a function of δ. As a result, $\lim_{\delta\to 0} \frac{\rho^C(\delta)}{\rho^R(\delta)} = 2$ by recalling that $\rho^R(\delta) \sim [2\ln(1/\delta)]^{-1}$ as presented in Section 2.2, where $\rho^R(\delta)$ denotes the real phase transition of the BP. This means that, in the extreme undersampling regime the complex BP can recover signals that are two times denser than the signals that are recovered by the real BP.

4.4 Conclusion

In this chapter, a new phase transition of BP for CS is discovered in the complex setting, which is positioned well above the real one with similar properties. Its connection to the existing result of block-sparse CS was studied and its theoretical derivation was discussed. The new phase transition complements the existing phase transition theory in the real setting and provides insights into performance of CS applications involving complex data.

Chapter 5

Robustly Stable Signal Recovery with Structured Matrix Perturbation

The sparsity-undersampling tradeoff of CS has been studied in Chapter 4 in terms of phase transition. Notice that the sensing matrix is assumed to be exactly known there. However, this is not always the case in practice as discussed in Subsection 2.4.2. This chapter is concerned with the latter case where perturbations (or uncertainties) exist in the sensing matrix. In standard formulation of CS, the signal $\boldsymbol{x}^o \in \mathbb{R}^N$ of interest is acquired by the linear measurements

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^o + \boldsymbol{e}, \tag{5.1}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$ is the sensing matrix (or linear operator), typically with $M \ll N$, $\boldsymbol{y} \in \mathbb{R}^M$ is the vector of measurements, and $\boldsymbol{e} \in \mathbb{R}^M$ denotes the vector of measurement noises with bounded energy, i.e., $\left\|\boldsymbol{e}\right\|_2 \leq \epsilon$ for $\epsilon > 0$. Given $\boldsymbol{\Phi}$ and ϵ, the task of CS is to recover $\boldsymbol{x}^o$ from a significantly reduced number of measurements $\boldsymbol{y}$. Candès *et al.* [17, 58] show that if $\boldsymbol{x}^o$ is sparse, then it can be stably recovered under mild conditions on $\boldsymbol{\Phi}$ with the recovery error being at most proportional to the measurement noise level ϵ by solving an ℓ_1 minimization problem. Similarly, the largest (in amplitude) entries of a compressible signal can be stably recovered. More details are presented in Subsection 5.1.2. There have been recent active studies on the CS problem where the sensing matrix $\mathbf{\Phi}$ is unknown or subject to an unknown perturbation. A list of publications include [73] on unknown sensing matrix, [74–76] on general matrix perturbations and [78] considering also structures in perturbation.

This chapter is focused on the perturbed CS problem and mainly based on [7]. We study a structured matrix perturbation with each column of the perturbation matrix being a (unknown) constant times a (known) vector which defines the direction of perturbation. For certain structured matrix perturbation, we provide conditions for guaranteed signal recovery performance. Our analysis shows that robust stability (see definition in Subsection 5.1.1) can be achieved for a sparse signal under similar mild conditions as those for the standard CS problem by solving an ℓ_1 minimization problem incorporated with the perturbation structure. In the special noise free case, the recovery is exact for a sufficiently sparse signal with respect to the perturbation level. A similar result holds for a compressible signal under an additional assumption of small perturbation (depending on the number of largest entries to be recovered). A practical application problem of DOA estimation will be studied in Chapter 8, where the off-grid DOA estimation problem is formulated into our proposed signal recovery problem subject to the structured sensing matrix perturbation, showing the practical relevance of our proposed framework and solution. To verify the obtained results, two algorithms for positive-valued and general signals respectively are proposed to solve the resulting nonconvex ℓ_1 minimization problem. Numerical simulations confirm our robustly stable signal recovery results.

5.1 Preliminary Results

5.1.1 Definitions

In CS, a signal $\boldsymbol{x}^o \in \mathbb{R}^N$ of length N is called K-sparse if it has at most K nonzero entries, and it is called compressible if its entries obey a power law

$$|x^o|_{(j)} \leq C_q j^{-q}, \tag{5.2}$$

where $|x^o|_{(j)}$ is the jth largest entry (in amplitude) of $\boldsymbol{x}^o$ ($|x^o|_{(1)} \geq |x^o|_{(2)} \geq \cdots \geq |x^o|_{(N)}$), $q > 1$ and C_q is a constant that depends only on q. Let $\boldsymbol{x}^K$ be a vector that keeps the K largest entries (in amplitude) of $\boldsymbol{x}^o$ with the rest being zero. If $\boldsymbol{x}^o$ is compressible, then it can be well approximated by the sparse signal $\boldsymbol{x}^K$ in the sense that

$$\left\|\boldsymbol{x}^o - \boldsymbol{x}^K\right\|_2 \leq C'_q K^{-q+1/2} \tag{5.3}$$

where C'_q is a constant.

For the purpose of clarification of expression, we define formally some terminologies for signal recovery used in this chapter, including stability in standard CS, robustness and robust stability in perturbed CS.

Definition 5.1 ([58]). *In standard CS where $\boldsymbol{\Phi}$ is known* a priori*, consider a recovered signal $\widehat{\boldsymbol{x}}$ of $\boldsymbol{x}^o$ from measurements $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^o + \boldsymbol{e}$ with $\|\boldsymbol{e}\|_2 \leq \epsilon$. We say that $\widehat{\boldsymbol{x}}$ achieves stable signal recovery if*

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^o\|_2 \leq C_1^{stb} K^{-q+1/2} + C_2^{stb}\epsilon$$

holds for compressible signal $\boldsymbol{x}^o$ obeying (5.2) and an integer K, or if

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^o\|_2 \leq C_2^{stb}\epsilon$$

holds for K-sparse signal $\boldsymbol{x}^o$, with nonnegative constants C_1^{stb}, C_2^{stb}.

Definition 5.2. *In perturbed CS where $\boldsymbol{\Phi} = \boldsymbol{A} + \boldsymbol{E}$ with $\boldsymbol{A}$ known* a priori *and $\boldsymbol{E}$ unknown with $\|\boldsymbol{E}\|_F \leq \eta$, consider a recovered signal $\widehat{\boldsymbol{x}}$ of $\boldsymbol{x}^o$ from measurements $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^o + \boldsymbol{e}$ with $\|\boldsymbol{e}\|_2 \leq \epsilon$. We say that $\widehat{\boldsymbol{x}}$ achieves robust signal recovery if*

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^o\|_2 \leq C_1^{rbt} K^{-q+1/2} + C_2^{rbt}\epsilon + C_3^{rbt}\eta$$

holds for compressible signal $\boldsymbol{x}^o$ obeying (5.2) and an integer K, or if

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^o\|_2 \leq C_2^{rbt}\epsilon + C_3^{rbt}\eta$$

holds for K-sparse signal $\boldsymbol{x}^o$, with nonnegative constants C_1^{rbt}, C_2^{rbt} and C_3^{rbt}.

Definition 5.3. *In perturbed CS where $\boldsymbol{\Phi} = \boldsymbol{A} + \boldsymbol{E}$ with $\boldsymbol{A}$ known* a priori *and $\boldsymbol{E}$ unknown with $\|\boldsymbol{E}\|_F \leq \eta$, consider a recovered signal $\widehat{\boldsymbol{x}}$ of $\boldsymbol{x}^o$ from measurements $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^o + \boldsymbol{e}$ with $\|\boldsymbol{e}\|_2 \leq \epsilon$. We say that $\widehat{\boldsymbol{x}}$ achieves robustly stable signal recovery if*

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^o\|_2 \leq C_1^{rs}(\eta) K^{-q+1/2} + C_2^{rs}(\eta)\epsilon$$

holds for compressible signal $\boldsymbol{x}^o$ obeying (5.2) and an integer K, or if

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^o\|_2 \leq C_2^{rs}(\eta)\epsilon$$

holds for K-sparse signal $\boldsymbol{x}^o$, with nonnegative constants C_1^{rs}, C_2^{rs} depending on η.

Remark 5.1.

(1) In the case where $\boldsymbol{x}^o$ is compressible, the defined stable, robust, or robustly stable signal recovery is in fact for its K largest (in amplitude) entries. The first term $O\left(K^{-q+1/2}\right)$ in the error bounds above represents, by (5.3), the best approximation error (up to a scale) that can be achieved when we know everything about $\boldsymbol{x}^o$ and select its K largest entries.

(2) The Frobenius norm of $\boldsymbol{E}$, $\|\boldsymbol{E}\|_F$, can be replaced by any other norm in Definitions 5.2 and 5.3 since the norms are equivalent.

(3) By robust stability, we mean that the signal recovery is stable for any fixed matrix perturbation level η according to Definition 5.3.

It should be noted that the stable recovery in standard CS and the robustly stable recovery in perturbed CS are exact in the noise free, sparse signal case while there is no such a guarantee for the robust recovery in perturbed CS.

5.1.2 Stable Signal Recovery of Standard CS

The task of standard CS is to recover the original signal $\boldsymbol{x}^o$ via an efficient approach given the sensing matrix $\boldsymbol{\Phi}$, acquired sample $\boldsymbol{y}$ and upper bound ϵ for the measurement noise. This chapter is focused on the ℓ_1 norm minimization approach. The restricted isometry property (RIP) (see Definition 2.2 of Chapter 2) has become a dominant tool to such analysis. Based on the RIP, the following theorem holds.

Theorem 5.1 ([17]). *Assume that $\delta_{2K}(\boldsymbol{\Phi}) < \sqrt{2}-1$ and $\|\boldsymbol{e}\|_2 \leq \epsilon$. Then an optimal solution $\boldsymbol{x}^*$ to the basis pursuit denoising (BPDN) problem*

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1, \text{ subject to } \|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \epsilon \tag{5.4}$$

satisfies

$$\|\boldsymbol{x}^*-\boldsymbol{x}^o\|_2 \leq C_0^{std} K^{-1/2} \left\|\boldsymbol{x}^o-\boldsymbol{x}^K\right\|_1 + C_1^{std}\epsilon \tag{5.5}$$

where $C_0^{std} = \frac{2\left[1+\left(\sqrt{2}-1\right)\delta_{2K}(\boldsymbol{\Phi})\right]}{1-\left(\sqrt{2}+1\right)\delta_{2K}(\boldsymbol{\Phi})}$, $C_1^{std} = \frac{4\sqrt{1+\delta_{2K}(\boldsymbol{\Phi})}}{1-\left(\sqrt{2}+1\right)\delta_{2K}(\boldsymbol{\Phi})}$.

Theorem 5.1 states that a K-sparse signal $\boldsymbol{x}^o$ ($\boldsymbol{x}^K = \boldsymbol{x}^o$) can be stably recovered by solving a computationally efficient convex optimization problem provided $\delta_{2K}(\boldsymbol{\Phi}) < \sqrt{2}-1$. The same conclusion holds in the case of compressible signal $\boldsymbol{x}^o$ since

$$K^{-1/2}\left\|\boldsymbol{x}^o-\boldsymbol{x}^K\right\|_1 \leq C_q'' k^{-q+1/2} \tag{5.6}$$

according to (5.2) and (5.3) with C_q'' being a constant. In the special noise free, K-sparse signal case, such a recovery is exact. The RIP condition in Theorem 5.1 can be satisfied provided $M \geq O\left(K \log\left(N/K\right)\right)$ with a large probability if the sensing matrix $\boldsymbol{\Phi}$ is i.i.d. subgaussian distributed [115]. Note that the RIP condition for the stable signal recovery in standard CS has been relaxed in [116,117] but it is beyond the scope of this chapter.

5.1.3 Robust Signal Recovery in Perturbed CS

In standard CS, the sensing matrix $\boldsymbol{\Phi}$ is assumed to be exactly known. Such an ideal assumption is not always the case in practice. Consider that the true sensing matrix is $\boldsymbol{\Phi} = \boldsymbol{A} + \boldsymbol{E}$ where $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ is the known nominal sensing matrix and $\boldsymbol{E} \in \mathbb{R}^{M\times N}$ represents the unknown matrix perturbation. Unlike the additive noise term $\boldsymbol{e}$ in the observation model in (5.1), a multiplicative "noise" $\boldsymbol{E}\boldsymbol{x}^o$ is introduced in perturbed CS and is more difficult to analyze since it is correlated with the signal of interest. Denote $\left\|\boldsymbol{E}\right\|_2^{(K)}$ the largest spectral norm taken over all K-column submatrices of $\boldsymbol{E}$, and similarly define $\left\|\boldsymbol{\Phi}\right\|_2^{(K)}$. The following theorem is stated in [74].

Theorem 5.2 ([74]). *Assume that there exist constants* $\varepsilon_{\boldsymbol{E},\boldsymbol{\Phi}}^{(K)}$, ϵ *and* $\epsilon_{\boldsymbol{E},\boldsymbol{x}^o}$ *such that* $\frac{\left\|\boldsymbol{E}\right\|_2^{(K)}}{\left\|\boldsymbol{\Phi}\right\|_2^{(K)}} \leq \varepsilon_{\boldsymbol{E},\boldsymbol{\Phi}}^{(K)}$, $\left\|\boldsymbol{e}\right\|_2 \leq \epsilon$ *and* $\left\|\boldsymbol{E}\boldsymbol{x}^o\right\|_2 \leq \epsilon_{\boldsymbol{E},\boldsymbol{x}^o}$. *Assume that* $\delta_{2K}\left(\boldsymbol{\Phi}\right) < \frac{\sqrt{2}}{\left(1+\varepsilon_{\boldsymbol{E},\boldsymbol{\Phi}}^{(2K)}\right)^2} - 1$ *and* $\left\|\boldsymbol{x}^o\right\|_0 \leq K$. *Then an optimal solution* $\boldsymbol{x}^*$ *to the BPDN problem with the nominal sensing matrix* $\boldsymbol{A}$, *denoted by N-BPDN,*

$$\min_{\boldsymbol{x}} \left\|\boldsymbol{x}\right\|_1, \ \textit{subject to} \ \left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right\|_2 \leq \epsilon + \epsilon_{\boldsymbol{E},\boldsymbol{x}^o} \tag{5.7}$$

achieves robust signal recovery with

$$\left\|\boldsymbol{x}^* - \boldsymbol{x}^o\right\|_2 \leq C^{ptb}\epsilon + C^{ptb}\epsilon_{\boldsymbol{E},\boldsymbol{x}^o} \tag{5.8}$$

where $C^{ptb} = \frac{4\sqrt{1+\delta_{2K}(\boldsymbol{\Phi})}\left(1+\varepsilon_{\boldsymbol{E},\boldsymbol{\Phi}}^{(2K)}\right)}{1-\left(\sqrt{2}+1\right)\left[(1+\delta_{2K}(\boldsymbol{\Phi}))\left(1+\varepsilon_{\boldsymbol{E},\boldsymbol{\Phi}}^{(2K)}\right)^2-1\right]}$.

Remark 5.2.

(1) *The relaxation of the inequality constraint in (5.7) from ϵ to $\epsilon + \epsilon_{\boldsymbol{E},\boldsymbol{x}^o}$ is to ensure that the original signal $\boldsymbol{x}^o$ is a feasible solution to N-BPDN. Theorem 5.2 is a little different from that in [74], where the multiplicative "noise" $\boldsymbol{E}\boldsymbol{x}^o$ is bounded using $\varepsilon_{\boldsymbol{E},\boldsymbol{\Phi}}^{(K)}$, $\delta_K(\boldsymbol{\Phi})$ and $\|\boldsymbol{\Phi}\boldsymbol{x}^o\|_2$ rather than a constant $\epsilon_{\boldsymbol{E},\boldsymbol{x}^o}$.*

(2) *Theorem 5.2 is applicable only to the small perturbation case where $\varepsilon_{\boldsymbol{E},\boldsymbol{\Phi}}^{(2K)} < \sqrt[4]{2} - 1$ since $\delta_{2K}(\boldsymbol{\Phi}) \geq 0$.*

(3) *Theorem 5.2 generalizes Theorem 5.1 for the K-sparse signal case. As the perturbation $\boldsymbol{E} \to 0$, Theorem 5.2 coincides with Theorem 5.1 for the K-sparse signal case.*

Theorem 5.2 states that, for a small matrix perturbation $\boldsymbol{E}$, the signal recovery of N-BPDN that is based on the nominal sensing matrix $\boldsymbol{A}$ is robust to the perturbation with the recovery error growing at most linearly with the perturbation level. Note that, in general, the signal recovery in Theorem 5.2 is unstable according to the definition of stability in this chapter since the recovery error cannot be bounded within a constant (independent of the noise) times the noise level as some perturbation occurs. A result on general signals in [74] is omitted that shows the robust recovery of a compressible signal. The same problem is studied and similar results are reported in [75] based on the greedy algorithm CoSaMP [21].

5.2 SP-CS: CS Subject to Structured Perturbation

5.2.1 Problem Description

In this chapter we consider a structured perturbation in the form $\boldsymbol{E} = \boldsymbol{B}\boldsymbol{\Delta}^o$ where $\boldsymbol{B} \in \mathbb{R}^{M \times N}$ is known *a priori*, $\boldsymbol{\Delta}^o = \operatorname{diag}(\boldsymbol{\beta}^o)$ is a bounded uncertain term with $\boldsymbol{\beta}^o \in [-r, r]^N$ and $r > 0$, i.e., each column of the perturbation is on a known direction. In addition, we assume that each column of $\boldsymbol{B}$ has unit norm to avoid the scaling problem between $\boldsymbol{B}$ and $\boldsymbol{\Delta}^o$ (in fact, the D-RIP condition on matrix $[\boldsymbol{A}, \boldsymbol{B}]$ in Subsection 5.2.2 implies that columns of both $\boldsymbol{A}$ and $\boldsymbol{B}$ have approximately unit norms). As a result, the observation model in (5.1) becomes

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^o + \boldsymbol{e}, \quad \boldsymbol{\Phi} = \boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^o \tag{5.9}$$

with $\boldsymbol{\Delta}^o = \operatorname{diag}(\boldsymbol{\beta}^o)$, $\boldsymbol{\beta}^o \in [-r, r]^N$ and $\|\boldsymbol{e}\|_2 \leq \epsilon$. Given $\boldsymbol{y}$, $\boldsymbol{A}$, $\boldsymbol{B}$, r and ϵ, the task of SP-CS is to recover $\boldsymbol{x}^o$ and possibly $\boldsymbol{\beta}^o$ as well.

Remark 5.3.

(1) The perturbed CS model in (5.9) is inspired by the off-grid DOA estimation problem that will be studied in Chapter 8. While this chapter is focused on the theoretical analysis, its application to the DOA estimation will be presented in Chapter 8.

(2) Without loss of generality, we assume that $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{e}$ are all in the real domain unless otherwise stated.

(3) If $x_j^o = 0$ for some $j \in \{1, \cdots, N\}$, then β_j^o has no contributions to the observation $\boldsymbol{y}$ and hence it is impossible to recover β_j^o. As a result, the recovery of $\boldsymbol{\beta}^o$ in this chapter refers only to the recovery on the support of $\boldsymbol{x}^o$.

5.2.2 Main Results

In this chapter, a vector $\boldsymbol{v}$ is called $2K$-duplicately (D-) sparse if $\boldsymbol{v} = \left[\boldsymbol{v}_1^T, \boldsymbol{v}_2^T\right]^T$ with $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ being of the same dimension and jointly K-sparse (each being K-sparse and sharing the same support). The concept of duplicate (D-) RIP is defined as follows.

Definition 5.4. *Define the $2K$-duplicate (D-) RIC of a matrix $\boldsymbol{\Phi}$, denoted by $\bar{\delta}_{2K}(\boldsymbol{\Phi})$, as the smallest number such that*

$$\left(1 - \bar{\delta}_{2K}(\boldsymbol{\Phi})\right) \|\boldsymbol{v}\|_2^2 \leq \|\boldsymbol{\Phi}\boldsymbol{v}\|_2^2 \leq \left(1 + \bar{\delta}_{2K}(\boldsymbol{\Phi})\right) \|\boldsymbol{v}\|_2^2$$

holds for all $2K$-D-sparse vectors $\boldsymbol{v}$. $\boldsymbol{\Phi}$ is said to satisfy the $2K$-D-RIP with constant $\bar{\delta}_{2K}(\boldsymbol{\Phi})$ if $\bar{\delta}_{2K}(\boldsymbol{\Phi}) < 1$.

According to the definition above, D-RIP is a special case of RIP but concerns about only a type of structured sparse signals. With respect to the perturbed observation model in (5.9), let $\boldsymbol{\Psi} = [\boldsymbol{A}, \boldsymbol{B}]$. The main results of this chapter are stated in the following theorems. The proof of Theorem 5.3 is provided in Appendix A.1 and proofs of Theorems 5.4 and 5.5 are in Appendix A.2.

Theorem 5.3. *In the noise free case where $\boldsymbol{e} = \boldsymbol{0}$, assume that $\|\boldsymbol{x}^o\|_0 \leq K$ and $\bar{\delta}_{4K}(\boldsymbol{\Psi}) < 1$. Then an optimal solution $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ to the perturbed combinatorial optimization problem*

$$\min_{\boldsymbol{x}\in\mathbb{R}^N, \boldsymbol{\beta}\in[-r,r]^N} \|\boldsymbol{x}\|_0, \text{ subject to } \boldsymbol{y} = (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta})\,\boldsymbol{x} \tag{5.10}$$

with $\boldsymbol{\Delta} = diag(\boldsymbol{\beta})$ recovers $\boldsymbol{x}^o$ and $\boldsymbol{\beta}^o$.

Theorem 5.4. *Assume that $\bar{\delta}_{4K}(\boldsymbol{\Psi}) < \left(\sqrt{2(1+r^2)}+1\right)^{-1}$, $\|\boldsymbol{x}^o\|_0 \leq K$ and $\|\boldsymbol{e}\|_2 \leq \epsilon$. Then an optimal solution $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ to the perturbed (P-) BPDN problem*

$$\min_{\boldsymbol{x}\in\mathbb{R}^N, \boldsymbol{\beta}\in[-r,r]^N} \|\boldsymbol{x}\|_1, \text{ subject to } \|\boldsymbol{y} - (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta})\,\boldsymbol{x}\|_2 \leq \epsilon \tag{5.11}$$

achieves robustly stable signal recovery with

$$\|\boldsymbol{x}^* - \boldsymbol{x}^o\|_2 \leq C\epsilon, \tag{5.12}$$

$$\|(\boldsymbol{\beta}^* - \boldsymbol{\beta}^o) \odot \boldsymbol{x}^o\|_2 \leq \mathcal{C}\epsilon, \tag{5.13}$$

where

$$C = \frac{4\sqrt{1+\bar{\delta}_{4K}(\boldsymbol{\Psi})}}{1 - \left(\sqrt{2(1+r^2)}+1\right)\bar{\delta}_{4K}(\boldsymbol{\Psi})},$$

$$\mathcal{C} = \frac{\left[2 + \sqrt{1+r^2}\,\|\boldsymbol{\Psi}\|_2\, C\right]}{\sqrt{1-\bar{\delta}_{4K}(\boldsymbol{\Psi})}}.$$

Theorem 5.5. *Assume that* $\bar{\delta}_{4K}(\mathbf{\Psi}) < \left(\sqrt{2(1+r^2)}+1\right)^{-1}$ *and* $\|\boldsymbol{e}\|_2 \leq \epsilon$. *Then an optimal solution* $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ *to the P-BPDN problem in (5.11) satisfies that*

$$\|\boldsymbol{x}^* - \boldsymbol{x}^o\|_2 \leq \left(C_0 K^{-1/2} + C_1\right) \left\|\boldsymbol{x}^o - \boldsymbol{x}^K\right\|_1 + C_2\epsilon, \tag{5.14}$$

$$\left\|(\boldsymbol{\beta}^* - \boldsymbol{\beta}^o) \odot \boldsymbol{x}^K\right\|_2 \leq \left(\mathcal{C}_0 K^{-1/2} + \mathcal{C}_1\right) \left\|\boldsymbol{x}^o - \boldsymbol{x}^K\right\|_1 + \mathcal{C}_2\epsilon, \tag{5.15}$$

where

$$\begin{aligned} C_0 &= 2\left[1 + \left(\sqrt{2(1+r^2)} - 1\right)\bar{\delta}_{4K}(\mathbf{\Psi})\right]/a, \\ C_1 &= 2\sqrt{2}r\bar{\delta}_{4K}(\mathbf{\Psi})/a, \\ \mathcal{C}_0 &= \sqrt{1+r^2}\,\|\mathbf{\Psi}\|_2\, C_0/b, \\ \mathcal{C}_1 &= \left[\sqrt{1+r^2}C_1 + 2r\right]\|\mathbf{\Psi}\|_2/b \end{aligned}$$

with $a = 1 - \left(\sqrt{2(1+r^2)}+1\right)\bar{\delta}_{4K}(\mathbf{\Psi})$, $b = \sqrt{1 - \bar{\delta}_{4K}(\mathbf{\Psi})}$ *and* $C_2 = C$, $\mathcal{C}_2 = \mathcal{C}$ *with* $C, \mathcal{C}$ *as defined in Theorem 5.4.*

Remark 5.4. *In general, the robustly stable signal recovery cannot be concluded for compressible signals since the error bound in (5.14) may be very large in the case of large perturbation by* $C_1 = O(r)$. *If the perturbation is small with* $r = O\left(K^{-1/2}\right)$, *then the robust stability can be achieved for compressible signals by (5.6) provided that the D-RIP condition in Theorem 5.5 is satisfied.*

5.2.3 Interpretation of the Main Results

Theorem 5.3 states that for a K-sparse signal $\boldsymbol{x}^o$, it can be recovered by solving a combinatorial optimization problem provided $\bar{\delta}_{4K}(\mathbf{\Psi}) < 1$ when the measurements are exact. Meanwhile, $\boldsymbol{\beta}^o$ can be recovered. Since the combinatorial optimization problem is NP-hard and that its solution is sensitive to measurement noise [118], a more reliable approach, ℓ_1 minimization, is explored in Theorems 5.4 and 5.5.

Theorem 5.4 states the robustly stable recovery of a K-sparse signal $\boldsymbol{x}^o$ in SP-CS with the recovery error being at most proportional to the noise level. Such robust stability is obtained by solving an ℓ_1 minimization problem incorporated with the perturbation structure provided that the D-RIC is sufficiently small with respect to the perturbation level in terms of r. Meanwhile, the perturbation parameter $\boldsymbol{\beta}^o$ can be stably recovered on the support of $\boldsymbol{x}^o$. As the D-RIP condition is satisfied in Theorem 5.4, the signal recovery error of perturbed CS is constrained by the noise level ϵ, and the influence of the perturbation is limited to the coefficient before ϵ. For example, if $\bar{\delta}_{4K}\left(\boldsymbol{\Psi}\right) = 0.2$, then $\left\|\boldsymbol{x}^* - \boldsymbol{x}^o\right\|_2 \leq 8.48\epsilon, 8.50\epsilon, 11.0\epsilon$ corresponding to $r = 0.01, 0.1, 1$, respectively. In the special noise free case, the recovery is exact. This is similar to that in standard CS but in contrast to the existing robust signal recovery result in Subsection 5.1.3 where the recovery error exists once a matrix perturbation appears. Another interpretation of the D-RIP condition in Theorem 5.4 is that the robustly stable signal recovery requires that $r < \sqrt{\frac{1}{2}\left(\bar{\delta}_{4K}\left(\boldsymbol{\Psi}\right)^{-1} - 1\right)^2 - 1}$ for a fixed matrix $\boldsymbol{\Psi}$. Using the aforementioned example where $\bar{\delta}_{4K}\left(\boldsymbol{\Psi}\right) = 0.2$, the perturbation is required to satisfy $r < \sqrt{7}$. As a result, our robustly stable signal recovery result of SP-CS applies to the case of large perturbation if the D-RIC of $\boldsymbol{\Psi}$ is sufficiently small while the existing result does not as demonstrated in Remark 5.2.

Theorem 5.5 considers general signals and is a generalized form of Theorem 5.4. In comparison with Theorem 5.1 in standard CS, one more term $C_1\left\|\boldsymbol{x}^o - \boldsymbol{x}^K\right\|_1$ appears in the upper bound of the recovery error. The robust stability does not hold generally for compressible signals as illustrated in Remark 5.4 while it is true under an additional assumption $r = O\left(K^{-1/2}\right)$.

The results in this chapter generalize that in standard CS. Without accounting for the symbolic difference between $\delta_{2K}\left(\boldsymbol{\Phi}\right)$ and $\bar{\delta}_{4K}\left(\boldsymbol{\Psi}\right)$, the conditions in Theorems 5.1 and 5.5 coincide, as well as the upper bounds in (5.5) and (5.14) for the recovery errors, as the perturbation vanishes or equivalently $r \rightarrow 0$. As mentioned before, the RIP condition for guaranteed stable recovery in standard CS has been relaxed. Similar techniques may be adopted to possibly relax the D-RIP condition in SP-CS. While this chapter is focused on the ℓ_1 minimization approach, it is also possible to modify other algorithms in standard CS and apply them to SP-CS to provide similar recovery guarantees.

5.2.4 When is the D-RIP satisfied?

Existing works studying the RIP are mainly focused on random matrices. In standard CS, $\boldsymbol{\Phi}$ has the K-RIP with constant δ with a large probability provided that $M \geq C_\delta K \log(N/K)$ and $\boldsymbol{\Phi}$ has properly scaled i.i.d. subgaussian distributed entries, where the constant C_δ depending on δ and the distribution [115]. The D-RIP can be considered as a model-based RIP introduced in [119]. Suppose that $\boldsymbol{A}$, $\boldsymbol{B}$ are mutually independent and both are i.i.d. subgaussian distributed (the true sensing matrix $\boldsymbol{\Phi} = \boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^o$ is also i.i.d. subgaussian distributed if $\boldsymbol{\beta}^o$ is independent of $\boldsymbol{A}$ and $\boldsymbol{B}$). The model-based RIP is determined by the number of subspaces of the structured sparse signals that are referred to as the D-sparse ones in this chapter. For $\boldsymbol{\Psi} = [\boldsymbol{A}, \boldsymbol{B}]$, the number of $2K$-dimensional subspaces for $2K$-D-sparse signals is $\begin{pmatrix} N \\ K \end{pmatrix}$. Consequently, $\boldsymbol{\Psi}$ has the $2K$-D-RIP with constant δ with a large probability also provided that $M \geq C_\delta K \log(N/K)$ by [119, Theorem 1] or [120, Theorem 3.3]. So, in the case of a high dimensional system and $r \to 0$, the D-RIP condition on $\boldsymbol{\Psi}$ in Theorem 5.4 or 5.5 can be satisfied when the RIP condition on $\boldsymbol{\Phi}$ (after proper scaling of its columns) in standard CS is met. It means that the perturbation in SP-CS gradually strengthens the D-RIP condition for robustly stable signal recovery but there exists no gap between SP-CS and standard CS in the case of high dimensional systems.

It is noted that there is another way to stably recover the original signal $\boldsymbol{x}^o$ in SP-CS. Given the sparse signal case as an example where $\boldsymbol{x}^o$ is K-sparse. Let $\boldsymbol{z}^o = \begin{bmatrix} \boldsymbol{x}^o \\ \boldsymbol{\beta}^o \odot \boldsymbol{x}^o \end{bmatrix}$, and it is $2K$-sparse. The observation model can be written as $\boldsymbol{y} = \boldsymbol{\Psi}\boldsymbol{z}^o + \boldsymbol{e}$. Then $\boldsymbol{z}^o$ and hence, $\boldsymbol{x}^o$, can be stably recovered from the problem[1]

$$\min_{\boldsymbol{z}} \left\|\boldsymbol{z}\right\|_1, \text{ subject to } \left\|\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{z}\right\|_2 \leq \epsilon \tag{5.16}$$

provided that $\delta_{4K}(\boldsymbol{\Psi}) < \sqrt{2} - 1$ by Theorem 5.1. It looks like that we transformed the perturbation into a signal of interest. Denote TPS-BPDN the problem in (5.16).

[1] It is hard to incorporate the knowledge $\boldsymbol{\beta}^o \in [-r, r]^N$ into the problem in (5.16).

In a high dimensional system, the condition $\delta_{4K}\left(\mathbf{\Psi}\right) < \sqrt{2}-1$ requires about twice as many as the measurements that makes the D-RIP condition $\bar{\delta}_{4K}\left(\mathbf{\Psi}\right) < \sqrt{2}-1$ hold by [119, Theorem 1] corresponding to the D-RIP condition in Theorem 5.4 or 5.5 as $r \to 0$. As a result, for a considerable range of perturbation level, the D-RIP condition in Theorem 5.4 or 5.5 for P-BPDN is weaker than that for TPS-BPDN since it varies slowly for a moderate perturbation (as an example, $\bar{\delta}_{4K}\left(\mathbf{\Psi}\right) < 0.414, 0.413, 0.409$ corresponds to $r = 0, 0.1, 0.2$ respectively). Numerical simulations in Subsection 5.4 can verify our conclusion.

5.2.5 Relaxation of the Optimal Solution

In Theorem 5.5 (Theorem 5.4 is a special case), $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ is required to be an optimal solution to P-BPDN. Naturally, we would like to know if the requirement of the optimality is necessary for a "good" recovery in the sense that a good recovery validates the error bounds in (5.14) and (5.15) under the conditions in Theorem 5.5. Generally speaking, the answer is negative since, regarding the optimality of $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$, only $\left\|\boldsymbol{x}^*\right\|_1 \leq \left\|\boldsymbol{x}^o\right\|_1$ and the feasibility of $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ are used in the proof of Theorem 5.5 in Appendix A.2. Denote $\mathcal{D}$ the feasible domain of P-BPDN, i.e.,

$$\mathcal{D} = \left\{(\boldsymbol{x}, \boldsymbol{\beta}) : \boldsymbol{\beta} \in [-r, r]^N, \left\|\boldsymbol{y} - (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta})\,\boldsymbol{x}\right\|_2 \leq \epsilon \text{ with } \boldsymbol{\Delta} = \operatorname{diag}(\boldsymbol{\beta})\right\}. \tag{5.17}$$

We have the following corollary.

Corollary 5.1. *Under the assumptions in Theorem 5.5, any $(\boldsymbol{x}, \boldsymbol{\beta}) \in \mathcal{D}$ that meets $\left\|\boldsymbol{x}\right\|_1 \leq \left\|\boldsymbol{x}^o\right\|_1$ satisfies that*

$$\begin{aligned}\left\|\boldsymbol{x} - \boldsymbol{x}^o\right\|_2 &\leq \left(C_0 K^{-1/2} + C_1\right)\left\|\boldsymbol{x}^o - \boldsymbol{x}^K\right\|_1 + C_2\epsilon,\\ \left\|(\boldsymbol{\beta} - \boldsymbol{\beta}^o) \odot \boldsymbol{x}^K\right\|_2 &\leq \left(\mathcal{C}_0 K^{-1/2} + \mathcal{C}_1\right)\left\|\boldsymbol{x}^o - \boldsymbol{x}^K\right\|_1 + \mathcal{C}_2\epsilon\end{aligned}$$

with $C_j, \mathcal{C}_j$, $j = 0, 1, 2$, as defined in Theorem 5.5.

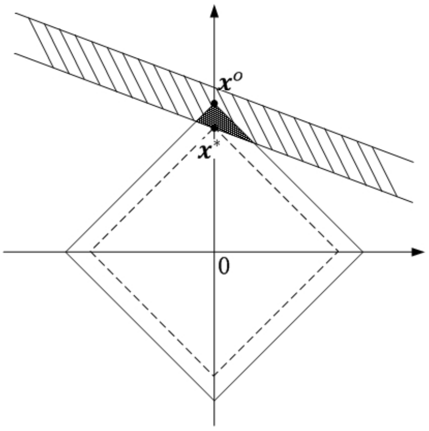


Figure 5.1: Illustration of Corollary 5.2. The shaded band area refers to the feasible domain of BPDN. The triangular area, the intersection of the feasible domain and the ℓ_1 ball $\{\boldsymbol{x} : \|\boldsymbol{x}\|_1 \leq \|\boldsymbol{x}^o\|_1\}$, is the set of all good recoveries.

Corollary 5.1 generalizes Theorem 5.5 and its proof follows directly from that of Theorem 5.5. It shows that a good recovery in SP-CS is not necessarily an optimal solution to P-BPDN. A similar result holds in standard CS that generalizes Theorem 5.1, and the proof of Theorem 5.1 in [17] applies directly to such case.

Corollary 5.2. *Under the assumptions in Theorem 5.1, any $\boldsymbol{x}$ that meets $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon$ and $\|\boldsymbol{x}\|_1 \leq \|\boldsymbol{x}^o\|_1$ satisfies that*

$$\|\boldsymbol{x} - \boldsymbol{x}^o\|_2 \leq C_0^{std} K^{-1/2} \left\|\boldsymbol{x}^o - \boldsymbol{x}^K\right\|_1 + C_1^{std}\epsilon \tag{5.18}$$

with C_0^{std}, C_1^{std} as defined in Theorem 5.1.

An illustration of Corollary 5.2 is presented in Fig. 5.1, where the shaded band area refers to the feasible domain of BPDN in (5.4) and all points in the triangular area, the intersection of the feasible domain of BPDN and the ℓ_1 ball $\{\boldsymbol{x} : \|\boldsymbol{x}\|_1 \leq \|\boldsymbol{x}^o\|_1\}$, are good candidates for recovery of $\boldsymbol{x}^o$. The reason why one seeks for the optimal solution $\boldsymbol{x}^*$ is to guarantee that the inequality $\|\boldsymbol{x}\|_1 \leq \|\boldsymbol{x}^o\|_1$ holds since $\|\boldsymbol{x}^o\|_1$ is generally unavailable *a priori*. Corollary 5.2 can explain why a satisfactory recovery can be obtained in practice using some algorithm that may not produce an optimal solution to BPDN, e.g., rONE-L1 proposed in Chapter 3. Corollaries 5.1 and 5.2 are useful for checking the effectiveness of an algorithm in the case when the output cannot be guaranteed to be optimal.[2] Namely, an algorithm is called effective in solving some ℓ_1 minimization problem if it can produce a feasible solution with its ℓ_1 norm no larger than that of the original signal.

5.3 Algorithms for P-BPDN

5.3.1 Special Case: Positive Signals

This subsection studies a special case where the original signal $\boldsymbol{x}^o$ is positive-valued (except zero entries). Such a case has been studied in standard CS [32, 33]. By incorporating the positiveness of $\boldsymbol{x}^o$, P-BPDN is modified into the positive P-BPDN (PP-BPDN) problem

$$\min_{\boldsymbol{x},\boldsymbol{\beta}} \mathbf{1}^T\boldsymbol{x}, \text{ subject to } \begin{cases} \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}\right)\boldsymbol{x}\right\|_2 \leq \epsilon, \\ \boldsymbol{x} \succeq \mathbf{0}, \\ r\mathbf{1} \succeq \boldsymbol{\beta} \succeq -r\mathbf{1}, \end{cases}$$

where $\succeq$ operates elementwise for vectors and $\mathbf{0}$, $\mathbf{1}$ are column vectors composed of 0, 1 respectively with proper dimensions. It is noted that the robustly stable signal recovery results in this chapter apply directly to the solution to PP-BPDN in such case. This subsection shows that the nonconvex PP-BPDN problem can be transformed into a convex one and hence its optimal solution can be efficiently obtained. Denote $\boldsymbol{p} = \boldsymbol{\beta} \odot \boldsymbol{x}$. A new, convex problem (P_1) is introduced as follows.

$$(P_1) \quad \min_{\boldsymbol{x},\boldsymbol{p}} \mathbf{1}^T\boldsymbol{x}, \text{ subject to } \begin{cases} \left\|\boldsymbol{y} - \boldsymbol{\Psi}\begin{bmatrix}\boldsymbol{x} \\ \boldsymbol{p}\end{bmatrix}\right\|_2 \leq \epsilon, \\ \boldsymbol{x} \succeq \mathbf{0}, \\ r\boldsymbol{x} \succeq \boldsymbol{p} \succeq -r\boldsymbol{x}. \end{cases}$$

[2] It is common when the problem to be solved is nonconvex, such as P-BPDN as discussed in Section 5.3 and ℓ_p $(0 \leq p < 1)$ minimization approaches [121–123] in standard CS. In addition, Corollaries 5.1 and 5.2 can be readily extended to the ℓ_p $(0 \leq p < 1)$ minimization approaches.

Theorem 5.6. *Problems PP-BPDN and* (P_1) *are equivalent in the sense that, if* $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ *is an optimal solution to PP-BPDN, then there exists* $\boldsymbol{p}^* = \boldsymbol{\beta}^* \odot \boldsymbol{x}^*$ *such that* $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ *is an optimal solution to* (P_1)*, and that, if* $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ *is an optimal solution to* (P_1)*, then there exists* $\boldsymbol{\beta}^*$ *with* $\beta_j^* = \begin{cases} p_j^*/x_j^*, & \textit{if } x_j^* > 0; \\ 0, & \textit{otherwise} \end{cases}$ *such that* $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ *is an optimal solution to PP-BPDN.*

Proof. We only prove the first part of Theorem 5.6 using contradiction. The second part follows similarly. Suppose that $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ with $\boldsymbol{p}^* = \boldsymbol{\beta}^* \odot \boldsymbol{x}^*$ is not an optimal solution to (P_1). Then there exists $(\boldsymbol{x}', \boldsymbol{p}')$ in the feasible domain of (P_1) such that $\|\boldsymbol{x}'\|_1 < \|\boldsymbol{x}^*\|_1$. Define $\boldsymbol{\beta}'$ as $\beta_j' = \begin{cases} p_j'/x_j', & \text{if } x_j' > 0; \\ 0, & \text{otherwise} \end{cases}$. It is easy to show that $(\boldsymbol{x}', \boldsymbol{\beta}')$ is a feasible solution to PP-BPDN. By $\|\boldsymbol{x}'\|_1 < \|\boldsymbol{x}^*\|_1$ we conclude that $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ is not an optimal solution to PP-BPDN, which leads to contradiction. ■

Theorem 5.6 states that an optimal solution to PP-BPDN can be efficiently obtained by solving the convex problem (P_1).

5.3.2 AA-P-BPDN: Alternating Algorithm for P-BPDN

For general signals, P-BPDN in (5.11) is nonconvex. A simple method is to solve a series of BPDN problems with

$$\boldsymbol{x}^{(j+1)} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1, \text{ subject to } \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^{(j)}\right)\boldsymbol{x}\right\|_2 \leq \epsilon, \tag{5.19}$$

$$\boldsymbol{\beta}^{(j+1)} = \arg\min_{\boldsymbol{\beta}\in[-r,r]^N} \left\|\boldsymbol{y} - (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta})\boldsymbol{x}^{(j+1)}\right\|_2 \tag{5.20}$$

starting from $\boldsymbol{\beta}^{(0)} = \boldsymbol{0}$, where the superscript ${}^{(j)}$ indicates the jth iteration and $\boldsymbol{\Delta}^{(j)} = \operatorname{diag}\left(\boldsymbol{\beta}^{(j)}\right)$. Denote AA-P-BPDN the alternating algorithm defined by (5.19) and (5.20). To analyze AA-P-BPDN, we first present the following two lemmas.

Lemma 5.1. *For a matrix sequence* $\left\{\boldsymbol{\Phi}^{(j)}\right\}_{j=1}^{\infty}$ *composed of fat matrices, let* $\mathcal{D}^j = \left\{\boldsymbol{v} : \left\|\boldsymbol{y} - \boldsymbol{\Phi}^{(j)}\boldsymbol{v}\right\|_2 \leq \epsilon\right\}$*,* $j = 1, 2, \cdots$*, and* $\mathcal{D}^* = \{\boldsymbol{v} : \|\boldsymbol{y} - \boldsymbol{\Phi}^*\boldsymbol{v}\|_2 \leq \epsilon\}$ *with* $\epsilon > 0$*.*

If $\mathbf{\Phi}^{(j)} \rightarrow \mathbf{\Phi}^{*}$, as $j \rightarrow +\infty$, then for any $\boldsymbol{v} \in \mathcal{D}^{*}$ there exists a sequence $\left\{\boldsymbol{v}^{(j)}\right\}_{j=1}^{\infty}$ with $\boldsymbol{v}^{(j)} \in \mathcal{D}^{(j)}$, $j = 1, 2, \cdots$, such that $\boldsymbol{v}^{(j)} \rightarrow \boldsymbol{v}$, as $j \rightarrow +\infty$.

Lemma 5.1 studies the variation of feasible domains $\mathcal{D}^{j}$, $j = 1, 2, \cdots$, of a series of BPDN problems whose sensing matrices $\mathbf{\Phi}^{(j)}$, $j = 1, 2, \cdots$, converge to $\mathbf{\Phi}^{*}$. It states that the sequence of the feasible domains also converges to $\mathcal{D}^{*}$ in the sense that for any point in $\mathcal{D}^{*}$, there exists a sequence of points, each of which belongs to one $\mathcal{D}^{j}$, that converges to the point. To prove Lemma 5.1, we first show that it holds for any interior point of $\mathcal{D}^{*}$ by constructing such a sequence. Then we show that it also holds for a boundary point of $\mathcal{D}^{*}$ by that for any boundary point there exists a sequence of interior points of $\mathcal{D}^{*}$ that converges to it. The detailed proof is given in Appendix A.3.

Lemma 5.2. *An optimal solution $\boldsymbol{x}^{*}$ to the BPDN problem in (5.4) satisfies that $\boldsymbol{x}^{*} = \mathbf{0}$, if $\|\boldsymbol{y}\|_2 \leq \epsilon$, or $\|\boldsymbol{y} - \mathbf{\Phi}\boldsymbol{x}^{*}\|_2 = \epsilon$, otherwise.*

Proof. It is trivial for the case where $\|\boldsymbol{y}\|_2 \leq \epsilon$. Consider the other case where $\|\boldsymbol{y}\|_2 > \epsilon$. Note first that $\boldsymbol{x}^{*} \neq \mathbf{0}$. We use contradiction to show that the equality $\|\boldsymbol{y} - \mathbf{\Phi}\boldsymbol{x}^{*}\|_2 = \epsilon$ holds. Suppose that $\|\boldsymbol{y} - \mathbf{\Phi}\boldsymbol{x}^{*}\|_2 < \epsilon$. Introduce $f(\theta) = \|\boldsymbol{y} - \theta\mathbf{\Phi}\boldsymbol{x}^{*}\|_2$. Then $f(0) > \epsilon$, and $f(1) < \epsilon$. There exists θ_0, $0 < \theta_0 < 1$, such that $f(\theta_0) = \epsilon$ since $f(\theta)$ is continuous on the interval $[0, 1]$. Hence, $\boldsymbol{x}' = \theta_0\boldsymbol{x}^{*}$ is a feasible solution to BPDN in (5.4). We conclude that $\boldsymbol{x}^{*}$ is not optimal by $\|\boldsymbol{x}'\|_1 = \theta_0 \|\boldsymbol{x}^{*}\|_1 < \|\boldsymbol{x}^{*}\|_1$, which leads to contradiction. ■

Lemma 5.2 studies the location of an optimal solution to the BPDN problem. It states that the optimal solution locates at the origin if the origin is a feasible solution, or at the boundary of the feasible domain otherwise. This can be easily observed from Fig. 5.1. Based on Lemmas 5.1 and 5.2, we have the following results for AA-P-BPDN.

Theorem 5.7. *Any accumulation point $(\boldsymbol{x}^{*}, \boldsymbol{\beta}^{*})$ of the sequence $\left\{\left(\boldsymbol{x}^{(j)}, \boldsymbol{\beta}^{(j)}\right)\right\}_{j=1}^{\infty}$ is a stationary point of AA-P-BPDN in the sense that*

$$\boldsymbol{x}^{*} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1, \text{ subject to } \|\boldsymbol{y} - (\boldsymbol{A} + \boldsymbol{B}\mathbf{\Delta}^{*})\,\boldsymbol{x}\|_2 \leq \epsilon, \tag{5.21}$$

$$\boldsymbol{\beta}^{*} = \arg\min_{\boldsymbol{\beta}\in[-r,r]^{N}} \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}\right)\boldsymbol{x}^{*}\right\|_{2} \tag{5.22}$$

with $\boldsymbol{\Delta}^{*} = diag\left(\boldsymbol{\beta}^{*}\right)$.

Theorem 5.8. *An optimal solution* $(\boldsymbol{x}^{*}, \boldsymbol{\beta}^{*})$ *to P-BPDN in (5.11) is a stationary point of AA-P-BPDN.*

Theorem 5.7 studies the property of the solution $\left(\boldsymbol{x}^{(j)}, \boldsymbol{\beta}^{(j)}\right)$ produced by AA-P-BPDN. It shows that $\left(\boldsymbol{x}^{(j)}, \boldsymbol{\beta}^{(j)}\right)$ is arbitrarily close to a stationary point of AA-P-BPDN as the iteration index j is large enough.[3] Hence, the output of AA-P-BPDN can be considered as a stationary point provided that an appropriate termination criterion is set. Theorem 5.8 tells that an optimal solution to P-BPDN is a stationary point of AA-P-BPDN. So, it is possible for AA-P-BPDN to produce an optimal solution to P-BPDN. The proofs of Theorems 5.7 and 5.8 are provided in Appendix A.4 and Appendix A.5 respectively.

Remark 5.5. *The following formulation of P-BPDN can possibly provide an efficient approach to an optimal solution to P-BPDN. Let* $\boldsymbol{x} = \boldsymbol{x}_{+} - \boldsymbol{x}_{-}$ *where* $\boldsymbol{x}_{+} \succeq \boldsymbol{0}$, $\boldsymbol{x}_{-} \succeq \boldsymbol{0}$ *and* $\boldsymbol{x}_{+} \odot \boldsymbol{x}_{-} = \boldsymbol{0}$. *Then we have* $|\boldsymbol{x}| = \boldsymbol{x}_{+} + \boldsymbol{x}_{-}$ *where* $|\cdot|$ *applies elementwise. Denote* $\boldsymbol{p} = \boldsymbol{\beta} \odot \boldsymbol{x}$. *A convex problem can be cast as follows:*

$$\begin{aligned} &\min_{\boldsymbol{x}_{+}, \boldsymbol{x}_{-}, \boldsymbol{p}} \boldsymbol{1}^{T}\left(\boldsymbol{x}_{+} + \boldsymbol{x}_{-}\right), \\ &\textit{subject to} \begin{cases} \left\| \boldsymbol{y} - \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{A} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{+} \\ \boldsymbol{x}_{-} \\ \boldsymbol{p} \end{bmatrix} \right\|_{2} \leq \epsilon, \\ \boldsymbol{x}_{+} \succeq \boldsymbol{0}, \\ \boldsymbol{x}_{-} \succeq \boldsymbol{0}, \\ r\left(\boldsymbol{x}_{+} + \boldsymbol{x}_{-}\right) \succeq \boldsymbol{p} \succeq -r\left(\boldsymbol{x}_{+} + \boldsymbol{x}_{-}\right). \end{cases} \end{aligned} \tag{5.23}$$

[3] It is shown in the proof of Theorem 5.7 in Appendix A.4 that the sequence $\left\{\left(\boldsymbol{x}^{(j)}, \boldsymbol{\beta}^{(j)}\right)\right\}_{j=1}^{\infty}$ is bounded. And it can be shown, for example, using contradiction, that for a bounded sequence $\{a_j\}_{j=1}^{\infty}$, there exists an accumulation point of $\{a_j\}_{j=1}^{\infty}$ such that a_j is arbitrarily close to it as j is large enough.

The above convex problem can be considered as a convex relaxation of P-BPDN since it can be shown (like that in Theorem 5.6) that an optimal solution to P-BPDN can be obtained based on an optimal solution to the problem in (5.23) incorporated with an additional nonconvex constraint $\boldsymbol{x}_+ \odot \boldsymbol{x}_- = \mathbf{0}$. An interesting phenomenon has been observed through numerical simulations that an optimal solution to the problem in (5.23) still satisfies the constraint $\boldsymbol{x}_+ \odot \boldsymbol{x}_- = \mathbf{0}$. Based on such an observation, an efficient approach to P-BPDN is to firstly solve (5.23), and then check whether its solution, denoted by $\left(\boldsymbol{x}_+^*, \boldsymbol{x}_-^*, \boldsymbol{p}^*\right)$, satisfies $\boldsymbol{x}_+ \odot \boldsymbol{x}_- = \mathbf{0}$. If it does, then $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ is an optimal solution to P-BPDN where $\boldsymbol{x}^* = \boldsymbol{x}_+^* - \boldsymbol{x}_-^*$ and $\beta_j^* = \begin{cases} p_j^*/x_j^*, & \text{if } x_j^* \neq 0; \\ 0, & \text{otherwise.} \end{cases}$ Otherwise, we may turn to AA-P-BPDN again. But we note that it is still an open problem whether an optimal solution to (5.23) always satisfies the constraint $\boldsymbol{x}_+ \odot \boldsymbol{x}_- = \mathbf{0}$. In addition, the convex relaxation in (5.23) does not apply to the complex signal case as in the DOA estimation studied in Chapter 8.

5.3.3 Effectiveness of AA-P-BPDN

As reported in the last subsection, it is possible for AA-P-BPDN to produce an optimal solution to P-BPDN. But it is not easy to check the optimality of the output of AA-P-BPDN because of the nonconvexity of P-BPDN. Instead, we study the effectiveness of AA-P-BPDN in solving P-BPDN in this subsection with the concept of effectiveness as defined in Subsection 5.2.5. By Corollary 5.1, a good signal recovery $\widehat{\boldsymbol{x}}$ of $\boldsymbol{x}^o$ is not necessarily an optimal solution. It requires only that $\left(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\beta}}\right)$, where $\widehat{\boldsymbol{\beta}}$ denotes the recovery of $\boldsymbol{\beta}^o$, be a feasible solution to P-BPDN and that $\|\widehat{\boldsymbol{x}}\|_1 \leq \|\boldsymbol{x}^o\|_1$ holds. As shown in the proof of Theorem 5.7 in Appendix A.4, that $\left(\boldsymbol{x}^{(j)}, \boldsymbol{\beta}^{(j)}\right)$ for any $j \geq 1$ is a feasible solution to P-BPDN and that the sequence $\left\{\left\|\boldsymbol{x}^{(j)}\right\|_1\right\}_{j=1}^{\infty}$ is monotone decreasing and converges. So, the effectiveness of AA-P-BPDN in solving P-BPDN can be assessed via numerical simulations by checking whether $\left\|\boldsymbol{x}^{AA}\right\|_1 \leq \|\boldsymbol{x}^o\|_1$ holds with $\boldsymbol{x}^{AA}$ denoting the output of AA-P-BPDN. The effectiveness of AA-P-BPDN is verified in Section 5.4 via numerical simulations, where we observe that the inequality $\left\|\boldsymbol{x}^{AA}\right\|_1 \leq \left\|\boldsymbol{x}^o\right\|_1$ holds in all experiments (over 3700 trials).

5.4 Numerical Simulations

This section demonstrates the robustly stable signal recovery results of SP-CS, as well as the effectiveness of AA-P-BPDN in solving P-BPDN in (5.11), via numerical simulations. AA-P-BPDN is implemented in Matlab with problems in (5.19) and (5.20) being solved using CVX [124]. AA-P-BPDN is terminated as $\frac{\left|\left\|\boldsymbol{x}^{(j)}\right\|_1 - \left\|\boldsymbol{x}^{(j-1)}\right\|_1\right|}{\left\|\boldsymbol{x}^{(j-1)}\right\|_1} \leq 1 \times 10^{-6}$ or the maximum number of iterations, set to 200, is reached. PP-BPDN is also implemented in Matlab and solved by CVX.

We first consider general signals. The sparse signal case is mainly studied. The variation of the signal recovery error is studied with respect to the noise level, perturbation level and number of measurements respectively. Besides AA-P-BPDN for P-BPDN in SP-CS, performances of three other approaches are also studied. The first one assumes that the perturbation is known *a priori* and recovers the original signal $\boldsymbol{x}^o$ by solving, namely, the oracle (O-) BPDN problem

$$\min_{\boldsymbol{x}} \left\|\boldsymbol{x}\right\|_1, \text{ subject to } \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^o\right)\boldsymbol{x}\right\|_2 \leq \epsilon.$$

The O-BPDN approach produces the best recovery result of SP-CS within the scope of ℓ_1 minimization of CS since it exploits the exact perturbation (oracle information). The second one corresponds to the robust signal recovery of perturbed CS as described in Subsection 5.1.3 and solves N-BPDN in (5.7) where $\epsilon_{\boldsymbol{E},\boldsymbol{x}^o} = \left\|\boldsymbol{B}\boldsymbol{\Delta}^o\boldsymbol{x}^o\right\|_2$ is used though it is not available in practice. The last one refers to the other approach to SP-CS that seeks for the signal recovery by solving TPS-BPDN in (5.16) as discussed in Subsection 5.2.5.

The first experiment studies the signal recovery error with respect to the noise level. We set the signal length $N = 200$, sample size $M = 80$, sparsity level $K = 10$

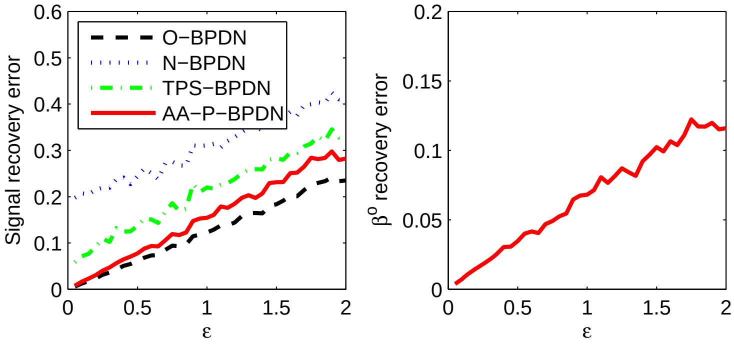


Figure 5.2: Signal and perturbation recovery errors with respect to the noise level ϵ with parameter settings $(N, M, K, r) = (200, 80, 10, 0.1)$. Both signal and $\boldsymbol{\beta}^o$ recovery errors of AA-P-BPDN for P-BPDN in SP-CS are proportional to ϵ.

and perturbation parameter $r = 0.1$. The noise level ϵ varies from 0.05 to 2 with interval 0.05. For each combination of (N, M, K, r, ϵ), the signal recovery error, as well as $\boldsymbol{\beta}^o$ recovery error (on the support of $\boldsymbol{x}^o$), is averaged over $R = 50$ trials. In each trial, matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are generated from Gaussian distribution and each column of them has zero mean and unit norm after proper scaling. The sparse signal $\boldsymbol{x}^o$ is composed of unit spikes with random signs and locations. Entries of $\boldsymbol{\beta}^o$ are uniformly distributed in $[-r, r]$. The noise $\boldsymbol{e}$ is zero mean Gaussian distributed and then scaled such that $\|\boldsymbol{e}\|_2 = \epsilon$. Using the same data, the four approaches, including O-BPDN, N-BPDN, TPS-BPDN and AA-P-BPDN for P-BPDN, are used to recover $\boldsymbol{x}^o$ respectively in each trial. The simulation results are shown in Fig. 5.2. It can be seen that both signal and $\boldsymbol{\beta}^o$ recovery errors of AA-P-BPDN for P-BPDN in SP-CS are proportional to the noise, which is consistent with our robustly stable signal recovery result. The error of N-BPDN grows linearly with the noise but a large error still exhibits in the noise free case. Except the ideal case of O-BPDN, our proposed P-BPDN has the smallest error.

The second experiment studies the effect of the structured perturbation. Experiment settings are the same as those in the first experiment except that we set $(N, M, K, \epsilon) = (200, 80, 10, 0.5)$ and vary $r \in \{0.05, 0.1, \cdots, 1\}$. Fig. 5.3 presents our simulation results. A nearly constant error is obtained using O-BPDN in standard CS since the perturbation is assumed to be known in O-BPDN. The error of

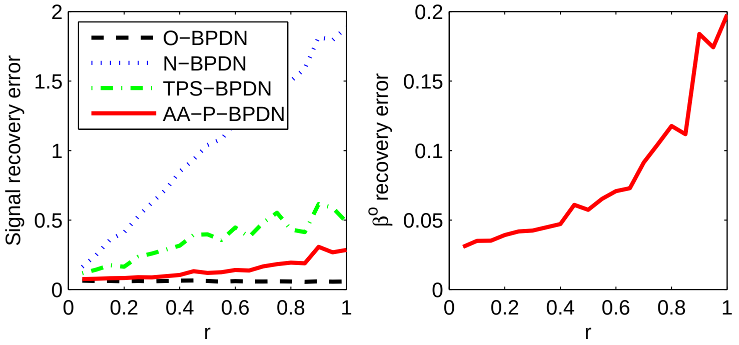


Figure 5.3: Signal and perturbation recovery errors with respect to the perturbation level in terms of r with parameter settings $(N, M, K, \epsilon) = (200, 80, 10, 0.5)$. The error of AA-P-BPDN for P-BPDN in SP-CS slowly increases with the perturbation level and is quite close to that of the ideal case of O-BPDN for a moderate perturbation.

AA-P-BPDN for P-BPDN in SP-CS slowly increases with the perturbation level and is quite close to that of O-BPDN for a moderate perturbation. Such a behavior is consistent with our analysis. Besides, it can be observed that the error of N-BPDN grows linearly with the perturbation level. Again, our proposed P-BPDN has the smallest error except O-BPDN.

The third experiment studies the variation of the recovery error with the number of measurements. We set $(N, K, r, \epsilon) = (200, 10, 0.1, 0.2)$ and vary $M \in \{30, 35, \cdots, 100\}$. Simulation results are presented in Fig. 5.4. Signal recovery errors of all four approaches decrease as the number of measurements increases. Again, it is observed that O-BPDN of the ideal case achieves the best result followed by our proposed P-BPDN. For example, to obtain the signal recovery error of 0.05, about 55 measurements are needed for O-BPDN while the numbers are, respectively, 65 for AA-P-BPDN and 95 for TPS-BPDN. It is impossible for N-BPDN to achieve such a small error in our observation because of the existence of the perturbation.

We next consider a compressible signal that is generated by taking a fixed sequence $\{2.8843 \cdot j^{-1.5}\}_{j=1}^{N}$ with $N = 200$, randomly permuting it, and multiplying by a random sign sequence (the coefficient 2.8843 is chosen such that the compressible signal has the same ℓ_2 norm as the sparse signals in the previous experiments). It is sought to be recovered from $M = 70$ noisy measurements with $\epsilon = 0.2$ and $r = 0.1$.

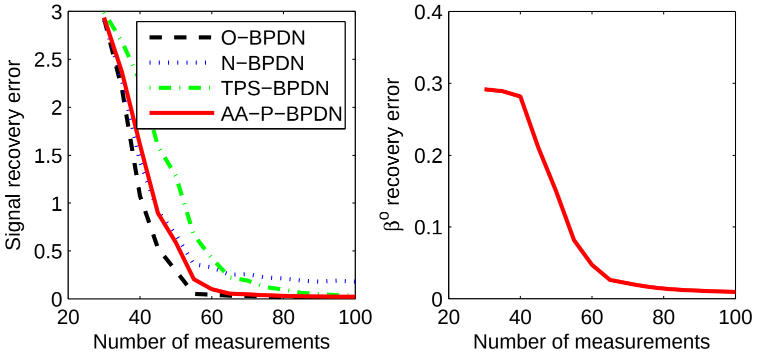


Figure 5.4: Signal and perturbation recovery errors with respect to the number of measurements with parameter settings $(N, K, r, \epsilon) = (200, 10, 0.1, 0.2)$. AA-P-BPDN for P-BPDN in SP-CS has the best performance except the ideal case of O-BPDN.

Give experiment results in one instance as an example. The signal recovery error of AA-P-BPDN for P-BPDN in SP-CS is about 0.239, while errors of O-BPDN, N-BPDN and TPS-BPDN are about 0.234, 0.361 and 0.314 respectively.

For the special positive signal case, an optimal solution to PP-BPDN can be efficiently obtained. An experiment result is shown in Fig. 5.5, where a sparse signal of length $N = 200$, composed of $K = 10$ positive unit spikes, is exactly recovered from $M = 50$ noise free measurements with $r = 0.1$ by solving (P_1).

5.5 Conclusion

This chapter studied the CS problem in the presence of measurement noise and a structured matrix perturbation. A concept named as robust stability for signal recovery was introduced. It was shown that the robust stability can be achieved for a sparse signal by solving an ℓ_1 minimization problem P-BPDN under mild conditions. In the presence of measurement noise, the recovery error is at most proportional to the noise level and the recovery is exact in the special noise free case. A general result for compressible signals was also reported. An alternating algorithm named as AA-P-BPDN was proposed to solve the nonconvex P-BPDN problem, and numerical simulations were carried out, verifying our theoretical analysis. A

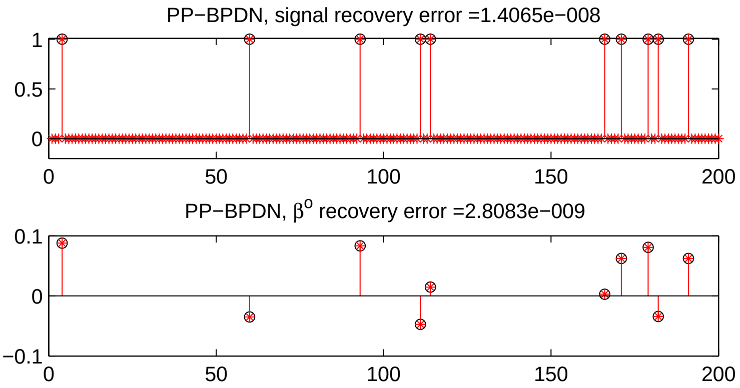


Figure 5.5: Exact recovery of a positive sparse signal from noise-free measurements with $(M, N, K, r, \epsilon) = (200, 50, 10, 0.1, 0)$. PP-BPDN is solved by solving (P_1). $\boldsymbol{\beta}^o$ and its recovery are shown only on the support of $\boldsymbol{x}^o$. Black circles: original signal and $\boldsymbol{\beta}^o$; red stars: recoveries.

practical application of the proposed perturbed CS framework and solution to the DOA estimation will be provided in Chapter 8.

Chapter 6

Bayesian Compressed Sensing with New Sparsity-Inducing Prior

Besides the ℓ_1 optimization methods that have been studied in the previous chapters, sparse Bayesian learning (SBL) [39] is another popular approach to the sparse signal recovery in CS, see a list of publications [38, 41–51, 54]. This chapter and the following two are built upon this approach. In this chapter, we introduce a new sparsity-inducing prior named as Gaussian shifted-truncated-gamma (G-STG) prior that generalizes the Laplace prior in [46] and the Gaussian-gamma prior in [54]. The extended flexibility of the new prior promotes its capability of modeling sparse signals. In fact, we show that the Gaussian-gamma prior cannot work in the main algorithm of this chapter. From the perspective of MAP estimation, the G-STG prior corresponds to a nonconvex objective function in optimization methods in general. For signal recovery we propose an iterative algorithm based on an evidence procedure and a fast greedy algorithm inspired by the algorithm in [57]. We show that similar theoretical guarantees shown in [40] for the basic SBL also hold for the new SBL method. Specifically, every local optimum of the SBL cost function is achieved at a sparse solution and the global optimum is achieved at the maximally sparse solution. Moreover, the proposed algorithm produces a sparser solution than existing

SBL methods. We provide simulation results with 1D synthetic signals and 2D images to verify our analysis and compare the proposed method with state-of-the-art ones.

6.1 G-STG Prior

6.1.1 Mathematical Formulation

We introduce the hierarchical Gaussian shifted-truncated-gamma (G-STG) prior for a sparse signal $\boldsymbol{x} \in \mathbb{R}^N$ as follows:

$$p\left(\boldsymbol{x}|\boldsymbol{\alpha}\right) = \mathcal{N}\left(\boldsymbol{x}|\mathbf{0}, \boldsymbol{\Lambda}\right), \tag{6.1}$$

$$p\left(\boldsymbol{\alpha}; \tau, \epsilon, \eta\right) = \prod_{i=1}^{N} p\left(\alpha_i; \tau, \epsilon, \eta\right) = \prod_{i=1}^{N} \Gamma_\tau\left(\alpha_i|\epsilon, \eta\right), \tag{6.2}$$

where $\boldsymbol{\Lambda} = \operatorname{diag}\left(\boldsymbol{\alpha}\right)$, $\boldsymbol{\alpha} \in \mathbb{R}^N$, $p\left(\alpha_i; \tau, \epsilon, \eta\right)$ is a shifted-truncated-gamma (STG) distribution for $\alpha_i \geq 0$ with

$$\Gamma_\tau\left(\alpha_i|\epsilon, \eta\right) = \frac{\eta^\epsilon}{\Gamma_{\eta\tau}\left(\epsilon\right)}\left(\alpha_i + \tau\right)^{\epsilon-1} \exp\left\{-\eta\left(\alpha_i + \tau\right)\right\}, \tag{6.3}$$

$\epsilon \geq 0$ is the shape parameter, $\eta \geq 0$ is the rate parameter, $\tau \geq 0$ is the threshold parameter and $\Gamma_\tau\left(\epsilon\right) = \int_\tau^\infty t^{\epsilon-1}e^{-t}dt$ denotes an incomplete gamma function. The first layer of the prior is a commonly used Gaussian prior that leads to convenient computations as shown later. In the second layer α_i, $i = 1, \cdots, N$, are assumed to be independent, and further $\alpha_i + \tau$, $i = 1, \cdots, N$, are i.i.d. truncated gamma distributed (that is why we say that $p\left(\alpha_i; \tau, \epsilon, \eta\right)$ is an STG distribution). By $p\left(\alpha_i\right) \propto \left(\alpha_i + \tau\right)^{\epsilon-1} \exp\left\{-\eta\alpha_i\right\}$, $i = 1, \cdots, N$, it is obvious that α_i is favored to be zero in the second layer if $\epsilon \leq 1$, resulting in that x_i is favored to be zero. Thus the hierarchical prior is a sparsity-inducing prior. In general, there is no explicit expression for the marginal distribution

$$\begin{aligned} p\left(\boldsymbol{x};\tau,\epsilon,\eta\right) &= \prod_{i=1}^{N} p\left(x_i;\tau,\epsilon,\eta\right) \\ &= \prod_{i=1}^{N} \int_0^{\infty} \mathcal{N}\left(x_i|0,\alpha_i\right)\Gamma_\tau\left(\alpha_i|\epsilon,\eta\right)d\alpha_i. \end{aligned} \tag{6.4}$$

In the following we study some special cases and show that the G-STG prior generalizes those in [46, 54].

1) $\epsilon = 1$*:* In this case, the second layer is reduced to an exponential prior independent of τ since $\Gamma_\tau\left(\alpha_i|1,\eta\right) = \eta e^{-\eta\alpha_i}$. Then, the G-STG prior coincides with the Laplace prior in [46] with $p\left(x_i;\tau,1,\eta\right) = \sqrt{\eta/2}\exp\left(-\sqrt{2\eta}\left|x_i\right|\right)$.

2) $\tau = 0$*:* The second layer becomes a gamma prior for each α_i. As a result, the proposed prior becomes the Gaussian-gamma prior in [54] and $p\left(x_i\right) = \frac{2^{3/4-\epsilon/2}}{\sqrt{\pi}\Gamma(\epsilon)}\eta^{\frac{2\epsilon+1}{4}}\left|x_i\right|^{\epsilon-\frac{1}{2}}\mathcal{K}_{\epsilon-\frac{1}{2}}\left(\sqrt{2\eta}\left|x_i\right|\right)$ where $\mathcal{K}_\nu\left(\cdot\right)$ is the modified Bessel function of the second kind and order $\nu \in \mathbb{R}$. In addition, we have that $p\left(0\right) = +\infty$ if $\epsilon \leq \frac{1}{2}$ and $p\left(0\right) < +\infty$ if $\epsilon > \frac{1}{2}$. Though the G-STG prior generalizes the Gaussian-gamma prior, it should be noted that the main algorithm based on the G-STG prior proposed in this chapter works differently from that in [54].

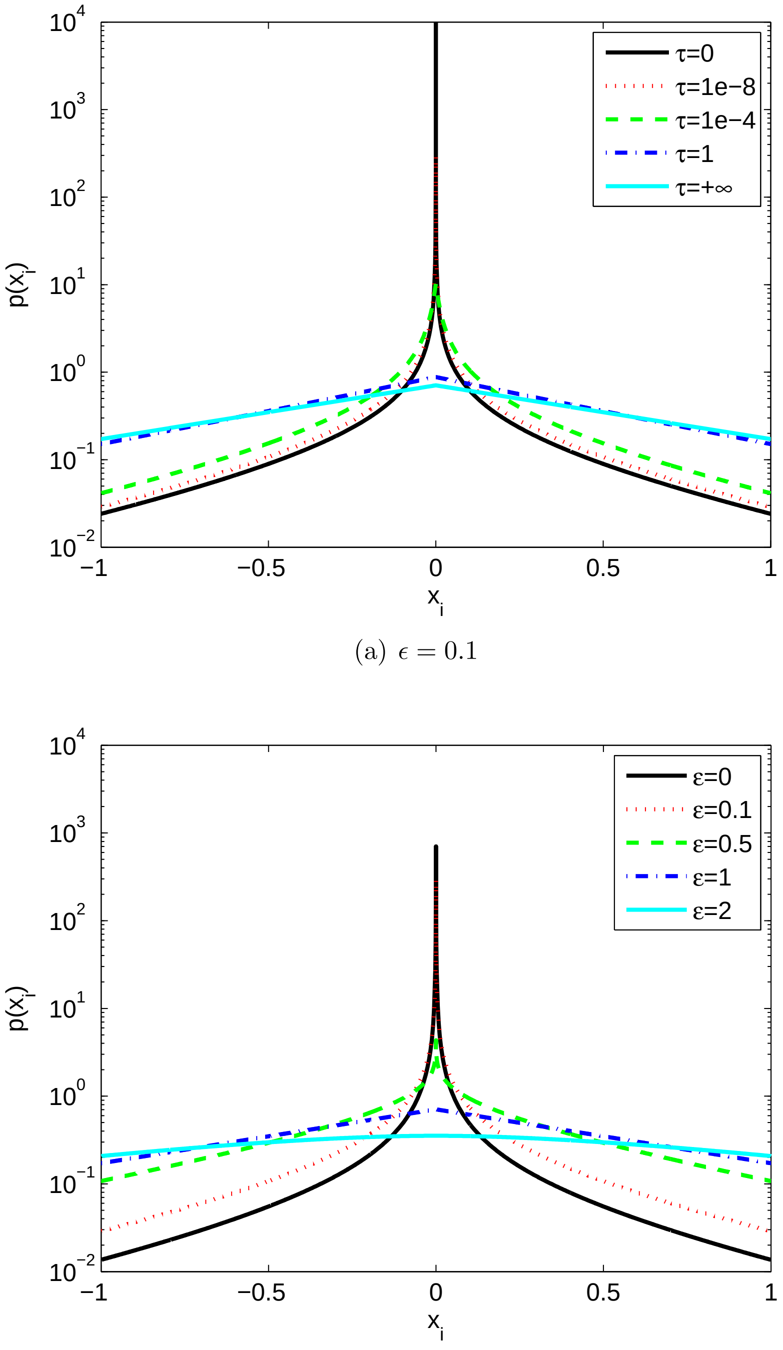
3) $\tau \to +\infty$*:* By l'Hospital's rule it can be shown that $\lim_{\tau\to+\infty}\Gamma_\tau\left(\alpha_i|\epsilon,\eta\right) = \eta e^{-\eta\alpha_i}$, i.e., the prior for α_i in the second layer approaches an exponential prior in such a case. Consequently, the proposed G-STG prior coincides with the Laplace prior as in *Case 1*.

To visualize the variation of the G-STG prior with respect to the two parameters τ and ϵ, we plot the PDF $p\left(x_i;\tau,\epsilon,\eta\right)$ in Fig. 6.1 with $\eta = 1$. Fig. 6.1(a) is for the case of $\epsilon = 0.1$ and varying τ. Obviously, the G-STG prior is a sparsity-inducing prior with the PDFs highly peaked at the origin, especially when $\tau \to 0$. The main difference between the cases $\tau = 0$ and $\tau = 1 \times 10^{-8}$ is near the origin where $p\left(x_i\right)$ approaches infinity for $\tau = 0$ while it is always finite for $\tau > 0$. As τ gets larger, less density concentrates near the origin and the resulting prior gets closer to the Laplace prior that corresponds to $\tau = +\infty$. Fig. 6.1(b) is for the case of $\tau = 1 \times 10^{-8}$ and varying ϵ. It is shown that the G-STG prior gets less sparsity-inducing as ϵ gets larger, and that it ceases to promote sparsity as $\epsilon > 1$. From the perspective of MAP estimation, the G-STG prior corresponds to a nonconvex optimization method as $\epsilon < 1$ since the term $-\log p(\boldsymbol{x})$ is nonconvex in such a case.

6.1.2 Intuitive Interpretation and Threshold Parameter Setting

Strictly speaking, a continuous prior is not suitable for sparse signals since any vector generated from a continuous prior is only approximately sparse (the probability of a zero-valued entry is zero). In the following, we provide an intuitive explanation about why the G-STG prior works for sparse signals by setting the threshold parameter τ according to the noise level.

We consider a Gaussian ensemble sensing matrix $\boldsymbol{A}$ (the entries of $\boldsymbol{A}$ are i.i.d. Gaussian $\mathcal{N}(0, M^{-1})$ where the variance is set to M^{-1} to make columns of $\boldsymbol{A}$ have expected unit norm). Then, consider a compressible signal $\boldsymbol{z}$ and the observation model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{z}$. In the Bayesian framework, we assume that $\boldsymbol{z}$ is distributed according to some sparsity-inducing prior. Here we adopt the Gaussian-gamma prior in [54], i.e., we assume that $p(\boldsymbol{z}|\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{\mathcal{B}})$ and $p(\boldsymbol{\beta}; \epsilon, \eta) = \prod_{i=1}^{N} \Gamma(\beta_i|\epsilon, \eta)$ where $\boldsymbol{\mathcal{B}} = \operatorname{diag}(\boldsymbol{\beta})$ and $\Gamma(\beta_i|\epsilon, \eta)$ refers to $\Gamma_\tau(\beta_i|\epsilon, \eta)$ with $\tau = 0$. However, theoretical results [23, 58] state that only significant entries of $\boldsymbol{z}$ can be recovered while insignificant ones play as noises. So we write $\boldsymbol{z}$ into $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{w}$ where $\boldsymbol{x}$ denotes the significant, "recoverable" component and $\boldsymbol{w}$ refers to the insignificant, "unrecoverable" part. Then the observation model becomes $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x} + \boldsymbol{w}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ with $\boldsymbol{e} = \boldsymbol{A}\boldsymbol{w}$. By the structure of $\boldsymbol{A}$, $\boldsymbol{e}$ is a zero-mean AWGN with the noise variance $\sigma^2 = M^{-1}\|\boldsymbol{w}\|_2^2$. We see that only the power of $\boldsymbol{w}$ is reflected in the noise. That is, for any vector $\widetilde{\boldsymbol{w}}$ satisfying $\|\widetilde{\boldsymbol{w}}\|_2^2 = \|\boldsymbol{w}\|_2^2$, $\boldsymbol{A}\widetilde{\boldsymbol{w}}$ and $\boldsymbol{e}$ are identically distributed. So we may replace $\boldsymbol{w}$ by $\widetilde{\boldsymbol{w}}$ in the observation model (i.e., $\boldsymbol{x}+\widetilde{\boldsymbol{w}}$ and $\boldsymbol{z}$ are indistinguishable for CS approaches). Then we may model $\widetilde{\boldsymbol{w}}$ as an i.i.d. zero-mean Gaussian



(a) $\epsilon = 0.1$

(b) $\tau = 1 \times 10^{-8}$

Figure 6.1: PDFs of the G-STG prior in the case of (a) $\epsilon = 0.1$ and varying τ and (b) $\tau = 1 \times 10^{-8}$ and varying ϵ with $\eta = 1$.

vector with variance $\tau = N^{-1} \|\boldsymbol{w}\|_2^2 = (M/N)\,\sigma^2$. In addition, $\widetilde{\boldsymbol{w}}$ is independent of $\boldsymbol{x}$. So, under the assumption of a Gaussian ensemble matrix $\boldsymbol{A}$, a compressible signal $\boldsymbol{z}$ is equivalent to the sum of its significant part $\boldsymbol{x}$ plus a white Gaussian noise $\widetilde{\boldsymbol{w}}$. Applying the Gaussian-gamma prior for $\boldsymbol{z}$ to $\boldsymbol{x} + \widetilde{\boldsymbol{w}}$, we obtain that $\beta_i = \alpha_i + \tau$, $i = 1, \cdots, N$, where $\boldsymbol{\alpha}$ is as defined in (6.1). Then we get $p(\alpha_i) = p(\beta_i - \tau | \beta_i \geq \tau)$ as a conditional distribution, resulting in that $p(\boldsymbol{\alpha})$ is in the exact form of (6.2).

In this chapter, we mainly consider the observation model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ where $\boldsymbol{x}$ is a sparse signal and $\boldsymbol{e}$ is a zero-mean AWGN with known variance σ^2. The same sparsity-inducing prior for $\boldsymbol{x}$ can be obtained by a reverse procedure and the details are omitted. So we can set the threshold parameter $\tau = (M/N)\,\sigma^2$ in the G-STG prior. Though this setting is only based on intuition without rigorous analysis, it indeed leads to good performance as to be reported via simulations in Section 6.4, where it is also observed that this setting applies to other sensing matrix ensembles besides the Gaussian one.

6.2 Sparse Bayesian Learning for Signal Recovery

6.2.1 Bayesian Formulation

In SBL, the signal of interest and noise are modeled as random variables. Under the common assumption of zero-mean AWGNs, i.e., $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, where σ^2 is the noise variance, the PDF of the compressive measurements is

$$p\left(\boldsymbol{y}|\boldsymbol{x}; \sigma^2\right) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}, \sigma^2 \boldsymbol{I}\right). \tag{6.5}$$

In this chapter we assume that the noise variance σ^2 is known *a priori*. This assumption has been widely made in the CS literature, e.g., [58, 125]. Moreover, it is shown in [41] that to estimate σ^2 jointly with the signal recovery process (e.g., in [39]) can lead to very inaccurate estimate.

The G-STG prior introduced in Section 6.1 is adopted as the sparsity-inducing prior for the sparse signal $\boldsymbol{x}$. The hyperparameters τ and ϵ are chosen manually according to the reasoning in Section 6.1 and Subsection 6.3.2. Numerical simulations will be provided in Section 6.4 to illustrate their performance. To estimate η from the measurements, we assume a gamma hyperprior for η: $p\left(\eta; c, d\right) = \Gamma\left(\eta|c, d\right)$, where we let $c, d \to 0$ to obtain a uniform hyperprior (over a logarithmic scale). So the joint PDF of the observation model is $p\left(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\alpha}, \eta; \sigma^2, \tau, \epsilon, c, d\right) = p\left(\boldsymbol{y}|\boldsymbol{x}; \sigma^2\right) p\left(\boldsymbol{x}|\boldsymbol{\alpha}\right) p\left(\boldsymbol{\alpha}|\eta; \tau, \epsilon\right) p\left(\eta; c, d\right)$, where $\boldsymbol{y}$ is the observation, $\boldsymbol{x}$ is the unknown signal of interest, $\boldsymbol{\alpha}$ and η are unknown parameters, and σ^2, τ, ϵ, c and d are fixed. The task is to estimate $\boldsymbol{x}$.

6.2.2 Bayesian Inference

Note that the exact Bayesian inference is intractable since $p\left(\boldsymbol{x}|\boldsymbol{y}\right)$ is computationally intractable. Some approximations have to be made. Following from [39], we decompose the posterior $p\left(\boldsymbol{x}, \boldsymbol{\alpha}, \eta|\boldsymbol{y}\right)$ into two terms as

$$p\left(\boldsymbol{x}, \boldsymbol{\alpha}, \eta|\boldsymbol{y}\right) = p\left(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\alpha}\right) p\left(\boldsymbol{\alpha}, \eta|\boldsymbol{y}\right). \tag{6.6}$$

The first term $p\left(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\alpha}\right)$ is the posterior for $\boldsymbol{x}$ given the hyperparameter $\boldsymbol{\alpha}$, which will be later shown to be a Gaussian PDF. Then, we compute the most probable estimates of $\boldsymbol{\alpha}$ and η, say $\boldsymbol{\alpha}_{MP}$ and η_{MP}, that maximize the second term $p\left(\boldsymbol{\alpha}, \eta|\boldsymbol{y}\right)$. We use $\boldsymbol{\alpha}_{MP}$ to obtain the posterior for $\boldsymbol{x}$. From the perspective of signal estimation, this is equivalent to requiring

$$p\left(\boldsymbol{x}|\boldsymbol{y}\right) = \int p\left(\boldsymbol{x}, \boldsymbol{\alpha}, \eta|\boldsymbol{y}\right) d\boldsymbol{\alpha}\, d\eta \approx p\left(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\alpha}_{MP}\right), \tag{6.7}$$

where $p\left(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\alpha}_{MP}\right)$ refers to $p\left(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\alpha}\right)$ at $\boldsymbol{\alpha}_{MP}$. Similar procedures have been adopted in [40,46]. We will provide theoretical evidence to show that this approach leads to desirable properties in Section 6.2.3. Simulation results presented in Section 6.4 also suggest that the signal recovery based on this approximation is very effective.

Since $p(\boldsymbol{y}|\boldsymbol{x})$ and $p(\boldsymbol{x}|\boldsymbol{\alpha})$ are both Gaussian, it is easy to show that the posterior for $\boldsymbol{x}$ and the marginal distribution for $\boldsymbol{y}$ are both Gaussian with $p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, $p(\boldsymbol{y}|\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0},\boldsymbol{C})$, where

$$\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{y}, \tag{6.8}$$

$$\boldsymbol{\Sigma} = \left(\sigma^{-2}\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{\Lambda}^{-1}\right)^{-1}, \tag{6.9}$$

$$\boldsymbol{C} = \sigma^2\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^T. \tag{6.10}$$

The maximization of $p(\boldsymbol{\alpha},\eta|\boldsymbol{y})$ is equivalent to that of $p(\boldsymbol{y},\boldsymbol{\alpha},\eta) = p(\boldsymbol{y}|\boldsymbol{\alpha})\,p(\boldsymbol{\alpha}|\eta)$ $p(\eta)$ by the relation $p(\boldsymbol{\alpha},\eta|\boldsymbol{y}) = p(\boldsymbol{y},\boldsymbol{\alpha},\eta)/p(\boldsymbol{y})$. We consider $\log\eta$ as the hidden variable instead of η since the uniform hyperprior is assumed over a logarithmic scale. By $p(\log\eta) = \eta p(\eta)$ we see that the hyperprior for η leads to a noninformative prior by setting $c = d = 0$. So the log-likelihood function is

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\alpha},\log\eta)\\
&= \log p(\boldsymbol{y},\boldsymbol{\alpha},\log\eta)\\
&= -\frac{1}{2}\log|\boldsymbol{C}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y}\\
&\quad + (\epsilon-1)\sum_{i=1}^{N}\log(\alpha_i+\tau) - \eta\sum_{i=1}^{N}(\alpha_i+\tau)\\
&\quad + (N\epsilon+c)\log\eta - N\log\Gamma_{\eta\tau}(\epsilon) - d\eta + C_1,
\end{aligned} \tag{6.11}$$

where C_1 is a constant. The maximizer of $\mathcal{L}$ will be analyzed in Subsection 6.2.3. In the following we provide an iterative procedure to maximize $\mathcal{L}$ by recognizing the identities $\log|\boldsymbol{C}| = \log|\boldsymbol{\Lambda}| + M\log\sigma^2 - \log|\boldsymbol{\Sigma}|$ and $\boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y} = \sigma^{-2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}\|_2^2 + \boldsymbol{\mu}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu}$.

6.2.2.1 Update of $\boldsymbol{\alpha}$

For α_i, $i = 1, \cdots, N$, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \alpha_i} &= -\frac{1}{2\alpha_i} + \frac{E\left\{x_i^2\right\}}{2\alpha_i^2} + \frac{\epsilon - 1}{\alpha_i + \tau} - \eta \\ &= -\frac{f\left(\alpha_i\right)}{2\alpha_i^2\left(\alpha_i + \tau\right)}, \end{aligned} \tag{6.12}$$

where $f\left(t\right) = 2\eta t^3 + \left(3 - 2\epsilon + 2\eta\tau\right) t^2 + \left(\tau - E\left\{x_i^2\right\}\right) t - \tau E\left\{x_i^2\right\}$ for $t \in \mathbb{R}$ is a cubic function and $E\left\{x_i^2\right\} = \mu_i^2 + \Sigma_{ii}$, where Σ_{ii} is the ith diagonal entry of $\boldsymbol{\Sigma}$. We need the following lemma.

Lemma 6.1. *For a cubic function $g\left(t\right) = \lambda_1 t^3 + \lambda_2 t^2 + \lambda_3 t + \lambda_4$, if $\lambda_1, \lambda_2 > 0$ and $\lambda_4 < 0$, then $g\left(t\right) = 0$ has a unique root on $\left(0, +\infty\right)$.*

Proof. By $g(0) = \lambda_4 < 0$ and $\lim_{t \to +\infty} g\left(t\right) = +\infty$ there exists at least one root in $(0, +\infty)$. We show that this root is unique using contradiction. Suppose there exists more than one positive root. Then there must exist three positive roots and that $g\left(t\right)$ has two positive stationary points. That is, the two solutions of $\frac{dg(t)}{dt} = 3\lambda_1 t^2 + 2\lambda_2 t + \lambda_3 = 0$ are both positive, resulting in that $\lambda_2 < 0$ (contradiction). ∎

By Lemma 6.1, it is easy to show that the maximum of $\mathcal{L}$ is achieved at the unique positive root, say $\alpha_i^* > 0$, of $f\left(\alpha_i\right) = 0$. We note that explicit expressions are available for the roots of a cubic function and hence α_i, $i = 1, \cdots, N$, can be efficiently updated.

6.2.2.2 Update of η

In general, there is no explicit expression for updating η. Since the first and second derivatives of $\mathcal{L}$ with respect to $\log \eta$ can be easily computed, $\mathcal{L}$ can be efficiently maximized with respect to $\log \eta$ using numerical methods, e.g., gradient ascending method or Newton's method. In addition, the computational complexity hardly depends on the CS problem dimension.

Remark 6.1. *The computation of the incomplete gamma function $\Gamma_{\eta\tau}(\epsilon)$ is involved in the update of η. This term can be efficiently computed using functions provided in Matlab if ϵ is properly bounded away from zero. But a numerical integration is needed if $\epsilon = 0$. In Section 6.4, we observe through simulations that the update of η may take considerably long time in the case of $\epsilon = 0$. But such differences are negligible in the case of a high-dimensional CS problem since the computation of η hardly depends on the problem dimension unlike other computations, such as the update of $\boldsymbol{\alpha}$.*

As a result, an iterative algorithm can be implemented to obtain $\boldsymbol{\alpha}_{MP}$ and η_{MP} by iteratively updating $\boldsymbol{\Sigma}$ in (6.9), $\boldsymbol{\mu}$ in (6.8), $\boldsymbol{\alpha}$ and η. It is easy to show that this algorithm can be implemented using an EM algorithm [126]. So the likelihood $\mathcal{L}$ increases monotonically at each iteration and the algorithm is guaranteed to converge. After convergence, the signal $\boldsymbol{x}$ is estimated using its posterior mean $\boldsymbol{\mu}$. One shortcoming of the iterative algorithm is that at each iteration a high-dimensional matrix inversion has to be calculated for updating $\boldsymbol{\Sigma}$ though this computation can be possibly alleviated using the Woodbury matrix identity.

6.2.3 Analysis of Global and Local Maxima

We analyze the global and local maxima of the likelihood $\mathcal{L}$ in (6.11) in this subsection. Our analysis is rooted in [40] and shows that the theoretical results on the basic SBL in [40] can be extended to our case with necessary modifications. In the following, we assume that $c = d = 0$ and $\eta > 0$ is fixed (it is a similar case if η is chosen to maximize $\mathcal{L}$ as well). Thus we may write $\mathcal{L}$ (with respect to $\boldsymbol{\alpha}$) as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha}) = &-\frac{1}{2}\log|\boldsymbol{C}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y} \\ &+ (\epsilon - 1)\sum_{i=1}^{N}\log(\alpha_i + \tau) - \eta\sum_{i=1}^{N}\alpha_i + C_4,\end{aligned} \tag{6.13}$$

where C_4 is a constant independent of $\boldsymbol{\alpha}$. We first consider the global maxima in the noise free case.

Theorem 6.1. *Let $\tau \geq 0$, $0 \leq \epsilon \leq 1$, $\boldsymbol{e} = 0$ and $\sigma^2 = 0$. Assume that $\boldsymbol{x}^0$ satisfying $\|\boldsymbol{x}^0\|_0 < M < N$ is the maximally sparse solution to the linear equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. Then, there exists some $\boldsymbol{\alpha}^0$ with $\|\boldsymbol{\alpha}^0\|_2 < +\infty$ and $\|\boldsymbol{\alpha}^0\|_0 = \|\boldsymbol{x}^0\|_0$ such that at $\boldsymbol{\alpha}^0$, $\mathcal{L}$ is globally maximized and the corresponding posterior mean is $\boldsymbol{\mu} = \boldsymbol{x}^0$.*

Proof. Note first that $\lim_{\sigma^2 \to 0} \boldsymbol{\mu} = \boldsymbol{\Lambda}^{\frac{1}{2}} \left(\boldsymbol{A}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{y} = \boldsymbol{\Lambda}^{\frac{1}{2}} \left(\boldsymbol{A}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{A}\boldsymbol{x}^0$ by (6.8). Since $\|\boldsymbol{x}^0\|_0 < M$ there must exist $\boldsymbol{\alpha}^0 \in \mathbb{R}^N$, $\|\boldsymbol{\alpha}^0\|_2 < +\infty$ and $\|\boldsymbol{\alpha}^0\|_0 = \|\boldsymbol{x}^0\|_0$, such that $\boldsymbol{\mu} = \boldsymbol{x}^0$ at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^0$. In fact, $\boldsymbol{\alpha}^0$ can be any vector satisfying that for $j = 1, \cdots, N$, $\alpha_j^0 > 0$ if $x_j^0 \neq 0$, and $\alpha_j^0 = 0$ otherwise. In the following we show that the global maximum of $\mathcal{L}$ is achieved at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^0$. Let $C_5 = \|\boldsymbol{\alpha}^0\|_\infty < \infty$ and denote $C_6 > 0$ the minimum nonzero α_i, $i = 1, \cdots, N$. Following from the proof of [40, Theorem 1], we have $|\boldsymbol{C}| = \left|\sigma^2 \boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^T\right| \to 0$ and $\boldsymbol{y}^T \boldsymbol{C}^{-1} \boldsymbol{y} \leq \|\boldsymbol{x}^0\|_2^2 / C_6$ if $\boldsymbol{\alpha} \to \boldsymbol{\alpha}^0$. In addition, we have $\sum_{i=1}^N \log(\alpha_i + \tau) \leq N \log(C_5 + \tau)$ and $\sum_{i=1}^N \alpha_i \leq NC_5$. So we have $\mathcal{L} = +\infty$ at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^0$, which completes the proof. ∎

Theorem 6.1 shows that the global maximum of the objective function is achieved at the maximally sparse solution. Thus the proposed approach has no structural errors and the remaining question of whether the algorithm can produce this solution is a convergence issue. Further, we have the following result which is similar to that in [40, Theorem 2].

Theorem 6.2. *Let $\tau \geq 0$ and $0 \leq \epsilon \leq 1$. Every local maximum of $\mathcal{L}$ is achieved at a sparse $\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\|_0 \leq M$ that leads to a sparse posterior mean $\boldsymbol{\mu}$ with $\|\boldsymbol{\mu}\|_0 \leq M$, regardless of the existence of noise.*

Proof. Note that $\|\boldsymbol{\mu}\|_0 \leq \|\boldsymbol{\alpha}\|_0$ since $\mu_i \to 0$ as $\alpha_i \to 0$, $i = 1, \cdots, N$. So we need only to show that every local maximum of $\mathcal{L}$ is achieved at a sparse $\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\|_0 \leq M$. Let $q(\boldsymbol{\alpha}) = -\frac{1}{2} \log |\boldsymbol{C}| + (\epsilon - 1) \sum_{i=1}^N \log(\alpha_i + \tau) - \eta \sum_{i=1}^N \alpha_i$, which is convex with respect to $\boldsymbol{\alpha} \in \mathbb{R}_+^N$ if $\epsilon \leq 1$. Suppose that $\boldsymbol{\alpha}^*$ is a local maximum point of $\mathcal{L}$. We may construct a closed, bounded convex polytope $\mathcal{P} \subset \mathbb{R}_+^N$ following from [40] (we omit the details) such that $\boldsymbol{\alpha}^* \in \mathcal{P}$ and if $\boldsymbol{\alpha} \in \mathcal{P}$, then the second term of $\mathcal{L}$, $-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y}$, equals a constant $C_7 = -\frac{1}{2}\boldsymbol{y}^T\left(\sigma^2\boldsymbol{I} + \boldsymbol{A}\text{diag}\left(\boldsymbol{\alpha}^*\right)\boldsymbol{A}^T\right)^{-1}\boldsymbol{y}$. In addition, all extreme points of $\mathcal{P}$ are sparse with support size no more than M. As a result, $\boldsymbol{\alpha}^*$ is a local maximum point of $q\left(\boldsymbol{\alpha}\right)$ with respect to $\boldsymbol{\alpha} \in \mathcal{P}$. By the convexity of $q\left(\boldsymbol{\alpha}\right)$, $\boldsymbol{\alpha}^*$ must be an extreme point of $\mathcal{P}$ with $\left\|\boldsymbol{\alpha}^*\right\|_0 \leq M$. ■

Theorem 6.2 states that all local maxima of $\mathcal{L}$ are achieved at sparse solutions. Since we can locally maximize $\mathcal{L}$ efficiently in practice, based on Theorem 6.2, we introduce a fast algorithm in Section 6.3 that searches for a sparse solution that locally maximizes $\mathcal{L}$.

6.3 Fast Algorithm

6.3.1 Fast Greedy Algorithm

Based on Theorem 6.2 in Section 6.2.3, the following algorithm aims to find a sparse solution that locally maximizes $\mathcal{L}$. We consider the contribution of a single basis vector $\boldsymbol{a}_j$ (the jth column of $\boldsymbol{A}$), $j = 1, \cdots, N$, and determine whether it should be included in the model (or in the active set) for the maximization of $\mathcal{L}$. Denote $\boldsymbol{C}_{-j} = \sigma^2\boldsymbol{I} + \sum_{i\neq j}\alpha_i\boldsymbol{a}_i\boldsymbol{a}_i^T = \boldsymbol{C} - \alpha_j\boldsymbol{a}_j\boldsymbol{a}_j^T$ that is independent of the jth basis vector $\boldsymbol{a}_j$. Then we have $\left|\boldsymbol{C}\right| = \left|\boldsymbol{C}_{-j}\right|\left|1 + \alpha_j\boldsymbol{a}_j^T\boldsymbol{C}_{-j}^{-1}\boldsymbol{a}_j\right|$ and $\boldsymbol{C}^{-1} = \boldsymbol{C}_{-j}^{-1} - \left(\alpha_j^{-1} + \boldsymbol{a}_j^T\boldsymbol{C}_{-j}^{-1}\boldsymbol{a}_j\right)^{-1}\boldsymbol{C}_{-j}^{-1}\boldsymbol{a}_j\boldsymbol{a}_j^T\boldsymbol{C}_{-j}^{-1}$. Using the two identities above, we rewrite the log-likelihood function (with respect to $\boldsymbol{\alpha}$) into

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\alpha}\right) &= -\frac{1}{2}\log\left|\boldsymbol{C}_{-j}\right| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{C}_{-j}^{-1}\boldsymbol{y} \\ &\quad + \left(\epsilon - 1\right)\sum_{i=1}^{N}\log\left(\alpha_i + \tau\right) - \eta\sum_{i=1}^{N}\alpha_i \\ &\quad - \frac{1}{2}\log\left|1 + \alpha_j\boldsymbol{a}_j^T\boldsymbol{C}_{-j}^{-1}\boldsymbol{a}_j\right| + \frac{q_j^2}{2\left(\alpha_j^{-1} + s_j\right)} + C_2 \\ &= \mathcal{L}\left(\boldsymbol{\alpha}_{-j}\right) + \ell\left(\alpha_j\right) + C_3,\end{aligned} \tag{6.14}$$

where $\mathcal{L}(\boldsymbol{\alpha}_{-j})$ denotes $\mathcal{L}(\boldsymbol{\alpha})$ after removing the contribution of the jth basis vector, $\ell(\alpha_j) = -\frac{1}{2}\log|1+\alpha_j s_j| + \frac{q_j^2}{2\left(\alpha_j^{-1}+s_j\right)} + (\epsilon-1)\log\left(\frac{\alpha_j}{\tau}+1\right) - \eta\alpha_j$ with $s_j = \boldsymbol{a}_j^T\boldsymbol{C}_{-j}^{-1}\boldsymbol{a}_j$ and $q_j = \boldsymbol{a}_j^T\boldsymbol{C}_{-j}^{-1}\boldsymbol{y}$, and C_2, C_3 are constants independent of $\boldsymbol{\alpha}$. Note that $\ell(\alpha_i)$ has been modified by a constant such that $\ell(0)=0$. In the following we compute the maximum point, say α_j^*, of $\ell(\alpha_j)$ on $[0,+\infty)$. If $\alpha_j^*>0$, then the basis vector $\boldsymbol{a}_j$ should be preserved in the active set since it gives positive contribution to the likelihood. Otherwise, it should be removed from the active set (by setting $\alpha_j=0$). We consider only the general case $\tau>0$, $0\le\epsilon\le1$ and $\eta>0$ since it is simple for other cases. First we have

$$\begin{aligned}\frac{d\,\ell(\alpha_j)}{d\,\alpha_j} &= -\frac{s_j}{2(1+\alpha_j s_j)} + \frac{q_j^2}{2(1+\alpha_j s_j)^2} + \frac{\epsilon-1}{\alpha_j+\tau} - \eta \\ &= -\frac{h(\alpha_j)}{2(\alpha_j+\tau)(1+\alpha_j s_j)^2},\end{aligned} \tag{6.15}$$

where $h(\alpha_j) = c_1\alpha_j^3 + c_2\alpha_j^2 + c_3\alpha_j + c_4$ is a cubic function of α_j with

$$c_1 = 2\eta s_j^2, \tag{6.16}$$

$$c_2 = (3-2\epsilon)s_j^2 + 4\eta s_j + 2\eta\tau s_j^2, \tag{6.17}$$

$$c_3 = (5-4\epsilon)s_j + 2\eta - q_j^2 + \tau\left(4\eta s_j + s_j^2\right), \tag{6.18}$$

$$c_4 = 2 - 2\epsilon + \tau\left(s_j + 2\eta - q_j^2\right). \tag{6.19}$$

To compute the maximum point of $\ell(\alpha_j)$, we need the following result.

Lemma 6.2 ([127]). *For a cubic function $g(t) = \lambda_1 t^3 + \lambda_2 t^2 + \lambda_3 t + \lambda_4$, let the discriminant $\Delta = 18\lambda_1\lambda_2\lambda_3\lambda_4 - 4\lambda_2^3\lambda_4 + \lambda_2^2\lambda_3^2 - 4\lambda_1\lambda_3^3 - 27\lambda_1^2\lambda_4^2$. If $\Delta>0$, then $g(t)=0$ has three distinct real roots. If $\Delta=0$, then the equation has three real roots including a multiple root. If $\Delta<0$, then the equation has one real root and two complex conjugate roots.*

Note that $s_j>0$ for $j=1,\cdots,N$ since $\boldsymbol{C}_{-j}$ and $\boldsymbol{C}_{-j}^{-1}$ are positive definite, and then $c_1, c_2 > 0$. We divide our discussions into three scenarios.

6.3.1.1 $c_4 < 0$

This case is the same as the update of $\boldsymbol{\alpha}$ in Subsection 6.2.2. By Lemma 6.1, $h\left(\alpha_j\right) = 0$ has a unique solution α_j^* on $(0, +\infty)$. Then it is easy to show that $\ell\left(\alpha_j\right)$ increases monotonically on $\left(0, \alpha_j^*\right]$ and decreases monotonically on $\left(\alpha_j^*, +\infty\right)$. Thus $\ell\left(\alpha_j\right)$ obtains the maximum at $\alpha_j^* > 0$.

6.3.1.2 $c_4 \geq 0$, $c_3 < 0$ and $\Delta > 0$

Let Δ denote the determinant of $h\left(\alpha_j\right)$. We see that $h\left(\alpha_j\right) = 0$ has three real distinct roots by Lemma 6.2 since $\Delta > 0$. The two stationary points of $h\left(\alpha_j\right)$ lie on different sides of the y-axis since $\frac{d\,h(\alpha_j)}{d\,\alpha_j} = 3c_1\alpha_j^2 + 2c_2\alpha_j + c_3$ with $c_3 < 0$. Thus the three roots include one negative root and two distinct positive roots since $h(0) = c_4 \geq 0$. Denote $\alpha_j' > 0$ the largest root. We see that $\ell\left(\alpha_j\right)$ decreases from $\alpha_j = 0$ to some point, then increases until $\alpha_j = \alpha_j'$ and decreases again. As a result, $\ell\left(\alpha_j\right)$ obtains the maximum at 0 or α_j'. So we have

- if $\ell\left(\alpha_j'\right) > 0$, then the maximum point is $\alpha_j^* = \alpha_j' > 0$;
- if $\ell\left(\alpha_j'\right) \leq 0$, then $\alpha_j^* = 0$;

6.3.1.3 $c_4 \geq 0$, and $c_3 \geq 0$ or $\Delta \leq 0$

In this scenario, the maximum of $\ell\left(\alpha_j\right)$ is obtained at $\alpha_j^* = 0$. We divide our discussions into three cases: *1)* $\Delta < 0$, *2)* $\Delta = 0$, and *3)* $\Delta > 0$ and $c_3 \geq 0$. In *Case 1*, $h\left(\alpha_j\right) = 0$ has only one (negative) real root. In *Case 2*, the equation has three negative roots (two of them coincide), or one negative root and a multiple positive root. In *Case 3* the equation has three distinct negative roots. So in all the cases $\ell'\left(\alpha_j\right) \leq 0$ for $\alpha_j \geq 0$, resulting in that $\ell\left(\alpha_j\right)$ decreases monotonically on $[0, +\infty)$.

Based on the analysis above, we can compute efficiently α_j^* (given s_j and q_j) at which the likelihood is maximized with respect to a single basis vector $\boldsymbol{a}_j$, $j = 1, \cdots, N$.

So, the likelihood consistently increases if we update a single α_j at one time with the basis $\boldsymbol{a}_j$, $j = 1, \cdots, N$, properly chosen. As a result, a greedy algorithm can be implemented. At the beginning, no basis vectors are included in the model (i.e., the active set is empty or all $\alpha_j = 0$). Then choose a vector $\boldsymbol{a}_j$ at each iteration such that it gives the largest likelihood increment by updating α_j from its current value α_j^0 to the maximum point α_j^*. If $\alpha_j^0 = 0$ and $\alpha_j^* > 0$, then the basis vector $\boldsymbol{a}_j$ is added to the model with $\alpha_j = \alpha_j^*$. If $\alpha_j^0 > 0$ and $\alpha_j^* > 0$, then α_j is re-estimated in the model. If $\alpha_j^0 > 0$ and $\alpha_j^* = 0$, then $\boldsymbol{a}_j$ is removed from the model with $\alpha_j = 0$. After that, update η as in Subsection 6.2.2. The process is repeated until convergence (that is guaranteed since $\mathcal{L}$ increases monotonically). Based on the results of [57], we see that $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, s_j and q_j, $j = 1, \cdots, N$, can be efficiently updated without matrix inversions. Consequently, the greedy algorithm is computationally efficient.

Remark 6.2. *The main differences between the proposed algorithm and those in [46, 57] are the updates of* $\boldsymbol{\alpha}$ *and* η. *Since roots of a cubic equation have explicit expressions and the update of* η *hardly depends on the problem dimension, the proposed greedy algorithm has the same computational complexity as those in [46, 57] at each iteration.*

6.3.2 Analysis of Basis Selection Condition

We study the basis selection condition of the fast algorithm in more details in this subsection. Based on the analysis in Subsection 6.3.1, we have the following result.

Proposition 6.1. *Suppose that the log-likelihood* $\mathcal{L}$ *has been locally maximized in the fast algorithm. If* $\alpha_j > 0$ *for some basis vector* $\boldsymbol{a}_j$, $j = 1, \cdots, N$, *then* $q_j^2 > \min\left\{s_j + 2\eta + \frac{2-2\epsilon}{\tau}, \left(5 - 4\epsilon\right) s_j + 2\eta + \tau\left(4\eta s_j + s_j^2\right)\right\}$ *where* s_j, q_j *are as defined in Subsection 6.3.1.*

Proof. Note that the inequalities $q_j^2 > s_j + 2\eta + \frac{2-2\epsilon}{\tau}$ and $q_j^2 > \left(5 - 4\epsilon\right) s_j + 2\eta + \tau\left(4\eta s_j + s_j^2\right)$ are equivalent to $c_4 < 0$ and $c_3 < 0$ respectively. Let us suppose that the conclusion does not hold. Then we have $c_4 \geq 0$ and $c_3 \geq 0$. It follows from the analysis in Subsection 6.3.1.3 that the likelihood increases if $\boldsymbol{a}_j$ is removed from the model, leading to contradiction. ■

Remark 6.3. *Proposition 6.1 provides a necessary condition for the basis vectors in the final active set. Note that this condition is generally insufficient since additional requirements are needed in the case of* $c_4 \geq 0$ *(e.g.,* $\Delta > 0$*) according to the analysis in Subsection 6.3.1. In a special case where* $q_j^2 > s_j + 2\eta + \frac{2-2\epsilon}{\tau}$ *holds, we can conclude that* $\boldsymbol{a}_j$ *is in the model according to Subsection 6.3.1.1.*

The basis selection condition concerns the sparsity level of the solution of the fast algorithm. We have illustrated that different settings of τ and ϵ lead to different sparsity-inducing priors in Section 6.1. In the following we will see how the parameters affect the sparsity of the algorithm solution. We discuss the effects of τ and ϵ separately.

Let us first fix $\tau > 0$ and vary $\epsilon \in [0, 1]$. It is easy to show that as ϵ decreases both the terms $s_j + 2\eta + \frac{2-2\epsilon}{\tau}$ and $(5 - 4\epsilon) s_j + 2\eta + \tau \left(4\eta s_j + s_j^2\right)$ increase. Thus the necessary condition in Proposition 6.1 becomes stronger. As a result, the solution of the greedy algorithm will be sparser, which is consistent with the fact that a smaller ϵ leads to a more sparsity-inducing prior as shown in Section 6.1. In the case of $\epsilon = 1$, the inequality turns to be $q_j^2 > s_j + 2\eta$ that coincides with the result in [46]. So, the proposed algorithm will produce a sparser solution than that of [46] by simply setting $\epsilon < 1$.

It is not obvious for the case of fixed $\epsilon < 1$ and varying τ (the case $\epsilon = 1$ has been discussed before) since, as τ decreases, the first term $s_j + 2\eta + \frac{2-2\epsilon}{\tau}$ increases while the second term $(5 - 4\epsilon) s_j + 2\eta + \tau \left(4\eta s_j + s_j^2\right)$ decreases. To make a correct conclusion, we observe that $\ell\left(\alpha_j\right)$ in Subsection 6.3.1 is a strictly increasing function with respect to $\tau > 0$ for any fixed $\alpha_j > 0$ and keeps equal to zero at $\alpha_j = 0$. As a result, as τ decreases, it is less likely that the maximum of $\ell\left(\alpha_j\right)$ is achived at a positive point, resulting in that the solution of the greedy algorithm gets sparser. This is consistent with that a smaller τ leads to a more sparsity-inducing prior as shown in Section 6.1. In fact, the greedy algorithm produces a zero solution if $\tau = 0$ and $\epsilon < 1$ since in such a case the maximum of the likelihood is always achieved at the origin with respect to every basis vector. So intuitively, a smaller τ should be used to obtain a more sparsity-inducing prior but too small τ may lead to inaccuracy for the greedy algorithm. Numerical simulations in Section 6.4 will illustrate that the recommended τ in Subsection 6.1.2 is a good choice.

Remark 6.4. *In the case of $\tau = 0$, the proposed G-STG prior coincides with the Gaussian-gamma prior in [54]. The greedy algorithm using this prior developed in [54] is claimed to follow from the same framework in [57] and maximize the likelihood sequentially. However, it should be noted that the algorithm in [54] does not really maximize the likelihood since, if it does, then it should produce a zero solution as discussed above. Specifically, the authors of [54] compute only a local maximum point of $\ell\left(\alpha_j\right)$ which cannot guarantee to increase the likelihood $\mathcal{L}$ while the global maximum of $\ell\left(\alpha_j\right)$ is always obtained at the origin. Hence the algorithm in [54] is technically incorrect because of the inappropriate basis update scheme. Moreover, the algorithm in [54] has not been shown to provide guaranteed convergence.*

6.4 Numerical Simulations

In this section, we present numerical results to illustrate the performance of the proposed method (we consider only the fast algorithm in Subsection 6.3.1). We consider both one-dimensional synthetic signals and two-dimensional images, and compare with existing methods, including ℓ_1 optimization (BP or BPDN), reweighted (RW-) ℓ_1 optimization [128], StOMP [129], the basic BCS [57] and BCS with the Laplace prior (denoted by Laplace) [46]. BCS, Laplace and the proposed method are SBL methods. ℓ_1 optimization is a convex optimization method. RW-ℓ_1 is related to nonconvex optimization. StOMP is a greedy method. The Matlab codes of BP, StOMP and BCS are obtained from the SparseLab package[1], and that of BPDN is from the

[1] Available at http://sparselab.stanford.edu.

ℓ_1-magic package[2]. When composing this thesis, we have also repeated all simulations by implementing BP and BPDN using a recent CS algorithm SPGL1 [130], which results in slightly different accuracy since the same problem is solved. Its computational speed will also be reported. The code of Laplace is available at https://netfiles.uiuc.edu/dbabacan/www/links.html. The number of iterations is set to 5 for RW-ℓ_1 (i.e., 5 ℓ_1 minimization problems are solved iteratively). To make a fair comparison, we use the same convergence criterion in the proposed method as in BCS and Laplace with the stopping tolerance set to 1×10^{-8}.

The performance metrics adopted include the relative mean squared error (RMSE, calculated by $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 / \|\boldsymbol{x}\|_2^2$), the support size ($\|\widehat{\boldsymbol{x}}\|_0$), the number of iterations (for the three BCS methods) and the CPU time, where $\widehat{\boldsymbol{x}}$ and $\boldsymbol{x}$ denote the recovered and original signals, respectively.

6.4.1 One-Dimensional Synthetic Signals

6.4.1.1 Performance with respect to τ

An explicit determination of τ has been recommended in Subsection 6.1.2. In the following, we show that, indeed, this setting leads to good performance in the signal recovery. In our simulation, we set the signal length $N = 512$ and the number of nonzero entries $K = 20$, and vary the sample size M from 40 to 140 with step size of 5. The nonzero entries of the sparse signal are randomly located with the amplitudes following from a zero-mean unit-variance Gaussian distribution. We consider two matrix ensembles for the sensing matrix $\boldsymbol{A}$, including Gaussian ensemble and uniform spherical ensemble (with columns uniformly distributed on the sphere $\mathbb{S}^{N-1}$). To obtain the desired SNR, zero-mean AWGNs are added to the linear measurements where the noise variance is set to $\sigma^2 = (K/M)\, 10^{-\mathrm{SNR}/10}$. We set $\mathrm{SNR} = 25\,\mathrm{dB}$. The noisy measurements are used in the following signal recovery process. In the proposed algorithm, we set $\epsilon = 0.01$ which results in both fast and accurate recovery

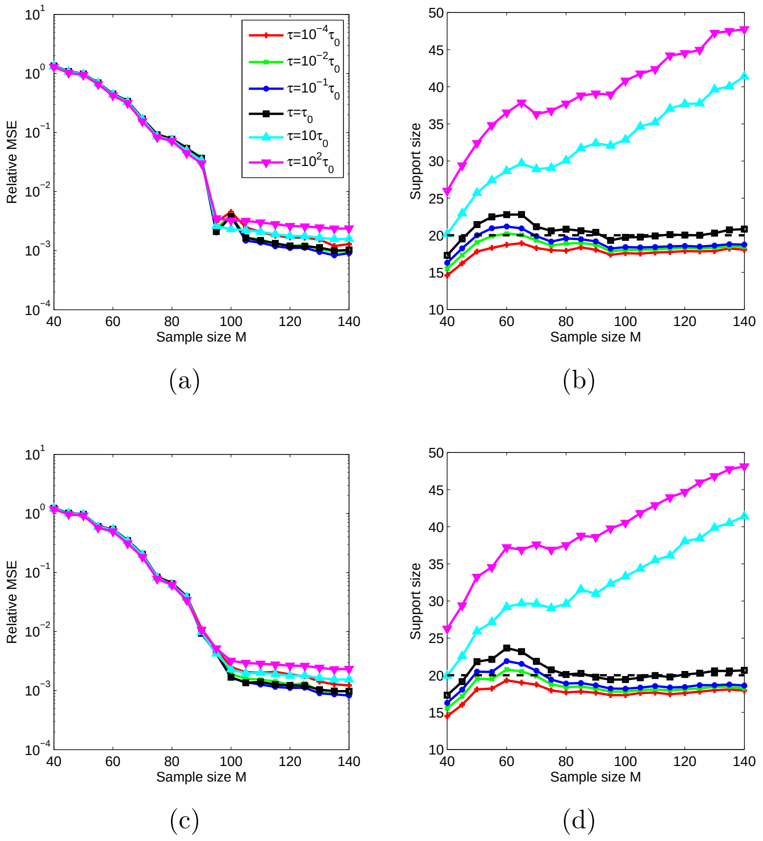
[2] Available at http://users.ece.gatech.edu/~justin/l1magic.

(this will be illustrated in the next experiment). Denote $\tau_0 = (M/N)\,\sigma^2$. We set $\tau = \theta\tau_0$ and consider six values of $\theta = 10^{-4},\ 10^{-2},\ 10^{-1},\ 1,\ 10$ and 100. Thus $\theta = 1$ leads to the recommended value of τ in Subsection 6.1.2. For each M, 100 random problems are generated and solved respectively using the proposed method with different τ. The metrics are averaged results over the 100 trials.

Our simulation results are presented in Fig. 6.2. Fig. 6.2(a) and Fig. 6.2(b) plot the RMSEs and support sizes of the proposed algorithm with Gaussian sensing matrices. It is shown that the recommended τ leads to approximately the smallest error with a reasonable number of measurements while the errors are almost the same when the sample size is small for different τ's. Fig. 6.2(b) shows that the recommended τ results in the most accurate estimation of the support size in most cases. In addition, it is shown that a sparser solution is obtained if a smaller τ is used in the algorithm as expected. Almost identical performance is shown in Fig. 6.2(c) and 6.2(d) by using the uniform spherical ensemble. Thus, we consider only the uniform spherical ensemble in the following experiments.

6.4.1.2 Performance with respect to ϵ

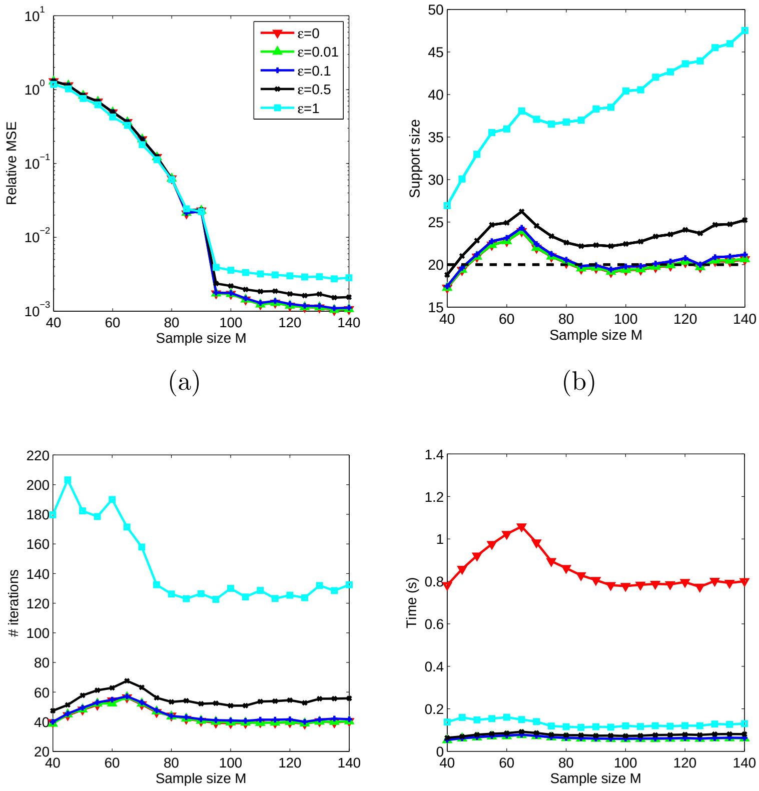
We study now the performance of the proposed algorithm with respect to ϵ. We repeat the simulation above using the recommended τ and consider five values of $\epsilon = 0,\ 0.01,\ 0.1,\ 0.5$ and 1. Note that the case $\epsilon = 1$ corresponds to the Laplace prior. Our simulation results are presented in Fig. 6.3. It is shown in Fig. 6.3(a) that the signal recovery error decays as the sample size increases in general. As the sample size is small the estimation errors differ slightly. But with a reasonable number of measurements a smaller ϵ results in a smaller error. It is shown in Fig. 6.3(b) that a smaller ϵ leads to a sparser solution as expected and more accurate support size estimation. Another advantage of adopting a small ϵ can be observed in Fig. 6.3(c) where it is shown that a smaller ϵ leads to less number of iterations. In general, the time consumption is proportional to the number of iterations since the computational workload is approximately the same at each iteration. Fig. 6.3(d)



(a) (b)

(c) (d)

Figure 6.2: Performance of the proposed algorithm with respect to different settings of τ with (a) & (b) Gaussian ensemble, and (c) & (d) uniform spherical ensemble.



(a) (b)

(c) (d)

Figure 6.3: Performance of the proposed algorithm with respect to different settings of ϵ.

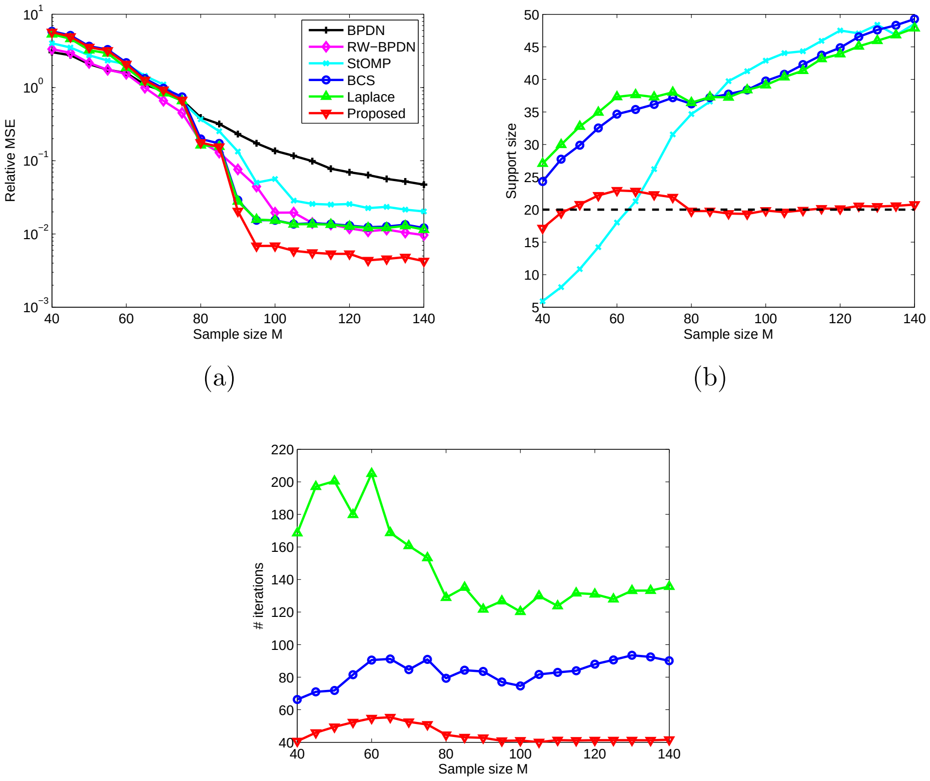
shows an exception at $\epsilon = 0$ as illustrated in Remark 6.1. In this case, the update of η takes most of the computational time in our simulation. Since it is shown in Figs. 6.3(a) – 6.3(c) that the performance at $\epsilon = 0$ and 0.01 is hardly distinguishable, we use $\epsilon = 0.01$ in the rest simulations.

6.4.1.3 Comparison with existing methods

We consider two simulation setups. In the first case we repeat the simulations above (i.e., we fix the SNR $= 25\,\text{dB}$ and vary M). In each trial, all methods share the same data. We adopt the FAR thresholding strategy in StOMP. Our simulation results are presented in Fig. 6.4. It is shown in Fig. 6.4(a) that the reconstruction errors of the three SBL methods (BCS, Laplace and our proposed method) are very close to each other and larger than those of BPDN, RW-BPDN and StOMP if the sample size is small. With a reasonable sample size it can be seen that our proposed method has the smallest error. Fig. 6.4(b) shows the average support size of the recovered signal. The results of BPDN and RW-BPDN are omitted since they are global optimization methods and their numerical solutions have no exact zero entries. In general, the estimated support sizes of StOMP, BCS and Laplace increase with the sample size. As expected the proposed method produces sparser solutions than BCS and Laplace. It is shown that the proposed method can accurately estimate the support of the sparse signal in most cases and has the best performance. Fig. 6.4(c) plots the number of iterations of the three SBL methods, where it is shown that the proposed one uses the least number of iterations and thus is the fastest one in computational speed. On average, StOMP uses the least computation time (about 0.01s), followed by the proposed method, BCS and Laplace (from 0.06 to 0.1s), and then BPDN (about 1s) and RW-BPDN (about 2s). Note also that when implemented with SPGL1, the speed of BPDN is comparable with the proposed method.

In the next simulation we set the sample size $M = 120$ and vary the SNR from 0 to 50dB with step size of 5dB. The simulation results are presented in Fig. 6.5. Fig. 6.5(a) shows that the proposed method has consistently the smallest signal recovery error. Fig. 6.5(b) shows that the proposed method produces the sparsest solution and the most accurate support size estimation. Fig. 6.5(c) shows that among the three SBL methods the proposed one uses the least number of iterations at all SNR levels. In the low SNR regime, it can be 6 and 3 times less in comparison with Laplace and BCS respectively, leading to that the proposed method is much faster than Laplace and BCS. As in the last simulation, in general SPGL1 and the proposed method is comparable in speed. The difference is that the computational time of SPGL1 increases with respect to the SNR while it decreases for the proposed method.

In summary, the proposed method has improved performance for sparse signals in comparison with existing ones. It outperforms its SBL peers in both signal recovery accuracy and computational speed.

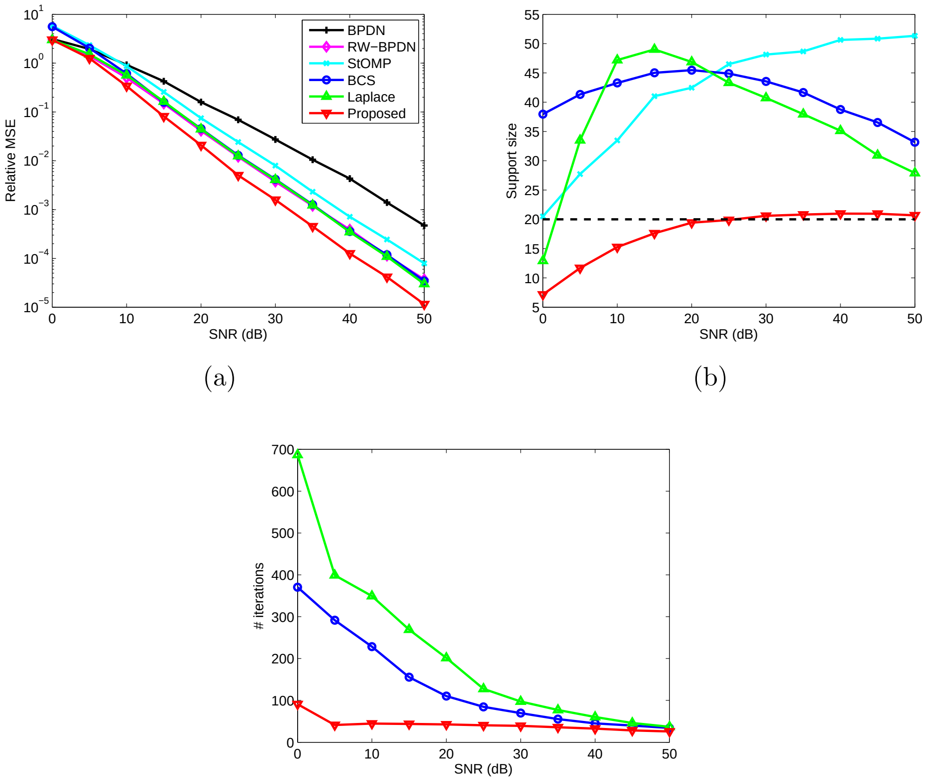


(a)

(b)

(c)

Figure 6.4: Performance comparison of the proposed algorithm with existing ones with $(N, K, \mathrm{SNR}) = (512, 20, 25\mathrm{dB})$.



(a)

(b)

(c)

Figure 6.5: Performance comparison of the proposed algorithm with existing ones with $(N, M, K) = (512, 120, 20)$.

6.4.2 Images

In this section, we revisit the widely used multiscale CS reconstruction [131] of the 512×512 Mondrian image in SparseLab. We use the same simulation setup, i.e., we choose the "symmlet8" wavelet as the sparsifying basis with a coarsest scale $j_0 = 4$ and a finest scale $j_1 = 6$. The number of wavelet samples is $N = 4096$ and the sample size of CS methods is $M = 2713$. The parameters of BP and StOMP with the FDR and FAR thresholding strategies (denoted by FDR and FAR respectively) are set as in SparseLab. Since the wavelet expansion of the Mondrian image is compressible but not exactly sparse, we set $\sigma^2 = 0.01 Var\left(\boldsymbol{y}\right)$ in Laplace and our proposed method as in BCS, where $Var\left(\boldsymbol{y}\right)$ denotes the variance of the entries of $\boldsymbol{y}$.

Table 6.1 presents the experimental results over 100 trials. Linear reconstruction from 4096 wavelet samples has a reconstruction error of 0.1333 that represents a lower bound of the error of the considered CS methods. The global optimization method BP has the smallest error among the CS methods, followed by BCS, Laplace, the proposed method and StOMP. The presented results verify again that the proposed method produces a sparser solution than BCS and Laplace. In fact, it produces the sparsest solution among all the methods. So it is reasonable that the proposed method has a slightly worse reconstruction error in comparison with BCS and Laplace since the original signal is not exactly sparse. We note that the proposed method is faster than BCS and Laplace. FDR uses the least time but has the worst accuracy. In comparison with FAR, the proposed method is slightly slower but more accurate. Finally, it can be observed that the proposed method has the most stable performance among the CS methods except BP by comparing the standard deviation of the metrics. When implemented with SPGL1, BP takes about four times computational time of the proposed method, which is consistent with an observation in the last simulation that SPGL1 takes longer time as the SNR increases. Fig. 6.6 shows a typical example of reconstructed images where faithful reconstructions of the Mondrian image can be observed.

In summary, the proposed method has satisfactory performance for compressible

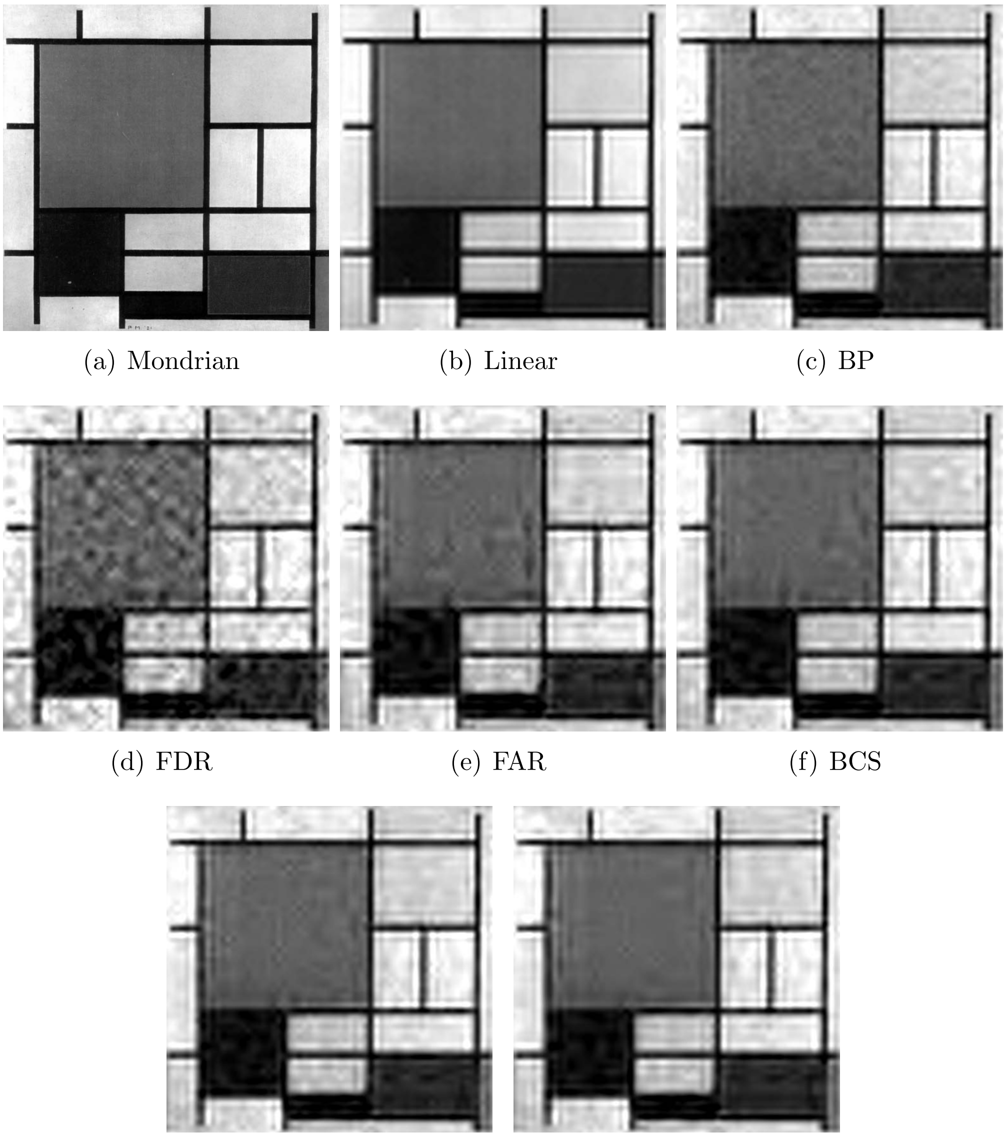
Table 6.1: Averaged Relative MSEs, CPU Times and Number of Nonzero Entries (mean $\pm$ standard deviation) for Multiscale CS Reconstruction of the Mondrian Image.

	RMSE	Time (s)	# Nonzeros
Linear	0.1333	—	4096
BP	0.1393 ± 0.0008	42.2 ± 4.03	4096 ± 0
FDR	0.1999 ± 0.0487	8.84 ± 2.12	2155 ± 122
FAR	0.1499 ± 0.0033	17.0 ± 4.35	1142 ± 41
BCS	0.1423 ± 0.0023	27.2 ± 5.92	1305 ± 67
Laplace	0.1429 ± 0.0013	25.7 ± 6.11	1218 ± 65
Proposed	0.1448 ± 0.0011	21.3 ± 4.09	1049 ± 21

signals and there exists a tradeoff between its recovery accuracy and estimated signal sparsity. In particular, a smaller ϵ leads to a sparser signal estimate which may be less accurate, while a larger ϵ (as for Laplace where $\epsilon = 1$) leads to a less sparse but more accurate signal estimate since more basis vectors contribute to the estimate.

6.5 Conclusion

The sparse signal recovery problem in CS was studied in this chapter. Within the framework of Bayesian CS, a new hierarchical sparsity-inducing prior was introduced and efficient signal recovery algorithms were developed. Similar theoretical results on the global and local optimizers of the proposed method were proven as that for the basic SBL. The main algorithm was shown to produce sparser solutions than its existing SBL peers. Numerical simulations were carried out to demonstrate the performance of the proposed sparsity-inducing prior and solution. The proposed G-STG prior preserves the general structure of existing hierarchical sparsity-inducing priors and can be implemented in other SBL-based methods with ease.



(a) Mondrian (b) Linear (c) BP

(d) FDR (e) FAR (f) BCS

(g) Laplace (h) Proposed

Figure 6.6: The 512 × 512 Mondrian image (a) and its reconstructions using (b) linear reconstruction (RMSE = 0.1333) from $N = 4096$ wavelet samples and a multiscale CS scheme from $M = 2713$ linear measurements by (c) BP (RMSE = 0.1391, time = 44.7s and # nonzeros = 4096), (d) StOMP with FDR thresholding (RMSE = 0.1751, time = 10.3s and # nonzeros = 2014), (e) StOMP with FAR thresholding (RMSE = 0.1529, time = 15.7s and # nonzeros = 1088), (f) BCS (RMSE = 0.1448, time = 24.1s and # nonzeros = 1293), (g) Laplace (RMSE = 0.1427, time = 25.5s and # nonzeros = 1229) and (h) our proposed method (RMSE = 0.1440, time = 19.8s and # nonzeros = 1033).

Chapter 7

Variational Bayesian Algorithm for Quantized Compressed Sensing

As Chapter 6, this chapter is built upon SBL. However, in this chapter we are concerned about the CS problem with quantized measurements, known as the quantized CS problem. Notice that quantization is inevitable in practical considerations though most studies of CS are focused on the standard formulation where the compressive measurements have infinite bit precision. Two scenarios are mainly studied in the literature which are categorized according to the number of bits per measurement used and include multi-bit CS (≥ 2 bits) [1, 58–61, 132] and 1-bit CS (a single bit) [62–68]. Since only the sign information of the measurements is retained in the 1-bit case and signal scaling is lost, it is quite different from the multi-bit CS and has been separately studied. In this chapter, we introduce a Bayesian framework for quantized CS that unifies the multi- and 1-bit cases. The new framework deals with quantization errors and measurement noises separately, allows data saturation in multi-bit CS, and does not need the signal sparsity information. Based on the new problem formulation, we propose an algorithm within the Bayesian CS framework where the quantization errors are modeled as random variables and jointly estimated with the signal of interest. The performance of the proposed algorithm is studied by extensive numerical simulations in various scenarios. This chapter is mainly based on [9].

7.1 A New Framework for Quantized CS

Multi- and 1-bit CS problems are typically studied separately in the literature due to their big difference. In this section, we propose a Bayesian framework that unifies both the cases. The new framework is applicable to various scenarios including noiseless/noisy environment and unsaturated/saturated quantizer. Its relations with existing methods are studied through an MAP interpretation.

7.1.1 A Unified Observation Model

In quantized CS, the observed samples are noisy linear measurements of the original signal after quantization:

$$\boldsymbol{z} = \mathcal{Q}\left(\boldsymbol{y}\right), \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}, \tag{7.1}$$

where $\boldsymbol{x}$ is the signal of interest, $\boldsymbol{A}$ is the sensing matrix, $\boldsymbol{n}$ is the measurement noise vector, $\boldsymbol{y}$ is the pre-quantized noisy measurement vector, $\mathcal{Q}$ denotes a quantizer and $\boldsymbol{z}$ is the observation vector. Notice that the noise is not denoted by $\boldsymbol{e}$ as in the previous chapters. Later we will see that $\boldsymbol{e}$ is reserved for the quantization error. A quantizer $\mathcal{Q}\left(v\right)$ for a scalar $v \in \mathbb{R}$ is defined as

$$\mathcal{Q}\left(v\right) = \begin{cases} v_0, & \text{if } v \in \left(u_0, u_1\right), \\ v_1, & \text{if } v \in \left[u_1, u_2\right), \\ \cdots, & \cdots, \\ v_{L-1}, & \text{if } v \in \left[u_{L-1}, u_L\right), \end{cases} \tag{7.2}$$

where L denotes the number of the quantization levels and typically satisfies $L = 2^B$ with B denoting the bit depth (bits per quantized measurement), $u_0 < u_1 < \cdots < u_L$, and $v_i \in \left[u_i, u_{i+1}\right)$ for $i = 0, \cdots, L-1$. The quantizer $\mathcal{Q}\left(v\right)$ is called unsaturated if $\left(u_0, u_L\right)$ is a finite interval, or saturated otherwise. For a vector $\boldsymbol{v}$, $\mathcal{Q}\left(\boldsymbol{v}\right)$ operates elementwise. Multi-bit CS refers to the case $B \geq 2$ while 1-bit CS corresponds to $B = 1$.

7.1.1.1 Multi-bit CS

We consider first a multi-bit quantizer where $B \geq 2$. Denote $\mathcal{D}_y$ the domain of $\boldsymbol{y}$. Then we have

$$\mathcal{D}_y = \mathcal{Q}^{-1}\left(\boldsymbol{z}\right) := \left\{\boldsymbol{y} \in \mathbb{R}^M | \mathcal{Q}\left(\boldsymbol{y}\right) = \boldsymbol{z}\right\}. \tag{7.3}$$

We introduce an auxiliary variable $\boldsymbol{e} = \boldsymbol{z} - \boldsymbol{y}$ denoting the quantization error with its domain

$$\mathcal{D}_e = \boldsymbol{z} - \mathcal{D}_y := \left\{\boldsymbol{z} - \boldsymbol{y} | \boldsymbol{y} \in \mathcal{D}_y\right\}. \tag{7.4}$$

Note that $\mathcal{D}_e$ is unbounded when data saturation occurs.

7.1.1.2 1-bit CS

In the case of 1-bit quantizer we set $u_0 = -\infty$, $u_1 = 0$ and $u_2 = +\infty$. The sign information of $\boldsymbol{y}$ is preserved in the quantized measurement $\boldsymbol{z}$. But the scaling information of $\boldsymbol{y}$ and $\boldsymbol{x}$ is lost. Without loss of generality, we let the 1-bit quantizer for a scalar $v \in \mathbb{R}$ be

$$\mathcal{Q}\left(v\right) = \varsigma \operatorname{sgn}\left(v\right),$$

where $\varsigma \to 0_+$ (ς is an arbitrarily small positive number) and $\operatorname{sgn}\left(\cdot\right)$ is the sign function. For convenience, we set $\operatorname{sgn}\left(0\right) = 1$ (the choice is arbitrary and can be replaced by $\operatorname{sgn}\left(0\right) = -1$). Then we have $\boldsymbol{z} \to \boldsymbol{0}$. To solve the signal scaling problem we impose a constraint that $\boldsymbol{y}$ has fixed unit norm, i.e.,

$$\left\|\boldsymbol{y}\right\|_s = 1 \tag{7.5}$$

with $s \geq 1$. Different from the multi-bit quantizer case we have in such a case that

$$\mathcal{D}_y = \left\{\boldsymbol{y} \in \mathbb{R}^M | \operatorname{sgn}\left(\boldsymbol{y}\right) = \operatorname{sgn}\left(\boldsymbol{z}\right), \left\|\boldsymbol{y}\right\|_s = 1\right\}, \tag{7.6}$$

$$\mathcal{D}_e = \left\{\boldsymbol{e} \in \mathbb{R}^M | \operatorname{sgn}\left(\boldsymbol{e}\right) = -\operatorname{sgn}\left(\boldsymbol{z}\right), \left\|\boldsymbol{e}\right\|_s = 1\right\}. \tag{7.7}$$

As a result, an observation model that unifies the multi- and 1-bit CS problems can be written into

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} + \boldsymbol{n}, \quad \boldsymbol{e} \in \mathcal{D}_e, \tag{7.8}$$

which is the observation model to be used in this chapter to recover $\boldsymbol{x}$.

Remark 7.1. *In practice, one is able to know the domain of* $\boldsymbol{e}$ *but it is difficult to characterize its exact relationship with* $\boldsymbol{x}$*. Thus, the dependence of* $\boldsymbol{e}$ *on* $\boldsymbol{x}$ *is dropped when we write (7.1) into (7.8) which is the observation model we use for the signal recovery, i.e., the only information we attempt to exploit during the signal recovery process is its domain. We note that the signal recovery performance may be improved if the dependence can be properly exploited which, however, is rather difficult.*

7.1.2 Bayesian Formulation of Quantized CS

In this subsection we formulate the quantized CS problem from a Bayesian perspective based on the observation model in (7.8). According to Remark 7.1 we treat $\boldsymbol{e}$ as a random variable independent of $\boldsymbol{x}$. The joint probability density function (PDF) $p\left(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{e}\right)$ is decomposed as

$$p\left(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{e}\right) = p\left(\boldsymbol{z} | \boldsymbol{x}, \boldsymbol{e}\right) p\left(\boldsymbol{x}\right) p\left(\boldsymbol{e}\right).$$

We define the three distributions on the right hand side as follows.

7.1.2.1 Noise model

Under an assumption of white Gaussian measurement noise, i.e., $\boldsymbol{n} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$ where σ^2 is the noise variance and $\boldsymbol{I}$ denotes an identity matrix of proper dimension, we have

$$p\left(\boldsymbol{z} | \boldsymbol{x}, \boldsymbol{e}; \sigma^2\right) = \mathcal{N}\left(\boldsymbol{z} | \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}, \sigma^2 \boldsymbol{I}\right). \tag{7.9}$$

7.1.2.2 Sparse signal model

A sparse prior is needed for the sparse signal $\boldsymbol{x}$ of interest. Here we do not give an explicit distribution to the sparse signal $\boldsymbol{x}$ but denote $p(\boldsymbol{x})$ its PDF. Then we let $f(\boldsymbol{x}) = -C_1 \log p(\boldsymbol{x}) + C_2$ where C_1 and C_2 are proper constants. The only thing that we assume for $p(\boldsymbol{x})$ is that it favors entries of $\boldsymbol{x}$ being zero. As an example, a commonly used sparse prior for $\boldsymbol{x}$ is a Laplace prior [133, 134]: $p(\boldsymbol{x}) = \lambda^N \exp\left\{-\lambda \|\boldsymbol{x}\|_1\right\}$ with λ being a positive constant. In such a case, we have $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ that has been extensively studied in deterministic optimization methods.

7.1.2.3 Quantization error model

We assume a uniform, noninformative prior for $\boldsymbol{e}$:

$$\boldsymbol{e} \sim U(\mathcal{D}_e) \tag{7.10}$$

since the only information of $\boldsymbol{e}$ that we use is $\boldsymbol{e} \in \mathcal{D}_e$.

Remark 7.2. *The uniform prior may not characterize well the quantization error in the case of a very small bit depth B. But it is noted that a sophisticated prior needs more information besides the domain $\mathcal{D}_e$ which is difficult to obtain.*

To obtain an MAP estimator of $\boldsymbol{x}$ requires to integrate out $\boldsymbol{e}$ from $p(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{e})$ that is computationally intractable. We propose to estimate $\boldsymbol{x}$ and $\boldsymbol{e}$ simultaneously using their joint MAP estimator:

$$\begin{aligned}
\{\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{e}}\} &= \arg\max_{\boldsymbol{x},\boldsymbol{e}} \log p(\boldsymbol{x}, \boldsymbol{e}|\boldsymbol{z}) \\
&= \arg\max_{\boldsymbol{x},\boldsymbol{e}} \log p(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{e}) \\
&= \arg\max_{\boldsymbol{x},\boldsymbol{e}} \log \left\{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{e})\, p(\boldsymbol{x})\, p(\boldsymbol{e})\right\} \\
&= \arg\min_{\boldsymbol{x},\boldsymbol{e}\in\mathcal{D}_e} \left\{f(\boldsymbol{x}) + \frac{C_1}{2\sigma^2}\|\boldsymbol{z} - \boldsymbol{e} - \boldsymbol{A}\boldsymbol{x}\|_2^2\right\}.
\end{aligned} \tag{7.11}$$

An equivalent form of the problem in (7.11) is

$$\min_{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{e}}} f\left(\tilde{\boldsymbol{x}}\right), \text{ subject to } \begin{cases} \left\|\boldsymbol{z}-\tilde{\boldsymbol{e}}-\boldsymbol{A}\tilde{\boldsymbol{x}}\right\|_2 \leq \epsilon, \\ \tilde{\boldsymbol{e}} \in \mathcal{D}_e, \end{cases} \tag{7.12}$$

where ϵ is a proper scalar that controls the noise energy. The first constraint in (7.12) is to ensure the data consistency against the measurement noise. In multi-bit CS, the second one concerns data consistency due to quantization. In 1-bit CS, an additional signal scaling constraint is included in the second constraint that prevents an optimal solution for $\boldsymbol{x}$ from $\mathbf{0}$. Before proceeding to our algorithm within the framework of Bayesian CS, we study in the next subsection relations of the proposed Bayesian framework with existing methods.

7.1.3 Relations with Existing Methods in Quantized CS

We first note that problem (7.12) is equivalent to the problem

$$\min_{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}} f\left(\tilde{\boldsymbol{x}}\right), \text{ subject to } \begin{cases} \left\|\tilde{\boldsymbol{y}}-\boldsymbol{A}\tilde{\boldsymbol{x}}\right\|_2 \leq \epsilon, \\ \tilde{\boldsymbol{y}} \in \mathcal{D}_y. \end{cases} \tag{7.13}$$

In the following we show that many existing problem formulations of quantized CS are special cases of or related to (7.13).

7.1.3.1 Multi-bit CS

We consider the case of ℓ_1 optimization where $f\left(\boldsymbol{x}\right) = \left\|\boldsymbol{x}\right\|_1$. In the noise free case where $\epsilon = 0$, the problem in (7.13) can be written into

$$\min_{\tilde{\boldsymbol{x}}} \left\|\tilde{\boldsymbol{x}}\right\|_1, \text{ subject to } \boldsymbol{A}\tilde{\boldsymbol{x}} \in \mathcal{D}_y, \tag{7.14}$$

which has been studied in [132]. Further, by assuming that $\mathcal{Q}$ is a uniform unsaturated quantizer the above problem can be written into

$$\min_{\tilde{\boldsymbol{x}}} \left\|\tilde{\boldsymbol{x}}\right\|_1, \text{ subject to } \left\|\boldsymbol{z}-\boldsymbol{A}\tilde{\boldsymbol{x}}\right\|_\infty \leq \frac{r}{2} \tag{7.15}$$

which is studied in [1,59] with r denoting the quantization bin width. While existing methods that account for measurement noise typically mix it up with the quantization error, e.g., in [60], problem (7.13) extends existing noise free formulations to the noisy case by dealing with the two uncertainties separately.

Remark 7.3. *Under the assumption that all quantization errors are independent and uniformly distributed in a common interval* $\left[-\frac{r}{2},\frac{r}{2}\right]$, *it is shown in [59] that the* ℓ_∞ *norm in problem (7.15) is not the best choice for the signal recovery. But it is unclear whether the result in [59] can be extended to the case of a general quantizer where the above assumption fails. It is noted that our problem formulation does not require this assumption and applies to an arbitrary quantizer. That is, by losing some optimality, we have obtained the universality.*

7.1.3.2 1-bit CS

In 1-bit CS (7.13) becomes

$$\min_{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}} f\left(\tilde{\boldsymbol{x}}\right), \text{ subject to } \begin{cases} \left\|\tilde{\boldsymbol{y}}-\boldsymbol{A}\tilde{\boldsymbol{x}}\right\|_2 \leq \epsilon, \\ \operatorname{sgn}\left(\tilde{\boldsymbol{y}}\right)=\operatorname{sgn}\left(\boldsymbol{z}\right), \\ \left\|\tilde{\boldsymbol{y}}\right\|_s = 1. \end{cases} \tag{7.16}$$

In the noise free case, it can be written into

$$\min_{\tilde{\boldsymbol{x}}} f\left(\tilde{\boldsymbol{x}}\right), \text{ subject to } \begin{cases} \operatorname{sgn}\left(\boldsymbol{A}\tilde{\boldsymbol{x}}\right)=\operatorname{sgn}\left(\boldsymbol{z}\right), \\ \left\|\boldsymbol{A}\tilde{\boldsymbol{x}}\right\|_s = 1. \end{cases} \tag{7.17}$$

This problem with the settings $f\left(\boldsymbol{x}\right)=\left\|\boldsymbol{x}\right\|_1$ and $s=1$ can be shown to be convex and has been studied in [67]. So by (7.16) we extend (7.17) to the noisy case while the authors of [67] state in [68] that “it was unclear how to modify the above convex program to account for possible noise.”

The third constraint in (7.16) serves to prevent the optimal solution for $\boldsymbol{x}$ from $\mathbf{0}$. If replacing it by $\|\tilde{\boldsymbol{x}}\|_2 = 1$, then problem (7.16) can be shown to be equivalent to the problem

$$\min_{\tilde{\boldsymbol{x}}} f\left(\tilde{\boldsymbol{x}}\right), \text{ subject to } \begin{cases} \left\|\left(\operatorname{sgn}\left(\boldsymbol{z}\right) \odot \boldsymbol{A}\tilde{\boldsymbol{x}}\right)_{-}\right\|_2 \leq \epsilon, \\ \|\tilde{\boldsymbol{x}}\|_2 = 1, \end{cases} \tag{7.18}$$

where $(v)_{-} = \max\{-v, 0\}$ for a scalar v and operates elementwise for a vector. In the noise free case (7.18) becomes

$$\min_{\tilde{\boldsymbol{x}}} f\left(\tilde{\boldsymbol{x}}\right), \text{ subject to } \begin{cases} \operatorname{sgn}\left(\boldsymbol{A}\tilde{\boldsymbol{x}}\right) = \operatorname{sgn}\left(\boldsymbol{z}\right), \\ \|\tilde{\boldsymbol{x}}\|_2 = 1, \end{cases} \tag{7.19}$$

which with $f\left(\boldsymbol{x}\right) = \|\boldsymbol{x}\|_1$ is the earliest formulation of the 1-bit CS problem introduced in [62] and solved using RFPI in [62] and RSS in [65]. Assume that the signal sparsity information is known instead of the noise energy, another formulation of (7.18) is to pose $\left\|\left(\operatorname{sgn}\left(\boldsymbol{z}\right) \odot \boldsymbol{A}\tilde{\boldsymbol{x}}\right)_{-}\right\|_2$ as the objective function and $f\left(\tilde{\boldsymbol{x}}\right) \leq S$ as a constraint, where the constant S refers to the sparsity information. Such kind of formulations have been studied in [62,64,66]. In addition, the convex program in [68] is related by observing that $\operatorname{sgn}^T\left(\boldsymbol{z}\right)\boldsymbol{A}\tilde{\boldsymbol{x}} = \left\|\left(\operatorname{sgn}\left(\boldsymbol{z}\right) \odot \boldsymbol{A}\tilde{\boldsymbol{x}}\right)_{+}\right\|_1 - \left\|\left(\operatorname{sgn}\left(\boldsymbol{z}\right) \odot \boldsymbol{A}\tilde{\boldsymbol{x}}\right)_{-}\right\|_1$, where $(v)_{+} = \max\{v, 0\}$.

Remark 7.4. *By (7.18) we see that the effective noise is $\left(sgn\left(\boldsymbol{z}\right) \odot \boldsymbol{A}\boldsymbol{x}\right)_{-}$ in 1-bit CS where by “effective noise” we refer to a noise that has the minimum energy and leads to the same measurement. Since its energy is much smaller than that of the true noise $\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$, from this point of view, we may say that 1-bit CS is robust to the measurement noise.*

7.2 Q-VMP: Variational Message Passing for Quantized CS

7.2.1 Model Selection

We assume that the noise variance σ^2 is known. Though it can be estimated by assuming an inverse Gamma prior for it in the case where it is unknown as in [46], its estimate is inaccurate due to an "identifiability issue" as addressed in [41]. For the sparse signal $\boldsymbol{x}$, we adopt a three-layer, Gaussian-Gamma-Gamma, hierarchical prior introduced in [54]:

$$p(\boldsymbol{x};\epsilon,c,d)=\iint p(\boldsymbol{x}|\boldsymbol{\alpha})\,p(\boldsymbol{\alpha}|\eta;\epsilon)\,p(\eta;c,d)\;d\boldsymbol{\alpha}\,d\eta$$

where

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\alpha}) &= \mathcal{N}(\boldsymbol{x}|\mathbf{0},\mathbf{\Lambda}), && (7.20)\\ p(\boldsymbol{\alpha}|\eta;\epsilon) &= \prod_{i=1}^{N}\Gamma(\alpha_i|\epsilon,\eta), && (7.21)\\ p(\eta;c,d) &= \Gamma(\eta|c,d) && (7.22)\end{aligned}$$

with $\mathbf{\Lambda}=\mathrm{diag}(\boldsymbol{\alpha})$ and constants ϵ, c, d. For a Gamma distributed variable $u\sim\Gamma(c,d)$, its PDF is $\Gamma(u|c,d)=\frac{d^c}{\Gamma(c)}u^{c-1}\exp(-du)$ with $\Gamma(c)$ being the Gamma function. By [54] the constants ϵ, c, d satisfy that $0\leq\epsilon\leq 1$, $c,d\geq 0$. In this chapter, we adopt $c=1$, $d=0$ to make the prior for η in (7.22) noninformative (flat on $\mathbb{R}_+$). Further, we choose $\epsilon=0$ since a smaller ϵ leads to a sparser prior and an estimator that approximates a hard-thresholding rule according to [54]. Readers are referred to [54] for more properties of the Gaussian-Gamma-Gamma prior and its relations with other sparse estimation techniques.

In 1-bit CS, we let $\boldsymbol{y}$ have unit ℓ_2 norm in (7.5), leading to that $\|\boldsymbol{e}\|_2=1$ in (7.7).

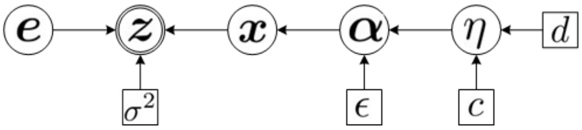


Figure 7.1: Directed graphical model that encodes the joint PDF in (7.23) of the Bayesian model. Nodes denoted with circles correspond to random variables, while nodes denoted with squares correspond to parameters of the model. Doubly circled $\boldsymbol{z}$ is the observation while single circled nodes represent hidden variables.

As a result, we have the joint PDF of the observation model (7.8):

$$p\left(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{e}, \boldsymbol{\alpha}, \eta\right) = p\left(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{e}\right) p\left(\boldsymbol{x}|\boldsymbol{\alpha}\right) p\left(\boldsymbol{\alpha}|\eta\right) p\left(\eta\right) p\left(\boldsymbol{e}\right) \tag{7.23}$$

with the distributions on the right hand side as defined respectively by (7.9), (7.20), (7.21), (7.22) and (7.10). A directed graphical model that encodes the factorization of the joint PDF in (7.23) is shown in Fig. 7.1.

7.2.2 Q-VMP Algorithm

It is well known that Bayesian inference is based on the posterior distribution $p\left(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{\alpha}, \eta|\boldsymbol{z}\right) = p\left(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{e}, \boldsymbol{\alpha}, \eta\right)/p\left(\boldsymbol{z}\right)$. However, such an exact posterior distribution is intractable since $p\left(\boldsymbol{z}\right) = \int \cdots \int p\left(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{e}, \boldsymbol{\alpha}, \eta\right) \, d\boldsymbol{x} \, d\boldsymbol{e} \, d\boldsymbol{\alpha} \, d\eta$ cannot be expressed explicitly.

A variational inference approach [135, 136] is adopted in this chapter. Denote $\boldsymbol{V} = \{\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{\alpha}, \eta\}$ the set of all unknown variables to be estimated. The goal in variational inference is to find a tractable distribution $q\left(\boldsymbol{V}\right)$ that closely approximates the true posterior distribution $p\left(\boldsymbol{V}|\boldsymbol{z}\right)$. To do this, some family of distributions that has enough flexibility is firstly chosen to represent $q\left(\boldsymbol{V}\right)$. Then the task is to find a member of the family that minimizes the Kullback-Leibler (KL) divergence between the true posterior $p\left(\boldsymbol{V}|\boldsymbol{z}\right)$ and the variational approximation $q\left(\boldsymbol{V}\right)$. A commonly used variational distribution $q\left(\boldsymbol{V}\right)$ is such that disjoint groups of variables are independent, i.e., $q\left(\boldsymbol{V}\right)$ has a factorized form $q\left(\boldsymbol{V}\right) = q\left(\boldsymbol{x}\right) q\left(\boldsymbol{e}\right) q\left(\boldsymbol{\alpha}\right) q\left(\eta\right)$. Variational message passing (VMP) is proposed in [137] for the variational inference using a message passing procedure on a graphical model. In VMP, the variational distributions $q(\boldsymbol{x})$, $q(\boldsymbol{e})$, $q(\boldsymbol{\alpha})$, $q(\eta)$ are iteratively updated to monotonically decrease the KL divergence and thus has guaranteed convergence. Readers are referred to [137] for more details of VMP. The updates of $q(\boldsymbol{x})$, $q(\boldsymbol{\alpha})$, $q(\eta)$ are similar to those in [54] because of the similarity between quantized and conventional CS. $q(\boldsymbol{e})$ is given complete flexibility in multi-bit CS as $q(\boldsymbol{x})$, $q(\boldsymbol{\alpha})$ and $q(\eta)$. We constrain $q(\boldsymbol{e})$ in 1-bit CS such that

$$q(\boldsymbol{e}) = \delta\left(\boldsymbol{e} - \boldsymbol{e}^0\right) \tag{7.24}$$

due to a computational issue to be discussed in Remark 7.6, where $\delta(\cdot)$ is the delta function and $\boldsymbol{e}^0 \in \mathbb{R}^M$ is to be estimated. Note that (7.24) is equivalent to the complete flexibility in the noise free case to be illustrated in Subsection 7.2.3.

Remark 7.5. *In the 1-bit case the convergence of the resulting algorithm is not a direct result of [137] due to the adoption of a degenerate distribution for $\mathbf{e}$. Instead, we may consider $\mathbf{e}$ as an unknown deterministic parameter in such a case. Then the resulting algorithm can be interpreted as a variational EM algorithm [136] and during the iterations the (marginal) likelihood $p(\boldsymbol{z}; \boldsymbol{e})$ is guaranteed to monotonically increase and thus convergence is guaranteed. Readers are referred to [136] for the details.*

7.2.2.1 Updates of $q(\boldsymbol{x})$, $q(\boldsymbol{\alpha})$ and $q(\eta)$

According to [137] we have that

$$\begin{aligned} q(\boldsymbol{x}) &\propto \exp\left\{\langle \ln p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{e})\rangle_{q(e)} \langle \ln p(\boldsymbol{x}|\boldsymbol{\alpha})\rangle_{q(\alpha)}\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}, \end{aligned}$$

and thus $q(\boldsymbol{x})$ is a Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \langle \boldsymbol{x}\rangle_{q(\boldsymbol{x})} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{A}^T\left(\boldsymbol{z} - \langle \boldsymbol{e}\rangle_{q(e)}\right), \tag{7.25}$$

$$\boldsymbol{\Sigma} = \left(\sigma^{-2}\boldsymbol{A}^T\boldsymbol{A} + \left\langle \boldsymbol{\Lambda}^{-1} \right\rangle_{q(\boldsymbol{\alpha})}\right)^{-1}. \tag{7.26}$$

For $\boldsymbol{\alpha}$ we have

$$\begin{aligned} q(\boldsymbol{\alpha}) &\propto \exp\left\{ \langle \ln p(\boldsymbol{x}|\boldsymbol{\alpha}) \rangle_{q(\boldsymbol{x})} \langle \ln p(\boldsymbol{\alpha}|\boldsymbol{\eta}) \rangle_{q(\eta)} \right\} \\ &\propto \prod_{n=1}^{N} \alpha_n^{\epsilon-\frac{3}{2}} \exp\left\{ -\frac{1}{2}\alpha_n^{-1} \left\langle x_n^2 \right\rangle_{q(\boldsymbol{x})} - \alpha_n \langle \eta \rangle_{q(\eta)} \right\}, \end{aligned}$$

where $\langle x_n^2 \rangle_{q(\boldsymbol{x})} = \mu_n^2 + \Sigma_{nn}$. The expression on the right hand side is the product of generalized inverse Gaussian (GIG) PDFs and thus we have for any $i \in \mathbb{R}$ [138]:

$$\left\langle \alpha_n^i \right\rangle_{q(\boldsymbol{\alpha})} = \left(\frac{\left\langle x_n^2 \right\rangle_{q(\boldsymbol{x})}}{2 \langle \eta \rangle_{q(\eta)}} \right)^{\frac{i}{2}} \frac{\mathcal{K}_{\epsilon+i-\frac{1}{2}}\left(\sqrt{2 \langle \eta \rangle_{q(\eta)} \left\langle x_n^2 \right\rangle_{q(\boldsymbol{x})}} \right)}{\mathcal{K}_{\epsilon-\frac{1}{2}}\left(\sqrt{2 \langle \eta \rangle_{q(\eta)} \left\langle x_n^2 \right\rangle_{q(\boldsymbol{x})}} \right)}, \tag{7.27}$$

where $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind and order $\nu \in \mathbb{R}$. The case of $i = -1$ in (7.27) gives the evaluation of $\langle \boldsymbol{\Lambda}^{-1} \rangle_{q(\boldsymbol{\alpha})}$ used in (7.26), and the case of $i = 1$ gives the calculation of $\langle \alpha_n \rangle_{q(\boldsymbol{\alpha})}$ used in a later expression in (7.28). The update of $q(\eta)$ can be shown to be $q(\eta) = \Gamma\left(\eta | N\epsilon + c, \sum_{n=1}^{N} \langle \alpha_n \rangle_{q(\boldsymbol{\alpha})} + d\right)$. The first moment of η used in (7.27) is given as

$$\langle \eta \rangle_{q(\eta)} = \frac{N\epsilon + c}{\sum_{n=1}^{N} \langle \alpha_n \rangle_{q(\boldsymbol{\alpha})} + d}. \tag{7.28}$$

7.2.2.2 Update of $q(\boldsymbol{e})$ in multi-bit CS

In multi-bit CS we have

$$\begin{aligned} q(\boldsymbol{e}) &\propto \exp\left\{ \langle \ln p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{e}) \rangle_{q(\boldsymbol{x})} \right\} p(\boldsymbol{e}) \\ &\propto \exp\left\{ -\frac{1}{2}\sigma^{-2} \left\langle \|\boldsymbol{z} - \boldsymbol{e} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \right\rangle_{q(\boldsymbol{x})} \right\} I_{\boldsymbol{e}}(\mathcal{D}_e) \\ &\propto \exp\left\{ -\frac{1}{2}\sigma^{-2} \|\boldsymbol{e} - (\boldsymbol{z} - \boldsymbol{A}\boldsymbol{\mu})\|_2^2 \right\} I_{\boldsymbol{e}}(\mathcal{D}_e), \end{aligned} \tag{7.29}$$

where $I_{\boldsymbol{e}}(\mathcal{D}_e)$ is an indicator function that equals to 1 if $\boldsymbol{e} \in \mathcal{D}_e$ or 0 otherwise. Hence, $q(\boldsymbol{e})$ is the product of PDFs of truncated Gaussian distributions, i.e., for each $m = 1, \cdots, M$, $q(e_m)$ is the PDF of a truncated Gaussian distribution. As a result, the first moment of e_m, $m = 1, \cdots, M$, used in (7.25) can be given in closed form after some derivations using the PDF $\phi(\cdot)$ and cumulative distribution function (CDF) $\Phi(\cdot)$ of a standard Gaussian distribution:

$$\langle e_m \rangle_{q(\boldsymbol{e})} = \sigma \frac{\phi(l_{e_m}) - \phi(u_{e_m})}{\Phi(u_{e_m}) - \Phi(l_{e_m})} + \mu_{e_m}, \tag{7.30}$$

where $\mu_{e_m} = (\boldsymbol{z} - \boldsymbol{A\mu})_m$, l_{e_m} and u_{e_m} satisfy that $\mathcal{D}_{e_m} = [\sigma l_{e_m} + \mu_{e_m}, \sigma u_{e_m} + \mu_{e_m}]$ with $\mathcal{D}_{e_m}$ denoting the domain of e_m, $\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\}$ and $\Phi(u) = \int_{-\infty}^{u} \phi(t)\, dt$ for $u \in \mathbb{R}$.

Remark 7.6. *Consider the case where $q(\boldsymbol{e})$ is given the complete flexibility in 1-bit CS. Note that entries of a point in $\mathcal{D}_e$ are no longer independent of each other in such a case, leading to that $q(\boldsymbol{e})$ is the PDF of a truncated multi-variable Gaussian distribution with $\boldsymbol{e}$ constrained in a nonconvex set $\mathcal{D}_e$ defined in (7.7). As a result, the calculation of $\langle \boldsymbol{e} \rangle_{q(\boldsymbol{e})}$ is in general computationally intractable in our considered CS problems where the dimension of $\boldsymbol{e}$ is large.*

7.2.2.3 Update of $q(\boldsymbol{e})$ in 1-bit CS

According to [137], this is equivalent to finding an MAP estimator of $\boldsymbol{e}$ with its posterior distribution defined in (7.29). So we have

$$\langle \boldsymbol{e} \rangle_{q(\boldsymbol{e})} = \boldsymbol{e}^0 \tag{7.31}$$

with

$$\begin{aligned}
\boldsymbol{e}^0 &= \arg\max_{\boldsymbol{e} \in \mathbb{R}^M} \left\{ \exp\left\{ -\frac{1}{2}\sigma^{-2} \left\| \boldsymbol{e} - (\boldsymbol{z} - \boldsymbol{A\mu}) \right\|_2^2 \right\} I_{\boldsymbol{e}}(\mathcal{D}_e) \right\} \\
&= \arg\min_{\boldsymbol{e} \in \mathcal{D}_e} \left\| \boldsymbol{e} - (\boldsymbol{z} - \boldsymbol{A\mu}) \right\|_2 \\
&= \mathcal{P}_{\mathcal{D}_e}(\boldsymbol{z} - \boldsymbol{A\mu}),
\end{aligned}$$

where $\mathcal{D}_e$ is defined in (7.7) and $\mathcal{P}_{\mathcal{D}}(\boldsymbol{v})$ denotes a projection of a point $\boldsymbol{v}$ onto a set $\mathcal{D}$. The calculation of $\mathcal{P}_{\mathcal{D}_e}(\cdot)$ with the nonconvex set $\mathcal{D}_e$ is provided in the following lemma.

Lemma 7.1. *For a vector $\boldsymbol{v} \in \mathbb{R}^M$, let $\overline{\boldsymbol{v}} = -sgn(\boldsymbol{z}) \odot \boldsymbol{v}$. Denote $\mathcal{I}$ the index set of all positive entries of $\overline{\boldsymbol{v}}$. Let $\mathcal{I}^c$ be its complementary set. If $\mathcal{I}$ is nonempty, then let $\boldsymbol{e}^* \in \mathbb{R}^M$ with $\boldsymbol{e}^*_{\mathcal{I}} = \frac{\boldsymbol{v}_{\mathcal{I}}}{\|\boldsymbol{v}_{\mathcal{I}}\|_2}$ and $\boldsymbol{e}^*_{\mathcal{I}^c} = \mathbf{0}$. Otherwise, let $i_0 = \arg\max_i(\overline{v}_i)$ and $\boldsymbol{e}^*$ such that $e^*_{i_0} = -sgn(z_{i_0})$ and $e^*_i = 0$ whenever $i \neq i_0$. Then $\boldsymbol{e}^* = \mathcal{P}_{\mathcal{D}_e}(\boldsymbol{v})$ with $\mathcal{D}_e$ as defined in (7.7).*

Proof. It is easy to show the following equivalences:

$$\begin{aligned} & \boldsymbol{e}^* = \mathcal{P}_{\mathcal{D}_e}(\boldsymbol{v}) \\ & \Leftrightarrow \boldsymbol{e}^* = \arg\min_{\boldsymbol{e}\in\mathcal{D}_e} \|\boldsymbol{e} - \boldsymbol{v}\|_2 \\ & \Leftrightarrow \boldsymbol{e}^* = \arg\max_{\boldsymbol{e}\in\mathcal{D}_e} \boldsymbol{v}^T\boldsymbol{e} \\ & \Leftrightarrow -\mathrm{sgn}(\boldsymbol{z}) \odot \boldsymbol{e}^* = \arg\max_{\boldsymbol{w}} f(\boldsymbol{w}) = \overline{\boldsymbol{v}}^T\boldsymbol{w}, \\ & \qquad \text{subject to } \|\boldsymbol{w}\|_2 = 1 \text{ and } \boldsymbol{w} \succeq \mathbf{0}. \end{aligned} \tag{7.32}$$

1) $\mathcal{I}$ is nonempty. Note that $\|\boldsymbol{w}_{\mathcal{I}}\|_2 \leq 1$ and $\overline{\boldsymbol{v}}_{\mathcal{I}^c} \preceq \mathbf{0}$. By the Cauchy inequality,

$$\begin{aligned} f(\boldsymbol{w}) &= \overline{\boldsymbol{v}}_{\mathcal{I}}^T\boldsymbol{w}_{\mathcal{I}} + \overline{\boldsymbol{v}}_{\mathcal{I}^c}^T\boldsymbol{w}_{\mathcal{I}^c} \\ &\leq \|\boldsymbol{w}_{\mathcal{I}}\|_2 \|\overline{\boldsymbol{v}}_{\mathcal{I}}\|_2 + \overline{\boldsymbol{v}}_{\mathcal{I}^c}^T\boldsymbol{w}_{\mathcal{I}^c} \\ &\leq \|\overline{\boldsymbol{v}}_{\mathcal{I}}\|_2 . \end{aligned} \tag{7.33}$$

It is readily verified that the equality holds if $\boldsymbol{w}$ is in the form of $-\mathrm{sgn}(\boldsymbol{z}) \odot \boldsymbol{e}^*$.

2) $\mathcal{I}$ is empty, i.e., $\overline{\boldsymbol{v}} \preceq \mathbf{0}$. We prove the following result: $f(\boldsymbol{w}) \leq \|\boldsymbol{w}\|_2 \max(\overline{\boldsymbol{v}}) = \max(\overline{\boldsymbol{v}})$. It is obvious that the equality holds if $\boldsymbol{w}$ is in the form of $-\mathrm{sgn}(\boldsymbol{z}) \odot \boldsymbol{e}^*$.

The case of $M = 1$ is trivial. We next prove the case of $M = 2$ and then use induction to complete the proof. When $M = 2$, substitute $w_1 = \sqrt{1 - w_2^2}$ into

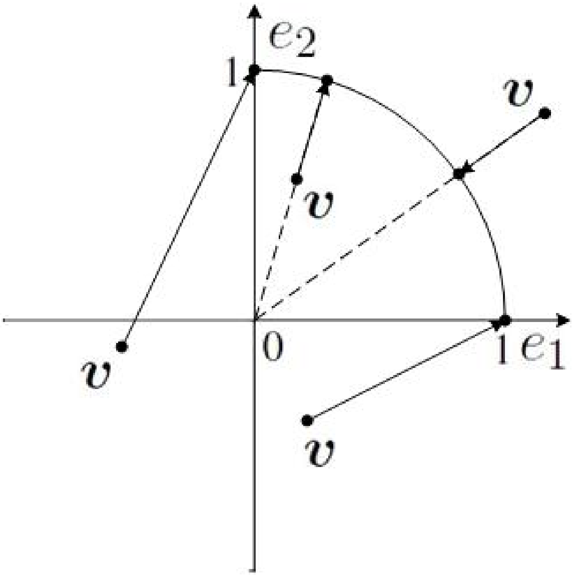


Figure 7.2: An illustration of Lemma 7.1 with nonnegative entries of $\boldsymbol{e}$. The unit circle in the first quadrant composes of $\mathcal{D}_e$. Projections of four possible $\boldsymbol{v}$'s are shown.

$f(\boldsymbol{w})$ and then

$$g(w_2) := f\left(\sqrt{1-w_2^2}, w_2\right) = \sqrt{1-w_2^2}\overline{v}_1 + w_2\overline{v}_2. \tag{7.34}$$

It is easy to show that $g'(w_2) \leq 0$ if $0 \leq w_2 \leq \frac{|\overline{v}_2|}{\|\overline{\boldsymbol{v}}\|_2}$, and $g'(w_2) \geq 0$ if $\frac{|\overline{v}_2|}{\|\overline{\boldsymbol{v}}\|_2} \leq w_2 < 1$. So the maximum of $g(w_2)$ can only be obtained at the boundary of the interval $[0,1]$, i.e., $f(\boldsymbol{w}) \leq \max(g(0), g(1)) = \max(\overline{\boldsymbol{v}})$.

Suppose the lemma holds when $M = n-1$ with $n > 3$. We next show that it holds when $M = n$. Denote $\boldsymbol{w}_{-1} = [w_2, \cdots, w_n]^T$, $\overline{\boldsymbol{v}}_{-1} = [\overline{v}_2, \cdots, \overline{v}_n]^T$. By $w_1^2 + \|\boldsymbol{w}_{-1}\|_2^2 = \|\boldsymbol{w}\|_2^2$ and applying the results when $M = n-1$ and $M = 2$ consecutively,

$$\begin{aligned} f(\boldsymbol{w}) &= \overline{v}_1 w_1 + \overline{\boldsymbol{v}}_{-1}^T \boldsymbol{w}_{-1} \\ &\leq \overline{v}_1 w_1 + \max(\overline{\boldsymbol{v}}_{-1}) \|\boldsymbol{w}_{-1}\|_2 \\ &\leq \|\boldsymbol{w}\|_2 \max(\overline{\boldsymbol{v}}) = \max(\overline{\boldsymbol{v}}). \end{aligned} \tag{7.35}$$

■

Lemma 7.1 tells how to calculate the projection onto the nonconvex set $\mathcal{D}_e$ defined in (7.7). An illustration of Lemma 7.1 is presented in Fig. 7.2, where we consider the two dimensional case with both entries of $\boldsymbol{e}$ nonnegative. The unit circle in the first quadrant composes of $\mathcal{D}_e$. Projections of four possible $\boldsymbol{v}$'s are shown. The resulting algorithm is summarized in Algorithm 7.1, named as variational message passing with quantization (Q-VMP).

```
Algorithm 7.1: Q-VMP
Input: sensing matrix $\boldsymbol{A}$, quantized measurement $\boldsymbol{z}$, domain of quantization error $\mathcal{D}_e$, and noise variance $\sigma^2$.
1. initialize $\langle\alpha_n^{-1}\rangle_{q(\boldsymbol{\alpha})}$, $n = 1, \cdots, N$, $\langle\eta\rangle_{q(\eta)}$ and $\langle\boldsymbol{e}\rangle_{q(\boldsymbol{e})}$;
2. while not converged do
3.     update $\boldsymbol{\Sigma}$ according to (7.26);
4.     update $\boldsymbol{\mu}$ according to (7.25);
5.     update $\langle\alpha_n^{-1}\rangle_{q(\boldsymbol{\alpha})}$ and $\langle\alpha_n\rangle_{q(\boldsymbol{\alpha})}$, $n = 1, \cdots, N$, according to (7.27);
6.     update $\langle\eta\rangle_{q(\eta)}$ according to (7.28);
7.     update $\langle\boldsymbol{e}\rangle_{q(\boldsymbol{e})}$ according to (7.30) in multi-bit CS and (7.31) in 1-bit CS, respectively;
8. end while
Output: recovered signal $\widehat{\boldsymbol{x}} = \boldsymbol{\mu}$.
```

7.2.3 The Noise Free Case

In this subsection we consider Q-VMP in the noise free case. We first consider the data consistency. A consistent recovery means that the observation can be reproduced from the recovered signal. Empirical results suggest that a consistent recovery results in less errors [59, 65]. A theoretical proof is provided in [66] on the 1-bit case. The following analysis can be applied to both multi- and 1-bit CS. Taking $\sigma^2 \to 0$ at both sides of (7.25) and (7.26) gives

$$\begin{aligned}\boldsymbol{\mu} &\to \left\langle\boldsymbol{\Lambda}^{-1}\right\rangle_{q(\boldsymbol{\alpha})}^{-\frac{1}{2}} \left(\boldsymbol{A}\left\langle\boldsymbol{\Lambda}^{-1}\right\rangle_{q(\boldsymbol{\alpha})}^{-\frac{1}{2}}\right)^{\dagger} (\boldsymbol{z}-\boldsymbol{e}),\\ \boldsymbol{\Sigma} &\to \left\langle\boldsymbol{\Lambda}^{-1}\right\rangle_{q(\boldsymbol{\alpha})}^{-1} - \left\langle\boldsymbol{\Lambda}^{-1}\right\rangle_{q(\boldsymbol{\alpha})}^{-\frac{1}{2}} \left(\boldsymbol{A}\left\langle\boldsymbol{\Lambda}^{-1}\right\rangle_{q(\boldsymbol{\alpha})}^{-\frac{1}{2}}\right)^{\dagger} \boldsymbol{A}\left\langle\boldsymbol{\Lambda}^{-1}\right\rangle_{q(\boldsymbol{\alpha})}^{-1}.\end{aligned}$$

Thus we have

$$\boldsymbol{A}\boldsymbol{\mu} \to \boldsymbol{z}-\boldsymbol{e} \in \mathcal{D}_y,$$

i.e., $\mathcal{Q}(\boldsymbol{A}\boldsymbol{\mu}) \to \boldsymbol{z}$, which indicates that the recovered signal reproduces the observation at each iteration.

We next consider the update of $q(\boldsymbol{e})$ in such a case. As $\sigma^2 \to 0$ we see that $q(\boldsymbol{e})$ degenerates into a single-point distribution by (7.29), that coincides with the stricter assumption in (7.24) in 1-bit CS.

7.2.4 Pruning a Basis Function

The most difficult computation of Q-VMP is the calculation of $\boldsymbol{\Sigma}$ that is the inverse of an $N \times N$ matrix. Using the Woodbury matrix identity, we have

$$\boldsymbol{\Sigma} = \left\langle \boldsymbol{\Lambda}^{-1} \right\rangle_{q(\boldsymbol{\alpha})}^{-1} - \left\langle \boldsymbol{\Lambda}^{-1} \right\rangle_{q(\boldsymbol{\alpha})}^{-1} \boldsymbol{A}^T \boldsymbol{C}^{-1} \boldsymbol{A} \left\langle \boldsymbol{\Lambda}^{-1} \right\rangle_{q(\boldsymbol{\alpha})}^{-1}$$

with $\boldsymbol{C} = \sigma^2 \boldsymbol{I} + \boldsymbol{A} \langle \boldsymbol{\Lambda}^{-1} \rangle_{q(\boldsymbol{\alpha})}^{-1} \boldsymbol{A}^T$ being an $M \times M$ matrix. Hence, to calculate $\boldsymbol{\Sigma}$ needs $O(\min\{N^3, N^2 M\})$ operations. It is noted that if Q-VMP produces some $\langle \alpha_n^{-1} \rangle_{q(\boldsymbol{\alpha})} \to +\infty$ with $n \in \{1, \cdots, N\}$, then the corresponding basis $\boldsymbol{A}_n$ can be removed from the model. To further speed up Q-VMP, we prune a basis $\boldsymbol{A}_n$ from the model (to reduce N) when the corresponding parameter $\langle \alpha_n^{-1} \rangle_{q(\boldsymbol{\alpha})}$ is larger than a certain threshold $\tau_{pruning}$. Similar basis pruning approaches have been used in [39, 54].

7.3 Numerical Simulations

In this section, we study the performance of the proposed observation model and Q-VMP algorithm in comparison with existing ones by numerical simulations.

7.3.1 Experimental Setup

Model Parameters in Q-VMP: We set $\epsilon = 0$, $c = 1$ and $d = 0$ in the Gaussian-Gamma-Gamma prior as discussed in Subsection 7.1.2.

Quantizer: In multi-bit CS, a uniform unsaturated quantizer is defined in (7.2) with $L = 2^B$, equispaced $u_0, u_1 \cdots, u_L$ and $v_i = (u_i + u_{i+1})/2$, $i = 1, \cdots, L-1$. In addition, we let $u_L = \|\boldsymbol{y}\|_\infty$ and $u_0 = -\|\boldsymbol{y}\|_\infty$ in each trial. For a saturated quantizer, we set $u_0 = -\infty$, $u_L = +\infty$.

CS problem generation: In our experiment, we set $N = 500$, $K = 10$, and vary the bit budget (total bits of all measurements) in $\{50, 100, \cdots, 1000\}$. In each trial, a K-sparse signal of length N is generated with Gaussian distributed nonzero entries and then scaled to unit norm. Entries of the sensing matrix $\boldsymbol{A}$ are generated independently according to a Gaussian distribution $\mathcal{N}(0, M^{-1})$. Thus the noise free measurement $\boldsymbol{y}^0 = \boldsymbol{A}\boldsymbol{x}$ has unit norm in expectation. To obtain a desired SNR, a white Gaussian measurement noise $\boldsymbol{n}$ is added with the noise variance $\sigma^2 = M^{-1}10^{-\frac{\text{SNR}}{10}}$. Then the quantized measurement $\boldsymbol{z} = \mathcal{Q}(\boldsymbol{y})$ is preserved for the following signal recovery.

Performance metrics: Three metrics are considered, including reconstruction SNR (RSNR), sparsity level of the recovered signal and computational speed. RSNR is defined as

$$\text{RSNR} = -20 \log_{10} \|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2$$

where $\widehat{\boldsymbol{x}}$ denotes the recovered signal of $\boldsymbol{x}$. The sparsity level is measured by the support size of the recovered signal. The computational speed is measured by the CPU time usage. All results are averaged over 200 trials.

7.3.2 Model Efficiency

We first study the efficiency of the observation model in (7.8) introduced in this chapter for quantized CS. We consider the multi-bit CS problem with a uniform unsaturated quantizer as an example. In existing methods that account for measurement noise, e.g., in [60], the quantization error and the noise are typically coupled and treated as a Gaussian noise (only the energy information is used). Then the quantized CS problem is transformed into a conventional one. We refer to this formulation as existing method hereafter. In this subsection we compare the signal recovery performance of the proposed formulation in (7.8) with the existing one. Naturally, we use the proposed Q-VMP algorithm for our formulation. A corresponding algorithm for the existing formulation is thus VMP introduced in [54] for conventional CS. The latter algorithm can be considered as a simplified version of Q-VMP with the quantization error $\boldsymbol{e}$ fixed throughout the algorithm. In addition, we consider the performance of conventional CS for comparison where the true-valued measurements are used. The conventional CS problem can be considered as one in quantized CS by using oracle information of the quantization error. So its performance acts as an upper boundary of the quantized CS problem.

In our experiment, we set SNR = 30dB and the bit depth $B = 4$ that leads to the number of quantized measurements varying from 12 to 250. Q-VMP is implemented as follows.

Q-VMP: We initialize $\langle \alpha_n^{-1} \rangle_{q(\boldsymbol{\alpha})} = 1/\left|\boldsymbol{A}_n^T \boldsymbol{z}\right|$, $n = 1, \cdots, N$, $\langle \eta \rangle_{q(\eta)} = 1$ and $\langle \boldsymbol{e} \rangle_{q(\boldsymbol{e})} = \boldsymbol{0}$. We set $\tau_{pruning} = 10^4$. Q-VMP is terminated if $\frac{\left\|\tilde{\boldsymbol{\alpha}}^j - \tilde{\boldsymbol{\alpha}}^{j-1}\right\|_2}{\left\|\tilde{\boldsymbol{\alpha}}^{j-1}\right\|_2} < 10^{-5}$ or the maximum number of iterations, set to 2000, is reached, where the superscript j indicates the iteration and $\tilde{\boldsymbol{\alpha}} = \left[\left\langle \alpha_1^{-1} \right\rangle_{q(\boldsymbol{\alpha})}^{-1}, \cdots, \left\langle \alpha_N^{-1} \right\rangle_{q(\boldsymbol{\alpha})}^{-1}\right]^T$.

The VMP algorithm for the other two cases is similarly implemented. The true noise variance is used in Q-VMP and conventional CS. For VMP with the existing formulation, we set the noise variance to $r^2/12 + \sigma^2$ where $r = u_1 - u_0$ denotes the quantization bin width. This noise variance corresponds to a Gaussian noise whose energy is comparable with that of $\boldsymbol{e} + \boldsymbol{n}$ under the assumption that $\boldsymbol{e}$ is uniformly distributed and independent of $\boldsymbol{n}$. Reconstruction SNRs of the three methods are depicted in Fig. 7.3. It can be seen that the VMP algorithm based on the proposed observation model is consistently better than that with the existing formulation though there is still a large gap between it and the oracle-aided case. So it confirms that the proposed model and framework improve the signal recovery accuracy by decoupling the quantization error from measurement noise.

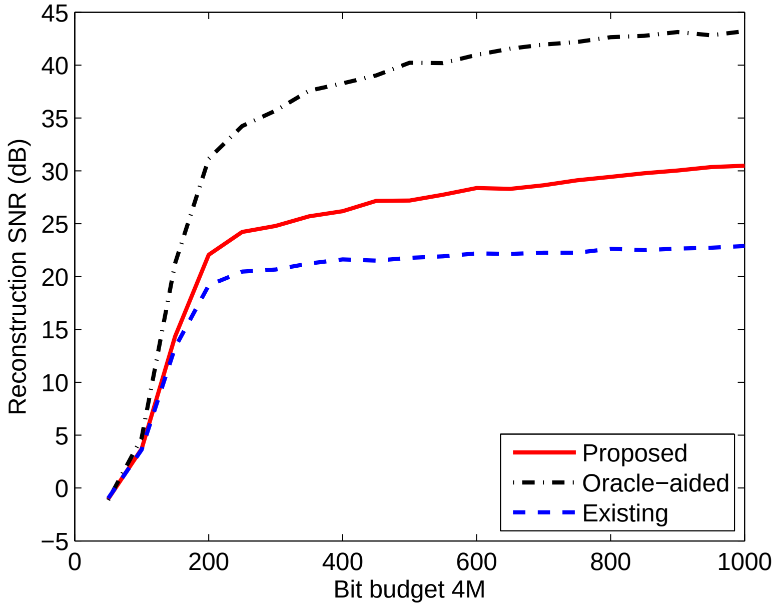


Figure 7.3: Reconstruction SNRs of VMP algorithms implemented respectively based on the proposed observation model in (7.8), an existing one that couples the quantization error and measurement noise, and conventional CS (oracle-aided quantized CS) as an upper boundary.

7.3.3 Performance Comparison in Multi-bit CS

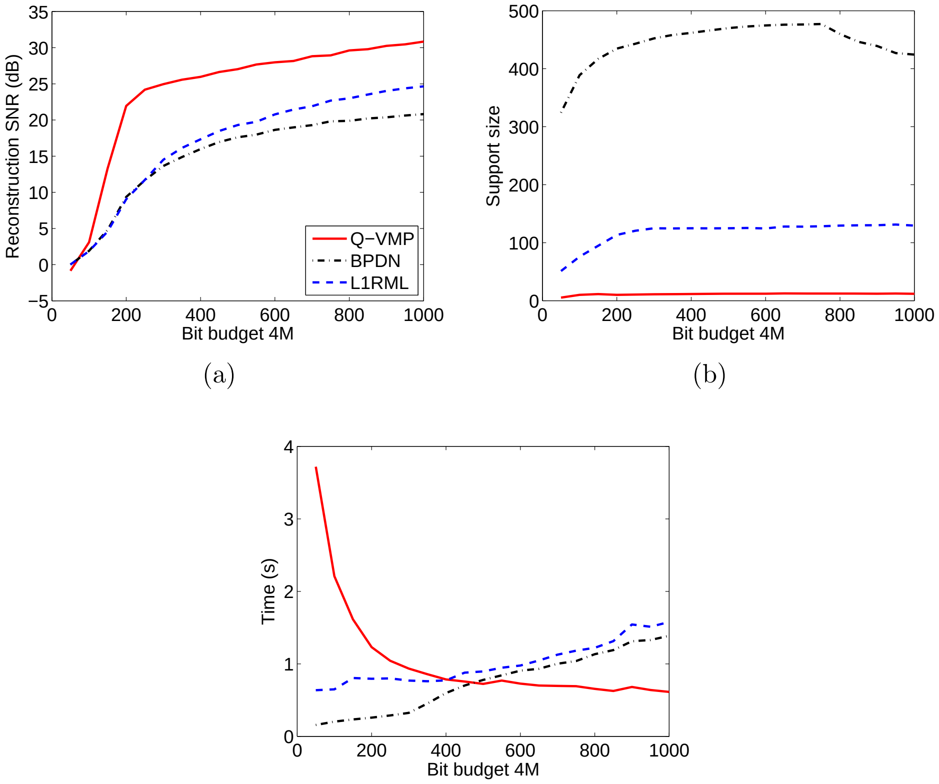
7.3.3.1 Unsaturated quantizer

In multi-bit CS, we first consider the case of a uniform unsaturated quantizer. As in the last subsection, we set SNR $=$ 30dB and $B = 4$. Besides Q-VMP, we also use BPDN [58] and L1RML [61] to recover the signal for comparison. Q-VMP is implemented as in the last subsection. The other two are implemented as follows.

BPDN: BPDN takes $\boldsymbol{z}$ as the measurement vector and is implemented using ℓ_1-magic[1]. We let $\epsilon = \|\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x}\|_2$ in our implementation to achieve the best result though it is unavailable in practice.

L1RML: The problem $\min_{\tilde{\boldsymbol{x}}} \{\lambda \|\tilde{\boldsymbol{x}}\|_1 - \log f_{ML}(\boldsymbol{A}\tilde{\boldsymbol{x}})\}$ is solved, where $f_{ML}(\cdot)$ is the likelihood function of the observation with $-\log f_{ML}(\boldsymbol{A}\tilde{\boldsymbol{x}})$ being a convex function of $\tilde{\boldsymbol{x}}$, and $\lambda > 0$ is a regularization parameter. In general, a larger λ leads to a

[1] Available at http://users.ece.gatech.edu/∼justin/l1magic



(a)

(b)

(c)

Figure 7.4: Performance comparison of Q-VMP, BPDN and L1RML with bit depth $B = 4$. SNR = 30dB. (a) Averaged reconstruction SNR; (b) Averaged support size of recovered signal; (c) Averaged CPU time.

sparser solution. Since there are no available guidelines for the choice of λ so far, we choose λ such that the recovered signal has the optimal RSNR. Additionally, we set $\tau = \frac{\sigma^2}{\|\boldsymbol{A}\|_2^2}$, $\epsilon = 10^{-4}$ and $\beta = 0.5$. Readers are referred to [61] for their interpretations.

The experimental results are shown in Fig. 7.4, where red solid lines denote Q-VMP, black dashed dot lines denote BPDN, and blue dashed lines denote L1RML. Fig. 7.4(a) depicts the averaged reconstruction SNRs of the three algorithms. A significant improvement of the reconstruction SNR can be observed using the proposed Q-VMP. It is over 6dB in comparison with L1RML and about an amplitude for BPDN. Moreover, Fig. 7.4(b) shows that Q-VMP produces the sparsest solution. It is noted that L1RML can produce a sparser solution by setting a larger value of λ as in [61] but at the cost of a lower RSNR. Fig. 7.4(c) shows that the speed of Q-VMP is comparable with that of BPDN and L1RML. Implemented with the basis pruning approach, Q-VMP is faster when more measurements are acquired since it is observed in such a case that the basis pruning approach works more efficiently.

7.3.3.2 Saturated Quantizer

We next consider the case of a saturated quantizer. We adopt the same experimental setup but a saturated quantizer where a noisy measurement falls in each quantization interval with the same probability. Since both the sensing matrix and measurement noise are Gaussian in the experiment, the noisy measurements are i.i.d. Gaussian $\mathcal{N}\left(0, M^{-1} + \sigma^2\right)$. Then it is easy to get the quantizer. As a result, 12.5% of the measurements are saturated in expectation. BPDN is inappropriate in such a case. We compare Q-VMP only with L1RML. The averaged reconstruction SNRs of Q-VMP and L1RML are presented in Fig. 7.5 (red solid lines). Q-VMP obtains a RSNR of about 10dB higher than L1RML when sufficient measurements are acquired. The performance of the two algorithms on support size and speed is similar to that in the uniform quantizer case and is omitted.

The experiment above may shed light on the optimal quantizer design for Q-VMP. By comparing the performance of Q-VMP in the two quantizer scenarios, it can be

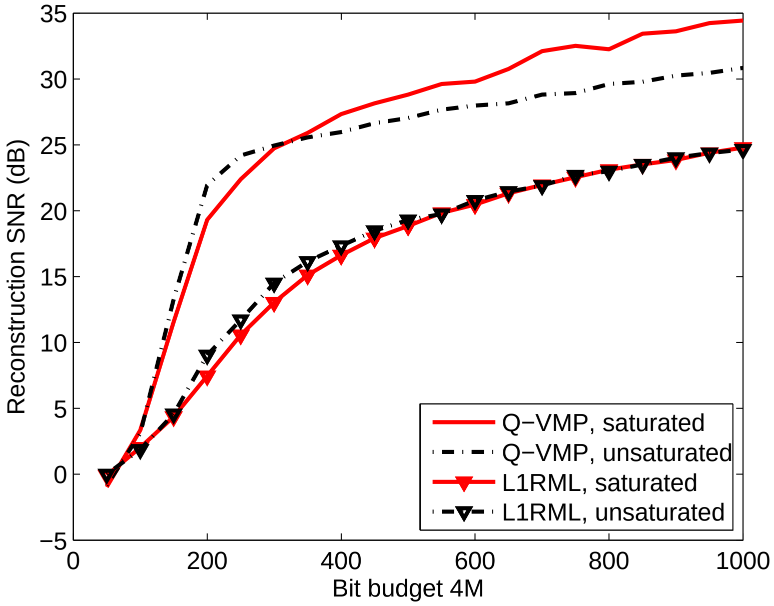


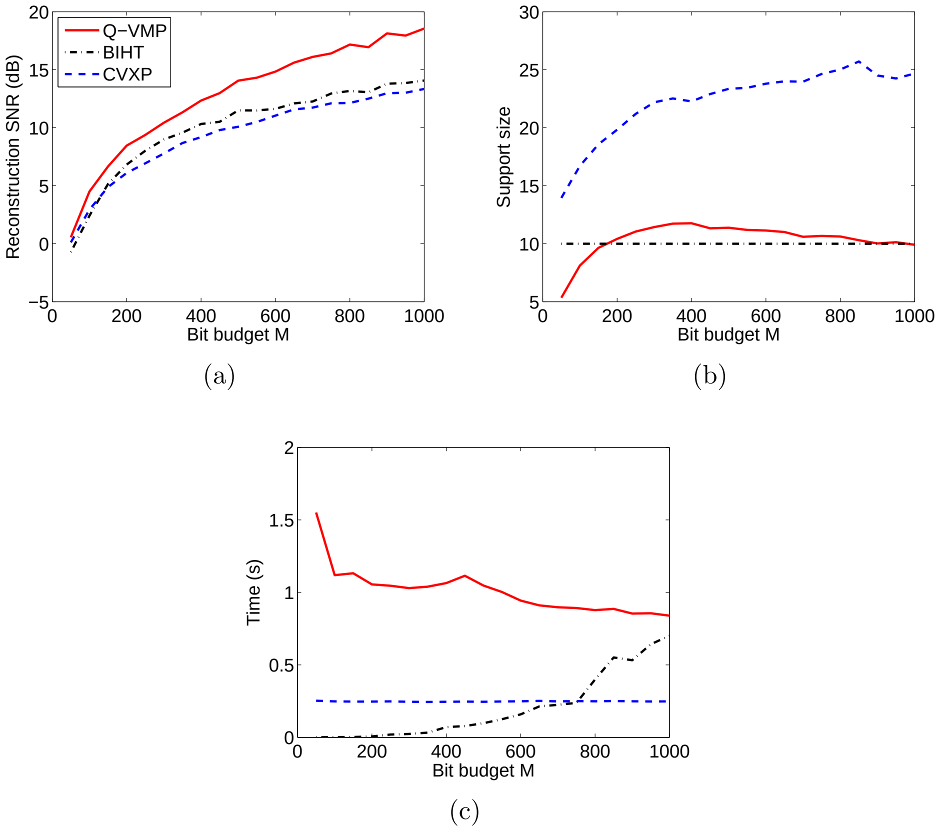
Figure 7.5: Reconstruction SNRs of Q-VMP and L1RML with a saturated quantizer, as well as those with the unsaturated quantizer in Fig. 7.4.

seen from Fig. 7.5 that the saturated quantizer outperforms the uniform unsaturated one when more measurements are taken for Q-VMP while it is not so clear for L1RML. We pose the problem of the optimal quantizer design for Q-VMP as a future work.

7.3.4 Performance Comparison in 1-bit CS

The bit-depth $B = 1$ in 1-bit CS. We set SNR $=$ 10dB. In such a case, 9.75% measurements flip their signs due to the noise in expectation. We compare Q-VMP with the state-of-the-art algorithms BIHT [66] and the convex programming approach in [68], denoted by CVXP. The three algorithms are implemented as follows.

Q-VMP: We initialize $\langle\alpha_n^{-1}\rangle_{q(\boldsymbol{\alpha})} = \sqrt{M}/\left|\boldsymbol{A}_n^T\text{sgn}\left(\boldsymbol{z}\right)\right|$, $n = 1,\cdots,N$, $\langle\eta\rangle_{q(\eta)} = 1$ and $\langle\boldsymbol{e}\rangle_{q(\boldsymbol{e})} = -\text{sgn}\left(\boldsymbol{z}\right)/\sqrt{M}$. As addressed in Remark 7.4, the effective noise level in 1-bit CS is much lower than the true one. We empirically find that it is a good choice to set the noise variance in Q-VMP to $10^{-3}\sigma^2$ (Q-VMP is slow in the case of a very small noise variance, which is the reason why we consider a lower SNR



(a)

(b)

(c)

Figure 7.6: Performance comparison of Q-VMP, BIHT and CVXP in 1-bit CS. SNR = 10dB. (a) Averaged reconstruction SNR; (b) Averaged support size of recovered signal; (c) Averaged CPU time.

in 1-bit CS). We set $\tau_{pruning} = 10^4$ and terminate Q-VMP as in multi-bit CS. The recovered signal is finally scaled to unit norm for comparison with the original one.

BIHT: The oracle information of K is used in the hard thresholding operation, i.e., BIHT is certain to return a reconstruction with K nonzero entries. BIHT is terminated if the Hamming error of the current recovery is below the expected Hamming error or the maximum number of iterations, set to 1000, is reached. Readers are referred to [66] for the definition of the Hamming error.

CVXP: The oracle information of $\|\boldsymbol{x}\|_1$ is used to achieve the best result. We implement it using CVX [124] (its speed can be accelerated in the future using more sophisticated solvers).

Our experimental results are presented in Fig. 7.6. In all figures, red solid lines denote Q-VMP, black dashed dot lines denote BIHT, and blue dashed lines denote CVXP. It is shown in Fig. 7.6(a) that the proposed Q-VMP outperforms consistently the other two algorithms in the recovery accuracy. From Fig. 7.6(b), it can be seen that Q-VMP produces a sparser solution than CVXP while BIHT uses this oracle information. Fig. 7.6(c) shows that the computational speed is a disadvantage of Q-VMP.

7.4 Conclusion

The problem of sparse signal recovery from noisy quantized compressive measurements was studied in this chapter. A Bayesian framework was presented that unifies the multi- and 1-bit CS problems and can be applied to various scenarios including noiseless/noisy environment and unsaturated/saturated quantizer. An algorithm was proposed based on variational Bayesian inference under the proposed framework. The quantization error is decoupled from the measurement noise, modeled as a random variable and estimated jointly with the signal of interest, leading to improved performance of the algorithm. Numerical simulations were provided to demonstrate its improved signal recovery accuracy over the existing results.

Part II

Applications

Chapter 8

Off-Grid Direction of Arrival Estimation

Chapters 3–7 have been focused on the theoretical analysis and algorithm design of CS. In this chapter and the next the applications of CS to direction of arrival (DOA) estimation and magnetic resonance imaging (MRI) will be studied, respectively. The DOA estimation is a classical problem in array signal processing [84] with many practical applications. Conventional covariance-based approaches, e.g., MUSIC [86,87], require a large number of snapshots for its success. Its research has recently been advanced owing to the development of methods based on sparse signal representation or CS [88,89,139,140]. While these methods have shown advantages over conventional ones, there are still difficulties in practical situations where the true DOAs are not on the discretized sampling grid. The resulting DOA estimation problem is referred to as off-grid DOA estimation in this thesis. To deal with the problem, this chapter studies an off-grid model which takes into account effects of the off-grid DOAs and has a smaller modeling error. Based on the model we introduce two new approaches to the problem solving. One is based on the results of Chapter 5. We show that the DOA estimation problem can be transformed into the perturbed CS problem studied in Chapter 5 under the off-grid model. Then the ℓ_1 optimization approach presented in Chapter 5 can be applied straightforward. But one shortcoming of this optimization approach is the high computational complexity. The other has reduced computational workload and adopts SBL to do the estimation in an iterative manner. It can be applied to both single snapshot and multi-snapshot cases. Numerical simulations show that both the proposed approaches have improved accuracy and can maintain high estimation accuracy even under a very coarse sampling grid. This chapter is mainly based on [7] and [11].

8.1 Problem Description and Off-Grid Model

8.1.1 Problem Description

Consider K narrowband far-field sources $s_k(t)$, $k = 1, \cdots, K$, impinging on an array of M omnidirectional sensors from directions d_k, $k = 1, \cdots, K$. According to [84], the time delays at different sensors can be represented by simple phase shifts, leading to the observation model:

$$\boldsymbol{y}(t) = \boldsymbol{A}(\boldsymbol{d})\boldsymbol{s}(t) + \boldsymbol{e}(t), \quad t = 1, \cdots, T, \tag{8.1}$$

where $\boldsymbol{y}(t) = [y_1(t), \cdots, y_M(t)]^T \in \mathbb{C}^M$, $\boldsymbol{d} = [d_1, \cdots, d_K]^T \in [0°, 180°)^K$, $\boldsymbol{s}(t) = [s_1(t), \cdots, s_K(t)]^T \in \mathbb{C}^K$, $\boldsymbol{e}(t) = [e_1(t), \cdots, e_M(t)]^T \in \mathbb{C}^M$, and $y_m(t)$ and $e_m(t)$, $m = 1, \cdots, M$, denote the output and measurement noise of the mth sensor at time t respectively. The matrix $\boldsymbol{A}(\boldsymbol{d}) = [\boldsymbol{a}(d_1), \cdots, \boldsymbol{a}(d_K)] \in \mathbb{C}^{M\times K}$ is an array manifold matrix and $\boldsymbol{a}(d_k) \in \mathbb{C}^M$ is called steering vector of the kth source which is determined by the geometry of the sensor array and is known. The entry $\boldsymbol{a}_m(d_k)$ contains the delay information of the kth source to the mth sensor. Note that the number of sources K may be unknown in practice. So, the objective of the DOA estimation is to find the unknown DOA vector $\boldsymbol{d}$ given the sensor measurements $\boldsymbol{y}(t)$ and the mapping $\boldsymbol{d} \rightarrow \boldsymbol{A}(\boldsymbol{d})$.

8.1.2 Off-Grid DOA Estimation Model

Since the DOAs of interest lie in a continuous range and CS is applicable to recovery of discrete signals, discretization of the DOA range is inevitable for applying CS. In [88, 89] a fixed sampling grid is selected which serves as the set of all candidates of DOA estimates. By assuming that all the true (unknown) DOAs are exactly on the selected grid, a CS problem can be formulated where the DOAs of interest constitute the support of the sparse signal to be recovered.

In practical situations, the true DOAs may not be on the selected grid. Using linear approximation an off-grid model was firstly proposed in [78] to account for the general case of off-grid DOA. Let $\widetilde{\boldsymbol{d}} = \left\{\widetilde{d}_1, \cdots, \widetilde{d}_N\right\}$ be a fixed sampling grid in the DOA range $[0, \pi)$, where N denotes the grid number and typically satisfies $N \gg M > K$. Suppose $d_k \notin \left\{\widetilde{d}_1, \cdots, \widetilde{d}_N\right\}$ for some $k \in \{1, \cdots, K\}$ and that $\widetilde{d}_{n_k}$, $n_k \in \{1, \cdots, N\}$, is the nearest grid point to d_k. We approximate the steering vector $\boldsymbol{a}(d_k)$ using linearization:

$$\boldsymbol{a}(d_k) \approx \boldsymbol{a}\left(\widetilde{d}_{n_k}\right) + \boldsymbol{b}\left(\widetilde{d}_{n_k}\right)\left(d_k - \widetilde{d}_{n_k}\right) \tag{8.2}$$

with $\boldsymbol{b}\left(\widetilde{d}_{n_k}\right) = \boldsymbol{a}'\left(\widetilde{d}_{n_k}\right)$. Denote $\boldsymbol{A} = \left[\boldsymbol{a}\left(\widetilde{d}_1\right), \cdots, \boldsymbol{a}\left(\widetilde{d}_N\right)\right]$, $\boldsymbol{B} = \left[\boldsymbol{b}\left(\widetilde{d}_1\right), \cdots, \boldsymbol{b}\left(\widetilde{d}_N\right)\right]$ and $\boldsymbol{\Phi}(\boldsymbol{\beta}) = \boldsymbol{A} + \boldsymbol{B}\mathrm{diag}(\boldsymbol{\beta})$, where $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_N]^T$, for $n = 1, \cdots, N$,

$$\begin{aligned} &\beta_n = d_k - \widetilde{d}_{n_k}, \quad x_n(t) = s_{n_k}(t), \quad \text{if } n = n_k \text{ for any } k \in \{1, \cdots, K\}; \\ &\beta_n = 0, \quad x_n(t) = 0, \quad \text{otherwise,} \end{aligned} \tag{8.3}$$

with $n_k \in \{1, \cdots, N\}$ and $\widetilde{d}_{n_k}$ being the nearest grid to a source d_k, $k \in \{1, \cdots, K\}$. By absorbing the approximation error into the measurement noise the observation model in (8.1) can be written into

$$\boldsymbol{y}(t) = \boldsymbol{\Phi}(\boldsymbol{\beta})\boldsymbol{x}(t) + \boldsymbol{e}(t), \quad t = 1, \cdots, T, \tag{8.4}$$

which is the off-grid model to be used in this chapter.

It should be noted that the off-grid model in (8.4) is closely related to the on-grid one that can be obtained by setting $\boldsymbol{\beta} = \mathbf{0}$ in (8.4) ($\boldsymbol{\Phi}(\mathbf{0}) = \boldsymbol{A}$). In fact, the off-grid model can be considered as the first order approximation of the true observation model while the on-grid one is the zeroth order approximation. As a result, the off-grid model has a much smaller modeling error than the on-grid one. Such an advantage is twofold. First, by adopting the same sampling grid the off-grid model results in higher accuracy, especially in the high SNR regime where the modeling error is the dominant modeling uncertainty. Second, a coarser sampling grid can be adopted in the off-grid model to achieve a considerably reduced computational workload with a comparable modeling accuracy.

To estimate the DOA vector $\boldsymbol{d}$ we need to find the support of the sparse signals $\boldsymbol{x}(t)$, $t = 1, \cdots, T$, as well as the off-grid difference $\boldsymbol{\beta}$.

8.2 Off-Grid DOA Estimation Using P-BPDN

8.2.1 Off-Grid DOA Estimation Under SP-CS

In this section we show that the off-grid DOA estimation problem can be formulated into a CS problem subject to structured matrix perturbation which has been studied in Chapter 5 and coined as SP-CS. For simplicity, we consider the single-snapshot and noise-free case with a uniform linear array (ULA). Correspondingly, we have $a_m(d_k) = \exp\left\{j\pi\left(m - \frac{M+1}{2}\right)\cos d_k\right\}$, $k = 1, \cdots, K$, $m = 1, \cdots, M$, where $j = \sqrt{-1}$. Denote $\boldsymbol{\theta} = [\cos d_1, \ldots, \cos d_K]^T \in (-1, 1]^K$ and the corresponding steering vector $\widetilde{\boldsymbol{a}}(\theta_k) = \boldsymbol{a}(d_k)$. We consider a mathematically equivalent problem of the estimation of $\boldsymbol{\theta}$ rather than that of $\boldsymbol{d}$ in this section following from [7]. Without loss of generality, let $\widetilde{\boldsymbol{\theta}} = \left\{\frac{1}{N} - 1, \frac{3}{N} - 1, \cdots, 1 - \frac{1}{N}\right\}$ be a uniform sampling grid in the continuous range $(-1, 1]$ of $\boldsymbol{\theta}$ (assume N is an even number). In existing standard CS based methods $\widetilde{\boldsymbol{\theta}}$ actually serves as the set of all candidate DOA estimates. As a result, their estimation accuracy is limited by the grid density since for some θ_j,

$j \in \{1, \cdots, K\}$, the best estimate of θ_j is its nearest grid point in $\widetilde{\boldsymbol{\theta}}$. It can be easily shown that a lower bound for the mean squared estimation error of each θ_j is $\mathrm{LB} = \frac{1}{3N^2}$ by assuming that θ_j is uniformly distributed in one or more grid intervals.

In the following we transform the off-grid model into the one studied in Chapter 5. Similar to (8.2), consider the Taylor expansion

$$\widetilde{\boldsymbol{a}}\left(\theta_j\right) = \widetilde{\boldsymbol{a}}\left(\widetilde{\theta}_{l_j}\right) + \widetilde{\boldsymbol{b}}\left(\widetilde{\theta}_{l_j}\right)\left(\theta_j - \widetilde{\theta}_{l_j}\right) + \boldsymbol{R}_j \tag{8.5}$$

with $\widetilde{\boldsymbol{b}}\left(\widetilde{\theta}_{l_j}\right) = \widetilde{\boldsymbol{a}}'\left(\widetilde{\theta}_{l_j}\right)$ and $\boldsymbol{R}_j$ being a remainder term with respect to θ_j. Denote $\kappa = \frac{\pi}{2}\sqrt{\frac{M^2-1}{3}}$, $\widetilde{\boldsymbol{A}} = \frac{1}{\sqrt{M}}\left[\widetilde{\boldsymbol{a}}\left(\widetilde{\theta}_1\right), \cdots, \widetilde{\boldsymbol{a}}\left(\widetilde{\theta}_N\right)\right]$, $\widetilde{\boldsymbol{B}} = \frac{1}{\kappa\sqrt{M}}\left[\widetilde{\boldsymbol{b}}\left(\widetilde{\theta}_1\right), \cdots, \widetilde{\boldsymbol{b}}\left(\widetilde{\theta}_N\right)\right]$, and for $l = 1, \cdots, N$,

$$\begin{aligned} &\beta_l^o = \kappa\left(\theta_j - \widetilde{\theta}_{l_j}\right), \ x_l^o = s_j, \text{ if } l = l_j \text{ for any } j \in \{1, \cdots, K\}; \\ &\beta_l^o = 0, \qquad\qquad\quad x_l^o = 0, \ \text{ otherwise,} \end{aligned}$$

with $l_j \in \{1, \cdots, N\}$ and $\widetilde{\theta}_{l_j}$ being the nearest grid to a source θ_j, $j \in \{1, \cdots, K\}$. It is easy to show that $\left\|\boldsymbol{x}^o\right\|_0 \leq K$, each column of $\widetilde{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{B}}$ has unit norm, and $\boldsymbol{\beta}^o \in [-r, r]^N$ with $r = \frac{\kappa}{N} = \frac{\pi}{2N}\sqrt{\frac{M^2-1}{3}}$. In addition, we let $\boldsymbol{e} = \boldsymbol{R}\boldsymbol{s}$ with $\boldsymbol{R} = \left[\boldsymbol{R}_1, \cdots, \boldsymbol{R}_K\right]$. We set $\epsilon = \frac{\sqrt{K}\|\boldsymbol{s}\|_2\pi^2}{8N^2}\sqrt{\frac{3M^4-10M^2+7}{15}}$ such that $\|\boldsymbol{e}\|_2 \leq \epsilon$ (the information of K and $\|\boldsymbol{s}\|_2$ is used). The reason is as follows. By (8.5), we have for $l = 1, \cdots, M$, $j = 1, \cdots, K$,

$$R_{lj} = \frac{A_{lj}''(\xi)}{2}\left(\theta_j - \widetilde{\theta}_{l_j}\right)^2, \tag{8.6}$$

where ξ is between θ_j and $\widetilde{\theta}_{l_j}$, $A_{lj}''(\xi) = -\frac{\pi^2}{\sqrt{M}}\left(l - \frac{M+1}{2}\right)^2 \exp\left\{i\pi\left(l - \frac{M+1}{2}\right)\xi\right\}$, and $\left|\theta_j - \widetilde{\theta}_{l_j}\right| \leq \frac{1}{N}$. It follows that for $j = 1, \cdots, K$,

$$\left\|\boldsymbol{R}_j\right\|_2 \leq \frac{1}{2}\max\left\|\boldsymbol{A}_j''\right\|_2 \cdot \frac{1}{N^2} = \frac{\pi^2}{8N^2}\sqrt{\frac{3M^4-10M^2+7}{15}}.$$

Finally, it gives the expression of ϵ by observing that

$$\left\|\boldsymbol{e}\right\|_2 = \left\|\boldsymbol{R}\boldsymbol{s}\right\|_2 \leq \left\|\boldsymbol{R}\right\|_{\mathrm{F}} \left\|\boldsymbol{s}\right\|_2 \leq \sqrt{K} \left\|\boldsymbol{R}_1\right\|_2 \left\|\boldsymbol{s}\right\|_2 .$$

After that, the DOA estimation model can be written into the form of our studied model in (5.9) of Chapter 5. The only differences are that $\widetilde{\boldsymbol{A}}$, $\widetilde{\boldsymbol{B}}$, $\boldsymbol{x}^o$ and $\boldsymbol{e}$ are in the complex domain rather than the real domain and that $\boldsymbol{e}$ denotes a modeling error term rather than the measurement noise. It is noted that the robust stability results in SP-CS apply straightforward to such complex signal case with few modifications. The objective becomes the recovery of $\boldsymbol{x}^o$ (its support actually) and $\boldsymbol{\beta}^o$. According to Theorem 5.4 $\boldsymbol{x}^o$ and $\boldsymbol{\beta}^o$ can be stably recovered if the D-RIP condition is satisfied. Denote $\widehat{\boldsymbol{x}}$ the recovered $\boldsymbol{x}^o$, $\widehat{\boldsymbol{\beta}}$ the recovered $\boldsymbol{\beta}^o$, and $\mathcal{I}$ the support of $\widehat{\boldsymbol{x}}$. Then we obtain the recovered $\boldsymbol{\theta}$: $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\mathcal{I}} + \kappa^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{I}}$, where $\boldsymbol{v}_{\mathcal{I}}$ keeps only entries of a vector $\boldsymbol{v}$ on the index set $\mathcal{I}$. The empirical results in Subsection 8.2.2 will illustrate the merits of applying the SP-CS framework to estimate DOAs.

8.2.2 Numerical Simulations

This subsection studies the empirical performance of the application of the SP-CS framework in the DOA estimation. We consider the case of $N = 90$ and $K = 2$. Numerical calculations show that the D-RIP condition $\bar{\delta}_{4K}\left(\boldsymbol{\Psi}\right) < \left(\sqrt{2\left(1 + r^2\right)} + 1\right)^{-1}$ in Theorem 5.4 is satisfied if $M \geq 145$. Though it ceases to be a "compressed" sensing problem when $M \geq N$, it still makes sense in SP-CS since there are $2N$ variables to be estimated and hence the P-BPDN problem is still underdetermined as $M < 2N$. As noted in Subsection 5.2.3, the D-RIP condition can be possibly relaxed using recent techniques in standard CS, which may reduce the required M value. In addition, the RIP condition is a sufficient condition for guaranteed signal recovery accuracy while its conservativeness in standard CS has been studied in [30]. We next choose a much smaller $M = 30$ ($r \approx 0.302$ in such a case) and show the empirical performance of the proposed SP-CS framework on such off-grid DOA estimation.

The experimental setup is as follows. In each trial, the complex source signal $\boldsymbol{s}$ is generated with both entries having unit amplitude and random phases. θ_1 and θ_2 are generated uniformly from intervals $\left[\frac{2}{N}, \frac{4}{N}\right]$ and $\left[\frac{12}{N}, \frac{14}{N}\right]$ respectively ($5.1^\circ \sim 7.7^\circ$ apart in the DOA domain). The P-BPDN problem in (5.11) is solved using AA-P-BPDN presented in Subsection 5.3.2 whose settings are the same as those in Section 5.4. Our experimental results of the estimation error $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ for both sources are presented in Fig. 8.1 where 1000 trials are used. It can be seen that P-BPDN performs well on the off-grid DOA estimation. All estimation errors lie in the interval $\left[-\frac{1}{N}, \frac{1}{N}\right]$ with most very close to zero. To achieve a possibly comparable mean squared estimation error, a grid of length at least $N = 360$ has to be used in standard CS based methods according to the lower bound mentioned in Subsection 8.2.1. An example of performance of SP-CS and standard CS on DOA estimation is shown in Fig. 8.2, where the two approaches share the same data set and $N = 360$ is set in standard CS. From the upper two sub-figures, it can be seen that SP-CS performs well on both source signal and $\boldsymbol{\beta}^o$ recoveries. From the lower left one, however, it can be seen that two nonzero entries are presented in the recovered signal around the location of each source when using standard CS. Such a phenomenon is much clearer in the last sub-figure, where it can be observed that a single peak exhibits at a place very close to the true location of source 1 using the proposed SP-CS framework while two peaks occur at places further away from the true source in standard CS.

8.3 Off-Grid DOA Estimation Using SBL

The ℓ_1 optimization-based off-grid DOA estimation approach within the framework of structured CS presented in the last section has strong theoretical motivation. However, the AA-P-BPDN algorithm is computationally inefficient since a series of BPDN problems need to be solved iteratively. In this section, we propose a computationally efficient algorithm to solve the off-grid DOA estimation problem under the framework of SBL, where the sparse source signal and the grid offset are

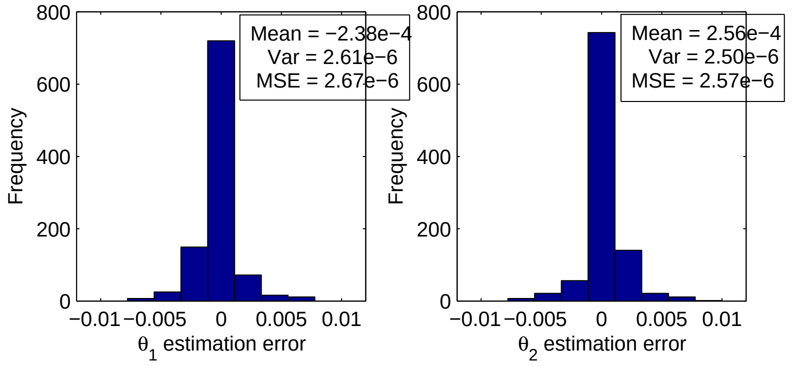


Figure 8.1: Histogram of $\boldsymbol{\theta}$ estimation error for both sources using P-BPDN for SP-CS. Statistics including mean, variance and mean squared error (MSE) are shown.

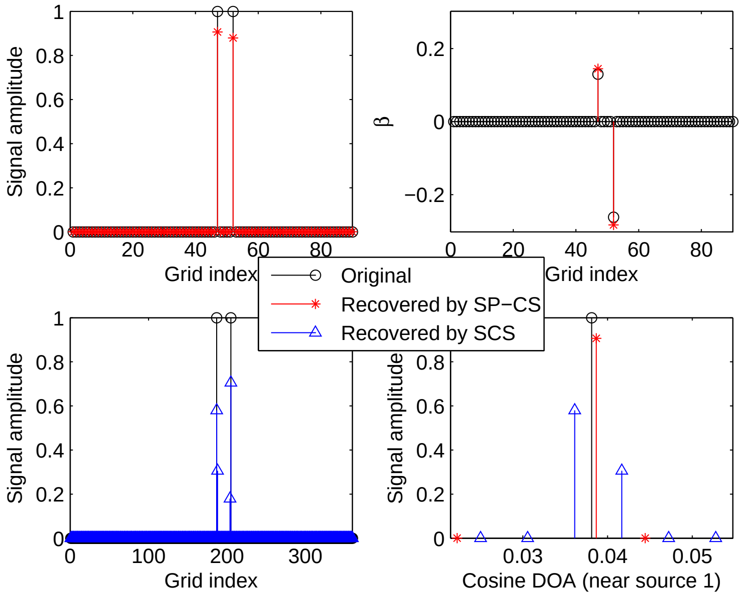


Figure 8.2: Performance comparison of SP-CS and standard CS (SCS) on DOA estimation. Upper left: signal recovery in SP-CS; upper right: $\boldsymbol{\beta}^o$ recovery in SP-CS (shown only on the signal support); lower left: signal recovery in standard CS; lower right: signal amplitude versus θ (near the location of source 1) in SP-CS and standard CS.

modeled as random variables and jointly estimated in an iterative manner. This section is mainly based on the work in [11]. For convenience, we follow [11] and consider the parameter estimation in the DOA domain and hence the observation model in (8.4) is adopted. As a result, the notations adopted here may have slightly different meanings from those in the last section. In addition, we assume that the number of sources K is *a priori* known. In fact, the presented approach can still be applied when K is unknown. In the latter case, we may let $K = K_{max}$ where K_{max} denotes the maximum number of sources detectable using the sensor array adopted. For example, $K_{max} = M - 1$ if an M-element ULA is adopted.

Without loss of generality, let $\widetilde{\boldsymbol{\theta}} \in [0^\circ, 180^\circ)^N$ be a uniform grid in the continuous DOA range $[0^\circ, 180^\circ)$ with a grid interval $r = \widetilde{\theta}_2 - \widetilde{\theta}_1 \propto N^{-1}$. We derive our algorithm in the multiple measurement vector (MMV) case. The single measurement vector (SMV) scenario is a special case by simply setting $T = 1$. Denote $\boldsymbol{Y} = [\boldsymbol{y}(1), \cdots, \boldsymbol{y}(T)] \in \mathbb{C}^{M\times T}$, $\boldsymbol{X} = [\boldsymbol{x}(1), \cdots, \boldsymbol{x}(T)] \in \mathbb{C}^{N\times T}$ and $\boldsymbol{E} = [\boldsymbol{e}(1), \cdots, \boldsymbol{e}(T)] \in \mathbb{C}^{M\times T}$. The off-grid DOA estimation model in (8.4) becomes

$$\boldsymbol{Y} = \boldsymbol{\Phi}\left(\boldsymbol{\beta}\right)\boldsymbol{X} + \boldsymbol{E} \tag{8.7}$$

with $\boldsymbol{\Phi}\left(\boldsymbol{\beta}\right) = \boldsymbol{A} + \boldsymbol{B}\text{diag}\left(\boldsymbol{\beta}\right) \in \mathbb{C}^{M\times N}$ and $\boldsymbol{\beta} \in \left[-\frac{1}{2}r, \frac{1}{2}r\right]^N$. The matrix $\boldsymbol{X}$ of interest is jointly sparse (or row-sparse), i.e., all columns of $\boldsymbol{X}$ are sparse and share the same support.

8.3.1 Sparse Bayesian Formulation

8.3.1.1 Noise model

Under an assumption of white (circular symmetric) complex Gaussian [141] noises, we have

$$p\left(\boldsymbol{E}|\alpha_0\right) = \prod_{t=1}^{T} \mathcal{CN}\left(\boldsymbol{e}(t)|\boldsymbol{0}, \alpha_0^{-1}\boldsymbol{I}\right) \tag{8.8}$$

where $\alpha_0 = \sigma^{-2}$ denotes the noise precision with σ^2 being the noise variance, the probability density function (PDF) of a (circular symmetric) complex Gaussian distributed random variable $\boldsymbol{u} \sim \mathcal{CN}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is [141]

$$\mathcal{CN}\left(\boldsymbol{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\pi^N \left|\boldsymbol{\Sigma}\right|} \exp\left\{-\left(\boldsymbol{u}-\boldsymbol{\mu}\right)^H \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{u}-\boldsymbol{\mu}\right)\right\}. \tag{8.9}$$

Then we have

$$p\left(\boldsymbol{Y}|\boldsymbol{X}, \alpha_0, \boldsymbol{\beta}\right) = \prod_{t=1}^{T} \mathcal{CN}\left(\boldsymbol{y}(t)|\boldsymbol{\Phi}\left(\boldsymbol{\beta}\right)\boldsymbol{x}(t), \alpha_0^{-1}\boldsymbol{I}\right). \tag{8.10}$$

In this chapter we assume that the noise precision α_0 is unknown. A Gamma hyperprior is assumed for α_0 since it is a conjugate prior of the Gaussian distribution:

$$p\left(\alpha_0; c, d\right) = \Gamma\left(\alpha_0|c, d\right) \tag{8.11}$$

where $\Gamma\left(\alpha_0|c, d\right) = \left[\Gamma\left(c\right)\right]^{-1} d^c {\alpha_0}^{c-1} \exp\left\{-d\alpha_0\right\}$ with $\Gamma\left(\cdot\right)$ being the Gamma function. We set $c, d \to 0$ as in [38, 39] to obtain a broad hyperprior.

Remark 8.1. *The noise term in fact contains two components: the measurement noise and the modeling error. In the low and moderate SNR regimes which are of main interest the measurement noise is the dominant uncertainty and thus the modeling error can be neglected.*

8.3.1.2 Sparse signal model

A sparse prior is needed for the jointly sparse matrix $\boldsymbol{X}$ of interest. We assume that the signals among snapshots are independent and adopt the following two-stage hierarchical prior: $p\left(\boldsymbol{X}; \rho\right) = \int p\left(\boldsymbol{X}|\boldsymbol{\alpha}\right) p\left(\boldsymbol{\alpha}; \rho\right) d\boldsymbol{\alpha}$, where $\rho > 0$, $\boldsymbol{\alpha} \in \mathbb{R}^N$, $\boldsymbol{\Lambda} = \operatorname{diag}\left(\boldsymbol{\alpha}\right)$ and

$$p\left(\boldsymbol{X}|\boldsymbol{\alpha}\right) = \prod_{t=1}^{T} \mathcal{CN}\left(\boldsymbol{x}(t)|\boldsymbol{0}, \boldsymbol{\Lambda}\right), \tag{8.12}$$

$$p\left(\boldsymbol{\alpha};\rho\right)=\prod_{n=1}^{N}\Gamma\left(\alpha_n|1,\rho\right). \tag{8.13}$$

It is easy to show that all columns of $\boldsymbol{X}$ are independent and share the same prior. According to [46], for $t=1,\cdots,T$ both $\Re\left\{\boldsymbol{x}(t)\right\}$ and $\Im\left\{\boldsymbol{x}(t)\right\}$ are identically Laplace distributed and strongly peaked at the origin. As a result, the two-stage hierarchical prior is a sparse prior that favors most rows of $\boldsymbol{X}$ being zero.

8.3.1.3 Off-grid distance model

We assume a uniform prior for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}\sim U\left(\left[-\frac{1}{2}r,\frac{1}{2}r\right]^N\right). \tag{8.14}$$

The prior is noninformative in the sense that the only information of $\boldsymbol{\beta}$ we use is its boundedness.

By combining the stages of the hierarchical Bayesian model, the joint PDF is

$$p\left(\boldsymbol{X},\boldsymbol{Y},\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta}\right)=p\left(\boldsymbol{Y}|\boldsymbol{X},\alpha_0,\boldsymbol{\beta}\right)p\left(\boldsymbol{X}|\boldsymbol{\alpha}\right)p\left(\boldsymbol{\alpha}\right)p\left(\alpha_0\right)p\left(\boldsymbol{\beta}\right) \tag{8.15}$$

with the distributions on the right hand side as defined by (8.10), (8.12), (8.13), (8.11) and (8.14) respectively.

8.3.2 Bayesian Inference

An evidence procedure [55] is exploited to perform the Bayesian inference since the exact posterior distribution $p\left(\boldsymbol{X},\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{Y}\right)$ cannot be explicitly calculated. Similar approaches have been used in standard Bayesian CS methods [38,46]. First it is easy to show that the posterior distribution of $\boldsymbol{X}$ is a complex Gaussian distribution:

$$p\left(\boldsymbol{X}|\boldsymbol{Y},\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta}\right)=\prod_{t=1}^{T}\mathcal{CN}\left(\boldsymbol{x}(t)|\boldsymbol{\mu}(t),\boldsymbol{\Sigma}\right) \tag{8.16}$$

with

$$\boldsymbol{\mu}(t) = \alpha_0 \boldsymbol{\Sigma} \boldsymbol{\Phi}^H \boldsymbol{y}(t), \quad t = 1, \cdots, T, \tag{8.17}$$

$$\boldsymbol{\Sigma} = \left(\alpha_0 \boldsymbol{\Phi}^H \boldsymbol{\Phi} + \boldsymbol{\Lambda}^{-1}\right)^{-1}. \tag{8.18}$$

Calculations of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}(t)$, $t = 1, \cdots, T$, need estimates of the hyperparameters α_0, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In an evidence procedure, they are estimated using an MAP estimate that maximizes $p\left(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{Y}\right)$. It can be easily observed that to maximize $p\left(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{Y}\right)$ is equivalent to maximizing the joint PDF $p\left(\boldsymbol{Y}, \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = p\left(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{Y}\right) p\left(\boldsymbol{Y}\right)$ since $p\left(\boldsymbol{Y}\right)$ is independent of the hyperparameters. An expectation-maximization (EM) algorithm is implemented that treats $\boldsymbol{X}$ as a hidden variable and turns to maximizing $E\left\{\log p\left(\boldsymbol{X}, \boldsymbol{Y}, \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right\}$, where $p\left(\boldsymbol{X}, \boldsymbol{Y}, \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ is given in (8.15) and $E\left\{\cdot\right\}$ denotes an expectation with respect to the posterior of $\boldsymbol{X}$ as given in (8.16) using the current estimates of the hyperparameters.

Denote $\boldsymbol{\mathcal{U}} = \left[\boldsymbol{\mu}(1), \cdots, \boldsymbol{\mu}(T)\right] = \alpha_0 \boldsymbol{\Sigma} \boldsymbol{\Phi}^H \boldsymbol{Y}$, $\underline{\boldsymbol{X}} = \boldsymbol{X}/\sqrt{T}$, $\underline{\boldsymbol{Y}} = \boldsymbol{Y}/\sqrt{T}$, $\underline{\boldsymbol{\mathcal{U}}} = \boldsymbol{\mathcal{U}}/\sqrt{T}$ and $\underline{\rho} = \rho/T$. Following a similar procedure as in [39], it is easy to obtain the following updates of $\boldsymbol{\alpha}$ and α_0:

$$\alpha_n^{new} = \frac{\sqrt{1 + 4\underline{\rho} E\left\{\left\|\underline{\boldsymbol{X}}^n\right\|_2^2\right\}} - 1}{2\underline{\rho}}, \quad n = 1, \cdots, N, \tag{8.19}$$

$$\alpha_0^{new} = \frac{M + (c-1)/T}{E\left\{\left\|\underline{\boldsymbol{Y}} - \boldsymbol{\Phi} \underline{\boldsymbol{X}}\right\|_{\mathrm{F}}^2\right\} + d/T}, \tag{8.20}$$

where $E\left\{\left\|\underline{\boldsymbol{X}}^n\right\|_2^2\right\} = \left\|\underline{\boldsymbol{\mathcal{U}}}^n\right\|_2^2 + \Sigma_{nn}$, $E\left\{\left\|\underline{\boldsymbol{Y}} - \boldsymbol{\Phi}\underline{\boldsymbol{X}}\right\|_{\mathrm{F}}^2\right\} = \left\|\underline{\boldsymbol{Y}} - \boldsymbol{\Phi}\underline{\boldsymbol{\mathcal{U}}}\right\|_{\mathrm{F}}^2 + \alpha_0^{-1} \sum_{n=1}^N \gamma_n$ with $\gamma_n = 1 - \alpha_n^{-1} \Sigma_{nn}$.

For $\boldsymbol{\beta}$, its estimate maximizes $E\left\{\log p\left(\boldsymbol{Y} | \boldsymbol{X}, \alpha_0, \boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}\right)\right\}$ by (8.15) and thus minimizes

$$\begin{aligned}
&E\left\{\frac{1}{T}\sum_{t=1}^{T}\left\|\boldsymbol{y}(t)-\left(\boldsymbol{A}+\boldsymbol{B}\mathrm{diag}\left(\boldsymbol{\beta}\right)\right)\boldsymbol{x}(t)\right\|_2^2\right\}\\
&=\frac{1}{T}\sum_{t=1}^{T}\left\|\boldsymbol{y}(t)-\left(\boldsymbol{A}+\boldsymbol{B}\mathrm{diag}\left(\boldsymbol{\beta}\right)\right)\boldsymbol{\mu}(t)\right\|_2^2\\
&\quad+Tr\left\{\left(\boldsymbol{A}+\boldsymbol{B}\mathrm{diag}\left(\boldsymbol{\beta}\right)\right)\boldsymbol{\Sigma}\left(\boldsymbol{A}+\boldsymbol{B}\mathrm{diag}\left(\boldsymbol{\beta}\right)\right)^H\right\}\\
&=\boldsymbol{\beta}^T\boldsymbol{P}\boldsymbol{\beta}-2\boldsymbol{v}^T\boldsymbol{\beta}+C
\end{aligned} \tag{8.21}$$

where C is a constant term independent of $\boldsymbol{\beta}$, $\boldsymbol{P}$ is a positive semi-definite matrix and

$$\boldsymbol{P}=\Re\left\{\overline{\boldsymbol{B}^H\boldsymbol{B}}\odot\left(\underline{\boldsymbol{\mathcal{U}}}\cdot\underline{\boldsymbol{\mathcal{U}}}^H+\boldsymbol{\Sigma}\right)\right\}, \tag{8.22}$$

$$\boldsymbol{v}=\Re\left\{\frac{1}{T}\sum_{t=1}^{T}\mathrm{diag}\left(\overline{\boldsymbol{\mu}(t)}\right)\boldsymbol{B}^H\left[\boldsymbol{y}(t)-\boldsymbol{A}\boldsymbol{\mu}(t)\right]-\mathrm{diag}\left(\boldsymbol{B}^H\boldsymbol{A}\boldsymbol{\Sigma}\right)\right\}. \tag{8.23}$$

Denote $\boldsymbol{\Delta}=\mathrm{diag}\left(\boldsymbol{\beta}\right)$. The last equality of (8.21) is based on the following two equalities:

$$\begin{aligned}
&\left\|\boldsymbol{y}-\left(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{\Delta}\right)\boldsymbol{\mu}\right\|_2^2\\
&=\left\|\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}\right)-\boldsymbol{B}\cdot\mathrm{diag}\left(\boldsymbol{\mu}\right)\cdot\boldsymbol{\beta}\right\|_2^2\\
&=\boldsymbol{\beta}^T\mathrm{diag}^H\left(\boldsymbol{\mu}\right)\boldsymbol{B}^H\boldsymbol{B}\cdot\mathrm{diag}\left(\boldsymbol{\mu}\right)\boldsymbol{\beta}\\
&\quad-2\Re\left\{\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}\right)^H\boldsymbol{B}\cdot\mathrm{diag}\left(\boldsymbol{\mu}\right)\cdot\boldsymbol{\beta}\right\}+C_1\\
&=\boldsymbol{\beta}^T\left(\overline{\boldsymbol{B}^H\boldsymbol{B}}\odot\boldsymbol{\mu}\boldsymbol{\mu}^H\right)\boldsymbol{\beta}\\
&\quad-2\Re\left\{\mathrm{diag}\left(\overline{\boldsymbol{\mu}}\right)\boldsymbol{B}^H\left(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}\right)\right\}^T\boldsymbol{\beta}+C_1,\\
&Tr\left\{\left(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{\Delta}\right)\boldsymbol{\Sigma}\left(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{\Delta}\right)^H\right\}\\
&=2Tr\left\{\Re\left\{\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{\Delta}\boldsymbol{B}^H\right\}\right\}+Tr\left\{\boldsymbol{B}\boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta}\boldsymbol{B}^H\right\}+C_2\\
&=2\Re\left\{Tr\left\{\boldsymbol{B}^H\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{\Delta}\right\}\right\}+Tr\left\{\boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta}\boldsymbol{B}^H\boldsymbol{B}\right\}+C_2\\
&=2\Re\left\{\mathrm{diag}\left(\boldsymbol{B}^H\boldsymbol{A}\boldsymbol{\Sigma}\right)\right\}^T\boldsymbol{\beta}+\boldsymbol{\beta}^T\left(\boldsymbol{\Sigma}\odot\overline{\boldsymbol{B}^H\boldsymbol{B}}\right)\boldsymbol{\beta}+C_2,
\end{aligned}$$

where C_1, C_2 are constants independent of $\boldsymbol{\beta}$, and the equality

$$Tr\left\{\operatorname{diag}^H(\boldsymbol{u})\,\boldsymbol{Q}\cdot\operatorname{diag}(\boldsymbol{v})\cdot\boldsymbol{R}^T\right\}=\boldsymbol{u}^H(\boldsymbol{Q}\odot\boldsymbol{R})\,\boldsymbol{v}$$

for vectors $\boldsymbol{u},\boldsymbol{v}$ and matrices $\boldsymbol{Q},\boldsymbol{R}$ of proper dimensions is used. Note that $\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}\in\mathbb{R}$ for a positive semi-definite matrix $\boldsymbol{S}$ of proper dimension and thus $\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}=\Re\left\{\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}\right\}=\boldsymbol{\beta}^T\cdot\Re\boldsymbol{S}\cdot\boldsymbol{\beta}$ since $\boldsymbol{\beta}$ is real-valued. Then (8.21) is obtained by observing that both $\overline{\boldsymbol{B}^H\boldsymbol{B}}\odot\boldsymbol{\mu}\boldsymbol{\mu}^H$ and $\boldsymbol{\Sigma}\odot\overline{\boldsymbol{B}^H\boldsymbol{B}}$ are positive semi-definite.

As a result, we have

$$\boldsymbol{\beta}^{new}=\arg\min_{\boldsymbol{\beta}\in\left[-\frac{1}{2}r,\frac{1}{2}r\right]^N}\left\{\boldsymbol{\beta}^T\boldsymbol{P}\boldsymbol{\beta}-2\boldsymbol{v}^T\boldsymbol{\beta}\right\}. \tag{8.24}$$

Remark 8.2. *Though an explicit expression of $\boldsymbol{\beta}^{new}$ cannot be given, by recognizing that $\boldsymbol{\beta}$ is jointly sparse with $\boldsymbol{x}$, the dimension of $\boldsymbol{\beta}$ can be reduced to K in the computation and hence $\boldsymbol{\beta}^{new}$ can be efficiently calculated. We provide the details in Subsection 8.3.5.*

The proposed off-grid SBL algorithm, abbreviated as OGSBL, is implemented as follows. After initializations of the hyperparameters $\boldsymbol{\alpha}$, α_0 and $\boldsymbol{\beta}$, we calculate $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}(t)$, $t=1,\cdots,T$, using the current values of the hyperparameters according to (8.18) and (8.17) respectively. Then we update $\boldsymbol{\alpha}$, α_0 and $\boldsymbol{\beta}$ according to (8.19), (8.20) and (8.24) respectively. The process is repeated until some convergence criterion is satisfied. OGSBL has guaranteed convergence since the function $p\left(\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{Y}\right)$ is guaranteed to increase at each iteration by the property of EM algorithm [126].

8.3.3 OGSBL-SVD

In this subsection we recall a subspace-based idea in [88] that uses the SVD of the measurement matrix $\boldsymbol{Y}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^H$ to reduce the computation of the signal reconstruction process and the sensitivity to the measurement noise. Then we incorporate it into our OGSBL algorithm. Consider the noise-free case where $\boldsymbol{Y} = \boldsymbol{\Phi X}$ with $K \leq T$. We have $\text{Rank}\,(\boldsymbol{Y}) \leq \text{Rank}\,(\boldsymbol{X}) \leq K$. Let $\boldsymbol{V} = [\boldsymbol{V}_1\ \boldsymbol{V}_2]$, where $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are matrices that consist of the first K and the rest $T-K$ columns of $\boldsymbol{V}$ respectively. Then we have that $\boldsymbol{Y}_{SV} = \boldsymbol{Y}\boldsymbol{V}_1 \in \mathbb{C}^{M\times K}$ preserves all signal information. In a general case where noises exist, by the SVD we have $\boldsymbol{Y}\boldsymbol{V} = [\boldsymbol{Y}_{SV}\ \boldsymbol{Y}\boldsymbol{V}_2]$, where the first part $\boldsymbol{Y}_{SV}$ preserves most signal information and is to be used in the following signal recovery process while the second part is abandoned. Denote $\boldsymbol{X}_{SV} = \boldsymbol{X}\boldsymbol{V}_1$ and $\boldsymbol{E}_{SV} = \boldsymbol{E}\boldsymbol{V}_1$. Then we have

$$\boldsymbol{Y}_{SV} = \boldsymbol{\Phi X}_{SV} + \boldsymbol{E}_{SV}. \tag{8.25}$$

In (8.25), $\boldsymbol{Y}_{SV}$, $\boldsymbol{X}_{SV}$ and $\boldsymbol{E}_{SV}$ can be viewed as the new matrices of sensor measurements, source signals and measurement noises respectively. The joint sparsity still holds in $\boldsymbol{X}_{SV}$. We do not exploit possible correlations that exist between columns of $\boldsymbol{X}_{SV}$ (and in $\boldsymbol{E}_{SV}$), i.e., we still assume that $\boldsymbol{X}_{SV}$ (and $\boldsymbol{E}_{SV}$) have independent columns.[1] It is then straightforward to apply the proposed OGSBL algorithm to estimate $\boldsymbol{X}_{SV}$, $\boldsymbol{\beta}$ and then the DOAs. We use OGSBL-SVD to refer to the resulting algorithm.

Based on implementation details to be introduced in Subsection 8.3.5, it can be shown that OGSBL-SVD has a computational complexity of order $O\left(MN^2\right)$ per iteration while that for OGSBL is $O\left(\max\left(MN^2, MNT\right)\right)$ per iteration. An additional computational workload of order $O\left(\max\left(M^2T, MT^2\right)\right)$ is for the SVD of $\boldsymbol{Y}$ in OGSBL-SVD. Since it is empirically found that OGSBL-SVD converges much faster than OGSBL, the whole computational workload of OGSBL-SVD is less than that of OGSBL in general.[2]

[1] The correlations between columns of the signal matrix ($\boldsymbol{X}_{SV}$ in our case) have recently been studied in [49].

[2] A possible exception happens in the case of $T \gg N$ where the computation for the SVD is quite heavy. A modified approach in such a case is to partition $\boldsymbol{Y}$ firstly into blocks with each of about N columns, then operate the SVD on each block and keep the resulting signal subspaces, and finally do another SVD on the new signal matrix composed of all signal subspaces. A model similar to (8.25) can be cast.

8.3.4 Source Power and DOA Estimation

We use the estimated source powers from different directions to form a spectrum of the proposed algorithm. In the following we derive a formula to estimate the source powers. We take OGSBL-SVD as an example. The case of OGSBL is similar with some modifications and thus is omitted. Let $\widehat{\boldsymbol{X}} = \boldsymbol{X}_{SV}\boldsymbol{V}_1^H$ be an estimate of the signal $\boldsymbol{X}$. Then consider $\widehat{\boldsymbol{X}}$ row by row and we have $\widehat{\boldsymbol{X}}^n \sim \mathcal{CN}\left(\widehat{\boldsymbol{\mathcal{U}}}^n\boldsymbol{V}_1^H, \widehat{\Sigma}_{nn}\boldsymbol{V}_1\boldsymbol{V}_1^H\right)$ where we use $\widehat{\boldsymbol{\mathcal{U}}}$ and $\widehat{\boldsymbol{\Sigma}}$ to denote the final estimates of the mean and covariance of $\boldsymbol{X}_{SV}$ respectively. We use the expectation as an estimate of the power $\wp_n$ from direction $\widetilde{\theta}_n$ (with a modification of β_n):

$$\begin{aligned}\widehat{\wp}_n &= E\left\{\wp_n\right\} = \frac{1}{T}E\left\{\left\|\widehat{\boldsymbol{X}}^n\right\|_2^2\right\}\\ &= \frac{1}{T}\left(\left\|E\left\{\widehat{\boldsymbol{X}}^n\right\}\right\|_2^2 + E\left\{\left\|\widehat{\boldsymbol{X}}^n - E\left\{\widehat{\boldsymbol{X}}^n\right\}\right\|_2^2\right\}\right)\\ &= \frac{1}{T}\left(\left\|\widehat{\boldsymbol{\mathcal{U}}}^n\boldsymbol{V}_1^H\right\|_2^2 + Tr\left\{\widehat{\Sigma}_{nn}\boldsymbol{V}_1\boldsymbol{V}_1^H\right\}\right)\\ &= \frac{\left\|\widehat{\boldsymbol{\mathcal{U}}}^n\right\|_2^2}{T} + \frac{K\widehat{\Sigma}_{nn}}{T}.\end{aligned} \tag{8.26}$$

Like other spectral-based methods, the DOAs are estimated using the locations of the highest peaks of the spectrum. Suppose that the grid indices of the highest K peaks of $\widehat{\boldsymbol{\wp}}$ are $\widehat{n}_k$, $k = 1, \cdots, K$. The estimated K DOAs will be $\widehat{\theta}_k = \widetilde{\theta}_{\widehat{n}_k} + \widehat{\beta}_{\widehat{n}_k}$, $k = 1, \cdots, K$.

8.3.5 Implementation Details

This subsection presents some details of our implementations of OGSBL and OGSBL-SVD. At each iteration of OGSBL or OGSBL-SVD, an $N \times N$ matrix inversion is required when updating $\boldsymbol{\Sigma}$ according to (8.18). By $M < N$ the Woodbury matrix identity is applied to give $\boldsymbol{\Sigma} = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{\Phi}^H\boldsymbol{C}^{-1}\boldsymbol{\Phi}\boldsymbol{\Lambda}$ with $\boldsymbol{C} = \alpha_0^{-1}\boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^H \in \mathbb{C}^{M\times M}$. By the fact that $\boldsymbol{\beta}$ is jointly sparse with $\boldsymbol{x}(t)$ whose K nonzero entries correspond to the locations of the K sources, we calculate only entries of $\boldsymbol{\beta}$ that correspond to locations of the maximum K entries of $\boldsymbol{\alpha}$ and set others to zero. As a result, $\boldsymbol{\beta}$, $\boldsymbol{P}$ and $\boldsymbol{v}$ can be truncated into dimension of K or $K \times K$. We still use $\boldsymbol{\beta}$, $\boldsymbol{P}$ and $\boldsymbol{v}$ hereafter to denote their truncated versions for simplicity. By (8.24) and $\frac{\partial}{\partial \beta}\left\{\boldsymbol{\beta}^T \boldsymbol{P} \boldsymbol{\beta} - 2\boldsymbol{v}^T\boldsymbol{\beta}\right\} = 2\left(\boldsymbol{P}\boldsymbol{\beta} - \boldsymbol{v}\right)$ we have $\boldsymbol{\beta}^{new} = \check{\boldsymbol{\beta}}$ if $\boldsymbol{P}$ is invertible and $\check{\boldsymbol{\beta}} = \boldsymbol{P}^{-1}\boldsymbol{v} \in \left[-\frac{1}{2}r, \frac{1}{2}r\right]^K$. Otherwise, we update $\boldsymbol{\beta}$ elementwise, i.e., at each step we update one β_n by fixing up the other entries of $\boldsymbol{\beta}$. For $n = 1, \cdots, K$, first we let

$$\check{\beta}_n = \frac{v_n - \left(\boldsymbol{P}_n\right)_{-n}^T \boldsymbol{\beta}_{-n}}{P_{nn}}, \tag{8.27}$$

where $\boldsymbol{u}_{-n}$ is $\boldsymbol{u}$ without the nth entry for a vector $\boldsymbol{u}$. Then by constraining $\beta_n \in \left[-\frac{1}{2}r, \frac{1}{2}r\right]$ we have

$$\beta_n^{new} = \begin{cases} \check{\beta}_n, & \text{if } \check{\beta}_n \in \left[-\frac{1}{2}r, \frac{1}{2}r\right]; \\ -\frac{1}{2}r, & \text{if } \check{\beta}_n < -\frac{1}{2}r; \\ \frac{1}{2}r, & \text{otherwise.} \end{cases} \tag{8.28}$$

It is easy to show that the objective function is guaranteed to decrease at each step with β_n defined in (8.28).

We terminate OGSBL and OGSBL-SVD if $\frac{\left\|\boldsymbol{\alpha}^{i+1} - \boldsymbol{\alpha}^i\right\|_2}{\left\|\boldsymbol{\alpha}^i\right\|_2} < \tau$ or the maximum number of iterations is reached, where τ is a user-defined tolerance and the superscript i refers to the iteration.

8.3.6 Numerical Simulations

In this section, we present our numerical results for the DOA estimation. A standard ULA of $M = 8$ sensors is considered. The origin is set at the middle point of the ULA to reduce the approximation error in (8.2). So, $A_{mn} = \exp\left\{j\pi\left(m - \frac{M+1}{2}\right)\cos\widetilde{\theta}_n\right\}$ and $B_{mn} = -j\pi\left(m - \frac{M+1}{2}\right)\sin\widetilde{\theta}_n \cdot A_{mn}$, $m = 1, \cdots, M$, $n = 1, \cdots, N$, with $j = \sqrt{-1}$. A uniform sampling grid $\{0^\circ, r, 2r, \cdots, 180^\circ - r\}$ is considered with r being the grid interval. The number of snapshots is set to $T = 200$ in the case of MMV. We consider only OGSBL-SVD in the MMV case since it is empirically observed to converge faster and be more accurate in comparison with OGSBL. In OGSBL-SVD, we set $\rho = 0.01$ and $c = d = 1 \times 10^{-4}$. We initialize $\alpha_0 = \frac{100K}{\sum_{t=1}^{K} Var\left\{(\boldsymbol{Y}_{SV})_t\right\}}$, $\boldsymbol{\alpha} = \frac{1}{MK} \sum_{t=1}^{K} \left| \boldsymbol{A}^H (\boldsymbol{Y}_{SV})_t \right|$ and $\boldsymbol{\beta} = \mathbf{0}$, where $|\cdot|$ applies elementwise. We set $\tau = 10^{-3}$ and the maximum number of iterations to 1000. We note that the proposed algorithm is insensitive to the initializations of α_0, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, as well as to ρ if ρ is not too large. As reported in [41], the estimate of α_0 can be inaccurate in some cases. But we have observed minimal effects on the result of DOA estimation. All experiments are carried out in Matlab v.7.7.0 on a PC with a Windows XP system and a 3GHz CPU.

8.3.6.1 Comparison with ℓ_1-SVD

We take ℓ_1-SVD in [88] as a representative of on-grid model based methods and compare OGSBL-SVD with it in terms of estimation bias, mean squared error (MSE) and computational time.

We first study the case where the true DOAs are exactly on the grid, i.e., the standard CS formulation describes accurately the true observation model in (8.1). Apart from OGSBL-SVD and ℓ_1-SVD, we also consider the on-grid form of OGSBL-SVD, denoted by SBL-SVD. SBL-SVD is based on the on-grid model and hence is a special case of OGSBL-SVD by imposing $\beta = \mathbf{0}$ during the iterations. A Monte Carlo method is used to estimate the bias of the DOA estimation. Two sources are considered with one being held fixed at direction 50°. The other source is from 54°, 55° and so on. For each pair of DOAs, the bias is calculated as the average of 100 trials. The grid interval is set to $r = 1°$ in all algorithms. Fig. 8.3 presents the biases of ℓ_1-SVD, SBL-SVD and OGSBL-SVD. Asymptotic biases exhibit in all three methods. When the two DOAs are close to each other, ℓ_1-SVD has the largest bias while those of SBL-SVD and OGSBL-SVD are relatively small. Compared to OGSBL-SVD, SBL-SVD utilizes the additional knowledge of on-grid DOAs and has a smaller bias.

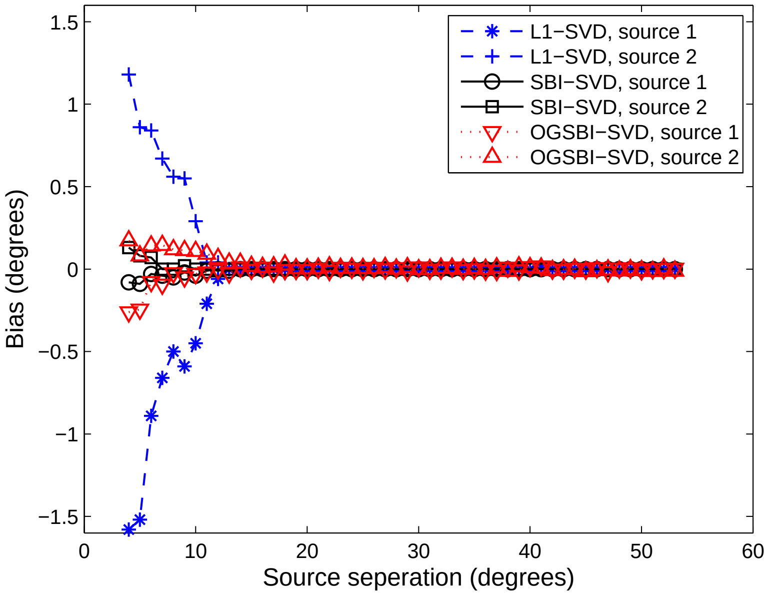
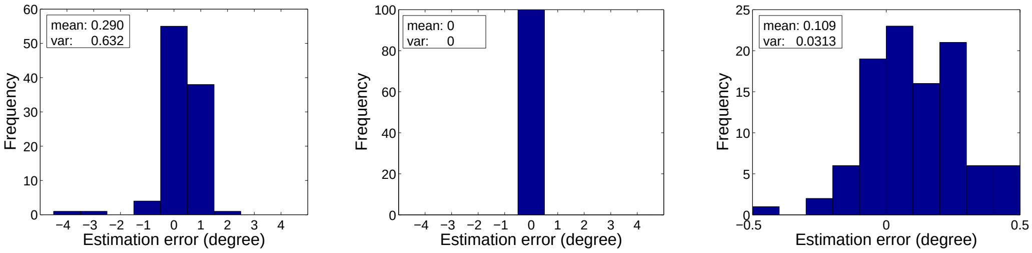


Figure 8.3: Bias comparison of ℓ_1-SVD, SBL-SVD and OGSBL-SVD in localizing two on-grid sources as a function of angular separation with source 1 fixed at 50° and SNR = 10 dB. Significant bias exhibits when using ℓ_1-SVD for closely spaced sources while small asymptotic biases occur in SBL-SVD and OGSBL-SVD.

To provide a better illustration of the DOA estimation errors, we present in Fig. 8.4 the histograms of the estimation errors of source 2 when it comes from 60°. For ℓ_1-SVD, the errors range from $-4°$ to $2°$ though it estimates the DOA accurately in over half of the 100 trials. A remarkable result is that SBL-SVD finds the DOA without error in all trials. All errors of OGSBL-SVD are within half the grid interval while it does not exploit the information that the true DOAs are on the sampling grid. The advantages of SBL-SVD and OGSBL-SVD over ℓ_1-SVD can also be recognized from the means and variances of the estimation errors shown in Fig. 8.4.

We then consider a general case of arbitrarily located sources. Similarly, we consider two sources with one being held fixed at direction 50.3°. Fifty positions of the other source are considered, which are respectively generated uniformly from intervals $[54°, 55°]$, $[55°, 56°]$ and so on. For each pair of DOAs, the bias is calculated as the average of 100 trials. We consider OGSBL-SVD and ℓ_1-SVD with grid interval $r = 1°$. Fig. 8.5 presents the bias for each of the DOA estimations as a function



(a) ℓ_1-SVD (b) SBL-SVD (c) OGSBL-SVD

Figure 8.4: Histograms of the DOA estimation errors of (a) ℓ_1-SVD, (b) SBL-SVD and (c) OGSBL-SVD for source 2. Source 1: $50°$; source 2: $60°$. SNR $=$ 10 dB. Grid interval $r = 1°$. The errors of ℓ_1-SVD range from $-4°$ to $2°$ while SBL-SVD finds the DOAs without error in all the trials. All errors of OGSBL-SVD are within half the grid interval without exploiting the information that the true DOAs are on the sampling grid.

of the angular separation with SNR $=$ 10dB. A significant bias occurs all the time for both sources when using ℓ_1-SVD since the on-grid model cannot describe the true observation model accurately. As a result, the best DOA estimate that ℓ_1-SVD can obtain is the grid point nearest to the true DOA. It can be seen that the bias of source 1 is around $-0.3°$ when source 2 is far away from source 1 and the bias of source 2 depends on its distance to the nearest grid point. Moreover, a large bias exhibits for closely spaced sources for ℓ_1-SVD. For OGSBL-SVD, only a small asymptotic bias exists similar to the on-grid case due to the joint estimate of the grid offset.

Next, we compare OGSBL-SVD with ℓ_1-SVD in terms of mean squared error (MSE) and computational time with respect to the grid interval r and SNR. In our experiment, we consider SNR $=$ 10 and 0dB, and $r = 0.5°, 1°, 2°$ and $4°$. In each trial, $K = 2$ sources θ_1, θ_2 are uniformly generated within intervals $[58°, 62°]$ and $[86°, 90°]$ respectively. For each combination (SNR, r), the MSE is averaged over $R = 200$ trials:

$$\text{MSE} = \frac{1}{RK} \sum_{i=1}^{R} \sum_{k=1}^{K} \left(\theta_k^i - \widehat{\theta}_k^i \right)^2, \tag{8.29}$$

where the superscript i refers to the ith trial. According to the results in the last section, there exists a lower bound for the MSE of ℓ_1-SVD regardless of the SNR since

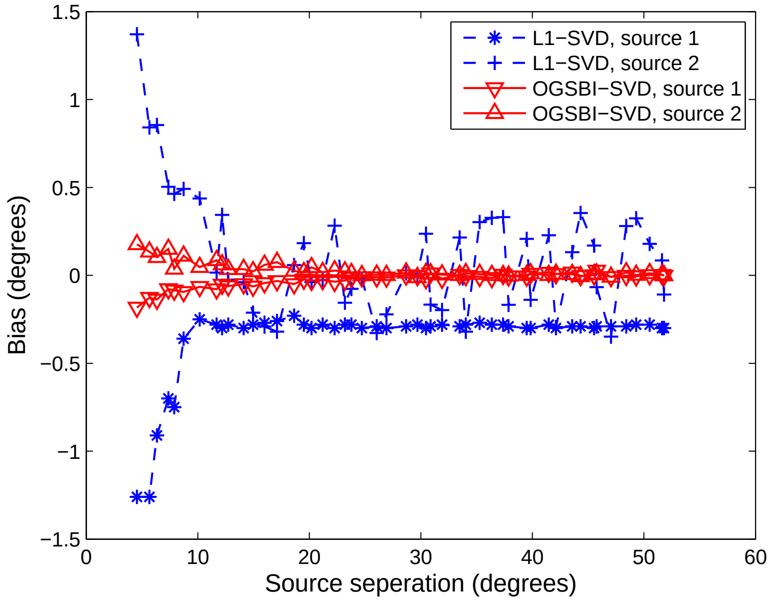


Figure 8.5: Bias comparison of OGSBL-SVD and ℓ_1-SVD in localizing two sources as a function of angular separation with source 1 fixed at 50.3° and SNR = 10 dB. A small asymptotic bias occurs in OGSBL-SVD while a significant bias exhibits all the time when using ℓ_1-SVD.

the best DOA estimate that ℓ_1-SVD can obtain is the grid point nearest to the true DOA. By assuming that the true DOA is uniformly distributed, the lower bound is LB $= r^2/12$. Fig. 8.6 presents our experimental results. In all scenarios under consideration, OGSBL-SVD is more accurate than ℓ_1-SVD. Moreover, OGSBL-SVD can exceed the lower bound of ℓ_1-SVD in most scenarios. The phenomenon is significant in the high SNR regime or in the coarse sampling grid case where the on-grid model is poor in describing the true observation model while the modeling error can be overcome to a large extent by the off-grid model used in this chapter.

Table 8.1 presents the averaged CPU times of OGSBL-SVD and ℓ_1-SVD (excluding the SVD process that takes about 0.003s in our case) with respect to SNR and r.[3] For both OGSBL-SVD and ℓ_1-SVD, their CPU times decrease as the grid gets coarser. OGSBL-SVD is faster than ℓ_1-SVD at $r = 2°$ and 4°. One drawback of the proposed method is that it is slow in the case of a dense sampling grid. In practice,

[3] The code of ℓ_1-SVD is provided by the author of [88]. We note that its speed can be accelerated using state-of-the-art algorithms for CS.

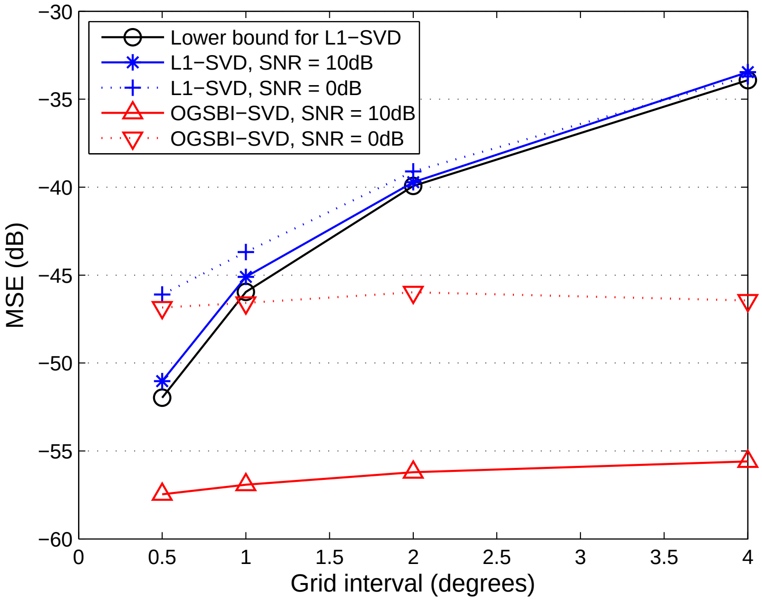


Figure 8.6: MSEs of OGSBL-SVD and ℓ_1-SVD. The lower bound is for ℓ_1-SVD regardless of the SNR.

we recommend to use a coarser grid with $r = 2°$ for the proposed algorithm since it can give an accurate yet fast DOA estimation.

8.3.6.2 Comparison with STLS

The off-grid model has recently been used in [78] for the DOA estimation. In [78], a sparse total least-squares (STLS) approach is proposed. In the SMV case, STLS

Table 8.1: Averaged Time Consumptions of ℓ_1-SVD and OGSBL-SVD with Respect to SNR and r. Time unit: sec.

SNR = 10dB				
	$r = 0.5°$	$r = 1°$	$r = 2°$	$r = 4°$
ℓ_1-SVD	0.601	0.413	0.324	0.291
OGSBL-SVD	10.2	0.782	0.096	0.025
SNR = 0dB				
	$r = 0.5°$	$r = 1°$	$r = 2°$	$r = 4°$
ℓ_1-SVD	0.413	0.295	0.218	0.190
OGSBL-SVD	10.9	0.831	0.104	0.024

seeks to solve the nonconvex optimization problem

$$\min_{\boldsymbol{x},\boldsymbol{\beta}} \left\{ \|\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y} - (\boldsymbol{A} + \boldsymbol{B}\text{diag}(\boldsymbol{\beta}))\, \boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1 \right\}, \tag{8.30}$$

where $\boldsymbol{x}$ is the sparse source signal of interest, $\boldsymbol{y}$ is the noisy measurement, $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\beta}$ are the same as defined in the off-grid model, and $\lambda > 0$ is a regularization parameter. From the Bayesian perspective, this is equivalent to seeking for an MAP solution of $(\boldsymbol{x}, \boldsymbol{\beta})$ by assuming that the measurement noise is white Gaussian, $\boldsymbol{x}$ is Laplacian and $\boldsymbol{\beta}$ is Gaussian. It is noted that the last assumption for $\boldsymbol{\beta}$ cannot properly capture the property of $\boldsymbol{\beta}$. A local minima of the problem in (8.30) is achieved in [78] by an alternating approach, i.e., alternatively solving $\boldsymbol{x}$ with a fixed $\boldsymbol{\beta}$, which requires a solution to an ℓ_1-regularized least square problem, and solving $\boldsymbol{\beta}$ with a fixed $\boldsymbol{x}$, which requires a solution to an N dimensional linear system. As argued in [88] the SVD used in OGSBL-SVD can alleviate the sensitivity to the measurement noise in the MMV case that is not used in [78]. To make a fair comparison, we consider only the SMV case when comparing our method with STLS though a similar problem can be cast for STLS in the MMV case. In our implementation of OGSBL, we initialize $\alpha_0 = \frac{100}{Var\{\boldsymbol{y}\}}$ and $\boldsymbol{\alpha} = \frac{1}{MK} \left|\boldsymbol{A}^H \boldsymbol{y}\right|$. The rest settings are the same as those for OGSBL-SVD in the MMV case.

In our experiment, we consider two DOAs from 63.2° and 90.3° with SNR $= 20$dB. We consider $r = 2^\circ$ and 4° for both OGSBL and STLS. The parameter λ in (8.30) is tuned to our best such that STLS achieves the smallest error. Table 8.2 presents the averaged MSEs and CPU times of STLS and OGSBL over $R = 200$ trials. OGSBL obtains more accurate DOA estimations than STLS in both the scenarios with remarkably less computational times. We also note that it is possible to accelerate STLS using state-of-the-art algorithms for CS.

Table 8.2: Averaged MSEs and CPU Times of STLS and OGSBL in the SMV Case with Respect to r.

	MSE (dB)		Time (sec)	
	$r = 2°$	$r = 4°$	$r = 2°$	$r = 4°$
STLS	-36.5	-36.6	5.31	1.77
OGSBL	-45.1	-43.3	0.098	0.028

8.4 Conclusion

In this chapter, we studied the off-grid DOA estimation model firstly proposed in [78] for reducing the modeling error due to discretization of a continuous range. An ℓ_1 optimization-based approach was firstly presented to solve the problem under the framework of perturbed CS we have studied in Chapter 5. To alleviate the computational burden, an SBL algorithm was then proposed in which the sparse signal and the grid offset are jointly estimated iteratively. Numerical simulations were provided to validate our approaches.

Chapter 9

Sparse MRI for Motion Correction

The application of CS to the DOA estimation has been studied in the last chapter to resolve the basis mismatch problem that exists in the literature. In this chapter, we propose a novel application of CS to motion correction in magnetic resonance imaging (MRI) while CS has been widely applied to MRI for imaging acceleration. Imaging artifacts due to patient motion remain a major challenge in many MRI applications. In this chapter, we consider MR image reconstruction from k-space data corrupted by 2D translational motion. The problem is ill-posed since there exist infinitely many image candidates and possible motions which result in the same k-space data. Existing methods solve the problem by acquiring more k-space data and/or adopting a specific sampling sequence, see [95–99].

In this chapter, we present a sparsity-based approach to rigid-body motion correction for the first time. The presented approach relies on the assumption that little motion occurs during a single readout line, which can be approximately satisfied in practice and is greatly relaxed in comparison with those in [97, 99]. We model the motion occurring during the scanning process as unknown parameters in the sensing system and seek after the motion estimate such that the compensated MR image is maximally sparse/compressible among the infinite candidates. The idea is inspired by our results in Chapter 5 where it is shown that uncertain system parameters can be accurately estimated along with the sparse signal in one particular situation. We present efficient iterative algorithms to jointly estimate the motion and the image content. The proposed method has a lot of merits, such as no need of additional data and loose requirement for the sampling sequence. Promising results are presented to demonstrate its performance. This chapter is mainly based on [12].

9.1 Problem Formulation of Motion Correction

Consider that a translational motion $\boldsymbol{\beta}_{\boldsymbol{k}}$ occurs when acquiring the k-space measurement at $\boldsymbol{k}$. So it holds at the moment $\overline{m}\left(\boldsymbol{r}\right) = m^o\left(\boldsymbol{r} - \boldsymbol{\beta}_{\boldsymbol{k}}\right)$, where $\boldsymbol{m}^o$ is the MR image of interest, $\overline{\boldsymbol{m}}$ is the translated image and $\boldsymbol{r}$ denotes the coordinate in the image domain. According to the relationship between an MR image and its k-space measurements, it holds

$$\begin{aligned}\overline{M}\left(\boldsymbol{k}\right) &= \iint \overline{m}\left(\boldsymbol{r}\right) e^{-i2\pi\langle \boldsymbol{k},\, \boldsymbol{r}\rangle} d\boldsymbol{r}\\ &= e^{-i2\pi\langle \boldsymbol{k},\, \boldsymbol{\beta}_{\boldsymbol{k}}\rangle} M^o\left(\boldsymbol{k}\right),\end{aligned} \tag{9.1}$$

where $\boldsymbol{M}^o = \mathcal{F}\boldsymbol{m}^o$ and $\overline{\boldsymbol{M}}$ denote the motion-free and motion-corrupted k-space data, respectively, with $\mathcal{F}$ denoting the 2D Fourier transform operator. It is shown in (9.1) that the translational motion leads to a phase error and unaltered amplitude. To represent (9.1) more compactly, let $\boldsymbol{\beta} = \left[\cdots, \boldsymbol{\beta}_{\boldsymbol{k}_j}, \cdots\right]^T$ and then write (9.1) into

$$\overline{\boldsymbol{M}} = \mathcal{T}_{\boldsymbol{\beta}}\mathcal{F}\boldsymbol{m}^o, \tag{9.2}$$

where $\mathcal{T}_{\boldsymbol{\beta}}$ denotes a linear operator caused by the translational motion $\boldsymbol{\beta}$ and is referred to as the translational operator hereafter. In particular, let $\boldsymbol{\Lambda}_{\boldsymbol{\beta}}$ be a matrix of the same dimension as $\overline{\boldsymbol{M}}$ with its element $\Lambda_{\boldsymbol{\beta}}\left(\boldsymbol{k}\right) = e^{-i2\pi\langle \boldsymbol{k},\, \boldsymbol{\beta}_{\boldsymbol{k}}\rangle}$. Then $\mathcal{T}_{\boldsymbol{\beta}}\boldsymbol{M}^o = \boldsymbol{\Lambda}_{\boldsymbol{\beta}} \odot \boldsymbol{M}^o$ where $\odot$ denotes the Hadamard product. In addition, it holds that $\mathcal{T}_{\boldsymbol{\beta}}^{-1} = \mathcal{T}_{-\boldsymbol{\beta}}$ and $\mathcal{T}_{\boldsymbol{0}} = \mathcal{I}$ where $\mathcal{I}$ denotes the identity operator. For a particular motion $\boldsymbol{\beta}$, the linear system of equations in (9.2) relates the MR image $\boldsymbol{m}^o$ of interest and the acquired k-space data $\overline{\boldsymbol{M}}$. Our objective is to reconstruct the image $\boldsymbol{m}^o$ and possibly the motion $\boldsymbol{\beta}$ given the corrupted k-space data $\overline{\boldsymbol{M}}$.

The challenge is that the problem of recovering $\boldsymbol{m}^o$ from $\overline{\boldsymbol{M}}$ is ill-posed since there exist infinite number of solutions to (9.2). In particular, for any $\boldsymbol{\beta}$ there exists a solution $\boldsymbol{m} = \mathcal{F}^{-1}\mathcal{T}_{-\boldsymbol{\beta}}\overline{\boldsymbol{M}}$ such that (9.2) holds. In this regard, the full sampled k-space data can be considered as "compressive" measurements in the language of CS. To choose the correct one among the infinite candidates, as a result, additional information has to be exploited.

9.2 Sparsity-Driven Motion Correction

9.2.1 Sparsity-Based Formulation

We first consider that the MR image is directly constructed from the corrupted k-space data $\overline{\boldsymbol{M}}$ without considering the translational motion. Then, the constructed image is

$$\widehat{\boldsymbol{m}} = \mathcal{F}^{-1}\overline{\boldsymbol{M}} = \mathcal{F}^{-1}\left(\boldsymbol{\Lambda}_{\boldsymbol{\beta}} \odot \boldsymbol{M}^o\right) = \left(\mathcal{F}^{-1}\boldsymbol{\Lambda}_{\boldsymbol{\beta}}\right) \otimes \boldsymbol{m}^o, \tag{9.3}$$

where $\otimes$ denotes the circular convolution operation. So the obtained image is the true image $\boldsymbol{m}^o$ after a convolution with $\mathcal{F}^{-1}\boldsymbol{\Lambda}_{\boldsymbol{\beta}}$. That explains how the imaging artifacts come from the translational motion. Due to the imaging artifacts, it is natural to conjecture that the translational motion will reduce the sparsity/compressibility of MR images, i.e., the true image is the maximally sparse/compressible solution (under an appropriate basis) to (9.2). Examples of a Shepp-Logan phantom and simulated human brain are presented in Fig. 9.1. In comparison with the motion-free images (in col 1), severe artifacts are present in the motion-corrupted ones (in col 2). Numerically, it can be shown that the ℓ_1 norms (a commonly used sparsity metric) of the motion-corrupted images (under an Haar wavelet basis) are larger than those of the motion-free ones.

Based on the conjecture above, we propose to reconstruct the motion-free MR image using the maximally sparse solution that satisfies the data consistency constraint

(9.2), i.e., by solving a BP like optimization problem:

$$\min_{\boldsymbol{m},\boldsymbol{\beta}} \left\|\mathcal{W}\boldsymbol{m}\right\|_1, \text{ subject to } \mathcal{T}_{\boldsymbol{\beta}}\mathcal{F}\boldsymbol{m} = \overline{\boldsymbol{M}} \text{ and } \boldsymbol{\beta} \in \mathcal{D}_{\boldsymbol{\beta}}, \tag{9.4}$$

where $\mathcal{W}$ is a sparsifying operator (e.g., a wavelet transform), $\|\cdot\|_1$ denotes the ℓ_1 norm (sum of amplitude of all elements) that is commonly used to promote sparsity and $\mathcal{D}_{\boldsymbol{\beta}}$ denotes the domain of $\boldsymbol{\beta}$. Given an appropriate constant C, an equivalent (LASSO-like) formulation of (9.4) is

$$\min_{\boldsymbol{m},\boldsymbol{\beta}} \left\|\overline{\boldsymbol{M}} - \mathcal{T}_{\boldsymbol{\beta}}\mathcal{F}\boldsymbol{m}\right\|_{\mathrm{F}}, \text{ subject to } \left\|\mathcal{W}\boldsymbol{m}\right\|_1 \leq C \text{ and } \boldsymbol{\beta} \in \mathcal{D}_{\boldsymbol{\beta}}, \tag{9.5}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm and $C < \left\|\mathcal{W}\mathcal{F}^{-1}\overline{\boldsymbol{M}}\right\|_1$ is a constant that needs to be tuned in practice and can be set to $C = \left\|\mathcal{W}\boldsymbol{m}^o\right\|_1$ in an ideal case. The motions $\boldsymbol{\beta}_{\boldsymbol{k}}$ with respect to the coordinate $\boldsymbol{k}$ are not independent. For example, the motion should be piecewise smooth when sorted chronologically. In this chapter, we assume that the same motion is shared among every readout line, which can be approximately satisfied in practice and greatly reduces the number of unknown parameters. Note that the assumption is greatly relaxed comparing to that in [97,99] where it is assumed that little motion exists during every echotrain (each comprising several readout lines). Moreover, $\boldsymbol{\beta}$ should be properly bounded in practice. The *a priori* knowledge narrows the selection of $\boldsymbol{\beta}$ and defines $\mathcal{D}_{\boldsymbol{\beta}}$. To the best of our knowledge, (9.4) and (9.5) present the first sparsity-based formulations for the MRI motion correction problem though CS has been widely applied to MRI for imaging acceleration.

9.2.2 An Alternating Algorithm

Due to the nonconvexity with respect to $\boldsymbol{\beta}$, problem (9.5) is nonconvex. We propose an alternating algorithm that starts with $\boldsymbol{\beta}^{(0)} = \boldsymbol{0}$ and iteratively carries out the following steps:

1) solving

$$\widetilde{\boldsymbol{m}}^{(j+1)} = \arg\min_{\boldsymbol{m}} \left\| \overline{\boldsymbol{M}} - \mathcal{T}_{\boldsymbol{\beta}^{(j)}} \mathcal{F} \boldsymbol{m} \right\|_{\mathrm{F}}, \text{ subject to } \|\mathcal{W}\boldsymbol{m}\|_1 \leq C. \tag{9.6}$$

2) solving

$$\boldsymbol{\beta}^{(j+1)} = \arg\min_{\boldsymbol{\beta}} \left\| \overline{\boldsymbol{M}} - \mathcal{T}_{\boldsymbol{\beta}} \mathcal{F} \widetilde{\boldsymbol{m}}^{(j+1)} \right\|_{\mathrm{F}}, \text{ subject to } \boldsymbol{\beta} \in \mathcal{D}_{\boldsymbol{\beta}}. \tag{9.7}$$

To interpret the algorithm above, denote two sets $S_1 = \{\boldsymbol{m} : \|\mathcal{W}\boldsymbol{m}\|_1 \leq C\}$ and $S_2 = \left\{\boldsymbol{m} : \overline{\boldsymbol{M}} = \mathcal{T}_{\boldsymbol{\beta}} \mathcal{F} \boldsymbol{m}, \boldsymbol{\beta} \in \mathcal{D}_{\boldsymbol{\beta}}\right\}$. We refer to S_1 as the *sparse domain* and S_2 as the *Fourier domain*, respectively. Denote $\mathcal{P}_1$ and $\mathcal{P}_2$ projections onto S_1 and S_2, respectively. Let $\boldsymbol{m}^{(j)} = \mathcal{F}^{-1}\mathcal{T}_{-\boldsymbol{\beta}^{(j)}} \overline{\boldsymbol{M}}$. Then it holds that $\widetilde{\boldsymbol{m}}^{(j+1)} = \mathcal{P}_1 \boldsymbol{m}^{(j)}$ following from *Step 1*. At *Step 2*, the equality $\boldsymbol{m}^{(j+1)} = \mathcal{P}_2 \widetilde{\boldsymbol{m}}^{(j+1)}$ holds. So the algorithm above is equivalent to the recursion

$$\boldsymbol{m}^{(j+1)} = \mathcal{P}_2 \mathcal{P}_1 \boldsymbol{m}^{(j)}, \tag{9.8}$$

starting with $\boldsymbol{m}^{(0)} = \widehat{\boldsymbol{m}}$ where $\widehat{\boldsymbol{m}}$ is as defined in (9.3). Since $\boldsymbol{m}^{(0)} \in S_2$, the recursion attempts to find a point in S_2 nearest to S_1, or equivalently, the maximally sparse image that is consistent with the k-space observation. The constant C is generally unavailable in advance and needs to be tuned in the algorithm. A simple method is to set C such that the obtained image has the least ℓ_1 norm in the sparsifying domain (which is consistent with our objective). The remaining tasks are to compute $\mathcal{P}_1$ and $\mathcal{P}_2$.

9.2.3 Computation of $\mathcal{P}_1$

To compute $\mathcal{P}_1$ (i.e., to solve (9.6)) is a convex problem. We consider the case where the sparsifying transform $\mathcal{W}$ is a unitary operator and has a fast operation (e.g., an Haar wavelet). A fast algorithm can be developed as follows. Denote the set $S_3 = \{\boldsymbol{w} : \|\boldsymbol{w}\|_1 \leq C\}$ and $\mathcal{P}_3$ the projection onto the set. It is easy to show that

$\mathcal{P}_1\boldsymbol{m} = \mathcal{W}^{-1}\mathcal{P}_3\left(\mathcal{W}\boldsymbol{m}\right)$. The following lemma presents a fast approach to computing $\mathcal{P}_3$ which has a computational complexity of $O\left(N\log N\right)$, where N denotes the size of the variable.

Lemma 9.1. *It holds that* $\mathcal{P}_3\boldsymbol{v} = \boldsymbol{v}$ *if* $\left\|\boldsymbol{v}\right\|_1 \leq C$. *Otherwise, denote* $\boldsymbol{v}^k$ *and* $|v|_{(k)}$ *the best* k*-term approximation and the* k*th largest (in amplitude) entry of* $\boldsymbol{v}$*, respectively, for an integer* k. *Let* $k^* = \max\left\{k : \left\|\boldsymbol{v}^k\right\|_1 - k\,|v|_{(k)} < C\right\}$ *and* $\lambda = \frac{\left\|\boldsymbol{v}^{k^*}\right\|_1 - C}{k^*}$. *Then* $\mathcal{P}_3\boldsymbol{v} = S_\lambda\left(\boldsymbol{v}\right)$*, where* $S_\lambda\left(\cdot\right)$ *denotes a soft thresholding operator defined in (3.4).*

Proof. Notice that

$$\mathcal{P}_3\boldsymbol{v} = \arg\min_{\boldsymbol{x}} \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{v}\right\|_2^2, \text{ subject to } \left\|\boldsymbol{x}\right\|_1 \leq C. \tag{9.9}$$

It is trivial for the case when $\left\|\boldsymbol{v}\right\|_1 \leq C$. For the case $\left\|\boldsymbol{v}\right\|_1 > C$, it is easy to show that $\left\|\boldsymbol{x}^*\right\|_1 = C$ for the optimal solution $\boldsymbol{x}^*$. We next consider the Lagrangian function

$$f\left(\boldsymbol{x}, \lambda\right) = \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{v}\right\|_2^2 + \lambda\left(\left\|\boldsymbol{x}\right\|_1 - C\right). \tag{9.10}$$

So it holds that $\boldsymbol{x}^* = S_\lambda\left(\boldsymbol{v}\right)$ for some $\lambda > 0$. Suppose $\left\|\boldsymbol{x}^*\right\|_0 = k$, i.e., $\boldsymbol{x}^*$ has k nonzero entries. It follows from $C = \left\|\boldsymbol{x}^*\right\|_1 = \left\|\boldsymbol{v}^k\right\|_1 - k\lambda$ that $\lambda = \frac{\left\|\boldsymbol{v}^k\right\|_1 - C}{k}$. The remaining task is to determine the integer k.

By $\boldsymbol{x}^* = S_\lambda\left(\boldsymbol{v}\right)$ and $\left\|\boldsymbol{x}^*\right\|_0 = k$, we have $|v|_{(k+1)} \leq \lambda < |v|_{(k)}$. After substituting λ into $\lambda = \frac{\left\|\boldsymbol{v}^k\right\|_1 - C}{k}$, we have the inequality

$$k\,|v|_{(k+1)} \leq \left\|\boldsymbol{v}^k\right\|_1 - C < k\,|v|_{(k)}. \tag{9.11}$$

In the following we show that such k is unique. To do this, let $g\left(k\right) = \left\|\boldsymbol{v}^k\right\|_1 - k\,|v|_{(k)} - C$. It is easy to show that $g\left(k\right)$ is monotonically increasing with respect to k. Since $g\left(k+1\right) = \left\|\boldsymbol{v}^k\right\|_1 - k\,|v|_{(k+1)} - C$, the inequality (9.11) is equivalent to $g\left(k\right) < 0 \leq g\left(k+1\right)$, which by the monotonicity of $g\left(k\right)$ determines a unique k that can be expressed as in the lemma. ■

9.2.4 Approximate Computation of $\mathcal{P}_2$

The computation of $\mathcal{P}_2$ is equivalent to solving the optimization problem in (9.7). Under the assumption that a common displacement is shared among each readout line and the motions among different readout lines are independent, the number of unknown variables in $\boldsymbol{\beta}$ is reduced to twice the number of readout lines (each displacement has 2 degrees of freedom in the 2D case) and the displacement in each line can be separately calculated. However, the exact calculation of $\mathcal{P}_2$, or equivalently, to obtain the global optimum of (9.7), is generally infeasible since the Fourier domain S_2 is severely nonconvex. The $\boldsymbol{\beta}$ solution is typically trapped at a local minimum if optimization methods, e.g., a gradient method, are used to solve (9.7). Inspired by a navigator-based method in [98], we propose a practically efficient algorithm as follows.

In (9.7), the term $\mathcal{F}\widetilde{\boldsymbol{m}}^{(j)}$, denoted by $\boldsymbol{M}^{(j)}$, can be considered as the current estimate of the motion-free k-space data $\boldsymbol{M}^o$ (without accounting for the difference between $\widetilde{\boldsymbol{m}}^{(j)}$ and $\boldsymbol{m}^{(j)}$). To solve (9.7) is in fact to estimate the translational motion between the motion-corrupted observation $\overline{\boldsymbol{M}}$ and the current motion-free estimate $\boldsymbol{M}^{(j)}$. So we may consider each readout line of $\overline{\boldsymbol{M}}$ as a navigator and the associated line of $\boldsymbol{M}^{(j)}$ as the reference. Then the motion within each readout line can be estimated using the navigator-based method in [98]. We omit the details. At last, we point out that the algorithm performance can be improved in practice by a modification of the current motion-free k-space estimate: $\boldsymbol{M}^{(j)} = \left|\overline{\boldsymbol{M}}\right| \odot \operatorname{sgn}\left(\mathcal{F}\widetilde{\boldsymbol{m}}^{(j)}\right)$, instead of using $\boldsymbol{M}^{(j)} = \mathcal{F}\widetilde{\boldsymbol{m}}^{(j)}$, where the absolute operator $|\cdot|$ does an element-wise operation. The underlying reason is obvious according to (9.1). The translational motion changes only the phase information of the k-space data and keeps the amplitude. So intuitively, the modified estimate is closer to the true motion-free data and leads to better performance.

9.2.5 Sparse RAAR for Motion Correction

According to (9.8), the algorithm in Subsection 9.2.1 is implemented by iteratively projecting onto the two sets S_1 and S_2. It is related to iterative projection algorithms [142] for phase retrieval [143], which aims to find an intersection point of two sets and are built upon combining projections onto the two sets in some fashion. In fact, the recursion (9.8) is exactly the error reduction (ER) algorithm [142] without accounting for the differences of the sets. One drawback of ER is its slow convergence. In our setting, we want to solve a feasibility problem:

$$\text{find } \boldsymbol{m} \in S_1 \cap S_2, \tag{9.12}$$

i.e., to find a point lying in both the sparse domain S_1 and the Fourier domain S_2. The relaxed averaged alternating reflections (RAAR) algorithm introduced in [144] has fast convergence speed and stable performance. RAAR starts with some initial point $\boldsymbol{m}^{(0)}$ and is defined by the recursion

$$\boldsymbol{m}^{(j+1)} = \left(\frac{\theta}{2}\left(\mathcal{R}_1\mathcal{R}_2 + \mathcal{I}\right) + (1-\theta)\,\mathcal{P}_2\right)\boldsymbol{m}^{(j)}, \tag{9.13}$$

where $\theta \in [0,\ 1]$ is a constant, and reflectors $\mathcal{R}_1 = 2\mathcal{P}_1 - \mathcal{I}$ and $\mathcal{R}_2 = 2\mathcal{P}_2 - \mathcal{I}$. By defining and computing $\mathcal{P}_1$ and $\mathcal{P}_2$ as before, (9.13) defines an algorithm for the motion correction problem, named as sparse RAAR (SRAAR), where $\widehat{\boldsymbol{m}}$ in (9.3) is used as the initial point.

9.3 Numerical Simulations

The proposed method is validated using a Shepp-Logan phantom and simulated human brain data obtained from BrainWeb [1] with both of size 256×256. Continuous translational motions are randomly generated among the readout lines. Motion artifacts are introduced by applying varying linear phase shifts to the motion-free

[1] http://www.bic.mni.mcgill.ca/brainweb

k-space data according to (9.1). An Haar wavelet is selected as the sparsifying transform. SRAAR is applied to the motion-corrupted k-space data to reconstruct the MR images with the setting $\theta = 0.9$. SRAAR is terminated after a fixed number of iterations.

Simulation results are presented in Fig. 9.1. For the phantom (row 1), almost exact reconstruction is obtained using SRAAR. A small amount of motions (within 5 pixels along both the readout and phase-encode directions) are studied in rows 2 (without noise) and 3 (noise added) with the human brain. It can be seen that even a small amount of motions may cause severe imaging artifacts. After the sparsity-based motion correction with SRAAR, only few artifacts remain. In the presence of both noises and a large amount of motions (3 times as large as those in rows 2 and 3), it is shown in row 4 that SRAAR may have difficulties to produce a good result though most artifacts are removed. We note that SRAAR is computationally efficient in general. In the above, each iteration takes about 1s in Matlab v.7.7.0 on a PC with a 3GHz CPU, i.e., each reconstructed image is obtained within few minutes. Moreover, SRAAR can be greatly accelerated by estimating motions in the readout lines in parallel when computing $\mathcal{P}_2$.

9.4 Conclusion

A first sparsity-based approach to motion correction was introduced in this chapter. An efficient algorithm was proposed partially inspired by phase retrieval and existing navigator-based methods. The promising results presented in this chapter indicate good potential for practical application of the proposed method. The current work considers only translational motion and encourages further studies of more complicated motions. The results of Chapter 5 reveal that unknown system parameters can be estimated even in the undersampling case, suggesting that imaging acceleration is also possible in the presence of motions.

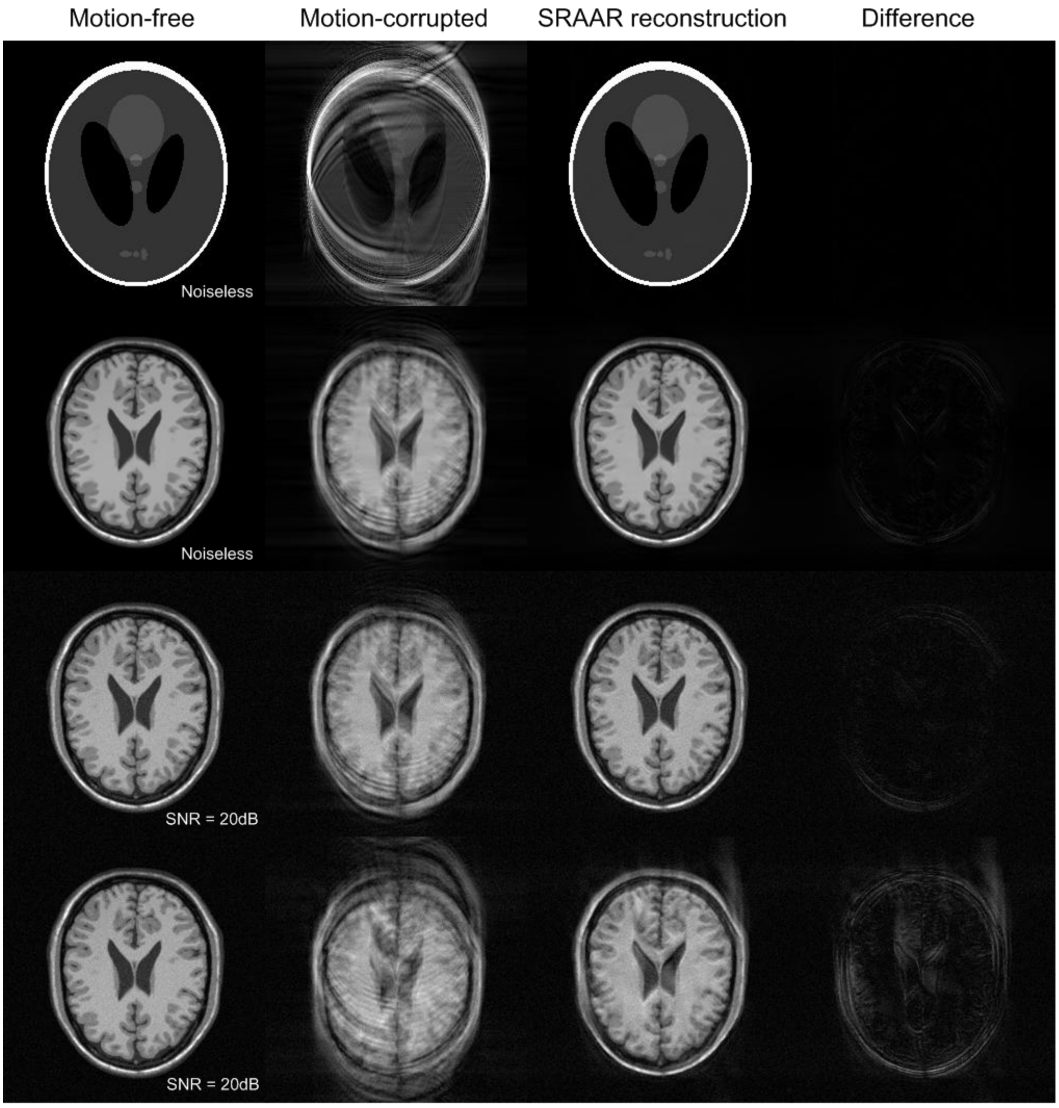


Figure 9.1: Simulation results of sparsity-based motion correction on a Shepp-Logan phantom (row 1, 100 iterations), simulated human brain (row 2, 200 iterations), and noisy brain with small motions (row 3, 200 iterations) and large motions (row 4, 400 iterations).

Chapter 10

Conclusion and Future Work

10.1 Conclusion

As an emerging technique which enables sub-Nyquist sampling and processing, CS has attracted attentions of many researchers in the past few years. In this thesis, we have studied several problems of CS ranging from theoretical analysis and algorithm design to practical applications. The contributions of the thesis can be summarized in the following aspects.

- We have discovered a new phase transition curve for the complex CS problem. The complex phase transition curve determines precisely the sparsity-undersampling tradeoff of ℓ_1 minimization in the complex setting and is positioned well above the known one in the real domain.

- We have analyzed the CS problem subject to structured matrix perturbations which has practical relevance to the DOA estimation problem. Sufficient conditions in terms of RIP were provided such that the formulated ℓ_1 minimization problems give accurate or even exact signal recoveries.

- We have proposed efficient algorithms for the sparse signal recovery in CS, including ONE-L1 algorithms for the convex relaxation method and novel

SBL algorithms based on a new sparsity-inducing prior. Their performances were demonstrated for 1D synthetic data and 2D images.

- We have presented a unified Bayesian framework and algorithm for CS problems with multiple-bit and 1-bit quantized measurements. The novel framework and algorithm can be applied to various scenarios including noiseless/noisy environment and unsaturated/saturated quantizer.
- We have proposed a fast and accurate approach to the DOA estimation by applying CS. The new approach resolves the problem that the true DOAs do not lie on the discretized sampling grid and can maintain high estimation accuracy under a coarse sampling grid.
- We have applied CS to MRI and presented the first sparsity-based approach to rigid-body motion correction in MRI. The new method solves the problem by exploiting the image sparsity in the wavelet domain and does not need to acquire additional k-space data.

10.2 Future Work

We list some possible directions for future studies.

- **Analysis of noisy 1-bit CS:** A convex formulation of noisy 1-bit CS problem has been studied in [68] with guaranteed signal recovery accuracy proven. The formulation requires the knowledge of the signal's ℓ_1 norm which cannot be known *a priori* in practice. In Chapter 7 we have introduced a different convex formulation (problem (7.16) with $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ and $s = 1$) as follows:

$$\min_{\boldsymbol{x},\boldsymbol{y}} \|\boldsymbol{x}\|_1, \text{ subject to } \begin{cases} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon, \\ \operatorname{diag}(\boldsymbol{b})\, \boldsymbol{y} \geq 0, \\ \boldsymbol{b}^T \boldsymbol{y} = 1, \end{cases} \tag{10.1}$$

where $\boldsymbol{x}$ is the signal of interest, ϵ refers to the noise level and $\boldsymbol{b}$ denotes the vector of sign measurements. This formulation uses the noise information rather than the signal sparsity and can be of independent interest. One future work is to analyze its theoretical guarantees.

- **Efficient algorithm for quantized CS:** One drawback of the Q-VMP algorithm in Chapter 7 for quantized CS is its high computational complexity due to an inversion of a high dimensional matrix at each iteration though it has been greatly alleviated with a basis pruning approach adopted in that chapter. So another future work is to develop fast alternatives to the current implementation.

- **Algorithm for motion correction:** In Chapter 9 we have presented the SRAAR algorithm for the motion correction in MRI. A shortcoming of the algorithm is the need of specifying a constant which determines the *sparse* domain and is typically unavailable in practice. While the process of parameter tuning can be time-consuming, it is of great interest to design efficient algorithms which do not need the constant. A possibly useful result is presented in our recent work [145].

- **Continuous DOA estimation:** Most of the existing CS-based methods for the DOA estimation are based on the standard CS formulation which can be considered as a zeroth-order approximation of the true observation model via discretizing the continuous DOA range. Our methods in Chapter 8 rely on a first-order approximation via discretization and linear approximation. One future work is to do DOA estimation directly in the continuous range without discretization. Recently, there have been some results in the case of a single snapshot (a.k.a. line spectral estimation), where one needs to estimate the frequencies of some sinusoids from their superposition [146–148]. The so-called atomic norm [149] is used in existing results which is a generalization of the ℓ_1 norm to the continuous setting. After submission of this thesis, we have developed the first discretization-free sparse method in [150], named as the sparse and parametric approach (SPA), for continuous DOA estimation in the presence of multiple snapshots, which can be applied to uniform/sparse linear arrays. While SPA is a statistical inference method, its connection to atomic norm-based methods has also be derived (see [151] and future publications).

- **Compressive phase retrieval:** Phase retrieval [143] refers to a classical problem of recovering a signal from its Fourier magnitude measurements. This problem arises in many imaging techniques where the phase information of the frequency data are not available to the detecting and sensing devices, e.g., X-ray crystallography [143]. It is also related to the motion correction in MRI that we have studied in Chapter 9 where the phase information, though not lost, is severely distorted. Conventional approaches to phase retrieval do not consider the image structure and solve the problem by oversampling in the frequency domain. Inspired by the success of CS which exploits the signal sparsity to reduce the number of measurements, sparsity-based phase retrieval has been studied recently, known as compressive phase retrieval (CPR) [152]. Though the CPR problem exhibits many similarities to CS the key difference is the nonlinearity of the measurements in CPR. One of the future directions is to design efficient algorithms and analyze their performance for CPR. Our preliminary results are presented in [145, 153].

Appendix A

Proofs of Results in Chapter 5

A.1 Proof of Theorem 5.3

Denote $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\beta} \odot \boldsymbol{x} \end{bmatrix}$ and similarly define $\boldsymbol{z}^o$ and $\boldsymbol{z}^*$. Then the problem in (5.10) can be rewritten into

$$\min_{\boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{\beta} \in [-r,r]^N} \left\| \boldsymbol{x} \right\|_0, \text{ subject to } \boldsymbol{y} = \boldsymbol{\Psi} \boldsymbol{z}. \tag{A.1}$$

Let $\bar{\delta}_K = \bar{\delta}_K \left(\boldsymbol{\Psi} \right)$ hereafter for brevity.

First note that $\boldsymbol{x}^o$ is K-sparse and $\boldsymbol{z}^o$ is $2K$-D-sparse. Since $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ is a solution to the problem in (A.1), we have $\left\| \boldsymbol{x}^* \right\|_0 \leq \left\| \boldsymbol{x}^o \right\|_0 \leq K$ and, hence, $\boldsymbol{z}^*$ is $2K$-D-sparse. By $\boldsymbol{y} = \boldsymbol{\Psi} \boldsymbol{z}^o = \boldsymbol{\Psi} \boldsymbol{z}^*$ we obtain $\boldsymbol{\Psi} \left(\boldsymbol{z}^o - \boldsymbol{z}^* \right) = \boldsymbol{0}$ and thus $\boldsymbol{z}^o - \boldsymbol{z}^* = \boldsymbol{0}$ by $\bar{\delta}_{4K} < 1$ and the fact that $\boldsymbol{z}^o - \boldsymbol{z}^*$ is $4K$-D-sparse. We complete the proof by observing that $\boldsymbol{z}^o - \boldsymbol{z}^* = \begin{bmatrix} \boldsymbol{x}^o - \boldsymbol{x}^* \\ \boldsymbol{\beta}^o \odot \boldsymbol{x}^o - \boldsymbol{\beta}^* \odot \boldsymbol{x}^* \end{bmatrix} = \boldsymbol{0}$.

A.2 Proof of Theorems 5.4 and 5.5

We only present the proof of Theorem 5.5 since Theorem 5.4 is a special case of Theorem 5.5. We first show the following lemma.

Lemma A.1. *We have*

$$|\langle \boldsymbol{\Psi v}, \boldsymbol{\Psi v}' \rangle| \leq \bar{\delta}_{2(K+K')} \left\|\boldsymbol{v}\right\|_2 \left\|\boldsymbol{v}'\right\|_2$$

for all $2K$*-D-sparse* $\boldsymbol{v}$ *and* $2K'$*-D-sparse* $\boldsymbol{v}'$ *supported on disjoint subsets.*

Proof. Without loss of generality, assume that $\boldsymbol{v}$ and $\boldsymbol{v}'$ are unit vectors with disjoint supports as above. Then by the definition of D-RIP and $\left\|\boldsymbol{v} \pm \boldsymbol{v}'\right\|_2^2 = \left\|\boldsymbol{v}\right\|_2^2 + \left\|\boldsymbol{v}'\right\|_2^2 = 2$ we have $2\left(1 - \bar{\delta}_{2(K+K')}\right) \leq \left\|\boldsymbol{\Psi v} \pm \boldsymbol{\Psi v}'\right\|_2^2 \leq 2\left(1 + \bar{\delta}_{2(K+K')}\right)$. Consequently, $|\langle \boldsymbol{\Psi v}, \boldsymbol{\Psi v}' \rangle| \leq \frac{1}{4}\left|\left\|\boldsymbol{\Psi v} + \boldsymbol{\Psi v}'\right\|_2^2 - \left\|\boldsymbol{\Psi v} - \boldsymbol{\Psi v}'\right\|_2^2\right| \leq \bar{\delta}_{2(K+K')}$, which completes the proof. ∎

Using the notations $\boldsymbol{z}$, $\boldsymbol{z}^o$, $\boldsymbol{z}^*$ and $\bar{\delta}_K$ in Appendix A.1, P-BPDN in (5.11) can be rewritten into

$$\min_{\boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{\beta} \in [-r,r]^N} \left\|\boldsymbol{x}\right\|_1, \text{ subject to } \left\|\boldsymbol{y} - \boldsymbol{\Psi z}\right\|_2 \leq \epsilon. \tag{A.2}$$

Let $\boldsymbol{h} = \boldsymbol{x}^* - \boldsymbol{x}^o$ and decompose $\boldsymbol{h}$ into a sum of K-sparse vectors $\boldsymbol{h}_{T_0}, \boldsymbol{h}_{T_1}, \boldsymbol{h}_{T_2}, \cdots$, where T_0 denotes the set of indices of the K largest entries (in absolute value) of $\boldsymbol{x}^o$, T_1 the set of the K largest entries of $\boldsymbol{h}_{T_0^c}$ with T_0^c being the complementary set of T_0, T_2 the set of the next K largest entries of $\boldsymbol{h}_{T_0^c}$ and so on. We abuse notations $\boldsymbol{z}_{T_j}^* = \begin{bmatrix} \boldsymbol{x}_{T_j}^* \\ \boldsymbol{\beta}_{T_j}^* \odot \boldsymbol{x}_{T_j}^* \end{bmatrix}$, $j = 0, 1, 2, \cdots$, and similarly define $\boldsymbol{z}_{T_j}^o$. Let $\boldsymbol{f} = \boldsymbol{z}^* - \boldsymbol{z}^o$ and $\boldsymbol{f}_{T_j} = \boldsymbol{z}_{T_j}^* - \boldsymbol{z}_{T_j}^o$ for $j = 0, 1, 2, \cdots$. For brevity we write $T_{01} = T_0 \cup T_1$. To bound $\left\|\boldsymbol{h}\right\|_2$, in the first step we show that $\left\|\boldsymbol{h}_{T_{01}^c}\right\|_2$ is essentially bounded by $\left\|\boldsymbol{h}_{T_{01}}\right\|_2$, and then in the second step we show that $\left\|\boldsymbol{h}_{T_{01}}\right\|_2$ is sufficiently small.

The first step follows directly from the proof of Theorem 1.3 in [17]. So we have

$$\left\|\boldsymbol{h}_{T_{01}^c}\right\|_2 \leq \left\|\boldsymbol{h}_{T_0}\right\|_2 + 2K^{-1/2} e_0 \tag{A.3}$$

with $e_0 \equiv \left\|\boldsymbol{x}^o - \boldsymbol{x}^K\right\|_1$.

In the second step, we bound $\|\boldsymbol{h}_{T_{01}}\|_2$ by utilizing its relationship with $\|\boldsymbol{f}_{T_{01}}\|_2$. Note that $\boldsymbol{f}_{T_j}$ for each $j = 0, 1, \cdots$ is $2K$-D-sparse. By $\boldsymbol{\Psi f}_{T_{01}} = \boldsymbol{\Psi f} - \sum_{j\geq 2} \boldsymbol{\Psi f}_{T_j}$ we have

$$\begin{aligned}
\|\boldsymbol{\Psi f}_{T_{01}}\|_2^2 &= \langle \boldsymbol{\Psi f}_{T_{01}}, \boldsymbol{\Psi f}\rangle - \sum_{j\geq 2} \left\langle \boldsymbol{\Psi f}_{T_{01}}, \boldsymbol{\Psi f}_{T_j}\right\rangle \\
&\leq \|\boldsymbol{\Psi f}_{T_{01}}\|_2 \cdot \|\boldsymbol{\Psi f}\|_2 + \bar{\delta}_{4K} \|\boldsymbol{f}_{T_0}\|_2 \sum_{j\geq 2} \left\|\boldsymbol{f}_{T_j}\right\|_2 \\
&\quad + \bar{\delta}_{4K} \|\boldsymbol{f}_{T_1}\|_2 \sum_{j\geq 2} \left\|\boldsymbol{f}_{T_j}\right\|_2 \qquad \text{(A.4)} \\
&\leq \|\boldsymbol{f}_{T_{01}}\|_2 \left(2\epsilon\sqrt{1+\bar{\delta}_{4K}} + \sqrt{2}\bar{\delta}_{4K} \sum_{j\geq 2} \left\|\boldsymbol{f}_{T_j}\right\|_2 \right). \qquad \text{(A.5)}
\end{aligned}$$

We used Lemma A.1 in (A.4). In (A.5), we used the D-RIP, and inequalities $\|\boldsymbol{f}_{T_0}\|_2 + \|\boldsymbol{f}_{T_1}\|_2 \leq \sqrt{2}\,\|\boldsymbol{f}_{T_{01}}\|_2$ and

$$\|\boldsymbol{\Psi f}\|_2 \leq \|\boldsymbol{y} - \boldsymbol{\Psi z}^*\|_2 + \|\boldsymbol{y} - \boldsymbol{\Psi z}^o\|_2 \leq 2\epsilon. \qquad \text{(A.6)}$$

By noting that $\boldsymbol{\beta}^o, \boldsymbol{\beta}^* \in [-r, r]^N$ and

$$\boldsymbol{f} = \begin{bmatrix} \boldsymbol{h} \\ \boldsymbol{\beta}^* \odot \boldsymbol{h} + (\boldsymbol{\beta}^* - \boldsymbol{\beta}^o) \odot \boldsymbol{x}^o \end{bmatrix} \qquad \text{(A.7)}$$

we have

$$\left\|\boldsymbol{f}_{T_j}\right\|_2 \leq \sqrt{1+r^2}\left\|\boldsymbol{h}_{T_j}\right\|_2 + 2r\left\|\boldsymbol{x}^o_{T_j}\right\|_2, \quad j = 0, 1, \cdots. \qquad \text{(A.8)}$$

Meanwhile,

$$\sum_{j\geq 2} \left\|\boldsymbol{x}^o_{T_j}\right\|_2 \leq \sum_{j\geq 2} \left\|\boldsymbol{x}^o_{T_j}\right\|_1 = \left\|\boldsymbol{x}^o_{T_{01}^c}\right\|_1 \leq e_0. \qquad \text{(A.9)}$$

Applying the D-RIP, (A.5), (A.8) and then (A.3) and (A.9) it gives

$$\begin{aligned}
&\left(1 - \bar{\delta}_{4K}\right) \|\boldsymbol{f}_{T_{01}}\|_2^2 \leq \|\boldsymbol{\Psi f}_{T_{01}}\|_2^2 \\
\leq &\|\boldsymbol{f}_{T_{01}}\|_2 \Big\{ 2\sqrt{1+\bar{\delta}_{4K}}\epsilon + \sqrt{2\left(1+r^2\right)}\bar{\delta}_{4K} \|\boldsymbol{h}_{T_0}\|_2 \\
&\qquad + 2\sqrt{2}\bar{\delta}_{4K}\left[\sqrt{1+r^2}K^{-1/2} + r\right] e_0 \Big\},
\end{aligned}$$

and thus $\left\|\boldsymbol{h}_{T_{01}}\right\|_2 \leq \left\|\boldsymbol{f}_{T_{01}}\right\|_2 \leq c_1\epsilon + c_0\left\|\boldsymbol{h}_{T_{01}}\right\|_2 + \left(c_2K^{-1/2} + c_3\right)e_0$ with $c_0 \equiv \frac{\sqrt{2(1+r^2)\bar{\delta}_{4K}}}{1-\bar{\delta}_{4K}}$, $c_1 \equiv \frac{2\sqrt{1+\bar{\delta}_{4K}}}{1-\bar{\delta}_{4K}}$, $c_2 \equiv 2c_0$ and $c_3 \equiv \frac{2\sqrt{2}\bar{\delta}_{4K}r}{1-\bar{\delta}_{4K}}$. Hence, we get a bound $\left\|\boldsymbol{h}_{T_{01}}\right\|_2 \leq (1-c_0)^{-1}\left[c_1\epsilon + \left(c_2K^{-1/2} + c_3\right)e_0\right]$, which together with (A.3) gives

$$\begin{aligned}\left\|\boldsymbol{h}\right\|_2 &\leq \left\|\boldsymbol{h}_{T_{01}}\right\|_2 + \left\|\boldsymbol{h}_{T_{01}^c}\right\|_2 \leq 2\left\|\boldsymbol{h}_{T_{01}}\right\|_2 + 2K^{-1/2}e_0 \\ &\leq \frac{2c_1}{1-c_0}\epsilon + \left[\left(\frac{2c_2}{1-c_0}+2\right)K^{-1/2} + \frac{2c_3}{1-c_0}\right]e_0,\end{aligned} \tag{A.10}$$

which concludes (5.14).

By (A.6), (A.7), (A.10) and the RIP we have

$$\begin{aligned}&\left(1-\bar{\delta}_{4K}\right)^{1/2}\left\|\left(\boldsymbol{\beta}_{T_0}^{*} - \boldsymbol{\beta}_{T_0}^{o}\right)\odot\boldsymbol{x}_{T_0}^{o}\right\|_2 \\ \leq&\left\|\boldsymbol{\Psi}\begin{bmatrix}\mathbf{0}\\ \left(\boldsymbol{\beta}_{T_0}^{*} - \boldsymbol{\beta}_{T_0}^{o}\right)\odot\boldsymbol{x}_{T_0}^{o}\end{bmatrix}\right\|_2 \\ =&\left\|\boldsymbol{\Psi}\left(\boldsymbol{f} - \begin{bmatrix}\boldsymbol{h}\\ \boldsymbol{\beta}^{*}\odot\boldsymbol{h}\end{bmatrix} - \begin{bmatrix}\mathbf{0}\\ \left(\boldsymbol{\beta}_{T_0^c}^{*} - \boldsymbol{\beta}_{T_0^c}^{o}\right)\odot\boldsymbol{x}_{T_0^c}^{o}\end{bmatrix}\right)\right\|_2 \\ \leq&2\epsilon + \sqrt{1+r^2}\left\|\boldsymbol{\Psi}\right\|_2\left\|\boldsymbol{h}\right\|_2 + 2r\left\|\boldsymbol{\Phi}\right\|_2 e_0 \\ \leq&c_4\epsilon + \left(c_5K^{-1/2} + c_6\right)e_0\end{aligned}$$

with $c_4 \equiv 2 + \frac{2\sqrt{1+r^2}\left\|\boldsymbol{\Psi}\right\|_2 c_1}{1-c_0}$, $c_5 \equiv \sqrt{1+r^2}\left(\frac{2c_2}{1-c_0}+2\right)\left\|\boldsymbol{\Psi}\right\|_2$ and $c_6 \equiv \left(\frac{2\sqrt{1+r^2}c_3}{1-c_0} + 2r\right)$ $\left\|\boldsymbol{\Psi}\right\|_2$, and thus

$$\left\|\left(\boldsymbol{\beta}_{T_0}^{*} - \boldsymbol{\beta}_{T_0}^{o}\right)\odot\boldsymbol{x}_{T_0}^{o}\right\|_2 \leq \left[c_4\epsilon + \left(c_5K^{-1/2} + c_6\right)e_0\right]/\sqrt{1-\bar{\delta}_{4K}}, \tag{A.11}$$

which concludes (5.15). We complete the proof by noting that the above results make sense if $c_0 < 1$, i.e., $\bar{\delta}_{4K} < \left(\sqrt{2\left(1+r^2\right)}+1\right)^{-1}$.

A.3 Proof of Lemma 5.1

We first consider the case where $\boldsymbol{v}$ is an interior point of $\mathcal{D}^*$, i.e., it holds that $\left\|\boldsymbol{y} - \boldsymbol{\Phi}^*\boldsymbol{v}\right\|_2 = \epsilon_0 < \epsilon$. Let $\eta = \epsilon - \epsilon_0$. Construct a sequence $\left\{\boldsymbol{v}^{(j)}\right\}_{j=1}^{\infty}$ such that

$\left\|\boldsymbol{v}^{(j)}-\boldsymbol{v}\right\|_2 \leq 1/j$. It is obvious that $\boldsymbol{v}^{(j)} \rightarrow \boldsymbol{v}$. We next show that $\boldsymbol{v}^{(j)} \in \mathcal{D}^j$ as j is large enough. By $\boldsymbol{\Phi}^{(j)} \rightarrow \boldsymbol{\Phi}^*$, $\boldsymbol{v}^{(j)} \rightarrow \boldsymbol{v}$ and that the sequence $\left\{\boldsymbol{v}^{(j)}\right\}_{j=1}^{\infty}$ is bounded, there exists a positive integer j_0 such that, as $j \geq j_0$, $\left\|\boldsymbol{\Phi}^*-\boldsymbol{\Phi}^{(j)}\right\|_2\left\|\boldsymbol{v}^{(j)}\right\|_2 \leq \eta/2$ and $\left\|\boldsymbol{\Phi}^*\right\|_2\left\|\boldsymbol{v}-\boldsymbol{v}^{(j)}\right\|_2 \leq \eta/2$. Hence, as $j \geq j_0$,

$$\begin{aligned}
&\left\|\boldsymbol{y}-\boldsymbol{\Phi}^{(j)}\boldsymbol{v}^{(j)}\right\|_2\\
=&\left\|\left(\boldsymbol{y}-\boldsymbol{\Phi}^*\boldsymbol{v}\right)+\left(\boldsymbol{\Phi}^*-\boldsymbol{\Phi}^{(j)}\right)\boldsymbol{v}^{(j)}+\boldsymbol{\Phi}^*\left(\boldsymbol{v}-\boldsymbol{v}^{(j)}\right)\right\|_2\\
\leq&\epsilon_0+\eta/2+\eta/2=\epsilon,
\end{aligned}$$

from which we have $\boldsymbol{v}^{(j)} \in \mathcal{D}^j$ for $j \geq j_0$. By re-selecting arbitrary $\boldsymbol{v}^{(j)} \in \mathcal{D}^j$ for $j < j_0$ we obtain the conclusion.

For the other case where $\boldsymbol{v}$ is a boundary point of $\mathcal{D}^*$, there exists a sequence $\left\{\boldsymbol{v}_{(l)}\right\}_{l=1}^{\infty} \subset \mathcal{D}^*$ with all $\boldsymbol{v}_{(l)}$ being interior points of $\mathcal{D}^*$ such that $\boldsymbol{v}_{(l)} \rightarrow \boldsymbol{v}$, as $l \rightarrow +\infty$. According to the first part of the proof, for each $l = 1, 2, \cdots$, there exists a sequence $\left\{\boldsymbol{v}_{(l)}^{(j)}\right\}_{j=1}^{\infty}$ with $\boldsymbol{v}_{(l)}^{(j)} \in \mathcal{D}^j$, $j = 1, 2, \cdots$, such that $\boldsymbol{v}_{(l)}^{(j)} \rightarrow \boldsymbol{v}_{(l)}$, as $j \rightarrow +\infty$. The sequence $\left\{\boldsymbol{v}_{(j)}^{(j)}\right\}_{j=1}^{\infty}$ is what we expected since

$$\left\|\boldsymbol{v}_{(j)}^{(j)}-\boldsymbol{v}\right\|_2 \leq \left\|\boldsymbol{v}_{(j)}^{(j)}-\boldsymbol{v}_{(j)}\right\|_2+\left\|\boldsymbol{v}_{(j)}-\boldsymbol{v}\right\|_2 \rightarrow 0,$$

as $j \rightarrow +\infty$.

A.4 Proof of Theorem 5.7

We first show the existence of an accumulation point. It follows from the inequality

$$\left\|\boldsymbol{y}-\left(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{\Delta}^{(j)}\right)\boldsymbol{x}^{(j)}\right\|_2 \leq \left\|\boldsymbol{y}-\left(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{\Delta}^{(j-1)}\right)\boldsymbol{x}^{(j)}\right\|_2 \leq \epsilon \tag{A.12}$$

that $\boldsymbol{x}^{(j)}$ is a feasible solution to the problem in (5.19), and thus $\left\|\boldsymbol{x}^{(j+1)}\right\|_1 \leq \left\|\boldsymbol{x}^{(j)}\right\|_1$ for $j = 1, 2, \cdots$. Then we have $\left\|\boldsymbol{x}^{(j)}\right\|_1 \leq \left\|\boldsymbol{x}^{(1)}\right\|_1 \leq \left\|\boldsymbol{A}^{\dagger}\boldsymbol{y}\right\|_1$ for $j = 1, 2, \cdots$, since $\boldsymbol{A}^{\dagger}\boldsymbol{y}$ is a feasible solution to the problem in (5.19) at the first iteration with the superscript † denoting the pseudo-inverse operator. This together with $\boldsymbol{\beta}^{(j)} \in [-r,r]^N$, $j = 1, 2, \cdots$, leads to that the sequence $\left\{\left(\boldsymbol{x}^{(j)}, \boldsymbol{\beta}^{(j)}\right)\right\}_{j=1}^{\infty}$ is bounded. Thus, there exists an accumulation point $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ of $\left\{\left(\boldsymbol{x}^{(j)}, \boldsymbol{\beta}^{(j)}\right)\right\}_{j=1}^{\infty}$.

For the accumulation point $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ there exists a subsequence $\left\{\left(\boldsymbol{x}^{(j_l)}, \boldsymbol{\beta}^{(j_l)}\right)\right\}_{l=1}^{\infty}$ of $\left\{\left(\boldsymbol{x}^{(j)}, \boldsymbol{\beta}^{(j)}\right)\right\}_{j=1}^{\infty}$ such that $\left(\boldsymbol{x}^{(j_l)}, \boldsymbol{\beta}^{(j_l)}\right) \to (\boldsymbol{x}^*, \boldsymbol{\beta}^*)$, as $l \to +\infty$. By (5.20), we have, for all $\boldsymbol{\beta} \in [-r,r]^N$,

$$\left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^{(j_l)}\right)\boldsymbol{x}^{(j_l)}\right\|_2 \le \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}\right)\boldsymbol{x}^{(j_l)}\right\|_2, \tag{A.13}$$

at both sides of which by taking $l \to +\infty$, we have, for all $\boldsymbol{\beta} \in [-r,r]^N$,

$$\left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^*\right)\boldsymbol{x}^*\right\|_2 \le \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}\right)\boldsymbol{x}^*\right\|_2, \tag{A.14}$$

which concludes (5.22).

For (5.21), we first point out that $\left\|\boldsymbol{x}^{(j)}\right\|_1 \to \left\|\boldsymbol{x}^*\right\|_1$, as $j \to +\infty$, since $\left\{\left\|\boldsymbol{x}^{(j)}\right\|_1\right\}_{j=1}^{\infty}$ is decreasing and $\boldsymbol{x}^*$ is one of its accumulation points. As in Lemma 5.1, let $\mathcal{D}^j = \left\{\boldsymbol{x} : \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^{(j)}\right)\boldsymbol{x}\right\|_2 \le \epsilon\right\}$ and $\mathcal{D}^* = \left\{\boldsymbol{x} : \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^*\right)\boldsymbol{x}\right\|_2 \le \epsilon\right\}$. By $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^{(j_l)} \to \boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^*$, as $l \to +\infty$, and Lemma 5.1, for any $\boldsymbol{x} \in \mathcal{D}^*$ there exists a sequence $\left\{\boldsymbol{x}_{(l)}\right\}_{l=1}^{\infty}$ with $\boldsymbol{x}_{(l)} \in \mathcal{D}^{j_l}$, $l = 1, 2, \cdots$, such that $\boldsymbol{x}_{(l)} \to \boldsymbol{x}$, as $l \to +\infty$. By (5.19), we have, for $l = 1, 2, \cdots$, $\left\|\boldsymbol{x}^{(j_l+1)}\right\|_1 \le \left\|\boldsymbol{x}_{(l)}\right\|_1$, at both sides of which by taking $l \to +\infty$, we have $\left\|\boldsymbol{x}^*\right\|_1 \le \left\|\boldsymbol{x}\right\|_1$. Finally, (5.21) is concluded since the inequality holds for arbitrary $\boldsymbol{x} \in \mathcal{D}^*$.

A.5 Proof of Theorem 5.8

We need to show that an optimal solution $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ satisfies (5.21) and (5.22). It is obvious for (5.21). For (5.22), we discuss two cases based on Lemma 5.2. If $\left\|\boldsymbol{y}\right\|_2 \le \epsilon$, then $\boldsymbol{x}^* = \boldsymbol{0}$ and, hence, (5.22) holds for any $\boldsymbol{\beta}^* \in [-r,r]^N$. If $\left\|\boldsymbol{y}\right\|_2 > \epsilon$, $\left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^*\right)\boldsymbol{x}^*\right\|_2 = \epsilon$ holds by (5.21) and Lemma 5.2. Next we use contradiction to show that (5.22) holds in such case.

Suppose that (5.22) does not hold as $\left\|\boldsymbol{y}\right\|_2 > \epsilon$. That is, there exists $\boldsymbol{\beta'} \in [-r, r]^N$ such that

$$\left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}'\right)\boldsymbol{x}^*\right\|_2 < \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}^*\right)\boldsymbol{x}^*\right\|_2 = \epsilon \tag{A.15}$$

holds with $\boldsymbol{\Delta'} = \operatorname{diag}\left(\boldsymbol{\beta}'\right)$. Then by Lemma 5.2 we see that $\boldsymbol{x}^*$ is a feasible but not optimal solution to the problem

$$\min_{\boldsymbol{x}} \left\|\boldsymbol{x}\right\|_1, \text{ subject to } \left\|\boldsymbol{y} - \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\Delta}'\right)\boldsymbol{x}\right\|_2 \leq \epsilon. \tag{A.16}$$

Hence, $\left\|\boldsymbol{x}'\right\|_1 < \left\|\boldsymbol{x}^*\right\|_1$ holds for an optimal solution $\boldsymbol{x}'$ to the problem in (A.16). Meanwhile, $(\boldsymbol{x}', \boldsymbol{\beta}')$ is a feasible solution to the P-BPDN problem in (5.11). Thus $(\boldsymbol{x}^*, \boldsymbol{\beta}^*)$ is not an optimal solution to the P-BPDN problem in (5.11) by $\left\|\boldsymbol{x}'\right\|_1 < \left\|\boldsymbol{x}^*\right\|_1$, which leads to contradiction.

Author's Publications

Preprints:

1. Z. Yang and L. Xie, "On sparse parametric methods for line spectral estimation from complete and incomplete data," submitted to *IEEE Transactions on Signal Processing*, November 2013.

2. Z. Yang, L. Xie, and C. Zhang, "A discretization-free sparse and parametric approach for linear array signal processing," submitted to *IEEE Transactions on Signal Processing*, July 2013.

3. Z. Yang, L. Xie, and C. Zhang, "Bayesian compressed sensing with new sparsity-inducing prior," Available online at http://arxiv.org/abs/1208.6464, 2013.

4. Z. Yang, C. Zhang, and L. Xie, "Robust compressive phase retrieval via l1 minimization with application to image reconstruction," Available online at http://arxiv.org/abs/1302.0081, 2013.

Journal papers:

1. Z. Yang, L. Xie, and C. Zhang, "Variational Bayesian algorithm for quantized compressed sensing," *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2815–2824, 2013.

2. A. Maleki, L. Anitori, Z. Yang, and R. Baraniuk, “Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP),” *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4290–4308, 2013.

3. Z. Yang, L. Xie, and C. Zhang, “Off-grid direction of arrival estimation using sparse Bayesian inference,” *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 38–43, 2013.

4. Z. Yang, C. Zhang, and L. Xie, “Robustly stable signal recovery in compressed sensing with structured matrix perturbation,” *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4658–4671, 2012.

5. Z. Yang, C. Zhang, and L. Xie, “On phase transition of compressed sensing in the complex domain,” *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 47–50, 2012.

6. Z. Yang, C. Zhang, J. Deng, and W. Lu, “Orthonormal expansion ℓ_1-minimization algorithms for compressed sensing,” *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6285–6290, 2011.

Conference papers:

1. C. Zhang, Z. Yang, and L. Xie, “The l_1 measure of image reconstruction subject to motion corrupted k-space data,” in *10th IEEE International Conference on Control and Automation (ICCA)*, 2013, pp. 1133–1137.

2. Z. Yang, C. Zhang, and L. Xie, “Sparse MRI for motion correction,” in *10th IEEE International Symposium on Biomedical Imaging (ISBI)*, 2013, pp. 962–965.

3. Z. Yang, L. Xie, and C. Zhang, “Accurate signal recovery in quantized compressed sensing,” in *15th IEEE International Conference on Information Fusion (FUSION)*, 2012, pp. 2531–2536.

4. Z. Yang, C. Zhang, and L. Xie, "Stable signal recovery in compressed sensing with a structured matrix perturbation," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2012, pp. 2737–2740.

5. J. Deng, Z. Yang, C. Zhang, and W. Lu, "Orthonormal expansion ℓ_1-minimization for compressed sensing in MRI," in *18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2297–2300.

6. Z. Yang and C. Zhang, "Sparsity-undersampling tradeoff of compressed sensing in the complex domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 3668–3671.

Bibliography

[1] E. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

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

[3] Z. Yang, C. Zhang, J. Deng, and W. Lu, "Orthonormal expansion ℓ_1-minimization algorithms for compressed sensing," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6285–6290, 2011.

[4] Z. Yang, C. Zhang, and L. Xie, "On phase transition of compressed sensing in the complex domain," *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 47–50, 2012.

[5] Z. Yang and C. Zhang, "Sparsity-undersampling tradeoff of compressed sensing in the complex domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 3668–3671.

[6] A. Maleki, L. Anitori, Z. Yang, and R. Baraniuk, "Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP)," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4290–4308, 2013.

[7] Z. Yang, C. Zhang, and L. Xie, "Robustly stable signal recovery in compressed sensing with structured matrix perturbation," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4658–4671, 2012.

[8] ——, “Stable signal recovery in compressed sensing with a structured matrix perturbation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2737–2740.

[9] Z. Yang, L. Xie, and C. Zhang, “Variational Bayesian algorithm for quantized compressed sensing,” *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2815–2824, 2013.

[10] ——, “Accurate signal recovery in quantized compressed sensing,” in *15th IEEE International Conference on Information Fusion (FUSION)*, 2012, pp. 2531–2536.

[11] ——, “Off-grid direction of arrival estimation using sparse Bayesian inference,” *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 38–43, 2013.

[12] Z. Yang, C. Zhang, and L. Xie, “Sparse MRI for motion correction,” in *10th IEEE International Symposium on Biomedical Imaging (ISBI)*, 2013, pp. 962–965.

[13] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM journal on scientific computing*, vol. 20, no. 1, pp. 33–61, 1998.

[14] J. F. Claerbout and F. Muir, “Robust modeling with erratic data,” *Geophysics*, vol. 38, no. 5, pp. 826–844, 1973.

[15] D. L. Donoho and M. Elad, “Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[16] E. Candès, “Compressive sampling,” in *Proceedings of the International Congress of Mathematicians*, vol. 3, 2006, pp. 1433–1452.

[17] ——, “The restricted isometry property and its implications for compressed sensing,” *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[18] S. Foucart, "Sparse recovery algorithms: Sufficient conditions in terms of restricted isometry constants," in *Approximation Theory XIII: San Antonio 2010*. Springer, 2012, pp. 65–77.

[19] T. Cai, L. Wang, and G. Xu, "New bounds for restricted isometry constants," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4388–4394, 2010.

[20] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[21] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[22] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[23] T. Blumensath and M. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[24] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[25] M. A. Davenport and M. B. Wakin, "Analysis of orthogonal matching pursuit using the restricted isometry property," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4395–4401, 2010.

[26] Q. Mo and Y. Shen, "A remark on the restricted isometry property in orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3654–3656, 2012.

[27] T. Zhang, “Sparse recovery with orthogonal matching pursuit under RIP,” *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6215–6221, 2011.

[28] M. Maleki, “Approximate message passing algorithms for compressed sensing,” Ph.D. dissertation, 2011.

[29] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *Available at http://arxiv.org/abs/1011.3027*, 2010.

[30] 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.

[31] D. Donoho and J. Tanner, “Counting the faces of randomly-projected hypercubes and orthants, with applications,” *Discrete and Computational Geometry*, vol. 43, no. 3, pp. 522–541, 2010.

[32] M. Stojnic, “Various thresholds for ℓ_1-optimization in compressed sensing,” *Available online at http://arxiv.org/abs/0907.3666*, 2009.

[33] D. 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.

[34] D. Donoho and J. Tanner, “Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing,” *Philosophical Transactions of the Royal Society A*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[35] D. L. Donoho and J. Tanner, “Precise undersampling theorems,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 913–924, 2010.

[36] M. Lustig, D. Donoho, and J. Pauly, “Sparse MRI: The application of compressed sensing for rapid MR imaging,” *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[37] M. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 6, pp. 2275–2284, 2009.

[38] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[39] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[40] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[41] ——, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3704–3716, 2007.

[42] D. Wipf and S. Nagarajan, "A unified Bayesian framework for MEG/EEG source imaging," *Neuroimage*, vol. 44, no. 3, pp. 947–966, 2009.

[43] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 92–106, 2009.

[44] L. He and L. Carin, "Exploiting structure in wavelet-based Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 9, pp. 3488–3497, 2009.

[45] S. D. Babacan, L. Mancera, R. Molina, and A. K. Katsaggelos, "Non-convex priors in Bayesian compressed sensing," in *17th European Signal Processing Conference (EUSIPCO)*, 2009, pp. 110–114.

[46] S. Babacan, R. Molina, and A. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 53–63, 2010.

[47] D. Baron, S. Sarvotham, and R. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 269–280, 2010.

[48] D. Shutin and B. H. Fleury, "Sparse variational Bayesian SAGE algorithm with application to the estimation of multipath wireless channels," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3609–3623, 2011.

[49] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912–926, 2011.

[50] L. Yu, H. Sun, J.-P. Barbot, and G. Zheng, "Bayesian compressive sensing for cluster structured sparse signals," *Signal Processing*, vol. 92, no. 1, pp. 259–269, 2012.

[51] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "A novel hierarchical Bayesian approach for sparse semisupervised hyperspectral unmixing," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 585–599, 2012.

[52] H. Ishwaran and J. S. Rao, "Spike and slab variable selection: Frequentist and Bayesian strategies," *Annals of Statistics*, pp. 730–773, 2005.

[53] C. P. Robert and G. Casella, *Monte Carlo statistical methods.* Citeseer, 2004, vol. 319.

[54] N. L. Pedersen, D. Shutin, C. N. Manchón, and B. H. Fleury, "Sparse estimation using Bayesian hierarchical prior modeling for real and complex models," *Available at http://arxiv.org/pdf/1108.4324v1*, 2011.

[55] D. J. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[56] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, 2003.

[57] M. Tipping and A. Faul, “Fast marginal likelihood maximisation for sparse Bayesian models,” in *9th International Workshop on Artificial Intelligence and Statistics*, vol. 1, no. 3, 2003.

[58] E. Candès, J. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[59] L. Jacques, D. K. Hammond, and J. M. Fadili, “Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine,” *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 559–571, 2011.

[60] J. N. Laska, P. T. Boufounos, M. A. Davenport, and R. G. Baraniuk, “Democracy in action: Quantization, saturation, and compressive sensing,” *Applied and Computational Harmonic Analysis*, vol. 31, no. 3, pp. 429–443, 2011.

[61] A. Zymnis, S. Boyd, and E. Candes, “Compressed sensing with quantized measurements,” *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 149–152, 2010.

[62] P. T. Boufounos and R. G. Baraniuk, “1-bit compressive sensing,” in *42nd IEEE Annual Conference on Information Sciences and Systems (CISS)*, 2008, pp. 16–21.

[63] J. N. Laska and R. G. Baraniuk, “Regime change: Bit-depth versus measurement-rate in compressive sensing,” *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3496–3505, 2012.

[64] P. T. Boufounos, “Greedy sparse signal reconstruction from sign measurements,” in *IEEE Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, 2009, pp. 1305–1309.

[65] J. N. Laska, Z. Wen, W. Yin, and R. G. Baraniuk, “Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements,” *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5289–5301, 2011.

[66] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.

[67] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1275–1297, 2013.

[68] Y. PLAN and R. VERSHYNIN, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 482–494, 2013.

[69] T. Blumensath and M. Davies, "Compressed sensing and source separation," in *International Conference on Independent Component Analysis and Signal Separation*, 2007, pp. 341–348.

[70] T. Xu and W. Wang, "A compressed sensing approach for underdetermined blind audio source separation with sparse representation," in *15th IEEE Workshop on Statistical Signal Processing*, 2009, pp. 493–496.

[71] J. Zheng and M. Kaveh, "Directions-of-arrival estimation using a sparse spatial spectrum model with uncertainty," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2848–2551.

[72] L. Zhang, M. Xing, C. Qiu, J. Li, and Z. Bao, "Achieving higher resolution isar imaging with limited pulses via compressed sampling," *Geoscience and Remote Sensing Letters*, vol. 6, no. 3, pp. 567–571, 2009.

[73] S. Gleichman and Y. C. Eldar, "Blind compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6958–6975, 2011.

[74] M. Herman and T. Strohmer, "General deviants: An analysis of perturbations in compressed sensing," *IEEE J. Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 342–349, 2010.

[75] M. Herman and D. Needell, “Mixed operators in compressed sensing,” in *44th IEEE Annual Conference on Information Sciences and Systems (CISS)*, 2010, pp. 1–6.

[76] Y. Chi, L. Scharf, A. Pezeshki, and A. Calderbank, “Sensitivity to basis mismatch in compressed sensing,” *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2182–2195, 2011.

[77] D. Chae, P. Sadeghi, and R. Kennedy, “Effects of basis-mismatch in compressive sampling of continuous sinusoidal signals,” in *2nd IEEE International Conference on Future Computer and Communication (ICFCC)*, vol. 2, 2010, pp. 739–743.

[78] H. Zhu, G. Leus, and G. Giannakis, “Sparsity-cognizant total least-squares for perturbed compressive sampling,” *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2002–2016, 2011.

[79] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, “Single-pixel imaging via compressive sampling,” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, 2008.

[80] J. Provost and F. Lesage, “The application of compressed sensing for photoacoustic tomography,” *IEEE Transactions on Medical Imaging*, vol. 28, no. 4, pp. 585–594, 2009.

[81] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, “Compressed channel sensing: A new approach to estimating sparse multipath channels,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, 2010.

[82] M. Mishali and Y. C. Eldar, “Wideband spectrum sensing at sub-Nyquist rates [applications corner],” *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 102–135, 2011.

[83] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, “Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays,” *IEEE*

Journal of Selected Topics in Signal Processing, vol. 2, no. 3, pp. 275–285, 2008.

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

[85] R. R. Edelman and S. Warach, "Magnetic resonance imaging," *New England Journal of Medicine*, vol. 328, no. 10, pp. 708–716, 1993.

[86] R. Schmidt, "A signal subspace approach to multiple emitter location spectral estimation," Ph.D. dissertation, 1981.

[87] P. Stoica and N. Arye, "Music, maximum likelihood, and Cramer-Rao bound," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 5, pp. 720–741, 1989.

[88] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[89] M. Hyder and K. Mahata, "Direction-of-arrival estimation using a mixed $\ell_{2,0}$ norm approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4646–4655, 2010.

[90] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 72–82, 2008.

[91] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "k-t FOCUSS: A general compressed sensing framework for high resolution dynamic MRI," *Magnetic Resonance in Medicine*, vol. 61, no. 1, pp. 103–116, 2009.

[92] R. Otazo, D. Kim, L. Axel, and D. K. Sodickson, "Combination of compressed sensing and parallel imaging for highly accelerated first-pass cardiac perfusion MRI," *Magnetic Resonance in Medicine*, vol. 64, no. 3, pp. 767–776, 2010.

[93] J. P. Haldar, D. Hernando, and Z.-P. Liang, "Compressed-sensing MRI with random encoding," *IEEE Transactions on Medical Imaging*, vol. 30, no. 4, pp. 893–903, 2011.

[94] V. M. Patel, R. Maleh, A. C. Gilbert, and R. Chellappa, "Gradient-based image recovery methods from incomplete Fourier measurements," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 94–105, 2012.

[95] M. Medley, H. Yan, and D. Rosenfeld, "An improved algorithm for 2-D translational motion artifact correction," *IEEE Transactions on Medical Imaging*, vol. 10, no. 4, pp. 548–553, 1991.

[96] R. A. Zoroofi, Y. Sato, S. Tamura, H. Naito, and L. Tang, "An improved method for MRI artifact correction due to translational motion in the imaging plane," *IEEE Transactions on Medical Imaging*, vol. 14, no. 3, pp. 471–479, 1995.

[97] Y. M. Kadah, A. A. Abaza, A. S. Fahmy, A.-B. M. Youssef, K. Heberlein, and X. P. Hu, "Floating navigator echo (FNAV) for in-plane 2D translational motion estimation," *Magnetic resonance in medicine*, vol. 51, no. 2, pp. 403–407, 2004.

[98] W. Lin, F. Huang, P. Börnert, Y. Li, and A. Reykowski, "Motion correction using an enhanced floating navigator and GRAPPA operations," *Magnetic Resonance in Medicine*, vol. 63, no. 2, pp. 339–348, 2010.

[99] J. Mendes, E. Kholmovski, and D. L. Parker, "Rigid-body motion correction with self-navigation MRI," *Magnetic Resonance in Medicine*, vol. 61, no. 3, pp. 739–747, 2009.

[100] E. Candès and J. Romberg, "ℓ1-magic: Recovery of sparse signals via convex programming," *Avaiable online at http://users.ece.gatech.edu/ justin/l1magic/downloads/l1magic.pdf*, 2005.

[101] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *IEEE J. Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2008.

[102] E. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for ℓ_1-regularized minimization with applications to compressed sensing," *Available online at http://www.caam.rice.edu/ zhang/reports/tr0707.pdf*, 2007.

[103] S. Becker, J. Bobin, and E. Candes, "NESTA: A fast and accurate first-order method for sparse recovery," *SIAM J. Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011.

[104] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

[105] J. Nocedal and S. Wright, *Numerical Optimization.* New York: Springer Verlag, 2006.

[106] D. Bertsekas, *Constrained optimization and Lagrange multiplier methods.* Boston: Academic Press, 1982.

[107] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, "A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation," *SIAM Journal on Scientific Computing*, vol. 32, no. 4, pp. 1832–1857, 2010.

[108] 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.

[109] K. Bredies and D. Lorenz, "Linear convergence of iterative soft-thresholding," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 813–837, 2008.

[110] A. Maleki and D. L. Donoho, "Optimally tuned iterative reconstruction algorithms for compressed sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 330–341, 2010.

[111] M. Stojnic, “Block-length dependent thresholds in block-sparse compressed sensing,” *Available online at http://arxiv.org/abs/0907.3679*, 2009.

[112] M. Stojnic, F. Parvaresh, and B. Hassibi, “On the reconstruction of block-sparse signals with an optimal number of measurements,” *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 3075–3085, 2009.

[113] Y. Eldar and H. Rauhut, “Average case analysis of multichannel sparse recovery using convex relaxation,” *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519, 2010.

[114] S. Foucart and R. Gribonval, “Real versus complex null space properties for sparse vector recovery,” *Comptes Rendus Mathematique*, vol. 348, no. 15, pp. 863–865, 2010.

[115] 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.

[116] S. Foucart and M. Lai, “Sparsest solutions of underdetermined linear systems via ℓ_1-minimization for $0 < q < 1$,” *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 395–407, 2009.

[117] T. Cai, L. Wang, and G. Xu, “Shifting inequality and recovery of sparse signals,” *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1300–1308, 2010.

[118] D. Donoho, M. Elad, and V. Temlyakov, “Stable recovery of sparse overcomplete representations in the presence of noise,” *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[119] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, “Model-based compressive sensing,” *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[120] T. Blumensath and M. Davies, “Sampling theorems for signals from the union of finite-dimensional linear subspaces,” *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[121] R. Chartrand, “Exact reconstruction of sparse signals via nonconvex minimization,” *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.

[122] R. Chartrand and V. Staneva, “Restricted isometry properties and nonconvex compressive sensing,” *Inverse Problems*, vol. 24, p. 035020, 2008.

[123] R. Saab, R. Chartrand, and O. Yilmaz, “Stable sparse approximations via nonconvex optimization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 3885–3888.

[124] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming,” *Available online at http://cvxr.com/cvx*, 2008.

[125] E. J. Candes and Y. Plan, “A probabilistic and RIPless theory of compressed sensing,” *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7235–7254, 2011.

[126] G. McLachlan and T. Krishnan, *The EM algorithm and extensions.* John Wiley & Sons, 1997.

[127] R. S. Irving, *Integers, polynomials, and rings: A course in algebra.* Springer, 2004.

[128] 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.

[129] D. Donoho, Y. Tsaig, I. Drori, and J. Starck, “Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit,” *Available online at http://www.cs.tau.ac.il/∼idrori/StOMP.pdf*, 2006.

[130] 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.

[131] Y. Tsaig and D. L. Donoho, "Extensions of compressed sensing," *Signal Processing*, vol. 86, no. 3, pp. 549–571, 2006.

[132] W. Dai and O. Milenkovic, "Information theoretical and algorithmic approaches to quantized compressive sensing," *IEEE Transactions on Communications*, vol. 59, no. 7, pp. 1857–1866, 2011.

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

[134] T. Park and G. Casella, "The Bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[135] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[136] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, 2008.

[137] J. Winn and C. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, pp. 661–694, 2005.

[138] B. Jørgensen, *Statistical properties of the generalized inverse Gaussian distribution.* Springer New York, 1982, vol. 21.

[139] J. M. Kim, O. K. Lee, and J. C. Ye, "Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing," *IEEE Transactions on Information Theory*, vol. 58, no. 1, pp. 278–301, 2012.

[140] K. Lee, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3613–3641, 2012.

[141] N. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction)," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 152–177, 1963.

[142] S. Marchesini, "A unified evaluation of iterative projection algorithms for phase retrieval," *Review of Scientific Instruments*, vol. 78, no. 1, pp. 011 301–011 301, 2007.

[143] R. Millane, "Phase retrieval in crystallography and optics," *JOSA A*, vol. 7, no. 3, pp. 394–411, 1990.

[144] D. R. Luke, "Relaxed averaged alternating reflections for diffraction imaging," *Inverse Problems*, vol. 21, no. 1, p. 37, 2005.

[145] Z. Yang, C. Zhang, and L. Xie, "Robust compressive phase retrieval via l1 minimization with application to image reconstruction," *Available online at http://arxiv.org/abs/1302.0081*, 2013.

[146] B. N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Transactions on Signal Processing, DOI: 10.1109/TSP.2013.2273443*, 2013.

[147] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics, DOI: 10.1002/cpa.21455*, 2013.

[148] G. Tang, B. N. Bhaskar, and B. Recht, "Near minimax line spectral estimation," *Available online at http://arxiv.org/abs/1303.4348*, 2013.

[149] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[150] Z. Yang, L. Xie, and C. Zhang, "A discretization-free sparse and parametric approach for linear array signal processing," *Submitted to IEEE Transactions on Signal Processing*, July 2013.

[151] Z. Yang and L. Xie, "On sparse parametric methods for line spectral estimation from complete and incomplete data," *Submitted to IEEE Transactions on Signal Processing*, November 2013.

[152] M. Moravec, J. Romberg, and R. Baraniuk, "Compressive phase retrieval," in *Proceedings of SPIE, the International Society for Optical Engineering*, 2007, pp. 670 120–1.

[153] C. Zhang, Z. Yang, and L. Xie, "The l_1 measure of image reconstruction subject to motion corrupted k-space data," in *10th IEEE International Conference on Control and Automation (ICCA)*, 2013, pp. 1133–1137.